



CHALMERS
UNIVERSITY OF TECHNOLOGY

LLM-retrieval based scientific knowledge grounding

Downloaded from: <https://research.chalmers.se>, 2026-03-17 01:19 UTC

Citation for the original published paper (version of record):

Reder, G., Collins, C., Rehim, A. et al (2025). LLM-retrieval based scientific knowledge grounding. CEUR Workshop Proceedings, 3977

N.B. When citing this work, cite the original published paper.

LLM-retrieval based scientific knowledge grounding

Gabriel K. Reder^{1,*}, Carl Collins², Abbi Abdel Rehim¹, Larisa Soldatova² and Ross D. King^{1,3,4}

¹University of Cambridge, Cambridge, CB3 0AS, United Kingdom

²Goldsmiths, University of London, London SE14 6AD, United Kingdom

³Chalmers University of Technology, Gothenburg 412 96, Sweden

⁴The Alan Turing Institute, London NW1 2DB, United Kingdom

Abstract

The automated high-throughput laboratory offers unprecedented potential for scientific discovery, yet effectively linking studies to existing knowledge remains a significant challenge. As the general body of scientific knowledge grows, so too does the burden of contextualizing a new experiment. While ontologies and databases serve as structured common repositories, their rigid schemas are often incompatible with the unstructured or semi-structured formats of most laboratories. In this study we investigate the integration of large language models (LLMs) with ontology-based vector databases to anchor semi-structured scientific experiments into knowledge bases via automated retrieval. Our approach extracts scientific entities from unstructured experimental texts, and grounds them to relevant ontology terms. We automate knowledge grounding, which enhances the integration of unstructured experimental data into established formal scientific languages. We have tested our method on a diverse selection of experimental yeast biology papers focused on *Saccharomyces cerevisiae*, a foundational model system that has driven major discoveries in molecular and cellular biology, and observed strong pipeline performance. We argue that such a knowledge grounding approach is a critical component for the new wave of efficient artificial intelligence (AI) driven automated laboratories that integrate LLMs with high-throughput experimentation and data-driven discovery.

Keywords

Artificial Intelligence, Large Language Models, Information Extraction for RKGs/SKGs, Ontologies, Knowledge Engineering, *Saccharomyces cerevisiae*

1. Introduction

The rise of automation in the laboratory presents a unique opportunity to generate scientific findings at an unprecedented rate; however, a key issue remains overlooked. Empirical science relies on the integration of experiments, data, and discoveries with existing knowledge. Linking to background knowledge is a major challenge for the experimentalist, one that grows larger as the scientific corpus expands. As laboratory automation accelerates this expansion, the linking problem grows, risking the creation of vast quantities of isolated observations stranded on knowledge islands.

Take, for example, the most fundamental kernel of the experimental process: the hypothesis. In the classical scientific method, an experiment is conducted to test a specific hypothesis, and knowledge is generated by its subsequent confirmation or rejection [1]. For knowledge to effectively spread within the scientific community, hypotheses must be clearly and universally understood. Achieving this requires the use of consistent vocabulary and terminology shared by other scientists. This consistency enables the scientific community to fully comprehend, replicate, and validate experimental findings, strengthening the reliability of the conclusions. If knowledge linking is successful, a single experiment may contribute to the general body of scientific knowledge. This process becomes even more important when laboratory automation is introduced. Unlike the human scientist, a laboratory robot has no understanding of the material it handles in the service of a hypothesis. As such, the hypothesis, and all

2nd International Workshop on Natural Scientific Language Processing and Research Knowledge Graphs (NSLP 2025), co-located with ESWC 2025, June 01–02, 2025, Portorož, Slovenia

*Corresponding author.

✉ gr513@cam.ac.uk (G. K. Reder); c.collins@gold.ac.uk (C. Collins); aar52@cam.ac.uk (A. A. Rehim); l.soldatova@gold.ac.uk (L. Soldatova); rk663@cam.ac.uk (R. D. King)

ORCID 0000-0001-8918-0789 (G. K. Reder); 0000-0001-6489-3029 (L. Soldatova); 0000-0001-7208-4387 (R. D. King)



© 2025 Copyright © 2025 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

sample metadata, must be sufficiently descriptive and exact to understand the results from an automated laboratory.

Typically, hypotheses are written in natural language with constraints around key scientific concepts. These constraints are implicit and can lead to confusion. For example, a scientist may contextualize their hypothesis by using the word “yeast” when in fact they mean “*Saccharomyces cerevisiae*”. This choice encodes the scope of the experiment’s findings and its implications for knowledge acquisition. These choices, shaped by experiential understanding of terminology, can vary considerably among scientists, types of publications, and research venues. Robust and systematic nomenclatures for scientific concepts, materials, and actions are needed to avoid confusion. This codification of shared nomenclature takes the form of knowledge bases.

Knowledge bases, specifically ontologies and databases, have proven to be highly effective as terminal repositories of community knowledge and observations, respectively. Public databases have been quintessential in advancing the life sciences since their adoption across the research community [2, 3, 4]. The Protein Data Bank (PDB) [5] is a prominent example of a successful knowledge repository. By establishing a communal repository for standardized data, the PDB and its associated standards have been invaluable to advancements in protein biology, from benchtop research to clinical applications, notably by enabling the development of computational models for protein design [6]. Databases serve as centralized repositories for findings from various laboratories, enforcing a standardized structure across all entries. In this sense, scientific database structures are syntactical templates for how findings should be communicated. Ontologies take this codification further by defining controlled vocabulary terms (entities) and the formal relations between them. Ontologies and knowledge graphs provide highly robust shared languages for information sharing [7, 8]. The knowledge structures (schemas) of databases and ontologies are key to their communal utility. Such schemas embody a robust, widely accepted community standard that facilitates effective communication of knowledge, enabling others to understand and replicate findings.

Formalization requires effort, and effective knowledge base utilization requires strict adherence to their structures. Yet scientists mostly work with unstructured or unlinked formats such as text documents, spreadsheets, notebooks, and raw data. Conforming these to knowledge base structures is a laborious process, especially if performed manually. As such, experimenters usually communicate results in a semi-structured format, the scientific paper, leaving some burden of linking to the reader. The significant increase in experiment execution and data generation rates from automated laboratories makes manual knowledge linking infeasible necessitating computational approaches.

Generative artificial intelligence (AI) models, specifically large language models (LLMs), represent an exciting opportunity to speed the adoption and implementation of automation for discovery science, and their capabilities hint at promising abilities to aid in the knowledge linking process. They excel at working with unstructured inputs and outputs; however, they do not natively interface well with predefined knowledge schemas [8]. They have, however, proven adept at interfacing with structured databases and application programming interfaces (APIs) when these structures are included in LLM prompts [9]. Further, LLMs have been shown to excel in question answering applications when the scope of answer is explicitly specified in input prompts. An exciting avenue of LLM usage combining database access with prompt scoping is retrieval augmented generation (RAG) [10]. In this formulation, the results of a database query-potentially initiated by an upstream LLM-are fed into a downstream LLM prompt. Since the LLM is prompted to answer from the query results rather than its internal knowledge, such an approach grounds answers to specific databases and limits hallucination.

In this work, we investigate the use of a combination of LLMs, database retrieval, and ontologies to ground unstructured scientific inputs to formal knowledge bases. Specifically, we follow a retrieval augmented generation approach utilizing LLMs coupled to vector database stores of ontologies to ground terms in hypotheses to ontology identifiers. LLMs were used to extract hypotheses from a pool of carefully selected yeast research papers and extract scientific entities before grounding them to ontology terms. We tested our pipeline on a diverse selection of papers spanning a wide range of research domains within the model organism *Saccharomyces cerevisiae* and found that this approach effectively automates the grounding of knowledge from unstructured scientific experimental data.

