GENE SET FUNCTION DISCOVERY WITH LLM-BASED AGENTS AND KNOWLEDGE RETRIEVAL

Daniela Pinto,[*] Aécio Santos,[†] Juliana Freire[*†], Sarah Keegan,[‡] Wenke Liu,[‡] David Fenyö[‡]
New York University
{daniela.pinto,aecio.santos,juliana.freire}@nyu.edu,
{sarah.keegan,wenke.liu,david.fenyo}@nyulangone.org

ABSTRACT

Advancements in high-throughput technologies have generated complex biomedical datasets, posing significant challenges for knowledge discovery. Traditional tools like Gene Set Enrichment Analysis (GSEA) and over-representation analysis (ORA) map gene sets to known pathways but are limited in their ability to uncover novel biological-mechanisms, often relying on manual interpretation to synthesize insights. While large language models (LLMs) aid in summarization, they lack transparency, adaptability to new knowledge, and integration with computational tools. To address these challenges, we introduce `Discovera`[1], an agentic system that combines LLMs with established computational bioinformatics pipelines, and retrieval-augmented generation (RAG) to support mechanistic discovery. `Discovera` bridges the gap between computation and interpretation, enabling users to explore hypotheses grounded in data and literature. We demonstrate the utility of `Discovera` in the context of endometrial carcinoma research, where it supports functional enrichment analysis and the summarization of potential mechanisms of action for gene sets associated with an observed phenotype.

1 INTRODUCTION

Advancements in high-throughput technologies, such as omics data analysis, create new opportunities in biomedical research, including the ability to investigate thousands of genes in a single experimental setting. However, they also pose significant challenges in data exploration and knowledge discovery. Critical tasks in this process include exploring, discovering, interpreting, and summarizing the potential mechanisms of action for gene sets correlated with an observed phenotype[2] and proposing new causal relations in the context of a particular research question, e.g., regarding a disease. Traditional tools like GSEA (Subramanian et al., 2005) and ORA (Yu et al., 2012) are foundational in annotating gene lists with curated pathway-level knowledge. While effective at identifying biological processes associated with known gene lists, these methods struggle to reveal novel mechanisms, requiring researchers to manually search for and synthesize insights from disparate sources – a process that is both time-consuming and error-prone. Furthermore, effective application of these methods often requires the ability to code and choose various parameters, making it difficult for biologists who are not familiar with these computational methods to leverage them.

Recent approaches based on large language models (LLMs) have shown promising results in summarizing known and novel gene relationships (Hu et al., 2025). However, despite the broad knowledge offered by LLMs, their abilities are limited by the data seen at training time: their knowledge cannot be easily expanded or revised (Lewis et al., 2020). Moreover, LLMs lack transparency and cannot always provide insight into their outputs (Huang et al., 2024). They are also prone to halluci-

[*]Corresponding authors affiliated to the Center for Data Science at NYU, New York, NY, USA.
[†]Corresponding authors affiliated to the Department of Computer Science at NYU, New York, NY, USA.
[‡]Corresponding authors affiliated to the Department of Biochemistry & Molecular Pharmacology, and the Institute for Systems Genetics at NYU Grossman School of Medicine, New York, NY, USA.

[1]Publicly available at https://github.com/VIDA-NYU/discovera.

[2]These gene sets are often identified through methods such as Independent Component Analysis (ICA), Shapley values, correlations, among others.

nations, producing confident but factually incorrect information (Ji et al., 2023; Maynez et al., 2020) To address these issues, this paper explores an approach that combines traditional omics analysis tools with LLMs reasoning and evidence retrieval from existing scientific literature. Specifically, we introduce `Discovera`, an LLM-based agent system that assists biomedical researchers in uncovering possible mechanisms of action for genes correlated with an observed phenotype. We also present a use case that shows how `Discovera` can be used in endometrial carcinoma (EC) research, illustrating its potential to enhance functional discovery for gene sets.

Our solution aims to empower biomedical researchers to perform data-driven knowledge discovery, regardless of technical expertise, by enabling interactive hypothesis generation and mechanistic insights about complex diseases.

2 Agentic Gene Set Function Discovery and Explanation

To address the challenges of interpretability, scalability, and evidence-grounded reasoning in functional genomics, we propose a modular LLM-based agent system, `Discovera`, that supports biomedical researchers in interactively exploring gene sets and generating mechanistic hypotheses. As illustrated in Figure 1, the system includes a chatbot user interface and integrates tools to perform data analyses and retrieve data from a knowledge base. Next, we detail each component.

