CAT Bridge: An Efficient Toolkit for Gene-Metabolite Association Mining from Multi-Omics Data

Authors

Bowen Yang (0009-0006-9859-731X)[1,2], Tan Meng (0009-0005-2451-2320)[1], Xinrui Wang (0009-0006-1733-3831)[1], Jun Li (0009-0002-8602-1306)[1], Shuang Zhao (0000-0002-1943-8542)[3], Yingheng Wang (0000-0002-4828-4757)[4], Shu Yi (0009-0009-1815-5717)[1], Yi Zhou (0009-0001-1516-1822)[1], Yi Zhang (0009-0000-8468-2747)[1], Liang Li (0000-0002-9347-2108)[2,3,*], Li Guo (0000-0001-6100-3481)[1,*]

Affiliations

[1] *Peking University Institute of Advanced Agricultural Sciences, Shandong Provincial Key Laboratory of Precision Molecular Crop Design and Breeding, Shandong Laboratory of Advanced Agricultural Sciences at 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*

Author Email:

by8@ualberta.ca, mengtan_big_data@163.com, 1120230736@mail.nankai.edu.cn, jun.li@pku-iaas.edu.cn, szhao1@ualberta.ca, ys3702289@outlook.com, yw2349@cornell.edu, zhouyi_zoe@cau.edu.cn, 2022308310509@cau.edu.cn, liang.li@ualberta.ca, li.guo@pku-iaas.edu.cn

*Correspondence should be addressed to Guo Li. Tel: +86 13325308113; Email: li.guo@pku-iaas.edu.cn

*Correspondence may also be addressed to Liang Li. Tel: +1 (780) 492-3250; Email: liang.li@ualberta.ca

Running title: CAT Bridge

Abstract

Background: With advancements in sequencing and mass spectrometry technologies, multi-omics 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 non-linear 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 multi-omics 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 multi-omics analysis. Additionally, CAT Bridge features an artificial intelligence (AI) 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* (*E. 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 (CCM) method outperforms traditional approaches in longitudinal multi-omics analyses. CAT Bridge simplifies access to various established methods for longitudinal multi-omics analysis, and enables researchers to swiftly identify associated gene-metabolite pairs for further validation.

Keywords: Web server, gene-metabolite association, Multi-omics, Time series data, Causality

Background

Recent advancements in sequencing and mass spectrometry (MS) technologies have made the acquisition of multi-omics data more cost-efficient and feasible, multi-omics data analysis is crucial for understanding intricate biological mechanisms from a more comprehensive perspective than single omics, 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 two classes: knowledge-driven approaches and data-driven approaches [5]. Knowledge-driven approaches have shown their inadequacies for non-model 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 non-linear 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

85 cholesterol is detrimental. However, the actual cause is underlying diseases like cancer, which
86 cause both low cholesterol and increased mortality [16]. Therefore, integrating data-driven and
87 knowledge-driven methodologies may offer a more comprehensive and accurate interpretation
88 of multi-omics data.

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

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

104 Materials and Methods

105 Overview of CAT Bridge

106 The workflow of CAT Bridge consists of three primary steps, (1) data processing; (2) statistical
107 modeling (causality estimation and correlation coefficient calculation); (3) visualization and
108 interpretation (**Figure 2A**). Users are required to upload two processed files, gene expression
109 and metabolite concentration matrices, and specify a metabolite of interest as the target. After
110 data processing, seven different causality and correlation algorithms are applied to measure the
111 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 (**Figure 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 Spearman Correlation Coefficient (Spearman) and Pearson Correlation Coefficient (Pearson). However, such correlation-based methods have substantial limitations [9, 10] because they overlook the non-linearity 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), Cross-Correlation Function (CCF) for calculating correlation coefficient (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 multi-omics 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 multi-omics studies, causal relationships may provide a more accurate depiction of the association between genes and metabolites.

140 **Table 1. Comparison of statistical methods by linearity and time lag consideration**

Statistical Method	Linearity	Time Lag Consideration
Pearson Correlation Coefficient (Pearson)	Non-linear	No
Spearman Correlation Coefficient (Spearman)	Non-linear	No
Convergent Cross Mapping (CCM)	Non-linear	Yes
Granger Causality Test (Granger)	Linear	Yes
Canonical Correlation Analysis (CCA)	Linear	No
Dynamic-Time-Warping (DTW)	Non-linear	Yes
Cross-Correlation Function (CCF)	Linear	Yes

141 **Heuristic ranking of candidate genes**

142 Fold change (FC) is another measurement frequently used in omics analyses to identify
143 differentially expressed genes [24]. CAT Bridge pinpoints the peak time of the target
144 metabolite and calculates each gene's log2 normalized FC of this peak time point and the
145 subsequent decline time point (that is, the next sampling time point after the peak time point)
146 using DESeq2 [25]. Then, causality or correlation and FC are combined into a vector to
147 represent the gene-metabolite pair. After the min-max normalization of values (The details are
148 provided in the Supplementary Text), the magnitude of this vector is calculated as the CAT
149 score (Figure 2B). This score heuristically ranks the strength of association between each gene
150 and the metabolite. Users can filter putative genes based on thresholds (e.g., 0.5 for causality
151 and correlation, 1 for normalized FC) or manually review them in descending order.

152 **Knowledge-driven approaches and visualization**

153 Optionally, if users provide a gene function annotations file, typically derived from homology
154 annotations using tools like InterProScan [26] or eggNOG-mapper [27] for non-model
155 organisms), the CAT score will be adjusted with an additional value. This value is determined
156 by a scoring rule based on the gene's description. By default, genes annotated as enzymes

157 receive a score of 0.2, while those with unknown functions get a score of 0.1. Users can
158 customize this scoring rule based on their specific requirements, depending on the presence of
159 target annotations and their importance. Finally, the top 100 genes in heuristically ranking will
160 be evaluated by the GPT-3.5 Turbo based AI agent, to identify putative genes on the gene's
161 functional annotation and prior knowledge (The implementation methods are provided in the
162 Supplementary Text). To enhance data interpretation, the CAT Bridge workflow offers a visual
163 ranking of genes based on computation results, and also incorporates a spectrum of widely
164 utilized graphical outputs, such as heatmap, and principal component analysis (PCA) plot,
165 details can be found in the Supplementary Text.

