Universiteit Leiden
The Netherlands

Opleiding Informatica

Comparing Neurodegenerative Disorders:

Discovering patterns through protein-protein interaction network analysis

Emma C.T.D. Bouwman

Supervisors:
Katy Wolstencroft & Lu Cao

BACHELOR THESIS

Leiden Institute of Advanced Computer Science (LIACS)
www.liacs.leidenuniv.nl

25\09\2024

Abstract

Alzheimer's disease (AD), Huntington's disease (HD), and spinocerebellar ataxia (SCA), especially SCA1 and SCA3, are characterized by neuronal degeneration. This thesis aimed to identify overlapping patterns and common mechanisms among these disorders using protein-protein interaction networks (PPINs). PPINs for AD, HD, and SCA were constructed using data from KEGG, STRING, and WikiPathways. Network analysis, clustering, and functional enrichment analysis revealed shared biological processes, including responses to misfolded proteins and the endoplasmic reticulum-associated degradation (ERAD) pathway, across all three disorders. A consistent cluster was found in three different PPINs. So, despite their distinct genetic origins and clinical differences, these diseases share common mechanisms. This thesis demonstrates that PPINs can uncover patterns across different diseases that may help identify shared drug targets.

Contents

1 Introduction **1**
- 1.1 Neurodegenerative Diseases . . . 1
  - 1.1.1 Alzheimer's disease (AD) . . . 1
  - 1.1.2 Huntington's disease (HD) . . . 2
  - 1.1.3 Spinocerebellar ataxia (SCA) . . . 2
- 1.2 Protein-Protein Interaction Networks (PPINs) . . . 3
- 1.3 Ontologies . . . 4
- 1.4 Research Question . . . 4
- 1.5 Thesis overview . . . 4

2 Definitions **5**

3 Related Work **6**
- 3.1 Alzheimer's and Huntingtons's disease C-PPINs . . . 6
- 3.2 Spinocerebellar ataxia disease C-PPINs . . . 6
- 3.3 Overlap in neurodegenerative diseases . . . 7

4 Methods **7**
- 4.1 Tools and Resources . . . 9
- 4.2 Data Sources . . . 9
- 4.3 Pre-processing . . . 10
- 4.4 Workflow Setup . . . 11
- 4.5 Merging and Filtering . . . 11
- 4.6 Recreation of Alzheimer's and Huntington's C-PPINs . . . 11
- 4.7 Creation of the Spinocerebellar Ataxia Consensus networks . . . 12
  - 4.7.1 SCA 1 & SCA 3 . . . 12
- 4.8 Merging and comparing the C-PPINs . . . 13
  - 4.8.1 Network Analysis . . . 14
  - 4.8.2 Clustering Analysis . . . 15
  - 4.8.3 Enrichment Analysis . . . 15

5 Results **16**
- 5.1 Network Analysis . . . 16
- 5.2 Clustering Analysis . . . 20
- 5.3 Enrichment Analysis . . . 23

6 Discussion **25**

7 Conclusions and Further Research **26**

Acknowledgements **27**

References **27**

A Figures **33**
- A.1 Individual C-PPINs . . . 33
- A.2 Merged C-PPINs . . . 36
- A.3 Clusters . . . 38

1 Introduction

Finding a solution to neurodegenerative diseases has proven to be a significant challenge for humanity. These diseases result from the complex interaction between environmental and genetic factors during the aging process [SMHV+16].

Neurodegenerative disorders are characterized by their persistent and progressive nature. A crucial aspect of these diseases is the specific and symmetric degeneration of neurons within the motor, sensory, or cognitive systems [Mar99]. Examples of neurodegenerative diseases include Alzheimer's disease (AD), Huntington's disease (HD), and spinocerebellar ataxia (SCA)[ALC05][CCJS22].

These diseases exhibit considerable variability in their characteristics and underlying mechanisms. For instance, AD arises from a combination of genetic factors and environmental influences, such as lifestyle and aging [MC09]. In contrast, HD is caused solely by a specific genetic mutation in the HTT gene, making it a hereditary condition [BZMN11]. Similarly, SCA is primarily caused by genetic mutations, with different subtypes showing variability in the severity and progression of symptoms depending on the specific genetic mutation involved [Mü21].

Aditionally, there is a significant contrast in the prevalence of these diseases. AD is one of the most prevalent neurodegenerative diseases worldwide [PMG22]. However, despite the significant variation in HD prevalence across different geographical regions, being more common in the Western world compared to Asia or Africa [RWW+16], both HD and SCA are classified as rare neurodegenerative diseases [Roo10][vPIAvdW21].

Studying these diseases together is motivated by the recognition that, despite their differences, shared mechanisms might underlie their pathogenesis. Examining these diseases collectively can uncover common patterns and interconnections that may not be apparent when each disease is studied in isolation [ALC05]. This comparative approach has the potential to reveal new insights into the molecular processes common to these disorders, thereby enhancing understanding and paving the way for more effective treatment strategies.

1.1 Neurodegenerative Diseases

This subsection provides a detailed overview of three neurodegenerative diseases: Alzheimer's disease (AD), Huntington's disease (HD), and spinocerebellar ataxia (SCA).

1.1.1 Alzheimer's disease (AD)

AD is a progressive and irreversible neurodegenerative disorder, primarily known for causing memory loss, cognitive decline, and behavioral changes [BYBT10]. It is the most common form of dementia, with two main subtypes: early-onset AD (EOAD), diagnosed before age 65, and late-onset AD (LOAD), diagnosed after 65. Due to its prevalence LOAD is the focus of most research[BYBT10].

AD is mainly characterized by two abnormal processes in the brain: the formation of amyloid plaques and neurofibrillary tangles, both of which contribute to the death of brain cells[KK18]. The first protein that plays a part in AD is Amyloid Precursor Protein (APP)[SSK+21]. APP, a type I transmembrane protein [PBM+06] that normally plays a role in communication between

nerve cells. Under normal conditions, APP is broken down in a controlled manner. However, in AD, it is incorrectly cleaved by an enzyme called ß-secretase[ZTZX11]. This results in the production of amyloid-ß (Aß), a sticky protein that accumulates to form plaques and disrupt communication between neurons[LXL+21].

The second key protein is tau. Tau is a protein found in the structure of nerve cells, specifically in microtubules. In AD, tau becomes abnormally modified through a process called phosphorylation[BB98]. This leads to the formation of twisted bundles known as neurofibrillary tangles. These tangles disrupt the normal functioning of neurons and eventually contribute to their death[BB98]. Tau phosphorylation is likely influenced by Aß plaques, but the exact mechanism behind this interaction is not yet understood[SSK+21].

Genetic factors, including mutations in the APP gene and the presenilin genes (PSEN1 and PSEN2), can lead to Alzheimer's disease at a younger age[HHB+21]. For older people another genetic factor is important, namely the APOE $\epsilon 4$ allele[SSK+21].

1.1.2 Huntington's disease (HD)

Huntington's disease (HD) is a rare autosomal-dominant neurodegenerative disorder that affects the central nervous system. It is characterized by uncontrollable movements, behavioral changes, and progressive cognitive decline, including dementia [Roo10].

HD is caused by a mutation in the huntingtin (HTT) gene located on chromosome 4. This mutation leads to an abnormal expansion of the CAG sequence, resulting in the production of an excess amount of the amino acid glutamine[Roo10][VD98]. In individuals without HD, the HTT gene typically contains fewer than 35 CAG repeats. Those with more than 40 repeats are certain to develop the disease, while individuals with 35 to 40 repeats fall into a "reduced penetrance" range, meaning they may or may not show symptoms, and if symptoms do appear, they tend to be milder or onset later in life[Roo10].

Although the exact cellular function of the HTT protein remains unclear[RT11], current treatments aim to reduce the amount of mutant HTT using methods such as CRISPR/Cas9 and targeting HTT mRNA. However, there is still no cure for the disease [TGL19].

1.1.3 Spinocerebellar ataxia (SCA)

SCAs are genetic disorders that cause progressive degeneration of the cerebellum, leading to difficulties with balance and coordination. Although SCAs encompass a range of conditions, this section will only concentrate on SCA1 and SCA3.

Spinocerebellar ataxia type 1 (SCA1) SCA1 is an autosomal-dominant neurodegenerative disorder characterized by symptoms such as difficulty with balance and coordination, problems with eye movements, and occasionally peripheral nerve issues [MDDGG]. Similar to HD, SCA1 is caused by an expansion of CAG repeats, but in this case, the repeats occur in the ATXN1 gene instead of the HTT gene. This mutation results in an abnormally long polyglutamine (polyQ) chain in the ataxin-1 protein. The expanded polyQ chain causes toxic effects, leading to dysfunction and degeneration of nerve cells, particularly in the cerebellum, brainstem, and spinocerebellar tracts [CCJS22].

Neurodegeneration in SCA1 primarily affects Purkinje cells in the cerebellum. Although different SCA types vary genetically, they all share this common feature of Purkinje cell degeneration [HV19]. Despite these findings, the precise mechanism behind SCA1 development remains not fully understood [CCJS22].

Spinocerebellar ataxia type 3 (SCA3) SCA3, also known as Machado-Joseph disease, is the most common type of spinocerebellar ataxia [ETRM14]. Like Huntington's disease (HD), SCA3 is an autosomal-dominant neurodegenerative disorder caused by an expanded CAG repeat in the ATXN3 gene, which encodes the ataxin-3 protein. Healthy individuals have fewer than 44 CAG repeats, while those with SCA3 have between 52 and 86 repeats [ETRM14].
Clinically, SCA3 is characterized by progressive ataxia, which affects balance, gait, and speech. Additional symptoms may include muscle rigidity, dysarthria (difficulty with speech), ophthalmoplegia (abnormal eye movements), and dystonia. As with other polyglutamine (polyQ) diseases, there is currently no cure for SCA3, and treatment focuses on managing symptoms [ETRM14][MMA+14].

1.2 Protein-Protein Interaction Networks (PPINs)

As introduced earlier in this chapter, the motivation for studying diseases together lies in the potential for shared underlying mechanisms despite their differences. By analyzing diseases collectively, it is possible to uncover common patterns and interconnections that might not be evident when each disease is studied in isolation [ALC05]. This approach is especially valuable for rare diseases, where the limited number of cases makes isolated study challenging.

PPINs represent the complex processes occurring within a cell in a graph-like structure. In the network, proteins are represented as nodes, while the interactions between the proteins are depicted as edges [Ram10]. These interactions can be directed or undirected, depending on the relationship between proteins. For instance, when protein A activates protein B, a directed edge goes from A to B, while mutual binding results in bidirectional edges [BO04a].

In recent years, the study of protein-protein interactions within organisms has gained increasing importance, shifting focus from individual proteins to entire PPINs [Ram10][BO04a][MV07]. New interactions within PPINs can be discovered using various methods. High-throughput experiments, such as yeast two-hybrid assays[SATRT+14], are one approach, though they are prone to false positives and can be difficult to reproduce. Alternatively, information can be extracted from literature and databases using text mining techniques, or predicted through computational methods based on protein structures [Ram10].

PPINs are often analyzed using characteristics like node degree distribution, clustering coefficient, and characteristic path length, which help to classify them into topologies such as random, small-world, or scale-free networks [SATRT+14][Ram10]. Biological networks typically display a scale-free topology, where hub nodes, proteins with many interactions, play a central role[BO04a]. The main goal of PPIN analysis is to identify clusters of proteins that share common characteristics, such as being expressed simultaneously or functioning in the same biological process. Functional enrichment analysis can further elucidate the processes these protein clusters may be involved in [KPA+12][SSO+12].

1.3 Ontologies

This section covers three essential ontologies: the Human Disease Ontology (DO), the Human Phenotype Ontology (HPO), and the Gene Ontology (GO).

The Human Disease Ontology (DO) The DO provides a standardized system for classifying human diseases, simplifying data integration across biomedical databases. By 2022, the ontology had grown to include more than 10,800 diseases, ranging from rare to complex conditions. This framework is key to aligning genomic, clinical, and computational data, enabling more effective comparisons between diseases [SMS+22].

The Human Phenotype Ontology (HPO) Launched in 2008, HPO standardizes the way phenotypic abnormalities are described in relation to diseases. Widely used in both clinical and research settings, it integrates genomic and phenotypic data to improve diagnostics, gene-disease discovery, and precision medicine. By 2021, HPO contained over 15,000 terms, aiding in disease modeling and cross-species phenotype comparisons [KGM+21].

The Gene Ontology (GO) GO offers a comprehensive vocabulary for annotating gene functions, biological processes, and cellular components across all species. It enables consistent data comparison and integration, particularly in genomics and systems biology. With more than 6.8 million annotations across 87,000 species, GO is continuously updated to reflect new biological knowledge, making it an invaluable tool for studying shared pathways in various diseases [Lom05].

1.4 Research Question

This thesis is building upon the foundation of the created consensus PPINs of Huntington's disease [Hen23] [JJ36] and Alzheimer's disease [Boe23]. The goal of this thesis is hereby to discover the overlapping patterns within PPINs that contribute to the characterization of common features across different neurodegenerative conditions. AD, HD, and SCA [STW+98] serve as illustrative examples of distinct neurodegenerative diseases. By comparing the PPINs of these neurodegenerative disorders, this research aims to discover overlapping patterns within these PPINs. The research question of this thesis is therefore:

RQ: *How do protein-protein interaction networks (PPINs) demonstrate overlapping and common patterns in (rare) neurodegenerative disorders?*

This thesis aims to provide insights into the identification of shared mechanisms across a spectrum of neurodegenerative diseases. The ultimate goal of the research is to improve our understanding of the molecular processes of neurodegenerative disorders, thereby creating possibilities for more effective treatment strategies.

1.5 Thesis overview

The first chapter, Section 1 of this thesis provides background information and outlines the study's objectives. Section 2 includes definitions used throughout the thesis. Section 3 reviews related work on PPINs for AD, HD, SCA, and overlaps in neurodegenerative diseases. Section 4 explains the methodology for constructing and comparing the different PPINs. The results

are presented in Section 5. The discussion and conclusion are found in Sections 6 and 7, respectively. Appendix A contains images of the networks and clusters that are discussed. The complete directory, including Cytoscape files, data, images, and results, is available at GitHub - comparison_neurodegenerative_disorders. This bachelor's thesis was conducted at LIACS (Leiden Institute of Advanced Computer Science) under the guidance of Dr. K.J. Wolstencroft.

2 Definitions

In this thesis, several abbreviations and technical terms are frequently used in the context of protein-protein interaction networks (PPINs) and neurodegenerative diseases. Below is an explanation of how these abbreviations and terms are utilized throughout this study:

- **AD, HD, SCA, SCA1, SCA3**: These abbreviations refer to specific neurodegenerative diseases: Alzheimer's disease (AD), Huntington's disease (HD), spinocerebellar ataxia (SCA), spinocerebellar ataxia type 1 (SCA1), and spinocerebellar ataxia type 3 (SCA3). "SCA" is used as a general term to describe the broader group of spinocerebellar ataxias, while "SCA1" and "SCA3" refer to specific subtypes.

- **C-PPIN (Consensus Protein-Protein Interaction Network)**: C-PPIN refers to a consensus protein-protein interaction network created using the automated workflow that integrates protein interaction data from databases such as KEGG, Wikipathways, and STRING. These networks are central to the thesis as they represent the complex interactions between proteins within specific diseases or across multiple diseases.

- ***[Disease]* C-PPIN**: The notation *[Disease]* C-PPIN is used to refer to the consensus protein-protein interaction network for a particular disease. For example, "AD C-PPIN" denotes the network specific to Alzheimer's disease, consisting of protein interactions relevant to AD.