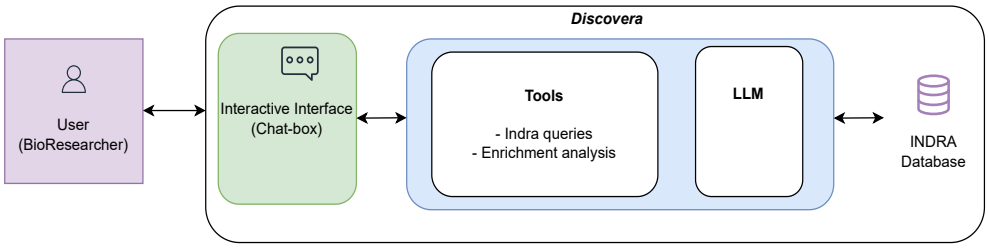

Figure 1: `Discovera`: Architecture for agent-assisted gene function discovery. The agent integrates gene set enrichment tools (e.g., GSEApy), knowledge synthesis via INDRA, and custom analysis modules. A chat interface enables interactive exploration, while structured prompts and tool wrappers guide the agent's reasoning and output generation.

Tools. Our prototype leverages a combination of existing packages and custom functions to compute enrichment statistics, integrate knowledge from various data sources, and distill meaningful insights. For functional gene analysis, we utilize the `GSEApy` package (Fang et al., 2022), which facilitates the calculation of enrichment statistics for both GSEA and ORA across a comprehensive collection of biological pathways. To integrate and synthesize knowledge from diverse sources, including biology-focused natural language processing systems and established databases, we employ `INDRA` (Integrated Network and Dynamical Reasoning Assembler), a knowledge base that integrates multiple cell-to-cell interaction databases through text mining of scientific literature (Gyori et al., 2017; Todorov et al., 2019). Additionally, we have developed custom functions to query specific data from `INDRA`, filter and subset relevant information, and compute supplementary metrics, such as correlations derived from gene expression and mutation data collected from a sample of interest. We provide a list of functions in Figure 4 (Appendix).

Agent. `Discovera` orchestrates and execute the primitives, including those listed in Figure 4, while facilitating seamless interaction with domain experts via a chat-like interface. Implementing an agent requires the definition of structured procedures and precise task descriptions, which are assembled into prompts for LLMs. For our prototype, we developed tool wrappers for each function, ensuring modular execution and a standardized interface. Each wrapper includes metadata specifying inputs, expected outputs, and contextual requirements. We also designed structured procedural guidelines that dictate when the LLM should request user input, incorporate additional context, and format outputs in a consistent and interpretable way. We leveraged `Archytas` (Archytas) an open-source framework for AI agents using the ReAct model (Yao et al., 2023), and integrated `Beaker` (Beaker, n.d.), a context-aware notebook system for chatbot UI development. This setup enables tool invocation, interactive markdown rendering, agent reasoning visualization, and seamless user feedback integration.

2.1 USE CASE: SUPPORTING ENDOMETRIAL CANCER FUNCTIONAL ANALYSIS

To demonstrate `Discovera` 's capabilities in advancing genomic research, we use a dataset from Dou et al. (2020) containing proteogenomic data and mutation status from an endometrial carcinoma cohort. This use case illustrates how the system supports users in performing enrichment analysis and exploring biological mechanisms beyond known pathways. By integrating dynamic code generation, up-to-date literature, and pathway enrichment, `Discovera` helps users uncover insights into potential drug targets and intervention pathways. Below, we detail a representative interaction between the user (👤) and the system (🤖), showcasing how the agent reasons through each analytical step and adapts to the user's goals.

In the first interaction, the user instructs `Discovera` to load a dataset containing a ranked list of genes, ordered by their correlation[3] with *beta-catenin mutation status*. `Discovera` then generates and executes code to load the data. Next, the user directs the system to identify the most statistically significant *biological pathway* using the built-in enrichment tool, `run_gsea()`. The ouptput is shown in Figure 2. Note that this analysis typically requires a ranked list of genes and known pathways (a collection of gene sets) as input. If no gene sets are specified, the system defaults to a curated collection including KEGG 2016, GO Biological Process 2023, Reactome Pathways 2024, and MSigDB Hallmark 2020.

