An LLM-based Knowledge Synthesis and Scientific Reasoning Framework for Biomedical Discovery

Oskar Wysocki[1,2], Magdalena Wysocka[2], Danilo S. Carvalho[2], Alex Bogatu[2], Danilo Gusicuma[1], Maxime Delmas[1], Harriet Unsworth[2], André Freitas[1,2,3]

[1]Idiap Research Institute, Switzerland
[2]National Biomarker Centre, CRUK-MI, Univ. of Manchester, United Kingdom
[3]Department of Computer Science, Univ. of Manchester, United Kingdom

Correspondence: firstname.lastname@idiap.ch[1] firstname.lastname@manchester.ac.uk[2]

Abstract

We present BioLunar, developed using the Lunar framework, as a tool for supporting biological analyses, with a particular emphasis on molecular-level evidence enrichment for biomarker discovery in oncology. The platform integrates Large Language Models (LLMs) to facilitate complex scientific reasoning across distributed evidence spaces, enhancing the capability for harmonizing and reasoning over heterogeneous data sources. Demonstrating its utility in cancer research, BioLunar leverages modular design, reusable data access and data analysis components, and a low-code user interface, enabling researchers of all programming levels to construct LLM-enabled scientific workflows. By facilitating automatic scientific discovery and inference from heterogeneous evidence, BioLunar exemplifies the potential of the integration between LLMs, specialised databases and biomedical tools to support expert-level knowledge synthesis and discovery.

1 Introduction

Contemporary biomedical discovery represents a prototypical instance of complex scientific reasoning, which requires the coordination of controlled *in-vivo*/*in-silico* interventions, complex multi-step data analysis pipelines and the interpretation of the results under the light of previous evidence (available in different curated databases and in the literature) (Paananen and Fortino, 2019; Nicholson and Greene, 2020). This intricacy emerges out of the inherent complexity of biological mechanisms underlying organism responses, which are defined by a network of multi-scale inter-dependencies (Bogdan et al., 2021). While more granular data is being generated by the evolution of instruments, assays and methods, and the parallel abundance of experimental interventions (Dryden-Palmer et al., 2020), there a practical barrier for integrating and cohering this evidence space into a specific context of analysis.

Within biomedical discovery, the language interpretation capabilities of Large Language Models (LLMs) can provide an integrative framework for harmonising and reasoning over distributed evidence spaces and tools, systematising and lowering the barriers to access and reason over multiple structured databases, textual bases such as PubMed, enriching the background knowledge through specialised ontologies and serving as interfaces to external analytical tools (e.g. mechanistic/perturbation models, gene enrichment models, etc). In this context, LLMs can serve as a linguistic analytical layer which can reduce the syntactic impedance across diverse functional components: once an adapter to an external component is built it can be integrated and reused in different contexts, creating a monotonic increase of functional components. Complementarily, from a Biomedical-NLP perspective, in order to address real-world problems, LLMs need to be complemented with mechanisms which can deliver contextual control (e.g. via Retrieval Augmented Generation: RAG: access the relevant background knowledge and facts) and perform the analytical tasks which are integral to contemporary biomedical inference ('toolforming').

Emerging LLM-focused coordination frameworks such as LangChain[1], Flowise[2] and Lunar[3] provide the capabilities to deliver a composition of functional components, some of them under a low-code/no-code use environment, using the abstraction of workflows. While there are general-purpose coordination frameworks, there is a lack of specialised components for addressing biomedical analyses.

In this paper we demonstrate BioLunar, a suite of components developed over the Lunar environment

[1] https://python.langchain.com
[2] https://github.com/FlowiseAI/Flowise
[3] https://lunarbase.ai

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations), pages 355–364

to support biological analyses. We demonstrate the key functionalities of the platform contextualised within a real-use case in the context of molecular-level evidence enrichment for biomarker discovery in oncology.

2 BioLunar

BioLunar enables the creation of LLM-based biomedical scientific workflows using software components with standardised APIs. A workflow is composed of components and subworkflows connected through input-output relationships, and are capable of handling multiple inputs. In the user interface, components are clustered according to their function (see Fig.1). Creating a workflow does not require programming knowledge since components are predefined and merely require data inputs or parameter settings. However, for users who wish to write custom code, 'Python Coder' and 'R Coder' components are provided, enabling the definition of custom methods. These custom components can be saved and subsequently accessed in the 'Custom' group tab.

In the paper we describe an exemplar *biomedical workflow* designed to integrate evidence and infer conclusions from bioinformatics pipeline results. Specifically, the *biomedical workflow* queries expert knowledge bases (KBs) that continuously compile clinical, experimental, and population genetic study outcomes, aligning them with assertions relevant to the significance of the observed gene or variant. It then employs Natural Language Inference (NLI) (via LLM) to integrate and harmonise the evidence space and interpreting the results, culminating in a comprehensive summary for the entire gene set input. This interpretation takes into account the bioanalytical context supplied by the user.

2.1 Exemplar Workflow

Next-generation sequencing (NGS) assays play a pivotal role in the precise characterisation of tumours and patients in experimental cancer treatments. NGS findings are essential to guide the design of novel biomarkers and cancer treatments. Nevertheless, the clinical elucidation of NGS findings subsequent to initial bioinformatics analysis often requires time-consuming manual analysis procedures which are vulnerable to errors. The interpretation of molecular signatures that are typically yielded by genome-scale experiments are often supported by pathway-centric approaches through which mechanistic insights can be gained by pointing at a set of biological processes. Moreover, gene and variant enrichment benefits from heterogeneous curated data sources which pose challenges to seamless integration. Furthermore, there are different levels of supporting evidence and therefore prioritising conclusions is crucial. Automating evidence interpretation, knowledge synthesis and leveraging evidence-rich gene set reports are fundamental for addressing the challenges in precision oncology and the discovery of new biomarkers.

2.2 User interface

The user interface facilitates an agile workflow construction by enabling users to select and arrange components via drag-and-drop from functionally grouped categories, such as, i.a.: 'Prompt Query' featuring NLI components, 'Knowledge Bases' components, 'Extractors' for retrieving files from zip archives or extracting text and tables from PDF files, and 'Coders', which allow for the creation of custom components using Python or R scripts.

