Γεφυρώνοντας μεγάλα ομικά, γενετικά και ιατρικά δεδομένα για την ευρεία εφαρμογή της Ιατρικής Ακριβείας στην Ελλάδα

DELIVERABLE D11.2

«Delivery of Database and User Interface – Software and Publication in a Scientific Journal»

Φορέας	Hellenic Pasteur Institute
Τύπος Παραδοτέου	Other
Ημερομηνία Υποβολής Παραδοτέου	15th February 2026
Ενότητα Εργασίας	Work Package 11 « Information Retrieval and Knowledge Extraction for Mutation–Cancer Associations from Biomedical Literature»

Ελλάδα 2.0
ΕΘΝΙΚΟ ΣΧΕΔΙΟ ΑΝΑΚΑΜΨΗΣ ΚΑΙ ΑΝΘΕΚΤΙΚΟΤΗΤΑΣ
Με τη χρηματοδότηση της Ευρωπαϊκής Ένωσης
NextGenerationEU

1 Introduction 4

2 Web interface implementation 6

3 Publication..... 13

Related Publications

1. Stavropoulos, S., Zacharopoulou, E., Georgakopoulos, S., Tasoulis, S., Plagianakos, V., & Hatzigeorgiou, A. G. (2024, August). Leveraging Large Language Models for Information Extraction: Identifying microRNA-Gene Interactions in Biomedical Literature. In *2024 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)* (pp. 1-7). IEEE.

2. Grigoriadis, D., Tsifintaris, M., Giannakakis, A., Pavlopoulos, G. A., & Perdikopanis, N. (2025). Public Omics Explorer (POE): Enabling integrative semantic search across GEO omics datasets based on PubMed publications. Computational and Structural Biotechnology Journal.

Ελλάδα 2.0
ΕΘΝΙΚΟ ΣΧΕΔΙΟ ΑΝΑΚΑΜΨΗΣ ΚΑΙ ΑΝΘΕΚΤΙΚΟΤΗΤΑΣ
Με τη χρηματοδότηση της Ευρωπαϊκής Ένωσης
NextGenerationEU

1 Introduction

Within WP11, and building directly on the text-mining/LLM pipeline of Deliverable 11.1, Deliverable 11.2 focused on operationalizing extracted knowledge into usable, searchable resources. Concretely, WP11.2 delivers (i) a database and lightweight web interface that host and expose the literature-derived gene–variant–cancer associations produced in WP11.1, and (ii) an integrated semantic retrieval platform (Public Omics Explorer, POE) that supports discovery of public omics datasets by leveraging the scientific narrative of the biomedical literature. Together, these components strengthen the WP11 objective of enabling evidence-backed retrieval and exploration across heterogeneous biomedical resources, supporting both non-technical users (e.g., biologists/clinicians) and technical users who require integration into downstream computational workflows.

The first component is a structured knowledge base of gene–variant–cancer associations extracted from PubMed abstracts. Each association is stored as a normalized, machine-actionable record linked to its supporting publication (PMID) and standard identifiers to preserve traceability and enable reuse. A lightweight web interface provides structured access to this content through three core views—Data, Methodology, and Statistics—balancing usability with transparency. The Data view supports a search-and-filter workflow, allowing users to query by gene (symbol or Ensembl ID), variant (rsID), and cancer type (including ontology identifiers where available), and to inspect results in a paginated table. To support reuse in analyses and reporting, the interface includes one-click export functionality both for the full dataset and for any filtered subset (XLSX export). The database currently contains 13,680 extracted associations supported by 7,319 unique publications (PMIDs), spanning 7,191 unique variants, 2,406 unique genes, and 73 cancer types.

In addition, WP11.2 incorporates the Public Omics Explorer (POE) as a complementary platform for unified, literature-driven discovery of public omics datasets. The motivation is the fragmentation of public resources: processed functional genomics datasets are typically accessed via GEO, raw sequencing reads via ENA, and the descriptive scientific context via PubMed. POE bridges these sources by semantically linking GEO datasets and ENA records through their associated PubMed publications and by enabling semantic search over the textual content of titles and abstracts. In practical terms, POE maintains an automated acquisition and indexing pipeline that periodically collects GEO metadata, GEO–ENA cross-references, and PubMed records, then produces biomedical text embeddings (SBioBERT) and indexes them for similarity search (FAISS). Users can submit natural-language queries, apply structured filters (e.g., organism, experiment type, sample type, library strategy, extracted molecule, publication year), retrieve relevant publications and the associated datasets, and access programmatic functionality through a RESTful API. By enabling literature-aware, semantic retrieval rather than keyword-only matching, POE supports hypothesis generation, meta-analysis, and exploratory research workflows over public omics data.

The scientific methodology that enabled the collection, processing, analysis, and development of WP11 resources has been published in the 2024 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB) under the title “Leveraging Large Language Models for Information Extraction: Identifying microRNA–Gene Interactions in Biomedical Literature” (DOI:

10.1109/CIBCB58642.2024.10702129). Finally, although the long-term scope of WP11 includes structuring additional evidence dimensions (e.g., pharmacological response and survival/clinical outcome evidence per cancer type and relationship category such as onset/recurrence/metastasis and good/poor response), the WP11.2 implementation is designed so that these fields can be incorporated through schema and interface extensions without changing the core interaction pattern.

Users may access the WP11.2 gene–variant–cancer database through https://dianalab.e-ce.uth.gr/microt_webserver/cancervar_db (or alternatively at .

2 Web interface implementation

2.1 Web-based Knowledge Base for Literature-Derived Gene–Variant–Cancer Associations

2.1.1 Description

Deliverable 11.2 implements a web-based interface that provides structured access to the literature-derived knowledge base produced in Deliverable 11.1. The interface was designed around typical user workflows in biomedical data exploration: searching for a specific gene or variant, narrowing results by cancer type, reviewing evidence-linked records, and exporting query results for downstream analysis. To support these use cases, the interface is organized into three main sections—Data, Methodology, and Statistics—each corresponding to a distinct layer of user interaction: retrieval of records, documentation of how records were produced, and high-level characterization of dataset content.

2.1.2 Impementation

The Data section implements a search-and-filter view that allows users to query associations by Gene (gene symbol or Ensembl ID), Variant (rsID), and Cancer type using either a textual term or a DOID identifier (Figure 1). Filters are designed to be composable so that users can apply one criterion (e.g., an rsID) or combine multiple criteria (e.g., a gene and cancer type) to refine results. Query results are displayed in a paginated table, enabling responsive browsing even when the underlying dataset contains thousands of records (Figure 2). Each row includes the PMID to preserve traceability to the originating publication. In addition to interactive browsing, the interface supports direct export to XLSX either for the currently filtered subset or for the complete dataset, providing a practical bridge to external analysis environments.

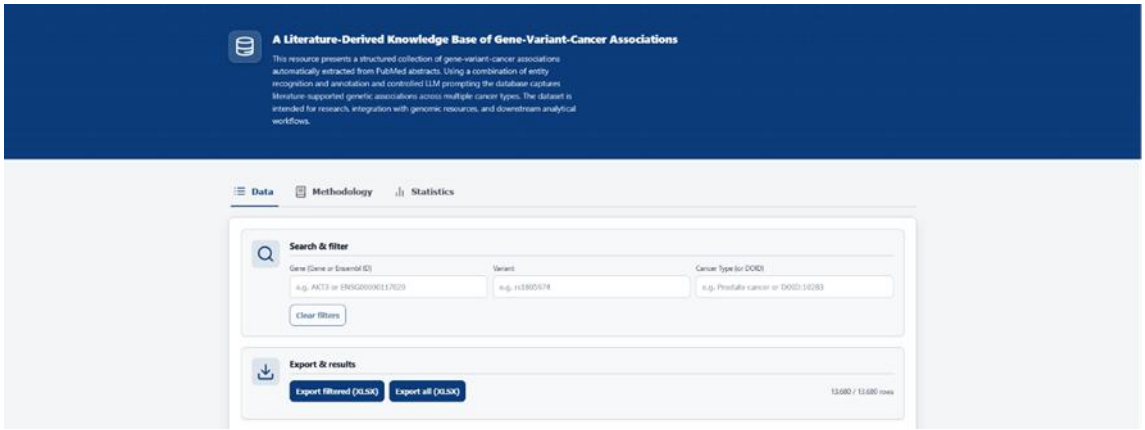

Figure 1. Main "Data" tab of the WP11.2 web interface showing the search panel with three primary filters (Gene, Variant, Cancer Type) and the associated export controls for downloading either the filtered result set or the entire dataset in XLSX format.

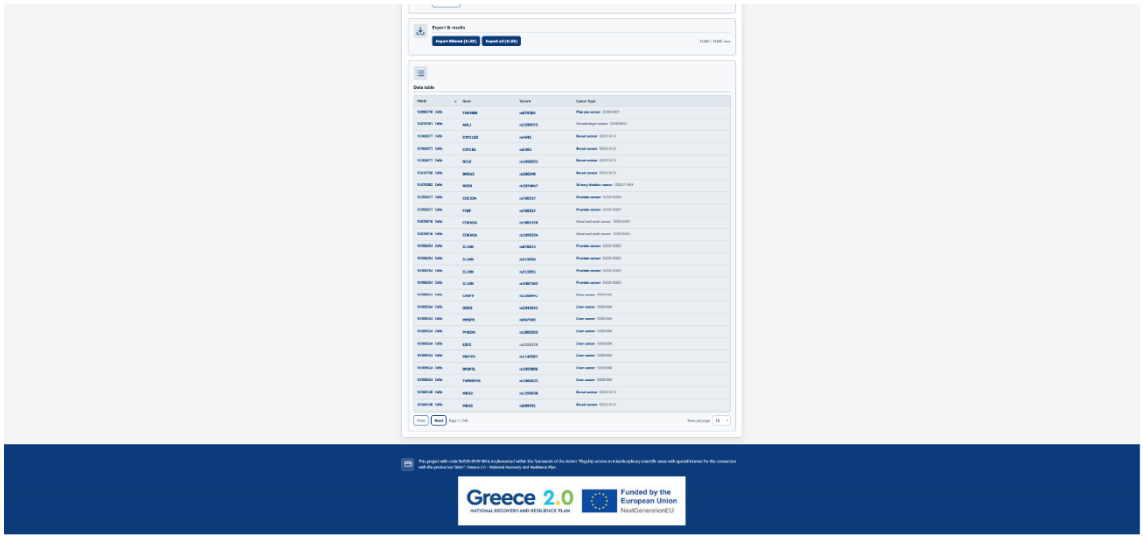

Figure 2. Paginated table view of the extracted associations, displaying key fields per record (PMID, Gene, Variant (rsID), Cancer Type, including ontology identifiers where available).

The Methodology section provides a concise narrative of the upstream workflow that generated the knowledge base, covering the main steps from literature collection and entity annotation to LLM-based extraction. This documentation is included directly in the interface so that users can understand, at a high level, what the entries represent and how they were derived, without needing to consult separate documents (Figure 3).

Figure 3. The "Methodology" tab summarising, in a user-facing form, the pipeline stages used to construct the database (literature collection, entity annotation/normalisation, LLM-based extraction, and resulting knowledge base), together with guidance on how the database should be cited when reused.

In parallel, the Statistics section summarizes dataset scale and composition using key counts (e.g., total associations, unique PMIDs, unique variants, unique cancer types, and unique genes) and includes descriptive plots that highlight dominant cancer categories or other aggregate patterns (Figure 4). This statistical view supports rapid orientation and helps users interpret the scope and coverage of the resource before running targeted queries.

Figure 4. The "Statistics" tab reporting dataset-wide summary counts (13,680 total associations; 7,319 PMIDs; 7,191 variants; 73 cancer types; 2,406 genes) and descriptive plots (e.g., top cancer types by number of extracted associations) to provide an at-a-glance view of coverage and distribution.

2.1.3 Infrastructure

The web interface Implementation uses a bundled Vue 3 single-page application that applies a modulepreload polyfill (when needed) and then boots the runtime and mounts the app to a single DOM container. The UI is structured as three top-level views (Data/Methodology/Statistics): the Data view drives parameterized queries and renders a paginated results table, while export is generated client-side as XLSX for either the active filtered subset or the full dataset. Methodology content is served as a static in-app page, and the Statistics view computes/loads dataset-wide aggregates and renders summary counts and plots for quick coverage assessment.

The current interface reflects the content produced in 11.1—namely gene–variant–cancer associations extracted from PubMed abstracts and linked to supporting PMIDs. At the same time, the implementation is structured so that additional annotation dimensions (for example, outcome-related metadata or therapeutic association fields) can be incorporated in future iterations without disrupting the existing search fields, table layout, or export mechanism. This

approach protects usability and backward compatibility while keeping the system open to future enrichment.

2.1.4 Availability

The database is accessible through https://dianalab.e-ce.uth.gr/microt_webserver/cancervar_db or alternatively at http://10.64.83.167/cancervar_db.

2.2 Public Omics Explorer (POE): An Integrated Platform for Semantic Search and Retrieval of Public Omics Datasets via the Biomedical Literature

2.2.1 Description

The exponential growth of publicly available omics datasets and biomedical literature creates both opportunities and challenges for knowledge discovery in the life sciences. While the Gene Expression Omnibus (GEO) hosts millions of high-throughput experiment datasets, the European Nucleotide Archive (ENA) stores the corresponding raw sequencing data, and PubMed provides a large body of related publications, unified exploration of these resources remains limited. We present the Public Omics Explorer (POE), a web platform that implements literature-driven dataset retrieval by semantically connecting GEO datasets and ENA records through their associated PubMed publications. POE collects and indexes GEO metadata, ENA cross-references and PubMed abstracts on a weekly basis. For semantic integration it uses the biomedical specialized SBioBERT model to produce dense vector representations from publication text. Embeddings are indexed with Facebook AI Similarity Search (FAISS) to enable high-precision, context-aware retrieval.

Users submit natural-language free-text queries that are converted into semantic queries to identify conceptually related datasets based on the content of linked publications. Structured filters allow refinement by organism, experiment type, library strategy, sample type, extracted molecule and publication year. POE also supports direct retrieval via accession identifiers (GSE IDs, PubMed IDs, DOIs) and provides a RESTful API for integration into computational pipelines.

By linking processed GEO data with raw ENA data through a shared publication context, POE facilitates hypothesis generation, meta-analysis and exploratory research. The application is freely available at https://nplab.gr/poe.

2.2.2 Relation to Work Package 11

Within Work Package 11, POE is leveraged and adapted as a core component for unified, semantic retrieval and indexing of biomedical information that connects literature and public omics datasets. POE is a web platform that implements literature-driven retrieval by semantically connecting Gene Expression Omnibus (GEO) datasets and raw sequencing records (ENA) through their associated PubMed publications. This addresses fragmentation across sources (GEO for

processed data, ENA for raw data, PubMed for descriptive context) and enables unified exploration based on publication content.

The implementation follows an automated acquisition and integration pipeline. GEO metadata and associated PubMed records are collected regularly, and cross-references to ENA are integrated to provide links to FASTQ files where available. Automated curation and metadata normalization harmonize records into a consistent schema enabling filtering and uniform downstream processing. Per-article bibliographic metadata (e.g., title, abstract, authors, publication dates, MeSH terms) are stored and linked to corresponding GEO/ENA records, creating an integrated article–dataset association graph.

