Universiteit Leiden
The Netherlands

Opleiding Informatica

Integrating Protein-Protein Interactions and Genetic Associations in Huntington's Disease

Suzanne Honders

Supervisors:
Dr. K.J. Wolstencroft & Dr. Lu Cao

BACHELOR THESIS

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

28/05/2025

Abstract

Huntington's disease (HD) is a genetically defined neurodegenerative disorder caused by a CAG repeat expansion in the HTT gene, yet its molecular pathology involves a far more complex network of interactions beyond the primary mutation. This thesis investigates the role of protein-protein interactions in the context of HD by constructing and analyzing a protein-protein interaction network (PPIN) using Cytoscape and data from the STRING database. Subnetworks centered around the huntingtin protein and other HD-relevant proteins were analyzed to identify structural patterns and key interactors. Additionally, the structure of the PPIN was used to integrate and explore data from a genome-wide association study (GWAS) on HD. This network-based approach provides a structured view of the HD-related protein landscape, summarizing key relationships between associated proteins.

Contents

1 Introduction

1.1 Huntington's Disease

Huntington's disease (HD) is a rare inherited neurodegenerative disease. It is caused by an autosomal dominant allele. The genetic abnormality is located on chromosome 4. The condition arises from an elongated trinucleotide repeat of CAG (whose length varies) within the HTT gene which is responsible for producing the huntingtin protein [BDG+15]. In individuals with HD, this repeat expansion leads to the production of a mutant form of the protein, known as mHtt (mutant huntingtin). This version of huntingtin contains an abnormally long polyglutamine sequences. The effect of the elongated polyglutamine sequences on the protein is the gain of toxic qualities and an increased risk of early breakage [BDG+15]. The number of CAG repeats correlates with disease risk and onset. Repeat lengths can be categorized as follows;

- Normal: 26 or fewer repeats – no risk of developing HD.
- Intermediate: 27–35 repeats – not associated with disease symptoms, but the repeat size may expand in offspring, potentially increasing their risk [Sem06].
- Pathogenic: 40 or more repeats – strongly associated with the development of HD in the individual [CN11].

1.1.1 Clinical Features of HD

HD symptoms are commonly categorized into three groups: motor, cognitive, and psychiatric. Individuals who carry the mutation that causes HD typically begin to develop symptoms in mid-adulthood. However, in some cases, symptoms can appear before the age of 20, a form known as juvenile Huntington's disease [Roo10]. The course of the disease is progressive and irreversible [GT18]. While medications and treatments are available to help manage symptoms, they do not prevent the ongoing decline in physical, cognitive, and behavioral functions [Cli17].
One of the common movement disorders associated with HD is chorea, characterized by involuntary, uncontrolled movements that appear sudden and unpredictable [Cli17]. These movements can affect various parts of the body, including the face, limbs, and trunk. Early signs may resemble restlessness or clumsiness, but as the disease advances, chorea often becomes more severe, interfering with daily activities such as walking or speaking [Roo10].
Cognitive impairments in HD often affect executive functions, including planning, organizing, and multitasking [oNDN21]. Memory and attention are also frequently impacted. Psychiatric symptoms commonly accompany HD and may even precede motor signs. Depression, irritability, anxiety, and apathy are prevalent, with some individuals also experiencing obsessive-compulsive behaviors or psychosis [BOW+08].

1.1.2 Molecular Mechanisms of HD

For this research, we will explore the protein-protein interactions associated with HD. While HD is known to be caused by a genetic mutation in the HTT gene, the exact molecular mechanisms by which this mutation leads to the wide range of symptoms are still not fully understood. In particular, the role of protein-protein interactions involving the mutant huntingtin (mHtt) protein remains underexplored [KF16]. By investigating how huntingtin interacts with other cellular proteins, the

aim is to uncover relevant genes, potential pathways and mechanisms that contribute to the disease's onset and progression. Gaining a deeper understanding of these interactions may offer valuable insights into the molecular mechanisms underlying HD.

1.2 Cytoscape

To visualize the protein-protein interactions associated with HD, a network was created using Cytoscape. Cytoscape is a widely used open-source platform for visualizing, analyzing, and integrating complex biological networks. It enables users to generate subnetworks and supports various analytical tools and plugins, including STRING enrichment analysis [Cyt23].

1.3 Genome-wide association study

Studies on HD use age at onset as a quantitative phenotype in genetic analysis to find HD modifiers. However, this is not ideal, age of onset is difficult to define clearly, and not always recorded [MD17]. The challenge arises because HD symptoms develop gradually and vary in type; motor, cognitive, or psychiatric making it difficult to determine when the disease truly begins [PLS+08]. To address this issue, the studies that created the GWAS (genome-wide association study) catalog was to create a better measure of how the disease progresses over time and use this new measure to search for genetic factors that influence how fast or slow HD progresses [MD17].
The research found that a region on chromosome 5 showed a statistically significant signal. This region includes the genes MSH3, DHFR, and MTRNR2L2. The strongest signal came from a single SNP in the MSH3 gene [MD17]. In this thesis, genes identified through GWAS were mapped onto a protein-protein interaction network to investigate how genetic risk loci relate to network structure and may contribute to HD pathology.

1.4 Related work

In previous years, other students from Leiden University conducted research into PPINs to better understand the molecular mechanisms underlying HD. One such study, by Chen Ji Rong Jiang, compared different databases to identify shared and unique interactions involving the huntingtin protein [Jia22]. The study applied enrichment analysis to confirm known HD-related processes such as oxidative phosphorylation and suggested that lesser-known areas of the network could offer novel research directions.
Another relevant study was conducted by Nina Anna Maria Henninger, who constructed a PPIN using data from KEGG, STRING, and WikiPathways, and incorporated gene expression information (Bachelor Thesis, Nina Henninger, Bioinformatics Bachelor, Leiden University, 2022/2023). Her analysis highlighted disrupted processes in HD, such as the regulation of biological quality, and proposed a possible link between the MYC gene and HD.
These previous studies laid important groundwork by building and analyzing HD-related PPINs. In contrast, this thesis integrates GWAS data with tissue-specific filtering to investigate whether combining genetic modifiers with brain-specific context can reveal novel insights into the molecular mechanisms of Huntington's disease.

1.5 Research question

Can a consensus protein-protein interaction network in Huntington's disease provide new insights into the disease mechanism through integration of genome-wide association study data?

1.6 Thesis overview

This bachelor thesis project is part of the bachelor "Bioinformatica" at LIACS and was supervised by Dr. K.J. Wolstencroft and Dr. Lu Cao.

This chapter contains the introduction; Section 2 includes the definitions; Section 3 outlines the methodology used to obtain the results; section 4 presents the results; Section 5 provides the conclusions; And section 6 offers suggestions for future research. All images that are used in this bachelor thesis can be found on GitLiacs: https://git.liacs.nl/s2662620/bachelor-thesis-2662620.

2 Definitions

2.1 PPIN

Proteins serve various roles in biomolecular systems, functioning as sensors, transporters, and structural components. Interactions between these proteins, protein-protein interactions, enable dynamic adaptation to changing environmental circumstances [Bar11]. The interactions between all proteins within a biological system are described in the protein-protein interaction network (PPIN). Even slight alterations in PPINs can have major consequences for the system and may lead to disease phenotypes [Bar11].

2.2 Nodes, edges, hubs

In PPINs, proteins are represented as nodes, and the interactions between proteins are shown as edges connecting these nodes. The number of edges per node, also known as degree, varies significantly across the network. While many proteins show low connectivity, a small number of nodes are highly connected [HZ06]. These highly connected nodes are referred to as hubs. Hubs often play central roles in maintaining the structure and function of the network. Their positioning, connectivity patterns, and associated biological functions can provide valuable insights into disease mechanisms and are therefore of particular interest for this research.

2.3 Enrichment analysis

Enrichment analysis can be performed on gene sets. Genes encode gene products, mainly proteins but non-coding RNA molecules as well. These products serve functions at the molecular, cellular, and organismal levels [CAea23]. Enrichment analysis is used to find pathways which are more enriched as would be expected to happen by chance [CAea23]. Studying upregulated and downregulated genes is particularly useful, as these may be associated with disease phenotypes.

2.4 Gene Ontology

The Gene Ontology (GO) is a standardized representation of biological knowledge. GO is divided into three categories; molecular function, cellular component, and biological process [AM00].

- Biological process refers to the broader biological objective or process that the gene product is involved in. For example cell division or immune response.
- Molecular function refers to the specific biochemical activity of a gene product, such as enzyme activity or binding to a particular molecule.
- Cellular component describes where in the cell the gene product is located or active. For example nucleus or plasma membrane.

By integrating these three categories, the GO enables researchers to create a comprehensive map of gene functions and their roles within cellular and organismal biology. This, in turn, supports a wide array of scientific applications, including functional genomics, systems biology, and the interpretation of experimental data [AM00].

