Agentic AI Integrated with Scientific Knowledge: Laboratory Validation in Systems Biology

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]Lead contact

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 integrating LLM-based agents with laboratory automation, guided by logical scaffolds incorporating symbolic relational learning, structured vocabularies, and experimental constraints. This integration reduces incoherence and improves reliability in automated workflows.

We couple this AI-driven approach to automated cell-culture and metabolomics platforms, enabling hypothesis validation and refinement, yielding a flexible discovery system.

We validate the system in *Saccharomyces cerevisiae*, identifying novel interactions, 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 enables a more reliable, machine-driven discovery process in systems biology.

Laboratory Automation | Systems Biology | Machine Learning | Inductive Logic Programming | Automation of Science | Large Language Models
Correspondence: ***rossk@chalmers.se***

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 we grant them is an important design question. In particular, real-world agency—the ability to directly control experiments—is vital for scientific discovery. Integration with laboratory 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 has been applied in fields such as materials development and chemistry (9, 10). Yet 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 impact of LLMs in scientific discovery, often via orchestrated multi-agent systems (12–19).

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 (20). 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 programs in high-level languages such as Python (17). However, these approaches generally rely on loosely typed message passing, shallow artifact 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. 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 ground-

ing them in formal logic, exploiting background knowledge and meeting transparent data-sharing requirements (12, 17–19, 21).

Here we introduce a hybrid approach that integrates language-driven flexibility with inductive logic and formal ontologies (see Fig. 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 prioritised 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, 22). All metadata, logs, and results can then be stored in a graph database that ensures transparency, reproducibility, and efficient reuse (23).

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.

Results

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, 22, 24–27). See Fig. 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 (28–32)—(ii) applying pattern mining to generate a logical rule (26, 27), and (iii) applying machine learning on metabolomics data to learn a testable implication associated to the rule (33). 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 (34). Automated experiments are executed for each hypothesis, producing and processing biological samples to generate time-series growth

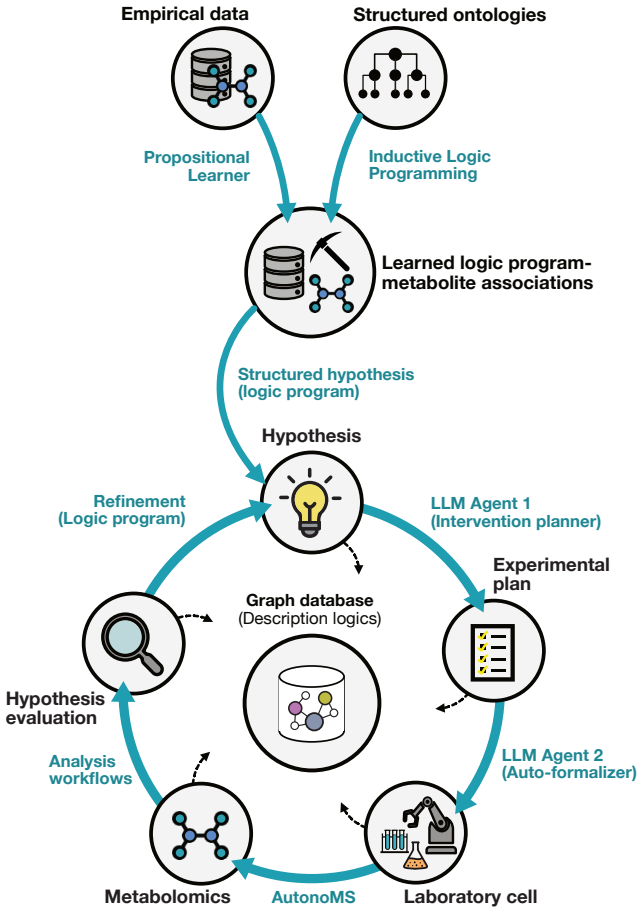

Fig. 1. An automated experimental framework enabling end-to-end biological discovery—from hypothesis generation to data integration. Hypotheses are generated by mining patterns from a Datalog database containing structured facts about yeast phenotype, physiology, and metabolism, using inductive logic programming. These are linked to metabolomics data via regression. Interventions to test each hypothesis are designed by a large language model (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 (22). All data—including results, intermediate outputs, and metadata are curated, analyzed, and integrated into a graph database for downstream access and reuse.

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.

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, utilising an extended version of the Ascomycete Phenotype Ontology (APO) (28, 29). 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).

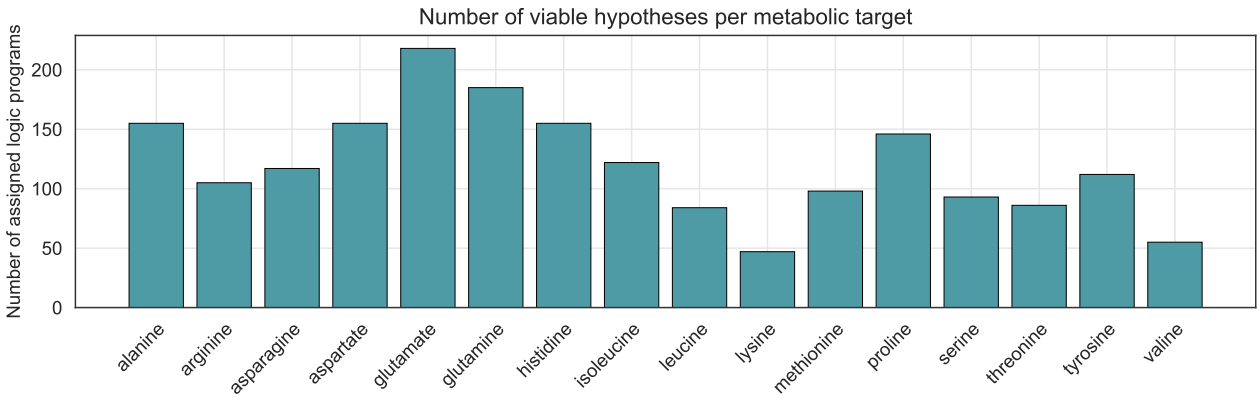

Fig. 2. Viable hypothesis counts per amino acid. Bar chart showing, for each amino acid, the number of logic programs with non-zero regression coefficients—i.e., the total number of viable hypotheses generated for that amino acid.

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 (29–31). Datalog's declarative structure enables clear expression of biological relationships and efficient querying across complex relational datasets (35).

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 (26, 27, 36). 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. (33). We propositionalised (37) 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 Fig. 2).

An overview of the hypothesis space can be seen in Sup. Note 1 and Sup. Table 1. A logic program and learned association can be seen below (also see *hypothesis 6* in table 1).

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

In this context, this program can be interpreted as: cells (A) which have decreased resistance to lithium (B) when treated with 0.1M 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. $\uparrow$ Arginine $\implies$ $\uparrow$ Lithium(+1) resistance).

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 Methods), and ran the experiment—producing time-series growth curves, endpoint metabolic profiles, effect sizes, and statistical evaluations.

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

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. "-", "+" indicates 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, 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 $\downarrow$ Arginine $\implies$ $\uparrow$ Caffeine resistance†	+	+	–	–	Yes
Hypothesis 2 $\uparrow$ Glutamate $\implies$ $\downarrow$ Spermine resistance	~	+	–	–	Yes
Hypothesis 3 $\uparrow$ Glutamate $\implies$ $\downarrow$ Formic acid resistance	~	+	–	+	No
Hypothesis 4 $\downarrow$ Glutamine $\implies$ $\downarrow$ Acetic acid resistance†	+	+	–	+	No
Hypothesis 5 $\uparrow$ Lysine $\implies$ $\downarrow$ Hyperosmotic stress resistance (at 30% (w/v) sucrose)	+	+	+	–	Yes
Hypothesis 6 $\downarrow$ Arginine $\implies$ $\downarrow$ Lithium(+1) resistance (at 0.1M lithium chloride)†	+	+	~	+	No
Hypothesis 7 $\uparrow$ Proline $\implies$ $\uparrow$ Lactic acid resistance	+	+	~	+	No
*Hypothesis 8** $\uparrow$ Aminoadipate $\implies$ $\uparrow$ Formic acid resistance (at 10mM formic acid)	–	+	+	+	Yes
*Hypothesis 9*** $\uparrow$ Proline $\implies$ $\downarrow$ Formic acid resistance	~	+	~	–	No

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 Sup. Note 7 and Sup. Table 2.

