IAN: **I**ntegrated Enrichment **An**alysis Using A Multi-Agent AI System

**Background**:

High throughput “Omics” experiments, aimed at understanding the biological systems, help researchers identify thousands of differentially active molecules between experimental conditions. To understand these large number of differentially active molecules and their relationship to the phenotype being investigated, researchers routinely perform enrichment analysis. Though natural language processing (NLP) based methods have been put forward to integrate enrichment analysis, based on the shared gene lists, they lack reasoning and intellectual integration with other relevant contextual data, to unravel hidden workings of the system. This has resulted in a treasure trove of high-quality data inviting to be mined and interpreted. Large Language Models perform extremely well in summarizing general text but have been struggling to impress scientists with their abstract writing skills. In this work we have developed IAN – an integrated enrichment analysis platform using a multi-agent artificial intelligence (AI) system, to strategically augment and guide Large Language Models “Omics” summarization abilities towards producing superior insightful interpretations resulting in novel discoveries.

**Methods**:

IAN is implemented as an R package, that is easy to setup and intuitive to use. IAN generates enrichment results for multiple “Omics” input sources (Seurat, DESeq2, custom DEG’s), using popular datasets for pathway analysis (KEGG pathways, WikiPathways and Reactome), along with transcription factor target datasets (ChEA) and Gene Ontology (BP) enrichments. Individual summaries of these enrichment results are generated using the Google’s popular LLM – Gemini, through a multi-agent architecture. The summaries are then contextually integrated and interpreted by the LLM, considering relevant protein-protein interaction data from the STRING database, using appropriately engineered prompts, to answer highly complex questions. The insightful results, along with potential explanations for the system under study and any novel observations, are provided as a HTML report that is easy to navigate and supported by the experimentally derived concepts.

**Results**:

Technical evaluations of our case studies show great potential for IAN to help researchers explore and decipher biological systems, using the vast information space, leveraging the understanding abilities of the LLMs. IAN is available at the projects GitHub page at: https://github.com/NIH-NEI/IANai , along with installation instructions and detailed example tutorials.

Elucidate or interpret molecular mechanisms

1. DESeq2 output, Seurat (findmarkers), Custom list of DEG, chipseq (proximal genes), genes of interest - proteomics,

2. Organism choice

3. Experimental Design

4. Up list, Down list, One list

5. Enrichr, ClusterProfiler, DAVID, GSEA, CTD (custom enrichment results)

6. Enrichr drugs, pathway summary

[29688259](https://pubmed.ncbi.nlm.nih.gov/29688259/) <https://pmc.ncbi.nlm.nih.gov/articles/PMC6137994/>