2.5 Reactome

Reactome is an openly accessible and collaboratively developed database that compiles information on biological pathways for a wide range of normal and disease-related biological processes. It aims to support research and education by providing advanced tools for visualizing, analyzing, and interpreting complex biological processes [Mea23].

2.6 Wikipathways

WikiPathways is an open, collaborative platform for collecting and curating biological pathways. It allows researchers and the broader scientific community to contribute, edit, and share pathway information, supporting the visualization and analysis of molecular interactions involved in both normal biological functions and disease processes [Aea23].

3 Methods

3.1 Data Gathering

Data from different sources were used to later filter and generate subnetworks.
Data sources used to collect data:

- STRING app disease database [Szk24] (version 11.5) — `https://string-db.org`
- GWAS Catalog [Bun24] (version from 2023) — `https://www.ebi.ac.uk/gwas/`
- Reactome [Gil24] (version 78, released 2024-01) — `https://reactome.org`
- WikiPathways [Mar24] (version 20230510) — `https://www.wikipathways.org`

The STRING app disease database was used to create the initial network. The data from GWAS catalog, Reactome and Wikipathways were used for comparison to the initial network.

3.2 Network Creation

The network was constructed using the STRING application integrated within Cytoscape. The data used to create the initial network were imported from public databases. To obtain the interaction data, the 'STRING: disease query' option was selected as the data source, with 'Huntington's disease' entered as the disease term. A full STRING network was chosen for the network type. The confidence score cutoff was set to 0.40, which is the default threshold. The maximum number of proteins was kept at the default value of 200. Using these settings, the network was imported into Cytoscape. The resulting initial HD network comprised a total of 1,200 nodes and 83,019 edges.
Following the import of the initial network, the network was extended to the desired size within the menu of the STRING application. The number of additional interactors to expand the network by was set to 1,000. This number was chosen to strike a balance between including too few nodes—potentially omitting relevant interactions—and excessive expansion, which could introduce noise and dilute meaningful disease-specific relationships. This approach aligns with recommendations from Szklarczyk et al. (2019) [SGL+19]. The organism selected for the interacting proteins was Homo sapiens. The selectivity parameter was kept at its default value of 0.5.
The network includes a node table and an edge table, which provide detailed information about the proteins (nodes) and the interactions between them (edges), respectively. The constructed network was analyzed as an undirected graph in the tools menu within the Cytoscape application, to characterize both global and node-specific properties. For each node, the degree (the number of undirected edges connected to a node) and radiality were calculated. At the network level, overall structural parameters were also determined, including the network diameter and network radius.

3.3 Network filtering

To identify meaningful patterns and structural properties within the HD network, the initial network was subjected to a series of filtering steps. These filtering steps aimed to reduce network complexity and focus the analysis on biologically relevant subnetworks, such as genes associated with the CNS or those related to the findings of the GWAS study.
The following subnetworks were created by network filtering:

- Subnetwork of first neighbors of HTT
- Subnetwork of second neighbors of HTT
- Subnetworks filtered by genes from the GWAS dataset
- Subnetworks filtered by nervous system–specific genes

3.3.1 Filtering for Central nervous system

To focus the analysis on genes relevant to the central nervous system (CNS), the initial HD network was filtered based on CNS-specific gene activity. The CNS, comprising the brain and spinal cord, is the primary site affected in HD. This focus helps identify proteins that are most relevant to disease mechanisms and progression in the affected tissue [MT18]. The node table in the network includes a column labeled "tissue nervous system," which provides a CNS-specific activity score for each protein. This score, ranging from 0 (no activity) to 5 (high activity), was used to filter proteins based on CNS relevance. Two CNS-specific subnetworks were constructed by applying

different thresholds to this score. The first subnetwork applied a threshold of 4.5 to include genes with moderate to high CNS activity. A stricter threshold of 4.9 was used for the second subnetwork, capturing only genes with strong CNS-specific activity. These thresholds were selected based on exploratory analysis. A threshold of 4.0 generated a large, less specific network (860 nodes), while a threshold of 5.0 yielded only 9 nodes—potentially excluding relevant proteins. The chosen thresholds of 4.5 and 4.9 thus represent broader and narrower CNS-specific subnetworks, respectively. The resulting subnetworks were analyzed to compare differences in network structure, including node counts, edge density, and the presence of key genes such as HTT, as well as to examine how varying the CNS activity threshold affects network composition.

3.3.2 Filtering for GWAS Catalog data

In the GWAS Catalog, one study with associated dataset was found to be associated with HD. These GWAS findings were compared to genes present in the initial HD network. TThis comparison revealed that MSH3 and GFRA1 were present in both datasets. To explore their roles within the network, three subnetworks were constructed. The first subnetwork included MSH3 and its first-degree neighbors. The second subnetwork included GFRA1 and its first-degree neighbors.The third subnetwork encompassed both MSH3 and GFRA1, along with nodes connecting them. These subnetworks were analyzed to assess the presence and positioning of key genes like HTT and to identify network hubs. The hubs identified in each subnetwork were compared to those in the initial HD network to determine if central genes were preserved or if distinct hub profiles emerged.

3.4 Reactome

The Reactome database was queried using the term "HTT" to identify relevant molecular events. Because Reactome is structured around molecular entities and reactions, "HTT" is a suitable query for identifying specific interactions. This search returned results across multiple categories, including protein, reaction, complex, and pathway. Among these, the reaction, complex, and pathway categories contained entries involving the MECP2 gene, suggesting a potential interaction between HTT and MECP2. In the reaction category, one result indicated a direct binding event where MECP2 binds to HTT (see Figure 10, Appendix) In the complex category, the result identified a molecular complex formed by MECP2 and HTT, labeled as "MECP2:HTT". The consistent appearance of MECP2 across result types highlights its potential relevance in HTT-associated pathways.

3.5 Wikipathways

The WikiPathways database was queried using the term "Huntington's disease" to identify relevant biological pathways. Since WikiPathways frequently annotates pathways with disease names, this broader term was considered appropriate. This search returned multiple results, including pathways that are specifically focused on HD as well as others that are more broadly related to the CNS or other neurodegenerative disorders. Among the results, two pathways were identified as being directly related to Huntington's disease: "Effect of omega-3 PUFA on Huntington's disease pathways" (see Figure 12, Appendix) and "ERK pathway in Huntington's disease" (see Figure 11, Appendix).

4 Results

4.1 Analysis initial HD network

The initial HD network, constructed from STRING database data, contains 1,200 nodes and 83,019 edges, indicating a high level of connectivity within the network. Among all nodes in the network, AKT1 was identified as the node with the highest degree, having a total of 822 direct interactions. Figure 1 provides a visual representation of the initial HD network, highlighting the top ten network hubs in distinct colors. Notably, these hubs are located in close proximity to one another within the network. The ten network hubs identified in the initial HD network are: AKT1, GAPDH, ACTB, TP53, TNF, BCL2, INS, CASP3, IL6, and MYC. In Table 1, the top ten network hubs nodes are listed, along with their ranking, display name, shared name, and degree. Figure 2 presents a visual representation of the hubs identified in the initial HD network, with node sizes scaled by degree. The figure illustrates that these hubs are highly interconnected.

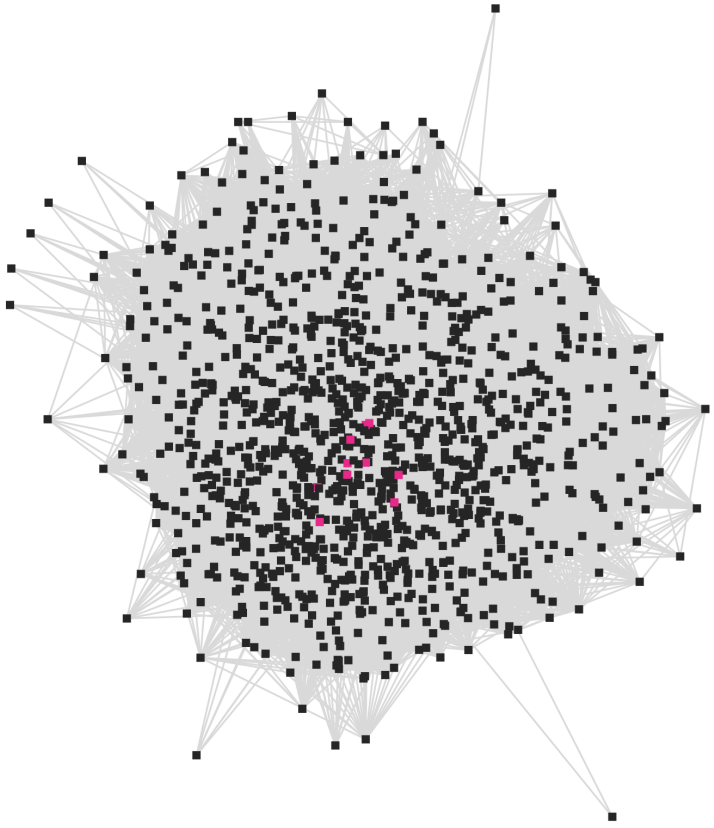

Figure 1: Initial HD PPIN with top ten hubs colored (STRING)