Components allow for individual execution, edition, or configuration adjustment via a visual interface. Workflows can be executed, saved, or shared. Each component has designated input and output capabilities, enabling seamless integration where the output from one can directly feed into another. Users have the flexibility to manually input values if no direct connection is established. Additionally, a component's output can feed into multiple components. The system's architecture supports effortless expansion, adding branches and components without affecting the existing workflow, thus facilitating scalable customization to meet changing requirements. The user interface with an example of a workflow is presented in Fig.1 and in demo video https://youtu.be/Hc6pAA_5Xu8.

2.3 Knowledge bases

The current framework integrates a diverse set of knowledge bases which are relevant for precision oncology. To identify gene mutations as biomarkers for cancer diagnosis, prognosis, and drug response, we integrated CIViC[4] and OncoKB[5]. CIViC provides molecular profiles (MPs) of genes, each linked to clinical evidence, with

[4] https://civicdb.org
[5] https://www.oncokb.org

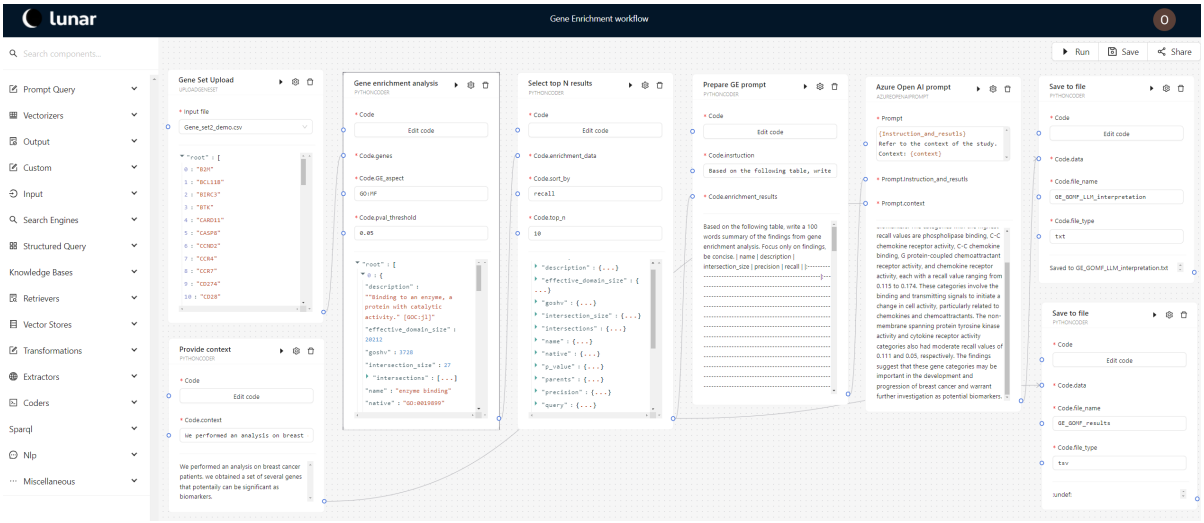


Figure 1: BioLunar interface. An exemplary workflow of Gene Enrichment with an input gene set, knowledge base query and LLM interpretation components.

a molecular score indicating evidence quality, assessed by human annotators. The Gene Ontology[6] (GO) offered gene function insights, and the Human Protein Atlas[7] supplied a list of potential drug targets and transcription factors. We employed COSMIC[8] for somatic mutation impacts in cancer, the largest resource in this field. Our analysis also included KEGG[9], Reactome[10], and WikiPathways[11] for pathway information, enriching our investigation with scientific literature via PubMed's API [12].

In the following subsections, we showcase examples of components, subworkflows, and workflows constructed using the BioLunar framework, motivated by the biomarker discovery/precision oncology themes.

2.4 Construction and reuse of specialised prompts

BioLunar employs standard LLM interfaces, allowing the use of different models according to users' preferences. The prompt components allows for the composition of specialised prompt chains which can be later reused, defining a pragmatic pathway for specialised Natural Language Inference (NLI) via prompt decomposition/composition. This approach allows for the creation of reasoning chains that combines user's instructions with the results of database queries and analyses from specialised tools within the context of the study. An instantiated example of the *Azure Open AI prompt* is described in Fig.1.

2.5 Subworkflow component

The *subworkflow* component enables the reuse of an existing workflow within another workflow, functioning as a component with specified inputs and outputs. This feature simplifies the composition of more complex workflows and avoids the repetition of defining identical steps for the same task. Subworkflows can be selected like other components from the left panel in the interface, offering access to all available workflows for easy integration. Examples of subworkflows are presented in Fig.2,3.

2.6 Gene Enrichment subworkflow

One example of a specialised subworkflow is the *Gene Enrichment subworkflow* (Fig.1,2A) begins with uploading the targeted gene sets. Then a component accesses a specific KB — such as Gene Ontology, KEGG, Reactome, or WikiPathways—defined by the user, using *gprofiler* API[13]. This component identifies gene groups with a statistically significant overlap with the input gene set, according to a Fisher's test, and calculates p-values, recall, and precision. The user then specifies a variable to rank these groups and selects the top N for further analysis. The output includes both a inter-

[6] https://geneontology.org
[7] https://www.proteinatlas.org
[8] https://cancer.sanger.ac.uk/cosmic
[9] https://www.kegg.jp/kegg/
[10] https://reactome.org
[11] https://www.wikipathways.org
[12] https://pubmed.ncbi.nlm.nih.gov
[13] https://biit.cs.ut.ee/gprofiler/page/apis

pretation performed by an NLI component (through LLM) and a table featuring the names, descriptions, and statistics of the top N selected groups.

2.7 Human Protein Atlas subworkflow

In the *Human Protein Atlas subworkflow*, given a gene set, an associated external KB is queried by selecting 'Transcription factors' from the HPA database using a dedicated query-database connector. A reusable component, 'Analyze overlap', then identifies genes that overlap and calculates relevant statistics. Similarly to the *Gene Enrichment subworkflow*, the results are interpreted by an prompt-based NLI component and presented alongside a table summarising the findings (Fig.2B,A.7).

