Universiteit Leiden

Master Computer Science

Filling Gaps in Huntington's Disease Understanding with Alzheimer's Disease Knowledge Graphs and LLMs

Name:	Mireia Palou i Tort
Student ID:	s3636046
Date:	29/08/2024
Specialisation:	Bioinformatics
1st supervisor:	Dr. Eleni Mina
2nd supervisor:	Dr. Katy Wolstencroft
3rd supervisor:	Dr. Núria Queralt Rosinach

Master's Thesis in Computer Science

Leiden Institute of Advanced Computer Science (LIACS)
Leiden University
Niels Bohrweg 1
2333 CA Leiden
The Netherlands

Abstract

Research on rare diseases is often impaired by a lack of data, unlike common diseases such as Alzheimer's disease, which receive more funding and are more investigated. Huntington's disease, a neurodegenerative disorder, shares multiple similarities with Alzheimer's. Utilizing computational tools such as knowledge graphs and Large Language Models, we gained new insights into Huntington's disease. We present a framework, with validated results, that identifies areas of interest to improve the understanding of the disease. Our findings suggest a potential interaction between *TP53* and *FTL* that could reduce iron accumulation in the brain, and we propose that regulating *BMP6* may help mitigate memory impairment. These insights offer new directions for future research on Huntington's disease.

Acknowledgements

I would like to thank my supervisors Dr. Eleni Mina, Dr. Núria Queralt-Rosinach and Dr. Katy Wolstencroft for their guidance this year. This project could not have been possible without their valuable feedback and support. I also want to thank the rest of the BioSemantics group at the LUMC for their great advice and good times every Thursday.
Lastly, I also want to thank all my family and friends, from Leiden to Barcelona for their continuous encouragement.

Contents

Chapter 1

Introduction

1.1 From common diseases to rare diseases

Rare diseases are a group of 6,000 disorders characterized by their very low prevalence. It is estimated that 85% of these diseases affect less than one individual in a million. In Europe, a disease is considered rare when it affects 1 person in 2,000, with approximately 30 million people diagnosed with such conditions [1,2].

For most of these diseases, there is no existing cure. The low prevalence of rare diseases often results in limited public awareness, reduced funding, and consequently less research. This underfunding results in a severe lack of medical knowledge, making it difficult for healthcare providers to offer accurate diagnoses or effective treatments. On average, a rare disease diagnosis takes up to 5 years. Moreover, even after a diagnosis, information on available treatments, procedures, or management strategies is often sparse or non-existent. Patients may struggle to find reliable guidance or support, causing them stress and anxiety [1,2].

The challenge in a rare disease study becomes obtaining appropriate data to use. The small sample sizes complicate any investigation and limit the knowledge that can be acquired. Through the years, common diseases have benefited from more attention and accumulated more elaborated and extensive data [3]. This project aims to investigate how existing knowledge of a common disease can be leveraged to identify new information on a rare disease. We propose to build an innovative framework that exploits the similarities between a common and a rare disease to gain new insight into the latter. This approach seeks to fill the gaps in research by utilizing established data,

to improve the understanding of the disease to further aid in identifying therapeutic interventions.

1.2 Neurodegenerative diseases and Iron

1.2.1 Huntington's disease

Huntington's disease (HD) is a rare neurodegenerative disease. Neurodegenerative diseases are characterized by the progressive loss of certain neuron types [4]. Huntington's is an autosomal dominant and fully penetrant disease. In the Western world, 4 to 10 people out of 100,000 have HD. [5,6]

The most prominent symptoms of this disease include motor dysfunction (such as chorea) as well as behavioral and psychiatric disturbances, most likely due to neuronal dysfunction and cell death. Although there have been several studies on Huntington's, there is still no cure. Nonetheless, some drugs are used to manage the symptoms. For instance, *tetrabenazine* and *deuterabenazine* are used to treat chorea. [5,6]

HD is caused by a mutation in the huntingtin (HTT) gene. This mutation consists of a CAG triplet repeat expansion, which encodes an expanded polyglutamine stretch in the huntingtin protein. It has been studied that the length of the CAG repeat is closely correlated with the age of onset, the age at which the patient starts being symptomatic [5,6]. Nevertheless, there are some effects due to environmental factors, such as socioeconomic status, diet or age. [7]

In recent years, numerous studies ([8–10]) have pointed to the correlation between neuron loss and iron dysregulation in Huntington's disease. The researchers reported altered iron levels in post-mortem brain tissue, specifically in the basal ganglia, a region known for its sensitivity to iron level changes. Other areas of the brain have also been studied, and it is clear that iron homeostasis is affected. Nevertheless, the precise role of iron dysregulation in the progression of HD remains unclear, but it is one of the focal points of study in this disease. [8–10]

1.2.2 Alzheimer's disease

Alzheimer's disease (AD) is also a neurodegenerative disease and it is considered to be the most common cause of dementia [11,12]. Over 55 million people are estimated to have Alzheimer's and other types of dementia. [11,12]

The most important symptom is cognitive loss. In general, the behavioral and psychiatric symptoms somewhat resemble those of Huntington's disease. However, motor dysfunction is not associated with it [4]. As they are both neurodegenerative diseases, characterized by neuron death, another similarity that has been studied is the pathogenic pathways. There are some underlying cellular mechanisms in common: protein processing and accumulation (amyloid plaques in AD, Huntingtin protein aggregates in HD) or changes in brain signaling molecules. [4]

Alzheimer's disease (AD) can be categorized into different types, such as familial, sporadic, and late-onset, each with distinct characteristics and potential causes. In the majority of cases, such as late-onset Alzheimer's, the disease is mainly associated with risk factors, including age and family history. Certain gene mutations have also been observed to have an important influence. Some of these genes are APP, PSEN-1, PSEN-2 (all 3 implicated in Familial Alzheimer's) [4,13], CLU (clusterin) [14] and APOE (both involved in sporadic and late-onset Alzheimer's) [4,13]. Other relevant genes include MAPT (microtubule-associated protein tau gene) [15] and BACE1 [16], which are related to the accumulation of tau protein and amyloid beta (Aβ) plaques in the brain respectively.