- **Individual C-PPIN**: Individual C-PPIN refers to a C-PPIN that represents the protein-protein interactions of a single disease. Examples include the "SCA C-PPIN", "SCA1 C-PPIN", "SCA3 C-PPIN", or the "HD C-PPIN". Each Individual C-PPIN is dedicated to the interaction network of one specific disease without combining data from other diseases.

- **Merged C-PPIN**: Merged C-PPIN refers to a consensus protein-protein interaction network that combines the data from two or more Individual C-PPINs. For example, "AD $\cap$ SCA3 C-PPIN" represents a merged network of Alzheimer's disease and spinocerebellar ataxia type 3, capturing interactions shared by both diseases, as well as unique interactions from each.

- **Intersection C-PPIN** and **Union C-PPIN**: These terms describe specific types of merged networks. The Intersection C-PPIN includes only the protein interactions that are common across the diseases being studied, while the Union C-PPIN includes all protein interactions present in any of the considered diseases, providing a comprehensive view of the combined interaction landscape.

3 Related Work

This chapter will discuss previous research on C-PPINs in AD, HD, and SCA, especially SCA1 and SCA3. In addition to that, the emphasis of this chapter also lies on the overlap between neurodegenerative diseases found in protein-protein network analysis research.

3.1 Alzheimer's and Huntingtons's disease C-PPINs

The HD C-PPIN was initially developed by Chen Jiang (C.J.), who combined consensus data from KEGG, WikiPathways, and STRING to create a detailed HD C-PPIN. This network was specifically designed to highlight both similarities and differences between these pathways, thereby improving our understanding of the HD pathway[JJ36].

Nina Henninger(N.H.) later expanded on C.J.'s network by incorporating the latest STRING data, as the KEGG and WikiPathways databases had not been updated. The new consensus network showed slight differences compared to the original network developed by C.J. After integrating this updated information, N.H. conducted a comparative analysis with a separate network created by Adam Labadorf *et al.* [LCM18]. The analysis revealed that the proteins ACTB, GAPDH, AKT1, TP53, and MYC were the most highly connected, with MYC being the only protein among them to show upregulation, so when overexpressed, could contribute to neurodegeneration [Hen23].

Finally, Aster de Boer (A.d.B) [Boe23] developed an AD C-PPIN using an automated workflow, as detailed in Section 4. In A.d.B.'s research, the HD consensus PPIN was merged with the AD C-PPIN to examine the overlap between these two networks. Data for the C-PPINs were sourced from KEGG, WikiPathways, and STRING, just like the previous studies. The intersection of the AD and HD C-PPIN revealed that approximately two-thirds of the network components overlapped, with key biological processes such as chemical synaptic transmission, autophagy, apoptosis, and nervous system development being enriched in the overlapping regions [Boe23]. A review of the literature further confirmed that these processes are significant in both AD and HD, suggesting that PPINs analyzes can be instrumental in identifying shared pathways between related diseases.

3.2 Spinocerebellar ataxia disease C-PPINs

In previous research, SCA C-PPINs were constructed to explore the underlying molecular mechanisms of the disease. Eidhof *et al.* conducted a notable study, with the objective of identifying shared biological processes among the genes mutated in SCA[EWS19]. The study integrated data from multiple databases, including GeneMANIA, HPRD, and BioGrid, to create a complete PPIN. The network analysis showed that genes linked to SCA form tightly connected groups involved in important processes like balancing calcium levels, sending signals between nerve cells, and responding to cellular stress. The study discovered that these patterns of protein interactions are consistent across different forms of SCA, regardless of how the disease is inherited. It also highlighted that problems with maintaining nerve cell function contribute to the disease's progression. The research provided important insights into possible treatments for SCA and its links to other neurodegenerative diseases like AD and HD[EWS19].

Another important study by Sousa et al. (2023) [SRV+23] aimed to find new treatment targets for SCA1 using the EvoPPI3 platform, a meta-database for protein-protein interactions (PPI).

EvoPPI3 combines PPI data from different sources, including patients, cell lines, and model organisms like fruit flies. Sousa *et al.* expanded the known network of the Ataxin-1 protein to include 909 interacting proteins, up from the previous 380. Their analysis identified 16 potential new treatment targets. Their work demonstrates how PPINs can help find new treatment options for SCA1 and shows how PPIN approaches can be useful for understanding and addressing other neurodegenerative diseases.

In addition to the previous two studies, another study also developed a SCA PPIN. This study, conducted by Lim *et al.* [LHS+06], aimed to understand specific neurodegenerative diseases that damage Purkinje cells, leading to imbalance in the cerebellum. SCA is one such disease. However, while the genetic mutations causing these diseases are known, the normal roles of the affected proteins remain unclear. To investigate these functions, the researchers constructed a PPIN for 54 proteins involved in 23 inherited ataxias. They expanded the network by incorporating known protein interactions from scientific literature, as well as evolutionarily conserved interactions. Using the yeast two-hybrid screening method, the study identified 770 new protein-protein interactions. The findings showed that many proteins associated with SCA share common interacting partners, some of which have been linked to neurodegeneration in animal models.

3.3 Overlap in neurodegenerative diseases

Several studies have explored the overlap between different neurodegenerative diseases using PPIN analyses. To emphasize the insights gained from studying overlapping neurodegenerative diseases, one relevant study is discussed in detail.

In the study titled *"Interactome Mapping Provides a Network of Neurodegenerative Disease Proteins and Uncovers Widespread Protein Aggregation in Affected Brains"* by Haenig *et al.*[HAT+20], an interactome map was created using a yeast two-hybrid interaction screening. The map focuses on neurodegenerative diseases, including 5,000 human proteins and approximately 30,000 candidate protein-protein interactions. This network revealed how proteins linked to various neurodegenerative diseases, such as AD, HD, and SCA1, are connected [HAT+20].

One of the key findings of this study is the identification of significant overlap between protein networks involved in different neurodegenerative diseases. Specifically, pathogenic proteins in AD, such as ß-amyloid and tau, were found to aggregate abnormally not only in AD but also in the brains of patients with other disorders such as HD and SCA. This supports the hypothesis that these diseases may share common disease mechanisms, particularly in terms of protein aggregation. For example, ATXN1, a mutated protein in SCA1, was found to aggregate in the brains of patients with Alzheimer's disease, further indicating an overlap between these conditions[HAT+20].

4 Methods

In this chapter, the methodology used for this study is thoroughly explained. It begins with a general overview of the data collection process and preprocessing steps common to all individual C-PPINs. Following this, the specific steps for recreating the AD, HD and the creation of the SCA C-PPINs are detailed, utilizing the automated workflow developed by A.d.B. [Boe23].

To select the diseases for this study, both the Human Phenotype Ontology and the Human Disease Ontology were employed to identify relevant disease phenotypes and classifications. This process led to the selection of spinocerebellar ataxia, specifically SCA1 and SCA3, because, like HD, these rare disorders are polyQ diseases, which makes them particularly relevant for this study. A detailed explanation of the disease mechanisms of SCA, is described in Section 1.

The chapter then presents the methodology for merging and comparing these networks, adhering to the structure outlined in A.d.B.'s thesis, which includes network, clustering, and functional enrichment analyses. This approach is inspired by a review article by Koh *et al.* [KPA+12], which provides a robust framework for such analyses. An overview of the entire methodology, including key steps and processes, is presented in Figure 1.

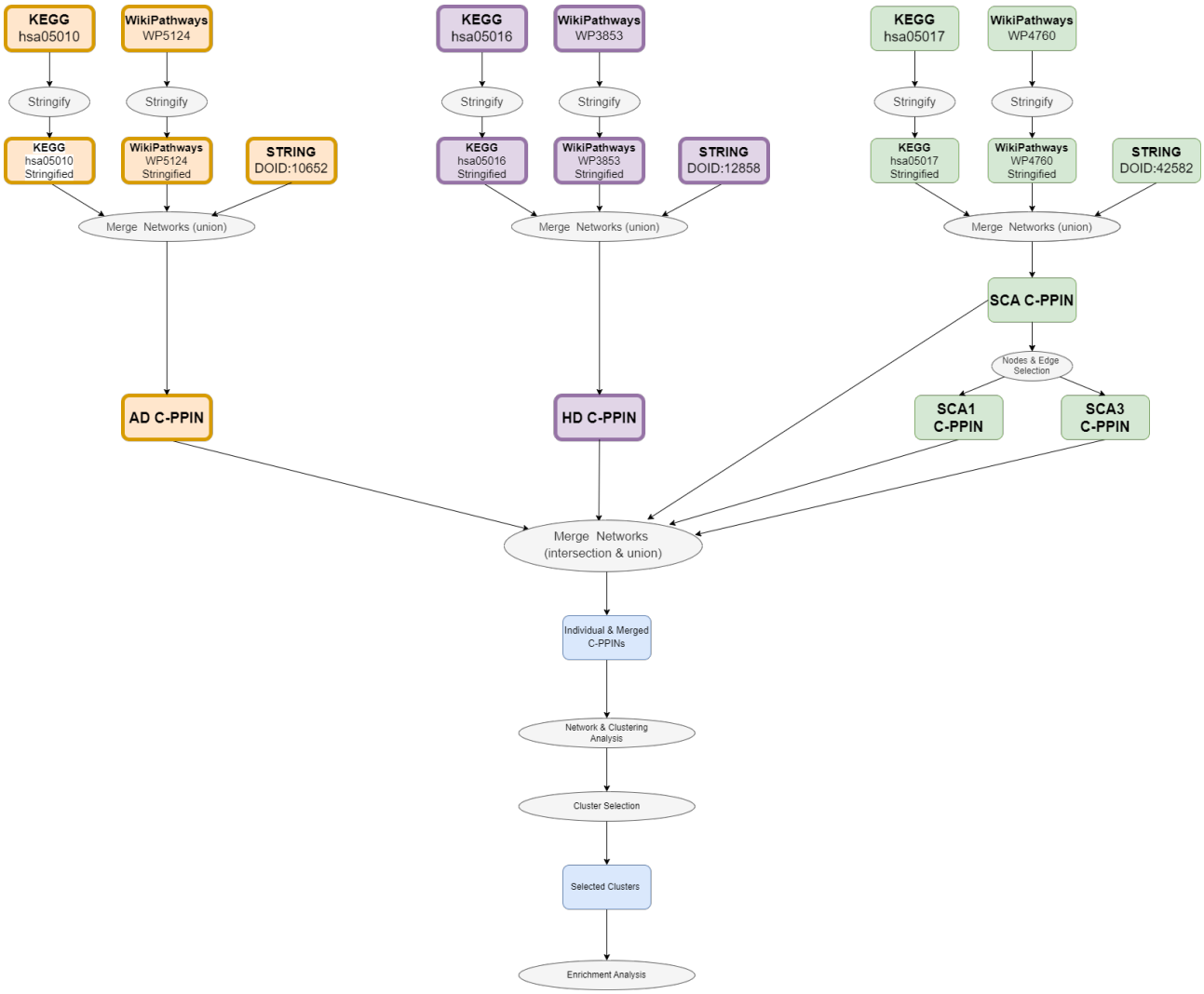


Figure 1: Overview of the methodology for creating, merging and comparing C-PPINs. The figure illustrates the key steps and processes involved in data collection, network creation, network analysis, clustering, and functional enrichment.

4.1 Tools and Resources

In this study, various databases, software, and resources were employed. The specific versions used are listed in Table 1. As these databases and software applications are frequently updated, the exact versions implemented in this research are detailed in the table.

Table 1: Overview of the databases and software applications used in this research, including the specific versions that were used.

Tools/resource	Version
Cytoscape	3.10.1 3.10.2
cyREST [OMK+15]	3.13.2
Wikipathways (CysApp) [KLEP14]	3.3.10
stringApp (CysApp)[DMGJ18]	2.0.3
NetworkAnalyzer (CysApp)[ARS+07]	4.5.0
MCODE (CysApp)[BH03]	2.0.3
BiNGO (CysApp)[MHK05]	3.0.5
STRING database	12
AD_hsa05010.gml	05/2024
WP5124.gpml	05/2024
HD_hsa05016.gml	05/2024
WP3853.gpml	05/2024
SCA_hsa05017.gml	06/2024
WP470.gpml	06/2024

4.2 Data Sources

The data sources that were used to create the C-PPINs include the Kyoto Encyclopedia of Genes and Genomes (KEGG), the Search Tool for the Retrieval of Interacting Genes/Proteins (STRING), and the WikiPathways database. These databases are explained in more detail below.

The KEGG database is a carefully organized collection of information about genes and genomes, providing a helpful tool for biological research. Each database entry is assigned a unique KEGG Identifier (KEGG ID). KEGG was initially developed to map genes to biological pathways, but it has since broadened its scope to include genomics as well as a variety of other omics disciplines. Human diseases, in particular, have become a major focus within the KEGG PATHWAY database. Therefore, the KEGG database was especially useful for this study, as it provides data on the AD, HD, and SCA pathways, making it a crucial data source [KFT+17] [KFS+23].

Another important data source used in this study is WikiPathways, an open-source database dedicated to biological pathways. WikiPathways gathers shared knowledge about these biological pathways, offering a valuable resource for research. A key advantage of WikiPathways is its open and collaborative nature, which aligns with the FAIR principles—ensuring that data is Findable, Accessible, Interoperable, and Reusable [SKH+18]. Additionally, WikiPathways has been extensively used to analyze and integrate experimental data from transcriptomics, proteomics, and metabolomics studies, making it a multifunctional tool for biological research and, therefore, a useful resource for this study [ABH+24] [SKH+18].

The last data source that was used for this study is the STRING database. The STRING database systematically gathers and integrates data on protein-protein interactions, including both physical interactions and functional associations. This information comes from various sources, such as text mining of scientific literature, computational predictions based on co-expression, conserved genomic context, interaction experiments, and curated databases of known complexes and pathways. These interactions are carefully evaluated, scored, and then applied to less-studied organisms using hierarchical orthology. [SGN+21] [SKK+23].

4.3 Pre-processing

For this study, assigning proteins to UniProt IDs was essential for merging the various datasets into one C-PPIN. The UniProt Knowledgebase (UniProtKB) provides a comprehensive and high-quality set of protein sequences with detailed annotations [BMO+21]. Using UniProt IDs ensures that proteins are consistently identified across different datasets, making it easier to create a reliable C-PPIN.

Before the workflow could be used, the gathered data needed to be pre-processed to create three networks that could be merged into a single consensus network. The script `kegg_data_converter.py` was used to convert proteins KEGG pathways to UniProt IDs. This script was originally created by A.d.B. For this study, the converter was slightly modified to make data retrieval easier by adding a function that allows direct use of KEGG pathway IDs. This function automatically constructs the appropriate URL, retrieves the data from the KEGG database in KGML format, and returns the pathway information. This was done to make it easier to convert the data when you want to create different disease C-PPINs.

In addition to converting KEGG data, the WikiPathways data also needed to be converted. The script `wikipathways_convert_uniprotIDs_NoriaYousufi.py` was created by Noria Yousufi (N.Y.) [You23] and remains unmodified for this study. The links to all the converters are provided in Table 2.

Table 2: Links to the converters used for transforming KEGG IDs and WikiPathways IDs to UniProt IDs, including a modified version for the KEGG ID conversion.

Converter	**Link**
KEGG to Uniprot	`kegg_data_converter.py`
KEGG to Uniprot - slightly modified	`kegg_data_converter_modified.py`
Wikipathways to Uniprot created by N.Y.	`wikipathways_convert_uniprotIDs_NoriaYousufi.py`

After the data from the different databases were loaded, the workflow utilized a Cytoscape app called STRINGapp. STRINGapp can convert any network in Cytoscape into a format that provides access to additional node attributes from the STRING database, as well as other features like enrichment analysis. This conversion process, known as "STRINGify," was applied to create a stringified version of the three different C-PPINs, allowing for enhanced analysis and integration of the networks[DMH+23]. Table 3 outlines the specific settings employed for this process.