Table 1: Top ten hub proteins in the initial HD network, ranked by degree

Top ten highest connected nodes			
Ranking	Display name	Shared name	Degree
1	AKT1	9606.ENSP00000451828	822
2	GAPDH	9606.ENSP00000380070	805
3	ACTB	9606.ENSP00000494750	789
4	TP53	9606.ENSP00000269305	744
5	TNF	9606.ENSP00000398698	656
6	BCL2	9606.ENSP00000381185	630
7	INS	9606.ENSP00000380432	611
8	CASP3	9606.ENSP00000311032	606
9	IL6	9606.ENSP00000385675	599
10	MYC	9606.ENSP00000478887	594

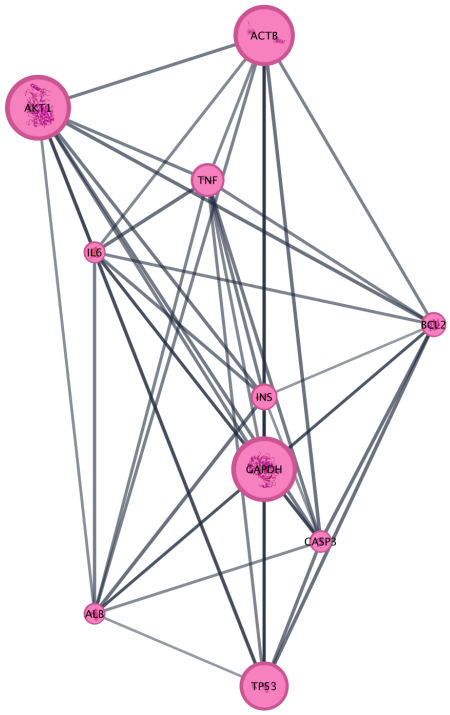

Figure 2: Top ten hubs from initial HD PPID together with their connecting edges (STRING)

4.1.1 HTT and the initial network

The huntingtin protein directly associated with the development of HD, has a degree of 275 within the initial HD network. Indicating that it directly interacts with 275 other proteins. Figure 3 illustrates a subnetwork comprising HTT and its first-degree neighbors. Consequently, this network contains a total of 276 nodes (the HTT protein and its 275 direct interactors) and includes 8,617 edges. The HTT first-neighbor network has an edge density of 0.226 (8,617 edges among 276 nodes), while the initial HD network has an edge density of 0.115 (83,019 edges among 1,200 nodes). This suggests a more densely interconnected structure in HTT's immediate interaction environment. In addition to these direct connections, HTT has 1,198 second-degree neighbors, which are proteins connected indirectly through one intermediate node. Despite its biological significance in the pathology of HD, HTT does not rank among the top ten highest-degree nodes in the initial HD network. Interestingly, the subnetwork composed of HTT and its second-degree neighbors closely mirrors the size of the initial HD network, with 1,199 nodes compared to 1,200 in the initial HD network. This subnetwork contains 83,012 edges, closely matching the full network's 83,019 edges. Figure 4 depicts the subnetwork of HTT and its second-degree neighbors.

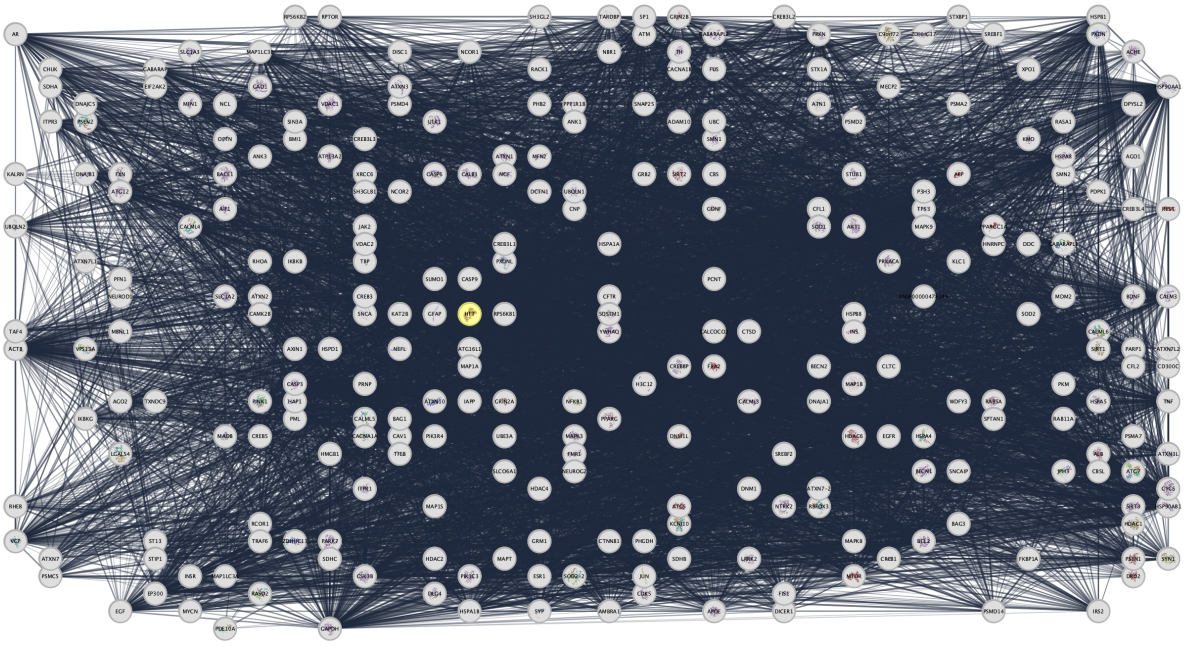

Figure 3: HD PPIN filtered for first-degree neighbors of HTT (STRING)

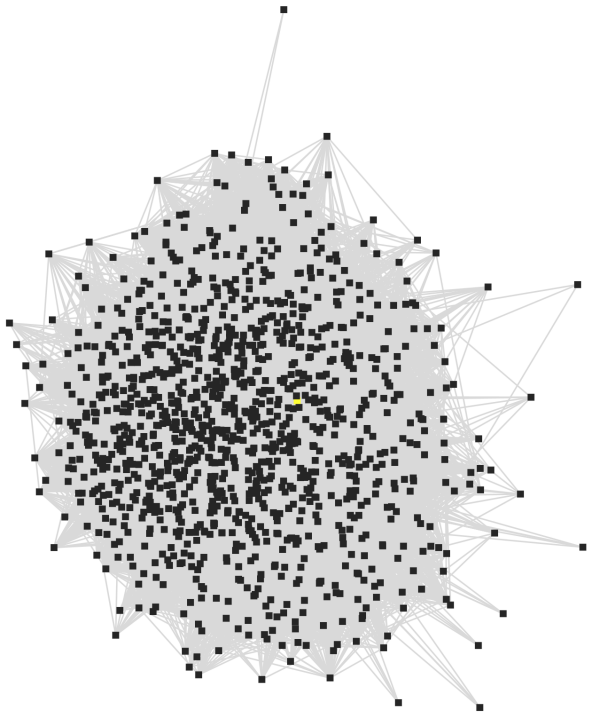

Figure 4: PPIN HD filtered for second-degree neighbors of HTT (STRING)

4.2 Analysis of the GWAS data

One study with a corresponding dataset related to HD is available in the GWAS Catalog. From this GWAS data, two overlapping genes, MSH3 and GFRA1, were identified within the initial HD network. These overlapping genes suggest a potential genetic association supported by both GWAS and protein-protein interaction data, highlighting their relevance to HD pathology. MSH3 has 23 first-degree neighbors in the initial HD network, while GFRA1 has 62 first-degree neighbors.
MSH3 and GFRA1, previously identified as common between the initial HD network and the GWAS dataset, were further analyzed with respect to HTT. These proteins are not direct (first-degree) neighbors of HTT; rather, they are second-degree neighbors, indicating an indirect interaction through one intermediary node. A subnetwork consisting of MSH3 and its first neighbors contains 24 nodes and 171 edges; this network is shown in Figure 5. Similarly, the subnetwork composed of GFRA1 and its first neighbors consists of 63 nodes and 1,106 edges and is displayed in Figure 6. Additionally, a focused subnetwork was constructed combining MSH3, GFRA1, and their connecting nodes. This subnetwork comprises 84 nodes and 1,487 edges. As expected, the HTT gene is not included in this subnetwork, due to the absence of direct connections with either MSH3 or GFRA1. This subnetwork is shown in Figure 7. Interestingly, despite the exclusion of HTT, the top ten hub genes (Table 2) in the subnetwork focused on MSH3, GFRA1, and their connecting components show overlap with those identified in the initial HD network (Table 1). Specifically, AKT1, TP53, and MYC are among the top ten most highly connected nodes in both the initial HD network and the subnetwork. This suggests that certain hub genes maintain a central role in the overall network structure, even when analysis is restricted to proteins indirectly associated with HTT.

