OXFORD (GIGA)n SCIENCE

GigaScience, 2024, **13**, 1–11
DOI: 10.1093/gigascience/giae083
TECH NOTE

CAT Bridge: an efficient toolkit for gene–metabolite association mining from multiomics data

Bowen Yang [1,2], Tan Meng [1], Xinrui Wang [1], Jun Li [1], Shuang Zhao [3], Yingheng Wang [4], Shu Yi [1], Yi Zhou [1], Yi Zhang [1], Liang Li [2,3,*], and Li Guo [1,*]

[1]Shandong Key Laboratory of Precision Molecular Crop Design and Breeding, Peking University Institute of Advanced Agricultural Sciences, Shandong Laboratory of Advanced Agricultural Sciences in Weifang, Weifang 261325, China
[2]Department of Chemistry, University of Alberta, Edmonton, AB T6G 2G2, Canada
[3]The Metabolomics Innovation Centre, University of Alberta, Edmonton, AB T6G 1C9, Canada
[4]Department of Computer Science, Cornell University, Ithaca, NY 14853, USA
*Correspondence address. Liang Li, Department of Chemistry, University of Alberta, Edmonton, AB T6G 2G2, Canada. E-mail: liang.li@ualberta.ca; Li Guo, Shandong Key Laboratory of Precision Molecular Crop Design and Breeding, Peking University Institute of Advanced Agricultural Sciences, Shandong Laboratory of Advanced Agricultural Sciences in Weifang, Weifang 261325, China. E-mail: li.guo@pku-iaas.edu.cn.

Abstract

Background: With advancements in sequencing and mass spectrometry technologies, multiomics data can now be easily acquired for understanding complex biological systems. Nevertheless, substantial challenges remain in determining the association between gene–metabolite pairs due to the nonlinear and multifactorial interactions within cellular networks. The complexity arises from the interplay of multiple genes and metabolites, often involving feedback loops and time-dependent regulatory mechanisms that are not easily captured by traditional analysis methods.

Findings: Here, we introduce Compounds And Transcripts Bridge (abbreviated as CAT Bridge, available at https://catbridge.work), a free user-friendly platform for longitudinal multiomics analysis to efficiently identify transcripts associated with metabolites using time-series omics data. To evaluate the association of gene–metabolite pairs, CAT Bridge is a pioneering work benchmarking a set of statistical methods spanning causality estimation and correlation coefficient calculation for multiomics analysis. Additionally, CAT Bridge features an artificial intelligence agent to assist users interpreting the association results.

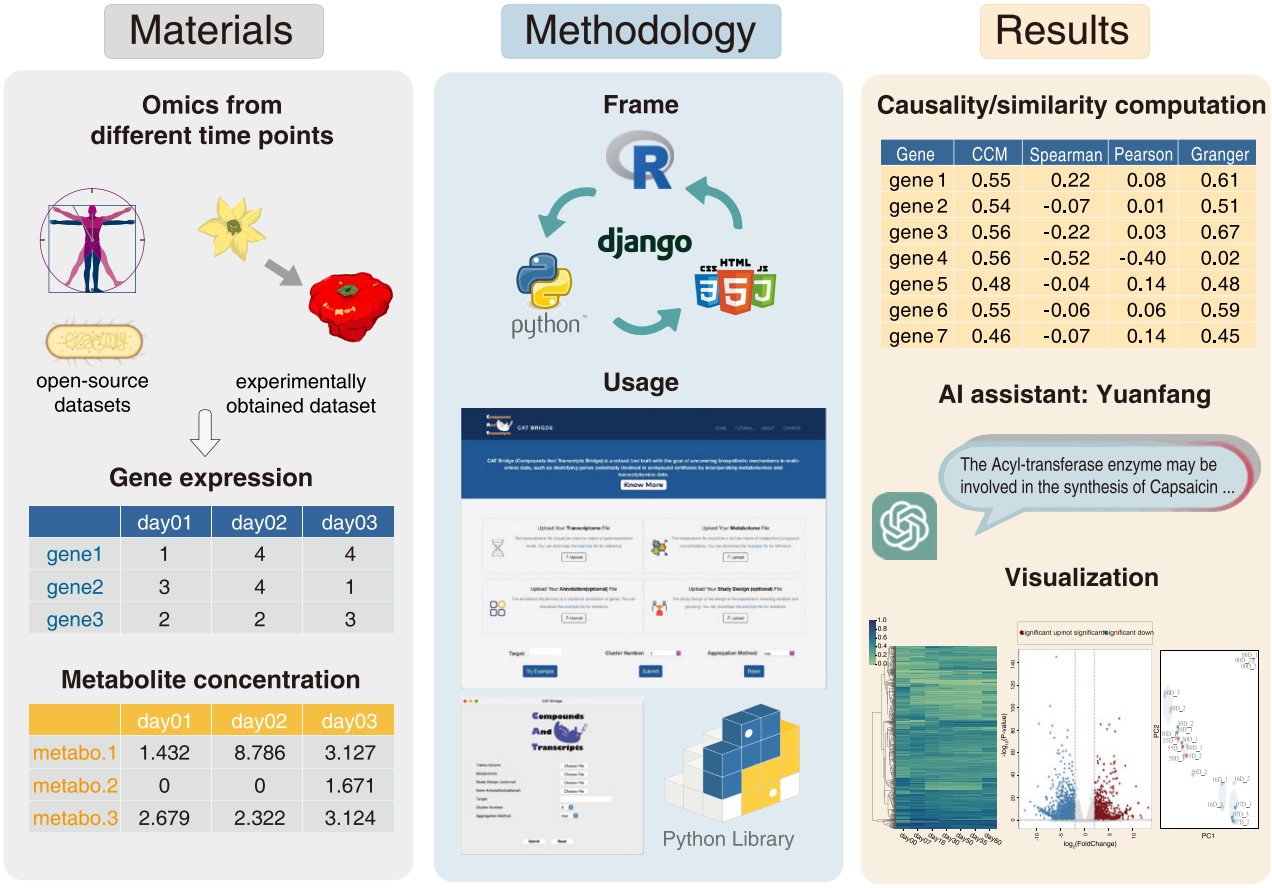
Conclusions: We applied CAT Bridge to experimentally obtained *Capsicum chinense* (chili pepper) and public human and *Escherichia coli* time-series transcriptome and metabolome datasets. CAT Bridge successfully identified genes involved in the biosynthesis of capsaicin in *C. chinense*. Furthermore, case study results showed that the convergent cross-mapping method outperforms traditional approaches in longitudinal multiomics analyses. CAT Bridge simplifies access to various established methods for longitudinal multiomics analysis and enables researchers to swiftly identify associated gene–metabolite pairs for further validation.

Keywords: web server, gene–metabolite association, multiomics, time-series data, causality

Background

Recent advancements in sequencing and mass spectrometry (MS) technologies have made the acquisition of multiomics data more cost-efficient and feasible. Multiomics data analysis is crucial for understanding intricate biological mechanisms from a more comprehensive perspective than single omics, and this holistic analysis allows us to explore the interplay between different molecular levels [1–4]. For integrated data analysis of transcriptomics and metabolomics, a crucial task is to examine the associated gene–metabolite pairs. Existing strategies bifurcate primarily into 2 classes: knowledge-driven approaches and data-driven approaches [5]. Knowledge-driven approaches have shown their inadequacies for nonmodel organisms due to the lack of knowledge, restrictions in revealing *de novo* mechanisms, and difficulties in quantifying and ranking their outcomes [5]. Data-driven strategies mainly depend on statistical methods that model the correlation of gene–metabolite pairs or sophisticated machine learning methods [6, 7]. However, due to the severe batch effects in omics data, machine learning approaches usually lack generalizability [8]. Meanwhile, they are also prone to overfit when applied to relatively small omics datasets, making them harder to transfer to different scenarios and less interpretable [5]. In terms of statistical methods, people usually calculate the correlation coefficient to match gene–compound pairs [9, 10] such as in the studies of the growth cycles of *Solanum lycopersicum* (tomato) [11] and *Oryza sativa* (rice) [12], where Pearson correlations were utilized to study metabolic regulatory networks by integrating transcriptomics and metabolomics data. However, these methods face reliability issues, especially when dealing with longitudinal omics data. This is because of the time lag in the expression of genes and metabolites and the inherent complexity of biological systems, which is a dynamical system with nonlinear interactions between different molecules [4, 13]. A more reliable solution is to use causality, which is inferred based on the ability of one time series (e.g., gene expression) to predict another (e.g., metabolite concentration), to replace the correlation coefficient [14, 15]. Furthermore, purely data-driven strategies can occasionally lead to biologically naive conclusions [9]. For example, a purely data-driven approach might incorrectly link low cholesterol levels with higher mortality rates, suggesting that lower cholesterol is detrimental. However,