Table 3: Parameter settings used to convert KEGG and WikiPathways networks into a string representation.

Parameter	**Setting**
Column for STRING query	uniprot
Include unmappable nodes	False
Map nodes to compounds	False
Species for the query	Homo Sapiens

4.4 Workflow Setup

First, the python libraries py4cytoscape and pandas were installed. Additionally, the automated workflow utilizes CyREST, a Cytoscape app that enables communication between the execution environment and Cytoscape. This setup allows the workflow to interact directly with Cytoscape, ensuring that network data can be manipulated and analyzed efficiently within the execution environment[DOM+15].

Furthermore, as noted in [MF], databases often assign a confidence score to each interaction, which helps users filter out less reliable data. This score ranges from 0 to 1, with higher values indicating greater confidence that the interaction is biologically relevant. STRING uses a special scoring system that combines different types of evidence into one confidence score for each predicted interaction [vMHJ+03]. Higher scores suggest more reliable data, while lower scores permit more interactions in the network. The confidence cutoff score was set at 0.4, as A.d.B.'s study showed no significant advantage in using a higher cutoff score, such as 0.7. Therefore, the choice was made to use the lower score to maximize the number of connections while still reducing the amount of noisy data. However, some noise may still be present in the data.

Additionally, A.d.B.'s study found that the clusters remained mostly the same even with an expansion of 2000 proteins. Since the aim of this study focuses more on the overlap between the SCA C-PPIN and the other C-PPINs, it was decided not to expand the networks.

4.5 Merging and Filtering

After preprocessing and setting up the workflow discussed in the previous sections, the data from KEGG, WikiPathways, and STRING were created into separate networks. Before merging these networks, some rows in the STRING and WikiPathways networks lacked canonical names because they represented elements other than proteins, such as micro-RNAs. These rows were removed to ensure a smooth merging process. The networks were then combined using canonical names (or UniProt IDs) to match corresponding nodes accurately. To track database overlaps in the merged network, a filter was applied to label each node based on its source database. This labeled data was integrated into the merged network, making it easier to examine how nodes are distributed across different databases.

4.6 Recreation of Alzheimer's and Huntington's C-PPINs

With the foundational workflow established and the merged network in place, the next step is to apply this approach to specific disease contexts. The following chapter focuses on the recreation of the AD and HD C-PPINs as the data used to create the C-PPINs is constantly

being updated. As explained at the beginning of this chapter, the automated workflow created by A.d.B. was used for the creation of the new AD and HD C-PPINs.

The data from the three different databases, KEGG, WikiPathways, and STRING, were used. The identifiers for AD and HD are presented in Table 4 and Table 5. The identifiers used to create the AD C-PPIN are the same as those in A.d.B.'s study.

Table 4: Identifiers and retrieval links for the KEGG, Wikipathway, and STRING databases used in recreating the AD consensus network.

Database	**AD Identifier**	**Link**
KEGG	hsa05010	https://www.kegg.jp/pathway/hsa05010
Wikipathways	WP5124	https://www.wikipathways.org/pathways/WP5124.html
STRING	DOID:10652	https://string-db.org/

Table 5: Identifiers and retrieval links for the KEGG, Wikipathway, and STRING databases used in recreating the HD consensus network.

Database	**HD Identifier**	**Link**
KEGG	hsa05016	https://www.kegg.jp/pathway/hsa05016
Wikipathways	WP3853	https://www.wikipathways.org/pathways/WP3853.html
STRING	DOID:12858	https://string-db.org/

4.7 Creation of the Spinocerebellar Ataxia Consensus networks

Just like the recreation of the AD and HD C-PPINs, the automated workflow created by A.d.B. was used for the creation of the SCA C-PPIN.

The data from the three different databases, KEGG, WikiPathways, and STRING, were used. The identifiers for the SCA C-PPIN are presented in Table 6.

Table 6: Identifiers and retrieval links for the KEGG, WikiPathways, and STRING databases used in recreating the SCA consensus network.

Database	**SCA Identifier**	**Link**
KEGG	hsa05017	https://www.kegg.jp/pathway/hsa05017
WikiPathways	WP4760	https://www.wikipathways.org/pathways/WP4760.html
STRING	DOID:42582	https://string-db.org/

4.7.1 SCA 1 & SCA 3

Since the SCA1 and SCA3 data in KEGG, Wikipathway, and STRING are in the SCA pathway, the SCA1 and SCA3 C-PPINs were created using the built-in select function in Cytoscape. The specific disease proteins were first selected: ATXN1 for SCA1 and ATXN3 for SCA3. After selecting the specific disease proteins, the following settings were applied:

- **Select → Nodes → First Neighbors of Selected Nodes → Undirected**
- **Select → Edges → Edges between Selected Nodes**

After selecting the specific nodes and edges, the new C-PPINs were created from the selection.

4.8 Merging and comparing the C-PPINs

Merging C-PPINs from diseases like AD, HD, and SCA is crucial for identifying both shared and unique molecular mechanisms. This approach can uncover common interaction hubs and new pathways that might be missed when analyzing networks separately[ALC05].

Therefore, after establishing the individual C-PPINs, new merged C-PPINs were generated using the built-in merge function in Cytoscape. The specific settings applied for creating the merged C-PPINs are summarized in Table 7.

Table 7: Parameter settings for merging C-PPINs in Cytoscape.

Parameter	**Setting**
Operation	Intersection
Matching Columns	name
How to Merge Columns	Default
Enable Merging Nodes/Edges	True

A list of the different individual C-PPINs and their merged networks can be found in Table 8.

Table 8: List of the individual and merged C-PPINs.

Category	**Networks**
Individual C-PPINs	AD, HD, SCA, SCA1, SCA3
Merged C-PPINs	AD ∩ HD AD ∩ SCA HD ∩ SCA AD ∩ SCA1 HD ∩ SCA1 AD ∩ SCA3 HD ∩ SCA3 AD ∩ HD ∩ SCA AD ∩ HD ∩ SCA1 AD ∩ HD ∩ SCA3 AD ∩ SCA1 ∩ SCA3 HD ∩ SCA1 ∩ SCA3 AD ∩ HD ∩ SCA1 ∩ SCA3 AD ∪ HD ∪ SCA1 ∪ SCA3 AD ∪ SCA1 ∪ SCA3 HD ∪ SCA1 ∪ SCA3

Before merging the individual C-PPINs, it was necessary to stringify them since the individual C-PPINs were not created in the same Cytoscape file. The C-PPINs were exported to .CYJS

files, and the settings for stringifying the different individual PPINs are detailed in Table 3. The .CYJS files can be found here. Additionally, to perform the network analysis (discussed in section 4.8.1), specific columns of the individual PPINs had to be converted to the correct type, as the CYJS file had converted them to doubles instead of integers. This conversion was done using the `column_type_converter.ipynb`. This step was completed prior to merging the C-PPINs.

4.8.1 Network Analysis

To perform the network analysis on the different C-PPINS in Table 8, the *Analyze Network* tool in Cytoscape was used, were the setting 'Analyze as Directed Graph' was set to False. In Table 8 the different parameters of the network analysis are being described. The defenitions of the different parameters are based on the defenitions in the review articles [Ram10][BO04b][MV07][DH07].

Table 9: Definitions of parameters used in the network analysis of individual and merged C-PPINs, shown in Table 8.

Parameter	**Definition**
Node	Represents a single protein in a PPIN.
Edge	A link between two nodes, signifying an interaction or relationship between proteins.
Node degree	The number of connections (edges) a node has with other nodes.
Node degree distribution	A probability distribution showing how the number of connections varies across nodes, which helps in understanding the network's structure.
Characteristic path length	The average distance (in terms of edges) between any two nodes in the network, indicating how closely connected the network is overall.
Average clustering coefficient	This metric reflects how connected a node's neighbors are. A higher clustering coefficient means that a node's neighbors are likely interconnected. The average across all nodes gives the network's clustering coefficient.
Network diameter	The maximum distance between the two most distant nodes, based on the shortest paths.
Network density	Indicates how many connections exist in the network relative to the total possible connections.
Network heterogeneity	Measures the variation in node connectivity. High heterogeneity points to the existence of hubs—nodes with a disproportionately high number of connections, common in biological networks.
Network centralization	Describes the variation in the centrality of nodes, showing how much more central some nodes are compared to others.
Betweenness	A measure of a node's importance based on how frequently it lies on the shortest paths between other nodes. Higher betweenness suggests a node is critical for network connectivity.
Connected components	Refers to the distinct, isolated sub-networks within the larger network. A fully connected network has one component, while a network with isolated nodes has multiple components.

4.8.2 Clustering Analysis

For the clustering analysis in this study, the same method utilized in A.d.B.'s research was utilized. There, the Cytoscape tool Molecular Complex Detection (MCODE) [BH03] was employed to identify clusters within various C-PPINs [Boe23]. MCODE is specifically designed to detect small clusters of densely connected nodes, though it does not assign every protein within the network to a cluster [KPA+12]. MCODE identifies clusters in a network by focusing on areas where nodes are densely connected. It starts with a dense "seed" node and expands outward to find larger dense regions[BH03][SSO+12]. One major advantage of MCODE compared to other clustering methods is its ability to focus on specific clusters of interest without analyzing the entire network. This makes it especially useful for studying how different clusters are connected within protein networks [BH03][KPA+12][SSO+12].

For every created C-PPIN a cluster analysis has been performed. In Table 10 the parameter settings that were applied in the clustering analysis can be found.

Table 10: Parameter settings for MCODE clustering analysis.

Parameter	**Setting**
Find Clusters	In Whole Network
Include Loops	False
Degree Cutoff	2
Haircut	True
Fluff	False
Node Density Cutoff	0.1
Node Score Cutoff	0.2
K-Core	2
Max.Depth	100

4.8.3 Enrichment Analysis

A common approach for enrichment analysis in protein sets involves using the Gene Ontology (GO) database, which classifies proteins into three main domains: Cellular Component, Biological Process, and Molecular Function. The Biological Process category[Edi], in particular, is useful for identifying processes that are statistically overrepresented within a protein cluster[HCI+04][SSO+12]. To maintain consistency with prior research, such as the methodology outlined in A.d.B.'s study, this approach was adopted in the current analysis.

In line with A.d.B.'s research, BiNGO, a Cytoscape-integrated tool[SSO+12], was used for the enrichment analysis. Enrichment analysis helps identify biological processes or functions that are overrepresented in a dataset compared to what would be expected by chance. The statistical test typically used is the hypergeometric test which determines whether the number of proteins or genes associated with a specific GO term is significantly higher than expected. BiNGO performs these statistical tests to identify significantly enriched GO terms in the selected clusters and visualizes them within the GO hierarchy[Edi][MHK05]. To refine the analysis and minimize false positives, statistical corrections such as the Bonferroni or Benjamini-Hochberg adjustments were applied[Sto02]. Additionally, the stringApp tool facilitated functional enrichment analysis

of STRING data, with results seamlessly integrated into Cytoscape for further analysis[DMH+23].

Following the cluster analysis, enrichment analysis was conducted on selected clusters. The focus was on clusters containing disease-related proteins from either AD or HD, specifically APP, APOE, MAPT, PSEN1, PSEN2, or HTT, as well as SCA-related proteins like ATXN1 or ATXN3. An overview of the selected clusters is provided in Table 18. Both BiNGO and stringApp were utilized for these analyses.

First, STRING enrichment was obtained for each cluster, concentrating on biological processes while filtering out redundant terms. Subsequently, BiNGO graphs were generated for each cluster, with the parameter settings detailed in Table 11.

As recommended in A.d.B.'s study, the most recent version of the GO database was downloaded for the BiNGO analysis to replace the outdated default version[KPA+12]. The latest version can be downloaded from `https://geneontology.org/docs/download-ontology/#go_basic`. Additionally, as noted in A.d.B.'s study, the default STRING identifiers could not be used directly in BiNGO; therefore, the canonical names of the proteins had to be manually entered into the parameter settings.

Table 11: Configuration parameters for BiNGO enrichment analysis.

Parameters	**BiNGO setting**
Retrieve data	Paste genes from text (copy the values from stringdb::canonical names column)
Assess	Overrepresentation
Visualization	True
Statistical Test	Hypergeometric test
Multiple testing correction	Benjamini & Hochberg False Discovery Rate (FDR) correction
Siginificance level	0.05
Categories to be visualized	Overrepresented categories after correction
Reference set	Use whole annotation as reference set
Ontology file	go-basic.obo
name_space	biological_process
Organism/annotation	Homo Sapiens
Discard evidence codes	–

5 Results

This chapter presents the results of the network analysis, clustering, and enrichment analysis of the individual and merged C-PPINs that were created. The Cytoscape files that were used can be found at `cytoscape_files`. Some images of the C-PPINs that were created can be found in Appendix A.1 and A.2. Images of selected clusters can be found in Appendix A.3. A full overview of the results is available on `result_data` and `images`.

5.1 Network Analysis

Figures 3 to 7 display the individual C-PPINs generated for each disease. Figures 8 to 10 illustrate the most relevant merged C-PPINs. Key disease-related proteins (APP, APOE, MAPT,

HTT, PSEN1, PSEN2, ATXN1, ATXN3) are enlarged, and the color scheme represents the confidence level of each protein's activity in the nervous system, with darker colors indicating higher confidence. Additionally, the different shapes indicate the database sources for the protein data.

Using the *Network analysis* tool from StringApp, network analysis was performed. The results of the individual C-PPINs analysis, including key topological properties such as the number of nodes, number of edges, average number of neighbors, network diameter, network radius, characteristic path length, clustering coefficient, network density, network heterogeneity, network centralization, and the number of connected components, can be found in Table 12. For the results of the network analysis across all merged C-PPINs, including the same metrics, see Tables 13 and 14.

Table 12: Network analysis results: Topological properties across the individual C-PPINs.

C-PPIN	**AD**	**HD**	**SCA**	**SCA1**	**SCA3**
Number of nodes:	2112	2030	780	93	85
Number of edges:	118820	105642	17121	867	1080
Avg. number of neighbors:	111.572	103.278	45.052	18.344	24.988
Network diameter:	5	5	6	2	2
Network radius:	3	3	3	1	1
Characteristic path length:	2.192	2.230	2.448	1.801	1.703
Clustering coefficient:	0.444	0.432	0.483	0.633	0.651
Network density:	0.053	0.051	0.060	0.199	0.297
Network heterogeneity:	0.98	0.945	1.01	0.719	0.554
Network centralization:	0.438	0.394	0.301	0.818	0.719
Connected components:	2	2	22	1	1
Analysis time (sec):	4.452	5.009	0.332	0.009	0.007

The reconstruction of the AD and HD C-PPINs resulted in some variations in the network analysis outcomes compared to the studies by A.d.B. [Boe23] and N.H. [Hen23]. Figures 3 and 4 display the reconstructed C-PPINs. As shown in Table 12, the updated C-PPINs have more nodes and edges than previously reported: the new AD C-PPIN now has 2112 nodes and 118,820 edges, compared to 2104 nodes and 95,608 edges. Similarly, the HD C-PPIN now includes 2030 nodes and 105,642 edges, up from 2023 nodes and 82,753 edges. Additionally, the characteristic path length for both the AD and HD networks has decreased relative to the values reported by A.d.B. and N.H. The characteristic path length of the AD and HD C-PPINs are both even shorter compared to the SCA C-PPIN, which has significantly fewer nodes and edges.

