Revolutionizing GWAS Data Exploration for Target Discovery and Drug Repurposing with LLMs and Knowledge Graphs

Ardigen: Błażej Szczerba, Marek Kudła

OVERVIEW

- We present a solution for the identification and interpretation of impactful variants and targets from GWAS studies, which enables regular biologists to access strong evidence of functional relevance for drug discovery..
- By providing empowering LLM-based natural language interface this solution unlocks professional bioinformatic analysis in a no-code fashion.
- By leveraging knowledge graphs as a source of functional insights, our solution handles with ease difficult task of result interpretation.

INTRODUCTION

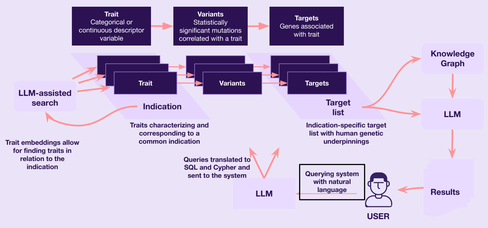
Genome-Wide Association Studies (GWAS) have become a cornerstone for identifying genetic variants that influence disease susceptibility and drug response. A notable example of high value provided by these studies is the successful repurposing of ustekinumab and risankizumab for Crohn's disease treatment. Despite multiple successes, the complexity, scale and difficulty in interpretation of GWAS data often present significant challenges for researchers aiming to translate genetic discoveries into therapeutic applications.

To address this, we showcase AI-powered solution designed to streamline the exploration and functional characterization of GWAS studies. It empowers researchers to interactively explore genetic data and uncover associations more efficiently. By leveraging LLM that is further enhanced with distilled knowledge gathered from over twenty large sources, Our solution enables rapid identification of meaningful genetic variants, prioritization of targets, and hypothesis generation in drug discovery and disease research.

With its user-friendly an intuitive no-code interface it allows biologists, clinicians, and researchers without coding experience to quickly analyze and interpret complex associations in genomic data, reducing the time and resources typically expended for this process. This efficiency and the comprehensive knowledge-based evidence support provided by the system enables faster decision-making in drug repurposing, supports novel therapeutic discoveries, and helps researchers make data-driven insights for better patient stratification in clinical settings.

TECHNOLOGY

Figure 1. Workflow of the analysis depicting genetic data flow within the system. Smart system components: LLM, knowledge graph and trait embeddings are responsible for advanced features of the system focusing on user experience through natural language interpretation, as well as functional analysis of the data and interpretation.



System

Solution (shown in Fig.1) is deployed as a data lake allowing for large-scale processing and computation. A LLM is responsible for translation of user input into system queries, enabling full no-code experience. Additionally, LLM is also used for integrating the results from various points of the system to present them in the form of comprehensive report with tables and visualizations.

Data

Summary statistics for all UK Biobank traits are available within the system. Data from other biobanks can be integrated into a single harmonized source. A comprehensive Knowledge Graph (Fig.2. shows encoded elements) contains the information that is used for interpreting and annotating the results.

Figure 2.Types of Knowledge Graph nodes and relationships between them

RESULTS AND DISCUSSION

Unlocking Data Exploration for All Biologists through AI-Powered No-code Solution

Scientific research is evolving, and so should the way we interact with data. The conviction that ease of access enables better research and can empower biologists and researchers to explore, analyze, and derive insights from genetic datasets—without the need for coding skills. This approach not only enables previously disenfranchised researchers, but also speeds up the process considerably, as well as provides more thoroughness and comprehensiveness than previously available. We demonstrate we streamlines the use of GWAS summary statistics data derived from Biobanks, making research that previously required collaboration with bioinformatician teams or acquired coding abilities more accessible to everybody.

Key differentiators of the system can be summarized as follows:

- **AI-Powered Comprehension of GWAS Summary Statistics Data** Infer biological insights faster, more thoroughly and accurately using models leading in text understanding and multidimensional embeddings of traits to help you formulate the analysis for particular indication.
- **Cloud-Based Data Lake Integration** Biobank data has been ingested into a secure and scalable Data Lake, enabling structured access for further exploration.
- **NLP-Powered Data Exploration** Navigate massive datasets using the language you would naturally use to describe your analysis goals, eliminating the programming barrier for analyzing data in depth. The system handles synonyms, and close-term matches seamlessly with AI-enhanced NLP
- **Intelligent Interpretation and Annotation with Comprehensive Knowledge Graph** Discover relationships, information and knowledge distilled from over twenty sources of relations between drugs, indications, targets and other entities that is used to analyze, annotate and interpret the data, allowing for connecting the dots, that would otherwise take months to complete.

