

User-driven development and evaluation of an agentic framework for analysis of large pathway diagrams

Marie Corradi¹, Ivo Djidrovski², Luiz Ladeira³, Bernard Staumont³, Anouk Verhoeven⁴,
 Julen Sanz Serrano⁴, Adrien Rougny⁵, Ahmad Vaez^{6,7,8}, Ahmed Hemedan⁹, Alexander Mazein¹⁰,
 Anna Niarakis¹¹, Arthur de Carvalho e Silva¹², Charles Auffray¹³, Egon Wilighagen¹⁴,
 Eliska Kuchovska¹⁵, Falk Schreiber^{16,17}, Irina Balaur¹⁸, Laurence Calzone¹⁹, Lisa Matthews²⁰,
 Lorenzo Veschini²¹, Marc E Gillespie²², Martina Kutmon²³, Matthias König^{24,25,26},
 Matti van Welzen²⁷, Noriko Hiroi²⁸, Oxana Lopata²⁹, Pierre Klemmer⁵, Rupert Overall²⁴,
 Tim Hofer²⁹, Venkata Satagopam⁵, Reinhard Schneider⁵, Marc Teunis¹, Liesbet Geris³
 and Marek Ostaszewski^{5,*}

¹Innovative Testing in Life Sciences and Chemistry, HU - University of Applied Sciences Utrecht, Netherlands

²Institute of Risk Assessment Sciences (IRAS), Utrecht University, Netherlands

³Biomechanics Research Unit, GIGA Institute, University of Liège, Belgium

⁴Department of Pharmaceutical and Pharmacological Sciences, Vrije Universiteit Brussel, Belgium

⁵Luxembourg Centre for Systems Biomedicine, University of Luxembourg, Luxembourg

⁶Unit of Genetic Epidemiology & Bioinformatics, Department of Epidemiology, University Medical Center Groningen, Netherlands

⁷Department of Bioinformatics, Isfahan University of Medical Sciences, Iran

⁸Medical Image and Signal Processing Research Center, Isfahan University of Medical Sciences, Iran

⁹Transversal Translational Medicine, Luxembourg Institute of Health, Luxembourg

¹⁰Department of Cancer Research, Luxembourg Institute of Health, Luxembourg

Department of Biology, Faculty of Science and Engineering, University of Toulouse, CBI, CNRS, France

¹²School of Biosciences, University of Birmingham, United Kingdom

¹³European Institute for Systems Biology and Medicine, France

¹⁴Translational Genomics, NUTRIM, FHML, Maastricht University, Netherlands

¹⁵IUF - Leibniz Research Institute for Environmental Medicine, Germany

¹⁶Department of Computer and Information Science, University of Konstanz, Germany

¹⁷Faculty of IT, Monash University, Australia

¹⁸Luxembourg National Data Service, Luxembourg

¹⁹Computational Systems Biology of Cancer, Institut Curie, U1331 INSERM, Mines ParisTech, PSL, France

²⁰Biochemistry and Molecular Pharmacology, NYU Grossman School of Medicine, United States

²¹Biocomplexity Institute, Department of Intelligent Systems Engineering, Indiana University at Bloomington, United States

Department of Pharmaceutical Sciences, St. John's University, United States

²³Maastricht Centre for Systems Biology and Bioinformatics (MaCSBio), Maastricht University, Netherlands

²⁴Institute of Biology, Humboldt University of Berlin, Germany

²⁵Institute of Structural Mechanics and Dynamics in Aerospace Engineering, University of Stuttgart, Germany

²⁶First Department of Medicine, Campus Lübeck, University Hospital Schleswig-Holstein, Germany

²⁷Department of Systems Biology and Bioinformatics, University of Rostock, Germany

²⁸Department of Electrical and Electronic Engineering, Kanagawa Institute of Technology, Japan

²⁹Department of Chemical Toxicology, Norwegian Institute of Public Health, Norway

*Corresponding author. marek.ostaszewski@uni.lu

Abstract

As biomedical knowledge keeps growing, resources storing available information multiply and grow in size and complexity. Such resources can be in the format of molecular interaction maps, which represent cellular and molecular processes under normal or pathological conditions. However, these maps can be complex and hard to navigate, especially to novice users. Large Language Models (LLMs), particularly in the form of agentic frameworks, have emerged as a promising technology to support this exploration. In this article, we describe a user-driven process of prototyping, development, and user testing of Lley, an LLM-based system for exploring these molecular interaction maps. By involving domain experts from the very first prototyping in the form of a hackathon and collecting both fine-grained and general feedback on more refined versions, we were able to evaluate the perceived utility and quality of the developed system, in particular for summarising maps and pathways, as well as prioritise the development of future features. We recommend continued user-driven development and benchmarking to keep the community engaged. This will also facilitate the transition towards

© The Author(s) 2026. Published by Oxford University Press.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License

(<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

Introduction

The ecosystem of biomedical knowledge repositories continues to grow in size and complexity. Structured information available in databases of pathway diagrams, knowledge graphs, or interaction databases can support hypothesis generation, inform experimental design, and facilitate interpretation of omics data (Mazein et al., 2024). However, navigating resources with different formats, interfaces, and granularities is a challenging and time-consuming task. Additionally, the sheer volume of these knowledge collections poses a challenge for navigating their individual components and relationships.

One such resource - disease maps - consists of comprehensive graphical representations of cellular and molecular processes associated with specific organ functions under normal or pathological conditions (Mazein et al., 2018; Staumont et al., 2025). Such maps are manually curated, provide information about source literature used for their construction, and follow the Systems Biology Graphical Notation (SBGN) (Le Novère et al., 2009) and Systems Biology Markup Language (SBML) (Keating et al., 2020) standards. They are hosted using the MINERVA Platform (Hoksza et al., 2019) which provides functionalities for visual exploration and programmatic access. Currently, a number of disease and physiological maps are available, describing pathways of neurodegeneration (Fujita et al., 2014), chronic immune and allergic diseases (Acencio et al., 2023, 2026; Lopata et al., 2025; Mazein et al., 2021), mechanisms of inflammation and viral infection (Ostaszewski et al., 2021; Serhan et al., 2020), and physiology of liver metabolism (Ladeira et al., 2025) or adverse outcome pathways (Mazein et al., 2025). Importantly, molecular interaction maps are distributed and differ in scope and granularity, and can benefit from unified access interfaces to reduce barriers to access and interpretation. This will facilitate visualisation, analysis and modelling of complex molecular mechanisms represented by these maps (Hemedan et al., 2024; Hoch et al., 2022, 2023; Zerrouk et al., 2022, 2024).

Large Language Models (LLMs) have emerged as a promising technology for summarising, analysing, and explaining structured knowledge. They range from general-purpose foundational models to applications in specialised tasks like building and exploring knowledge graphs or data analysis and hypothesis generation. Direct retrieval of knowledge from LLMs has been evaluated in, e.g., assisting pathway enrichment analysis (Azam et al., 2024), retrieving ageing-relevant knowledge under different hyperparameter configurations (Shakhbazov, 2025), or searching for drug combinations in Alzheimer's disease (Wang et al., 2026). For processing structured knowledge, LLM-based workflows have been proposed to generate knowledge graphs from the literature (James et al., 2026) and to process them for prioritising experimental results (Feng et al., 2025) and exploring clinically valid hypotheses (Khan et al., 2025). Additionally, frameworks have been proposed to convert biological databases into structured knowledge graphs (Lobentanzer et al., 2023) and to access information via chatbot interfaces (Lobentanzer et al., 2025). In parallel to specific use cases, LLM-supported architectures were proposed to assist with access to bioinformatic databases and summary of results (Sokolova et al., 2025), supporting collaborative construction of scientific hypotheses (Jiang et al., 2026), or comprehensive analysis of omics data (Nagarajan et al., 2025). Finally, LLMs were evaluated for their support in curation and annotation of bioinformatic knowledge (Devkota et al., 2025),

including pathway databases (Wu et al., 2025). However, the need for a dedicated solution addressing molecular interaction maps is still missing. As these maps are highly interactive resources with a range of different applications, an LLM-based system should be developed and evaluated by directly engaging the target user group.

In this article, we describe a user-driven process of prototyping, development, and user testing of Llemy, an LLM-based system for exploring large pathway diagrams. Llemy is an online system built for exploratory analysis of user-built repositories of molecular interaction maps. We evaluated Llemy's usability and its perceived performance by having a group of users conduct an exploratory analysis of various molecular interaction maps. We gathered information on the system's performance at the level of individual user prompts and collected overall user feedback via a survey completed after their work. Based on these results, we identify key tasks performed by the users, and assess their perceived performance. We demonstrate a need for a robust framework for open-ended evaluation of tasks involving interactive diagram exploration and analysis. The results of this evaluation provide insight into the design of future LLM-based systems for exploring and analysing molecular interaction maps and indicate the types of tasks on which such systems may underperform.

This article is structured as follows. In the Methods section, we introduce the hackathon in which Llemy was prototyped, the user-driven approach to its development, its architecture, and the evaluation process. In the Results section, we briefly describe the user interface, discuss use case examples, and analyse user feedback data. Finally, we discuss the results and provide an outlook for future work.

Llemy can be accessed via <https://llemy.vhp4safety.nl>, and its code repository is available at <https://github.com/ontox-project/Llemy> (see Data availability).

Methods

Llemy was developed in several steps, following a user-centric design. First, a prototype was developed during a two-day hackathon. This prototype was then refined and evaluated by a group of users involved in the development and analysis of molecular interaction maps.

Prototype development

The project was initiated during a two-day hackathon held at the University of Liège, Belgium, on May 8-9, 2025. The aim was to design a system that facilitates exploration of molecular interaction maps, with liver lipid and bile metabolism (Ladeira et al., 2025) as use cases. These maps were developed as a part of the ONTOX project (Vinken et al., 2021). The hackathon group involved hepatotoxicologists, curators of the maps, computational biologists, and LLM experts.