A central element is semantic search: publication text (title + abstract) is converted into dense vectors via the biomedical SBioBERT model, enabling retrieval by conceptual relevance rather than keyword matching. Embeddings are indexed with FAISS so that natural language user queries return the most relevant articles and, consequently, their linked omics datasets. Structured filters (organism, experiment type, library strategy, sample type, extracted molecule, publication year) and direct lookup via identifiers (GSE, PMID, DOI) are also provided.

POE is delivered as a web application accessible via modern browsers without local installation, and is complemented by a RESTful API for integration into computational workflows. Structured metadata are stored in a relational database (e.g., PostgreSQL), and vector representations and similarity indexing enable high-precision, context-aware retrieval. A regular update mechanism (e.g., weekly) ensures incorporation of new records and repository updates.

2.2.3 Methods

System overview and accessibility

POE is hosted as a web platform accessible through modern browsers without user-side installation or configuration (Figure 1). Metadata ingestion, embedding generation, indexing and result delivery are performed server-side.

Data sources and acquisition

POE integrates data from GEO, PubMed and ENA. GEO metadata are retrieved weekly via NCBI FTP and E-utilities. GSE and GSM records are parsed to extract fields such as organism, library strategy, experiment type, sample type and extracted molecule (Figure 2).

PubMed publications linked to GSE are collected where PMIDs are available. Metadata are retrieved programmatically via NCBI APIs and HTML parsing (BeautifulSoup) and include title, abstract, MeSH terms, authors/affiliations and publication dates. ENA metadata are incorporated to provide FASTQ links for GEO datasets when available.

Data processing and integration pipeline

Integration of GEO and literature is implemented via a fully automated weekly pipeline that updates GEO and PubMed stores, links GSE to PMIDs, generates embeddings for new/updated articles and updates the FAISS index.

For each new or updated GSE record, an article is mapped via PMID. Records without a valid link are retained in the GEO index but excluded from semantic embedding. Structured metadata are stored in PostgreSQL with an extensible schema for heterogeneous fields.

Embedding generation and vector representation

POE uses SentenceTransformers with the SBioBERT model, trained on PubMed abstracts and PMC full texts and further optimized for sentence-level tasks. SBioBERT produces 768-dimensional embeddings suitable for biomedical semantic similarity.

Before encoding, title and abstract are concatenated, cleaned of non-informative characters and truncated to a maximum of 256 tokens. Embeddings are stored and indexed with FAISS to enable fast retrieval of semantically similar content.

Semantic search and indexing

The semantic index is based on 768-dimensional embeddings and implemented with FAISS. By default, IndexFlatL2 is used, enabling exact nearest-neighbor search with full recall/precision at the cost of higher memory/time requirements. Alternative index types (IndexIVFFlat, IndexPQ, IndexHNSW) were evaluated as speed/memory trade-offs.

Back-end infrastructure

The back end is implemented in Python with FastAPI. Metadata and embeddings are stored in PostgreSQL with pgvector. User queries are embedded into the same space and used to search for semantically similar PubMed abstracts via FAISS; these results are then used to rank associated GEO datasets. The service is deployed with Uvicorn for asynchronous serving.

Front-end and user interface

The user interface is implemented as a single-page application using React, TypeScript and Tailwind CSS. It supports natural-language queries, browsing of matched PubMed records and exploration of linked datasets with filters (organism, library strategy, experiment type, publication year) and export/collection functionality. Direct ENA FASTQ links are displayed when available.

Programmatic access via RESTful API

POE provides a RESTful API with token-based authentication. Tokens expire after 24 hours and rate limiting is applied (1 request per 4 seconds). Endpoints include semantic PubMed search with filters, direct metadata retrieval by GSE/PMID/DOI/title, and ENA metadata retrieval with file links.

API documentation: https://github.com/DimitrisGrig/POE/wiki/API.

Software versions

POE was developed with Python 3.12.9. The back end uses FastAPI v0.115.11 and Uvicorn v0.34.0. Embeddings are generated with sentence-transformers v3.4.1 and SBioBERT. FAISS-cpu v1.10.0 and PostgreSQL v17.4 are used. The front end is based on React v19.0.0, TypeScript v5.7.2 and Tailwind CSS v4.0.9.

Update schedule

POE is updated weekly through automated schedulers orchestrating retrieval, embedding generation, indexing and deployment.

2.2.4 Results

Platform overview and coverage

POE is a fully functional web application for semantic exploration of omics datasets via linked literature. As of September 2025, POE indexes 262,173 GEO datasets (GSE) and 9,682,571 sample records (GSM), together with 130,548 PubMed publications. GEO records with valid PubMed identifiers are semantically embedded, enabling similarity search from text.

Literature-driven retrieval

Unlike keyword search, POE supports concept discovery through publication content. For example, queries such as “inflammation in Parkinson’s disease” can return relevant publications and linked GEO datasets even when terms are not explicitly present in dataset metadata.

Additional access modes

POE provides two direct modes: Dataset Search (by GSE ID or GEO title) and Article Search (by PubMed title or DOI), enabling rapid access to known studies without exploratory queries.

Availability

POE does not host or store raw omics data. It indexes metadata and provides external links to GEO, PubMed and ENA. The application is freely available at https://nplab.gr/poe.

3 Publications

The methodological foundations relevant to WP11—namely the literature collection, entity recognition and normalisation, and controlled LLM-based information extraction strategies—were presented at the 2024 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB) under the title “Leveraging Large Language Models for Information Extraction: Identifying microRNA–Gene Interactions in Biomedical Literature” (DOI: 10.1109/CIBCB58642.2024.10702129). This work informed parts of the broader WP11 extraction and structuring workflow.

In addition, the Public Omics Explorer (POE) platform integrated within WP11.2 has been published as: “Public Omics Explorer (POE): Enabling integrative semantic search across GEO omics datasets based on PubMed publications”, Grigoriadis D, Tsifintaris M, Giannakakis A, Pavlopoulos GA, Perdikopanis N., Computational and Structural Biotechnology Journal, Volume 27 (2025), Pages 4802–4812 (DOI: 10.1016/j.csbj.2025.11.004; PMID: 41282419).

2024 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB) | 979-8-3503-5663-2/24/$31.00 ©2024 IEEE | DOI: 10.1109/CIBCB58642.2024.10702129

Leveraging Large Language Models for Information Extraction: Identifying microRNA - Gene Interactions in Biomedical Literature

Steve Stavropoulos*
Department of Computer Science and Biomedical Informatics
University of Thessaly, Greece
stavropoulos@uth.gr

Elissavet Zacharopoulou*
Department of Computer Science and Biomedical Informatics
University of Thessaly, Greece
Hellenic Pasteur Institute, Greece
elzacharop@uth.gr

Spiros Georgakopoulos
Department of Mathematics
University of Thessaly, Greece
spirosgeorg@uth.gr

Sotiris Tasoulis
Department of Computer Science and Biomedical Informatics
University of Thessaly, Greece
stas@uth.gr

Vassilis Plagianakos
Department of Computer Science and Biomedical Informatics
University of Thessaly, Greece
vpp@uth.gr

Artemis G. Hatzigeorgiou
Department of Computer Science and Biomedical Informatics
University of Thessaly, Greece
Hellenic Pasteur Institute, Greece
arhatzig@uth.gr

***Abstract*—The rapid growth of biomedical literature necessitates efficient Information Extraction systems able to identify relevant knowledge for various biological applications, such as understanding gene regulation by microRNA (miRNA). In this study, we employed a Large Language Model, specifically GPT-3.5 (version 0301), in conjunction with BERN2 for miRNA-gene interaction extraction from paper titles and abstracts. We optimized our approach using an initial dataset of about a thousand molecular biology papers and subsequently evaluated its performance on a manually curated dataset of 400 papers, achieving an accuracy of 82-85%. Driven by the promising results and the practical utility of our method, we applied the system to a large dataset of 39,000 papers. The extracted miRNA-gene interactions, combined with a Natural Language Processing approach, were included in the TarBase v9 database. Our findings demonstrate the potential of Large Language Models in biomedical Information Extraction tasks and highlight the limitations of the current gene and miRNA recognition systems, which hinder further improvements in accuracy.**

***Index Terms*—Large Language Models, chatGPT, Information Extraction, MicroRNAs, Gene Regulation**

I. Introduction

The growing volume of biomedical literature has made manual curation of microRNA (miRNA) gene interactions increasingly impractical. Various Natural Language Processing (NLP) and Information Extraction techniques have been employed to address this issue, but their accuracy and efficiency remain challenging. In this paper, we demonstrate the potential of Large Language Models (LLMs), specifically GPT-3.5 (version 0301), in extracting miRNA-gene interactions from research articles.

MiRNAs characterized as non-protein coding single-stranded endogenous RNAs typically about 22 nucleotides in length, can actively regulate gene expression and play a crucial role in cell differentiation, proliferation, apoptosis, and tumorigenesis [1]. A worthy aspect of miRNA functionality is to bind to protein-coding messenger RNAs (mRNAs), degrade them, and induce their translational suppression [2]. The inaugural recognition of this capability dates back to 1993, with the divulgence that lin-4 negatively regulates the protein LIN-14 [3]. This groundbreaking discovery opened a path for scientists to explore how these mechanisms operate in different organisms and uncover their roles in biological processes, encompassing normal functions and those related to diseases, such as cancer susceptibility.

Over the last three decades, numerous scientific studies have been dedicated to experimentally validating binding interactions between miRNAs and mRNAs and exploring their endogenous activity. As the exploration of miRNA functionality continues to expand, the accompanying literature grows exponentially, presenting an escalating challenge in navigating and extracting relevant information from the literature. To illustrate the growth and trajectory of miRNA research, Figure 1 depicts the number of publications featuring miRNA references in the PubMed database over the years. Our search query was designed to capture various mentions of microRNAs in the titles or abstracts of papers, as well as using a MeSH term. The query was formulated as: "'mirna'[Title/Abstract] OR 'microrna'[Title/Abstract] OR 'mirnas'[Title/Abstract] OR 'micrornas'[Title/Abstract] OR 'MicroRNAs'[MeSH Terms]".

In this work, we utilize GPT-3.5 and BERN2 [4] for biomedical named entity recognition to extract miRNA-gene interactions from biomedical literature, resulting in a computational tool. We specifically chose BERN2 over PubTator because BERN2 is trained to recognize miRNAs, whereas PubTator

*These authors contributed equally to this work.

struggles with accurately identifying miRNAs in text. This tool efficiently extracts information from biomedical literature with high accuracy, surpassing current techniques in this field. We also have created a web interface for on demand information extraction. Our research contributes to the advancement of computational methods in biology, advocating for further refinement of language models and entity recognition systems.

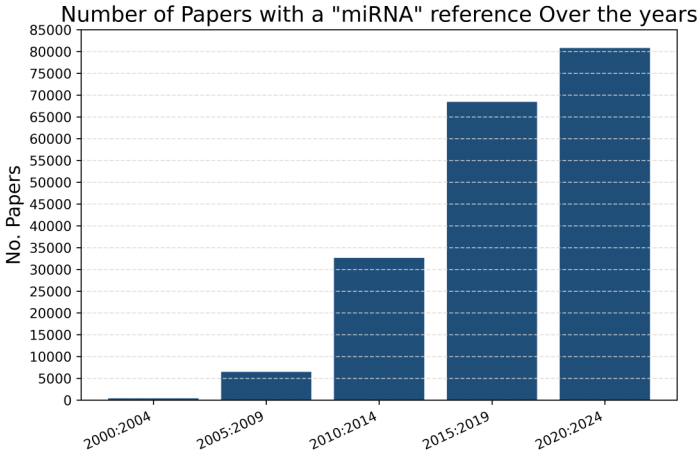

Fig. 1: Using a PubMed query, we generated a bar plot to show the annual scientific interest based on the number of papers published each year.

II. Related Work

Natural Language Processing (NLP) has gained significant traction in biomedical science, offering an efficient technique for automatically extracting essential information from the vast body of literature. Sentence segmentation, text summarization, Named Entity Recognition (NER) and Relation Extraction (RE) constitute methods that form a robust workflow for extracting critical relations between keywords of interest, such as gene-to-gene interactions or mutations and diseases relations. Biomedical NER and NEN approaches can be dictionary-based, rule-based, and statistical methods [5]–[7]. Dictionary-based methods are limited to cases where authors may refer to the names of genes and other molecules in distinct ways. Constructing rules or patterns for named entity identification in semantic approaches relies on an in-depth understanding of the domain. In contrast, statistical approaches treat NER as a classification problem by training statistical models, such as decision trees or Support Vector Machines (SVMs). Noteworthy tools in the realm of NEN include PubTator [8], a web-based annotation tool developed by the National Center for Biotechnology Information (NCBI) [9], and BERN2 [4], a neural network-based algorithm designed for biomedical named entity recognition. Deep learning models, such as long short-term memory (LSTM), convolutional neural network (CNN) [10], and bidirectional encoder representations from transformers (BERT) [11], have firmly positioned themselves as the state-of-the-art approaches in NLP tasks. However, these endeavors come with inherent challenges, including the potential for contextual ambiguities and a far from perfect accuracy.

The introduction of Large Language Models (LLMs), with their power of general understanding, has marked a significant evolution in automated Information Extraction, showing potential to surpass previous methods in effectiveness. In the field of Materials Chemistry, a fine-tuned version of GPT-3 was utilized, focusing on entity recognition and relation extraction. This application, as described in [12], highlights the model's effectiveness in discerning and categorizing relevant information within the domain. Over the last year, with advancements like GPT-3.5 and various other open-source models, there has been an increase in experimentation across various fields [13]. These instances illustrate the growing trend of employing LLMs in Information Extraction, underscoring their potential to significantly impact research methodologies across multiple disciplines.

III. Methodology

Before this work, we tried a classic NLP approach, involving a set of 4571 sentences classified as positive (2722 sentences) or irrelevant (1849 sentences). Employing machine learning techniques, we specifically utilized a BioBERT model [14] for sentiment analysis to discern the sentiment of these sentences. Following this, the Stanza model [15] was used to extract dependency trees, laying the foundation for further analysis. To capture intricate relationships between keywords, we applied the Term Frequency - Inverse Document Frequency (TF-IDF) statistical method and cosine similarity.

This comprehensive process identified critical terms within miRNA and gene references, establishing a threshold for accurate interaction extraction. BERN2 provides NCBI IDs [9] for genes and miRNAs, which we then annotated with Ensembl IDs [16] for genes and miRBase IDs [17] for miRNAs, respectively. The integration of our NLP pipeline with the BERN2 tool achieved only a 66.24% accuracy though. This relatively low accuracy, highlights the problems with the classic NLP approach and led us to explore the use of an LLM.

Using GPT-3.5 we saw significantly improved results, achieving an impressive 81.62% accuracy (measured in a dataset of about 1000 papers). Combining both methods and retaining only the common results yielded an even higher accuracy of 87.05% in extracting interactions from positive sentences. These advancements contributed to the TarBase v9 [18] database, providing a valuable resource for researchers.