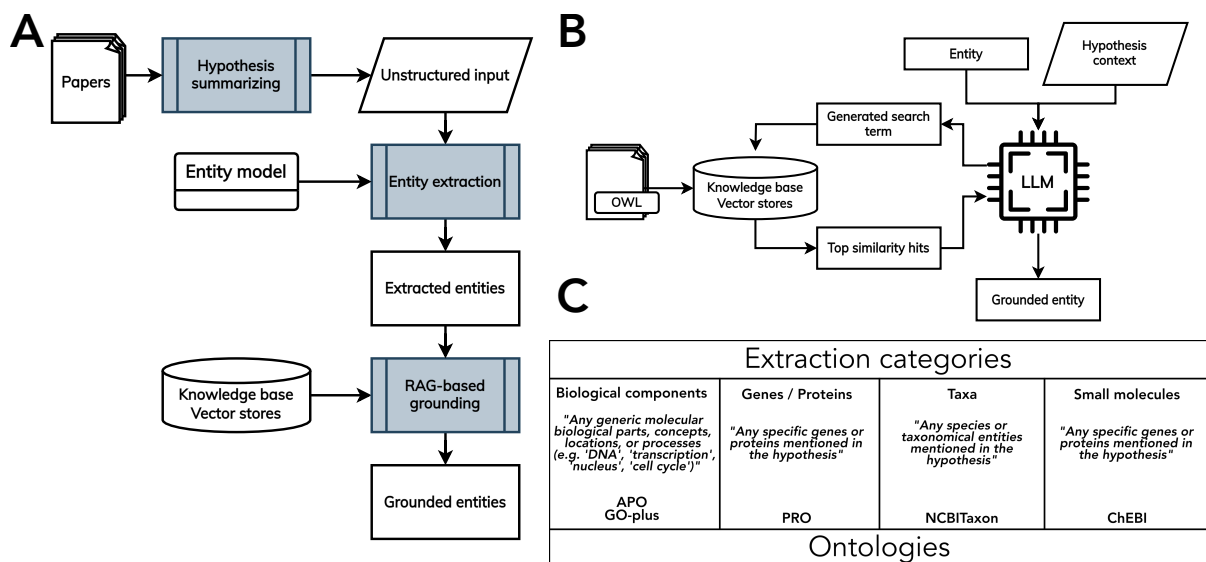


Figure 1: Pipeline, entity grounding scheme, and extraction knowledge model. Further details on pipeline implementation, technical environment, and ontologies used are available in the Methods section. (A) The tested pipeline flow. Computational modules involving a large language model (LLM) are shown in shaded gray. Papers were ingested to produce hypothesis summaries in the Summarization module. Entities were extracted from the unstructured hypotheses using a user-defined knowledge extraction schema in the Extraction module. Extracted entities were grounded using the Grounding module using LLM interaction with vector stores compiled from knowledge bases. (B) The grounding module involves a retrieval augmented generation (RAG) configuration to map extracted entities to ontology terms. Given the input extracted entity and the original hypothesis, the LLM creates a query search term and is presented with the closest ontology matches from the vector store. The LLM then decides on the grounding given the hypothesis. (C) The knowledge extraction model used in tests. Four extraction categories were tested, each linked to one or more ontologies. Extraction category model prompts passed to the LLM for extraction are shown in italics.

2. Methods

2.1. Pipeline overview

The automated pipeline developed and tested consists of three modules: (1) hypothesis extraction from the full text of a paper (summarization), (2) stratified entity extraction from the summarized hypothesis (extraction), and (3) grounding of entities to ontological knowledge bases (grounding). Summarization consists of presentation of the entire paper text to a LLM model which is prompted to consider the text and produce the main hypothesis tested by the paper. The extraction module takes an input text and extracts single entities/concepts of interest according to user-defined schema. The grounding module takes these entities and links them to the best found term in ontologies specified by the user to fall under a given schema category. Crucially, the entity grounding utilizes the LLMs fuzzy reasoning capabilities to make a decision based on scientific context.

2.2. Manual selection of yeast biology research papers and extraction of hypotheses

A total of 15 recent research publications were manually selected to eclectically cover the field of yeast (*Saccharomyces cerevisiae*) biology in an unbiased manner. To provide a framework for research paper selection, we used the Gene Ontology (GO) [11] as a guide for selecting papers representative of a broad range of scientific concepts. Papers were manually selected to partially or fully embody concepts represented by direct children nodes of the root 'Biological Process' GO term. These ranged from high level cellular processes such as cell division, reproduction, and homeostasis, to somewhat more restricted phenomena including metabolism, gene regulation, and localized processes such as chromatin reorganization, cellular component biogenesis and signaling. We excluded several GO terms

deemed to have no relevance to *S. cerevisiae*. Papers were also carefully selected to represent traditional biological research domains including genomics, molecular biology and biochemistry, cellular biology, systems biology, evolutionary biology and biotechnology. The 15 papers and their respective manually annotated GO terms are shown in Table 1.

2.3. Manual hypothesis summarization

Three expert reviewers were assembled to read the papers and decide on consensus human extracted hypotheses for comparison to the automated pipeline results. The 15 papers were divided into groups of 5 and each reviewer was assigned two of these groups for a total of 10 papers each. Each reviewer read the assigned papers and extracted hypotheses for each of them independently of the other reviewers. Reviewers were prompted to extract hypotheses that are clear, logically-sound, actionable, and reflective of the paper's aims. The three reviewers then compared individual results to produce a consensus human hypothesis for each of the papers. The consensus hypotheses are shown in Table 2.

2.4. Automated hypothesis summarization

Papers were fed to the pipeline as PDF files and tokenized into LLM-compatible text inputs. Together with the tokenized paper, the LLM was prompted to extract only the main hypothesis tested by the paper. To provide a degree of consistency across extraction, the LLM was prompted to phrase the hypothesis as a single sentence. Paper contents were tokenized and concatenated into a single annotated string before injection into the LLM summarization prompt. Original page breaks and counts in the PDF were maintained as annotation strings to give the LLM a sense of the paper's physical structure. The summary module flow was implemented using LangChain [12]. PDF files were processed using the PyMuPDF library [13]. Output model specification and verification were implemented using Pydantic [14]. The resulting extracted hypotheses are shown in Table 3.

2.5. Entity extraction and ontology selection

Input hypotheses were fed into the extraction module together with a user-defined knowledge extraction schema of desired entity categories and annotated descriptions of each category. The module flow was implemented using LangChain, and schemas were defined using Pydantic. Field descriptions in the extraction model's Pydantic class are fed to the LLM during the entity extraction task. Together with the class-level docstring, these define the entity extraction task for the LLM. For example, the class-level docstring "The entities extracted from a scientific hypothesis. The entities are divided into different categories, these field lists MUST be mutually exclusive. An entity cannot be in more than one list." together with the category-level description "Any specific genes or proteins mentioned in the hypothesis" were used to extract gene/protein entities. In principle, the module accepts any user-defined Pydantic schema. The example schema and linked ontologies used in our testing is shown in Figure 1 including the prompt annotations used to define each category.