Hypothesis 1 (see table 1) was consistent with acquired empirical data (shown in Fig. 3A). The addition of L-arginine alone caused a modest reduction in AUC (area under the growth curve) (-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 Fig. 3B). 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 10mM, $p < 0.05$) and significantly decreased L-glutamate accumulation ($p < 0.001$).

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

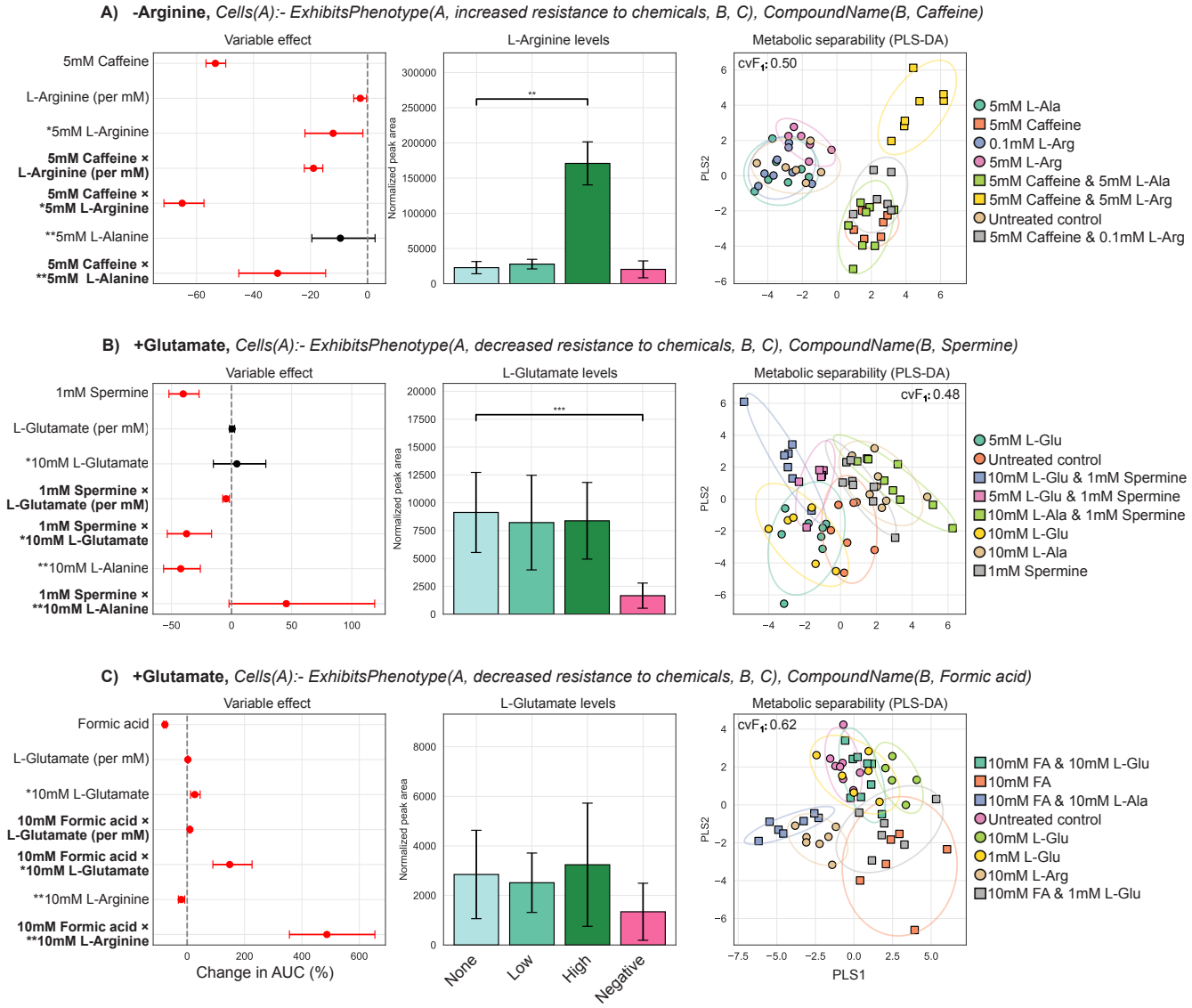

Fig. 3. Results from automated interaction experiments with examples of different outcomes. The title of each row denotes logic program that was tested along with the learned association that defines the intervention. The forest plot signifies the coefficients and confidence intervals (2.5% and 97.5%) 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. *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.*

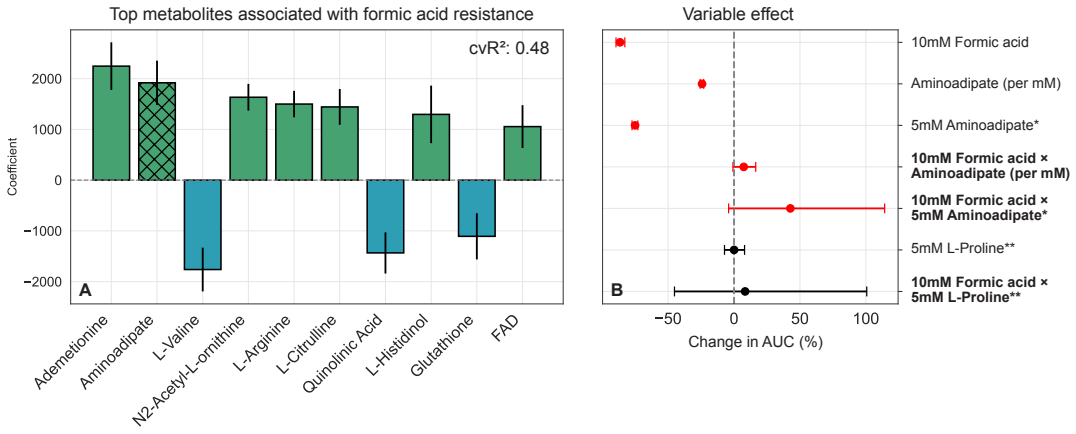

Fig. 4. Iterative hypothesis refinement. A. The bar-plot denotes the average magnitude of coefficients derived from a linear regressor when predicting for resistance to formic acid stress (using data derived from *hypothesis 4*, see Fig. 3B). Regression performance is presented as coefficient of determination 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 hypothesised logic program. **Effect scaled to correspond to an equimolar dose of the specificity control. **Negative control.*

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 10mM, $p < 0.01$), 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 5mM, $p < 0.05$) and decreased intracellular concentrations of arginine ($p < 0.05$). *Hypothesis 7* (involving L-proline and lactic acid) showed minimal impact 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 20mM, $p < 0.001$).

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 Fig. 3. Processed data is stored in the graph database and raw data is stored in online repositories (see details in *data availability*).

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

Regularized linear regression was applied to data from *hypothesis 3* (visualized in Fig. 3B) 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 Fig. 4A), a logic program was constructed using experiment metadata and the highest-ranked locally available compound—aminoadipate (see below).

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

The hypothesis was fed into the initial step of the pipeline, and the experiment was subsequently performed. Results can be seen in Fig. 4B. Complete regression coefficients can be found in Sup. Table 3.

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

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 then transformed the generated hypotheses, the experimental protocols, and the empirical data into OWL-DL statements and recorded these in a graph database (Fig. 5AB), 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).

