LLM Kernel: an evaluation framework for open-ended scientific interpretation

William Connell
Transcripta Bio
Palo Alto, CA
will@transcriptabio.com

Drishti Guin
Transcripta Bio
Palo Alto, CA
alex@transcriptabio.com

Clayton Mellina
Transcripta Bio
Palo Alto, CA
clayton@transcriptabio.com

Abstract

Large language models (LLMs) have demonstrated strong performance on structured tasks such as mathematics and scientific problem-solving, but their role in open-ended discovery science remains limited by the difficulty of validating their complex reasoning. Here we introduce **LLM Kernel**, a framework that makes an LLM's interpretation of data verifiable by prompting it to produce a quantitative similarity score directly linked to its qualitative reasoning trace. Applied to transcriptomics, an LLM kernel consistently outperforms standard numerical approaches in recovering known biological relationships, with performance improving as a function of compute. Ablation experiments show that performance depends on the model's biological knowledge of gene identities rather than mere approximation of statistical correlations. Furthermore, the framework's flexibility enables novel cross-modal comparisons: an LLM kernel can score the similarity between a natural language description of a disease and a numerical gene expression profile to identify relevant therapeutic compounds. LLM Kernel provides a scalable approach to quantitatively benchmark model reasoning, representing a step towards auditable AI for scientific interpretation.

1 Introduction

Modern science produces large datasets that capture system-level measurements across many disciplines, from physics and climate science to economics and biology. These datasets offer opportunities to investigate complex systems in their entirety, yet interpreting them remains a core challenge. Statistical methods are crucial for pattern identification but often reveal correlations without the mechanistic explanations necessary for testable hypotheses. The disconnect between data-driven pattern extraction and scientific understanding is a primary gap in discovery research.

Among the many fields facing this challenge, biology stands out. Genomics, transcriptomics, and proteomics generate high-dimensional data that record how living systems operate. Interpreting these data requires connecting them with the large, unstructured body of biological knowledge to turn patterns into mechanistic insight.

LLMs may be well-suited to assist in open-ended scientific discovery due to their training on broad human knowledge. While LLMs excel at solving math, coding, and information retrieval

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: NeurIPS 2025 AI for Science Workshop.

problems, their role in discovery research is less clear. A key challenge is validating their reasoning, or interpretations of data. As reasoning traces grow more complex, there is a need for scalable frameworks that assess the credibility of LLM outputs and connect them to empirical validation.

To address this, we introduce **LLM Kernel**, a framework for measuring the quality of reasoning traces. Our approach prompts an LLM to score the similarity between two data samples, producing an artifact with two key components: a qualitative interpretation of the pairwise comparison and a corresponding quantitative score that depends on the interpretation. The score can then be compared against intrinsic or extrinsic expectations of pairwise similarity to assess quality.

We apply the LLM kernel to transcriptomics, where interpreting high-dimensional features is a central challenge. Transcriptomic data is well-suited for this approach because its feature labels—gene identities—encode biological knowledge that LLMs can directly interpret. Unlike traditional kernels, the LLM kernel incorporates context from feature labels, enabling analyses that go beyond purely statistical similarity measures.

Our work builds on two lines of research: classical similarity metrics in statistics and emerging LLM-based approaches for scientific discovery. Traditional methods for comparing biological profiles use statistical measures like Pearson and Spearman correlation, cosine similarity, and other non-parametric approaches. In transcriptomics, a prominent example is the Connectivity Map, which applies a rank-based method to score similarity between gene expression signatures[1,2]. These similarity analyses support discovery by clustering samples for functional annotation and by ranking complementary profiles for drug repurposing[3]. However, such analyses are agnostic to biological mechanisms—they quantify statistical relationships without providing a biological rationale.

Concurrently, the use of LLMs to automate scientific workflows is rapidly growing. In this emerging field, LLMs act as reasoning engines for tasks ranging from literature synthesis and experimental design to database mining and computational analysis[4–8]. Typically, multi-agent systems embed LLMs in environments with access to databases and computational tools, enabling end-to-end automated analysis.

While automation promises to accelerate research, few systems have been rigorously benchmarked for the validity of their knowledge-based interpretations of biological data. Yet this is central to discovery science, where progress depends on generating explanations through integrated data interpretation[9–11]. A core challenge is that reasoning outputs are qualitative, making systematic validation difficult. Assessing the correctness and discovery value of a generated explanation can be as complex as the original scientific question. The LLM Kernel framework leverages pairwise relationships—often obtainable without full labels—to enable scalable validation of model reasoning. Each comparison yields a reasoning trace coupled to a verifiable score, providing a foundation for quantitative evaluation and novel similarity analyses.

2 LLM Kernel framework

We consider the general case of using an LLM to produce a scientific interpretation of input data. This data can take various forms depending on the application, including raw measurement data (e.g., transcriptomics or other omics) and abstracted text descriptions of natural processes (e.g., mechanistic descriptions of disease biology). Typical usage of an LLM may provide a data sample, d_i, along with an appropriate prompt, p, and return a scientific interpretation, g_i, of the data sample:

$$g_i = \mathrm{LLM}(d_i, p)$$

However, directly evaluating such interpretations remains challenging. A common approach is to score LLM responses on multiple-choice scientific benchmarks such as ScienceQA, or to use LLMs themselves to rate the quality of free-text outputs[12–14]. While useful for quantifying the extent of prior knowledge learned by the LLM, these approaches only weakly convey credibility on novel scientific interpretation tasks. In contrast, human evaluation relies on experts to assess LLM outputs, scoring responses along multiple axes such as novelty and discovery value. These scores may inform real-world applications where LLM-generated interpretations help guide experimental decisions. However, human evaluation scales poorly and may fail to consistently recognize truly novel insights.

LLM Kernel instead casts scientific interpretation as a comparative task in which an LLM is prompted to produce scientific interpretations along with a score, s, of the concordance between a pair of data samples:

$$g_{ij},\ s_{ij} = \mathrm{LLM}(d_i, d_j, p)$$

Such a pairwise score can be quantitatively compared to other information about the sample pair—including exogenous labels or statistical similarity metrics—providing a scalable, indirect evaluation of the LLM's interpretation that amounts to scoring consistency with known information. For instance, when comparing drug response data, LLM Kernel can quantify the extent to which the LLM's interpretation concords with known shared drug mechanisms.

Unlike classification, which restricts interpretation to a fixed set of predefined categories, similarity scoring allows explanations to emerge directly from the data. This design frees the model to explore patterns without bias from coarse labels (e.g., pathways or mechanisms), enabling novel associations.

As a result, the framework can quantify LLM performance without stipulating the specific content of an interpretation or constraining it to prior knowledge. This design also enables the direct, quantitative comparison of different LLM interpretation strategies. Finally, the framework is highly data-efficient, as the pairwise structure quadratically increases the effective number of available samples.

3 Results

3.1 LLM Kernel improves upon numerical similarity scores as a function of compute

We demonstrate that LLMs can reliably generate quantitative similarity scores between biological profiles. In our experimental setup, we prompt an LLM to score the similarity [-1, 1] of chemically perturbed gene expression profiles given a list of gene identities, differential expression log fold-changes (LFCs), and p-values (Fig. 1, Methods). These scores exhibit high self-consistency and a low frequency of failures (Fig. 1a, Fig. S1). Importantly, LLM-derived similarity scores align with and improve upon traditional numerical baseline measures of similarity. Using a drug response dataset designed to assess batch effects of the DRUG-seq assay across 14 transcriptionally diverse chemical perturbations, we observed concordance between LLM kernel scores and both (1) global patterns of original data (e.g., LFCs) and (2) local expected similarity structure (e.g., compound treatment labels)[15]. This concordance was quantified by Pearson correlation between LLM kernel scores and numerical similarity scores, with increased test-time computational budget improving alignment (Table 1, Fig. 2a, Methods). LLM kernel scores improved clustering quality, as measured by sample compound treatment labels. This improvement was especially notable at higher computational budgets, surpassing the baseline set by Spearman/Pearson correlations on the LFC data (Table 1, Fig. 2b, Methods). Finally, LLM kernel scores improved local similarity structure, as measured by neighborhood recall (Area Under Recall@k, AURecall@k) of sample annotations. Again, we observed improvements with increased computational budget (Table 1, Fig. 2c, Methods).

