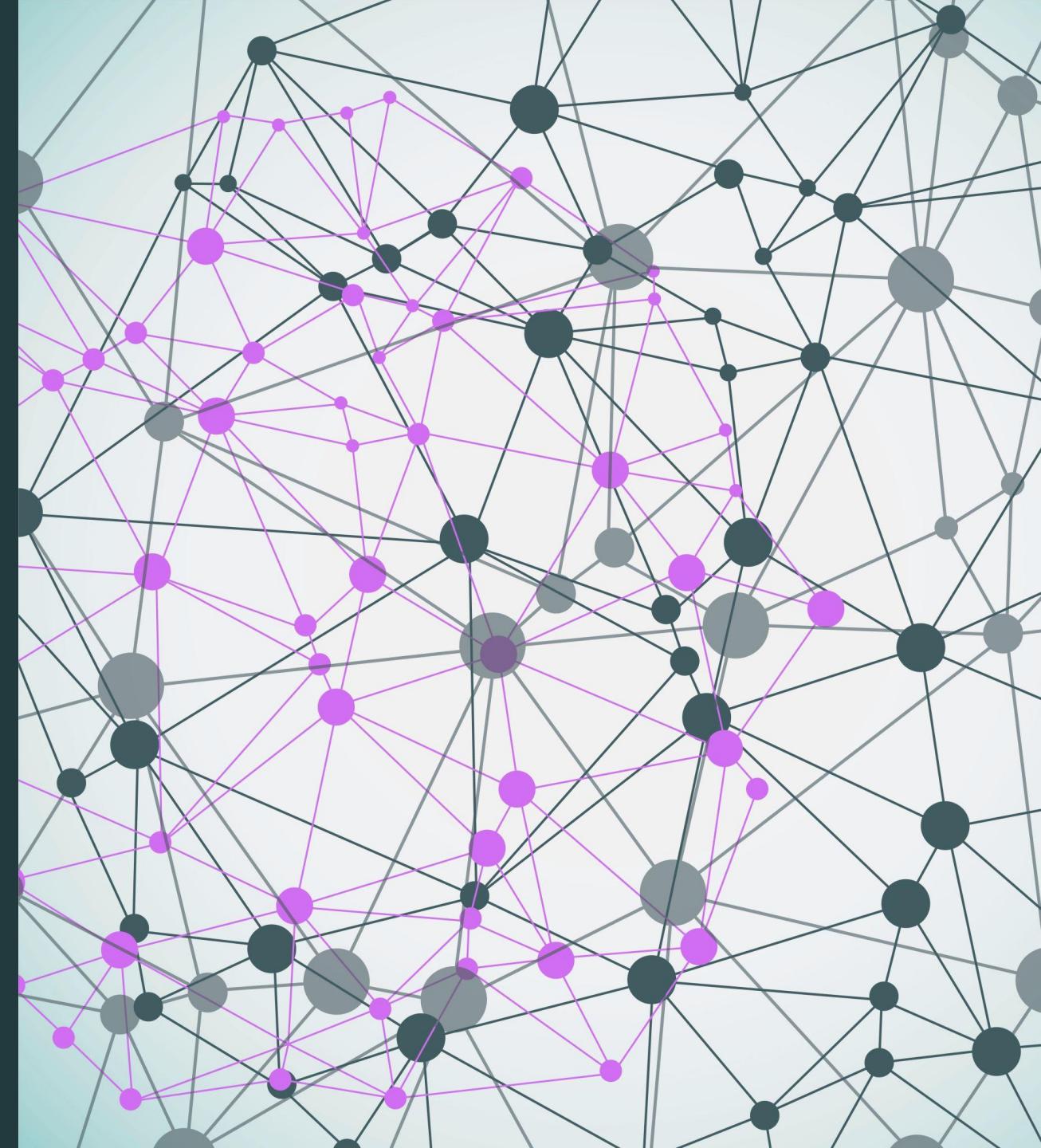


GENEVIC

GENetic data Exploration and
Visualization via Intelligent
interactive Console

Author & Presenter: Anindita Nath
Bioinformatics and System Medicine
Laboratory
SBMI, UTHealth Houston
Wednesday, March 20, 2024



Motivation

- ✓ surge in biomedical knowledge and tools
- ✓ knowledge discovery a challenge, requires deep domain and data science expertise
- ✓ Generative AI (Chat GPT, GPT-4) ‘hallucination effect’, unreliable in specialized fields.

GENEVIC: key features

- ✓ enable dynamic engagement with the PolyGenic Score(PGS) Catalog
 - enhancing data mining and discerning intricate patterns
 - broadens access to domain expertise
- ✓ integrates bioinformatics APIs, STRING and ENRICHr:
 - enriching polygenic score analysis with specialized knowledge
 - identification of genetic markers and pathways.
- ✓ supports literature searches from major sites

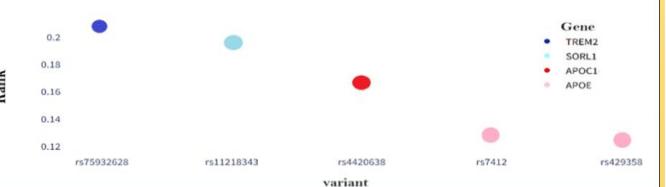
- ✓ facilitates detailed analysis and informed decision-making
- ✓ democratizes access to genetic research insights
- ✓ simplifying data for diverse research backgrounds
- ✓ promoting multidisciplinary collaboration

Overview and Functionality

(i) Prompts to backend for processing

Backend	Tech Stack
PGS RANK DATABASE <ul style="list-style-type: none"> Curated Database Derivative of Polygenic Score Catalog (web DB) Ranked variants for each trait 	Prompt based Learning Template 
RDBMS database structure <ul style="list-style-type: none"> SQLite used for development 	Auto-translate natural language prompt to SQL Code or Python code as required
	Display results from the custom database in a table or chart format as desired by user