Another important observation is that the network analysis results for the AD ∩ HD ∩ SCA1 ∩ SCA3 and AD ∩ SCA1 ∩ SCA3 C-PPINs are completely identical (see Table 13 and Table 14). This means that all measured network properties, including the number of nodes, number of edges, average number of neighbors, network diameter, network radius, characteristic path length, clustering coefficient, network density, network heterogeneity, network centralization, and connected components are the same for both networks. The analysis revealed that the inclusion of the HD C-PPIN in the AD ∩ HD ∩ SCA1 ∩ SCA3 C-PPIN does not introduce any

unique nodes or edges. The nodes and edges contributed by HD are already present in the AD ∩ SCA1 ∩ SCA3 C-PPIN.

Figure 2 presents a Venn diagram comparing the AD ∩ HD ∩ SCA1 ∩ SCA3 and AD ∩ SCA1 ∩ SCA3 C-PPINs. The smaller circles represent the SCA C-PPINs, which have more in common with the HD C-PPIN, as they are both PolyQ diseases. It can be seen that the HD C-PPIN does not add any additional nodes to the intersection of all C-PPINs. While this diagram does not replicate the networks in detail, it serves as a visual aid to illustrate the overlap between these C-PPINs.

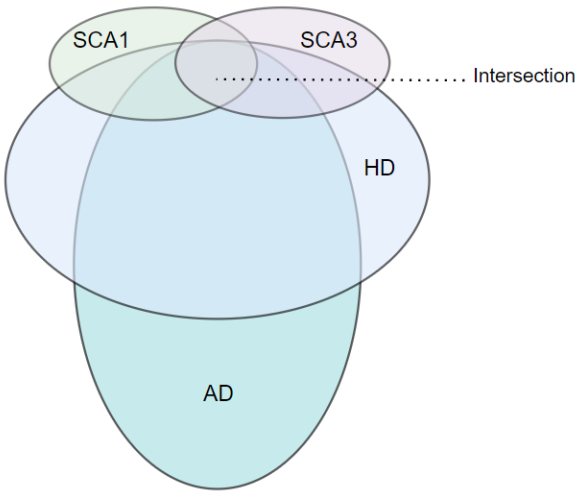


Figure 2: Venn diagram comparing the overlap of C-PPINs for AD, HD, SCA1, and SCA3. The diagram shows that the HD C-PPIN does not introduce any unique nodes or edges to the shared intersection of AD, SCA1, and SCA3. The stronger overlap between HD and the SCA C-PPINs reflects their shared classification as PolyQ diseases.

Additionally, since the HD C-PPIN did not introduce any new nodes in the intersection, union networks were created. As shown in Table 13, the AD ∪ SCA1 ∪ SCA3 and HD ∪ SCA1 ∪ SCA3 C-PPINs contain slightly more nodes than the individual AD or HD C-PPINs. Moreover, the AD ∪ HD ∪ SCA1 ∪ SCA3 C-PPIN contains significantly more nodes than any of the individual C-PPINs, though not as many as the combined total across all networks, which is 5,023 nodes. Images of all union networks are available here.

Table 13: Network analysis results I: Topological properties across the merged C-PPINs.

C-PPIN	**Number of nodes**	**Number of edges**	**Avg. number of neighbors**	**Network diameter**	**Network radius**	**Characteristic path length**
AD ∩ HD	1499	81895	109,022	4	2	2,135
AD ∩ SCA	437	12260	56,005	4	2	2,154
HD ∩ SCA	509	13699	53,906	5	3	2,206
AD ∩ SCA1	51	521	20,431	2	1	1,591
HD ∩ SCA1	64	676	21,125	2	1	1,665
AD ∩ SCA3	60	712	23,600	3	2	1,642
HD ∩ SCA3	77	937	24,312	3	2	1,708
AD ∩ HD ∩ SCA	424	11885	56,052	5	3	2,152
AD ∩ HD ∩ SCA1	49	500	20,408	2	1	1,575
AD ∩ HD ∩ SCA3	60	709	23,600	3	2	1,642
AD ∩ SCA1 ∩ SCA3	31	267	17,226	2	1	1,426
HD ∩ SCA1 ∩ SCA3	39	354	18,154	2	1	1,522
AD ∩ HD ∩ SCA1 ∩ SCA3	31	267	17,226	2	1	1,426
AD ∪ HD ∪ SCA1 ∪ SCA3	2674	142843	105,581	5	3	2,29
AD ∪ SCA1 ∪ SCA3	2167	119413	109,266	5	3	2,220
HD ∪ SCA1 ∪ SCA3	2063	105942	101,890	5	3	2,248

Table 14: Network analysis results II: Topological properties across the merged C-PPINs.

C-PPIN	**Clustering coefficient**	**Network density**	**Network heterogeneity**	**Network centralization**	**Connected components**	**Analysis time (sec)**
AD ∩ HD	0,468	0,073	0,874	0,463	1	3,507
AD ∩ SCA	0,592	0,128	0,778	0,373	1	0,235
HD ∩ SCA	0,550	0,106	0,818	0,347	2	0,195
AD ∩ SCA1	0,688	0,409	0,427	0,616	1	0,003
HD ∩ SCA1	0,642	0,335	0,473	0,686	1	0,004
AD ∩ SCA3	0,688	0,400	0,479	0,568	1	0,004
HD ∩ SCA3	0,661	0,320	0,531	0,671	1	0,006
AD ∩ HD ∩ SCA	0,595	0,133	0,769	0,375	1	0,164
AD ∩ HD ∩ SCA1	0,692	0,425	0,414	0,599	1	0,003
AD ∩ HD ∩ SCA3	0,688	0,400	0,479	0,568	1	0,006
AD ∩ SCA1 ∩ SCA3	0,777	0,574	0,362	0,455	1	0,001
HD ∩ SCA1 ∩ SCA3	0,739	0,478	0,395	0,550	1	0,003
AD ∩ HD ∩ SCA1 ∩ SCA3	0,777	0,574	0,362	0,455	1	0,002
AD ∪ HD ∪ SCA1 ∪ SCA3	0,427	0,04	1,042	0,386	3	8,435
AD ∪ SCA1 ∪ SCA3	0,449	0,050	1,000	0,429	2	6,273
HD ∪ SCA1 ∪ SCA3	0,436	0,049	0,958	0,390	2	4,040

5.2 Clustering Analysis

In total, 21 MCODE clustering analyses were performed across all the generated C-PPINs listed in Table 8. The number of clusters identified in each C-PPIN varied significantly, ranging from a maximum of 38 clusters to a minimum of 2 clusters. Due to the large number of C-PPINs, it was essential to filter the clustering results. Following the approach outlined in A.d.B.'s study, the focus was on the clusters containing key disease-related proteins, namely APP, MAPT, APOE, PSEN1, PSEN2, HTT, ATXN1, and ATXN3.

Tables 15, 16, and 17 provide detailed overviews of the clusters containing these disease-related proteins within the individual, intersection, and union C-PPINs, respectively. The clusters are listed in descending order based on their clustering score within each network. A higher clustering score indicates greater confidence that the group of proteins forms a functional complex [PR14].

Looking at Table 15 the clustering results for the updated AD C-PPIN and HD C-PPIN differ from those reported in A.d.B.'s and N.H.'s studies. The total number of clusters in the AD C-PPIN is now 30, down from the original 35, while the HD C-PPIN now contains 33 clusters instead of 30.

Another significant observation that can be seen in Table 15 is the absence of clusters in the AD C-PPIN containing the ATXN1 disease protein, despite its presence in the overall network.

Importantly, Table 16 show three clusters that are completely identical across multiple C-PPINs, having the same node and edge count and containing the same disease proteins from AD, HD, SCA1, and SCA3:

- AD $\cap$ SCA1 $\cap$ SCA3 - Cluster 1
- HD $\cap$ SCA1 $\cap$ SCA3 - Cluster 1
- AD $\cap$ HD $\cap$ SCA1 $\cap$ SCA3 - Cluster 1

Figures 11, 12, and 13 illustrate these three clusters, each containing the four disease proteins APP, HTT, ATXN1, and ATXN3. These clusters share a clustering score of 17,158, which, while relatively low on a scale of 0 to 100, should not be disregarded.

Finally, given the large number of clusters, enrichment analysis was conducted only on the clusters that contained disease proteins from AD or HD, along with proteins from SCA, SCA1, or SCA3. The selected clusters for this analysis can be found in Table 18.

Table 15: Clustering analysis results: overview of the number of clusters, clusters containing disease-related proteins, node and edge counts, and clustering scores across individual C-PPINs

C-PPIN	**Total #Clusters**	**Cluster**	**Disease Proteins**	**Node Count**	**Edge Count**	**Clustering Score**
AD	30	Cluster 1	APOE	284	13899	97,357
		Cluster 2	APP	194	6754	69,689
		Cluster 4	MAPT	125	813	13
		Cluster 5	HTT	144	762	10,643
		Cluster 10	ATXN3	94	335	6,194
		Cluster 11	PSEN1	40	113	5,795
		Cluster 29	PSEN2	6	7	2,8
HD	33	Cluster 1	APOE	260	105642	100,919
		Cluster 2	APP	135	3286	48,701
		Cluster 3	HTT	189	2439	25,83
		Cluster 5	PSEN2, ATXN3	164	882	10,81
		Cluster 11	MAPT	75	258	6,973
		Cluster 17	ATXN1	4	6	4,000
SCA	17	Cluster 2	APP APOE, MAPT	84	1232	29,518
		Cluster 3	HTT, PSEN1	48	338	14,340
		Cluster 4	ATXN3	61	312	10,167
SCA1	5	Cluster 1	APP, HTT, ATXN3	27	147	11,077
		Cluster 2	PSEN1, ATXN1	22	94	8,857
SCA3	5	Cluster 1	APP, MAPT	22	166	15,810
		Cluster 2	HTT, ATXN1, ATXN3	18	108	12,588

Table 16: Clustering analysis results: overview of the number of clusters, clusters containing disease-related proteins, node and edge counts, and clustering scores across intersection C-PPINs

C-PPIN	Total #Clusters	Cluster	Disease Proteins	Node Count	Edge Count	Clustering Score
AD ∩ HD	20	Cluster 1	APOE	269	14574	107,873
		Cluster 2	APP	122	2347	38,793
		Cluster 3	HTT	153	1609	21
		Cluster 4	PSEN2	178	1230	13,898
		Cluster 7	PSEN1	65	292	9,094
		Cluster 8	MAPT	65	190	5,938
		Cluster 9	ATXN3	84	221	5,325
AD ∩ SCA	10	Cluster 2	APP, APOE, MAPT	85	1249	29,667
		Cluster 3	HTT, PSEN1	48	332	14,085
		Cluster 4	ATXN3	46	228	10,133
		Cluster 8	PSEN2	22	53	5,048
		Cluster 9	ATXN1	18	37	4
HD ∩ SCA	12	Cluster 2	APP	80	1070	27,013
		Cluster 3	APOE, MAPT, HTT, PSEN1	48	329	14
		Cluster 4	ATXN3	52	265	10,392
		Cluster 11	ATXN1	11	18	3,6
AD ∩ SCA1	3	Cluster 1	APP, HTT, ATXN3	19	101	11,222
		Cluster 2	PSEN1, ATXN1	18	93	10,941
HD ∩ SCA1	4	Cluster 1	APP, HTT, ATXN3	20	105	11,053
		Cluster 2	PSEN1, ATXN1	26	126	10,08
AD ∩ SCA3	4	Cluster 1	APP, MAPT	19	137	15,222
		Cluster 2	HTT, ATXN1, ATXN3	16	97	12,933
HD ∩ SCA3	5	Cluster 1	APP, MAPT	18	128	15,059
		Cluster 2	ATXN1, ATXN3	18	116	13,647
AD ∩ HD ∩ SCA	10	Cluster 2	APP	80	1064	26,911
		Cluster 3	APOE, MAPT, HTT, PSEN1	47	320	13,913
		Cluster 4	ATXN3	47	269	11,696
		Cluster 8	PSEN2	22	42	4
		Cluster 10	ATXN1	13	21	3,5
AD ∩ HD ∩ SCA1	3	Cluster 1	APP, HTT, ATXN3	18	96	11,294
		Cluster 2	PSEN1, ATXN1	18	94	11,059
AD ∩ HD ∩ SCA3	4	Cluster 1	APP, MAPT	19	137	15,222
		Cluster 2	HTT, ATXN1, ATXN3	16	97	12,933
AD ∩ SCA1 ∩ SCA3	2	Cluster 1	APP, HTT, ATXN1, ATXN3	20	163	17,158
HD ∩ SCA1 ∩ SCA3	2	Cluster 1	APP, HTT, ATXN 1, ATXN3	20	163	17,158
AD ∩ HD ∩ SCA1 ∩ SCA3	2	Cluster 1	APP, HTT, ATXN1, ATXN3	20	163	17,158

Table 17: Clustering analysis results: overview of the number of clusters, clusters containing disease-related proteins, node and edge counts, and clustering scores across union C-PPINs

C-PPIN	Total #Clusters	Cluster	Disease Protein	Node Count	Edge Count	Clustering Score
AD ∪ HD ∪ SCA1 ∪ SCA3	38	Cluster 1	APOE	284	14427	97,187
		Cluster 2	APP	196	6943	70,749
		Cluster 3	HTT	238	3394	25,3
		Cluster 4	MAPT	129	847	13,094
		Cluster 11	PSEN1	72	205	5,746
		Cluster 17	PSEN2, ATXN1	37	77	4,278
		Cluster 22	ATXN3	14	22	3,385
AD ∪ SCA1 ∪ SCA3	34	Cluster 1	APOE	284	13899	97,357
		Cluster 2	APP	191	6749	70,737
		Cluster 4	MAPT	125	813	13
		Cluster 9	HTT	113	436	7,768
		Cluster 10	PSEN1	50	146	5,959
		Cluster 13	ATXN3	79	188	4,744
		Cluster 17	PSEN2, ATXN1	27	53	4
HD ∪ SCA1 ∪ SCA3	35	Cluster 1	APOE	260	13762	100,919
		Cluster 2	APP	135	3286	48,701
		Cluster 3	HTT	189	2439	25,83
		Cluster 5	PSEN2, ATXN3	164	882	10,81
		Cluster 7	PSEN1	96	434	9,116
		Cluster 11	MAPT	75	258	6,973
		Cluster 24	ATXN1	3	3	3

5.3 Enrichment Analysis

The enrichment analyzed was performed on 18 selected clusters. These clusters were selected form the clustering analysis. The clusters that were chosen are the clusters that both have a SCA disease protein (either ATXN1 or ATXN3), and a disease protein from AD (APP, APOE, MAPT, PSEN1 or PSEN2) or HD (HTT). See Table 18 for an overview of the the selected clusters. The clusters are ordered in clustering score, from a bigger clustering score to a lower clustering score.

Following the example in A.d.B's study two ways were applied for the enrichment analysis; a STRING enrichment and a BiNGO enrichment:

- The STRING enrichment analysis was performed on each cluster. STRING enrichment orders the different processes by their false discovery rate (FDR). That is a statistical measure used in the context of multiple hypothesis testing. It represents the proportion of false positives among all the hypotheses that have been identified as significant. In other words, FDR quantifies the expected rate at which incorrect conclusions are drawn from the set of hypotheses that were considered significant[Sto02].
- In this study, BiNGO analysis results are visualized as a directed acyclic network representing sections of the Gene Ontology (GO) related to biologicial processes [Edi] [MHK05]

[KPA+12]. Each node corresponds to a GO term, with more general processes positioned higher and more specific processes lower in the hierarchy. Nodes that are not enriched appear in white and act as connectors, while the color of each node indicates the level of enrichment-light yellow for significance and dark orange for higher significance. Node size represents the number of proteins associated with each term. Due to the hierarchical nature of the network, a node that leads to a lower-level node with similar attributes often indicates that the lower node is a major contributor to the observed enrichment. Thus, the most significant findings are shown by the large orange nodes at the bottom of the graph. This hierarchical layout was applied consistently across all clusters listed in Table 18.