2.6. Retrieval Augmented Generation (RAG) grounding of hypotheses

Extracted entities from the hypotheses were grounded to ontologies in database vector store formats using similarity search to term names from intermediate LLM-generated search terms. OWL [15] files were downloaded for our chosen ontologies of interest and compiled to Faiss vector store databases [16]. The following ontologies and versions were chosen for each schema category:

Biological components:

- Gene ontology (GO-plus, version 2024-09-08) [11]
- Ascomycete phenotype ontology (APO, version 2024-09-18) [17]

Genes/proteins:

- Protein ontology (PRO, version 2024-08-08) [18]

Taxa:

- NCBI Taxon (NCBITaxon, version 2024-07-01) [19]

Small molecules:

- Chemical Entities of Biological Interest (ChEBI, version 2024-07-27) [20]

The langchain-rdf [21] library was used to parse OWL files into a Faiss-compatible format. Input user YAML files were used to link extraction schema categories with specific vector databases. We note that a single vector database associated with an extraction category may contain one or more stored ontologies according to the user's preferences. In our case, the 'Biological components' category was linked to a vector store database containing terms from the GO and APO ontologies. For each entity, the following procedure was performed. The LLM was first prompted to look at the hypothesis and the individual entity to decide on appropriate search terms. This term could be the same as the extracted entity or a modified entity term given the context of the hypothesis. Both cases were observed as output in our pipeline runs. For example, during grounding of paper 1 hypothesis terms, the LLM decided to use the search term "protein synthesis rate" when grounding the term "protein production rate". Finally, the generated search term was used as input for a similarity search against the vector database linked to the entity's schema category. Searches were based on ontology term names and the top 5 results were presented to the LLM in a subsequent prompt. Given the top 5 search hits, the input hypothesis, and the original entity, the LLM was prompted to select the best search term. The entire body of extraction and grounding results is shown in the supplementary information.

2.7. Evaluation

The panel of three scientists compared the LLM-extracted hypothesis to the consensus human hypothesis for each of the 15 papers. Each reviewer individually assigned an evaluation score between 1 and 3 for the LLM extraction performance based on the following criteria:

3 = The LLM hypothesis accurately summarizes the paper's aims and agrees with the human consensus hypothesis

2 = The LLM hypothesis differs slightly from the human consensus hypothesis but captures some or all of the paper's aims

1 = The LLM hypothesis does not correctly summarize the paper's aims at all

For each paper, the individual scores were averaged across the three reviewers' scores to produce an average hypothesis evaluation score.

The three scientists also produced an individual entity score and a grounding score for each entity extracted from each of the hypotheses. The entity score evaluated the quality of the entity extraction, specifically how reasonable the extracted entity is. Reasonable entities were sought to be self-contained, understandable, coherent scientific concepts/units that should map to ontology terms. The grading criteria followed a 1-3 scale where:

3 = the entity is reasonable in the right category (e.g. is extracted as a "Taxa" when it should be).

2 = reasonable entity in the wrong category or semi-reasonable entity in the right category.

1 = nonsensical/overly complex entity.

The grounding score was designed to evaluate the effectiveness of grounding based on the extracted entity. In other words, in the context of the hypothesis, does the grounding do a good job of finding the appropriate term in the ontology? The following 1-3 scale was used:

- 3 = The grounding term fits the entity very well in its context.
- 2 = The grounding term is related to the entity but not an exact fit.
- 1 = The grounding term is unrelated to the entity.

Evaluation scores and summaries are shown in Figure 2. The individual reviewer scores can be found in the supplementary information.

2.8. Environment, models, and code availability

Gpt-4o (version 2024-08-06) was used as the LLM model for all pipeline tasks. The pipeline can currently use any OpenAI model or model available through Ollama [22]. All text embedding was performed with the FlagEmbedding (bge-small-en-v1.5) embedding model [23].

Vector store creation was performed on an Amazon Web Services (AWS) EC2 r5.4xlarge instance with 16 vCPUs and 128 GB memory. Currently, the pipeline requires substantial memory to create vector stores involving large ontology files (hundreds of MB or more). Once created, vector stores can be reused without rebuilding. The prebuilt vector stores used in this work's tests are available in this work's Zenodo repository <https://doi.org/10.5281/zenodo.14014577>.

The complete pipeline is available in the ragnosis GitHub repository <https://github.com/gkreder/ragnosis>.

3. Results and Discussion

In this work, we aimed to assess the capability of current LLM-retrieval systems to automatically link unstructured scientific content to structured knowledge bases. We tested our pipeline on 15 representative research papers in *Saccharomyces cerevisiae* biology, evaluating its ability to extract hypotheses, identify key entities, and accurately ground these entities within selected ontologies. Our findings demonstrate that this approach shows immense promise, closely matching human performance on critical tasks.

3.1. Results

The automated hypothesis extraction module generated concise and accurate single-sentence hypotheses for each paper, aligning closely with the human consensus hypotheses (Table 2 and Table 3). Human evaluation of the results on a basic scale of 1-3 are shown in Figure 2A. Further details on the automated hypothesis extraction, human consensus extraction, and evaluation can be found in the Methods section.

The entity extraction module was run on the hypotheses extracted using the automated pipeline. For this test, entities were extracted into four schema categories: (1) biological components, (2) genes/proteins, (3) taxa, and (4) small molecules. A total of 123 entities were extracted across all papers. On a per-category basis the entity counts were the following: biological components: 73, genes/proteins: 27, taxa: 14, small molecules: 9. Extracted entities were grounded using the pipeline with the compiled ontology vector stores. The prompts used to guide extraction for each category, along with the ontologies associated with them, can be found in Figure 1C. Human evaluation scores, ranging from 1 to 3 on basic scales, were assigned by a three-person review panel for both entity extraction and grounding. These average scores across reviewers are shown in Figure 2B, while per-reviewer distributions for each score are presented in Figure 2C. Full extraction and grounding results can be found in the supplementary information.

3.2. Discussion

The hypothesis extraction demonstrated strong consistency between the automated outputs and the consensus derived from human evaluators, mirroring broader findings that LLMs excel in summarization