2.8 CIVIC subworkflow

This *subworkflow* exemplifies a more complex composition of components (Fig.3). This subworkflow initiates by querying the CIVIC database for input genes, yielding, among other things, gene descriptions in clinical contexts, and their variants and molecular profiles (MPs), which are essential for the final interpretation. Additionally, users specify the analysis context, including aspects such as cancer types or subtypes, treatments, populations, etc. Initially, gene descriptions are analysed by a prompt-based NLI component within this defined context. Subsequently, MPs scored below a predefined threshold (set at a MP score of 10) are tagged as *less known*, reflecting lower scientific evidence and ranking by CIVIC annotators. The evidence supporting these lesser-known MPs is then interpreted by a prompt-based NLI component, considering the broader analysis context. Conversely, evidence from *well-known* MPs, scoring above 10, undergoes a similar interpretation process.

For genes without identified MPs in CIVIC, a sequence of components perform further evidence retrieval from PubMed. An NLI module generates context-based keywords for PubMed queries, which are combined with the names of genes lacking MPs. A 'PubMed search' component then retrieves N publications, including metadata, citation counts and MeSH terms (used later for context alignment validation). The abstracts of these publications are interpreted by an NLI module in the context of the analysis.

All clinical evidence interpretations are then succinctly summarised by via a prompt component, taking into account the context of the analysis. These interpretations, along with tabular results, constitute the output.

2.9 Bioworkflow - comprehensive analysis for a set of genes.

The exemplar *bioworkflow* composes multiple subworkflows (Fig.4), each dedicated to a specific multi-step and specialised task, which are typically defined by the composition of heterogeneous components, most commonly connectors and query instance components to specialised databases (e.g. CIVIC, HPA, PubMed, OncoKB), external specialised analytical tools (toolformers for gene enrichment analysis) and chains of specialised interpretation prompts (e.g. selection, filtering, extraction, summarisation). This setup forms a comprehensive workflow which exemplifies the close dialogue between LLMs and genomic analysis, encompassing gene enrichment, comparison with reference gene sets, and access to evidence within

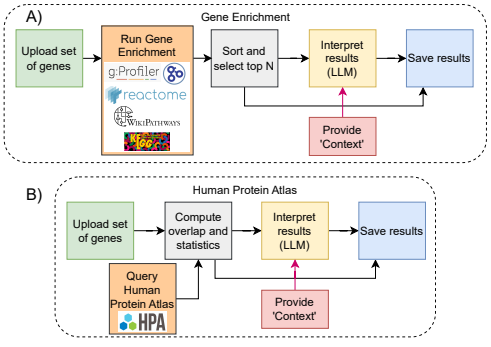


Figure 2: A) Gene Enrichment workflow - uses the *gprofiler* API to access i.a. Gene Ontology, KEGG, WikiPathways, Reactome; B) Human Protein Atlas workflow. Compares and interprets the input and reference gene sets.

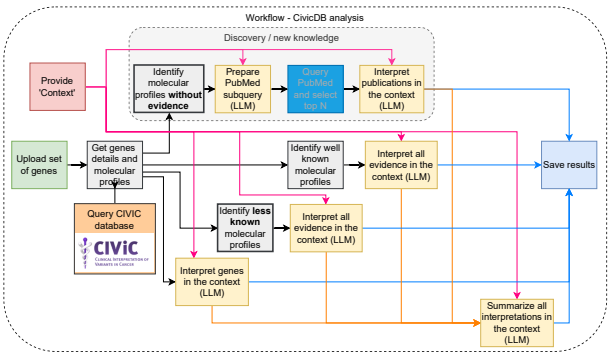


Figure 3: CIVIC evidence analysis workflow. prompt-based NLI components are fed by both the results and context of the analysis in order to produce relevant evidence-based conclusions.

an experimental medicine setting. Additionally, it queries PubMed publications within the CIVIC component to seek evidence for molecular profiles not yet described. Its componentised architecture facilitates the extensibility of the workflow with new sources, prompts and external tools. Conclusions drawn from each subworkflow are interpreted within the analysis context, being integrated in a comprehensive summary. All findings are compiled in a report, exported as a PDF file.

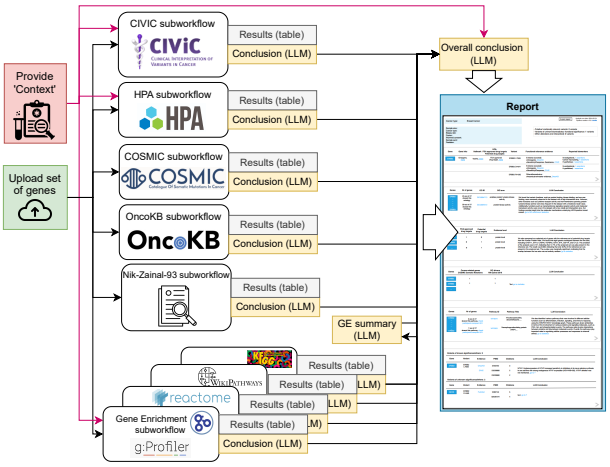


Figure 4: Diagram of the Bioworkflow.

2.10 Software description

BioLunar uses the *LunarVerse* backend for its operations. LunarVerse is downloaded and installed by the setup script included with the demonstration code. Some of its components need user specific configuration to work, such as private API keys, which are defined in a configuration file indicated in the setup instructions. LunarVerse is distributed under a *open software* license. The workflow can also be operated via a graphical interface (*LunarFlow*)

Running a workflow can be done in two ways: i) directly, by calling the LunarVerse engine on a specified workflow descriptor file; ii) through the Web interface, by pressing the “Run” button.

The first way is the default one in the demonstration code. It returns a copy of the workflow descriptor, with all component output fields filled, which is then used to extract and filter the desired outputs, based on the component labels. It is also the best way to automate multiple workflow runs and to integrate their outputs into other systems.The supporting code is available at `https://github.com/neuro-symbolic-ai/lunar-bioverse-demo`.

2.11 Report

The *Bioworkflow*, as outlined in point 2.9, generates a report in PDF (Fig.5) format that begins by outlining the context of the study, analysis details, dates, and software versions at the top. The report is enhanced with hyperlinks for easy navigation to specific sections.

A "General Statistics" table provides a comprehensive overview of key metrics aggregated from all components, aiming to consolidate information for each gene throughout the analysis, with hyperlinks directing to the report sections where this information originates.