<https://github.com/isglobal-brge/CTDquerier/tree/master/R>

6. IDs to Gene names

7. Enrichment results

8. Integrate

9. Explore

- key genes

- key pathways

- mechanistic understanding

- potential markers – why

- potential targets – why

- Transcriptional factors – regulatory mechanisms/networks

- microRNA, lncRNAs other noncoding rnas role, involvement

- potential downstream experiments to validate

- potential for diagnostics - how

- potential repurposable drugs

- tissue specificity

- kinase involvement

- ligand involvement

- histone modifications

- novel targets, markers – not represented in the pathways  
10. Network

11. Report with reasoning

- Top 10 of each enrichment – table, bar graph

- Save full results, filtered results, bar graph

- Summarize, concepts, genes

- Regulatory network – explain

- Explore key genes, potential markers, targets – how explain

- Repurposing known compounds – how explain

- Network file, graph

Integrated Enrichment Analysis, Reasoning, Exploration

1. **Initial Setup & Data:** Provide experimental context, DEG list, pathway/GO enrichment results, ChEA data, and STRING data.
2. **Pathway Analysis:** Identify enriched DEGs for each pathway and generate pathway-DEG network.
3. **Hub Gene Identification:** Identify hub genes based on pathway/GO enrichment and STRING connectivity, and generate hub gene-GO term/pathway network.
4. **Transcription Factor Analysis:** Identify key TFs, their target DEGs, and generate a TF-DEG regulatory network, adding to the system representation network.
5. **Non-coding RNA Analysis:** Identify non-coding RNAs, their target DEGs, and add non-coding RNA-DEG interactions to both the regulatory and system representation networks.
6. **Synthesis & Model:** Summarize key findings, propose a mechanistic model, and identify potential markers/targets.
7. **Hypothesis Generation:** Suggest experiments to validate the model, identify diagnostics, novel targets, and repurposable drugs.

Review, Evaluate, Guide, Support reasoning in LLMs

Zero-shot prompting – no example output

One-shot prompting – one example output

Few-shot prompting – more than one example output

Active prompting – iterative human intervention, feedback

CoT (Chain-of-thoughts) – Intermediate steps in the example

One-shot CoT, Few-shot CoT

Auto-CoT

Least-to-most prompting

ToT (Tree-of-thoughts)

IoT (Iterations-of-thoughts)

CQoT (Critical-Questions-of-Thoughts) <https://www.alphaxiv.org/abs/2412.15177>

RoT (Reviewer-of-Thoughts)

Internal Review Agent (IRA)

produce a series of intermediate steps that mimic the logical progression of cognitive thinking

CQoT

What is the claim here ?

What limits or conditions exist on the claim ?

What fact, data, evidence or premises support the claim ?

What connection, justification bridges the data and the claim ?

What counter arguments or exceptions exist to the claim ?

What additional support or justification exist for the claim ?

Mark as pass/fail

Provide prompt as premise and conclusions

Access the validity of the responses using CQoT

- only human or mouse organism are supported

- only custom gene list, deseq2 and seurat findmarker are supported

- only entrezid, ensembl or symbol are supported

- only kegg, reactome, wikipathway, chea2022, stringdb and go are used

- only mouse or human specific chea data is used from results

- only google gemini is supported

- transparent – original outputs, prepared output, prompts, response, parameters used – parameter tracking

- system retries 3 times for agent failures

- pathway comparison provided

-network modules – generate integrated network, analyze integrated network to putforward system model

- genes not represented in pathways – identified and used for novel insight generation

- default response report saving for all agents

- default text file saving for all enrichment analysis, original and filtered

Todo

- add all gemini parameter as user input

- add html report

-retry if the llm failed to respond

* -coauthor thoughts on
  + Layout of report
  + Any other presentation thoughts
  + Install test
  + Tutorial test
  + Personal data test

FAQs

* check for 429 http return codes and perform retry with exponential backoff, the normal congestion control approach.
* Above if error running multi agent

1. Make sure all the genes seen, discussed, reported are grounded in the provided deg.
2. Make sure all the terms discussed, reported are grounded in the provided data.
3. Analyze the pathway results in the context of the provided chea results. Examine if the transcription factor in chea list is also found in our pathway results or the transcription factor targets are found in our pathway results.
4. Analyze the pathway results in the context of the provided string interaction data.
5. Highlight any new context specific interactions identified
6. If we get this error “Request failed with status code: 503, message: The model is overloaded. Please try again later.”, use serial code

Data:

Custom DEG from rnaseq: Uveitis - [33503442](https://pubmed.ncbi.nlm.nih.gov/33503442/)

**Identifying RNA Biomarkers and Molecular Pathways Involved in Multiple Subtypes of Uveitis** [**https://pubmed.ncbi.nlm.nih.gov/33503442/**](https://pubmed.ncbi.nlm.nih.gov/33503442/)

**DESeq2:** Processed DEG using deseq2

[**https://pubmed.ncbi.nlm.nih.gov/35727985/**](https://pubmed.ncbi.nlm.nih.gov/35727985/)

**Single-cell analyses highlight the proinflammatory contribution of C1q-high monocytes to Behçet's disease**