# Self-driven Biological Discovery through Automated Hypothesis Generation and Experimental Validation

**Daniel Brunnsåker[a,✉], Alexander H. Gower[a], Prajakta Naval[b], Erik Y. Bjurström[b], Filip Kronström[a], Ievgeniia A. Tiukova[b,c], and Ross D. King[a,d,e]**

[a]Department of Computer Science and Engineering, Chalmers University of Technology and University of Gothenburg, SE-41296 Gothenburg, Sweden
[b]Department of Life Sciences, Chalmers University of Technology, SE-41296 Gothenburg, Sweden
[c]Division of Industrial Biotechnology, KTH Royal Institute of Technology, SE-10691 Stockholm, Sweden
[d]Department of Chemical Engineering and Biotechnology, University of Cambridge, Cambridge CB3 0AS, United Kingdom
[e]The Alan Turing Institute, London NW1 2DB, United Kingdom

**Modern scientific challenges demand faster, more reliable research workflows. A primary goal for applying artificial intelligence in science is the automation of research, offering faster discoveries and new insights. Large Language Models (LLMs) are revolutionizing AI, achieving impressive results on tasks previously requiring human intelligence. By using explicit mathematical models and logic to generate hypotheses, LLMs can be provided a logical scaffold to reason around, reducing hallucinations and enhancing the reliability of its outputs.**

**Here we demonstrate a method that uses LLMs to automate parts of experimental design, using relational learning-derived hypotheses and physical laboratory constraints. We integrate this methodology with an automated laboratory cell and metabolomics platform, presenting a flexible and efficient approach to automated scientific discovery in a user-defined hypothesis space. We evaluate the methodology on multiple interaction experiments in *Saccharomyces cerevisiae*, revealing, among other findings, antagonistic effects between glutamate and spermine.**

**We record the hypotheses, experimental designs, and empirical data in a graph database, using controlled vocabularies. We extend existing ontologies and present a novel way of representing scientific hypotheses using description logics.**

**We also offer a proof-of-concept demonstrating how metabolomics data can be used to further refine generated hypotheses.**

**Laboratory Automation | Systems Biology | Machine Learning | Inductive Logic Programming | Automation of Science | Large Language Models**

**Correspondence:** ***danbru@chalmers.se***

## Introduction

Scientific discovery has historically relied on human scientists generating hypotheses, which then undergo experimental validation and analysis. However, automation and artificial intelligence are rapidly transforming how humans perform science. These technologies allow the analysis and understanding of data that would otherwise be impossibly complex and, when integrated with robotics and automation, can enable the full automation of the scientific process (1). This approach has previously been applied in fields such as functional genomics, materials development, and chemistry (2–7).

Systems biology poses significant challenges to the way humans have traditionally conducted scientific research. Even comparatively simple model organisms, such as *Saccharomyces cerevisiae* (baker's yeast) and *Escherichia coli*, exhibit complexity far exceeding that which humans can reliably interpret or experimentally validate within a reasonable timeframe (3). To achieve reliable and scalable scientific discovery under these conditions, the deployment of integrated automated experimental and computational methods becomes essential (3, 8).

Recently, with the advent of extreme model scaling, large language models (LLMs) are revolutionizing science by achieving impressive results in tasks previously requiring human intelligence and interaction, including scientific discovery (9). Prior to this, automated laboratory systems were primarily model-driven, relying on hypotheses generated in a tightly controlled experimental space, often based on precise logic, and with empirical data collection limited to few modalities (2, 3, 10, 11). Consequently, the scientific scope was restricted not only by the narrow biological coverage afforded by limited modalities, but also by the rigidity of manually encoded logic. Unlike modern LLM-augmented approaches, these systems offered little adaptability in problem formulation or experimental scope; a serious drawback for tackling the complexity of systems biology (12). However, care must be taken when using LLMs in science. Hallucinations and inconsistent outputs are still significant challenges, even when augmented with sophisticated reasoning frameworks (13). Despite the increasingly widespread use of LLMs in scientific discovery (4, 7), this issue is seldom properly addressed. Additionally, despite growing reliance on LLMs for hypothesis generation, these hypotheses are rarely subjected to rigorous evaluation, and the underlying data often remain siloed within individual labs rather than being shared with the scientific community, undermining collective progress. Thus, there is a need for methods that combine the rigor of logic-driven approaches with the flexibility of LLMs, underpinned by transparent, precise data-sharing protocols.

In this work, we introduce a hybrid approach combining LLMs with the precision of relational learning, inductive logic programming (ILP), and rigorous ontologies to gen-

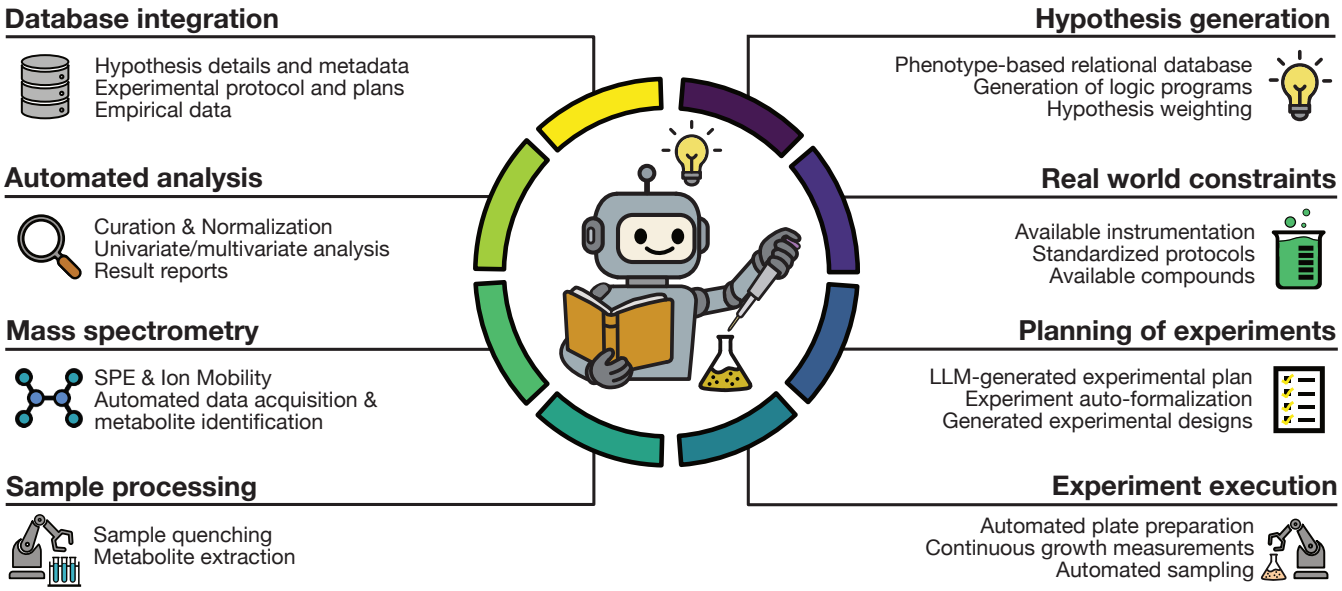


**Fig. 1.** Automated biological discovery. An automated experimental pipeline combines robotic liquid handlers, the EVE platform for culture and sampling, ion-mobility based metabolomics to iteratively generate and test cell-biology hypotheses. Hypotheses are derived by mining yeast phenotype and metabolism data from community databases using inductive logic programming. Interventions to test the hypothesis are then designed by an LLM, and subsequently auto-formalized into a machine readable protocol and executed. Each run yields complete growth and metabolite profiles, which are curated, analyzed and summarized. All results, intermediate data and metadata are recorded in a graph database.

erate hypotheses grounded in community knowledge, empirical data, and relational databases—mitigating hallucinations while preserving the flexibility of language-driven problem formulation. Agents then systematically filter and prioritize these hypotheses based on experimental results, before designing and executing tests using automated laboratory infrastructure (such as robotic liquid handlers, automated cultivation systems, and mass-spectrometry–based metabolomics) (10, 14). After execution, hypothesis details, experimental metadata, runtime logs, and empirical data are integrated into a structured graph database, ensuring efficient data reuse, transparency, and reproducibility, facilitating further research by clearly linking each hypothesis with its supporting evidence.

We apply and validate this framework across multiple interaction experiments in *S. cerevisiae*, demonstrating its ability to enable high-throughput, reproducible and scalable scientific discovery in a complex biological system.