Subsequent sections categorise genes into various tables based on biological aspects and the KBs consulted. These include Molecular Function for genes sharing ontologies, drug target checks based on the Human Protein Atlas, assessments of cancer-related genes, Pathway Analysis and Mapping via WikiPathways, and classification of gene alterations by clinical relevance. By correlating genes with known functional information, the workflow identifies statistically significant enriched terms and summarizes these findings using LLM, which also furnishes evidence.

LLM interprets each table, offering textual conclusions relevant to the analysis context. A final summary, crafted using LLM, synthesizes all results within the given context. Importantly, all LLM interpretations are grounded in concrete evidence, with sources cited alongside the narrative. This approach underscores the rigor of the analysis by highlighting distinct sources that substantiate the relevance of each gene and variant.

3 Case study

To demonstrate the capabilities of the *Bioworkflow*, we analyzed outputs in two different scenarios, each producing a distinct set of genes from separate bioinformatics analyses. We entered these gene sets along with their analysis contexts into the *Bioworkflow* and executed it. Subsequently, we qualitatively assessed the output reports (see Fig.A.8,A.9), considering both the statistical data and the interpretations provided by the prompt-based NLI modules.

In Scenario 1, the user aims to explore the unique molecular characteristics of HER2-low breast cancer to determine if it constitutes a distinct category within breast cancer types, where the input genes are ERBB2, ESR1, PIK3CA, CBFB, SF3B. The

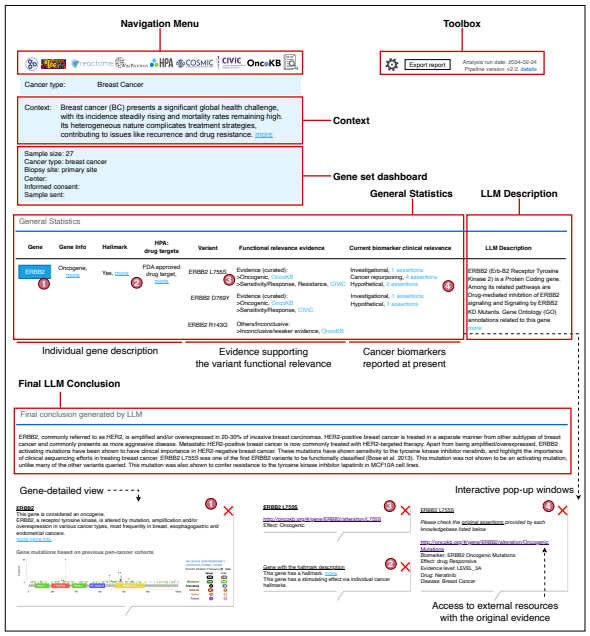


Figure 5: The BioLunar report's overview, produced by *Bioworkflow*.

report shows genomic alterations and genomic signatures that were identified, including ERBB2 amplification, mutations in PIK3CA and ESR1, which are important biomarkers in the selection of breast cancer treatment. For the remaining two genes, evidence was found confirming that these are new, significantly mutated genes for which there is preclinical evidence of actionability in clinical practice.

In Scenario 2, the user aims to discover new genes that could lead to more accurate breast cancer diagnoses, enhancing treatment strategies and addressing the disease's complexity. His numerical analysis resulted in a set of genes (DIXDC1, DUSP6, PDK4, CXCL12, IRF7, ITGA7, NEK2, NR3C1) that require investigation. The report informs that none of the genes is an oncogene (confirmation according to OncoKB), two of the genes are potential drug targets and one is FDA approved drug targets. According to the KEGG-based enrichment analysis, these genes were mainly enriched through several signaling pathways including tumor necrosis factor (TNF) signaling pathway. Using LLMs in conjunction with a PubMed search component, papers were searched in PubMed that describe various gene variants and the genes have been indicated as prospective biomarkers associated with breast cancer.

Note that in scenario 2, for genes lacking molecular profiles in the KB, a search in PubMed was conducted. This approach enables the workflow to automatically uncover and search for non-obvious and previously unknown relationships. Essentially, if a gene is absent from the database, it suggests that its relevance is relatively novel and not yet documented. Therefore, seeking out the most recent publications that describe this gene within the analysis context represents a significant advantage, provided by the workflow that integrates various components.

4 Related Work

Bioinformatics Pipelines Over the past decade, three scientific workflow management systems such as Galaxy (gal, 2022), Snakemake (Köster and Rahmann, 2012), and Nextflow (Di Tommaso et al., 2017), have been instrumental to bioinformaticians to systematise their complex analytical processes. Nextflow targets bioinformaticians and facilitates gene enrichment analysis, annotate biological sequences, and perform gene expression analysis by including modules supported by various bioinformatics tools. These workflow systems are currently centred around the composition of specialised bioinformatics software, configuration parameters and supporting datasets, facilitating reuse and reproducibility. In contrast, this paper explores the concept on using LLMs within a specialised workflow environment to support the interpretation and integration of multiple analytical processes.

5 Conclusion

In this paper we provided a demonstration of a scientific workflow based on LLMs to support specialised gene analyses using oncology and gene enrichment as a driving motivational scenario. The framework is built using the Lunar framework and allows for the composition of specialised analytical workflows, integrating external databases (Retrieval Augmented Generation), external tools (ToolFormers) and contextualised chains of LLM-based interpretation. The paper highlights that a workflow environment with specialised components for RAG, ToolFormers and a set of specialised prompts-based Natural Language Inference can serve as the foundation for streamlining and automating complex analytical process within a biomedical setting. . We showcase analytical applications within the biomedical domain, particularly in oncology, constructively progressing towards more complex gene analysis workflows. The

developed *bioworkflow* demonstrates the LLMs can be instrumental in enabling a complex end-to-end highly-specialised analytical workflow, in a reproducible manner, supporting the integration of heterogeneous evidence, synthesising conclusions and while simultaneously documenting and linking to the data sources within a comprehensive output report. The proposed workflow is based on a low-code paradigm that enables domain experts, regardless of their programming skills, to construct and scientific workflows enabled by generaqtive AI amethods.

Limitations

- The current demonstration uses external LLM-based APIs but can be adapted to open source LLM models.
- The LLM-based inferences require a critical supporting quantitative evaluation and hallucinations are possible. The current workflow is motivated by a hypothesis generation process, which is fully human supervised and does not have direct clinical applications.