User requirements and the system architecture were established based on user prompts prepared by the domain experts. These prompts were then used in iterative prototyping and evaluation. The prompts included i) queries with scope beyond the contents of the map, and ii) queries containing false premises, designed to evaluate the ability to identify and address incorrect assumptions (see Supplementary H1).

The final hackathon prototype consisted of two separate agents collecting information in parallel: one from the map hosted on the MINERVA Platform, and one from Perplexity using deep research

(accessed on 08.05.2025). Then, a third synthesis agent (OpenAI GPT4.1) combined and summarised the findings. Commercial LLMs were chosen for quick prototyping. The whole prototype was containerised in a Docker application with the possibility to deploy it locally and provide API keys for both LLMs. This architecture was inspired by the O-QT assistant (Djidrovski et al., 2025).

System architecture

Llemy was built entirely in Python, using GPT-4.1 nano, Streamlit (v1.50.0) for the front-end, and LangChain (v0.3.27) as the agentic backend. Note that it is possible for a user to change the model to another LLM in the OpenAI family, replacing `MODEL_NAME="gpt-4.1-nano"` by their LLM of choice in `workflow.py` (line 23). Llemy is deployed publicly on a cloud platform (<https://llemy.vhp4safety.nl/>), and works with user-provided API keys. It is also available as a Docker container for local deployment. In the deployed version, API keys are stored only for the duration of a user session to ensure security. To avoid context rate limits, a follow-up prompt using the previous answer is not possible in the current version as of the publication of this paper. However, the entire chat history from a given session is visible. Llemy combines user prompt with map data and instructions output, focus, and inclusion of literature references. The full instructions are available in the file `workflow.py` on the GitHub repository (<https://github.com/ontox-project/Llemy>).

User study design

Llemy was evaluated by a group of 25 users recruited via Disease Maps Community channels (<https://disease-maps.io/>). The users worked with the system by performing queries to a selected pathway or disease map, using either a pre-constructed set of prompts (see Supplementary U1), or prompts provided by the users. Users were provided with API keys and asked to evaluate the system's performance after each prompt. Performance metrics were accuracy, conciseness, and reliability, evaluated on a five-point scale. Users were also able to provide free-text comments on the system's performance for each metric. The API keys for user testing were provided by ELIXIR Luxembourg Node. At the end of their testing session, the users were asked to fill out a questionnaire summarising their overall assessment of the performance of Llemy (see Supplementary U2).

Two separate datasets were collected for analysis. One consisted of individual prompts, system replies, and user feedback on the accuracy, conciseness, and reliability of the responses (Prompt Dataset). The other dataset contained user replies to the final questionnaire (Summary Dataset). They were treated as separate in downstream analyses.

Prompt Dataset analysis

The Prompt Dataset contains 157 entries, including the user prompt, the response of the system, the response time, and the evaluation of the response in terms of accuracy, conciseness, and reliability, rated on a 5-point scale (see Supplementary Table U3). One prompt ("Describe how I can bake an apple cake") was intentionally designed to test the system's ability to remain within the scope of its source maps. The system correctly responded by explaining its own purpose, and this entry was removed from the subsequent analysis as irrelevant.

To understand the prompt landscape, 114 unique prompts were categorised into three categories. The categories were established post hoc based on the content of the prompts. Each prompt was assigned a category based on a consensus among three independent experts. Score distributions were summarised, both aggregated across all categories and stratified by prompt category, to examine potential variation in perceived system performance across different categories. Free-text comments provided for each prompt evaluation were grouped by associated metric score and analysed qualitatively. There were no comments for conciseness provided by the users.

Summary Dataset analysis

The Summary Dataset gathers a general evaluation of the system from 19 unique users. The data collection was fully anonymous. Users were asked about their experience using molecular interaction maps and LLMs. They were then asked to estimate how long they spent testing the app, and what type of prompts they entered. Further feedback on ease of use, productivity and overall assessment were collected using both a 5-point scale and free-text comments.

Statistical analysis

All analysis was performed in R version 4.5.2. To analyse the impact of response time on performance metric, a cumulative link mixed model (ordinal package, `clmm` method, `link = "probit"`, `nAGQ = 10`) was fitted on Accuracy, Conciseness and Reliability metrics (jointly), with a random intercept per user to account for within-user correlation. A cumulative link mixed model was chosen because of high inter-metric correlation: Spearman correlation between Accuracy and Conciseness is 0.53, between Conciseness and Reliability is 0.43, and between Accuracy and Reliability is 0.71.

To examine differences in Accuracy, Conciseness, and Reliability across Categories, post-hoc pairwise comparisons, using the FSA package, were conducted using Dunn's test with Holm correction for multiple comparisons. This non-parametric test was applied separately for each dependent variable to identify which specific category pairs differed significantly while controlling for family-wise error rate.

Results

Implementation of Llemy

Llemy consists of two parallel processes (Figure 1) following the prompt provided by the user, and the selected map. First, the user prompt is enriched with instructions for the LLM ensuring contextualised output, scientific focus, and inclusion of literature references (see Supplementary P1). Second, a retrieval agent uses the MINERVA API (Hoksza et al., 2019) to get contents of the map selected by the user: elements (nodes), reactions (edges), and annotations. The system provides access to public maps available under a permissive licence shared via a repository of molecular interaction maps, MINERVA Net (<https://minerva-net.lcsb.uni.lu/>) (Gawron et al., 2023), and a set of predefined prompts. The contents are concatenated into a single text object, including element identifiers, reaction types and identifiers of supporting literature. Importantly, this transformation removes SBML properties from analysed diagrams to narrow the context of the LLM for more accurate answers. Because of this, downstream computational workflows would have to retrieve MINERVA identifiers provided by

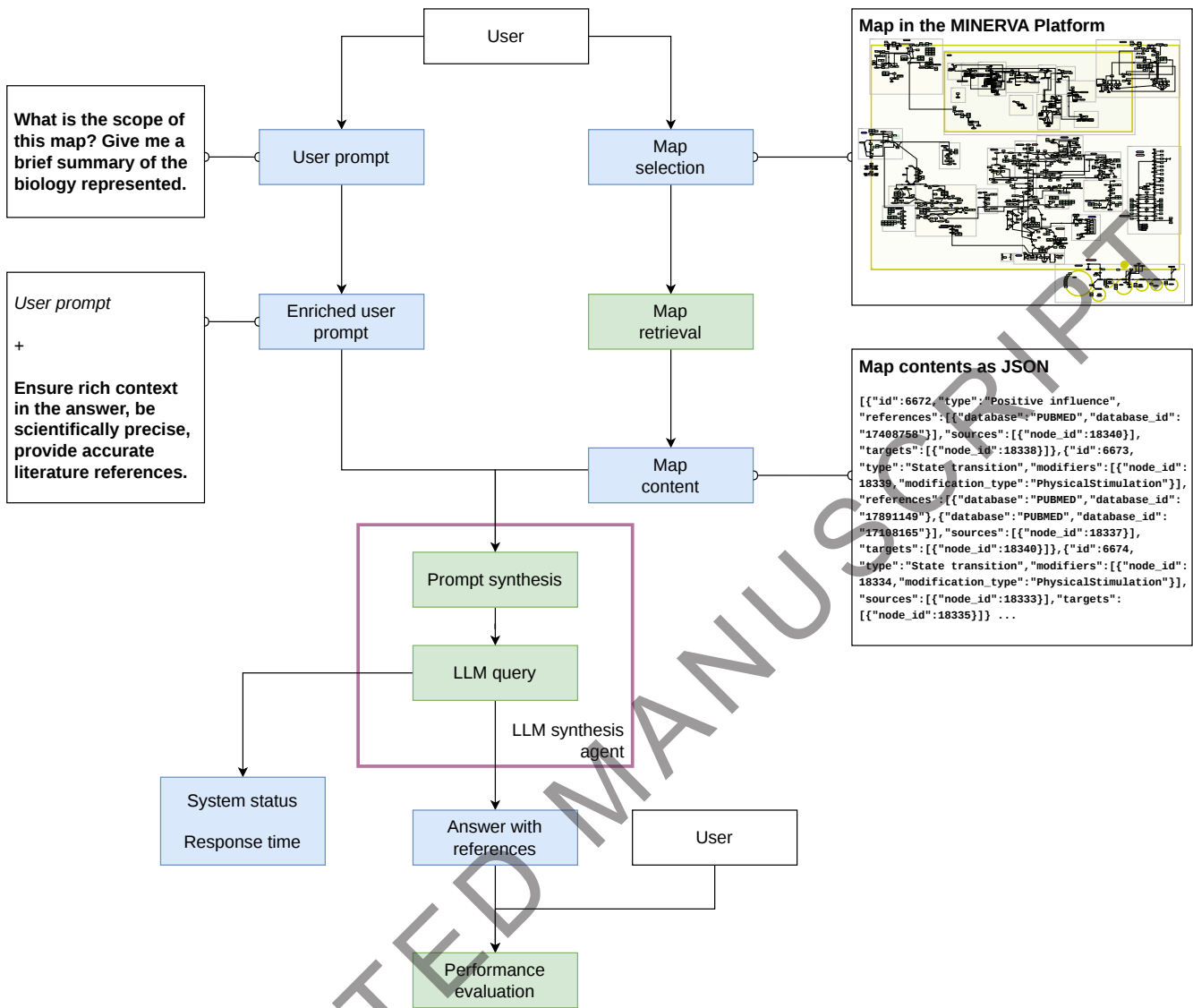


Figure 1 The workflow of Llemetry. Users provide their prompt and select a map they want to process. Then, the prompt is enriched with instructions for the content of the answer (see full system prompt in Supplementary P1), and, in parallel, map data is fetched from the corresponding MINERVA Platform. Both enriched prompt and processed map data are combined and used as the LLM prompt. The answer is provided to the user for evaluation, and system performance logs are collected. Green: process, blue: data.

in the system response, and use them with MINERVA API calls to access full content of the map of interest,

This text object is then given to a “synthesiser agent” as a context, together with the enriched user prompt. The response is formatted post-hoc by adding links to specific parts of the map of interest: when the LLM cites an element or reaction identifier in a given sentence, they are isolated and collated to a clickable link after the end of that sentence. A robust error-handling and logging layer is integrated throughout the workflow: API responses are normalised into explicit states (e.g., **success**, **error**, **no_data_found**, **skipped**), retrieval errors are passed to the synthesiser as structured context, and both the synthesised response and raw map payload are returned for traceability. This makes each answer auditable and allows users

to verify exactly which MINERVA-derived data supported the final output. Response times for each query were also calculated.