## Results

### Automatic Systems Biology Research.

We established an automated experimental and computational pipeline by integrating robotic liquid handling stations (a Hamilton Microlab Star and an Agilent Bravo), an automated laboratory cell (Eve) for cell culture and sampling, and mass spectrometry–based metabolomics (an Agilent 6560 IM-MS coupled to a 365-RapidFire) along with computational tools in R and Python (10, 14–16). This pipeline is designed to automatically generate and experimentally validate hypotheses about cell biology. Testable hypotheses are generated by (i) data-mining semantically meaningful facts about *S. cerevisiae* phenotype and metabolism—extracted from genome-scale metabolic models and the Saccharomyces Genome Database (SGD)—(ii) applying inductive logic programming to generate the descriptive basis, and (iii) applying supervised learning on existing metabolomics data to learn an actionable implication (17–21). It then systematically filters and prioritizes the hypotheses. An LLM is used to design the experimental interventions that test those hypotheses and auto-formalize them into a machine readable format (22). For each hypothesis tested, time-series growth data and endpoint metabolic profiles are generated for all biological samples. Following each experimental run, the generated data is curated for outliers, processed, analysed, and tested, creating a user-readable summary of each experiment.

Each step of the hypothesis generation and all experimental details are recorded, including queries to and responses from the LLM. Experimental data and metadata are saved in a graph database, using controlled vocabularies, utilising an extended version of the Ascomycete Phenotype Onotology (APO) (17). The hypotheses and their implications are represented using description logics. This enables efficient reuse of data in the case an experiment relevant to a generated hypothesis has already been executed, which was the case for one hypothesis which we demonstrate below.

### Generated hypothesis space.

To generate testable, data-driven hypotheses, we constructed a curated Datalog database of ~60,000 yeast phenotype, physiology, and metabolic relations. ILP-based pattern mining on this database yielded 735 semantically distinct logic programs, 390 of which mapped uniquely to positive examples (gene deletion strains) from the metabolomics dataset of Mülleder et al. (21). When multiple propositionalised programs described the same examples, we prioritized the

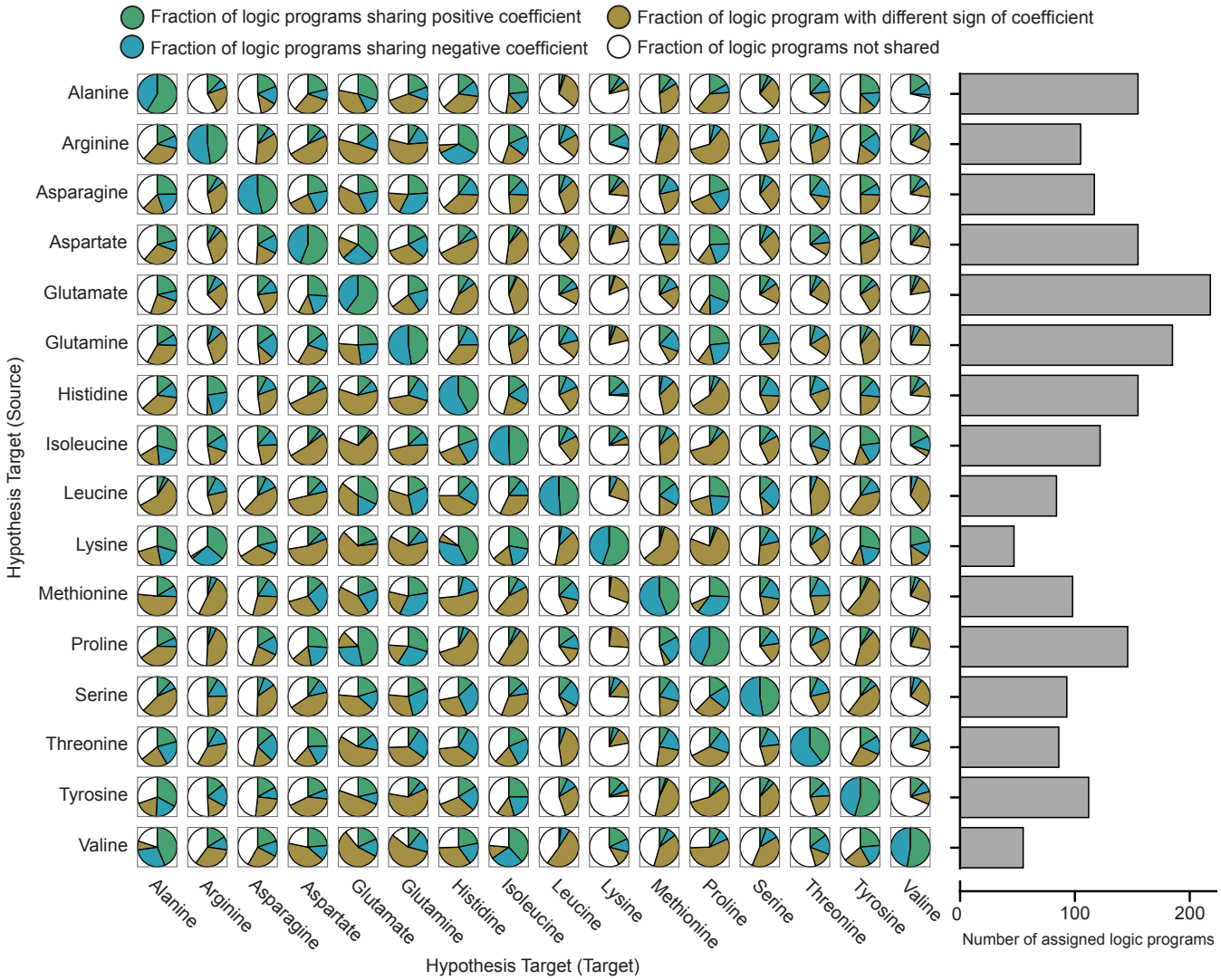


**Fig. 2.** Asymmetric hypothesis-space overlap between pairs of amino acids. Each pie chart shows—as a fraction of the total number of logic programs that are assigned to the amino acid on the vertical axis—the fraction of logic programs that also are assigned to the amino acid on the horizontal axis. For those that are shared, the blue, green, and yellow segments represent respectively positive, negative, and differing signs of coefficient in the linear regression model for prediction of abundance. The bar chart shows the total number of assigned logic programs for each amino acid on the vertical axis.

most detailed ones using a scored vocabulary. We then fit the propositionalised programs to metabolite measurements via regularised linear regression, producing 1933 independent hypotheses across 16 metabolic targets (see Fig. 2 and Sup. Table 1). An example logic program and learned association can be seen below (also see *hypothesis 6* in table 1).

$$
\begin{aligned}
&- \mathsf{Arginine,\ Cells(A)} :=\\
&\quad \mathsf{ExhibitsPhenotype(A,\ Decreased\ Resistance,\ B,\ C),}\\
&\quad \mathsf{CompoundName(B,\ Lithium(+1)),}\\
&\quad \mathsf{Condition(C,\ Treatment:\ 0.1M\ Lithium\ Chloride).}
\end{aligned}
$$

In this context, this program can be interpreted as: cells (A) which have decreased resistance to lithium (B) when treated with 0.1M of lithium chloride (C) are associated with lower levels of intracellular arginine.

### 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). Each investigation began by selecting a metabolic target (an amino acid) and hypothesis-generation parameters (e.g., sampling temperature, context). For each target, the framework selected the highest-weighted hypothesis, automatically chose an appropriate negative control (based on learned associations), and performed the corresponding experiment, yielding time-series growth curves, endpoint metabolic profiles, effect sizes, and statistical evaluations (see Methods). Experiments were run in up to 10 biological replicates per condition. We demonstrate this with five metabolic targets: L-glutamate, L-arginine, L-proline, L-glutamine, and L-lysine. An overview of all investigations is presented in Table 1, with full statistical details in Sup. Note 6 and Sup. Table 2.