Acknowledgements

This work was partially funded by The Ark foundation, by the European Union's Horizon 2020 research and innovation program (grant no. 965397) through the Cancer Core Europe DART project, and by the Manchester Experimental Cancer Medicine Centre and the NIHR Manchester Biomedical Research Centre.

References

2022. The galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2022 update. *Nucleic Acids Research*, 50(W1):W345–W351.

Paul Bogdan, Gustavo Caetano-Anollés, Anna Jolles, Hyunju Kim, James Morris, Cheryl A Murphy, Catherine Royer, Edward H Snell, Adam Steinbrenner, and Nicholas Strausfeld. 2021. Biological Networks across Scales—The Theoretical and Empirical Foundations for Time-Varying Complex Networks that Connect Structure and Function across Levels of Biological Organization. *Integrative and Comparative Biology*, 61(6):1991–2010.

Paolo Di Tommaso, Maria Chatzou, Evan W Floden, Pablo Prieto Barja, Emilio Palumbo, and Cedric Notredame. 2017. Nextflow enables reproducible computational workflows. *Nature biotechnology*, 35(4):316–319.

K.D. Dryden-Palmer, C.S. Parshuram, and W.B. Berta. 2020. Context, complexity and process in the implementation of evidence-based innovation: a realist informed review. *BMC Health Services Research*, 20(81):1472–6963.

Johannes Köster and Sven Rahmann. 2012. Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics*, 28(19):2520–2522.

David N. Nicholson and Casey S. Greene. 2020. Constructing knowledge graphs and their biomedical applications. *Computational and Structural Biotechnology Journal*, 18:1414–1428.

Jussi Paananen and Vittorio Fortino. 2019. An omics perspective on drug target discovery platforms. *Briefings in Bioinformatics*, 21(6):1937–1953.

A Appendix

Scenario 1

Context:
The analysis focuses on HER2-low breast cancer (HLBC), a subtype that challenges traditional classifications based on HER2 expression and ERBB2 amplification. Despite being operationally defined, HLBCs constitute a significant portion of breast cancers, particularly among estrogen receptor-positive tumors. This study aims to elucidate the molecular characteristics of HLBCs, examining their mutational and transcriptional profiles. The research also investigates potential heterogeneity within HLBCs and compares their genomic landscape with HER2-positive and HER2-negative breast cancers. By providing insights into the distinct molecular features of HLBCs, this analysis seeks to establish whether they represent a unique entity in breast cancer pathology.

List of genes:
ERBB2, ESR1, PIK3CA, CBFB, SF3B1

Scenario 2

Context:
Breast cancer (BC) presents a significant global health challenge, with its incidence steadily rising and mortality rates remaining high. Its heterogeneous nature complicates treatment strategies, contributing to issues like recurrence and drug resistance. Biomarkers play a crucial role in diagnosing and managing BC, aiding in personalized treatment approaches. However, existing biomarkers have limitations, necessitating the exploration of novel markers, particularly in the realm of molecular and genetic analysis. This study focuses on identifying genes with potential diagnostic utility in breast cancer, aiming to contribute to the development of more effective biomarkers and therapies, including immunotherapies, to combat this disease.

List of genes:
DIXDC1, DUSP6, PDK4, CXCL12, IRF7, ITGA7, NEK2, NR3C1

Figure A.6: User-defined context of the analysis, including aspects like cancer types or subtypes, treatments, populations, for Scenario 1 and 2.

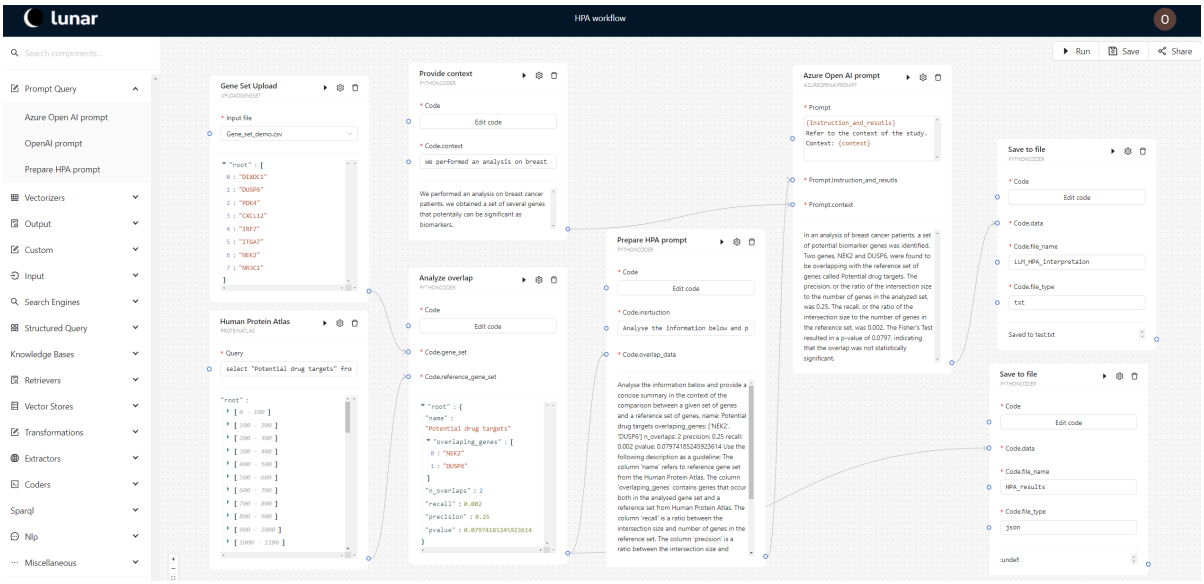
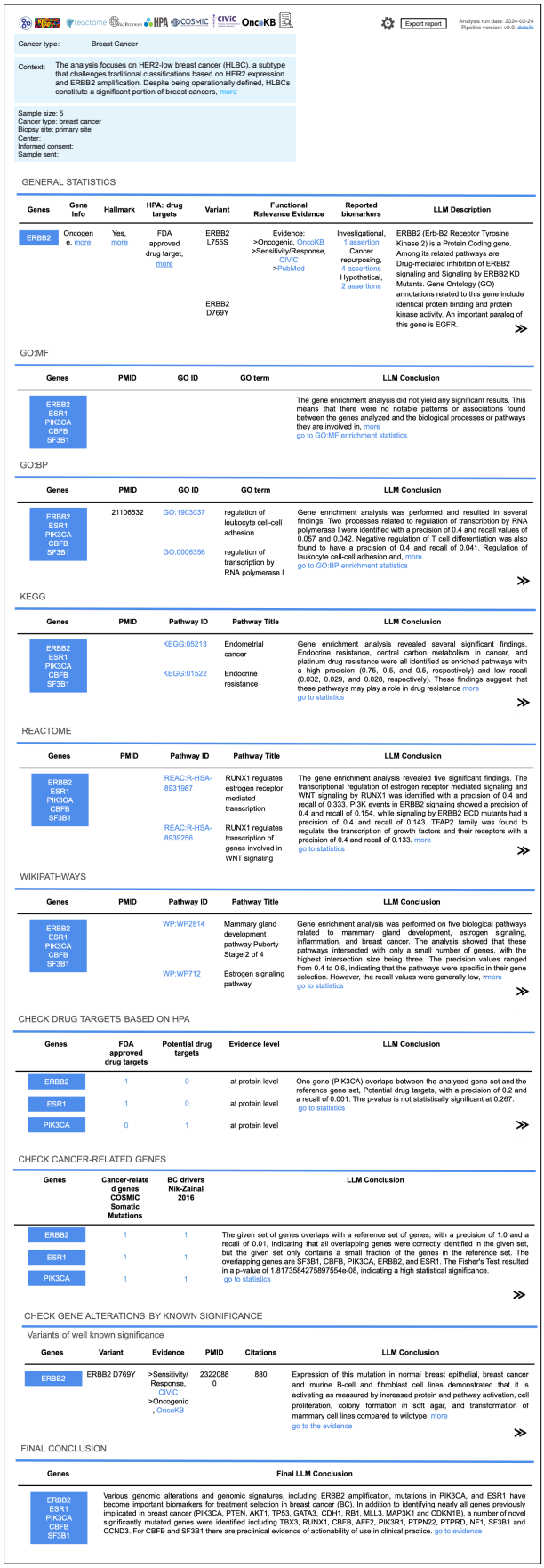


