Agentic AI integrated with scientific knowledge: laboratory validation in systems biology

Downloaded from: https://research.chalmers.se, 2026-11-06 22:01 UTC

Citation for the original published paper (version of record):
Brunnsåker, D., Gower, A., Naval, P. et al (2026). Agentic AI integrated with scientific knowledge: laboratory validation in systems biology. Journal of the Royal Society Interface, 23(240). http://dx.doi.org/10.1098/rsif.2026.0043

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

research.chalmers.se offers the possibility of retrieving research publications produced at Chalmers University of Technology. It covers all kind of research output: articles, dissertations, conference papers, reports etc. since 2004. research.chalmers.se is administrated and maintained by Chalmers Library

(article starts on next page)

INTERFACE

royalsocietypublishing.org/rsif

CC BY

Research

Cite this article: Brunnsåker D, Gower AH, Naval P, Bjurström EY, Kronström F, Tiukova I, King R. 2026 Agentic AI integrated with scientific knowledge: laboratory validation in systems biology. *J. R. Soc. Interface* **23**: 20260043.
https://doi.org/10.1098/rsif.2026.0043

Received: 12 January 2026
Accepted: 15 May 2026

Subject Category:
Life Sciences–Engineering interface

Subject Areas:
computational biology, systems biology, bioengineering

Keywords:
laboratory automation, systems biology, machine learning, inductive logic programming, automation of science, large language models

Author for correspondence:
Ross King
e-mail: rk663@cam.ac.uk

Supplementary material is available online at https://doi.org//10.6084/m9.figshare.c.8495771.

THE ROYAL SOCIETY PUBLISHING

Agentic AI integrated with scientific knowledge: laboratory validation in systems biology

Daniel Brunnsåker[1], Alexander Howard Gower[1], Prajakta Naval[2], Erik Yuusuke Bjurström[2], Filip Kronström[1], Ievgeniia Tiukova[2,3] and Ross King[1,4]

[1]Department of Computer Science and Engineering and [2]Department of Life Sciences, Chalmers University of Technology, Gothenburg, Sweden
[3]Division of Industrial Biotechnology, KTH Royal Institute of Technology, Stockholm, Sweden
[4]Department of Chemical Engineering and Biotechnology, University of Cambridge, Cambridge, UK

DB, 0000-0002-5167-0536; EYB, 0000-0002-8863-1207; RK, 0000-0001-7208-4387

Automation is transforming scientific discovery by enabling systematic exploration of complex hypotheses. Large language models (LLMs) perform well across diverse tasks and promise to accelerate research, but often struggle with logical structures. Here, we present a framework for biological discovery integrating LLM-based agents with laboratory automation, guided by logical scaffolds incorporating symbolic relational learning, structured vocabularies and experimental constraints. This integration improves coherence and reliability in automated workflows. We couple this AI-driven approach to automated cell-culture and metabolomics platforms, enabling integrated hypothesis validation and refinement, yielding a flexible discovery system. The system identified novel interactions in *Saccharomyces cerevisiae*, including glutamate-induced growth inhibition in spermine-treated cells and aminoadipate's partial rescue of formic-acid stress. All hypotheses, experiments and data are captured in a graph database employing controlled vocabularies. Existing ontologies are extended, and a novel representation of scientific hypotheses is presented using description logics. This work demonstrates the potential for a reliable machine-driven discovery process in systems biology.

1. Introduction

Artificial intelligence (AI) and laboratory automation are rapidly transforming science. AI enables the analysis of data that would otherwise be intractable, and when integrated with laboratory robotics, can automate scientific research—a capability essential for addressing the increasingly complex problems found in modern science [1–3]. For example, comparatively simple model organisms, such as *Saccharomyces cerevisiae* (baker's yeast) and *Escherichia coli*, exhibit complexity far beyond what humans can interpret or validate experimentally within a reasonable timeframe [4]. Reliable, scalable scientific discovery under these conditions requires integrated automated experimentation and computational analysis [4,5].

The effectiveness of a scientist (human or AI) is determined by their scientific ideas, agency (the tools and techniques available to them) and collaborative capacity. For AI scientists, deciding what agency to grant them is an important design question. In particular, real-world agency—the ability to directly control experiments—is vital for scientific discovery, and integration with laboratory

© 2026 The Authors. Published by the Royal Society under the terms of the Creative Commons Attribution License http://creativecommons.org/licenses/by/4.0/, which permits unrestricted use, provided the original author and source are credited.

Downloaded from royalsocietypublishing.org/rsif/article-pdf/doi/10.1098/rsif.2026.0043/6150761/rsif20260043.pdf
by Chalmers University of Technology user
on 21 August 2026

robotics was a focus of the previous generation of robot scientists. These model-driven systems generated hypotheses in tightly controlled experimental spaces and collected empirical data in only a few data modalities [4,6–8]. Several variations of this approach have been applied in fields such as materials development and chemistry [9,10]. The next generation of AI scientists must contend with increasingly complex scientific questions that demand more powerful tools for hypothesis generation, multimodal data integration, experimental design and analysis [1].

One promising direction is to grant AI scientists access to large language models (LLMs), which have enabled AI systems to achieve impressive results in tasks that previously required human intelligence, including scientific discovery [3,11]. LLM-augmented approaches offer adaptability in how problems are formulated and in expanding the experimental scope—a major advantage in the natural sciences. Consequently, we are increasingly seeing the widespread use and effect of LLMs in scientific discovery, often via orchestrated multi-agent systems [12–22].

The adaptability of LLM-augmented approaches is particularly beneficial in systems biology research, which involves complex interaction networks and vast spaces of experimental conditions and data modalities [23]. However, LLM outputs are often inconsistent, difficult to verify and rarely subjected to rigorous experimental evaluation—issues compounded by siloed data practices that limit transparency and reuse [18]. A further weakness of LLMs remains their inability, or at best inefficiency and unreliability, when interacting with formal languages and logical structures. Recent agentic systems mitigate this issue by granting the LLM agency to write and execute programmes in high-level languages such as Python [17]. However, these approaches generally rely on loosely typed message passing, shallow artefact tracking and lack unified domain representations. As a result, they can generate brittle, unvalidated outputs that are difficult to systematically audit [18].

Agentic systems for scientific discovery should make the best use of available scientific data and models, and report hypotheses and findings in a way that enables good collaboration and recording of knowledge [20,24]. In many cases, including in systems biology, there is a wealth of structured data encoded according to well-defined rules. There is thus a need for frameworks that retain the flexibility of LLMs while grounding them in formal logic, exploiting background knowledge and meeting transparent data-sharing requirements [12,17–19,25].

Here, we introduce a hybrid approach that integrates language-driven flexibility with inductive logic and formal ontologies (see figure 1). The framework combines LLMs with the precision of relational learning (inductive logic programming (ILP)) and community-adopted scientific ontologies to generate hypotheses grounded in empirical data and structured biological knowledge. The hypotheses are prioritized based on prior experimental results and can, thanks to their structured format, be automatically tested on laboratory automation platforms such as robotic liquid handlers, automated cultivation systems and mass-spectrometry-based metabolomics [8,26]. All metadata, logs and results can then be stored in a graph database that ensures transparency, reproducibility and efficient reuse [27]. Human intervention was limited to laboratory administration (preparation of stock solutions and long-term metabolomics reference samples), interfacing tasks (occasional plate transfers, script and data transfers between non-networked machines) and safety review of proposed experiments.

By integrating symbolic learning with ontologies and statistical inference, the framework overcomes key limitations of standard agentic LLM systems: it improves traceability, reduces speculative outputs and enables higher degrees of auditability across the entire discovery loop.

This framework was evaluated through automated experimental testing of several data-driven hypotheses. For example, we observed glutamate-induced synergistic growth inhibition in spermine-treated cells and arginine-mediated enhancement of caffeine toxicity in caffeine-treated cells. We further applied it to test a metabolomics-informed prediction that led to the discovery of aminoadipate-mediated rescue of formic acid stress. These results demonstrate the framework's potential for automated, scalable, reproducible discovery in biological systems.

2. Results

2.1. Automated systems biology research

We built an automated experimental and computational framework by integrating LLMs; structured community databases; robotic liquid handling stations; an automated cell-culturing and sampling platform; an ion mobility–based mass spectrometry metabolomics workflow; and computational tools in R, Python and Prolog [8,26,28,29]. See figure 1 for a graphical overview of the process.

This framework is designed to automatically generate and experimentally evaluate hypotheses about cell biology. Testable hypotheses are generated by (i) data-mining semantically meaningful facts about *S. cerevisiae* phenotype and metabolism—extracted from structured community databases [30–34]—(ii) applying pattern mining to generate a logical rule [28,29], and (iii) applying machine learning on metabolomics data to learn a testable implication associated with the rule [35]. The generated hypotheses are filtered and prioritized based on prior experimental data. LLM agents then design experimental interventions to test these hypotheses and further formalize them into a machine-readable format with executable real-world constraints [36]. Automated experiments are executed for each hypothesis, producing and processing biological samples to generate time-series growth data and endpoint metabolic profiles via ion-mobility-based mass spectrometry. After each run, the data are automatically curated and analysed, and presented in a concise, user-readable summary.

