PAG-Agent: a biologist-oriented research assistant for context-aware pathway prioritization and interpretation

Quang-Huy Nguyen[1,2,5], Zeru Zhang[5], Zhandos Sembay[3], Duc-Hau Le[2,*], Jake Y. Chen[3,*], Wei-Shinn Ku[1,*], Hao Chen[4,*], Zongliang Yue[5,*]

[1]Department of Computer Science and Software Engineering, Auburn University, Auburn, AL 36849, United States
[2]School of Information and Communications Technology, Hanoi University of Science and Technology, Hanoi, Vietnam
[3]Department of Biomedical Informatics and Data Science, University of Alabama at Birmingham, Birmingham, AL 35233, United States
[4]College of Forestry and Wildlife Sciences, Auburn University, Auburn, AL 36849, United States
[5]Department of Health Outcomes Research and Policy, Auburn University, Auburn, AL 36849, United States
*To whom the correspondence should be addressed.
Email: hauld@soict.hust.edu.vn, jakechen@uab.edu, wzk0004@auburn.edu; hzc0134@auburn.edu, zzy0065@auburn.edu

Abstract

Researchers often face two challenges after identifying significant pathways. First, they must determine which pathways are most relevant to the biological question under investigation. Second, they must integrate biological knowledge, literature evidence, and study-specific context to interpret these findings. Existing enrichment methods primarily prioritize pathways based on statistical significance, whereas pathway interpretation often requires multiple tools and resources. Although large language models (LLMs) offer new opportunities for biological interpretation, hallucinated citations and inaccurate reasoning remain important concerns. Here, we present PAG-Agent, a biologist-oriented virtual research assistant for pathway prioritization and interpretation. PAG-Agent introduces context-aware Over-Representation Analysis (cORA), which incorporates experimental conditions and research objectives to identify biologically relevant pathways beyond statistical significance alone, while integrating analytical tasks, post-analysis tasks, and literature-supported reasoning within a unified workflow. Using a benchmark of 60 cancer-related gene sets derived from bulk and single-cell transcriptomic studies, cORA consistently outperformed conventional enrichment methods and LLM-based baselines across similarity thresholds, achieving the highest accuracy (0.913), hit rate (0.902), and their harmonic mean (0.898). Ablation experiments demonstrated that both biologically informed module detection and contextual prioritization contributed to cORA performance. We further demonstrated PAG-Agent using Alzheimer's disease and chronic myeloid leukemia case studies and showed that the platform identified biologically relevant pathways across diverse analytical settings. PAG-Agent also outperformed competing LLMs in retrieving contextually relevant scientific references across five evaluation scenarios. Together, these results demonstrate that PAG-Agent supports context-aware pathway prioritization and literature-grounded biological interpretation within a unified workflow.

1 Introduction

Advances in high-throughput omics technologies have enabled researchers to profile molecular changes across diverse biological systems and disease conditions. These studies routinely generate lists of differentially expressed genes or gene products from bulk and single-cell experiments. Although informative, gene-level results alone rarely explain the biological mechanisms underlying the observed phenotypes. Pathway analysis (PA) therefore serves as a key step for translating molecular changes into biological processes and functional modules using curated knowledge bases such as KEGG [1] and Gene Ontology (GO) [2,3]. However, PA often produces large collections of statistically significant pathways. Researchers therefore face challenges when translating these findings into biological insight.

Traditional PA methods prioritize pathways primarily based on statistical significance. As a result, highly annotated and broadly represented biological processes are frequently ranked among the top results [4, 5]. However, these pathways are not necessarily the most informative for a particular study. Researchers must therefore determine which significantly enriched pathways are most relevant to the biological question under investigation. We refer to this process as **pathway prioritization**. In practice, researchers often rely on their own domain knowledge, making pathway prioritization a largely manual and time-consuming process.

A plethora of PA platforms primarily support analytical tasks, including data preprocessing, differential expression analysis (DEA), PA, consensus analysis, meta-analysis, and visualization [6–19]. However, researchers often perform additional post-analysis tasks to translate prioritized pathways into biological insight, including gene annotation, pathway annotation, literature validation, and hypothesis generation. Together, analytical and post-analysis tasks constitute a typical workflow that extends beyond statistical analysis (Fig. 1a). In practice, these capabilities are often distributed across separate platforms and resources. Researchers frequently move between different tools and external resources to complete these tasks. One important post-analysis task is **pathway interpretation**, which builds upon pathway annotation to understand the biological functions of prioritized pathways, their relationships to the studied phenotype, and the evidence supporting these associations. As a result, researchers must integrate information from multiple databases, knowledge resources, and published studies, which demands substantial biological expertise and manual effort, particularly for early-career researchers, computational biologists, and experimental researchers who may be less familiar with PA and functional interpretation [20–23].

Large language models (LLMs) have emerged as promising tools for addressing the aforementioned challenges. Recent studies have demonstrated that LLMs can perform gene-set enrichment analysis and generate biologically meaningful functional annotations [5, 24]. These capabilities suggest that LLMs may be useful for both pathway prioritization and pathway interpretation. However, existing approaches still focus primarily on inferring relevant pathways from gene sets themselves without explicitly incorporating experimental conditions, research objectives, and other study-specific context. As a result, they provide limited support for context-aware pathway prioritization. LLMs have also been increasingly applied to post-analysis tasks. However, AI hallucinations can lead to fabricated citations and inaccurate biological interpretations. Researchers have proposed retrieval-augmented generation (RAG) systems to improve factual accuracy by incorporating external knowledge during inference [25, 26]. However, these approaches often require substantial computational resources and continuous maintenance of large knowledge repositories.

Here, we present PAG-Agent, a web-based and biologist-oriented virtual research assistant for pathway prioritization and interpretation. PAG-Agent supports both bulk and single-cell transcriptomic data and integrates the analytical and post-analysis workflow through natural language interactions (Fig. 1a-c). To support reliable post-analysis, PAG-Agent assists users with gene and pathway annotation, literature validation, and hypothesis generation while incorporating lightweight retrieval and in-context learning

strategies to improve the reliability of literature-supported responses. To support pathway prioritization, PAG-Agent introduces context-aware Over-Representation Analysis (cORA), which incorporates experimental conditions and research objectives to identify biologically relevant pathways beyond statistical significance alone (Fig. 1d). Together, PAG-Agent provides an integrated framework for context-aware pathway prioritization and pathway interpretation.

2 Results

2.1 PAG-Agent overview

Fig. 1 summarizes the architecture, workflow, and major capabilities of PAG-Agent through two main modules: **Click Interaction** (Fig. 3a and Supplementary Fig. 1) and **Chat Interaction** (Supplementary Fig. 5–29). Through Click Interaction, users specified analysis settings, including the analysis type (consensus analysis or meta-analysis), data type (bulk or single-cell RNA sequencing data), input type (expression matrix, gene list, or gene list and fold change), preprocessing strategy, DEA method, and pathway database. Through Chat Interaction, users performed the analytical and post-analysis tasks using natural language. These tasks included DEA, PA, pathway-level consensus analysis, meta-analysis, gene and pathway interpretation, hypothesis generation, academic writing refinement, and citation retrieval for validation. Supplementary Fig. 30 provides a detailed flow diagram of the PAG-Agent pipeline. Depending on the input type, users interacted with PAG-Agent to perform specific analyses and generate visualizations. Fig. 1b shows representative publication-ready outputs generated by the platform, including MA plots, volcano plots, forest plots, bar charts, and KEGG pathway maps. Fig. 1c illustrates a simplified dialogue between a human biologist and PAG-Agent for PA, pathway annotation, and literature validation. Fig. 1d shows the cORA workflow for context-aware pathway analysis and prioritization. PAG-Agent was accessible through a web portal (`https://huynguyen250896.github.io/pag-agent-homepage/`).

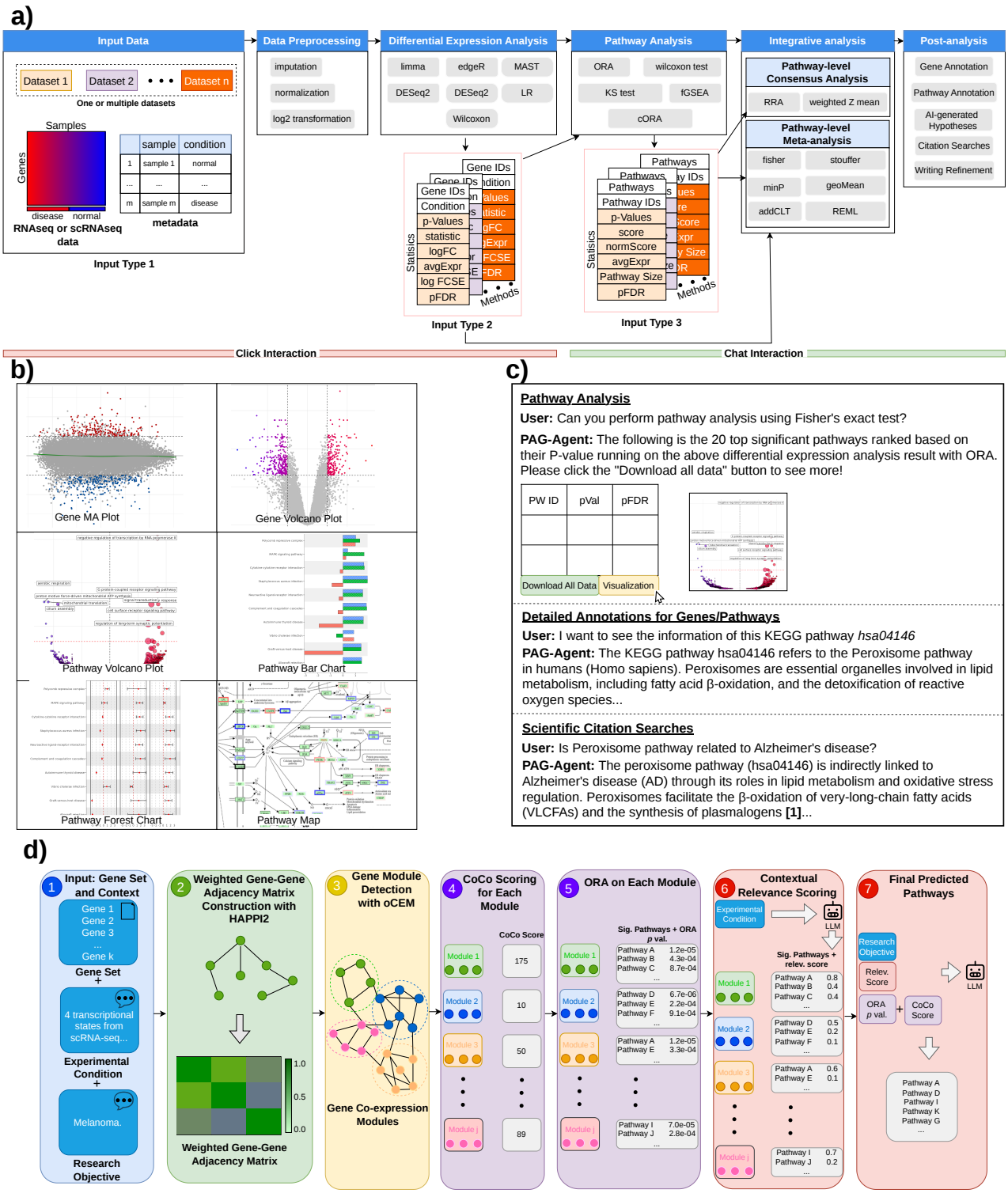

Figure 1: Agentic workflow of PAG-Agent. **a)** PAG-Agent integrates the analytic and post-analysis workflow: input data uploading, data preprocessing, differential expression analysis, pathway analysis, integrative analysis, and post-analysis tasks through Click Interaction and Chat Interaction modules. **b)** PAG-Agent produces publication-ready graphs as raster (.png) and vector (.svg and .pdf) images. **c)** Simplified example of a dialogue between a human biologist and PAG-Agent during a Chat Interaction session in English (see Supplementary Information Section 2 for details). **d)** cORA workflow for context-aware pathway analysis and prioritization using experimental conditions and research objectives.