Figure A.7: Human Protein Atlas workflow in the BioLunar interface.



Cancer type: Breast Cancer

Context: The analysis focuses on HER2-low breast cancer (HLBC), a subtype that challenges traditional classifications based on HER2 expression and ERBB2 amplification. Despite being operationally defined, HLBCs constitute a significant portion of breast cancers, more

Sample size: 5
Cancer type: breast cancer
Biopsy site: primary site
Center:
Informed consent:
Sample sent:

GENERAL STATISTICS

Genes	Gene Info	Hallmark	HPA: drug targets	Variant	Functional Relevance Evidence	Reported biomarkers	LLM Description
ERBB2	Oncogene, more	Yes, more	FDA approved drug target, more	ERBB2 L755S ERBB2 D769Y	Evidence: >Oncogenic, OncoKB >Sensitivity/Response, CIVIC >PubMed	Investigational, 1 assertion Cancer repurposing, 4 assertions Hypothetical, 2 assertions	ERBB2 (Erb-B2 Receptor Tyrosine Kinase 2) is a Protein Coding gene. Among its related pathways are Drug-mediated inhibition of ERBB2 signaling and Signaling by ERBB2 KD Mutants. Gene Ontology (GO) annotations related to this gene include identical protein binding and protein kinase activity. An important paralog of this gene is EGFR. ≫

GO:MF

Genes	PMID	GO ID	GO term	LLM Conclusion
ERBB2 ESR1 PIK3CA CBFB SF3B1				The gene enrichment analysis did not yield any significant results. This means that there were no notable patterns or associations found between the genes analyzed and the biological processes or pathways they are involved in, more go to GO:MF enrichment statistics

GO:BP

Genes	PMID	GO ID	GO term	LLM Conclusion
ERBB2 ESR1 PIK3CA CBFB SF3B1	21106532	GO:1903037 GO:0006356	regulation of leukocyte cell-cell adhesion regulation of transcription by RNA polymerase I	Gene enrichment analysis was performed and resulted in several findings. Two processes related to regulation of transcription by RNA polymerase I were identified with a precision of 0.4 and recall values of 0.057 and 0.042. Negative regulation of T cell differentiation was also found to have a precision of 0.4 and recall of 0.041. Regulation of leukocyte cell-cell adhesion and, more go to GO:BP enrichment statistics ≫

KEGG

Genes	PMID	Pathway ID	Pathway Title	LLM Conclusion
ERBB2 ESR1 PIK3CA CBFB SF3B1		KEGG:05213 KEGG:01522	Endometrial cancer Endocrine resistance	Gene enrichment analysis revealed several significant findings. Endocrine resistance, central carbon metabolism in cancer, and platinum drug resistance were all identified as enriched pathways with a high precision (0.75, 0.5, and 0.5, respectively) and low recall (0.032, 0.029, and 0.028, respectively). These findings suggest that these pathways may play a role in drug resistance more go to statistics ≫

REACTOME

Genes	PMID	Pathway ID	Pathway Title	LLM Conclusion
ERBB2 ESR1 PIK3CA CBFB SF3B1		REAC:R-HSA-8931987 REAC:R-HSA-8939256	RUNX1 regulates estrogen receptor mediated transcription RUNX1 regulates transcription of genes involved in WNT signaling	The gene enrichment analysis revealed five significant findings. The transcriptional regulation of estrogen receptor mediated signaling and WNT signaling by RUNX1 was identified with a precision of 0.4 and recall of 0.333. PI3K events in ERBB2 signaling showed a precision of 0.4 and recall of 0.154, while signaling by ERBB2 ECD mutants had a precision of 0.4 and recall of 0.143. TFAP2 family was found to regulate the transcription of growth factors and their receptors with a precision of 0.4 and recall of 0.133. more go to statistics ≫

WIKIPATHWAYS

Genes	PMID	Pathway ID	Pathway Title	LLM Conclusion
ERBB2 ESR1 PIK3CA CBFB SF3B1		WP:WP2814 WP:WP712	Mammary gland development pathway Puberty Stage 2 of 4 Estrogen signaling pathway	Gene enrichment analysis was performed on five biological pathways related to mammary gland development, estrogen signaling, inflammation, and breast cancer. The analysis showed that these pathways intersected with only a small number of genes, with the highest intersection size being three. The precision values ranged from 0.4 to 0.6, indicating that the pathways were specific in their gene selection. However, the recall values were generally low, rmore go to statistics ≫