Downloaded from royalsocietypublishing.org/rsif/article-pdf/doi/10.1098/rsif.2026.0043/6150761/rsif20260043.pdf
by Chalmers University of Technology user
on 21 August 2026

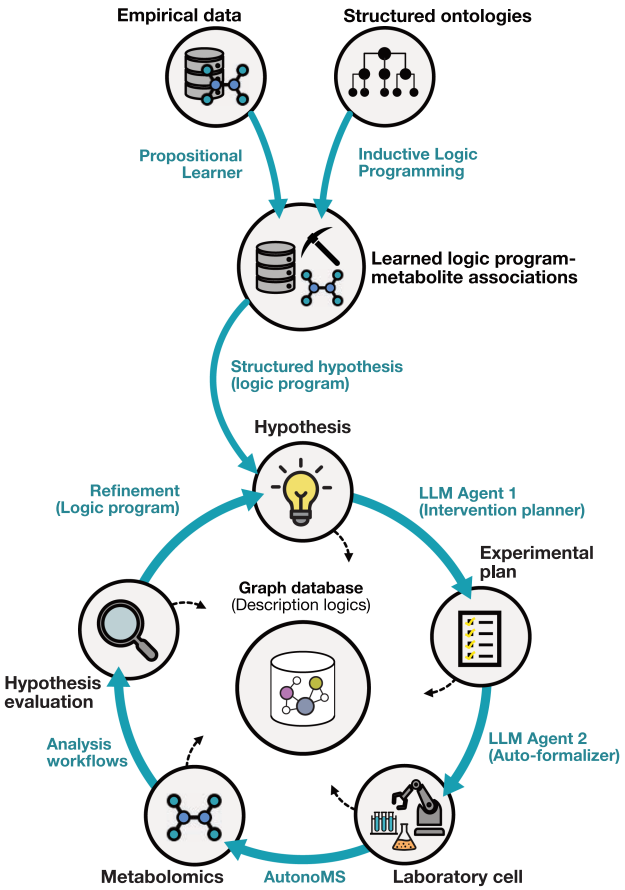

Figure 1. An automated experimental framework enabling end-to-end biological discovery—from hypothesis generation to data integration. Hypotheses are generated by mining patterns from a Datalog database containing structured facts about yeast phenotype, physiology and metabolism, using ILP. These are linked to metabolomics data via regression. Interventions to test each hypothesis are designed by an LLM, which incorporates real-world experimental constraints. The interventions are auto-formalized into machine-readable protocols and executed by an automated laboratory cell. Time-series growth data are collected throughout the experiments, followed by automated sample processing and acquisition of endpoint metabolic profiles via ion-mobility mass spectrometry, mediated by AutonoMS [26]. All data, including results, intermediate outputs and metadata, are curated, analysed and integrated into a graph database for downstream access and reuse.

Each step of the process and accompanying experimental results are recorded in a graph database, including all LLM queries and responses, to ensure transparency and facilitate efficient reuse of data. Experimental data and metadata are saved using controlled vocabularies, utilizing an extended version of the ascomycete phenotype ontology (APO) [30,31]. Hypotheses and their implications are represented using description logics. This approach allows reusing existing data when a generated hypothesis matches a previously executed experiment, as demonstrated for one hypothesis below (see *hypothesis 9* in table 1).

2.2. Linking metabolites to traits for hypothesis discovery

To generate testable, data-driven hypotheses based on prior scientific knowledge, we constructed a Datalog database of approximately 60 000 phenotypical, physiological and metabolic relations for *S. cerevisiae*. This database integrates curated datasets from several different sources, structured by established ontologies, and encoded as logical facts to enable structured reasoning [31–33]. Datalog's declarative structure enables clear expression of biological relationships and efficient querying across complex relational datasets [37].

To identify patterns that could explain observed phenotypes, we applied ILP, a methodology well-suited to learning interpretable and well-defined logic programs (i.e. rules) from structured, relational data [28,29,38]. ILP-based pattern mining on this structured database yielded 735 distinct logic programs, 390 of which mapped uniquely to positive examples (gene deletion strains) from the metabolomics dataset of Mülleder *et al.* [35]. We propositionalized [39] each program into a binary feature (denoting whether the program covers a given example) and fitted a linear regression model to link these logic programs to metabolite measurements through regression coefficients. This produced 1933 independent hypotheses across 16 proteinogenic amino acids (see figure 2).

An overview of the hypothesis space can be seen in supplementary material, note S1 and table S1. A logic program and learned association can be seen below (also see *hypothesis 6* in table 1).

Downloaded from royalsocietypublishing.org/rsif/article-pdf/doi/10.1098/rsif.2026.0043/6150761/rsif20260043.pdf
by Chalmers University of Technology user
on 21 August 2026

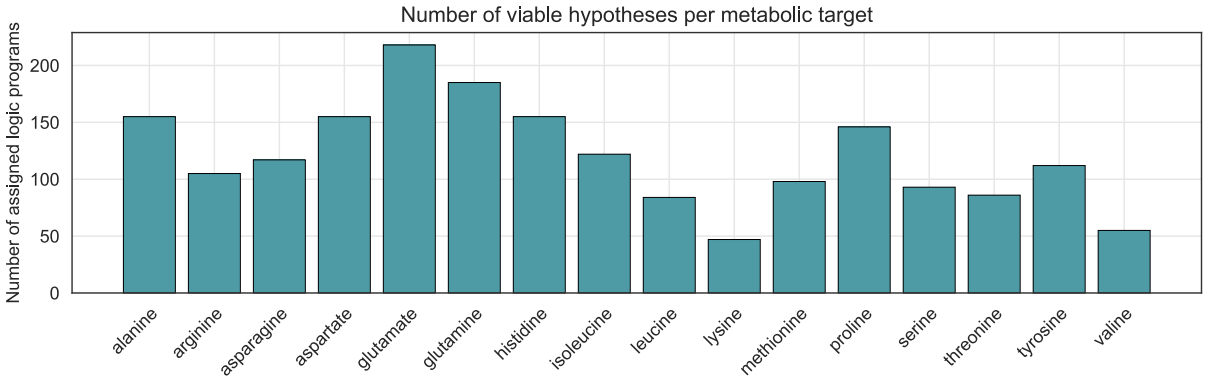

Figure 2. Viable hypothesis counts per amino acid. Bar chart showing, for each amino acid, the number of logic programs with non-zero regression coefficients—that is, the total number of viable hypotheses generated for that amino acid.

Table 1. Summary of hypotheses selected for testing. The evaluation shows, for each hypothesis, whether the change in intracellular accumulation and the effect on resistance were (i) observed and (ii) in the predicted direction, and whether the negative control showed specificity. '−' and '+' indicate the observed or predicted sign of the change. '~' denotes an inconclusive change in accumulation. **Iterative hypothesis generated through data generated in hypothesis 4. **Hypothesis tested using already acquired data. †Experiment tests inverse hypothesis, that is, nutrient supplementation rather than reduction, and reversed effect.*

generated hypothesis	change in accumulation		effect on resistance		
	observed	predicted	observed	predicted	intervention specificity
hypothesis 1 ↓ arginine ⟹ ↑ caffeine resistance†	+	+	−	−	yes
hypothesis 2 ↑ glutamate ⟹ ↓ spermine resistance	~	+	−	−	yes
hypothesis 3 ↑ glutamate ⟹ ↓ formic acid resistance	~	+	−	+	no
hypothesis 4 ↓ glutamine ⟹ ↓ acetic acid resistance†	+	+	−	+	no
hypothesis 5 ↑ lysine ⟹ ↓ hyperosmotic stress resistance (at 30% (w/v) sucrose)	+	+	+	−	yes
hypothesis 6 ↓ arginine ⟹ ↓ lithium(+1) resistance (at 0.1 M lithium chloride)†	+	+	~	+	no
hypothesis 7 ↑ proline ⟹ ↑ lactic acid resistance	+	+	~	+	no
*hypothesis 8** ↑ aminoadipate ⟹ ↑ formic acid resistance (at 10 mM formic acid)	−	+	+	+	yes
*hypothesis 9*** ↑ proline ⟹ ↓ formic acid resistance	~	+	~	−	no

```
– Arginine, Cells(A) :=
    ExhibitsPhenotype(A, Decreased Resistance, B, C),
    CompoundName(B, Lithium(+1)),
    Condition(C, Treatment : 0.1M Lithium Chloride).
```

In this context, this program can be interpreted as: cells (A) which have decreased resistance to lithium (B) when treated with 0.1 M lithium chloride (C) are associated with lower intracellular arginine levels. The link between arginine levels and the resistance phenotype suggests the testable hypothesis that arginine supplementation will restore lithium resistance (i.e. ↑ Arginine ⟹ ↑ Lithium(+1) resistance).

2.3. Automated experiments identify interaction effects

To evaluate the utility of the framework, we conducted automated investigations on chemical stressor and nutrient intervention interactions (the most common type of hypothesis generated in the previous step). For each amino acid, the framework selected the highest weighted hypothesis, chose an appropriate negative control from the model's learned associations (see §4), and ran the experiment—producing time-series growth curves, endpoint metabolic profiles, effect sizes and statistical evaluations.