The amyloid beta plaques have been proposed to be a key part of the pathogenesis of Alzheimer's disease [17]. They are linked to cell function disruption and recent studies ([18,19]) point to iron contributing to their formation. In fact, the elevated levels of iron and consequent ferroptosis is one of the most prominent hypothesis for neuronal loss in Alzheimer's disease. [18,19]

1.2.3 Iron dysregulation in the brain

Iron plays an important role in the brain by supporting multiple processes such as neuronal development, oxygen transportation and respiration. Although it is found throughout the brain, the highest levels are found in the basal ganglia. Iron binds to transferrin (protein) and enters the brain through the blood-brain barrier (BBB). This barrier is a regulator of iron transport from the blood to the brain. Another important protein is ferritin, a regulatory protein that stores one-third of the iron in the brain. [20-22]

Iron homeostasis is essential for all the associated processes to work correctly. When iron levels are affected, either due to a deficiency or an overload, it can cause significant neurological consequences. As briefly mentioned in the diseases introductions, a lot of

research points to the implication of iron accumulation in neurodegenerative diseases [8–10, 18, 19]. Many factors could induce changes in iron levels, like aging, however, the mechanisms causing it are still being studied. It is also not clear how to reverse these changes. While several therapeutic approaches have been aimed, present cures can only alleviate symptoms. [20, 21]

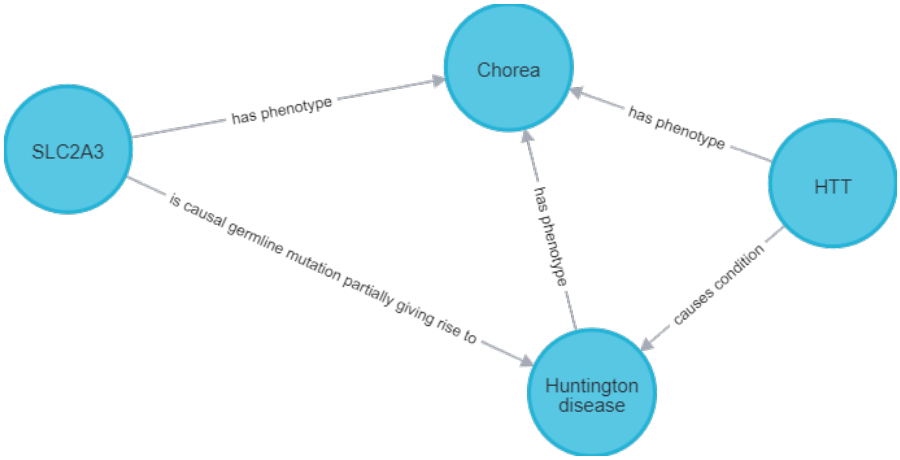
1.3 Knowledge graphs

As data generation accelerates across various domains, including biomedicine, effective tools for representing, managing and leveraging this data are essential [23]. Knowledge graphs have emerged as a state-of-the-art solution for representing complex information, offering advanced capabilities in data integration and exploitation. Unlike traditional storage methods, such as relational databases, knowledge graphs offer two main advantages: flexibility in the data format and smooth integration of data sources or new data. Knowledge graphs structure the data homogeneously leading to less ambiguity and effortless usage, as well as allowing more adaptability [24]. Knowledge graphs support complex querying and analysis. This makes them invaluable for research, enabling efficient retrieval of relevant data and facilitating the exploration of the relationships within the data. By leveraging the structured, semantic nature of knowledge graphs, researchers can gain deeper insights and drive advances in fields such as biomedicine. [23, 24]

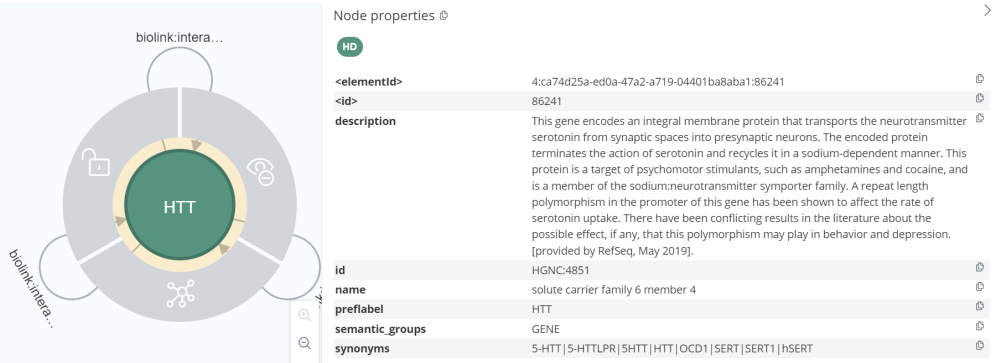
All types of graphs consist of two main elements: nodes, which represent the entities, and edges, which in knowledge graphs relate nodes through a semantic relationship. Each pair of nodes connected by an edge forms a triplet, expressed as *(subject, predicate, object)* [25]. In the biological domain, nodes represent concepts such as genes, diseases, body parts, phenotypes, etc. Edges capture the relationships through a textual label: *associated*, *interacts with*, *has phenotype*, etc. It is also possible to store additional information in a knowledge graph. The inherent flexibility allows each node and edge to possess a set of attributes of any category or data type, further enriching the informational context within the graph.

A simple example is depicted in figure 1.1a. It consists of a small subset of nodes related to Huntington's disease: 2 genes (*SLC2A3* and *HTT*), Huntington's disease and the phenotype *Chorea*. There are 3 different types of relationships: *has phenotype*, *causes condition* and *is casual germline mutation partially giving rise to*. There are 5 distinct triplets in the graph, for instance *(Huntington disease, has phenotype, Chorea)*. In

addition to the information shown, each node and edge contains additional information in the form of attributes, an example of node *HTT* is shown in figure 1.1b.



(a) Small knowledge graph. The 4 nodes (genes *SLC2A3* and *HTT*, Huntington's disease and phenotype *chorea*) are connected through edges with different relationship types. The direction of the arrow depicts the triplet structure (ex: HTT, causes condition, Huntington's disease).



(b) Properties of a node. Knowledge graphs allow the storage of attributes. In this example, there is the node for gene *HTT*, with all the information related to it.

Figure 1.1: Kowledge graph examples. (a) Knowledge graph visualization (b) Detailed properties of a node.