To evaluate PAG-Agent, we combined benchmark-scale and case-study analyses. For context-aware pathway prioritization, we constructed a benchmark comprising 60 independent cancer-related gene sets from RummaGEO (`https://rummageo.com/`), spanning diverse biological conditions and experimental settings (see Section Methods for details). For end-to-end workflow demonstration, we analyzed three Alzheimer's disease datasets: *GSE153873* [27], *GSE61196* [28], and *GSE5281* [29] (Input Type 1 in Fig. 1a). We selected Alzheimer's disease as a representative case study because multiple neurodegeneration-related pathways share genes and biological mechanisms with the disease [30–33], allowing expected biological signals to be assessed across different analytical settings. In addition, to further demonstrate PAG-Agent's ability to perform PA using a single gene list and fold change data (Input Type 2 in Fig. 1a and Supplementary Fig. 3) and conduct meta-analysis using gene/pathway lists and their statistics (Input Type 2 and Input Type 3 in Fig. 1a, and Supplementary Fig. 4), we generated example DEA and PA result files. Detailed methods for generating these files were provided in the Supplementary Information Section 2.

2.2 cORA achieves non-generic, significant pathways across 60 cancer-related gene sets

We next evaluated whether incorporating experimental conditions and research objectives improves pathway prioritization beyond conventional enrichment statistics and general-purpose LLM reasoning. We constructed a benchmark consisting of 60 independent cancer-related gene sets derived from bulk and single-cell transcriptomic studies across diverse biological conditions and experimental settings (Supplementary Table 3). For each study, the differentially expressed genes served as input, while significant pathway lists (p-value ≤ 0.05) reported in RummaGEO served as reference pathways. We used PAG-Agent to automatically infer experimental conditions and research objectives from study descriptions provided in RummaGEO. Fig. 2a compares cORA against g:Profiler and four LLM-based baselines across different similarity thresholds, including GPT-4o-mini, Gemini-2.5-Flash, Qwen2.5-7B, and GSAI (GPT4o; "*gpt-4o-2024-08-06*") [5], using accuracy (ACC), hit rate (HIT), and their harmonic mean (F1). At all similarity score thresholds, cORA achieved the highest F1 among all evaluated methods. cORA consistently outperformed g:Profiler, GPT-4o-mini, Gemini-2.5-Flash, Qwen2.5-7B, and GSAI, with statistically significant improvements across most similarity thresholds (one-sided paired Wilcoxon signed-rank test, $p \leq 0.05$). The performance advantage was most pronounced at lower similarity thresholds and remained evident as the similarity threshold became more stringent. These results indicate that incorporating experimental conditions and research objectives improves pathway prioritization beyond both conventional enrichment statistics and LLM reasoning. cORA-identified pathways remained more consistent with study-reported biological findings under increasingly stringent matching criteria.

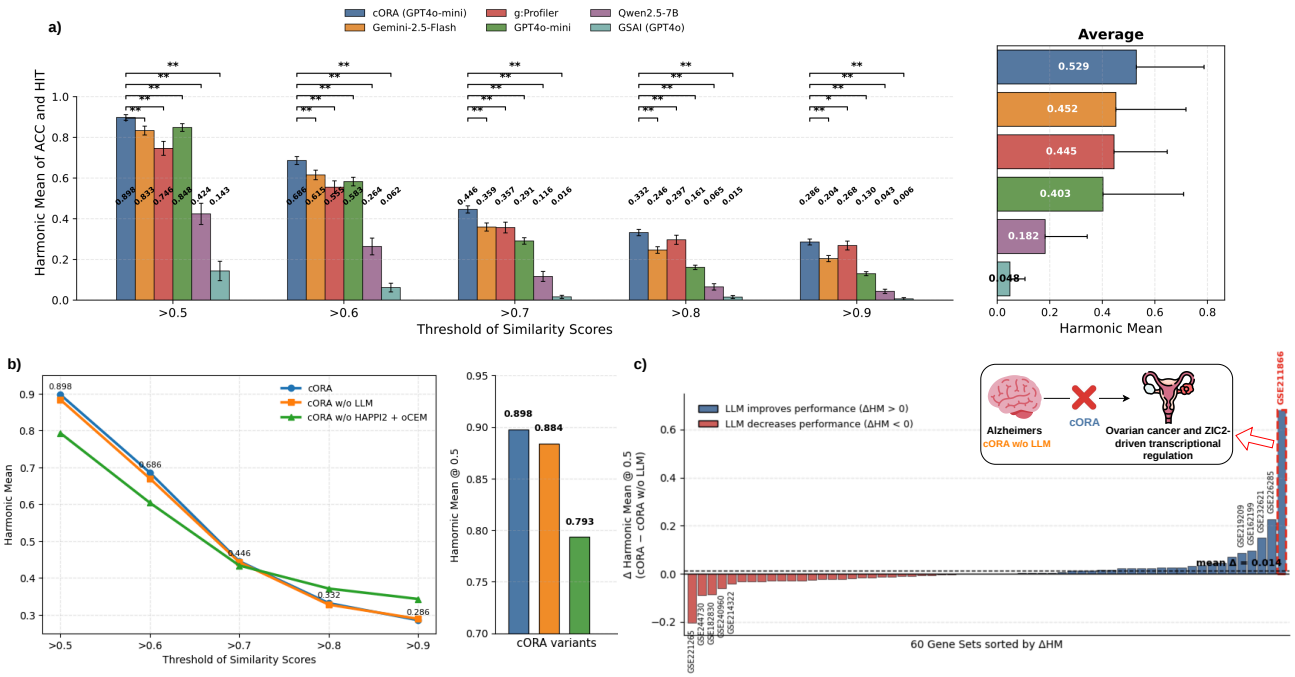

Figure 2: Comparisons between cORA and baseline methods (a) and between cORA and its variants (b,c) across 60 cancer-related gene sets. **a)** Harmonic means of ACC and HIT (F1) across similarity score thresholds for cORA, g:Profiler, GPT-4o-mini, Gemini-2.5-Flash, Qwen2.5-7B, and GSAI. HIT quantifies the recovery of reference pathways, whereas ACC quantifies the agreement between predicted pathways and reference pathways. The right panel summarizes average F1 performance across similarity thresholds. **b)** Ablation analysis of cORA after removing the HAPPI2–oCEM component or GPT-4o-mini. The right panel summarizes average F1 performance at a similarity threshold of 0.5. **c)** Study-level contribution of GPT-4o-mini to cORA at a similarity threshold of 0.5. Error bars represent standard deviations across benchmark gene sets. $^*p \leq 0.05$; $^{**}p \leq 0.01$ (one-sided paired Wilcoxon signed-rank test).

To assess the contribution of individual components, we performed ablation experiments by removing either the HAPPI2–oCEM module detection component or the GPT-4o-mini-based contextual prioritization component (Fig. 2b). HAPPI2 [35], our high-confidence protein–protein interaction resource, provided prior biological knowledge for constructing the gene-gene interaction network. oCEM [36], our overlapping gene module detection framework, identified functional modules that allowed genes to participate in multiple biological processes. Together, HAPPI2 and oCEM enabled biologically informed module detection before pathway prioritization. Removing either component reduced performance across all similarity thresholds. Excluding the HAPPI2–oCEM component produced the largest performance decrease, highlighting the importance of their combination for pathway prioritization. Removing GPT-4o-mini also reduced performance, indicating that contextual information contributed additional value beyond statistical enrichment alone.

We further examined the contribution of GPT-4o-mini to cORA at the study level by comparing performance with and without contextual prioritization (Fig. 2c). GPT-4o-mini improved performance for most benchmark studies, with several studies showing substantial gains in harmonic mean. However, a smaller subset of studies exhibited modest performance decreases after contextual prioritization. These observations suggest that contextual information benefits pathway prioritization in most settings but contributes unevenly across biological contexts. The quality and specificity of contextual descriptions likely influence the effectiveness of contextual prioritization. More explicit project descriptions written directly

by domain experts or project leaders may provide richer contextual information and further improve prioritization performance. Despite this variability, the overall improvement across the benchmark indicates that contextual prioritization provides a net benefit for identifying biologically relevant pathways. For example, in GSE211866, cORA's contextual prioritization replaced the broadly enriched Alzheimer disease pathway with pathways more consistent with the biological context of ZIC2-overexpressing ovarian cancer cells, including *Glycolysis/Gluconeogenesis*, *Mannose type O-glycan biosynthesis*, and *Glycosaminoglycan biosynthesis* (Supplementary Table 4).

2.3 PAG-Agent supports the end-to-end analytical and post-analysis workflow

2.3.1 PAG-Agent supports modality-specific preprocessing and differential expression analysis

Fig. 3 illustrates the user interface of PAG-Agent's Click Interaction and Chat Interaction modules. In the Click Interaction module, we selected the example dataset *GSE153873-RNAseq*, which meant we used the *GSE153873* datasets consisting of the expression data matrix, metadata, and mapping files. The mapping file served as a reference for gene identifier conversion, allowing PAG-Agent to map source IDs in the input data to a unique list of Entrez Gene IDs. While DEA did not require Entrez Gene IDs, this conversion was a prerequisite for PA and integrative analysis (i.e., consensus analysis and meta-analysis).