(ii) Output generated by backend displayed to user on the interface

User Interface	GENETIC Output																		
	<p>prompt I: Plot variant and genes against ranks for the top 5 ranked genes, ranked high to low, in Alzheimer. If duplicate, show only once.</p> <p>output I:</p> <pre># Query the variants and their ranks for the top 5 ranked genes in Alzheimer's disease sql_query = "SELECT DISTINCT gene, rsID, rank FROM pgs_snpmeta WHERE trait = 'Alzheimer' ORDER BY rank DESC LIMIT 5" variants_df = execute_sql(sql_query) # Replace NAN with 0 variants_df = variants_df.replace(np.nan, 0) # Observe the query result observe("variants_df", variants_df)</pre> <table border="1"> <thead> <tr> <th>rsID</th> <th>gene</th> <th>rank</th> </tr> </thead> <tbody> <tr> <td>rs75932628</td> <td>TREM2</td> <td>0.2083</td> </tr> <tr> <td>rs11218343</td> <td>SORL1</td> <td>0.197</td> </tr> <tr> <td>rs4420638</td> <td>APOC1</td> <td>0.1667</td> </tr> <tr> <td>rs7412</td> <td>APOE</td> <td>0.1285</td> </tr> <tr> <td>rs429358</td> <td>APOE</td> <td>0.125</td> </tr> </tbody> </table> <p>Plot the variants against the ranks</p> <pre>fig = px.scatter(snpmeta_df, x='rsID', y='rank', color='gene', title='Variants vs Ranks for Top 5 ranked Genes for Alzheimer') fig.update_layout(xaxis_title='variant', yaxis_title='Rank') # Show the plot to the user show(fig)</pre> 	rsID	gene	rank	rs75932628	TREM2	0.2083	rs11218343	SORL1	0.197	rs4420638	APOC1	0.1667	rs7412	APOE	0.1285	rs429358	APOE	0.125
rsID	gene	rank																	
rs75932628	TREM2	0.2083																	
rs11218343	SORL1	0.197																	
rs4420638	APOC1	0.1667																	
rs7412	APOE	0.1285																	
rs429358	APOE	0.125																	

(iii) Output generated by backend displayed to user on the interface

Backend	Tech Stack	Literature Search
	OpenAI	LangChain
	automate call to respective APIs	PubMed
	 STRING protein interaction network graph	 arXiv gene set enrichment analysis
	prompt II(a): protein interaction	prompt III(a): Search for articles with APOE and Alzheimer's in PubMed.
	prompt II(b): enrichment analysis	prompt III(b): https://pubmed.ncbi.nlm.nih.gov/30140
	output II(a): Query me: (one human protein or multiple space separated proteins)	output III(a): Here are some articles related to APOE and Alzheimer's disease:
	<div style="border: 1px solid black; padding: 5px; display: inline-block;">TREM2 SORL1 APOC1 APOE</div>	<ol style="list-style-type: none"> [APOE and Alzheimer's disease: advances in genetics, pathophysiology, and therapeutic approaches] (https://pubmed.ncbi.nlm.nih.gov/33340485/) [ApoE in Alzheimer's disease: pathophysiology and therapeutic strategies] (https://pubmed.ncbi.nlm.nih.gov/36348357/) [Interplay between innate immunity and Alzheimer disease: APOE and TREM2 in the spotlight] (https://pubmed.ncbi.nlm.nih.gov/30140051/)
	output II(b):	output III(b): The article titled 'Interplay between innate immunity and Alzheimer disease: APOE and TREM2 in the spotlight' discusses the role of APOE and TREM2 in the pathogenesis of Alzheimer's disease. It highlights the immunomodulatory function of APOE and its connection to triggering receptor expressed on myeloid cells 2 (TREM2), which is expressed by microglia in the central nervous system. The article provides insights into the interplay between neuroinflammation and Alzheimer's disease. You can find the article [here] (https://pubmed.ncbi.nlm.nih.gov/30140051/).

Custom Database: PolyGenicScore(PGS) Rank Database

- ✓ PGS score files from PGS Catalog (Release: August 4th, 2023), human reference genome (GRCh38 build), local database for easier access.
- ✓ Trait specific data using R's "Quincunx" package, querying the catalog's REST API with ontology IDs
- ✓ essential columns: effect allele, effect weight, and SNP rsID/SNP coordinates, and harmonized variant labeling discrepancies.
- ✓ merged phenotype specific PGS files into a single dataset using R's "dplyr" and "bind_rows", missing values as 'NA'
- ✓ ranked variants using the *Dowdall* method, an alternate Borda method, based on their absolute effect weights and annotated the variants using ANNOVAR:
 - assigning the reciprocal of ranks (RR) to each variant
 - top rank -> RR is 1, 2nd rank -> RR is 1/2, 3rd rank -> RR is 1/3, so on.
 - For each variant query (Q), calculate the mean of the reciprocals of its ranks (MRR) across multiple scores. weights unavailable -> MRR of 0.

PolyGenicScore(PGS) Rank Database (contd.)

Phenotype	#ScoreFiles	SNPInfo
Alzheimer	23	w:[(-0.95)-1.64]; m:0.0013; md: 5.82e-07; sd: 0.0390
schizophrenia	5	w:[(-0.04)-0.05]; m:5.61e-05; md:8.36e-06; sd:0.0018
cognition	5	w:[(-3.60)-1.68]; m:(-0.0005); md:-7e-08;sd:0.0321

Table 1: Summary of statistics of the phenotypes used to develop the PGS Rank database.

Phenotype denotes disease or trait name, #ScoreFiles denote the count of score files for each PGS Score ID corresponding to the phenotype, SNP info refers to the aggregated information regarding the SNPs or variants in each score file. w is the weight range, m, md, sd denote the mean, median, and standard deviation of weights, respectively.