3.2 LLM Kernel leverages prior biological knowledge for similarity scoring

As previously noted, we observed that LLM kernel scores improved clustering quality over Spearman/Pearson correlations as measured by true compound treatment labels, suggesting that the LLM leveraged biological prior knowledge in interpretation and scoring. To probe this further—and in particular to test whether the model was simply approximating conventional numerical methods—we conducted a series of ablation experiments. Our initial prompt explicitly requested reasoning about gene set and pathway overlap (*numerical-prior*) (Box S1, Methods). When we ablated gene identity, both Global Correlation and local expected similarity structures (Purity, ARI, AMI, AURecall@k) decreased (Table 2, Methods). This indicates that the model does not solely rely on internal approximations of common numerical methods, such as simple gene list overlap, but rather integrates unstructured biological knowledge into its scoring mechanism. Under the *numerical-prior* prompt, removal of gene identity information resulted in a 9.01% performance drop in Global Correlation, suggesting that this component of the LLM kernel score is attributable to biological knowledge captured by the model (Table 2).

Further probing revealed that the model performs a directional comparison of LFC magnitudes rather than a simple overlap analysis of gene feature labels. We ablated directional information of the LFCs,

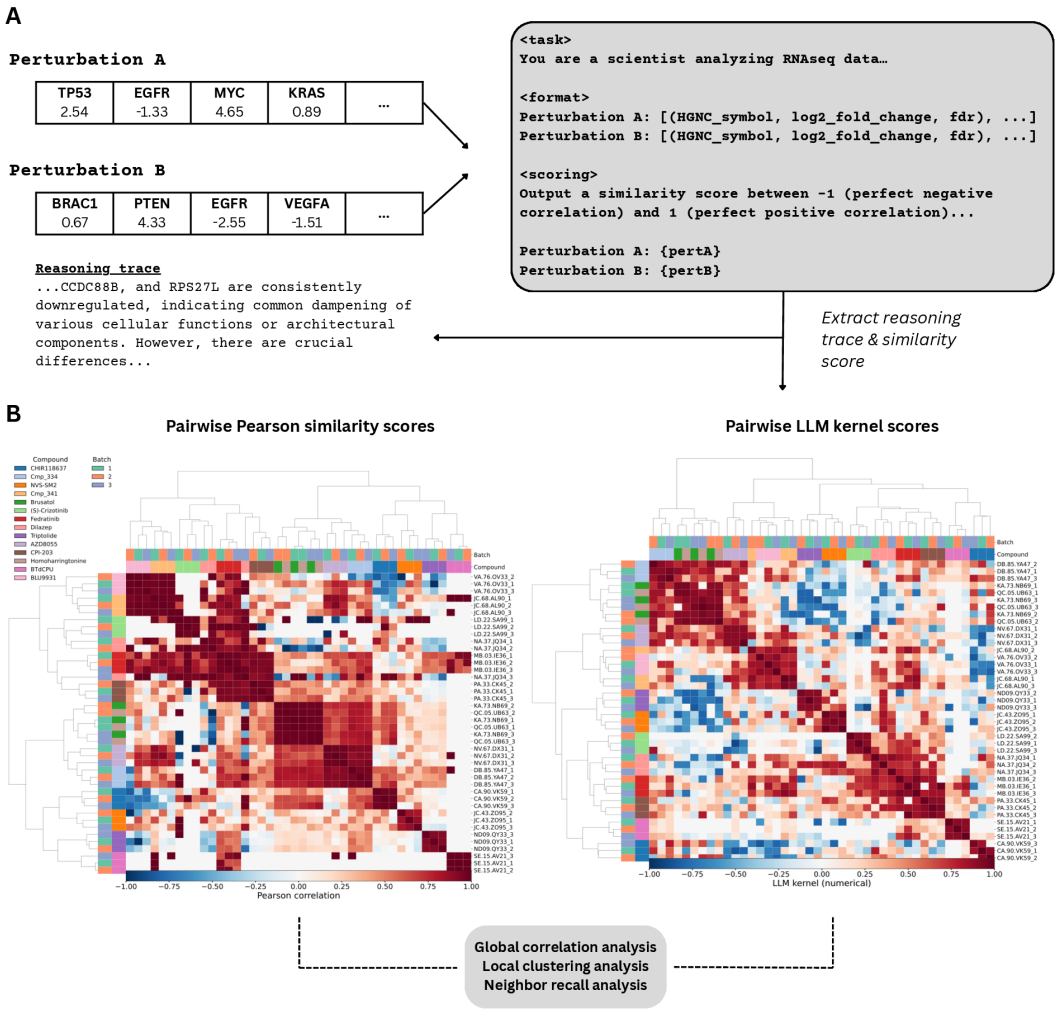

Figure 1: **Overview of LLM Kernel and analysis methods. (A)** Transcriptomes from perturbations A and B are injected into a prompt requesting a quantitative similarity comparison of their features. The reasoning trace is extracted, and the resulting similarity score is incorporated into a pairwise similarity matrix. **(B)** Heatmaps of the LFC similarity matrix (left) and the LLM kernel similarity matrix (right), with samples annotated by batch identity and compound treatment. Global and local analyses are then applied to compare these matrices against intrinsic and extrinsic notions of similarity.

and found performance fell further, providing evidence that the method employs directional logic crucial for discerning compensatory or additive biological effects (Table 2, Methods).

To test whether the model relied on superficial heuristics—such as noticing that some genes have large effect sizes, without knowing which genes they are—we shuffled the association between gene identity and numerical values. This preserved the overall distribution of values but broke the link between each gene and its true effect. The model's performance completely collapsed after shuffling, indicating that it relies on correctly associating specific genes with their statistical importance (Table 2, Methods).

3.3 LLM Kernel adapts to different notions of similarity based on natural language priors

The LLM kernel can reflect diverse notions of similarity by incorporating different prompts that induce varying priors. We modified the original, *numerical-prior* prompt to one that focused on interpreting functional and causal information about genes and biological processes (*functional-prior*)

Table 1: **LLM Kernel performance.** Global Correlation measures the Pearson correlation between each method and the Pearson LFC similarity matrix. For Purity, ARI, AMI and AURecall@k, a one-sample t-test tests whether the LLM kernel (numerical) method differs significantly from baseline methods (Pearson and Spearman). All LLM kernel results are computed under a fixed budget of 6144, and metrics are reported as the mean $\pm$ SEM (n=5). **Bold** indicates the best performing method. Significance levels: * $10^{-2} < p \leq 5 \times 10^{-2}$; ** $10^{-3} < p \leq 10^{-2}$; *** $10^{-4} < p \leq 10^{-3}$; **** $p \leq 10^{-4}$.

Method	Global Corr.	Purity	ARI	AMI	AURecall@k
LLM kernel (numerical)	$0.61 \pm 0.01^{***}$	$\mathbf{0.86 \pm 0.02}$	$\mathbf{0.75 \pm 0.03}$	$\mathbf{0.83 \pm 0.02}$	$\mathbf{0.96 \pm 0.00}$
Pearson	**1.00**	0.81^{*}	0.69	0.81	0.93^{***}
Spearman	0.93^{****}	0.81^{*}	0.65^{*}	0.80	0.92^{***}

Table 2: **LLM Kernel ablation results.** Global Correlation measures the Pearson correlation between each ablation method and the Pearson LFC similarity matrix. Significance for the LLM kernel (numerical) reflects LFC correlation significance; ablation samples significance reflects a two-sample t-test vs. the LLM kernel (numerical). For Purity, ARI, AMI and AURecall@k, an independent two-sample t-test tests whether ablation methods differ significantly from the LLM kernel (numerical) method. All LLM kernel results are computed under a fixed budget of 6144, and metrics are reported as the mean $\pm$ SEM (n=5). **Bold** indicates the best performing method. Significance levels: * $10^{-2} < p \leq 5 \times 10^{-2}$; ** $10^{-3} < p \leq 10^{-2}$; *** $10^{-4} < p \leq 10^{-3}$; **** $p \leq 10^{-4}$.