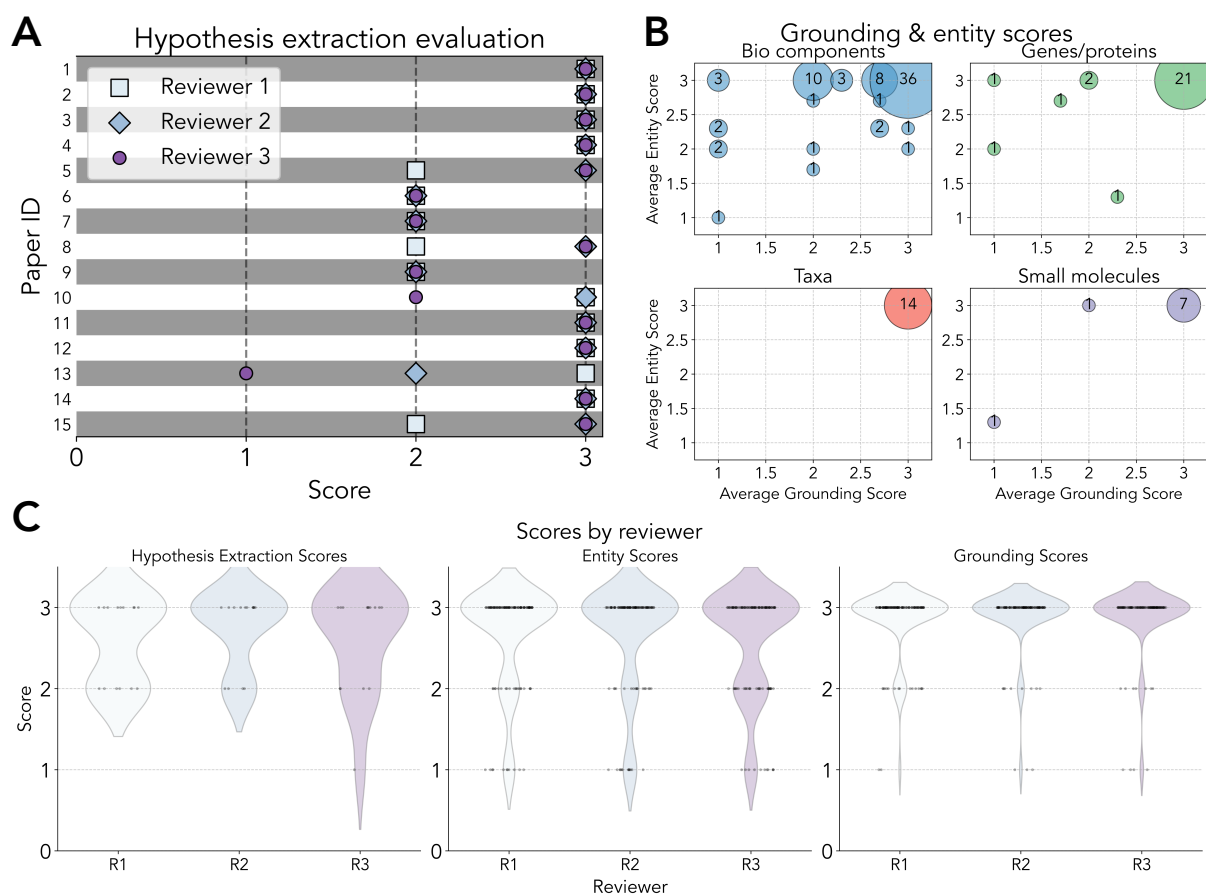


Figure 2: Evaluations of summarization, extraction, and grounding. Human evaluation scores from the automated pipeline results. All scores were based on separate 1-3 scales of increasing quality and assigned separately by three human reviewers. Evaluation criteria and the extraction categories are described in the Methods section. (A) Hypothesis summarization scores per reviewer. Scores are reported individually. (B) Entity and grounding scores. Scores are reported as the mean per-entity score across the three reviewers. Bubble size is scaled and labeled according to the number of entities occupying the given coordinate at the bubble’s center. Entities are visualized by extraction category. (C) Score distributions per reviewer by score type.

tasks [24]. Indeed, this task required no retrieval or grounding from the LLMs. Rather, it served as a means of producing condensed unstructured scientific input to the downstream grounding task. Nevertheless, the performance of the automated hypothesis extraction was impressive. Our primary goal was to test entity extraction and knowledge grounding on unstructured input, and hypothesis extraction effectively served this purpose.

Analysis of the hypothesis extraction process highlighted several key issues. The most significant is defining what constitutes a hypothesis. Our tests focused on extracting hypotheses from papers describing completed scientific studies. Both in human extraction and LLM prompting, efforts were made to avoid incorporating the study’s results into the hypotheses. However, this was somewhat unavoidable, as a paper’s narrative is naturally influenced by the study’s findings. Papers describe studies with a priori aims that range from entirely speculative to purely confirmational, and the mechanistic granularity of hypotheses tested can vary dramatically. Our human and automated results reflect this heterogeneity. At times, the LLM incorporated more of the study’s results into its hypotheses (as seen in paper 7), while in other cases, the human evaluators included more mechanistic detail (as in paper 4). A related issue concerns the question of novelty. Hypotheses inherently rely on existing knowledge, with their novelty stemming from the ability to shed new light on that knowledge, rather than simply rephrasing it. This subtlety may be more challenging for LLMs to capture. For example, the LLM-extracted hypothesis from paper 9 appears to suggest that the novelty of the work centers on the characterization of the Spt6 protein, when in reality, the authors aimed to shed light on the

Cdc73 subunit of Paf1. Both human and LLM hypotheses describe the proposed relationship between the proteins, but the LLM hypothesis frames Spt6 as the focus of discovery. In brief, modern scientific studies do not uniformly adhere to the classical scientific method. They include engineering applications and untargeted screens in addition to classical experiments. Such heterogeneity must be taken into consideration in follow up automation development.

We found entity performance to be effective, regardless of the type of hypothesis used as input. Terms tended to fit the categories and were generally well-scoped. Notably, the extraction process is flexible, allowing the user to specify their desired input schema categories to guide identification of relevant entities. This user-defined input schema for entity extraction has been successfully demonstrated in previous work, such as in [25], and our results further validate its effectiveness and viability. A key consideration is that the flexibility of knowledge schema choice leads to variations in performance depending on the input knowledge model. This is not a question of computational extraction performance, but rather schema design.

As shown in Figure 2, entity extraction performance varies significantly across categories. The ‘Biological components’ category exhibits the weakest performance, primarily due to its inherent vagueness compared to more specific categories like ‘Taxa’, particularly in the context of our paper set focused solely on yeast biology. This touches on the larger question of knowledge modeling in the life sciences and beyond. The design and optimal use of descriptive and appropriate knowledge schema for scientific domains is a fruitful field in itself [26, 27, 28]. This work does not aim to rigorously compare schemas; rather, we emphasize that our retrieval approach facilitates the linking of unstructured scientific texts to any knowledge schema. The effectiveness of this linking is inherently constrained by the quality of the chosen schema. Entity extraction also suffers from the vague definition of an entity. A few times, entire phrases were extracted as an entity, for example “spectrum of beneficial mutations” from paper 3 or “cell wall protein mannosylation” from paper 12. Such compound entities could likely be further decomposed into more modular knowledge units. Conversely, the pipeline extracted overly specific entities at times, such as “serine 15” from paper 10. In some instances, the grounding scheme successfully overcame these vague or overly specific extractions, as described below. A straightforward approach would be to base the knowledge extraction schema on the knowledge base itself. Ontologies effectively capture the hierarchical structure of encoded knowledge, and basing extracted categories on ontology levels would undoubtedly enhance grounding performance. We believe this approach will be especially well-suited for grounding applications on static knowledge bases. Often, experiments will involve unexplored areas of knowledge and existing knowledge bases will provide an incomplete structure of the scientist’s domain. Such cases will require user-defined knowledge schema and perhaps iterated cycles of knowledge base modification from experimental results. We envision automated closed-loop cycles of experimental grounding and knowledge base improvement involving intermediate test schema.

The most significant neurosymbolic performance enhancement in our pipeline comes from retrieval augmented generation (RAG) grounding of extracted terms, yielding highly promising results. Previous approaches have utilized LLMs for schema-aligned entity extraction, however they typically rely on traditional deterministic methods for grounding within ontologies or knowledge bases. In contrast, our method and concurrent new approaches [29] integrate LLMs directly in the grounding process, marking a shift toward more adaptable, context-sensitive knowledge grounding. This is achieved through RAG, presenting the LLM with a choice of possible grounding terms and leaving the final choice to the LLM given the input context. The grounding evaluation scores, shown in Figure 2, largely reflect that the automated LLM-retrieval approach performs well on this task. The LLM’s fuzzy logic is crucial in selecting among the database hits. A clear example of this occurred across multiple papers, where the pipeline accurately grounded protein entities to their correct species designation within the PRO ontology. For instance, when grounding the entity “CLN3”, it correctly selected “S000000038 CLN3 (yeast)” instead of alternative terms such like “protein CLN3 (mouse)”, “G1/S-specific cyclin CLN3 (Candida albicans SC5314)”, or “protein CLN3 (human)”, all of which are also present in the ontology. Such logic would be challenging to implement deterministically across different use cases, but it leverages the strengths of LLMs, where they excel in handling complex,

context-dependent reasoning. As noted earlier, the grounding scheme effectively compensated for poorly extracted entities by utilizing the hypothesis context. For instance, it grounded “cell wall protein mannosylation” from paper 12 to “GO_0035268 protein mannosylation,” and “serine 15” from paper 10 to “MOD_00696 phosphorylated residue.” Full details of the extraction and grounding results are provided in the supplementary information.