1.3.1 Constructing knowledge graphs: BioKnowledge Reviewer

The construction of Knowledge graphs requires advanced computational tools and methodologies. The BioKnowledge Reviewer, a library developed by Queralt et al. [26] was built to create a structured review in the form of knowledge graphs. These graphs organize and integrate information regarding a research question, so it can be used for computational analysis. The authors used as a proof-of-concept the NGLY1 Deficiency (DOID:0060728). Their method integrated all gathered information on this disease into a knowledge graph allowing easier access. The organized structure enables researchers to explore relationships between entities, identify patterns, and obtain new insights based on the information. The knowledge can be obtained by navigating the graph or posing specific queries. Notable is also the processing of the data. The capacity to merge multiple databases facilitates a more comprehensive analysis. [26]

The mechanism is shown in Figure 1.2. The BioKnowledge Reviewer gathers information from different databases, such as the monarch initiative [27] or transcriptomics data (gene expression data), and creates a graph for each. Once these graphs are constructed, they are merged and can be used for further analysis. [26]

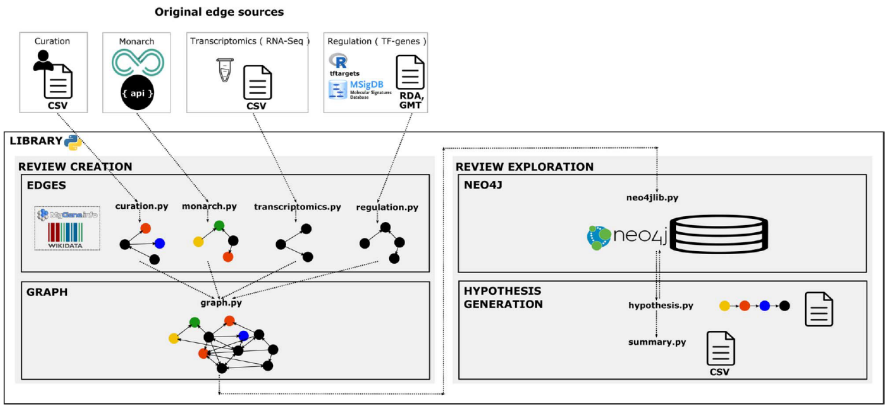


Figure 1.2: Architecture of the BioKnowledge Reviewer. At the top we see how each of the data sources is structured into subgraphs, which are then merged to create the knowledge graph. On the right we have that the graph is imported into neo4j [28] to perform hypothesis generation

1.3.2 Knowledge graph analysis through visualization

To effectively leverage knowledge graphs, visualization is an essential component that aids in understanding and interpreting the underlying data structures. Neo4j is a graph database management that enables visualization and analysis of the data through an integrated terminal [28]. It does so utilizing the *Cypher* query language, which is similar to SQL and is intuitive due to its use of ASCII-Art patterns. [29]

Neo4j can be used to retrieve interesting subgraphs and learn about the graph's topology. For instance, to retrieve the shortest path between two nodes, the minimal amount of nodes in between two specific nodes. A *shortest path query* example is shown in figure 1.3.

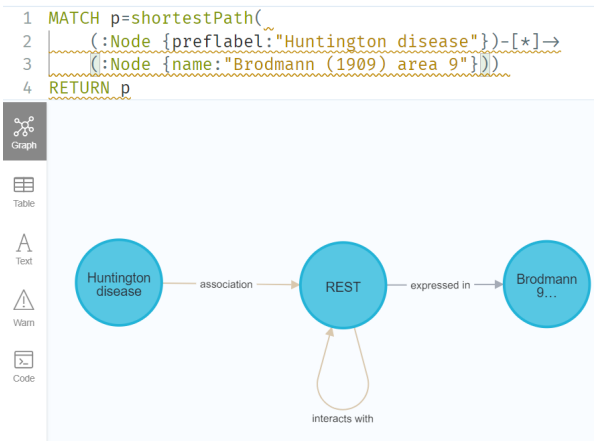


Figure 1.3: Shortest path: Example from *Huntington's disease* to *Broadmann (1909) area 9*. The shortest path is of length 2, with only 1 node in between (*REST*).

1.4 Large Language Models (LLMs)

Large Language Models (LLMs) are a class of artificial intelligence models trained on vast amounts of data. These models are popularly used to recognize, summarize, translate, predict or generate text, among other tasks. LLMs are based on transformers, a specific type of neural network based on self-attention mechanisms to learn complex patterns and relationships within the data. This allows LLMs to capture semantic relationships and understand better the context within the text. [30, 31]

The strengths of LLMs have made them particularly popular in recent years, leading

to the development of numerous models across a great variety of domains [31]. One emerging area of interest is knowledge discovery, where LLMs have shown considerable potential. Their advanced capabilities enable them to uncover new insights and generate novel content by analyzing and interpreting extensive datasets [32].

To tailor LLMs for specific tasks they are often fine-tuned. This critical step involves training the model with task-relevant data. Fine-tuning has been shown to enhance the model's performance by improving prediction accuracy and the reliability of generated content while mitigating issues such as hallucinations (false predictions) [33]. Their capability to be used in different fields shows a key advantage, their adaptability to different types of data. Although the majority of models are used to exploit traditional textual data, we also find new cases using structured data such as knowledge graphs. [34]

Despite their capabilities, LLMs can be limited by high computational costs, data biases, and ethical considerations such as the generation of misleading or harmful content. [30,31]

1.5 Related work

1.5.1 Computational tools for disease similarity identification

Identifying similarities between diseases can lead to repurposing treatments or a new insight into the pathology [35]. However, there is not a gold standard to declare two diseases are similar. It can be due to a similar molecular cause, treatment by similar drugs, similar biomarkers for diagnosis, or even similar phenotypes. [36].

There have been numerous attempts to quantify how similarity is measured. In 2019, Cheng et al. [36] performed an extensive review and classified all methods into 4 categories:

- *phenotype-based*: The methods utilize qualitative associations between phenotypes and diseases.
- *molecule-based*: These methods involve protein-protein interactions (PPIs) or co-expression data.
- *hierarchy-based*: Methods that use hierarchical structures of disease-related ontologies to calculate the similarity of terms based on distance.
- *hybrid methods*: Combinations of molecular and hierarchical methods.

The classical methods prove useful for common diseases but are clearly limited by the disease vocabulary and annotations, making them ineffective for rare or newly discovered diseases [37]. This limitation highlights the necessity for more advanced and comprehensive methods to measure disease similarity.

1.5.2 Transfer learning for knowledge graph completion

Transfer learning in biomedicine

Transfer learning is a Machine Learning technique that repurposes knowledge from one task to improve learning in another. It can be applied to any domain, typically where there is some similarity and a scarcity of data [38].

In biomedicine, we find significant findings in the field of computer vision. Pre-trained models are used to improve the analysis of imaging data, such as magnetic resonance imaging (MRI) scans [39]. Additionally, transfer learning has also been applied for drug repurposing. During the COVID-19 pandemic, this technique gained prominence as researchers utilized existing therapeutic data from related diseases to identify potential treatments for the novel virus [40]. Zhang et al. [40] performed this repurposing using knowledge graphs. Their approach consisted of building up the information found in their graphs, a process known as knowledge graph completion [41].