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

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 Fig. 5C—shows no decrease in resistance to formic acid stress, see Fig. 4B, which is inconsistent with the hypothesis. Therefore, this hypothesis can be de-prioritised for experimental validation, conserving valuable resources.

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 minimisation 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 (38–40). 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) (41).

Another previously uncharacterized finding is the interaction between L-arginine and caffeine, likely 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 (42, 43).

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 down-regulated under this stressor (44), 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 (45).

Testing these hypotheses highlighted mechanistic assumptions that proved limited. For example, treatments intended to increase intracellular glutamate sometimes failed to do so. In other cases (e.g., *hypotheses 2* and *3*), resistance phenotypes were likely 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 (46). To capitalize on these insights, by archiving all outcomes under existing ontologies, the framework builds a knowledge base that supports future systems biology research and cross-study insights. Thus, the platform treats falsification as a valuable contribution rather than an uninformative failure.

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 3). Its modular design also allows incorporation of additional data types, like transcriptomics and proteomics, with only minimal adjustments to the hypothesis-generation pipeline (47).

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 for transferring sample plates between work stations. 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 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, 48). More advanced LLMs, such as complex reasoning models, could also allow more sophisticated hypotheses to be explored.

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.

Methods

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) (28, 30–32, 49, 50).

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 3 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 (35).

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 (26, 27). It is used to find frequent relational patterns using the deleted genes as positive examples, as in Brunnsåker et al. (47). 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 3.

The resulting patterns were propositionalised into a tabular dataset of logic programs covering 4678 single-gene deletants from the Mülleder et al. metabolomics dataset (33). When multiple propositionalised programs described the same examples, we prioritized the most detailed ones using a scored vocabulary.

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. (33). Associations were learned using an ElasticNet algorithm (51). 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 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 the results.

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) (28, 29), and chemical species are defined using terms from Chemical Entities of Biological Interest (ChEBI) (32).

APO was designed to record hypotheses about genotype–phenotype relations, and statements generally have four components: (1) the type of mutant; (2) the observable phenotype; (3) a qualifier on this observable (e.g. “increased"); and (4) 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. 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

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

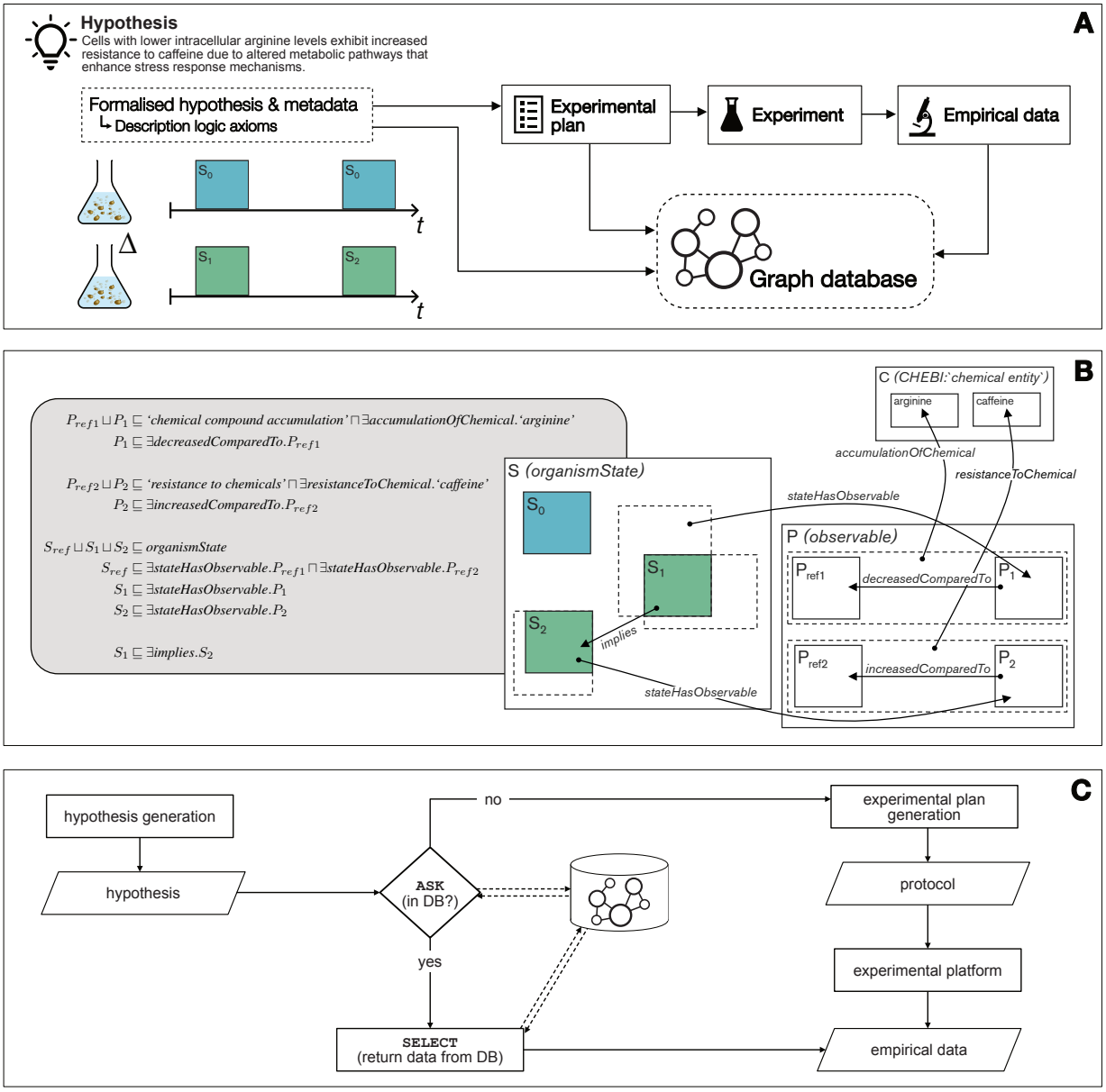

Fig. 5. Graph Database and description logics. A. The hypothesis, in its original form as a logic program and also with the LLM output, is formalised into description logic axioms using an extended version of the Ascomycete Phenotype Ontology (APO); the generated experimental plans are stored using terms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

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) (52). 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. (53). 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 6.

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 *code availability*).

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) (34).

Using the formalised, machine-readable experimental protocol, the experimental variables are automatically identified, and a robust plate-layout is generated using PLAID (54). 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 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 sulfate, 2% glucose, and 75 µg/mL ampicillin, but no amino acids). The strain used to represent a reference genotype was selected to be Δho from the prototrophic strain collection constructed in Mülleder et al. (55)—due to the inertness of the deletion in non-mating-type contexts.

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 predefined 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 20h. See Sup. Note 2 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. (5). 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 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 (0.2uM, VWR, 97052-124) and separated using positive pressure filtering (set at 40psi for 60 seconds).

Study QCs are prepared manually through overnight cultivation of Δho 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. (22).

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 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 (LOESS, statsmodels, v.0.14.4) (56). AUC is calculated using the composite trapezoidal rule on the smoothed curves

(numpy, v.1.24.3) (57).

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 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 Fig. 3. The experimental variables, such as concentrations, are extracted from the auto-formalized protocols.

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 (30). CCS-values for relevant metabolites were acquired from AllCCS (58). 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 (24, 25). 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 (22). Peaks are identified based on both m/z (mass-to-charge ratio) and CCS ($\AA^2$) (59). 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 4 and Sup. Note 5.

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 (60). Missing values are imputed with multivariate iterative imputation using RandomForest (61). 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.* (62). 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) (63). 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) (64). 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) (65). AUC of treated samples are used as the dependent variable. Example results can be seen in Fig. 3.

As part of the evaluation, metabolomics data were collected for each individual cultivation in the study, producing 976 separate biological injections, including 336 sQC 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. See *data availability* for details.