Further development will be undertaken to improve the performance of the extraction and grounding modules. One notable improvement will come from basing candidate ontology term presentation to the LLM on semantic search to more than the term name in the ontology. Knowledge base terms often contain annotations, synonyms, and relations to other terms. Basing the vector store search on these rather than purely term names will likely further improve the quality of the candidates for the LLM to choose from. We also note that the choice of ontologies to search was somewhat arbitrary and based on our test entity extraction schema. Some choices were straightforward, such as grounding entities in the "Taxa" category to the NCBITaxon ontology. Choosing GO-Plus and APO for the “biological components” category likely omitted other ontologies with potentially relevant terms. More comprehensive study of the design of these elements warrants further investigation. Such development will build on prior and ongoing efforts to robustly systematize and encode scientific knowledge, hypotheses, protocols, data, and findings [30, 31, 32, 33]. Further expansion beyond hypotheses must also be explored. An especially promising direction will be in grounding raw and processed data from experiments to knowledge base terms. For example, sample identifiers in experimental data may be automatically contextualized to other experiments through iterated rounds of context presentation and grounding. Such a scheme would increase the insight generated from a single experiment and allow for richer meta-analysis.

Based on our human evaluations, we found that the extraction and grounding pipeline performed well for knowledge grounding and represents a promising direction for future research. Reflective of the scientific process itself, the evaluators varied in their assessments but were in general agreement (Figure 2). Our tests were conducted on a small, representative set of papers spanning a broad range of topics within yeast biology. However, there is no reason to believe that our approach would not be suitable for larger-scale work that encompasses a broader range of scientific knowledge, inputs and outputs, and applications - given appropriate resources. Beyond its use in mining public data, we envision such a pipeline as a crucial component of a fully automated self-driving laboratory where AI agents generate hypotheses, conduct experiments, and analyze data in iterative cycles of closed-loop discovery. Approaches that integrate laboratory robotics with symbolic systems and knowledge bases have been and continue to be developed [34, 35, 36]. However, experimental knowledge is often inherently unstructured, making such neurosymbolic systems crucial for the future of automated laboratories, as they leverage the strengths of both logical and subsymbolic generative approaches. By leveraging these advanced systems, we are approaching a new phase in scientific research, where automation and intelligent knowledge grounding can open valuable opportunities for discovery and innovation.

Acknowledgments

This work was supported by the Engineering and Physical Sciences Research Council [grant numbers EP/X032418/1, EP/X033740/1] and the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Alice Wallenberg Foundation.

Declaration on Generative AI

During the preparation of this work, the author(s) used GPT-4o in order to: Grammar and spelling check, Paraphrase and reword. After using these tool(s)/service(s), the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication’s content.

References

- [1] K. Popper, *The Logic of Scientific Discovery*, Routledge Classics, 2nd ed ed., Taylor and Francis, Abingdon, Oxon, 2005.
- [2] H. J. Imker, 25 Years of Molecular Biology Databases: A Study of Proliferation, Impact, and Maintenance, *Frontiers in Research Metrics and Analytics* 3 (2018). doi:10.3389/frma.2018.00018.
- [3] D. J. Rigden, X. M. Fernández, The 2024 Nucleic Acids Research database issue and the online molecular biology database collection, *Nucleic Acids Research* 52 (2024) D1–D9. doi:10.1093/nar/gkad1173.
- [4] A. D. Baxevanis, A. Bateman, The Importance of Biological Databases in Biological Discovery, *Current Protocols in Bioinformatics* 50 (2015). doi:10.1002/0471250953.bi0101s50.
- [5] S. K. Burley, H. M. Berman, J. M. Duarte, Z. Feng, J. W. Flatt, B. P. Hudson, R. Lowe, E. Peisach, D. W. Piehl, Y. Rose, A. Sali, M. Sekharan, C. Shao, B. Vallat, M. Voigt, J. D. Westbrook, J. Y. Young, C. Zardecki, Protein Data Bank: A Comprehensive Review of 3D Structure Holdings and Worldwide Utilization by Researchers, Educators, and Students, *Biomolecules* 12 (2022) 1425. doi:10.3390/biom12101425.
- [6] S. K. Burley, C. Bhikadiya, C. Bi, S. Bittrich, H. Chao, L. Chen, P. A. Craig, G. V. Crichlow, K. Dalenberg, J. M. Duarte, S. Dutta, M. Fayazi, Z. Feng, J. W. Flatt, S. Ganesan, S. Ghosh, D. S. Goodsell, R. K. Green, V. Guranovic, J. Henry, B. P. Hudson, I. Khokhriakov, C. L. Lawson, Y. Liang, R. Lowe, E. Peisach, I. Persikova, D. W. Piehl, Y. Rose, A. Sali, J. Segura, M. Sekharan, C. Shao, B. Vallat, M. Voigt, B. Webb, J. D. Westbrook, S. Whetstone, J. Y. Young, A. Zalevsky, C. Zardecki, RCSB Protein Data Bank (RCSB.org): Delivery of experimentally-determined PDB structures alongside one million computed structure models of proteins from artificial intelligence/machine learning, *Nucleic Acids Research* 51 (2023) D488–D508. doi:10.1093/nar/gkac1077.
- [7] R. Hoehndorf, P. N. Schofield, G. V. Gkoutos, The role of ontologies in biological and biomedical research: A functional perspective, *Briefings in Bioinformatics* 16 (2015) 1069–1080. doi:10.1093/bib/bbv011.
- [8] R. Stevens, Ontology-based knowledge representation for bioinformatics, *Briefings in Bioinformatics* 1 (2000) 398–414. doi:10.1093/bib/1.4.398.
- [9] T. Schick, J. D.-Y. R. Dessì, R. Raileanu, M. Lomeli, E. Hambro, L. Zettlemoyer, N. Cancedda, T. Scialom, Toolformer: Language Models Can Teach Themselves to Use Tools (????).
- [10] S. Wu, Y. Xiong, Y. Cui, H. Wu, C. Chen, Y. Yuan, L. Huang, X. Liu, T.-W. Kuo, N. Guan, C. J. Xue, Retrieval-Augmented Generation for Natural Language Processing: A Survey, 2024. arXiv:2407.13193.
- [11] The Gene Ontology Consortium, The Gene Ontology Resource: 20 years and still GOing strong, *Nucleic Acids Research* 47 (2019) D330–D338. doi:10.1093/nar/gky1055.
- [12] H. Chase, LangChain, 2022.
- [13] Pymupdf/PyMuPDF, PyMuPDF, 2024.
- [14] S. Colvin, E. Jolibois, H. Ramezani, A. Garcia Badaracco, T. Dorsey, D. Montague, S. Matveenko, M. Trylesinski, S. Runkle, D. Hewitt, A. Hall, Pydantic, 2024.
- [15] G. Antoniou, F. van Harmelen, Web Ontology Language: OWL, in: S. Staab, R. Studer (Eds.), *Handbook on Ontologies*, Springer, Berlin, Heidelberg, 2009, pp. 91–110. doi:10.1007/978-3-540-92673-3_4.
- [16] M. Douze, A. Guzhva, C. Deng, J. Johnson, G. Szilvasy, P.-E. Mazaré, M. Lomeli, L. Hosseini, H. Jégou, The Faiss library, 2024. doi:10.48550/arXiv.2401.08281. arXiv:2401.08281.
- [17] S. R. Engel, R. Balakrishnan, G. Binkley, K. R. Christie, M. C. Costanzo, S. S. Dwight, D. G. Fisk, J. E. Hirschman, B. C. Hitz, E. L. Hong, C. J. Krieger, M. S. Livstone, S. R. Miyasato, R. Nash, R. Oughtred, J. Park, M. S. Skrzypek, S. Weng, E. D. Wong, K. Dolinski, D. Botstein, J. M. Cherry, Saccharomyces Genome Database provides mutant phenotype data, *Nucleic Acids Research* 38 (2010) D433–D436. doi:10.1093/nar/gkp917.
- [18] C. Chen, H. Huang, K. E. Ross, J. E. Cowart, C. N. Arighi, C. H. Wu, D. A. Natale, Protein