RATIONALE

The Strength of Genetic Data in Informing Target Selection for Drug Discovery

The most important aspect of the Drug Development process is the decision on which biological process to influence in order to achieve the therapeutics goal. Historically, the path to this decision has been lengthy and complex, relying on serendipitous discoveries and accumulated multi-year research experience of scientific labs investigating various related phenomena. In order to modulate this process a specific target needs to be found within its mechanism, to which a specific drug needs to be developed, typically using the traditional small molecule approach. A significant challenge in these efforts has been the need to identify a target based on an incomplete and irregular body of evidence, characterized by fragmented data, isolated research findings, and a general scarcity of information. This has led to instances where the leading candidate target failed in clinical trials after being invalidated as actually unrelated to the process of interest, largely due to confusion between cause and effect, as well as difficulties in distinguishing between primary and secondary effects.

This has been slowly changing in the last three decades due to the availability of large biological data sets enabled by the rapid availability of cheap NGS capability. One of the most notable consequences of this transformation was development of GWAS analysis made possible by sequencing genomes for large cohorts of patients, allowing for establishing a link between the specific indications and traits and the genomic variants shared by the patients that are linked to these traits. This information provides a very strong, often causative link between targets and indications, allowing for efficient identification of targets for therapeutic modulation.

After initial limited availability and inadequate quality of the data, last decade have shown emergence of large national biobanks that can be systematically explored for drug discovery purposes. The results of first trials with targets selected with the help of GWAS were aggregated, revealing a significant improvement of chances in the drug discovery trials. [1,2,3]

"The most important step in drug discovery - identify of drug target. Historically choosing the wrong target, was the biggest source of failure. And what we learnt in the last decade is that genetics, studying the DNA of humans is the best way to find the targets and increase probability of success of about two and a half fold."

Jake Rubens, CEO Quotient Therapeutics, The Flagship Pioneering Company

While it was realized very fast that this approach holds enormous promise, it still requires a significant amount of know-how to access the data, process it and interpret it. This still is a considerable hurdle that is slowing down efforts to adopt genetic data in the drug discovery process. Furthermore, while analysis of single traits is relatively simple, it is still rare to see fully integrated approaches analyzing multiple traits in relation to a particular disease for better characterization resulting in a broader list of novel targets. One of the serious restrictions is the necessity of coding and bioinformatics know-how required to run a larger investigative study. By eliminating these requirements we democratize the access to genetic data, enabling all specialists with deep biological knowledge to contribute directly to the drug discovery regardless of technical acumen.

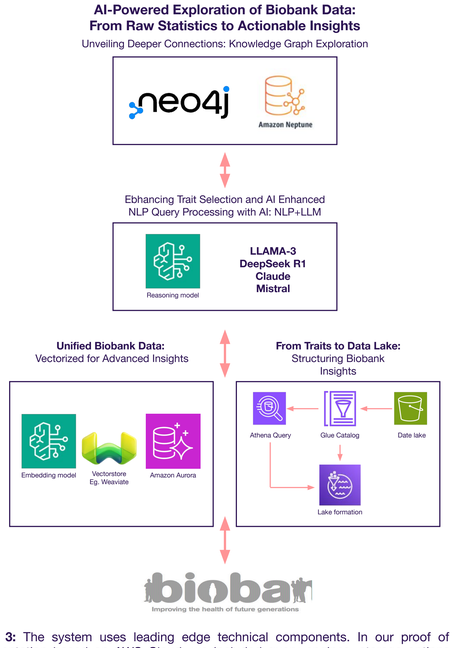


Figure 3: The system uses leading edge technical components. In our proof of concept, implementation based on AWS Cloud, we included query engines, storage options and LLM systems. Being cloud/system agnostic lies in our roots, so options to use self-hosted LLMs the queries within the system can be kept contained within deployed solution, ensuring full confidentiality.