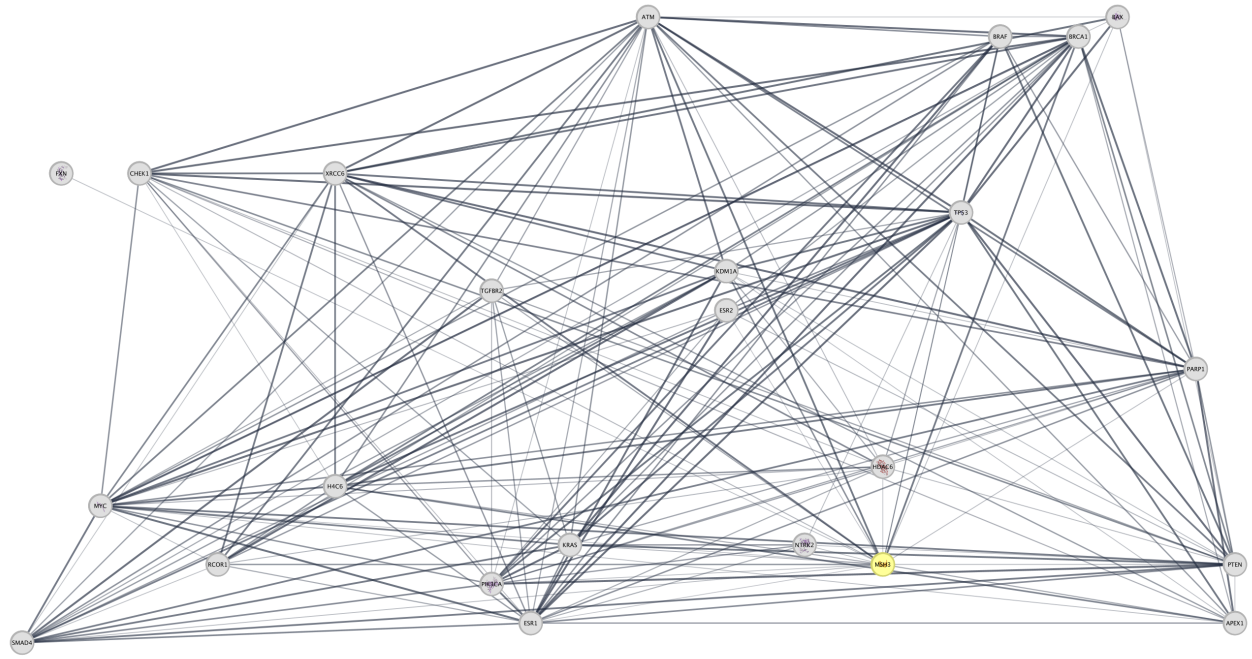

Figure 5: HD PPIN filtered for first-degree neighbors of MSH3 (STRING)

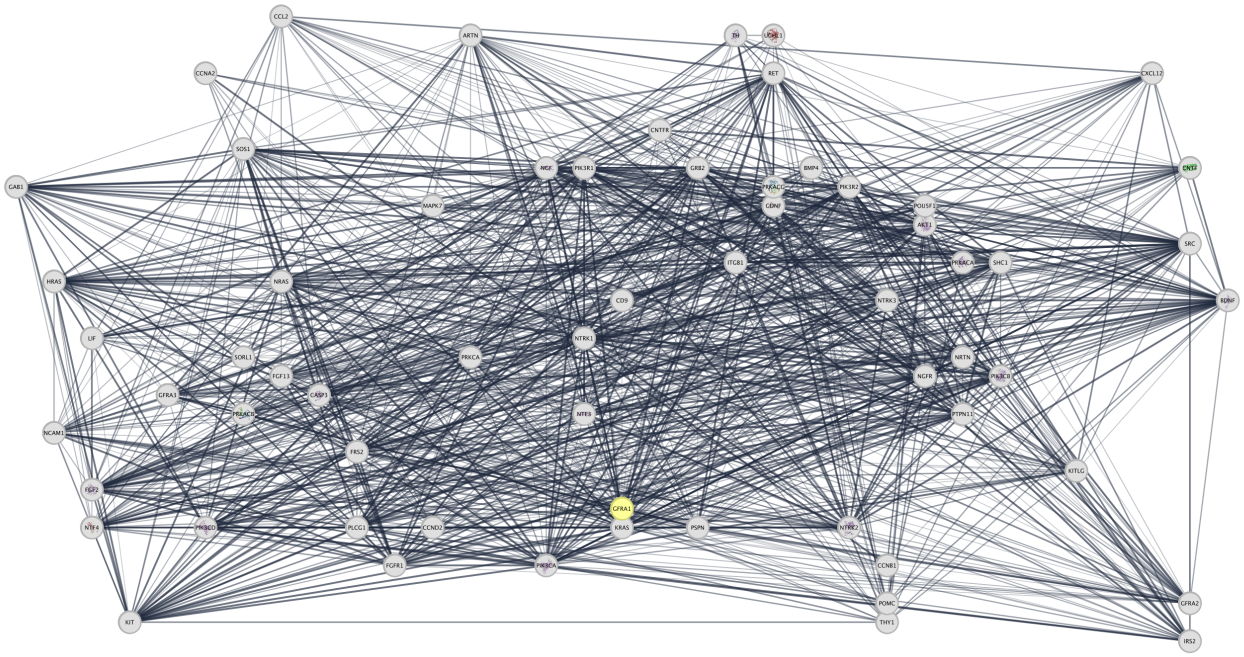

Figure 6: HD PPIN filtered for first-degree neighbors of GFRA1 (STRING)

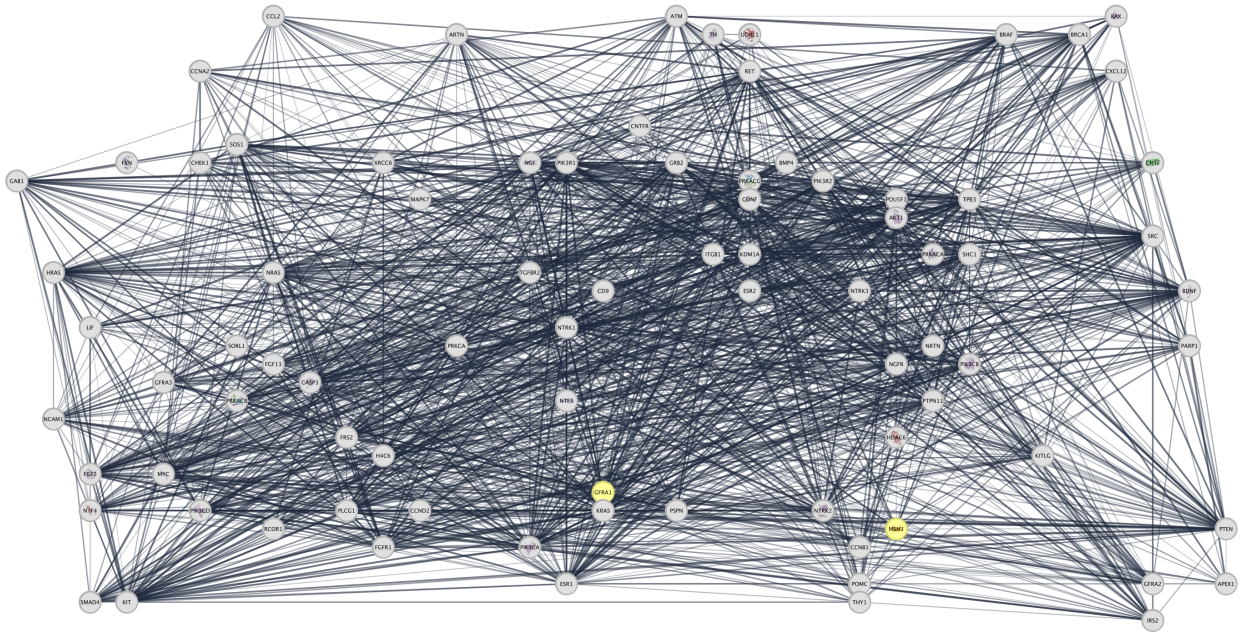

Figure 7: HD PPIN filtered for shared first-degree neighbors of MSH3 and GFRA1 (STRING)

Table 2: Top ten most connected nodes in the subnetwork of first-degree neighbors of MSH3 and GFRA1

Top ten highest connected nodes			
Ranking	Display name	Shared name	Degree
1	TP53	9606.ENSP00000269305	66
2	AKT1	9606.ENSP00000451828	64
3	GFRA1	9606.ENSP00000347591	62
4	SRC	9606.ENSP00000362680	61
5	KRAS	9606.ENSP00000256078	60
6	PTEN	9606.ENSP00000361021	58
7	MYC	9606.ENSP00000478887	58
8	PIK3CA	9606.ENSP00000263967	57
9	BDNF	9606.ENSP00000414303	57
10	NGF	9606.ENSP00000358525	56

4.2.1 STRING enrichment analysis

A STRING enrichment analysis was performed on the subnetwork constructed from MSH3, GFRA1, and their connecting nodes. The results revealed several significant GO terms associated with biological processes, cellular components, and molecular functions for both genes.