Method	Global Corr.	Purity	ARI	AMI	AURecall@k
LLM kernel (numerical)	$\mathbf{0.61 \pm 0.01}^{***}$	$\mathbf{0.86 \pm 0.02}$	$\mathbf{0.75 \pm 0.03}$	$\mathbf{0.83 \pm 0.02}$	$\mathbf{0.96 \pm 0.00}$
Gene ID ablation	$0.56 \pm 0.01^{***}$	$0.79 \pm 0.01^{**}$	$0.64 \pm 0.03^{*}$	$0.73 \pm 0.03^{*}$	$0.94 \pm 0.00^{***}$
Direction ablation	$0.42 \pm 0.01^{***}$	$0.71 \pm 0.01^{***}$	$0.43 \pm 0.02^{***}$	$0.55 \pm 0.03^{***}$	$0.89 \pm 0.00^{***}$
Gene shuffle ablation	$0.00 \pm 0.01^{***}$	$0.66 \pm 0.03^{***}$	$0.39 \pm 0.03^{***}$	$0.48 \pm 0.04^{***}$	$0.85 \pm 0.01^{***}$

to score sample similarity (Box S2, Methods). We created an artificial challenge task by simulating strong batch effects between compound sample replicates in the DRUG-seq calibration dataset (Fig. S3a, Methods). We observed that under strong simulated batch effects, the *numerical-prior* prompt offered a slight improvement over baseline approaches. In contrast, the *functional-prior* prompt demonstrated substantial improvement over both baseline and *numerical-prior* approaches (Fig. 3a, Fig. S3b,c). We further hypothesized that the *functional-prior* prompt may strongly rely on biological priors associated with gene identities. To test this, we again performed a gene ID ablation analysis, finding a significant reduction in expected similarity scores (Fig. 3a, Fig. S3b,c). Across local similarity metrics, ablation reduced the performance of the *numerical-prior* prompt by an average of

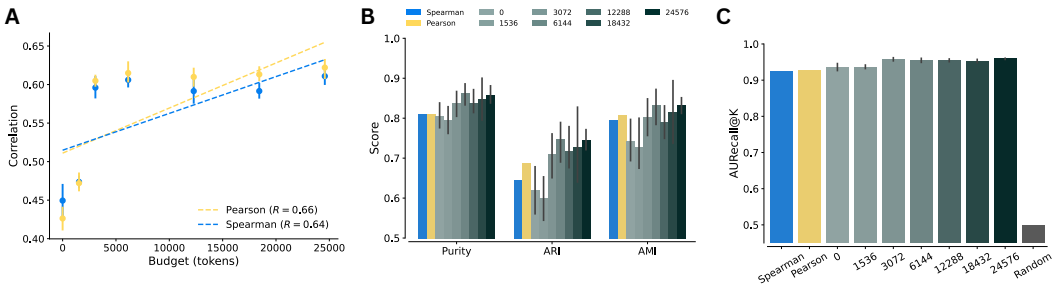

Figure 2: **Global and local similarity analyses.** Comparative evaluation of LFC similarity matrices (Pearson and Spearman) and LLM kernel similarity matrices across reasoning compute budgets. Budgets were sampled more densely at the lower end of the range to capture steeper performance changes. **(A)** Pearson correlation with Pearson similarity ($p = 1.48 \times 10^{-5}$) and Spearman similarity ($p = 3.69 \times 10^{-5}$). **(B)** Clustering metrics: Purity, Adjusted Rand Index (ARI), and Adjusted Mutual Information (AMI). **(C)** Retrieval metric: Area Under the Recall@k curve (AURecall@k). All values are reported as mean $\pm$ standard error of the mean (SEM) (n= 5).

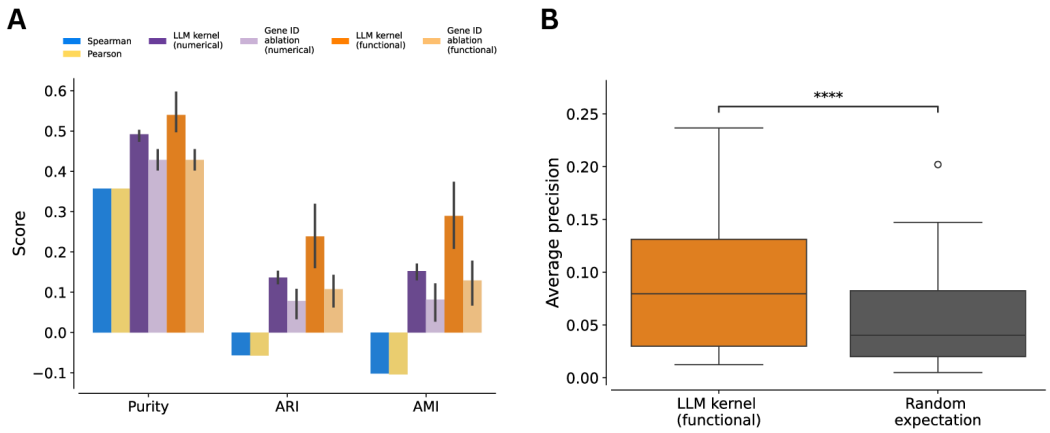

Figure 3: **LLM Kernel enables flexible notions of similarity and cross-modal comparisons.** **(A)** Local similarity structure analysis of batch-simulated transcriptomic data. LFC-based similarity matrices (Spearman and Pearson correlations) and LLM kernel similarity matrices generated using a *numerical-prior* prompt, a *functional-prior* prompt, and each prompt under gene ID ablation. **(B)** Evaluation of average precision for inverse ranking of positive control drug transcriptomes against matched textual disease descriptions. LLM kernel (functional) similarity scores vs. random expectation in recovering known drug–disease relationships (**** $p \leq 10^{-4}$).

24.6%, while the *functional-prior* prompt showed a larger average decrease of 37.7%. An example *functional-prior* reasoning trace can be found in Box S3.

3.4 LLM Kernel enables cross-modal similarity scoring

Given that natural language is used to induce an LLM kernel, we hypothesized that this method could score similarity between different modalities, a novel capability beyond standard numerical approaches. To demonstrate this, we selected 20 diseases with a strong genetic basis and identified relevant approved or investigational drugs (Table S1, Methods). We generated text-based disease signatures, detailing the functional basis and hypothesized transcriptome changes for each disease. We then scored the similarity between these textual disease signatures and the transcriptome profiles of chemical perturbations from positive control drugs and random compounds (Methods). Our hypothesis was that drugs approved for a specific disease would show a complementary (low) similarity score, similar to using negative cosine similarity in transcriptomic screens to identify compounds that induce complementary transcriptomic profiles[3]. Our results showed that the average precision for ranking drugs across the 20 diseases was significantly higher than random (Fig. 3b). This indicates a non-random retrieval of positive control compounds for their corresponding diseases, highlighting a novel capability of cross-modal similarity scoring unique to LLM kernels.

4 Discussion

In this work, we present a method for quantitatively evaluating the reasoning trace of LLMs, addressing a challenge in benchmarking LLMs for scientific interpretation. LLM Kernel prompts an LLM to perform a pairwise comparison between samples and assign a quantitative similarity score. Applied to transcriptomic data, these scores capture a notion of similarity informed by biological knowledge, complementing purely statistical measures.

Our ablation studies support this conclusion. The significant decrease in performance when gene identities are removed indicates the model leverages knowledge of gene function and pathway relationships to inform scoring, rather than simply approximating statistical patterns. In our cross-modal experiment, the model compared natural language disease descriptions with gene expression profiles—a task requiring both modalities to be mapped into a shared functional representation. This native multimodality is a feature of the underlying LLM, differing from prior cross-modal approaches

that require custom architectures to bridge data types. These results suggest the LLM kernel functions by comparing latent representations of biological concepts learned from its training data.

We applied LLM Kernel to evaluate reasoning in a simple, single-step paradigm and found reliable gains in similarity scoring performance as a function of compute, suggesting that the quality of the biological interpretation is tunable. Future work could extend this approach to more complex reasoning pipelines or multi-agent systems.

This framework has several limitations. Language model-based pairwise comparison introduces significant computational costs, making it more resource-intensive than standard numerical methods. Our cross-modal experiment tested a curated list of 20 genetically driven malignancies, pre-filtered for strong molecular bases. Performance will likely differ when applied to more complex, polygenic diseases with less distinct molecular phenotypes. Input prompt structure may also shape interpretation: although we used shuffled gene lists to avoid ordering bias, the overall prompt and data formatting may influence outputs[16]. Interestingly, our *functional-prior* kernel results highlight that sensitivity to prompt design may be a desirable feature, as it suggests prompt tuning provides an effective means to induce diverse notions of similarity.

While interpretations partially externalize model logic, our experiments do not distinguish whether quantitative scores arise directly from the input data, from subsequent reasoning tokens, or from a combination of both. Our finding that similarity scores degrade under compute constraints suggests a dependence on compute volume, rather than a direct link between interpretation quality and performance. Understanding the relationship between compute volume and the quality of reasoning content remains a central question in NLP research, and our framework may provide a means to explore this in future work.