Knowledge graph completion with LLMs

Knowledge graph completion consists of enriching the graph by finding new relationships between existing nodes or predicting new nodes to connect to the graph [41]. There are several techniques to exploit a knowledge graph and gain new knowledge. A common approach consists of first embedding the graph (transforming it into a numerical vector), and then using it to perform various metrics or feed into an algorithm of choice [42]. There are two main embedding approaches: translational distance methods and semantic matching methods. Translational distance methods, such as TransE and TransH, use the distance to give a scoring. Semantic matching methods focus on semantic similarity, a well-known method is RESCAL. [42, 43]

Completing a knowledge graph is severely limited by computational power. When looking for new information the method needs to be efficient, as the search space can become significantly extensive. Another relevant limitation is the introduction of new entities or relationships. Newer methods aim to tackle these challenges by leveraging

computational techniques from Machine Learning and Artificial Intelligence, which have proven to be useful and computationally efficient in many tasks [44].

The most relevant example in literature of Knowledge graph completion with LLMs is KG-BERT [34]. It is based on the well-known LLM called BERT (Bidirectional Encoder Representations from Transformers). BERT's main advantage is its pre-training on representations from unlabeled text, using context from both left and right, making it bidirectional. It allows an easy adaptation of the pre-trained model and fine-tuning to obtain state-of-the-art solutions for specific problems, such as knowledge graph completion [45]. The architecture of a BERT-based model follows a systematic structure. The input consists of sentences preceded by a special empty token ([CLS]) that are tokenized to fit the model correctly. When a label is predicted, the initial token contains the prediction and the rest remains unchanged [45]. The authors of KG-BERT [34] propose a multi-label classification task. The model takes as input two entities and outputs a label with the type of relationship that should connect both entities. An illustration of the process is provided in Figure 1.4.

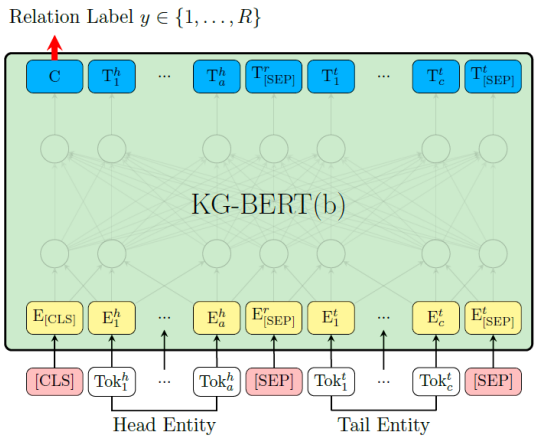


Figure 1.4: Workflow of KG-BERT for relation prediction. [34] At the bottom we have that the 2 entities, tokenized, are inputted into the model. The output given (at the top) is the relation label that relates both.

To the best of our knowledge, there is no case of using LLMs for knowledge graph completion in biomedicine. Nevertheless, there are several popular models such as BioBERT, fine-tuned with PubMed abstracts and PMC full-text articles, that can easily be repurposed. [46]

1.6 Research Questions and goals

The primary research question addressed in this project is: **Can we gain new knowledge on a rare disease by using information on a common disease stored in a knowledge graph, through the use of Large Language Models?**

The goal of this research was to exploit existing knowledge on a common disease, Alzheimer's, to gain valuable novel information on a rare disease, Huntington's. Following current research, we focused specifically on uncovering if it was possible to generate new insights regarding the role of iron in Huntington's disease and identifying potential therapeutic treatments.

In particular, we investigated whether we could harness the semantic structures inherent in a knowledge graph to facilitate the inference of new relationships within the graph. The goal was to construct a computation and analytical tool that would leverage the data on a knowledge graph to promote transfer learning. We aimed to extract the most relevant information of the common disease, for our use case, and gain new insight into the rare disease that could be added to complete the graph.

To achieve the knowledge graph completion, we employed BioBERT, a state-of-the-art Large Language Model. We hypothesized that fine-tuning BioBERT with data related to the common disease would improve the model's ability to generate accurate and relevant predictions. Furthermore, we proposed that the predictions would be supported by evidence in the existing literature, thereby ratifying the model's efficacy in discovering meaningful connections. To test this hypothesis, the results of the model's predictions were validated first by examining the disease knowledge graphs; then by manually reviewing the scientific literature.

Chapter 2

Methods

A simple workflow of the methodology can be seen in Figure 2.1. It is split into 5 main tasks that aim to achieve the objectives presented in the previous section:

1. Gathering data: Using the BioKnowledge Reviewer, gather all information on the 2 diseases.
2. Creation of the Knowledge Graphs: Use the gathered data and create the two graph visualizations for HD and AD.
3. Knowledge graph examination: Initial comparative analysis of the knowledge graphs. Finding areas of interest to complete.
4. Knowledge graph completion: Fine-tuning of the BioBERT model, with Alzheimer's data, so it performs relationship predictions to complete the Huntington's Knowledge graph.
5. Analysis of the predictions: Filter out the most significant results and investigate them in literature and within the context of the graph.

The code, data and results for this thesis can be found on GitHub[1].

[1] https://github.com/mirepalou/kg_adhd

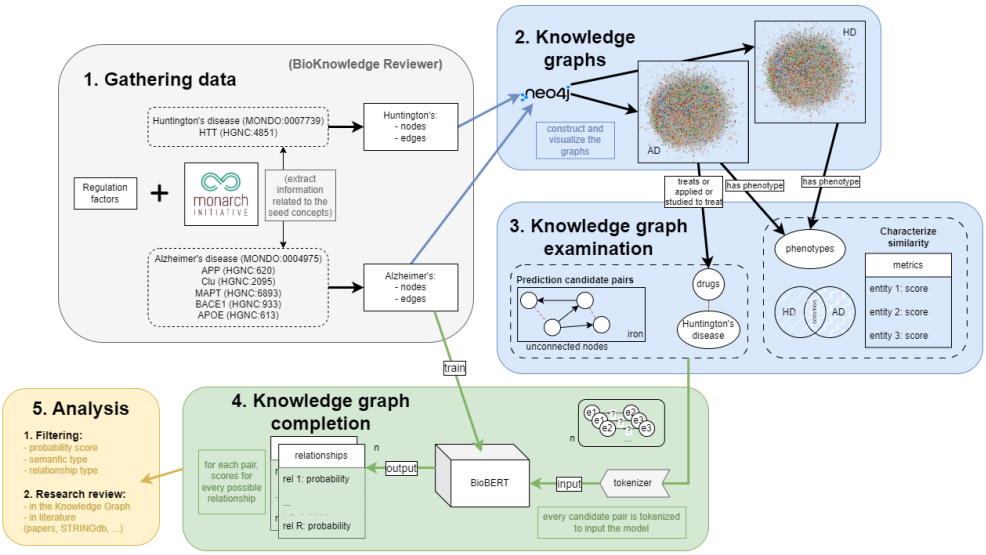


Figure 2.1: Workflow of this project. Each box represents one of the main tasks of this project. From gathering the data, to constructing the knowledge graphs, exploring them, and finally inferring new relationships and analyzing them.