166 **Plant Material Cultivation and Sampling**

167 To test the effectiveness of CAT Bridge across different species, especially its applicability to
168 non-model organisms, we collected transcriptome sequencing and metabolic profiling data
169 from *C. chinense*, focusing on one of its trademark natural products, capsaicin. *C. chinense*
170 seedlings were divided into three groups, each containing 15 seedlings, grown in a greenhouse
171 of Peking University Institute of Advanced Agricultural Sciences with a controlled
172 environment of 25°C temperature, a light-dark cycle of 16 hours light and 8 hours dark, and
173 70% relative humidity. The fruits of *C. chinense* were sampled at seven distinct time points,
174 starting from the day of flowering, i.e. 0 day post-anthesis (DPA) during which flowers were
175 collected. Followed by fruit harvest on days 7, 16, 30, 50, 55, and 60 DPA. For each time point,
176 we sampled 15 fruits from each group of seedlings, yielding a total of 45 sampling per time
177 point. For each group, we utilized 1.0 mL of 70% aqueous methanol per sample, with a sample
178 weight of approximately 50 mg (aside from the 7 DPA samples which averaged 27.3mg). The
179 samples were ground and freeze-dried in liquid nitrogen. Each sample was then extracted using
180 1.0 mL of 70% aqueous methanol for every 50 mg of sample. Following extraction, the samples
181 underwent ultrasonic treatment using an Ultrasonic Cell Disruptor SCIENTZ-IID (Ningbo
182 Scientz Biotechnology Co., LTD., China) at a frequency of 40 kHz for 10 minutes at room
183 temperature. Standards were prepared as follows: a mixed standard solution, ranging from 20-
184 50 μg/mL, was prepared using MS-grade methanol. For the amino acid standard solution, a 1

185 mg/mL stock solution was initially prepared in water, and then diluted with 50% methanol to
186 achieve a final concentration of 50 μg/mL. Three biological sample replicates were utilized in
187 the subsequent transcriptome and metabolome analyses.

188 **Metabolome Profiling using HPLC-MS and Data Pre-processing**

189 The metabolome profiling was carried out using untargeted metabolomics based on liquid
190 chromatography coupled with mass spectrometry (LC-MS). The samples were filtered through
191 a 0.22 μm membrane and transferred into the lining tube of a sampling vial. Subsequent
192 centrifugation was carried out at 12000 rcf and 4℃ for 10 minutes. The processed samples
193 were then analyzed using Thermo Scientific Orbitrap Exploris™ 240 (Thermo Fisher Scientific,
194 USA). Chromatographic separation was achieved on a T3 C18 (1.7 μm, 2.1 mm × 150 mm
195 column, USA) maintained at 40℃. The mobile phase consisted of A: 1% formic acid in water
196 and B: 1% formic acid in acetonitrile, with a flow rate of 300 μL/min. A 3 μL sample was
197 injected at an autosampler temperature of 10℃. The elution gradient was set as follows: 0-2.5
198 min, 3-10% B; 2.5-6 min, 10-44% B; 6-14 min, 44-80% B; 14-20 min, 80-95% B; 20-23 min,
199 95% B; 23-23.1 min, 95-3% B; 23.1-28 min, 3% B. MS was performed using both positive and
200 negative ion scans, with a precursor ion scan mode. The auxiliary gas heater temperature was
201 set at 350℃, and the ion transfer tube temperature was also maintained at 350℃. The sheath
202 gas flow rate and auxiliary gas flow rate were set to 35 arb and 15 arb, respectively. The
203 voltages were set to 3.5 KV for the positive spectrum and 3.2 KV for the negative spectrum.
204 For MS1, the scan resolution was 60000, with a scan range of 80-1200. For MS2, the scan
205 resolution was 15000, with a stepped collision energy of 20, 40, and 60 eV. Metabolite
206 identification and quantification were performed using the Compound Discoverer software 3.3
207 (Thermo Fisher Scientific, USA).

208 **RNA extraction and transcriptome sequencing**

209 Total RNA was isolated from the above collected plant materials using Trizol Reagent (Thermo
210 Fisher, USA) following manufacturer recommended protocol. The quality of RNA extracts was
211 evaluated using RNA Nano 6000 Assay Kit of the Bioanalyzer 2100 system (Agilent

212 Technologies, USA) following manufacturer's recommendation and samples with a RIN
213 value >7 were used in downstream sequencing library construction and sequencing. The library
214 construction was conducted using Illumina True-seq transcriptome kit (Illumina, USA)
215 following standard protocols. Transcriptome sequencing was carried out by Novogene Co., Ltd.
216 The sequencing reads were procured from the Illumina NovaSeq 6000 platform. For pre-
217 processing, fastp [28] was employed to conduct quality control and clean the data.
218 Subsequently, these reads were mapped to the *C. chinense* cultivar PI159236 genome [29]
219 using STAR [30]. StringTie [31] was utilized to quantify and assess the expression levels of
220 the genes that were successfully mapped. The biological function annotation of genes is
221 obtained through the eggNOG-mapper.

222 **Acquisition and processing of public datasets**

223 We also collected human and *Escherichia coli* (*E. coli*) multi-omics data from public datasets
224 to further examine the performance of CAT Bridge.
225 The human data was sourced from a published aging study [32] that sampled transcriptome
226 and metabolome data from 28 younger (20 to 25 years) and 54 older (55 to 66 years) female
227 human donors. 13 time points are included in them where both the transcriptome and the
228 metabolome data were detected. We obtained the transcripts expression levels and the
229 concentrations of glucose and fructose 6-phosphate for 13 donors, if multiple donor data were
230 available at a single time point, the average value was used.
231 The dataset for *E. coli* originated from a study on the interactions between metabolites and
232 genes within the bacterium [33]. Researchers manipulated the culture conditions of *E. coli*,
233 alternating between growth and starvation phases. They then collected the transcriptome and
234 metabolome under these varying conditions to systematically explore how metabolites
235 influence transcription factors. The dataset includes 29 time points where both transcriptomic
236 and metabolomic information were concurrently obtained.
237
238
239

240 ## Results

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

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

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

258 In our inaugural case study, we leveraged experimentally obtained non-model organism data
259 to examine the performance of CAT Bridge. This data comprised the transcriptome and
260 metabolome of *C. chinense* at seven different developmental stages after bloom (**Figure 3A**).
261 Capsaicin, an important natural product produced by *C. chinense* that gives fruit pungency and
262 has potential anti-cancer and analgesic activity [34], was selected as the target metabolite for
263 this study. Time-series transcriptome and metabolic profiling of developing *C. chinense* fruits
264 were used as input data to test CAT Bridge.