**Table 1.** Summary of those generated hypotheses that were selected for experimentation. 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. "-", "+" indicates the observed or predicted sign of the change. "~" denotes an inconclusive change in accumulation. **Iterative hypothesis generated through data generated in hypothesis 4. †Experiment tests inverse hypothesis, i.e. nutrient supplementation rather than reduction, and reversed effect.*

| | | Change in accumulation | | Effect on resistance | | |
|---|---|---|---|---|---|---|
| | Generated Hypothesis | 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* | ↑ Lysine ⟹ ↓ Hyperosmotic stress resistance (at 30% (w/v) Sucrose) | + | + | + | – | Yes |
| *Hypothesis 5* | ↓ Glutamine ⟹ ↓ Acetic acid resistance† | + | + | - | + | No |
| *Hypothesis 6* | ↓ Arginine ⟹ ↓ Lithium(+1) resistance (at 0.1M Lithium chloride)† | + | + | ~ | + | No |
| *Hypothesis 7* | ↑ Proline ⟹ ↑ Lactic acid resistance | + | + | ~ | + | No |
| *Hypothesis 8* | ↑ Aminoadipate ⟹ ↑ Formic acid resistance (at 10mM Formic acid)* | - | + | + | + | Yes |

*Hypothesis 1* (regarding the antagonistic interaction between L-arginine and caffeine, shown in Fig. 3A) was consistent with acquired empirical data. With addition of L-arginine showing a modest, but significant, reduction in AUC (Area under growth curve) by itself ($-2.5\%$ per mM, $p < 0.05$), whereas caffeine supplementation showed a strong inhibitory effect ($-53.4\%$, $p < 0.001$). Their combined application resulted in a synergistic decrease relative to bliss independence ($-19.0\%$ per mM, $p < 0.001$), while the L-alanine negative control yielded a smaller decrease, confirming intervention specificity ($-31.6\%$ at 5mM, $p < 0.01$). Intracellular L-arginine accumulation increased significantly ($p < 0.01$) under high extracellular L-arginine supplementation, confirming the validity of the intervention.

The generated hypothesis (*hypothesis 2*) regarding an antagonistic interaction between glutamate and spermine, was partially consistent with empirical data (see Fig. 3B). Addition of L-glutamate by itself 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 accumulation did not increase under supplementation, suggesting that a downstream metabolite might be the actual mediator of changes in resistance. The negative control (L-alanine) increased resistance to spermine ($+45.5\%$ at 10mM, $p < 0.05$) and significantly decreased accumulation of L-glutamate ($p < 0.001$).

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 Fig. 3C, 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 negative control (L-arginine) also had a much larger effect on resistance under treatment compared to bliss independence ($+486\%$ at 10 mM, $p < 0.001$), almost completely nullifying the effect of the stressor.

In regards to *hypothesis 4* (a suggested interaction between intracellular L-lysine accumulation and hyperosmotic stress tolerance when exposed to 30% (w/v) sucrose), L-lysine showed a strong rescuing and specific effect in regards to the stressor ($+8.9\%$ per mM, $p < 0.001$), with the control (L-valine) showing a significant, but modest rescuing effect ($+22.5\%$ at 10 mM, $p < 0.05$). L-lysine levels were significantly elevated under the higher supplementation condition ($p < 0.05$).

L-glutamine and acetic acid (*hypothesis 5*) showed, in contrast to the predicted outcome, a strong antagonistic effect ($-4.3\%$ per mM, $p < 0.05$). However, L-leucine showed an equally strong antagonistic effect ($-36.9\%$ at 10mM, $p < 0.01$), indicating the likely non-specificity of the intervention. L-glutamine was not significantly accumulated when supplemented, indicating a likely downstream mediator of reduced resistance.

*Hypothesis 6*, suggesting a rescuing effect of L-arginine supplementation on lithium-induced stress (at a lithium chloride concentration of 0.1M) was inconsistent with the data. 100 mM of lithium chloride proved a significant stress for the cells, reducing AUC ($-84.6\%$, $p < 0.001$). The intervention had little effect on the outcome of resistance, although did increase arginine accumulation at higher concentrations ($p < 0.05$). The negative control (L-alanine) showed a small antagonistic decrease in AUC relative to independent effects ($-15.9\%$ at 5mM, $p < 0.05$) and decreased intracellular concentrations of arginine ($p < 0.05$). *Hypothesis 7*, regarding L-proline and lactic acid, showed little to no effect on the outcome of treatment. Noting that the intervention and the negative control both had moderate, and significant in the case of the latter, effects on growth dynamics, with L-alanine showing a slight inhibitory effect ($-38.6\%$ at 20mM, $p < 0.001$) on growth.

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

### Iterative hypothesis refinement.

To demonstrate the usefulness of metabolomics data in refining hypotheses in this framework, we apply regularized linear regression (ElasticNet) to estimate metabolic factors likely to increase resistance to formic acid (using treated metabolic profiles and growth dynamics derived from *hypothesis 3*, represented in Fig. 3B). Using the results, a logic program was designed using the experiment metadata and the highest-valued—locally available—metabolite (aminoadipate), see below. The hypothesis is then fed into the initial step of the pipeline, and the experiment was subsequently performed. Results can be seen in Fig. 4. Complete regression coefficients can be found in Sup. Table 3.

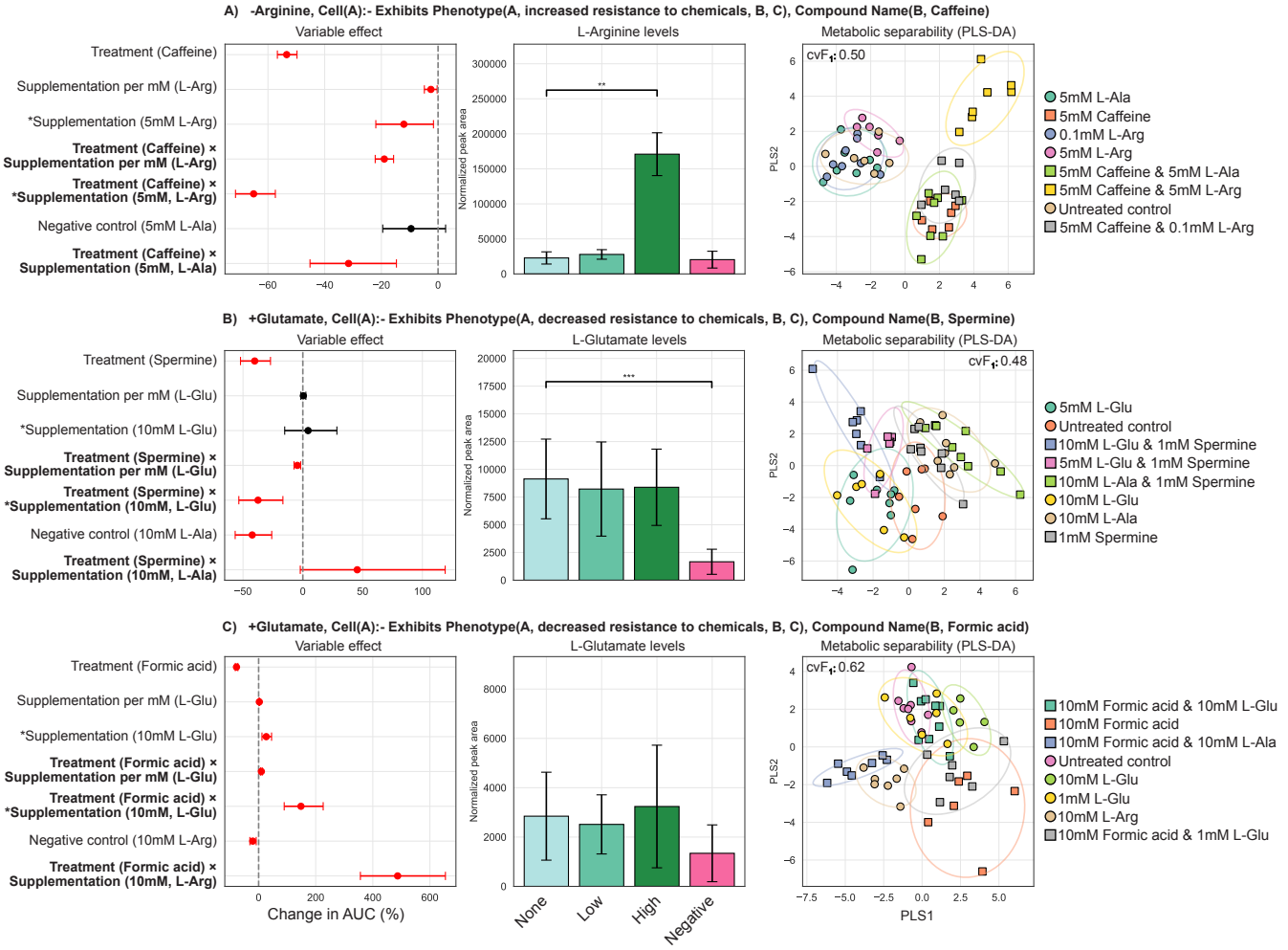


**Fig. 3.** Results from three of the interaction experiments with different outcomes. The title denotes logic program that was tested. The forest plot signifies the coefficients and confidence intervals (2.5% and 97.5%) derived from a generalized linear model using the area under the growth curve (AUC) as the response variable. Red highlights significant ($p < 0.05$) variable effects as calculated by a non-parametric bootstrap. The bar-plot shows the intracellular accumulation of the intervention compound, with the error-bars denoting the standard deviation. The PLS-DA loading plot visualizes the two-dimensional separability of the metabolic profiles of the experimental groups, confirming a metabolic change from the given interventions. The classification performance for the full five components can be seen in the corner. **Effect scaled to correspond to an equimolar dose of the negative control.*

$$
\begin{aligned}
&+\mathsf{Aminoadipate},\ \mathsf{Cells(A)} :=\\
&\quad \mathsf{ExhibitsPhenotype(A,\ Increased\ Resistance,\ B,\ C)},\\
&\quad \mathsf{CompoundName(B,\ Formic\ acid)},\\
&\quad \mathsf{Condition(C,\ Treatment:\ 10mM\ Formic\ acid)}.
\end{aligned}
$$

The results show 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 Fig. 4B), and as such, the dynamics might be subject to saturation effects.

