The Yeast Connectome:

Constructing a Connectome Database from Scientific Literature using GPT Models

A Thesis submitted for the completion of

requirements for the degree of

Bachelor of Science

(Research)

by

Mani Kumar R

Undergraduate Programme

Indian Institute of Science

Under the supervision of

Prof. Marek Mutwil

School of Biological Sciences, Nanyang Technological University, Singapore

Prof. Mohit Kumar Jolly

Department of Bioengineering, Indian Institute of Science

Acknowledgements

This thesis was carried out during and after my time at Nanyang Technological University, Singapore. Throughout this period, I delved into the fascinating realm of large language models. Additionally, my travels across Singapore gave me many new experiences and memories. This wouldn't have been possible without the support of Prof. Marek Mutwil, who kindly accepted me into his group and guided me during the project. Prof. Guillaume Thibault, for helping me to find the important protein-protein interactions in yeast. I am thankful to NTU - India Connect Research fellowship program for financial support during my project at NTU.

I would like express my sincere thanks to Prof Mohit Kumar Jolly for supporting me.

I would like acknowledge Ms. An-Nikol Kutevska for her contribution in data preparation for fine-tuning the GPT model and evaluation.

I am grateful to my parents and sister for their constant support, care and encouragement. I am grateful to my dearest friend Vedavalli for her constant support over the years.

I would also like to thank Kishore Vaigyanik Protsahan Yojana (KVPY) for the financial support offered for my undergraduate studies at IISc.

Certificate

This is to certify that the thesis titled "**The Yeast Connectome: Constructing a Connectome Database from Scientific Literature using GPT Models**" is original and was carried out by Mani Kumar R under the supervision of Prof. Marek Mutwil for the degree of Bachelor of Science with a Major in Biology at the Indian Institute of Science, Bangalore, India.

Prof Mohit Kumar Jolly
Department of Bioengineering,
Indian Institute of Science,
Bangalore-560012.

Declaration

I declare that this work titled "**The Yeast Connectome: Constructing a Connectome Database from Scientific Literature using GPT Models**" is original and was carried out by me during my internship at Prof. Marek Mutwil lab, School of Biological Sciences, Nanyang Technological University, Singapore. This thesis has not formed the basis for the award of any degree, diploma, associateship, membership, or similar title of any university or institution. Wherever I have referred to another author's work, I have rightly attributed it in the references. I have acknowledged the people who have helped me during this thesis project.

Mani Kumar R
Undergraduate Program
Indian Institute of Science, Bengaluru

Abstract

Yeast, particularly *Saccharomyces cerevisiae*, is important model organism in biology due to its eukaryotic nature and ease of manipulation in haploid and diploid forms. Many essential cellular processes are conserved between yeast and humans, making yeast a powerful tool for studying cell biology and disease. In addition to its role as a model organism, yeast has been used extensively in biotechnological research. Moreover, yeast has been used as a "cell-factory" to produce commercially important proteins such as insulin, human serum albumin, and hepatitis vaccines. Hence understanding the yeast biology, gene function information is crucial. On an average, 10,000 research articles on yeast were published every year. Indeed, the manual extraction of gene function information is a meticulous and resource-intensive task. Manual curation involves thorough examination and interpretation of research data, which can be quite slow and requires significant effort from researchers.

In this thesis project, we presented an innovative high-throughput pipeline that leverages OPENAI's Generative Pre-trained Transformer(GPT) Model (specifically GPT 3.5 turbo) for the systematic extraction and analysis of connectivity information from both full-texts and abstracts of 84,427 publications related to *Saccharomyces cerevisiae*. Using this approach, we extracted 34,32,749 relationships involving genes, molecules, compartments, stresses, organs, and other yeast entities. we created a comprehensive, searchable online connectome that link relevant keywords to their corresponding PubMed IDs, providing easy access to a vast knowledge network encompassing *Saccharomyces cerevisiae*. Our work reinforce the transformative potential of merging artificial intelligence with bioinformatics to deepen our understanding of complex biological systems. We also discuss significant nodes within yeast connectome, including the HSP104 and ATG8 proteins.

Contents

Acknowledgements i

Certificate ii

Declaration iii

Abstract iv

1 Introduction 1

1.1 Growing literature and Natural Language processing 1

1.2 Natural Language Processing(NLP) and Large Language Models (LLMs) 2

1.3 GPT-3.5 model for Gene function information extraction. 3

2 Methods 4

2.1 Retrieval and pre-processing of literature 4

2.2 Processing of texts using GPT-3.5 Turbo 4

2.3 Filtering and Construction of the database 5

2.4 API for yeast connectome. 7

3 Results 8

3.1 Data collection and analysis 8

3.2 Evaluation of GPT-3.5 Turbo models 9

3.3 Connectome analysis 10

3.4 User interactions 13

3.5 Comparative Analysis of Connectome and BioGRID Networks 17

3.6 Utilization of YeastConnectome in Gene Regulatory Network Analysis. . 19
3.6.1 Example 1 –"HSP104" 19
3.6.2 Example 2 – "ATG8" 21
3.7 Gene ontology (GO Term) prediction using GPT embeddings 23
3.7.1 What are embeddings? 23
3.7.2 Cosine similarity 24
3.7.3 GO Term prediction and evaluation 24

4 Discussion **25**

5 Limitations and Future directions **27**

5.1 Shortcomings of the approach and ways to mitigate 27
5.2 Future directions 28

Chapter 1

Introduction

Many essential cellular processes are the same in yeast and humans, making it a powerful model for studying cell biology and disease [1]. Studying the biology of yeast has enabled scientists to work out the connections between genes and proteins, and the functions they carry out in cells. An important feature of yeasts is that their cells, like humans, are eukaryotic. Yeast genetics has also been a powerful tool in studying human diseases [1]. The conservation of fundamental cellular processes between yeasts and humans allows for the use of yeast in high-throughput genetic screens to identify human disease genes and dissect molecular pathways regulating disease-related proteins [2]. Moreover, the mapping of yeast chromosomes has provided insights into genetic processes such as mutation repair, enzyme production, and cellular division regulation, which are relevant to human health issues like cancer and aging [3]. Yeast, particularly *Saccharomyces cerevisiae*, has long been one of the best-studied model organisms for basic biological research. The number of scientific publications on yeast shows its importance. Despite the many genes that have been described, a significant portion of the yeast genome still needs to be explored, as indicated by the research articles being published [4].

1.1 Growing literature and Natural Language processing

The number of research articles published every year on *Saccharomyces cerevisiae* have grown drastically from 1,782 articles in 1980's to 10,182 articles in 2022 [5]. It is becoming increasingly hard for researchers in keeping up with this growing literature [6]. Indeed, the process of gaining knowledge about gene function depends heavily on experimentally verified data, often referred to as "gold standard data". Creating this gold standard involved manually extracting functional information about genes from scientific articles.

This is a daunting task given the vast amount of literature available and labour intensive. The manual curation process is not only time consuming but also requires a high level of expertise to accurately interpret and extract relevant information.

BioGRID is a large repository with many interaction datasets, allowing the study of protein and genetic interactions between different organisms[7]. However, like many databases, it may not be able to keep up with real-time search dynamics due to inherent delays in data management [8]. The reactome database, while delineating pathways for many biological processes, is missing some species-specific functions and distinct cellular contexts [9]. Despite their usefulness, these tools still have limitations. Their reliance on curated data ensures accuracy, but can result in updates lagging behind the most recent material due to the labor-intensive nature of manual curation. Additionally, these databases may not fully cover the complex relationships between genes. Such relationships are important for a comprehensive understanding of complex phenotypes and diseases. Therefore, there is a need for automatic methods like Natural Language Processing(NLP) that can aid curation.

Natural language processing (NLP) methods have been employed for a long time to carry out a variety of text-mining tasks [10]. Through the use of supervised learning techniques, language models have been trained to become proficient in the specific tasks they were designed for [11]. In the realm of biological NLP (BioNLP), these models are used for extracting insights, summarizing, and analyzing data from biological text sources, thereby enhancing the efficiency and processes of research [12].

1.2 Natural Language Processing(NLP) and Large Language Models (LLMs)

Named entity recognition (NER) is an important technique in natural language processing (NLP), focusing on identifying and classifying entities. NER involves identifying key information in text and classifying it into a set of predetermined categories. An entity is something that is mentioned or mentioned consistently in a text, such as names of people, genes, proteins, quantities and defined categories. In the context of scientific literature, NER plays an important role. For example, SciNER, a NER model specifically designed to recognize named entities in scientific contexts has been developed [13]. Relationship extraction (RE) is another important process in NLP which involves the identification and classification of connections between entities in natural language text [14].

Large Language Models have demonstrated superior capabilities in natural language processing tasks and beyond. By capturing more complex relationships between words, large language models can lead to more accurate predictions [15]. Their ability to grasp contextual relationships and the nuances of language has revolutionized NLP tasks, enabling more natural and intuitive human-machine interactions. Additionally, large language models can process more data in less time, making them faster and more efficient [16]. In summary, both NER and RE play a central role in extracting valuable information from scientific literature. The development of LLM has significantly improved these processes, providing more accurate and efficient results.

In recent AI rise, large language models (LLMs) have received significant attention due to their outstanding ability to handle various natural language processing (NLP) tasks. These tasks include creating texts, summarizing, and answering questions. Among these models, Generative Pre-Trained Transformer (GPT) has become a global phenomenon. It is trained on very large datasets containing 175 billion parameters. GPT demonstrates its performance in understanding natural language and engaging in human-like conversations. It can produce responses that closely resemble human language [17]. OPENAI's ChatGPT has demonstrated performance comparable to humans. Specifically, in answering multiple-choice questions related to human genetics, ChatGPT achieved an accuracy of 68.2%, surpassing human responders who achieved 66.6% accuracy [18]. This shows ChatGPT's potential to compete with human counterparts and provide responses that mimic humans.

1.3 GPT-3.5 model for Gene function information extraction.

We have used the complex text mining capabilities of the Generative Pre-trained Transformer language model to improve our understanding of yeast biology. we developed an innovative, high-throughput text-mining pipeline that leverages OPENAI's GPT-3.5 turbo model to systematically extract and analyze gene functional information from a large corpus of research publications concerning *saccharomyces cerevisiae*. we processed more than 84,427 research abstracts and full text articles from leading journals(Figure 3.1). we extracted more than 34,32,749 relationships involving genes, molecules, compartments, stresses, organs, and other yeast entities. The database is powered with interactive web interface and made available to researchers and public. Yeast connectome is accessible at http://yeast.connectome.tools/. Additionally, the data can be accessed programatically using it's Application Programming Interface (API).

Chapter 2

Methods

2.1 Retrieval and pre-processing of literature

A comprehensive list of Yeast genes and their aliases was downloaded from the YeastMine database [19] and UniProt [20]. For each gene on this list, we initiated searches in PubMed using a specialized search query: "*(Saccharomyces cerevisiae[Title/Abstract] AND gene[tw]) OR (S cerevisiae[Title/Abstract] AND gene[tw]) OR (yeast[Title/Abstract] AND gene[tw])*". This query enables us to search for research papers featuring the gene name prominently in the text. We utilized the Bio.Entrez package (v1.81) to retrieve articles containing gene function information for *Saccharomyces cerevisiae* from PubMed [21], downloading abstracts and full-text papers. Additionally, the Elsevier API was accessed via NTU library to retrieve full-length articles where available.