265 Through examination using the CCM method for hypothetical ranking, BC332_05016
266 encoding an Acyl-transferase was ranked first, regardless of whether an annotation file was

267 provided. The Result suggests that this gene was more likely to be the synthetic gene associated
268 with capsaicin in *C. chinense* (**Figure 3B**). The complete heuristic ranking results are provided
269 in Supplementary Table S1. BLAST search revealed that BC332_05016 was homologous of
270 PUN1 (sequence identity: 100%), a.k.a AT3 (Acyl-transferase 3) or CS (capsaicin synthesis)
271 gene [35]. Moreover, when common thresholds were applied for screening, only CCM passed
272 the criteria. The causality modeled based on CCM was 0.55, implying a strong association
273 between BC332_05016 and capsaicin. By contrast, the conventional Pearson correlation
274 method produced a result of 0.08, which would fall below the commonly used threshold and
275 potentially lead to an overlook of this gene-metabolite pair (**Figure 3C**). The AI agent also
276 accurately found BC332_05016 among the top 100 genes based on functional annotation
277 (**Figure 3D**). Furthermore, CAT Bridge visualization result showed that capsaicinoids such as
278 nonivamide, dihydrocapsaicin, and homocapsaicin have high similarity to capsaicin (**Figure**
279 **3E**) and may play a significant role in response to the variable (**Figure 3F**). The rest of the
280 visualization results are provided in the Supplementary Figures (**Figure S1-S4**). These results
281 show that the CAT Bridge is a valuable tool in multi-omics analysis to reliably identify
282 associated gene-metabolite pairs.

283 **Case Study 2: Identifying association of gene-metabolite pairs in glycolysis**

284 For case study 2, we utilized an open-source multi-omics dataset generated by previous aging
285 research (Kuehne et al. 2017) [32]. After processing, 13 time points from the skin of younger
286 and older people used for testing, we particularly compared different association modeling
287 results from various statistical methods in CAT Bridge.

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

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

309 **Case Study 3: metabolite-transcription factor interactions and Acetyl-CoA regulation**

310 In the third study, we utilized a transcriptomic and metabolomic dataset from *E. coli* to
311 demonstrate the capabilities of CAT Bridge in analyzing simple organisms. The dataset
312 contains gene expression and metabolite concentration data across 29 time points, and it was
313 originally generated by Lempp et al. to investigate the allosteric regulation of transcription
314 factors (TFs) by metabolites [33]. They constructed a "literature network" of known
315 metabolite-TF interactions, including 16 interactions as activations of the TF by metabolites.
316 Using Kinetic Correlation, which accommodates time lags, Lempp et al. recovered 10 out of
317 the 16 activation interactions. We re-analyzed this dataset using CAT Bridge to test its
318 performance with simple organisms. Utilizing the functions integrated within CAT Bridge and
319 setting a time lag of 4, we calculated the association strengths between these molecules. The
320 CCM method once again performed best, reproducing 14 out of 16 interactions.

321 Focusing on Acetyl-CoA as the target metabolite and setting a time lag of 3, we investigated
322 the associated genes. The CCM-based methods fared the best, and the results showed strong
323 causality (greater than 0.9) with the aceE, aceF, and lpd genes, which are involved in the
324 conversion of pyruvate to Acetyl-CoA (**Figure 5A**) [36]. These genes were also ranked in the
325 top 100 in heuristic sorting and clustered together. Additionally, the acs and pta genes, which
326 are involved in the conversion of acetate to Acetyl-CoA, [36], also showed strong causality
327 (0.840 and 0.721, respectively) but did not rank in the top 100 in heuristic sorting (**Figure 5B**).
328 The complete heuristic ranking results are provided in Supplementary Table S2. This suggests
329 that in this study, the precursors for Acetyl-CoA may predominantly come from pyruvate.

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

334 **Comparison with other web-based tools**

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

343

344

345

346

347 **Table 2. Comparison of CAT Bridge with other web-based multi-omics tools.**

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

348 Symbols indicating feature assessments: '✓' denotes presence, '×' signifies absence.
349 OmicsAnalyst [41]; 3omics [42]; IntLIM [43]; CorDiffViz [44].

350 Discussion

351 In recent years, there has been a surge in multi-omics research. A critical aspect often
352 overlooked in such studies is the unique nature of the longitudinal experimental design.
353 Longitudinal omics analysis is particularly important in research on the developmental cycle
354 of plants and investigations related to chronic diseases and aging [45-48]. However, many
355 studies tend to use generic methodologies for analysis [11, 12]. This may inadvertently miss
356 key discoveries. CAT Bridge provides a platform specifically for longitudinal multi-omics
357 analysis, by drawing insights from disciplines where time series data is more prevalent and
358 benchmarking them with data. Through the gene-metabolite causality and correlation modeling
359 method, combined with visualization tools and AI assistance, researchers can more quickly
360 identify putative genes for experimental validation. We believe that discovering associated
361 gene-metabolite pairs will have practical applications in many fields. For example, in metabolic
362 engineering, a single metabolite is often regulated by multiple genes, by comparing the
363 association strength, one can infer which genes may play a dominant role under certain
364 conditions, facilitating the development of targeted metabolic engineering strategies.

365 Furthermore, in horticulture and breeding, identifying genes with causal relationships to key
366 natural products, particularly in non-model organisms, can significantly aid in molecular
367 breeding efforts, potentially leading to the development of new crop varieties with enhanced
368 yields or desired characteristics.

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

389 Aside from computational methods, the reliability of analytical results is also influenced by
390 experimental design and data acquisition methods. Increasing the number of sampling time
391 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 (CIL) 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 non-linear relationships, and constructing molecular networks based on temporal causality and metabolic pathways.

Availability of source code and requirements

Project name: CAT Bridge

Project home page: 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 multi-omics

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

Author Contribution

B Yang: Conceptualization, Software, Methodology, Visualization, Investigation, Writing—original draft. T Meng: Software. X Wang: Visualization. J Li: Investigation. S Zhao: Writing—review & editing. Y Wang: Methodology, Writing—review & editing. S Yi: Investigation. Y Zhou: Software. Y Zhang: Software. L Li: Supervision. L Guo: 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. LG is also supported by Taishan Scholars Program of Shandong Province.