### Structured acquisition of metabolomics data.

As part of the evaluation, metabolomics data was collected for each individual cultivation in the study, producing 976 separate biological injections, including 336 sQC injections. This generated a well-annotated dataset with identified peaks (using both CCS and m/z for matching, providing higher confidence identifications), complete with metadata regarding the full hypothesis details and experimental designs. The metabolomics data consists of data in 52 unique conditions, with 21 different nutrient compositions, and 8 treatments, including controls. Each condition has a minimum of 10 biological replicates along with associated growth dynamics.

### Database of hypotheses and experimental data.

To enable future iterations of this work using semantically rich data, and to fully record the hypotheses and experimental data, we extended and connected existing ontologies (e.g. APO), introducing several new draft terms where necessary. We transformed the generated hypotheses, the experimental protocols, and the empirical data into OWL-DL statements and recorded these in a database. The details of the dataset and ontology terms are described in the Methods section and are shown in Fig. 2.

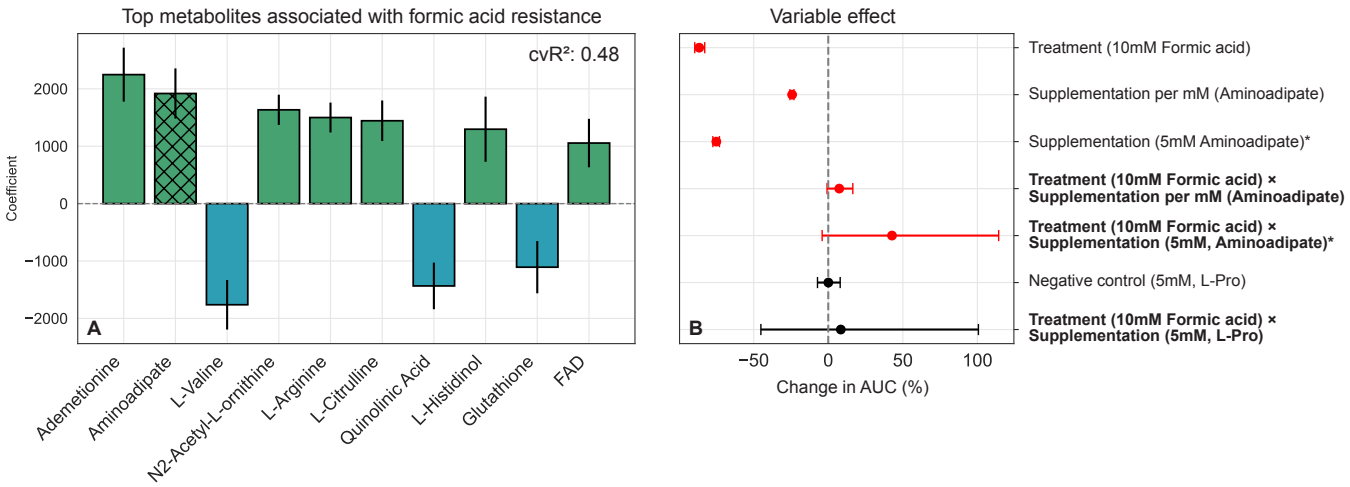


**Fig. 4.** Results from hypothesis refinement. **A.** The bar-plot denotes the average magnitude of feature importances derived from an ElasticNet regressor when predicting for resistance to formic acid stress (utilising data derived from *hypothesis 4*, see Fig. 3B). Regression performance is described in coefficient of determination across repeated cross validation ($cvR^2$). The hatched bar is the compounds selected to generate the experiment (aminoadipate). The error bars denote the standard deviation across folds. **B.** The forest plot signifies the coefficients and confidence intervals (2.5% and 97.5%) derived from a generalized linear model using the area under the growth curve (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 hypothesised logic program. **Effect scaled to correspond to an equimolar dose of the specificity control.*

### Efficient reuse of experimental data.

Despite efficiencies from the automation procedures we developed, conducting experiments is costly. Good scientific practice makes the most of available data, and often an assessment of a hypothesis can be reached, at least to some initial degree, without designing new experiments. To achieve this, we designed the graph database to be queried for empirical data relevant for a particular combination of chemical stressors and nutrient interventions. The following hypothesis (regarding a suggested antagonistic interaction between intracellular L-proline accumulation and resistance to formic acid stress) was generated by the initial stages of the automated procedure (see Sup. Table 1).

$$
\begin{aligned}
&+\mathsf{Proline},\ \mathsf{Cells(A)} :=\\
&\quad \mathsf{ExhibitsPhenotype(A,\ Decreased\ Resistance,\ B,\ C)},\\
&\quad \mathsf{CompoundName(B,\ Formic\ acid)}
\end{aligned}
$$

Before executing an experiment, the dataset 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 Fig. 5C—shows that proline supplementation showed a rescuing, but statistically insignificant, effect on formic acid stress resistance, see Fig. 4B. This evidence is inconsistent with the hypothesis, so we can de-prioritise this hypothesis for experimental study, saving valuable resources.

## Discussion

The future of research is moving towards automated solutions. Advances in robotics, machine learning and high-throughput technologies are transforming the scientific process (1). These innovations can accelerate the research cycle and drastically reduce variation introduced by human error and incomplete recording of experimental protocols, environmental conditions, and other subtle factors that can influence outcomes (23). In this automated paradigm, algorithms design experiments, robots execute protocols, and the results are automatically analysed, enabling exploration of vast experimental spaces that lie far beyond human throughput (2–4, 10). By codifying workflows and enforcing rigorous data-handling and reusability practices, these platforms could ensure that every result is traceable, reproducible, and extensible. Ultimately, by combining complex exploration with automated, quality-controlled systems, automation promises to transform biology into a truly scalable, collaborative, and reliable discipline.

Building on this vision, we designed and evaluated an automated framework that systematically produces and experimentally validates computationally generated, data-driven hypotheses, with comprehensive metadata recorded at every stage. With the method revealing, among others, antagonistic interactions between biologically essential compounds such as L-glutamate and spermine; phenomena not previously explored. Additionally, we demonstrate the utility of metabolomics data for evaluating interventions, suggesting alternative explanations for hypotheses and follow-up experiments. Mass spectrometry-based metabolomics is a promising method for automatic hypothesis generation and validation due to its relative speed, ease of automation and low cost (8, 14).

While some hypotheses were consistent with the data, negative results were equally informative, revealing significant interactions between the intervention and stressor compounds. Hypothesis falsification exposes the boundaries of our mechanistic assumptions, revealing, for instance, when intracellular glutamate levels fail to rise or when a downstream metabolite drives resistance. In this work, we used metabolite accumulations to generate hypotheses rather than functional fluxes, due to their greater availability. While flux-

omics offers a more direct mechanistic link (24), such data are rare, costly, and large-scale screens in *S. cerevisiae* are largely non-existent. Although concentrations are not directly predictive of flux, interventions based on metabolite levels will likely affect related reactions; modulating fluxes upstream and downstream of the target. By systematically archiving every outcome using existing ontologies, we continuously build a knowledge base that both refines our predictive framework and serves as a community resource for future systems-biology research and cross-study insights. In this way, the platform leverages falsification not as failure but rather as a critical contribution.

Several of the findings discovered as part of the evaluation are understudied 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 pathways, such as glutathione metabolism (25). The same goes for the interaction between L-arginine and caffeine, albeit potentially connected through separate regulation of the TORC1-complex. Caffeine acts as an inhibitor of TORC1 in yeast, and arginine is known to be tightly connected to the TORC1-complex (26, 27). Additionally, while L-glutamate has been previously implicated in formic acid tolerance, no study has yet to strongly connect aminoadipate to formic acid resistance (28). For some of the investigations, whilst the hypothesis itself predicted the wrong sign in change of resistance, the experiment itself still uncovered informative interactions, such as between L-lysine and sucrose-induced hyper-osmotic stress. Lysine related genes have been shown to be down-regulated when treated with this stressor (29), but little work has been done to draw a direct link between the two.