A. Data Collection and Preprocessing

We collected two datasets for this study. The first dataset consisted of about a thousand molecular biology papers to establish an optimal LLM setup. The second dataset comprised 400 manually curated papers, which served as the evaluation set. These papers were randomly selected to ensure diversity in their contexts, allowing us to assess the model's performance across a wide range of complexity. The manual curation process involved strict guidelines to ensure consistency in

the use of official symbols for miRNAs and genes. We u
BERN2 for gene and miRNA recognition and preprocessi

B. LLM-based Information Extraction

Initially, we employed a two-step approach to ext
miRNA-gene interactions using GPT-3.5. The first step
volved asking the LLM to describe the regulations in
own words. The second step required the LLM to output
information in a custom format (mirna1:gene1;mirna2:gen

The evaluation of the approximately one thousand pa
of our first dataset, employed a semi-automated annota
approach. This method did not encompass entire abstr
but rather focused on specific sentences extracted by an N
system. Subsequently, these sentences underwent a mar
curation process governed by relatively flexible guideli
For instance, maybe one paper was annotated as microRNA-124:gene_name and the next as miRNA-124:gene_name. This variability in terminology introduced challenges in accurately assessing the performance of the LLM-based system. To address this, the final stage of our evaluation involved a comparative analysis between the LLM's responses and the annotations provided by human curators, which was facilitated by the LLM itself. We presented both annotations to the LLM and asked it to determine if they were equivalent. This approach proved to be far more effective than manual methods, such as using regular expressions. This comparison aimed to ascertain the congruence between the LLM-generated data and the manually curated labels, and it proved to be an effective method for evaluating the system's accuracy. Utilizing an LLM for performance evaluation in the context of imprecise data has shown to be highly effective.

We conducted several iterations and optimizations, as seen in Figure 2.

Through iterative optimization, we observed that loading only the answer from the LLM's first step and asking for the result in our custom format, without loading the full conversation, reduced the cost by half while maintaining comparable (or even better) accuracy. The above approach led us to about 75% accuracy on our 400 paper manually curated evaluation set.

In addressing the challenges encountered with our latest method, we identified two primary sources of error. Firstly, BERN2's limitations significantly impacted the accuracy of our results (for example mir-452-5p would be identified as mir-4525 by BERN2 and result in a false miss), a hurdle that was particularly challenging to overcome. Secondly, it is noteworthy that the LLM encountered difficulties in excluding specific types of information that were outside the intended scope of our curated dataset, such as transcriptional level interactions. Despite our efforts to configure the LLM to disregard certain aspects, the model demonstrated challenges in effectively filtering out unwanted information. This observation prompted us to further examine the limitations of the LLM and explore alternative strategies to address these issues.

Consequently, we pivoted to a new approach, shifting from our previous custom output format to a more comprehensive

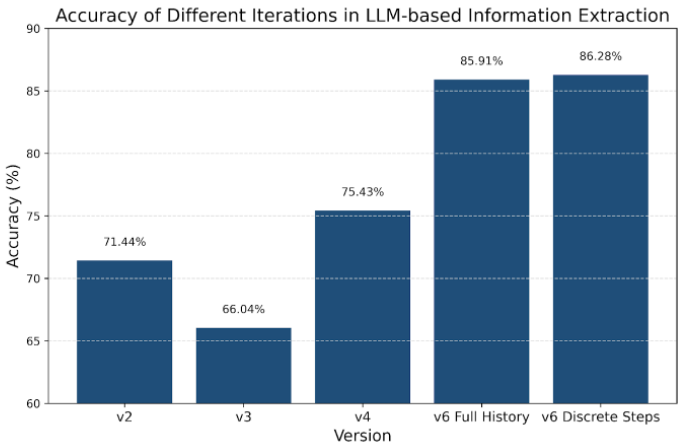

Fig. 2: The performance of an LLM depends a lot on the prompt used and this shows in our various iterations showing above. v6 uses JSON output and asks for more information, resulting in a much higher accuracy. The two v6 bars show the similar accuracy we got when giving the full history of the conversation to the LLM and when in each step we only give the last answer.

JSON structure. This change allowed us to harness the LLM's capability to extract a broader range of information from the abstracts. Specifically, we directed the LLM to generate JSON outputs where each key represented a discovered miRNA. Associated with each miRNA key, we requested a detailed list of regulated genes, including exhaustive information such as the type and level of regulation, tissue and cell line specifics, relevant diseases, employed methodologies, and the organism of study.

This change led to the following benefits:

1) Cost-efficiency: The new approach required only one step, reducing the cost and simplifying the parsing process.
2) Improved accuracy: By providing more specific output requirements, the LLM could better parse the paper, producing more accurate results.
3) Additional information: The JSON format allowed us to gather more information about the paper's results, including regulation type (up/down), regulation level, tissue type, cell line, disease, methodology used, and organism.

In applying our refined system to a dataset of 400 manually curated papers, we observed an overall accuracy of approximately 82-85%. This outcome underscores the limitations inherent in the current capabilities of GPT-3.5 and BERN2, the tools central to our methodology. Accuracy was measured by evaluating miRNA-gene interactions based on the normalization results provided by BERN2. Specifically, a match was considered accurate only if the extracted miRNA or gene name exactly matched the corresponding official ID name. Any deviation from the exact name, such as "hsa-miR-4525" instead of

"hsa-miR-452-5p," was considered incorrect. Notably, a significant portion of the inaccuracies encountered can be attributed to BERN2's limitations in accurately recognizing a broad spectrum of miRNAs and genes. This limitation has emerged as a critical bottleneck, hindering the further enhancement of our system's accuracy. While the LLM demonstrated robust performance, the dependency on BERN2 for initial entity recognition introduces a ceiling on the achievable accuracy, suggesting an area for future improvement and research in this domain.

C. Pipeline Workflow for microRNA-Gene Interaction Extraction

The LLM pipeline was implemented to enrich TarBase v9 with information about possible miRNA-gene interactions derived from the literature. Due to the vast number of papers included in PubMed, we first needed to narrow down the dataset to those containing the relevant information. We employed an Entrez query to obtain a list of approximately 39,000 PMIDs. The query was: "(MicroRNAs[MeSH Terms]) AND (RNA-Binding Proteins[MeSH Terms] OR RNA Messenger[MeSH Terms] OR Cell Line, Tumor[MeSH Terms]) AND (Journal Article[ptyp] OR Research Support, Non-U.S. Gov't[ptyp] OR Research Support, N.I.H., Extramural[ptyp]) NOT Review[ptyp] NOT Retracted Publication[ptyp]". Utilizing the BERN2 API, we retrieved the titles and abstracts of the papers for each PMID, as well as the IDs for the miRNAs and genes mentioned in their text. After acquiring these IDs, we queried the GPT-3.5 model with the corpus, capturing the analysis of the LLM response, particularly focusing on the JSON output format which includes any identified miRNA-gene interactions. The JSON format is: " "mirna_name": ["gene_name": "type": "up" or "down" or "unknown", "regulation_level": "the regulation level", "tissue_type": "the tissue type", "cell_line": ["cell line",...], "disease": "disease name", "methodology": ["methodology name (e.g. Luciferase)",...], "organism": "organism name",...], "mirna2": ...". Subsequently, we annotated the entities with Ensembl IDs for genes and miRBase IDs for miRNAs based on the NCBI IDs provided by BERN2. This process resulted in a finalized database containing PMID:miRNA_ID:gene_ID triplets. In evaluating the accuracy of the LLM, we consider only the interactions it identifies and assess these for correctness. Interactions that the LLM fails to identify are excluded from the accuracy evaluation. Therefore, the accuracy metric reflects the proportion of correctly identified interactions among those detected by the LLM.

D. Other Versions of GPT

In addition to our primary focus on version 0301 of GPT-3.5, we conducted tests on other iterations of this model. Notably, version 0613 exhibited the lowest performance, achieving approximately 79% accuracy. Meanwhile, version 1106 showed a marginal improvement over 0613, attaining an accuracy of about 80%. Version 0301, showed higher accuracy, ranging between 82-85%, but the other versions marginally excelled at consistently outputting correct JSON format. Additionally, we assessed GPT-4 and found its performance to be approximately 80% accurate. These findings indicate that the performance limitations are not inherently due to GPT-3.5. Instead, it suggests that future efforts to improve accuracy should focus on enhancing the entity recognition component.

E. Limitations in Current Entity Recognition: Implications for LLaMA Utilization

We also experimented with LLaMA2 [19], which indicated a basic comprehension of the scientific abstracts. However, the model encountered significant challenges in generating outputs in the specific format required for our parsing needs. This suggests that fine-tuning LLaMA could potentially make it suitable for our purposes. Nevertheless, a more pressing issue emerged during our research process, directly impacting the feasibility of effectively using LLaMA or any similar models.

The core obstacle lies in the current limitations of BERN2, the tool we employed for the identification of miRNAs and genes in scientific literature. BERN2's performance was not as robust as required, leading to inaccuracies in entity recognition. This deficiency in accurately identifying key biological entities casts doubt on the reliability of any subsequent analysis of the work performed by LLaMA. Given that precise entity recognition is crucial for the evaluation of the LLMs, addressing this fundamental flaw becomes a prerequisite for further advancements. Once this primary issue is resolved, we can revisit the possibility of fine-tuning open-source LLMs for more effective and reliable information extraction in our domain.

IV. Interface and Visualization

For efficient visualization of the outcomes derived from the large language model (LLM), a web-based user interface has been developed (see Figure 3). Upon inputting a PubMed ID (PMID), both the title and abstract of the associated research article are extracted via BERN2 API requests and displayed along with the information extracted by the LLM and the complete LLM response. Additionally, the interface highlights miRNAs and genes identified by the BERN2 system. The LLM is also employed to pinpoint significant sentences within the abstract, which are subsequently emphasized. Brief explanations for these sentences are provided and can be accessed by hovering the cursor over the highlighted text as shown in Figure 4.

V. Results

Our LLM-based approach demonstrated promising results in the field of miRNA-gene interaction extraction. When applied to a carefully curated evaluation dataset comprising 400 papers, the system achieved an overall accuracy of 82-85%. This level of precision marks a significant advancement over previous methods and highlights the potential of language models in biomedical information extraction.

Our system was implemented on a dataset encompassing 39,000 research articles, and the outcomes were integrated

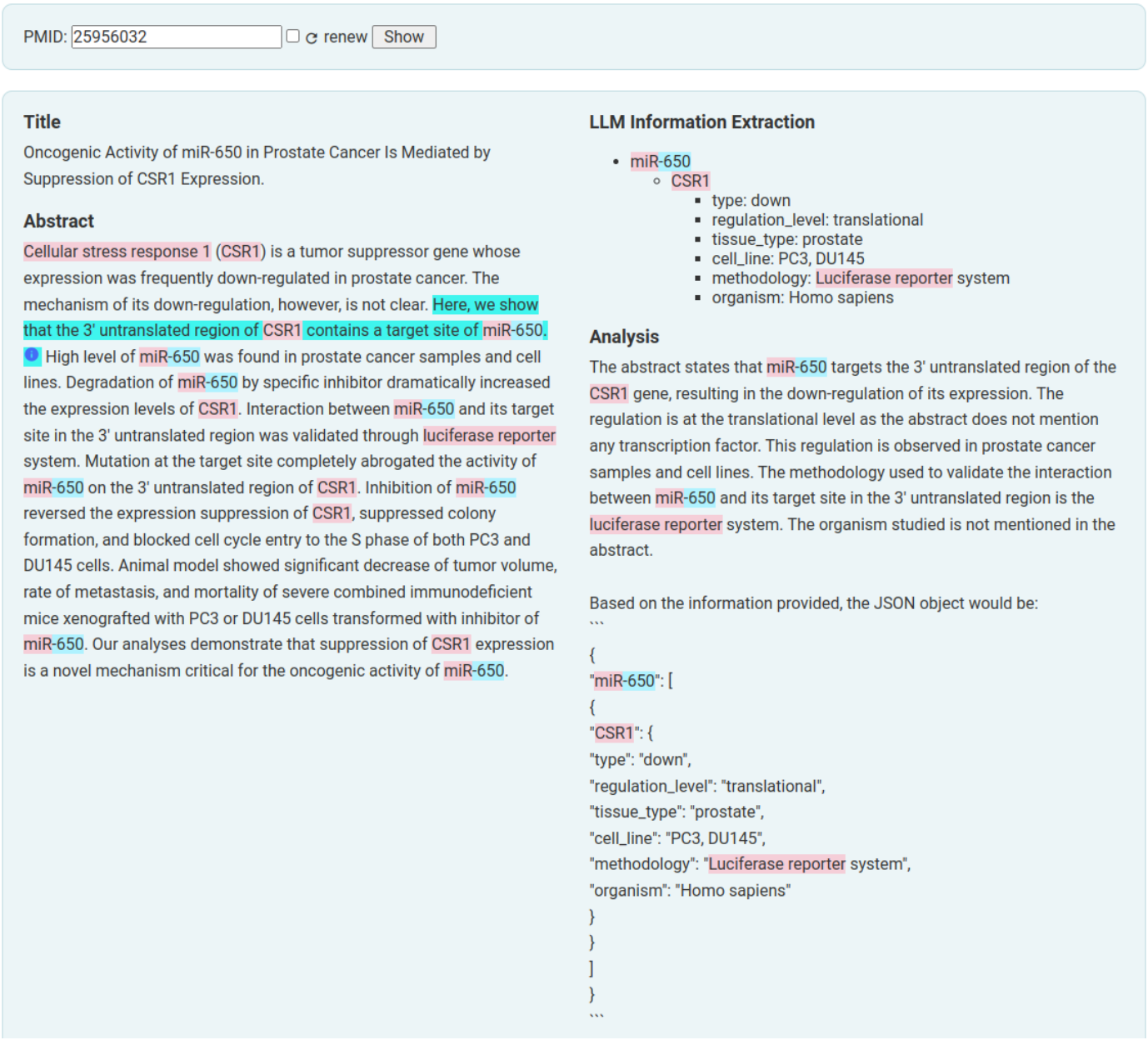

Fig. 3: Within the web interface, miRNAs are denoted by blue highlighting, while genes are marked in pink. As observed, the limitations associated with BERN2 identification become apparent.

with our conventional natural language processing (NLP) methodology to enhance precision and reliability. The findings from this execution have been incorporated into TarBase v9.

A key breakthrough in our methodology was the transition from the initial two-step process, which utilized a custom format for output, to a more streamlined single-step process with JSON output. This strategic shift not only simplified the data extraction pipeline but also enhanced the accuracy and richness of the information gathered. The JSON format allowed for a more detailed and structured representation of interactions, including comprehensive data on regulation types, levels, tissue and cell line specifics, associated diseases, and methodologies used.

Despite these advancements, the main source of errors in our system was identified as stemming from the BERN2 gene and miRNA recognition system. Notably, BERN2 demonstrated a significant challenge in accurately identifying several miRNAs and genes. This limitation in entity recognition posed a substantial barrier to achieving higher accuracy levels. The errors primarily occurred in the misidentification of miRNAs and genes in the abstracts, which subsequently led to inaccuracies in the interaction data extracted by the LLM.