Downloaded from royalsocietypublishing.org/rsif/article-pdf/doi/10.1098/rsif.2026.0043/6150761/rsif20260043.pdf
by Chalmers University of Technology user
on 21 August 2026

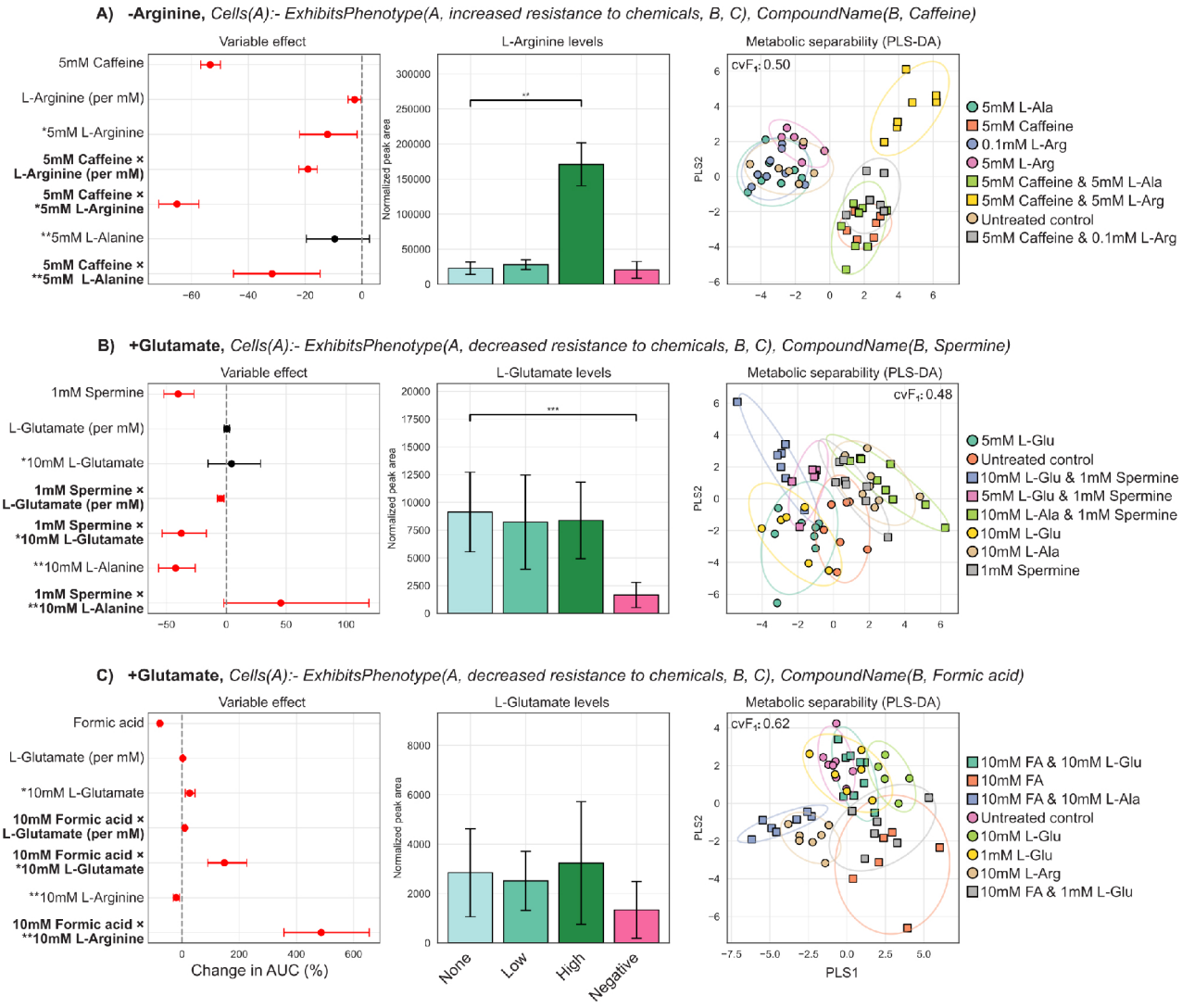

Figure 3. Results from automated interaction experiments with examples of different outcomes. The title of each row denotes the logic program that was tested along with the learned association that defines the intervention. The forest plot signifies the coefficients and confidence intervals (2.5% and 97.5%) derived from a generalized linear model using AUC as the response variable. Red highlights significant ($p < 0.05$) variable effects as calculated by a non-parametric bootstrap (5000 resamples). The bar-plot shows the intracellular accumulation of the intervention compound, with the error-bars representing the standard deviation. Asterisks in the bar-chart denote significance (*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$). *None*, *Low* and *High* denote the levels of concentrations of the intervention compound. The PLS-DA loading plot visualizes the two-dimensional separability of the metabolic profiles of the experimental groups, confirming a metabolic change from the supplied interventions. The classification performance for the full five PLS-DA components can be seen in the corner. **Effect scaled to correspond to an equimolar dose of the negative control. **Negative control.*

Investigations were performed on five proteinogenic amino acids: L-glutamate, L-arginine, L-proline, L-glutamine and L-lysine. Most investigations revealed significant interaction effects on growth in stressor-treated cells upon amino acid supplementation, several of which have not been previously discovered. An overview of all investigations is presented in table 1, with full statistical details in supplementary material, note S7 and table S2.

Hypothesis 1 (see table 1) was consistent with the acquired empirical data (shown in figure 3*A*). Addition of L-arginine alone caused a modest reduction in area under the growth curve (AUC) (−2.5% per mM, $p < 0.05$), whereas caffeine treatment showed an inhibitory effect (−53.4%, $p < 0.001$). Their combined application resulted in a synergistic decrease relative to independent effects (−19.0% per mM, $p < 0.001$), while the L-alanine negative control yielded a smaller decrease (−31.6% at 5mM, $p < 0.01$), confirming intervention specificity. Intracellular L-arginine accumulation increased significantly under high extracellular L-arginine supplementation ($p < 0.01$), confirming the validity of the intervention.

Hypothesis 2 was partially consistent with empirical data (see figure 3*B*). Addition of only L-glutamate yielded a small—but statistically insignificant—increase in AUC. However, it significantly sensitized the cells to spermine (−4.6% per mM, $p < 0.01$). Intracellular L-glutamate accumulation did not increase with supplementation, suggesting a downstream metabolite may be the actual mediator of changes in resistance. The L-alanine control increased resistance to spermine (+45.5% at 10 mM, $p < 0.05$) and significantly decreased L-glutamate accumulation ($p < 0.001$).

Downloaded from royalsocietypublishing.org/rsif/article-pdf/doi/10.1098/rsif.2026.0043/6150761/rsif20260043.pdf
by Chalmers University of Technology user
on 21 August 2026

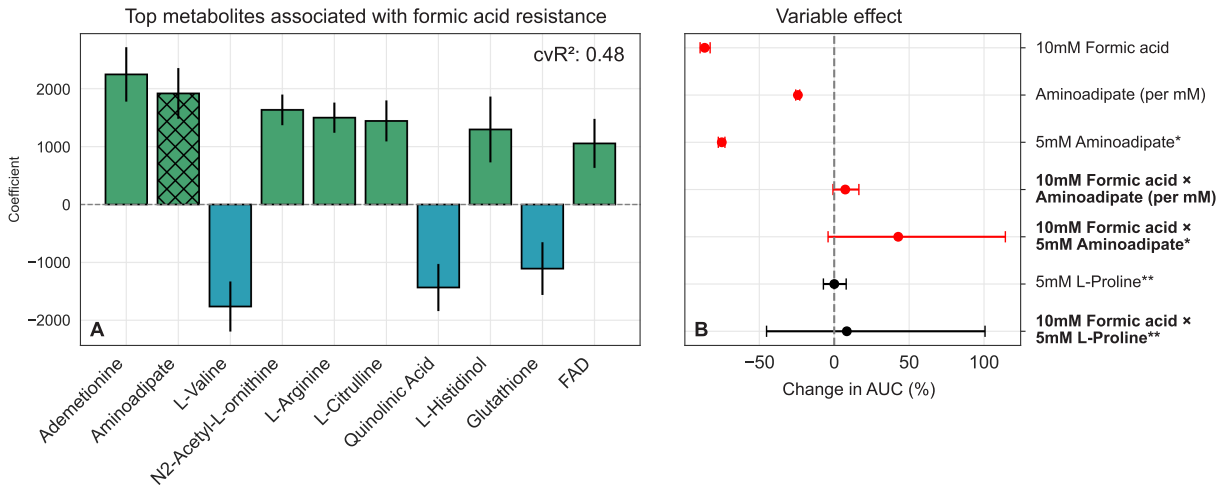

Figure 4. Iterative hypothesis refinement. (A) The bar-plot denotes the average magnitude of coefficients derived from a linear regressor when predicting for resistance to formic acid stress (using data derived from *hypothesis 4*, see figure 3*B*). Regression performance is presented as coefficient of determination across repeated cross validation (cvR^2). The hatched bar is the compound selected to generate the experiment (aminoadipate). The error bars denote the standard deviation across cross validation folds. (B) The forest plot signifies the coefficients and confidence intervals (2.5% and 97.5%) derived from a generalized linear model using AUC as the response variable. Red highlights significant ($p < 0.05$) variable effects as calculated by a non-parametric bootstrap (5000 resamples). The interaction terms are of special importance to the hypothesized logic programme. **Effect scaled to correspond to an equimolar dose of the specificity control. **Negative control.*

In a few select cases, the experimental validation of the hypothesis showed a non-predicted and non-specific interaction. For example, for *hypothesis 3*, as seen in figure 3*C*, the interaction effect between L-glutamate and formic acid was significant (+9.5% per mM, $p < 0.001$), but not in the predicted direction. The L-alanine negative control had a much larger effect on resistance under treatment compared to independent effects (+486% at 10 mM, $p < 0.001$), almost nullifying the effect of the stressor. L-glutamine and acetic acid (*hypothesis 4*) showed strong synergistic inhibition (−4.3% per mM, $p < 0.05$). However, L-leucine showed an equally strong effect (−36.9% at 10 mM, $p < 0.01$), indicating the non-specificity of the intervention. L-glutamine was not significantly accumulated when supplemented, indicating a downstream mediator of reduced resistance.

Regarding *hypothesis 5*, L-lysine exhibited a specific, strong rescue effect against sucrose-induced hyper-osmotic stress (+8.9% per mM, $p < 0.001$), while valine produced a modest rescuing effect (+22.5% at 10 mM, $p < 0.05$). Intracellular L-lysine levels increased under the highest supplementation level ($p < 0.05$).