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) (66); and terms for growth and metabolomics are taken primarily from the Genesis ontology (53).

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); the mapping files can be found in the GitHub repository (see *code availability*).

Resource availability

Lead contact.

Further information and requests for resources should be directed to and will be fulfilled by the lead contact, Ross D. King (rossk@chalmers.se).

Materials availability.

This study did not generate new unique reagents.

Data and code availability.

All data supporting the findings of this study are available in the supplementary information or from the lead contact upon request. Raw mass spectrometry data have been deposited at MassIVE under accession MSV000099724 (https://doi.org/doi:10.25345/C5HT2GR2T). All data and associated metadata can be downloaded from Zenodo at https://doi.org/10.5281/zenodo.17153885. 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.

Acknowledgements

The authors gratefully acknowledge the members of the Ross King Group at Chalmers University of Technology 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 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 (EPSRC) [EP/R022925/2, EP/W004801/1 and EP/X032418/1], the Chalmers AI Research Centre, and the Swedish Research Council Formas [2020-01690].

Author contributions.

Conceptualization, D.B., R.D.K. and A.H.G.; Data curation, D.B.; Formal analysis, D.B. and A.H.G.; Funding acquisition/Supervision, R.D.K. and I.A.T.; Investigation, D.B., P.N., and E.Y.B.; Methodology, D.B., A.H.G., and P.N.; Project administration, R.D.K., I.A.T., and D.B.; Resources, R.D.K.; Software, D.B., A.H.G., and F.K.; Validation, D.B., and A.H.G.; Visualization, D.B., and A.H.G.; Writing - Original draft, D.B., A.H.G., and R.D.K.; Writing - review & editing, D.B., A.H.G., P.N., E.Y.B, F.K., I.A.T., and R.D.K.

Declaration of interests.

The authors declare no competing interests.

Declaration of generative AI and AI-assisted technologies in the manuscript preparation process.

During the preparation of this work the authors used GPT4o in order to check for grammar errors and improve the clarity of the writing. After using this service, the authors reviewed and edited the content as needed and takes full responsibility for the content of the published article.

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, and Christopher G. Lucas et al. 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.
2. Alexander H. Gower, Konstantin Korovin, Daniel Brunnsåker, Filip Kronström, Gabriel K. Reder, Ievgeniia A. Tiukova, Ronald S. Reiserer, John P. Wikswo, and Ross D. King. The Use of AI-Robotic Systems for Scientific Discovery, June 2024. arXiv:2406.17835 [cs].
3. Hector Zenil, Jesper Tegnér, Felipe S. Abrahão, Alexander Lavin, Vipin Kumar, Jeremy G. Frey, Adrian Weller, Larisa Soldatova, Alan R. Bundy, and Nicholas R. Jennings et al. The Future of Fundamental Science Led by Generative Closed-Loop Artificial Intelligence, August 2023. arXiv:2307.07522 [cs].
4. Anthony Coutant, Katherine Roper, Daniel Trejo-Banos, Dominique Bouthinon, Martin Carpenter, Jacek Grzebyta, Guillaume Santini, Henry Soldano, Mohamed Elati, and Jan Ramon et al. 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. doi: 10.1073/pnas.1900548116.
5. 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.
6. 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.
7. Ross D. King, Jem Rowland, Stephen G. Oliver, Michael Young, Wayne Aubrey, Emma Byrne, Maria Liakata, Magdalena Markham, Pinar Pir, and Larisa N. Soldatova et al. The Automation of Science. *Science*, 324(5923):85–89, April 2009. doi: 10.1126/science.1165620.
8. Kevin Williams, Elizabeth Bilsland, Andrew Sparkes, Wayne Aubrey, Michael Young, Larisa N. Soldatova, Kurt De Grave, Jan Ramon, Michaela de Clare, and Worachart Sirawaraporn et al. 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. doi: 10.1098/rsif.2014.1289.
9. 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. doi: 10.1038/s41378-024-00798-y.
10. Nathan J. Szymanski, Bernardus Rendy, Yuxing Fei, Rishi E. Kumar, Tanjin He, David Milsted, Matthew J. McDermott, Max Gallant, Ekin Dogus Cubuk, and Amil Merchant et al. An autonomous laboratory for the accelerated synthesis of novel materials. *Nature*, 624(7990):86–91, December 2023. doi: 10.1038/s41586-023-06734-w.
11. 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].
12. Juraj Gottweis, Wei-Hung Weng, Alexander Daryin, Tao Tu, Anil Palepu, Petar Sirkovic, Artiom Myaskovsky, Felix Weissenberger, Keran Rong, and Ryutaro Tanno et al. Towards an AI co-scientist, February 2025.
13. Daniil A. Boiko, Robert MacKnight, Ben Kline, and Gabe Gomes. Autonomous chemical research with large language models. *Nature*, 624(7992):570–578, December 2023. doi: 10.1038/s41586-023-06792-0.
14. Abbi Abdel-Rehim, Hector Zenil, Oghenejokpeme Orhobor, Marie Fisher, Ross J. Collins, Elizabeth Bourne, Gareth W. Fearnley, Emma Tate, Holly X. Smith, and Larisa N. Soldatova et al. Scientific hypothesis generation by large language models: laboratory validation in breast cancer treatment. *Journal of The Royal Society Interface*, 22(227):20240674, June 2025. doi: 10.1098/rsif.2024.0674.
15. Gabriel K. Reder, Carl Collins, Abbi Abdel Rehim, Larisa Soldatova, and Ross D. King. LLM-retrieval based scientific knowledge grounding. In *2nd International Workshop on Natural Scientific Language Processing and Research Knowledge Graphs (NSLP 2025)*, 2025.
16. Wanlu Lei, Caterina Fuster-Barceló, Gabriel Reder, Arrate Muñoz-Barrutia, and Wei Ouyang. BioImage.IO Chatbot: a community-driven AI assistant for integrative computational bioimaging. *Nature Methods*, 21(8):1368–1370, August 2024. doi: 10.1038/s41592-024-02370-y.
17. Ali Essam Ghareeb, Benjamin Chang, Ludovico Mitchener, Angela Yiu, Caralyn J. Szostkiewicz, Jon M. Laurent, Muhammed T. Razzak, Andrew D. White, Michaela M. Hinks, and Samuel G. Rodriques. Robin: A multi-agent system for automating scientific discovery, May 2025.
18. Sebastian Lobentanzer, Shaohong Feng, Noah Bruderer, Andreas Maier, Cankun Wang, Jan Baumbach, Jorge Abreu-Vicente, Nils Krehl, Qin Ma, and Thomas Lemberger et al. A platform for the biomedical application of large language models. *Nature Biotechnology*, 43(2):166–169, February 2025. doi: 10.1038/s41587-024-02534-3.
19. Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha. The AI Scientist: Towards Fully Automated Open-Ended Scientific Discovery, September 2024.
20. Hiroaki Kitano. Systems Biology: A Brief Overview. *Science*, 295(5560):1662–1664, March 2002. doi: 10.1126/science.1069492.
21. Marcel Binz, Stephan Alaniz, Adina Roskies, Balazs Aczel, Carl T. Bergstrom, Colin Allen, Daniel Schad, Dirk Wulff, Jevin D. West, and Qiong Zhang et al. How should the advancement of large language models affect the practice of science? *Proceedings of the National Academy of Sciences*, 122(5):e2401227121, February 2025. doi: 10.1073/pnas.2401227121.
22. Gabriel K. Reder, Erik Y. Bjurström, Daniel Brunnsåker, Filip Kronström, Praphapan Lasin, Ievgeniia Tiukova, Otto I. Savolainen, James N. Dodds, Jody C. May, and John P. Wikswo et al. AutonoMS: Automated Ion Mobility Metabolomic Fingerprinting. *Journal of the American Society for Mass Spectrometry*, 35(3):542–550, March 2024. doi: 10.1021/jasms.3c00396.
23. Ross D. King, Maria Liakata, Chuan Lu, Stephen G. Oliver, and Larisa N. Soldatova. On the formalization and reuse of scientific research. *Journal of The Royal Society Interface*, 8(63):1440–1448, April 2011. doi: 10.1098/rsif.2011.0029.
24. 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. doi: 10.1021/acs.analchem.6b04889.
25. Jody C. May, Cody R. Goodwin, Nichole M. Lareau, Katrina L. Leaptrot, Caleb B. Morris, Ruwan T. Kurulugama, Alex Mordehai, Christian Klein, William Barry, and Ed Darland et al. 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. doi: 10.1021/ac4038448.
26. Ashwin Srinivasan. The Aleph Manual, 2007.
27. 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.
28. J. Michael Cherry, Eurie L. Hong, Craig Amundsen, Rama Balakrishnan, Gail Binkley, Esther T. Chan, Karen R. Christie, Maria C. Costanzo, Selina S. Dwight, and Stacia R. Engel et al. Saccharomyces Genome Database: the genomics resource of budding yeast. *Nucleic Acids Research*, 40(Database issue):D700–705, January 2012. doi: 10.1093/nar/gkr1029.
29. 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: ad-