2.1 Gathering data

The knowledge graphs are built following the BioKnowledge Reviewer [26] [2]. Recall that this library collects information from public data sources about genes, phenotypes, diseases, etc. The collected information is stored as nodes and edges, with their corresponding attributes, to construct the graph. In this project, the BioKnowledge Reviewer was used to retrieve information from two data sources: Monarch [27] (from the API version 3[3]), to gather data on the disease biology, and Molecular Signature database (MSigDB) [47] (the C3:TFT sub-collection v6.1), to collect data on transcription factors regulation

Monarch is a platform that integrates a large volume of cross-species biological data from several data sources. To this date, it stores over 10 million associations, that is relations between 2 biological entities [27]. The relationships are characterized by the *biolink Model* (v4.2.2) [48]. Each graph was created based on a list of key concepts or seeds. For Huntington's disease, the seeds were the disease identifier (MONDO:0007739) and the *HTT* gene (HGNC:182293). For Alzheimer's disease,

[2] hhttps://github.com/NuriaQueralt/bioknowledge-reviewer
[3] https://api.monarchinitiative.org/v3/docs

the seeds used were the disease (MONDO:0004975), and the following genes: *APP* (HGNC:620), *CLU* (HGNC:2095), *MAPT* (HGNC:6893), *BACE1* (HGNC:933) and *APOE* (HGNC:613). As discussed, Alzheimer's disease does not have a unique cause or a distinct set of genetic causes. Therefore, the set of genes was chosen based on the expertise of Dr. Elena Daoutsali.

Additionally, information on transcription factors (TFs) was added. These are proteins in charge of converting, or transcribing, DNA into RNA [49]. They regulate the expression of genes and are involved in multiple disease processes. Analyzing their activity can provide significant insight into our research [50]. Thus, it is very relevant for a study relating diseases with genetic causes to include transcription factors, as well as their relation with other data in the project. To include the transcription factors nothing was altered from the original BioKnowledge Reviewer as it already included regulatory relationships from MSigDB (a collection of annotated gene sets) and curated gene regulatory information [26, 47].

2.2 Building the knowledge graphs

After gathering all the data, it was formatted and loaded into a graph management system, Neo4j (v5.16) [28], which was selected due to its intuitive use. Initially, the BioKnowledge Reviewer library [26] was used to automatically format and load data for a single graph. However, some modifications were made to integrate multiple knowledge graphs within the same Neo4j instance. This involved custom formatting and data integration steps to maintain the integrity of each graph while allowing for simultaneous analysis.

All the instructions to install the correct version of Neo4j, other requirements, the data, and the commands to load the data can be found in the GitHub repository. [4]

2.3 Knowledge graph examination

Knowledge graph examination was performed using Python (through jupyter notebook [51]) and Neo4j [28]. The code is accessible in the GitHub repository of the thesis [4]. This step was done to characterize the data, get a deeper understanding of the information contained, and decide what is most interesting and should be explored.

[4] https://github.com/mirepalou/kg_adhd

We started with a general comparative analysis of the nodes and edges, followed by a topological analysis of the knowledge graphs. The results of this analysis are shown in the next chapter.

Given that the graphs were quite large, it would be computationally inefficient to perform complete knowledge graph completion, of all kinds and through the entire graph. This project focused on relationship prediction, finding a new connection on a previously unconnected pair. The search area became exponentially large as the graphs increased (note that $u = n * (n-1) - c$, where u are the number of unconnected pairs, n the number of nodes, and c the already established connections). Thus, it was necessary to target specific areas that would be interesting to enrich. The two areas chosen were the relationship of Huntington's with iron and the possible repurposing of therapeutic drugs present in the Alzheimer's graph.

The candidate drugs for repurposing to Huntington's disease were identified through queries in Neo4j. This process involved locating drugs within the graph that are used for treating Alzheimer's disease and evaluating their relevance for Huntington's disease according to the model. The exact query and the results (the drugs selected and their predicted significance) are shown in Chapter 3.

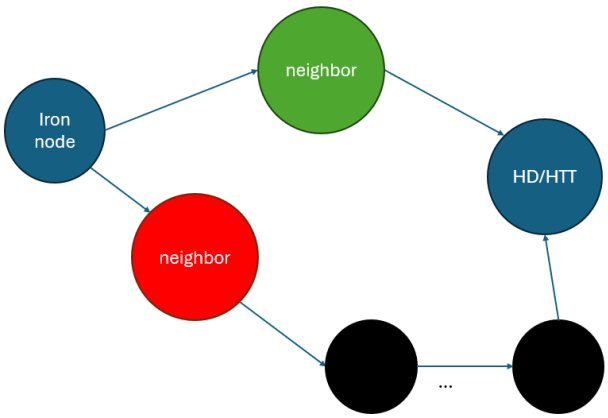


Figure 2.2: Candidate node selection for the iron subgraph. The subgraph of nodes is chosen from the *iron* nodes and their neighbors. Only the neighbors that are connected to the *Huntington's disease* or *HTT* node are kept (in green). The nodes that are not direct neighbors (in red) are filtered out.

As stated, for computational efficiency we only performed knowledge graph comple-

tion in the subgraph related to our initial hypothesis, the role of iron in HD. The procedure to obtain the iron subgraph of the HD graph is illustrated in figure 2.2. First, we selected all nodes that had the string `iron` in their name or description, as well as their first neighbors (nodes directly attached to them). In Neo4j we added a boolean attribute named *iron* to differentiate the nodes. To further restrict into a pre-established interesting area, relevant to our research, we constrained by distance to Huntington's disease. Aside from the *iron* nodes, only the neighbors that were also neighbors from the *Huntington's disease* node or the *HTT* gene node were considered. This consisted of a subgraph of 484 candidate nodes. Then a trivial recombination was done to create all possible pairs. Note that previously connected pairs as well as pairs where none of the nodes were related to iron were deleted for clarity.

2.4 Knowledge graph completion

2.4.1 Proposed model

Given the data's biomedical nature, KG-BERT, the BERT model for knowledge graphs completion [34], was adapted to suit our use case better. Instead of starting with the pre-trained standard BERT, BioBERT, a pre-trained BERT version that was fine-tuned using biomedical text (PubMed abstracts and PMC full-text articles) was selected [46]. A workflow of the procedure is provided in figure 2.3.