ontology on the semantic web for knowledge discovery, *Scientific Data* 7 (2020) 337. doi:10.1038/s41597-020-00679-9.

- [19] E. W. Sayers, E. E. Bolton, J. R. Brister, K. Canese, J. Chan, D. C. Comeau, C. M. Farrell, M. Feldgarden, A. M. Fine, K. Funk, E. Hatcher, S. Kannan, C. Kelly, S. Kim, W. Klimke, M. J. Landrum, S. Lathrop, Z. Lu, T. L. Madden, A. Malheiro, A. Marchler-Bauer, T. D. Murphy, L. Phan, S. Pujar, S. H. Rangwala, V. A. Schneider, T. Tse, J. Wang, J. Ye, B. W. Trawick, K. D. Pruitt, S. T. Sherry, Database resources of the National Center for Biotechnology Information in 2023 (????).
- [20] K. Degtyarenko, P. de Matos, M. Ennis, J. Hastings, M. Zbinden, A. McNaught, R. Alcántara, M. Darsow, M. Guedj, M. Ashburner, ChEBI: A database and ontology for chemical entities of biological interest, *Nucleic Acids Research* 36 (2008) D344–D350. doi:10.1093/nar/gkm791.
- [21] V. Emonet, LangChain RDF, 2024.
- [22] Ollama/ollama, Ollama, 2024.
- [23] S. Xiao, Z. Liu, P. Zhang, N. Muennighoff, D. Lian, J.-Y. Nie, C-Pack: Packed Resources For General Chinese Embeddings, in: *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '24*, Association for Computing Machinery, New York, NY, USA, 2024, pp. 641–649. doi:10.1145/3626772.3657878.
- [24] Y. Chang, X. Wang, J. Wang, Y. Wu, L. Yang, K. Zhu, H. Chen, X. Yi, C. Wang, Y. Wang, et al., A survey on evaluation of large language models, *ACM transactions on intelligent systems and technology* 15 (2024) 1–45.
- [25] J. H. Caufield, H. Hegde, V. Emonet, N. L. Harris, M. P. Joachimiak, N. Matentzoglou, H. Kim, S. Moxon, J. T. Reese, M. A. Haendel, P. N. Robinson, C. J. Mungall, Structured Prompt Interrogation and Recursive Extraction of Semantics (SPIRES): A method for populating knowledge bases using zero-shot learning, *Bioinformatics* 40 (2024) btae104. doi:10.1093/bioinformatics/btae104.
- [26] M. E. Aranguren, E. Antezana, M. Kuiper, R. Stevens, Ontology Design Patterns for bio-ontologies: A case study on the Cell Cycle Ontology, *BMC Bioinformatics* 9 (2008) S1. doi:10.1186/1471-2105-9-S5-S1.
- [27] S. Tartir, I. B. Arpinar, Ontology Evaluation and Ranking using OntoQA, in: *International Conference on Semantic Computing (ICSC 2007)*, 2007, pp. 185–192. doi:10.1109/ICSC.2007.19.
- [28] R. Jackson, N. Matentzoglou, J. A. Overton, R. Vita, J. P. Balhoff, P. L. Buttigieg, S. Carbon, M. Courtot, A. D. Diehl, D. M. Dooley, W. D. Duncan, N. L. Harris, M. A. Haendel, S. E. Lewis, D. A. Natale, D. Osumi-Sutherland, A. Ruttenberg, L. M. Schriml, B. Smith, C. J. Stoeckert Jr., N. A. Vasilevsky, R. L. Walls, J. Zheng, C. J. Mungall, B. Peters, OBO Foundry in 2021: Operationalizing open data principles to evaluate ontologies, *Database* 2021 (2021) baab069. doi:10.1093/database/baab069.
- [29] D. Shlyk, T. Groza, M. Mesiti, S. Montanelli, E. Cavalleri, REAL: A Retrieval-Augmented Entity Linking Approach for Biomedical Concept Recognition, in: D. Demner-Fushman, S. Ananiadou, M. Miwa, K. Roberts, J. Tsujii (Eds.), *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, Association for Computational Linguistics, Bangkok, Thailand, 2024, pp. 380–389. doi:10.18653/v1/2024.bionlp-1.29.
- [30] L. N. Soldatova, A. Rzhetsky, Representation of research hypotheses, *Journal of Biomedical Semantics* 2 (2011) S9. doi:10.1186/2041-1480-2-S2-S9.
- [31] L. N. Soldatova, R. D. King, An ontology of scientific experiments, *Journal of The Royal Society Interface* 3 (2006) 795–803. doi:10.1098/rsif.2006.0134.
- [32] S. J. Chalk, Scidata: a data model and ontology for semantic representation of scientific data, *Journal of cheminformatics* 8 (2016) 1–24.
- [33] A. Bandrowski, R. Brinkman, M. Brochhausen, M. H. Brush, B. Bug, M. C. Chibucos, K. Clancy, M. Courtot, D. Derom, M. Dumontier, L. Fan, J. Fostel, G. Fragoso, F. Gibson, A. Gonzalez-Beltran, M. A. Haendel, Y. He, M. Heiskanen, T. Hernandez-Boussard, M. Jensen, Y. Lin, A. L. Lister, P. Lord, J. Malone, E. Manduchi, M. McGee, N. Morrison, J. A. Overton, H. Parkinson, B. Peters, P. Rocca-Serra, A. Ruttenberg, S.-A. Sansone, R. H. Scheuermann, D. Schober, B. Smith, L. N. Soldatova, C. J. S. Jr, C. F. Taylor, C. Torniai, J. A. Turner, R. Vita, P. L. Whetzel, J. Zheng, The Ontology for Biomedical

