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

**Authors:**
Marie Corradi[1], Ivo Djidrovski[2], Luiz Ladeira[3], Bernard Staumont[3], Anouk Verhoeven[4], Julen Sanz Serrano[4], Adrien Rougny[5], Ahmad Vaez[6,7,8], Ahmed Hemedan[9], Alexander Mazein[10], Anna Niarakis[11], Arthur de Carvalho e Silva[12], Charles Auffray[13], Egon Wilighagen[14], Eliska Kuchovska[15], Falk Schreiber[16], Irina Balaur[17], Laurence Calzone[18], Lisa Matthews[19], Lorenzo Veschini[20], Marc E Gillespie[21], Martina Kutmon[22], Matthias König[23,24], Matti van Welzen[25], Noriko Hiroi[26], Oxana Lopata[5], Pierre Klemmer[5], Rupert Overall[23], Tim Hofer[26], Venkata Satagopam[5], Reinhard Schneider[5], Marc Teunis[1], Liesbet Geris[3], Marek Ostaszewski[5,*]

1. Innovative Testing in Life Sciences and Chemistry, HU - University of Applied Sciences Utrecht
2. Institute of Risk Assessment Sciences (IRAS), Utrecht University
3. Biomechanics Research Unit, GIGA Institute, University of Liège
4. Department of Pharmaceutical and Pharmacological Sciences, Vrije Universiteit Brussel (VUB), Vrije Universiteit Brussel
5. Luxembourg Centre for Systems Biomedicine, University of Luxembourg
5. Unit of Genetic Epidemiology & Bioinformatics, Department of Epidemiology, University Medical Center Groningen, University of Groningen
6. Department of Bioinformatics, Isfahan University of Medical Sciences, Isfahan, Iran
7. Medical Image and Signal Processing Research Center, Isfahan University of Medical Sciences, Isfahan, Iran
8. Transversal Translational Medicine, Luxembourg Institute of Health
9. Department of Cancer Research, Luxembourg Institute of Health
10. Department of Biology, Faculty of Science and Engineering, University of Toulouse, CBI, CNRS
11. School of Biosciences, University of Birmingham
12. European Institute for Systems Biology and Medicine
13. Translational Genomics, NUTRIM, FHML, Maastricht University
14. IUF - Leibniz Research Institute for Environmental Medicine
15. Department of Computer and Information Science, University of Konstanz
16. Luxembourg National Data Service
17. Computational Systems Biology of Cancer, Institut Curie, U1331 INSERM, Mines ParisTech, PSL
18. Biochemistry and Molecular Pharmacology, NYU Grossman School of Medicine
19. Biocomplexity Institute, Department of Intelligent Systems Engineering, Indiana University at Bloomington
20. Department of Pharmaceutical Sciences, St. John's University
21. Maastricht Centre for Systems Biology and Bioinformatics (MaCSBio), Maastricht University
22. Institute of Biology, Humboldt University of Berlin
23. Institute of Structural Mechanics and Dynamics in Aerospace Engineering, University of Stuttgart
24. Department of Systems Biology and Bioinformatics, University of Rostock
25. Department of Electrical and Electronic Engineering, Kanagawa Institute of Technology
26. Department of Chemical Toxicology, Norwegian Institute of Public Health

* Corresponding author

# 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 Llemy, 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 open-weight LLMs to support the needs of the open research environment in an ever-changing technology landscape.


# 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 (I. 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 of these resources is comprehensive graphical representations of cellular and molecular processes associated with specific organ functions under normal or pathological conditions (A. 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 standards of Systems Biology Graphical Notation (SBGN) (Le Novère et al., 2009) and Systems Biology Markup Language (SBML) (Keating et al., 2020). They are hosted using the MINERVA Platform (Hoksza et al., 2019), which provides functionalities for visual exploration and programmatic access. Currently a number of such 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; A. 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 (A. Mazein et al., 2025). Importantly, repositories hosting 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 such structured knowledge. Their applications range from general-purpose foundational models to 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 (Shakhbazau, 2025), or searching for drug combinations in Alzheimer's disease (E. J. Wang et al., 2026). For processing structured knowledge, LLM-based workflows have been proposed to generate knowledge graphs from the literature and to process them for prioritising experimental results (Feng et al., 2025) and exploring clinically valid hypotheses (Khan et al., 2025). In parallel, 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, assisting 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 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 GPT4.1-nano, Streamlit (v1.50.0) for the front-end, and LangChain (v0.3.27) as the agentic backend. 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, intentionally designed to test the system's ability to remain within the scope of its source maps, was removed from the subsequent analysis. 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.