2.2 Processing of texts using GPT-3.5 Turbo

We used a Python [22] scripts designed to extract relevant information from each article. This script systematically scans the text of each downloaded article, identifies sentences containing pertinent information related to the specific gene of interest, and extracts the target sentence along with the sentence immediately preceding and following it and excluding the extraneous information.

The method employed for processing texts in this study involves using OpenAI's GPT-3.5-turbo model. To ensure the accuracy of the information generated by the model, we programmatically feed the extracted information from the texts into the language model using its API [23], with the temperature parameter set to zero. This setting minimizes the randomness in the model's output, thereby reducing the likelihood of generating incorrect

or hallucinated information. The model's primary task is identifying entities within the text, such as genes and proteins and discerning their relationships. We engineered a specific prompt (Figure 2.3) and employed a one-shot learning methodology to accomplish this. This prompt guides the model to generate structured and easy outputs, with each line containing precisely two entities and their relationship. This approach allows for a more granular and detailed analysis of the relationships between various entities within the scientific texts.

```
messages= [

    {"role": "system", "content": "You are data scientist tasked with identifying entities and
relationships in scientific text. Please provide only one statement per line, and ensure that each
line contains exactly two entities. If a relationship involves more than two entities, please break
it down into multiple separate lines. "},

    {"role": "user", "content": "CLas inhibits callose deposition in the sieve pores and the
accumulation of reactive oxygen species to favor its cell-to-cell movement. CESA1 interacts with
CESA4 and 7. HIS3 mutant inhibits proline production."},

    {"role": "assistant", "content": "CLas: !inhibits! callose deposition\nCLas: !inhibits!
accumulation of reactive oxygen species\ncallose deposition: !inhibits! cell-to-cell
movement\nCESA1!interacts with!CESA4\nCESA1!interacts with!CESA7\nHIS3 mutant!inhibits!proline
production"},

    {"role":"user","content":"VERY SHORT, CONCISE SUMMARY CONTAINING ALL INFORMATION WITH TWO ACTORS
PER LINE: "+ article text}
  ]
```

Figure 2.1: GPT prompt

2.3 Filtering and Construction of the database

This study employed a multi-step approach to generate and evaluate edges for further analysis. Initially, we utilized the GPT-3.5 turbo model [23] to create edges. To filter out hallucinations, the spaCy [24] NLP library was used to check if the edges corresponded to the input text; any edges not found were classified as hallucinations. An example article text and GPT output are shown in Figure 2.2 and Figure 2.3 respectively. Additionally, edges with less than two or more than two exclamation marks were classified as bad edges. The identified bad edges were then reprocessed by feeding them into the GPT model. Finally, after this iterative process, we categorized the resulting edges as either "good" or "bad" (still exhibiting issues or inaccuracies). The implementation was done using Python with spaCy for NLP tasks and a custom script for edge filtering.

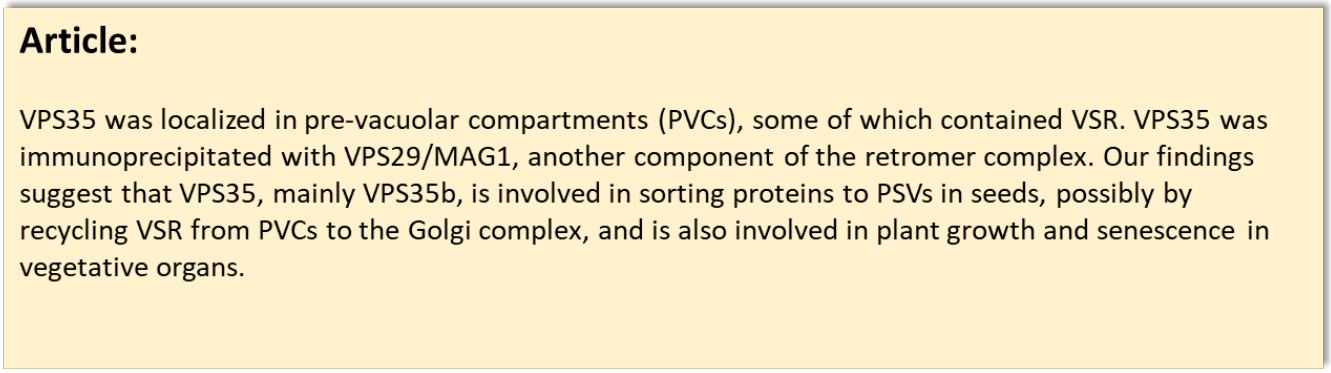

Figure 2.2: Example input article

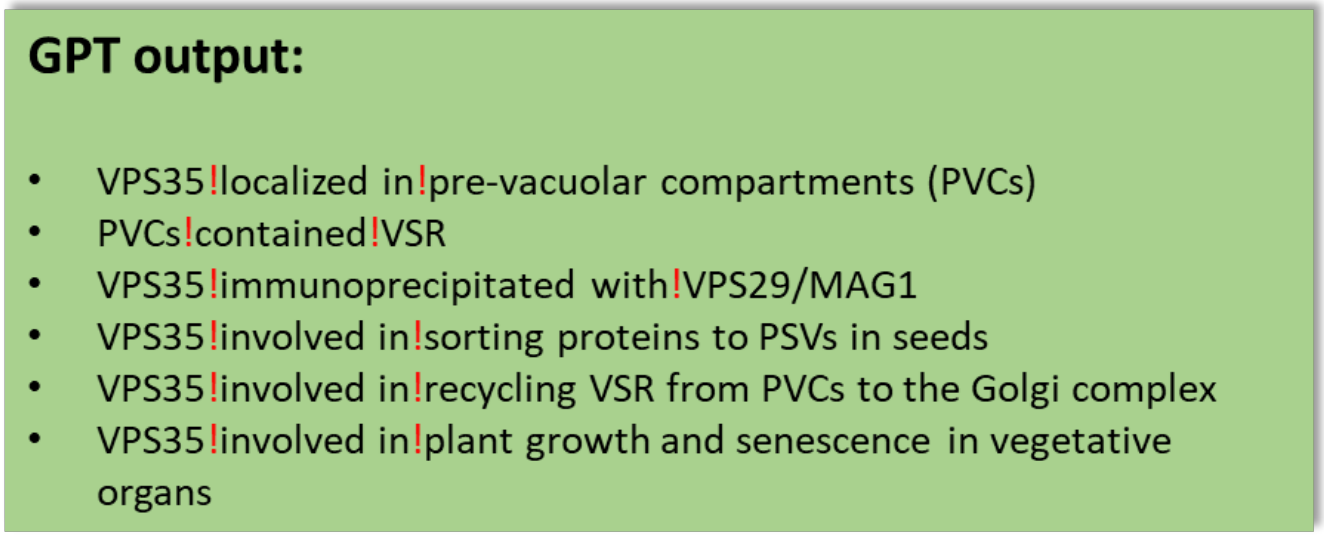

Figure 2.3: example GPT output

We constructed a comprehensive database to facilitate the analysis and visualization of network edges. Our approach involves selecting high-quality edges from the dataset. To create an efficient and user-friendly web application interface, we used the Python-Flask framework (v2.2.3). Additionally, Networkx (v3.1) is used for graph analysis and manipulation, enabling efficient handling of network structures. To enhance the visualization experience, we incorporated Cytoscape.js (v3.23). As our database management system, we opted for MongoDB, a document-oriented NoSQL database. MongoDB's flexibility allowed us to adapt schemas as needed while maintaining simplicity. The integration with PyMongo, a MongoDB's Python driver, facilitated seamless interaction with our database.

In summary, our approach involved edge selection, database hosting, framework utilization, and MongoDB integration, ensuring efficient data management and robust analysis capabilities for our research.

2.4 API for yeast connectome.

The Yeast Connectome includes a robust Application Programming Interface (API) designed to facilitate remote search queries for users. This API accepts GET requests and leverages the same package set as previously detailed. With each successful API call, users receive a JSON object containing relevant nodes, edges, and text summaries related to their search query. To utilize the API effectively, users need to append /api/search type /search query to the web address, replacing search type with the desired search category and search query with their specific query. For example, information related to gene MKK2 can be accessed by sending GET request to the url http://yeast.connectome.tools/api/exact/mkk2.

Chapter 3

Results

3.1 Data collection and analysis

Utilizing the OpenAI GPT-3.5 model, we successfully constructed the connectome forYeast (*Saccharomyces cerevisiae*), leveraging the power of GPT's natural language processing capabilities. We processed 84,427 publications (Figure 3.1), leading to the construction of YeastConnectome—a robust resource elucidating 3,432,749 relationships among genes, molecules, compartments, stresses, organs, and other yeast entities. The YeastConnectome is accessible through dedicated platform: http://yeast.connectome.tools. It provides a multifaceted platform for querying genes, metabolites, organs, and other entities using terms, author names, and PubMed IDs. The platform is completed by an "entities" catalog, which lists all the entities present in the database.