Received: April 8, 2024. **Revised:** August 8, 2024. **Accepted:** October 4, 2024



	day01	day02	day03
gene1	1	4	4
gene2	3	4	1
gene3	2	2	3

	day01	day02	day03
metabo.1	1.432	8.786	3.127
metabo.2	0	0	1.671
metabo.3	2.679	2.322	3.124

Gene	CCM	Spearman	Pearson	Granger
gene 1	0.55	0.22	0.08	0.61
gene 2	0.54	-0.07	0.01	0.51
gene 3	0.56	-0.22	0.03	0.67
gene 4	0.56	-0.52	-0.40	0.02
gene 5	0.48	-0.04	0.14	0.48
gene 6	0.55	-0.06	0.06	0.59
gene 7	0.46	-0.07	0.14	0.45

Figure 1: The architecture of the CAT Bridge project. Benchmark data come from publicly available datasets and our experimentally obtained *C. chinense* dataset, including gene expression and metabolite concentration matrices. The construction of CAT Bridge relies on Python, R, HTML/CSS, and JavaScript and provides 3 modes of usage: web server, standalone program, and Python library. To assist in the discovery of associated gene–metabolite pairs, it provides results from 3 perspectives: statistical analysis, AI agent-generated responses, and data visualization.

the actual cause is underlying diseases like cancer, which cause both low cholesterol and increased mortality [16]. Therefore, integrating data-driven and knowledge-driven methodologies may offer a more comprehensive and accurate interpretation of multiomics data.

To address the existing limitations, we have introduced Compounds And Transcripts Bridge (CAT Bridge), a comprehensive cross-platform toolkit that provides a novel analysis pipeline for integrative analysis linking upstream and downstream omics (typically transcriptomics and metabolomics). The novel pipeline encompasses 3 essential steps: data preprocessing, computing association between gene–metabolite pairs, and result presentation. For measuring the association, we integrated 7 different statistical algorithms on causality estimation and correlation coefficient calculation and benchmarked them on human, plant, and microorganism datasets. It also offers 3 ways to display results that are generated from both the data-driven approach and the knowledge-driven approach, including common omics statistical analysis and visualization, heuristic ranking of candidate genes based on causality or correlation, and an artificial intelligence (AI) agent driven by large language models (LLMs) to identify associated gene–metabolite pairs through prior knowledge (Fig. 1).

Besides, we offer 3 different access options for CAT Bridge, including a web server, a standalone application, and a Python library. We also provide a detailed tutorial and a sample dataset to help the users get started easily.

Materials and Methods

Overview of CAT Bridge

The workflow of CAT Bridge consists of 3 primary steps: (i) data processing, (ii) statistical modeling (causality estimation and correlation coefficient calculation), and (iii) visualization and interpretation (Fig. 2A). Users are required to upload 2 processed files, gene expression and metabolite concentration matrices, and specify a metabolite of interest as the target. After data processing, 7 different causality and correlation algorithms are applied to measure the relationship between each gene and the target metabolite. The algorithm chosen by the user will then generate a vector for each gene representing one-to-one associations between the gene and the target metabolite (Fig. 2B). Subsequently, a vector magnitude is employed for heuristic ranking, with the top 100 genes being reviewed by the AI agent, to utilize prior knowledge to offer inspiration to users. Finally, commonly used omics visualization will be employed to assist users in interpreting the overall expression pattern and facilitating the selection of potential candidate genes.

Gene–metabolite association computing

For gene–metabolite pair identification such as inferring the biosynthetic genes for particular metabolites, correlation is often used to imply association, such as the Spearman correlation coefficient (Spearman) and the Pearson correlation coefficient (Pear-

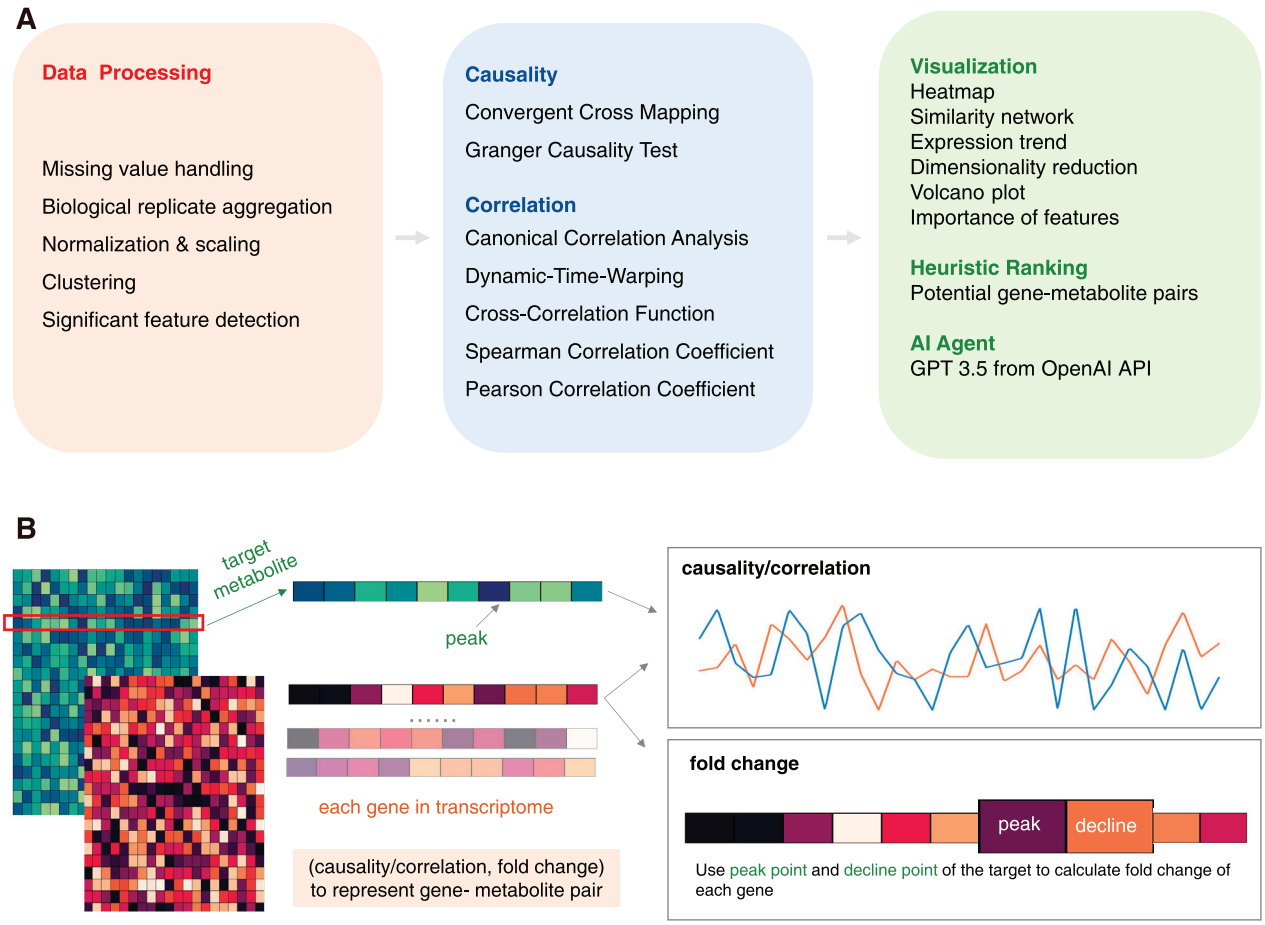


Figure 2: The features and overall workflow of CAT Bridge. (A) The workflow of CAT Bridge consists of 3 primary steps: data preprocessing, estimation of cause–effect relationships or computation of the correlation coefficient, and the presentation of results, which includes visualization, heuristic ranking, and responses from an AI agent. (B) The computation of CAT Bridge involves extracting the target metabolite from the metabolite concentration matrix and pinpointing the time point of its maximum concentration as the peak time. Next, the causality or correlation between each gene in the gene expression matrix and the target metabolite is obtained, along with the fold change between the peak and decline time points to compose a vector to represent the association between gene–metabolite pairs.