- Investigations, PLOS ONE 11 (2016) e0154556. doi:10.1371/journal.pone.0154556.
- [34] R. D. King, K. E. Whelan, F. M. Jones, P. G. K. Reiser, C. H. Bryant, S. H. Muggleton, D. B. Kell, S. G. Oliver, Functional genomic hypothesis generation and experimentation by a robot scientist, *Nature* 427 (2004) 247–252. doi:10.1038/nature02236.
- [35] A. Coutant, K. Roper, D. Trejo-Banos, D. Bouthinon, M. Carpenter, J. Grzebyta, G. Santini, H. Soldano, M. Elati, J. Ramon, C. Rouveirol, L. N. Soldatova, R. D. King, Closed-loop cycles of experiment design, execution, and learning accelerate systems biology model development in yeast, *Proceedings of the National Academy of Sciences* 116 (2019) 18142–18147. doi:10.1073/pnas.1900548116.
- [36] G. K. Reder, A. H. Gower, F. Kronström, R. Halle, V. Mahamuni, A. Patel, H. Hayatnagarkar, L. N. Soldatova, R. D. King, Genesis-DB: A database for autonomous laboratory systems, *Bioinformatics Advances* 3 (2023) vbad102. doi:10.1093/bioadv/vbad102.
- [37] A. Litsios, D. H. E. W. Huberts, H. M. Terpstra, P. Guerra, A. Schmidt, K. Buczak, A. Papagiannakis, M. Rovetta, J. Hekelaar, G. Hubmann, M. Exterkate, A. Miliadis-Argeitis, M. Heinemann, Differential scaling between G1 protein production and cell size dynamics promotes commitment to the cell division cycle in budding yeast, *Nature Cell Biology* 21 (2019) 1382–1392. doi:10.1038/s41556-019-0413-3.
- [38] S. Khondker, G.-S. Han, G. M. Carman, Protein kinase Hsl1 phosphorylates Pah1 to inhibit phosphatidate phosphatase activity and regulate lipid synthesis in *Saccharomyces cerevisiae*, *Journal of Biological Chemistry* 300 (2024) 107572. doi:10.1016/j.jbc.2024.107572.
- [39] K. J. Fisher, S. W. Buskirk, R. C. Vignogna, D. A. Marad, G. I. Lang, Adaptive genome duplication affects patterns of molecular evolution in *Saccharomyces cerevisiae*, *PLOS Genetics* 14 (2018) e1007396. doi:10.1371/journal.pgen.1007396.
- [40] J. N. Bröker, B. Müller, N. Van Deenen, D. Prüfer, C. Schulze Gronover, Upregulating the mevalonate pathway and repressing sterol synthesis in *Saccharomyces cerevisiae* enhances the production of triterpenes, *Applied Microbiology and Biotechnology* 102 (2018) 6923–6934. doi:10.1007/s00253-018-9154-7.
- [41] C. Camelo, F. Vilas-Boas, A. P. Cepeda, C. Real, J. Barros-Martins, F. Pinto, H. Soares, H. S. Marinho, L. Cyrne, Opi1p translocation to the nucleus is regulated by hydrogen peroxide in *Saccharomyces cerevisiae*: Opi1p regulation by hydrogen peroxide in *Saccharomyces cerevisiae*, *Yeast* 34 (2017) 383–395. doi:10.1002/yea.3240.
- [42] E. M. Parodi, J. M. Roesner, L. S. Huang, SPO73 and SPO71 Function Cooperatively in Prospore Membrane Elongation During Sporulation in *Saccharomyces cerevisiae*, *PLOS ONE* 10 (2015) e0143571. doi:10.1371/journal.pone.0143571.
- [43] A. Schummer, R. Maier, S. Gabay-Maskit, T. Hansen, W. W. D. Mühlhäuser, I. Suppanz, A. Fadel, M. Schuldiner, W. Girzalsky, S. Oeljeklaus, E. Zalckvar, R. Erdmann, B. Warscheid, Pex14p Phosphorylation Modulates Import of Citrate Synthase 2 Into Peroxisomes in *Saccharomyces cerevisiae*, *Frontiers in Cell and Developmental Biology* 8 (2020) 549451. doi:10.3389/fcell.2020.549451.
- [44] S. Kumar, M. Mashkoo, P. Balamurugan, A. Grove, Yeast Crf1p is an activator with different roles in regulation of target genes, *Yeast* 41 (2024) 379–400. doi:10.1002/yea.3939.
- [45] M. A. Ellison, S. Namjilsuren, M. K. Shirra, M. S. Blacksmith, R. A. Schusteff, E. M. Kerr, F. Fang, Y. Xiang, Y. Shi, K. M. Arndt, Spt6 directly interacts with Cdc73 and is required for Paf1 complex occupancy at active genes in *Saccharomyces cerevisiae*, *Nucleic Acids Research* 51 (2023) 4814–4830. doi:10.1093/nar/gkad180.
- [46] S. Kaps, K. Kettner, R. Migotti, T. Kanashova, U. Krause, G. Rödel, G. Dittmar, T. M. Kriegel, Protein Kinase Ymr291w/Tda1 Is Essential for Glucose Signaling in *Saccharomyces cerevisiae* on the Level of Hexokinase Isoenzyme ScHxk2 Phosphorylation*, *Journal of Biological Chemistry* 290 (2015) 6243–6255. doi:10.1074/jbc.M114.595074.
- [47] S. Sasaki, P. Schlarmann, K. Hanaoka, H. Nishii, H. Moriya, M. Muñiz, K. Funato, Protein sorting upon exit from the endoplasmic reticulum dominates Golgi biogenesis in budding yeast, *FEBS Letters* 598 (2024) 548–555. doi:10.1002/1873-3468.14830.

- [48] M. Molon, O. Woznicka, J. Zebrowski, Cell wall biosynthesis impairment affects the budding lifespan of the *Saccharomyces cerevisiae* yeast, *Biogerontology* 19 (2018) 67–79. doi:10.1007/s10522-017-9740-6.
- [49] J. B. Robertson, C. R. Davis, C. H. Johnson, Visible light alters yeast metabolic rhythms by inhibiting respiration, *Proceedings of the National Academy of Sciences* 110 (2013) 21130–21135. doi:10.1073/pnas.1313369110.
- [50] M. Mühlhofer, E. Berchtold, C. G. Stratil, G. Csaba, E. Kunold, N. C. Bach, S. A. Sieber, M. Haslbeck, R. Zimmer, J. Buchner, The Heat Shock Response in Yeast Maintains Protein Homeostasis by Chaperoning and Replenishing Proteins, *Cell Reports* 29 (2019) 4593–4607.e8. doi:10.1016/j.celrep.2019.11.109.
- [51] W.-C. Hsieh, B. M. Sutter, H. Ruess, S. D. Barnes, V. S. Malladi, B. P. Tu, Glucose starvation induces a switch in the histone acetylome for activation of gluconeogenic and fat metabolism genes, *Molecular Cell* 82 (2022) 60–74.e5. doi:10.1016/j.molcel.2021.12.015.

A. Tables

Table 1
Paper GO Terms

Paper	GO Term
1 [37]	Cellular Process (GO:0009987)
2 [38]	Metabolic Process (GO:0008152)
3 [39]	Chromatin Organization (GO:0006325)
4 [40]	Metabolic Process (GO:0008152)
5 [41]	Response to Stimulus (GO:0050896)
6 [42]	Developmental Process (GO:0032502)
7 [43]	Localization (GO:0051179)
8 [44]	Biological Regulation (GO:0065007)
9 [45]	Biological Regulation (GO:0065007)
10 [46]	Signaling (GO:0023052)
11 [47]	Cellular Component Organization or Biogenesis (GO:0016043)
12 [48]	Cellular Process (GO:0009987)
13 [49]	Biological Phase (GO:0044848)
14 [50]	Homeostatic Process (GO:0042592)
15 [51]	Response to Stimulus (GO:0050896)