Acknowledgement

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

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 has been deposited at the Metabolomics Workbench [58] with the study ID ST003172. The sequencing data of case study 2 was obtained from Gene Expression Omnibus (GEO) [59], with the accession number GSE85358. The sequencing data of case study 3 was obtained from

GEO with the accession code GSE131992, and metabolomics data was 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 and Arnold M. Multi-omics integration in biomedical research - A metabolomics-centric review. Anal Chim Acta. 2021;1141:144-62. doi:10.1016/j.aca.2020.10.038.
2. Hasin Y, Seldin M and Lusis A. Multi-omics approaches to disease. Genome Biology. 2017;18 1:83. doi:10.1186/s13059-017-1215-1.
3. Subramanian I, Verma S, Kumar S, Jere A and Anamika K. Multi-omics Data Integration, Interpretation, and Its Application. Bioinform Biol Insights. 2020;14:1177932219899051. doi:10.1177/1177932219899051.
4. Eicher T, Kinnebrew G, Patt A, Spencer K, Ying K, Ma Q, et al. Metabolomics and Multi-Omics Integration: A Survey of Computational Methods and Resources. Metabolites. 2020;10 5:202.
5. Zhou G, Ewald J and Xia J. OmicsAnalyst: a comprehensive web-based platform for visual analytics of multi-omics data. Nucleic Acids Research. 2021;49 W1:W476-W82. doi:10.1093/nar/gkab394.
6. Krassowski M, Das V, Sahu SK and Misra BB. State of the Field in Multi-Omics Research: From Computational Needs to Data Mining and Sharing. Front Genet. 2020;11:610798. doi:10.3389/fgene.2020.610798.
7. Athieniti E and Spyrou GM. A guide to multi-omics data collection and integration for translational medicine. Computational and Structural Biotechnology Journal. 2023;21:134-49. doi:https://doi.org/10.1016/j.csbj.2022.11.050.
8. Albaradei S, Thafar M, Alsaedi A, Van Neste C, Gojobori T, Essack M, et al. Machine learning and deep learning methods that use omics data for metastasis prediction. Comput Struct Biotechnol J. 2021;19:5008-18. doi:10.1016/j.csbj.2021.09.001.
9. Cavill R, Jennen D, Kleinjans J and Briedé JJ. Transcriptomic and metabolomic data integration. Briefings in Bioinformatics. 2015;17 5:891-901. doi:10.1093/bib/bbv090.
10. Chong J and Xia J. Computational Approaches for Integrative Analysis of the Metabolome and Microbiome. Metabolites. 2017;7 4:62.
11. Li Y, Chen Y, Zhou L, You S, Deng H, Chen Y, et al. MicroTom metabolic network: rewiring tomato metabolic regulatory network throughout the growth cycle. Molecular plant. 2020;13 8:1203-18.
12. Yang C, Shen S, Zhou S, Li Y, Mao Y, Zhou J, et al. Rice metabolic regulatory network spanning the entire life cycle. Molecular Plant. 2022;15 2:258-75.
13. Singh KS, van der Hooft JJJ, van Wees SCM and Medema MH. Integrative omics approaches for biosynthetic pathway discovery in plants. Natural Product Reports. 2022;39 9:1876-96. doi:10.1039/D2NP00032F.
14. Ye H, Deyle ER, Gilarranz LJ and Sugihara G. Distinguishing time-delayed causal interactions using convergent cross mapping. Scientific Reports. 2015;5 1:14750. doi:10.1038/srep14750.
15. Yuan AE and Shou W. Data-driven causal analysis of observational biological time series. Elife. 2022;11 doi:10.7554/eLife.72518.
16. Sattar N and Preiss D. Reverse Causality in Cardiovascular Epidemiological Research.

516 Circulation. 2017;135 24:2369-72. doi:doi:10.1161/CIRCULATIONAHA.117.028307.
517 17. Rockwood AL, Crockett DK, Oliphant JR and Elenitoba-Johnson KS. Sequence
518 alignment by cross-correlation. J Biomol Tech. 2005;16 4:453-8.
519 18. Skutkova H, Vitek M, Babula P, Kizek R and Provaznik I. Classification of genomic
520 signals using dynamic time warping. BMC Bioinformatics. 2013;14 10:S1.
521 doi:10.1186/1471-2105-14-S10-S1.
522 19. Seoane JA, Campbell C, Day IN, Casas JP and Gaunt TR. Canonical correlation
523 analysis for gene-based pleiotropy discovery. PLoS Comput Biol. 2014;10
524 10:e1003876. doi:10.1371/journal.pcbi.1003876.
525 20. Jiang MZ, Aguet F, Ardlie K, Chen J, Cornell E, Cruz D, et al. Canonical correlation
526 analysis for multi-omics: Application to cross-cohort analysis. PLoS Genet. 2023;19
527 5:e1010517. doi:10.1371/journal.pgen.1010517.
528 21. Yuan AE and Shou W. Data-driven causal analysis of observational biological time
529 series. eLife. 2022;11:e72518. doi:10.7554/eLife.72518.
530 22. Heerah S, Molinari R, Guerrier S and Marshall-Colon A. Granger-causal testing for
531 irregularly sampled time series with application to nitrogen signalling in Arabidopsis.
532 Bioinformatics. 2021;37 16:2450-60. doi:10.1093/bioinformatics/btab126.
533 23. Stokes PA and Purdon PL. A study of problems encountered in Granger causality
534 analysis from a neuroscience perspective. Proc Natl Acad Sci U S A. 2017;114
535 34:E7063-e72. doi:10.1073/pnas.1704663114.
536 24. Arora S, Pattwell SS, Holland EC and Bolouri H. Variability in estimated gene
537 expression among commonly used RNA-seq pipelines. Scientific Reports. 2020;10
538 1:2734. doi:10.1038/s41598-020-59516-z.
539 25. Love MI, Huber W and Anders S. Moderated estimation of fold change and dispersion
540 for RNA-seq data with DESeq2. Genome Biology. 2014;15 12:550.
541 doi:10.1186/s13059-014-0550-8.
542 26. Ye J, Coulouris G, Zaretskaya I, Cutcutache I, Rozen S and Madden TL. Primer-BLAST:
543 a tool to design target-specific primers for polymerase chain reaction. BMC
544 Bioinformatics. 2012;13:134. doi:10.1186/1471-2105-13-134.
545 27. Cantalapiedra CP, Hernández-Plaza A, Letunic I, Bork P and Huerta-Cepas J. eggNOG-
546 mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at
547 the Metagenomic Scale. Molecular Biology and Evolution. 2021;38 12:5825-9.
548 doi:10.1093/molbev/msab293.
549 28. Chen S, Zhou Y, Chen Y and Gu J. fastp: an ultra-fast all-in-one FASTQ preprocessor.
550 Bioinformatics. 2018;34 17:i884-i90. doi:10.1093/bioinformatics/bty560.
551 29. Kim S, Park J, Yeom SI, Kim YM, Seo E, Kim KT, et al. New reference genome
552 sequences of hot pepper reveal the massive evolution of plant disease-resistance genes
553 by retroduplication. Genome Biol. 2017;18 1:210. doi:10.1186/s13059-017-1341-9.
554 30. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast
555 universal RNA-seq aligner. Bioinformatics. 2013;29 1:15-21.
556 doi:10.1093/bioinformatics/bts635.
557 31. Pertea M, Pertea GM, Antonescu CM, Chang T-C, Mendell JT and Salzberg SL.