son). However, such correlation-based methods have substantial limitations [9, 10] because they overlook the nonlinearity and lag issues of gene expression leading to metabolite changes. Therefore, besides considering Spearman and Pearson, we have integrated into CAT Bridge various distinct statistical methods, including Granger causality (Granger) and convergent cross-mapping (CCM) by evaluating the predictability of metabolite concentration from gene expression to represent causality, as well as canonical correlation analysis (CCA), dynamic time warping (DTW), and cross-correlation function (CCF) for calculating correlation coefficients (the implementation methods are provided in the supplementary text). These algorithms were based on different assumptions so that some of them allow compatibility with time-series data and complex systems (Table 1). Among them, correlation-based strategies have been widely applied in genomics and multiomics analysis [17–20]. The CCM and Granger, which estimate causality from time-series data, are already used in some areas of biology such as ecology and neurobiology but are overlooked in the omics analysis [14, 21–23]. Furthermore, our benchmarking results (as detailed in the Results section) suggest that in longitudinal multiomics studies, causal relationships may provide a more accurate depiction of the association between genes and metabolites.

Table 1: Comparison of statistical methods by linearity and time lag consideration

Statistical method	Linearity	Time lag consideration
Pearson correlation coefficient (Pearson)	Nonlinear	No
Spearman correlation coefficient (Spearman)	Nonlinear	No
Convergent cross-mapping (CCM)	Nonlinear	Yes
Granger causality test (Granger)	Linear	Yes
Canonical correlation analysis (CCA)	Linear	No
Dynamic time warping (DTW)	Nonlinear	Yes
Cross-correlation function (CCF)	Linear	Yes

Heuristic ranking of candidate genes

Fold change (FC) is another measurement frequently used in omics analyses to identify differentially expressed genes [24]. CAT Bridge pinpoints the peak time of the target metabolite and calculates each gene's log2-normalized FC of this peak time point and the subsequent decline time point (i.e., the next sampling time point after the peak time point) using DESeq2 [25]. Then, causality

or correlation and FC are combined into a vector to represent the gene–metabolite pair. After the min–max normalization of values (details are provided in the supplementary text), the magnitude of this vector is calculated as the CAT score (Fig. 2B). This score heuristically ranks the strength of association between each gene and the metabolite. Users can filter putative genes based on thresholds (e.g., 0.5 for causality and correlation, 1 for normalized FC) or manually review them in descending order.

Knowledge-driven approaches and visualization

Optionally, if users provide a gene function annotations file (typically derived from homology annotations using tools like InterProScan [26] or eggNOG-mapper [27] for nonmodel organisms), the CAT score will be adjusted with an additional value. This value is determined by a scoring rule based on the gene's description. By default, genes annotated as enzymes receive a score of 0.2, while those with unknown functions get a score of 0.1. Users can customize this scoring rule based on their specific requirements, depending on the presence of target annotations and their importance. Finally, the top 100 genes in a heuristic ranking will be evaluated by the GPT-3.5 Turbo-based AI agent to identify putative genes on the gene's functional annotation and prior knowledge (implementation methods are provided in the supplementary text). To enhance data interpretation, the CAT Bridge workflow offers a visual ranking of genes based on computation results and also incorporates a spectrum of widely utilized graphical outputs, such as heatmap and principal component analysis (PCA) plot; details can be found in the supplementary text.

Plant material cultivation and sampling

To test the effectiveness of CAT Bridge across different species, especially its applicability to nonmodel organisms, we collected transcriptome sequencing and metabolic profiling data from *Capsicum chinense*, focusing on one of its trademark natural products, capsaicin. *C. chinense* seedlings were divided into 3 groups, each containing 15 seedlings, grown in a greenhouse at the Peking University Institute of Advanced Agricultural Sciences with a controlled environment of 25°C, a light/dark cycle of 16 hours light and 8 hours dark, and 70% relative humidity. The fruits of *C. chinense* were sampled at 7 distinct time points, starting from the day of flowering: 0 day post anthesis (DPA), during which flowers were collected, followed by fruit harvest on days 7, 16, 30, 50, 55, and 60 DPA. For each time point, we sampled 15 fruits from each group of seedlings, yielding a total of 45 samplings per time point. For each group, we utilized 1.0 mL of 70% aqueous methanol per sample, with a sample weight of approximately 50 mg (aside from the 7 DPA samples, which averaged 27.3 mg). The samples were ground and freeze-dried in liquid nitrogen. Each sample was then extracted using 1.0 mL of 70% aqueous methanol for every 50 mg of sample. Following extraction, the samples underwent ultrasonic treatment using an Ultrasonic Cell Disruptor SCIENTZ-IID (Ningbo Scientz Biotechnology) at a frequency of 40 kHz for 10 minutes at room temperature. Standards were prepared as follows: a mixed standard solution, ranging from 20 to 50 μg/mL, was prepared using MS-grade methanol. For the amino acid standard solution, a 1-mg/mL stock solution was initially prepared in water and then diluted with 50% methanol to achieve a final concentration of 50 μg/mL. Three biological sample replicates were utilized in the subsequent transcriptome and metabolome analyses.

Metabolome profiling using liquid chromatography (LC)-MS and data preprocessing

The metabolome profiling was carried out using untargeted metabolomics based on liquid chromatography coupled with mass spectrometry (LC-MS). The samples were filtered through a 0.22-μm membrane and transferred into the lining tube of a sampling vial. Subsequent centrifugation was carried out at 12,000 rcf and 4°C for 10 minutes. The processed samples were then analyzed using Thermo Scientific Orbitrap Exploris 240. Chromatographic separation was achieved on a T3 C18 (1.7 μm, 2.1 × 150 mm column) maintained at 40°C. The mobile phase consisted of A: 1% formic acid in water and B: 1% formic acid in acetonitrile, with a flow rate of 300 μL/min. A 3-μL sample was injected at an autosampler temperature of 10°C. The elution gradient was set as follows: 0–2.5 minutes, 3%–10% B; 2.5–6 minutes, 10%–44% B; 6–14 minutes, 44%–80% B; 14–20 minutes, 80%–95% B; 20–23 minutes, 95% B; 23–23.1 minutes, 95%–3% B; and 23.1–28 minutes, 3% B. MS was performed using both positive and negative ion scans, with a precursor ion scan mode. The auxiliary gas heater temperature was set at 350°C, and the ion transfer tube temperature was also maintained at 350°C. The sheath gas flow rate and auxiliary gas flow rate were set to 35 arb and 15 arb, respectively. The voltages were set to 3.5 KV for the positive spectrum and 3.2 KV for the negative spectrum. For MS1, the scan resolution was 60,000, with a scan range of 80–1,200. For MS2, the scan resolution was 15,000, with a stepped collision energy of 20, 40, and 60 eV. Metabolite identification and quantification were performed using the Compound Discoverer software 3.3 (Thermo Fisher Scientific).

RNA extraction and transcriptome sequencing

Total RNA was isolated from the above-collected plant materials using Trizol Reagent (Thermo Fisher) following the manufacturer-recommended protocol. The quality of RNA extracts was evaluated using the RNA Nano 6000 Assay Kit of the Bioanalyzer 2100 system (Agilent Technologies) following the manufacturer's recommendation, and samples with a RNA integrity number value >7 were used in downstream sequencing library construction and sequencing. The library construction was conducted using Illumina True-seq transcriptome kit following standard protocols. Transcriptome sequencing was carried out by Novogene. The sequencing reads were procured from the Illumina NovaSeq 6000 platform. For preprocessing, fastp [28] was employed to conduct quality control and clean the data. Subsequently, these reads were mapped to the *C. chinense* cultivar PI159236 genome [29] using STAR [30]. StringTie [31] was utilized to quantify and assess the expression levels of the genes that were successfully mapped. The biological function annotation of genes is obtained through the eggNOG-mapper.