Hypothesis 6 suggested a rescuing effect of L-arginine supplementation on lithium-induced stress, but this was inconsistent with the data. Exposure to 100 mM lithium chloride severely reduced AUC (−84.6%, $p < 0.001$). The intervention had little effect on resistance outcome, although it did increase intracellular L-arginine accumulation at higher levels of supplementation ($p < 0.05$). The L-alanine negative control showed a synergistic decrease in AUC relative to independent effects (−15.9% at 5 mM, $p < 0.05$) and decreased intracellular concentrations of arginine ($p < 0.05$). *Hypothesis 7* (involving L-proline and lactic acid) showed minimal effect on treatment outcome. Both the intervention and the negative control affected growth dynamics moderately, but only the L-alanine negative control reached statistical significance (−38.6% at 20 mM, $p < 0.001$).

All outcomes—whether supporting or refuting the initial hypothesis—are systematically recorded in a graph database.

Partial least-squares discriminant analysis (PLS-DA) revealed clear separation of controls, interventions and stressors through their metabolic profiles. Cross-validation metrics indicated high model reliability, underscoring the consistency and specificity of the metabolomics data. Evaluation metrics and low-dimensional representations of several investigations can be seen in figure 3. Processed data are stored in the graph database and raw data are stored in online repositories (see details in *Data accessibility*).

2.4. Iterative hypothesis refinement showed alleviation of formic acid stress by aminoadipate

Regularized linear regression was applied to data from *hypothesis 3* (visualized in figure 3*B*) to model growth dynamics as a function of the acquired formic acid-treated metabolic profiles. By interpreting the model's feature weights, this iterative analysis refined the initial hypothesis and led to a new, testable prediction. Based on the resulting ranked metabolites (see figure 4*A*), a logic program was constructed using experiment metadata and the highest ranked locally available compound—aminoadipate (see below).

The hypothesis was fed into the initial step of the pipeline, and the experiment was subsequently performed. Results can be seen in figure 4*B*. Complete regression coefficients can be found in supplementary material, table S3.

The results showed a partial rescue of formic acid stress (+7% per mM, $p < 0.05$) when combined with aminoadipate, confirming the hypothesis. Note that the supplement and treatment are independently highly toxic to the cells (see figure 4*B*), and as such, the dynamics might be subject to saturation effects.

Downloaded from royalsocietypublishing.org/rsif/article-pdf/doi/10.1098/rsif.2026.0043/6150761/rsif20260043.pdf
by Chalmers University of Technology user
on 21 August 2026

2.5. A database of hypotheses and experimental data enables the efficient reuse of experimental data

Although our automated procedures improve efficiency, conducting experiments remains costly. To enable reuse of empirical data for initial assessment of hypotheses, and to fully record the hypotheses and experimental data, existing ontologies (e.g. APO) were extended and connected. We transformed the generated hypotheses, the experimental protocols and the empirical data into OWL-DL statements and recorded these in a graph database (figure 5*A,B*), which can be queried for empirical data relevant for a particular hypothesis. The following hypothesis—suggesting an interaction between intracellular L-proline accumulation and resistance to formic acid stress—was generated by the initial stages of the automated procedure (see *hypothesis 9* in table 1).

```
+ Proline, Cells(A) :=

    ExhibitsPhenotype(A, Decreased Resistance, B, C),

    CompoundName(B, Formic acid)
```

Before executing any experiments, the database was queried for relevant empirical data. Proline was used as a negative control for *hypothesis 8*, meaning there are data from experiments conducted with proline supplementation and formic acid treatment. Returning these data from the database—see figure 5*C*—shows no decrease in resistance to formic acid stress, see figure 4*B*, which is inconsistent with the hypothesis. Therefore, this hypothesis can be de-prioritized for experimental validation, conserving valuable resources.

3. Discussion

The future of scientific research lies in automation: advances in robotics, AI and high-throughput technologies are transforming the scientific process [1,3]. These innovations will enable the minimization of human bias, accelerate research cycles and reduce variation introduced by human error and incomplete recording of experimental protocols, environmental conditions and other subtle factors that can influence outcomes and reproducibility [40–42]. In this automated paradigm, algorithms design experiments, robots execute protocols and the results are all automatically analysed, thus enabling exploration of vast experimental spaces beyond human throughput [4,7,8,13]. By codifying workflows and enforcing rigorous data-handling and reusability practices, such platforms ensure that every result is traceable, reproducible and extensible. Ultimately, combining complex exploration with automated, quality-controlled systems promises to transform biology into a truly scalable, collaborative and reliable discipline.

Building on this vision, we designed and evaluated an automated multi-agent framework that systematically produces and experimentally evaluates logic-derived, data-driven computational hypotheses, with metadata captured at every stage. We also demonstrate the utility of metabolomics data for evaluating interventions and for suggesting alternative explanations and follow-up experiments. Unlike standard agentic workflows—that orchestrate flexible yet fragile toolchains with limited semantic structure—our framework formalizes knowledge representation from the outset. This foundation enables consistent hypothesis generation, rigorous tracking of experimental logic and clearer links between data and interpretation.

Several hypotheses generated by our framework point to biologically relevant but previously underexplored interactions in *S. cerevisiae*. One example is the interaction between L-glutamate and spermine, despite these being two essential compounds connected to each other via several different reactions in glutathione metabolism (KEGG) [43].

Another previously uncharacterized finding is the interaction between L-arginine and caffeine, probably mediated by distinct regulation of the TORC1-complex. Caffeine acts as an inhibitor of TORC1 in yeast, and arginine is tightly linked to TORC1 activity [44,45].

Even when outcomes diverged from predictions, the data revealed significant interactions between interventions and stressors—for example, between L-lysine and sucrose-induced hyper-osmotic stress. Lysine-related genes are downregulated under this stressor [46], but direct functional links remain sparse.

Through iterative hypothesis refinement, we have demonstrated for the first time that aminoadipate confers resistance to formic acid stress, whereas L-glutamate's role in modulating tolerance has been reported previously [47].

Each hypothesis tested by the experiments has two components: a hypothesized effect of intervention (broadly, supplementation ⟶ accumulation), and a hypothesized resultant effect (accumulation ⟶ phenotype). In the case of *hypotheses* 2–4, the first part about the effect of intervention proved not to be realized, highlighting mechanistic assumptions that proved limited. For example, treatments intended to increase intracellular glutamate sometimes failed to do so. In some cases (e.g. *hypotheses* 2 and *3*), resistance phenotypes were probably driven by downstream metabolites rather than the supplemented compounds themselves. Interventions based on suggested metabolite concentrations may still modulate upstream or downstream reactions, explaining the disconnect between predicted concentrations and observed phenotypes [48]. Incorporating fluxomics, dynamic modelling approaches and nonlinear assumptions regarding concentrations could help address this limitation, by capturing more complex dynamics rather than relying solely on predictions of static accumulations. Though this could come at the cost of hypotheses that are harder to interpret and communicate [49].

In this work, we identified the sub-components of the hypotheses in a general form, and how each component can be turned into a testable prediction. To do this systematically for future projects, one could possibly use LLMs, though this has not been tested here.

Downloaded from royalsocietypublishing.org/rsif/article-pdf/doi/10.1098/rsif.2026.0043/6150761/rsif20260043.pdf
by Chalmers University of Technology user
on 21 August 2026

By storing all experimental outcomes in a knowledge base, including falsified hypotheses such as those above, this framework supports future systems biology research and cross-study insights.

The hypotheses generated and evaluated in this work are simple, but the framework can easily be extended to handle other types of hypotheses, such as those based on observed metabolite accumulations, changes in metabolite excretion rates, environmental stressors and drug-to-metabolite connections (relations present in the initial hypothesis generation steps, see supplementary material, note S3). The description logic scheme was designed to extend to compound genotypes and phenotypes, including multiple environmental stressors. In theory, such patterns are also potentially reachable by providing the hypothesis forms to the ILP engine, though the increased complexity will lead to greater computational demands. However, during testing, the LLMs struggled somewhat with such 'combinatorial' hypotheses, mistaking directionality and misinterpreting the intersection of conditions. Future work could reevaluate this, either with newer versions of LLMs, or by enforcing the LLM agents to interact directly with the knowledge graph to constrain their reasoning. Introducing these extra degrees of freedom in hypothesis design would also complicate the orchestration of experiments, but we expect the structure of the overall framework would remain similar.

The modular design also allows incorporation of additional data types, like transcriptomics and proteomics, with only minimal adjustments to the hypothesis-generation pipeline [49].

Although no intellectual input is needed from humans beyond defining the framework's scope and limitations, the workflow is not fully automated. The remaining manual steps are in principle addressable: plate transfers could be eliminated through more bespoke robotics or by designing laboratory layouts that minimize physical handoffs, while safety reviews could be automated by, for example, integrating additional agents that query chemical safety databases and verify compliance with internal regulations—although this remains an area of active development [50]. Moreover, adding a more comprehensive orchestration layer could further coordinate task scheduling, making it truly autonomous.

The flexibility of LLMs allowed for rapid prototyping and development. Although rule-based agents could have replaced LLMs in some of the steps, developing and testing them takes time and expert resource. In addition, LLM agents facilitated connections between the slightly abstract activities of ILP and hypothesis formation, and the concrete aspects of experimentation. The LLM agents did contribute potential mechanisms for hypotheses, probably from the latent knowledge of systems biology they contain. These additions were under-used in this work, but could be a promising avenue for research.

Looking forward, scaling this automated workflow will require strict selection of hypotheses to maximize information gain per run. Integrating richer logical, decision-theoretic frameworks, econometric modelling and metrics from information theory will enable the system to better formulate and prioritize experiments, saving resources and contributing to scientific knowledge in a faster and more ethically sound manner [1,8,51]. More advanced or domain-specific LLMs, such as reasoning models more tailored for scientific discovery, could also allow more sophisticated hypotheses to be explored [52].