Dynamic AI-Driven Trait Selection

Presented solution enables researchers to dynamically define trait ensembles by leveraging natural language input and advanced multidimensional similarity analysis. Users simply provide text descriptions, and the system identifies relevant traits from a comprehensive list of GWAS summary statistics sourced from multiple biobanks.

How It Works:

- **Natural Language Trait Matching**
  Users input text-based queries describing traits of interest, usually around the indication of interest. The system performs multidimensional similarity analysis, finding the most relevant GWAS traits from biobank datasets.
- **AI-Enhanced Trait Ranking**
  Foundational models (e.g., LLAMA3, Mistral) further refine the list, identifying hidden relationships between traits. Each trait is assigned a specificity score ranking its relevance towards the intended analysis focus.
- **Seamless Data Integration**
  The selected set of relevant traits is subsequently imported into a data lakehouse, making it readily available for downstream processing, querying and analysis.

This AI-powered trait selection transforms how researchers discover, rank, and integrate biobank data, streamlining the path from hypothesis to analysis. By leveraging NLP and machine learning, it enhances data accessibility, automates trait selection, and enables seamless exploration through follow up annotation with knowledge graphs. This approach accelerates time-to-insights, improves consistency and accuracy, and reduces manual effort, allowing researchers to focus on crafting hypothesis as opposed to wrangling with technical issues.

Figure 4: The POC allows for exploration of the UK Biobank summary statistics data, however the architecture supports data from all biobanks

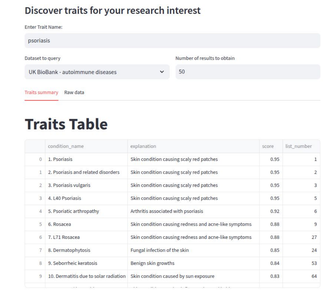


Fig 5. Exemplary output of the system where given a starting indication of psoriasis - most similar traits have been identified, along with the similarity score based on multidimensional embedding. This can be used for defining cohort of traits for further downstream analysis.

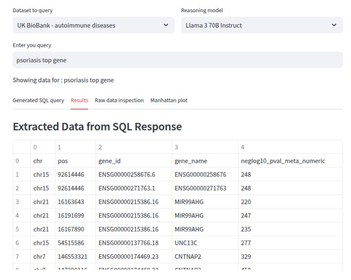


Fig 6. User interface section where the user provides an exemplary query using everyday language and the LLM reasoning model, like Llama 3 allows it to be transformed into query language, like SQL allowing non-coding user eg. biologist or genomic researcher to generate complex searches and analyses.

Adaptive Data Exploration with AI-Powered Queries

Flexible and dynamic environment is a important factor for researchers, that enables to explore pre-populated GWAS summary statistics Data Lakes, which can be generated using a trait selection tool. Users can refine their datasets, launch complex queries and analyze the results, exploring which from the different AI models best suits their analytical needs and personal preferences.

How It Works:

- **Choose from Pre-populated Data Lakes**
- **Adaptive AI Model Selection**
- **Intelligent Query Understanding**

Standard & Custom Functionalities

While the solution is highly flexible and can be coerced to perform most complex queries, we are also offering on-demand specialized tools that can be tailored to specific datasets and requirements.

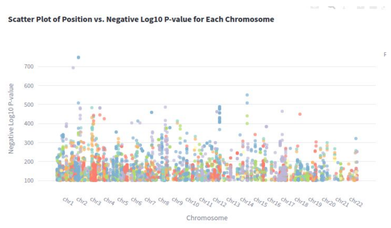


Fig 7. Example: Manhattan Plot Visualization – Effortlessly identify the most significant genes and variants across the genome for the selected trait set. This is one of the many plots that are possible to be created within the system. This approach enhances user experience by combining AI-assisted natural language queries with powerful visualization tools, allowing for dynamic creation of plots and bridging the gap between free data exploration and methodical data analysis.

Unveiling Gene-to-Disease Associations through Knowledge Graph Exploration

Building dynamic trait selections is a significant first step—but true discovery lies in exploring deeper connections. This capability enables researchers to investigate actual associations between genes, proteins, drugs and diseases, providing rich context for biological insights through the use of sizeable knowledge graph (KG).