Acquisition and processing of public datasets

We also collected human and *Escherichia coli* multiomics data from public datasets to further examine the performance of CAT Bridge.

The human data were sourced from a published aging study [32] that sampled transcriptome and metabolome data from 28 younger (20–25 years) and 54 older (55–66 years) female human donors. Thirteen time points were included in which both the transcriptome and the metabolome data were detected. We obtained the transcript expression levels and the concentrations of glucose and fructose 6-phosphate for 13 donors, and if multiple donor data were available at a single time point, the average value was used.

The dataset for *E. coli* originated from a study on the interactions between metabolites and genes within the bacterium [33]. Researchers manipulated the culture conditions of *E. coli*, alternating between growth and starvation phases. They then collected the transcriptome and metabolome under these varying conditions to systematically explore how metabolites influence transcription factors. The dataset includes 29 time points where both transcriptomic and metabolomic information was concurrently obtained.

Results

CAT Bridge provides a platform with a novel pipeline that allows for the rapid identification of putative genes for further investigation and validation. Three distinct usage modes are offered to cater to a wide range of user requirements. First, it features a web server, designed for user-friendliness and accessibility, and is open to all users without any login requirements. This is particularly beneficial to those who are less familiar with programming languages. Second, a standalone application is available for users handling large data files, as it lifts the constraints on file sizes. Finally, a Python library is available for bioinformaticians with complete features and customizable workflows (implementation of the software is provided in the supplementary text).

To showcase the utility and features of CAT Bridge, we applied it to both an experimentally obtained dataset collected from *C. chinense* and publicly available human and *E. coli* dataset in 2 case studies. Our analysis revealed that in the context of longitudinal multiomics, causality-based strategies tend to outperform those solely based on similarity. As such, we advocate for the adoption of causality rather than similarity in longitudinal multiomics analysis such as co-mining the transcriptomics and metabolomics data. The capsaicin dataset generated in this study has been made open source for user exploration.

Case study 1: Identifying genes associated with capsaicin biosynthesis in *C. chinense*

In our inaugural case study, we leveraged experimentally obtained nonmodel organism data to examine the performance of CAT Bridge. These data comprised the transcriptome and metabolome of *C. chinense* at 7 different developmental stages after bloom (Fig. 3A). Capsaicin, an important natural product produced by *C. chinense* that gives fruit pungency and has potential anticancer and analgesic activity [34], was selected as the target metabolite for this study. Time-series transcriptome and metabolic profiling of developing *C. chinense* fruits were used as input data to test CAT Bridge.

Through examination using the CCM method for hypothetical ranking, BC332_05,016 encoding an acyl-transferase was ranked first, regardless of whether an annotation file was provided. The result suggests that this gene was more likely to be the biosynthetic gene associated with capsaicin in *C. chinense* (Fig. 3B). The complete heuristic ranking results are provided in Supplementary Table S1. BLAST search revealed that BC332_05,016 was homologous of PUN1 (sequence identity: 100%), aka the AT3 (acyltransferase 3) or CS (capsaicin synthesis) gene [35]. Moreover, when common thresholds were applied for screening, only CCM passed the criteria. The causality modeled based on CCM was 0.55, implying a strong association between BC332_05,016 and capsaicin. By contrast, the conventional Pearson correlation method produced a result of 0.08, which would fall below the commonly used threshold and potentially lead to an overlook of this gene–metabolite pair (Fig. 3C). The AI agent also accurately found BC332_05,016 among the top 100 genes based on functional annotation (Fig. 3D). Furthermore, CAT Bridge visualization result showed that capsaicinoids such as nonivamide, dihydrocapsaicin, and homocapsaicin have high similarity to capsaicin (Fig. 3E) and may play a significant role in response to the variable (Fig. 3F). The rest of the visualization results are provided in the Supplementary Figs. S1–S4. These results show that the CAT Bridge is a valuable tool in multiomics analysis to reliably identify associated gene–metabolite pairs.

Case study 2: Identifying association of gene–metabolite pairs in glycolysis

For case study 2, we utilized an open-source multiomics dataset generated by previous aging research [32]. After processing, 13 time points from the skin of younger and older people used for testing, we compared different association modeling results from various statistical methods in CAT Bridge.

Using this dataset, Kuehne et al. [32] found glycolysis altered activity in the upper body when aging, and hexokinase 2 (HK2) and phosphofructokinase (PFKP), 2 enzyme genes, were reduced, whereas fructose bisphosphatase 1 (FBP1) and aldolase A (ALDOA) were increased in older skin. Because glycolysis is well researched, and the study's focus was not identifying gene–metabolite pairs, they only compared the fold changes in genes and metabolites without evaluating the strength of associations between gene–metabolite pairs.

To explore whether different statistical methods can identify known gene–metabolite pairs in glycolysis, we used this dataset to evaluate the relationship between key genes and glucose and fructose 6-phosphate; for fold change, we used the same comparison methods as Kuehne et al. [32], that is, an older group divided by the younger group. The results show that CCM still performed the best, accurately identifying a strong negative correlation between the concentration of glucose and the expression of HK2, a strong positive correlation between fructose 1,6-bisphosphate and PFKP, and a weak negative correlation with ALDOA. However, only Spearman yielded the expected relationship between fructose 1,6-bisphosphate and FBP1 (Fig. 4A, B). Additionally, previous studies did not identify hexokinase 1 (HK1) as a significant gene in the glucose change because its levels in the elderly were slightly higher than in younger individuals. However, CCM correctly identified this relationship. Taking the expression pattern of PFKP and fructose 1,6-bisphosphate as an example, we observed that changes in gene expression were always reflected in the metabolite concentration at the next time point, indicating a delay between metabolite and transcript responses. This illustrates the advantage of causal relationship modeling methods over traditional methods.

Case study 3: Metabolite–transcription factor interactions and acetyl-CoA regulation

In the third study, we utilized a transcriptomic and metabolomic dataset from *E. coli* to demonstrate the capabilities of CAT Bridge in analyzing simple organisms. The dataset contains gene expression and metabolite concentration data across 29 time points, and it was originally generated by Lempp et al. [33] to investigate the allosteric regulation of transcription factors (TFs) by metabolites. They constructed a "literature network" of known metabolite–TF interactions, including 16 interactions as activations of the TF by metabolites. Using kinetic correlation, which accommodates time lags, Lempp et al. recovered 10 out of the 16 activation in-