Several areas should be prioritized for further work. First, the framework's generalizability as a validation tool should be tested by extending it to other data modalities and scientific disciplines. Second, the validated reasoning traces themselves could be used as rich, structured features for augmenting human scientific interpretation or for downstream algorithmic learning. Finally, automated prompt optimization could be explored to create specialized kernels fine-tuned to assess specific types of biological reasoning, such as causality or compensatory pathway activation. In conclusion, LLM Kernel offers a framework for extracting knowledge-based similarity scores and quantitatively benchmarking LLM-based interpretations of data.

References

[1] Justin Lamb, Emily D Crawford, David Peck, Joshua W Modell, Irene C Blat, Matthew J Wrobel, Jim Lerner, Jean-Philippe Brunet, Aravind Subramanian, Kenneth N Ross, Michael Reich, Haley Hieronymus, Guo Wei, Scott A Armstrong, Stephen J Haggarty, Paul A Clemons, Ru Wei, Steven A Carr, Eric S Lander, and Todd R Golub. The connectivity map: using gene-expression signatures to connect small molecules, genes, and disease. *Science*, 313(5795): 1929–1935, September 2006.

[2] Aravind Subramanian, Rajiv Narayan, Steven M Corsello, David D Peck, Ted E Natoli, Xiaodong Lu, Joshua Gould, John F Davis, Andrew A Tubelli, Jacob K Asiedu, David L Lahr, Jodi E Hirschman, Zihan Liu, Melanie Donahue, Bina Julian, Mariya Khan, David Wadden, Ian C Smith, Daniel Lam, Arthur Liberzon, Courtney Toder, Mukta Bagul, Marek Orzechowski, Oana M Enache, Federica Piccioni, Sarah A Johnson, Nicholas J Lyons, Alice H Berger, Alykhan F Shamji, Angela N Brooks, Anita Vrcic, Corey Flynn, Jacqueline Rosains, David Y Takeda, Roger Hu, Desiree Davison, Justin Lamb, Kristin Ardlie, Larson Hogstrom, Peyton Greenside, Nathanael S Gray, Paul A Clemons, Serena Silver, Xiaoyun Wu, Wen-Ning Zhao, Willis Read-Button, Xiaohua Wu, Stephen J Haggarty, Lucienne V Ronco, Jesse S Boehm, Stuart L Schreiber, John G Doench, Joshua A Bittker, David E Root, Bang Wong, and Todd R Golub. A next generation connectivity map: L1000 platform and the first 1,000,000 profiles. *Cell*, 171(6):1437–1452.e17, November 2017.

[3] Marina Sirota, Joel T Dudley, Jeewon Kim, Annie P Chiang, Alex A Morgan, Alejandro Sweet-Cordero, Julien Sage, and Atul J Butte. Discovery and preclinical validation of drug indications using compendia of public gene expression data. *Sci. Transl. Med.*, 3(96):96ra77, August 2011.

[4] Daniil A Boiko, Robert MacKnight, Ben Kline, and Gabe Gomes. Autonomous chemical research with large language models. *Nature*, 624(7992):570–578, December 2023.

[5] Siddharth Narayanan, James D Braza, Ryan-Rhys Griffiths, Manu Ponnapati, Albert Bou, Jon Laurent, Ori Kabeli, Geemi Wellawatte, Sam Cox, Samuel G Rodriques, and Andrew D White. Aviary: training language agents on challenging scientific tasks. *arXiv [cs.AI]*, December 2024.

[6] Siddharth M Narayanan, James D Braza, Ryan-Rhys Griffiths, Albert Bou, Geemi Wellawatte, Mayk Caldas Ramos, Ludovico Mitchener, Samuel G Rodriques, and Andrew D White. Training a scientific reasoning model for chemistry. *arXiv [cs.LG]*, June 2025.

[7] Kexin Huang, Serena Zhang, Hanchen Wang, Yuanhao Qu, Yingzhou Lu, Yusuf Roohani, Ryan Li, Lin Qiu, Gavin Li, Junze Zhang, Di Yin, Shruti Marwaha, Jennefer N Carter, Xin Zhou, Matthew Wheeler, Jonathan A Bernstein, Mengdi Wang, Peng He, Jingtian Zhou, Michael Snyder, Le Cong, Aviv Regev, and Jure Leskovec. Biomni: A general-purpose biomedical AI agent. *bioRxivorg*, page 2025.05.30.656746, June 2025.

[8] Eric Wang, Samuel Schmidgall, Paul F Jaeger, Fan Zhang, Rory Pilgrim, Yossi Matias, Joelle Barral, David Fleet, and Shekoofeh Azizi. TxGemma: Efficient and agentic LLMs for therapeutics. *arXiv [cs.AI]*, April 2025.

[9] Juraj Gottweis, Wei-Hung Weng, Alexander Daryin, Tao Tu, Anil Palepu, Petar Sirkovic, Artiom Myaskovsky, Felix Weissenberger, Keran Rong, Ryutaro Tanno, Khaled Saab, Dan Popovici, Jacob Blum, Fan Zhang, Katherine Chou, Avinatan Hassidim, Burak Gokturk, Amin Vahdat, Pushmeet Kohli, Yossi Matias, Andrew Carroll, Kavita Kulkarni, Nenad Tomasev, Yuan Guan, Vikram Dhillon, Eeshit Dhaval Vaishnav, Byron Lee, Tiago R D Costa, José R Penadés, Gary Peltz, Yunhan Xu, Annalisa Pawlosky, Alan Karthikesalingam, and Vivek Natarajan. Towards an AI co-scientist. *arXiv [cs.AI]*, February 2025.

[10] Kyle Swanson, Wesley Wu, Nash L Bulaong, John E Pak, and James Zou. The virtual lab of AI agents designs new SARS-CoV-2 nanobodies. *Nature*, pages 1–3, July 2025.

[11] Hanchen Wang, Tianfan Fu, Yuanqi Du, Wenhao Gao, Kexin Huang, Ziming Liu, Payal Chandak, Shengchao Liu, Peter Van Katwyk, Andreea Deac, Anima Anandkumar, Karianne Bergen, Carla P Gomes, Shirley Ho, Pushmeet Kohli, Joan Lasenby, Jure Leskovec, Tie-Yan Liu, Arjun Manrai, Debora Marks, Bharath Ramsundar, Le Song, Jimeng Sun, Jian Tang, Petar Veličković, Max Welling, Linfeng Zhang, Connor W Coley, Yoshua Bengio, and Marinka Zitnik. Scientific discovery in the age of artificial intelligence. *Nature*, 620(7972):47–60, August 2023.

[12] Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and A Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Neural Inf Process Syst*, abs/2209.09513:2507–2521, September 2022.

[13] David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R Bowman. GPQA: A graduate-level google-proof Q&A benchmark. *arXiv [cs.AI]*, November 2023.

[14] Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Appl. Sci. (Basel)*, 11(14):6421, July 2021.

[15] Jingyao Li, Daniel J Ho, Martin Henault, Chian Yang, Marilisa Neri, Robin Ge, Steffen Renner, Leandra Mansur, Alicia Lindeman, Brian Kelly, Tayfun Tumkaya, Xiaoling Ke, Gilberto Soler-Llavina, Gopi Shanker, Carsten Russ, Marc Hild, Caroline Gubser Keller, Jeremy L Jenkins, Kathleen A Worringer, Frederic D Sigoillot, and Robert J Ihry. DRUG-seq provides unbiased biological activity readouts for neuroscience drug discovery. *ACS Chem. Biol.*, 17(6):1401–1414, June 2022.

[16] Haokun Liu, Yangqiaoyu Zhou, Mingxuan Li, Chenfei Yuan, and Chenhao Tan. Literature meets data: A synergistic approach to hypothesis generation. *arXiv [cs.AI]*, October 2024.

[17] Chaoyang Ye, Daniel J Ho, Marilisa Neri, Chian Yang, Tripti Kulkarni, Ranjit Randhawa, Martin Henault, Nadezda Mostacci, Pierre Farmer, Steffen Renner, Robert Ihry, Leandra Mansur, Caroline Gubser Keller, Gregory McAllister, Marc Hild, Jeremy Jenkins, and Ajamete Kaykas. DRUG-seq for miniaturized high-throughput transcriptome profiling in drug discovery. *Nat. Commun.*, 9(1):4307, October 2018.