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 Sup. Note 2). Its modular design also allows incorporation of additional data types, like transcriptomics and proteomics, with only minimal adjustments to the hypothesis-generation pipeline (30).

Although no intellectual input is needed from humans beyond defining the framework's scope and limitations, the workflow is not fully automated: it still requires minimal human intervention at select steps, such as transferring sample plates between stations and discarding supernatant after quenching. We also note that these actions are trivially automated should the method need to be scaled; moreover, adding a more comprehensive orchestration layer could coordinate task scheduling, making it truly autonomous. Looking forward, scaling this automated workflow would require strict selection of hypotheses to maximize information gain per run. Integrating decision-theoretic frameworks, econometric modelling and metrics from information theory could allow the system to better prioritize experiments, saving resources and contributing to scientific knowledge in a faster and more ethically sound manner (1, 10, 31). More recent—albeit more computationally expensive—LLMs, such as complex reasoning models, could also allow more sophisticated hypotheses to be explored .

In this work, we demonstrate a flexible approach to automatic hypothesis generation and experimental verification using LLMs, relational learning, and automated laboratory setups. Although the evaluated hypotheses are relatively simple, our results suggest that this method can dramatically increase the throughput and reliability of scientific discovery in systems biology.

## Methods

### Frequent pattern mining.

The relational database makes use of the Saccharomyces Genome Database (SGD), 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) (18, 32–35). 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 Sup. Note 2 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. The database was filtered to only include facts derived from systematic mutation sets. The database is represented in Datalog (36).

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 (19, 20). This to find frequent relational patterns using the deleted genes as positive examples, as in Brunnsåker et al. (30). 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 Sup. Note 2.

The resulting patterns were propositionalised into a tabular dataset of logic programs covering 4678 single-gene deletants from the Mülleder et al. metabolomics dataset (21).

### Hypothesis weighting.

In order to link logic programs with a biological observable, the propositionalised logic programs were used as independent variables for predicting amino acid accumulations, as measured in Mülleder et al. (21). Associations were learned using an ElasticNet algorithm (37). Note that this makes the implicit assumption that it is an accumulation that underlies the hypothesis, not the functional flux. Learned coefficients were then averaged across 10-fold cross-validation. Each

logic program was tied to its coefficient, providing a weight (and an 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 $\infty$-norm. Amino acids with a negative $cvR^2$ were removed from further consideration.

### Semantic recording of hypotheses.

The generated hypotheses have their origins in data downloaded from the Saccharomyces Genome Database (SGD), which records data using terms taken from ontologies: phenotypes are recorded using terminology defined in the Ascomycete Phenotype Ontology (APO) (17), and chemical species are defined using terms from Chemical Entities of Biological Interest (ChEBI) (38).

APO was designed to record hypotheses about genotype–phenotype relations, and statements generally have four components: (1) the type of mutant; (2) the phenotype; (3) a qualifier on the phenotype (e.g. "increased"); and (4) the experiment type[1]. The implicit definition of these qualifiers is for a mutant with respect to a wild-type strain.

Being insufficient for recording hypotheses concerning arbitrary changes in genotype *or* phenotype, we extend 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." Fig. 5A,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 axioms introduced to extend APO were edited using Protégé (v5.6.4) (39). 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. (40). An example set of description logic axioms that record the hypothesis stated above, and the resulting OWL-DL statements (serialised in TriG format), is provided in Sup. Note 5.

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

### Generative experimental design.

A standardized context (background, protocols and lab constraints) is 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. 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 setup in use, this final experimental plan is at this stage manually assessed for safety before being passed on.

A third LLM-agent auto-formalises 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-formalised via the OpenAI API using the GPT-4o model (v.2024-08-06) through the official Python client library (openai v.1.44.1) (22).

Using the formalised, machine-readable experimental protocol, the experimental variables are automatically identified, and a robust plate-layout is generated using PLAID (41). 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; Hamilton Company) is generated. An example scheme can be seen in Sup. Table 4.

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 Sup. Table 5.

### Automated cultivation and sample processing.

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 sulfate, 2% glucose, and 75 µg/mL ampicillin, but no amino acids). The strain used to represent a reference genotype was selected to be $\Delta ho$ from the prototrophic strain collection constructed in Mülleder et al. (42), due to the inertness of the deletion in non-mating-type contexts. Each hypothesis is tested on separate microwell plates, allowing for

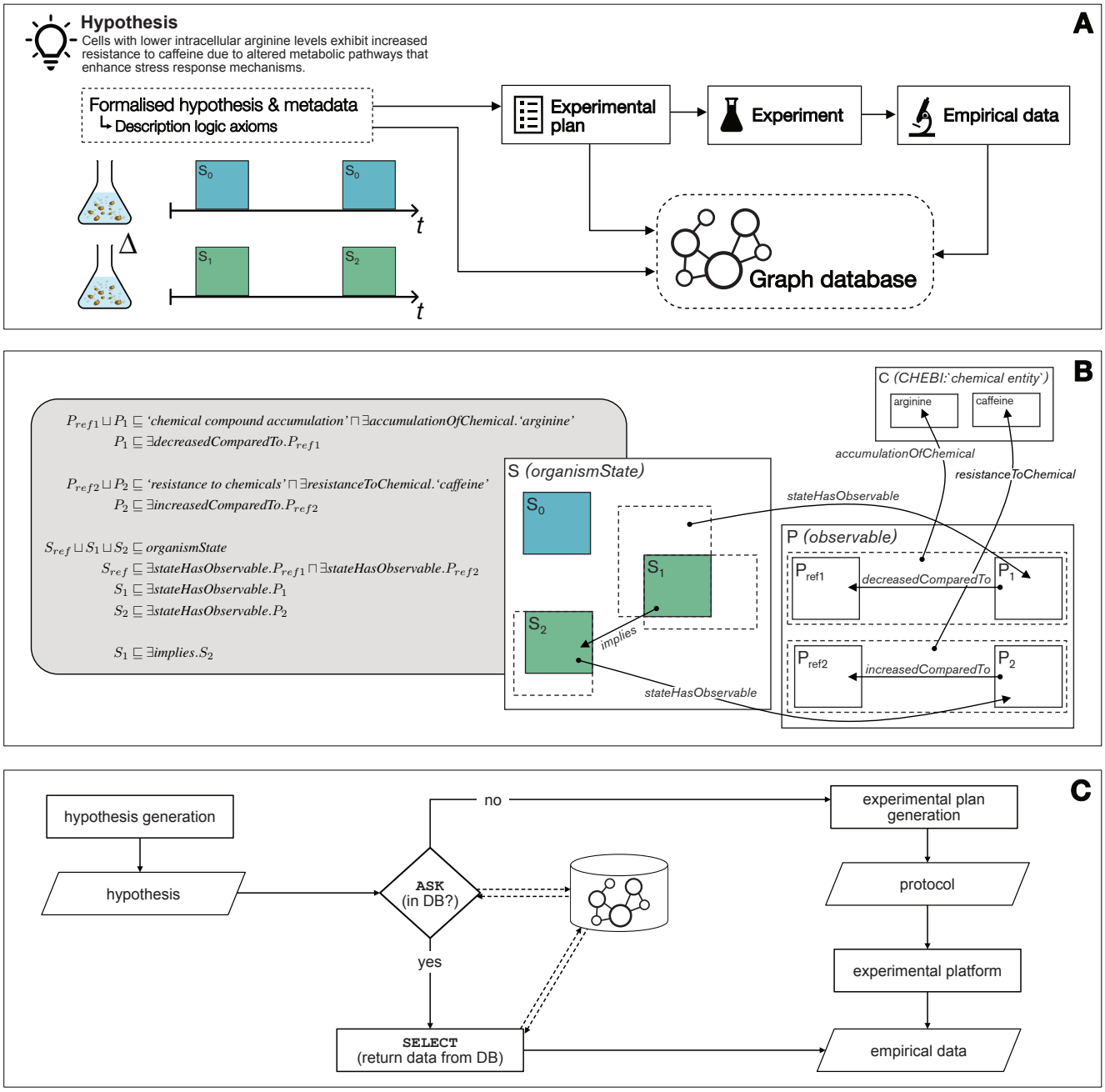


**Fig. 5.** Graph Database and description logics. **A.** The hypothesis, in its original form as a logic program and also with the LLM output, is formalised into description logic axioms using an extended version of the Ascomycete Phenotype Ontology (APO); the generated experimental plans are stored using terms from the Ontology for Biomedical Investigations (OBI); and empirical data is 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.