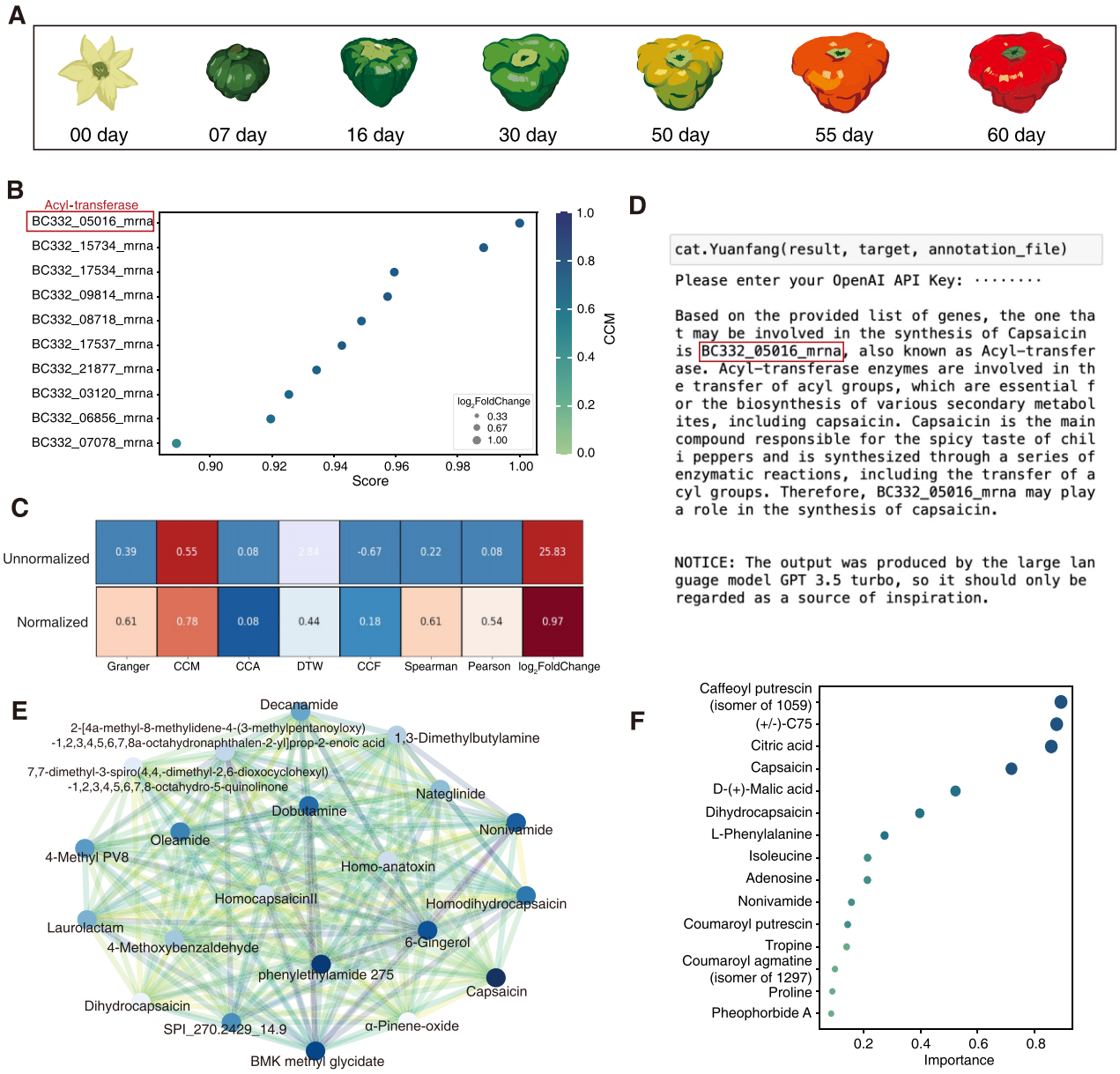


Figure 3: Application of CAT Bridge to mine the transcript–capsaicin association. (A) Diagram showing the time points sampled for transcriptome and metabolic profiling during fruit development of *C. chinense* in case study 1. (B) Heuristic ranking produced using the CCM-based method. (C) Comparative values across different methods. For unnormalized values: red indicates a strong association; original denotes medium association; blue suggests values that are below the commonly used threshold, show no association, or are negatively associated (depending on the method); light blue means this method does not adhere to a common threshold. For normalized values: red signifies values that are high after min–max normalization; blue represents low normalized values. (D) Interpretation results derived from the AI agent. (E) The correlation network of capsaicin. (F) The significance of metabolites.

teractions. We reanalyzed this dataset using CAT Bridge to test its performance with simple organisms. Utilizing the functions integrated within CAT Bridge and setting a time lag of 4, we calculated the association strengths between these molecules. The CCM method once again performed best, reproducing 14 out of 16 interactions.

Focusing on acetyl-CoA as the target metabolite and setting a time lag of 3, we investigated the associated genes. The CCM-based methods fared the best, and the results showed strong causality (greater than 0.9) with the aceE, aceF, and lpd genes, which are involved in the conversion of pyruvate to acetyl-CoA (Fig. 5A) [36]. These genes were also ranked in the top 100 in heuristic sorting and clustered together. Additionally, the acs and pta genes, which are involved in the conversion of acetate to acetyl-CoA, [36], also showed strong causality (0.840 and 0.721, respectively) but did not rank in the top 100 in heuristic sorting (Fig. 5B). The complete heuristic ranking results are provided in Supplementary Table S2. This suggests that in this study, the precursors for acetyl-CoA may predominantly come from pyruvate.

These results demonstrate the CAT Bridge's potential to extract meaningful insights from multiomics data across diverse species. By integrating time-series analysis methods, particularly CCM, it offers superior performance in longitudinal omics compared to common methods.

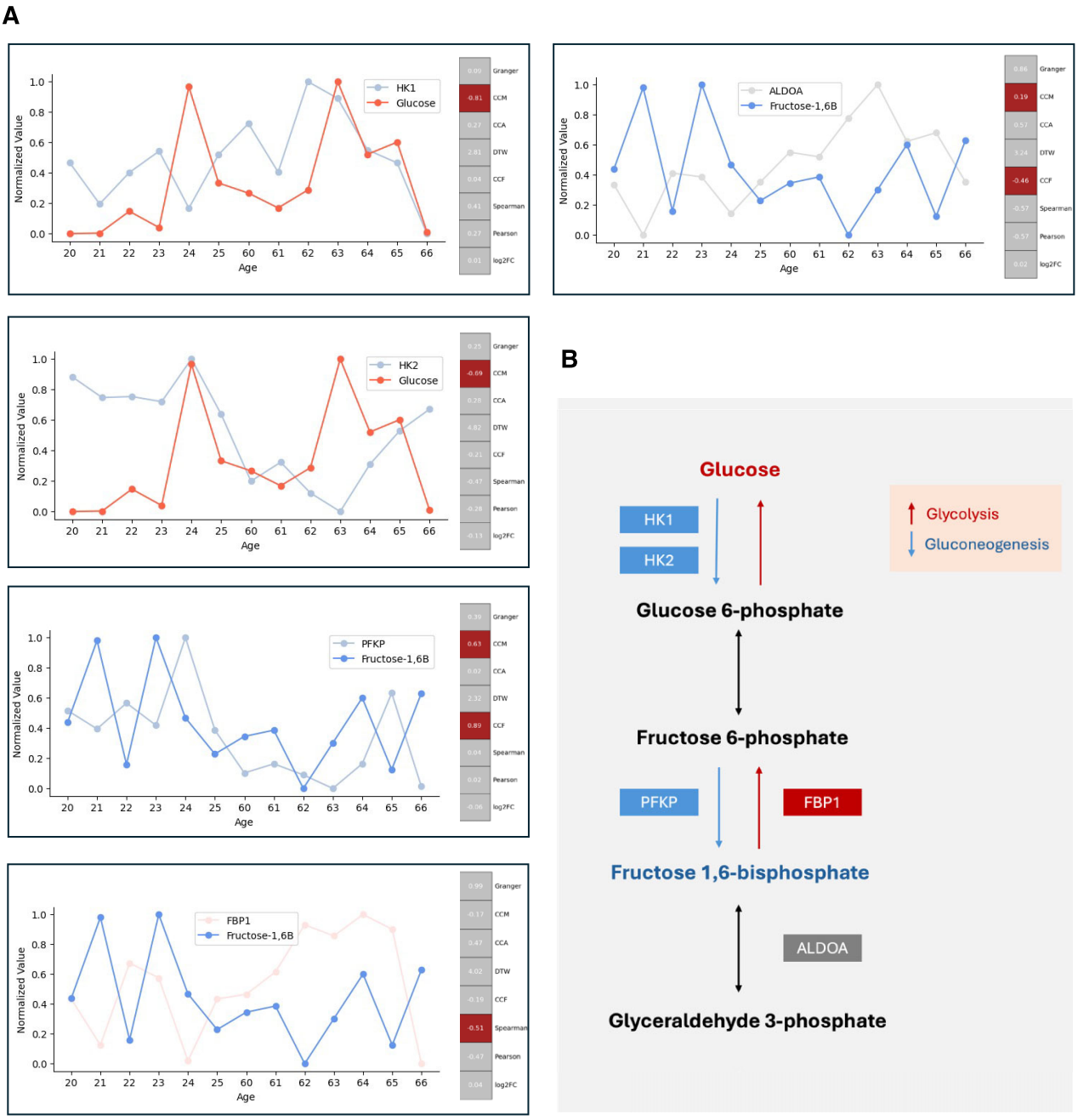


Figure 4: Application of CAT Bridge to mine gene–metabolite associations in the glycolysis pathway. (A) Left side: the expression patterns of functional genes and their corresponding target metabolites. Right side: the strength of associations was evaluated using different methods. Red denotes evaluation outcomes that align with known pathway information under typical thresholds (e.g., a correlation coefficient <−0.5 indicating a strong negative association), whereas gray signifies inconsistent or unsuitable. (B) Part of the glycolysis pathway.