[18] Stephen M Canham, Yuan Wang, Allen Cornett, Douglas S Auld, Daniel K Baeschlin, Maude Patoor, Philip R Skaanderup, Ayako Honda, Luis Llamas, Greg Wendel, Felipa A Mapa, Peter Aspesi, Jr, Nancy Labbé-Giguère, Gabriel G Gamber, Daniel S Palacios, Ansgar Schuffenhauer, Zhan Deng, Florian Nigsch, Mathias Frederiksen, Simon M Bushell, Deborah Rothman, Rishi K Jain, Horst Hemmerle, Karin Briner, Jeffery A Porter, John A Tallarico, and Jeremy L Jenkins. Systematic chemogenetic library assembly. *Cell Chem. Biol.*, 27(9):1124–1129, September 2020.

5 Methods

5.1 Data sources

5.1.1 DRUG-seq calibration dataset

A publicly available dataset designed to evaluate DRUG-seq assay reproducibility was utilized. DRUG-seq is a high-throughput transcriptomic profiling method for measuring gene expression changes induced by small molecules or drugs in cells[17]. The original dataset was generated from nine 384-well plates, comprising three independent batches of U2OS cells plated on different days, with three replicate plates per batch[15]. Each plate was treated with 14 compounds across an eight-point dose response ($3.2\,\mathrm{nM}$ to $10\,\mu\mathrm{M}$), with three replicates per dose. The original experiment demonstrated strong assay fidelity across batches and plates (Fig. 1a). Differential expression scores and corresponding statistics were generated for drug treatment samples using the publicly available analysis pipeline (`https://github.com/Novartis/DRUG-seq/tree/main/data/Novartis_drugseq_U2OS_MoABox`). For downstream analysis, $3.16\,\mu\mathrm{M}$ treatment samples of all 14 compounds across three batches were used, yielding three replicates per compound treatment.

5.1.2 Simulated batch effect DRUG-seq calibration dataset

To benchmark the robustness of LLM kernels against technical confounding, synthetic batch effects were introduced into the DRUG-seq calibration dataset. This evaluated the sensitivity of LLM kernel scoring to systematic numerical biases by simulating realistic batch effects that degrade transcriptome signal fidelity. A two-component simulation model was developed to reflect both global and gene-specific sources of batch-induced variation:

1. **Systematic shift (B_i)**: A scalar offset applied uniformly to the full transcriptome vector of all samples in batch i, representing global technical variation (e.g., plate effects, reagent lot differences).
2. **Sparse gene-specific shift (G_{ij})**: Random offsets applied to a subset of genes within batch i, modeling heterogeneous gene-level sensitivity to batch-specific conditions.

The model was parameterized as follows. For each batch i, the systematic shift was drawn as $B_i \sim \mathcal{N}(0,\, 3^2)$ and applied uniformly across all genes in that batch. Gene-specific effects were specified by an indicator variable $Z_j \sim \text{Bernoulli}(0.4)$, denoting whether gene j was affected. For each gene j in batch i, the shift was then drawn as $G_{ij} \sim Z_j \cdot \mathcal{N}(0,\, 1.5^2)$.

These parameters were chosen to introduce strong yet biologically plausible confounding while maintaining interpretability. Original experimental batches defined batch identities. The noised LFC matrix was normalized by dividing each value by the batch effect strength ($\sigma = 3$). This post hoc rescaling ensured that the overall dynamic range of the data remained comparable to the original dataset, facilitating downstream comparisons (Fig. S3a).

5.1.3 NIBR MoA Box dataset

We used the Novartis Institutes for BioMedical Research (NIBR) Mechanism-of-Action (MoA) Box dataset, a large, high-quality, publicly available DRUG-seq resource, for cross-modal comparisons between textual disease descriptions and perturbation transcriptomes[18]. The dataset comprises 4,755 compounds curated for strong target annotation, broad target diversity and specificity, low redundancy, and overall bioactivity. Raw count files were reprocessed using an internal RNA-seq analysis pipeline.

5.1.4 ChEMBL indications and compounds

MoA Box compounds were mapped to disease indications for cross-modal LLM kernel similarity scoring. ChEMBL (v33) compound structures, indications (MeSH terms), and maximum approval phase were extracted. MoA Box compound structures were aligned to corresponding ChEMBL compounds with associated indications, and compound–indication pairs were filtered to those with maximum approval phase ≥ 3. The resulting set of unique indications ($n = 574$) was provided to a language model (Google Gemini Flash 2.5), which was prompted to identify genetically driven malignancies. This process yielded 20 indications with a strong molecular basis for disease (Table S1).

We focused on genetically driven malignancies because many reference indications were broad (e.g., *Postoperative Nausea and Vomiting*, *Burns*, *Smoking Cessation*) or represented complex diseases lacking clear molecular phenotypes.

5.2 LLM kernel methods

5.2.1 Data preprocessing

For each sample, differentially expressed genes (DEG) were filtered to those with a false discovery rate (FDR) adjusted p-value < 0.05 and an absolute $\log_2$ fold change (LFC) value ≥ 0.5. The DEG vector was then sorted in descending order by absolute LFC, and a maximum of the top 1000 genes were selected.

Each DEG vector was randomly shuffled, and a list containing the HGNC gene symbol, LFC, and FDR-adjusted p-value was injected into the prompt for each perturbation $\{A, B\}$ in a tuple format: `[(HGNC_symbol, log2_fold_change, FDR), ...]`. This method resulted in variable-length lists up to 1000 genes, with ordering independent of DEG effect size.

5.2.2 *Numerical-prior* prompt

A prompt was designed to instruct the LLM to compare the similarity of two transcriptomic vectors. This prompt requested independent interpretations of both perturbations, a similarity analysis of the perturbations, and a similarity score in the range $[-1, 1]$ (Box S1). A structured output was implemented to capture each component of the response.

For individual perturbation analysis, the requested outputs were:

- `reasoning: str`
- `key_genes: list`
- `affected_pathways: list`
- `biological_impact: str`

For similarity analysis, the requested outputs were:

- `overlapping_genes: list`
- `overlapping_pathways: list`
- `similarity_reasoning: str`
- `similarity_score: float,` $[-1, 1]$

This *numerical-prior* prompt biased the response towards direct consideration of key genes, pathways, and their explicit overlap between perturbations. Models `gemini-2.5-flash-preview-05-20` or `gemini-2.5-flash` were queried via Google Vertex API under default settings, with maximum thinking budget modified as appropriate for experiments via the `thinkingBudget` API parameter.

5.2.3 *Functional-prior* prompt

A secondary prompt was designed to assess functional similarity between transcriptome vectors, emphasizing mechanistic interpretation rather than gene extraction and overlap reasoning in the comparison (Box S2). For each perturbation, the model was prompted to infer upstream drivers (identify likely direct and secondary targets explaining expression changes), characterize downstream effects (map affected biological pathways and cellular processes), and assess mechanistic similarity (compare causal mechanisms, functional concordance, and pathway-level effects). A structured output was implemented to capture each component of the response.

For individual perturbation analysis, the requested outputs were:

- `primary_targets: list`
- `secondary_targets: list`
- `upstream_regulators: list`

- `affected_pathways: list`
- `functional_themes: list`
- `causal_reasoning: str`

For similarity analysis, the requested outputs were:

- `shared_targets: list`
- `shared_functional_themes: list`
- `mechanistic_comparison: str`
- `similarity_score: float [-1, 1]`

Models `gemini-2.5-flash-preview-05-20` or `gemini-2.5-flash` were queried via Google Vertex API under default settings, with maximum thinking budget modified as appropriate for experiments.

5.3 Data postprocessing

5.3.1 Baseline similarity scores

Spearman and Pearson correlation coefficients were computed between preprocessed DEG vectors using `pandas.DataFrame.corr()`. Undefined sample comparisons (i.e., samples with no overlapping gene set) were imputed with 0.

5.3.2 LLM Kernel scores

Each perturbation pair $\{A, B\}$ was tested once, with randomized order in the LLM query. Up to three attempts were made to retrieve a valid API response; nonetheless, a nonzero failure rate was observed in each experiment (Fig. S1c). Failures typically occurred when responses exceeded the maximum token limit before completing the structured output. The LLM kernel score was extracted from each successful response, and a symmetric similarity matrix was constructed. For failed queries, similarity scores were imputed as 0, except for self-scored pairs $\{A, A\}$, where diagonal elements were set to 1 following self-consistency analysis.

5.3.3 Self-consistency analysis

Two self-consistency metrics were computed to assess whether each method preserved the property that samples should have perfect similarity with themselves (Fig. S1a,b). *Hard self-consistency* was defined as the proportion of diagonal elements in the similarity matrix equal to 1.0. *Soft self-consistency* was defined as the average value of diagonal elements (excluding failed comparisons). For both metrics, a value of 1.0 corresponds to perfect self-consistency. These measures served as quality control, with lower values indicating potential systematic biases or methodological issues in similarity matrix generation.