Table 18: Selected clusters containing both SCA disease proteins and AD or HD disease proteins across individual, intersection and union C-PPINs.

C-PPINs	**Total #Clusters**	**Cluster**	**Disease Proteins**	**Node Count**	**Edge Count**	**Clustering Score**
HD	33	Cluster 5	PSEN2, ATXN3	164	882	10,81
SCA1	5	Cluster 1	APP, HTT, ATXN3	27	147	11,077
		Cluster 2	PSEN1, ATXN1	22	94	8,857
SCA3	5	Cluster 2	HTT, ATXN1, ATXN3	18	108	12,588
AD ∩ SCA1	3	Cluster 1	APP, HTT, ATXN3	19	101	11,222
		Cluster 2	PSEN1, ATXN1	18	93	10,941
HD ∩ SCA1	4	Cluster 1	APP, HTT, ATXN3	20	105	11,053
		Cluster 2	PSEN1, ATXN1	26	126	10,08
AD ∩ SCA3	4	Cluster 2	HTT, ATXN1, ATXN3	16	97	12,933
AD ∩ HD ∩ SCA1	3	Cluster 1	APP, HTT, ATXN3	18	96	11,294
		Cluster 2	PSEN1, ATXN1	18	94	11,059
AD ∩ HD ∩ SCA3	4	Cluster 2	HTT, ATXN1, ATXN3	16	97	12,933
AD ∩ SCA1 ∩ SCA3	2	Cluster 1	APP, HTT, ATXN1, ATXN3	20	163	17,158
HD ∩ SCA1 ∩ SCA3	2	Cluster 1	APP, HTT, ATXN1, ATXN3	20	163	17,158
AD ∩ HD ∩ SCA1 ∩ SCA3	2	Cluster 1	APP, HTT, ATXN1, ATXN3	20	163	17,158
AD ∪ HD ∪ SCA1 ∪ SCA3	38	Cluster 17	PSEN2, ATXN1	37	77	4,278
AD ∪ SCA1 ∪ SCA3	34	Cluster 17	PSEN2, ATXN1	27	53	4
HD ∪ SCA1 ∪ SCA3	35	Cluster 5	PSEN2, ATXN3	164	882	10,81

First, examining the images in Figures 11 to 13, the STRING results for the three clusters— AD ∩ SCA1 ∩ SCA3 - (Cluster 1), HD ∩ SCA1 ∩ SCA3 (Cluster 1) and AD ∩ HD ∩ SCA1 ∩ SCA3 (Cluster 1), discussed in Section 5— reveal that they share the same GO biological processes. These include the *response to topologically incorrect proteins*, *positive regulation of the catabolic process*, the *ERAD pathway*, *regulation of protein localization*, and *chaperone mediated autophagy*. The corresponding FDR values are listed alongside each biological process. The BiNGO hierarchical graphs of the three clusters appear large and broad, with many yellow nodes, indicating significant but less strongly enriched processes. There are few orange nodes at the ends, which would suggest higher significance. Notably, all three graphs share one orange leaf node with the same term, namely *mitochondrial outer membrane translocase complex assembly*. See Figure 14.

Images of all clusters analyzed using STRING and BiNGO are available at `cluster_images`, showing GO biological processes and their associated FDR values. The results of the BiNGO enrichment analysis can be found at `BiNGO_results`.

6 Discussion

Since the AD and HD C-PPINs were recreated, the results of the network analysis show minor differences compared to those in A.d.B.'s and N.H.'s studies [Boe23][Hen23]. As expected, the recreated C-PPINs contain a higher number of nodes and edges due to the continuous updates in data sources like KEGG, WikiPathways, and STRING [KFS+23][SKH+18][SKK+23]. This increase in the number of edges has resulted in a decrease in the characteristic path length for both the AD and HD networks, making them more cohesive and indicating that proteins within these networks are more tightly connected than in earlier created C-PPINs [JTA+00].

A notable observation is that the HD C-PPIN does not introduce any new nodes in the AD ∪ HD ∪ SCA1 ∪ SCA3 network. This can be explained by the properties of AD. Since AD is the only non-polyQ disease in this analysis [BB98][SSK+21], it is expected that the nodes and edges present in the AD ∪ SCA1 ∪ SCA3 C-PPINs are also found in HD, as HD is a polyQ disease like SCA1 and SCA3 [Roo10][CCJS22]. Additionally, the HD ∪ SCA1 ∪ SCA3 C-PPIN is larger than the AD ∪ HD ∪ SCA1 ∪ SCA3 network, suggesting that HD, SCA1, and SCA3 share overlapping nodes and edges due to their common polyQ mechanism.

Another significant observation is the difference in clusters compared to those identified in A.d.B.'s and N.H.'s analyses. Changes in network composition likely led to the formation of new clusters. Additionally, the absence of clusters in the AD C-PPIN containing the ATXN1 protein, despite its presence in the network, may be due to how MCODE identifies clusters by focusing on areas with dense connectivity. As a result, a cluster represents a subnetwork of proteins that are highly physically connected and may be involved in related biological processes. This approach works well for proteins that physically interact; however, for proteins involved in signal transduction, where connections are more dynamic and unstable, capturing them in clusters is more difficult. For ATXN1, its exclusion from a cluster may indicate unstable interactions within the AD C-PPIN, making it less likely to be part of a densely connected group[MHK05]. It is also worth noting that the MCODE clustering algorithm is sensitive to noise, such as false-positive interactions [KPA+12].

One of the most significant findings of this study is the identification of three clusters across multiple C-PPINs that exhibit consistent clustering results, including identical edge and node counts, the same disease-associated proteins, and a uniform clustering score of 17,158. Although this score is not particularly high compared to the maximum of 100, the presence of the same cluster across three different networks suggests that these diseases share common mechanisms. This finding also aligns with the interactome mapping study discussed in Section 3.3 [HAT+20].

Additionally, the STRING enrichment analysis reveals that the three clusters share similar biological processes. These include the *response to topologically incorrect proteins*, *positive regulation of the catabolic process*, the *ERAD pathway*, *regulation of protein localization*, and *chaperone mediated autophagy*. The response to topologically incorrect proteins, as well as the ERAD (Endoplasmic Reticulum-Associated Degradation) pathway, plays a critical role in managing misfolded proteins, a key factor in neurodegenerative diseases [CK15]. Misfolded proteins can form toxic aggregates in diseases like AD, HD and SCA. The ERAD pathway helps to prevent this by targeting misfolded proteins for degradation[MHJS05].

7 Conclusions and Further Research

In this study, multiple consensus protein-protein interaction networks were recreated using an existing automated workflow. The primary objective of this thesis was to identify overlapping patterns within C-PPINs that contribute to the characterization of common features across different neurodegenerative conditions, focusing on Alzheimer's disease, Huntington's disease, and spinocerebellar ataxia (particularly SCA1 and SCA3). By comparing the PPINs of these diseases, the goal was to discover shared mechanisms that could provide insights into their molecular processes and potentially inform new treatment strategies.

The findings suggest that these neurodegenerative diseases share significant biological mechanisms, especially in processes related to protein misfolding. The identification of three consistent clusters across different C-PPINs further supports the notion of overlapping mechanisms between these diseases.

However, while this study focused on three polyglutamine (polyQ) diseases and one tauopathy, future research could explore overlaps between different disease classes or focus on only one disease class. Future studies could expand upon the techniques used here, integrating newer data sources and more diverse disease types to further explore these overlaps or examine the differences instead. The world of neurodegenerative diseases still holds many unanswered questions, and the opportunities for discovery remain extensive.

Acknowledgements

I would like to express my sincere gratitude to my supervisor, Katy Wolstencroft, for her time, feedback, and patience throughout the entire process of creating this thesis.

I would also like to extend my thanks to Lu Cao for providing the necessary guidance and support.

References

[ABH+24] Ayushi Agrawal, Hasan Balcı, Kristina Hanspers, Susan L. Coort, and Marvin Martens *et al.* Wikipathways 2024: next generation pathway database. *Nucleic Acids Research*, 52:D679–D689, 1 2024.

[ALC05] Richard A. Armstrong, Peter L. Lantos, and Nigel J. Cairns. Overlap between neurodegenerative disorders. *Neuropathology*, 25(2):111–124, 4 2005.

[ARS+07] Yassen Assenov, Fidel Ramírez, Sven-Eric Schelhorn, Thomas Lengauer, and Mario Albrecht. Computing topological parameters of biological networks. *Bioinformatics*, 24(2):282–284, 11 2007.

[BB98] J.-P Brion and Jean-Pierre Brion. Neurofibrillary tangles and alzheimer's disease, 1998.

[BH03] Gary D Bader and Christopher Wv Hogue. An automated method for finding molecular complexes in large protein interaction networks, 2003.

[BMO+21] Alex Bateman, Maria Jesus Martin, Sandra Orchard, Michele Magrane, and Rahat Agivetova *et al.* Uniprot: the universal protein knowledgebase in 2021. *Nucleic Acids Research*, 49:D480–D489, 1 2021.

[BO04a] Albert László Barabási and Zoltán N. Oltvai. Network biology: understanding the cell's functional organization. *Nature Reviews Genetics*, 5(2):101–113, 2 2004.

[BO04b] Albert László Barabási and Zoltán N. Oltvai. Network biology: Understanding the cell's functional organization, 2 2004.

[Boe23] Aster M. de Boer. Identifying overlapping processes of alzheimer's and huntington's disease with a protein-protein interaction network, 2023.

[BYBT10] Lynn M. Bekris, Chang En Yu, Thomas D. Bird, and Debby W. Tsuang. Review article: Genetics of alzheimer disease, 12 2010.

[BZMN11] D. Bano, F. Zanetti, Y. Mende, and P. Nicotera. Neurodegenerative processes in huntington's disease, 11 2011.

[CCJS22] Jiu Ming Chen, Shi Kai Chen, Pei Pei Jin, and Shun Chang Sun. Identification of the ataxin-1 interaction network and its impact on spinocerebellar ataxia type 1. *Human Genomics*, 16, 12 2022.

[CK15] Aaron Ciechanover and Yong Tae Kwon. Degradation of misfolded proteins in neurodegenerative diseases: therapeutic targets and strategies, 3 2015.

[DH07] Jun Dong and Steve Horvath. Understanding network concepts in modules. *BMC Systems Biology*, 1, 6 2007.

[DMGJ18] Nadezhda T. Doncheva, John H. Morris, Jan Gorodkin, and Lars J. Jensen. Cytoscape StringApp: Network Analysis and Visualization of Proteomics Data. *Journal of Proteome Research*, 18(2):623–632, 11 2018.

[DMH+23] Nadezhda T. Doncheva, John H. Morris, Henrietta Holze, Rebecca Kirsch, and Katerina C. Nastou *et al.* Cytoscape stringapp 2.0: Analysis and visualization of heterogeneous biological networks. *Journal of Proteome Research*, 22:637–646, 2 2023.

[DOM+15] Barry Demchak, Keiichiro Ono, Tanja Muetze, Georgi Kolishovski, and Paul Shannon. Cyrest: Turbocharging cytoscape access for external tools via a restful api. *F1000Research*, 4, 2015.

[Edi] Maria Victoria Schneider Editor. In silico systems biology.

[ETRM14] Melvin M. Evers, Lodewijk J.A. Toonen, and Willeke M.C. Van Roon-Mom. Ataxin-3 protein and rna toxicity in spinocerebellar ataxia type 3: Current insights and emerging therapeutic strategies, 2014.

[EWS19] Ilse Eidhof, Bart P. Van De Warrenburg, and Annette Schenck. Integrative network and brain expression analysis reveals mechanistic modules in ataxia. *Journal of Medical Genetics*, 56:283–292, 5 2019.

[HAT+20] Christian Haenig, Nir Atias, Alexander K. Taylor, Arnon Mazza, and Martin H. Schaefer *et al.* Interactome mapping provides a network of neurodegenerative disease proteins and uncovers widespread protein aggregation in affected brains. *Cell Reports*, 32, 8 2020.

[HCI+04] M. A. Harris, J. Clark, A. Ireland, J. Lomax, and M. Ashburner *et al.* The gene oncology (go) database and informatics resource. *Nucleic Acids Research*, 32, 1 2004.

[Hen23] Nina Anna Maria Henninger. Extending consensus knowledge in huntington's disease protein interaction networks, 2023.

[HHB+21] Harald Hampel, John Hardy, Kaj Blennow, Christopher Chen, and George Perry *et al.* The amyloid- pathway in alzheimer's disease, 10 2021.

[HV19] Miaozhen Huang and Dineke S. Verbeek. Why do so many genetic insults lead to purkinje cell degeneration and spinocerebellar ataxia? *Neuroscience Letters*, 688:49–57, 2019. The Cerebellum in Health and Disease.

[JJ36] Chen Ji and Rong Jiang. Opleiding informatica finding consensus knowledge in the huntington's disease pathway, 2636.

[JTA+00] H. Jeong, B. Tombor, R. Albert, Z. N. Oltvai, and A.-l. Barabási. The large-scale organization of metabolic networks. *Nature*, 407(6804):651–654, 10 2000.

[KFS+23] Minoru Kanehisa, Miho Furumichi, Yoko Sato, Masayuki Kawashima, and Mari Ishiguro-Watanabe. Kegg for taxonomy-based analysis of pathways and genomes. *Nucleic Acids Research*, 51:D587–D592, 1 2023.

[KFT+17] Minoru Kanehisa, Miho Furumichi, Mao Tanabe, Yoko Sato, and Kanae Morishima. Kegg: New perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Research*, 45:D353–D361, 1 2017.

[KGM+21] Sebastian Köhler, Michael Gargano, Nicolas Matentzoglu, Leigh C. Carmody, and David Lewis-Smith *et al.* The human phenotype ontology in 2021. *Nucleic Acids Research*, 49:D1207–D1217, 1 2021.

[KK18] I. A. Kuznetsov and A. V. Kuznetsov. How the formation of amyloid plaques and neurofibrillary tangles may be related: A mathematical modelling study. volume 474. Royal Society Publishing, 2 2018.

[KLEP14] Martina Kutmon, Samad Lotia, Chris T Evelo, and Alexander R Pico. WikiPathways App for Cytoscape: Making biological pathways amenable to network analysis and visualization. *F1000Research*, 3:152, 9 2014.

[KPA+12] Gavin C.K.W. Koh, Pablo Porras, Bruno Aranda, Henning Hermjakob, and Sandra E. Orchard. Analyzing protein-protein interaction networks. *Journal of Proteome Research*, 11:2014–2031, 4 2012.

[LCM18] Adam Labadorf, Seung H. Choi, and Richard H. Myers. Evidence for a pan-neurodegenerative disease response in huntington's and parkinson's disease expression profiles. *Frontiers in Molecular Neuroscience*, 10, 1 2018.

[LHS+06] Janghoo Lim, Tong Hao, Chad Shaw, Akash J. Patel, and Gábor Szabó *et al.* A protein-protein interaction network for human inherited ataxias and disorders of purkinje cell degeneration. *Cell*, 125:801–814, 5 2006.

[Lom05] Jane Lomax. Get ready to GO! A biologist's guide to the Gene Ontology. *Briefings in Bioinformatics*, 6(3):298–304, 1 2005.