For **GO biological process**, the enriched terms were:

- MSH3: cellular response to stimulus, system development and multicellular organism development
- GFRA1: cellular response to stimulus, system development, multicellular organism development and enzyme-linked receptor protein signaling pathway.

For **GO cellular component**, the enriched terms were:

- MSH3: protein containing-complex
- GFRA1: protein containing-complex, axon and somatodendritic compartment

For **GO molecular function**, the enriched terms were:;

- MSH3: protein binding and enzyme binding
- GFRA1: protein binding and signaling receptor binding

MSH3 was identified as the most significant genetic modifier of HD progression in the GWAS study by Moss et al. It was found to have a strong association with somatic expansion of the CAG repeat in HTT[MD17]. The enriched GO terms for MSH3, including "cellular response to stimulus" and "protein binding," are consistent with its known role in DNA mismatch repair and stress response mechanisms implicated in HD. Although GFRA1 was not identified as significant in the same study, it was mentioned among sub-threshold signals. The enriched functions for GFRA1, such as neurodevelopmental signaling and axonal localization, suggest potential roles in neuronal maintenance or degeneration in HD, although Moss et al. did not highlight it as a primary modifier [MD17].

4.3 Central nervous system

Two subnetworks were created to focus on CNS relevance. The first subnetwork was generated using a threshold of 4.5, while the second subnetwork applied a stricter threshold of 4.9.

4.3.1 Subnetwork filtered nervous system 4.5

This subnetwork contained 768 nodes and 34,986 edges. This subnetwork is shown in Figure 8. Node coloring is based on the nervous system activity score. A gradient from green to red is used to visually represent relative CNS relevance, with green representing lower scores and red indicating higher CNS relevance. HTT was included in this filtered network, indicating its relevance within the nervous system context. Moreover, GFRA1 was also present in the network. In contrast, MSH3 was not included, suggesting it does not meet the CNS activity threshold for CNS activity and may be less involved in nervous system-specific mechanisms.

Additionally, the top ten highest degree genes in this filtered network differed from those in the initial HD network, indicating a reorganization of key hubs when the network is refined to include only CNS-relevant proteins. The top ten highest connected nodes are: ACTB, TP53, CASP3, ALB, JUN, STAT3, EGFR, CTNNB1, MAPK3, and HIF1A. The nodes ACTB, TP53, and CASP3 are also present within the top ten highest connected genes in the initial HD network. This shift highlights the impact of tissue-specific filtering on the network's structural composition and the identification of potentially important nodes within the CNS context.

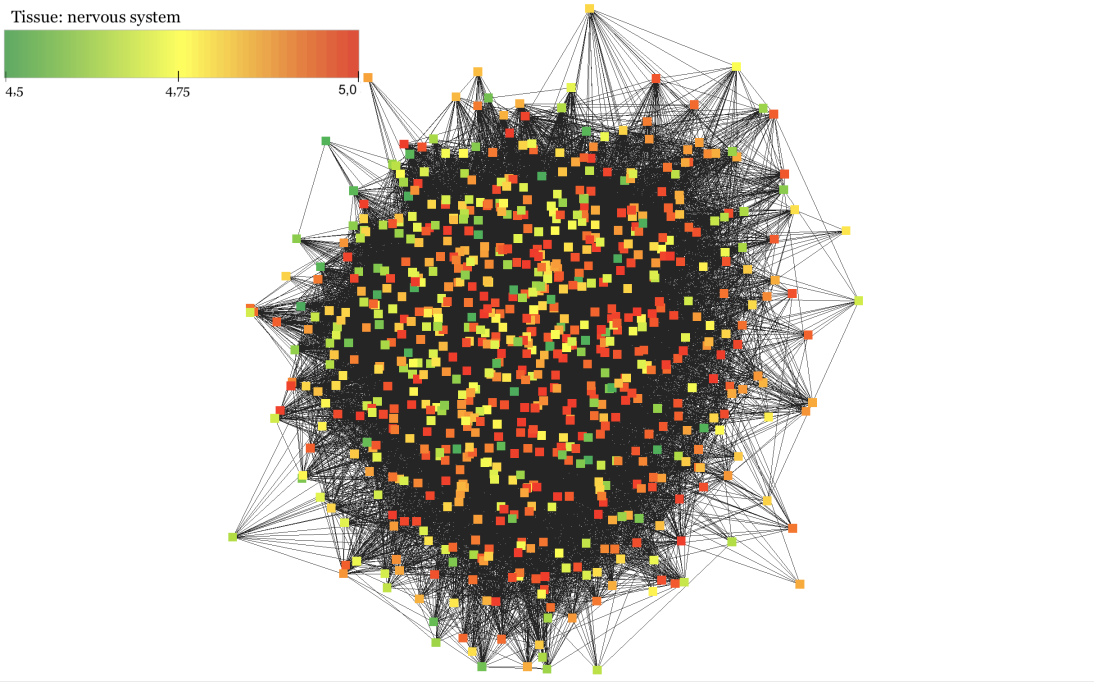

Figure 8: HD PPIN filtered for tissue nervous system cutoff 4.5, with coloring based on tissue nervous system score (STRING)

4.3.2 Subnetwork filtered nervous system 4.9

The resulting subnetwork contained 308 nodes and 7,541 edges, as shown in Figure 9. HTT was still included in the network, though it was not among the top ten highest-degree nodes. In this more stringently filtered network, neither MSH3 nor GFRA1 was present, indicating that both fell below the higher CNS relevance threshold applied here.
Furthermore, the top ten most highly connected genes in this subnetwork differed from those identified in the initial HD network and the 4.5-threshold filtered network. The top ten most highly connected nodes in this subnetwork are: GAPDH, ACTB, TP53, ALB, CTNNB1, HIF1A, HSP90AA1, HSP90AB1, FOS, and MTOR.
When compared with the initial HD network, ACTB, TP53, and GAPDH are shared between both networks, indicating their central roles. In comparison with the subnetwork filtered for CNS relevance at the 4.5 threshold, ACTB and TP53 remain common to both networks, highlighting their continued importance in the CNS context despite the more restrictive filtering.

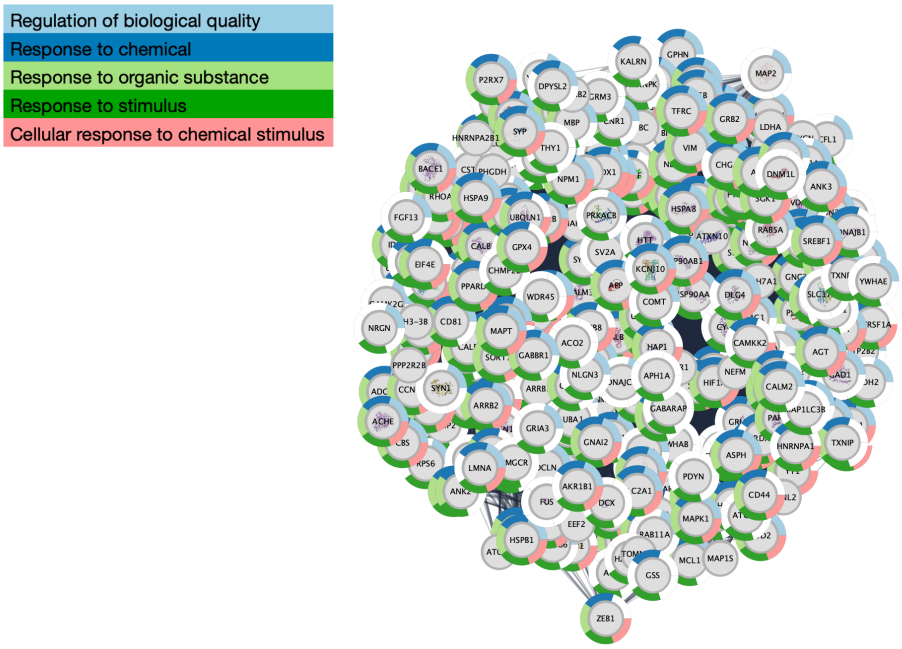

Figure 9: HD PPIN filtered for tissue nervous system cutoff 4.9, with STRING enrichment analysis coloring (STRING)

A STRING enrichment analysis for Gene Ontology (GO) biological processes was performed on the subnetwork filtered for nervous system relevance at the 4.9 threshold. The analysis revealed several significantly enriched GO biological process terms:

- GO:0065008 – Regulation of biological quality
- GO:0042221 – Response to chemical
- GO:0010033 – Response to organic substance
- GO:0050896 – Response to stimulus
- GO:0070887 – Cellular response to chemical stimulus

The ancestor charts from QuickGO can be found in Figure 12 of the Appendix. Among the identified GO terms, cellular response to chemical stimulus (GO:0070887) emerged as the most significant. This term was associated with nine of the top ten hubs from the initial HD network, with HSP90AA1 being the only exception. The next most significant terms were regulation of biological quality (GO:0065008) and response to chemical (GO:0042221). The term response to stimulus (GO:0050896) was the least significant among those considered. Notably, the term response to organic substance (GO:0010033) has since been marked obsolete and is no longer in active use in current GO annotations.
This result aligns with established characteristics of HD pathology. The enrichment of terms related to chemical and stimulus responses corresponds with prior findings that HD-affected neurons exhibit

altered responses to a wide range of chemical and environmental stressors. Notably, key pathological mechanisms in HD, such as oxidative stress, mitochondrial impairment, and abnormal glutamate signaling, are all associated with cellular responses to chemical stimuli[ZVC10]. Therefore, the identification of these GO terms among the hubs in this subnetwork network is an expected and biologically plausible finding. It supports the robustness of the network approach used in this study and reinforces the relevance of stress response pathways in HD progression.