CHECK DRUG TARGETS BASED ON HPA

Genes	FDA approved drug targets	Potential drug targets	Evidence level	LLM Conclusion
ERBB2	1	0	at protein level	One gene (PIK3CA) overlaps between the analysed gene set and the reference gene set, Potential drug targets, with a precision of 0.2 and a recall of 0.001. The p-value is not statistically significant at 0.267. go to statistics
ESR1	1	0	at protein level	
PIK3CA	0	1	at protein level	≫

CHECK CANCER-RELATED GENES

Genes	Cancer-related genes COSMIC Somatic Mutations	BC drivers Nik-Zainal 2016	LLM Conclusion
ERBB2	1	1	The given set of genes overlaps with a reference set of genes, with a precision of 1.0 and a recall of 0.01, indicating that all overlapping genes were correctly identified in the given set, but the given set only contains a small fraction of the genes in the reference set. The overlapping genes are SF3B1, CBFB, PIK3CA, ERBB2, and ESR1. The Fisher's Test resulted in a p-value of 1.817358427589755 4e-08, indicating a high statistical significance. go to statistics
ESR1	1	1	
PIK3CA	1	1	≫

CHECK GENE ALTERATIONS BY KNOWN SIGNIFICANCE

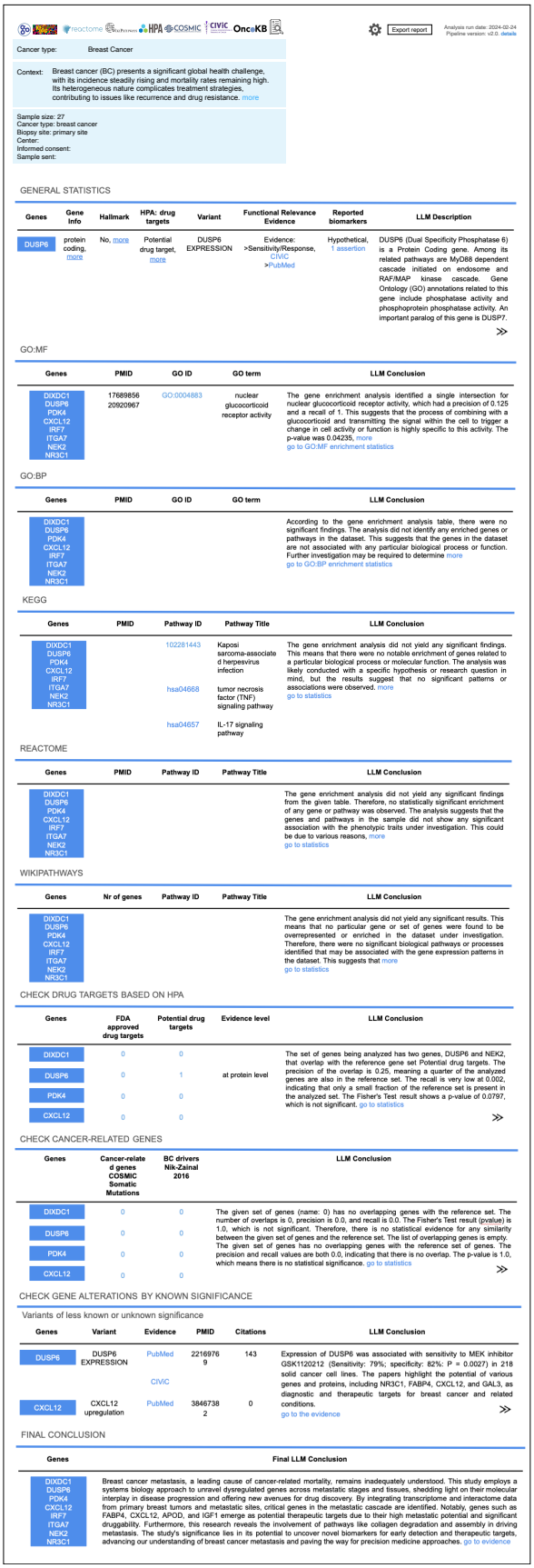
Variants of well known significance

Genes	Variant	Evidence	PMID	Citations	LLM Conclusion
ERBB2	ERBB2 D769Y	>Sensitivity/Response, CIVIC >Oncogenic, OncoKB	23220880	880	Expression of this mutation in normal breast epithelial, breast cancer and murine B-cell and fibroblast cell lines demonstrated that it is activating as measured by increased protein and pathway activation, cell proliferation, colony formation in soft agar, and transformation of mammary cell lines compared to wildtype. more go to the evidence ≫

FINAL CONCLUSION

Genes	Final LLM Conclusion
ERBB2 ESR1 PIK3CA CBFB SF3B1	Various genomic alterations and genomic signatures, including ERBB2 amplification, mutations in PIK3CA, and ESR1 have become important biomarkers for treatment selection in breast cancer (BC). In addition to identifying nearly all genes previously implicated in breast cancer (PIK3CA, PTEN, AKT1, TP53, GATA3, CDH1, RB1, MLL3, MAP3K1 and CDKN1B), a number of novel significantly mutated genes were identified including TBX3, RUNX1, CBFB, AFF2, PIK3R1, PTPN22, PTPRD, NF1, SF3B1 and CCND3. For CBFB and SF3B1 there are preclinical evidence of actionability of use in clinical practice. go to evidence

Figure A.8: The BioLunar report, produced by *Bioworkflow* for Scenario 1



Cancer type: Breast Cancer

Context: Breast cancer (BC) presents a significant global health challenge, with its incidence steadily rising and mortality rates remaining high. Its heterogeneous nature complicates treatment strategies, contributing to issues like recurrence and drug resistance. more

Sample size: 27
Cancer type: breast cancer
Biopsy site: primary site
Center:
Informed consent:
Sample sent:

GENERAL STATISTICS

Genes	Gene Info	Hallmark	HPA: drug targets	Variant	Functional Relevance Evidence	Reported biomarkers	LLM Description
DUSP6	protein coding, more	No, more	Potential drug target, more	DUSP6 EXPRESSION	Evidence: >Sensitivity/Response, CIVIC >PubMed	Hypothetical, 1 assertion	DUSP6 (Dual Specificity Phosphatase 6) is a Protein Coding gene. Among its related pathways are MyD88 dependent cascade initiated on endosome and RAF/MAP kinase cascade. Gene Ontology (GO) annotations related to this gene include phosphatase activity and phosphoprotein phosphatase activity. An important paralog of this gene is DUSP7. ≫

GO:MF

Genes	PMID	GO ID	GO term	LLM Conclusion
DIXDC1 DUSP6 PDK4 CXCL12 IRF7 ITGA7 NEK2 NR3C1	17688956 20920967	GO:0004883	nuclear glucocorticoid receptor activity	The gene enrichment analysis identified a single intersection for nuclear glucocorticoid receptor activity, which had a precision of 0.125 and a recall of 1. This suggests that the process of combining with a glucocorticoid and transmitting the signal within the cell to trigger a change in cell activity or function is highly specific to this activity. The p-value was 0.04235, more go to GO:MF enrichment statistics

GO:BP

Genes	PMID	GO ID	GO term	LLM Conclusion
DIXDC1 DUSP6 PDK4 CXCL12 IRF7 ITGA7 NEK2 NR3C1				According to the gene enrichment analysis table, there were no significant findings. The analysis did not identify any enriched genes or pathways in the dataset. This suggests that the genes in the dataset are not associated with any particular biological process or function. Further investigation may be required to determine more go to GO:BP enrichment statistics

KEGG

Genes	PMID	Pathway ID	Pathway Title	LLM Conclusion
DIXDC1 DUSP6 PDK4 CXCL12 IRF7 ITGA7 NEK2 NR3C1		102281443 hsa04668 hsa04657	Kaposi sarcoma-associated herpesvirus infection tumor necrosis factor (TNF) signaling pathway IL-17 signaling pathway	The gene enrichment analysis did not yield any significant findings. This means that there were no notable enrichment of genes related to a particular biological process or molecular function. The analysis was likely conducted with a specific hypothesis or research question in mind, but the results suggest that no significant patterns or associations were observed. more go to statistics

REACTOME

Genes	PMID	Pathway ID	Pathway Title	LLM Conclusion
DIXDC1 DUSP6 PDK4 CXCL12 IRF7 ITGA7 NEK2 NR3C1				The gene enrichment analysis did not yield any significant findings from the given table. Therefore, no statistically significant enrichment of any gene or pathway was observed. The analysis suggests that the genes and pathways in the sample did not show any significant association with the phenotypic traits under investigation. This could be due to various reasons, more go to statistics

WIKIPATHWAYS

Genes	Nr of genes	Pathway ID	Pathway Title	LLM Conclusion
DIXDC1 DUSP6 PDK4 CXCL12 IRF7 ITGA7 NEK2 NR3C1				The gene enrichment analysis did not yield any significant results. This means that no particular gene or set of genes were found to be overrepresented or enriched in the dataset under investigation. Therefore, there were no significant biological pathways or processes identified that may be associated with the gene expression patterns in the dataset. This suggests that more go to statistics

CHECK DRUG TARGETS BASED ON HPA

Genes	FDA approved drug targets	Potential drug targets	Evidence level	LLM Conclusion
DIXDC1	0	0	at protein level	The set of genes being analyzed has two genes, DUSP6 and NEK2, that overlap with the reference gene set Potential drug targets. The precision of the overlap is 0.25, meaning a quarter of the analyzed genes are also in the reference set. The recall is very low at 0.002, indicating that only a small fraction of the reference set is present in the analyzed set. The Fisher's Test result shows a p-value of 0.0797, which is not significant. go to statistics
DUSP6	0	1		
PDK4	0	0		
CXCL12	0	0		≫

CHECK CANCER-RELATED GENES

Genes	Cancer-related genes COSMIC Somatic Mutations	BC drivers Nik-Zainal 2016	LLM Conclusion
DIXDC1	0	0	The given set of genes (name: 0) has no overlapping genes with the reference set. The number of overlaps is 0, precision is 0.0, and recall is 0.0. The Fisher's Test result (pvalue) is 1.0, which is not significant. Therefore, there is no statistical evidence for any similarity between the given set of genes and the reference set. The list of overlapping genes is empty. The given set of genes has no overlapping genes with the reference set of genes. The precision and recall values are both 0.0, indicating that there is no overlap. The p-value is 1.0, which means there is no statistical significance. go to statistics ≫
DUSP6	0	0	
PDK4	0	0	
CXCL12	0	0	

CHECK GENE ALTERATIONS BY KNOWN SIGNIFICANCE

Variants of less known or unknown significance

Genes	Variant	Evidence	PMID	Citations	LLM Conclusion
DUSP6	DUSP6 EXPRESSION	PubMed CIVIC	22169769	143	Expression of DUSP6 was associated with sensitivity to MEK inhibitor GSK1120212 (Sensitivity: 79%; specificity: 82%; P = 0.0027) in 218 solid cancer cell lines. The papers highlight the potential of various genes and proteins, including NR3C1, FABP4, CXCL12, and GAL3, as diagnostic and therapeutic targets for breast cancer and related conditions. go to the evidence ≫
CXCL12	CXCL12 upregulation	PubMed	38467382	0	

FINAL CONCLUSION

Genes	Final LLM Conclusion
DIXDC1 DUSP6 PDK4 CXCL12 IRF7 ITGA7 NEK2 NR3C1	Breast cancer metastasis, a leading cause of cancer-related mortality, remains inadequately understood. This study employs a systems biology approach to unravel dysregulated genes across metastatic stages and tissues, shedding light on their molecular interplay in disease progression and offering new avenues for drug discovery. By integrating transcriptome and interactome data from primary breast tumors and metastatic sites, critical genes in the metastatic cascade are identified. Notably, genes such as FABP4, CXCL12, APOD, and IGF1 emerge as potential therapeutic targets due to their high metastatic potential and significant druggability. Furthermore, this research reveals the involvement of pathways like collagen degradation and assembly in driving metastasis. The study's significance lies in its potential to uncover novel biomarkers for early detection and therapeutic targets, advancing our understanding of breast cancer metastasis and paving the way for precision medicine approaches, go to evidence

Figure A.9: The BioLunar report, produced by *Bioworkflow* for Scenario 2.