Availability

- ✓ Streamlit community cloud at <https://genevic-anath2024.streamlit.app>, ready to use without installation.
 - Streamlit's 1 GB data limit, PGS Rank DB is pared down to top 100 genes each for Alzheimer's, schizophrenia, and cognition
 - adequate for this pilot project but should be expanded later with architecture of higher capacity
- ✓ runnable locally, as detailed in the installation guide,
<http://tinyurl.com/InstalaltionGuide>
 - prerequisite being Python 3.10 or higher version

Home Page



Home

PGSChat

GeneAPIChat

LiteratureSearch

Select a task from above 

Click Settings  for Open AI credentials

Settings

Azure OpenAI Credentials

ChatGPT deployment name:

gpt-35-turbo

GPT-4 deployment name (if not specified, default to ChatGPT's):

gpt-35-turbo

Azure OpenAI Endpoint:

<https://resourcenamehere.openai.azure.com/>

Azure OpenAI Key:

Click on the respective functionalities to go to the respective page:
PGSChat, GeneAPIChat and Literature Search

- ✓ Users first require an active subscription with Azure OpenAI (refer <http://tinyurl.com/AzureOpenAIInstructions>),
- ✓ Deploy a model incorporating the ChatGPT engine,
- ✓ Enter their account details in the 'Settings' section of the application's interface.

Home

PGSChat

GeneAPIChat

LiteratureSearch

Select a task from above 🤖

Click Settings 🤖 for Open AI credentials

Settings

Azure OpenAI Credentials

ChatGPT deployment name:
gpt-35-turbo

GPT-4 deployment name (if not specified, default to ChatGPT's):
gpt-35-turbo

Azure OpenAI Endpoint:
[https://resourcenamehere.openai..](https://resourcenamehere.openai.com)

Azure OpenAI Key:

Resources

[GENEVIC Architecture](#)

[Tutorial](#)

[Install and Run GENEVIC](#)

[Azure OpenAI Instructions](#)

[PGS Catalog Website](#)

[PGS Rank Database](#)

[Test Cases](#)

[Code Repository](#)

UTHouston
McWilliams School of Biomedical Informatics

Meet the Team

[Bioinfo & Systems Medicine Lab](#)

Organization: MSBMI, UTHH

Contact Us



GENEVIC: GENetic data Exploration and Visualization Intelligent interactive Console

Ask where you want to go:

Click on heading below 🤖 to go the respective task page!



GeneAPI Chat

Explore external Bioinformatics websites via automated web API calls
Demo APIs explored: STRING and ENRICHR
Generate gene-gene interaction network, one or more gene names as input
Entire functionality of STRING API replicated as is
Interactive in-app display of the network
Perform gene enrichment analysis with reference gene set libraries, given gene list as input
Visualize the network graph
Download the enrichment results as CSV and/or the visualizations in known image formats

Auto-changing block:
Clicking on the header will take you to respective functionality section

User prompts: simple relevant keywords or question to navigate to the respective functionality interface: **Few-shot prompt classification**



Resources:

Supplementary materials,
[https://github.com/anath2110/
GENEVIC_Supplementary.git](https://github.com/anath2110/GENEVIC_Supplementary.git)

Video Demo to aid in navigating the 3 GENEVIC functionality interfaces

Contact research and development team:

- Lab link
- School Website
- Feedback Form

PGS Chat



Ask me a question

Show top 5 ranked variants, top to bottom, along with their other information in Alzheimer. If duplicate, show once only.

Submit

User prompt: intuitive, simple, natural language

Results from the PGS Rank database:
Top 5 variants and corresponding genes for Alzheimer ranked top to bottom

SQL Code

```
SELECT DISTINCT
    snpcoord,
    gene,
    func,
    rank,
    trait
FROM
    pgs_snpmeta
WHERE
    trait = 'Alzheimer'
ORDER BY
    rank DESC
LIMIT 5;
```

Natural Language queries/prompt auto-translated to SQL code

	snpcoord	gene	func	rank	trait
0	6_41161514	TREM2	exonic	0.2083	Alzheimer
1	11_121564878	SORL1	intronic	0.197	Alzheimer
2	19_44919689	APOC1(dist=345)	downstream	0.1667	Alzheimer
3	19_44908822	APOE	exonic	0.1285	Alzheimer
4	19_44908684	APOE	exonic	0.125	Alzheimer