Overall, the results underscore the potential and current limitations of using advanced language models in the extraction of

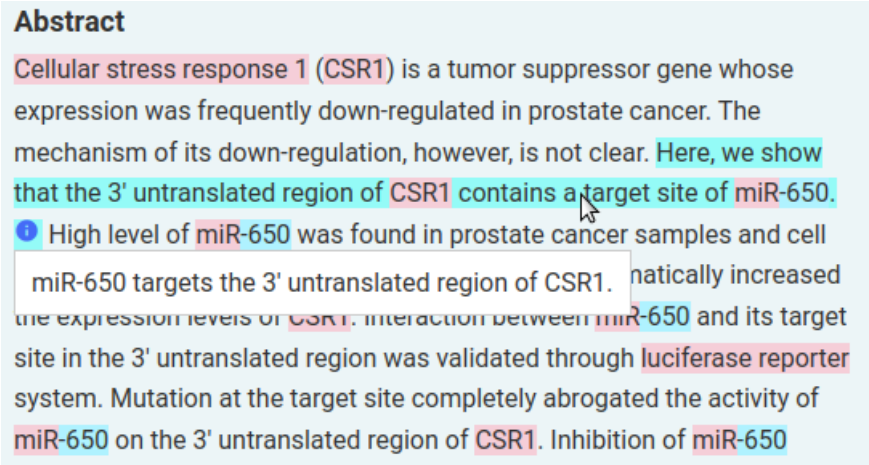

Fig. 4: Significant sentences are emphasized and succinct explanations are displayed upon hovering the cursor over the respective sentence.

complex biological interactions from scientific literature. The findings from this study not only contribute to the ongoing development of automated information extraction systems but also highlight critical areas for future research, particularly in enhancing entity recognition capabilities.

VI. Discussion and Conclusion

This study represents a significant stride in the application of language models, specifically GPT-3.5, for extracting miRNA-gene interactions from scientific literature. Our results, demonstrating an 82-85% accuracy rate on a curated dataset of 400 papers, underscore the potential of LLMs in automating complex biological data extraction. However, this research also highlights the challenges and limitations inherent in current technologies. The primary challenge was the inconsistent performance of the BERN2 entity recognition system, which significantly impacted our system's overall accuracy, indicating a need for more precise tools in automated entity recognition.

A. Methodological Innovations and Their Implications

The transition from a two-step extraction process to a single-step, JSON-based output format was a key methodological innovation. This change not only improved the efficiency and accuracy of our system but also provided a more structured and comprehensive dataset. This approach could serve as a model for future endeavors in similar domains, where rich data extraction is crucial.

B. Future Directions

The outcomes of this study pave the way for several future research directions. Firstly, there is a clear need for improved entity recognition systems that can more accurately process specialized scientific terminologies. Secondly, exploring ways to refine LLMs' focus on specific interaction types could enhance the precision of information extraction. Lastly, applying our methodology to other domains within molecular biology could further validate its effectiveness and adaptability.

C. Broader Impacts

This research contributes to the broader field of computational biology by demonstrating the practical application and challenges of using advanced LLMs in scientific research. The insights gained from this study could inform the development of more nuanced and accurate tools for data extraction, ultimately aiding in the advancement of scientific discovery and understanding.

In conclusion, our study demonstrates the potential of using GPT-3.5 for extracting miRNA-gene interactions from scientific literature. The transition to a single-step, JSON-based output significantly enhanced the efficiency and richness of the data extracted. While challenges with the BERN2 entity recognition system highlighted current technological limitations, our findings emphasize the need for improved entity recognition systems and more precise language models in computational biology. Future research should focus on refining these tools to enhance accuracy and efficiency in data extraction, suggesting a promising direction for further advancement in this field.

Acknowledgment

Funded by the European Union - NextGenerationEU through Greece 2.0—National Recovery and Resilience Plan, under the call "Flagship actions in interdisciplinary scientific fields with a special focus on the productive fabric" (ID 16618), project name "Bridging big omic, genetic and medical data for Precision Medicine implementation in Greece" (project code TAEDR-0539180).

References

[1] J. Winter, S. Jung, S. Keller, R. I. Gregory, and S. Diederichs, "Many roads to maturity: microrna biogenesis pathways and their regulation." *Nature cell biology*, vol. 11, no. 3, pp. 228–34, 2009.

[2] D. P. Bartel, "Micrornas: genomics, biogenesis, mechanism, and function." *Cell*, vol. 116, no. 2, pp. 281–97, 2004.

[3] R. C. Lee, R. L. Feinbaum, and V. Ambros, "The c. elegans heterochronic gene lin-4 encodes small rnas with antisense complementarity to lin-14." *Cell*, vol. 75, no. 5, pp. 843–54, 1993.

[4] M. Sung, M. Jeong, Y. Choi, D. Kim, J. Lee, and J. Kang, "Bern2: an advanced neural biomedical named entity recognition and normalization tool." *Bioinformatics (Oxford, England)*, vol. 38, no. 20, pp. 4837–4839, 2022.

[5] K. M. Hettne, R. H. Stierum, M. J. Schuemie, P. J. M. Hendriksen, B. J. A. Schijvenaars, E. M. v. Mulligen, J. Kleinjans, and J. A. Kors, "A dictionary to identify small molecules and drugs in free text." *Bioinformatics (Oxford, England)*, vol. 25, no. 22, pp. 2983–91, 2009.

[6] R. Leaman and G. Gonzalez, "Banner: an executable survey of advances in biomedical named entity recognition." *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, vol. Unknown Volume, pp. 652–63, 2008.

[7] X. Wang, "Rule-based protein term identification with help from automatic species tagging," in *Computational Linguistics and Intelligent Text Processing*, A. Gelbukh, Ed. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, pp. 288–298.

[8] C.-H. Wei, A. Allot, R. Leaman, and Z. Lu, "Pubtator central: automated concept annotation for biomedical full text articles." *Nucleic acids research*, vol. 47, no. W1, pp. W587–W593, 2019.

[9] E. W. Sayers, E. E. Bolton, J. R. Brister, K. Canese, J. Chan, D. Comeau, R. Connor, K. Funk, C. Kelly, S. Kim, T. Madej, A. Marchler-Bauer, C. Lanczycki, S. Lathrop, Z. Lu, F. Thibaud-Nissen, T. Murphy, L. Phan, Y. Skripchenko, T. Tse, J. Wang, R. Williams, B. Trawick, K. Pruitt, and S. Sherry, "Database resources of the national center for biotechnology information," *Nucleic Acids Research*, vol. 50, no. D1, pp. D20–D26, 12 2021. [Online]. Available: https://doi.org/10.1093/nar/gkab1112

[10] Y. Zhang, H. Lin, Z. Yang, J. Wang, Y. Sun, B. Xu, and Z. Zhao, "Neural network-based approaches for biomedical relation classification: A review," *Journal of Biomedical Informatics*, vol. 99, 2019.

[11] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," 2019.

[12] A. Dunn, J. Dagdelen, N. Walker, S. Lee, A. S. Rosen, G. Ceder, K. Persson, and A. Jain, "Structured information extraction from complex scientific text with fine-tuned large language models," 2022.

[13] D. Xu, W. Chen, W. Peng, C. Zhang, T. Xu, X. Zhao, X. Wu, Y. Zheng, and E. Chen, "Large language models for generative information extraction: A survey," 2023.

[14] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang, "Biobert: a pre-trained biomedical language representation model for biomedical text mining." *Bioinformatics (Oxford, England)*, vol. 36, no. 4, pp. 1234–1240, 2020.

[15] P. Qi, Y. Zhang, Y. Zhang, J. Bolton, and C. D. Manning, "Stanza: A python natural language processing toolkit for many human languages," 2020.

[16] P. W. Harrison, M. R. Amode, O. Austine-Orimoloye, A. Azov, M. Barba, I. Barnes, A. Becker, R. Bennett, A. Berry, J. Bhai, S. K. Bhurji, S. Boddu, P. R. Branco Lins, L. Brooks, S. Ramaraju, L. Campbell, M. C. Martinez, M. Charkhchi, K. Chougule, A. Cockburn, C. Davidson, N. De Silva, K. Dodiya, S. Donaldson, B. El Houdaigui, T. Naboulsi, R. Fatima, C. G. Giron, T. Genez, D. Grigoriadis, G. Ghattaoraya, J. G. Martinez, T. Gurbich, M. Hardy, Z. Hollis, T. Hourlier, T. Hunt, M. Kay, V. Kaykala, T. Le, D. Lemos, D. Lodha, D. Marques-Coelho, G. Maslen, G. Merino, L. Mirabueno, A. Mushtaq, S. Hossain, D. Ogeh, M. P. Sakthivel, A. Parker, M. Perry, I. Piližota, D. Poppleton, I. Prosovetskaia, S. Raj, J. Pérez-Silva, A. Salam, S. Saraf, N. Saraiva-Agostinho, D. Sheppard, S. Sinha, B. Sipos, V. Sitnik, W. Stark, E. Steed, M.-M. Suner, L. Surapaneni, K. Sutinen, F. F. Tricomi, D. Urbina-Gómez, A. Veidenberg, T. A. Walsh, D. Ware, E. Wass, N. Willhoft, J. Allen, J. Alvarez-Jarreta, M. Chakiachvili, B. Flint, S. Giorgetti, L. Haggerty, G. Ilsley, J. Keatley, J. Loveland, B. Moore, J. Mudge, G. Naamati, J. Tate, S. Trevanion, A. Winterbottom, A. Frankish, S. E. Hunt, F. Cunningham, S. Dyer, R. Finn, F. Martin, and A. Yates, "Ensembl 2024," *Nucleic Acids Research*, vol. 52, no. D1, pp. D891–D899, 11 2023. [Online]. Available: https://doi.org/10.1093/nar/gkad1049

[17] A. Kozomara, M. Birgaoanu, and S. Griffiths-Jones, "miRBase: from microRNA sequences to function," *Nucleic Acids Research*, vol. 47, no. D1, pp. D155–D162, 11 2018. [Online]. Available: https://doi.org/10.1093/nar/gky1141

[18] G. Skoufos, P. Kakoulidis, S. Tastsoglou, E. Zacharopoulou, V. Kotsira, M. Miliotis, G. Mavromati, D. Grigoriadis, M. Zioga, A. Velli, I. Koutou, D. Karagkouni, S. Stavropoulos, F. S. Kardaras, A. Lifousi, E. Vavalou, A. Ovsepian, A. Skoulakis, S. K. Tasoulis, S. V. Georgakopoulos, V. P. Plagianakos, and A. G. Hatzigeorgiou, "TarBase-v9.0 extends experimentally supported miRNA–gene interactions to cell-types and virally encoded miRNAs," *Nucleic Acids Research*, vol. 52, no. D1, pp. D304–D310, 11 2023. [Online]. Available: https://doi.org/10.1093/nar/gkad1071

[19] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, D. Bikel, L. Blecher, C. C. Ferrer, M. Chen, G. Cucurull, D. Esiobu, J. Fernandes, J. Fu, W. Fu, B. Fuller, C. Gao, V. Goswami, N. Goyal, A. Hartshorn, S. Hosseini, R. Hou, H. Inan, M. Kardas, V. Kerkez, M. Khabsa, I. Kloumann, A. Korenev, P. S. Koura, M.-A. Lachaux, T. Lavril, J. Lee, D. Liskovich, Y. Lu, Y. Mao, X. Martinet, T. Mihaylov, P. Mishra, I. Molybog, Y. Nie, A. Poulton, J. Reizenstein, R. Rungta, K. Saladi, A. Schelten, R. Silva, E. M. Smith, R. Subramanian, X. E. Tan, B. Tang, R. Taylor, A. Williams, J. X. Kuan, P. Xu, Z. Yan, I. Zarov, Y. Zhang, A. Fan, M. Kambadur, S. Narang, A. Rodriguez, R. Stojnic, S. Edunov, and T. Scialom, "Llama 2: Open foundation and fine-tuned chat models," 2023.

Computational and Structural Biotechnology Journal 27 (2025) 4802–4812

Contents lists available at ScienceDirect

Computational and Structural Biotechnology Journal

journal homepage: www.elsevier.com/locate/csbj

Software/web server article

Public Omics Explorer (POE): Enabling integrative semantic search across GEO omics datasets based on PubMed publications

Dimitris Grigoriadis [a,*], Margaritis Tsifintaris [b], Antonis Giannakakis [b], Georgios A. Pavlopoulos [c,d,**,1], Nikos Perdikopanis [e,*,1]

[a] Department of Bioinformatics, Genekor Medical S.A, Athens 15344, Greece
[b] Department of Molecular Biology and Genetics, Democritus University of Thrace, Alexandroupolis 68100, Greece
[c] Institute for Fundamental Biomedical Research, BSRC "Alexander Fleming", Vari 16672, Greece
[d] Department of Computational Biology, Mohamed bin Zayed University of Artificial Intelligence (MBZUAI), Abu Dhabi, United Arab Emirates
[e] Department of Informatics and Telecommunications, National and Kapodistrian University of Athens, Athens, Greece

ARTICLE INFO

Keywords:
Semantic search
GEO datasets
Omics datasets
PubMed integration
BERT embeddings
Dataset discovery

ABSTRACT

The exponential growth of publicly available omics datasets and biomedical literature has created both opportunities and challenges for data-driven discovery in life sciences. While the Gene Expression Omnibus (GEO) hosts millions of high-throughput experimental datasets, the European Nucleotide Archive (ENA) stores the corresponding raw sequencing data, and PubMed contains an extensive body of related scientific publications, integrated exploration of these resources remains limited. We present Public Omics Explorer (POE), a web-based platform that performs literature-informed dataset retrieval, semantically linking GEO datasets and ENA records through their associated PubMed publications. POE automatically collects and indexes GEO metadata, ENA cross-references, and PubMed abstracts on a daily basis. For semantic embedding, POE employs the biomedical-specialized SBioBERT model, which generates dense vector representations from publication text. These embeddings are indexed using Facebook AI Similarity Search (FAISS) to enable high-precision, context-aware retrieval. Users can search using free-text natural language queries, which are processed through semantic search to identify conceptually relevant datasets based on linked publication content. Structured filters allow refinement by organism, experiment type, library strategy, sample type, extracted molecule, and publication year. In addition to semantic queries, POE supports direct retrieval of datasets via accession identifiers (GSE IDs, PubMed IDs, DOIs) and offers a programmatic RESTful API for integration into computational pipelines and automated workflows. By linking processed data in GEO with raw data in ENA through shared publication context, POE facilitates hypothesis generation, meta-analysis, and exploratory research. The application is freely available at https://nplab.gr/poe.

1. Introduction

In the era of multi-omics research, repositories such as the NCBI Gene Expression Omnibus (GEO) [1] aggregate massive volumes of data. As of September 2025, GEO hosts more than 9.6 million samples and is expanding exponentially. However, despite this wealth, discovering datasets relevant to a specific biological query remains challenging. The core obstacle is the heterogeneity and often unstructured nature of metadata, which are typically expressed as free text using inconsistent vocabularies, despite the existence of well-documented submission guidelines since 2002 [2] and the establishment of the FAIR principles (Findable, Accessible, Interoperable, Reusable) [3].

Early efforts to improve metadata consistency focused on manual curation. Notable examples include the BioSamples [4] and METAGENOTE [5] databases, as well as tools like ArcheGEO [6] GEOmetacuration [7] and Zooma (https://www.ebi.ac.uk/spot/zooma/). The CEDAR Workbench [8] further supports the creation of semantically structured metadata templates based on biomedical ontologies [9].

* Corresponding authors.
** Corresponding author at: Institute for Fundamental Biomedical Research, BSRC "Alexander Fleming", Vari 16672, Greece.
E-mail addresses: d.grigoriadis@genekor.com (D. Grigoriadis), pavlopoulos@fleming.gr (G.A. Pavlopoulos), nikosp@uoa.gr (N. Perdikopanis).
[1] These authors contributed equally as senior/last authors.

https://doi.org/10.1016/j.csbj.2025.11.004
Received 20 September 2025; Received in revised form 31 October 2025; Accepted 1 November 2025
Available online 3 November 2025

While these initiatives have improved metadata quality and searchability, their reliance on human effort limits scalability and long-term sustainability.