Comparison with other web-based tools

Table 2 displays the function coverage comparisons between CAT Bridge and other data-driven multiomics analysis web-based tools, including OmicsAnalyst [5], 3omics [37], IntLIM [38, 39], and CorDiffViz [40]. In association identification, IntLIM, 3omics, and CorDiffViz integrate either Pearson or Spearman correlations or both to aid in the discovery of feature relationships. What sets CAT Bridge apart is its assembly of various algorithms that handle time-series data and causality, as well as incorporate an AI agent to inspire users. Notably, the performance of CCM has been found from 3 previous case studies to be potentially more suitable for longitudinal multiomics analysis compared to traditional methods.

Discussion

In recent years, there has been a surge in multiomics research. A critical aspect often overlooked in such studies is the unique nature of the longitudinal experimental design. Longitudinal omics analysis is particularly important in research on the developmental cycle of plants and investigations related to chronic diseases and aging [41–44]. However, many studies tend to use generic

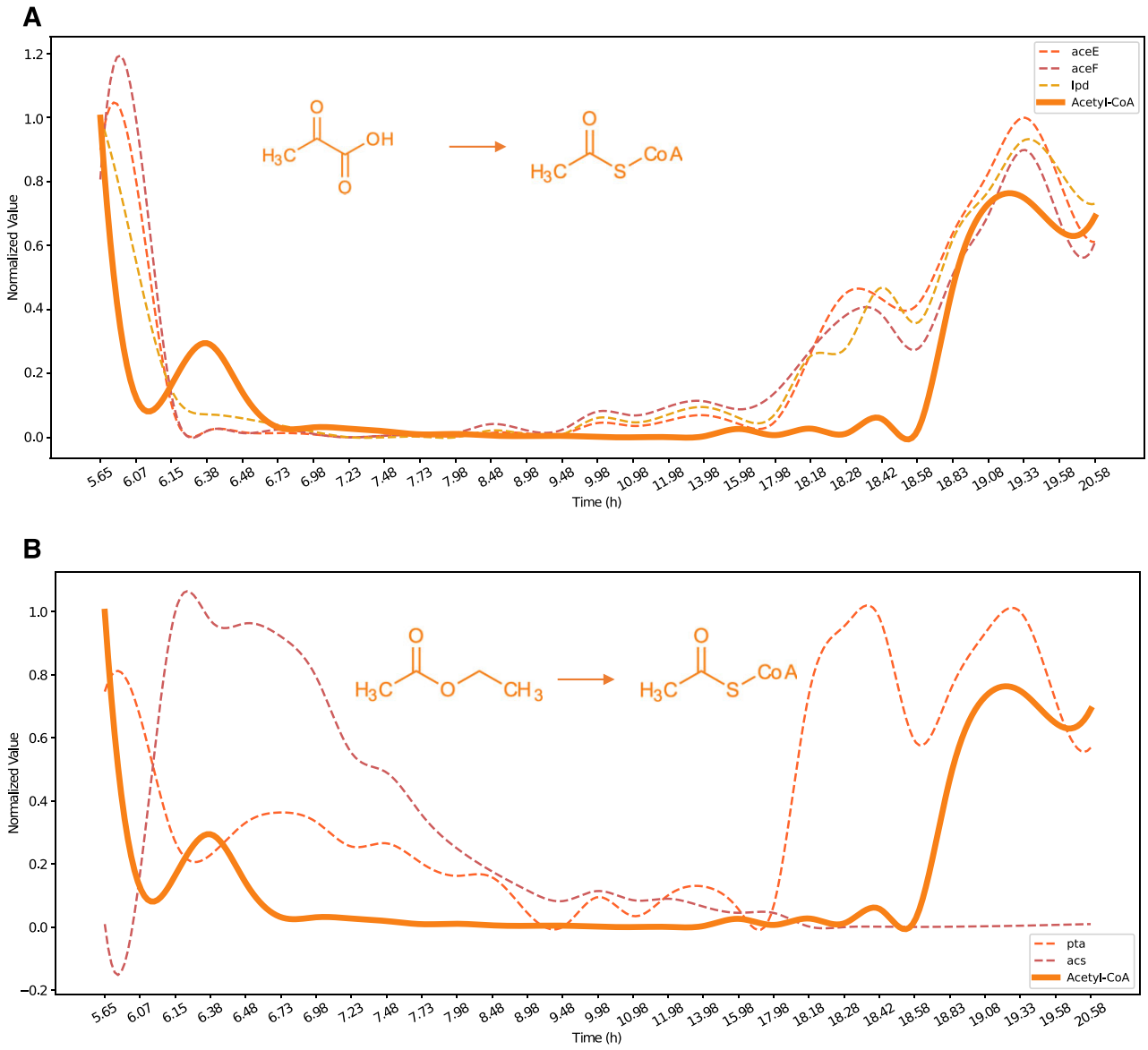


Figure 5: Expression patterns of acetyl-CoA synthesis genes and acetyl-CoA concentration. (A) Genes involved in the conversion of pyruvate to acetyl-CoA. (B) Genes involved in the conversion of acetate to acetyl-CoA.

Table 2: Comparison of CAT Bridge with other web-based multiomics tools

	CAT Bridge	OmicsAnalyst	3omics	IntLIM	CorDiffViz
Preprocessing	✓	✓	×	×	×
Visual analytics	✓	✓	✓	✓	✓
Cross-omics association	✓	×	✓	✓	✓
Cross-platform	✓	×	×	✓	✓
Longitudinal study	✓	×	×	×	×
Third-party database connectivity	×	×	×	×	×
Enrichment analysis	×	×	✓	×	×
AI agent	✓	×	×	×	×

Symbols indicating feature assessments: "✓" denotes presence; "×" signifies absence. OmicsAnalyst [45]; 3omics [46]; IntLIM [47]; CorDiffViz [48].

methodologies for analysis [11, 12]. This may inadvertently miss key discoveries. CAT Bridge provides a platform specifically for longitudinal multiomics analysis by drawing insights from disciplines where time-series data are more prevalent and benchmarking them with data. Through the gene–metabolite causality and correlation modeling method, combined with visualization tools and AI assistance, researchers can more quickly identify putative genes for experimental validation. We believe that discovering associated gene–metabolite pairs will have practical applications in many fields. For example, in metabolic engineering, a single metabolite is often regulated by multiple genes, and by comparing the association strength, one can infer which genes may play a dominant role under certain conditions, facilitating the development of targeted metabolic engineering strategies. Furthermore, in horticulture and breeding, identifying genes with causal relationships to key natural products, particularly in nonmodel organisms, can significantly aid in molecular breeding efforts, potentially leading to the development of new crop varieties with enhanced yields or desired characteristics.

In 3 case studies, CCM showed its superiority compared to other methods, which is probably due to its modeling capability of dynamical systems. Thus, it can better capture the complex nonlinear interactions within biological systems [21]. We advocate for modeling cause-and-effect relationships in longitudinal omics analyses instead of more widely used Pearson or Spearman correlations. However, this does not mean that CCM is always appropriate. Factors such as sampling intervals and the number of samples also need to be considered. More precise methods for estimating causality, as well as postprocessing for a vector representing gene–metabolite pairs, are both required to explore and validate by using more data. We tested the performance of CAT Bridge on the gene-to-metabolite and metabolite-to-gene tasks in this research, and since CAT Bridge is derived by integrating algorithms originating from different fields and its theoretical basis suggests it has generalizability, we believe that it can be used for investigating associations between other different molecular levels (such as gene to protein). Nonetheless, further validating its performance across these different interactions is necessary for the conclusion due to the variations in complexity and regulatory timing at different molecular levels. Additionally, LLM-based AI agents have shown a wide range of applications in various fields, but the hallucination of knowledge deficiency remains an issue that can mislead users [49, 50]. Although we have enhanced the credibility of the AI agent response via an appropriate prompt and low temperature [50–52], this cannot be regarded as a substitute for professional knowledge and experimental verification and merely serves as a starting point for verification. The definitive conclusions still require manual assessment by researchers.