558 StringTie enables improved reconstruction of a transcriptome from RNA-seq reads.
559 Nature Biotechnology. 2015;33 3:290-5. doi:10.1038/nbt.3122.
560 32. Kuehne A, Hildebrand J, Soehle J, Wenck H, Terstegen L, Gallinat S, et al. An
561 integrative metabolomics and transcriptomics study to identify metabolic alterations in
562 aged skin of humans in vivo. BMC Genomics. 2017;18 1:169. doi:10.1186/s12864-
563 017-3547-3.
564 33. Lempp M, Farke N, Kuntz M, Freibert SA, Lill R and Link H. Systematic identification
565 of metabolites controlling gene expression in E. coli. Nature Communications. 2019;10
566 1:4463. doi:10.1038/s41467-019-12474-1.
567 34. Fattori V, Hohmann MS, Rossaneis AC, Pinho-Ribeiro FA and Verri WA. Capsaicin:
568 Current Understanding of Its Mechanisms and Therapy of Pain and Other Pre-Clinical
569 and Clinical Uses. Molecules. 2016;21 7 doi:10.3390/molecules21070844.
570 35. Kim S, Park M, Yeom S-I, Kim Y-M, Lee JM, Lee H-A, et al. Genome sequence of the
571 hot pepper provides insights into the evolution of pungency in Capsicum species.
572 Nature Genetics. 2014;46 3:270-8. doi:10.1038/ng.2877.
573 36. Chiang C-J, Ho Y-J, Hu M-C and Chao Y-P. Rewiring of glycerol metabolism in
574 Escherichia coli for effective production of recombinant proteins. Biotechnology for
575 Biofuels. 2020;13 1:205. doi:10.1186/s13068-020-01848-z.
576 37. Kuo TC, Tian TF and Tseng YJ. 3Omics: a web-based systems biology tool for analysis,
577 integration and visualization of human transcriptomic, proteomic and metabolomic data.
578 BMC Syst Biol. 2013;7:64. doi:10.1186/1752-0509-7-64.
579 38. Siddiqui JK, Baskin E, Liu M, Cantemir-Stone CZ, Zhang B, Bonneville R, et al.
580 IntLIM: integration using linear models of metabolomics and gene expression data.
581 BMC Bioinformatics. 2018;19 1:81. doi:10.1186/s12859-018-2085-6.
582 39. Eicher T, Spencer KD, Siddiqui JK, Machiraju R and Mathé EA. IntLIM 2.0:
583 identifying multi-omic relationships dependent on discrete or continuous phenotypic
584 measurements. Bioinformatics Advances. 2023;3 1 doi:10.1093/bioadv/vbad009.
585 40. Yu S, Drton M, Promislow DEL and Shojaie A. CorDiffViz: an R package for
586 visualizing multi-omics differential correlation networks. BMC Bioinformatics.
587 2021;22 1:486. doi:10.1186/s12859-021-04383-2.
588 41. OmicsAnalyst. https://www.omicsanalyst.ca. Accessed 19 May 2024.
589 42. 3omics. https://3omics.cmdm.tw. Accessed 19 May 2024.
590 43. IntLIM. https://intlim.ncats.io. Accessed 19 May 2024.
591 44. CorDiffViz. https://diffcornet.github.io/CorDiffViz/demo.html. Accessed 19 May 2024.
592 45. Kudryashova KS, Burka K, Kulaga AY, Vorobyeva NS and Kennedy BK. Aging
593 Biomarkers: From Functional Tests to Multi-Omics Approaches. Proteomics. 2020;20
594 5-6:e1900408. doi:10.1002/pmic.201900408.
595 46. Cellerino A and Ori A. What have we learned on aging from omics studies? Seminars
596 in Cell & Developmental Biology. 2017;70:177-89.
597 doi:https://doi.org/10.1016/j.semcdb.2017.06.012.
598 47. Allegri M, Gregori MD, Minella CE, Klersy C, Wang W, Sim M, et al. 'Omics'
599 biomarkers associated with chronic low back pain: protocol of a retrospective

longitudinal study. BMJ Open. 2016;6 10:e012070. doi:10.1136/bmjopen-2016-012070.

48. Mars RAT, Yang Y, Ward T, Houtti M, Priya S, Lekatz HR, et al. Longitudinal Multi-omics Reveals Subset-Specific Mechanisms Underlying Irritable Bowel Syndrome. Cell. 2020;182 6:1460-73.e17. doi:https://doi.org/10.1016/j.cell.2020.08.007.

49. Mittelstadt B, Wachter S and Russell C. To protect science, we must use LLMs as zero-shot translators. Nature Human Behaviour. 2023;7 11:1830-2. doi:10.1038/s41562-023-01744-0.

50. Rosoł M, Gąsior JS, Łaba J, Korzeniewski K and Młyńczak M. Evaluation of the performance of GPT-3.5 and GPT-4 on the Polish Medical Final Examination. Scientific Reports. 2023;13 1:20512. doi:10.1038/s41598-023-46995-z.

51. Antaki F, Milad D, Chia MA, Giguère C-É, Touma S, El-Khoury J, et al. Capabilities of GPT-4 in ophthalmology: an analysis of model entropy and progress towards human-level medical question answering. British Journal of Ophthalmology. 2023:bjo-2023-324438. doi:10.1136/bjo-2023-324438.

52. Miotto M, Rossberg N and Kleinberg B. Who is GPT-3? An exploration of personality, values and demographics. arXiv preprint arXiv:220914338. 2022.

53. Zhao S, Li H, Han W, Chan W and Li L. Metabolomic Coverage of Chemical-Group-Submetabolome Analysis: Group Classification and Four-Channel Chemical Isotope Labeling LC-MS. Anal Chem. 2019;91 18:12108-15. doi:10.1021/acs.analchem.9b03431.

54. CAT Bridge (Compounds And Transcripts Bridge). https://catbridge.work. Accessed 19 Mar 2024.

55. Yang, Bowen. CAT Bridge.zip. figshare. Software. 2024. https://doi.org/10.6084/m9.figshare.25044854.v3.

56. Yang, Bowen. CAT Bridge. Software Heritage archive. 2024. https://archive.softwareheritage.org/swh:1:rev:69cf3a9cc195870d8e012a3b32e2c60a9e1a89e6;origin=https://github.com/Bowen999/CAT-Bridge;visit=swh:1:snp:176c4df277b17d55d315228e446158b5b3a0674b.

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.