Llemetry is deployed as a service on the VHP4Safety cloud (<https://llemetry.vhp4safety.nl/>). Upon landing on the page (see Figure 2), the user is prompted to enter an OpenAI API key. To provide user feedback, a password is required to ensure privacy, and consent to collect prompt and response data to analyse the performance of the system. If the user previously agreed to the recording of their data, a brief feedback form is displayed to rate the accuracy, conciseness, and reliability of the answer on a 1 to 5 scale, and provide a free text comment on the result of a particular query. Llemetry can be used without providing feedback.

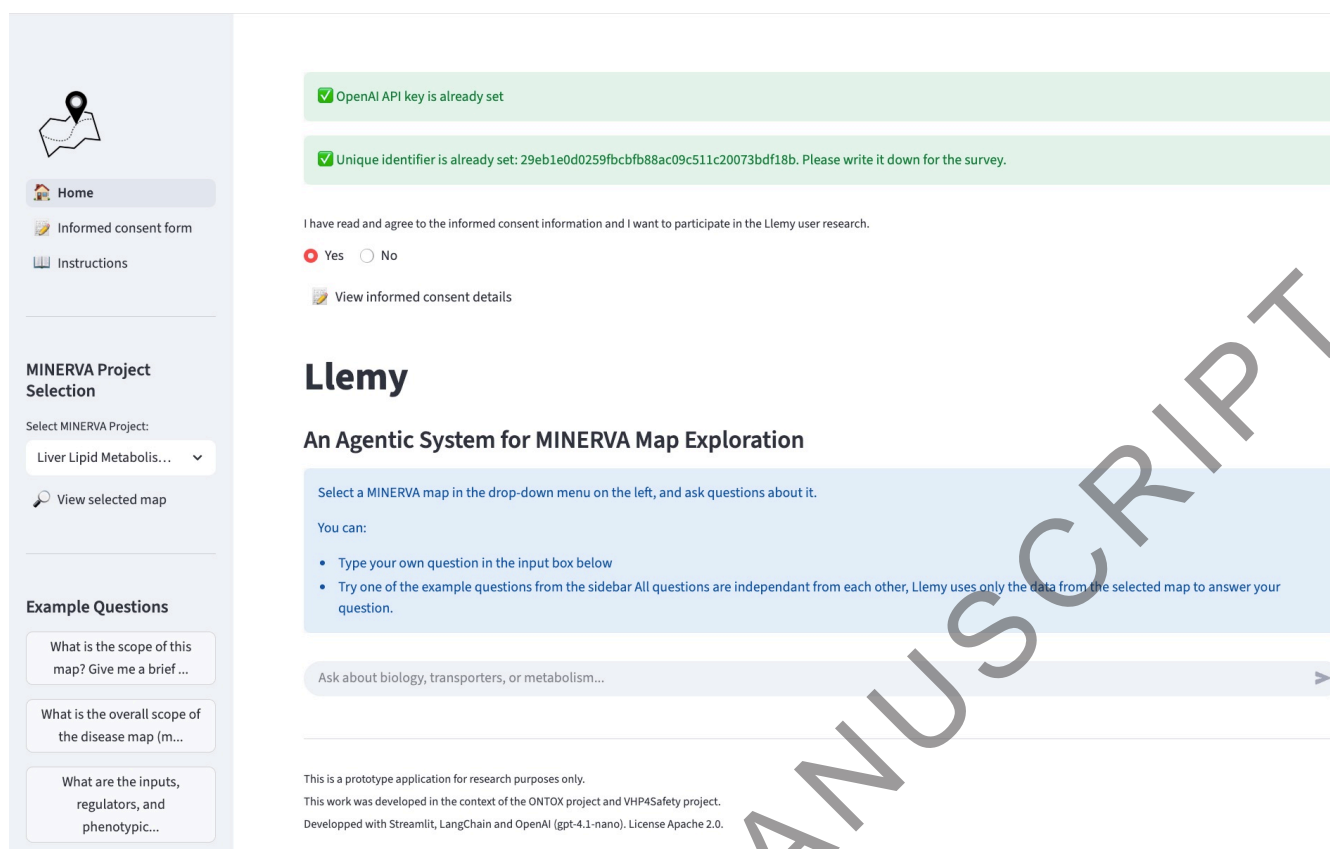


Figure 2 Screenshot of Llemy's welcome page. The left-side panel comprises i) a section with links to the home page, informed consent form, and user manual; ii) a map selection drop-down menu, and a link to the selected map; iii) a list of example prompts (see Supplementary U1 for a complete list).

Use cases

The purpose of Llemy is to facilitate exploration of complex molecular interaction maps. Such maps can have thousands of elements and interactions, constructed from hundreds of literature references, which makes working with these resources a challenge, despite dedicated software supporting exploratory and analytical tasks. We identify three main groups of tasks that Llemy can facilitate when working with large molecular interaction maps. These are to i) summarise large content into understandable reports, ii) find specific elements, or iii) analyse certain properties of the maps (see Table 1). Below, we discuss each of these categories based on concrete examples, which are available via Supplementary E1. Each cited query below has a corresponding example file with a complete output of Llemy and the corresponding map from which the content was taken.

Summarise

These tasks support initial exploration of large maps by providing digestible reports about different aspects of the map. User queries may involve creating an overview of the contents of a given map to a new user, allowing to grasp key diagrams and pathways in a given map and evaluate its scope at a glance, e.g. "summarise the biology of COVID-19" (summarise_6, Supplementary E1). Another set of queries in this task asks for the role of a single molecule across different pathways, e.g. "summarise the role of NAMPT in ageing" (summarise_1), which may gather map information about

multiple instances of multi-context and multi-purpose proteins in multiple map locations. Finally, queries about microenvironment, tissue specificity, temporal resolution or inter-organelle interactions allow users to quickly understand the context of a given map and its utility for data interpretation and modelling, e.g. "summarise inter-organelle interactions, crossing e.g. nucleus, mitochondria, ER." (summarise_8).

Find

These tasks extend search functionalities of the MINERVA Platform beyond simple lookup of target molecules. LLM support offers natural language-based queries of the content of maps, which would be difficult to implement as a conventional search function. User queries may be similar to existing functionalities, e.g. "list all drugs in the map" (find_2, Supplementary E1), but Llemy supports advanced inquiries like "find stress pathways in the map" (find_5), which requires a combination of the map contents, and the LLM reasoning of what may be considered a stress pathway. Finally, queries like "find the initial event for COVID pathology, and its endpoint" (find_7) or "find the inputs, regulators, and phenotypic outputs" (find_6) allow users to identify map elements that are functionally related to each other without time-consuming tracking of diagram connectivity.

Analyse

These tasks support reasoning and calculation-related queries, offering computational support over the content of molecular interaction maps. Given the limitations of LLM-based reasoning

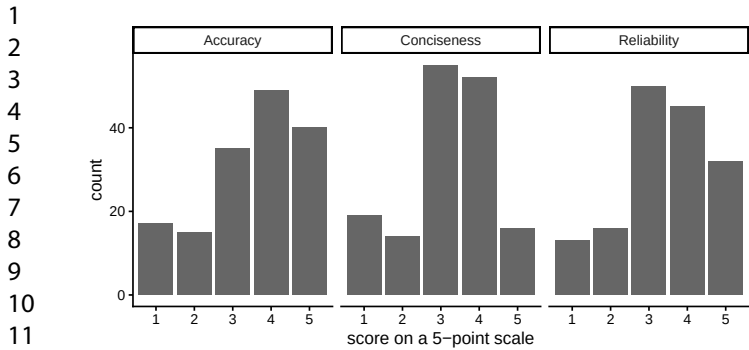


Figure 3 Score distribution by evaluation criterion.

in some models, such support may have a moderate extent; nevertheless, it allows basic hypothesis building and may give users a better understanding of the modelling potential of a map, and an inspiration to construct advanced computational workflows. Queries like “identify pathways affected by blocking autophagy; explain your findings” (analyse_2, Supplementary E1) or “define measurable downstream changes of NAMPT attenuation; propose the best biomarker of NAMPT perturbation” (analyse_4) are essentially modelling questions, however as the input that Llemmy operates on is reduced, results of such queries should be followed by dedicated, reproducible computational workflows.

User study
User feedback for individual prompts

We assessed the perceived performance of the system across three metrics - accuracy, conciseness, and reliability - based on user feedback (see Figure 3). Medians for these three metrics were 4, 3, and 4, respectively. We performed a cumulative linked mixed model (see Methods) to check if any of these metrics were affected by the response time. Response time had a significant negative effect on ratings ($\beta = -0.34, SE = 0.06, z = -6.06, p < 0.001$), indicating that longer response times were associated with lower perceived quality. This effect was uniform across Accuracy, Conciseness and Reliability. On average, users rated Conciseness significantly lower than Accuracy ($\beta = -0.37, p = 0.002$), while Reliability did not differ from Accuracy ($\beta = -0.12, p = 0.34$).