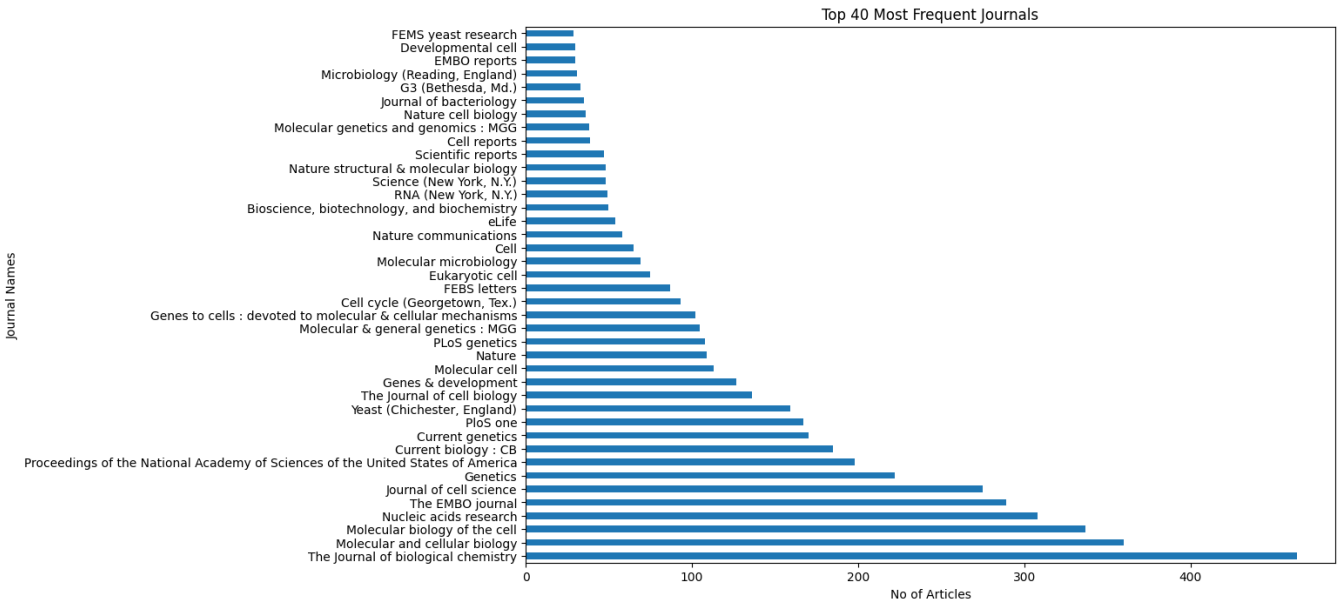

Figure 3.1: The 40 Most Frequent Journals

3.2 Evaluation of GPT-3.5 Turbo models

The OPENAI's GPT 3.5 turbo can be fine tuned for specific purpose to obtain higher quality results than prompting using our own data [25]. They also provide function calling method to obtain more reliable, structured data back from the model in a specific formats like JSON etc [26]. Among those, We tried to identify the most effective Generative Pre-trained Transformer (GPT) model for Named Entity Recognition (NER) and biomedical Relationship Extraction (RE). We used OpenAI's function calling method to improve the accuracy, which was applied via a custom-built function. The function used for function calling is shown in the Figure 3.2. Theis function contains the information/format to generate structured JSON data, detailing 'Entity1', 'Entity2', and 'Relationship' between them.

```
function={
    "name": "get_all_relationships",
    "description": "A funcction that takes in a text from scientic literature as input and provide
relationships between exactly two entities.",
    "parameters": {
      "type": "object",
      "properties": {
        "Relationships":{
          "type":"array",
          "description":"A list of all identified relationships from the text.",
          "items":{
        "Entity1": {
          "type": "string","description": "Name of the entity from which the relationship is associated.",},
        "Relationship":{
          "type": "string","description": "The type of relation between the two entities.",},
        "Entity22": {
          "type": "string","description": "Name of the entity to which the relationship is associated.",}
          }
       },"required":["Entity-1","relationship","Entity-2"],
      },

    },"required": ["Relationships"],
}
```

Figure 3.2: Function used for function calling with GPT model.

The evaluation is focused on the model's ability to discern biological entities within texts—such as genes, proteins, and the relationships among these entities. These relationships can be interactions between genes (e.g., gene X interacts with gene Y, protein A binds to gene B), and more. The GPT-3.5 Turbo model, previously used during the development of the PlantConnectome database, demonstrated an accuracy of approximately

75% when tasked with analyzing outputs for 50 randomly chosen plant abstracts [27].

To fine-tune the model, we manually curated 50 abstracts and example outputs in JSONL format. The resulting JSONL file is provided to GPT for fine-tuning to increase it's performance. Then, We evaluated the fine-tuned model based on it's performance on 20 random articles. Similarly, the function calling method is also evaluated on 20 random articles. The comparative accuracy of the models is shown in Table 3.1. It is evident that, GPT-3.5 turbo model without any function calling or fine-tuning, performs better. Hence, we continued with GPT-3.5 turbo.

Model	**GPT-3.5 turbo**	**Fine-tuned GPT-3.5**	**GPT-3.5 with Function calling**
%Correct	74.6%	69.3%	49.6%
%incorrect	25.4%	30.7%	50.4%

Table 3.1: Accuracy of different methods.

3.3 Connectome analysis

Our comprehensive analysis identified the 20 most frequent entities, edges, genes, and gene-to-gene interactions in the YeastConnectome database. The bar graphs shown below illustrate these results. "Saccharomyces cerevisiae" and "RAD51" are major entities in the yeast connectome with more than 3000 occurences (Figure 3.3). The connectome mostly have "interact with" edges between entities with more than 1.1 million occurences (Figure 3.4) indicating the complexity and level of interactions within the yeast's genetic network.

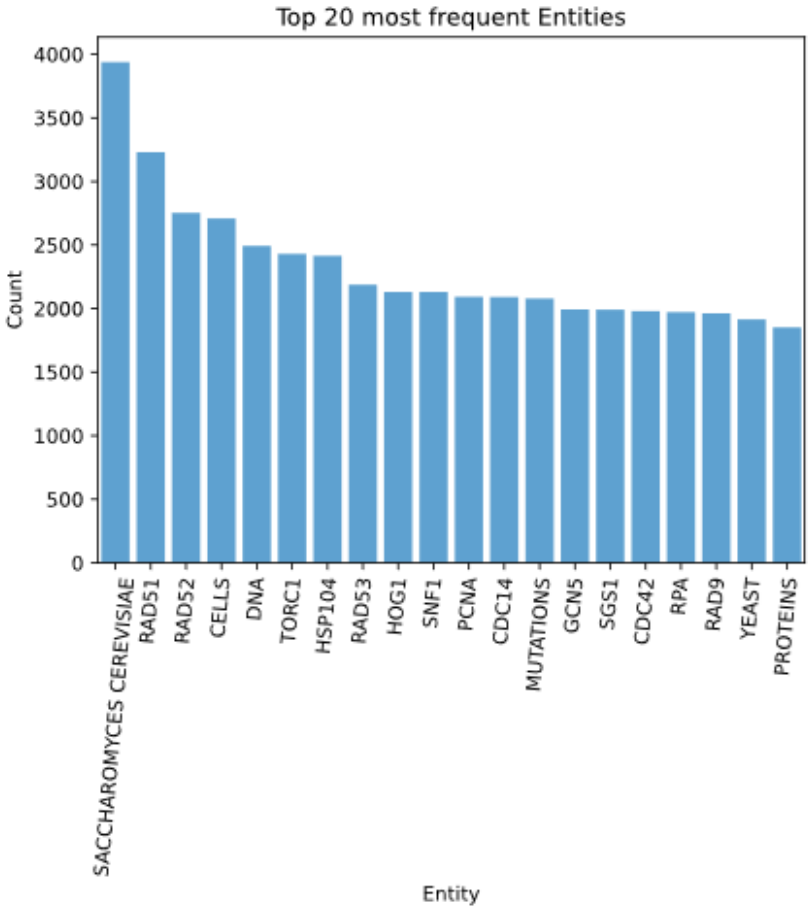

Figure 3.3: The distribution of 20 most frequent entities.

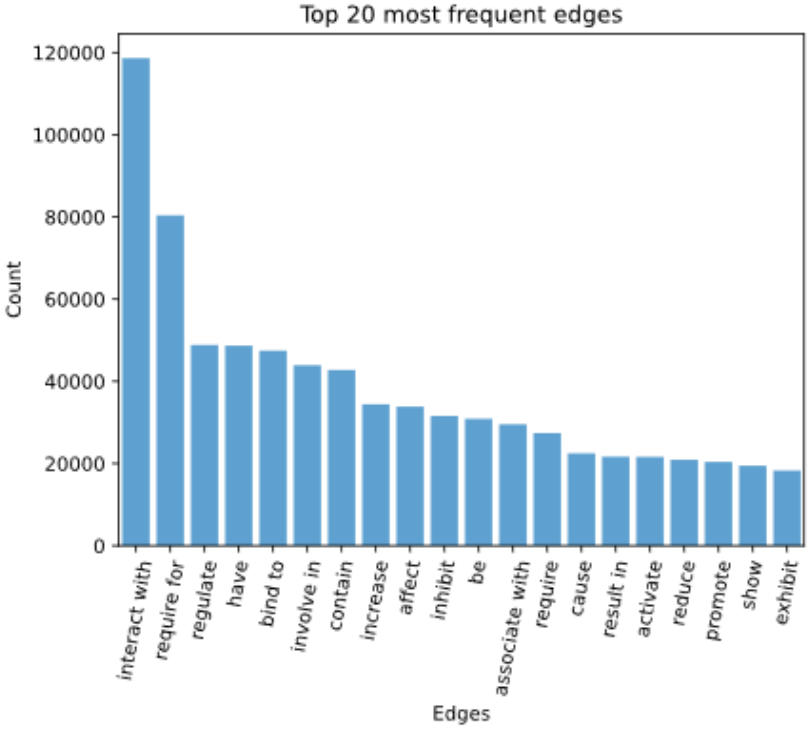

Figure 3.4: The distribution of 20 most frequent edges.

Among the genes, “GAL4”, “RAD51” and “SIR2” were prominently characterized in more than 25000 relationships (Figure 3.5), reflecting their important roles in yeast biology. “GAL4” is a transcription factor in yeast that positively regulates galactose-induced gene expression. The “Gal4” family, includes more than 50 members in the*Saccharomyces cerevisiae*, such as Oaf1, Pip2, Pdr1, Pdr3 and Leu3 [28]. Specifically, GAL4 plays an important role in activating genes involved in galactose metabolism. On the other hand, “RAD51” is a strand exchange protein required for recombinant repair of DNA double-strand breaks (DSBs) during vegetative growth and meiosis in yeast. RAD51 forms helical filaments with DNA, searching for similarities. It’s phosphorylation by Cdc28p at G2/M phase promotes DNA binding, strand invasion, and primer elongation [29]. The “interact with” edges are more prominent between the genes (Figure 3.6), followed by “bind to” and “regulate” edges, representing the protein - protein and gene - protein interactions in yeast.

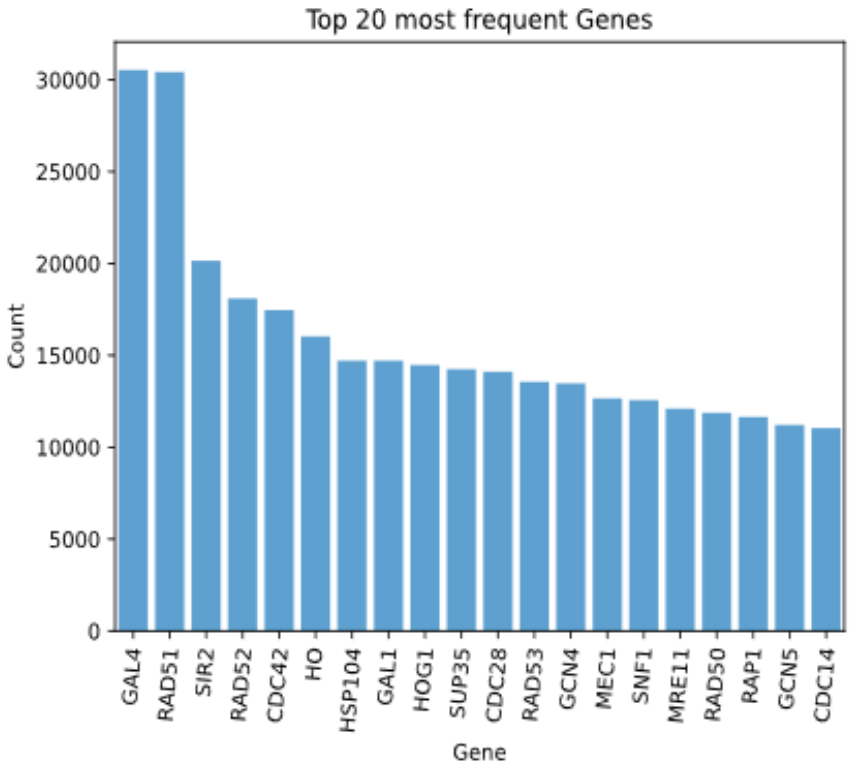

Figure 3.5: The distribution of 20 most frequent genes.