In this work, we demonstrate a flexible multi-agent framework for end-to-end hypothesis generation and experimental validation—combining LLMs, relational learning and automated laboratory workflows. Our approach leverages mass-spectrometry metabolomics—scalable, low-cost and automation friendly—to provide a rich information source to drive discovery, and deliver high-throughput, reproducible insights, firmly rooted in logic and community-adopted ontologies. We believe this platform lays the groundwork for a more reliable, machine-driven discovery process in systems biology, operating at unprecedented scale and rigour.

4. Methods

4.1. Frequent pattern mining

The relational database used for hypothesis generation makes use of the Saccharomyces Genome Database (SGD), ChEBI, STITCH-DB and the Yeast consensus model (Yeast9) as sources of relational data. All of the data used to create the database were downloaded using either AllianceGenome, STITCH-DB or the Metabolic Atlas (all retrieved 2024-12-09) [30,32–34,53,54].

The database consists of relations covering exhibited phenotypes of single-gene deletant mutants and ontological descriptors of genetic and chemical perturbations. Additionally, it contains relations regarding protein–chemical interactions, and metabolism connected to the genes of interest (see supplementary material, note S3, for a list of available relations). Only observables that are actionable within our experimental capacity were considered. This includes descriptors such as resistance to chemicals, compound accumulations and environmental sensitivities. Relations were filtered to only include facts derived from systematic mutation sets. The database is represented in Datalog, as its declarative structure enables clear and concise expression of biological relationships [37].

In order to extract biologically relevant patterns from the constructed database, relational learning was applied in the form of frequent pattern mining. These patterns were learned using the induce features mode in the ILP-engine aleph (v.5), utilizing a simplified version of the data-mining algorithm WARMR [28,29]. It is used to find frequent relational patterns using the deleted genes as positive examples, as in Brunnsåker *et al.* [49]. The 25 genes with the most recorded phenotypes were used to construct the bottom clause for the search (a bottom clause is the most specific clause that covers a given example(s), serving as a starting point for generalization). Complete parameters and a more detailed description of the search can be found in supplementary material, note S3.

Downloaded from royalsocietypublishing.org/rsif/article-pdf/doi/10.1098/rsif.2026.0043/6150761/rsif20260043.pdf
by Chalmers University of Technology user
on 21 August 2026

The resulting patterns were propositionalized into a tabular dataset of logic programs covering 4678 single-gene deletants from the Mülleder *et al.* metabolomics dataset [35]. When multiple propositionalized programs described the same examples, we prioritized the most detailed ones using a scored vocabulary.

4.2. Hypothesis weighting

In order to link logic programs with a biological observable, the propositionalized logic programs were used as independent variables for predicting amino acid accumulations, as measured in Mülleder *et al.* [35]. Associations were learned using an ElasticNet algorithm [55].

Note that this makes the implicit assumption that it is an accumulation that underlies the hypothesis, not the functional flux, and that the relationship between the logic program(s) and the accumulation is linear—a simplification that may not hold in all biological contexts but aids interpretability. Learned coefficients were then averaged across tenfold cross-validation. Each logic program was tied to the coefficients, providing a weight (and a testable association) for each logic program and amino acid present in the dataset. Logic programs with zero-valued coefficients are removed from the pool of hypotheses for each metabolite observable. Coefficients were scaled to unit ∞-norm. Amino acids with a negative cvR^2 were removed from further consideration. Example hypotheses can be seen in §2.

4.3. Semantic recording of hypotheses

The generated hypotheses have their origins in data downloaded from the SGD, which records data using terms taken from ontologies: phenotypes are recorded using terminology defined in the APO [30,31], and chemical species are defined using terms from Chemical Entities of Biological Interest (ChEBI) [34].

APO was designed to record hypotheses about genotype–phenotype relations, and statements generally have four components: (i) the type of mutant, (ii) the observable phenotype, (iii) a qualifier on this observable (e.g. 'increased') and (iv) the experiment type.[1] The implicit definition of qualifiers in APO is for a mutant with respect to a wild-type strain, and is therefore insufficient for recording hypotheses concerning arbitrary changes in genotype or phenotype. For this work, we extended APO to include qualifying relations between collections of genotype and phenotype changes, called *organism states*. We also created relations to specify phenotypes that relate to chemical species, defined externally in ChEBI. These extensions are necessary to express the types of hypotheses handled in this work, for example: 'cells with lower intracellular arginine levels exhibit increased resistance to caffeine'. Figure 5*A*,*B* shows how this specific hypothesis is modelled using our framework.

The basic form of hypotheses in systems biology is $A \rightarrow B$, and for a hypothesis to be testable, A should be an intervention achievable with the available laboratory equipment and consumables, and B is measurable. In our new terminology, A and B are *organism states*. The specific nature of the implication, $A \rightarrow B$, will vary in different domains, and as such the *implies* relation should be seen as a top-level term. In this work, given the prompt provided to the LLM, we can interpret *implies*(A, B) as

> intervention to achieve state A will lead to an observation of state B, where an absence of intervention A will not [and optionally where an alternative intervention A' (negative control) also will not lead to an observation of state B.]

For now, we delegate the assessment of the truth of this relation for each hypothesis to the statistical analyses described.

The extension to APO was not formally evaluated on expressiveness, coverage or any other desirable characteristics for the ontology. Rather, this was a 'pragmatic' extension which facilitated this work. The axioms introduced to extend APO were edited using Protégé (v5.6.4) [56]. We transform each hypothesis to OWL-DL statements using Python scripts. The resulting quads are stored in datasets on Apache Jena Fuseki server (v.4.5.0), building on the database system developed by Reder *et al.* [57]. An example set of description logic axioms that record the hypothesis stated above, and the resulting OWL-DL statements (serialized in TriG format), is provided in supplementary material, note S6.

4.4. Generative experimental design

A standardized context (background, protocols and lab constraints) was used to prompt an LLM-agent for an initial, detailed explanation of a generated hypothesis and a basic experimental plan to test it. A negative control is automatically selected from a list of amino acids that have zero-valued coefficients for the generated hypothesis. The contexts, explanation and prompts are saved in the graph database for each hypothesis. Templates can be found in the GitHub repository (see *Data accessibility*).

To ensure increased robustness, several hypothesis-plan variations are then generated with the same context, and a second LLM-agent picks the best one based on relevance to yeast physiology, safety and context fidelity. Note that in order to comply with safety regulations of the laboratory set-up in use, this final experimental plan is at this stage manually assessed for safety before being passed on.

A third LLM-agent auto-formalizes the chosen experimental plan into a JSON-file, which is then validated against the original context and hypothesis. If it fails any check, the JSON is regenerated up to three times. Experimental plans are generated and auto-formalized via the OpenAI API using the GPT-4o model (v.2024-08-06) through the official Python client library (openai v.1.44.1) [36].

[1]Note that experiment type will be absent in the case of a hypothesis not yet linked to any empirical data.

Downloaded from royalsocietypublishing.org/rsif/article-pdf/doi/10.1098/rsif.2026.0043/6150761/rsif20260043.pdf
by Chalmers University of Technology user
on 21 August 2026