5.3.4 Hierarchical clustering

Agglomerative hierarchical clustering was applied to the processed data matrix to identify groups of similar samples. Pairwise Euclidean distances between rows of the matrix were computed using the `pdist` function from SciPy. These distances were provided as input to the `linkage` function with Ward's method, which iteratively merged clusters to minimize total within-cluster variance. The resulting linkage matrix captured the hierarchical structure of the data and was used for dendrogram visualization and cluster analysis.

5.4 Statistical analysis

5.4.1 Global Correlation

To assess global agreement between similarity matrices, we extracted the upper triangles from both the baseline LFC similarity matrix (computed using Pearson or Spearman correlation) and the LLM kernel similarity matrix. The Pearson correlation coefficient was then calculated between these two upper triangles to quantify overall concordance.

5.4.2 Local clustering analysis: Purity, ARI, AMI

Hierarchical clustering was performed to group samples by compound treatment, with an expected outcome of 14 clusters (each containing 3 replicates, corresponding to the 14 compounds). The `fcluster` function from `SciPy`, with the `maxclust` criterion, was applied to the linkage matrix (derived from the similarity matrix) to assign samples to clusters.

Clustering quality was evaluated using three complementary metrics:

Purity Measures the extent to which clusters align with ground-truth labels. For each cluster, the most frequent true label is identified, these maximum counts are summed across clusters, and the total is normalized by the number of samples. Values range from 0 to 1, with higher scores indicating better alignment.

Adjusted Rand Index (ARI) Quantifies the agreement between clustering assignments and true labels, corrected for chance. ARI ranges from -1 to 1, with 0 indicating agreement at the level expected by chance, and higher values signifying stronger agreement (implemented in `scikit-learn`).

Adjusted Mutual Information (AMI) Captures the mutual information between clustering assignments and true labels, adjusted for chance. AMI ranges from -1 to 1, with 0 indicating agreement at the level expected by chance, and higher values signifying stronger agreement (implemented in `scikit-learn`).

Across all three metrics, higher values indicate superior clustering performance of compound treatment groups.

5.4.3 Local neighborhood analysis: AURecall@k

To evaluate how well a similarity matrix preserves replicate sample relationships, we computed recall@k metrics. For each sample, its k-nearest neighbors were identified from the similarity matrix and compared to the expected set of replicates. The recall@k score was defined as the proportion of expected replicates recovered among the k-nearest neighbors, averaged across samples. Scores were calculated for $k = 1, \ldots, 42$, and the area under the recall@k curve (AURecall@k) was then computed as a comprehensive summary measure. Higher values indicate that replicate samples from the same compound treatment are more frequently identified as nearest neighbors, reflecting better preservation of expected local structure in the similarity matrix.

5.5 Ablation experiments

Gene ID ablation HGNC symbols were replaced with standardized random identifiers (e.g., `GENE_001`, `GENE_002`). This mapping was fixed across all samples so that each gene maintained the same masked ID. For example, input data appeared as: `[(GENE_678, log2_fold_change, FDR), ...]`.

Direction ablation Directional information was removed by taking the absolute values of $\log_2$ fold-change (LFC) values before injection into the LLM kernel.

Gene shuffle ablation The relationship between gene identity and statistics was disrupted by randomly shuffling HGNC symbols relative to their corresponding LFC and adjusted p-values. Shuffling was performed independently for each sample, ensuring that the overall set of gene identities and statistics was preserved but their associations were randomized.

5.5.1 Analysis

For each ablation and baseline method, a similarity matrix was computed. LLM kernel methods and their ablations were run in triplicate, each with a thinking budget of 6144.

Global Correlation Pearson correlation was calculated between each method's similarity matrix and a reference Pearson correlation similarity matrix. Global correlations were compared to the Pearson LFC similarity matrix using a Student's t-test.

Local similarity Clustering and retrieval metrics—Purity, ARI, AMI, and AURecall@k—were computed. Each method's scores were compared against the numerical LLM kernel baseline using a Student's t-test.

Code availability A framework for the LLM Kernel method is available at `https://github.com/transcriptabio/llm-kernel`.

A Supplementary Material

A.1 Supplementary Figures

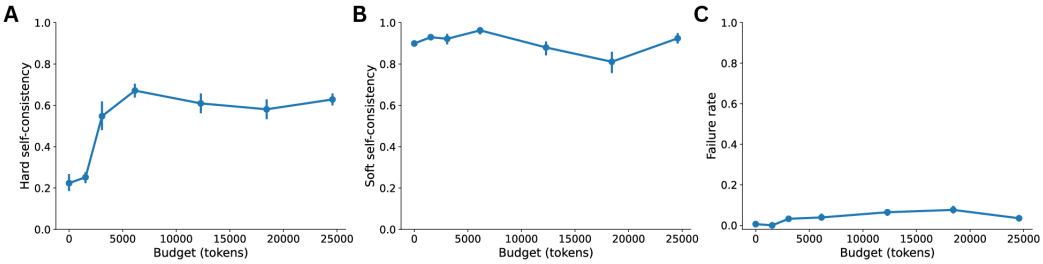

Figure S1: **LLM Kernel response success metrics.** **(A)** Hard self-consistency, **(B)** soft self-consistency, and **(C)** response failure rate across thinking budgets.

A.2 Supplementary Tables

Table S1: **Genetic driven malignancies.** Indications queried for cross-modal LLM kernel similarity scoring (n=20).

Disease
Leukemia, Myelogenous, Chronic, BCR-ABL Positive
Leukemia, Lymphocytic, Chronic, B-Cell
Leukemia, Myeloid, Acute
Leukemia, Promyelocytic, Acute
Lymphoma, Mantle-Cell
Lymphoma, Follicular
Lymphoma, Large B-Cell, Diffuse
Lymphoma, T-Cell, Peripheral
Multiple Myeloma
Carcinoma, Non-Small-Cell Lung
Breast Neoplasms
Melanoma
Astrocytoma
Carcinoma, Renal Cell
Wilms Tumor
Neuroendocrine Tumors
Adenocarcinoma
Myelodysplastic Syndromes
Lymphoma, Non-Hodgkin
Leukemia, Mast-Cell

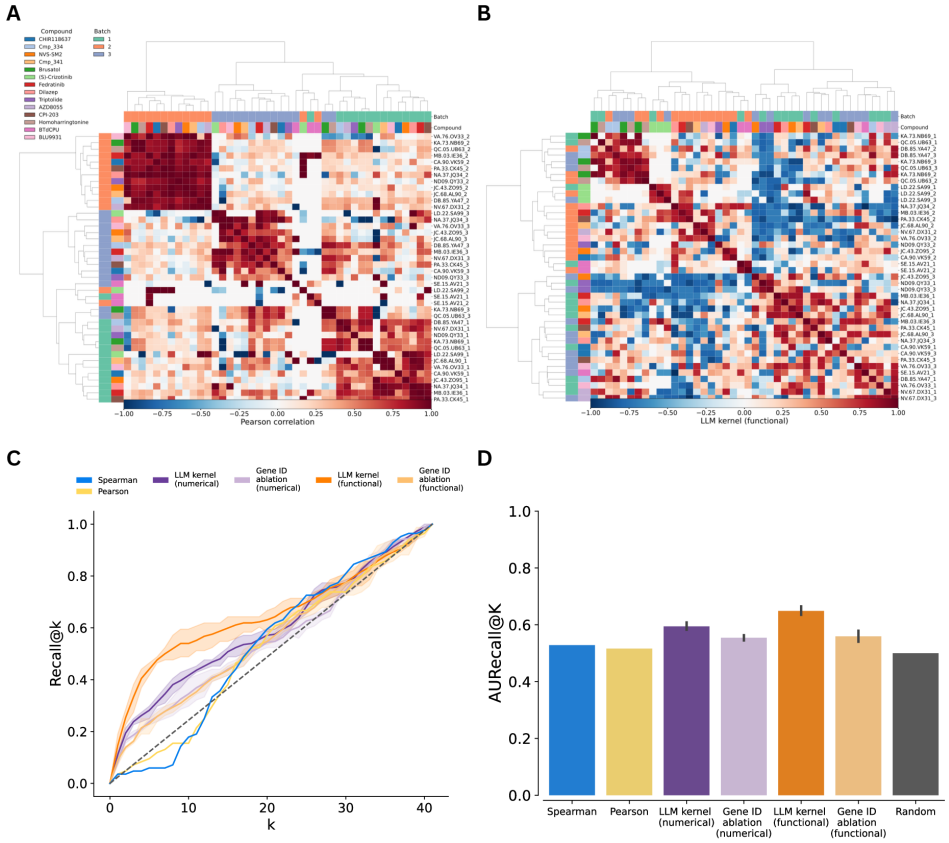

Figure S2: **Batch-simulated DRUG-seq dataset analysis. (A)** Heatmaps of batch-simulated dataset LFC similarity matrix (left) and LLM kernel similarity matrix (right) with samples annotated by batch identity and compound treatment. **(B)** Recall@k lineplot and **(C)** AURecall@k barplot for LFC-based similarity matrices (Spearman and Pearson correlations) compared to LLM kernel similarity matrices generated using a *numerical-prior* prompt, a *functional-prior* prompt, each prompt under gene ID ablation, and random expectation (0.5).