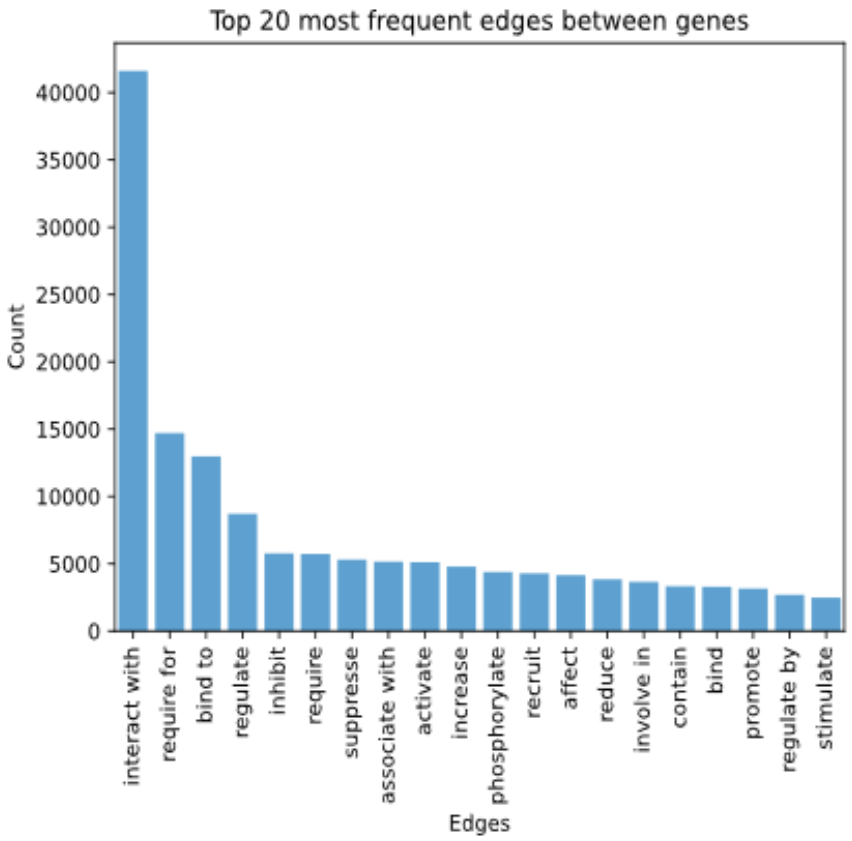

Figure 3.6: The distribution of 20 most frequent edges between genes.

3.4 User interactions

Yeast connectome web interface provide users with numerous search options. Users can initiate searches by gene, molecule, compartment, stress, cell type, organ, or other related terms. Search by "Pubmed ID" and search by "author name" is also made available (Figure 3.7). While searching for a gene or molecule, users can choose one for the following search options:

1. Search by **"Word"**: This will find all entities that contains the search query.
    - For instance, if "MKK2" is searched, this search will find the following entities "MKK2", "MKK2 genes" etc.

2. **"Exact"** search: This will finds the entity that exactly matches the search query.
    - if "MKK2" is searched, this search will find "MKK2". But, not "MKK2 genes".

3. Search by **"Aliases"**: This will find all gene aliases that are associated with the search query.

    - For example, if "MKK2" is searched, this search will find the following entities: "MKK2", "MKK1" AND "YPL140C".

4. Search by **"substring"**: This will find all entities that contain the search query as a substring.

    - if "GAL" is searched, this will return entities like "GAL4" and "GAL42".

5. **"Non-alphanumeric"** search: This will find all entities that contain the search query followed by a non-alphanumeric character.

    - For instance, if "YOR304C" is searched, this search will find the following entities: "YOR304C-A" and "YOR304C/A"

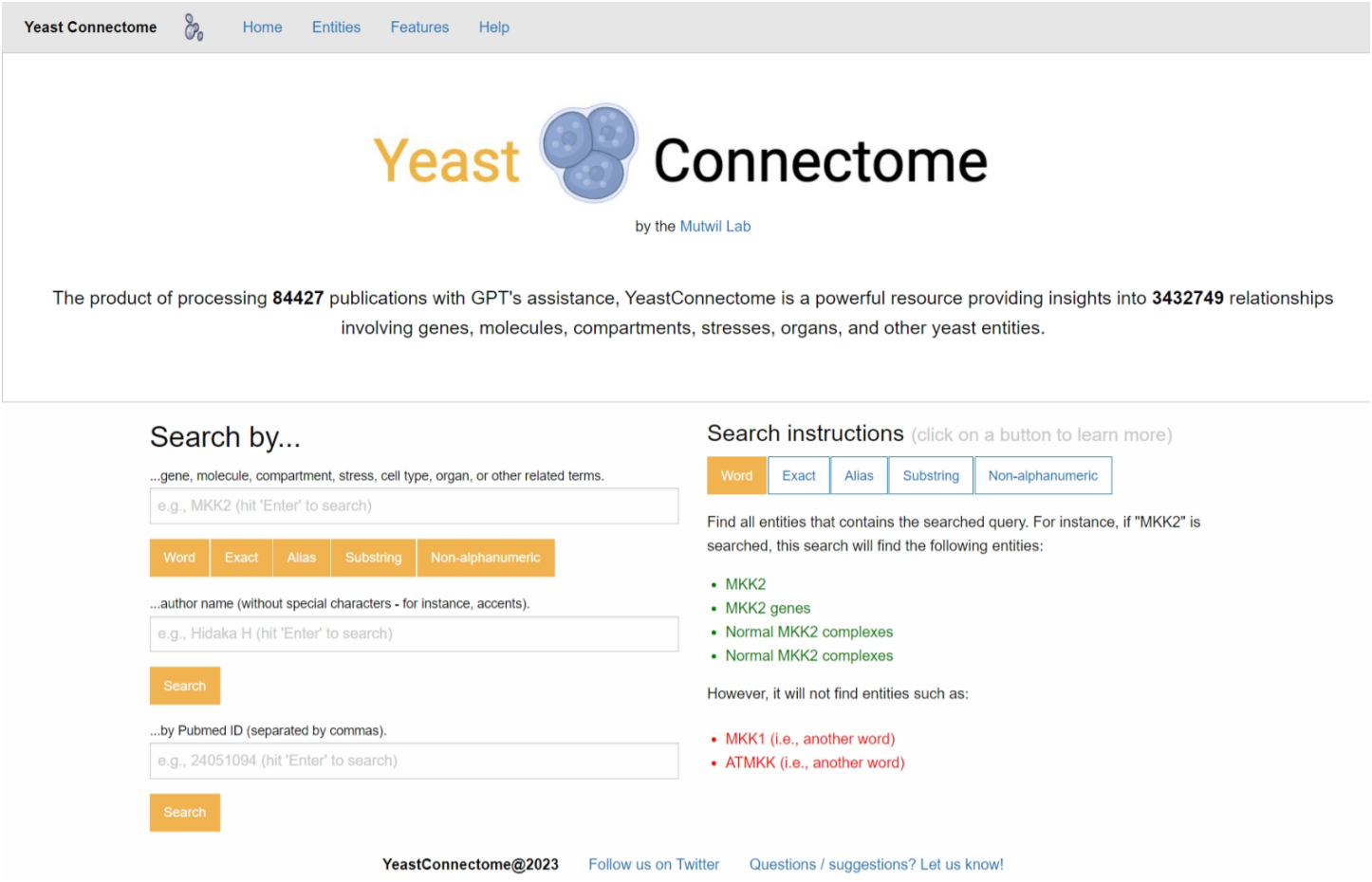

Figure 3.7: Web inferface of Yeast connectome

KnowledgeNetwork, a visual representation of Connectome allows users to interact with the network dynamically. Users can interact with the network by selecting nodes, which activates a tooltip displaying options for further exploration, such as removing nodes, isolating neighborhoods, and visiting pages corresponding to features as shown in the Figure 3.9. The network is customizable; users can remove nodes or clusters and filter relationships, such as “binds to”, “link”, or “encode for”, etc (Figure 3.8). Given the large size of the network, we limit the display to 500 nodes for optimal visibility. However, users can download the full version as a tab-delimited file for advanced analysis. In addition to the knowledge network, a textual summary is provided, organized at the node level, containing the entity name, link type, and PubMed IDs of the publications from which the relationship was derived. This summary makes it easier to verify relationships, especially when considering potential inaccuracies inherent in GPT results. The network is also available in table form at the bottom of the results page. Additionally, we extend the functionality of connections by allowing searches through the API and returning a JSON object with relevant network and functional information. This feature is aimed at researchers who need programmatic access to databases.