The agent demonstrates *proactive task execution* by dynamically generating the code required to load the data and visualize the results, even when no predefined functions are available for those steps. This *on-demand code generation* enables the system to adapt fluidly to user input, reducing the need for manual intervention. By combining flexible code synthesis with access to specialized analysis tools like `run_gsea()`, the system offers both *usability* and *adaptability*. It supports efficient, reproducible workflows while lowering the technical barrier for users, allowing complex analyses to be executed with minimal configuration.

🤖 I will perform Gene Set Enrichment Analysis (GSEA) using the loaded dataset. Since the user did not specify a gene set, I will use a predefined list of gene sets.
Result:

Pathway/Process	ES	NES	NOM p-val	FDR q-val	Lead Genes
Negative Regulation Of Wnt Signaling Pathway (GO:0030178)	0.730867	2.3146	0	0.00657817	LRP4, NOTUM, APCDD1, DKK4, CTNNBIP1, WIF1, CSNK1A1
Wnt Signaling Pathway (KEGG_2016)	0.675749	2.19035	0	0.0202827	NOTUM, CTNNB1, DKK4, LEF1, CTNNBIP1, WIF1, CSNK1A1, MMP7, PLCB4

[···]

The key genes from the most significant pathway, **"Negative Regulation of Wnt Signaling Pathway"** (GO:0030178), are:

LRP4, NOTUM, APCDD1, DKK4, CTNNBIP1,WIF1, CSNK1A1

Figure 2: Top enriched pathways identified by GSEA using KEGG, GO, Reactome, and MSigDB gene sets.

As noted in Section 1, enrichment analysis tools like GSEA annotate gene lists using curated pathway databases (e.g., "Negative Regulation of Wnt Signaling Pathway"; see Figure 2), highlighting leading genes that drive known biological signals. However, such methods are limited to existing pathway knowledge. To go beyond this and explore novel mechanisms, `Discovera` enables users to examine the top correlated genes, 15 in this case, by retrieving relationships documented in up-to-date scientific literature. This is done using `INDRA` via `get_bulk_pairs()`, which returns known gene-gene interactions, the supporting statements, and source links. This helps users uncover connections not yet captured in curated pathways. The results are illustrated in Figure 3.

Finally, `Discovera` empowers the user to synthesize all retrieved information, including knowledge graphs, curated pathways, and literature-backed gene relationships, into well, supported hypotheses about potential drug targets or intervention pathways. Through interaction with the integrated LLM, the system distills complex, multi-source data into coherent, testable insights. Crucially, each hypothesis is grounded in excerpts from peer-reviewed scientific literature, allowing users to inspect the original evidence, assess credibility, and plan targeted follow-up analyses.

[3]Other ranking metrics, such as Independent Component Analysis (ICA) scores or log2 fold change, can also be used depending on the analysis goals.

The GSEA results highlighted the 'Negative Regulation Of Wnt Signaling Pathway' as a significant pathway. The Wnt signaling pathway is known to play a crucial role in cell proliferation, differentiation, and migration, which are key processes in cancer development, including endometrial carcinoma.

- **'CTNNB1', 'AMOT'**
  - Excerpts: AMOT phosphorylation promotes neurogenesis by inhibiting YAP and promoting beta-catenin activity. AMOT proteins sequester YAP independent of TNKS1/2 containing beta-catenin degradasomes.
  - Analysis: The interaction between AMOT and beta-catenin may influence the Hippo signaling pathway, which is involved in regulating cell growth and apoptosis. Dysregulation of this pathway can lead to uncontrolled cell proliferation, a hallmark of cancer.

[···]
The integration of GSEA results, known pathway involvement, and documented gene relationships suggest that these gene interactions play significant roles in processes related to endometrial carcinoma, such as cell proliferation, adhesion, and EMT. Targeting these pathways could provide insights into potential therapeutic strategies for managing endometrial carcinoma.

Figure 3: Integration of RAG, LLM Reasoning, and Biomedical Tool Outputs.

Discussion. To assess the reliability of our agent's predictions and mitigate the risk of hallucinated relationships, we conducted an initial evaluation using a random sample of 15 genes. First, we queried `INDRA` to extract documented gene-gene interactions based on curated scientific literature. Among the selected gene pairs, only one exhibited a previously established relationship. Subsequently, we posed the same queries to our agent, specifically investigating whether it could infer potential associations with endometrial carcinoma. If the agent generated a response, we further prompted it to evaluate the plausibility of its own prediction. In all cases, the agent assigned a confidence score of 0.6 or lower. In contrast, for previously validated interactions, confidence scores were consistently above 0.7.