[LXL+21] Ziyi Luo, Hao Xu, Liwei Liu, Tymish Y. Ohulchanskyy, and Junle Qu. Optical imaging of beta-amyloid plaques in alzheimer's disease, 8 2021.

[Mar99] Joseph B. Martin. Molecular basis of the neurodegenerative disorders. *New England Journal of Medicine*, 340(25):1970–1980, 1999. PMID: 10379022.

[MC09] Lucia Migliore and Fabio Coppedè. Genetics, environmental factors and the emerging role of epigenetics in neurodegenerative diseases, 7 2009.

[MDDGG] Antoni Matilla-Duen˜as, Duen˜ Duen˜as, Robert Goold, and Paola Giunti. Clinical, genetic, molecular, and pathophysiological insights into spinocerebellar ataxia type 1.

[MF] Cheryl L Meyerkord and Haian Fu. Protein-protein interactions methods and applications second edition methods in molecular biology 1278.

[MHJS05] Birgit Meusser, Christian Hirsch, Ernst Jarosch, and Thomas Sommer. Erad: the long road to destruction, 2005.

[MHK05] Steven Maere, Karel Heymans, and Martin Kuiper. Bingo: A cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. *Bioinformatics*, 21:3448–3449, 8 2005.

[MMA+14] Adriana Moro, Renato P. Munhoz, Walter O. Arruda, Salmo Raskin, Mariana Moscovich, and Hélio A.G. Teive. Spinocerebellar ataxia type 3: Subphenotypes in a cohort of brazilian patients. *Arquivos de Neuro-Psiquiatria*, 72:659–662, 9 2014.

[MV07] Oliver Mason and Mark Verwoerd. Graph theory and networks in biology, 2007.

[Mü21] Ulrich Müller. Spinocerebellar ataxias (scas) caused by common mutations, 10 2021.

[OMK+15] Keiichiro Ono, Tanja Muetze, Georgi Kolishovski, Paul Shannon, and Barry Demchak. CyREST: Turbocharging Cytoscape Access for External Tools via a RESTful API. *F1000Research*, 4:478, 8 2015.

[PBM+06] Christina Priller, Thomas Bauer, Gerda Mitteregger, Bjarne Krebs, Hans A. Kretzschmar, and Jochen Herms. Synapse formation and function is modulated by the amyloid precursor protein. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, 26:7212–7221, 2006.

[PMG22] Philip V. Peplow, Bridget Martinez, and Thomas A. Gennarelli. *Prevalence, Needs, Strategies, and Risk Factors for Neurodegenerative Diseases*, volume 173, pages 3–8. Humana Press Inc., 2022.

[PR14] Clara Pizzuti and Simona E. Rombo. Algorithms and tools for protein-protein interaction networks clustering, with a special focus on population-based stochastic methods, 5 2014.

[Ram10] Karthik Raman. Construction and analysis of protein–protein interaction networks. *Automated Experimentation*, 2(1):2, 1 2010.

[Roo10] Raymund A.C. Roos. Huntington's disease: A clinical review, 2010.

[RT11] Christopher A Ross and Sarah J Tabrizi. Huntington's disease: from molecular pathogenesis to clinical treatment, 2011.

[RWW+16] Michael D. Rawlins, Nancy S. Wexler, Alice R. Wexler, Sarah J. Tabrizi, Ian Douglas, Stephen J.W. Evans, and Liam Smeeth. The prevalence of huntington's disease, 3 2016.

[SATRT+14] Nahid Safari-Alighiarloo, Mohammad Taghizadeh, Mostafa Rezaei-Tavirani, Bahram Goliaei, and Ali Asghar Peyvandi. Protein-protein interaction networks (ppi) and complex diseases, 2014.

[SGN+21] Damian Szklarczyk, Annika L. Gable, Katerina C. Nastou, David Lyon, and Rebecca Kirsch *et al.* The string database in 2021: Customizable protein-protein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic Acids Research*, 49:D605–D612, 1 2021.

[SKH+18] Denise N. Slenter, Martina Kutmon, Kristina Hanspers, Anders Riutta, and Jacob Windsor *et al.* Wikipathways: A multifaceted pathway database bridging metabolomics to other omics research. *Nucleic Acids Research*, 46:D661–D667, 1 2018.

[SKK+23] Damian Szklarczyk, Rebecca Kirsch, Mikaela Koutrouli, Katerina Nastou, and Farrokh Mehryary *et al.* The string database in 2023: protein-protein association networks and functional enrichment analyses for any sequenced genome of interest. *Nucleic Acids Research*, 51:D638–D646, 1 2023.

[SMHV+16] José V. Sánchez-Mut, Holger Heyn, Enrique Vidal, Sebastián Morán, and Sergi *et al.* Sayols. Human DNA methylomes of neurodegenerative diseases show common epigenomic patterns. *Translational Psychiatry*, 6(1):e718, 1 2016.

[SMS+22] Lynn M. Schriml, James B. Munro, Mike Schor, Dustin Olley, and Carrie McCracken *et al.* The human disease ontology 2022 update. *Nucleic Acids Research*, 50:D1255–D1261, 1 2022.

[SRV+23] André Sousa, Sara Rocha, Jorge Vieira, Miguel Reboiro-Jato, Hugo López-Fernández, and Cristina P. Vieira. On the identification of potential novel therapeutic targets for spinocerebellar ataxia type 1 (sca1) neurodegenerative disease using evoppi3. *Journal of integrative bioinformatics*, 20, 6 2023.

[SSK+21] Philip Scheltens, Bart De Strooper, Miia Kivipelto, Henne Holstege, Gael Chételat, Charlotte E. Teunissen, Jeffrey Cummings, and Wiesje M. van der Flier. Alzheimer's disease, 4 2021.

[SSO+12] Rintaro Saito, Michael E. Smoot, Keiichiro Ono, Johannes Ruscheinski, Peng Liang Wang, Samad Lotia, Alexander R. Pico, Gary D. Bader, and Trey Ideker. A travel guide to cytoscape plugins, 11 2012.

[Sto02] John D Storey. A direct approach to false discovery rates, 2002.

[STW+98] S.K. Singhrao, P. Thomas, J.D. Wood, J.C. MacMillan, and J.W. Neal *et al.* Huntingtin protein colocalizes with lesions of neurodegenerative diseases: An investigation in huntington's, alzheimer's, and pick's diseases. *Experimental Neurology*, 150(2):213–222, 1998.

[TGL19] Sarah J. Tabrizi, Rhia Ghosh, and Blair R. Leavitt. Huntingtin lowering strategies for disease modification in huntington's disease, 3 2019.

[VD98] Jean Paul G. Vonsattel and Marian DiFiglia. Huntington Disease. *Journal of Neuropathology Experimental Neurology*, 57(5):369–384, 5 1998.

[vMHJ+03] Christian von Mering, Martijn Huynen, Daniel Jaeggi, Steffen Schmidt, Peer Bork, and Berend Snel. String: A database of predicted functional associations between proteins, 1 2003.

[vPIAvdW21] Teije van Prooije, Norlinah Mohamed Ibrahim, Shahrul Azmin, and Bart van de Warrenburg. Spinocerebellar ataxias in asia: Prevalence, phenotypes and management, 11 2021.

[You23] Noria Yousufi. Functional profiling of alzheimer's through a consensus network, 2023.

[ZTZX11] Yun Wu Zhang, Robert Thompson, Han Zhang, and Huaxi Xu. App processing in alzheimer's disease, 2011.

A Figures

A.1 Individual C-PPINs

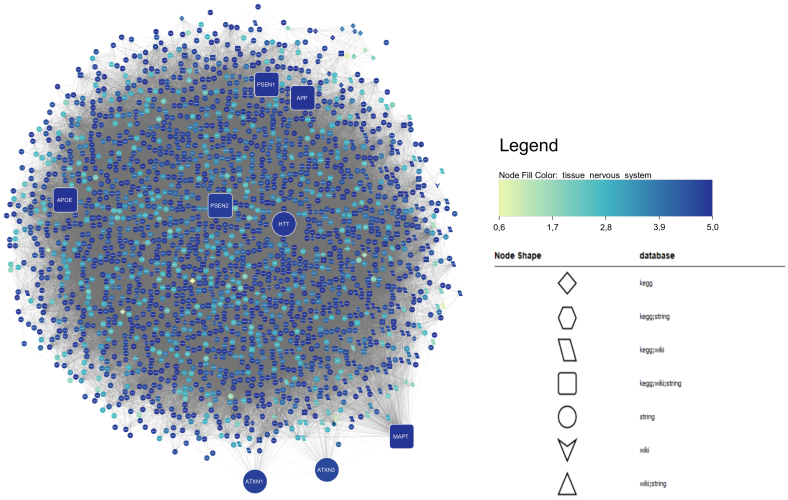


Figure 3: **AD C-PPIN** - Dark blue nodes represent proteins with higher certainty of activity in the nervous system. The shapes indicate the database sources, with the disease-related proteins APP, MAPT, PSEN1, PSEN2, APOE, HTT, ATXN1, and ATXN3 enlarged for emphasis. Proteins associated with AD were sourced from KEGG, WikiPathways, and STRING, while HTT, ATXN1, and ATXN3 were exclusively derived from STRING interactions.

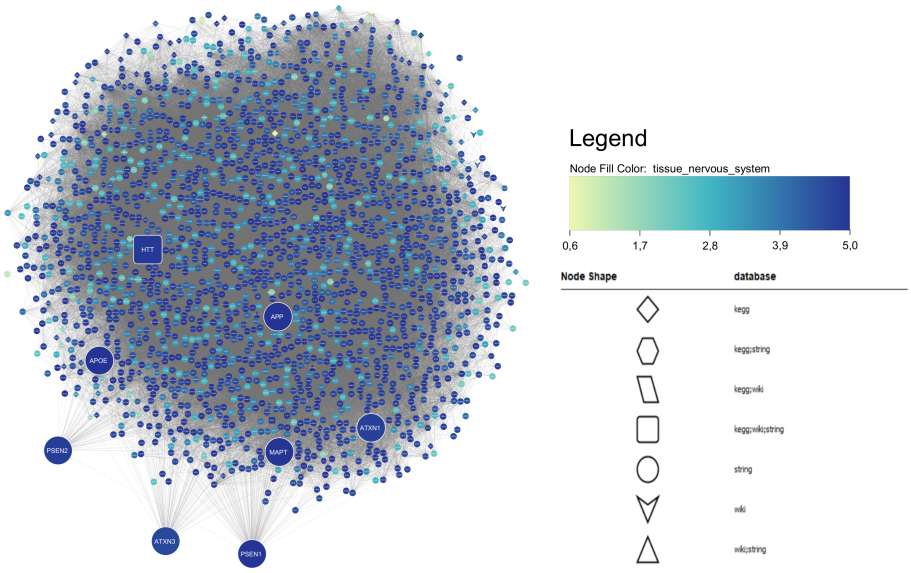


Figure 4: **HD C-PPIN** - Dark blue nodes represent proteins with higher certainty of activity in the nervous system. The shapes indicate the database sources, with the disease-related proteins APP, MAPT, PSEN1, PSEN2, APOE, HTT, ATXN1, and ATXN3 enlarged for emphasis. Proteins associated with HD were sourced from KEGG, WikiPathways, and STRING, while proteins associated with AD and SCA were exclusively derived from STRING interactions.

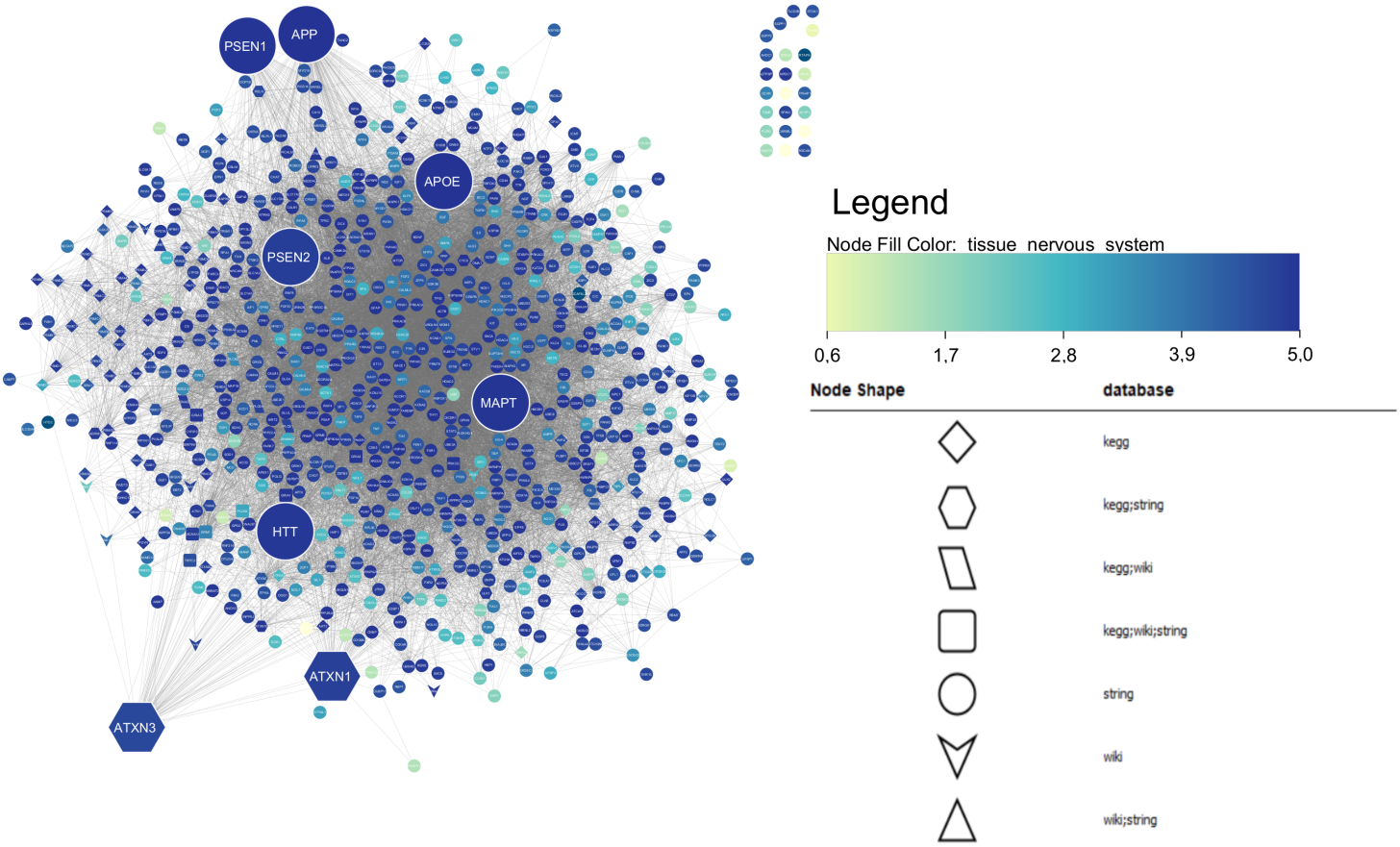


Figure 5: **SCA C-PPIN** - Dark blue nodes represent proteins with higher certainty of activity in the nervous system. The shapes indicate the database sources, with the disease-related proteins APP, MAPT, PSEN1, PSEN2, APOE, HTT, ATXN1, and ATXN3 enlarged for emphasis. Proteins associated with SCA were sourced from KEGG and STRING, while the proteins from HD and AD were exclusively derived from STRING interactions.