Aside from computational methods, the reliability of analytical results is also influenced by experimental design and data acquisition methods. Increasing the number of sampling time points and setting a reasonable interval between them can enhance the credibility of the results. On the data acquisition front, it is recommended to annotate the transcriptome with an updated, high-quality reference genome and employ advanced metabolomics techniques such as chemical isotope labeling liquid LC-MS [53] to ensure a high coverage and more accurate relative quantification of metabolome.

Future work will involve determining response times between molecules at different levels, visualizing these lagged nonlinear relationships, and constructing molecular networks based on temporal causality and metabolic pathways.

Availability of Source Code and Requirements

Project name: CAT Bridge
Project homepage: https://catbridge.work [54]
GitHub page: https://github.com/Bowen999/CAT-Bridge
Operating system(s): Platform independent
Programming language: Python, R, HTM, CSS, JavaScript
License: CC0 1.0 Public Domain Dedication
Biotools: cat_bridge
RRID: SCR_025410

Additional Files

Supplementary Text 1. Extended experimental procedure.
Supplementary Table S1. Heuristic ranking results of case study 1.
Supplementary Table S2. Heuristic ranking results of case study 3.
Supplementary Fig. S1. PCA plot from the test results of case study 1. (A) PCA plot of transcriptomics. (B) PCA plot of metabolomics. (C) PCA plot of integrated multiomics.
Supplementary Fig. S2. Heatmap from the test results of case study 1. (A) Heatmap of transcriptomics. (B) Heatmap of metabolomics.
Supplementary Fig. S3. Volcano plot (peak vs decline points) from the test result of case study 1.
Supplementary Fig. S4. VIP plot of gene from the test result of case study 1.

Abbreviations

AI: artificial intelligence; CCA: canonical correlation analysis; CCF: cross-correlation function; CCM: convergent cross-mapping; DPA: day postanthesis; DTW: dynamic time warping; FC: fold change; LC: liquid chromatography; LLM: large language model; MS: mass spectrometry; PCA: principal component analysis; TF: transcription factor.

Acknowledgments

We thank the Bioinformatics Platform at Peking University Institute of Advanced Agricultural Sciences for providing the high-performance computing resources.

Author Contributions

B.Y.: conceptualization, software, methodology, visualization, investigation, writing—original draft. T.M.: software. X.W.: visualization. J.L.: investigation. S.Z.: writing—review & editing. Y.W.: methodology, writing—review & editing. S.Y.: investigation. Y. Zhou: software. Y. Zhang: software. L.L.: supervision. L.G.: conceptualization, writing—review & editing, supervision, funding acquisition. All authors have read and agreed to the published version of the manuscript.

Funding

This work was supported by the Key R&D Program of Shandong Province (Grant No. ZR202211070163) and Natural Science Foundation for Distinguished Young Scholars (Grant No. ZR2023JQ010) of Shandong Province. L.G. is also supported by Taishan Scholars Program of Shandong Province.

Data Availability

The CAT Bridge web server [54] is open and free for all users and there is no login requirement. The source code used for CAT Bridge is available on FigShare [55]. An archival copy of the code is available via Software Heritage [56].

The sequencing data for case study 1 is available in the Small Read Archive (SRA) [57] under the BioProject accession code PRJNA1030882, and the metabolome data for this case study have been deposited at the Metabolomics Workbench [58] with the study ID ST003172. The sequencing data of case study 2 were obtained from Gene Expression Omnibus (GEO) [59], with the accession number GSE85358. The sequencing data of case study 3 were obtained from GEO with the accession code GSE131992, and metabolomics data were sourced from the MetaboLights database [60] with the accession code MTBLS1044.

Competing Interests

The authors have declared no competing interests.

References

1. Wörheide MA, Krumsiek J, Kastenmüller G et al. Multi-omics integration in biomedical research—a metabolomics-centric review. Anal Chim Acta 2021;1141:144–62. https://doi.org/10.1016/j.aca.2020.10.038.
2. Hasin Y, Seldin M, Lusis A. Multi-omics approaches to disease. Genome Biol 2017;18(1):83. https://doi.org/10.1186/s13059-017-1215-1.
3. Subramanian I, Verma S, Kumar S et al. Multi-omics data integration, interpretation, and its application. Bioinform Biol Insights 2020;14:1177932219899051. https://doi.org/10.1177/1177932219899051.
4. Eicher T, Kinnebrew G, Patt A, et al. Metabolomics and multi-omics integration: a survey of computational methods and resources. Metabolites 2020;10(5):202. https://doi.org/10.3390/metabo10050202.
5. Zhou G, Ewald J, Xia J. OmicsAnalyst: a comprehensive web-based platform for visual analytics of multi-omics data. Nucleic Acids Res 2021;49(W1):W476–82. https://doi.org/10.1093/nar/gkab394.
6. Krassowski M, Das V, Sahu SK, et al. state of the field in Multi-Omics research: from computational needs to data mining and sharing. Front Genet 2020;11:610798. https://doi.org/10.3389/fgene.2020.610798.
7. Athieniti E, Spyrou GM. A guide to multi-omics data collection and integration for translational medicine. Comput Struct Biotechnol J 2023;21:134–49. https://doi.org/https://doi.org/10.1016/j.csbj.2022.11.050.
8. Albaradei S, Thafar M, Alsaedi A, et al. Machine learning and deep learning methods that use omics data for metastasis prediction. Comput Struct Biotechnol J 2021;19:5008–18. https://doi.org/10.1016/j.csbj.2021.09.001.
9. Cavill R, Jennen D, Kleinjans J, et al. Transcriptomic and metabolomic data integration. Briefings Bioinf 2015;17(5):891–901. https://doi.org/10.1093/bib/bbv090.
10. Chong J, Xia J. Computational approaches for integrative analysis of the metabolome and microbiome. Metabolites 2017;7(4):62. https://doi.org/10.3390/metabo7040062.
11. Li Y, Chen Y, Zhou L, et al. MicroTom metabolic network: rewiring tomato metabolic regulatory network throughout the growth cycle. Mol Plant 2020;13(5):1203–18. https://doi.org/10.1016/j.molp.2020.06.005.
12. Yang C, Shen S, Zhou S, et al. Rice metabolic regulatory network spanning the entire life cycle. Mol Plant 2022;15(2):258–75. https://doi.org/10.1016/j.molp.2021.10.005.
13. Singh KS, van der Hooft JJJ, van Wees SCM, et al. Integrative omics approaches for biosynthetic pathway discovery in plants. Nat Prod Rep 2022;39(9):1876–96. https://doi.org/10.1039/D2NP00032F.
14. Ye H, Deyle ER, Gilarranz LJ, et al. Distinguishing time-delayed causal interactions using convergent cross mapping. Sci Rep 2015;5(1):14750. https://doi.org/10.1038/srep14750.
15. Yuan AE, Shou W. Data-driven causal analysis of observational biological time series. eLife 2022;11:e72518. https://doi.org/10.7554/eLife.72518.
16. Sattar N, Preiss D. Reverse causality in cardiovascular epidemiological research. Circulation 2017;135(24):2369–72. https://doi.org/10.1161/CIRCULATIONAHA.117.028307.
17. Rockwood AL, Crockett DK, Oliphant JR et al. Sequence alignment by cross-correlation. J Biomol Tech 2005;16(4):453–58. https://pubmed.ncbi.nlm.nih.gov/16522868.
18. Skutkova H, Vitek M, Babula P, et al. Classification of genomic signals using dynamic time warping. BMC Bioinf 2013;14(10):S1. https://doi.org/10.1186/1471-2105-14-S10-S1.
19. Seoane JA, Campbell C, Day IN, et al. Canonical correlation analysis for gene-based pleiotropy discovery. PLoS Comput Biol 2014;10(10):e1003876. https://doi.org/10.1371/journal.pcbi.1003876.
20. Jiang MZ, Aguet F, Ardlie K, et al. Canonical correlation analysis for multi-omics: application to cross-cohort analysis. PLoS Genet 2023;19(5):e1010517. https://doi.org/10.1371/journal.pgen.1010517.
21. Yuan AE, Shou W. Data-driven causal analysis of observational biological time series. eLife 2022;11:e72518. https://doi.org/10.7554/eLife.72518.
22. Heerah S, Molinari R, Guerrier S, et al. Granger-causal testing for irregularly sampled time series with application to nitrogen signalling in Arabidopsis. Bioinformatics 2021;37(16):2450–60. https://doi.org/10.1093/bioinformatics/btab126.
23. Stokes PA, Purdon PL. A study of problems encountered in Granger causality analysis from a neuroscience perspective. Proc Natl Acad Sci U S A 2017;114(34):E7063–72. https://doi.org/10.1073/pnas.1704663114.
24. Arora S, Pattwell SS, Holland EC, et al. Variability in estimated gene expression among commonly used RNA-seq pipelines. Sci Rep 2020;10(1):2734. https://doi.org/10.1038/s41598-020-59516-z.
25. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol 2014;15(12):550. https://doi.org/10.1186/s13059-014-0550-8.
26. Ye J, Coulouris G, Zaretskaya I, et al. Primer-BLAST: a tool to design target-specific primers for polymerase chain reaction. BMC Bioinf 2012;13:134. https://doi.org/10.1186/1471-2105-13-134.
27. Cantalapiedra CP, Hernández-Plaza A, Letunic I, et al. eggNOG-mapper v2: functional annotation, orthology assignments, and domain prediction at the metagenomic scale. Mol Biol Evol 2021;38(12):5825–29. https://doi.org/10.1093/molbev/msab293.
28. Chen S, Zhou Y, Chen Y, et al. fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics 2018;34(17):i884–90. https://doi.org/10.1093/bioinformatics/bty560.
29. Kim S, Park J, Yeom SI, et al. New reference genome sequences of hot pepper reveal the massive evolution of plant