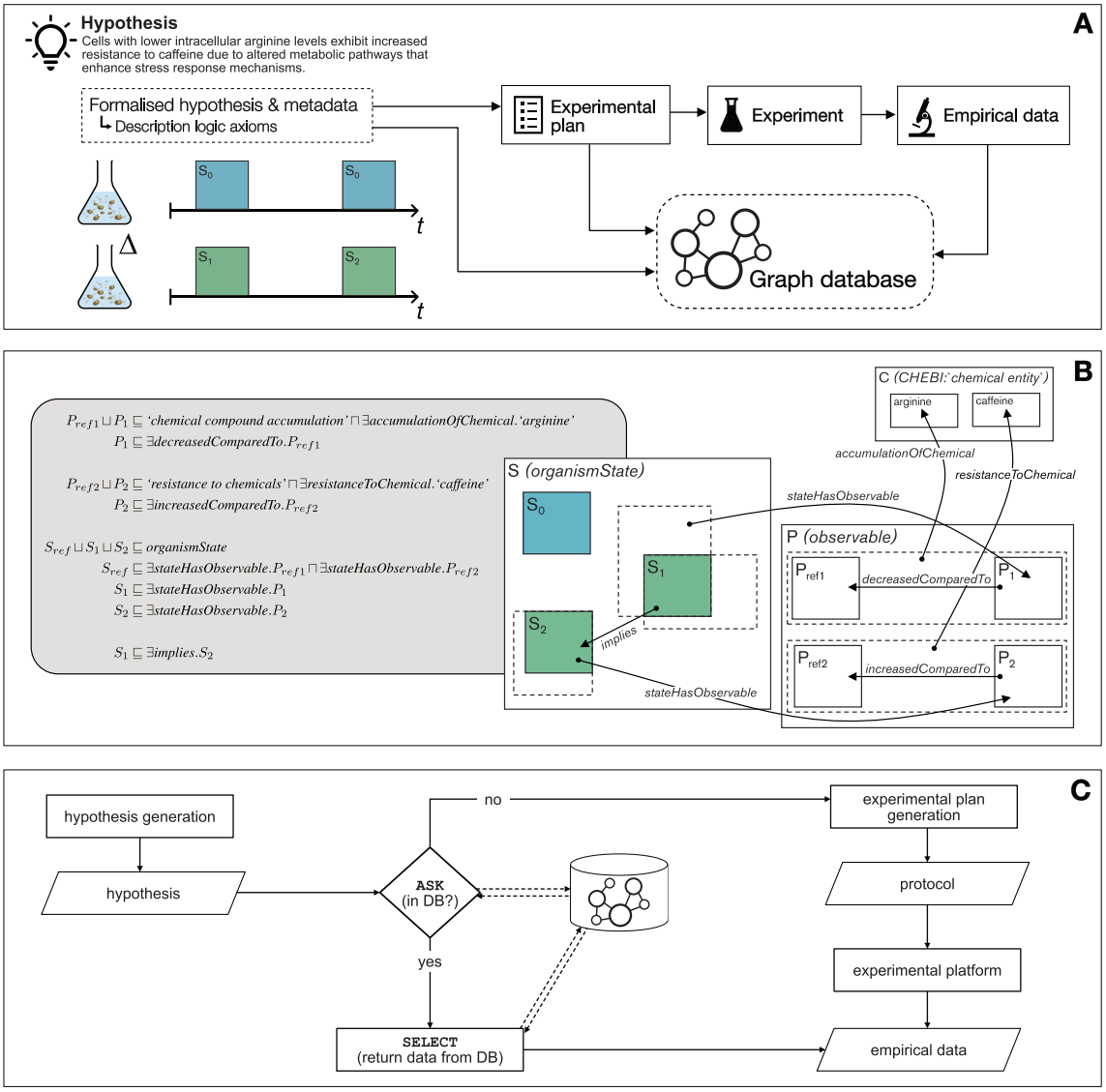

Figure 5. Graph database and description logics. (A) The hypothesis, in its original form as a logic program and also with the LLM output, is formalized into description logic axioms using an extended version of the APO; the generated experimental plans are stored using terms from the ontology for biomedical investigations (OBI); and empirical data are stored using terms from the genesis ontology. (B) *(L)* DL axioms expressing the same hypothesis shown in (A); *(R)* a visual representation of the concept inclusions expressed in the statements (with a subset of the relations shown, for clarity). (C) To make efficient use of data already collected, an `ASK` SPARQL query is executed on the graph database to identify if there are empirical data that can be relevant to the hypothesis; if yes, then a `SELECT` query is executed to return the data, else if no, the full automated experimental generation procedure is followed.

Using the formalized, machine-readable experimental protocol, the experimental variables are automatically identified, and a robust plate-layout is generated using PLAID [58]. Using the layout, a library of stock compounds and the experimental protocol, a dispensing scheme with aspiration and dispensing targets and volumes for use with Hamilton Venus Software (v.5.0.1), is generated. An example scheme can be seen in supplementary material, table S4.

A mass spectrometry run-list is automatically produced from the supplied protocol and user-defined parameters. The injection order is randomized to mitigate instrumental variation and systematic effects. Study quality control samples (sQCs) were placed systematically across the entire run-list, before and after biological samples, allowing for normalization and batch correction. Blocks of samples were separated by blank (extraction solvent) injections in order to assess analyte carry-over and background signal. An example run-list can be seen in supplementary material, table S5.

4.5. Automated sample preparation and cultivation

The generated plate-layout and dispensing schemes are executed on a Hamilton Microlab Star (using pre-prepared stocks of the available compounds) producing the microwell plate used for the experiment. Unless otherwise specified, experiments use minimal YNB medium (with ammonium sulphate, 2% glucose and 75 μg mL^{-1} ampicillin, without amino acids). The strain used to represent a reference genotype was Δ*ho* from the prototrophic strain collection constructed in Mülleder *et al.* [59]—owing to the inertness of the deletion in non-mating-type contexts.

Downloaded from royalsocietypublishing.org/rsif/article-pdf/doi/10.1098/rsif.2026.0043/6150761/rsif20260043.pdf
by Chalmers University of Technology user
on 21 August 2026

Each of the generated hypotheses were tested on separate microwell plates, allowing for up to 10 biological replicates per experimental condition, depending on the suggested intervention. Intervention and perturbation compounds present in the library are dissolved in sterile milliQ-water.

To run the cultivation and sampling, an Overlord (v.2.0) script (the orchestration software used to control the robotics and instrumentation in Eve) is generated programmatically using pre-defined parameters, such as cultivation time and plate-reader configurations [4,5,8]. For all of the experiments performed in this project, cultivation duration was set to 20 hours. See supplementary material, note S2, for a complete description of the cultivation protocols used.

4.6. Automated sample processing

Sampling for mass spectrometry analysis is automatically performed after experiment termination. Quenched samples are prepared by robotically passing the cultivation plate to a Bravo Automated Liquid Handling Platform and transferring the samples into −80°C methanol (99% purity). The ratio between sample and methanol is kept at 1:1 v/v. The quenching plate is then subsequently passed to a centrifuge, where it is centrifuged at 2240 × g for 10 minutes. The supernatant is then discarded and the pellet used for subsequent metabolite extraction.

Metabolite extraction is performed using a standardized, automated variation of the extraction protocol used in Brunnsåker *et al.* [5]. Extraction is robotically performed in a Hamilton Microlab Star, where 150 μL of 80% ethanol (preheated to boiling temperatures) is dispensed into the sample plate, which is placed on a Hamilton HeaterShaker add-on set at 100°C to avoid drops in temperature. The sample plate is shaken vigorously at 1600 rpm for 2 minutes to expedite resuspension of the quenched cells. The supernatant is transferred to a filter plate (0.2 μM, VWR, 97052-124) and separated using positive pressure filtering (set at 40 psi for 60 seconds).

The sQCs are prepared manually through overnight cultivation of Δ*ho* in reference conditions (30°C, YNB without amino acids). Samples are quenched in 1/1 (v/v) −80°C pure methanol, and subsequently, extracted in 80% ethanol heated to boiling temperatures through a water bath. The boiling ethanol is poured over the cells, and vortexed for 1 minute, and put back in the hot water bath for 4 minutes. The pellet is discarded, and the supernatant is stored in −80°C pending analysis. This is similar to protocols used in Brunnsåker *et al.* [5] and Reder *et al.* [26].

4.7. Growth metrics and hypothesis testing

Growth curves are constructed from raw optical density (OD) readings at 600 nm using a BMG Omega Polarstar. The growth curves are then automatically processed to remove outliers. Outlier detection is based on the median absolute deviation (MAD) of log-transformed AUC, total variation and final OD values. Wells exceeding a threshold of 3 MADs for any parameter are identified, excluded and reported. Curves are normalized to background by subtracting the average value of the three closest blanks on the microwell plate. The remaining curves are smoothed using locally estimated scatterplot smoothing (statsmodels, v.0.14.4). AUC is calculated using the composite trapezoidal rule on the smoothed curves (numpy, v.1.24.3). The effect of treatment and intervention on growth dynamics are automatically tested using log-transformed AUC as the response variable, and fitting a generalized linear model (statsmodels, v.0.14.4) that included terms for treatment, the concentration of the intervention, an interaction term and a separate flag for negative controls. To account for skewness and unequal variances, a non-parametric bootstrap (5000 resamples) is used to compute 95% confidence intervals and empirical p-values for each main and interacting effect. Example results can be seen in figure 3. The experimental variables, such as concentrations, are extracted from the auto-formalized protocols.

4.8. Automated metabolomics data acquisition and analysis

An internal collision cross-section (CCS) library was created using metabolites present in the Yeast9 Genome scale metabolic model as compounds of interest [32]. CCS values for relevant metabolites were acquired from AllCCS [60]. In case of missing experimental records for common adducts of amino acids (as these are the main targets of the generated hypotheses), predicted CCS values from AllCCS were used.

Mass spectrometry analysis is performed using an Agilent RapidFire-365 and an Agilent 6560 DTIMS-QToF system. The instrument data acquisition and peak picking is done automatically via AutonoMS, using a run-list generated from the hypothesis metadata and experimental protocol as the input [26]. Peaks are identified based on both m/z (mass-to-charge ratio) and CCS (Å^2) [61]. The cartridge in use for all experiments was hydrophilic interaction liquid chromatography. Details regarding mobile phases, method parameters, acquisition settings, demultiplexing parameters and peak picking parameters can be seen in supplementary material, notes S4 and S5.

The AutonoMS-output (a tabular representation of identified metabolites with an associated peak area) is initially curated using sensor outputs from the RapidFire system, removing injections with insufficient injected sample (greater or equal to 600 ms). Peaks with an average area below the included blanks (extraction solvent), along with peaks with more than 1/3 missing detections across study samples, are excluded from further analysis. In case of several identified adducts per metabolite, the one with the lowest relative standard deviation in the sQCs is used. Samples are normalized using probabilistic quotient normalization, utilizing the sQCs as the reference spectra [62]. Missing values are imputed with multivariate iterative imputation using RandomForest [63]. Samples outside of the 95% Hotelling T2 ellipse (using PCA-derived principal components) and high

Downloaded from royalsocietypublishing.org/rsif/article-pdf/doi/10.1098/rsif.2026.0043/6150761/rsif20260043.pdf
by Chalmers University of Technology user
on 21 August 2026

unexplained residual variances (beyond 99th percentile) are treated as outliers, similar to quality control recommendations by González-Domínguez *et al.* [64].

To estimate the validity of the LLM-suggested interventions (confirming intracellular accumulation), pairwise comparisons of the target metabolites across untreated conditions is performed using a Wilcoxon rank-sum test (scipy, v.1.15.1). Metabolic separability is classified using (5-component) PLS-DA, treating each experimental group as a separate class (mixOmics, v.6.22.0) [65]. Classification performance is evaluated using the macro F_1 score.

Metabolites associated with resistance to treatment are determined with ElasticNetCV, using repeated cross-validation (scikit-learn v.1.2.2) [66]. AUC of treated samples are used as the dependent variable. Example results can be seen in figure 3.