**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 `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

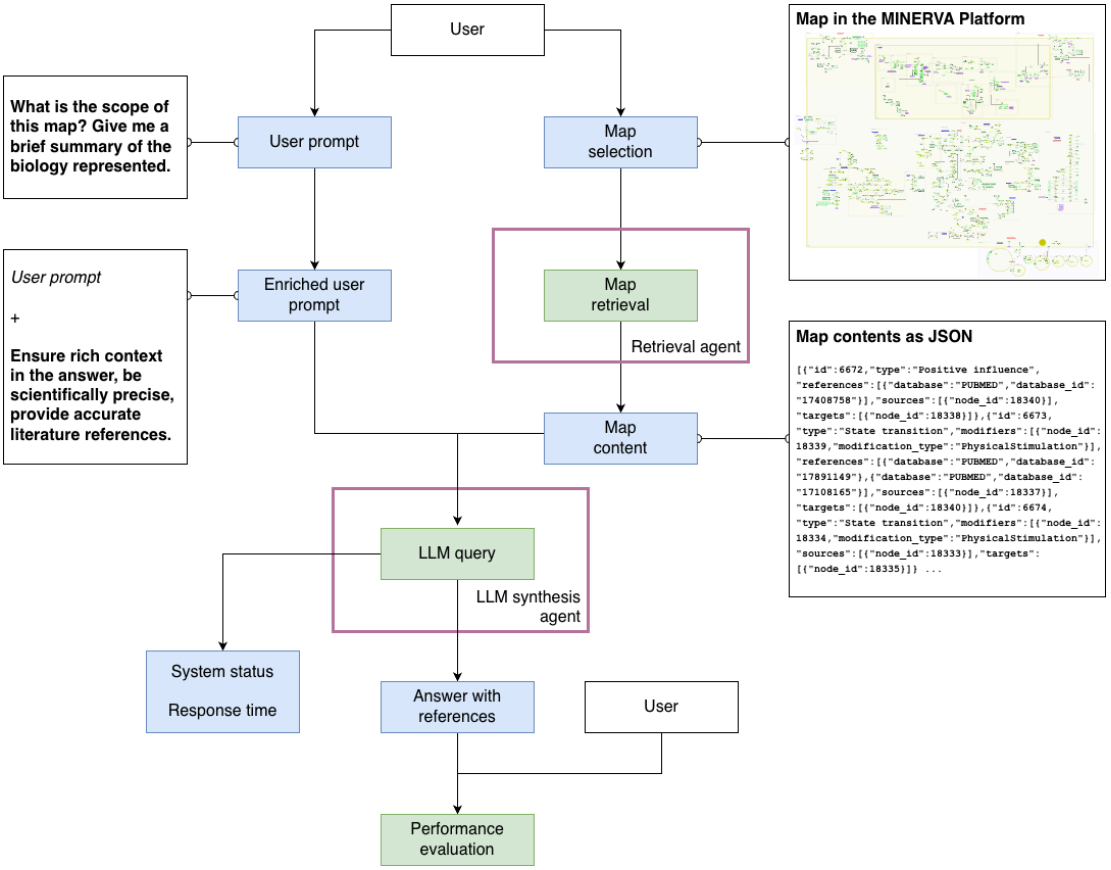


**Figure 1**: The workflow of Llemy. Users provide their prompt and select a map they want to process. Then, the prompt is enriched with instructions for the content of the answer (see full system prompt in Supplementary P1), and, in parallel, map data is fetched from the corresponding MINERVA Platform. 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.

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. 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.

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 and carbon emissions for each query were also calculated; for the latter, unfortunately, the calculation mechanism failed in 90% of the cases so we were unable to draw any conclusions.