4.4 Reactome

According to Reactome pathway data, MECP2 is reported to bind directly to HTT, indicating a potential functional relevance to HD. MECP2 was identified within the initial HD network, where it had a degree of 132, placing it outside the top ten most highly connected nodes. Furthermore, MECP2 was also found in both the first and second neighbor subnetworks of HTT, further supporting its possible involvement in HD-related protein interactions. However, MECP2 was not present in the first neighbor subnetworks of MSH3 and GFRA1, nor in the combined subnetwork connecting these two GWAS-linked genes from the GWAS dataset. Additionally, MECP2 was absent from the CNS-filtered subnetworks generated using threshold scores of 4.5 and 4.9. These observations indicate that while MECP2 may have a direct molecular interaction with HTT, its expression or functional involvement appears limited in CNS-specific contexts and does not feature prominently in subnetworks derived from GWAS-associated genes.

4.5 WikiPathways

Among the WikiPathways results, two pathways were directly associated with HD: "Effect of omega-3 PUFA on Huntington's disease pathways" and "ERK pathway in Huntington's disease."

4.5.1 ERK pathway in Huntington's disease

In the WikiPathways entry for the "ERK pathway in Huntington's disease," the protein MAPK1 is listed. This protein is also present in the initial HD network. Interestingly, although MAPK1 is part of the full HD network, it is neither a first- nor second-degree neighbor of HTT. This makes MAPK1 the only protein from the initial HD network that is not represented in the HTT second neighbor subnetwork, accounting for the difference in node count between the two (1,200 vs. 1,199 nodes). Furthermore, MAPK1 was identified in both CNS-relevant subnetworks created using the 4.5 and 4.9 tissue score thresholds, indicating a likely relevance in the nervous system context. However, it was not present in any of the subnetworks constructed from GWAS Catalog-derived proteins (MSH3 and GFRA1). Suggesting that MAPK1 may play a role in HD pathology through pathways specific to CNS signaling rather than through genetic associations captured in the GWAS studies.

In addition to MAPK1, the following proteins from "the ERK pathway in Huntington's disease" (WikiPathways) were also identified in the initial HD network: RAF1, MAP3K1, MAP2K1, CASP7, CASP3, GRM1, EGFR, NTRK2, BDNF, and EGF. Among these, CASP3 stands out with a degree of 625, ranking 8th among the top ten highest degree nodes in the initial HD network (Table 1). Table 3 presents a comparison of these proteins across the initial HD network and its subnetworks. The subnetwork based on MSH3, GFRA1, and their connecting nodes is not included in the table, as proteins not present in the individual subnetworks of MSH3 and GFRA1 cannot appear in their

combined subnetwork. Proteins present in both separate subnetworks will automatically also be present in the combined subnetwork of MSH3 and GFRA1. For example, NTRK2 is found in both individual subnetworks and is therefore expected to be included in their combined subnetwork as well. Of the proteins listed in Table 3, only NTRK2 appears in both individual subnetworks of MSH3 and GFRA1, and is therefore also present in their combined subnetwork.
Additionally, all listed proteins except for MAPK1, which was previously noted as excluded—are expected to be included in the HTT second neighbor subnetwork. Since the second neighbor network encompasses all of HTT's first-degree neighbors and their immediate interactions, any protein found in HTT's first neighbor subnetwork will, by definition, also appear in the second-degree subnetwork..

Table 3: Presence of ERK Pathway Proteins in the HD Network and Subnetworks. Proteins annotated in the "ERK pathway in Huntington's disease" (WikiPathways) are listed with their display names. The table compares their inclusion in the initial HD network and relevant derived subnetworks. The "Degree" column reflects each protein's degree in the initial HD network. HTT1 refers to the HTT first-neighbor subnetwork; HTT refers to the HTT second-neighbor subnetwork. An "X" indicates presence, and a "–" indicates absence of the protein in the respective subnetwork.

Name	Degree	HTT1	HTT2	MSH3	GFRA1	CNS4.5	CNS4.9
RAF1	157	-	X	-	-	X	-
MAPK3	540	X	X	-	-	X	-
MAP2K1	219	-	X	-	-	-	-
CASP7	113	-	X	-	-	-	-
CASP3	625	X	X	-	X	X	-
GRM1	96	X	X	-	-	-	-
EGFR	574	X	X	-	-	X	-
NTRK2	219	X	X	X	X	X	X
BDNF	415	X	X	-	X	X	-
EGF	447	X	X	-	-	-	-

4.5.2 Effect of omega-3 PUFA on Huntington's disease pathways

The "Effect of omega-3 PUFA on Huntington's disease pathways" offers a more detailed and comprehensive map than the "ERK pathway in Huntington's disease". It includes a wider range of proteins and interactions relevant to HD, with the ERK pathway fully encompassed within it. One key interaction shown in the pathway is from mHTT to BCL2. BCL2 is included in the initial HD network, where it has a degree of 637, making it the 6th most connected node (Table 1). It is also found in the first neighbor subnetwork of HTT, and consequently in the second neighbor subnetwork of HTT. However, BCL2 is not present in the subnetworks based on MSH3 and GFRA1, nor in their combined subnetwork. Furthermore, BCL2 is absent from both CNS-focused subnetworks, indicating it may have limited relevance in CNS-specific or GWAS-derived HD contexts. As BCL2 is not included in the MSH3 and GFRA1 subnetworks, it is not among the top ten highest ranked nodes based on degree within these subnetwork (Table 2). This absence reflects the changing network structure when focusing on GWAS data relevance. Another component of this pathway

to which mHTT is linked is the phosphorylation of BAD. BAD is also present in the initial HD network, where it has a degree of 49—relatively low compared to other proteins discussed previously. It was identified only in the second neighbor subnetwork of HTT, which aligns with expectations given the near-complete overlap in node count between the initial HD network and the HTT second neighbor subnetwork (1,200 vs. 1,199 nodes).

5 Conclusions

This thesis examined the protein-protein interaction landscape of Huntington's disease by analyzing networks centered on the huntingtin protein and other HD-relevant genes. The initial HD network, constructed using STRING data, was large and densely interconnected, comprising 1,200 nodes and over 83,000 edges. This high level of connectivity underscores the value of network biology in studying complex diseases such as HD. A tissue-specific subnetwork focused on the central nervous system revealed distinct structural properties compared to the full HD network, underlining the value of context-specific filtering in the study of neurological diseases. Subnetwork analysis using genes from a GWAS-based study showed limited overlap with the initial HD network; only two proteins, MSH3 and GFRA1, were found in both the GWAS dataset and the PPIN. This suggests that while genetic studies can identify potential risk loci, these genes may not correspond to highly connected proteins within disease-relevant interaction networks.
Overall, this thesis addressed the research question: *Can a consensus protein-protein interaction network in Huntington's disease provide new insights into the disease mechanism through integration of genome-wide association study data?* By combining STRING-derived interaction data with CNS-specific filtering and GWAS-based genetic findings, the study demonstrated that integrating protein interaction networks with genetic and tissue-specific information can yield complementary insights not evident from individual data sources alone. These findings emphasize the importance of integrating diverse data sources to gain a more comprehensive understanding of the molecular mechanisms underlying HD.

6 Further Research

Genes identified in this study as potentially involved in HD may also contribute to other neurodegenerative disorders, and conversely, genes associated with those diseases could offer valuable insights into HD mechanisms. Given that more prevalent neurodegenerative conditions have been studied in greater molecular detail, comparative analysis of pathways across these diseases may reveal shared mechanisms relevant to HD. For instance, the pathways returned from the WikiPathways search were generally broad and not specific to HD, but they may highlight common molecular processes active in the CNS. Future research could focus on exploring these shared pathways in greater detail, as well as validating candidate genes through experimental approaches to better clarify their roles in HD pathology. Additionally, subsequent studies could examine how the HD network evolves over time or varies across different brain cell types, such as neurons and glial cells. Another promising direction would be to assess whether any existing drugs target key proteins identified in this study, potentially enabling drug repurposing for HD treatment.