Download CSV

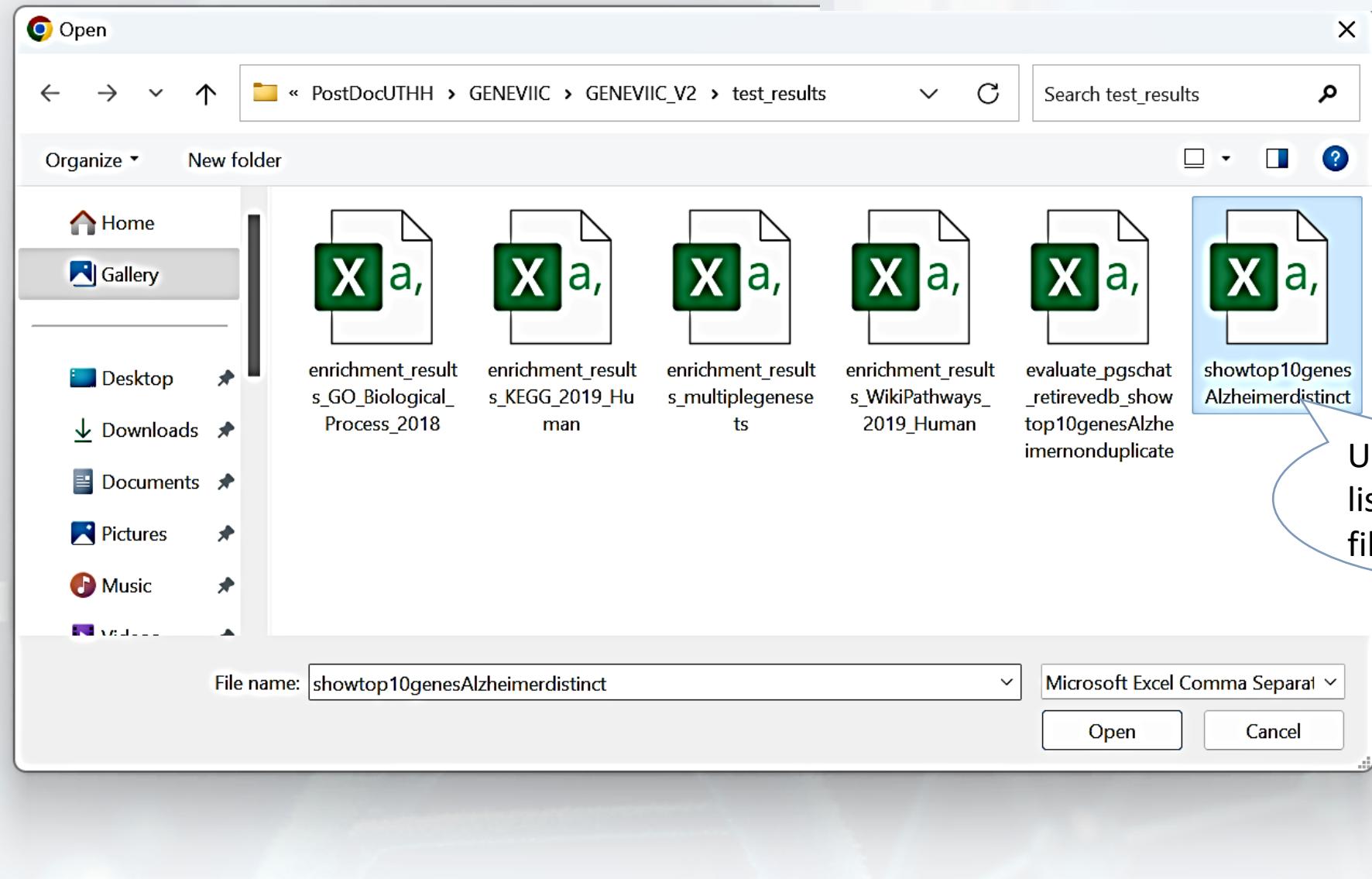
Download and save the top 10 genes information as csv file in local client location

Query ChatGPT directly

Ask me a question

Functional annotation of the genes

User prompt:
*Functional
annotation
of the genes*



Functional annotation is the process of assigning biological functions to genes based on experimental evidence or computational predictions. Here is a brief functional annotation of the genes you mentioned:

1. TREM2 (Triggering Receptor Expressed on Myeloid Cells 2): It is involved in the regulation of immune responses and phagocytosis. Mutations in TREM2 have been associated with increased risk of Alzheimer's disease.
2. SORL1 (Sortilin-Related Receptor 1): It plays a role in the trafficking and processing of amyloid precursor protein (APP), which is involved in the production of amyloid-beta plaques in Alzheimer's disease.
3. APOC1 (Apolipoprotein C1): It is a component of lipoprotein particles and plays a role in lipid metabolism. Variations in APOC1 have been associated with increased risk of Alzheimer's disease.
4. APOE (Apolipoprotein E): It is involved in lipid transport and metabolism. APOE has been extensively studied in relation to Alzheimer's disease, as certain variants of this gene are associated with increased risk.
5. BIN1 (Bridging Integrator 1): It is involved in membrane remodeling and endocytosis. BIN1 has been implicated in Alzheimer's disease, as it interacts with tau protein and affects its aggregation.
6. CYP27C1 (Cytochrome P450 Family 27 Subfamily C Member 1): It encodes an enzyme involved in the metabolism of vitamin D. Its specific role in relation to Alzheimer's disease is not well understood.
7. ABI3 (Abl Interactor 3): It is involved in signal transduction and regulation of cell adhesion. ABI3 has been associated with immune-related disorders, but its role in Alzheimer's disease is not well characterized.
8. DSG2 (Desmoglein 2): It is a component of desmosomes, which are cell adhesion structures. DSG2 mutations have been associated with arrhythmogenic right ventricular cardiomyopathy, but its role in Alzheimer's disease is unclear.

Please note that the functional annotations provided here are a general overview and may not encompass all known functions or associations of these genes. Further research and experimental evidence are required to fully understand the functions and roles of these genes in specific biological processes or diseases.

Chat
GPT/GPT-4
provides
the answer
to the
query
using its
vast pre-
trained
knowledge

Ask me a question

Plot variants corresponding to their ranks for 5 top ranked genes, ranked from top to bottom in Alzheimer. If duplicate, show once.

Press Ctrl+Enter to apply

Submit

Visualize custom database

Question: Plot variants against ranks for the top 5 ranked genes, ranked high to low, in Alzheimer. use pgssnpmeta only. No duplicates

Thought 1

To plot variants against ranks for the top 5 ranked genes in Alzheimer, we need to retrieve the necessary data from the pgssnpmeta table. We will filter the data based on the reported_trait column to select only the variants associated with Alzheimer. Then, we will sort the data based on the rank column in descending order to get the top-ranked genes. Finally, we will plot the variants against ranks for the top 5 genes.