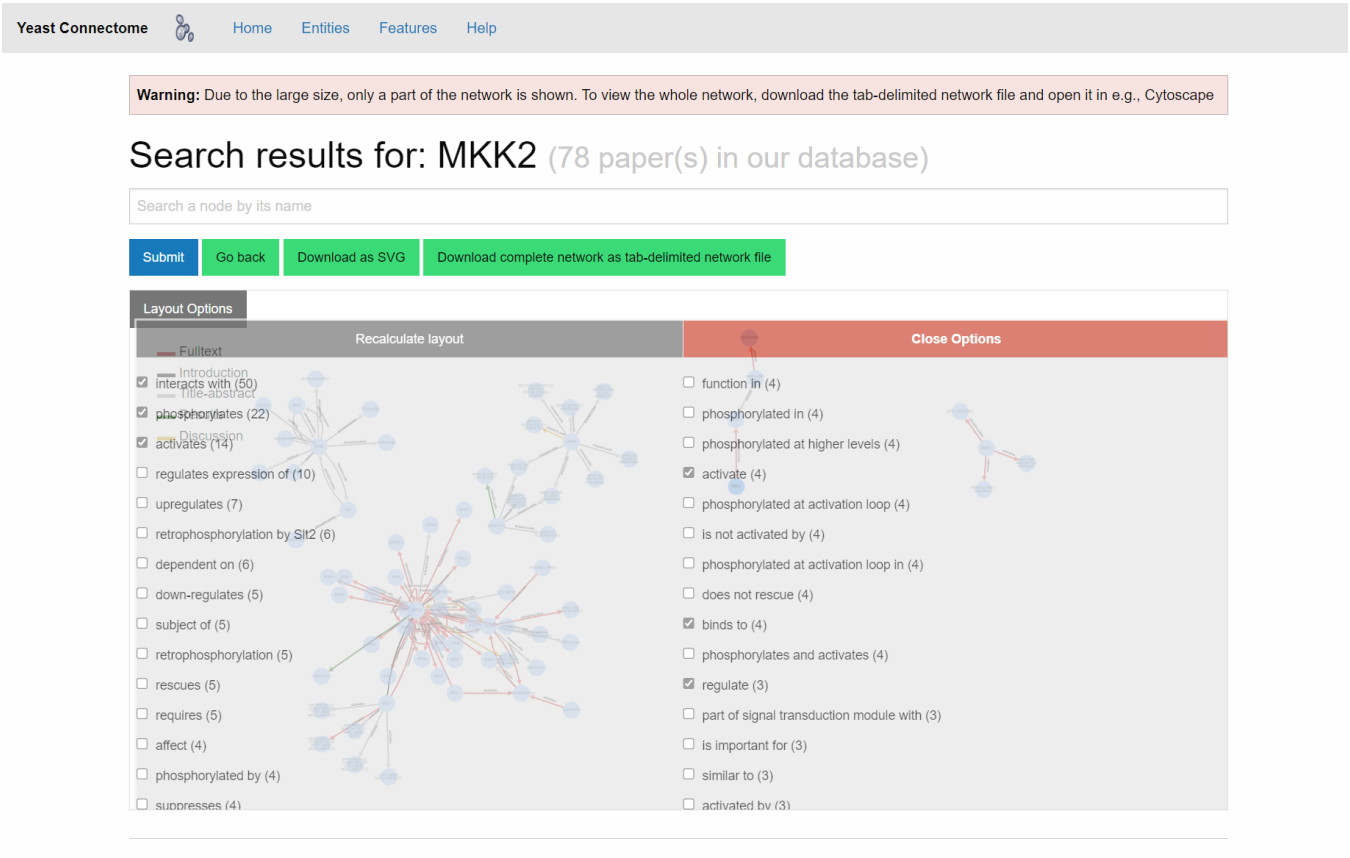

Figure 3.8: Yeast connectome - Layout options

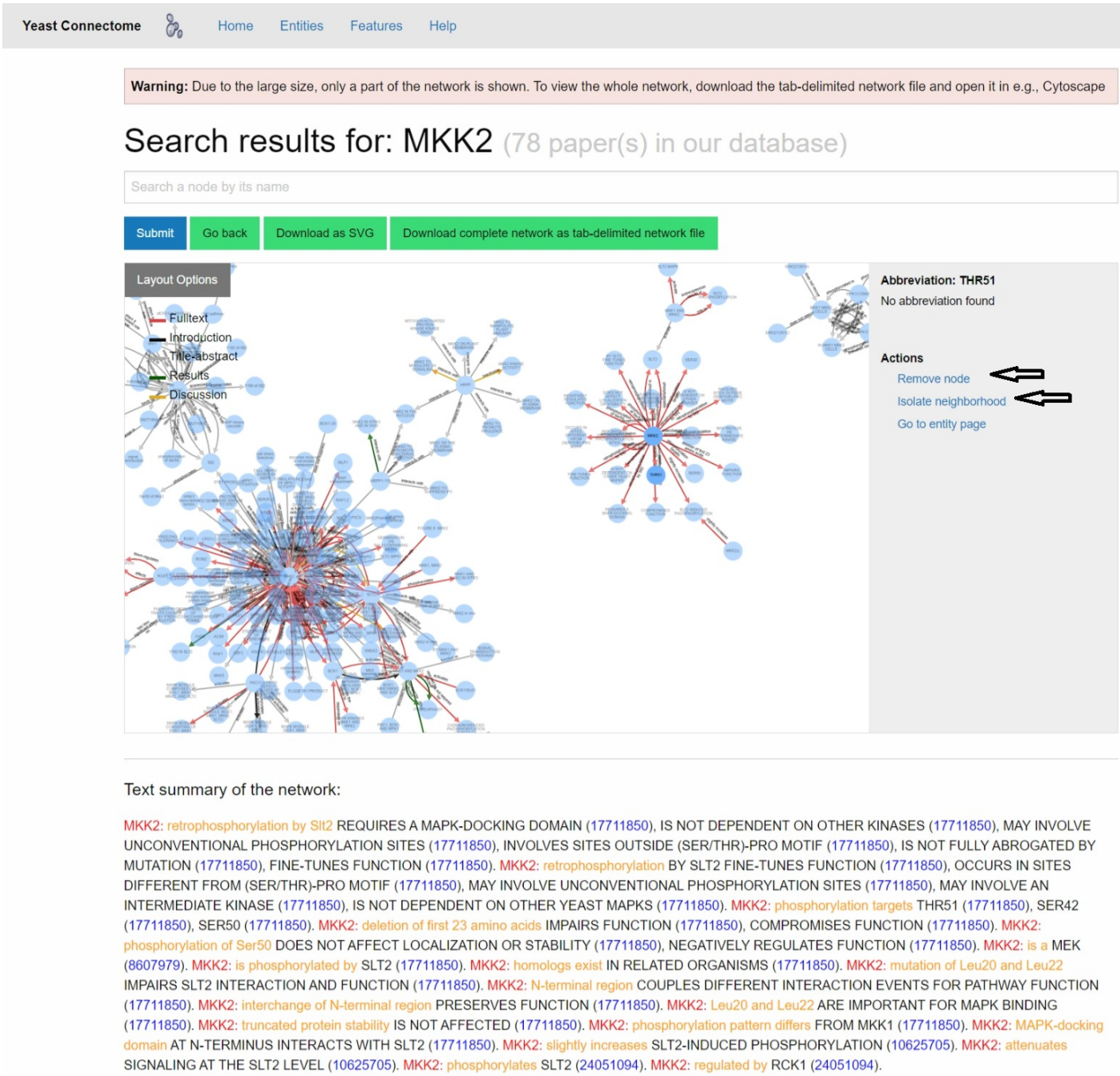

Figure 3.9: Yeast connectome - Node interactions

3.5 Comparative Analysis of Connectome and BioGRID Networks

Our comparative analysis focuses on the coverage and accuracy of the gene regulatory network (GRN). We paired data from BioGRID's protein-protein interaction (PPI) network with data from YeastConnectome. This comparison shows partial overlap; Exactly 3,848 of the 237,415 "interact with" edges in YeastConnectome coincide with edges in BioGRID (Figure 3.10). Notably, YeastConnectome identified 233,567 additional interaction edges not present in BioGRID, highlighting its comprehensive detection capabilities. Additionally, among the overlapping edges, 1,607 had distinct interaction types, such as "interact with", "phosphorylates" and "dephosphorylate". This highlights YeastConnectome's ability to capture a broader range of interactions (Figure 3.11).
We compared connectome's protein - protein interaction (PPI) edges with BioGRID and found an overlap of 3,784 edges out of total 69,160 PPI edges found in yeast connectome. Almost 65,376 protein- protein interaction were newly found in connectome (Figure 3.12).

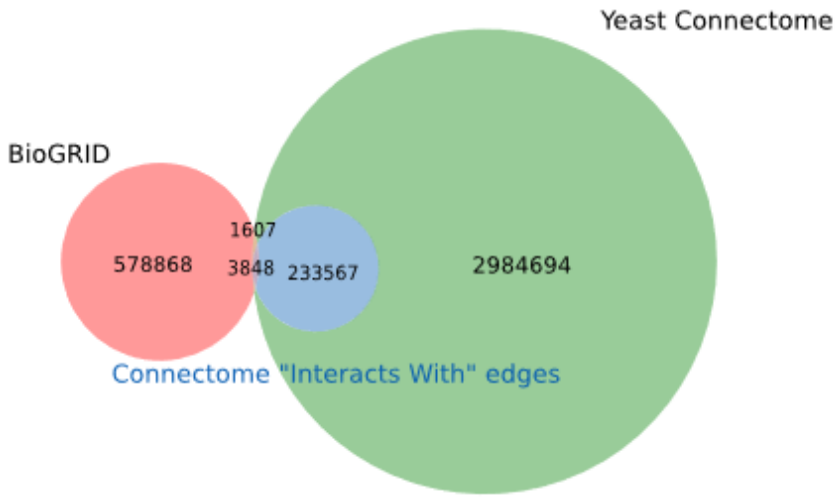

Figure 3.10: Venn diagram showing the intersection of Yeast Connectome's "interacts with" edges and BioGRID's protein-protein interaction networks.

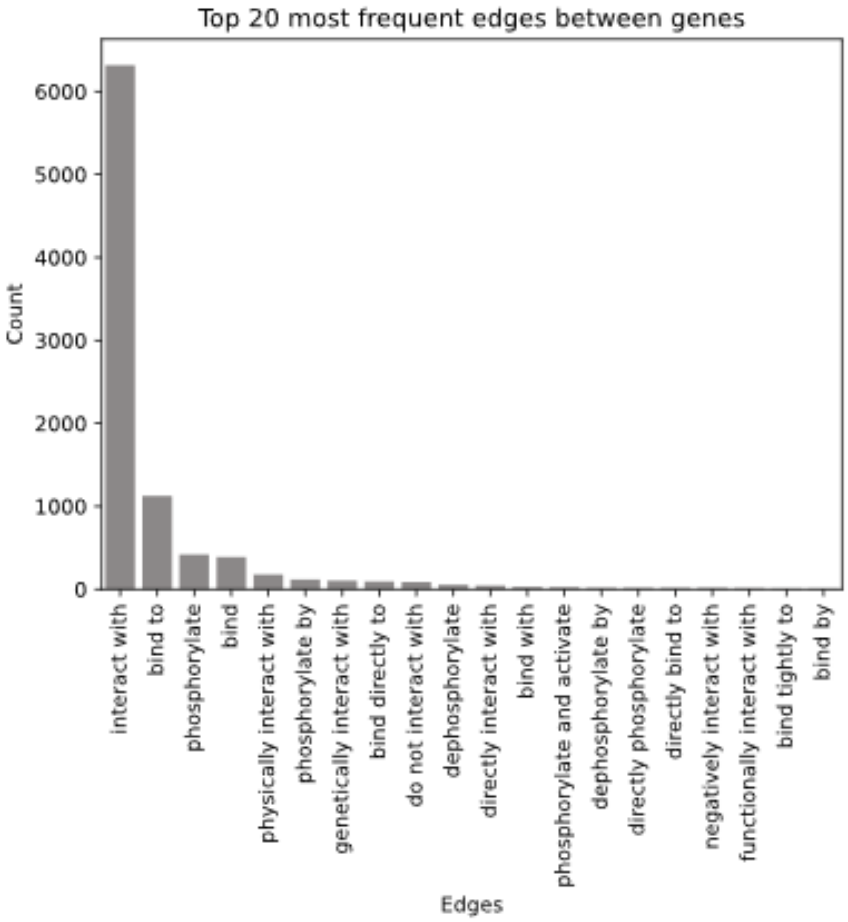

Figure 3.11: Top 20 edges found in overlap between Yeast connectome and BioGRID.