Individual prompts were grouped into three categories (see Methods), as shown in Table 1, together with the number of prompts per category. Stratified analysis of evaluation metrics using post-hoc Dunn test (see Methods), revealed no significant performance differences across categories. Mean accuracy, conciseness, and reliability were the highest for the "Summarise" task (Figure 4), and the "Find" category exhibited a broader distribution of scores, showing lower evaluations for all scores, possibly reflecting the complexity of querying molecular interaction maps with an LLM when the content of the map is in the prompt.

Qualitative analysis of individual user comments revealed distinct patterns associated with score levels. For higher accuracy scores (4-5), users acknowledged the system's ability to provide comprehensive summaries, correctly identify pathway connections across submaps, and appropriately acknowledge limitations when specific entities were absent from maps. For lower accuracy scores (1-2), users most frequently cited fundamental factual errors, failure to locate existing

Table 1 Categories of prompts in the Prompt Dataset, and their short definitions.

Category	N	%	Definition
Summarise	65	41.7	Summary of a map, mechanisms involving specific pathways, elements or interactions (e.g., autophagy, bile acid metabolism, signalling pathways), or mechanisms missing from the map.
Find	46	29.5	Retrieval of specific map elements, elements with specific properties (e.g., kinases, ligands, enzymes, transcription factors, protein interactors), or upstream or downstream targets (e.g., receptors, organelles, phenotypes) outside the map.
Analyse	45	28.8	Analysis of the map or its parts (e.g., count specific elements, calculate connectivity of the network) or reason about the consequences of perturbing the molecular network (e.g., drug effects, disease-related perturbations).

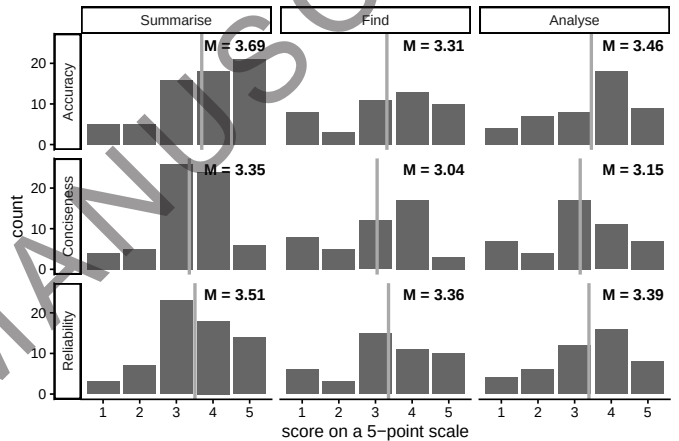


Figure 4 Distribution of tasks performed by the users testing Llemmy, grouped into three categories. M = average.

map content, and fabricated or mismatched map reaction references. Common criticisms included misunderstanding map element types, missing available annotations, and instances where the system failed to recognise entities present under synonyms. There were no individual comments on the conciseness of results. Reliability scores were strongly influenced by the quality of the reaction hyperlinks to the source content, coming from maps hosted in the MINERVA platform. High reliability scores (4-5) corresponded to systematic and verifiable referencing, allowing systematic verification. Low reliability scores (1-2) were associated with broken or non-resolving links and inconsistent behaviour across repeated queries, notably varying structure and scope of provided references to the source map, from insufficient to unnecessarily repetitive.

Several cross-cutting themes emerged from user feedback. First, synonym and nomenclature handling presented recurring challenges when the system failed to recognise entities represented by HGNC (HUGO Gene Name Committee) names rather than common abbreviations. Second, users valued contextual awareness, noting that responses sometimes missed map-specific or domain-specific framing (e.g., failing to mention organ context for an organ-specific map). Third, users at intermediate score levels (3) noted that answers

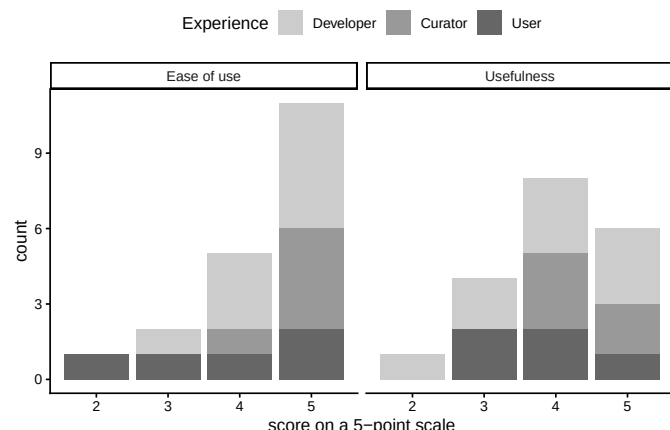


Figure 5 Usefulness and usability based on experience with molecular interaction maps. Developer - participants building workflows including maps; Curator - participants creating maps; User - participants using the maps for knowledge retrieval and data visualisation. In case of multiple roles, the highest level of experience - Developer, then Curator, then User - is taken into account.

were accurate but incomplete, and challenging to assess due to a lack of proper references to the source material.

Detailed, representative examples of queries where Llemmy underperformed are available as described in Supplementary E1, “challenging-cases”. Summarised individual user comments having particular score levels are listed in Supplementary U4.

Overall user feedback

The user survey showed a varied distribution of users, curators, and developers of molecular interaction maps. The relatively low number of responses in the “user” group can be explained by the fact that most users also had experience as curators or developers. Almost all users had prior experience with LLMs. All users tested the system for more than 15 min, half of them spent more than an hour, and 75% estimated that they saved time by using Llemmy. Another important feedback was concerning the variability of the output for similar queries, where, on a 1 to 5 scale, all users except one reported variability of 3 and higher. Figure 5 summarises usefulness and usability scores based on experience with molecular interaction maps.

Discussion

Advantages of the LLM-based approach

The potential of LLMs to accelerate research in the field of molecular pathways and their regulation is a topic of intense investigation. In this work, we explore the capability of LLM-based systems to process, summarise, analyse, and suggest extensions of structured, focused knowledge repositories describing molecular pathways involved in human physiology and pathology. We introduce Llemmy, a system for summarising knowledge graphs based on user prompts. Our approach parallels previously published LLM solutions to refine knowledge graphs from literature (Feng et al., 2025; James et al., 2026; Khan et al., 2025). However, in the present work the sources of the system are manually curated diagrams. This allows Llemmy to directly reference diagram elements and interactions, thereby improving the

system’s transparency and facilitating exploration of the original content based on the system’s response.

User-focused analysis

Our approach focused on the user perspective in the design and evaluation of the system. The system was designed and prototyped during a hackathon, when domain experts directly formulated prompts they would ask, and evaluated replies on-the-fly. The finalised system was tested by an independent group of 25 users at the level of individual prompts, and via a survey regarding their overall user experience. Based on 156 individual prompts, we proposed three categories posed to the system and showed that tasks with higher complexity are asked less frequently and receive less positive feedback. Although no statistical differences were observed, we’ve noticed comparatively higher scores in the “Summarise” category. Based on these categories, Llemmy workflows can be tailored to users’ needs. Thus, the present work parallels the functionality of the Alvessa system proposed in (Sokolova et al., 2025), but it focuses on the nature of the task rather than on specific resources to be queried based on the system prompt.

One of our findings was a negative impact of response time on performance metrics, i.e. longer response times corresponded to worse performance. Notably, literature reports in this regard are mixed (Hu et al., 2025; Kim et al., 2026; Maslych et al., 2025; Wang and Lo, 2025), so our results should not be overinterpreted. Importantly, specific examples of Llemmy output for the same query, for the same map (see “variability”, Supplementary E1) indicate multiple factors impacting varying response time. In highlighted examples, Llemmy answers differ in length, content, and structure, but their length is not correlated with the response time. In one of the answers, the system provides a link to some related proteins in the map, but not the others. In some examples, no links are provided. In this light, longer response times can be explained by the impact of MINERVA API when providing the map or variable response time by the LLM, but it is unlikely that it correlates with the complexity of the answer.

User perception of an LLM-based system

A summary user survey completed by 19 of 25 participants rated the system as highly accessible; however, the utility rating was more moderate, particularly among map users. This may be explained by the fact that the users less familiar with the content processed by Llemmy can give more favourable evaluations to its answers. All users generally rated the usability of Llemmy as high (see Figure 5), with over 80% rating it a 4 or 5. It is worthwhile to note that users-only were quite split. Over 70% of users found Llemmy useful (score of 4 or more).

Reproducibility of results and the sustainability of the system are key considerations in developing systems such as Llemmy. All users except one reported a marked variability of the system output, of at least 3 points on the 1 to 5 scale, in line with other reports (Wang and Wang, 2025). To improve this outcome, open-weight models may be considered, contrary to the current commercial LLM. Open-weight LLM models show promising overall performance when compared to commercial solutions (Stachura et al., 2025), however, use of open LLMs comes with the need for a dedicated infrastructure and associated costs. Moreover, tests of locally deployed, open-access, and open-weight LLMs are necessary to compare the tradeoff between quality and cost of use, which in turn requires benchmark datasets. Our open-ended user survey may inform the design of such a system,

similar to foundational model benchmarks (Laurent et al., 2024) used in Sokolova et al. (2025).

Roadmap for the system and its evaluation

Llemy and similar LLM-powered systems have the potential to reduce the complexity barrier when working with complex diagrams. Based on the user feedback in both the Summary and Prompt datasets, we drafted a roadmap to support future improvements. We attributed priorities based on the frequency with which users requested the feature. In the short term, systems architecture should be updated to improve response time and accuracy of the provided references. Afterwards, dedicated workflows for summarising, querying, and analysing the map should be introduced, with task granularity matching the user study. Finally, programmatic integration of Llemy functionalities with the GUI of the MINERVA Platform using plugins (Hoksza et al., 2019) and workflows using the Model Context Protocol (Hou et al., 2025; Kuehl et al., 2025) will streamline the interaction of LLMs with the content of maps. This should be supported by the use of open-weight LLMs to address the needs of the open-access research environment.