vances in genome annotation, expanded biochemical pathways, and other key enhancements. *Genetics*, 229(3):iyae185, March 2025. doi: 10.1093/genetics/iyae185.

30. 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, and Albert Tafur Rangel et al. Yeast9: a consensus genome-scale metabolic model for S. cerevisiae curated by the community. *Molecular Systems Biology*, 20(10):1134–1150, October 2024. doi: 10.1038/s44320-024-00060-7.
31. 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. doi: 10.1093/nar/gkv1277.
32. 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. doi: 10.1093/nar/gkv1031.
33. 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.
34. OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, and Sam Altman et al. GPT-4 Technical Report, March 2024. arXiv:2303.08774 [cs].
35. 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.
36. Stephen Muggleton, Luc De Raedt, David Poole, Ivan Bratko, Peter Flach, Katsumi Inoue, and Ashwin Srinivasan. ILP turns 20. *Machine Learning*, 86(1):3–23, January 2012. doi: 10.1007/s10994-011-5259-2.
37. Stefan Kramer, Nada Lavrač, and Peter Flach. Propositionalization Approaches to Relational Data Mining. In *Relational Data Mining*, pages 262–91. Springer, Berlin, Heidelberg, 2001. ISBN 978-3-662-04599-2.
38. 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.
39. Marcus R. Munafò, Brian A. Nosek, Dorothy V. M. Bishop, Katherine S. Button, Christopher D. Chambers, Nathalie Percie du Sert, Uri Simonsohn, Eric-Jan Wagenmakers, Jennifer J. Ware, and John P. A. Ioannidis. A manifesto for reproducible science. *Nature Human Behaviour*, 1(1):0021, January 2017. doi: 10.1038/s41562-016-0021.
40. John P. A. Ioannidis. Why Most Published Research Findings Are False. *PLOS Medicine*, 2 (8):e124, August 2005. doi: 10.1371/journal.pmed.0020124.
41. 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.
42. 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.
43. 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.
44. 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.
45. 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.
46. 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. doi: 10.3389/fphar.2022.805782.
47. 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. doi: 10.1093/bioinformatics/btae050.
48. 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.
49. 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.
50. 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.
51. 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.
52. 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.
53. 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.
54. 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.
55. 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. doi: 10.1038/nbt.2442.
56. 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. doi: 10.1080/01621459.1988.10478639.
57. Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, and Nathaniel J. Smith et al. Array programming with NumPy. *Nature*, 585(7825):357–362, September 2020. doi: 10.1038/s41586-020-2649-2.
58. 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. doi: 10.1021/acs.analchem.3c02267.
59. Kendra J. Adams, Brian Pratt, Neelanjan Bose, Laura G. Dubois, Lisa St. John-Williams, Kevin M. Perrott, Karina Ky, Pankaj Kapahi, Vagisha Sharma, and Michael J. MacCoss et al. Skyline for Small Molecules: A Unifying Software Package for Quantitative Metabolomics. *Journal of Proteome Research*, 19(4):1447–1458, April 2020. doi: 10.1021/acs.jproteome.9b00640.
60. 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. doi: 10.1021/ac051632c.
61. 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.
62. 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. doi: 10.1021/acs.analchem.3c03660.
63. Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, and Jonathan Bright et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature Methods*, 17 (3):261–272, March 2020. doi: 10.1038/s41592-019-0686-2.
64. 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. doi: 10.1371/journal.pcbi.1005752.
65. Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, and Vincent Dubourg et al. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12(85): 2825–2830, 2011.
66. Anita Bandrowski, Ryan Brinkman, Mathias Brochhausen, Matthew H. Brush, Bill Bug, Marcus C. Chibucos, Kevin Clancy, Mélanie Courtot, Dirk Derom, and Michel Dumontier et al. The Ontology for Biomedical Investigations. *PLOS ONE*, 11(4):e0154556, April 2016. doi: 10.1371/journal.pone.0154556.

Supplementary Information

Agentic AI Integrated with Scientific Knowledge: Laboratory Validation in Systems Biology

Daniel Brunnsåker[1], Alexander H. Gower[1], Prajakta Naval[2], Erik Y. Bjurström[2], Filip Kronström[1], Ievgeniia A. Tiukova[2,3], Ross D. King[1,4]