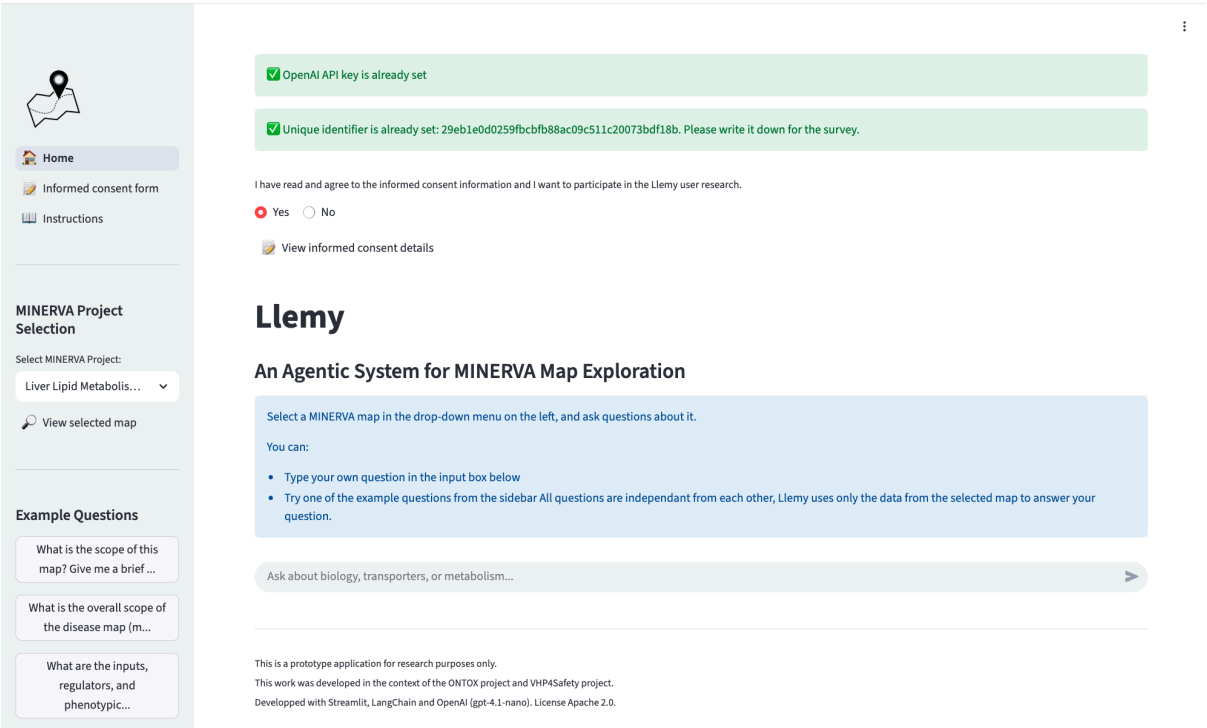


**Figure 2**: Screenshot of Llemy's welcome page. The left-side panel comprises i) a section with links to the home page, informed consent form, and user manual; ii) a map selection drop-down menu, and a link to the selected map; iii) a list of example prompts (see Supplementary U1 for a complete list).

Llemy is deployed as a service on the VHP4Safety cloud (https://llemy.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. Llemy can be used without providing feedback.

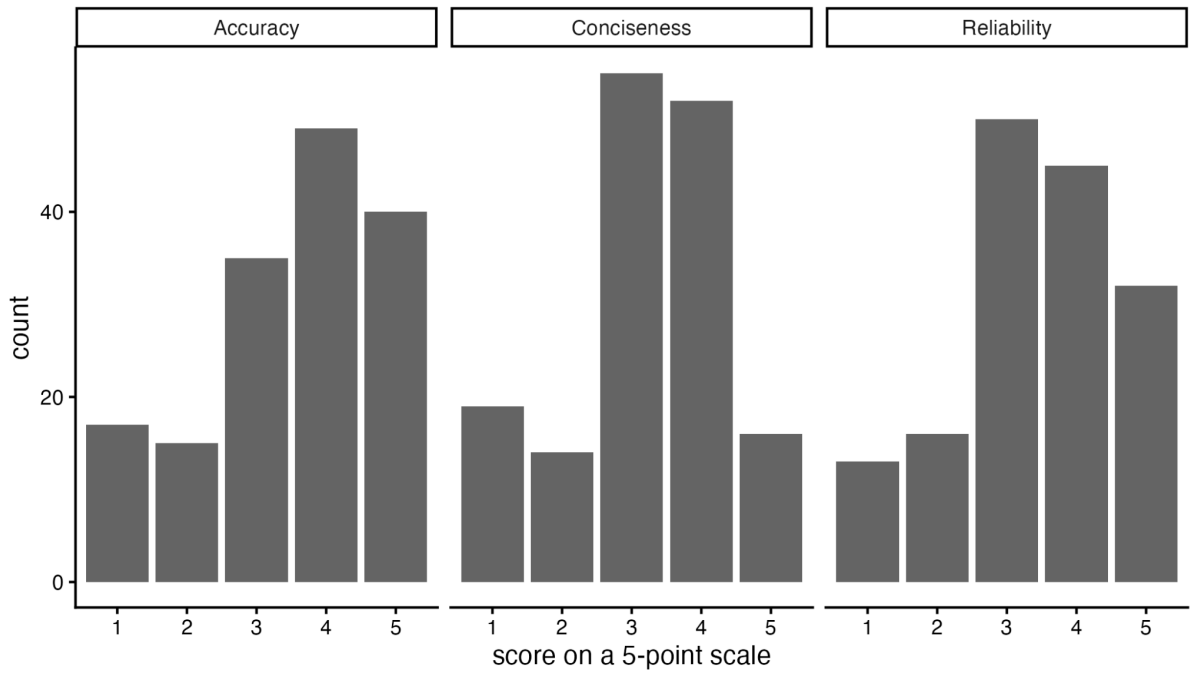


**Figure 3**: Score distribution by evaluation criterion.

## 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.

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

| Prompt 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 having specific properties (e.g., kinases, ligands, enzymes, transcription factors, protein interactors), or upstream or downstream targets (e.g., receptors, organelles, phenotypes) outside of 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) |