References

[Aea23] Ayushi Agrawal and Balcı et al. Wikipathways 2024: next generation pathway database. *Nucleic Acids Research*, 52 (D1):D679–D689, 11 2023.

[AM00] Blake JA et al. Ashburner M, Ball CA. Gene ontology: tool for the unification of biology. *Nat Genet.*, 25(1):25–29, 2000.

[Bar11] Gulbahce N. Loscalzo Barabási, AL. Network medicine: a network-based approach to human disease. *Nature Reviews Genetics*, 12:56–68, 2011.

[BDG+15] Gillian P Bates, Ray Dorsey, James F Gusella, Michael R Hayden, Chris Kay, Blair R Leavitt, Martha Nance, Christopher A Ross, Rachael I Scahill, Ronald Wetzel, et al. Huntington disease. *Nature reviews Disease primers*, 1(1):1–21, 2015.

[BOW+08] L. J. Beglinger, J. J. O'Rourke, C. Wang, D. R. Langbehn, K. Duff, J. C. Stout, and J. S. Paulsen. Depression in huntington's disease: prevalence and clinical correlates. *Journal of Neuropsychiatry and Clinical Neurosciences*, 20(4):441–446, 2008.

[Bun24] A. et al. Buniello. Gwas catalog, 2024. Accessed March 2025.

[CAea23] The Gene Ontology Consortium, Suzi A Aleksander, and Balhoff et al. The gene ontology knowledgebase in 2023. *Genetics*, 224(1):iyad031, 03 2023.

[Cli17] Mayo Clinic. Huntington's disease – symptoms and causes, n.d. [Accessed 2025 May 17].

[CN11] Wright GEB Caron NS. Huntington disease., 1998 Oct 23 [Updated 2020 Jun 11].

[Cyt23] Cytoscape Consortium. Cytoscape: An open source platform for complex network analysis and visualization, 2023.

[Gil24] M. et al. Gillespie. Reactome pathway database, 2024. Accessed March 2025.

[GT18] Rhia Ghosh and Sarah J. Tabrizi. Chapter 17 - huntington disease. In Daniel H. Geschwind, Henry L. Paulson, and Christine Klein, editors, *Neurogenetics, Part I*, volume 147 of *Handbook of Clinical Neurology*, pages 255–278. Elsevier, 2018.

[HZ06] Xionglei He and Jianzhi Zhang. Why do hubs tend to be essential in protein networks? *PLOS Genetics*, 2(6):1–9, 06 2006.

[Jia22] Chen Jiang. Constructing a protein-protein interaction network for huntington's disease and integration of gene expression data, 2022.

[KF16] C. Kenney and H.H. Fernández. Protein interactions of mutant huntingtin in huntington's disease: mechanisms and therapeutic opportunities. *Journal of Cellular and Molecular Medicine*, 20(1):3–19, 2016.

[Mar24] M. et al. Martens. Wikipathways: Pathways for the people, 2024. Accessed March 2025.

[MD17] Langbehn D et al. Moss DJH, Pardiñas AF. Identification of genetic variants associated with huntington's disease progression: a genome-wide association study. *The Lancet. Neurology*, 2017.

[Mea23] Marija Milacic and Beavers et al. The reactome pathway knowledgebase 2024. *Nucleic Acids Research*, 52 (D1):D672–D678, 11 2023.

[MT18] Patrick McColgan and Sarah J. Tabrizi. Huntington's disease: A clinical review. *European Journal of Neurology*, 25(1):24–34, 2018.

[oNDN21] National Institute of Neurological Disorders and Stroke (NINDS). Huntington's disease information page, 2021.

[PLS+08] J.S. Paulsen, D.R. Langbehn, J.C. Stout, E. Aylward, C.A. Ross, M. Nance, M. Guttman, S. Johnson, M. MacDonald, L.J. Beglinger, K. Duff, E. Kayson, K. Biglan, I. Shoulson, D. Oakes, and M. Hayden. Detection of huntington's disease decades before diagnosis: The predict-hd study. *Journal of Neurology, Neurosurgery Psychiatry*, 79(8):874–880, 2008.

[Roo10] Raymund A.C. Roos. Huntington's disease: A clinical review. *Orphanet Journal of Rare Diseases*, 5(1):40, 2010.

[Sem06] Creighton S. Warby S. Hayden M. R. Semaka, A. Predictive testing for huntington disease: interpretation and significance of intermediate alleles. *Clinical genetics*, 70(4):283–294, 2006.

[SGL+19] D. Szklarczyk, A.L. Gable, D. Lyon, A. Junge, S. Wyder, J. Huerta-Cepas, M. Simonovic, N.T. Doncheva, J.H. Morris, P. Bork, L.J. Jensen, and C. von Mering. String v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Research*, 47(D1):D607–D613, 2019.

[Szk24] D. et al. Szklarczyk. String: Functional protein association networks, 2024. Accessed March 2025.

[ZVC10] Chiara Zuccato, Marta Valenza, and Elena Cattaneo. Molecular mechanisms and potential therapeutical targets in huntington's disease. *Physiological Reviews*, 90(3):905–981, 2010.

Appendix

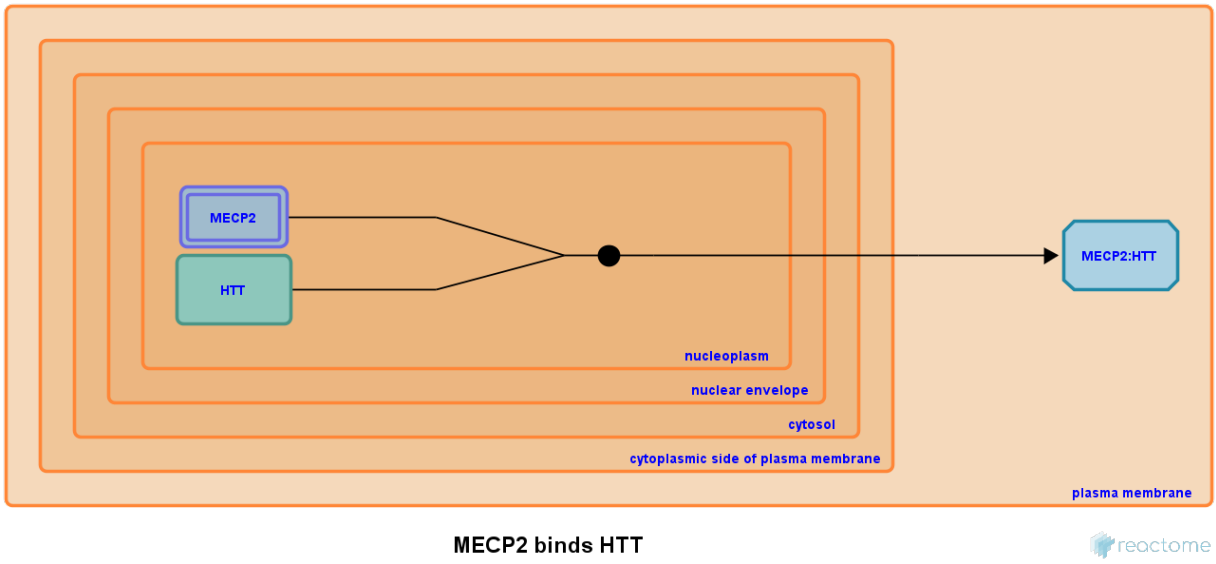

Figure 10: MECP2 binds HTT part of Regulation of MECP2 expression and activity (Homo sapiens) from Reactome