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

December 4, 2025

Supplementary Note 1: Overview of hypothesis space

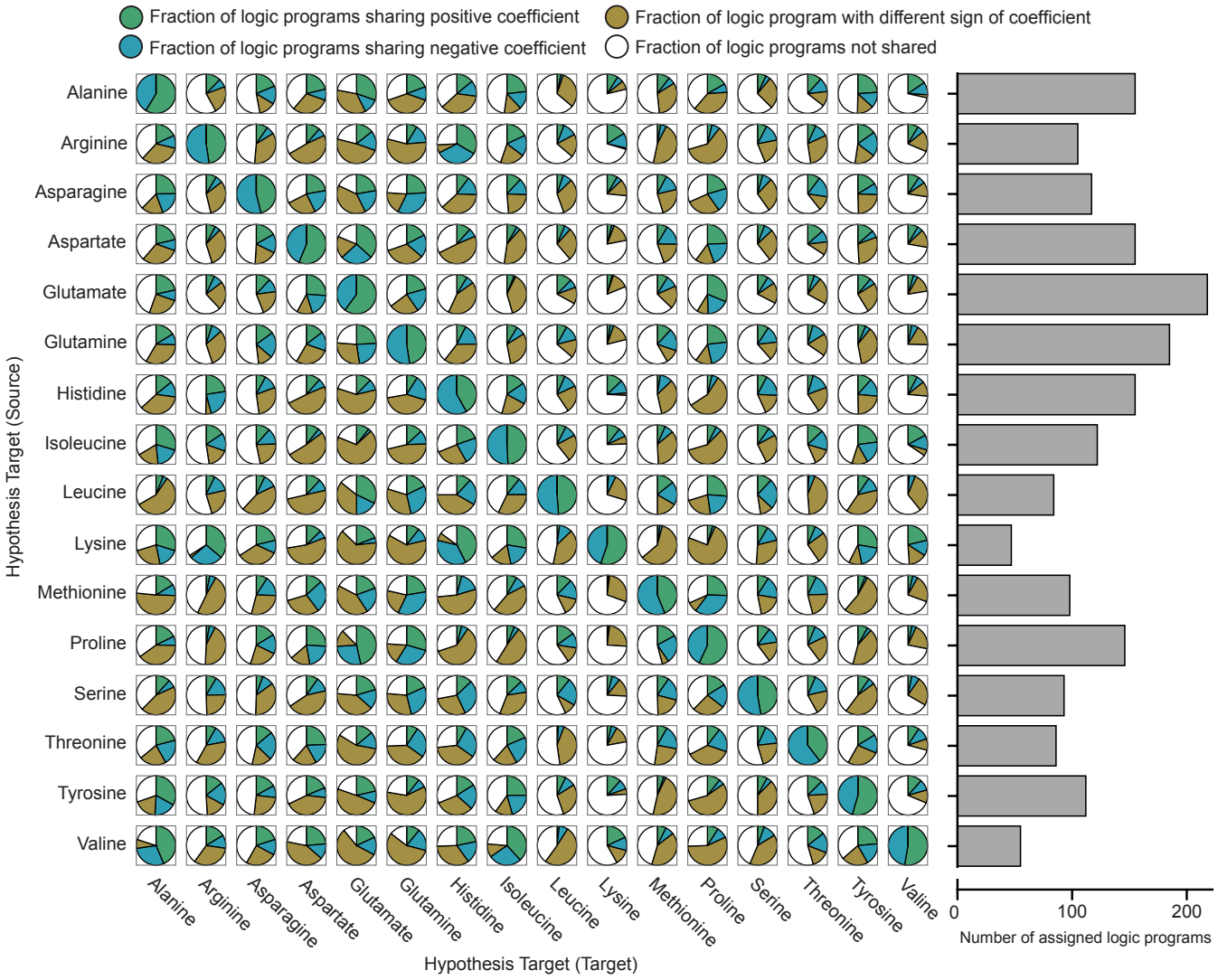

Fig. 6. Asymmetric hypothesis-space overlap between pairs of amino acids. Each pie chart shows—as a fraction of the total number of logic programs that are assigned to the amino acid on the vertical axis—the fraction of logic programs that also are assigned to the amino acid on the horizontal axis. For 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.

Supplementary Note 2: Cultivation procedures

After plate preparation in a Hamilton Microlab Star, growth profiling was performed using the automated laboratory cell "Eve". The plate is transferred from Eve's Cytomat™ automated incubator (30°C) to a Teleshaker Magnetic Shaking System, where it is shaked for 30s at 800rpm, divided evenly between clockwise and counter-clockwise double-orbital shaking. After shaking the plate is transferred to a BMG Polarstar plate reader, where it undergoes optical density measurements at 600nM (the temperature in the plate reader was kept at a constant 30°C). After measuring, the plate is returned to the incubator. The protocol is automatically repeated every 20min for 20h (or the time specified in the generated protocol). At experiment termination, the microwell plate is transferred to an Agilent Bravo liquid handling station, depending on user preference.

Runtime-logs were saved for all of the cultivations done in the study. Note that in some cases several plates were run in parallel, slightly complicating the interpretation of the logs. These records—and others—can be found in the accompanying GitHub repository or by querying the graph database.

Supplementary Note 3: Allowed relations & Search parameters

In order to extract valid logic programs from the constructed database, relational learning was applied in the form of frequent pattern mining using a simplified version of the data-mining algorithm WARMR in aleph (version 5, https://www.cs.ox.ac.uk/activities/programinduction/Aleph/aleph.html). The search was performed using the deletant strains provided in (33) as positive examples. WARMR is a data mining algorithm that uses a level-wise search algorithm, iteratively adding logical conditions until some constraint (defined by the search parameters) has been achieved. When utisiling WARMR the search is restricted to a subset of first-order logic (due to the difficulties of propositionalising higher-order patterns).

The relations allowed in the parameter search are the following:

$$\begin{aligned}
&\mathsf{Genotype}(+\mathsf{ORF})\\
&\mathsf{ExhibitsPhenotype}(+\mathsf{ORF},\ \#\mathsf{State},\ -\mathsf{Value},\ -\mathsf{Cond}),\\
&\mathsf{CompoundName}(+\mathsf{Value},\ \#\mathsf{Name}),\\
&\mathsf{CompoundModulatesTarget}(+\mathsf{Value},\ \#\mathsf{Action},\ -\mathsf{ORF}),\\
&\mathsf{ParticipatesInMetabolism}(+\mathsf{ORF},\ \#\mathsf{Type},\ \#\mathsf{Metabolite}),\\
&\mathsf{Condition}(+\mathsf{Cond},\ \#\mathsf{Description})
\end{aligned}$$

The parameters used for the pattern-search in Aleph, can be seen below, along with a short description.

Parameter	**Value**	**Description**
i	20	Upper bound on layers of new variables
clauselength	4	Upper bound on number of literals in an acceptable clause
search	ar	Uses WARMR as the basis of the search
pos_fraction	0.001	Rules must cover at least this fraction of the total examples
max_features	4097	Upper bound on the maximum number of boolean features
noise	inf	Upper bound on the number of negative examples allowed
nodes	200000	Upper bound on the nodes to be explored when searching for an acceptable clause

Supplementary Note 4: Parameters for metabolomics acquisition

The RapidFire method used typically involved a sample sipping time of 600ms, followed by a loading-phase lasting 4000ms. It was then followed by a 7000ms sample elution into the MS, and lastly, 1500ms for equilibration. The sipper was washed between injections in organic (Methanol) and acqeous (Water) solvents. One injection has an analysis-time of 13.1 seconds.

Mobile phases were prepared with 10 mM ammonium formate and 0.176% formic acid. Phase A comprised 90% acetonitrile, 5% methanol and 5% water, while Phase B comprised 50% acetonitrile, 5% methanol and 45% water. The protocol in use was a modified version of the one derived from Mülleder et al. (33). Mass spectrometry data was collected in IM-QTOF mode with 4-bit multiplexed introduction of the ion packets into the drift tube. The Agilent 6560 was operated in the 100–1700 m/z range at a frame rate of 1.1 frames/s and a gas temperature of 325 °C. Analysed adducts include $[M+H]^+$, $[M+Na]^+$, $[M-H_2O+H]^+$ and $[M+NH_4]^+$. The mass spectrometer was run in positive mode.

Table 2. The Agilent 6560-IM-MS acquisition parameters used for the data collection in positive mode.

Parameter	Value
Ion Source	Agilent Dual AJS ESI
MS Abs. Threshold	200
Ion Mobility Mode	IMS QTOF
Component Model	G6560B
MS Rel. Threshold (%)	0.01
Min Range (m/z)	50
Max Range (m/z)	1700
Frame Rate (frames/sec)	1.1
IM Transient Rate (transients/frame)	14
Max Drift Time (ms)	60
Trap Fill Time (μs)	3900
Trap Release Time (μs)	100
Gas Temperature (C)	325
Gas Flow (l/min)	11
Nebulizer (psi)	45
Sheath Gas Temp (C)	275
Sheath Gas Flow (l/min)	12
VCap (V)	4000
Nozzle Voltage (V)	1000
Drift Tube Entrance Voltage (V)	1274
Drift Tube Exit Voltage (V)	224

Supplementary Note 5: AutonoMS parameters

Table 3. AutonoMS method parameters. Encompasses both RapidFire method settings, preprocessing parameters and peak picking parameters.

Parameter	Value
Sipper Height (mm)	1
Wash Between Sips	1
No. Flushes After Plates	0
Pump1FlowRate (ml/min)	1.25
Pump2FlowRate (ml/min)	0.01
Pump3FlowRate (ml/min)	1.25
Pump 1 [B, C, D] Line %	[100, 0, 0]
Pump 2 [B, C, D] Line %	[100, 0, 0]
Pump 3 [B, C, D] Line %	[100, 0, 0]
Plate Configuration	96-well-plate
Missed Sip Tolerance	10000
Aspirate Cycle Duration (ms)	600
Load/Wash Cycle Duration (ms)	4000
Extra Wash Cycle Duration (ms)	0
Elute Cycle Duration (ms)	7000
Reequilibrate Cycle Duration (ms)	1500
Chromatography/infusion moving average	3
Minimum pulse coverage (%)	50
Moving average window smoothing size (drift)	3
Signal intensity lower threshold	5
Resolving power (IM)	30
Resolving power (TOF)	30000
High selectivity extraction	Yes
Method match tolerance (m/z)	0.005

Supplementary Note 6: Example hypothesis in description logic

Hypothesis 1 ($\downarrow$ Arginine $\implies$ $\uparrow$ Caffeine resistance).