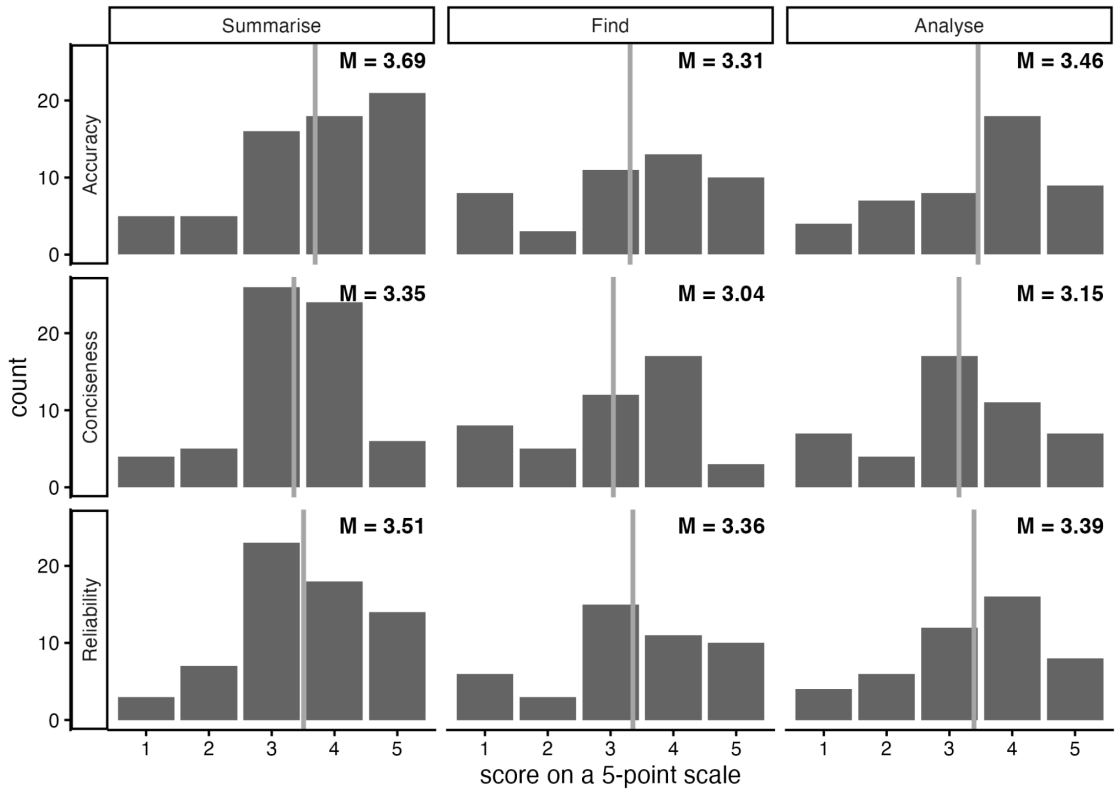


**Figure 4**: Distribution of tasks performed by the users testing Llemy, grouped into three categories. M = average.

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

Overall, we summarised individual user comments having particular score levels as listed in Supplementary U4. No individual comments on the conciseness of the system were submitted.

## Overall user feedback

Results of 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 Llemy. 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.