User interaction is an important part of the Llemy project, and it can be improved by acting directly on provided feedback. Analysis of individual user prompts and answers, especially for the low-scoring cases (see “challenging-cases”, Supplementary E1) provides a valuable window into improvements needed for the system. First, we need to improve input limitations that cause errors where requested content is inaccessible directly from the prompt, like specific identifiers or synonyms. Second, we have to address reasoning limitations, likely related to the chosen model, which may cause Llemy to stumble on multi-step problems, like connecting a synonym to its official symbol and then looking for that symbol in the content of the map. Finally, we need to fix mishandling of MINERVA references which causes repetitive artifacts in the output.

These issues can be addressed by supporting term disambiguation and feedback for prompts that cannot be answered, as well as better linking of references to original publications or map content used in the answer. External services for grounding biomedical terms (James et al., 2026) can assist in this task. Better transparency into requested data and its use in answering user queries will improve understanding and evaluation of the system’s behaviour. These steps will lead to an improved user evaluation system, likely combining benchmarks specific to proposed tasks with an open-ended scoring system.

The user-based evaluation system applied for Llemy requires refinement and further development, together with the changing landscape of technology and its possible applications. Open-ended user testing in computational biology is a limited, but expanding area (Mitchener et al., 2025), which can inform the development of highly interactive and multipurpose resources like molecular interaction maps. The framework proposed in this article is limited to three metrics of performance, which nevertheless allowed us to post-hoc identify main tasks explored by the user community. This informed the roadmap of Llemy development, and can be expanded into a set of task-specific evaluations, combining structured and open-ended scenarios of map exploration and analysis. Evaluators from communities of interest should be iteratively engaged when such refinements become available, including pathway and network interaction databases, like Reactome (Ragueneau et al., 2026), WikiPathways (Agrawal et al., 2024) or SIGNOR (Lo Surdo et al., 2026), and synergising with relevant efforts (Mohammadi et al.,

2025). The timeline of future user testing depends on system development, but can be supported by an engaged community, tailoring LLM support to user needs.

Study limitations

Our study faces a number of limitations. First, the limited number of users and the recruitment process may have affected the survey’s overall evaluation and conclusions. This can be seen in the relatively low number of participants in the “user” group, i.e., people with no prior knowledge about the development or analysis of molecular interaction maps. Second, Llemy features an inconsistent output, which affects the evaluation by its users. The answers vary for prompts having the same content and structure, including missing references, varying answer size, or overwhelming and unnecessary repetitions. This is a known behaviour of LLM-based systems for complex tasks (Wang and Wang, 2025), and is beyond the scope of our current work. Next, due to technical limitations, the size of the content processed by Llemy is restricted by the LLM capabilities, resulting in a limited scope of user queries. A potential mitigation for this issue would be to move from a long-context architecture to a (graph) retrieval-based approach, using solutions akin to those developed by Lobentanzer et al. (2025). Finally, since LLM providers continuously update model behaviour, the results should be interpreted as a snapshot of the tested model, its version and evaluation period. A model change may affect answer quality and scope. Llemy is designed so the underlying LLM can be exchanged, however any substantial model change should trigger re-evaluation using the same framework before. Use of GPT-4.1-nano was a pragmatic choice for a user-facing prototype because of its low latency and lower cost use. Future work should compare other models to quantify the trade-off between speed, cost, and response quality.

Conclusions

Llemy shows a promising use case for the usage of LLMs to interact with biological databases, particularly as a summarisation tool for an entry point into molecular interaction maps. The user-driven approach we took is important for evaluating the needs and perception of LLM-based systems in science, and yielded valuable insights on further development, in particular for more advanced tasks. With a rapidly changing landscape of available LLMs, it is important to construct sustainable systems that can be benchmarked when underlying models change. User-driven development and open-ended benchmarking systems can support these LLM comparisons, especially for complex tasks using interactive tools. Future plans also include re-running the study with an extended framework, as well as recruiting more uniformly distributed participants for a more generalisable evaluation and a better coverage of possible tasks.

Conflicts of interest

The authors declare that they have no competing interests.

Funding

This work was supported by the ONTOX project (<https://ontox-project.eu/>) which has received funding from the European Union’s Horizon 2020 Research and Innovation programme [grant agreement No 963845]. ONTOX is part of the ASPIS project

cluster (<https://aspis-cluster.eu/>). This work was supported by the European Union's Horizon Europe Framework Program (HEU/2022-2027) ERC INSTant-CARMa [grant number 101088919]. This work was supported by the Netherlands Organization for Scientific Research (NWO) under NWA project VHP4Safety [grant number 1292.19.272]. This work was supported by the ELIXIR Luxembourg Node (<https://elixir-luxembourg.org/>).

Data availability

Llemy's code and documentation are available at GitHub (<https://github.com/ontox-project/Llemy>), under the Apache 2.0 License (<https://www.apache.org/licenses/LICENSE-2.0>). A public instance is hosted at <https://llemy.vhp4safety.nl/>.

Author contributions statement

Conceptualization: MC, ID, LL, BS, AV, JSS, MT, LG and MO. **Methodology:** MC, ID, LL, BS, AV, JSS, MT, LG and MO. **Software:** MC, ID and MO. **Validation:** MC, ID, LL, BS, AV, JSS, AR, AV, AH, AM, AN, ACS, CA, EW, EK, FS, IB, LC, LM, LV, MEG, MK, MKönig, MW, NH, OL, PK, RO, TH, MT, LG and MO. **Resources:** MC, ID, LL, BS, AV, JSS, VS, RS, MT, LG and MO. **Writing - original draft:** MC, ID, LL, BS, AV, JSS, MT, LG and MO. **Writing - review and editing:** MC, ID, LL, BS, AV, JSS, AR, AV, AH, AM, AN, ACS, CA, EW, EK, FS, IB, LC, LM, LV, MEG, MK, MKönig, MW, NH, OL, PK, RO, TH, VS, RS, MT, LG and MO. **Supervision:** MT, LG, MO. **Funding acquisition:** VS, RS, LG, and MT.

Acknowledgments

Online browsing is supported by the MINERVA team (<https://minerva.uni.lu>) at the Bioinformatics Core of the Luxembourg Centre for Systems Biomedicine. The work presented in this article was supported by the ELIXIR Luxembourg tools and services. (<https://elixir-luxembourg.org>). We would like to thank Ozan Cinar for his support and advice in deploying Llemy on the VHP4Safety platform.

References

M. L. Acencio, M. Ostaszewski, A. Mazein, P. Rosenstiel, K. Aden, N. Mishra, V. Andersen, P. Sidiropoulos, A. Banos, A. Filia, S. Rahmouni, A. Finckh, W. Gu, R. Schneider, and V. Satagopam. The SYSCID map: a graphical and computational resource of molecular mechanisms across rheumatoid arthritis, systemic lupus erythematosus and inflammatory bowel disease. *Frontiers in Immunology*, 14: 1257321, 2023. ISSN 1664-3224. doi: 10.3389/fimmu.2023.1257321.

M. L. Acencio, O. Lopata, A. A. Hemedan, N. Dand, S. Weidinger, L. Paternoster, S. J. Brown, J. Rastrick, C. H. Smith, M. Hübenthal, R. Ramessur, M. Ndlovu, S. Ghosh, X. Wang, R. Schneider, V. Satagopam, and M. Ostaszewski. The Inflammatory Skin Disease map: an interactive computational resource focused on molecular mechanisms in psoriasis and atopic dermatitis. *The British Journal of Dermatology*, 194

(3):597–599, Feb. 2026. ISSN 1365-2133. doi: 10.1093/bjd/ljaf491.

A. Agrawal, H. Balci, K. Hanspers, S. L. Coort, M. Martens, D. N. Slenter, F. Ehrhart, D. Digles, A. Waagmeester, I. Wassink, T. Abbassi-Daloui, E. N. Lopes, A. Iyer, J. M. Acosta, L. G. Willighagen, K. Nishida, A. Riutta, H. Basaric, C. T. Evelo, E. L. Willighagen, M. Kutmon, and A. R. Pico. WikiPathways 2024: next generation pathway database. *Nucleic Acids Research*, 52(D1):D679–D689, Jan. 2024. ISSN 1362-4962. doi: 10.1093/nar/gkad960.

P. Devkota, S. D. Mohanty, and P. Manda. Optimizing large language models for ontology-based annotation: a study on gene ontology in biomedical texts. *BioData Mining*, 19(1):5, Dec. 2025. ISSN 1756-0381. doi: 10.1186/s13040-025-00507-z.

I. Djidrovski, R. Pieters, J. Legler, and M. Teunis. O-QT assistant: a multi-agent AI system for streamlined chemical hazard assessment and read-across analysis using the OECD QSAR toolbox API. *Computational Toxicology*, page 100395, Dec. 2025. ISSN 24681113. doi: 10.1016/j.comtox.2025.100395. URL <https://linkinghub.elsevier.com/retrieve/pii/S2468111325000556>.

Y. Feng, J. Wang, R. He, L. Zhou, and Y. Li. A retrieval-augmented knowledge mining method with deep thinking LLMs for biomedical research and clinical support. *GigaScience*, 14: g1af109, Jan. 2025. ISSN 2047-217X. doi: 10.1093/gigascience/g1af109.