Action:

```
# Import necessary libraries
import plotly.express as px

# Query the necessary data
sql_query = "SELECT * FROM pgssnpmeta WHERE trait = 'Alzheimer' AND gene != '' GROUP BY rsID ORDER BY rank DESC LIMIT 5"
snpmeta_df = execute_sql(sql_query)

# Replace NAN with 0
snpmeta_df = snpmeta_df.replace(np.nan, 0)

# Observe the query result
observe("pgssnpmeta data for Alzheimer", snpmeta_df)
```

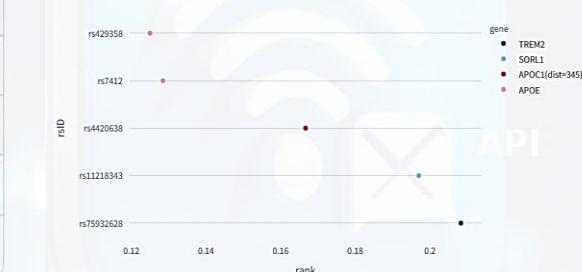
pgssnpmeta data for Alzheimer

	PGS_Score_ID	rsID	gene	func	rank	snpcoord	trait
0	PGS000898	rs75932628	TREM2	exonic	0.2083	6_41161514	Alzheimer
1	PGS002249	rs11218343	SORL1	intronic	0.197	11_121564878	Alzheimer
2	PGS001348	rs4420638	APOC1(dist=345)	downstream	0.1667	19_44919689	Alzheimer
3	PGS000334	rs7412	APOE	exonic	0.1285	19_44908822	Alzheimer
4	PGS003574	rs429358	APOE	exonic	0.125	19_44908684	Alzheimer

Action:

```
# Plot variants against ranks for the top 5 ranked genes in Alzheimer
fig = px.scatter(snpmeta_df, x='rank', y='rsID', color='gene', title='Variants vs Ranks for Top 5 Ranked Genes in Alzheimer')
show(fig)
```

Variants vs Ranks for Top 5 Ranked Genes in Alzheimer



Ask me a question

Plot the PGS score IDs and their count in European Ancestry for Alzheimer

Submit

User prompt: even small typos lead to correct results

PGS Score IDs and Count in European Ancestry for Alzheimer



Answer:

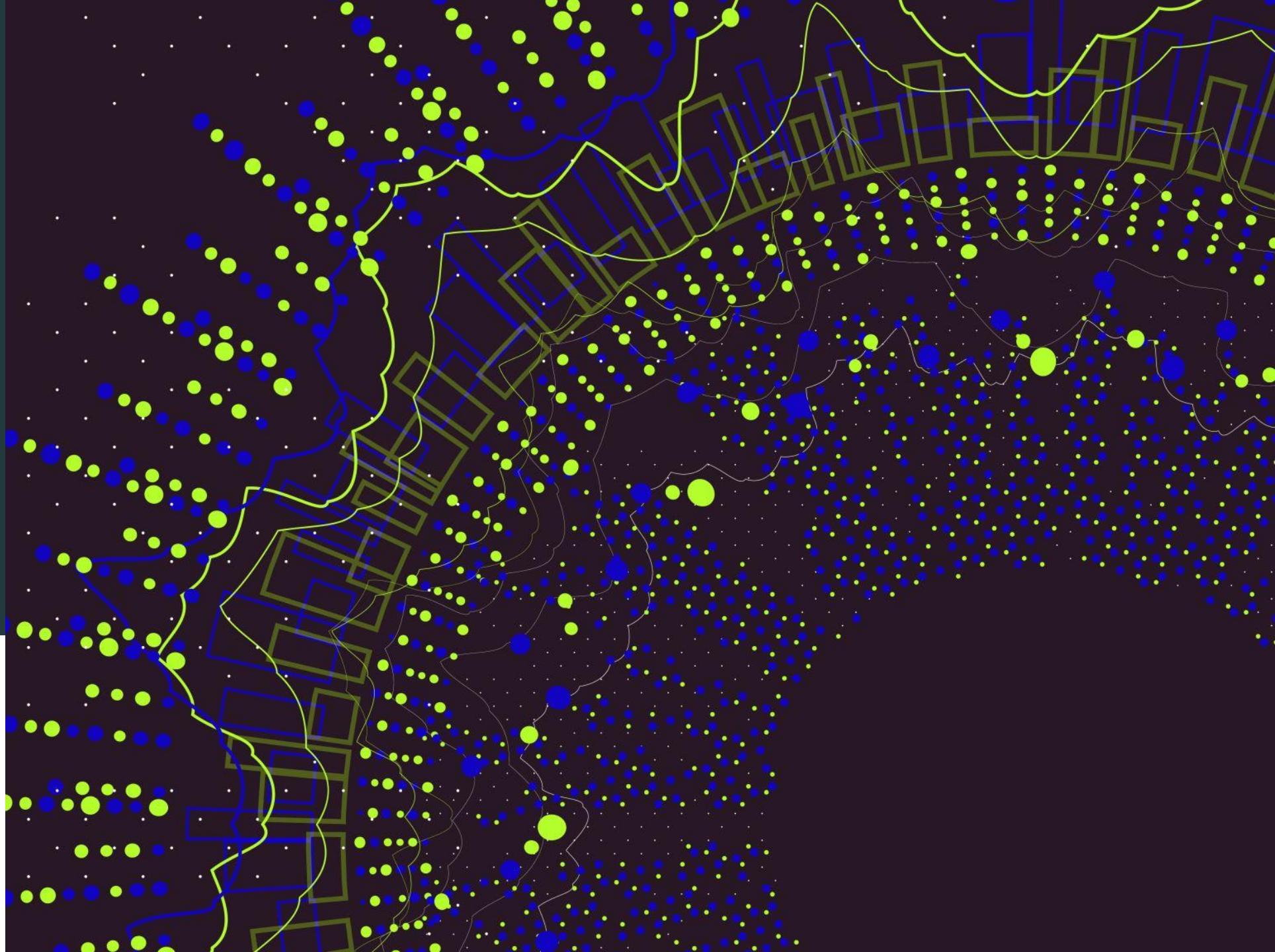
The PGS score IDs and their count in European Ancestry for Alzheimer are as follows:

- PGS000025: 3
- PGS000026: 1
- PGS000053: 1
- PGS000334: 3
- PGS000779: 1
- PGS000823: 2
- PGS000876: 1
- PGS000898: 2
- PGS000945: 2
- PGS001347: 2

Results from the PGS Rank database: Total Count of each PGS Score ID for Alzheimer in European Ancestry; auto-decides the relevant type of graph and generates the Python code for visualization as well from natural language prompts

These counts represent the number of PGS scores available for each PGS score ID in the European Ancestry for Alzheimer.