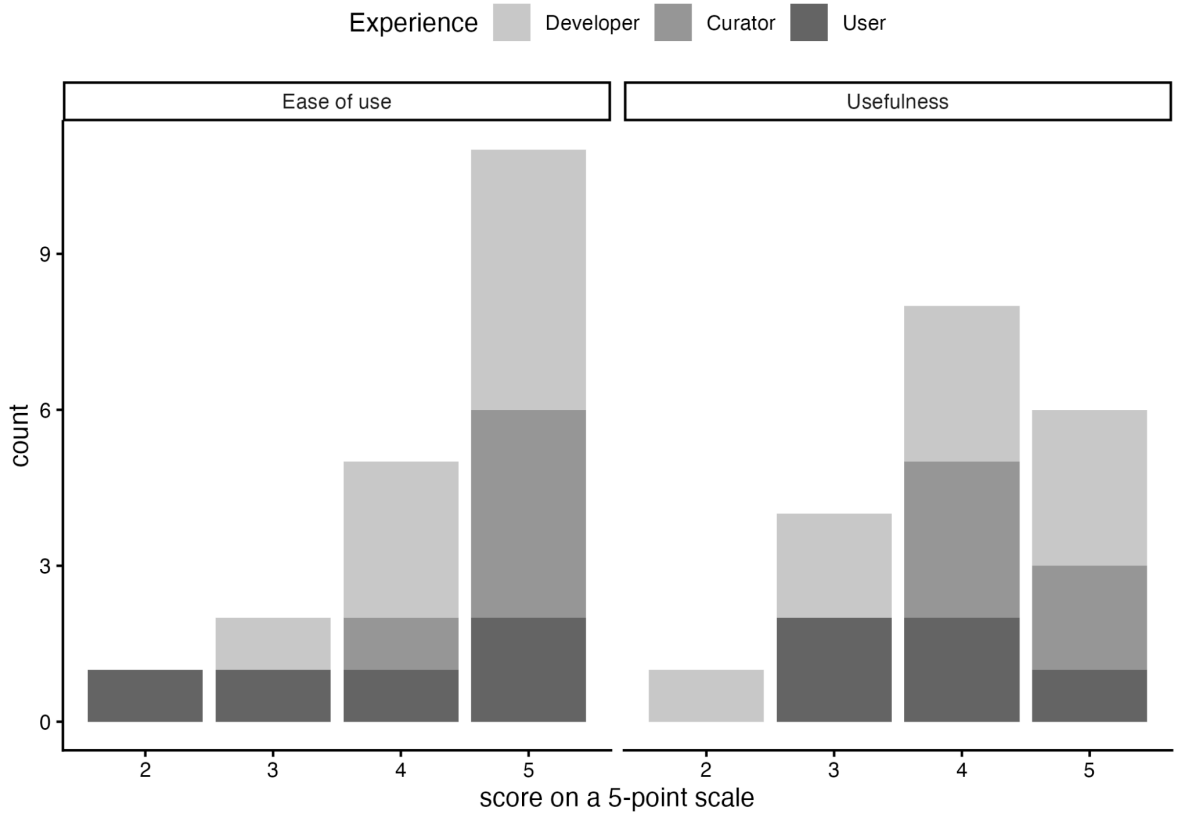


**Figure 5**: Usefulness and usability based on experience with molecular interaction maps. Developer - participants building workflows including maps; Curator - participants creating maps; User - participants using the maps for knowledge retrieval and data visualisation. In case of multiple roles, the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

# Discussion

**Advantages of the LLM-based approach**

The potential of LLMs to accelerate research is a topic of intense investigation - including in the field of molecular pathways and their regulation. In this work, we explore the capability of LLM-based systems to process, summarise, analyse, and extend structured, focused knowledge repositories describing molecular pathways involved in human physiology and pathology. We introduce Llemy, a system for summarising knowledge graphs based on user prompts. Our approach resembles proposed LLM solutions to refine knowledge graphs from literature (Feng et al., 2025; Khan et al., 2025), but the sources of the system are manually curated diagrams. This allows Llemy 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 then tested by an independent group of 25 users at the level of individual prompts, and via a survey of their overall evaluation. Based on 156 individual prompts, we proposed three categories of prompts 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, Llemy workflows can be tailored to the tasks performed by users. This is similar to the functionality of the Alvessa system proposed in (Sokolova et al., 2025), but focuses on the nature of the task rather than on specific resources to be queried based on the system prompt.

**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 Llemy can give more favourable evaluations to its answers. All users generally rated the usability of Llemy 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 Llemy useful (score of 4 or more).

Reproducibility of results and the sustainability of the system are key considerations in developing systems such as Llemy. 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 (J. J. Wang & Wang, 2025). In order to improve this situation, open-weight models may be considered, contrary to a commercial LLM used currently. Importantly, open-weight LLM models show promising 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**

Overall, Llemy and similar 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 use of open 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. Llemy needs improvements 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. 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 mulitpurpose 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 (Rague neau 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 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 (J. J. Wang & Wang, 2025), and is beyond the scope of our current work. Finally, 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).

# 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.

## Funding

This work was performed in the context of the ONTOX project (https://ontox-project.eu/) which has received funding from the European Union's Horizon 2020 Research and Innovation programme under grant agreement No 963845. ONTOX is part of the ASPIS project cluster (https://aspis-cluster.eu/). This work was also funded by the European Union's Horizon Europe Framework Program (HEU/2022-2027) ERC INSTant CARMA (101088919) and received funding from the Netherlands Organization for Scientific Research (NWO) under NWA project VHP4Safety 1292.19.272 and from the ELIXIR Luxembourg Node.

## Declaration of competing interests

The authors declare they have no competing interests.

## Disclosure of Ethical Statements

N/A

## Consent for publication

All authors agree to the publication of this study.

## CRediT author 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 & 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.

## Acknowledgements

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).

## 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/.

## List of abbreviations

API: Application Programming Interface;

GUI: Graphical User Interface;

HGNC: HUGO Gene Name Committee

LLM: Large Language Model;

MCP: Model Context Protocol;

MINERVA: Molecular Interaction NEtwoRk VisuAlization;

SBGN: Systems Biology Graphical Notation;

## References

Acencio, M. L., Lopata, O., Hemedan, A. A., Dand, N., Weidinger, S., Paternoster, L., Brown, S. J., Rastrick, J., Smith, C. H., Hübenthal, M., Ramessur, R., Ndlovu, M., Ghosh, S., Wang, X., Schneider, R., Satagopam, V., & Ostaszewski, M. (2026). 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. https://doi.org/10.1093/bjd/ljaf491

Acencio, M. L., Ostaszewski, M., Mazein, A., Rosenstiel, P., Aden, K., Mishra, N., Andersen, V., Sidiropoulos, P., Banos, A., Filia, A., Rahmouni, S., Finckh, A., Gu, W., Schneider, R., & Satagopam, V. (2023). 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. https://doi.org/10.3389/fimmu.2023.1257321