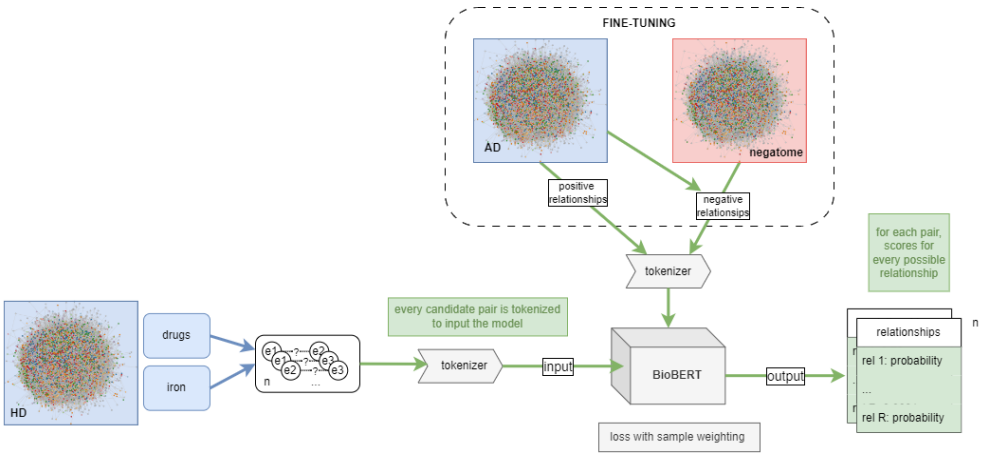


Figure 2.3: Workflow of the proposed model On the top, the Alzheimer's graph is used to fine-tune the model. On the left, relevant areas of the HD graph are selected to be enriched. They are sent to the model that outputs a set of probabilities for each connection.

Fine-tuning

Before the model could be used for predictions, it needed to be fine-tuned (indicated by the dashed box in Figure 2.3). This process involved training the model further to improve its performance on the specific task of interest. In this case, the BioBERT model was fine-tuned with the AD graph data, to ensure the transfer learning from the common to the rare disease. The goal was to enhance the LLM's performance by adding data on AD. The model was forced to learn the patterns in the AD graph that might be relevant to Huntington's disease. This pseudo-bias process, this adjustment of the learning process, is meant to guide the model to make predictions on the rare disease that reflect similar biological processes or pathways observed in the common disease.

All the triplets on the Alzheimer's graph were taken as positive samples. To ensure the model could discriminate correctly, a negative sample space was created. We implemented a new relationship category (*no interaction*) that related entities that are known to not be connected. To avoid false negatives the *Negatome* [52], a curated collection of non-interacting protein pairs was used. Moreover, to maintain balance and diversity in the training data, random unconnected nodes from the Alzheimer's graph were sampled and paired accordingly [53]. For every triplet, the two entities' names (subject and object) were concatenated and tokenized using the pre-trained BioBERT tokenizer. The maximum sequence length was set to 128 tokens. This is an indispensable process for any LLM project as it embeds the training data so it can be passed on to the model. The relationships were mapped into numerical labels ($\in \{1, ..., R\}$, where R is the total number of relationships).

To prevent overfitting the data was separated into two datasets: train and test. As customary the test data was used to check if the model performed similarly with new unseen data. The split was done using the python library *sklearn*' method *train_test_split* [54]. The data was separated randomly, 80% of it became the training data (2,221,722 samples) and 20% the test (555,431 samples).

The model to fine-tune was the pre-trained BioBERT model for sequence classification. We utilized the *BioBERT v1.1.*, which is available in *Hugging Face* (a platform with machine learning tools) [55]. *AdamW* [56], from the Python library *PyTorch*, was set as the optimizer (it fine-tunes parameters during training). Its adaptative weight decay (set to default 0.01) allows for better regularization of the loss. We considered it to be well-suited for a problem large in terms of data as it is computationally efficient. The

learning rate was set to *1e-5*. The model was trained in batches of size 32, therefore in each iteration 32 samples were selected. The batches were randomly created to break any unwanted dependencies in the data. As this is a multi-classification task, to compute the loss we used the *Cross entropy* (shown in equation 2.1). To avoid any problems with imbalanced data, the loss was set up so it took into account the weight of each class [57]. The training was done for 3 epochs: the model is trained on the full train dataset 3 times, each time adjusting the weights.

$$-\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{R}(y_{i,j}\cdot log(p_{i,j})) \tag{2.1}$$

(where N is the number of samples, R the number of classes/relationship types, $y_{i,j}$ the true label and $p_{i,j}$ the probability) [58]

After training, the test set was used to assess the quality of the model.

Hyperparameter tuning

The grid search approach was used to optimize the hyperparameters of the model. It consists of systematically testing a range of values for the different hyperparameters. This approach allowed us to have more control over the values tested as well as perform tests in a parallelized way [59]. In table 2.1 we have an overview of the tested parameters. In the case of the *batch size* and the *maximum sequence length*, the values were chosen based on the best-reported values of the BioBERT model [46]. Moreover, for the batch size, one has to be attentive to the computational limitations, and as our model is quite intensive we selected sizes from the lower range (1-32). For the *learning rate* and the *weight decay*, we based it on the common default parameters of the selected optimizer *AdamW* [56].

Hyperparameter	Description	Search space
learning rate	speed at which the search space is explored	5e-5, 3e-5, **1e-5**
batch size	amount of samples per iteration	8,16,**32**
max sequence length	length input sequence	64, **128**, 256
weight decay	shrinking weights in optimization to prevent overfitting	**0.01**, 0.001

Table 2.1: Hyperparameter tuning. Each row contains a parameter that is tuned and its description [45, 56]. The search space indicates the range of parameters evaluated to optimize model performance, in bold the values representing the initial configurations.

Evaluation

The model was evaluated using several metrics. This was done to avoid any bias when interpreting the results. By analyzing the predictions from different points of view we got a better picture of how well it performed. The metrics used and their mathematical expressions were the following:

- precision $= \frac{\text{true positives}}{(\text{true positives}+\text{false positives})}$
- recall $= \frac{\text{true positives}}{(\text{true positives}+\text{false negatives})}$
- f1 score $= \frac{2 \cdot \text{precision} \cdot \text{recall}}{(\text{precision}+\text{recall})}$
- accuracy $= \frac{\text{true positives}+\text{true negatives}}{(\text{true positives}+\text{false positives}+\text{true negatives}+\text{false negatives})}$

These metrics were calculated for each class/relationship type. Then, to adapt to the multi-class case, they were averaged in 2 ways: using the general arithmetical mean and using the class weights. The results are shown in the following chapter.

The best combination was **learning rate 1e-5, batch size 8, maximum sequence length 128 and weight decay 0.01**. This combination was used to construct the final model for prediction.