Table 2
Human Extracted Hypotheses

Paper	Human Hypotheses
1 [37]	Progression through the START Checkpoint of the cell cycle is dependent on an increased production rate of Cln3 which does not scale relative to cell size increase.
2 [38]	In <i>Saccharomyces cerevisiae</i> , the Hsl1 protein kinase inhibits the activity of the Pah1-encoded phosphatidate phosphatase (PAP) by phosphorylation at sites Ser-748 and Ser-773 of the PAP protein, leading to a reduced conversion of phosphatidate (PA) into diacyl-glycerol (DAG) and increase in the synthesis of membrane phospholipids.
3 [39]	Whole genome duplication (WGS) in <i>Saccharomyces cerevisiae</i> provides an immediate fitness advantages via accumulation of structural variants enabling adaptive strategies to environment, however this comes at a cost of slowing long-term adaptability, for example via recessive mutations, aneuploidies, and copy-number variants.
4 [40]	Increased production of pentacyclic triterpenes in <i>Saccharomyces cerevisiae</i> can be engineered through overexpression of mevalonate (MVA) pathway genes such as ERG13 (HMGS) and HMG1 (HMGR), overexpression of lupeol synthase, deletion of the negative regulator of the MVA pathway, ROX1, and repression of the sterol synthesis pathway gene ERG7.
5 [41]	Under hydrogen peroxide induced stress in <i>Saccharomyces cerevisiae</i> , the Opi1p protein translocates to the nucleus to repress transcription of inositol upstream activating sequence (UAS-INO)-containing genes for the cell to adapt its cell membrane to reduce permeability to exogenous hydrogen peroxide.
6 [42]	SPO73 acts in a pathway involving SPO71, VPS13 and SPO1 to regulate proper prospore membrane elongation in <i>Saccharomyces cerevisiae</i> .
7 [43]	The peroxisomal import membrane protein, Pex14p, is regulated by phosphorylation in <i>saccharomyces cerevisiae</i> .
8 [44]	The <i>Saccharomyces cerevisiae</i> transcription factor, Crf1p, activates the transcription of the ribosomal biogenesis genes, UTP22 and HMO1, as a mechanism to fine-tune responses to mTORC1 signaling.
9 [45]	The Cdc73 subunit of the Paf1 complex in <i>Saccharomyces cerevisiae</i> interacts with subunits of the Pol II elongation complex in an unknown manner.
10 [46]	One or all of the kinases, Ymr291w/Tda1, PKA, Sch9, and Snf1 phosphorylate hexokinase ScHxk2 at phosphorylation site serine 15 to activate transcription of glucose-repressible genes under low external glucose levels in <i>Saccharomyces cerevisiae</i>
11 [47]	The number of Golgi cisternae produced and maintained in a <i>Saccharomyces cerevisiae</i> is tightly regulated by the GPI-anchored protein sorting process in the endoplasmic reticulum, particularly to the endoplasmic reticulum exit sites (ERESs).
12 [48]	In <i>Saccharomyces cerevisiae</i> , genes related to cell wall biosynthesis and maintenance, are direct determinants of the budding lifespan of the cell, as measured by its reproductive longevity or budding cycle capacity.
13 [49]	In <i>Saccharomyces cerevisiae</i> , visible light affects the Yeast Respiratory Oscillator (YRO) through impacting oxidative state, leading to negative effects on metabolism.
14 [50]	The heat shock response program in <i>Saccharomyces cerevisiae</i> involves changes to the expression of genes encoding proteins with a wide range of functions in addition to molecular chaperones.
15 [51]	Under glucose starvation and carbon scarcity conditions in <i>Saccharomyces cerevisiae</i> , Acetyl-CoA driven histone acetylation is shifted to focus on up-regulation of transcriptional programs focused on starvation survival and metabolism regulation, namely gluconeogenesis and fat metabolism.

Table 3
LLM Extracted Hypotheses

Paper	LLM Hypotheses
1 [37]	The differential scaling between protein production rate and cell size dynamics leads to a temporal increase in Cln3 concentration, which promotes commitment to the cell division cycle in budding yeast.
2 [38]	The phosphorylation of Pah1 by the septin-associated protein kinase Hsl1 inhibits phosphatidate phosphatase (PAP) activity, thereby promoting membrane phospholipid synthesis at the expense of triacylglycerol synthesis in <i>Saccharomyces cerevisiae</i> .
3 [39]	Whole genome duplication (WGD) in haploid <i>Saccharomyces cerevisiae</i> populations provides a direct fitness advantage, but results in a reduced adaptation rate due to a shift in the spectrum of beneficial mutations and the accumulation of recessive deleterious mutations, indicating a trade-off between immediate fitness gains and long-term adaptability.
4 [40]	Upregulating the mevalonate pathway and repressing sterol synthesis in <i>Saccharomyces cerevisiae</i> enhances the production of triterpenes, specifically by overexpressing MVA pathway genes, deleting the ROX1 gene, and using a copper-regulated promoter to redirect metabolic flux from sterol biosynthesis to the production of pentacyclic triterpenes such as lupeol and β -amyrin.
5 [41]	Hydrogen peroxide (H ₂ O ₂) in concentrations that induce adaptation regulates the translocation of the transcriptional repressor Opi1p to the nucleus in <i>Saccharomyces cerevisiae</i> , affecting the expression of UASINO-containing genes, and this process involves the oxidation of cysteine residue 159 of Opi1p.
6 [42]	SPO73 and SPO71 function cooperatively to promote prospore membrane elongation during sporulation in <i>Saccharomyces cerevisiae</i> , acting in opposition to SPO1, while VPS13 plays a distinct role in prospore membrane development.
7 [43]	Phosphorylation of Pex14p at serine 266 modulates the peroxisomal import of citrate synthase 2 (Cit2p) in <i>Saccharomyces cerevisiae</i> , thereby helping the cells to adjust their carbohydrate metabolism according to the nutritional conditions.
8 [44]	Crf1p functions as an activator of the ribosomal biogenesis genes UTP22 and HMO1, and its presence is required to prevent rebinding of Ifh1p to some gene promoters after mTORC1 inhibition, thereby optimizing transcriptional responses during mTORC1 inhibition.
9 [45]	Spt6 directly interacts with Cdc73 and is required for Paf1 complex occupancy at active genes in <i>Saccharomyces cerevisiae</i> , suggesting that Spt6 plays a critical role in recruiting or retaining Paf1C on transcribing RNA polymerase II.
10 [46]	Ymr291w/Tda1 is essential for the phosphorylation of hexokinase isoenzyme ScHxk2 at serine 15 in <i>Saccharomyces cerevisiae</i> , and this phosphorylation is crucial for glucose signaling, particularly under conditions of limited external glucose availability.
11 [47]	Protein sorting in the endoplasmic reticulum, particularly the sorting of GPI-anchored proteins into selective ER exit sites, plays a critical role in maintaining Golgi biogenesis by affecting the number of Golgi cisternae in budding yeast.
12 [48]	Disturbances in cell wall biosynthesis, particularly in chitin biosynthesis and cell wall protein mannosylation, affect the budding lifespan of <i>Saccharomyces cerevisiae</i> yeast, with specific mutations leading to either a reduction or increase in lifespan, suggesting that cell wall integrity and composition play a significant role in determining yeast longevity.
13 [49]	Visible light, particularly blue and green wavelengths, significantly modulates the period and amplitude of the yeast respiratory oscillation (YRO) by inhibiting respiration through light absorption by cytochromes, which affects electron transport and oxidative phosphorylation, and this modulation is similar to the effects observed with the electron transport inhibitor sodium azide.
14 [50]	The heat shock response (HSR) in yeast is modular and involves the upregulation of genes to counterbalance increased protein turnover due to stress, maintaining protein homeostasis by replenishing proteins through increased synthesis and chaperoning, with distinct programs activated depending on the severity of the stress.
15 [51]	Glucose starvation induces a switch in the histone acetylome, mediated by the histone deacetylase Rpd3p and the acetyltransferase Gcn5p, to activate genes involved in gluconeogenesis and fat metabolism, thereby promoting cell survival under conditions of limited acetyl-CoA.