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

Azam, M., Chen, Y., Arowolo, M. O., Liu, H., Popescu, M., & Xu, D. (2024). A comprehensive evaluation of large language models in mining gene relations and pathway knowledge. *Quantitative Biology*, *12*(4), 360–374. https://doi.org/10.1002/qub2.57

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

Djidrovski, I., Pieters, R., Legler, J., & Teunis, M. (2025). 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*, 100395. https://doi.org/10.1016/j.comtox.2025.100395

Feng, Y., Wang, J., He, R., Zhou, L., & Li, Y. (2025). A retrieval-augmented knowledge mining method with deep thinking LLMs for biomedical research and clinical support. *GigaScience*, *14*, giaf109. https://doi.org/10.1093/gigascience/giaf109

Fujita, K. A., Ostaszewski, M., Matsuoka, Y., Ghosh, S., Glaab, E., Trefois, C., Crespo, I., Perumal, T. M., Jurkowski, W., Antony, P. M. A., Diederich, N., Buttini, M., Kodama,

A., Satagopam, V. P., Eifes, S., Del Sol, A., Schneider, R., Kitano, H., & Balling, R. (2014). Integrating pathways of Parkinson's disease in a molecular interaction map. *Molecular Neurobiology*, *49*(1), 88–102. https://doi.org/10.1007/s12035-013-8489-4

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

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

Hoch, M., Smita, S., Cesnulevicius, K., Lescheid, D., Schultz, M., Wolkenhauer, O., & Gupta, S. (2022). Network- and enrichment-based inference of phenotypes and targets from large-scale disease maps. *NPJ Systems Biology and Applications*, *8*(1), 13. https://doi.org/10.1038/s41540-022-00222-z

Hoch, M., Smita, S., Cesnulevicius, K., Schultz, M., Lescheid, D., Wolkenhauer, O., & Gupta, S. (2023). 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. https://doi.org/10.1186/s12950-023-00335-0

Hoksza, D., Gawron, P., Ostaszewski, M., Smula, E., & Schneider, R. (2019). MINERVA API and plugins: Opening molecular network analysis and visualization to the community. *Bioinformatics*, *35*(21), 4496–4498. https://doi.org/10.1093/bioinformatics/btz286

Hou, X., Zhao, Y., Wang, S., & Wang, H. (2025). *Model Context Protocol (MCP): Landscape, Security Threats, and Future Research Directions* (Version 3). arXiv. https://doi.org/10.48550/ARXIV.2503.23278

Jiang, H., Shi, S., Yao, Y., Jiang, C., & Li, Q. (2026). 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. https://doi.org/10.1109/TVCG.2025.3633887

Keating, S. M., Waltemath, D., König, M., Zhang, F., Dräger, A., Chaouiya, C., Bergmann, F. T., Finney, A., Gillespie, C. S., Helikar, T., Hoops, S., Malik-Sheriff, R. S., Moodie, S. L., Moraru, I. I., Myers, C. J., Naldi, A., Olivier, B. G., Sahle, S., Schaff, J. C., … SBML Level 3 Community members. (2020). SBML Level 3: An extensible format for the exchange and reuse of biological models. *Molecular Systems Biology*, *16*(8), e9110. https://doi.org/10.15252/msb.20199110

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

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

Ladeira, L., Verhoeven, A., Van Ertvelde, J., Jiang, J., Gamba, A., Sanz-Serrano, J., Vanhaecke, T., Heusinkveld, H. J., Jover, R., Vinken, M., Geris, L., & Staumont, B. (2025). Unlocking liver physiology: Comprehensive pathway maps for mechanistic understanding. *Frontiers in Toxicology*, *7*, 1619651. https://doi.org/10.3389/ftox.2025.1619651

Laurent, J. M., Janizek, J. D., Ruzo, M., Hinks, M. M., Hammerling, M. J., Narayanan, S., Ponnapati, M., White, A. D., & Rodriques, S. G. (2024). *LAB-Bench: Measuring Capabilities of Language Models for Biology Research* (Version 3). arXiv. https://doi.org/10.48550/ARXIV.2407.10362

Le Novère, N., Hucka, M., Mi, H., Moodie, S., Schreiber, F., Sorokin, A., Demir, E., Wegner, K., Aladjem, M. I., Wimalaratne, S. M., Bergman, F. T., Gauges, R., Ghazal, P., Kawaji, H., Li, L., Matsuoka, Y., Villéger, A., Boyd, S. E., Calzone, L., … Kitano, H.