638 ## Figure Legends

639 **Figure 1. The architecture of CAT Bridge project**

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

646 **Figure 2. The features and overall workflow of CAT Bridge**

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

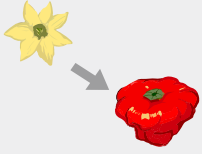
656 **Figure 3**. **Application of CAT Bridge to mine transcript-capsaicin association**

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

666 **Figure 4. Application of CAT Bridge to mine transcript-metabolite associations in**
667 **glycolysis pathway**
668 (A) Left side: the expression patterns of functional genes and their corresponding target
669 metabolites. Right side: the strength of associations was evaluated using different methods, red
670 denotes evaluation outcomes that align with known pathway information under typical
671 thresholds (e.g., a correlation coefficient < -0.5 indicating a strong negative association),
672 whereas grey signifies inconsistent or unsuitable. (B) Part of the glycolysis pathway.
673 **Figure 5. Expression Patterns of Acetyl-CoA Synthesis Genes and Acetyl-CoA**
674 **Concentration**
675 (A) Genes involved in the conversion of pyruvate to Acetyl-CoA. (B) Genes involved in the
676 conversion of acetate to Acetyl-CoA.

Materials

Omics from different time points



open-source datasets

experimentally obtained dataset

Gene expression

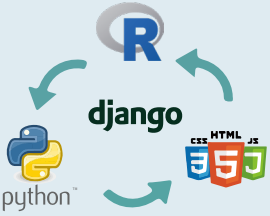
	day01	day02	day03
gene1	1	4	4
gene2	3	4	1
gene3	2	2	3

Metabolite concentration

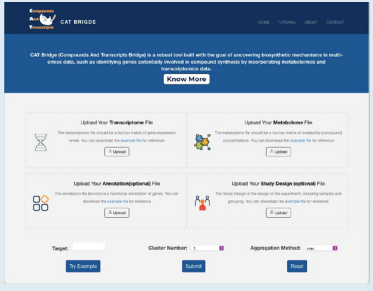
	day01	day02	day03
metabo.1	1.432	8.786	3.127
metabo.2	0	0	1.671
metabo.3	2.679	2.322	3.124

Methodology

Frame



Usage



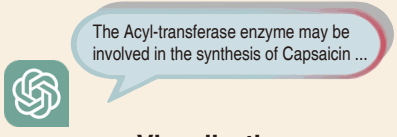
Python Library

Results

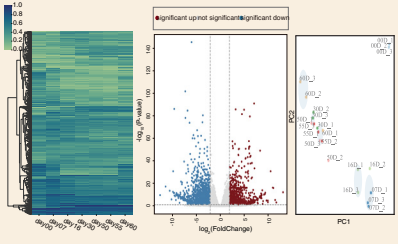
Causality/similarity computation

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

AI assistant: Yuanfang



Visualization



A

Data Processing

Missing value handling
Biological replicate aggregation
Normalization & scaling
Clustering
Significant feature detection

Causality

Convergent Cross Mapping
Granger Causality Test

Correlation

Canonical Correlation Analysis
Dynamic-Time-Warping
Cross-Correlation Function
Spearman Correlation Coefficient
Pearson Correlation Coefficient

Visualization

Heatmap
Similarity network
Expression trend
Dimensionality reduction
Volcano plot
Importance of features