To address these constraints, automated approaches have been developed using heuristics, rule-based logic, and machine learning. Tools such as metaSRA [10], ALE [11] and ReGEO [12] utilize machine learning techniques to infer ontology terms or extract semantic structure. Others like tidyGEO [13] employ rule-based or interactive transformation methods to convert free-text metadata into structured, ontology-aligned fields. Complementary platforms such as GeoQuery [14], GEOmetadb [15], SRA-Explorer (https://sra-explorer.info/), and SpiderSeqR [16] facilitate programmatic or SQL-like access to GEO, improving access to curated data subsets.

In parallel, an alternative line of tools focuses not on metadata but on expression-level similarity. Signature-based search tools, including LCE (Lung Cancer Explorer) [17] Expression Atlas [18], LINCS [19], and ARCHS4 [20], support dataset retrieval based on gene expression profiles. Broader platforms like OmicsDI [21] and DisGeNET [22] support retrieval of datasets based on phenotype similarity, genetic variants, or interaction networks.

Advances in semantic search have enabled more meaningful retrieval, moving beyond keyword-based approaches to techniques that consider user intent and contextual understanding. Tools like PubTator3 [23], LitCovid + GEO [1], and LitSense [24] facilitate literature-level search using semantic embeddings or concept-based representations. In addition, BioASQ [25] offers a large-scale biomedical question-answering and information retrieval benchmark, focusing on semantic matching between PubMed abstracts and ontology-derived concepts. While these tools have advanced semantic exploration of biomedical text, they operate exclusively at the publication level, without linking retrieved articles to underlying experimental datasets. Related infrastructures, including EBI BioStudies [26], Semantic Scholar (https://www.semanticscholar.org/,) and TogoGenome [27], enrich dataset and entity records with semantic annotations linked to biomedical ontologies such as Medical Subject Headings (MeSH) and Unified Medical Language System (UMLS).

Despite the often explicit linkage between GEO datasets and associated publications in PubMed or PMC, few tools have systematically used the content of these articles as a primary source of semantic information for dataset discovery. Biochat [28], a prototype released as a preprint on bioRxiv, attempted to calculate similarity between GEO datasets using vector representations of associated article abstracts. However, it was never peer-reviewed and the service is now defunct. Similarly, LitCovid + GEO relied on keyword matching and manual curation, while PubTator3 and LitSense focused on extracting and indexing biomedical entities from literature without integrating dataset retrieval capabilities. Even official PubMed-GEO linkouts remain underutilized in semantic or conceptual search frameworks.

To address these limitations, we developed the Public Omics Explorer (POE) - a web application that enables semantic search across GEO datasets based on PubMed publications. POE systematically processes the titles and abstracts of articles linked to GEO datasets as input for publication-content-aware semantic search. It constructs a unified embedding space using SBioBERT, a biomedical-specialized transformer model pretrained on large-scale biomedical corpora to effectively capture domain-specific terminology and context. The model generates dense vector representations of each article's content, which are indexed using Facebook AI Similarity Search (FAISS) to enable high-precision, context-aware retrieval of semantically related literature and associated datasets.

In addition to semantic search, POE offers structured filtering across multiple metadata dimensions, including organism, experiment type, library strategy, sample type, extracted molecule, and publication year. For users with prior knowledge of specific datasets or publications, the platform supports targeted retrieval via accession identifiers such as GSE IDs, PubMed IDs, or DOIs, providing direct links to the corresponding records hosted in external repositories like GEO and ENA [29] and their mirrored SRA records.

POE also includes a RESTful API and programmatic interface that allow integration into external pipelines, applications, or institutional infrastructures, supporting scalable and reproducible workflows in computational biology. Importantly, POE does not store or replicate omics datasets locally, but instead facilitates semantic access and redirection to their sources. The platform is updated daily through automated scheduling, ensuring that new GEO metadata ENA/SRA cross-references and PubMed records are continuously incorporated into the semantic index.

2. Methods

2.1. System overview and accessibility

POE is hosted as an online platform accessible through modern web browsers without the need for installation or user-side configuration (Fig. 1). All computational infrastructure, including metadata ingestion, semantic embedding, indexing, and result delivery, is maintained on the server side.

2.2. Data sources and acquisition

POE integrates data from three major biomedical resources: GEO, PubMed, and the European Nucleotide Archive (ENA) [29]. GEO metadata is retrieved daily through NCBI's FTP server and E-utilities interface. During each cycle, POE downloads and parses metadata for both study-level records (GSE) and sample-level records (GSM), extracting key fields such as organism, library strategy, experiment type, sample type, and molecule extracted. This metadata forms the core of the system's searchable omics dataset repository (Fig. 2).

PubMed articles linked to GEO datasets are also collected during this process. When GSE records contain associated PubMed identifiers (PMIDs), the corresponding publications are programmatically retrieved. Metadata is extracted using a hybrid strategy that includes both NCBI API access and HTML parsing with the BeautifulSoup library [30]. For each article, POE collects the title, abstract, MeSH terms, author list, affiliations, publication dates, reported substances, and a list of similar articles. This comprehensive metadata profile is used to semantically index the literature and establish links to relevant omics datasets. Additionally, POE integrates metadata from the ENA, particularly to retrieve FASTQ file links associated with GEO datasets.

2.3. Data processing and integration pipeline

The integration of omics datasets with literature information in POE is managed by a custom-developed, fully automated pipeline executed daily. This pipeline updates the GEO and PubMed metadata stores, links GSE entries to associated publications, generates vector embeddings for newly indexed abstracts, and constructs a FAISS index to support semantic search.

For each new or updated GSE entry, the system attempts to map a PubMed article using the provided PMID. If a valid link is found, the corresponding publication metadata is fetched and associated with the dataset. Entries without a valid PubMed link are retained in the GEO index but excluded from semantic embedding. All structured metadata is normalized and stored in a PostgreSQL database. The schema is designed to be flexible and extensible, enabling advanced filtering across heterogeneous metadata fields and seamless integration with the semantic search backend.

2.4. Comparative evaluation of POE and standard GEO keyword search

Following the description of POE's architecture and data integration pipeline, we conducted a systematic evaluation to quantify its

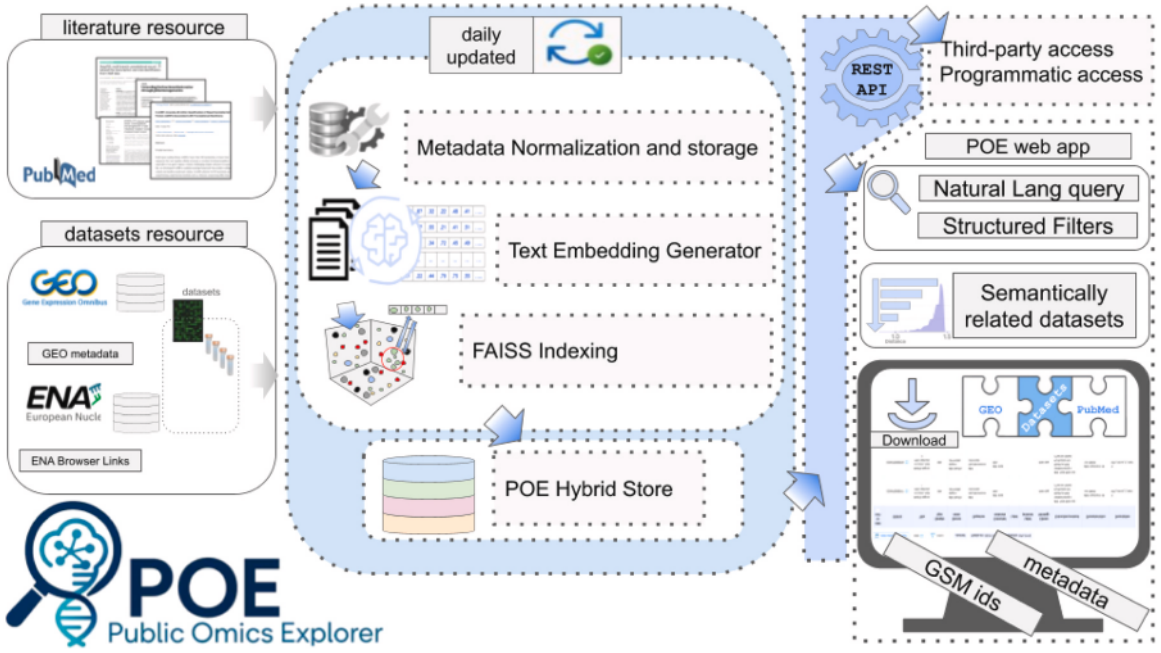

Fig. 1. POE High-level system overview. POE integrates literature and dataset resources by systematically linking PubMed abstracts with GEO metadata and ENA cross-references. Metadata are normalized, embedded with SBioBERT, and indexed with FAISS to enable semantic similarity search. All information is stored in a hybrid database that combines metadata, embeddings, and dataset links. The system is automatically updated on a daily basis. Users interact with POE primarily through the web platform, which supports natural language queries, semantic search, structured filters, and dataset export. For programmatic integration, POE also provides a REST API for third-party access.

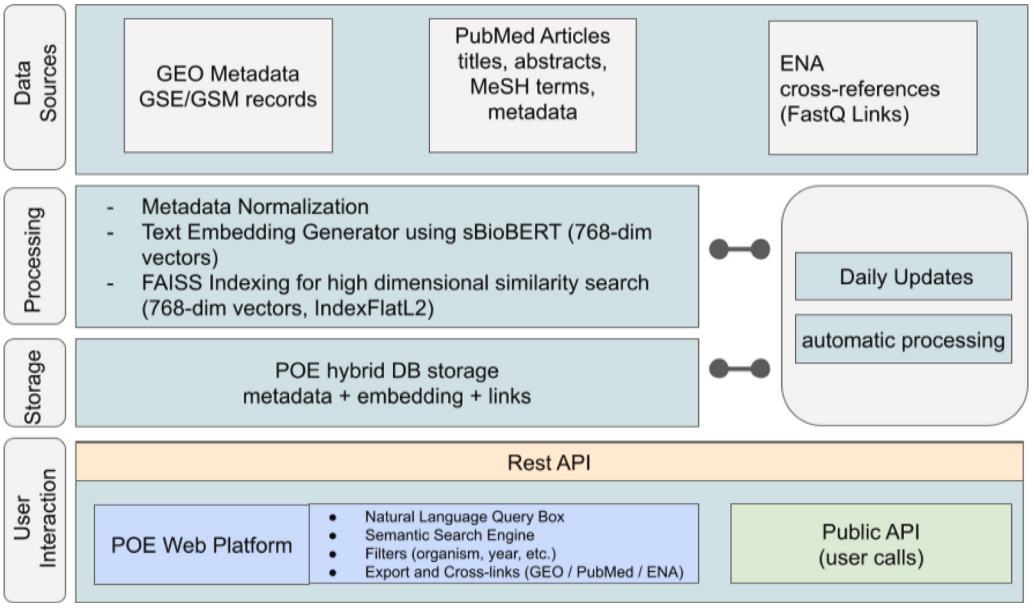

Fig. 2. Technical workflow of the POE. POE integrates GEO, PubMed, and ENA metadata into a hybrid database that stores metadata, semantic embeddings, and dataset links. Publication text is embedded with SBioBERT (768-dimensional vectors) and indexed with FAISS to enable high-dimensional similarity search. All access to the system is mediated through a central REST API. The POE Web Platform interacts with this API via a private internal interface to deliver a user-friendly environment for natural language queries, semantic search, structured filters, and dataset export with cross-links to GEO, PubMed, and ENA. In addition, a public REST API is available for external users and programmatic integration into third-party workflows, extending POE's reach beyond the web interface.

performance relative to standard keyword-based GEO searches. The goal of this evaluation was to assess the added value of semantic search for dataset discovery. To ensure comparability, the analysis was restricted to PubMed-linked GEO datasets, which represent the subset directly supported by POE's indexing strategy. We compared POE's semantic search results with those retrieved from the GEO Datasets and GEO Profiles databases using an identical set of 30 natural-language biological queries. (see Supplementary Table 1).

These queries were organized into five thematic categories: a)*Epigenetics, b)Immune regulation and cancer, c)Neurological disease, d)Non-coding RNA biology* and e)*Regulatory genomics and stem-cell development*-to capture a broad spectrum of biological contexts.

Representative examples include *"chromatin accessibility in stem-cell fate decisions"* (Epigenetics), *"neuroinflammation signatures in*

Alzheimer's disease" (Neurological disease), and *"miRNA promoter variants in cancer"* (Non-coding RNA biology).

Each query was formulated manually by two domain experts to reflect realistic biological research questions that scientists might pose when exploring public omics data.

Queries were expressed in natural-language form, typically as concise phrases of 4–10 words, without Boolean operators or database-specific syntax.

Topics were selected to represent a balance between well-studied and emerging biological areas, covering diseases, molecular mechanisms, and regulatory processes.

This formulation ensured that the benchmark reflected authentic user behavior and evaluated POE's strength in handling open-ended, concept-level queries rather than keyword matching.

This diversity ensured that the evaluation reflected dataset discovery across both established and emerging areas of biomedical research.

2.4.1. Quantitative analysis

For each biological query, identical searches were executed in both systems. Detailed results of each system can be found in Supplementary material 1. The GEO search results were subsequently partitioned into two categories: (i) datasets associated with a PubMed identifier and (ii) datasets lacking a PubMed identifier. Since POE exclusively indexes datasets linked to PubMed publications, the comparative analysis was restricted to the PubMed-linked subset of GEO results.

Formally, for each query *i let:*

$G_i = \{ g \in \text{GEO results} \mid g \text{ is associated with a PubMed identifier} \}$

denote the set of PubMed-linked GEO datasets, and

$\mathbf{P_i} = \{ \mathbf{p} \in \mathbf{POE\ results} \}$

denote the corresponding POE results.

Since POE typically yields a larger number of results than GEO ($|P_i|>|G_i|$), we applied a truncation procedure to ensure equal-sized sets for pairwise comparison. Specifically, we defined

$$P_i^* = top_{|Gi|}(P_i)$$

where $top_{|Gi|}(P_i)$ corresponds to the $|G_i|$ elements of P_i, ranked according to their semantic similarity scores to the query.

All subsequent analyses were therefore performed on the aligned pairs (G_i, P_i^*). To further characterize the relative contributions of each system, we defined:

$$\text{GEO-unique}_i = G_i \setminus P_i^*,\ \text{POE-unique}_i = P_i^* \setminus Gi,\ \text{Intersection}_i = G_i \cap P_i^*$$

These categories respectively capture datasets identified only by GEO, only by POE, and by both systems for query i. The relative sizes of these sets were used to quantify overlap and divergence between the two search strategies. (see Supplementary Table 2).

2.4.2. Qualitative relevance assessment

To complement this overlap analysis, we performed a manual evaluation of the unique results from each system, that is, datasets belonging to GEO-unique or POE-unique. Each PubMed-linked dataset in these categories was scored for relevance by evaluating the associated publication title and abstract on a 1–10 scale. For statistical robustness, we restricted this evaluation to queries with more than four unique datasets, which resulted in the inclusion of 9 out of the 30 queries (see supplementary table 3).