(2009). The Systems Biology Graphical Notation. *Nature Biotechnology*, *27*(8), 735–741. https://doi.org/10.1038/nbt.1558

Lo Surdo, P., Iannuccelli, M., Karis, K., Meo, E., Omidi, P., Tosoni, M., Graziosi, S., Panni, S., Fiorillo, M. T., Licata, L., Sacco, F., Gyori, B. M., & Perfetto, L. (2026). SIGNOR 4.0: The 2025 update with focus on phosphorylation data. *Nucleic Acids Research*, *54*(D1), D682–D690. https://doi.org/10.1093/nar/gkaf1237

Lobentanzer, S., Aloy, P., Baumbach, J., Bohar, B., Carey, V. J., Charoentong, P., Danhauser, K., Doğan, T., Dreo, J., Dunham, I., Farr, E., Fernandez-Torras, A., Gyori, B. M., Hartung, M., Hoyt, C. T., Klein, C., Korcsmaros, T., Maier, A., Mann, M., … Saez-Rodriguez, J. (2023). Democratizing knowledge representation with BioCypher. *Nature Biotechnology*, *41*(8), 1056–1059. https://doi.org/10.1038/s41587-023-01848-y

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

Lopata, O., Acencio, M. L., Wang, X., Hemedan, A. A., Chao, M. J., Jelinsky, S. A., Tran, F., Rosenstiel, P., Li Yim, A. Y. F., Schneider, R., Satagopam, V., & Ostaszewski, M. (2025). 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. https://doi.org/10.1016/j.csbj.2025.09.008

Mazein, A., Ivanova, O., Balaur, I., Ostaszewski, M., Berzhitskaya, V., Serebriyskaya, T., Ligon, T., Hasenauer, J., De Meulder, B., Overall, R. W., Roy, L., Knowles, R. G., Wheelock, C. E., Dahlen, S.-E., Chung, K. F., Adcock, I. M., Roberts, G., Djukanovic, R., Pellet, J., … eTRIKS Consortium. (2021). AsthmaMap: An interactive knowledge repository for mechanisms of asthma. *The Journal of Allergy and Clinical*

*Immunology*, *147*(3), 853–856. https://doi.org/10.1016/j.jaci.2020.11.032

Mazein, A., Lopata, O., Reiche, K., Sewald, K., Alb, M., Sakellariou, C., Gogesch, P., Morgan, H., Neuhaus, V., Pham, N.-N., Sommer, C., Perkins, E., Fogal, B., Shoaib, M., Schneider, R., Satagopam, V., & Ostaszewski, M. (2025). 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. https://doi.org/10.3389/fimmu.2025.1601670

Mazein, A., Ostaszewski, M., Kuperstein, I., Watterson, S., Le Novère, N., Lefaudeux, D., De Meulder, B., Pellet, J., Balaur, I., Saqi, M., Nogueira, M. M., He, F., Parton, A., Lemonnier, N., Gawron, P., Gebel, S., Hainaut, P., Ollert, M., Dogrusoz, U., … Auffray, C. (2018). Systems medicine disease maps: Community-driven comprehensive representation of disease mechanisms. *NPJ Systems Biology and Applications*, *4*, 21. https://doi.org/10.1038/s41540-018-0059-y

Mazein, I., Rougny, A., Mazein, A., Henkel, R., Gütebier, L., Michaelis, L., Ostaszewski, M., Schneider, R., Satagopam, V., Jensen, L. J., Waltemath, D., Wodke, J. A. H., & Balaur, I. (2024). Graph databases in systems biology: A systematic review. *Briefings in Bioinformatics*, *25*(6), bbae561. https://doi.org/10.1093/bib/bbae561

Mitchener, L., Laurent, J. M., Andonian, A., Tenmann, B., Narayanan, S., Wellawatte, G. P., White, A., Sani, L., & Rodriques, S. G. (2025). *BixBench: A Comprehensive Benchmark for LLM-based Agents in Computational Biology* (Version 3). arXiv. https://doi.org/10.48550/ARXIV.2503.00096

Mohammadi, H., Almodaresi, F., Hogue, G. F. J., Wright, A., Orlic-Milacic, M., Li, N. T., Mawani, A., & Stein, L. (2025). React-to-Me: A Conversational Interface for Interactive Exploration of the Reactome Pathway Knowledgebase. *bioRxiv: The Preprint Server for Biology*, 2025.12.12.693752. https://doi.org/10.64898/2025.12.12.693752

Nagarajan, V., Shi, G., Horai, R., Yu, C.-R., Gopalakrishnan, J., Yadav, M., Liew, M. H., Gentilucci, C., & Caspi, R. R. (2025). IAN: An Intelligent System for Omics Data

Analysis and Discovery. *bioRxiv: The Preprint Server for Biology*, 2025.03.06.640921. https://doi.org/10.1101/2025.03.06.640921

Ostaszewski, M., Niarakis, A., Mazein, A., Kuperstein, I., Phair, R., Orta-Resendiz, A., Singh, V., Aghamiri, S. S., Acencio, M. L., Glaab, E., Ruepp, A., Fobo, G., Montrone, C., Brauner, B., Frishman, G., Monraz Gómez, L. C., Somers, J., Hoch, M., Kumar Gupta, S., … COVID-19 Disease Map Community. (2021). COVID19 Disease Map, a computational knowledge repository of virus-host interaction mechanisms. *Molecular Systems Biology*, *17*(10), e10387. https://doi.org/10.15252/msb.202110387