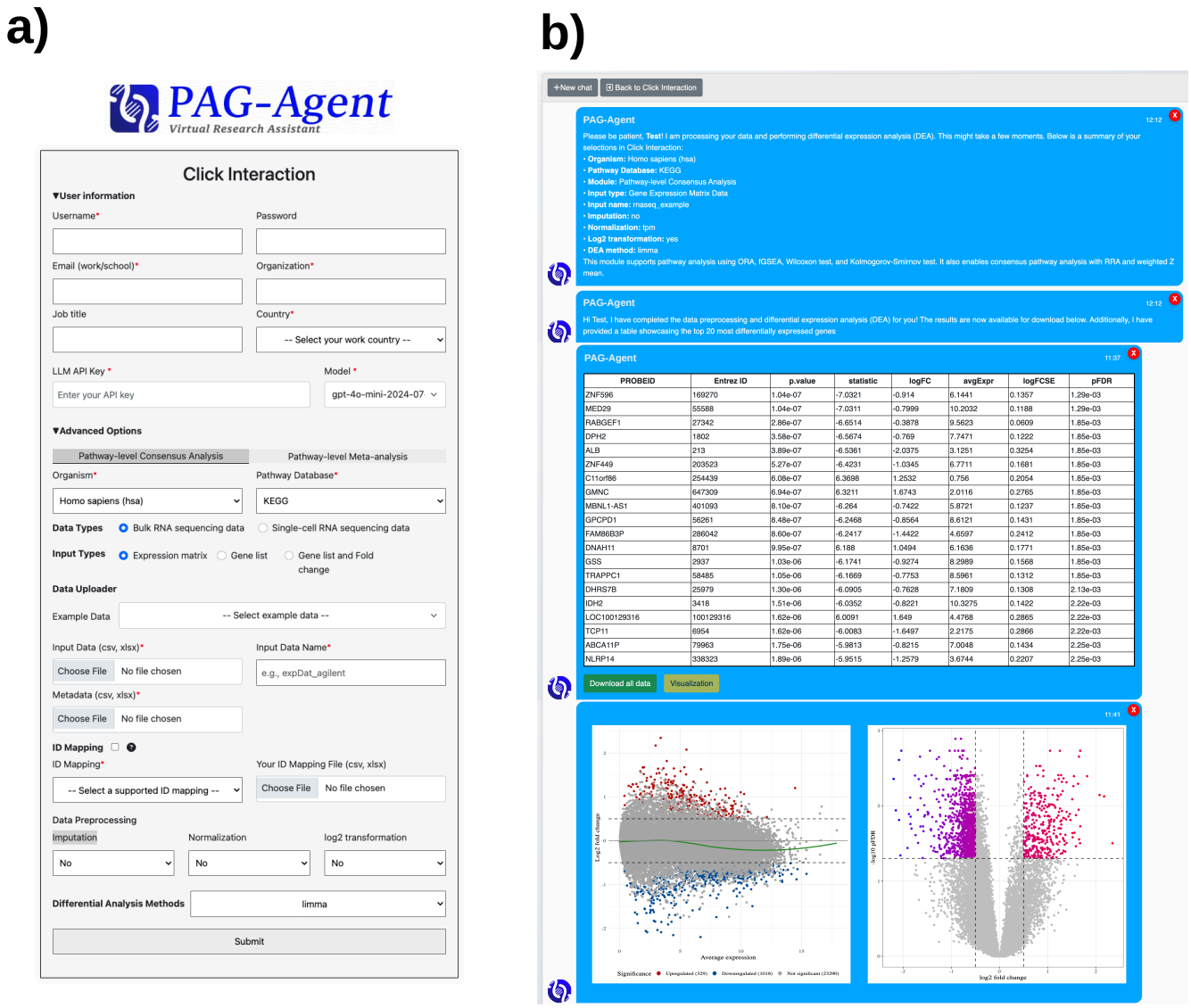

Figure 3: PAG-Agent's Click Interaction (a) and Chat Interaction (b). **a)** User interface of PAG-Agent's Click Interaction (see Supplementary Information Section 1 for detailed descriptions of the fields). **b)** After submission, PAG-Agent preprocesses and performs DEA on *GSE153873*. A preview table shows the top 20 differentially expressed genes, ranked in increasing order of FDR-adjusted p-value (pFDR). All outputs are available for download via "Download all data", providing a ZIP file containing the preprocessed data and DEA result files in CSV format, along with publication-ready visualizations, including MA and Volcano plots, in PNG, SVG, and PDF formats. Visualizations of the analysis outcomes can also be viewed using "Visualization". An MA plot displays the average expression levels (x-axis) against the $\log_2$ fold-change (y-axis), with most points centered around zero, indicating no change; deviations suggest differential expression. The colored points are the DE genes with pFDR < 0.05 and absolute $\log_2$ fold-change > 0.5. A volcano plot shows $\log_2$ fold-change (x-axis) versus the negative $\log_{10}$ p-value (y-axis), highlighting genes with significant expression changes. The colored points are the DE genes with pFDR < 0.05 and absolute log2 fold-change > 0.5.

PAG-Agent provided preprocessing strategies tailored to various expression data types, ensuring compatibility with the specific assay techniques used to generate them (see Methods). For this example, PAG-Agent normalized *GSE153873* (RNAseq) from raw counts to transcripts per million (TPM), applied $\log_2$ transformation to the normalized data, and performed DEA using limma [37] on the preprocessed

dataset (Fig. 3a). For both *GSE61196* (Agilent) and *GSE5281* (Affymetrix), we applied quantile normalization and $\log_2$ transformation for data preprocessing, followed by limma for DEA. After submitting data, PAG-Agent redirected users to separate chat sessions, each corresponding to a single input. The platform provided a preview table of the DEA results displaying the top 20 differentially expressed genes (Fig. 3b). Further, users could click on "Visualization" to generate the MA and Volcano plots, providing a visual summary of the DEA results. All outputs from this step were available for download via the "Download all data" button, which provided a ZIP file containing the preprocessed data and DEA result files in CSV format, along with the MA and volcano plots in PNG, SVG, and PDF formats. The outcomes of this step for *GSE61196* and *GSE5281* were shown in Supplementary Fig. 6 and 7, respectively.

PAG-Agent visualized a gene network retrieved from KEGG. It mapped DEA results (i.e., log2 fold-change values) onto the network. Users applied this feature to a single dataset or integrate multiple DEA results from different datasets. Supplementary Fig. 30 shows when users could leverage this PAG-Agent's feature. To illustrate, we used the gene network of the KEGG Alzheimer disease pathway (`https://www.genome.jp/pathway/hsa05010`). Fig. 4a shows the mapping of log2 fold-change values of differentially expressed genes onto components of the Alzheimer's disease pathway. Most significant genes belonged to components associated with the PI3K complex, proteasome, and Amyloid β. Supplementary Fig. 26 illustrates the integration of the three DEA results from the three Alzheimer's datasets analyzed using limma into the network.

2.3.2 PAG-Agent consistently identifies neurodegeneration-related pathways across Alzheimer's datasets

We continued the current chat session by prompting PAG-Agent to perform PA with KEGG on the three DEA outcomes obtained above, using four available methods: over-representation analysis (ORA), fast Gene Set Enrichment Analysis (fGSEA) [38], Kolmogorov-Smirnov (KS) test, and Wilcoxon test. Each method predicted a list of pathways significantly impacted between the two studied phenotypes (i.e., healthy versus Alzheimer's disease). As shown in Fig. 4b, pathways associated with neurodegenerative diseases ranked prominently, including Alzheimer disease (hsa05012), Parkinson's disease (hsa05012), Huntington's disease (hsa05016), Prion disease (hsa05020), and Pathways of neurodegeneration - multiple diseases (hsa05022). For the *GSE61196* dataset, ORA did not identify any significant pathways (adjusted p-values > 0.05, Benjamini-Hochberg multiple correction [39]), whereas the other methods produced expected results (Supplementary Fig. 12-15). Similarly, for the *GSE5281* dataset, all four PA methods yielded expected results (Supplementary Fig. 16-19).

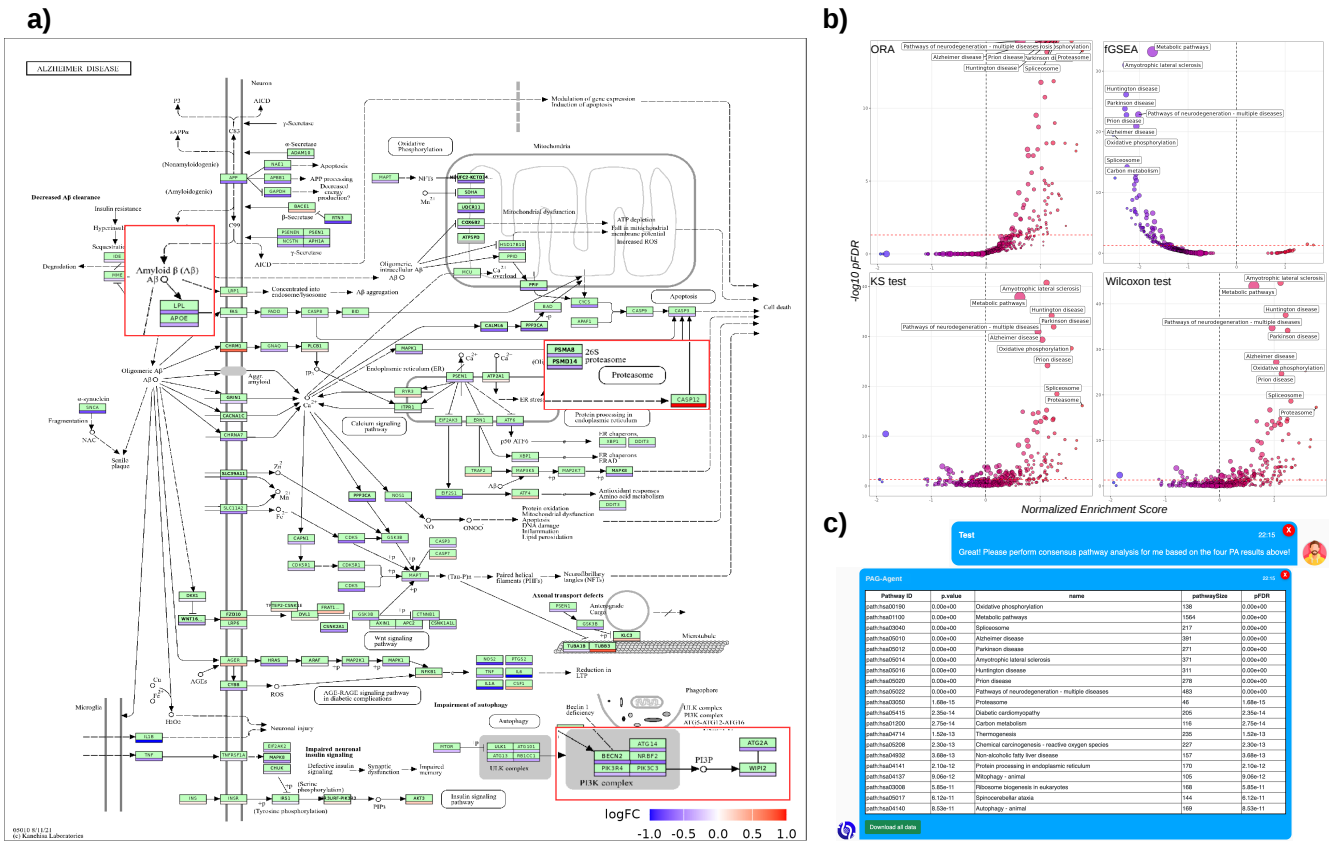

Figure 4: Visualizations of analysis results for *GSE153873*. **a)** Gene network of the KEGG Alzheimer's disease pathway (hsa05010) with expression changes. Green nodes represent genes involved in the pathway, with bars beneath each node indicating the direction of regulation (up- or down-regulated). **b)** Volcano plots displaying pathway analysis (PA) results for *GSE153873* using over-representation analysis (ORA), fast Gene Set Enrichment Analysis (fGSEA), Kolmogorov-Smirnov (KS) test, and Wilcoxon test. **c)** Pathway-level consensus analysis outcomes using weighted Z mean applied to the four PA results (**b**), presented as a preview table showing the top 20 significant pathways ranked by pFDR. Dialogues between a human biologist and PAG-Agent for reproducing these results are detailed in Supplementary Information Section 2.