Scoring followed a structured rubric comprising five criteria:

- **C1: Conceptual match,** alignment between the query concept and the study focus (2 = direct, 1 = partial, 0 = keyword-only).
- **C2: Biological context,** correspondence of species, tissue, disease, or biological system (2 = full match, 1 = partial, 0 = mismatch). **C3: Methods/data fit,** appropriateness of the experimental methods and data type to the query (2 = ideal, 1 = related, 0 = irrelevant).
- **C4: Contribution of findings,** degree to which the study provides insights directly addressing the query (2 = direct, 1 = indirect/limited, 0 = none/theoretical).
- **C5: Focus/explicitness,** whether the query topic was explicitly targeted versus only peripheral within a broader theme (2 = explicit, 1 = broader, 0 = peripheral).

The sum of these five criteria yielded a relevance score ranging from 0 (no relevance) to 10 (highly relevant). This rubric enabled a structured and reproducible comparison of the scientific relevance of results obtained through semantic search (POE) versus keyword-based search (GEO).

Manual relevance scoring was independently performed by two evaluators with expertise in bioinformatics, who rated each retrieved dataset on a 1–10 relevance scale according to its biological connection to the query.

Inter-rater agreement was assessed using Cohen's κ ($\kappa = 0.79$), indicating substantial consistency between evaluators.

A two-tailed Wilcoxon signed-rank test comparing the total scores of the two experts confirmed that their ratings did not differ significantly ($p = 0.097 > 0.05$), demonstrating strong agreement and reliability in the manual assessments.

Consequently, the final relevance score for each dataset was calculated as the mean of both expert ratings.

To evaluate the significance of the observed differences between POE and GEO relevance scores across the benchmark queries, a two-tailed Wilcoxon signed-rank test was performed.

This non-parametric test was chosen for its robustness to non-normal distributions of ordinal data and revealed that POE achieved statistically higher relevance than GEO ($p < 0.05$).

These analyses enhance the methodological transparency and statistical rigor of the evaluation framework.

2.5. Embedding generation and vector representation

To support semantic search across the literature, POE incorporates a transformer-based embedding model implemented via the Sentence-Transformers library [31]. Specifically, POE employs SBioBERT, a biomedical-specialized model pretrained on large-scale corpora including PubMed abstracts and PMC full-text articles, with additional sentence-level fine-tuning to capture domain-specific terminology and context more effectively. SBioBERT generates 768-dimensional dense vector embeddings optimized for biomedical semantic similarity (Table 1). The model is available via Hugging Face: https://huggingface.co/dmis-lab/sentence-biobert.

SBioBERT was selected as the backbone embedding model of POE due to its strong performance in representing biomedical concepts, enabling accurate retrieval of semantically related literature and associated datasets. Through the web interface or RESTful API, users can seamlessly query the system using natural language, benefiting from SBioBERT's ability to capture fine-grained biomedical context.

Before encoding, each PubMed title and abstract is concatenated, lowercased, stripped of non-informative characters, and truncated or padded to a maximum of 256 tokens. The resulting embeddings are stored and indexed using the FAISS library, enabling fast and meaningful retrieval of semantically similar content tailored to biomedical abstracts.

2.6. Semantic search and indexing

The 768-dimensional SBioBERT embeddings form the basis of POE's semantic index, which is implemented using FAISS to enable fast and efficient retrieval. The index currently uses the IndexFlatL2 configuration, which performs exact nearest-neighbor search with perfect recall

Table 1
SBioBERT embedding model used in POE. SBioBERT generates 768-dimensional embeddings trained on biomedical corpora, optimized for fine-grained semantic search in specialized biomedical queries.

Model name	Embedding size	Training data sources	Domain focus	Recommended use cases
SBioBERT	768	PubMed abstracts, PMC articles, and specialized biomedical datasets with sentence-level fine-tuning	Biomedical domain (sentence-level tasks)	Fine-grained biomedical concept matching, narrow terminology searches

and precision, serving as the ground truth benchmark. While IndexFlatL2 provides the highest retrieval accuracy, it requires relatively higher memory usage and query time.

To evaluate trade-offs in performance, we benchmarked alternative FAISS index types, including IndexIVFFlat, IndexPQ, and IndexHNSW. IndexIVFFlat leverages an inverted file structure to reduce the search space, offering faster query times and lower memory requirements at the cost of approximate results. IndexPQ applies product quantization to compress vector representations, providing substantial memory and time efficiency but with a noticeable drop in retrieval quality. IndexHNSW employs a hierarchical navigable small-world graph that balances speed and accuracy, enabling efficient approximate search with moderate resources.

Although SBioBERT was originally optimized for sentence-level semantic understanding, in POE it is applied to variable-length biomedical text segments, including both PubMed abstracts and user-submitted queries.

To handle multi-sentence queries, all sentences are concatenated and processed as a single input sequence with a maximum token length of 256, allowing cross-sentence dependencies to be captured within a unified contextual window.

For multiword biomedical entities (e.g., Parkinson's disease, oxidative stress response, transcription factor binding sites), SBioBERT's WordPiece tokenizer decomposes complex terms into subword units while preserving their semantic integrity.

Embeddings corresponding to tokens belonging to the same entity or query are aggregated using mean pooling, generating a dense, context-aware representation that retains information about compositional biomedical terms.

This implementation enables POE to adapt SBioBERT beyond its original sentence-level scope, providing robust semantic encoding for realistic, multi-sentence biomedical queries encountered in literature-informed dataset retrieval.

The following table summarizes performance across these index types:

We systematically assessed these configurations by executing 100 representative semantic search queries derived from biomedical abstracts. For each query, the top 10 nearest neighbors were retrieved, and Recall@10 and Precision@10 were computed by comparing approximate methods against the IndexFlatL2 baseline. Additionally, average query time and memory usage were measured to reflect realistic usage scenarios.

As summarized in Table 2, approximate methods offered substantial gains in speed and memory efficiency, but IndexFlatL2 consistently achieved the highest accuracy and most reliable semantic retrieval. Given the importance of precise dataset matching in biomedical discovery, IndexFlatL2 was selected as the default index type in POE.

Table 2
Performance comparison of FAISS index types used in POE. IndexFlatL2 provides exact nearest-neighbor search with perfect recall and precision but requires higher memory and query time. Approximate methods (IVFFlat, PQ, HNSW) offer faster performance and reduced memory usage with varying trade-offs in retrieval quality. For biomedical dataset discovery, where accuracy is critical, POE uses IndexFlatL2 as the default configuration.

Index type	Memory Usage (MB)	Avg. Query Time (ms)	Recall@10 (%)	Precision@10 (%)
IndexFlatL2	150	15–20	100	100
IndexIVFFlat	40–50	1–5	90	88
IndexPQ	10–15	1–3	84	78
IndexHNSW	70–90	2–6	92	90

All benchmarking was conducted on a machine with CPU Intel Core i3–10100 CPU @ 3.60 GHz, 8 cores, 64 GB RAM, using a dataset with more than 250,000 GEO-linked articles.

2.7. Backend infrastructure

POE's backend is implemented in Python using the FastAPI web framework. Metadata, records, and vector embeddings are stored in a PostgreSQL database with the pgvector extension enabled. All sensitive information, including user credentials and API tokens, is securely managed using hashed passwords.

The semantic search engine is integrated with the database and embedding pipeline, such that a user's query is embedded into the same 768-dimensional space and used to identify semantically similar PubMed abstracts through FAISS. These results are then used to identify and rank associated GEO datasets. The entire backend service is deployed with unicorn, providing asynchronous I/O and fast response times for search queries and API endpoints.

2.8. Front-end and user interface

The POE user interface is implemented as a single-page web application using React (https://react.dev/reference/react), TypeScript [32], and Tailwind CSS (https://tailwindcss.com/docs/). It provides a responsive and accessible experience optimized for both desktop and mobile environments. The frontend communicates with the backend exclusively through RESTful API endpoints.

Users can submit natural language queries, browse semantically matched PubMed articles, and explore linked datasets. Structured filtering is available across multiple metadata fields, including organism, library strategy, experiment type, and publication year. Each result item is enriched with cross-links to external resources, including GEO, PubMed, and ENA, and may be exported in tabular format for downstream analysis. In addition, ENA FASTQ file links retrieved during data acquisition are presented directly in the interface, enabling users to access raw sequencing data alongside processed dataset metadata.

The current version of POE provides interactive result filtering, and summary statistics summarizing dataset counts, organism distribution, and experiment types retrieved per query.

It also features an interactive histogram of similarity scores, displaying the geometric distance (L2) of each retrieved result from the query embedding, allowing users to visually assess the semantic proximity of datasets and publications.

These elements collectively support intuitive data exploration and rapid evaluation of search relevance.

2.9. Collection basket for dataset management

POE includes an interactive "collection basket" feature that enables users to curate and temporarily store selected datasets during a search session. Datasets can be added to the basket from any search results page or filtered view. This functionality allows researchers to compile relevant datasets from multiple queries before initiating downloads or exporting metadata. The basket supports exporting the list of selected

GEO or ENA accessions in standard formats, facilitating integration with external analysis pipelines or download managers.

2.10. Programmatic access via RESTful API and downstream analysis

In addition to the web-based user interface, POE provides a secure, token-based RESTful API that enables programmatic access to its semantic search capabilities and linked dataset records. All requests require a valid bearer token in the Authorization header; tokens expire after 24 h, and the API is rate-limited to one request every four seconds. Available endpoints include:

(i) Semantic PubMed search, which uses FAISS-based vector similarity to retrieve publications relevant to a free-text query with optional metadata filters (organism, library strategy, experiment type, molecule, publication year).
(ii) Direct dataset/article lookup, allowing retrieval of metadata for specific GEO Series (GSE) accessions, PubMed IDs, DOIs, or article titles; and
(iii) ENA metadata download, which returns a tab-delimited file containing metadata and raw data file links for a given ENA project accession.

Python code examples are provided in the online documentation, enabling integration of POE's capabilities into external pipelines, reproducible workflows, and automated dataset publication matching.

API documentation is available at https://github.com/DimitrisGrig/POE/wiki/API.

Beyond programmatic querying, POE results can be exported in standardized TSV or JSON formats, which preserve key metadata fields such as PubMed identifiers, GEO accessions, and ENA project links.

These outputs can be readily parsed in analytical environments such as R or Python using libraries like pandas for downstream filtering, aggregation, and study selection.

The retrieved dataset identifiers can then be used to fetch associated expression matrices or sequencing files through GEOquery, GEOparse, or ENA utilities, enabling subsequent analysis steps such as differential expression, meta-analysis, or data reprocessing.

In this way, POE acts as a semantic discovery layer that guides researchers toward relevant datasets, which can then be programmatically integrated into established analytical pipelines.

2.11. Tool and library versioning

POE was developed using Python 3.12.9. The backend employs FastAPI v0.115.11 for API management and Uvicorn v0.34.0 as the asynchronous web server. Semantic embeddings are generated using sentence-transformers v3.4.1 with the SBioBert model [33]. FAISS-cpu v1.10.0 is used for similarity indexing, and PostgreSQL v17.4 serves as the metadata database. The frontend interface is built with React v19.0.0, TypeScript v5.7.2, and Tailwind CSS v4.0.9, ensuring a modern and responsive user experience.

2.12. System update schedule

POE is updated daily via automated schedulers that orchestrate the end-to-end data retrieval, embedding, indexing, and deployment processes.

3. Results

3.1. Platform overview and dataset coverage

Public Omics Explorer (POE) is a fully operational web application that enables semantic exploration of omics datasets through their linked biomedical literature.As of September 2025, POE indexes a total of 262,173 GEO datasets (GSE records) and 9682,571 sample-level entries (GSM), integrating them with a collection of 130,548 PubMed publications. All GEO entries with valid PubMed identifiers are indexed with semantic embeddings, enabling vector-based similarity search from text queries to experimental datasets.

Users interact with the platform through a user-friendly interface (available at https://nplab.gr/poe) that requires no installation or authentication (Fig. 3). The POE web interface was iteratively refined through informal testing with graduate students and researchers in molecular biology and bioinformatics. User feedback emphasized clarity of search results, intuitive filtering controls, and smooth navigation between literature and dataset views.

These sessions confirmed that POE can be effectively used by non-computational scientists to explore datasets starting from natural-language queries.

Queries can be initiated using free-text natural language input, which is embedded into a shared semantic space alongside PubMed titles and abstracts. POE retrieves the most semantically similar articles, each linked to associated GEO datasets, enriched metadata, and direct links to GEO, PubMed, and ENA records.

At any point, users can refine their results through structured filters that include organism, sample type, library strategy, experiment type, publication year, and molecule extracted. Each result is displayed with summarized metadata and export options, allowing easy transition from conceptual queries to actionable datasets. Additionally, results can be further filtered based on a visually adjustable L2 similarity threshold.

3.2. Literature-informed dataset retrieval

Unlike conventional keyword-based search engines, POE supports concept-level discovery by leveraging the scientific context of publications. For example, a user query such as "inflammation in Parkinson's disease" (see Supplementary Table 4 for POE result set) retrieves PubMed abstracts discussing neuroinflammatory pathways and neurodegeneration. From these, POE surfaces semantically associated GEO datasets, even when the metadata of those datasets do not explicitly mention either "Inflammation" or "Parkinson's disease."

This capability significantly enhances discovery by bridging the gap between implicit biological context in publications and structured experimental datasets. It enables researchers to identify datasets based on the relevance of the publication content, rather than reliance on exact metadata matches.