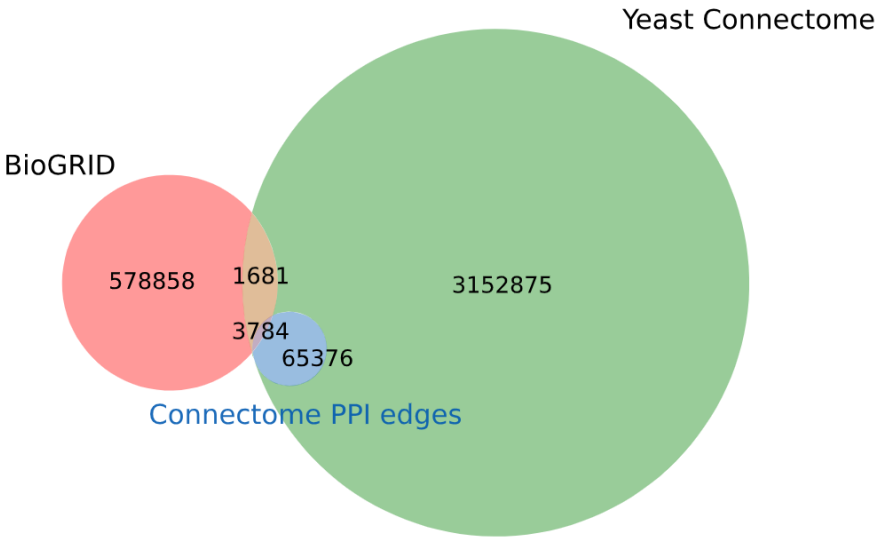

Figure 3.12: Venn diagram showing the intersection of Yeast Connectome's protein-protein interaction networks and BioGRID's protein-protein interaction networks.

To evaluate the accuracy of edges processed with GPT, we performed a manual evaluation of 100 random edges. Our evaluation shows that 89 out of 100 edges are correct (Figure 3.13), validating Connectome's reliability as an accurate and valuable addition to BioGRID.

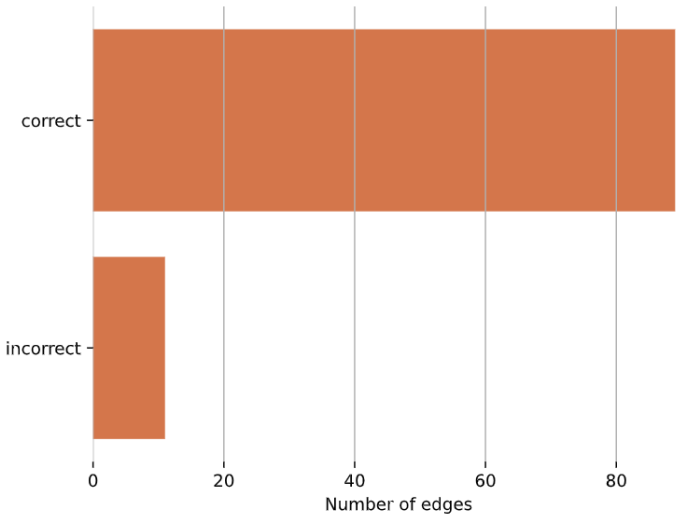

Figure 3.13: Number of correct edges out of 100 randomly selected relationships from yeast connectome.

3.6 Utilization of YeastConnectome in Gene Regulatory Network Analysis.

The Yeast Connectome is a powerful resource that consolidates vast amounts of information from research abstracts, providing a comprehensive tool for the yeast research community. Here, we showcase the utility of the Yeast Connectome in studying gene regulatory networks, protein complexes, metabolic pathways, and stress responses.

3.6.1 Example 1 –"HSP104"

Hsp104, a pivotal protein disaggregase in S. cerevisiae, plays an instrumental role in conferring thermotolerance. It achieves this by working with Hsp40 (Ydj1) and Hsp70 (Ssa1) [30] to promote the disassembly, resolubilization, and subsequent refolding of aggregated proteins following stress conditions. The Yeast Connectome reveals the extensive reach of Hsp104. When "HSP104" is queried, the Connectome maps out a network sourced from a robust compilation of 793 papers. Refining the search to "interacts with" within the "Layout Options" unveils a more focused network from 136 papers, promi-

nently featuring interactions with Hsp40 (Ydj1) and Hsp70 (Ssa1), denoted as “HSP70/40 PAIR”, “HSP70/40”, “HSP SYSTEM”, “YDJ1”, and “SSA1” (Figure 3.14).

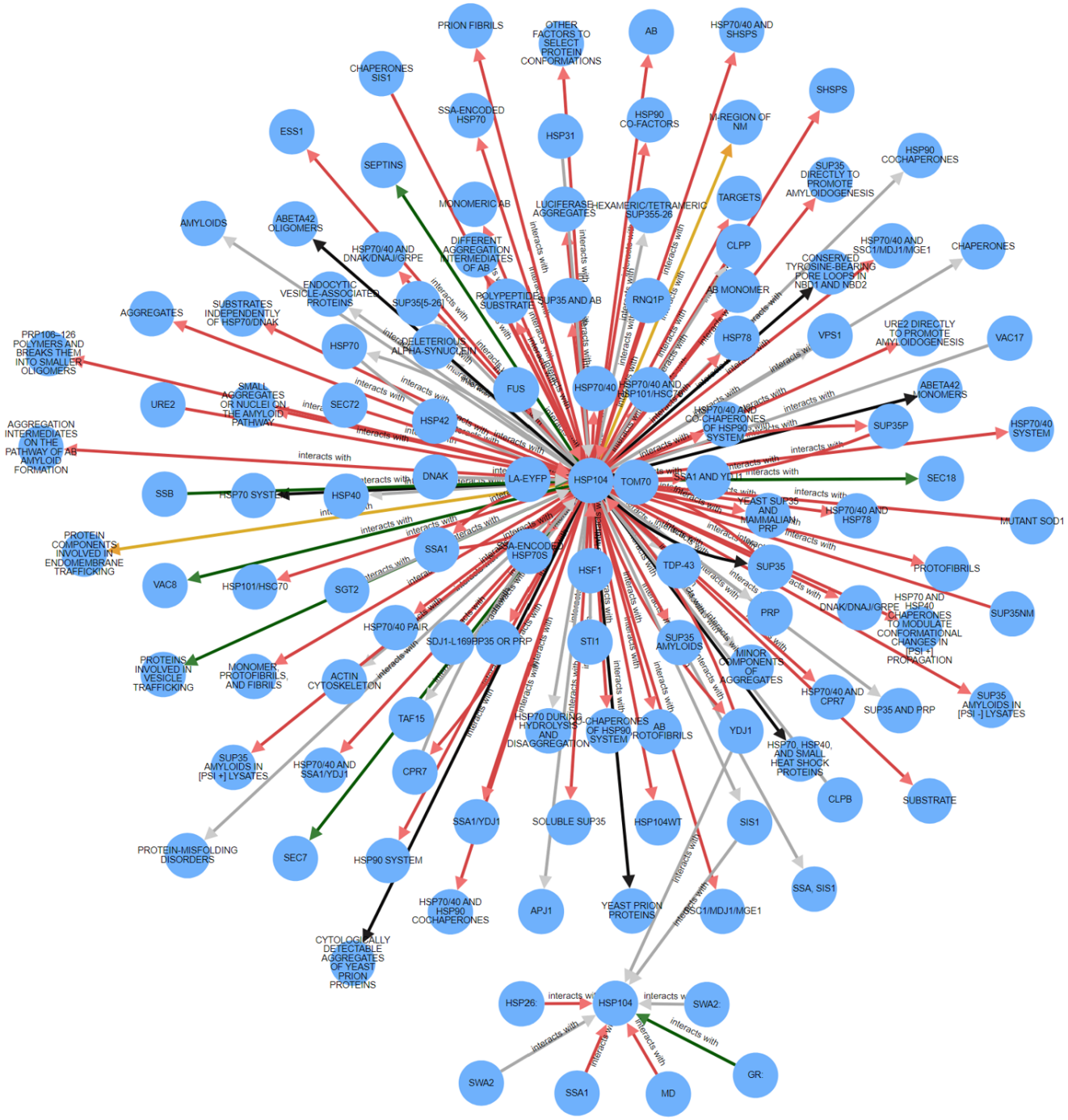

Figure 3.14: HSP104 interaction network.

Notably, the Connectome also highlights Hsp104’s interaction with various protein aggregate model substrates, such as “SUP35”, “LUCIFERASE AGGREGATES”, “ABETA42 MONOMERS”, “SUP35 AND AB”, and “PRION FIBRILS”. This indicates Hsp104’s broad role in protein homeostasis and stress response. While these interactions are

well-established, the Yeast Connectome offers an expedited overview of the diverse contexts and conditions under which Sup35, among others, was identified as an interactor. For instance, detailed interactions such as "HSP104 interacts with HEXAMERIC/TETRAMERIC SUP355-26" [31] and "SOLUBLE SUP35" [32] are readily accessible through one-click links to the publications, underscoring the tool's utility in providing a swift and comprehensive overview of protein interactions.

3.6.2 Example 2 – "ATG8"

ATG8 (LC3), a ubiquitin-like protein pivotal to the formation of cytoplasm-to-vacuole transport vesicles and autophagosomes, is a key player in the autophagy pathway [33] [34]. The Yeast Connectome, a comprehensive resource, provides detailed insights into ATG8's interactions, as sourced from an extensive collection of 682 papers. Specifically, 41 papers within the Connectome detail ATG8 "binds to" interactions, shedding light on its dynamic associations in the cellular context (Figure 3.15). Notably, ATG8, in conjunction with ATF4, mediates the delivery of vesicles and autophagosomes to the vacuole via the microtubule cytoskeleton. Additionally, the connectomes spotlight the indispensable roles of ATG3 and ATF7, both of which are crucial, post-ATG4, for conjugating ATG8 with phosphatidylethanolamine (PE).