Next, users were required to perform at least two different PA methods in the chat session before running pathway-level consensus analysis. For *GSE153873* and *GSE5281*, we used all four previously generated PA outcomes. For *GSE61196*, we used three PA outcomes, excluding ORA-based results due to the absence of significant pathways. We selected consensus analysis with weighted Z mean (weightedZMean) [40] as an example (see Methods for a complete list of supported consensus analysis methods). Fig. 4b, Supplementary Fig. 21, and Supplementary Fig. 22 illustrate the top 20 significant pathways with high consensus for *GSE153873*, *GSE61196*, and *GSE5281*, respectively.

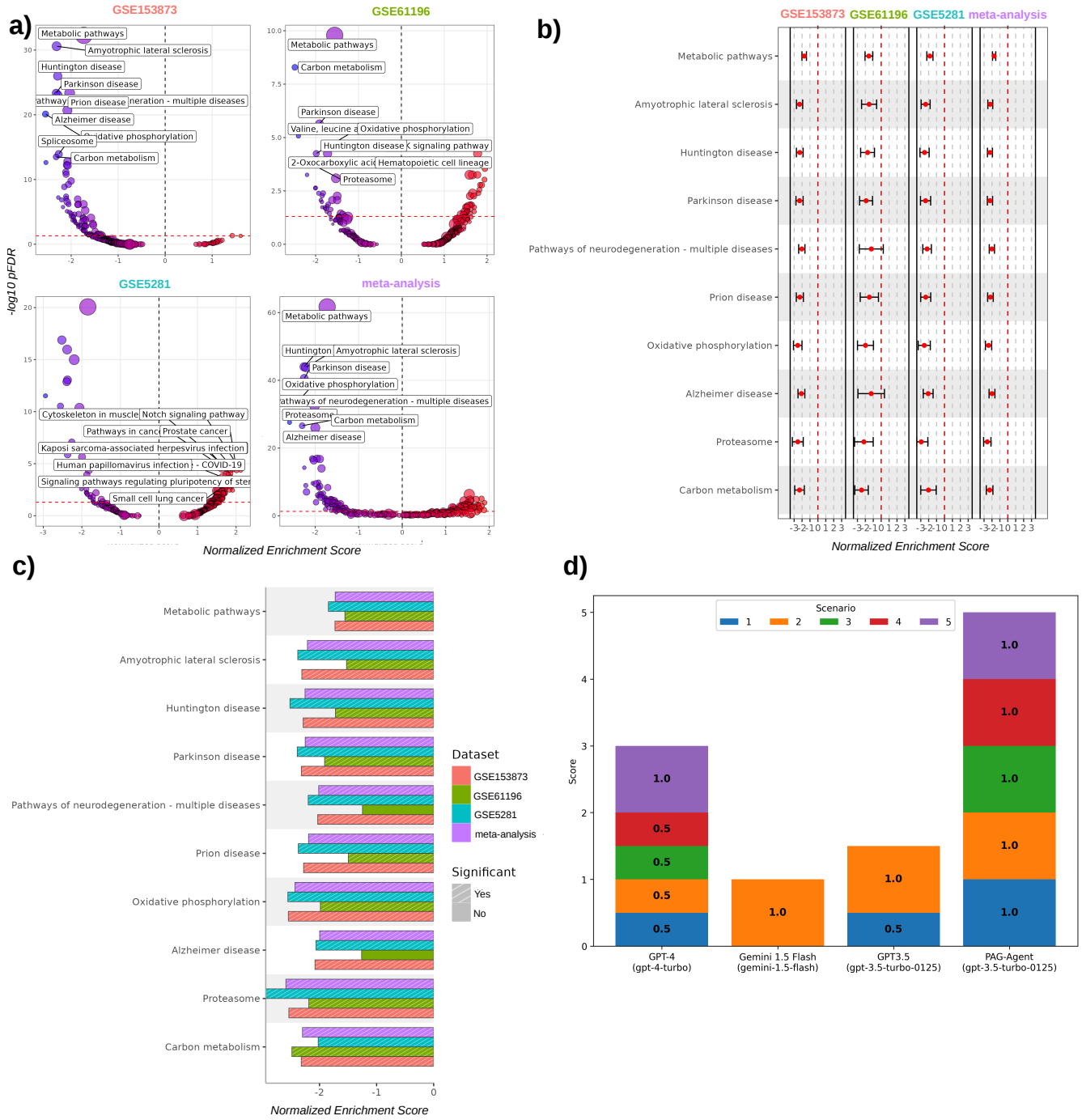

Figure 5: Visualizations of pathway-level meta-analysis results and assessment of PAG-Agent and competing LLMs in providing relevant references. **a-c)** Four Volcano plots, four Forest plots, and bar charts represent the PA results with fast Gene Set Enrichment Analysis (fGSEA) for *GSE153873*, *GSE61196*, *GSE5281*, and the meta-analysis result at the pathway level. The dialogues between a human biologist and PAG-Agent to reproduce the results can be found in Supplementary Information Section 2. **d)** Stacked bar chart showcases the capabilities of PAG-Agent and six competing LLMs in providing relevant references across five common scenarios. In the first three scenarios, users seek scientific citations to validate associations between significantly perturbed pathways and specific conditions, with requests expressed in various styles. The fourth scenario tests the ability of the LLMs to retrieve desired information alongside relevant citations. The final scenario challenges the LLMs to incorporate relevant citations into a long, user-written paragraph (see Methods and Supplementary Table 2 for more details).

In addition, PAG-Agent supported pathway-level meta-analysis. Users could not perform meta-analysis within the current chat session. Instead, they had to return to the Click Interaction module and select Pathway-level Meta-analysis (Supplementary Fig. 1c). Pathway-level meta-analysis required analysis outcomes as input at the gene or pathway level. For gene-level outcomes, users provided a list of DEA outcomes. For pathway-level outcomes, they provided a list of PA outcomes. To demonstrate this functionality, we asked PAG-Agent to perform meta-analysis using Stouffer's method [41] (see Methods for a complete list of supported meta-analysis methods). We inputted a set of three fGSEA-based PA outcomes generated earlier for *GSE153873* (RNAseq), *GSE61196* (Agilent), and *GSE5281* (Affymetrix) into PAG-Agent to execute meta-analysis at the pathway level. PAG-Agent simplified the result presentation via Volcano plots (Fig. 5a), Forest plots (Fig. 5b), and Bar charts (Fig. 5c). The complete meta-analysis results for gene- and pathway-level inputs were available in Supplementary Fig. 23 and 24, respectively.

In summarize, PAG-Agent performed a total of 16 pathway-related analyses across the three Alzheimer's datasets. Six pathways associated with neurodegenerative disorders, including Pathways of neurodegeneration - multiple diseases, Alzheimer's disease, Huntington's disease, Parkinson's disease, Amyotrophic lateral sclerosis, and Prion disease, were consistently significant in 15 out of 16 analyses.

2.3.3 PAG-Agent facilitates literature-grounded interpretation and hypothesis generation

Detailed annotations for genes or pathways provided critical biological context for interpreting results and selecting candidates for follow-up studies. Traditionally, computational scientists performing DEA and PA would deliver lists of significant genes or pathways to biologists without additional context. Consequently, experimentalists had to identify relevant knowledge bases, extract, and analyze the required information to address their queries. Common questions included understanding the biological role of a gene or pathway, identifying related pathways, determining which drugs targeted a specific gene or pathway, and exploring associations between genes or pathways and the studied phenotypes. This process was particularly challenging for novice biologists unfamiliar with bioinformatics. For example, in Fig. 3b, *ALB* (Entrez ID 213) ranked high on the DEA results, while PAG-Agent identified metabolic pathways (hsa01100) as significant (Fig. 4). However, even domain experts struggled to immediately provide accurate biological context or justify associations with Alzheimer's disease. PAG-Agent, leveraging the combined power of GPT-4o-mini and in-context learning techniques (Methods), rapidly generated detailed and accurate annotations of these genes and pathways as well as made hypothesis on their relation with the studied conditions, as shown in Supplementary Fig. 27 and 28. These AI-generated hypotheses provided biologists with interpretable results or actionable starting points for further investigation. Notably, we encouraged LLMs to generate creative and novel hypotheses, exploring plausible connections between discoveries and phenotypes. We tolerated outputs containing hallucinations when they introduced potential insights that had not been previously considered.

A key aspect of comparative studies involved determining whether any previous articles reported similar findings. Supplementary Fig. 29 illustrates PAG-Agent, integrated with a search engine (see Methods), retrieved relevant references linking Oxidative Phosphorylation to Alzheimer's disease. It further summarized these references and presented them in an academic format suitable for immediate use. To benchmark its performance, we compared PAG-Agent against six competing LLMs across five common scenarios (see Methods and Supplementary Table 2). Importantly, we intentionally downgraded GPT-4o-mini (version 'gpt-4o-mini') to GPT-3.5 Turbo (version 'gpt-3.5-turbo-0125') as the backend for PAG-Agent to highlight the effectiveness of our proposed strategy for mitigating hallucinations. As shown in Fig. 5d, PAG-Agent consistently outperformed the competing LLMs. Open-weight models, including Phi3.5, Llama 3.1, and Mistral, failed all five scenarios. Scenarios 3, 4, and 5 posed the greatest challenges. While ChatGPT-4 (version 'gpt-4-turbo') scored 0.5, 0.5, and 1, respectively, PAG-Agent achieved perfect scores in all three scenarios. Generating a complete response per scenario took PAG-Agent 3–5 seconds, depending on re-

sponse length and internet speed, comparable to GPT-3.5 Turbo via its API. Notably, GPT-3.5 Turbo without our proposed strategy performed well in only one scenario and moderately in another, achieving a total score of 1.5 out of 5. This result underscored how our strategy substantially reduced AI hallucinations without increasing runtime or computational cost.

English was the global language of science and academic writing [42]. However, non-native English speakers faced significant challenges in manuscript writing, requiring more time and effort [43]. Studies highlighted that these difficulties stem primarily from limited language proficiency [43–45]. Common struggles included difficulties in self-expression, slower writing processes, and restricted vocabulary [44]. These challenges arise not only from linguistic factors such as grammar, vocabulary, and sentence structure but also from meta-linguistic factors, including logical sentence connections, paragraph development, and overall organization. Advancements in internet tools and computational technologies made it easier for researchers to address linguistic issues [46]. However, meta-linguistic challenges remained a significant barrier. Poor English proficiency could also delay research publication [47]. PAG-Agent, powered by advanced LLMs, assisted researchers in refining text and improving overall writing quality. By enhancing clarity and coherence, it promoted equality among researchers, regardless of linguistic background or nationality.

3 Discussion

Pathway-level analysis often extends beyond statistical enrichment. Researchers must determine which significant pathways are most relevant to the biological question under investigation and subsequently integrate biological knowledge, literature evidence, and study-specific context to interpret these findings. Existing pathway-analysis platforms provide many of the statistical components required for these tasks, whereas recent LLM-based approaches facilitate biological interpretation. However, pathway prioritization and pathway interpretation are typically addressed through separate tools and workflows. PAG-Agent addresses this gap by combining context-aware pathway prioritization, biological interpretation, literature-supported reasoning, and research assistance within a unified platform.