As part of the evaluation, metabolomics data were collected for each individual cultivation in the study, producing 976 separate biological injections, including 336 sQCs injections. The metabolomics dataset spans 52 unique conditions, with 21 different nutrient compositions, and 8 treatments, including controls. Each condition has a minimum of 10 biological replicates and associated growth dynamics.

4.9. Linking hypotheses and experimental plans to data

We also record the experimental plans, and the resulting empirical data from completed experiments, using controlled vocabularies in a graph database. These plans and data are linked to the hypothesis through a combination of terms from existing ontologies and new terms. In the case of experimental protocols, the terms are taken primarily from the OBI [67]; and terms for growth and metabolomics are taken primarily from the Genesis ontology [57].

The JSON files containing the experimental protocols, and the tabular files containing growth and metabolomics data, undergo pre-processing using Python and are then mapped to relevant terms from existing ontologies using RMLMapper (v7.3.3); to extend this work to new experimental settings, including alternative hypothesis forms and experimental protocols, new RMLMapper files and SPARQL queries would need to be written. The mapping files, and SPARQL queries, can be found in the GitHub repository (see *Data accessibility*).

Ethics. This work did not require ethical approval from a human subject or animal welfare committee.

Data accessibility. Raw mass spectrometry data have been deposited at MassIVE under accession MSV000099724 (https://doi.org/doi:10.25345/C5HT2GR2T). All data, code and associated metadata can be downloaded from Zenodo at [68]. Code, supplementary information and experimental results can also be found on GitHub at https://github.com/DanielBrunnsaker/GenExp.

Supplementary material is available online [69].

Declaration of AI use. During the preparation of this work, the authors used AI in order to check for grammar errors and improve the clarity of the writing.

Authors' contributions. D.B.: conceptualization, data curation, formal analysis, investigation, methodology, project administration, software, validation, visualization, writing—original draft, writing—review and editing; A.H.G.: conceptualization, formal analysis, methodology, software, validation, visualization, writing—original draft, writing—review and editing; P.N.: investigation, methodology, writing—review and editing; E.Y.B.: investigation, writing—review and editing; F.K.: software, writing—review and editing; I.T.: funding acquisition, project administration, supervision, writing—review and editing; R.K.: conceptualization, funding acquisition, project administration, resources, supervision, writing—original draft, writing—review and editing.

All authors gave final approval for publication and agreed to be held accountable for the work performed therein.

Conflict of interest declaration. We declare we have no competing interests.

Funding. This work was partially supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation, the UK Engineering and Physical Sciences Research Council (grant no. EP/R022925/2, EP/W004801/1, EP/X032418/1), the Chalmers AI Research Centre (CHAIR), and the Swedish Research Council Formas (grant no. 2020-01690).

Acknowledgements. The authors thank the Ralser Lab at Charité–Universitätsmedizin Berlin for providing the yeast strains used in this study.

References

1. Musslick S *et al.* 2025 Automating the practice of science: opportunities, challenges, and implications. *Proc. Natl Acad. Sci. USA* **122**, e2401238121. (doi:10.1073/pnas.2401238121)
2. Gower AH, Korovin K, Brunnsåker D, Kronström F, Reder GK, Tiukova IA, Reiserer RS, Wikswo JP, King RD. 2024 The use of AI-robotic systems for scientific discovery. *arXiv*. (doi:10.48550/arXiv.2406.17835)
3. Zenil H *et al.* 2023 The future of fundamental science led by generative closed-loop artificial intelligence. *arXiv*. (doi:10.48550/arXiv.2307.07522)
4. Coutant A *et al.* 2019 Closed-loop cycles of experiment design, execution, and learning accelerate systems biology model development in yeast. *Proc. Natl Acad. Sci. USA* **116**, 18142–18147. (doi:10.1073/pnas.1900548116)
5. Brunnsåker D, Reder GK, Soni NK, Savolainen OI, Gower AH, Tiukova IA, King RD. 2023 High-throughput metabolomics for the design and validation of a diauxic shift model. *npj Syst. Biol. Appl.* **9**, 1–9. (doi:10.1038/s41540-023-00274-9)
6. King RD, Whelan KE, Jones FM, Reiser PGK, Bryant CH, Muggleton SH, Kell DB, Oliver SG. 2004 Functional genomic hypothesis generation and experimentation by a robot scientist. *Nature* **427**, 247–252. (doi:10.1038/nature02236)
7. King RD *et al.* 2009 The automation of science. *Science* **324**, 85–89. (doi:10.1126/science.1165620)
8. Williams K *et al.* 2015 Cheaper faster drug development validated by the repositioning of drugs against neglected tropical diseases. *J. R. Soc. Interface* **12**, 20141289. (doi:10.1098/rsif.2014.1289)

Downloaded from royalsocietypublishing.org/rsif/article-pdf/doi/10.1098/rsif.2026.0043/6150761/rsif20260043.pdf
by Chalmers University of Technology user
on 21 August 2026

9. Zhang P *et al*. 2024 Automating life science labs at the single-cell level through precise ultrasonic liquid sample ejection: PULSE. *Microsyst. Nanoeng.* **10**, 1–16. (doi:10.1038/s41378-024-00798-y)
10. Szymanski NJ *et al*. 2023 An autonomous laboratory for the accelerated synthesis of novel materials. *Nature* **624**, 86–91. (doi:10.1038/s41586-023-06734-w)
11. Microsoft Research AI4Science, Microsoft Azure Quantum. 2023 The impact of large language models on scientific discovery: a preliminary study using GPT-4. *arXiv*. (doi:10.48550/arXiv.2311.07361)
12. Gottweis J *et al*. 2025 Towards an AI co-scientist. *arXiv*. (doi:10.48550/arXiv.2502.18864)
13. Boiko DA, MacKnight R, Kline B, Gomes G. 2023 Autonomous chemical research with large language models. *Nature* **624**, 570–578. (doi:10.1038/s41586-023-06792-0)
14. Abdel-Rehim A *et al*. 2025 Scientific hypothesis generation by large language models: laboratory validation in breast cancer treatment. *J. R. Soc. Interface* **22**, 20240674. (doi:10.52843/cassyni.9gxrjf)
15. Reder GK, Collins C, Rehim AA, Soldatova L, King RD. 2025 LLM-retrieval based scientific knowledge grounding. In *Proc. 2nd Int. Workshop on Natural Scientific Language Processing and Research Knowledge Graphs (NSLP 2025)*.
16. Lei W, Fuster-Barceló C, Reder G, Muñoz-Barrutia A, Ouyang W. 2024 BioImage.IO Chatbot: a community-driven AI assistant for integrative computational bioimaging. *Nat. Methods* **21**, 1368–1370. (doi:10.1038/s41592-024-02370-y)
17. Ghareeb AE *et al*. 2025 Robin: a multi-agent system for automating scientific discovery. *arXiv*. (doi:10.48550/arXiv.2505.13400)
18. Lobentanzer S *et al*. 2025 A platform for the biomedical application of large language models. *Nat. Biotechnol.* **43**, 166–169. (doi:10.1038/s41587-024-02534-3)
19. Lu C, Lu C, Lange RT, Foerster J, Clune J, Ha D. 2024 The AI scientist: towards fully automated open-ended scientific discovery. *arXiv*. (doi:10.48550/arXiv.2408.06292)
20. Ghafarollahi A, Buehler MJ. 2025 SciAgents: automating scientific discovery through bioinspired multi-agent intelligent graph reasoning. *Adv. Mater.* **37**, 2413523. (doi:10.1002/adma.202413523)
21. Ghafarollahi A, Buehler MJ. 2024 ProtAgents: protein discovery via large language model multi-agent collaborations combining physics and machine learning. *Digit. Discov.* **3**, 1389–1409. (doi:10.1039/D4DD00013G)
22. Ghafarollahi A, Buehler MJ. 2025 Sparks: multi-agent artificial intelligence model discovers protein design principles. *arXiv*. (doi:10.48550/arXiv.2504.19017)
23. Kitano H. 2002 Systems biology: a brief overview. *Science* **295**, 1662–1664. (doi:10.1126/science.1069492)
24. Buehler MJ. 2024 Accelerating scientific discovery with generative knowledge extraction, graph-based representation, and multimodal intelligent graph reasoning. *Mach. Learn. Sci. Technol.* **5**, 035083. (doi:10.1088/2632-2153/ad7228)
25. Binz M *et al*. 2025 How should the advancement of large language models affect the practice of science? *Proc. Natl Acad. Sci. USA* **122**, e2401227121. (doi:10.1073/pnas.2401227121)
26. Reder GK *et al*. 2024 AutonoMS: automated ion mobility metabolomic fingerprinting. *J. Am. Soc. Mass. Spectrom.* **35**, 542–550. (doi:10.1021/jasms.3c00396)
27. King RD, Liakata M, Lu C, Oliver SG, Soldatova LN. 2011 On the formalization and reuse of scientific research. *J. R. Soc. Interface* **8**, 1440–1448. (doi:10.1098/rsif.2011.0029)
28. Srinivasan A. 2007 *The Aleph manual*. https://www.cs.ox.ac.uk/activities/programinduction/Aleph/aleph.html.
29. King RD, Srinivasan A, Dehaspe L. 2001 Warmr: a data mining tool for chemical data. *J. Comput. Aided Mol. Des.* **15**, 173–181. (doi:10.1023/A:1008171016861)
30. Cherry JM *et al*. 2012 Saccharomyces genome database: the genomics resource of budding yeast. *Nucleic Acids Res.* **40**, D700–D705. (doi:10.1093/nar/gkr1029)
31. Engel SR, Aleksander S, Nash RS, Wong ED, Weng S, Miyasato SR, Sherlock G, Cherry JM. 2025 Saccharomyces genome database: advances in genome annotation, expanded biochemical pathways, and other key enhancements. *Genetics* **229**, iyae185. (doi:10.1093/genetics/iyae185)
32. Zhang C *et al*. 2024 Yeast9: a consensus genome-scale metabolic model for *S. cerevisiae* curated by the community. *Mol. Syst. Biol.* **20**, 1134–1150. (doi:10.1038/s44320-024-00060-7)
33. Szklarczyk D, Santos A, von Mering C, Jensen LJ, Bork P, Kuhn M. 2016 STITCH 5: augmenting protein-chemical interaction networks with tissue and affinity data. *Nucleic Acids Res.* **44**, D380–D384. (doi:10.1093/nar/gkv1277)
34. Hastings J *et al*. 2016 ChEBI in 2016: improved services and an expanding collection of metabolites. *Nucleic Acids Res.* **44**, 1214–1219. (doi:10.1093/nar/gkv1031)
35. Mülleder M, Calvani E, Alam MT, Wang RK, Eckerstorfer F, Zelezniak A, Ralser M. 2016 Functional metabolomics describes the yeast biosynthetic regulome. *Cell* **167**, 553–565. (doi:10.1016/j.cell.2016.09.007)
36. OpenAI *et al*. 2024 GPT-4 technical report. *arXiv*. (doi:10.48550/arXiv.2303.08774)
37. Dehaspe L, Toivonen H. 1999 Discovery of frequent DATALOG patterns. *Data Min. Knowl. Discov.* **3**, 7–36. (doi:10.1023/A:1009863704807)
38. Muggleton S, De Raedt L, Poole D, Bratko I, Flach P, Inoue K, Srinivasan A. 2012 ILP turns 20. *Mach. Learn.* **86**, 3–23. (doi:10.1007/s10994-011-5259-2)
39. Kramer S, Lavrač N, Flach P. 2001 Propositionalization approaches to relational data mining. In *Relational data mining* (eds S Džeroski, N Lavrač), pp. 262–291. Berlin, Heidelberg: Springer. (doi:10.1007/978-3-662-04599-2_11)
40. Roper K, Abdel-Rehim A, Hubbard S, Carpenter M, Rzhetsky A, Soldatova L, King RD. 2022 Testing the reproducibility and robustness of the cancer biology literature by robot. *J. R. Soc. Interface* **19**, 20210821. (doi:10.1098/rsif.2021.0821)
41. Munafó MR *et al*. 2017 A manifesto for reproducible science. *Nat. Hum. Behav.* **1**, 0021. (doi:10.1038/s41562-016-0021)
42. Ioannidis JPA. 2005 Why most published research findings are false. *PLoS Med.* **2**, e124. (doi:10.1371/journal.pmed.0020124)
43. Kanehisa M, Furumichi M, Sato Y, Kawashima M, Ishiguro-Watanabe M. 2023 KEGG for taxonomy-based analysis of pathways and genomes. *Nucleic Acids Res.* **51**, 587–592. (doi:10.1093/nar/gkac963)
44. Dikicioglu D, Dereli Eke E, Eraslan S, Oliver SG, Kirdar B. 2018 *Saccharomyces cerevisiae* adapted to grow in the presence of low-dose rapamycin exhibit altered amino acid metabolism. *Cell Commun. Signal.* **16**, 85. (doi:10.1186/s12964-018-0298-y)
45. Reinke A, Chen JCY, Aronova S, Powers T. 2006 Caffeine targets TOR complex I and provides evidence for a regulatory link between the FRB and kinase domains of Tor1p. *J. Biol. Chem.* **281**, 31616–31626. (doi:10.1016/S0021-9258(19)84075-9)
46. Erasmus DJ, van der Merwe GK, van Vuuren HJ. 2003 Genome-wide expression analyses: metabolic adaptation of *Saccharomyces cerevisiae* to high sugar stress. *FEMS Yeast Res.* **3**, 375–399. (doi:10.1016/S1567-1356(02)00203-9)
47. Zeng L, Si Z, Zhao X, Feng P, Huang J, Long X, Yi Y. 2022 Metabolome analysis of the response and tolerance mechanisms of *Saccharomyces cerevisiae* to formic acid stress. *Int. J. Biochem. Cell Biol.* **148**, 106236. (doi:10.1016/j.biocel.2022.106236)