3 Conclusion & Future Work

For future work, we aim to enhance the usability and interpretability of our system by integrating additional features that improve access to relevant scientific literature. A key direction is retrieving paper abstracts and, where possible, full-text content to provide richer contextual information for each documented interaction. This will enable users to better understand the source and reliability of the extracted relationships. Given the inherent limitations of context windows in language models, we will explore advanced summarization techniques to condense relevant information while preserving key details. Additionally, interactive querying mechanisms will be developed to allow users to dynamically retrieve and refine information, facilitating a more comprehensive and user-driven exploration of gene interactions and supporting evidence.

Finally, a crucial aspect is the development of a rigorous evaluation protocol to assess the accuracy, reliability, and scientific value of the system's outputs. Ensuring the truthfulness and insightfulness of extracted relationships is essential for their utility in biomedical research. To this end, we plan to extend and test the usability of our system across diverse biological and clinical use cases, allowing for a more comprehensive assessment of its applicability. Additionally, we will engage with domain experts, including biomedical researchers and computational biologists, to gather qualitative and quantitative feedback on the `Discovera` 's effectiveness. Their insights will be instrumental in defining appropriate evaluation metrics, such as precision, recall, and domain relevance, as well as identifying potential biases or limitations in the generated outputs.

Acknowledgments

This work was supported by the DARPA Automating Scientific Knowledge Extraction and Modeling (ASKEM) program, Agreement No. HR0011262087, and ARPA-H. Freire was partially supported by NSF awards CMMI-2146306, IIS-2106888, and OAC-2411221. Pinto acknowledges support through the NSF under award 1922658. The views and findings expressed are those of the authors and should not be interpreted as representing the official views or policies of the DARPA, ARPA-H, NSF, or the U.S. Government. NSF awards CMMI-2146306, CMMI-2146312.

REFERENCES

Archytas. Archytas: A tools interface for ai agents. https://github.com/jataware/archytas.

Beaker. Beaker-kernel: Contextually-aware notebooks with built-in ai assistant, n.d. URL https://github.com/jataware/beaker-kernel. Accessed: 2025-02-23.

Y. Dou, E. A. Kawaler, D. Cui Zhou, M. A. Gritsenko, C. Huang, L. Blumenberg, A. Karpova, V. A. Petyuk, S. R. Savage, S. Satpathy, W. Liu, Y. Wu, C. F. Tsai, B. Wen, Z. Li, S. Cao, J. Moon, Z. Shi, M. Cornwell, M. A. Wyczalkowski, R. K. Chu, S. Vasaikar, H. Zhou, Q. Gao, R. J. Moore, K. Li, S. Sethuraman, M. E. Monroe, R. Zhao, D. Heiman, K. Krug, K. Clauser, R. Kothadia, Y. Maruvka, A. R. Pico, A. E. Oliphant, E. L. Hoskins, S. L. Pugh, S. J. I. Beecroft, D. W. Adams, J. C. Jarman, A. Kong, H. Y. Chang, B. Reva, Y. Liao, D. Rykunov, A. Colaprico, X. S. Chen, A. Czekański, M. Jędryka, R. Matkowski, M. Wiznerowicz, T. Hiltke, E. Boja, C. R. Kinsinger, M. Mesri, A. I. Robles, H. Rodriguez, D. Mutch, K. Fuh, M. J. Ellis, D. DeLair, M. Thiagarajan, D. R. Mani, G. Getz, M. Noble, A. I. Nesvizhskii, P. Wang, M. L. Anderson, D. A. Levine, R. D. Smith, S. H. Payne, K. V. Ruggles, K. D. Rodland, L. Ding, B. Zhang, T. Liu, D. Fenyö, and Clinical Proteomic Tumor Analysis Consortium. Proteogenomic characterization of endometrial carcinoma. *Cell*, 180(4):729–748.e26, 2020. doi: 10.1016/j.cell.2020.01.026. URL https://doi.org/10.1016/j.cell.2020.01.026. Epub 2020 Feb 13.

Zhuoqing Fang, Xinyuan Liu, and Gary Peltz. Gseapy: a comprehensive package for performing gene set enrichment analysis in python. *Bioinformatics*, 38(10):3005–3007, 2022. doi: 10.1093/bioinformatics/btac757. URL https://doi.org/10.1093/bioinformatics/btac757.