A key contribution of PAG-Agent is cORA, which incorporates experimental conditions and research objectives into pathway prioritization. Across 60 cancer-related gene sets derived from bulk and single-cell transcriptomic studies, cORA consistently achieves higher Accuracy, HIT rate, and F1 score than conventional ORA and four LLM-based baselines. These results suggest that contextual information can improve the identification of biologically relevant pathways beyond statistical significance alone. Ablation experiments further show that both biologically informed module detection and contextual prioritization contribute to cORA performance. HAPPI2 and oCEM provide a biologically informed foundation for identifying functional gene modules, whereas GPT-4o-mini incorporates study-specific context to prioritize pathways relevant to the biological question under investigation. We also observe substantial variability in the contribution of contextual prioritization across studies. This observation suggests that the quality and specificity of contextual descriptions influence pathway prioritization performance. More explicit project descriptions written directly by domain experts may therefore further improve context-aware pathway prioritization.

A central design objective of PAG-Agent is to improve the reliability of literature-supported responses without relying on large, manually maintained knowledge repositories. Retrieval-augmented generation (RAG) systems have emerged as a common strategy for reducing hallucinations by incorporating external knowledge during inference. However, maintaining large collections of scientific materials requires continuous updates and additional computational resources. PAG-Agent instead combines lightweight information retrieval with in-context learning. Relevant references are dynamically retrieved from external sources and incorporated into prompts as contextual evidence for response generation. This strategy enables access to

recent scientific findings while maintaining a lightweight architecture. The strong performance of PAG-Agent relative to both proprietary and open-weight LLMs suggests that retrieval quality and contextual grounding may be more important than model size alone for literature-supported biological interpretation.

Overall, PAG-Agent lowers the technical barriers associated with pathway prioritization and interpretation. By combining context-aware pathway prioritization, biological interpretation, literature support, and scientific communication within a unified workflow, PAG-Agent enables researchers to move more efficiently from transcriptomic data to biological insight.

4 Methods

4.1 Large language model behind PAG-Agent

PAG-Agent supports multiple LLMs through their corresponding Application Programming Interfaces (APIs), including 'gpt-3.5-turbo-0125', 'gpt-4o-mini-2024-07-18', 'gpt-5-nano-2025-08-07', 'gemini-2.0-flash-lite', and 'gemini-2.5-flash-lite'. These models allow adjustment of the 'temperature' parameter, which controls response variability. Lower temperatures generally produce more reproducible and conservative outputs [48, 49]. The effect of the temperature parameter on LLM analyses is beyond the scope of this study. Unless otherwise specified, all analyses reported in this study were conducted using 'gpt-4o-mini-2024-07-18'. PAG-Agent uses a temperature of 1.5 for gene and pathway annotation tasks and 0.7 for all other tasks. A higher temperature encourages broader biological exploration and more creative hypothesis generation.

4.2 Prompt engineering

The collaborative functioning of eight engineered prompts underpins the smooth operation of PAG-Agent (Supplementary Fig. 31). These prompts included: (i) **User Intent Detection**: This prompt detects user intents, each related to a specific task. (ii) **Pathway Analysis Method Detection**: This prompt identifies the pathway analysis method requested by the user. (iii) **Meta-analysis Method Detection**: This prompt determines the meta-analysis method requested by the user. (iv) **Gene Annotation**: This prompt places a gene to its biological context. (v) **KEGG Pathway Annotation**: This prompt provides information related to a specific KEGG pathway. (vi) **Gene Ontology Term Annotation**: This prompt provides information about a specific Gene Ontology (GO) term. (vii) **Citation Retrieval**: This prompt finds relevant and valid citations based on user-provided context. (viii) **Writing Refinement**: This prompt refines user-submitted academic writing.

We engineer these prompts to identify predefined keywords that indicate user intents for PAG-Agent. **User Intent Detection** includes 10 preset user intents, each linked to a list of keywords. For example, the pathway analysis intent includes keywords: "*Users mention the following keywords: "pathway analysis", "enrichment analysis", "functional analysis", "gene set analysis", "gene set enrichment analysis", "fgsea", "ora", "over-representation analysis", "ks", "ks-test", "ks test", "kolmogorov-smirnov", "kolmogorov-smirnov test", "wilcox", "wilcox-test", "wilcox test", "wilcoxon", "wilcoxon-test", "wilcoxon test", "hypergeometric test", "hypergeometric", "fisher", "fisher's exact test", "fisher's exact", or "fisher exact test*"". When a user submits a query like "*Can you perform ora for me?*", PAG-Agent detects the keyword "ora" and maps it to the PA intent. The system then proceeds to the **Pathway Analysis Method Detection** prompt to determine the specific method (over-representation analysis in this example). Supplementary Fig. 31 outlines the hierarchical organization of these prompts.

To minimize hallucinations in responses from the **Gene Annotation**, **KEGG Pathway Annotation**, **Gene Ontology Term Annotation**, and **Citation Retrieval** prompts, we integrate recent and relevant information through in-context learning strategies into the system prompts. For **Gene Annotation**, we

use species-specific annotation packages from Bioconductor (`https://bioconductor.org/packages/3.20/data/annotation/`; release 3.20.0) to retrieve biological data. For instance, when querying the human gene *DPH2* (Entrez ID: 1802) with a prompt like "*I want to get the information of the gene 1802 within the context of Alzheimer's disease*", the R annotation package org.Hs.eg.db fetches data from NCBI (`https://www.ncbi.nlm.nih.gov/gene/1802`) and provides relevant context for PAG-Agent to generate accurate responses. A similar approach is applied to **KEGG Pathway Annotation** and **Gene Ontology Term Annotation**, using the R packages KEGGREST (version 1.46.0) and GO.db, respectively. For **Citation Retrieval**, we leverage the DuckDuckGo Search API (`https://api.duckduckgo.com/`), accessed via the Python library duckduckgo_search (version 7.1.0). This enables the retrieval of valid citations, reducing hallucinations and enhancing the reliability of responses.

4.3 Assessment of PAG-Agent's capability to provide relevant references

Providing scientific references is crucial for enhancing the reliability of analysis results generated by statistical methods. While LLMs can provide citations, their responses often include hallucinated references. To address this, we design PAG-Agent to rapidly deliver scientific citations that are contextually relevant and valid. To evaluate our strategy for efficient citation retrieval, we intentionally downgrade 'gpt-4o-mini' to 'gpt-3.5-turbo-0125'. Six LLMs are selected for evaluation: GPT-3.5 Turbo (version 'gpt-3.5-turbo-0125'), ChatGPT-4 (version 'gpt-4-turbo'), Google's Gemini 1.5 Flash, Meta's Llama 3.1 8B, Mistral 7B from the Mistral AI team, and Microsoft's Phi-3.5 3.8B. We access ChatGPT-4 and Gemini 1.5 Flash through their official platforms (`https://chatgpt.com/` and `https://gemini.google.com/app`), while Llama 3.1, Mistral, and Phi-3.5 are queried via the API endpoint of Ollama (`https://ollama.com/`). The 'gpt-3.5-turbo-0125' version of GPT-3.5 Turbo is utilized through its API. The evaluation involves five common scenarios, including but not limited to Alzheimer's disease. In the first three scenarios, users seek scientific citations to validate associations between significantly perturbed pathways and specific conditions, with requests expressed in various styles. The fourth scenario tests the ability of the LLMs to retrieve desired information alongside relevant citations. The final scenario challenges the LLMs to incorporate relevant citations into a long, user-written paragraph. The prompt strategy for each situation is detailed in Supplementary Table 2. The system prompt used for the six competing LLMs is: "*You are a useful assistant. Please provide related citations and include detailed summaries for each*". In contrast, PAG-Agent utilizes this system prompt in combination with relevant references retrieved from DuckDuckGo. Importantly, the tempurature parameter cannot be controlled for ChatGPT-4 and Gemini 1.5 Flash, as they are accessed via their web interfaces.

LLM-generated responses are classified into three categories: 'fully match', if all citations are real, valid, and contextually relevant; 'partially match', if the response includes real and contextually relevant references but not suitable as scientific citations (e.g., online news or non-peer-reviewed sources, excluding open-access preprint repositories like arXiv or bioRxiv); and 'mismatch', if the LLM hallucinates citations or fails to address the task adequately. To enable comparison, scores of 1, 0.5, and 0 are assigned to responses classified as 'fully match', 'partially match', and 'mismatch', respectively.

4.4 Software architecture

The PAG-Agent platform architecture comprises front-end and back-end components. The front-end is built with jQuery and JavaScript, featuring UI components for data uploading and browsing, as well as modules for data fetching and operations. The back-end consists of three core components: (i) the Python FastAPI framework, which manages client requests and application logic; (ii) Nginx (`https://www.nginx.com`), which serves static files and acts as a reverse proxy; and (iii) R analysis workers, which perform data preprocessing, differential expression analysis, pathway analysis, integrative analysis, visualizations,

and context retrieval for annotations. Data is stored in MongoDB (`https://www.mongodb.com`), while Redis (`https://redis.io/`) supports high-throughput session management and real-time features. This architecture is designed to handle complex biological data analyses with reliability, scalability, and high performance.

4.5 Input and data management

The PAG-Agent platform supports bulk and single-cell transcriptomic datasets through three input types: (i) a gene expression matrix (Supplementary Fig. 2), (ii) a set of DEA outcomes (Supplementary Fig. 3 and 4), and (iii) a set of pathway analysis outcomes (Supplementary Fig. 4). All input data can be provided in either .csv (comma-separated) or Excel .xlsx formats. For expression matrix input, the dataset can include two separate files: one for the expression matrix and another for metadata. The metadata file must contain two main columns: the first lists sample identifiers, and the second specifies their corresponding conditions (e.g., normal or disease). The platform also supports conversions from other gene identifiers to Entrez Gene IDs. Users can either select an available mapping file or import a custom mapping file (Supplementary Fig. 2) for this ID-to-Entrez conversion.

PAG-Agent includes a file manager for uploading, managing, and downloading expression data. Users can upload files directly from their local machines. To free storage space, files uploaded by anonymous users are automatically deleted after 24 hours. Registered users, identified by username, password, and email, can save data permanently and access their chat sessions across multiple devices. Each chat session receives a unique identifier. This identifier allows users to access past interactions with PAG-Agent without re-uploading input data or reselecting parameters in the Click Interaction module, provided the input data remains available. This feature ensures the reproducibility of analysis outcomes.

4.6 Data preprocessing

Data preprocessing is a critical step for all statistical analyses. Zhang et al. [50] have demonstrated that pathway analysis is highly influenced by the normalization procedures applied to gene expression measurements. PAG-Agent provides three preprocessing options depending on input data.

Imputation methods, including mean, median, and zero imputation, are available to handle missing values by replacing them with the average, median, or zero value of the corresponding gene across all samples. Users should select the most appropriate method based on the dataset characteristics. Mean or median imputation is generally suitable for Affymetrix and Agilent data (continuous data) to account for random missing values or reduce skewness. Zero imputation, in contrast, is often more suitable for count-based transcriptomic data, including bulk RNA-seq and scRNA-seq data, where zero values may reflect biological absence or technical sparsity.