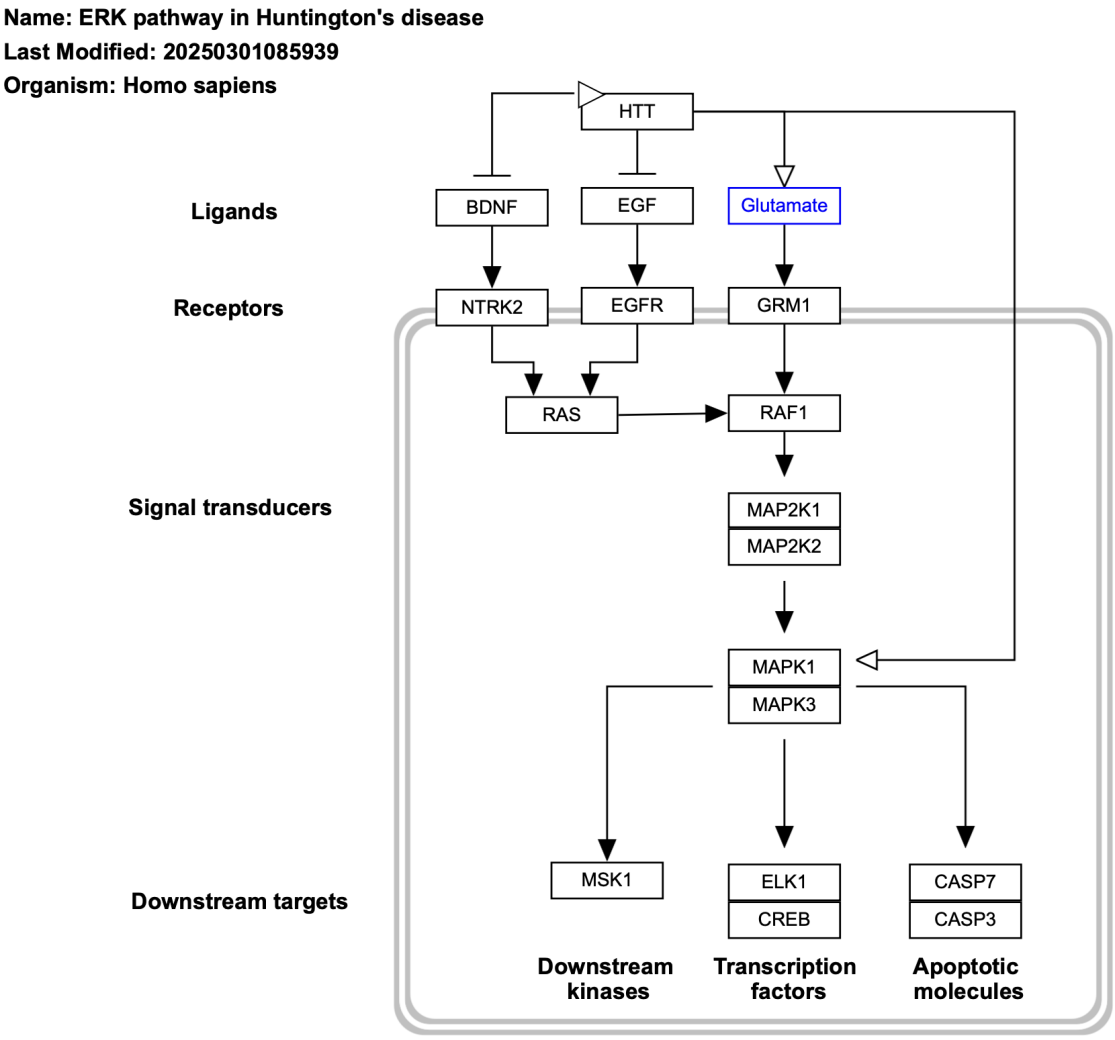

Figure 11: ERK pathway in Huntington's disease from Wikipathways

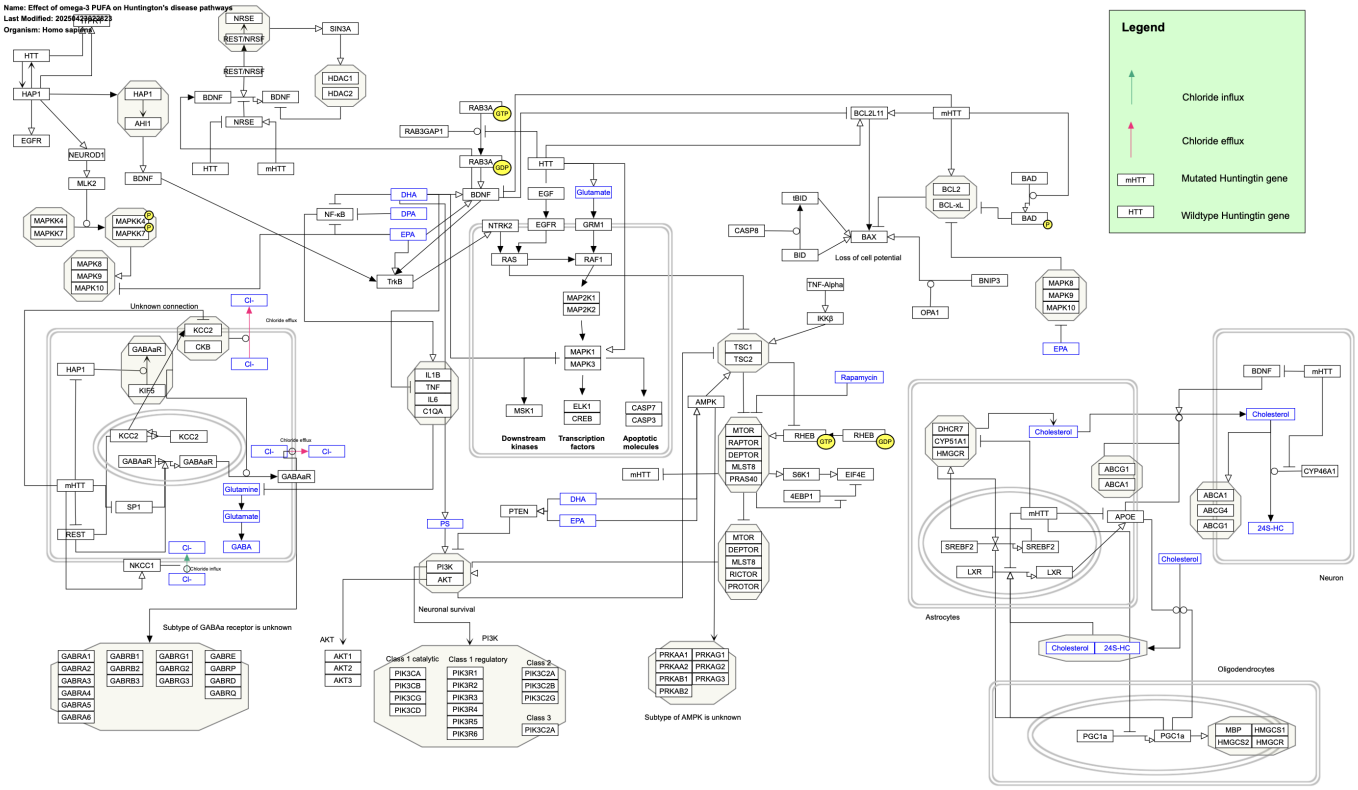

Figure 12: Effect of omega-3 PUFA on Huntington's disease pathways from Wikipathways

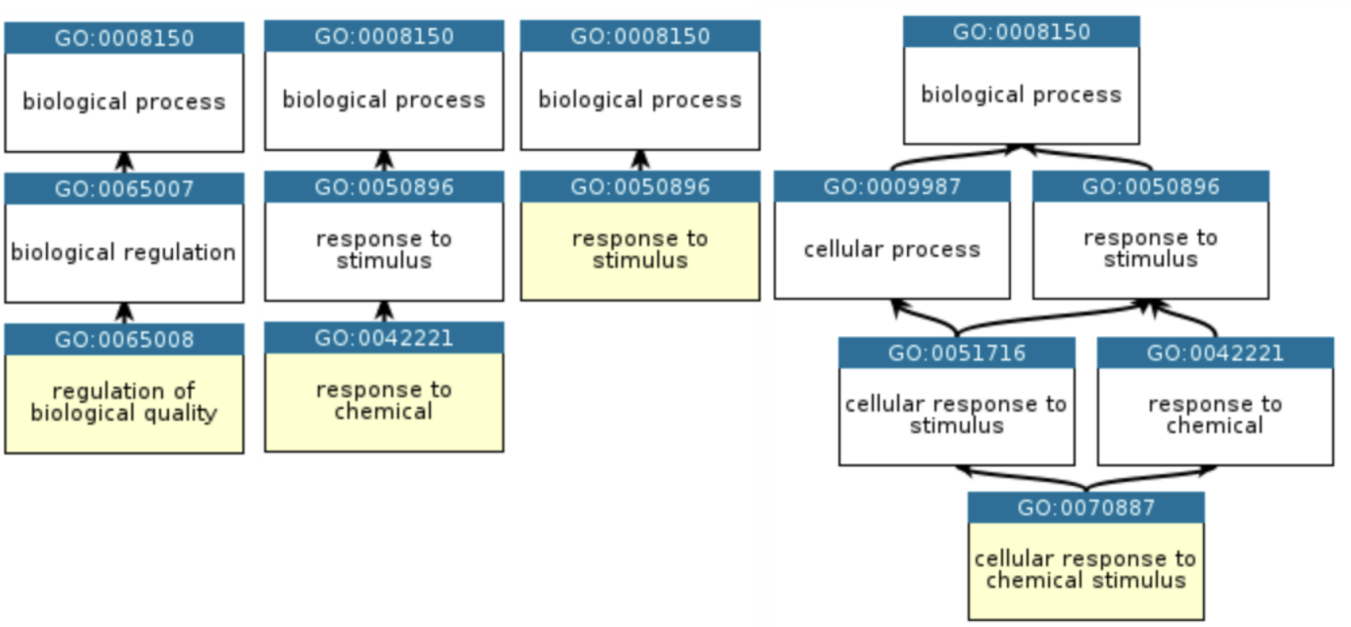

Figure 13: Anchestor charts for GO terms from Quick GO