Gene API Chat



Explore external websites related to genetic data and information ⓘ

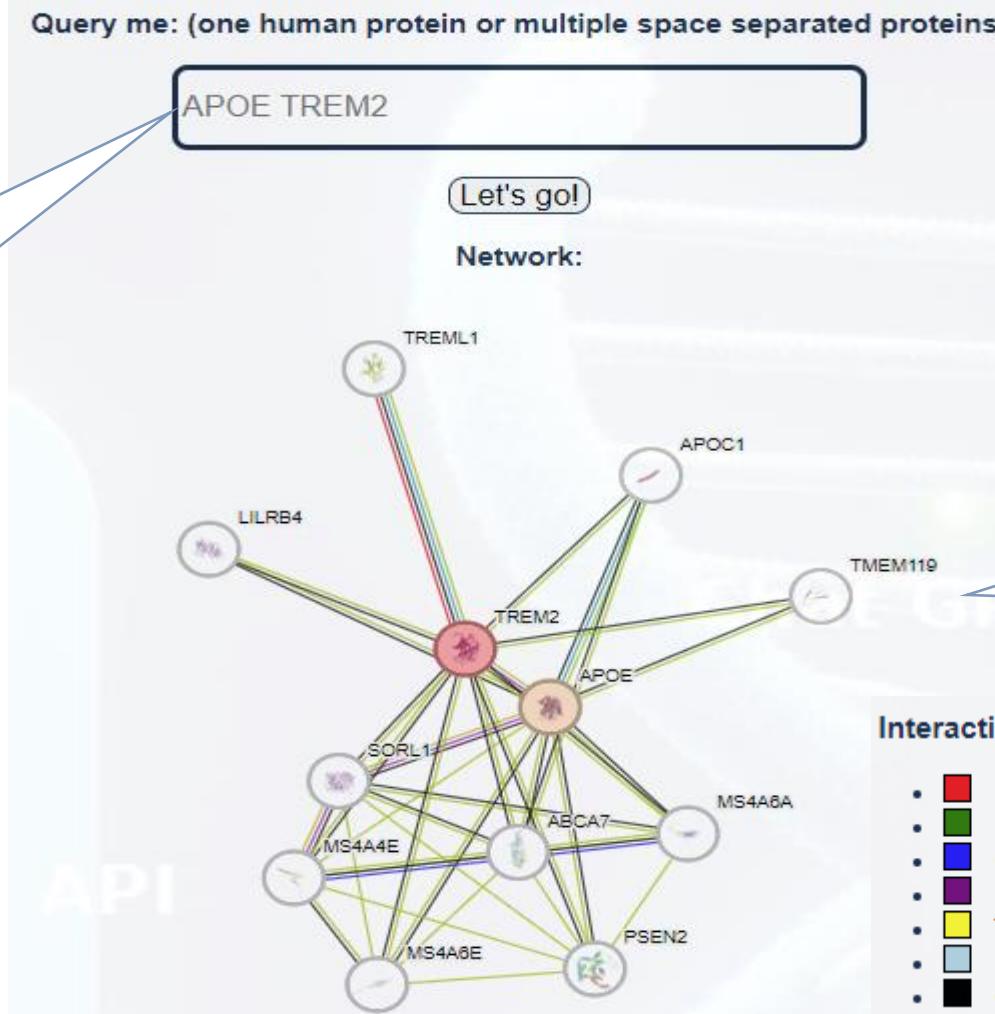
Your question or command:

protein interaction

Submit

User prompt: keywords like protein interaction, gene interaction, gene network, interaction network or fully formed questions with these keywords calls STRING API.

User enters the protein or protein names separated by space



Replicated the STRING API web interface

if no interaction found, selects 10 closest neighbors

Types of evidences are highlighted in different colors

Home

PGSChat

GeneAPIChat

LiteratureSearch

Settings

Explore external websites related to genetic data and information

Your question or command:

enrich

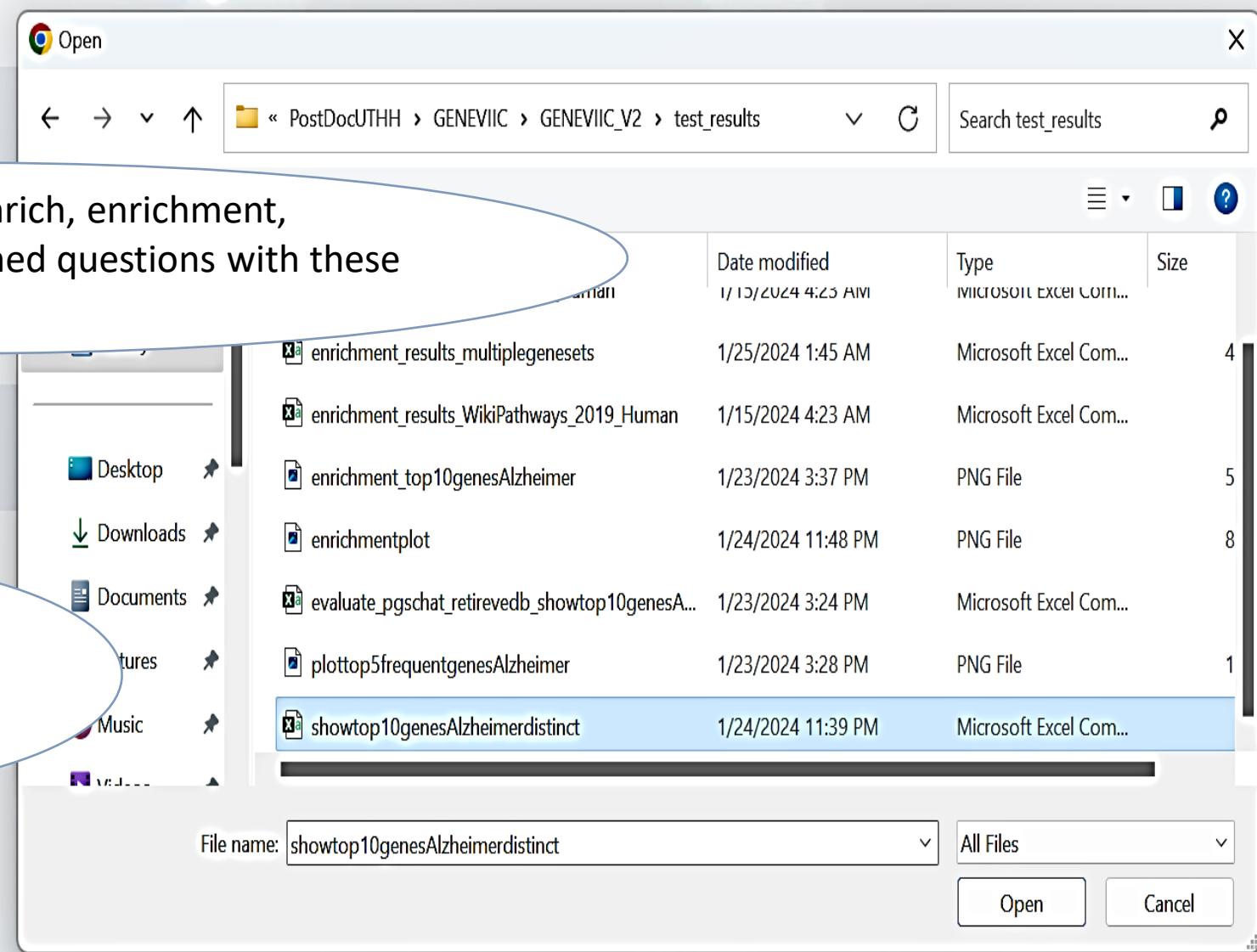
User prompt: keywords like enrich, enrichment, analysis, pathway or fully formed questions with these keywords calls ENRICHR API.

Upload your gene list

Drag and drop file here

Limit 200MB per file

Upload the gene list for top 10 Alzheimer disease saved previously as csv file



Explore external websites related to genetic data and information

Your question or command:

enrich



Submit

Upload your gene list



Drag and drop file here
Limit 200MB per file

Browse files



showtop10genesAlzheimerdistinct.csv 509.0B



Choose gene set libraries

GO_Biological_P... 

KEGG_2019_Hu... 

WikiPathways_2... 

Enrich

Enrichment analysis results saved to test_results\enrichment_results_multiplegenesets.csv

Do you want to visualize the enrichment results?

Choose from the available gene set reference libraries:
either one or more at a time

Enrichment analysis saved in local directory under root/test_results as a csv file



enrichment_results_multiplegenesets.csv 43.4KB



Choose type of visualization

Network Graph

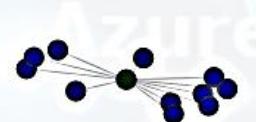
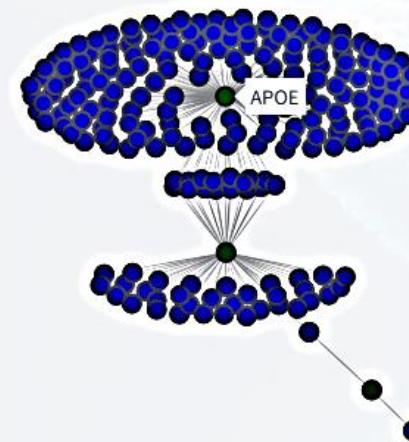


Enter p-value cutoff

0.05



Visualize Enrichment



Literature Search



Home

PGSChat

GeneAPIChat

LiteratureSearch

Click Settings 🤖 for Open AI credentials

Settings

Find literature evidence in PubMed, Google Scholar or Arxiv

Mention the portal you want to search in along with your search query

Alert! Alert!

You need to specify Open AI (here, Azure's) and SERP API credentials to proceed.

👉 Click on Settings on the left sidebar!

Steps to navigate this section:

- o [Azure OpenAI Instructions](#)
- o [Create a free SERP API account to get the private key to search in Google Scholar](#)
- o Use prompts like "articles with APOE and Alzheimer in PubMed"
- o Type or copy paste link of an article in the prompt/user input to retrieve its abstract
- o For advanced questions such as forecasting, you can use GPT-4 (if available) as the engine

You:

Hi!



Welcome to your Literature Review Assistant.

Enter your query above, e.g. 'Search for articles with gene APOE and Alzheimer in Pubmed.'

Errors
and Steps
to rectify

Find literature evidence from PubMed, Google Scholar or Arxiv.

Mention the portal you want to search in along with your search query.

You:

Search for articles with gene APOE and Alzheimer in Pubmed

Search in PubMed for keywords
'APOE' and Alzheimer'

Hi!



" Welcome to your Literature Review Assistant.

Enter your query above, e.g. 'Search for articles with gene APOE and Alzheimer in Pubmed.'