For bulk transcriptomic data, normalization options include TPM normalization and quantile normalization. TPM normalization is recommended for bulk RNA-seq data, whereas quantile normalization is recommended for microarray data. For single-cell RNA-seq data, PAG-Agent provides quality-control filtering based on the number of detected genes, total UMI counts, and mitochondrial gene fraction. After filtering, the platform applies library-size normalization followed by natural-log transformation to improve comparability across cells.

For bulk transcriptomic data, log2 transformation is applied to stabilize variance and reduce skewness. For single-cell RNA-seq data, PAG-Agent applies natural-log transformation following library-size normalization as part of the standard preprocessing workflow.

4.7 Parameter setting for differential expression analysis

Differential expression analysis (DEA) is a fundamental approach for identifying genes that are differentially expressed between distinct conditions or phenotypes. Numerous robust techniques have been developed for differential analysis [51–53]. In this work, PAG-Agent supports multiple differential expression analysis methods tailored to the characteristics of bulk and single-cell transcriptomic data. For bulk transcriptomic datasets, the platform provides limma [37], DESeq2 [52], and edgeR [54]. For scRNA-seq datasets, PAG-Agent supports the Wilcoxon rank-sum test, MAST [55], logistic regression, and DESeq2.

Users can select the most appropriate method based on the assay technology and data representation. Limma is typically applied to continuous data, such as microarray expression or TPM-normalized bulk RNA-seq data, whereas DESeq2 and edgeR are suited for count-based bulk RNA-seq data or scRNA-seq data. For single-cell RNA-seq analyses, the Wilcoxon rank-sum test and logistic regression provide robust non-parametric and discriminative alternatives, while MAST explicitly models zero inflation and cellular detection rates commonly observed in single-cell measurements. The primary output of these methods includes a list of genes with their corresponding fold-changes, p-values, adjusted p-values. Adjusted p-values, calculated using the Benjamini-Hochberg procedure [39], are used to identify significantly differentially expressed genes.

4.8 Parameter setting for pathway-level analyses

The platform supports several widely used pathway analysis (PA) methods, including ORA, fGSEA [38] (default), the KS test, and the Wilcoxon test, each designed to detect different data patterns. In this work, the default method refers to cases where the User Intent Detection prompt identifies a user intent for PA, but the Pathway Analysis Method Detection prompt cannot determine the preferred method (e.g., due to typos or unspecified method names). In such cases, PAG-Agent defaults to using fGSEA.

PAG-Agent currently supports 10 organisms: *H. sapiens*, *M. musculus*, *R. norvegicus*, *D. rerio*, *D. melanogaster*, *C. elegans*, *S. cerevisiae*, *A. thaliana*, *S. pombe*, and *P. falciparum*. Gene sets for these species from KEGG (Release 112.1) and GO (2024-11-03) are integrated into the platform. Previous studies have highlighted the adverse impact of smaller gene sets on pathway analysis performance [4, 50]. Damian and Gorfine [56] attributed this to larger gene variances in smaller gene sets, which affect the accuracy of test statistics for enrichment. To address this, PAG-Agent excludes gene sets containing fewer than 10 genes.

Existing web-based tools for pathway analysis face several limitations: (i) they cannot combine, compare, or contrast results from multiple methods, (ii) they lack the ability to integrate results across datasets, and (iii) they do not support comprehensive visualization of gene networks and expression changes simultaneously. PAG-Agent addresses these shortcomings through two key strategies: pathway-level consensus analysis and pathway-level meta-analysis.

Pathway-level consensus analysis, also known as multi-method analysis, combines results from different PA methods, such as the KS test and fGSEA, applied to a single gene expression dataset. This approach enhances robustness by identifying pathways consistently enriched across methods. PAG-Agent provides two methods for consensus analysis: weighted Z mean (default) and robust rank aggregation (RRA) [57].

Pathway-level meta-analysis, also known as multi-method and multi-cohort analysis, integrates pathway- or gene-level results across multiple datasets, offering a unified perspective and improving reliability. Pathway-level results are generated from (i) a single dataset using different combinations of DEA and PA methods or (ii) multiple datasets from different experiments analyzed with similar or varied approaches. Similarly, gene-level results combines DEA results within the same context. PAG-Agent supports several established meta-analysis methods: Fisher's method [58], Stouffer's method [41] (default), minimum p-value [59], geometric mean, addCLT [60], and restricted (residual) maximum likelihood (REML) [61].

These methods are chosen for their ability to account for heterogeneity across studies, yielding more accurate combined statistics.

Importantly, identification through these integrative analyses does not necessarily imply greater biological significance for a specific gene or pathway. Instead, these methods provide a robust foundation for generating hypotheses that can be explored further in experimental studies.

4.9 context-aware Over-Representation Analysis (cORA)

4.9.1 cORA workflow

To improve the biological relevance of pathway prioritization, PAG-Agent introduces the cORA workflow. Given a set of differentially expressed genes D, we first construct a weighted gene–gene interaction network $G = (V, E, W)$, where V denotes genes, E denotes interactions retrieved from HAPPI2, and W denotes the corresponding HAPPI confidence scores. HAPPI2 is a comprehensive human protein–protein interaction (PPI) database that integrates both experimentally validated and computationally predicted physical and functional interactions with associated confidence scores [35]. The interaction network is subsequently represented as a weighted adjacency matrix A, where each entry A_{ij} corresponds to the confidence score assigned to the interaction between genes i and j. Thus, A captures both the connectivity and confidence of known functional relationships among genes. Next, we apply oCEM [36] to the weighted adjacency matrix to identify overlapping gene modules:

$$\mathcal{M} = \mathrm{oCEM}(G, \Theta),$$

where Θ denotes the parameters of oCEM and $\mathcal{M} = \{M_1, \ldots, M_K\}$ represents the set of detected overlapping gene co-expression modules. Specifically, oCEM uses independent component analysis to decompose the interaction network into partially overlapping functional modules, allowing genes to participate in multiple biological processes. This property better reflects the pleiotropic nature of biological systems compared with non-overlapping clustering approaches. For each module M_i, we perform ORA against KEGG:

$$\mathcal{P}_i = \mathrm{ORA}(M_i, \alpha),$$

where α denotes the significance threshold and $\mathcal{P}_i$ represents the set of enriched pathways associated with module M_i. Because multiple enriched pathways may represent related or partially redundant biological processes, PAG-Agent further prioritizes these candidates using contextual information available from the original study. Specifically, PAG-Agent leverages experimental conditions C_E to estimate the contextual relevance of candidate pathways using GPT-4o-mini with a temperature of zero:

$$R_i = \mathrm{LLM}(\mathcal{P}_i, C_E),$$

where R_i denotes the relevance scores assigned to pathways in $\mathcal{P}_i$. The model evaluates the consistency between each pathway and the experimental context. Experimental conditions assess pathway relevance within the biological setting of the study, whereas research objectives guide the selection of pathways that best address the underlying scientific question. Finally, PAG-Agent uses GPT-4o-mini with a temperature of zero to select a representative pathway for each module by jointly considering pathway significance, module quality, contextual relevance, and research objectives:

$$P_i^* = \mathrm{LLM}\big(\mathcal{P}_i, S_i, R_i, C_R\big)$$

where S_i denotes pathway significance derived from ORA and weighted by the CoCo score, a network-based measure of module cohesiveness and topological quality in which higher values indicate more biologically

coherent modules [62]; R_i denotes contextual relevance scores; and C_R denotes the research objectives. The resulting representative pathway P_i^* is selected as the pathway that best explains the biological question under investigation.

4.9.2 Benchmark design

Evaluating context-aware pathway prioritization requires benchmark datasets in which biologically relevant pathways can be linked to experimentally derived gene sets. Existing PA benchmarks primarily evaluate statistical enrichment performance and provide limited support for assessing pathway prioritization in the context of specific biological questions. To address this limitation, we constructed a benchmark dataset from RummaGEO (`https://rummageo.com/`), a large collection of gene sets derived from publicly available transcriptomic studies. For each study, we collected the differentially expressed gene sets and associated study descriptions reported in RummaGEO. We then used PAG-Agent to automatically infer experimental conditions and research objectives from the corresponding study descriptions. Next, we applied a series of quality-control criteria to retain only studies with sufficient information for evaluating pathway prioritization (see Supplementary Information Section 5).

After filtering, the final benchmark consisted of 60 independent studies spanning diverse biological conditions and experimental settings (Supplementary Table 3). For each study, the differentially expressed genes served as the input, while the corresponding significant pathway lists ($p \leq 0.05$) reported in RummaGEO were treated as reference pathways for evaluation. We compared cORA against a conventional ORA-based prioritization strategy and four LLM-based baselines, including 'gpt-4o-mini-2024-07-18', 'gemini-2.5-flash-lite', 'Llama-3.1-8B', 'Qwen2.5-7B', and 'Phi-3.5-mini'. This benchmark enabled systematic assessment of whether incorporating experimental conditions and research objectives improves pathway prioritization beyond conventional enrichment statistics and general-purpose LLM reasoning.

Direct comparison of pathway names can underestimate performance because biologically equivalent pathways may be described using different terminology. For example, a reference pathway such as 'Regulation of apoptosis' and a predicted pathway such as 'Apoptotic process regulation' may describe highly related biological processes despite not being exact string matches. To account for such semantic variation, we employed MedCPT, a biomedical text encoder, to generate embedding representations for both reference and predicted pathways. Similarity between two pathways was calculated using cosine similarity between their corresponding embedding vectors. Using these similarity scores, we quantified the agreement between predicted pathways and reference pathways using HIT rate, accuracy (ACC), and F1 score. For each benchmark study, let Y denote the set of reference pathways reported in the original study and P denote the set of pathways predicted by a method. A predicted pathway ($p \in P$) was considered a match if its maximum similarity to at least one reference pathway ($y \in Y$) exceeded a predefined similarity threshold (τ). We evaluated performance across similarity thresholds ranging from 0.5 to 0.9. We primarily focus on results at $\tau = 0.5$, which allows biologically related pathways with different names to be considered matches. We defined the HIT rate as

$$\text{HIT} = \frac{|\{y \in Y : \exists p \in P, \, \text{sim}(p, y) \geq \tau\}|}{|Y|}$$

which measures the proportion of reference pathways recovered by a method. We defined accuracy (ACC) as

$$\text{ACC} = \frac{|\{p \in P : \exists y \in Y, \, \text{sim}(p, y) \geq \tau\}|}{|P|}$$

which measures the proportion of predicted pathways supported by the reference annotations. To summarize overall performance, we calculated the harmonic mean of ACC and HIT (F1):

$$F1 = \frac{2 \times \text{ACC} \times \text{HIT}}{\text{ACC} + \text{HIT}}$$

Higher values of ACC, HIT, and F1 indicate better agreement between predicted pathways and study-reported biological findings.

Data availability

The three Alzheimer's disease datasets, GSE5281, GSE61196, and GSE153873, were downloaded from NCBI GEO [63] and used as example datasets in this study. We additionally constructed a benchmark dataset from RummaGEO (`https://rummageo.com/`), from which 60 studies satisfying the filtering criteria described in Supplementary Information Section 5 were retained for evaluation (Supplementary Table 3).

Code availability

The complete prompt engineering details and code for reproducing evaluation task results are available on Github `https://github.com/aimed-lab/PAG-Agents` under the MIT License. The PAG-Agent website (`https://huynguyen250896.github.io/pag-agent-homepage/`) is free, open to all users, and does not require login.