A. DL Axioms.

$$P_{ref1} \sqcup P_1 \sqsubseteq \text{`chemical compound accumulation'} \sqcap \exists \textit{accumulationOfChemical}.\text{`arginine'} \quad \textbf{(1)}$$

$$P_1 \sqsubseteq \exists \textit{decreasedComparedTo}.P_{ref1} \quad \textbf{(2)}$$

$$P_{ref2} \sqcup P_2 \sqsubseteq \text{`resistance to chemicals'} \sqcap \exists \textit{resistanceToChemical}.\text{`caffeine'} \quad \textbf{(3)}$$

$$P_2 \sqsubseteq \exists \textit{increasedComparedTo}.P_{ref2} \quad \textbf{(4)}$$

$$S_{ref} \sqcup S_1 \sqcup S_2 \sqsubseteq \textit{organismState} \quad \textbf{(5)}$$

$$S_{ref} \sqsubseteq \exists \textit{stateHasObservable}.P_{ref1} \sqcap \exists \textit{stateHasObservable}.P_{ref2} \quad \textbf{(6)}$$

$$S_1 \sqsubseteq \exists \textit{stateHasObservable}.P_1 \quad \textbf{(7)}$$

$$S_2 \sqsubseteq \exists \textit{stateHasObservable}.P_2 \quad \textbf{(8)}$$

$$S_1 \sqsubseteq \exists \textit{implies}.S_2 \quad \textbf{(9)}$$

B. RDF/OWL Statements (TriG format).

```
@prefix hypo:  <http://hypo.project-genesis.io#> .
@prefix obo:   <http://purl.obolibrary.org/obo/> .
@prefix owl:   <http://www.w3.org/2002/07/owl#> .
@prefix rdfs:  <http://www.w3.org/2000/01/rdf-schema#> .
@prefix dct: <http://purl.org/dc/terms/> .

<http://hypo.project-genesis.io/hypothesis-metadata> {
    <http://hypo.project-genesis.io/H-0196d121-3310-771d-ab08-bb5f9f4973d4>
            dct:created  "2025-03-13T15:39:00"^^<http://www.w3.org/2001/XMLSchema#dateTime> ;
            dct:creator  "Genesis" .
}

<http://hypo.project-genesis.io/H-0196d121-3310-771d-ab08-bb5f9f4973d4> {
    hypo:S-0196d121-3400-715f-95a9-0430c564ffdb
            rdfs:subClassOf  [ a                  owl:Restriction ;
                               owl:onProperty     hypo:implies ;
                               owl:someValuesFrom hypo:S-0196d121-3500-7ad8-bf51-ec89040bdb84
                             ] .
}

<http://hypo.project-genesis.io/states> {
    hypo:S-REF-0196d120-c332-7b22-b220-572b040a96a8
            rdfs:subClassOf  [ a                  owl:Restriction ;
                               owl:onProperty     hypo:stateHasObservable ;
                               owl:someValuesFrom hypo:P-REF-0196d121-33e2-76a8-b698-330c8911c9ae
                             ] ;
            rdfs:subClassOf  [ a                  owl:Restriction ;
                               owl:onProperty     hypo:stateHasObservable ;
                               owl:someValuesFrom hypo:P-REF-0196d121-34ed-799c-b5ea-296d85f551ce
                             ] .

    hypo:S-0196d121-3500-7ad8-bf51-ec89040bdb84
            rdfs:subClassOf  hypo:organismState ;
            rdfs:subClassOf  [ a                  owl:Restriction ;
                               owl:onProperty     hypo:stateHasObservable ;
                               owl:someValuesFrom hypo:P-0196d121-34f7-7dc8-a499-7a7664f3e710
                             ] .

    hypo:S-0196d121-3400-715f-95a9-0430c564ffdb
            rdfs:subClassOf  hypo:organismState ;
            rdfs:subClassOf  [ a                  owl:Restriction ;
                               owl:onProperty     hypo:stateHasObservable ;
                               owl:someValuesFrom hypo:P-0196d121-33f6-7e22-b286-acc2a7a6227e
                             ] .
}

<http://hypo.project-genesis.io/phenotypes> {
    hypo:P-REF-0196d121-33e2-76a8-b698-330c8911c9ae
            rdfs:subClassOf  obo:APO_0000095 ;                                  # 'chemical compound accumulation'
            rdfs:subClassOf  [ a                  owl:Restriction ;
                               owl:onProperty     hypo:accumulationOfChemical ;
```

```
                              owl:someValuesFrom  obo:CHEBI_29016                # 'arginine'
                            ] .

    hypo:P-REF-0196d121-34ed-799c-b5ea-296d85f551ce
            rdfs:subClassOf  obo:APO_0000087 ;                                   # 'resistance to chemicals'
            rdfs:subClassOf  [ a                  owl:Restriction ;
                               owl:onProperty     hypo:resistanceToChemical ;
                               owl:someValuesFrom  obo:CHEBI_27732               # 'caffeine'
                             ] .

    hypo:P-0196d121-33f6-7e22-b286-acc2a7a6227e
            rdfs:subClassOf  obo:APO_0000095 ;                                   # 'chemical compound accumulation'
            rdfs:subClassOf  [ a                  owl:Restriction ;
                               owl:onProperty     hypo:accumulationOfChemical ;
                               owl:someValuesFrom  obo:CHEBI_29016               # 'arginine'
                             ] ;
            rdfs:subClassOf  [ a                  owl:Restriction ;
                               owl:onProperty     hypo:decreasedComparedTo ;
                               owl:someValuesFrom  hypo:P-REF-0196d121-33e2-76a8-b698-330c8911c9ae
                             ] .

    hypo:P-0196d121-34f7-7dc8-a499-7a7664f3e710
            rdfs:subClassOf  obo:APO_0000087 ;                                   # 'resistance to chemicals'
            rdfs:subClassOf  [ a                  owl:Restriction ;
                               owl:onProperty     hypo:resistanceToChemical ;
                               owl:someValuesFrom  obo:CHEBI_27732               # 'caffeine'
                             ] ;
            rdfs:subClassOf  [ a                  owl:Restriction ;
                               owl:onProperty     hypo:increasedComparedTo ;
                               owl:someValuesFrom  hypo:P-REF-0196d121-34ed-799c-b5ea-296d85f551ce
                             ] .
}
```

Supplementary Note 7: Growth statistics from performed investigations

Term	Estimate	CI 2.5%	CI 97.5%	% Change	p-value
Intercept	11.0678	11.0075	11.1247		0.0002
Caffeine	-0.7636	-0.8366	-0.6883	-53.4013	0.0002
L-Arginine (per mM)	-0.02573	-0.04959	-0.003280	-2.5404	0.01180
L-Arginine at 5.0 mM	-0.1287	-0.2479	-0.01640	-12.07287	0.01180
Caffeine:L-Arginine (per mM)	-0.2103	-0.2507	-0.1707	-18.9652	0.0002
Caffeine:L-Arginine (at 5.0 mM)	-1.05146	-1.2535	-0.8533	-65.05719	0.0002
L-Alanine (at 5.0 mM)	-0.1000	-0.2171	0.02621	-9.5239	0.05179
Caffeine:L-Alanine (at 5.0 mM)	-0.3794	-0.6010	-0.1589	-31.5732	0.0010

Table 4. Summary statistics of *Hypothesis 1* (L-Arginine, caffeine and L-Alanine) with AUC of the time-series growth curve as the dependent variable in different treatment conditions.