How it works:

- **Comprehensive KG Built from Over Twenty Data Sources**
  We have integrated publicly available datasets from more than 20 biomedical databases This ensures that researchers can access diverse, up-to-date information on relationships between indications, drugs, targets, phenotypes, pathways and localization.
- **AI-Powered Querying for Deeper Exploration**
  Just like in the trait selection tool, users can input queries in natural language that is further processed by the system to create a Cypher query used to interact with the knowledge graph, allowing it to return the data that is further processed to return structured, relevant insights.

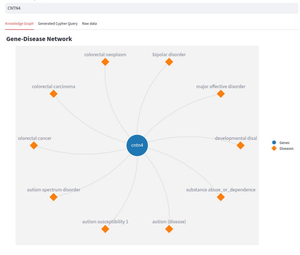


Fig 8. This view presents fragment of the knowledge graph highlighting associations between a particular target with known indications. This is just one of the simplest relations that can result from the queries submitted to the system

- **From Traits to Knowledge Graph Insights**
  The knowledge graph allows researchers to go beyond raw data and explore how selected traits, genes, and diseases interconnect. By linking genetic variants to known targets, drugs, biological pathways, and phenotypic outcomes, we provide a holistic perspective on biomedical relationships.
- **Analysis, Annotation and Interpretation of Genetic Results**
  The knowledge graph is particularly well suited for analysis of hidden relations present within the target list. Customarily, such lists do not convey much information on their own and are the subject of functional analysis and biological interpretation. Here, the system uses KG to annotate the list with all relevant associations from the possible six types of nodes (drugs, pathways, phenotypes, etc., see Fig. 2.), while the subsequent filtering and processing with the use of LLM compiles the list of insights available in the data.

CONCLUSIONS

The power of GWAS approach lies in its statistical power, where large cohorts allow for detecting causative variants with high certainty. This has resulted in widely recognized contribution of genetic data to 2.5x increase in success probability for targets selected with its use. [2]
The true advantage of the presented solution comes from its smart features where the deep learning components are orchestrated to annotate and describe the data. This brings elements of explainability to the process that was previously the subject of additional follow-up analysis, often supported by manual research.
With Knowledge Graph as a tool in their belt, researchers can navigate complex biological networks effortlessly, turning raw data and scientific questions into actionable knowledge. Furthermore, the use of knowledge graphs has also been shown to increase the probability of successful outcome in target identification, with approximately 3x improvement over traditional methods. [4]

Future developments and improvements

- Integrating additional data modalities to bring independent lines of evidence to the process
- Orchestrating interaction between genetic data and KG, ensuring truly multiplicative synergy in increasing success of right target identification
- Expanding the set of ready-to-go visualizations built in into the reporting system
- Dedicated workflows for drug repurposing
- Adding additional modes of validation, as opposed to discovery, with stricter filtering criteria at a cost of novelty of targets

REFERENCES

[1] Ochoa D, Karim M, Ghoussaini M, Hulcoop DG, McDonagh EM, Dunham I. Human genetics evidence supports two-thirds of the 2021 FDA-approved drugs. Nat Rev Drug Discov. 2022 Aug;21(8):551. doi: 10.1038/d41573-022-00120-3. PMID: 35804044.
[2] Minikel EV, Painter JL, Dong CC, Nelson MR. Refining the impact of genetic evidence on clinical success. Nature. 2024 May;629(8012):624-629. doi: 10.1038/s41586-024-07316-0. Epub 2024 Apr 17. PMID: 38632401; PMCID: PMC11096124.
[3] Nelson MR, Tipney H, Painter JL, Shen J, Nicoletti P, Shen Y, Floratos A, Sham PC, Li MJ, Wang J, Cardon LR, Whittaker JC, Sanseau P. The support of human genetic evidence for approved drug indications. Nat Genet. 2015 Aug;47(8):856-60. doi: 10.1038/ng.3314. Epub 2015 Jun 29. PMID: 26121088.
[4] Liu C, Xiao K, Yu C, Lei Y, Lyu K, Tian T, Zhao D, Zhou F, Tang H, Zeng J. A probabilistic knowledge graph for target identification. PLoS Comput Biol. 2024 Apr 5;20(4):e1011945. doi: 10.1371/journal.pcbi.1011945. PMID: 38578805; PMCID: PMC11034645.