Benjamin M Gyori, John A Bachman, Kartik Subramanian, Jeremy L Muhlich, Lucian Galescu, and Peter K Sorger. From word models to executable models of signaling networks using automated assembly. *Molecular Systems Biology*, 13(11):954, Nov 2017. doi: 10.15252/msb.20177651. URL https://doi.org/10.15252/msb.20177651. Published under the terms of the CC BY 4.0 license.

Mengzhou Hu, Sahar Alkhairy, Ingoo Lee, Rudolf T. Pillich, Dylan Fong, Kevin Smith, Robin Bachelder, Trey Ideker, and Dexter Pratt. Evaluation of large language models for discovery of gene set function. *Nature Methods*, 22(1):82–91, 2025. ISSN 1548-7105. doi: 10.1038/s41592-024-02525-x. URL https://doi.org/10.1038/s41592-024-02525-x.

Yue Huang, Lichao Sun, Haoran Wang, Siyuan Wu, Qihui Zhang, Yuan Li, Chujie Gao, Yixin Huang, Wenhan Lyu, Yixuan Zhang, Xiner Li, Hanchi Sun, Zhengliang Liu, Yixin Liu, Yijue Wang, Zhikun Zhang, Bertie Vidgen, Bhavya Kailkhura, Caiming Xiong, Chaowei Xiao, Chunyuan Li, Eric P. Xing, Furong Huang, Hao Liu, Heng Ji, Hongyi Wang, Huan Zhang, Huaxiu Yao, Manolis Kellis, Marinka Zitnik, Meng Jiang, Mohit Bansal, James Zou, Jian Pei, Jian Liu, Jianfeng Gao, Jiawei Han, Jieyu Zhao, Jiliang Tang, Jindong Wang, Joaquin Vanschoren, John Mitchell, Kai Shu, Kaidi Xu, Kai-Wei Chang, Lifang He, Lifu Huang, Michael Backes, Neil Zhenqiang Gong, Philip S. Yu, Pin-Yu Chen, Quanquan Gu, Ran Xu, Rex Ying, Shuiwang Ji, Suman Jana, Tianlong Chen, Tianming Liu, Tianyi Zhou, William Yang Wang, Xiang Li, Xiangliang Zhang, Xiao Wang, Xing Xie, Xun Chen, Xuyu Wang, Yan Liu, Yanfang Ye, Yinzhi Cao, Yong Chen, and Yue Zhao. Position: TrustLLM: Trustworthiness in large language models. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 20166–20270. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/huang24x.html.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Comput. Surv.*, 55(12), March 2023. ISSN 0360-0300. doi: 10.1145/3571730. URL https://doi.org/10.1145/3571730.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33: 9459–9474, 2020.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. On faithfulness and factuality in abstractive summarization. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 1906–1919, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.173. URL https://aclanthology.org/2020.acl-main.173/.

A. Subramanian, P. Tamayo, V.K. Mootha, S. Mukherjee, B.L. Ebert, M.A. Gillette, A. Paulovich, S.L. Pomeroy, T.R. Golub, E.S. Lander, and J.P. Mesirov. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*, 102(43):15545–15550, 2005. doi: 10.1073/pnas.0506580102. URL https://doi.org/10.1073/pnas.0506580102.

P. V. Todorov, B. M. Gyori, J. A. Bachman, and P. K. Sorger. Indra-ipm: interactive pathway modeling using natural language with automated assembly. *Bioinformatics*, 35(21):4501–4503, Nov 2019. doi: 10.1093/bioinformatics/btz289.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models, 2023. URL https://arxiv.org/abs/2210.03629.

Guangchuang Yu, Li-Gen Wang, Yanyan Han, and Qing-Yu He. clusterprofiler: an r package for comparing biological themes among gene clusters. *Omics: a journal of integrative biology*, 16 (5):284–287, 2012.

A Appendix

`query_bulk_pairs``(genes)`
Queries INDRA API for multiple source-target relationships..

`excerpts``(genes)`
Extract specific excerpts where there are documented relationships.

`rank_gsea``(source, gene_sets, gene_col, rank_col, min_size, max_size)`
Performs gene set enrichment analysis to find statistically significant gene sets enriched in a dataset.

`get_correlations``(proteomics, mutation)`
Give proteomics and mutation data, get correlation information that will serve as ranked list

`relationships``(genes)`
Summarize the types and frequencies of relationships documented between gene pairs.

Figure 4: Predefined selected toolset.