K. A. Fujita, M. Ostaszewski, Y. Matsuoka, S. Ghosh, E. Glaab, C. Trefois, I. Crespo, T. M. Perumal, W. Jurkowski, P. M. A. Antony, N. Diederich, M. Buttini, A. Kodama, V. P. Satagopam, S. Eifes, A. Del Sol, R. Schneider, H. Kitano, and R. Balling. Integrating pathways of Parkinson's disease in a molecular interaction map. *Molecular Neurobiology*, 49(1):88–102, Feb. 2014. ISSN 1559-1182. doi: 10.1007/s12035-013-8489-4.

P. Gawron, E. Smula, R. Schneider, and M. Ostaszewski. Exploration and comparison of molecular mechanisms across diseases using MINERVA Net. *Protein Science: A Publication of the Protein Society*, 32(2):e4565, Feb. 2023. ISSN 1469-896X. doi: 10.1002/pro.4565.

A. A. Hemedan, V. Satagopam, R. Schneider, and M. Ostaszewski. Cohort-specific boolean models highlight different regulatory modules during Parkinson's disease progression. *iScience*, 27(10):110956, Oct. 2024. ISSN 2589-0042. doi: 10.1016/j.isci.2024.110956.

M. Hoch, S. Smita, K. Cesnulevicius, D. Lescheid, M. Schultz, O. Wolkenhauer, and S. Gupta. Network- and enrichment-based inference of phenotypes and targets from large-scale disease maps. *NPJ systems biology and applications*, 8(1):13, Apr. 2022. ISSN 2056-7189. doi: 10.1038/s41540-022-00222-z.

M. Hoch, S. Smita, K. Cesnulevicius, M. Schultz, D. Lescheid, O. Wolkenhauer, and S. Gupta. Network analyses reveal new insights into the effect of multicomponent Tr14 compared to single-component diclofenac in an acute inflammation model. *Journal of Inflammation*, 20(1):12, Mar. 2023. ISSN 1476-9255. doi: 10.1186/s12950-023-00335-0.

D. Hoksza, P. Gawron, M. Ostaszewski, E. Smula, and R. Schneider. MINERVA API and plugins: opening molecular network analysis and visualization to the community. *Bioinformatics*, 35(21):4496–4498, Nov. 2019. ISSN 1367-4811. doi: 10.1093/bioinformatics/btz286.

X. Hou, Y. Zhao, S. Wang, and H. Wang. Model Context Protocol (MCP): Landscape, Security Threats, and Future Research Directions, 2025. URL <https://arxiv.org/abs/2503.23278>. Version Number: 3.

Z. Hu, L. Song, J. Zhang, Z. Xiao, T. Wang, Z. Chen, N. J. Yuan, J. Lian, K. Ding, and H. Xiong. Explaining Length Bias in LLM-Based Preference Evaluations. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 6763–6794, Suzhou, China, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.findings-emnlp.358. URL <https://aclanthology.org/2025.findings-emnlp.358>.

F. James, D. Pratt, C. Churas, and A. Luna. textToKnowledgeGraph: generation of molecular interaction knowledge graphs using large language models for exploration in Cytoscape. *Bioinformatics*, 42(2):btag031, Feb. 2026. ISSN 1367-4811. doi: 10.1093/bioinformatics/btag031. URL <https://academic.oup.com/bioinformatics/article/doi/10.1093/bioinformatics/btag031/8430289>.

H. Jiang, S. Shi, Y. Yao, C. Jiang, and Q. Li. HypoChainer: A Collaborative System Combining LLMs and Knowledge Graphs for Hypothesis-Driven Scientific Discovery. *IEEE transactions on visualization and computer graphics*, 32(1):298–308, Jan. 2026. ISSN 1941-0506. doi: 10.1109/TVCG.2025.3633887.

S. M. Keating, D. Waltemath, M. König, F. Zhang, A. Dräger, C. Chaouiya, F. T. Bergmann, A. Finney, C. S. Gillespie, T. Helikar, S. Hoops, R. S. Malik-Sheriff, S. L. Moodie, I. I. Moraru, C. J. Myers, A. Naldi, B. G. Olivier, S. Sahle, J. C. Schaff, L. P. Smith, M. J. Swat, D. Thieffry, L. Watanabe, D. J. Wilkinson, M. L. Blinov, K. Begley, J. R. Faeder, H. F. Gómez, T. M. Hamm, Y. Inagaki, W. Liebermeister, A. L. Lister, D. Lucio, E. Mjolsness, C. J. Proctor, K. Raman, N. Rodriguez, C. A. Shaffer, B. E. Shapiro, J. Stelling, N. Swainston, N. Tanimura, J. Wagner, M. Meier-Schellersheim, H. M. Sauro, B. Palsson, H. Bolouri, H. Kitano, A. Funahashi, H. Hermjakob, J. C. Doyle, M. Hucka, and SBML Level 3 Community members. SBML Level 3: an extensible format for the exchange and reuse of biological models. *Molecular Systems Biology*, 16(8):e9110, Aug. 2020. ISSN 1744-4292. doi: 10.15252/msb.20199110.

T. Khan, M. Toufiq, M. Yurieva, N. Indrawattana, A. Jittmittraphap, N. Kosoltanapiwat, P. Pumirat, P. Sukphopetch, M. Vanaporn, K. Palucka, B. Syed Ahamed Kabeer, D. Rinchai, and D. Chaussabel. Automating candidate gene prioritization with large language models: from naive scoring to literature-grounded validation. *Bioinformatics*, 41(10):btaf541, Oct. 2025. ISSN 1367-4811. doi: 10.1093/bioinformatics/btaf541.

K. Kim, G. Shams, and K. K. Kim. From Seconds to Sentiments: Differential Effects of Chatbot Response Latency on Customer Evaluations. *International Journal of Human-Computer Interaction*, 42(1):597–612, Jan. 2026. ISSN 1044-7318, 1532-7590. doi: 10.1080/10447318.2025.2508915. URL <https://www.tandfonline.com/doi/full/10.1080/10447318.2025.2508915>.

M. Kuehl, D. P. Schaub, F. Carli, L. Heumos, M. Hellmig, C. Fernández-Zapata, N. Kaiser, J. Schaul, A. Kulaga, N. Usanov, S. Verse Community, C. F. Krebs, U. Panzer, S. Bonn, S. Lobentanzer, J. Saez-Rodriguez, and V. G. Puelles. BioContextAI is a community hub for agentic biomedical systems. *Nature Biotechnology*, 43(11):1755–1757, Nov. 2025. ISSN 1546-1696. doi: 10.1038/s41587-025-02900-9.

L. Ladeira, A. Verhoeven, J. Van Ertvelde, J. Jiang, A. Gamba, J. Sanz-Serrano, T. Vanhaecke, H. J. Heusinkveld, R. Jover, M. Vinken, L. Geris, and B. Staumont. Unlocking liver physiology: comprehensive pathway maps for mechanistic understanding. *Frontiers in Toxicology*, 7: 1619651, July 2025. ISSN 2673-3080. doi: 10.3389/ftox.2025.1619651. URL <https://www.frontiersin.org/articles/10.3389/ftox.2025.1619651/full>.

J. M. Laurent, J. D. Janizek, M. Ruzo, M. M. Hinks, M. J. Hammerling, S. Narayanan, M. Ponnappati, A. D. White, and S. G. Rodrigues. LAB-Bench: Measuring Capabilities of Language Models for Biology Research, 2024. URL <https://arxiv.org/abs/2407.10362>. Version Number: 3.

N. Le Novère, M. Hucka, H. Mi, S. Moodie, F. Schreiber, A. Sorokin, E. Demir, K. Wegner, M. I. Aladjem, S. M. Wimalaratne, F. T. Bergman, R. Gauges, P. Ghazal, H. Kawaji, L. Li, Y. Matsuoka, A. Villéger, S. E. Boyd, L. Calzone, M. Courtot, U. Dogrusoz, T. C. Freeman, A. Funahashi, S. Ghosh, A. Jouraku, S. Kim, F. Kolpakov, A. Luna, S. Sahle, E. Schmidt, S. Watterson, G. Wu, I. Goryanin, D. B. Kell, C. Sander, H. Sauro, J. L. Snoep, K. Kohn, and H. Kitano. The Systems Biology Graphical Notation. *Nature Biotechnology*, 27(8):735–741, Aug. 2009. ISSN 1546-1696. doi: 10.1038/nbt.1558.

P. Lo Surdo, M. Iannuccelli, K. Karis, E. Meo, P. Omid, M. Tosoni, S. Graziosi, S. Panni, M. T. Fiorillo, L. Licata, F. Sacco, B. M. Gyori, and L. Perfetto. SIGNOR 4.0: the 2025 update with focus on phosphorylation data. *Nucleic Acids Research*, 54(D1):D682–D690, Jan. 2026. ISSN 1362-4962. doi: 10.1093/nar/gkaf1237.

S. Lobentanzer, P. Aloy, J. Baumbach, B. Bohar, V. J. Carey, P. Charoentong, K. Danhauser, T. Doğan, J. Dreo, I. Dunham, E. Farr, A. Fernandez-Torras, B. M. Gyori, M. Hartung, C. T. Hoyt, C. Klein, T. Korcsmaros, A. Maier, M. Mann, D. Ochoa, E. Pareja-Lorente, F. Popp, M. Preusse, N. Probul, B. Schwikowski, B. Sen, M. T. Strauss, D. Turei, E. Ulusoy, D. Waltemath, J. A. H. Wodke, and J. Saez-Rodriguez. Democratizing knowledge representation with BioCypher. *Nature Biotechnology*, 41(8):1056–1059, Aug. 2023. ISSN 1546-1696. doi: 10.1038/s41587-023-01848-y.

S. Lobentanzer, S. Feng, N. Bruderer, A. Maier, BioChatter Consortium, C. Wang, J. Baumbach, J. Abreu-Vicente, N. Krehl, Q. Ma, T. Lemberger, and J. Saez-Rodriguez. A platform for the biomedical application of large language models. *Nature Biotechnology*, 43(2):166–169, Feb. 2025. ISSN 1546-1696. doi: 10.1038/s41587-024-02534-3.