References

[1] Minoru Kanehisa, Miho Furumichi, Mao Tanabe, Yoko Sato, and Kanae Morishima. KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Research*, 45(D1):D353–D361, 2017.

[2] Michael Ashburner, Catherine A. Ball, Judith A. Blake, David Botstein, Heather Butler, J. Michael Cherry, Allan P. Davis, Kara Dolinski, Selina S. Dwight, Janan T. Eppig, Midori A. Harris, David P. Hill, Laurie Issel-Tarver, Andrew Kasarskis, Suzanna Lewis, John C. Matese, Joel E. Richardson, Martin Ringwald, Gerald M. Rubin, and Gavin Sherlock. Gene Ontology: tool for the unification of biology. *Nature Genetics*, 25:25–29, 2000.

[3] The Gene Ontology Consortium, Suzi A. Aleksander, James Balhoff, Seth Carbon, J. Michael Cherry, Harold J. Drabkin, Dustin Ebert, Marc Feuermann, Pascale Gaudet, Nomi L. Harris, David P. Hill, Raymond Lee, Huaiyu Mi, Sierra Moxon, Christopher J. Mungall, Anushya Muruganugan, Tremayne Mushayahama, Paul W. Sternberg, Paul D. Thomas, Kimberly Van Auken, Jolene Ramsey, Deborah A. Siegele, Rex L. Chisholm, Petra Fey, Maria Cristina Aspromonte, Maria Victoria Nugnes, Federica Quaglia, Silvio Tosatto, Michelle Giglio, Suvarna Nadendla, Giulia Antonazzo, Helen Attrill, Gil dos Santos, Steven Marygold, Victor Strelets, Christopher J. Tabone, Jim Thurmond, Pinglei Zhou, Saadullah H. Ahmed, Praoparn Asanitthong, Diana Luna Buitrago, Meltem N. Erdol, Matthew C. Gage, Mohamed Ali Kadhum, Kan Yan Chloe Li, Miao Long, Aleksandra Michalak, Angeline Pesala, Armalya Pritazahra, Shirin C. C. Saverimuttu, Renzhi Su, Kate E. Thurlow, Ruth C. Lovering, Colin Logie, Snezhana Oliferenko, Judith Blake, Karen Christie, Lori Corbani, Mary E. Dolan, Harold J. Drabkin, David P. Hill, Li Ni, Dmitry Sitnikov, Cynthia Smith, Alayne Cuzick, James Seager, Laurel Cooper, Justin Elser, Pankaj Jaiswal, Parul Gupta, Pankaj Jaiswal, Sushma Naithani, Manuel Lera-Ramirez, Kim Rutherford, Valerie Wood, Jeffrey L. De Pons, Melinda R. Dwinell, G. Thomas Hayman, Mary L. Kaldunski, Anne E. Kwitek, Stanley J. F. Laulederkind,

Marek A. Tutaj, Mahima Vedi, Shur-Jen Wang, Peter D'Eustachio, Lucila Aimo, Kristian Axelsen, Alan Bridge, Nevila Hyka-Nouspikel, Anne Morgat, Suzi A. Aleksander, J. Michael Cherry, Stacia R. Engel, Kalpana Karra, Stuart R. Miyasato, Robert S. Nash, Marek S. Skrzypek, Shuai Weng, Edith D. Wong, Erika Bakker, Tanya Z. Berardini, Leonore Reiser, Andrea Auchincloss, Kristian Axelsen, Ghislaine Argoud-Puy, Marie-Claude Blatter, Emmanuel Boutet, Lionel Breuza, Alan Bridge, Cristina Casals-Casas, Elisabeth Coudert, Anne Estreicher, Maria Livia Famiglietti, Marc Feuermann, Arnaud Gos, Nadine Gruaz-Gumowski, Chantal Hulo, Nevila Hyka-Nouspikel, Florence Jungo, Philippe Le Mercier, Damien Lieberherr, Patrick Masson, Anne Morgat, Ivo Pedruzzi, Lucille Pourcel, Sylvain Poux, Catherine Rivoire, Shyamala Sundaram, Alex Bateman, Emily Bowler-Barnett, Hema Bye-A-Jee, Paul Denny, Alexandr Ignatchenko, Rizwan Ishtiaq, Antonia Lock, Yvonne Lussi, Michele Magrane, Maria J. Martin, Sandra Orchard, Pedro Raposo, Elena Speretta, Nidhi Tyagi, Kate Warner, Rossana Zaru, Alexander D. Diehl, Raymond Lee, Juancarlos Chan, Stavros Diamantakis, Daniela Raciti, Magdalena Zarowiecki, Malcolm Fisher, Christina James-Zorn, Virgilio Ponferrada, Aaron Zorn, Sridhar Ramachandran, Leyla Ruzicka, and Monte Westerfield. The gene ontology knowledgebase in 2023. *Genetics*, 224(1):iyad031–iyad031, 2023.

[4] Christian H. Holland, Jovan Tanevski, Javier Perales-Patón, Jan Gleixner, Manu P. Kumar, Elisabetta Mereu, Brian A. Joughin, Oliver Stegle, Douglas A. Lauffenburger, Holger Heyn, Bence Szalai, and Julio Saez-Rodriguez. Robustness and applicability of transcription factor and pathway analysis tools on single-cell RNA-seq data. *Genome Biology*, 21:36, 2020.

[5] Mengzhou Hu, Sahar Alkhairy, Ingoo Lee, Rudolf T. Pillich, Dylan Fong, Kevin Smith, Robin Bachelder, Trey Ideker, and Dexter Pratt. Evaluation of large language models for discovery of gene set function. *Nature Methods*, 2024.

[6] John M. Elizarraras, Yuxing Liao, Zhiao Shi, Qian Zhu, Alexander R. Pico, and Bing Zhang. WebGestalt 2024: faster gene set analysis and new support for metabolomics and multi-omics. *Nucleic Acids Research*, 52(W1):W425–W421, 2024.

[7] Liis Kolberg, Uku Raudvere, Ivan Kuzmin, Priit Adler, Jaak Vilo, and Hedi Peterson. g: Profiler—interoperable web service for functional enrichment analysis and gene identifier mapping (2023 update). *Nucleic Acids Research*, 51(W1):W207–W212, 2023.

[8] Gabriele Sales, Enrica Calura, Paolo Martini, and Chiara Romualdi. Graphite Web: Web tool for gene set analysis exploiting pathway topology. *Nucleic Acids Research*, 41(W1):W89–W97, 2013.

[9] Yingyao Zhou, Bin Zhou, Lars Pache, Max Chang, Alireza Hadj Khodabakhshi, Olga Tanaseichuk, Christopher Benner, and Sumit K Chanda. Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. *Nature Communications*, 10:1523, 2019.

[10] Steven Xijin Ge, Eun Wo Son, and Runan Yao. iDEP: an integrated web application for differential expression and pathway analysis of RNA-Seq data. *BMC Bioinformatics*, 19:1–24, 2018.

[11] Martina Kutmon, Martijn P. van Iersel, Anwesha Bohler, Thomas Kelder, Nuno Nunes, Alexander R. Pico, and Chris T. Evelo. PathVisio 3: an extendable pathway analysis toolbox. *PLOS Computational Biology*, 11(2):e1004085, 2015.

[12] Clara WT Koh, Justin SG Ooi, Eugenia Ziying Ong, and Kuan Rong Chan. STAGEs: A web-based tool that integrates data visualization and pathway enrichment analysis for gene expression studies. *Scientific Reports*, 13(1):7135, 2023.

[13] Saman Farahmand, Corey OĆonnor, Jill A Macoska, and Kourosh Zarringhalam. Causal Inference Engine: a platform for directional gene set enrichment analysis and inference of active transcriptional regulators. *Nucleic Acids Research*, 47(22):11563–11573, 2019.

[14] Weijun Luo, Gaurav Pant, Yeshvant K. Bhavnasi, Steven G. Blanchard Jr, and Cory Brouwer. Pathview Web: user friendly pathway visualization and data integration. *Nucleic Acids Research*, 45(W1):W501–W508, 2017.

[15] Sarah Mubeen, Vinay S. Bharadhwaj, Yojana Gadiya, Martin Hofmann-Apitius, Alpha T. Kodamullil, and Daniel Domingo-Fernández. DecoPath: a web application for decoding pathway enrichment analysis. *NAR Genomics and Bioinformatics*, 3(3):lqab087, 2021.

[16] Haizhou Liu, Mengqin Yuan, Ramkrishna Mitra, Xu Zhou, Min Long, Wanyue Lei, Shunheng Zhou, Yu-e Huang, Fei Hou, Christine M. Eischen, and Wei Jiang. CTpathway: a CrossTalk-based pathway enrichment analysis method for cancer research. *Genome Medicine*, 14:118–118, 2022.

[17] Nico Gerstner, Tim Kehl, Kerstin Lenhof, Anne M´uller, Carolin Mayer, Lea Eckhart, Nadja Liddy Grammes, Caroline Diener, Martin Hart, Oliver Hahn, Jörn Walter, Tony Wyss-Coray, Eckart Meese, Andreas Keller, and Hans-Peter Lenhof. GeneTrail 3: advanced high-throughput enrichment analysis. *Nucleic Acids Research*, 48(W1):W515–W520, 2020.

[18] Po-Jung Huang, Yi-Chen Weng, Kuo-Yang Huang, Chi-Ching Lee, Yuan-Ming Yeh, Yu-Tong Chen, Cheng-Hsun Chiu, and Petrus Tang. ProFun: A web server for functional enrichment analysis of parasitic protozoan genes. *Journal of Microbiology, Immunology and Infection*, 57(3):509–517, 2024.

[19] Jianguo Xia, Erin E. Gill, and Robert EW Hancock. NetworkAnalyst for statistical, visual and network-based meta-analysis of gene expression data. *Nature Protocols*, 10(6):823–844, 2015.

[20] Sebastian Canzler, Jana Schor, Wibke Busch, Kristin Schubert, Ulrike E Rolle-Kampczyk, Hervé Seitz, Hennicke Kamp, Martin von Bergen, Roland Buesen, and Jörg Hackermüller. Prospects and challenges of multi-omics data integration in toxicology. *Archives of Toxicology*, 94:371–388, 2020.

[21] Lv Jin, Xiao-Yu Zuo, Wei-Yang Su, Xiao-Lei Zhao, Man-Qiong Yuan, Li-Zhen Han, Xiang Zhao, Ye-Da Chen, and Shao-Qi Rao. Pathway-based analysis tools for complex diseases: a review. *Genomics, Proteomics and Bioinformatics*, 12(5):210–220, 2014.

[22] Kathleen Gregory, Paul Groth, Helena Cousijn, Andrea Scharnhorst, and Sally Wyatt. Searching data: a review of observational data retrieval practices in selected disciplines. *Journal of the Association for Information Science and Technology*, 70(5):419–432, 2019.

[23] Sean Kandel, Jeffrey Heer, Catherine Plaisant, Jessie Kennedy, Frank Van Ham, Nathalie Henry Riche, Chris Weaver, Bongshin Lee, Dominique Brodbeck, and Paolo Buono. Research directions in data wrangling: Visualizations and transformations for usable and credible data. *Information Visualization*, 10(4):271–288, 2011.