A.3 Supplementary Boxes

***Numerical-prior* prompt**

```
You are a scientist analyzing RNA-seq data to compare the biological impact of
     two perturbations.

Each perturbation alters the expression of genes, which in turn affects broader
     biological pathways, transcriptional programs, or regulatory modules.
     However, not all perturbations have strong or clear biological effects.

You are given the output of differential expression analyses (from limma) for
     each perturbation. These are lists of significantly affected genes (FDR <=
     0.05), along with their log2 fold changes (abs(LFC) >= 0.5). The lists are
     unordered and vary in length:

Perturbation A: [(HGNC_symbol, log2_fold_change, fdr), ...]
Perturbation B: [(HGNC_symbol, log2_fold_change, fdr), ...]

Your task is to analyze these perturbations and provide:

1. Analysis of Perturbation A:
 - Reasoning about the biological impact, considering:
   * Direction and magnitude of gene expression changes
   * Key genes affected and their biological functions
```

```
    * Affected pathways and regulatory modules
    * Overall biological impact summary
  - If the changes are minimal or unclear, explicitly state this
  - If the biological significance is uncertain, acknowledge this uncertainty
  - Key genes affected (list of gene symbols)
  - Affected pathways (list of pathway names)
  - Biological impact (detailed summary, including any limitations or
      uncertainties)

2. Analysis of Perturbation B:
  - Reasoning about the biological impact, considering:
    * Direction and magnitude of gene expression changes
    * Key genes affected and their biological functions
    * Affected pathways and regulatory modules
    * Overall biological impact summary
  - If the changes are minimal or unclear, explicitly state this
  - If the biological significance is uncertain, acknowledge this uncertainty
  - Key genes affected (list of gene symbols)
  - Affected pathways (list of pathway names)
  - Biological impact (detailed summary, including any limitations or
      uncertainties)

3. Similarity Analysis:
  - Reasoning about how the perturbations are similar or different, considering:
    * Genes changing in the same direction contribute to positive similarity
    * Genes changing in opposite directions contribute to negative similarity
    * Magnitude of changes affects the strength of the similarity
    * Overall pattern of gene expression changes
    * Both shared and unique changes in each perturbation
    * Relative magnitudes of changes
  - If either perturbation has minimal effects, this should be reflected in the
      similarity score
  - If the biological significance is unclear, the similarity score should be more
      conservative

Guidelines for scoring:
- 1.0: Perfect positive correlation (identical changes in same direction)
- 0.5-0.9: Strong positive correlation (many shared changes in same direction)
- 0.0-0.4: Weak positive correlation (some shared changes but many differences)
- 0.0: No correlation (no overlapping genes, unrelated effects, or minimal
    effects)
- -0.4-0.0: Weak negative correlation (some changes in opposite directions)
- -0.9--0.5: Strong negative correlation (many changes in opposite directions)
- -1.0: Perfect negative correlation (identical changes in opposite directions)

Important guidelines:
1. If the biological functions of the affected genes are completely unrelated,
    the similarity score should be 0.0
2. If either perturbation has minimal or unclear effects, the similarity score
    should be more conservative
3. When in doubt about the biological significance, err on the side of caution
4. Do not overinterpret small or unclear changes
5. Explicitly acknowledge when the biological impact is uncertain or minimal

Perturbation A: {pert_a}
Perturbation B: {pert_b}
```

Box S1: ***Numerical-prior* prompt.** Direct prompt input with reference variables (`pert_a`, `pert_b`).

***Functional-prior* prompt**

```
You are a systems biologist analyzing RNA-seq data to infer the mechanistic
    impact of two perturbations.

Each perturbation causes differential gene expression, which reflects the
    engagement of upstream targets and regulatory pathways. These responses may
    be caused by a single molecular interaction or multiple targets, especially
    in the case of small molecules with polypharmacology.

You are provided with the results of differential expression analyses (from
    limma) for each perturbation. These are lists of significantly affected
    genes, filtered by:
- FDR <= 0.05
- Absolute log2 fold change >= 0.5
```

```
Format:
Perturbation A: [(HGNC_symbol, log2_fold_change, FDR), ...]
Perturbation B: [(HGNC_symbol, log2_fold_change, FDR), ...]

Your task is to reason about the **mechanistic drivers** of the observed
     responses and provide a structured biological interpretation for each
     perturbation, ultimately to score the functional similarity between the two
     perturbations. Do not score merely based on the overlap of genes, or the
     magnitude of the log2 fold changes, as there may be experimental noise in
     the data.

-----------------------------------------------------
1. Mechanistic Analysis of Perturbation A:
 - Consider the **upstream causes** of the observed expression changes, including:
   * Likely **direct targets** (e.g., enzymes, receptors, transcription factors),
       which may be modulated by the perturbation
   * Potential **secondary targets**, consistent with a polypharmacology profile
   * Engagement of known **regulatory networks**, signaling cascades, or
       transcriptional programs
 - Describe the **downstream consequences** reflected in the gene expression
     profile, including:
   * Key affected pathways or cellular processes
   * Overall biological theme (e.g., ER stress, apoptosis, metabolic
       reprogramming)
 - Acknowledge uncertainty if the target or causal mechanism is unclear
 - Output:
   * Hypothesized direct targets: [list of gene/protein names]
   * Affected biological systems: [list of pathway names or processes]
   * Mechanistic summary: [paragraph with causal reasoning]

2. Mechanistic Analysis of Perturbation B:
 - Use the same structure and reasoning as above

-----------------------------------------------------
3. Comparative Mechanistic Similarity:
 - Compare the **causal mechanisms and biological programs** engaged by
     perturbations A and B
 - Consider:
   * Do they affect similar pathways or cellular systems?
   * Are the **likely targets** overlapping or distinct?
   * Are the observed effects **functionally concordant**, **opposing**, or
       unrelated?
   * Could they result from **shared upstream regulators** or **distinct
       mechanisms converging on the same output**?

 - Avoid scoring based on gene-level overlap or log2 fold change magnitude
 - Prioritize **biological reasoning** about mechanistic similarity
 - If either perturbation's mechanism is unclear or biologically diffuse, use a
     conservative score

 - Output:
   * Mechanistic comparison: [paragraph]
   * Functional similarity score: [float from -1.0 to 1.0]
   * Justification: [paragraph]

-----------------------------------------------------
Scoring Guidelines:
- +1.0: Identical mechanisms and downstream programs
- +0.5 to +0.9: Strongly similar causal programs or affected pathways
- 0.0 to +0.4: Partial or unclear overlap in mechanisms or systems
- 0.0: Distinct or ambiguous biological mechanisms
- -0.4 to -0.0: Opposing effects on shared pathways
- -0.9 to -0.5: Strong inverse regulation of common biological systems
- -1.0: Identical programs activated in opposite directions

Evaluation principles:
1. Focus on **causal reasoning**, not just expression magnitude or overlap
2. There may be experimental noise in the data, so do not overinterpret the
     numerical values of log2 fold changes
3. Explicitly identify and interpret **candidate upstream drivers**
4. Acknowledge and accommodate the possibility of **polypharmacology**
5. Do not overinterpret weak or diffuse transcriptional signals
6. Be conservative in scoring when uncertainty is high

-----------------------------------------------------
Perturbation A: {pert_a}
Perturbation B: {pert_b}
```

Box S2: ***Functional-prior*** **prompt.** Direct prompt input with reference variables (`pert_a`, `pert_b`).