disease-resistance genes by retroduplication. Genome Biol 2017;18(1):210. https://doi.org/10.1186/s13059-017-1341-9.
30. Dobin A, Davis CA, Schlesinger F, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics 2013;29(1):15–21. https://doi.org/10.1093/bioinformatics/bts635.
31. Pertea M, Pertea GM, Antonescu CM, et al. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nat Biotechnol 2015;33(3):290–95. https://doi.org/10.1038/nbt.3122.
32. Kuehne A, Hildebrand J, Soehle J, et al. An integrative metabolomics and transcriptomics study to identify metabolic alterations in aged skin of humans in vivo. BMC Genomics 2017;18(1):169. https://doi.org/10.1186/s12864-017-3547-3.
33. Lempp M, Farke N, Kuntz M, et al. Systematic identification of metabolites controlling gene expression in E. coli. Nat Commun 2019;10(1):4463. https://doi.org/10.1038/s41467-019-12474-1.
34. Fattori V, Hohmann MS, Rossaneis AC, et al. Capsaicin: current understanding of its mechanisms and therapy of pain and other pre-clinical and clinical uses. Molecules 2016;21(7):844. https://doi.org/10.3390/molecules21070844.
35. Kim S, Park M, Yeom S-I, et al. Genome sequence of the hot pepper provides insights into the evolution of pungency in Capsicum species. Nat Genet 2014;46(3):270–78. https://doi.org/10.1038/ng.2877.
36. Chiang C-J, Ho Y-J, Hu M-C, et al. Rewiring of glycerol metabolism in Escherichia coli for effective production of recombinant proteins. Biotechnol Biofuels 2020;13(1):205. https://doi.org/10.1186/s13068-020-01848-z.
37. Kuo TC, Tian TF, Tseng YJ. 3Omics: a web-based systems biology tool for analysis, integration and visualization of human transcriptomic, proteomic and metabolomic data. BMC Syst Biol 2013;7:64. https://doi.org/10.1186/1752-0509-7-64.
38. Siddiqui JK, Baskin E, Liu M, et al. IntLIM: integration using linear models of metabolomics and gene expression data. BMC Bioinf 2018;19(1):81. https://doi.org/10.1186/s12859-018-2085-6.
39. Eicher T, Spencer KD, Siddiqui JK, et al. IntLIM 2.0: identifying multi-omic relationships dependent on discrete or continuous phenotypic measurements. Bioinform Adv 2023;3(1):vbad009. https://doi.org/10.1093/bioadv/vbad009.
40. Yu S, Drton M, Promislow DEL, et al. CorDiffViz: an R package for visualizing multi-omics differential correlation networks. BMC Bioinf 2021;22(1):486. https://doi.org/10.1186/s12859-021-04383-2.
41. Kudryashova KS, Burka K, Kulaga AY, et al. Aging biomarkers: from functional tests to multi-omics approaches. Proteomics 2020;20(5–6):e1900408. https://doi.org/10.1002/pmic.201900408.
42. Cellerino A, Ori A. What have we learned on aging from omics studies? Semin Cell Dev Biol 2017;70:177–89. https://doi.org/https://doi.org/10.1016/j.semcdb.2017.06.012
43. Allegri M, Gregori MD, Minella CE, et al.'Omics' biomarkers associated with chronic low back pain: protocol of a retrospective longitudinal study. BMJ Open 2016;6(10):e012070. https://doi.org/10.1136/bmjopen-2016-012070.
44. Mars RAT, Yang Y, Ward T, et al. Longitudinal multi-omics reveals subset-specific mechanisms underlying irritable bowel syndrome. Cell 2020;182(6):1460–73. e17. https://doi.org/https://doi.org/10.1016/j.cell.2020.08.007.
45. OmicsAnalyst. https://www.omicsanalyst.ca. Accessed 19 May 2024.
46. 3omics. https://3omics.cmdm.tw. Accessed 19 May 2024.
47. IntLIM. https://intlim.ncats.io. Accessed 19 May 2024.
48. CorDiffViz. https://diffcornet.github.io/CorDiffViz/demo.html. Accessed 19 May 2024.
49. Mittelstadt B, Wachter S, Russell C. To protect science, we must use LLMs as zero-shot translators. Nat Hum Behav 2023;7(11):1830–32. https://doi.org/10.1038/s41562-023-01744-0.
50. Rosoł M, Gąsior JS, Łaba J, et al. Evaluation of the performance of GPT-3.5 and GPT-4 on the. Polish Med Final Exam Sci Rep 2023;13(1):20512. https://doi.org/10.1038/s41598-023-46995-z.
51. Antaki F, Milad D, Chia MA, et al. Capabilities of GPT-4 in ophthalmology: an analysis of model entropy and progress towards human-level medical question answering. Br J Ophthalmol 2024;108:1371–78. https://doi.org/10.1136/bjo-2023-324438.
52. Miotto M, Rossberg N, Kleinberg B. Who is GPT-3? An exploration of personality, values and demographics. arXiv. 2022; 2209.14338. https://doi.org/10.48550/arXiv.2209.14338. Accessed 18 July 2024.
53. Zhao S, Li H, Han W, et al. Metabolomic coverage of chemical-group-submetabolome analysis: group classification and four-channel chemical isotope labeling LC-MS. Anal Chem 2019;91(18):12108–115. https://doi.org/10.1021/acs.analchem.9b03431.
54. CAT Bridge (Compounds And Transcripts Bridge). https://catbridge.work. Accessed 19 Mar 2024.
55. Yang B. CAT Bridge.zip. Figshare. 2024. https://doi.org/10.6084/m9.figshare.25044854.v3. Accessed 19 May 2024.
56. Yang B. CAT Bridge software heritage archive. 2024. https://archive.softwareheritage.org/browse/origin/directory/?origin_url=https://github.com/Bowen999/CAT-Bridge. Accessed 19 May 2024.
57. Sequence Read Archive. http://www.ncbi.nlm.nih.gov/sra. Accessed 19 May 2024.
58. Metabolomics Workbench. https://www.metabolomicsworkbench.org. Accessed 19 May 2024.
59. Gene Expression Omnibus. https://www.ncbi.nlm.nih.gov/geo/. Accessed 19 May 2024.
60. MetaboLights. https://www.ebi.ac.uk/metabolights/. Accessed 19 May 2024.

Received: April 8, 2024. **Revised:** August 8, 2024. **Accepted:** October 4, 2024