Predictions

When the model was fully trained it was used to predict on unconnected pairs. That is to predict which relationship would these 2 entities have, or if it is likely that these 2 entities should be connected. The predictions were made on the two areas of interest established during the graph exploration.

The results are taken from the logits outputted by the models. These are series of vectors that can be interpreted as the probability for each sample to be labeled as each relationship type. The logits are transformed using the *softmax* function (shown in equation 2.2). This ensures that the probabilities are bounded ($\in [0, 1]$) and normalized. The higher the probability, the more confidence in the specific class.

$$s(x_i) = \frac{e^{x_i}}{\sum_j e^{x_j}} \tag{2.2}$$

where x_i is a vector of logits [60]

GPU

An LLM model of this scale needed to be executed in a machine with considerable computational power. In these cases, the standard procedure consists of running the code in a GPU (Graphical Processing Unit). For this project, we used the GPUs on the Snellius national supercomputer. They consist of NVIDIA A100 that works in conjunction with Intel Xeon CPUs. Every node has 72 cores and 480 GiB of memory. [61]

2.5 Analysis of the predictions

To analyze the BioBERT model's predictions and enhance our understanding of Huntington's disease, we designed a systematic approach:

1. **Filtering of the predictions:**
    - Positive relationships: The pairs predicted to have no interaction were eliminated as we wanted novel information, candidate suggestions that should be further explored.
    - Quality control: minimum probability score threshold. This was done to ensure that the relationships were predicted with confidence and distinctively from other relationships.
    - Entity and relationship types: we gave more relevancy to predictions relating genes or phenotypes.
2. **Contextual analysis:**
    - Initial literature examination: only cases pertinent to the disease and its underlying mechanisms were chosen. The aim was to find predictions that contributed to the iron hypothesis and could give new directions for future research related to the disease.
    - Investigation in Neo4j: examined the context of the nodes within the knowledge graphs using pre-built functions *shortestPath* and *AllShortestPaths* [28]. This was performed individually for each entity, with the two entities together, and for both Knowledge graphs. It provided preliminary insights into the entities, such as their common interactors, phenotypes, etc.

- External databases: we consulted the STRING database [62] (platform with protein-protein interactions of 59,309,604 proteins from 12,535 organisms). Using the batch method, we supervised gene interactions from different sources. The given relationships allowed us to have a better distinction of which genes were interesting to explore further.

3. **Final literature review**

- We conducted an exhaustive literature review to identify previous mentions of the predicted relationships, as well as for each entity with the disease. Furthermore, we also investigated mentions relating to Alzheimer's disease to prove the transfer learning. This step provided hard evidence to support our final hypotheses and candidate suggestions.

Chapter 3

Results

In the following section, we delve into the practical results of this thesis, from the creation of the knowledge graphs to the predictions and their interpretation in relation to our research question.

3.1 Knowledge Graph examination

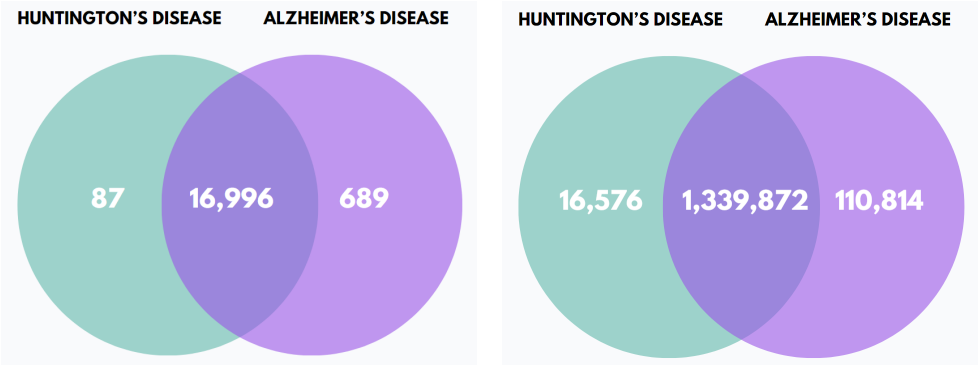
To review the topological properties of the knowledge graphs we checked the characteristics of both the entities and the relationships. This gave us an initial understanding of how similar the 2 diseases are.

Figure 3.1 contains Venn diagrams with the entities and relationships in common between the 2 graphs. We see that when the graphs were constructed a large overlap of information was included. Thus, it is safe to assume that there should also be a large overlap in the biological processes that are associated with both diseases. To have a more comprehensive view we thoroughly reviewed the entities and relationships.

3.1.1 Entities

In this project, the nodes represented biological entities and had a set of attributes that characterized them: *id*, *semantic_groups*, *preflabel* (preferred label), *synonyms*, *description* and *name*.

The category *semantic groups* was added by the BioKnowledge reviewer based on the identifier. There are 8 types: anatomical structure (ANAT), chemical entity (CHEM),



(a) Entities **(b)** Relationships

Figure 3.1: Venn Diagrams of entities and relationships. A considerable amount of the graph is shared. For the relationships, we do not include the new *no interaction* to compare in equal conditions.

disease/ disorder (DISO), gene/ genetic element (GENE), genomic entity (GENO), living beings (LIVB), physical object (PHYS), variant/ mutation (VARI) and not available (NA). In table 3.1 we show how the entities are distributed over these categories for both graphs, where the most popular category is *gene*. To assess the similarity of the graphs, the last row contains the amount present in both. The most significant observation is that almost all genes are shared between both graphs.

Graph	ANAT	DISO	GENE	GENO	PHYS	NA	Total
HD	10	78	16,213	8	4	770	17,083
AD	29	231	16,299	20	336	770	17,685
Both	4	26	16,193	2	1	770	16,996

Table 3.1: Distribution of node types in the knowledge graphs. In each column, we have the number of nodes present in the graph of that certain semantic group. The last row contains the number of nodes that both graphs have in common for that category.

Iron nodes

As stated, we aimed to enrich the subgraph of Huntington's disease related to *iron*. To define the *iron* nodes, the whole graph was inspected and only the nodes with the name or description containing the string `iron` were selected, giving **141 genes** (present in both graphs). Through some initial explorations, we found that 28 of these genes are connected to *APP*, *MAPT* and *CLU* (some of the relevant genes for

Alzheimer's). Moreover, 7 genes interacted with *HTT* (shown in figure 3.2).