Best Matched
top results



Here are some articles related to APOE and Alzheimer's disease:

1. [ApoE in Alzheimer's disease: pathophysiology and therapeutic strategies](<https://pubmed.ncbi.nlm.nih.gov/36348357/>)
2. [APOE and Alzheimer's disease: advances in genetics, pathophysiology, and therapeutic approaches](<https://pubmed.ncbi.nlm.nih.gov/33340485/>)
3. [ApoE Lipidation as a Therapeutic Target in Alzheimer's Disease](<https://pubmed.ncbi.nlm.nih.gov/32882843/>)

Search for articles with gene APOE and Alzheimer in Pubmed



Search in Google Scholar for keywords 'APOE' and Alzheimer'

search for articles with APOE and Alzheimer in Google Scholar



Here are some articles related to APOE and Alzheimer's disease:

1. [APOE and Alzheimer disease: a major gene with semi-dominant inheritance](<https://www.nature.com/articles/mp201152>)
2. [APOE and Alzheimer's disease: advances in genetics, pathophysiology, and therapeutic approaches]([https://www.thelancet.com/journals/laneur/article/PIIS1474-4422\(20\)30412-9/fulltext](https://www.thelancet.com/journals/laneur/article/PIIS1474-4422(20)30412-9/fulltext))
3. [Apolipoprotein E and Alzheimer disease.](<https://www.pnas.org/doi/abs/10.1073/pnas.92.11.4725>)
4. [Apolipoprotein E and Alzheimer disease: risk, mechanisms and therapy](<https://www.nature.com/articles/nrneurol.2012.263>)
5. [Apolipoprotein E and Alzheimer disease: pathobiology and targeting strategies](<https://www.nature.com/articles/s41582-019-0228-7>)

Search in arxiv for keywords 'APOE' and Alzheimer'

search for articles with APOE and Alzheimer in Arxiv



Here are some articles related to APOE and Alzheimer's disease from Arxiv:

1. Analyzing the effect of APOE on Alzheimer's disease progression using an event-based model for stratified populations
 - Authors: Vikram Venkatraghavan, Stefan Klein, Lana Fani, Leontine S. Ham, Henri Vrooman, M. Kamran Ikram, Wiro J. Niessen, Esther E. Bron
 - [Link to Article](<https://arxiv.org/abs/2009.07147>)
2. A deep belief network-based method to identify proteomic risk markers for Alzheimer disease
 - Authors: Ning An, Liuqi Jin, Huitong Ding, Jiaoyun Yang, Jing Yuan
 - [Link to Article](<https://arxiv.org/abs/2003.05247>)
3. Accounting for iron-related off-target binding effects of 18F-AV1451 PET in the evaluation of cognition and microstructure in APOE-e4+ MCI
 - Authors: Jason Langley, Daniel E. Huddleston, Ilana J. Bennett, Xiaoping P. Hu
 - [Link to Article](<https://arxiv.org/abs/2203.03847>)

Extract the abstract when user provides the link to an article

<https://pubmed.ncbi.nlm.nih.gov/33340485>

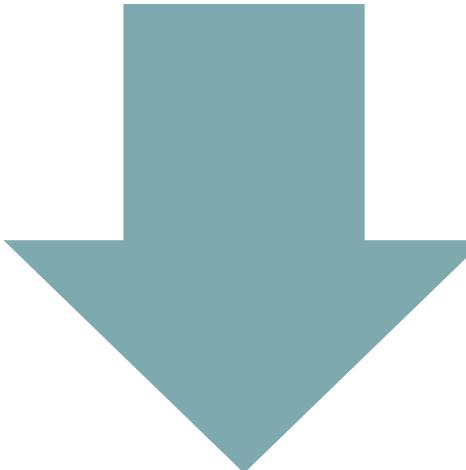


The article titled 'APOE and Alzheimer's disease: advances in genetics, pathophysiology, and therapeutic approaches' provides an overview of the role of APOE in Alzheimer's disease and discusses various therapeutic approaches targeting APOE. The article is published in The Lancet Neurology journal. You can access the article [here] (<https://pubmed.ncbi.nlm.nih.gov/33340485/>).

Pros & Cons

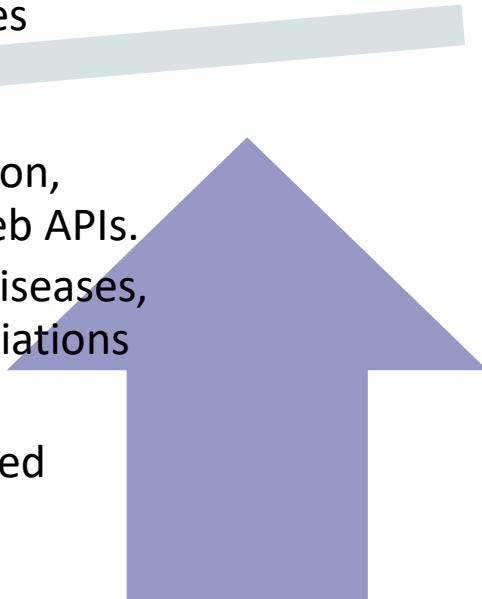
Contributions

- generative AI unifies and streamlines access, navigation, automate analysis of biomedical DBs and external web APIs.
- aids in extracting pivotal genetic variants related to diseases, building an extensive map of variant-gene-trait associations
- intuitive, user-friendly interface
- no biomedical knowledge or technical training required
- leverages standardized databases
- minimal "hallucination"
- customize data sources
- protect sensitive clinical data, Azure OpenAI's HIPAA-compliant infrastructure



Limitations

- real-time lags
- web-dependency, may crash altogether, needs to restarted
- only a prototype/pilot framework
- demonstrates capabilities using a limited PGS Rank DB
 - ✓ data for only three phenotypes
 - ✓ basic approach to ranking variants via PGS scores



Future Directions

- ✓ Better variant ranking scheme:
 - comprehensive data
 - robust weighing scheme
 - various factors:
 - ethnic background
 - genotype data
 - specific PGS scoring techniques
 - degree of sample overlap across PGS datasets for each phenotype
- ✓ broaden the database scope
- ✓ integrate additional biomedical web APIs
- ✓ enhance Literature Search functionalities:
 - auto-extraction of deeper insights using advanced NLP techniques
- ✓ integrate automated predictive modeling and analytics using generative AI