Downloaded from royalsocietypublishing.org/rsif/article-pdf/doi/10.1098/rsif.2026.0043/6150761/rsif20260043.pdf
by Chalmers University of Technology user
on 21 August 2026

48. Emwas AH, Szczepski K, Al-Younis I, Lachowicz JI, Jaremko M. 2022 Fluxomics—new metabolomics approaches to monitor metabolic pathways. *Front. Pharmacol.* **13**, 805782. (doi:10.3389/fphar.2022.805782)
49. Brunnsåker D, Kronström F, Tiukova IA, King RD. 2024 Interpreting protein abundance in *Saccharomyces cerevisiae* through relational learning. *Bioinformatics* **40**, btae050. (doi:10.1093/bioinformatics/btae050)
50. Tang X *et al*. 2025 Risks of AI scientists: prioritizing safeguarding over autonomy. *Nat. Commun.* **16**, 8317. (doi:10.1038/s41467-025-63913-1)
51. King RD, Scassa T, Kramer S, Kitano H. 2024 Stockholm declaration on AI ethics: why others should sign. *Nature* **626**, 716–716. (doi:10.1038/d41586-024-00517-7)
52. Buehler MJ. 2025 In situ graph reasoning and knowledge expansion using graph-preflexor. *Adv. Intell. Discov.* **1**, e202500006. (doi:10.1002/aidi.202500006)
53. The Alliance of Genome Resources Consortium. 2024 Updates to the alliance of genome resources central infrastructure. *Genetics* **227**, iyae049. (doi:10.1534/genetics.119.302523)
54. Li F, Chen Y, Anton M, Nielsen J. 2023 GotEnzymes: an extensive database of enzyme parameter predictions. *Nucleic Acids Res.* **51**, 583–586. (doi:10.1093/nar/gkac831)
55. Zou H, Hastie T. 2005 Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **67**, 301–320. (doi:10.1111/j.1467-9868.2005.00503.x)
56. Musen MA. 2015 The protégé project: a look back and a look forward. *AI Matters* **1**, 4–12. (doi:10.1145/2757001.2757003)
57. Reder GK, Gower AH, Kronström F, Halle R, Mahamuni V, Patel A, Hayatnagarkar H, Soldatova LN, King RD. 2023 Genesis-DB: a database for autonomous laboratory systems. *Bioinf. Adv.* **3**, vbad102. (doi:10.1093/bioadv/vbad102)
58. Francisco Rodríguez MA, Carreras Puigvert J, Spjuth O. 2023 Designing microplate layouts using artificial intelligence. *Artif. Intell. Life Sci.* **3**, 100073. (doi:10.1016/j.ailsci.2023.100073)
59. Mülleder M, Capuano F, Pir P, Christen S, Sauer U, Oliver SG, Ralser M. 2012 A prototrophic deletion mutant collection for yeast metabolomics and systems biology. *Nat. Biotechnol.* **30**, 1176–1178.
60. Zhang H, Luo M, Wang H, Ren F, Yin Y, Zhu ZJ. 2023 AllCCS2: curation of ion mobility collision cross-section atlas for small molecules using comprehensive molecular representations. *Anal. Chem.* **95**, 13913–13921.(doi:10.1021/acs.analchem.3c02267)
61. Adams KJ *et al*. 2020 Skyline for small molecules: a unifying software package for quantitative metabolomics. *J. Proteome Res.* **19**, 1447–1458. (doi:10.1021/acs.jproteome.9b00640)
62. Dieterle F, Ross A, Schlotterbeck G, Senn H. 2006 Probabilistic quotient normalization as robust method to account for dilution of complex biological mixtures. Application in 1h NMR metabonomics. *Anal. Chem.* **78**, 4281–4290. (doi:10.1021/ac051632c)
63. Wei R, Wang J, Su M, Jia E, Chen S, Chen T, Ni Y. 2018 Missing value imputation approach for mass spectrometry-based metabolomics data. *Sci. Rep.* **8**, 663. (doi:10.1038/s41598-017-19120-0)
64. González-Domínguez Á, Estanyol-Torres N, Brunius C, Landberg R, González-Domínguez R. 2024 QComics: recommendations and guidelines for robust, easily implementable and reportable quality control of metabolomics data. *Anal. Chem.* **96**, 1064–1072. (doi:10.1021/acs.analchem.3c03660)
65. Rohart F, Gautier B, Singh A, Cao KAL. 2017 MmixOmics: an R package for 'omics feature selection and multiple data integration. *PLoS Comput. Biol.* **13**, e1005752. (doi:10.1371/journal.pcbi.1005752)
66. Pedregosa F *et al*. 2011 Scikit-learn: machine learning in python. *J. Mach. Learn. Res.* **12**, 2825–2830.
67. Bandrowski A *et al*. 2016 The ontology for biomedical investigations. *PLoS One* **11**, e0154556. (doi:10.1371/journal.pone.0154556)
68. Brunnsåker, D. 2026. Agentic AI Integrated with Scientific Knowledge: Laboratory Validation in Systems Biology. Zenodo. (doi:10.5281/zenodo.19403642)
69. Brunnsåker D, Gower AH, Naval P, Bjurström EY, Kronström F, Tiukova I, King R. 2026 Supplementary material from: Agentic AI integrated with scientific knowledge: laboratory validation in systems biology. FigShare. (doi:10.6084/m9.figshare.c.8495771)

Downloaded from royalsocietypublishing.org/rsif/article-pdf/doi/10.1098/rsif.2026.0043/6150761/rsif20260043.pdf
by Chalmers University of Technology user
on 21 August 2026