Term	Estimate	CI 2.5%	CI 97.5%	% Change	p-value
Intercept	9.2949	9.1364	9.4773		0.0002
Spermine	-0.5147	-0.7362	-0.3152	-40.2329	0.0002
L-Glutamate (per mM)	0.0043	-0.0164	0.0251	0.4327	0.3413
L-Glutamate (at 10.0 mM)	0.0432	-0.1643	0.2507	4.4117	0.3413
Spermine:L-Glutamate (per mM)	-0.0469	-0.0766	-0.0182	-4.5859	0.0012
Spermine:L-Glutamate (at 10.0 mM)	-0.4694	-0.7658	-0.1822	-37.4649	0.0012
L-Alanine (at 10.0 mM)	-0.5499	-0.8338	-0.3010	-42.2992	0.0004
Spermine:L-Alanine (at 10.0 mM)	0.3750	-0.0216	0.7841	45.5027	0.0322

Table 5. Summary statistics of *Hypothesis 2* (L-Glutamate, spermine and L-Alanine) with AUC of the time-series growth curve as the dependent variable in different treatment conditions.

Term	Estimate	CI 2.5%	CI 97.5%	% Change	p-value
Intercept	10.7106	10.6006	10.8087		0.0002
Formic acid	-1.4950	-1.6995	-1.2983	-77.5751	0.0002
L-Glutamate (per mM)	0.0236	0.0110	0.0374	2.3924	0.0010
L-Glutamate (at 10.0 mM)	0.2364	0.1105	0.3739	26.6715	0.0010
Formic acid:L-Glutamate (per mM)	0.0910	0.0640	0.1182	9.5214	0.0002
Formic Acid:L-Glutamate (at 10.0 mM)	0.9095	0.6398	1.1822	148.3068	0.0002
L-Arginine (at 10.0 mM)	-0.2302	-0.3673	-0.0894	-20.5651	0.0008
Formic acid:L-Arginine (at 10.0 mM)	1.7698	1.5185	2.0208	486.9951	0.0002

Table 6. Summary statistics of *Hypothesis 3* (L-Glutamate, formic acid and L-Arginine) with AUC of the time-series growth curve as the dependent variable in different treatment conditions.

Term	Estimate	CI 2.5%	CI 97.5%	% Change	p-value
Intercept	10.8973	10.8142	11.0033		0.0002
30% sucrose	-1.9001	-2.0451	-1.7550	-85.0445	0.0002
L-Lysine (per mM)	-0.0828	-0.1024	-0.0637	-7.9499	0.0004
L-Lysine (at 10.0 mM)	-0.8284	-1.0240	-0.6370	-56.3241	0.0004
30% sucrose:L-Lysine (per mM)	0.0856	0.0627	0.1096	8.9324	0.0004
30% sucrose:L-Lysine (at 10.0 mM)	0.8556	0.6266	1.0958	135.2731	0.0004
L-Valine (at 10.0 mM)	-0.0853	-0.2400	0.0606	-8.1725	0.1230
30% sucrose:L-Valine (at 10.0 mM)	0.2030	-0.0287	0.4374	22.5089	0.0428

Table 7. Summary statistics of *Hypothesis 4* (L-Lysine, sucrose and L-Valine) with AUC of the time-series growth curve as the dependent variable in different treatment conditions.

Term	Estimate	CI 2.5%	CI 97.5%	% Change	p-value
Intercept	10.7598	10.6911	10.8380		0.0002
Acetic acid	-1.4278	-1.7564	-1.1741	-76.0175	0.0002
L-Glutamine (per mM)	-0.0078	-0.0193	0.0031	-0.7770	0.0824
L-Glutamine (at 10.0 mM)	-0.0780	-0.1926	0.0309	-7.5039	0.0824
Acetic acid:L-Glutamine (per mM)	-0.0437	-0.0784	-0.0049	-4.2788	0.0152
Acetic acid:L-Glutamine (at 10.0 mM)	-0.4373	-0.7840	-0.0490	-35.4223	0.0152
L-Leucine (at 10.0 mM)	-0.9040	-1.0583	-0.7693	-59.5059	0.0002
Acetic acid:L-Leucine (at 10.0 mM)	-0.4604	-0.8109	-0.0443	-36.8946	0.0166

Table 8. Summary statistics of *Hypothesis 5* (L-Glutamine, acetic acid and L-Leucine) with AUC of the time-series growth curve as the dependent variable in different treatment conditions.

Term	Estimate	CI 2.5%	CI 97.5%	% Change	p-value
Intercept	10.9362	10.8886	10.9845		0.0002
Lithium chloride	-1.8723	-1.9975	-1.7507	-84.6234	0.0002
L-Arginine (per mM)	-0.0417	-0.0607	-0.0232	-4.0813	0.0002
L-Arginine (at 5.0 mM)	-0.2083	-0.3034	-0.1159	-18.8076	0.0002
Lithium chloride:L-Arginine (per mM)	-0.0017	-0.0377	0.0352	-0.1678	0.4697
Lithium chloride:L-Arginine (at 5.0 mM)	-0.0084	-0.1887	0.1760	-0.8360	0.4697
L-Alanine (at 5.0 mM)	-0.1017	-0.1818	-0.0269	-9.6668	0.0048
Lithium chloride:L-Alanine (at 5.0 mM)	-0.1726	-0.3571	0.0140	-15.8520	0.0358

Table 9. Summary statistics of *Hypothesis 6* (L-Arginine, lithium chloride and L-Alanine) with AUC of the time-series growth curve as the dependent variable in different treatment conditions.

Term	Estimate	CI 2.5%	CI 97.5%	% Change	p-value
Intercept	10.6823	10.5761	10.7985		0.0002
(S)-lactic acid	-0.2747	-0.4346	-0.1291	-24.0235	0.0006
L-Proline (per mM)	-0.0064	-0.0144	0.0016	-0.6413	0.0584
L-Proline (at 20.0 mM)	-0.1287	-0.2885	0.0311	-12.0745	0.0584
(S)-lactic acid:L-Proline (per mM)	-0.0053	-0.0170	0.0066	-0.5290	0.1880
(S)-lactic acid:L-Proline (at 20.0 mM)	-0.1061	-0.3394	0.1326	-10.0656	0.1880
L-Alanine (at 20.0 mM)	-0.6429	-0.3365	-38.5959	0.0002	
(S)-lactic acid:L-Alanine (at 20.0 mM)	-0.1883	-0.4411	0.0615	-17.1662	0.0684

Table 10. Summary statistics of *Hypothesis 7* (L-Proline, lactic acid and L-Alanine) with AUC of the time-series growth curve as the dependent variable in different treatment conditions.

Term	Estimate	CI 2.5%	CI 97.5%	% Change	p-value
Intercept	10.9446	10.8981	10.9874		0.0002
Formic acid	-2.0004	-2.2581	-1.7573	-86.4722	0.0002
Aminoadipate (per mM)	-0.2776	-0.2958	-0.2617	-24.2421	0.0002
Aminoadipate (at 5.0 mM)	-1.3881	-1.4788	-1.3085	-75.0460	0.0002
Formic acid:Aminoadipate (per mM)	0.0768	-0.0083	0.1610	7.9840	0.0388
Formic acid:Aminoadipate (at 5.0 mM)	0.3841	-0.0415	0.8052	46.8243	0.0388
L-Proline (at 5.0 mM)	0.0011	-0.0745	0.0770	0.1108	0.4789
Formic acid:L-Proline (at 5.0 mM)	0.0225	-0.7586	0.6890	2.2737	0.4561

Table 11. Summary statistics of *Hypothesis 8* (aminoadipate, formic acid and L-Proline) with AUC of the time-series growth curve as the dependent variable in different treatment conditions.