3.2.1. Comparative performance: keyword vs. semantic search

The comparative evaluation showed that POE and GEO returned overlapping but distinct sets of datasets. Notably, POE consistently retrieved substantially more results than GEO, reflecting the broader coverage enabled by its semantic indexing of PubMed-linked datasets. For instance, for *"chromatin accessibility in stem cell fate decisions,"* POE retrieved 506 datasets compared to just 6 PubMed-linked (7 total) datasets from GEO, while for *"epigenetic control of lineage commitment,"* POE returned 485 versus 10 from GEO. In several cases, GEO returned only a handful of results-for example, $N = 2$ in *"miRNA promoter variants in cancer"* and $N = 1$ for *"pluripotency transcriptional networks"* versus 784 and 116 from POE, respectively-underscoring the limitations of keyword-based matching for complex biological queries. This broader recall of POE is advantageous, as it expands the discovery space and reduces the risk of missing potentially relevant studies. In contrast, GEO's keyword-based search is inherently constrained, since it requires all query terms to appear explicitly in the text of dataset annotations (unless more complex Boolean expressions with AND/OR are manually constructed, which substantially complicates the search process). As a result, many relevant datasets are not retrieved if even a single term is absent. A summary of the results for each query, including the number of POE results (total and with distance ≤ 1.0), GEO results with and without PubMed IDs, common GSE identifiers, and unique datasets per method,

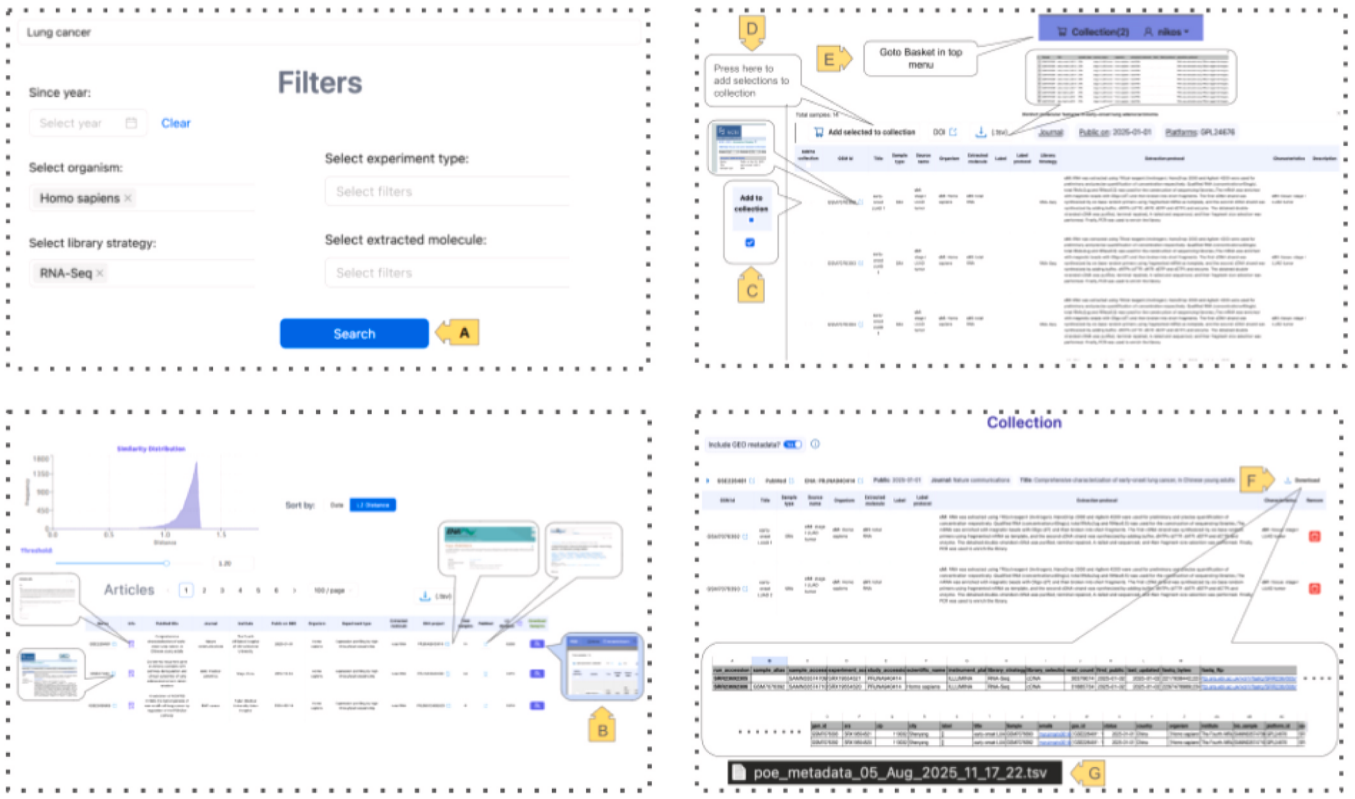

Fig. 3. POE semantic search interface (accessed via the "POE search" in the main menu). (A) Query construction through the web platform, with natural language input and structured filters (organism, experiment type, library strategy, and extracted molecule). (B) Search results are displayed as PubMed-linked articles with interactive previews, semantic similarity scores, and dataset associations. (C-D) Results can be organized into user-defined collections via the "Collection" menu. (E) Collected datasets and associated metadata are accessible in a dedicated collection view. (F-G) Users can export results, including PubMed-linked GEO datasets and metadata, in TSV format for downstream analysis.

is provided in Supplementary Table 2. The complete set of returned results for all queries across both systems is available in Supplementary_material_2.

The qualitative relevance assessment was performed on the unique datasets (counts shown in the last two columns of Supplementary Table 2) returned by each system, focusing on queries for which both POE and GEO contributed at least four unique results. In total, 9 queries met this criterion, ensuring that the comparison was based on a sufficiently large and balanced set of records to yield statistically meaningful insights. The qualitative scoring highlighted systematic differences between POE-unique and GEO-unique results (see Supplementary Table 5). On average, POE-unique datasets achieved higher relevance scores than GEO-unique datasets, indicating that semantic search preferentially retrieved studies more closely aligned with the biological intent of the queries. For example, in the case study queries *"chromatin accessibility in stem cell fate decisions"* and *"epigenetic control of lineage commitment,"* POE uniquely identified datasets where the associated publications explicitly investigated the query concepts (mean score 7 out of 10), while GEO returned superficially matched records with only marginal or incidental relevance (mean score 2.3). Even in more challenging queries, such as *"stress response pathways in cancer survival,"* POE still maintained a higher mean score (6.1 vs 5.0). This trend was consistent across most queries, demonstrating that POE not only increases recall but also reduces noise by surfacing qualitatively more relevant datasets that are often missed by keyword-only retrieval.

This issue is exemplified by the query *"inflammatory signaling in cardiovascular disease."* GEO returned a series of datasets unrelated to cardiovascular biology, such as *"Molecular Clusters and Tumor-Immune Drivers of IgM Monoclonal Gammopathies"* (focused on Waldenström macroglobulinemia), *"Differential DNA Methylation Associated with Multiple Sclerosis in an Underrepresented Population"* (neurological disease), and *"The effect of rewarming ischemia on tissue transcriptome signatures: a clinical observational study in lung transplantation."* These records were captured primarily because of incidental overlaps in generic terms such as "inflammatory" or "signaling," yet they lack any substantive connection to cardiovascular disease. Such mismatches illustrate the tendency of keyword-based retrieval in GEO to return biologically irrelevant datasets when queries involve complex, multi-word biological concepts.

In contrast, POE uniquely retrieved datasets directly aligned with the query's intent. Representative examples include *"Longitudinal profiling in patients undergoing cardiac surgery reveals postoperative changes in DNA methylation," "Identification of a gene network driving the attenuated response to lipopolysaccharide of monocytes from hypertensive coronary artery disease patients,"* and *"Whole-Blood Transcriptome Unveils Altered Immune Response in Acute Myocardial Infarction Patients With Aortic Valve Sclerosis."* These studies explicitly investigate inflammatory processes in cardiovascular contexts-ranging from cardiac surgery and hypertension to acute myocardial infarction-thereby providing highly relevant biological insights that are missed by simple keyword filtering.

This trend was consistent across most queries, demonstrating that POE not only increases recall but also reduces noise by surfacing qualitatively more relevant datasets that are often missed by keyword-only retrieval.

To further illustrate this difference in practical terms, the following case example highlights how POE enables cross-context discovery and hypothesis generation beyond explicit keyword matches.

3.2.2. Case example: cross-context discovery of mechanistic links (miR-21 in breast cancer)

To illustrate this difference, the query *"miR-21 breast cancer expression"* demonstrates POE's ability to uncover mechanistically related

datasets beyond the explicitly annotated disease context. While the GEO keyword search retrieved directly annotated datasets such as *GSE298584* (*"Fusobacterium nucleatum promotes metastasis of breast cancer via the miR-21–3p/FOXO3 axis"*), the POE semantic retrieval additionally surfaced functionally coherent studies including *GSE109592* (*miR-21/DUSP8 axis in colorectal carcinoma*) and *GSE117697* (*miR-21 in tumor-associated macrophages*).

Despite arising from distinct cancer models, these datasets converge on miR-21-mediated modulation of stromal and immune microenvironments, suggesting a shared pro-metastatic mechanism. This cross-context retrieval enabled the formulation of a new, data-driven hypothesis, that *miR-21 may promote breast cancer metastasis through conserved pathways regulating immune and stromal cell signaling*, similar to its role in other tumor types. Such examples highlight POE's capability to facilitate hypothesis generation by integrating semantically linked datasets across biological systems that traditional keyword search fails to associate.

3.2.3. Comparative context with existing semantic systems

To further contextualize POE within existing semantic search systems, we compared it with representative biomedical literature-based tools, including PubTator3, LitSense, and BioASQ.

These systems enable entity recognition, sentence-level similarity search, or question answering within the biomedical literature but do not link retrieved articles to experimental datasets.

In contrast, POE uniquely supports dataset-level semantic retrieval by connecting PubMed publications to GEO and ENA records through SBioBERT embeddings and FAISS-based similarity indexing (Table 3).

This functional distinction positions POE as a complementary resource to existing literature-centric systems, extending semantic discovery beyond publication text to the underlying omics data.

3.3. Additional access modes

3.3.1. Dataset search and article search

In addition to semantic queries, POE provides two direct access modes through the Dataset Search and Article Search menu options (Fig. 4). The Dataset Search allows retrieval of datasets using GEO Series accessions (GSE IDs) or GEO titles, while the Article Search enables retrieval based on PubMed titles or DOIs. These modes provide rapid access to known studies without requiring exploratory queries. All retrieved records include direct cross-links to GEO for processed data, ENA/SRA for raw sequencing data, and PubMed for the corresponding publication.

Table 3
Comparison of POE with representative biomedical semantic search and literature-linked systems. PubTator3, LitSense, and BioASQ operate exclusively at the publication level, supporting semantic annotation or information-retrieval tasks. In contrast, POE uniquely enables dataset-level semantic discovery by linking PubMed publications to GEO and ENA datasets through transformer-based embeddings and similarity indexing.

System	Scope	Core Function	Data Sources	Dataset Linking
PubTator3	Literature	Named-entity recognition and annotation for biomedical concepts (genes, diseases, chemicals)	PubMed, PMC	No
LitSense	Literature	Sentence-level semantic similarity and related-sentence retrieval	PubMed	No
BioASQ	Literature QA	Biomedical question-answering and large-scale IR benchmark using ontology-based concept matching	PubMed, MeSH, UMLS	No
POE	Literature + Omics	Semantic linking of PubMed abstracts to GEO and ENA datasets via SBioBERT embeddings and FAISS indexing	PubMed, GEO, ENA	Yes

3.3.2. Programmatic access

In addition to the web interface, POE offers a RESTful API that mirrors its core search and retrieval capabilities. This programmatic interface allows external tools, pipelines, and institutional platforms to query POE's semantic index, retrieve matched datasets, and access linked records automatically. The API supports semantic PubMed search with filters, direct retrieval by identifiers (GSE ID, PubMed ID, DOI, title), and ENA metadata download for given project accessions. Token-based authentication ensures secure access, and Python examples are provided for integration into reproducible bioinformatics workflows.

3.3.3. Limitations and future work

While POE demonstrates strong performance in linking biomedical literature with omics datasets, several limitations highlight areas for future improvement.

First, the system currently focuses on GEO datasets that include PubMed-linked publications, which may introduce a coverage bias toward well-annotated or frequently cited studies.

To mitigate this limitation, future updates will explore automated strategies to infer probable publication links for additional GEO records lacking explicit PubMed identifiers.

These strategies will combine metadata alignment (e.g., title, author, organism, and submission date matching) with semantic similarity analysis between GEO study descriptions and PubMed abstracts using SBioBERT embeddings.

Second, POE currently embeds titles and abstracts from PubMed rather than full-text articles. This design reflects both licensing constraints and computational efficiency requirements. Nonetheless, the integration of full-text semantic embeddings, where legally and technically feasible, represents a valuable direction for enhancing contextual understanding and retrieval depth.

Third, the platform currently integrates transcriptomic datasets from GEO, while data from other omics domains such as proteomics (PRIDE) [34] and metabolomics (MetaboLights) [35], are planned for inclusion. Extending POE to support cross-omics semantic search will further strengthen its role as a unifying discovery resource across diverse experimental modalities.

Finally, although the use of the IndexFlatL2 configuration in FAISS ensures exact nearest-neighbor retrieval and maximal accuracy, it is associated with higher memory and computational requirements compared to approximate methods. Future work will explore hybrid or hierarchical indexing strategies (e.g., IVFFlat or HNSW) to balance scalability with precision as the corpus continues to grow.

Collectively, these planned enhancements will expand POE's coverage, improve retrieval efficiency, and extend its applicability across multiple omics domains, while maintaining transparency and reproducibility through its public API and open configuration pipeline.

Although POE was not specifically designed as a FAIR-compliance framework, its architecture inherently aligns with the FAIR principles of data management.

All indexed entities include persistent identifiers (GEO accessions, PubMed IDs, and ENA project IDs), ensuring that datasets remain findable and uniquely referenced.

The platform provides accessibility through both an intuitive graphical interface and a public REST API, enabling programmatic integration into external analytical workflows.

By maintaining standardized metadata structures compatible with major repositories, POE promotes interoperability, while direct linking to original data providers ensures reusability of omics datasets.

Collectively, these features position POE as a FAIR-aligned semantic

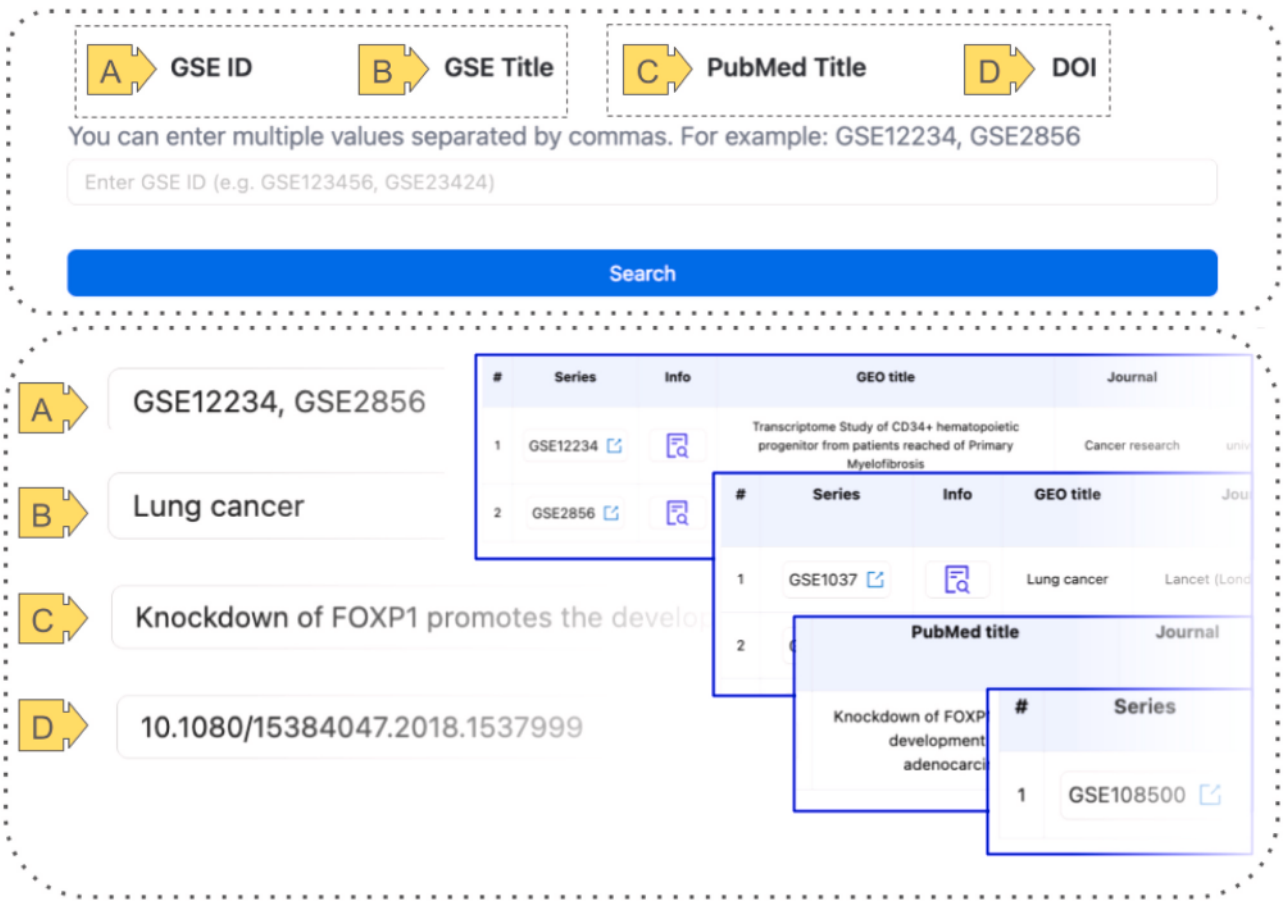

Fig. 4. Dataset and article search functionality in POE. POE provides two complementary entry points for semantic retrieval. Under the Dataset Search menu (A-B), users can search by GSE identifier or GEO title to directly retrieve datasets and metadata. Under the Article Search menu (C-D), users can search by PubMed title or DOI to link literature references with the corresponding GEO datasets. Together, these search modes provide complementary entry points that connect publications and datasets within POE.

search system that fosters transparent, accessible, and reusable biomedical data discovery.