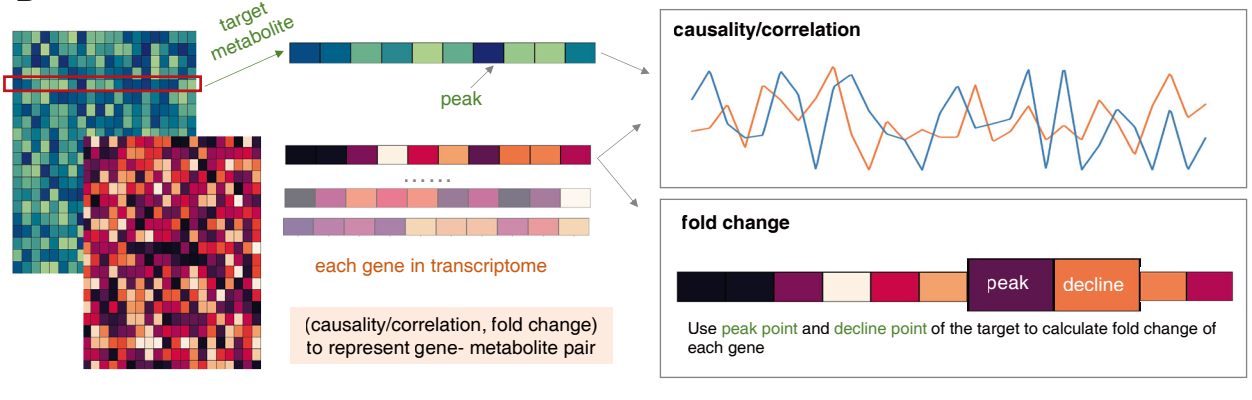
Heuristic Ranking

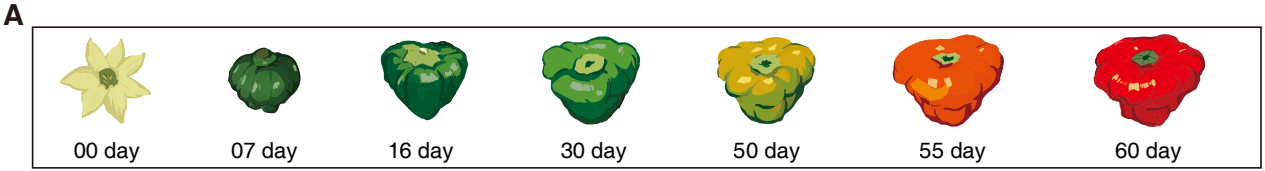
Potential gene-metabolite pairs

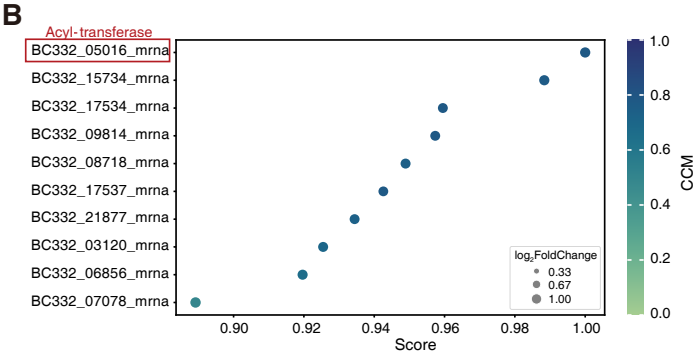
AI Agent

GPT 3.5 from OpenAI API

B



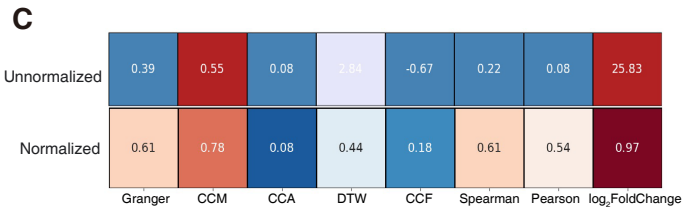


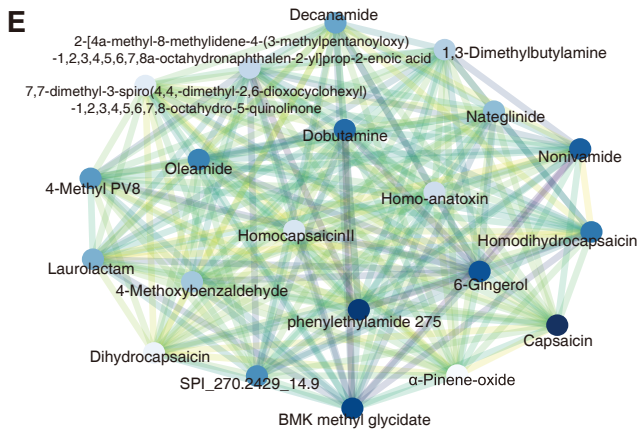


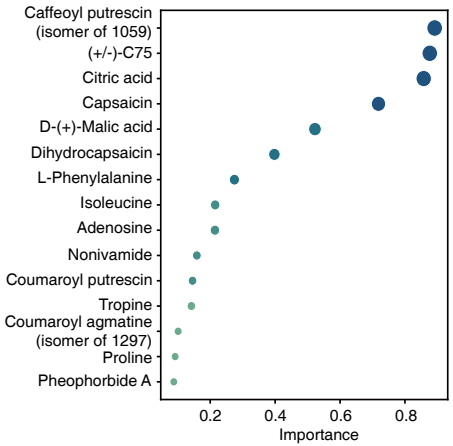
D

```
cat.Yuanfang(result, target, annotation_file)
```

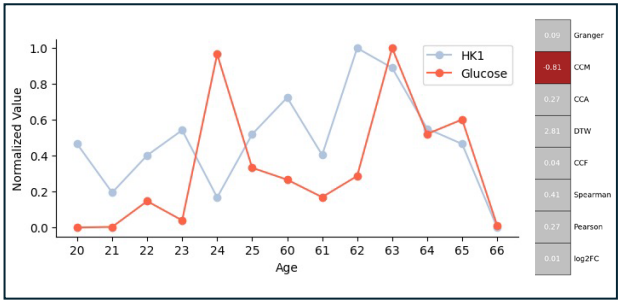
```
Please enter your OpenAI API Key: ........

Based on the provided list of genes, the one tha
t may be involved in the synthesis of Capsaicin
is BC332_05016_mrna, also known as Acyl-transfer
ase. Acyl-transferase enzymes are involved in th
e transfer of acyl groups, which are essential f
or the biosynthesis of various secondary metabol
ites, including capsaicin. Capsaicin is the main
compound responsible for the spicy taste of chil
i peppers and is synthesized through a series of
enzymatic reactions, including the transfer of a
cyl groups. Therefore, BC332_05016_mrna may play
a role in the synthesis of capsaicin.

NOTICE: The output was produced by the large lan
guage model GPT 3.5 turbo, so it should only be
regarded as a source of inspiration.
```

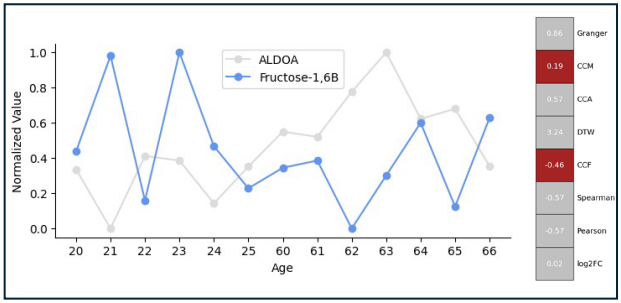


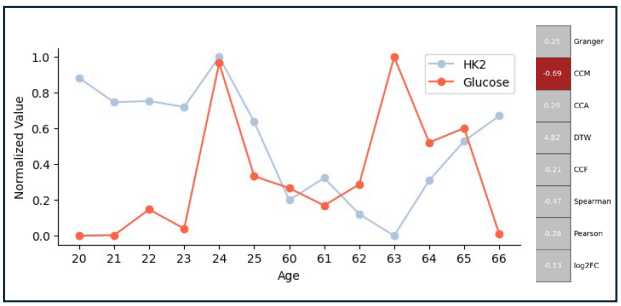


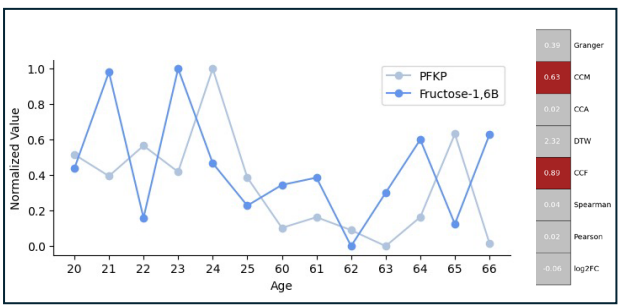


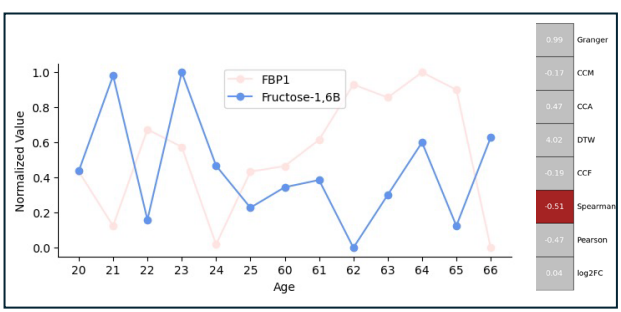
A


HK1
Glucose
Normalized Value
Age
Granger 0.09
CCM -0.81
CCA 0.27
DTW 2.81
CCF 0.04
Spearman 0.41
Pearson 0.27
log2FC 0.01