Rагueneau, E., Gong, C., Sinquin, P., Sevilla, C., Beavers, D., Grentner, A., Griss, J., Hogue, G. F. J., Li, N. T., Matthews, L., May, B., Milacic, M., Mohammadi, H., Petryszak, R., Rothfels, K., Shamovsky, V., Stephan, R., Tiwari, K., Weiser, J., … D'Eustachio, P. (2026). The Reactome Knowledgebase 2026. *Nucleic Acids Research*, *54*(D1), D673–D681. https://doi.org/10.1093/nar/gkaf1223

Serhan, C. N., Gupta, S. K., Perretti, M., Godson, C., Brennan, E., Li, Y., Soehnlein, O., Shimizu, T., Werz, O., Chiurchiù, V., Azzi, A., Dubourdeau, M., Gupta, S. S., Schopohl, P., Hoch, M., Gjorgevikj, D., Khan, F. M., Brauer, D., Tripathi, A., … Wolkenhauer, O. (2020). The Atlas of Inflammation Resolution (AIR). *Molecular Aspects of Medicine*, *74*, 100894. https://doi.org/10.1016/j.mam.2020.100894

Shakhbazau, A. (2025). Fine-tuning LLM hyperparameters to align semantic and physiological contexts of aging-related pathways. *Molecular Diversity*. https://doi.org/10.1007/s11030-025-11226-2

Sokolova, K., Kosenkov, D., Nallamotu, K., Vedula, S., Sokolov, D., Sapiro, G., & Troyanskaya, O. G. (2025). An Evidence-Grounded Research Assistant for Functional Genomics and Drug Target Assessment. *bioRxiv: The Preprint Server for Biology*, 2025.12.30.697073. https://doi.org/10.64898/2025.12.30.697073

Stachura, D., Konieczna, J., & Nowak, A. (2025). *Are Smaller Open-Weight LLMs Closing the Gap to Proprietary Models for Biomedical Question Answering?* (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2509.18843

Staumont, B., Ladeira, L., Gamba, A., Heusinkveld, H. J., Piersma, A., Fritsche, E., Masereeuw, R., Vanhaecke, T., Teunis, M., Luechtefeld, T. H., Hartung, T., Jover, R., Vinken, M., & Geris, L. (2025). Mapping physiology: A systems biology approach for the development of alternative methods in toxicology. *ALTEX*, *42*(2), 301–307. https://doi.org/10.14573/altex.2412241

Vinken, M., Benfenati, E., Busquet, F., Castell, J., Clevert, D.-A., De Kok, T. M., Dirven, H., Fritsche, E., Geris, L., Gozalbes, R., Hartung, T., Jennen, D., Jover, R., Kandarova, H., Kramer, N., Krul, C., Luechtefeld, T., Masereeuw, R., Roggen, E., … Piersma, A. H. (2021). Safer chemicals using less animals: Kick-off of the European ONTOX project. *Toxicology*, *458*, 152846. https://doi.org/10.1016/j.tox.2021.152846

Wang, E. J., Oğuztüzün, Ç., Xu, R., & Gao, Z. (2026). Comparative evaluation of large language models in retrieving known and predicting novel drug combinations. *Journal of Alzheimer's Disease: JAD*, 13872877261415635. https://doi.org/10.1177/13872877261415635

Wang, J. J., & Wang, V. X. (2025). *Assessing Consistency and Reproducibility in the Outputs of Large Language Models: Evidence Across Diverse Finance and Accounting Tasks* (Version 4). arXiv. https://doi.org/10.48550/ARXIV.2503.16974

Wu, G., Matthews, L., Boyer, N., Milacic, M., Beavers, D., Li, N. T., May, B., Rothfels, K., Shamovsky, V., Stephan, R., Gillespie, M., Hermjakob, H., D'Eustachio, P., & Stein, L. (2025). Application of Large Language Models for Annotating Genes into Reactome Pathways. *bioRxiv: The Preprint Server for Biology*, 2025.12.20.695723. https://doi.org/10.64898/2025.12.20.695723

Zerrouk, N., Aghakhani, S., Singh, V., Augé, F., & Niarakis, A. (2022). A Mechanistic Cellular Atlas of the Rheumatic Joint. *Frontiers in Systems Biology*, *2*, 925791. https://doi.org/10.3389/fsysb.2022.925791

Zerrouk, N., Alcraft, R., Hall, B. A., Augé, F., & Niarakis, A. (2024). Large-scale computational modelling of the M1 and M2 synovial macrophages in rheumatoid arthritis. *NPJ Systems Biology and Applications*, *10*(1), 10.

https://doi.org/10.1038/s41540-024-00337-5