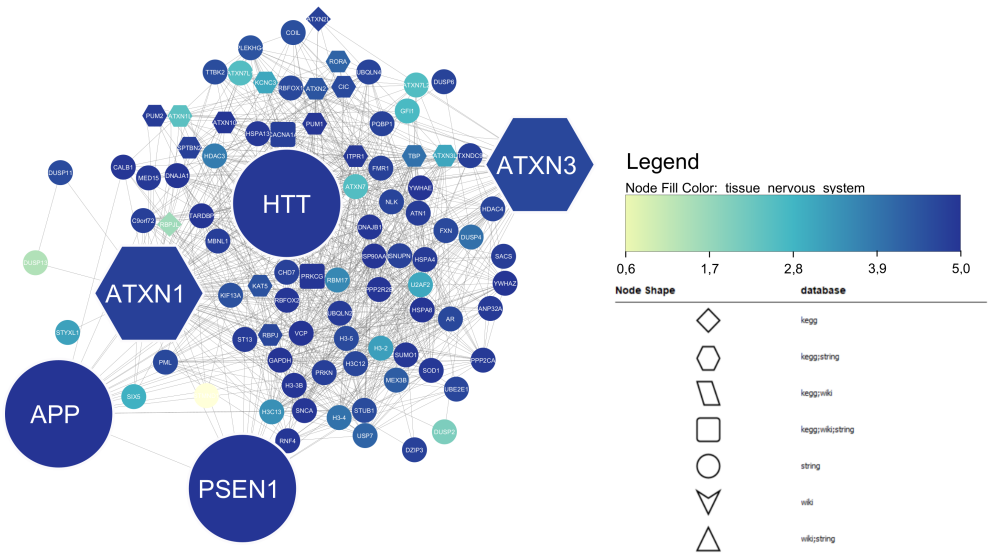


Figure 6: **SCA C-PPIN** - Dark blue nodes represent proteins with higher certainty of activity in the nervous system. The shapes indicate the database sources, with the disease-related proteins APP, MAPT, PSEN1, PSEN2, APOE, HTT, ATXN1, and ATXN3 enlarged for emphasis. Proteins associated with SCA were sourced from KEGG and STRING, while the proteins from HD and AD were exclusively derived from STRING interactions.

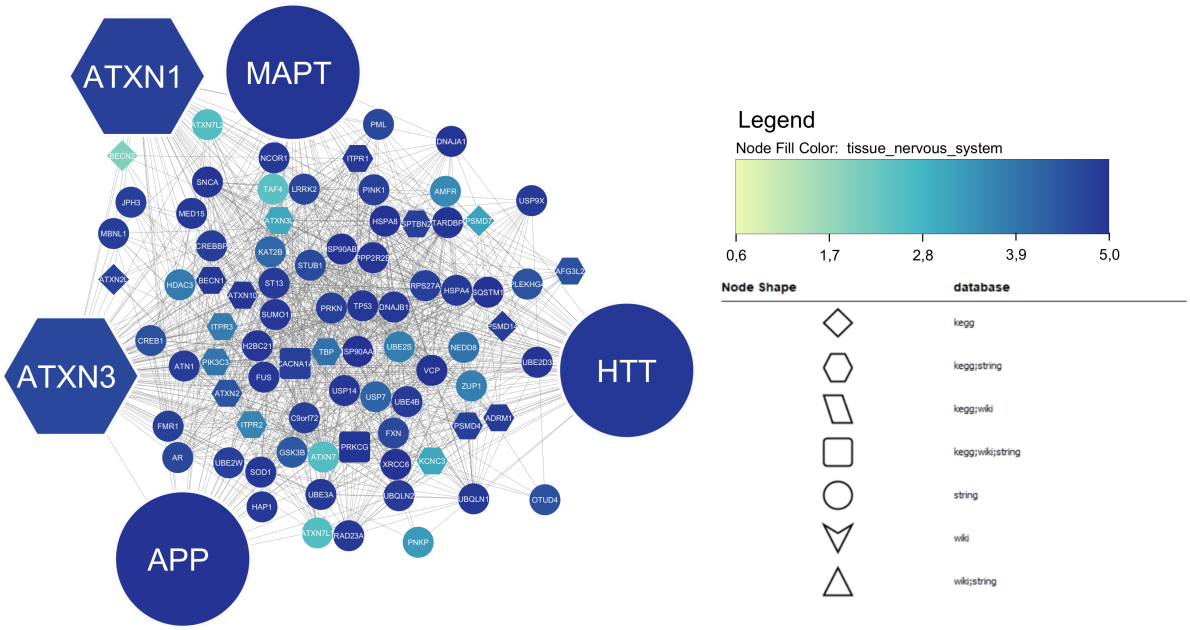


Figure 7: **SCA C-PPIN** - Dark blue nodes represent proteins with higher certainty of activity in the nervous system. The shapes indicate the database sources, with the disease-related proteins APP, MAPT, PSEN1, PSEN2, APOE, HTT, ATXN1, and ATXN3 enlarged for emphasis. Proteins associated with SCA were sourced from KEGG and STRING, while the proteins from HD and AD were exclusively derived from STRING interactions.

A.2 Merged C-PPINs

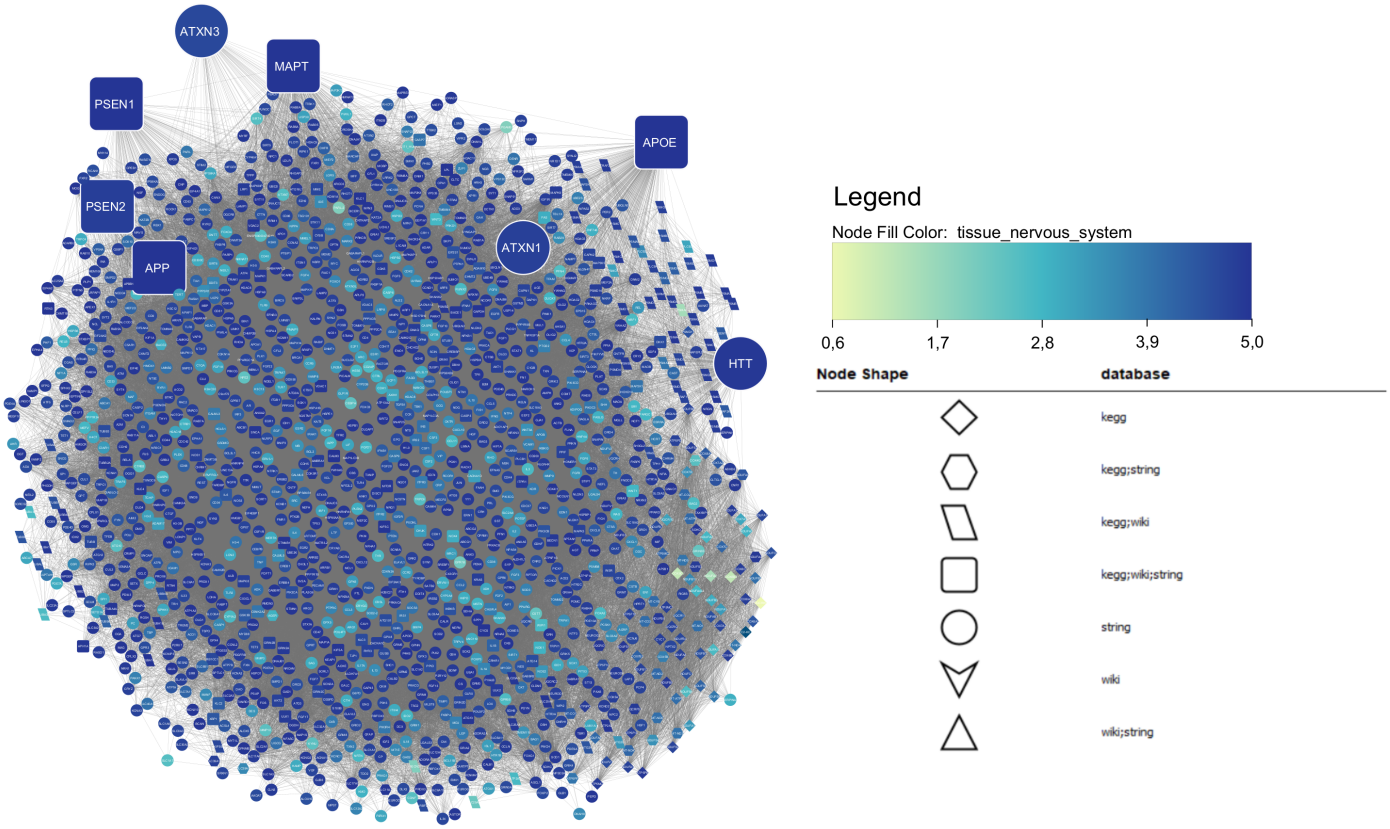


Figure 8: **AD ∩ HD C-PPIN** - Dark blue nodes represent proteins with higher certainty of activity in the nervous system. The shapes indicate the database sources, with the disease-related proteins APP, MAPT, PSEN1, PSEN2, APOE, HTT, ATXN1, and ATXN3 enlarged for emphasis. Proteins associated with AD were sourced from KEGG and STRING and WikiPathways, while the proteins from HD and SCA were exclusively derived from STRING interactions.

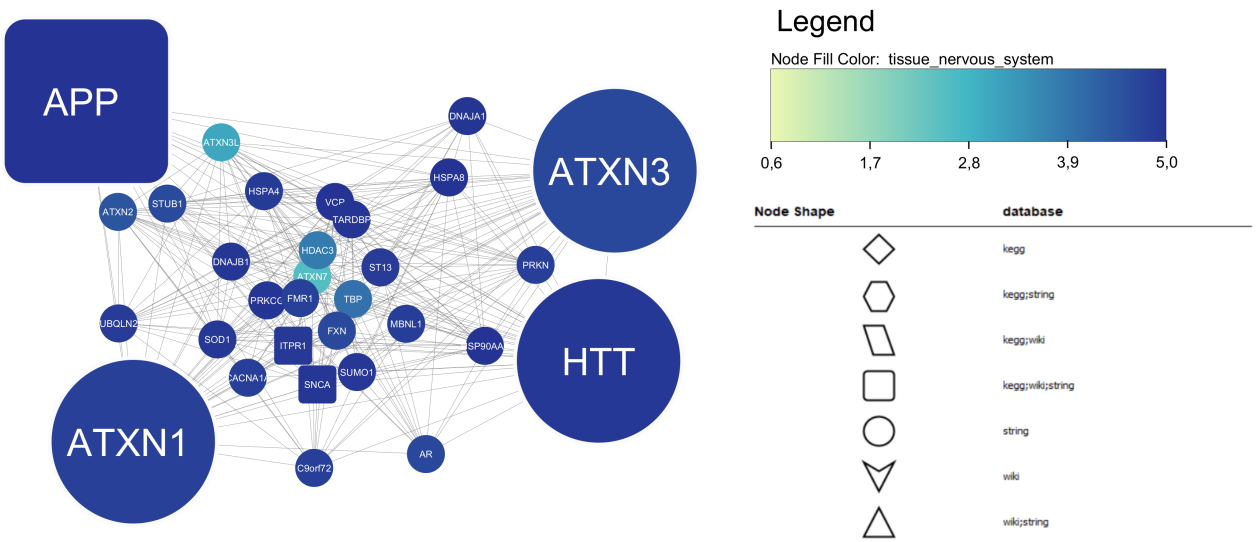


Figure 9: **AD ∩ HD ∩ SCA1 ∩ SCA3 C-PPIN** - Dark blue nodes represent proteins with higher certainty of activity in the nervous system. The shapes indicate the database sources, with the disease-related proteins APP, MAPT, PSEN1, PSEN2, APOE, HTT, ATXN1, and ATXN3 enlarged for emphasis. Proteins associated with AD were sourced from KEGG and STRING and WikiPathways, while the proteins from HD and SCA were exclusively derived from STRING interactions.

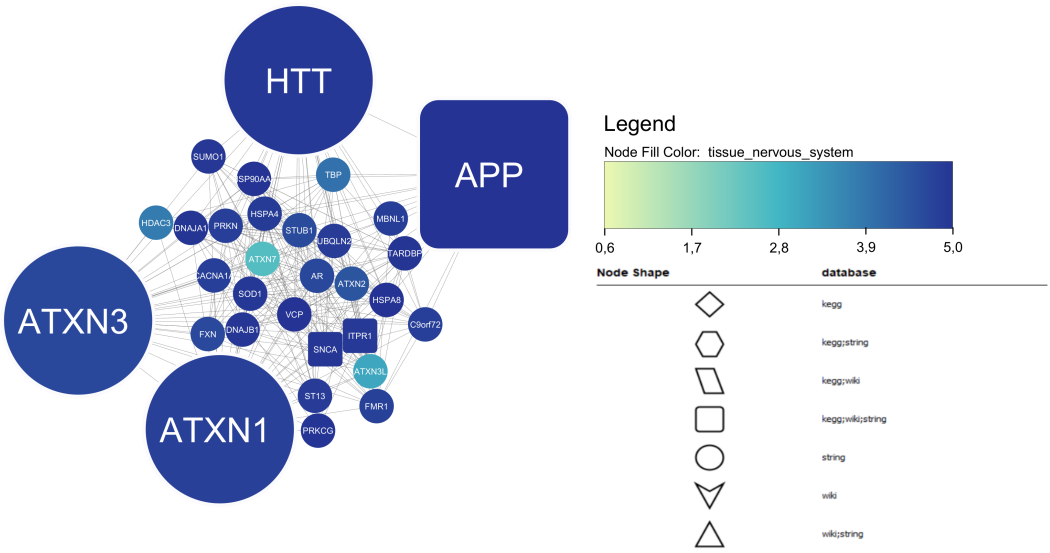
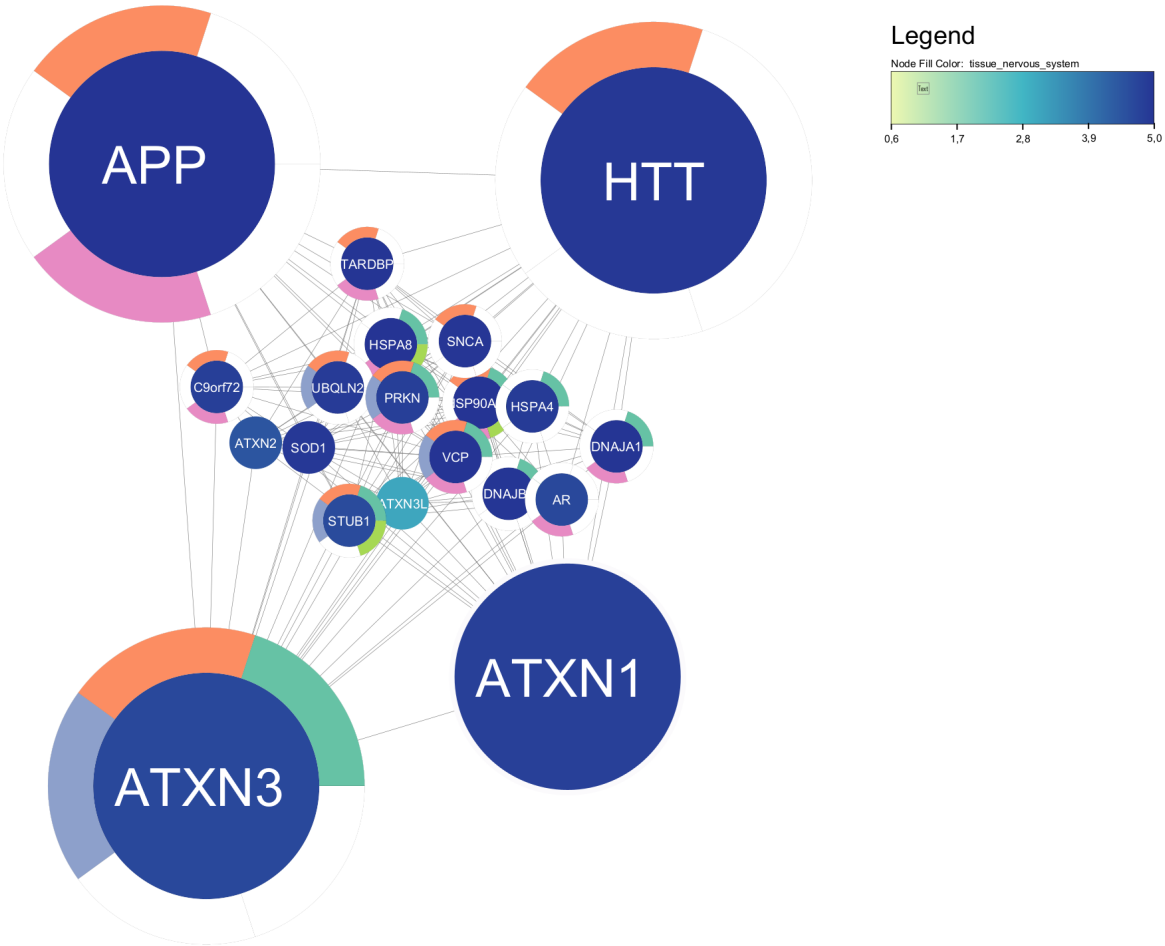