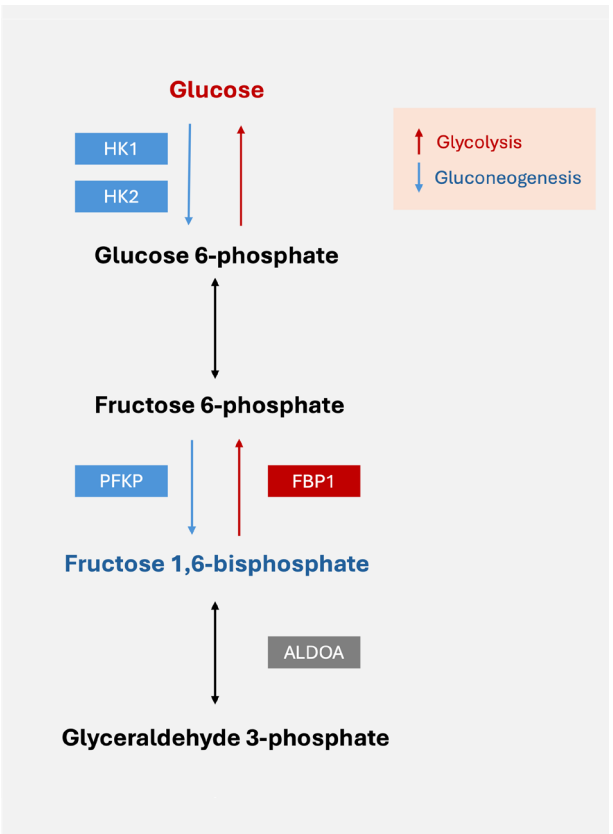

ALDOA
Fructose-1,6B
Normalized Value
Age
Granger 0.86
CCM 0.19
CCA 0.57
DTW 3.24
CCF -0.46
Spearman -0.57
Pearson -0.57
log2FC 0.02

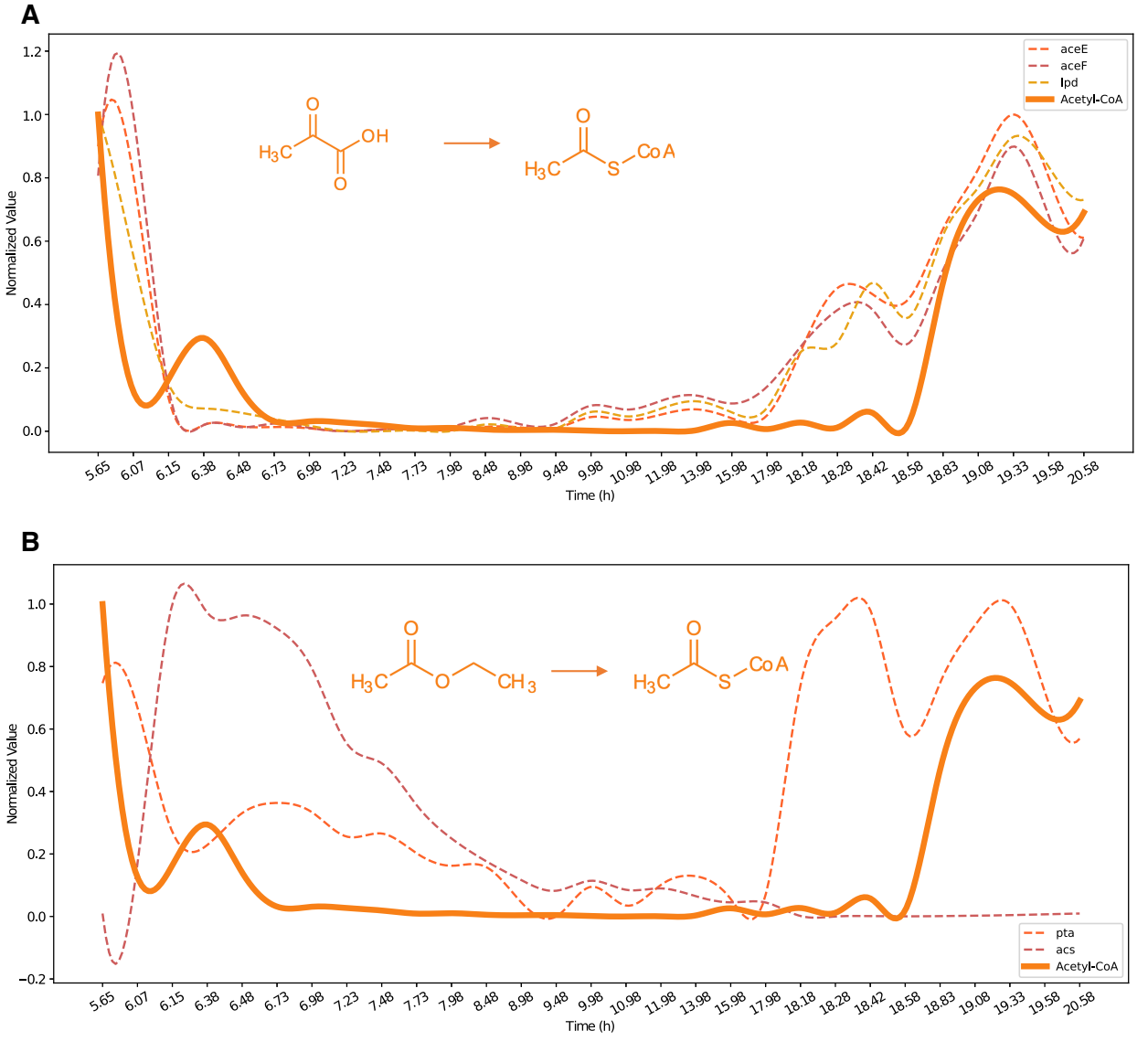

HK2
Glucose
Normalized Value
Age
Granger 0.25
CCM -0.69
CCA 0.28
DTW 4.82
CCF -0.21
Spearman -0.47
Pearson -0.28
log2FC -0.13


PFKP
Fructose-1,6B
Normalized Value
Age
Granger 0.39
CCM 0.63
CCA 0.02
DTW 2.32
CCF 0.89
Spearman 0.04
Pearson 0.02
log2FC -0.06


FBP1
Fructose-1,6B
Normalized Value
Age
Granger 0.99
CCM -0.17
CCA 0.47
DTW 4.02
CCF -0.19
Spearman -0.51
Pearson -0.47
log2FC 0.04

B


Glucose
HK1
HK2
Glycolysis
Gluconeogenesis
Glucose 6-phosphate
Fructose 6-phosphate
PFKP
FBP1
Fructose 1,6-bisphosphate
ALDOA
Glyceraldehyde 3-phosphate


A
aceE
aceF
lpd
Acetyl-CoA
Normalized Value
Time (h)
B
pta
acs
Acetyl-CoA
Normalized Value
Time (h)