Example *functional-prior* reasoning trace


```
"response": {
    "pert_a": {
        "primary_targets": [],
        "secondary_targets": [],
        "upstream_regulators": [
          "Integrated Stress Response",
          "Wnt signaling pathway",
          "Cell cycle regulators"
        ],
        "affected_pathways": [
          "Integrated Stress Response",
          "ER Stress",
          "Wnt signaling pathway",
          "Cell cycle regulation",
          "Lipid metabolism",
          "Extracellular matrix organization",
          "Muscle structure and function"
        ],
        "functional_themes": [
          "Cellular Stress (ER/ISR)",
          "Anti-proliferative effects",
          "Tissue remodeling",
          "Altered cytoskeletal dynamics"
        ],
        "causal_reasoning": "Perturbation A induces a cellular stress response,
             evidenced by the upregulation of key Integrated Stress Response
             (ISR) and ER stress markers such as DDIT3 (CHOP), SESN2, NUPR1, and
             CHAC1. Paradoxically, ATF4, a central ISR transcription factor, is
             downregulated, suggesting a complex or perhaps a distinct branch of
             the stress response, or an intricate interplay with other affected
             pathways. Beyond stress, there's a prominent downregulation of genes
             involved in Wnt signaling (WNT5B, FZD7), suggesting an inhibition of
             this pathway, and a broad suppression of cell cycle components
             (e.g., MCM6, ORC4, KIF20A, CENPF), indicating anti-proliferative
             effects. Genes related to muscle structure (DYSF, TNNT1, TNNT2,
             TPM1) and extracellular matrix organization (COL4A2, CTGF, SERPINE1)
             are also significantly downregulated, pointing towards tissue
             remodeling or altered cellular adhesion. Upregulation of specific
             keratin family members (KRT17, KRT75) and certain lncRNAs (SNHG1,
             SNHG15, SNHG6) further characterize the cellular adaptation to this
             perturbation. While direct molecular targets are not explicit from
             this data, the overall profile suggests a perturbation that broadly
             impacts cellular homeostasis, possibly by inducing proteotoxic
             stress, leading to a coordinated cellular response involving stress
             pathways, cell cycle inhibition, and changes in structural
             components."
        },
    "pert_b": {
        "primary_targets": [],
        "secondary_targets": [],
        "upstream_regulators": [
          "Integrated Stress Response",
          "ER Stress"
        ],
        "affected_pathways": [
          "Integrated Stress Response",
          "ER Stress",
          "Protein unfolding and degradation",
          "Metabolic reprogramming (gluconeogenesis)",
          "Cell cycle progression",
          "Extracellular matrix remodeling"
        ],
        "functional_themes": [
          "Severe Cellular Stress (ER/Proteotoxic Stress)",
          "Metabolic Adaptation",
          "Increased cell proliferation/growth signaling"
        ],
        "causal_reasoning": "Perturbation B elicits a remarkably strong and
             canonical Integrated Stress Response (ISR) and ER stress signature.
             This is overwhelmingly evident from the highly significant
             upregulation of multiple core ISR genes, including PPP1R15A
             (GADD34), DDIT3 (CHOP), GADD45B, ATF4P3, TRIB3, and TXNIP. PPP1R15A,
             in particular, is a potent hallmark of an active ISR, involved in
```

```
            the dephosphorylation of eIF2\u03b1 and translational recovery
            following stress. The coordinated induction of these factors points
            to a severe disruption of protein folding or ER homeostasis.
            Concurrently, genes involved in metabolic adaptation, such as PCK2
            (gluconeogenesis) and SLC2A3 (glucose transport), are upregulated,
            suggesting metabolic reprogramming in response to the stress.
            Upregulation of MMP3 indicates extracellular matrix remodeling,
            often a component of stress or repair responses. Multiple Small
            Nucleolar Host Genes (SNHG1, SNHG6, SNHG15, SNHG17) are also
            consistently upregulated, indicating changes in RNA processing.
            While direct molecular targets cannot be determined solely from this
            transcriptomic data, the observed gene expression profile strongly
            implicates a perturbation that severely challenges cellular
            proteostasis and activates a robust ER stress response, leading to
            downstream metabolic and structural adaptations."
    },
    "shared_targets": [],
    "shared_regulators": [
        "Integrated Stress Response",
        "ER Stress pathway"
        ],
        "shared_functional_themes": [
        "Cellular Stress Response"
        ],
    "mechanistic_comparison": "Both Perturbation A and Perturbation B induce a
         significant cellular stress response, with the upregulation of DDIT3
         being a common and strong indicator of Integrated Stress Response (ISR)
         activation in both cases. Both also show consistent upregulation of
         several SNHG family lncRNAs (SNHG1, SNHG6, SNHG15), SLC2A3, PEX16,
         RPL32P29, KHK, SLC12A7, and XAGE5. However, while the overarching theme
         of 'cellular stress' is shared, the specific molecular signatures and
         downstream consequences diverge significantly. Perturbation B presents a
         more robust and canonical ISR profile, characterized by the strong
         induction of PPP1R15A, GADD45B, ATF4P3, TRIB3, and TXNIP. In contrast,
         Perturbation A exhibits a more complex and mixed stress signature,
         notably with a downregulation of ATF4, and additional widespread effects
         on cell cycle inhibition, Wnt signaling suppression, and muscle/ECM
         components. A substantial number of genes involved in stress,
         metabolism, and cellular function are regulated in *opposite directions*
         between the two perturbations (e.g., GLRX, TM7SF2, ABTB1, C6orf48, EMC9,
         HDAC5, FOXO6, WIPI1, RASA3, YPEL3, MMP3, DYSF, CTGF, CCL3L3, TCF19,
         LA16c-83F12.6, ADCK3). This indicates that while both perturbations
         ultimately lead to cellular stress, their initial mechanistic triggers
         or the specific adaptive pathways engaged are largely distinct,
         resulting in functionally divergent downstream responses beyond the
         common core stress pathway.",
    "similarity_score": 0.4
}
"thoughts": "Alright, let's get this done. I've been given RNA-seq data for two
     perturbations, A and B, and I need to dive deep into their systems biology
     implications. My task is to analyze each one separately, looking for key
     targets, upstream regulators, affected pathways, and functional themes, and
     then compare them to assess their similarity. I've got a checklist to
     follow, and a high confidence score, so let's get started.

First, Perturbation A. I'll start by sifting through the differentially expressed
     genes, focusing on those with the most significant changes in expression.
     I'm looking for the big hitters \u2013 genes that are strongly up- or
     downregulated. Immediately, I'm noting things like the downregulation of
     DYSF, WNT5B, and MYEOV, and upregulation of keratin genes like KRT17 and
     KRT75. From this gene list, I can see a pattern beginning to emerge: it
     seems that Perturbation A might be interfering with the ER, along with
     suppressing the Wnt pathway and cell cycle activity. The upregulation of ER
     stress-related genes like DDIT3, CHAC1, and SESN2, alongside the
     downregulation of cell cycle components and muscle-related genes, is a clear
     indication of stress and possible interference with cellular processes. My
     hypothesis is that it's an activation of the ISR.

Moving on to Perturbation B, the analysis will follow the same rigorous approach.
     Scanning the data, I immediately see a significant upregulation of PPP1R15A,
     a key player in the integrated stress response, along with strong activation
     of DDIT3, GADD45B, and ATF4P3. This strongly suggests a scenario of ER
     stress or a similar type of cellular stress. The upregulated SNHG lncRNAs,
     coupled with metabolic changes such as increased PCK2 and SLC2A3, reinforce
     this perspective.

Then, the comparison. Both perturbations seem to induce an ER stress/ISR response
     based on shared DDIT3 upregulation. However, Perturbation B's induction is
```

```
    more pronounced and complete. There is the strong signal from the
    upregulation of key ISR genes, especially PPP1R15A. Looking closer, though,
    the contrasting gene expression patterns are striking. Perturbation A
    presents a more complex signature, with indications of cell cycle
    inhibition, Wnt suppression, and effects on lipid metabolism, muscle, and
    cytoskeleton. Perturbation B, on the other hand, seems to exhibit a more
    direct and strong ER stress response. Many genes show opposing regulation in
    the two perturbations. This opposing regulation reduces the overall
    mechanistic congruence.

Finally, the similarity score. Although the core theme of ISR is shared, the
    contrasting gene expression profiles suggest distinct initiating mechanisms
    or cellular responses. Given the shared core ISR component (DDIT3
    upregulation) but the significant differences in how various genes are
    modulated, the mechanistic similarity is moderate. Therefore, the similarity
    score will land at 0.4."
```

Box S3: **Example reasoning trace.** Example reasoning trace for the *functional-prior* prompt with structured output and summarized thinking.