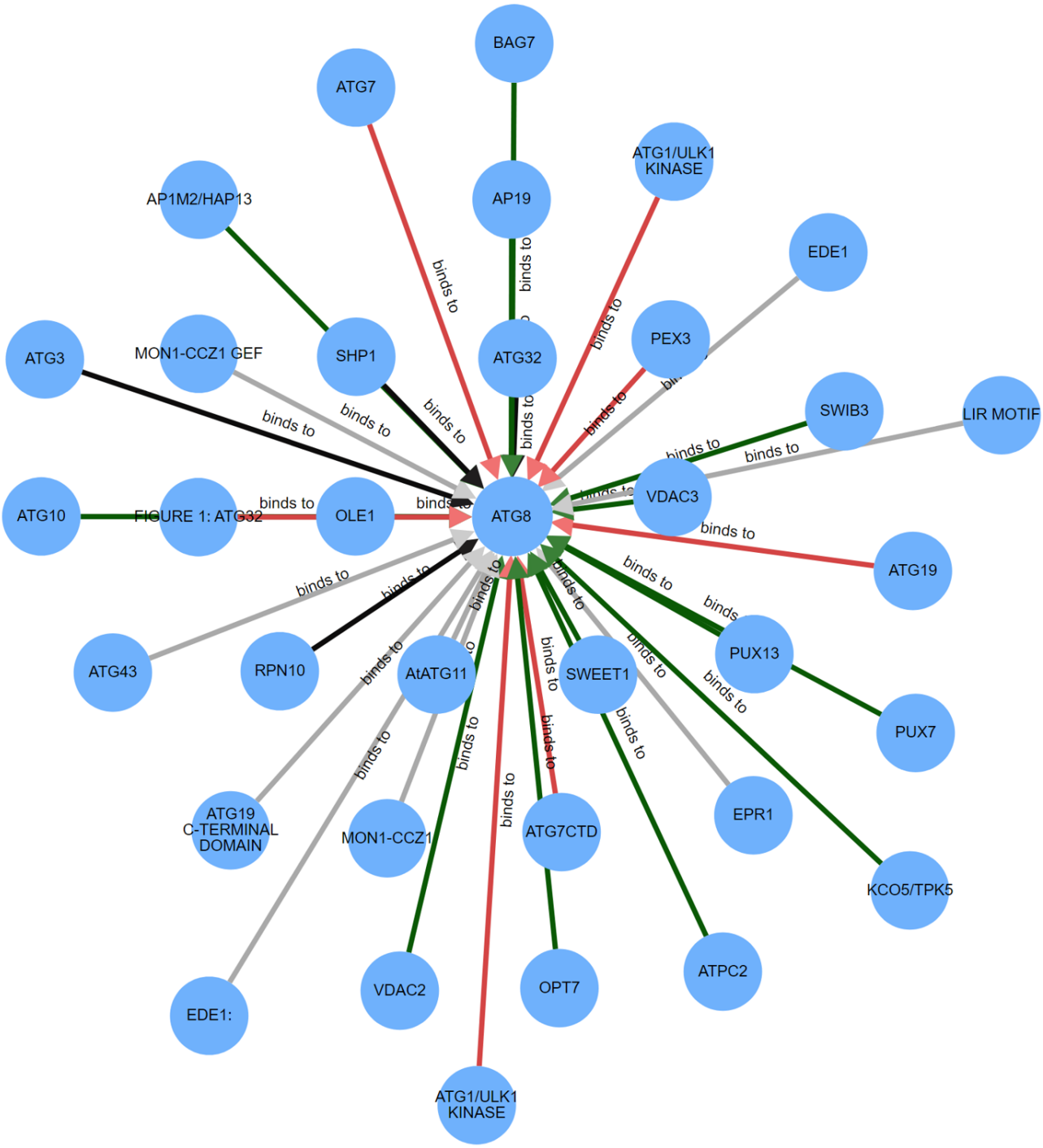

Figure 3.15: ATG8 interaction network.

Diverging from conventional protein-protein interaction databases, the Yeast Connectome offers a granular view of ATG8's interactions, accessible through direct links to the pertinent literature. For instance, it delineates that ATG8 "binds to" the "C-TERMINAL FLEXIBLE TAIL OF ATG7" [35] and the "HR OF ATG3" [36]. Moreover, the Connectome uncovers that ATG8 also "binds to" "GROWING MEMBRANE" [37], adding another layer to our understanding of its multifaceted role in cellular processes. Furthermore, the ATG8 connectome encapsulates valuable information and provides di-

rect links to publications associated with "is involved in" interactions, summarizing findings from 8 pivotal publications. This subset of the Connectome reveals ATG8's involvement in diverse biological processes, including "MACRONUCLEOPHAGY" [38], "HEMIFUSION OF LIPOSOMES" [34], "MEMBRANE TETHERING" [34], and "AGGREPHAGY" [39].

3.7 Gene ontology (GO Term) prediction using GPT embeddings

3.7.1 What are embeddings?

Word embeddings are numeric vector representations of words or phrases, capturing semantic and syntactic information. In the field of natural language processing (NLP), the quest to understand semantic closeness between text entities has led to the development of these embeddings. An example embeddings with 7 dimensions are shown in the Figure 3.16.

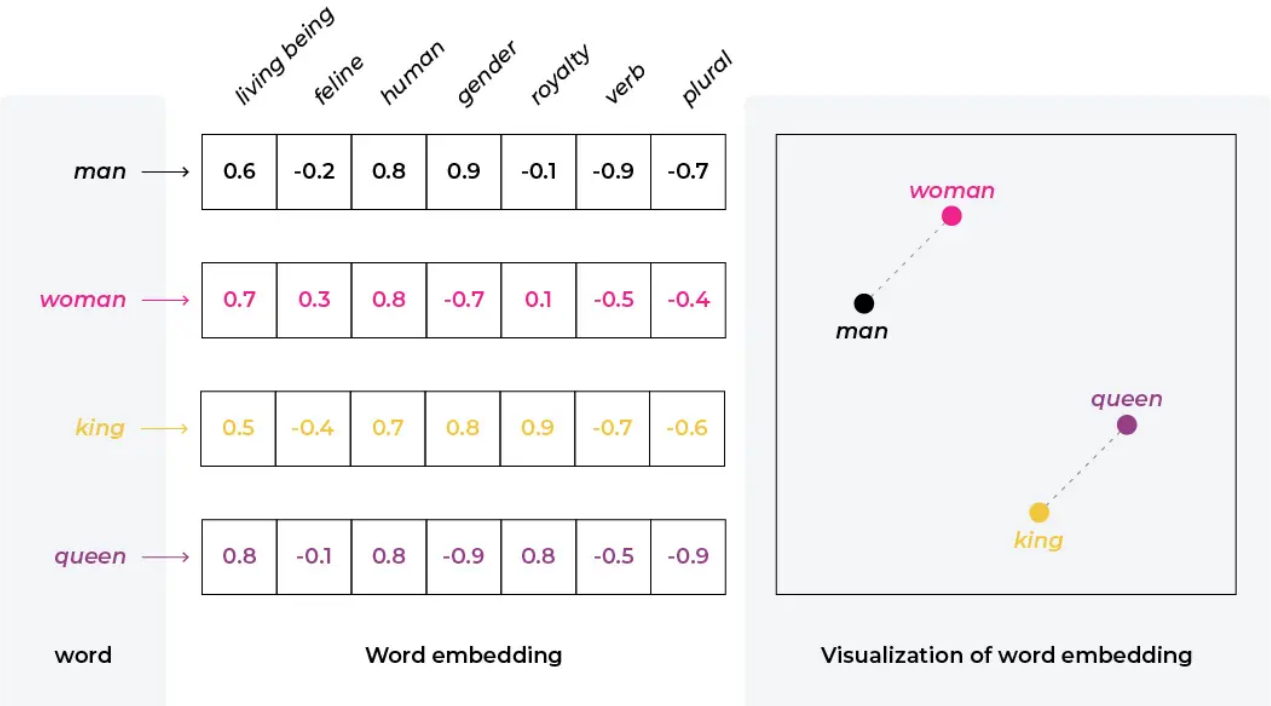

Figure 3.16: Figure showing example embeddings for words man, woman, king and queen. source: Google.

3.7.2 Cosine similarity

Cosine similarity is a measure used to measure the similarity between two vectors, regardless of their magnitude. In the context of GPT embedding, it quantifies the semantic distance between two text entities by calculating the cosine of the angle between their vector representations. The resulting value ranges from -1 to 1, where 1 represents identical direction, i.e., maximum semantic similarity, and -1 represents the exact opposite direction, indicating no similarity. In the example (Figure 3.16), we can observe that king and queen are closer to each other than king and women. Hence, we can get the semantic similarity or closeness between two sentences using embeddings.

3.7.3 GO Term prediction and evaluation

In the above sections, we discussed about embeddings. Here, we discuss a plausible method to predict GO terms using functional information from yeast connectomes. At first, we filtered the edges that contain the gene name so that the GO term can be predicted for that particular gene. After basic filtering, we are left with 1226785 edges. We then created embeddings for each edge using the "text-embedding-ada-002" model [40]. The list of all existing GO terms for yeast genes is downloaded from the UniProt database [20] and created embeddings for the same. Further, we compared the closeness of both embeddings based on the cosine similarity and assigned the GO term with the least cosine distance for each edge containing the gene name.

We evaluated the predicted GO terms with gold standard data from yeastmine [19] and found a match for 112440 out of a total of 1226785 GO predictions. Although, this is insignificant. However, the gold standard data may not contain GO terms for only some genes. We are working on improving the method and its integration with yeast connectome.

Chapter 4

Discussion

In this work, we have introduced a pioneering method to leveraging advanced natural language processing (NLP) techniques, specifically the OpenAI GPT-3.5 model, for extracting functional gene information from a large corpus of scientific literature. We have created the YeastConnectome by processing over 90879 research abstracts at a moderate cost (1,000 USD) within two weeks. The web platform is made available providing insights into 3,853,947 relationships involving genes, molecules, compartments, stresses, organs, and other yeast entities. The connectome, with over million relationships, is a tool for researchers exploring the complex interactions within Yeast.

The Connectome database with its 3.8 million relationships, is a powerful tool for researchers exploring the complex interactions within Yeast. Yeast Connectome's core component, KnowledgeNetwork, provides an intuitive visual interface for exploring these relationships. The ability to customize the network by filtering specific relationship types and isolating nodes provides interactivity and personalization to improve the user experience.

Evaluation of connectome accuracy and coverage compared to public repositories such as BioGRID demonstrates the effectiveness of our approach. Although overlap in interaction edges with BioGRID was observed, the connectome identified many additional interactions, including more specific interaction types such as "phosphorylates" and "inhibits". This suggests that the connectome has the potential to complement existing databases, expand our knowledge of gene interactions, and fill gaps in our understanding of yeast biology. However, relying on AI to extract data and identify relationships comes with challenges. Manual inspection of a subset of edges further confirmed the accuracy of the connectome with an edge accuracy of 89%. This high level of accuracy is essential to ensure the reliability of the extracted information and increase the confidence of re-

searchers using the connectome in their studies.

In conclusion, the yeast connectome illustrate the transformative impact of generative AI models on biological research. The YeastConnectome and its associated web platform provide a powerful tool for researchers to explore gene interactions in Yeast and advance our understanding of biological systems. Our method significantly expands the interaction data available by utilizing computational techniques to analyze full-text publications. However, this approach is restricted by the availability of literature, often locked behind paywalls. Additionally, the scalability of automated data extraction comes at the potential cost of precision compared to manual curation, which could impact the overall accuracy of the interaction data. We are actively addressing these limitations by refining our methodologies and seeking to improve access to recent scientific publications.

Chapter 5

Limitations and Future directions

5.1 Shortcomings of the approach and ways to mitigate

In the current field of natural language processing, pre-trained Generative Transformer (GPT) has shown impressive ability to understand and produce human-like texts. Our database's reliance on literature that is publicly available or accessible through institutional subscriptions means that some recent studies, especially those that are not Free access due to paywalls may not be integrated quickly. Therefore, there may be a delay in bringing the latest findings and research advances into our connected database. Furthermore, it is important to consider the inherent limitations of this technology, especially when applied to specialized areas such as scientific literature. A significant challenge lies in the model's difficulty in generating results that are both coherent and structured. The resulting knowledge graph may lack a structured representation, leading to reduced clarity. Additionally, GPT's tendency to misidentify entities and their interrelationships can lead to unreliable results, a problem that is observed when analyzing large texts.