O. Lopata, M. L. Acencio, X. Wang, A. A. Hemedan, M. J. Chao, S. A. Jelinsky, F. Tran, P. Rosenstiel, A. Y. F. Li Yim, R. Schneider, V. Satagopam, and M. Ostaszewski. Identification of shared and unique mechanisms of atopic dermatitis and ulcerative colitis by construction and computational analysis of disease maps. *Computational and Structural Biotechnology Journal*, 27: 4007–4018, 2025. ISSN 2001-0370. doi: 10.1016/j.csbj.2025.09.008.

M. Maslych, M. Katebi, C. Lee, Y. Hmaiti, A. Ghasemaghahi, C. Pumarada, J. Palmer, E. Segarra Martinez, M. Emporio, W. Snipes, R. P. McMahan, and J. J. LaViola Jr. Mitigating Response Delays in Free-Form Conversations with LLM-powered Intelligent Virtual Agents. In *Proceedings of the 7th ACM Conference on Conversational User Interfaces*, pages

- 1–15, Waterloo ON Canada, July 2025. ACM. ISBN 979-8-4007-1527-3. doi: 10.1145/3719160.3736636. URL <https://dl.acm.org/doi/10.1145/3719160.3736636>.
- A. Mazein, M. Ostaszewski, I. Kuperstein, S. Watterson, N. Le Novère, D. Lefaudeux, B. De Meulder, J. Pellet, I. Balaour, M. Saqi, M. M. Nogueira, F. He, A. Parton, N. Lemonnier, P. Gawron, S. Gebel, P. Hainaut, M. Ollert, U. Dogrusoz, E. Barillot, A. Zinovyev, R. Schneider, R. Balling, and C. Auffray. Systems medicine disease maps: community-driven comprehensive representation of disease mechanisms. *NPJ systems biology and applications*, 4:21, 2018. ISSN 2056-7189. doi: 10.1038/s41540-018-0059-y.
- A. Mazein, O. Ivanova, I. Balaour, M. Ostaszewski, V. Berzhitskaya, T. Serebriyskaya, T. Ligon, J. Hasenauer, B. De Meulder, R. W. Overall, L. Roy, R. G. Knowles, C. E. Wheelock, S.-E. Dahlen, K. F. Chung, I. M. Adcock, G. Roberts, R. Djukanovic, J. Pellet, P. Gawron, R. Balling, A. H. Maitland-van der Zee, R. Schneider, P. J. Sterk, C. Auffray, U-BIOPRED Study Group, and eTRIKS Consortium. AsthmaMap: An interactive knowledge repository for mechanisms of asthma. *The Journal of Allergy and Clinical Immunology*, 147(3):853–856, Mar. 2021. ISSN 1097-6825. doi: 10.1016/j.jaci.2020.11.032.
- A. Mazein, O. Lopata, K. Reiche, K. Sewald, M. Alb, C. Sakellariou, P. Gogesch, H. Morgan, V. Neuhaus, N.-N. Pham, C. Sommer, E. Perkins, B. Fogal, M. Shoaib, R. Schneider, V. Satagopam, and M. Ostaszewski. An explorable model of an adverse outcome pathway of cytokine release syndrome related to the administration of immunomodulatory biotherapeutics and cellular therapies. *Frontiers in Immunology*, 16:1601670, 2025. ISSN 1664-3224. doi: 10.3389/fimmu.2025.1601670.
- I. Mazein, A. Rougny, A. Mazein, R. Henkel, L. Gütebier, L. Michaelis, M. Ostaszewski, R. Schneider, V. Satagopam, L. J. Jensen, D. Waltemath, J. A. H. Wodke, and I. Balaour. Graph databases in systems biology: a systematic review. *Briefings in Bioinformatics*, 25(6):bbae561, Sept. 2024. ISSN 1477-4054. doi: 10.1093/bib/bbae561.
- L. Mitchener, J. M. Laurent, A. Andonian, B. Tenmann, S. Narayanan, G. P. Wellawatte, A. White, L. Sani, and S. G. Rodriques. BixBench: a Comprehensive Benchmark for LLM-based Agents in Computational Biology, 2025. URL <https://arxiv.org/abs/2503.00096>. Version Number: 3.
- H. Mohammadi, F. Almodaresi, G. F. J. Hogue, A. Wright, M. Orlic-Milacic, N. T. Li, A. Mawani, and L. Stein. React-to-Me: A Conversational Interface for Interactive Exploration of the Reactome Pathway Knowledgebase. *bioRxiv: The Preprint Server for Biology*, page 2025.12.12.693752, Dec. 2025. ISSN 2692-8205. doi: 10.64898/2025.12.12.693752.
- V. Nagarajan, G. Shi, R. Horai, C.-R. Yu, J. Gopalakrishnan, M. Yadav, M. H. Liew, C. Gentilucci, and R. R. Caspi. IAN: An Intelligent System for Omics Data Analysis and Discovery. *bioRxiv: The Preprint Server for Biology*, page 2025.03.06.640921, Mar. 2025. ISSN 2692-8205. doi: 10.1101/2025.03.06.640921.
- M. Ostaszewski, A. Niarakis, A. Mazein, I. Kuperstein, R. Phair, A. Orta-Resendiz, V. Singh, S. S. Aghamiri, M. L. Acencio, E. Glaab, A. Ruepp, G. Fobo, C. Montrone, B. Brauner, G. Frishman, L. C. Monraz Gómez, J. Somers, M. Hoch, S. Kumar Gupta, J. Scheel, H. Borlinghaus, T. Czauderna, F. Schreiber, A. Montagud, M. Ponce de Leon, A. Funahashi, Y. Hiki, N. Hiroi, T. G. Yamada, A. Dräger, A. Renz, M. Naveez, Z. Bocskei, F. Messina, D. Börnigen, L. Fergusson, M. Conti, M. Rameil, V. Nakonecni, J. Vanhoefer, L. Schmiester, M. Wang, E. E. Ackerman, J. E. Shoemaker, J. Zucker, K. Oxford, J. Teuton, E. Kocakaya, G. Y. Summak, K. Hanspers, M. Kutmon, S. Coort, L. Eijssen, F. Ehrhart, D. A. B. Rex, D. Slenter, M. Martens, N. Pham, R. Haw, B. Jassal, L. Matthews, M. Orlic-Milacic, A. Senff Ribeiro, K. Rothfels, V. Shamovsky, R. Stephan, C. Sevilla, T. Varusai, J.-M. Ravel, R. Fraser, V. Ortseifen, S. Marchesi, P. Gawron, E. Smula, L. Heirendt, V. Satagopam, G. Wu, A. Riutta, M. Golebiewski, S. Owen, C. Goble, X. Hu, R. W. Overall, D. Maier, A. Bauch, B. M. Gyori, J. A. Bachman, C. Vega, V. Grouès, M. Vazquez, P. Porras, L. Licata, M. Iannuccelli, F. Sacco, A. Nesterova, A. Yuryev, A. de Waard, D. Turei, A. Luna, O. Babur, S. Soliman, A. Valdeolivas, M. Esteban-Medina, M. Peña-Chilet, K. Rian, T. Helikar, B. L. Puniya, D. Modos, A. Treveil, M. Olbei, B. De Meulder, S. Ballereau, A. Dugourd, A. Naldi, V. Noël, L. Calzone, C. Sander, E. Demir, T. Korcsmaros, T. C. Freeman, F. Augé, J. S. Beckmann, J. Hasenauer, O. Wolkenhauer, E. L. Wilighagen, A. R. Pico, C. T. Evelo, M. E. Gillespie, L. D. Stein, H. Hermjakob, P. D'Eustachio, J. Saez-Rodriguez, J. Dopazo, A. Valencia, H. Kitano, E. Barillot, C. Auffray, R. Balling, R. Schneider, and COVID-19 Disease Map Community. COVID19 Disease Map, a computational knowledge repository of virus-host interaction mechanisms. *Molecular Systems Biology*, 17(10):e10387, Oct. 2021. ISSN 1744-4292. doi: 10.15252/msb.202110387.
- E. Ragueneau, C. Gong, P. Sinquin, C. Sevilla, D. Beavers, A. Gretnier, J. Griss, G. F. J. Hogue, N. T. Li, L. Matthews, B. May, M. Milacic, H. Mohammadi, R. Petryszak, K. Rothfels, V. Shamovsky, R. Stephan, K. Tiwari, J. Weiser, A. Wright, M. Gillespie, G. Wu, L. Stein, H. Hermjakob, and P. D'Eustachio. The Reactome Knowledgebase 2026. *Nucleic Acids Research*, 54(D1):D673–D681, Jan. 2026. ISSN 1362-4962. doi: 10.1093/nar/gkaf1223.
- C. N. Serhan, S. K. Gupta, M. Perretti, C. Godson, E. Brennan, Y. Li, O. Soehnlein, T. Shimizu, O. Werz, V. Chiurchiù, A. Azzi, M. Dubourdeau, S. S. Gupta, P. Schopohl, M. Hoch, D. Gjorgevikj, F. M. Khan, D. Brauer, A. Tripathi, K. Cesnulevicius, D. Lescheid, M. Schultz, E. Särndahl, D. Repsilber, R. Kruse, A. Sala, J. Z. Haeggström, B. D. Levy, J. G. Filep, and O. Wolkenhauer. The Atlas of Inflammation Resolution (AIR). *Molecular Aspects of Medicine*, 74:100894, Aug. 2020. ISSN 1872-9452. doi: 10.1016/j.mam.2020.100894.
- A. Shakhbazov. Fine-tuning LLM hyperparameters to align semantic and physiological contexts of aging-related pathways. *Molecular Diversity*, June 2025. ISSN 1573-501X. doi: 10.1007/s11030-025-11226-2.
- K. Sokolova, D. Kosenkov, K. Nallamotu, S. Vedula, D. Sokolov, G. Sapiro, and O. G. Troyanskaya. An Evidence-Grounded Research Assistant for Functional Genomics and Drug Target Assessment. *bioRxiv: The Preprint Server for Biology*, page 2025.12.30.697073, Dec. 2025. ISSN 2692-8205. doi: 10.64898/2025.12.30.697073.
- D. Stachura, J. Konieczna, and A. Nowak. Are Smaller Open-Weight LLMs Closing the Gap to Proprietary Models for Biomedical Question Answering?, 2025. URL <https://arxiv.org/abs/2509.18843>. Version Number: 1.