In summary, POE embodies the principles of Findable, Accessible, Interoperable, and Reusable (FAIR) data management, ensuring long-term transparency and sustainability in biomedical data access.

While POE currently focuses on GEO-hosted datasets such as RNA-seq, ChIP-seq, and ATAC-seq, its semantic search architecture can be extended to other repositories hosting complementary data types, including those not yet fully integrated with GEO.

For instance, CAGE-seq datasets central to transcription start site analysis, such as those used in DeepTSS and AdaptCAGE, are primarily hosted by the FANTOM consortium or in ArrayExpress.

In future iterations, POE's framework could be adapted to include such repositories, enhancing its utility across a broader spectrum of regulatory genomics tools.

Similarly, literature-linked discovery of datasets relevant to microRNA regulation and promoter activity could facilitate deeper exploration using platforms such as miRGen v4.

4. Conclusions

POE introduces a transformative approach for discovering and reusing public omics data by semantically linking GEO datasets to PubMed literature. Unlike traditional keyword-based searches that often miss relevant studies due to inconsistent metadata or varying terminology, POE performs a semantic search across a subset of PubMed articles explicitly linked to GEO records. By embedding article titles and abstracts into a high-dimensional vector space, the platform enables the retrieval of conceptually relevant publications and their associated datasets using natural language queries.

This paradigm shift significantly lowers the barrier to data-driven discovery. Instead of requiring expertise in GEO metadata structure or precise keyword formulation, researchers can begin with a biological question-such as "inflammation in Parkinson's disease"- and obtain relevant literature and datasets, even when dataset annotations do not explicitly contain those terms. This is particularly valuable in interdisciplinary contexts, where terminology often diverges across subfields. By leveraging the semantic content of publications as a proxy for dataset context, POE facilitates hypothesis generation, accelerates exploratory research, and supports meta-analyses, cross-condition comparisons, and validation studies. Furthermore, by employing the biomedical-optimized SBioBERT model - selected for its demonstrated superior performance in representing biomedical terminology and contextual relationships - POE adapts to the linguistic characteristics of biomedical literature, enabling accurate and context-aware retrieval for domain-specific searches.

A notable advantage of semantic search is its ability to accommodate queries expressed in natural language, closely resembling the way researchers formulate scientific questions. This makes it possible to retrieve datasets that align more precisely with the biological intent of the query. In contrast, keyword-based approaches typically require the construction of complex Boolean expressions with multiple AND and OR operators. Such queries can easily become overly restrictive-excluding relevant datasets overly permissive, yielding large sets of superficially matched results. This structural limitation of lexical search contributes to the lower qualitative relevance observed in GEO-unique results,

whereas semantic search can capture the conceptual meaning of a query without relying on exhaustive keyword enumeration.

Beyond search functionality, POE promotes data reuse and reproducibility by facilitating the discovery of datasets aligned by conceptual relevance rather than rigid annotation. Its intuitive web interface and natural language support ensure accessibility across research backgrounds, enabling experimental biologists, clinicians, students, and educators to explore public omics data without requiring advanced computational skills. By lowering both technical and conceptual barriers, POE democratizes access to high-value omics resources and fosters broader participation in data-driven science and systems biology.

The strength of this approach was further substantiated by comparative evaluation. Across 30 benchmark queries, POE consistently retrieved substantially more datasets than GEO, in some cases exceeding GEO by two orders of magnitude. Crucially, this broader recall was not achieved at the expense of specificity: manual scoring of unique results revealed that POE-unique datasets were systematically more relevant to the biological intent of the queries compared to GEO-unique datasets. The case of "inflammatory signaling in cardiovascular disease" exemplifies this contrast, where GEO returned records unrelated to cardiovascular biology due to incidental keyword matches, while POE uniquely identified studies directly investigating inflammatory mechanisms in cardiac surgery, coronary artery disease, and myocardial infarction. These findings demonstrate that semantic indexing not only expands the discovery space but also improves the signal-to-noise ratio of retrieved datasets. By reducing irrelevant hits and surfacing contextually aligned results, POE provides a clear functional advantage over keyword-based search, supporting its role as a transformative tool for hypothesis-driven dataset discovery in biomedicine.

Funding

This work was developed and deployed using the computational infrastructure and servers of the Department of Informatics and Telecommunications, National and Kapodistrian University of Athens, where Nikos Perdikopanis is based. No dedicated research funding was received for the development of the Public Omics Explorer (POE).

Publication fees were covered by Georgios A. Pavlopoulos through the Hellenic Foundation for Research and Innovation (H.F.R.I.), under the call "Greece 2.0 - Basic Research Financing Action (Horizontal support of all Sciences), Sub-action II", Grant ID: 16718-PRPFOR, and the "Greece 2.0 - National Recovery and Resilience Plan", Grant ID: TAEDR-0539180.

Author Statement

The authors confirm that the following statements accurately describe the individual contributions, funding information, and potential conflicts of interest related to this work.

CRediT authorship contribution statement

Georgios A. Pavlopoulos: Writing – review & editing, Validation, Project administration, Funding acquisition. **Antonis Giannakakis:** Writing – review & editing, Validation. **Dimitris Grigoriadis:** Writing – review & editing, Software, Project administration, Formal analysis, Data curation, Conceptualization. **Margaritis Tsifintaris:** Writing – review & editing. **Nikos Perdikopanis:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Resources, Project administration, Investigation.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors would like to thank all collaborators from partner universities and research institutes for their valuable contributions to the development and validation of the Public Omics Explorer (POE).

Dimitris Grigoriadis would like to personally thank Afroditi Orfanidou for her invaluable moral support during the development of this work. A special and heartfelt acknowledgment goes to Kleoniki, whose quiet presence was a constant source of inspiration throughout this journey.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.csbj.2025.11.004.

Data availability

POE does not host or store raw omics data. Instead, it indexes structured metadata and provides direct external links to corresponding entries in GEO, PubMed, and ENA.

POE is a freely available web application, accessible at https://nplab.gr/poe.

References

[1] Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, et al. NCBI GEO: archive for functional genomics data sets-update. Nucleic Acids Res 2013 Jan 1;41(D1):D991–5.
[2] Domrachev Edgar R, Lash M. AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res 2002 Jan 1;30(1):207–10.
[3] Wilkinson MD, Dumontier M, IjJ Aalbersberg, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 2016 Mar 15;3(1):160018.
[4] Courtot M, Cherubin L, Faulconbridge A, Vaughan D, Green M, Richardson D, et al. BioSamples database: an updated sample metadata hub. Nucleic Acids Res 2019 Jan;47(D1):D1172–8.
[5] Quiñones M, Liou DT, Shyu C, Kim W, Vujkovic-Cvijin I, Belkaid Y, et al. METAGENOTE: a simplified web platform for metadata annotation of genomic samples and streamlined submission to NCBI's sequence read archive. BMC Bioinforma 2020 Sept;21(1):378.
[6] Chua HE, Tucker-Kellogg L, Bhowmick SS. ArcheGEO. Proceedings of the 13th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics. New York, NY, USA: ACM; 2022.
[7] Li Z, Li J, Yu P. GEOMetaCuration: a web-based application for accurate manual curation of Gene Expression Omnibus metadata. Database (Oxf) 2018 Jan;2018.
[8] Gonçalves RS, O'Connor MJ, Martínez-Romero M, Egyedi AL, Willrett D, Graybeal J, et al. CEDAR Work OntolAssist Environ Author metadata that Descr Sci Exp arXiv [csDB 2019 May.
[9] Martínez Romero M. Using biomedical ontologies to improve metadata management in CEDAR project. Proceedings of MOL2NET 2016, International Conference on Multidisciplinary Sciences, 2nd edition. Basel, Switzerland: MDPI; 2016.
[10] Bernstein MN, Doan A, Dewey CN. MetaSRA: normalized human sample-specific metadata for the Sequence Read Archive. Bioinformatics 2017 Sept;33(18):2914–23.
[11] Giles CB, Brown CA, Ripperger M, Dennis Z, Roopnarinesingh X, Porter H, et al. ALE: automated label extraction from GEO metadata. BMC Bioinforma 2017 Dec;18(S14).
[12] Chen G, Ramírez JC, Deng N, Qiu X, Wu C, Zheng WJ, et al. Restructured GEO: restructuring Gene Expression Omnibus metadata for genome dynamics analysis. Database (Oxf) 2019 Jan;2019.
[13] Mecham A, Stephenson A, Quinteros BI, Brown GS, Piccolo SR. TidyGEO: preparing analysis-ready datasets from Gene Expression Omnibus. J Integr Bioinform 2024 Mar;21(1).
[14] Davis S, Meltzer PS. GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor. Bioinformatics 2007 July;23(14):1846–7.
[15] Zhu Y, Davis S, Stephens R, Meltzer PS, Chen Y. GEOmetadb: powerful alternative search engine for the gene expression omnibus. Bioinformatics 2008 Dec;24(23):2798–800.
[16] Sozanska AM, Fletcher C, Bihary D, Samarajiwa SA. SpiderSeqR: an R package for crawling the web of high-throughput multi-omic data repositories for data-sets and annotation. bioRxiv 2020 Apr.
[17] Cai L, Lin S, Zhou Y, Yang L, Ci B, Zhou Q, et al. Lung Cancer Explorer (LCE): an open web portal to explore gene expression and clinical associations in lung cancer. bioRxiv 2018 Feb.

[18] Papatheodorou I, Fonseca NA, Keays M, Tang YA, Barrera E, Bazant W, et al. Expression Atlas: gene and protein expression across multiple studies and organisms. Nucleic Acids Res 2018 Jan;46(D1):D246–51.
[19] Stathias V, Turner J, Koleti A, Vidovic D, Cooper D, Fazel-Najafabadi M, et al. LINCS Data Portal 2.0: next generation access point for perturbation-response signatures. Nucleic Acids Res 2020 Jan;48(D1):D431–9.
[20] Lachmann A, Torre D, Keenan AB, Jagodnik KM, Lee HJ, Wang L, et al. Massive mining of publicly available RNA-seq data from human and mouse. Nat Commun 2018 Dec;9(1).
[21] Perez-Riverol Y, Bai M, da Veiga Leprevost F, Squizzato S, Park YM, Haug K, et al. Discovering and linking public omics data sets using the Omics Discovery Index. Nat Biotechnol 2017 May;35(5):406–9.
[22] Piñero J, Ramírez-Anguita JM, Saüch-Pitarch J, Ronzano F, Centeno E, Sanz F, et al. The DisGeNET knowledge platform for disease genomics: 2019 update. Nucleic Acids Res 2020 Jan;48(D1):D845–55.
[23] Wei CH, Allot A, Lai PT, Leaman R, Tian S, Luo L, et al. PubTator 3.0: an AI-powered literature resource for unlocking biomedical knowledge. Nucleic Acids Res 2024 July;52(W1):W540–6.
[24] Yeganova L, Kim W, Tian S, Comeau DC, Wilbur WJ, Lu Z. LitSense 2.0: AI-powered biomedical information retrieval with sentence and passage level knowledge discovery. Nucleic Acids Res 2025 July;53(W1):W361–8.
[25] Tsatsaronis G, Balikas G, Malakasiotis P, Androutsopoulos I, Paliouras G, et al. An overview of the BIOASQ large-scale biomedical semantic indexing and question answering competition. BMC Bioinforma 2015 Apr 22;16:138.
[26] Sarkans U, Gostev M, Athar A, Behrangi E, Melnichuk O, Ali A, et al. The BioStudies database-one stop shop for all data supporting a life sciences study. Nucleic Acids Res 2018 Jan;46(D1):D1266–70.
[27] Katayama T, Kawashima S, Okamoto S, Moriya Y, Chiba H, Naito Y, et al. TogoGenome/TogoStanza: modularized Semantic Web genome database. Database (Oxf) 2019 Jan;2019.
[28] Khomtchouk BB, Dyomkin V, Vand KA, Assimes T, Gozani O. Biochat: a database for natural language processing of Gene Expression Omnibus data. bioRxiv 2018 Nov.
[29] Leinonen R, Akhtar R, Birney E, Bower L, Cerdeno-Tárraga A, Cheng Y, et al. The European Nucleotide Archive. Nucleic Acids Res 2011 Jan;39:D28–31 (Database issue).
[30] Beautiful Soup Documentation - Beautiful Soup 4.13.0 documentation [Internet]. [cited 2025 July 18]. Available from: ⟨https://www.crummy.com/software/BeautifulSoup/bs4/doc/⟩.
[31] Reimers N., Gurevych I. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks [Internet]. arXiv; 2019 [cited 2025 Aug 1]. Available from: ⟨http://arxiv.org/abs/1908.10084⟩.
[32] (a) Bierman G, Abadi M, Torgersen M. Understanding TypeScript. In: Jones R, editor. ECOOP 2014 - Object-Oriented Programming [Internet]. Berlin, Heidelberg: Springer Berlin Heidelberg; 2014. p. 257–81 [cited 2025 Sept 19].
(b)(Hutchison D., Kanade T., Kittler J., Kleinberg J.M., Kobsa A., Mattern F., et al., editors. Lecture Notes in Computer Science; vol. 8586). Available from: http://link.springer.com/10.1007/978-3-662-44202-9_11.
[33] Thakur N., Reimers N., Daxenberger J., Gurevych I. Augmented SBERT: Data Augmentation Method for Improving Bi-Encoders for Pairwise Sentence Scoring Tasks [Internet]. arXiv; 2021 [cited 2025 Sept 19]. Available from: http://arxiv.org/abs/2010.08240.
[34] Perez-Riverol Y, Bandla C, Kundu DJ, Kamatchinathan S, Bai J, Hewapathirana S, et al. The PRIDE database at 20 years: 2025 update. Nucleic Acids Res 2025 Jan 6;53(D1):D543–53.
[35] Yurekten O, Payne T, Tejera N, Amaladoss FX, Martin C, Williams M, et al. MetaboLights: open data repository for metabolomics. Nucleic Acids Res 2024 Jan 5;52(D1):D640–6.