We may mitigate these shortcomings by training the language model and fine-tuning it using scientific research articles. The another option is to replace GPT 3.5 model with GPT-4 model to process the text. Since, GPT-4 underwent an additional six months of training with human and AI feedback. It's accuracy is higher and can handle more complex tasks. For instance, in predicting outcomes of court cases, GPT-4 achieves a higher prediction accuracy rate of approximately 88% compared to GPT-3.5's 81% [41]. But, The accuracy comes with the cost. Use of GPT-3.5 is 20 times cheaper compared to GPT-4 [42].

5.2 Future directions

The future plans would be to create similar connectomes for several other model organisms like *c. elegans*, *Drosophila melanogaster*, *E. coli* etc . We can create similar knowledge graphs using biomedical literature that can be very useful for clinical researcher and clinicians. Our next goal is to fully automate the information extraction pipeline and continuously update the database with additional new information without any manual intervention.

Bibliography

[1] David Botstein and Gerald R Fink. "Yeast: An Experimental Organism for 21st Century Biology". In: *Genetics* 189.3 (Nov. 2011), pp. 695–704. ISSN: 1943-2631. DOI: 10.1534/genetics.111.130765. eprint: https://academic.oup.com/genetics/article-pdf/189/3/695/42138643/genetics0695.pdf. URL: https://doi.org/10.1534/genetics.111.130765.

[2] Preeti Dabas, Deepak Kumar, and Nimisha Sharma. "Yeast Genetics as a Powerful Tool to Study Human Diseases". In: *Yeast Diversity in Human Welfare*. Ed. by Tulasi Satyanarayana and Gotthard Kunze. Singapore: Springer Singapore, 2017, pp. 191–214. ISBN: 978-981-10-2621-8. DOI: 10.1007/978-981-10-2621-8_8. URL: https://doi.org/10.1007/978-981-10-2621-8_8.

[3] In: (). URL: https://science.jrank.org/pages/7440/Yeast-Biotechnology-yeast.html.

[4] Erwin Tantoso et al. "Did the early full genome sequencing of yeast boost gene function discovery?" In: *Biology Direct* 18 (2023). DOI: 10.1186/s13062-023-00403-8. URL: https://doi.org/10.1186/s13062-023-00403-8.

[5] In: (). URL: https://pubmed.ncbi.nlm.nih.gov/?term=Saccharomyces+Cerevisiae.

[6] Joseph Kamtchum-Tatuene and Joseline Zafack. *Keeping Up With the Medical Literature: Why, How, and When?* Nov. 2021. DOI: 10.1161/strokeaha.121.036141. URL: https://doi.org/10.1161/STROKEAHA.121.036141.

[7] Rose Oughtred et al. "The biogrid database: A comprehensive biomedical resource of curated protein, genetic, and chemical interactions". In: *Protein Science* 30.1 (Nov. 2020), pp. 187–200. DOI: 10.1002/pro.3978.

[8] Rose Oughtred et al. "The BioGRID Interaction Database: 2019 update". In: *Nucleic Acids Research* 47.D1 (Nov. 2018). DOI: 10.1093/nar/gky1079.

[9] Marija Milacic et al. "The reactome pathway knowledgebase 2024". In: *Nucleic Acids Research* 52.D1 (Nov. 2023). DOI: 10.1093/nar/gkad1025.

[10] Diksha Khurana et al. "Natural language processing: state of the art, current trends and challenges". In: *Multimedia Tools and Applications* 82.3 (July 2022), pp. 3713–3744. DOI: 10.1007/s11042-022-13428-4. URL: http://dx.doi.org/10.1007/s11042-022-13428-4.

[11] Rushdi Shams. In: (). URL: https://doi.org/10.48550/arXiv.1409.7612.

[12] Qijie Chen et al. "An extensive benchmark study on biomedical text generation and mining with chatgpt". In: *Bioinformatics* 39.9 (Sept. 2023). DOI: 10.1093/bioinformatics/btad557.

[13] Zhi Hong et al. "SciNER: Extracting Named Entities from Scientific Literature". In: *Computational Science – ICCS 2020*. Ed. by Valeria V. Krzhizhanovskaya et al. Cham: Springer International Publishing, 2020, pp. 308–321. ISBN: 978-3-030-50417-5.

[14] Shaina Raza and Brian Schwartz. "Entity and relation extraction from clinical case reports of COVID-19: A natural language processing approach". In: *BMC Medical Informatics and Decision Making* 23.1 (Jan. 2023). DOI: 10.1186/s12911-023-02117-3.

[15] Humza Naveed et al. *A Comprehensive Overview of Large Language Models*. 2024. arXiv: 2307.06435 [cs.CL].

[16] Vaishnav Manoj. *Demystifying language models: An overview of LLMS*. Sept. 2023. URL: https://medium.com/data-science-community-srm/demystifying-language-models-an-overview-of-llms-5cbc7600c3f8.

[17] Enkelejda Kasneci et al. "ChatGPT for good? On opportunities and challenges of large language models for education". In: *Learning and Individual Differences* 103 (2023), p. 102274. ISSN: 1041-6080. DOI: https://doi.org/10.1016/j.lindif.2023.102274. URL: https://www.sciencedirect.com/science/article/pii/S1041608023000195.

[18] Dat Duong and Benjamin D. Solomon. "Analysis of large-language model versus Human Performance for Genetics questions". In: *European Journal of Human Genetics* (May 2023). DOI: 10.1038/s41431-023-01396-8.

[19] Rama Balakrishnan et al. "YeastMine—an integrated data warehouse for Saccharomyces cerevisiae data as a multipurpose tool-kit". In: *Database* 2012 (Jan. 2012). DOI: 10.1093/database/bar062. URL: http://dx.doi.org/10.1093/database/bar062.

[20] R. Apweiler. "UniProt: the Universal Protein knowledgebase". In: *Nucleic Acids Research* 32.90001 (Jan. 2004), pp. 115D–119. DOI: 10.1093/nar/gkh131. URL: http://dx.doi.org/10.1093/nar/gkh131.

[21] Edith Motschall and Yngve Falck-Ytter. "Searching the MEDLINE Literature Database through PubMed: A Short Guide". In: *Oncology Research and Treatment* 28.10 (2005), pp. 517–522. DOI: 10.1159/000087186. URL: http://dx.doi.org/10.1159/000087186.

[22] Guido Van Rossum and Fred L Drake Jr. *Python reference manual*. Centrum voor Wiskunde en Informatica Amsterdam, 1995.

[23] URL: https://platform.openai.com/docs/models/gpt-3-5-turbo.

[24] Matthew Honnibal and Ines Montani. "spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing". To appear. 2017.

[25] URL: https://platform.openai.com/docs/guides/fine-tuning.

[26] URL: https://platform.openai.com/docs/guides/function-calling.

[27] Kevin Fo et al. "PlantConnectome: knowledge networks encompassing ¿100,000 plant article abstracts". In: (July 2023). DOI: 10.1101/2023.07.11.548541. URL: http://dx.doi.org/10.1101/2023.07.11.548541.

[28] Aug. 2023. URL: https://en.wikipedia.org/wiki/Gal4_transcription_factor.

[29] URL: https://www.yeastgenome.org/locus/S000000897.

[30] John R Glover and Susan Lindquist. "Hsp104, Hsp70, and Hsp40". In: *Cell* 94.1 (July 1998), pp. 73–82. DOI: 10.1016/s0092-8674(00)81223-4. URL: http://dx.doi.org/10.1016/s0092-8674(00)81223-4.

[31] Saravanakumar Narayanan et al. "Importance of low-oligomeric-weight species for prion propagation in the yeast prion system Sup35/Hsp104". In: *Proceedings of the National Academy of Sciences* 100.16 (July 2003), pp. 9286–9291. DOI: 10.1073/pnas.1233535100. URL: http://dx.doi.org/10.1073/pnas.1233535100.

[32] Yuji Inoue et al. "Hsp104 Binds to Yeast Sup35 Prion Fiber but Needs Other Factor(s) to Sever It". In: *Journal of Biological Chemistry* 279.50 (Dec. 2004), pp. 52319–52323. DOI: 10.1074/jbc.m408159200. URL: http://dx.doi.org/10.1074/jbc.m408159200.

[33] Yoshinobu Ichimura et al. "A ubiquitin-like system mediates protein lipidation". In: *Nature* 408.6811 (Nov. 2000), pp. 488–492. DOI: 10.1038/35044114. URL: http://dx.doi.org/10.1038/35044114.

[34] Hitoshi Nakatogawa, Yoshinobu Ichimura, and Yoshinori Ohsumi. "Atg8, a Ubiquitin-like Protein Required for Autophagosome Formation, Mediates Membrane Tethering and Hemifusion". In: *Cell* 130.1 (July 2007), pp. 165–178. DOI: 10.1016/j.cell.2007.05.021. URL: http://dx.doi.org/10.1016/j.cell.2007.05.021.

[35] Masaya Yamaguchi et al. “Atg7 Activates an Autophagy-Essential Ubiquitin-like Protein Atg8 through Multi-Step Recognition”. In: *Journal of Molecular Biology* 430.3 (Feb. 2018), pp. 249–257. DOI: 10.1016/j.jmb.2017.12.002. URL: http://dx.doi.org/10.1016/j.jmb.2017.12.002.

[36] Yuya Yamada et al. “The Crystal Structure of Atg3, an Autophagy-related Ubiquitin Carrier Protein (E2) Enzyme that Mediates Atg8 Lipidation”. In: *Journal of Biological Chemistry* 282.11 (Mar. 2007), pp. 8036–8043. DOI: 10.1074/jbc.m611473200. URL: http://dx.doi.org/10.1074/jbc.m611473200.

[37] Usha Nair and Daniel J. Klionsky. “Molecular Mechanisms and Regulation of Specific and Nonspecific Autophagy Pathways in Yeast”. In: *Journal of Biological Chemistry* 280.51 (Dec. 2005), pp. 41785–41788. DOI: 10.1074/jbc.r500016200. URL: http://dx.doi.org/10.1074/jbc.r500016200.

[38] Florian B. Otto and Michael Thumm. “Mechanistic dissection of macro- and micronucleophagy”. In: *Autophagy* 17.3 (Feb. 2020), pp. 626–639. DOI: 10.1080/15548627.2020.1725402. URL: http://dx.doi.org/10.1080/15548627.2020.1725402.

[39] Stephanie B.M. Miller, Axel Mogk, and Bernd Bukau. “Spatially Organized Aggregation of Misfolded Proteins as Cellular Stress Defense Strategy”. In: *Journal of Molecular Biology* 427.7 (Apr. 2015), pp. 1564–1574. DOI: 10.1016/j.jmb.2015.02.006. URL: http://dx.doi.org/10.1016/j.jmb.2015.02.006.

[40] URL: https://platform.openai.com/docs/guides/embeddings/embedding-models.

[41] June 2023. URL: https://ecoagi.ai/articles/compare-gpt-4-gpt-3.

[42] URL: https://openai.com/pricing.