Leveraging Large Language Models for Rapid Literature Review and Experimental Validation in Cancer Research

Oskar Wysocki[1,2], Magdalena Wysocka[2], Andre Freitas[1,2]

1. Idiap Research Institute, Martigny, Switzerland
2. digital Experimental Cancer Medicine Team, Cancer Research UK National Biomarker Centre, University of Manchester, UK;

Introduction

In the rapidly evolving field of oncology, understanding the intricate relationships within omics data is crucial for advancing personalized medicine and therapeutic strategies. Our innovative system leverages a collection of domain-specific components, including comprehensive databases, to analyze a given set of genes. By integrating Large Language Models (LLMs) to interpret the resulting data, we aim to unveil complex omics relationships, thereby facilitating groundbreaking discoveries in cancer research. This end-to-end workflow is designed not only to lower the barriers to evidence interpretation but also to establish a sophisticated reasoning infrastructure capable of navigating the vast expanse of biomedical evidence. Our approach seeks to empower researchers with tools for enhanced experimentation and demonstration, marking a significant leap forward in the utilization of AI for oncological innovation.

Purpose: Create an end-to-end workflow to support the interpretation/discovery of complex omics relations in oncology.

Results and Discussion

LLMs can lower the barriers for evidence interpretation

The system annotates the data with multiple resources based on the classification schemes. Among these resources, the system uses third-party knowledge bases that

(i) formalize information about the variants' effect by using predefined processes,
(ii) are based on published biomedical literature and
(iii) are committed to periodical updates.

Variant nomenclature, terminology and taxonomy systems used to annotate the data are also unified across knowledge bases with an automatic pipeline that is regularly updated. Additional filtering distinguishes knowledge base assertions supported by weaker or inconclusive evidence, as extracted from the corresponding metadata as appropriate. If none of the evidence is conclusive, then the system evaluates variants' relevance with using PubMed Searcher component and LLM.

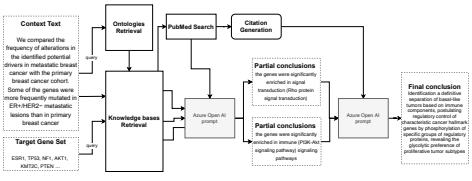

Figure 1. The main workflow of the system.

Results and Discussion

LLMs can provide a complex reasoning infrastructure over biomedical evidence

Core input:
- Gene-gene interaction signature, target subtypes

Query components:
- NL query + list of genes in CSV file
- SQL query

Outputs:
- Associated pathways and functions with citations: MSigDB, KEGG, WikiPathways, PathwayCommons, UniProt KB, NextProt
- List of most relevant evidence reported from curated DBs with citations: CIViC, OncoKB, COSMIC
- List of most relevant evidence reported from PubMed Abstracts

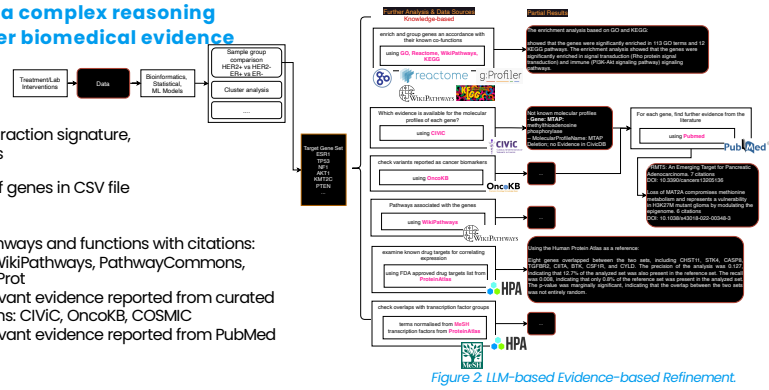

Figure 2. LLM-based Evidence-based Refinement.

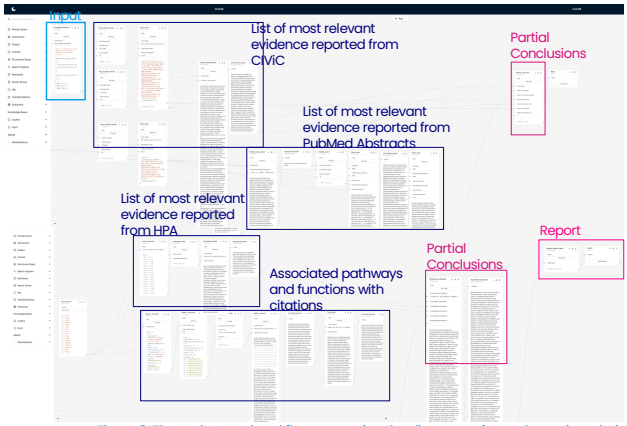

Figure 3. The end-to-end workflow supporting the discovery of complex omics relations.

Report

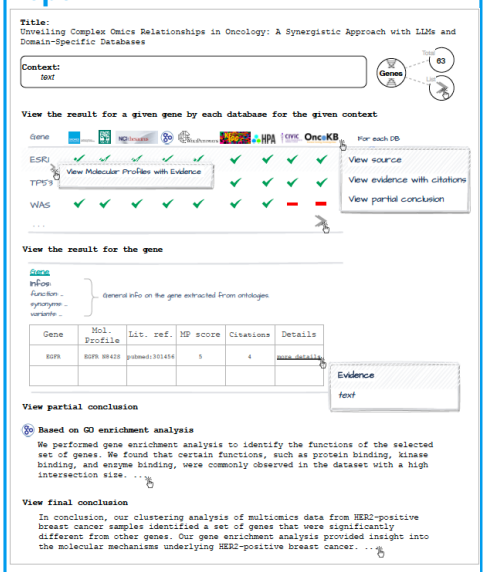

Figure 4. Interactive report containing list of most relevant evidence reported from curated DBs with citations, associated pathways and functions with citations.

Conclusions

In conclusion, our system represents a significant leap forward in the utilization of AI for oncological innovation. By combining the analytical power of LLMs with comprehensive domain-specific databases, we offer a new paradigm for understanding and treating cancer. This approach not only enhances the precision of oncology research but also holds the promise of transforming patient care through the development of more effective, personalized treatment plans.

Future Work

Building upon the significant strides made in leveraging Large Language Models (LLMs) for the interpretation of omics data in oncology, our future work aims to expand and refine the capabilities of this innovative system. Here are the key directions for our future research and development:

1. Integration of Emerging Data Types
2. Improvement of LLM Interpretative Abilities
3. Enhancing User Interfaces and Accessibility
4. Expanding Collaborative Networks

Acknowledgements

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 965397.