1 B. Staumont, L. Ladeira, A. Gamba, H. J. Heusinkveld,
2 A. Piersma, R. Masereeuw, T. Vanhaecke, M. Teunis, T. H.
3 Luechtefeld, T. Hartung, R. Jover, M. Vinken, and L. Geris.
4 Mapping Physiology: A Systems Biology Approach for the
5 Development of Alternative Methods in Toxicology. *ALTEX*,
6 42, Apr. 2025. doi: 10.14573/altex.2412241. URL <https://doi.org/10.14573/altex.2412241>.
7
8 M. Vinken, E. Benfenati, F. Busquet, J. Castell, D.-A. Clevert,
9 T. M. De Kok, H. Dirven, E. Fritsche, L. Geris, R. Gozalbes,
10 T. Hartung, D. Jennen, R. Jover, H. Kandarova, N. Kramer,
11 C. Krul, T. Luechtefeld, R. Masereeuw, E. Roggen, S. Schaller,
12 T. Vanhaecke, C. Yang, and A. H. Piersma. Safer chemicals
13 using less animals: kick-off of the European ONTOX project.
14 *Toxicology*, 458:152846, June 2021. ISSN 0300483X. doi:
15 10.1016/j.tox.2021.152846. URL <https://linkinghub.elsevier.com/retrieve/pii/S0300483X21001694>.
16
17 E. J. Wang, c. Oğuztüzün, R. Xu, and Z. Gao. Comparative
18 evaluation of large language models in retrieving known and
19 predicting novel drug combinations. *Journal of Alzheimer's*
20 *disease: JAD*, page 13872877261415635, Jan. 2026. ISSN 1875-
21 8908. doi: 10.1177/13872877261415635.
22
23 J. J. Wang and V. X. Wang. Assessing Consistency and
24 Reproducibility in the Outputs of Large Language Models:
25 Evidence Across Diverse Finance and Accounting Tasks, 2025.
26 URL <https://arxiv.org/abs/2503.16974>. Version Number: 4.
27
28 Y.-L. Wang and C.-W. Lo. The effects of response time on
29 older and young adults' interaction experience with Chatbot.
30 *BMC Psychology*, 13(1):150, Feb. 2025. ISSN 2050-7283. doi:
31 10.1186/s40359-025-02459-9. URL <https://bmcp psychology.biomedcentral.com/articles/10.1186/s40359-025-02459-9>.
32
33 G. Wu, L. Matthews, N. Boyer, M. Milacic, D. Beavers, N. T. Li,
34 B. May, K. Rothfels, V. Shamovsky, R. Stephan, M. Gillespie,
35 H. Hermjakob, P. D'Eustachio, and L. Stein. Application of
36 Large Language Models for Annotating Genes into Reactome
37 Pathways. *bioRxiv: The Preprint Server for Biology*, page
38 2025.12.20.695723, Dec. 2025. ISSN 2692-8205. doi: 10.64898/
39 2025.12.20.695723.
40
41 N. Zerrouk, S. Aghakhani, V. Singh, F. Augé, and A. Niarakis. A
42 Mechanistic Cellular Atlas of the Rheumatic Joint. *Frontiers in*
43 *Systems Biology*, 2:925791, July 2022. ISSN 2674-0702. doi: 10.
44 3389/fsysb.2022.925791. URL <https://www.frontiersin.org/articles/10.3389/fsysb.2022.925791/full>.
45
46 N. Zerrouk, R. Alcraft, B. A. Hall, F. Augé, and A. Niarakis.
47 Large-scale computational modelling of the M1 and M2
48 synovial macrophages in rheumatoid arthritis. *NPJ systems*
49 *biology and applications*, 10(1):10, Jan. 2024. ISSN 2056-7189.
50 doi: 10.1038/s41540-024-00337-5.

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51

What is the scope of this map? Give me a brief summary of the biology represented.

User prompt

+

Ensure rich context in the answer, be scientifically precise, provide accurate literature references.

User prompt

Enriched user prompt

Map selection

Map retrieval

Map content

Prompt synthesis

LLM query

LLM synthesis agent

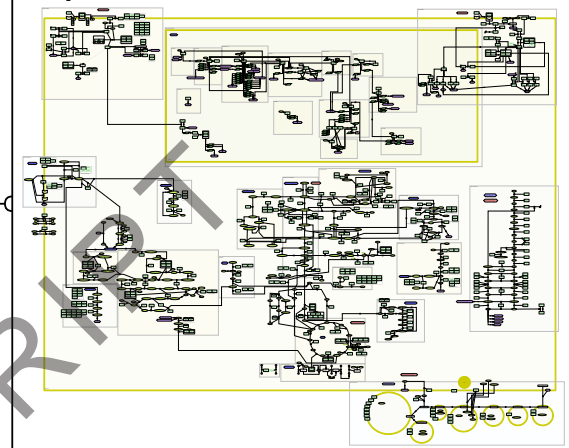
System status
Response time

Answer with references

User

Performance evaluation

Map in the MINERVA Platform



Map contents as JSON

```
[{"id":6672,"type":"Positive influence",
"references":[{"database":"PUBMED","database_id":
"17408758"}],"sources":[{"node_id":18340}],
"targets":[{"node_id":18338}],{"id":6673,
"type":"State transition","modifiers":[{"node_id":
18339,"modification_type":"PhysicalStimulation"}],
"references":[{"database":"PUBMED","database_id":
"17891149"}],"sources":[{"node_id":18337}],
"targets":[{"node_id":18340}],{"id":6674,
"type":"State transition","modifiers":[{"node_id":
18334,"modification_type":"PhysicalStimulation"}],
"sources":[{"node_id":18333},"targets":
[{"node_id":18335}]} ...
```

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60



-  Home
-  Informed consent form
-  Instructions

MINERVA Project Selection

Select MINERVA Project:

Liver Lipid Metabolis... ▼

 View selected map

Example Questions

- What is the scope of this map? Give me a brief ...
- What is the overall scope of the disease map (m...
- What are the inputs, regulators, and phenotypic...

✔ OpenAI API key is already set

✔ Unique identifier is already set: 29eb1e0d0259fbcfbfb88ac09c511c20073bdf18b. Please write it down for the survey.

I have read and agree to the informed consent information and I want to participate in the Llemy user research.

☒ Yes ☐ No

 View informed consent details

Llemy

An Agentic System for MINERVA Map Exploration

Select a MINERVA map in the drop-down menu on the left, and ask questions about it.

You can:

- Type your own question in the input box below
- Try one of the example questions from the sidebar All questions are independant from each other, Llemy uses only the data from the selected map to answer your question.

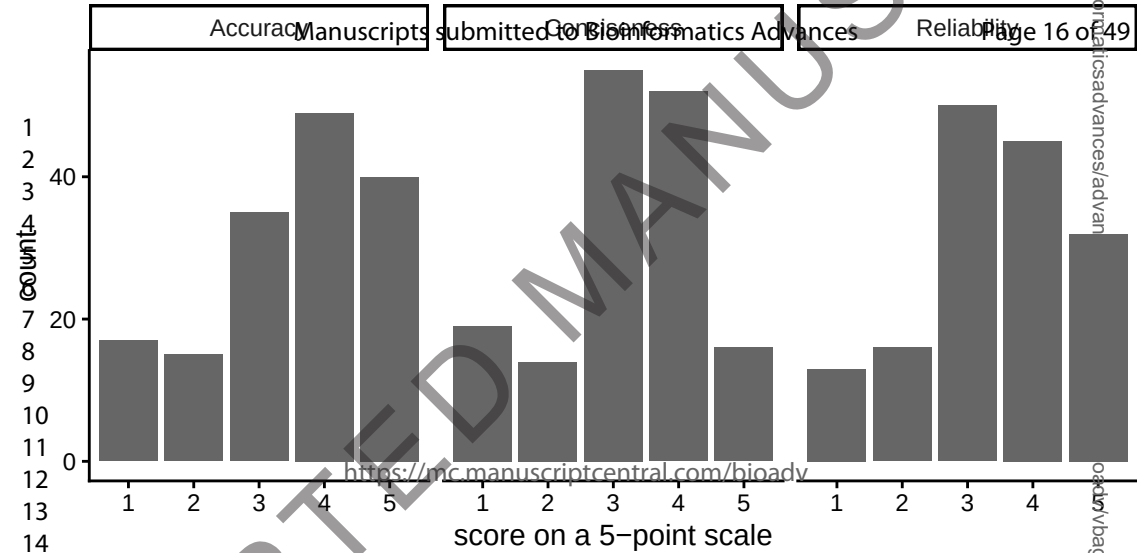
Ask about biology, transporters, or metabolism...



This is a prototype application for research purposes only.

This work was developed in the context of the ONTOX project and VHP4Safety project.

Developped with Streamlit, LangChain and OpenAI (gpt-4.1-nano). License Apache 2.0.
<https://mc.manuscriptcentral.com/bioadv>



Accuracy

M = 3.69**M = 3.31****M = 3.46**

Conciseness

M = 3.35**M = 3.04****M = 3.15**

Reliability

M = 3.51**M = 3.36****M = 3.39**

score on a 5-point scale