[24] Zhizheng Wang, Qiao Jin, Chih-Hsuan Wei, Shubo Tian, Po-Ting Lai, Qingqing Zhu, Chi-Ping Day, Christina Ross, Robert Leaman, and Zhiyong Lu. GeneAgent: self-verification language agent for gene-set analysis using domain databases. *Nature Methods*, 22:1677–1685, 2025.

[25] Yubo Wang, Xueguang Ma, and Wenhu Chen. Augmenting Black-box LLMs with Medical Textbooks for Biomedical Question Answering. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 1754–1770, Miami, Florida, USA, 2024. Association for Computational Linguistics.

[26] Minbyul Jeong, Jiwoong Sohn, Mujeen Sung, and Jaewoo Kang. Improving medical reasoning through retrieval and self-reflection with retrieval-augmented large language models. *Bioinformatics*, 40(Supplement_1):i119–i129, 2024.

[27] Raffaella Nativio, Yemin Lan, Greg Donahue, Simone Sidoli, Amit Berson, Ananth R. Srinivasan, Oksana Shcherbakova, Alexandre Amlie-Wolf, Ji Nie, Xiaolong Cui, Chuan He, Li-San Wang, Benjamin A. Garcia, John Q. Trojanowski, Nancy M. Bonini, and Shelley L. Berger. An integrated multi-omics approach identifies epigenetic alterations associated with Alzheimer's disease. *Nature Genetics*, 52(10):1024–1035, 2020.

[28] Arthur A Bergen, Sovann Kaing, Jacoline B Ten Brink, Netherlands Brain Bank, Theo G Gorgels, and Sarah F Janssen. Gene expression and functional annotation of human choroid plexus epithelium failure in Alzheimer' s disease. *BMC Genomics*, 16:1–15, 2015.

[29] Winnie S. Liang, Travis Dunckley, Thomas G. Beach, Andrew Grover, Diego Mastroeni, Douglas G. Walker, Richard J. Caselli, Walter A. Kukull, Daniel McKeel, John C. Morris, Christine Hulette, Donald Schmechel, Gene E. Alexander, Eric M. Reiman, Joseph Rogers, and Dietrich A. Stephan. Gene expression profiles in anatomically and functionally distinct regions of the normal aged human brain. *Physiological Genomics*, 28(3):311–322, 2007.

[30] Russell H. Swerdlow. Brain aging, Alzheimer's disease, and mitochondria. *Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease*, 1812(12):1630–1639, 2011.

[31] Aleksandra Maruszak and Cezary Żekanowski. Mitochondrial dysfunction and Alzheimer's disease. *Progress in Neuro-Psychopharmacology and Biological Psychiatry*, 35(2):320–330, 2011.

[32] Xiongwei Zhu, George Perry, Mark A Smith, and Xinglong Wang. Abnormal mitochondrial dynamics in the pathogenesis of Alzheimer's disease. *Journal of Alzheimer's Disease*, 33(S1):S253–S262, 2013.

[33] Henry W Querfurth and Frank M LaFerla. Mechanisms of disease. *New England Journal of Medicine*, 362(4):329–344, 2010.

[34] Jani Huuhtanen, Shady Adnan-Awad, Jason Theodoropoulos, Sofia Forstén, Rebecca Warfvinge, Olli Dufva, Jonas Bouhlal, Parashar Dhapola, Hanna Duàn, Essi Laajala, Tiina Kasanen, Jay Klievink, Mette Ilander, Taina Jaatinen, Ulla Olsson-Strömberg, Henrik Hjorth-Hansen, Andreas Burchert, Göran Karlsson, Anna Kreutzman, Harri Lähdesmäki, and Satu Mustjoki. Single-cell analysis of immune recognition in chronic myeloid leukemia patients following tyrosine kinase inhibitor discontinuation. *Leukemia*, 38:109–125, 2024.

[35] Jake Y. Chen, Ragini Pandey, and Thanh M. Nguyen. HAPPI-2: a comprehensive and high-quality map of human annotated and predicted protein interactions. *BMC Genomics*, 18:182–182, 2017.

[36] Quang-Huy Nguyen and Duc-Hau Le. oCEM: Automatic detection and analysis of overlapping co-expressed gene modules. *BMC Genomics*, 23:39, 2022.

[37] Matthew E. Ritchie, Belinda Phipson, Di Wu, Yifang Hu, Charity W. Law, Wei Shi, and Gordon K. Smyth. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, 43(7):e47–e47, 2015.

[38] Gennady Korotkevich, Vladimir Sukhov, Nikolay Budin, Boris Shpak, Maxim N. Artyomov, and Alexey Sergushichev. Fast gene set enrichment analysis. *BioRxiv*, page 060012, 2021.

[39] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300, 1995.

[40] Michael C Whitlock. Combining probability from independent tests: the weighted z-method is superior to fisher's approach. *Journal of evolutionary biology*, 18(5):1368–1373, 2005.

[41] Samuel A Stouffer, Edward A Suchman, Leland C DeVinney, Shirley A Star, and Robin M Williams Jr. The american soldier: Adjustment during army life. In Samuel A. Stouffer, Carl I. Hovland, Arthur A. Lumsdaine, and Fred D. Sheffield, editors, *Studies in social psychology in World War II*, volume 1. Princeton: Princeton University Press, 1949.

[42] John Flowerdew. Some thoughts on English for research publication purposes (ERPP) and related issues. *Language Teaching*, 48(2):250–262, 2015.

[43] Dong Wan Cho. Science journal paper writing in an EFL context: The case of Korea. *English for Specific Purposes*, 28(4):230–239, 2009.

[44] John Flowerdew. Problems in writing for scholarly publication in English: The case of Hong Kong. *Journal of Second Language Writing*, 8(3):243–264, 1999.

[45] Akiko Okamura. Two types of strategies used by Japanese scientists, when writing research articles in English. *System*, 34(1):68–79, 2006.

[46] Ann M. Rogerson and Grace McCarthy. Using Internet based paraphrasing tools: Original work, patchwriting or facilitated plagiarism? *International Journal for Educational Integrity*, 13:1–15, 2017.

[47] Rudolf Jaenisch. Paper trail: Inside the stem cell wars. *New Scientist*, 9, 2010.

[48] Nitish Shirish Keskar, Bryan McCann, Lav R Varshney, Caiming Xiong, and Richard Socher. Ctrl: A conditional transformer language model for controllable generation. *arXiv preprint*, 2019.

[49] Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. *arXiv preprint*, 2019.

[50] Yaru Zhang, Yunlong Ma, Yukuan Huang, Yan Zhang, Qi Jiang, Meng Zhou, and Jianzhong Su. Benchmarking algorithms for pathway activity transformation of single-cell RNA-seq data. *Computational and Structural Biotechnology Journal*, 18:2953–2961, 2020.

[51] M. Kathleen Kerr, Mitchell Martin, and Gary A. Churchill. Analysis of Variance for Gene Expression Microarray Data. *Journal of Computational Biology*, 7(6):819–837, 2000.

[52] Michael I. Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15:550, 2014.

[53] William Sealy Gosset. The Probable Error of a Mean. *Biometrika*, 6:1–25, 1908.

[54] Mark D. Robinson, Davis J. McCarthy, and Gordon K. Smyth. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140, 2010.

[55] Greg Finak, Andrew McDavid, Masanao Yajima, Jingyuan Deng, Vivian Gersuk, Alex K. Shalek, Chloe K. Slichter, Hannah W. Miller, M. Juliana McElrath, Martin Prlic, Peter S. Linsley, and Raphael Gottardo. MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biology*, 16:278, 2015.

[56] Doris Damian and Malka Gorfine. Statistical concerns about the GSEA procedure. *Nature Genetics*, 36(7):663; author reply 663, July 2004.

[57] Raivo Kolde, Sven Laur, Priit Adler, and Jaak Vilo. Robust rank aggregation for gene list integration and meta-analysis. *Bioinformatics*, 28(4):573–580, 2012.

[58] Ronald Aylmer Fisher. On the interpretation of χ^2 from contingency tables, and the calculation of p. *Journal of the Royal Statistical Society*, 85(1):87–94, 1922.

[59] Leonard H. C. Tippett. *The Methods of Statistics. An Introduction mainly for Workers in the Biological Sciences*. Williams & Norgate, London, 1931.

[60] Tin Nguyen and Sorin Draghici. *BLMA: A package for bi-level meta-analysis*. Bioconductor, 2017. R package.

[61] Robert R Corbeil and Shayle R Searle. Restricted maximum likelihood (REML) estimation of variance components in the mixed model. *Technometrics*, 18(1):31–38, 1976.

[62] Zongliang Yue, Madhura M. Kshirsagar, Thanh Nguyen, Chayaporn Suphavilai, Michael T. Neylon, Liugen Zhu, Timothy Ratliff, and Jake Y. Chen. PAGER: constructing PAGs and new PAG–PAG relationships for network biology. *Bioinformatics*, 31(12):i250–i257, 2015.

[63] Tanya Barrett, Stephen E. Wilhite, Pierre Ledoux, Carlos Evangelista, Irene F. Kim, Maxim Tomashevsky, Kimberly A. Marshall, Katherine H. Phillippy, Patti M. Sherman, Michelle Holko, Andrey Yefanov, Hyeseung Lee, Naigong Zhang, Cynthia L. Robertson, Nadezhda Serova, Sean Davis, and Alexandra Soboleva. NCBI GEO: archive for functional genomics data sets–update. *Nucleic Acids Research*, 41(D1):D991–D995, 2013.

Authors and Affiliations

Department of Computer Science and Software Engineering, Auburn University, Auburn, AL 36849, United States
Quang-Huy Nguyen
School of Information and Communications Technology, Hanoi University of Science and Technology, Hanoi, Vietnam
Quang-Huy Nguyen, Duc-Hau Le
Department of Biomedical Informatics and Data Science, University of Alabama at Birmingham, Birmingham, AL 35233, United States
Jake Y. Chen
College of Forestry and Wildlife Sciences, Auburn University, Auburn, AL 36849, United States
Hao Chen
Department of Health Outcomes Research and Policy, Auburn University, Auburn, AL 36849, United States
Quang-Huy Nguyen, Zeru Zhang, Zongliang Yue

Corresponding Authors

Correspondence to Duc-Hau Le (hauld@soict.hust.edu.vn), Jake Y. Chen (jakechen@uab.edu), Wei-Shinn Ku (wzk0004@auburn.edu), Hao Chen (hzc0134@auburn.edu), Zongliang Yue (zzy0065@auburn.edu).

Author contributions

Q-.H.N and D-.H.L designed the study. Q-.H.N conceived the idea, drafted the manuscript, conducted prompt engineering, developed the web application system (frontend and backend), and implemented PAG-Agent pipeline. Q-.H.N and Z.Z. deployed the web interface, Q-.H.N. and D-.H.L evaluated the analysis and conducted the scientific review of the LLM outputs. D-.H.L edited the writing and provided feedback on the web app's UI to improve clarity. D-.H.L, J.Y.C, H.C., Z.Y. co-supervised all aspects of the work. All authors approved the final version of this manuscript.

Competing interests

The authors declare no competing interests.

Funding

This research was funded by National Academies of Sciences, Engineering, and Medicine, grant number SCON-10001538 to Z.Y..