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 predefined parameters, such as cultivation time (3, 8, 10). For all of the experiments performed in this project, cultivation duration was set to 20h. See Sup. Note 1 for a complete description of the cultivation protocols used.

### 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 2240g 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. (8). Extraction is robotically performed in a Hamilton Microlab Star, where 150uL 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 temperature drops. 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 (0.2uM, VWR, 97052-124) and separated using positive pressure filtering (set at 40psi for 60 seconds).

Study QCs are prepared manually through overnight cultivation of $\Delta ho$ in reference conditions. 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. (8) and Reder et al. (14).

**Growth metrics and hypothesis testing.**
Growth curves are constructed from raw optical density (OD) readings at 600nm 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 area under the curve (AUC), total variation and final OD values. Wells exceeding a threshold of 3 MADs for either 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 (LOESS, statsmodels, v.0.14.4) (43). AUC is calculated using the composite trapezoidal rule on the smoothed curves (numpy, v.1.24.3) (44).

The impact 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 continuous supplement concentration, their interaction, 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 Fig. 3. The experimental variables, such as concentrations, are extracted from the generated protocols.

**Automated metabolomics 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 (18). CCS-values for relevant metabolites were acquired from AllCCS (45). 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 (15, 16). 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 (14). Peaks are identified based on both m/z (mass-to-charge ratio) and CCS ($Å^2$) (46). The cartridge in use for all experiments was HILIC (Hydrophilic Interaction Liquid Chromatography). Details regarding mobile phases, method parameters, acquisition settings, demultiplexing parameters and peak picking parameters can be seen in Sup. Note 3 and Sup. Note 4.

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 ($\geq 600ms$). 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 (RSD) in the sQC samples is used. Samples are normalized using Probabilistic Quotient Normalization (PQN), utilising the sQC-samples as the reference spectra (47). Missing values are imputed with multivariate iterative imputation using RandomForest (48). Samples outside of the 95% Hotelling T2 ellipse (using PCA-derived principal components) and high unexplained residual variances (beyond 99th percentile) are treated as outliers, similar to quality control recommendations suggested by González-Domínguez *et al.* (49). Outlier reports are included in the experiment summaries found in the GitHub repository.

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) (50). Metabolic separability is classified using (5-component) partial least-squares discriminant analysis (PLS-DA), treating each experimental group as a separate class (mixOmics, v.6.22.0) (51). 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) (52). AUC of treated samples are used as the dependent variable. Example results can be seen in Fig. 3.

**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 Ontology for Biomedical Investigations (OBI) (53); and terms for

growth and metabolomics are taken primarily from the Genesis ontology (40).

The JSON files containing the experimental protocols, and the tabular files containing growth and metabolomics data, are then mapped to relevant terms from existing ontologies using RMLMapper (v7.3.3); the mapping files can be found in the GitHub repository.

## Data availability

All data are included in the supplementary information or available from the authors. Mass-spectrometry data and associated metadata can be downloaded from Zenodo at https://doi.org/10.5281/zenodo.15730908.

## Code availability

Code (including cultivation, quenching and extraction scripts), contexts, complete supplemental information and experimental results can be found on GitHub at https://github.com/DanielBrunnsaker/GenExp.

## Bibliography


1. Sebastian Musslick, Laura K. Bartlett, Suyog H. Chandramouli, Marina Dubova, Fernand Gobet, Thomas L. Griffiths, Jessica Hullman, Ross D. King, J. Nathan Kutz, Christopher G. Lucas, Suhas Mahesh, Franco Pestilli, Sabina J. Sloman, and William R. Holmes. Automating the practice of science: Opportunities, challenges, and implications. *Proceedings of the National Academy of Sciences*, 122(5):e2401238121, February 2025. doi: 10.1073/pnas.2401238121. Publisher: Proceedings of the National Academy of Sciences.
2. Ross D. King, Jem Rowland, Stephen G. Oliver, Michael Young, Wayne Aubrey, Emma Byrne, Maria Liakata, Magdalena Markham, Pinar Pir, Larisa N. Soldatova, Andrew Sparkes, Kenneth E. Whelan, and Amanda Clare. The Automation of Science. *Science*, 324 (5923):85–89, April 2009. doi: 10.1126/science.1165620. Publisher: American Association for the Advancement of Science.
3. Anthony Coutant, Katherine Roper, Daniel Trejo-Banos, Dominique Bouthinon, Martin Carpenter, Jacek Grzebyta, Guillaume Santini, Henry Soldano, Mohamed Elati, Jan Ramon, Celine Rouveirol, Larisa N. Soldatova, and Ross 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(36):18142–18147, September 2019. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1900548116.
4. Daniil A. Boiko, Robert MacKnight, Ben Kline, and Gabe Gomes. Autonomous chemical research with large language models. *Nature*, 624(7992):570–578, December 2023. ISSN 1476-4687. doi: 10.1038/s41586-023-06792-0. Publisher: Nature Publishing Group.
5. Peiran Zhang, Zhenhua Tian, Ke Jin, Kaichun Yang, Wesley Collyer, Joseph Rufo, Neil Upreti, Xianjun Dong, Luke P. Lee, and Tony Jun Huang. Automating life science labs at the single-cell level through precise ultrasonic liquid sample ejection: PULSE. *Microsystems & Nanoengineering*, 10(1):1–16, November 2024. ISSN 2055-7434. doi: 10.1038/s41378-024-00798-y. Publisher: Nature Publishing Group.
6. Nathan J. Szymanski, Bernardus Rendy, Yuxing Fei, Rishi E. Kumar, Tanjin He, David Milsted, Matthew J. McDermott, Max Gallant, Ekin Dogus Cubuk, Amil Merchant, Haegyeom Kim, Anubhav Jain, Christopher J. Bartel, Kristin Persson, Yan Zeng, and Gerbrand Ceder. An autonomous laboratory for the accelerated synthesis of novel materials. *Nature*, 624 (7990):86–91, December 2023. ISSN 1476-4687. doi: 10.1038/s41586-023-06734-w. Publisher: Nature Publishing Group.
7. Juraj Gottweis, Wei-Hung Weng, Alexander Daryin, Tao Tu, Anil Palepu, Petar Sirkovic, Artiom Myaskovsky, Felix Weissenberger, Keran Rong, Ryutaro Tanno, Khaled Saab, Dan Popovici, Jacob Blum, Fan Zhang, Katherine Chou, Avinatan Hassidim, Burak Gokturk, Amin Vahdat, Pushmeet Kohli, Yossi Matias, Andrew Carroll, Kavita Kulkarni, Nenad Tomasev, Yuan Guan, Vikram Dhillon, Eeshit Dhaval Vaishnav, Byron Lee, Tiago R. D. Costa, José R. Penadés, Gary Peltz, Yunhan Xu, Annalisa Pawlosky, Alan Karthikesalingam, and Vivek Natarajan. Towards an AI co-scientist, February 2025. arXiv:2502.18864 [cs].
8. Daniel Brunnsåker, Gabriel K. Reder, Nikul K. Soni, Otto I. Savolainen, Alexander H. Gower, Ievgeniia A. Tiukova, and Ross D. King. High-throughput metabolomics for the design and validation of a diauxic shift model. *npj Systems Biology and Applications*, 9(1):1–9, April 2023. doi: 10.1038/s41540-023-00274-9.
9. Microsoft Research AI4Science and Microsoft Azure Quantum. The Impact of Large Language Models on Scientific Discovery: a Preliminary Study using GPT-4, December 2023. arXiv:2311.07361 [cs].
10. Kevin Williams, Elizabeth Bilsland, Andrew Sparkes, Wayne Aubrey, Michael Young, Larisa N. Soldatova, Kurt De Grave, Jan Ramon, Michaela de Clare, Worachart Sirawaraporn, Stephen G. Oliver, and Ross D. King. Cheaper faster drug development validated by the repositioning of drugs against neglected tropical diseases. *Journal of the Royal Society, Interface*, 12(104):20141289, March 2015. ISSN 1742-5662. doi: 10.1098/rsif.2014.1289.
11. Ross D. King, Kenneth E. Whelan, Ffion M. Jones, Philip G. K. Reiser, Christopher H. Bryant, Stephen H. Muggleton, Douglas B. Kell, and Stephen G. Oliver. Functional genomic hypothesis generation and experimentation by a robot scientist. *Nature*, 427(6971): 247–252, January 2004. ISSN 1476-4687. doi: 10.1038/nature02236. Publisher: Nature Publishing Group.
12. Hiroaki Kitano. Systems Biology: A Brief Overview. *Science*, 295(5560):1662–1664, March 2002. doi: 10.1126/science.1069492. Publisher: American Association for the Advancement of Science.
13. Ziwei Xu, Sanjay Jain, and Mohan Kankanhalli. Hallucination is Inevitable: An Innate Limitation of Large Language Models, February 2025. arXiv:2401.11817 [cs].
14. Gabriel K. Reder, Erik Y. Bjurström, Daniel Brunnsåker, Filip Kronström, Praphapan Lasin, Ievgeniia Tiukova, Otto I. Savolainen, James N. Dodds, Jody C. May, John P. Wikswo, John A. McLean, and Ross D. King. AutonoMS: Automated Ion Mobility Metabolomic Fingerprinting. *Journal of the American Society for Mass Spectrometry*, 35(3):542–550, March 2024. ISSN 1044-0305. doi: 10.1021/jasms.3c00396. Publisher: American Society for Mass Spectrometry. Published by the American Chemical Society. All rights reserved.
15. Brian T. Veach, Thilak K. Mudalige, and Peter Rye. RapidFire Mass Spectrometry with Enhanced Throughput as an Alternative to Liquid–Liquid Salt Assisted Extraction and LC/MS Analysis for Sulfonamides in Honey. *Analytical Chemistry*, 89(6):3256–3260, March 2017. ISSN 0003-2700. doi: 10.1021/acs.analchem.6b04889. Publisher: American Chemical Society.
16. Jody C. May, Cody R. Goodwin, Nichole M. Lareau, Katrina L. Leaptrot, Caleb B. Morris, Ruwan T. Kurulugama, Alex Mordehai, Christian Klein, William Barry, Ed Darland, Gregor Overney, Kenneth Imatani, George C. Stafford, John C. Fjeldsted, and John A. McLean. Conformational Ordering of Biomolecules in the Gas Phase: Nitrogen Collision Cross Sections Measured on a Prototype High Resolution Drift Tube Ion Mobility-Mass Spectrometer. *Analytical Chemistry*, 86(4):2107–2116, February 2014. ISSN 0003-2700. doi: 10.1021/ac4038448. Publisher: American Chemical Society.
17. Stacia R Engel, Suzi Aleksander, Robert S Nash, Edith D Wong, Shuai Weng, Stuart R Miyasato, Gavin Sherlock, and J Michael Cherry. Saccharomyces Genome Database: advances in genome annotation, expanded biochemical pathways, and other key enhancements. *Genetics*, 229(3):iyae185, March 2025. ISSN 1943-2631. doi: 10.1093/genetics/iyae185.
18. Chengyu Zhang, Benjamín J Sánchez, Feiran Li, Cheng Wei Quan Eiden, William T Scott, Ulf W Liebal, Lars M Blank, Hendrik G Mengers, Mihail Anton, Albert Tafur Rangel, Sebastián N Mendoza, Lixin Zhang, Jens Nielsen, Hongzhong Lu, and Eduard J Kerkhoven. Yeast9: a consensus genome-scale metabolic model for S. cerevisiae curated by the community. *Molecular Systems Biology*, 20(10):1134–1150, October 2024. ISSN 1744-4292. doi: 10.1038/s44320-024-00060-7. Num Pages: 1150 Publisher: John Wiley & Sons, Ltd.
19. Ashwin Srinivasan. The Aleph Manual, 2007.
20. Ross D. King, Ashwin Srinivasan, and Luc Dehaspe. Warmr: a data mining tool for chemical data. *Journal of Computer-Aided Molecular Design*, 15(2):173–81, February 2001. doi: 10.1023/A:1008171016861.
21. Michael Mülleder, Enrica Calvani, Mohammad Tauqeer Alam, Richard Kangda Wang, Florian Eckerstorfer, Aleksej Zelezniak, and Markus Ralser. Functional Metabolomics Describes the Yeast Biosynthetic Regulome. *Cell*, 167(2):553–65, October 2016. doi: 10.1016/j.cell.2016.09.007.
22. OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan

Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, C. J. Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. GPT-4 Technical Report, March 2024. arXiv:2303.08774 [cs].

23. Katherine Roper, A. Abdel-Rehim, Sonya Hubbard, Martin Carpenter, Andrey Rzhetsky, Larisa Soldatova, and Ross D. King. Testing the reproducibility and robustness of the cancer biology literature by robot. *Journal of The Royal Society Interface*, 19(189):20210821, April 2022. doi: 10.1098/rsif.2021.0821.
24. Abdul-Hamid Emwas, Kacper Szczepski, Inas Al-Younis, Joanna Izabela Lachowicz, and Mariusz Jaremko. Fluxomics - New Metabolomics Approaches to Monitor Metabolic Pathways. *Frontiers in Pharmacology*, 13, March 2022. ISSN 1663-9812. doi: 10.3389/fphar.2022.805782. Publisher: Frontiers.
25. Minoru Kanehisa, Miho Furumichi, Yoko Sato, Masayuki Kawashima, and Mari Ishiguro-Watanabe. KEGG for taxonomy-based analysis of pathways and genomes. *Nucleic Acids Research*, 51(1):587–92, January 2023. doi: 10.1093/nar/gkac963.
26. Duygu Dikicioglu, Elif Dereli Eke, Serpil Eraslan, Stephen G. Oliver, and Betul Kirdar. Saccharomyces cerevisiae adapted to grow in the presence of low-dose rapamycin exhibit altered amino acid metabolism. *Cell Communication and Signaling*, 16(1):85, November 2018. ISSN 1478-811X. doi: 10.1186/s12964-018-0298-y.
27. Aaron Reinke, Jenny C. Y. Chen, Sofia Aronova, and Ted Powers. Caffeine Targets TOR Complex I and Provides Evidence for a Regulatory Link between the FRB and Kinase Domains of Tor1p*. *Journal of Biological Chemistry*, 281(42):31616–31626, October 2006. ISSN 0021-9258. doi: 10.1016/S0021-9258(19)84075-9.
28. Lingjie Zeng, Zaiyong Si, Xuemei Zhao, Pixue Feng, Jinxiang Huang, Xiufeng Long, and Yi Yi. Metabolome analysis of the response and tolerance mechanisms of *Saccharomyces cerevisiae* to formic acid stress. *The International Journal of Biochemistry & Cell Biology*, 148:106236, July 2022. ISSN 1357-2725. doi: 10.1016/j.biocel.2022.106236.
29. Daniel J Erasmus, George K van der Merwe, and Hennie J.J. van Vuuren. Genome-wide expression analyses: Metabolic adaptation of Saccharomyces cerevisiae to high sugar stress. *FEMS Yeast Research*, 3(4):375–399, June 2003. ISSN 1567-1356. doi: 10.1016/S1567-1356(02)00203-9.
30. Daniel Brunnsåker, Filip Kronström, Ievgeniia A Tiukova, and Ross D King. Interpreting protein abundance in Saccharomyces cerevisiae through relational learning. *Bioinformatics*, 40(2):btae050, February 2024. ISSN 1367-4811. doi: 10.1093/bioinformatics/btae050.
31. Ross D. King, Teresa Scassa, Stefan Kramer, and Hiroaki Kitano. Stockholm declaration on AI ethics: why others should sign. *Nature*, 626(8000):716–716, February 2024. doi: 10.1038/d41586-024-00517-7. Bandiera_abtest: a Cg_type: Correspondence Publisher: Nature Publishing Group Subject_term: Machine learning, Ethics.
32. J. Michael Cherry, Eurie L. Hong, Craig Amundsen, Rama Balakrishnan, Gail Binkley, Esther T. Chan, Karen R. Christie, Maria C. Costanzo, Selina S. Dwight, Stacia R. Engel, Dianna G. Fisk, Jodi E. Hirschman, Benjamin C. Hitz, Kalpana Karra, Cynthia J. Krieger, Stuart R. Miyasato, Rob S. Nash, Julie Park, Marek S. Skrzypek, Matt Simison, Shuai Weng, and Edith D. Wong. Saccharomyces Genome Database: the genomics resource of budding yeast. *Nucleic Acids Research*, 40(Database issue):700–5, January 2012. doi: 10.1093/nar/gkr1029.
33. Damian Szklarczyk, Alberto Santos, Christian von Mering, Lars Juhl Jensen, Peer Bork, and Michael Kuhn. STITCH 5: augmenting protein–chemical interaction networks with tissue and affinity data. *Nucleic Acids Research*, 44(D1):D380–D384, January 2016. ISSN 0305-1048. doi: 10.1093/nar/gkv1277.
34. The Alliance of Genome Resources Consortium. Updates to the Alliance of Genome Resources central infrastructure. *Genetics*, 227(1):iyae049, May 2024. ISSN 1943-2631. doi: 10.1093/genetics/iyae049.
35. Feiran Li, Yu Chen, Mihail Anton, and Jens Nielsen. GotEnzymes: an extensive database of enzyme parameter predictions. *Nucleic Acids Research*, 51(1):583–6, January 2023. doi: 10.1093/nar/gkac831.
36. Luc Dehaspe and Hannu Toivonen. Discovery of frequent DATALOG patterns. *Data Mining and Knowledge Discovery*, 3(1):7–36, March 1999. doi: 10.1023/A:1009863704807.
37. Hui Zou and Trevor Hastie. Regularization and Variable Selection Via the Elastic Net. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 67(2):301–320, April 2005. ISSN 1369-7412. doi: 10.1111/j.1467-9868.2005.00503.x.
38. Janna Hastings, Gareth Owen, Adriano Dekker, Marcus Ennis, Namrata Kale, Venkatesh Muthukrishnan, Steve Turner, Neil Swainston, Pedro Mendes, and Christoph Steinbeck. ChEBI in 2016: Improved services and an expanding collection of metabolites. *Nucleic acids research*, 44(1):1214–9, January 2016. ISSN 1362-4962. doi: 10.1093/nar/gkv1031.
39. Mark A. Musen. The protégé project: a look back and a look forward. *AI Matters*, 1(4):4–12, June 2015. doi: 10.1145/2757001.2757003.
40. Gabriel K Reder, Alexander H Gower, Filip Kronström, Rushikesh Halle, Vinay Mahamuni, Amit Patel, Harshal Hayatnagarkar, Larisa N Soldatova, and Ross D King. Genesis-DB: a database for autonomous laboratory systems. *Bioinformatics Advances*, 3(1):vbad102, January 2023. ISSN 2635-0041. doi: 10.1093/bioadv/vbad102.
41. María Andreína Francisco Rodríguez, Jordi Carreras Puigvert, and Ola Spjuth. Designing microplate layouts using artificial intelligence. *Artificial Intelligence in the Life Sciences*, 3: 100073, December 2023. ISSN 2667-3185. doi: 10.1016/j.ailsci.2023.100073.