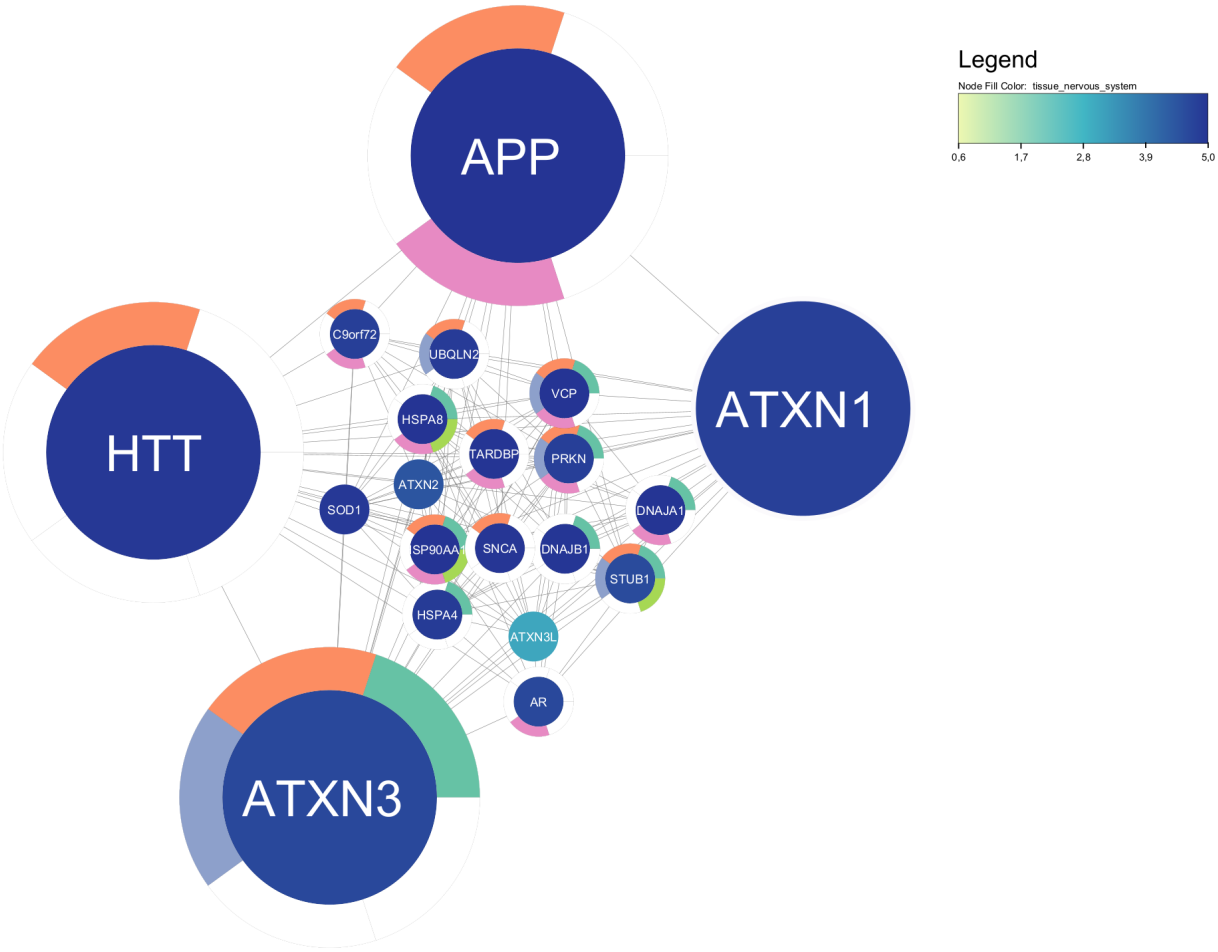
Figure 10: **AD ∩ SCA1 ∩ SCA3 C-PPIN** - Dark blue nodes represent proteins with higher certainty of activity in the nervous system. The shapes indicate the database sources, with the disease-related proteins APP, MAPT, PSEN1, PSEN2, APOE, HTT, ATXN1, and ATXN3 enlarged for emphasis. Proteins associated with AD were sourced from KEGG and STRING and WikiPathways, while the proteins from HD and SCA were exclusively derived from STRING interactions.

A.3 Clusters



category	chart color	term name	description	FDR value
GO Biological Process		GO:0035966	Response to topologically incorrect protein	1,97E-10
GO Biological Process		GO:0009896	Positive regulation of catabolic process	2,97E-9
GO Biological Process		GO:0036503	ERAD pathway	3,55E-5
GO Biological Process		GO:0032880	Regulation of protein localization	3,63E-5
GO Biological Process		GO:0061684	Chaperone-mediated autophagy	6,96E-5

Figure 11: **Cluster 1 AD ∩ HD ∩ SCA1 ∩ SCA3 C-PPIN** – Dark blue nodes indicate proteins with higher confidence of activity in the nervous system. Disease-related proteins (APP, MAPT, PSEN1, PSEN2, APOE, HTT, ATXN1, and ATXN3) are enlarged for emphasis. The five processes with the lowest FDR values from the STRING enrichment analysis are displayed.



category	chart color	term name	description	FDR value
GO Biological Process		GO:0035966	Response to topologically incorrect protein	1,97E-10
GO Biological Process		GO:0009896	Positive regulation of catabolic process	2,97E-9
GO Biological Process		GO:0036503	ERAD pathway	3,55E-5
GO Biological Process		GO:0032880	Regulation of protein localization	3,63E-5
GO Biological Process		GO:0061684	Chaperone-mediated autophagy	6,96E-5

Figure 12: **Cluster 1 AD ∩ SCA1 ∩ SCA3 C-PPIN** – Dark blue nodes indicate proteins with higher confidence of activity in the nervous system. Disease-related proteins (APP, MAPT, PSEN1, PSEN2, APOE, HTT, ATXN1, and ATXN3) are enlarged for emphasis. The five processes with the lowest FDR values from the STRING enrichment analysis are displayed.

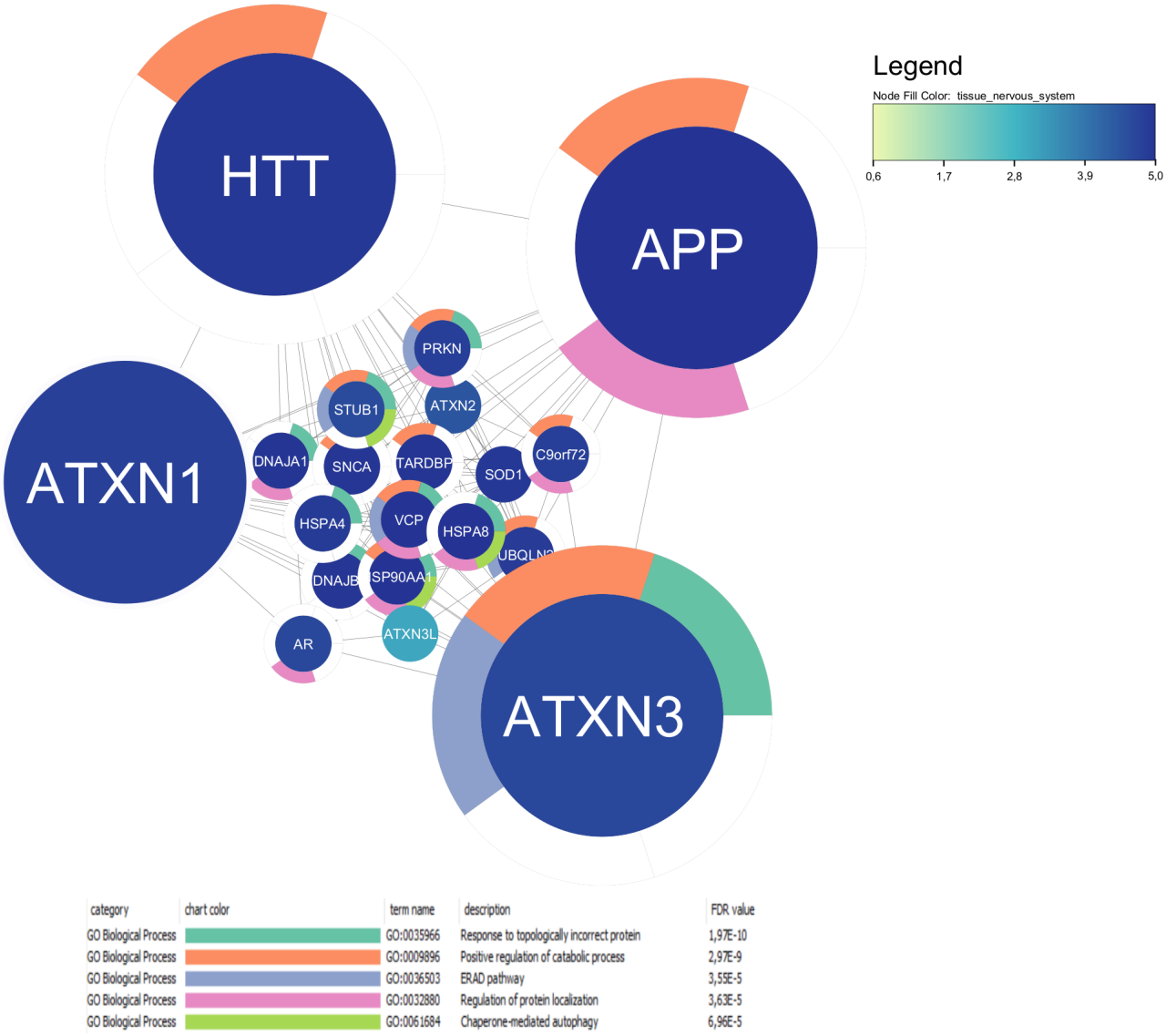


Figure 13: **Cluster 1 HD ∩ SCA1 ∩ SCA3 C-PPIN** – Dark blue nodes indicate proteins with higher confidence of activity in the nervous system. Disease-related proteins (APP, MAPT, PSEN1, PSEN2, APOE, HTT, ATXN1, and ATXN3) are enlarged for emphasis. The five processes with the lowest FDR values from the STRING enrichment analysis are displayed.

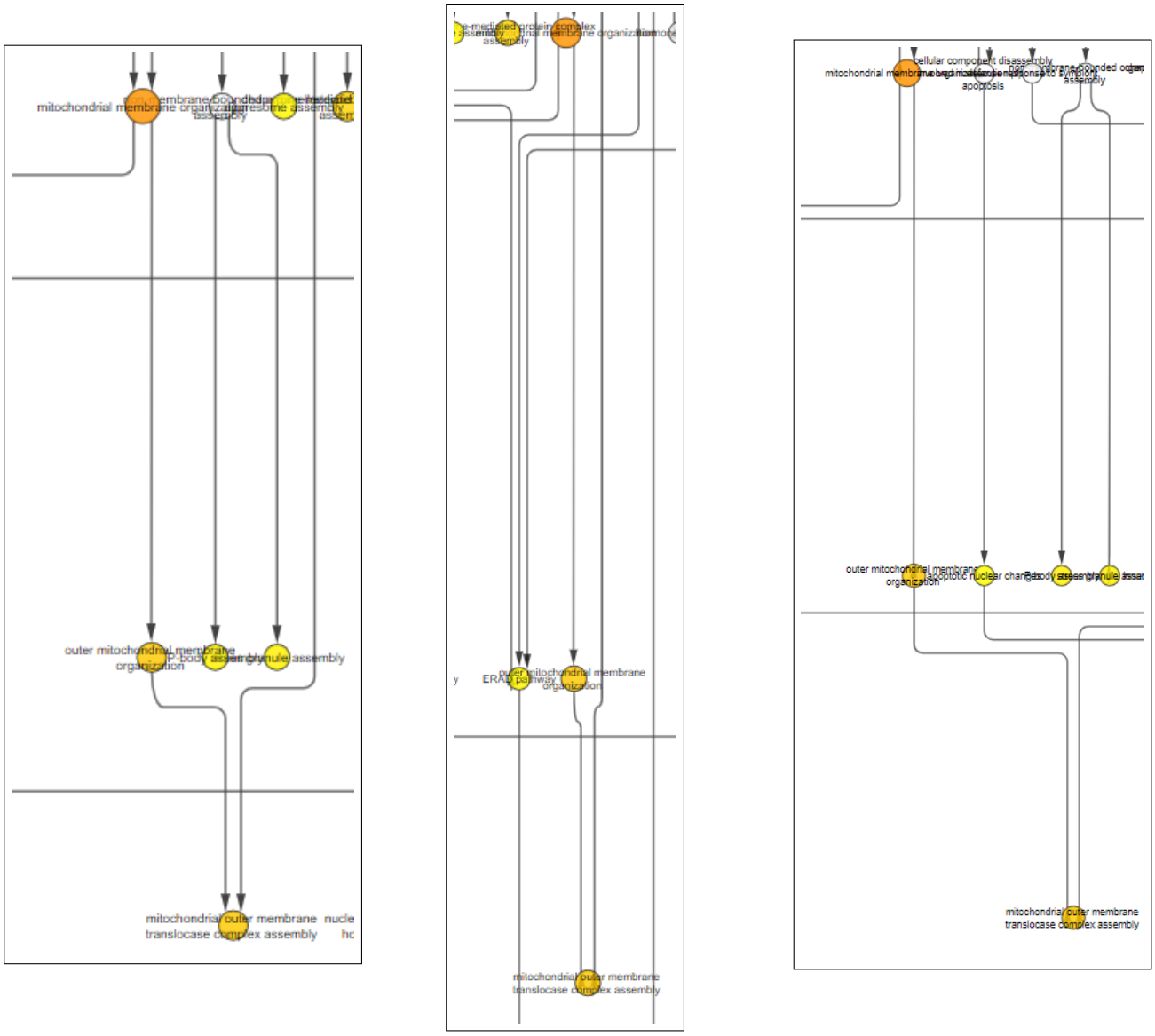


Figure 14: Enlarged BiNGO graph comparison of Cluster 1 across three C-PPINs. From left to right: AD ∩ HD ∩ SCA1 ∩ SCA3, AD ∩ SCA1 ∩ SCA3, and HD ∩ SCA1 ∩ SCA3. The GO Biological Process *mitochondrial outer membrane translocase complex assembly* is notably more enriched across all three clusters.