42. Michael Mülleder, Floriana Capuano, Pınar Pir, Stefan Christen, Uwe Sauer, Stephen G. Oliver, and Markus Ralser. A prototrophic deletion mutant collection for yeast metabolomics and systems biology. *Nature Biotechnology*, 30(12):1176–1178, December 2012. ISSN 1546-1696. doi: 10.1038/nbt.2442. Publisher: Nature Publishing Group.
43. William S. Cleveland, , and Susan J. Devlin. Locally Weighted Regression: An Approach to Regression Analysis by Local Fitting. *Journal of the American Statistical Association*, 83(403):596–610, September 1988. ISSN 0162-1459. doi: 10.1080/01621459.1988.10478639. Publisher: ASA Website _eprint: https://www.tandfonline.com/doi/pdf/10.1080/01621459.1988.10478639.
44. Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. Array programming with NumPy. *Nature*, 585(7825):357–362, September 2020. ISSN 1476-4687. doi: 10.1038/s41586-020-2649-2. Publisher: Nature Publishing Group.
45. Haosong Zhang, Mingdu Luo, Hongmiao Wang, Fandong Ren, Yandong Yin, and Zheng-Jiang Zhu. AllCCS2: Curation of Ion Mobility Collision Cross-Section Atlas for Small Molecules Using Comprehensive Molecular Representations. *Analytical Chemistry*, 95(37): 13913–13921, September 2023. ISSN 0003-2700. doi: 10.1021/acs.analchem.3c02267. Publisher: American Chemical Society.
46. Kendra J. Adams, Brian Pratt, Neelanjan Bose, Laura G. Dubois, Lisa St. John-Williams, Kevin M. Perrott, Karina Ky, Pankaj Kapahi, Vagisha Sharma, Michael J. MacCoss, M. Arthur Moseley, Carol A. Colton, Brendan X. MacLean, Birgit Schilling, and J. Will Thompson. Skyline for Small Molecules: A Unifying Software Package for Quantitative Metabolomics. *Journal of Proteome Research*, 19(4):1447–1458, April 2020. ISSN 1535-3893. doi: 10.1021/acs.jproteome.9b00640. Publisher: American Chemical Society.
47. Frank Dieterle, Alfred Ross, Götz Schlotterbeck, and Hans Senn. Probabilistic Quotient Normalization as Robust Method to Account for Dilution of Complex Biological Mixtures. Application in 1H NMR Metabonomics. *Analytical Chemistry*, 78(13):4281–4290, July 2006. ISSN 0003-2700. doi: 10.1021/ac051632c. Publisher: American Chemical Society.
48. Runmin Wei, Jingye Wang, Mingming Su, Erik Jia, Shaoqiu Chen, Tianlu Chen, and Yan Ni. Missing Value Imputation Approach for Mass Spectrometry-based Metabolomics Data. *Scientific Reports*, 8(1):663, January 2018. ISSN 2045-2322. doi: 10.1038/s41598-017-19120-0. Publisher: Nature Publishing Group.
49. Álvaro González-Domínguez, Núria Estanyol-Torres, Carl Brunius, Rikard Landberg, and Raúl González-Domínguez. QComics: Recommendations and Guidelines for Robust, Easily Implementable and Reportable Quality Control of Metabolomics Data. *Analytical Chemistry*, 96(3):1064–1072, January 2024. ISSN 0003-2700. doi: 10.1021/acs.analchem.3c03660. Publisher: American Chemical Society.
50. Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C. J. Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, and Paul van Mulbregt. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature Methods*, 17(3):261–272, March 2020. ISSN 1548-7105. doi: 10.1038/s41592-019-0686-2. Publisher: Nature Publishing Group.
51. Florian Rohart, Benoît Gautier, Amrit Singh, and Kim-Anh Lê Cao. mixOmics: An R package for 'omics feature selection and multiple data integration. *PLOS Computational Biology*, 13(11):e1005752, November 2017. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1005752. Publisher: Public Library of Science.
52. Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12(85):2825–2830, 2011. ISSN 1533-7928.
53. Anita Bandrowski, Ryan Brinkman, Mathias Brochhausen, Matthew H. Brush, Bill Bug, Marcus C. Chibucos, Kevin Clancy, Mélanie Courtot, Dirk Derom, Michel Dumontier, Liju Fan, Jennifer Fostel, Gilberto Fragoso, Frank Gibson, Alejandra Gonzalez-Beltran, Melissa A. Haendel, Yongqun He, Mervi Heiskanen, Tina Hernandez-Boussard, Mark Jensen, Yu Lin, Allyson L. Lister, Phillip Lord, James Malone, Elisabetta Manduchi, Monnie McGee, Norman Morrison, James A. Overton, Helen Parkinson, Bjoern Peters, Philippe Rocca-Serra, Alan Ruttenberg, Susanna-Assunta Sansone, Richard H. Scheuermann, Daniel Schober, Barry Smith, Larisa N. Soldatova, Christian J. Stoeckert Jr, Chris F. Taylor, Carlo Torniai, Jessica A. Turner, Randi Vita, Patricia L. Whetzel, and Jie Zheng. The Ontology for Biomedical Investigations. *PLOS ONE*, 11(4):e0154556, April 2016. ISSN 1932-6203. doi: 10.1371/journal.pone.0154556. Publisher: Public Library of Science.

**ACKNOWLEDGEMENTS**

The authors gratefully acknowledge the members of the Ross King Group at Chalmers University and Cambridge University for their insights and discussions. We also thank the members of the Chalmers Mass Spectrometry Infrastructure for their help and technical expertise. Additionally, we thank the Ralser Lab at Charité–Universitätsmedizin Berlin for providing the yeast strains used in this study. This work was partially supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation

**AUTHOR CONTRIBUTIONS**

D.B. and R.D.K. conceptualized the study. D.B. and A.H.G. designed the generation and analysis framework. D.B. and P.N. designed the experimental pipelines. A.H.G., F.K. and D.B. designed the ontologies and databases. D.B., P.N. and E.Y.B. executed the experiments. D.B., E.Y.B. and I.A.T. analysed the results. D.B., A.H.G. and R.D.K. wrote the manuscript.