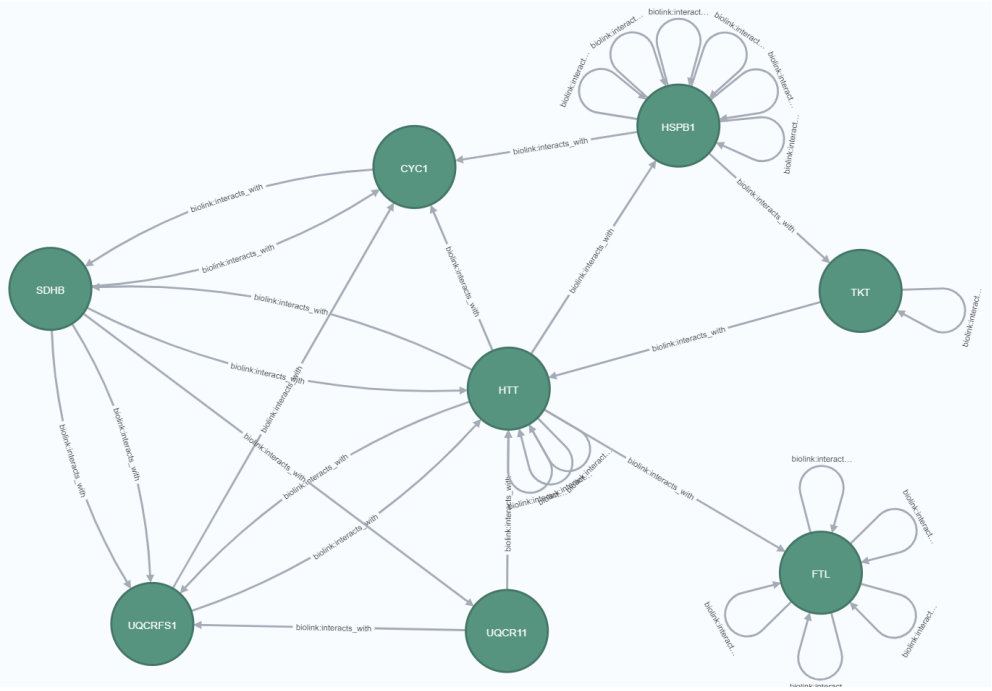


Figure 3.2: Fragment of the iron subgraph. Part of the iron subgraph showing the *iron* genes known to interact with gene *HTT*.

Recall that to obtain a subgraph aligned to the research question, we took the *iron* nodes and their first neighbors (direct connections). Moreover, only the neighbors connected to *Huntington's disease* or gene *HTT* were kept to avoid the search space increasing exponentially. This filtering was done based on the assumption that nodes that are closer are more relevant to the disease. In total, we obtained a subgraph of **484 nodes**. The semantic groups of these nodes are shown in table 3.2.

DISO	GENE	ANAT	PHYS	Total
39	434	10	1	484

Table 3.2: Distribution of node types in the iron subgraph. Presented are the number of nodes of each semantic group.

As explained in the methods section, the 484 nodes of the subgraph were paired creating 233,772 possible pairs. The ones that were already connected were excluded,

as well as the connections that were not between iron nodes and any other node. In total, we ended up with **56,288 possible new pairs**.

Statistically significant nodes

The goal of this step was to see if similar processes were involved in both diseases to characterize how similar they are. Nevertheless, in general, we see that important nodes seem to be quite general and non-specific to the study's diseases.

- **Degree (d)**: Number of edges connected to a node. **Degree centrality (d.c.)** ($\in [0,1]$) is the fraction of connections. [63].

Table 3.3 presents the top 5 most connected nodes for each graph. The nodes *protein binding*, *MYC* (transcription factor) and *TRIM67* are common to both graphs. All of them are involved in many cell functions, thus their high degree. Of particular relevance to our use case, *TRIM67* is involved in the regulation of neuron projection development. [64]

id	label	d	d. c.
GO:0005515	protein binding	10,984	0.643
HGNC:7533	MYC	5,028	0.294
HGNC:31859	TRIM67	4,438	0.260
HGNC:9514	PSG1	3,864	0.226
HGNC:18224	ZRANB1	3,851	0.225

(a) Huntington's disease

id	label	d	d. c.
GO:0005515	protein binding	11,024	0.623
HGNC:7553	MYC	5,031	0.284
GO:0005634	nucleus	4,795	0.271
GO:0005829	cytosol	4,782	0.270
HGNC:31859	TRIM67	4,450	0.252

(b) Alzheimer's disease

Table 3.3: Highest degree nodes. Top 5 nodes that accumulate the most connections in the graphs.

- **Closeness centrality (C)**: How close a node (u) is to all other nodes (n-1), using the shortest path ($d(u,v)$). We used equation 3.1: the higher the value, the more central the node is [65].

$$C(u) = \frac{n-1}{\sum_{v=1}^{n-1} d(u,v)} \in [0,1] \tag{3.1}$$

Results are shown in table 3.4. As before, the node *protein binding* is shared in both diseases. For Huntington's disease, the rest of the nodes consist of locations, which do not provide much insight into the disease. For Alzheimer's, we highlight node *APP*, which is known to be one of the genetic causes of the disease [4,13]. Being in the top

5 ratifies that this gene is associated with many of the disease's processes.

id	label	c
GO:0005515	protein binding	0.672
GO:0005829	cytosol	0.488
GO:0005737	cytoplasm	0.488
GO:0005634	nucleus	0.487
GO:0005886	plasma membrane	0.464

(a) Huntington's disease

id	label	c
GO:0005515	protein binding	0.692
HP:0000006	Autosomal dominant inheritance	0.436
HGNC:620	APP	0.434
UBERON:0003701	calcaneal tendon	0.432
UBERON:0003053	ventricular zone	0.429

(b) Alzheimer's disease

Table 3.4: Highest closeness centrality nodes. Each table shows the 5 most central nodes for each disease.

3.1.2 Relationships

Relationships are depicted by edges that connect 2 given nodes (subject and object). Each edge contains the following attributes: *subject_id*, *property_id*, *object_id*, *reference_uri*, *reference_supporting_text*, *reference_date*, *property_label*, *property_description*, *property_uri*.

Following the *biolink Model*, a certain subset of relationships/types of properties can categorize a connection. In table 3.5 we see the 25 types present in our graphs, and their distribution. Given that we saw that most of the nodes are genes, it is not surprising that almost all relationships were *interacts with*. However, this does mean we had a highly imbalanced dataset. Highlighted in gray are the most popular relationships. In bold we have marked 2 relationships we explore more in depth further.

Negative relationships

The number of negative relationships (label: *no interaction*) added to the graph was 1,328,063, the same as the most popular class (*interacts with*). As stated this was done to maintain balance and diversity.

Relationship	HD	AD	Both
active in	0	10,486	0
actively involved in	28	8,407	0
acts upstream of	0	59	0
acts upstream of negative effect	0	6	0
acts upstream of or within negative effect	0	1	0
acts upstream of positive effect	0	6	0
acts upstream of or within positive effect	0	7	0
acts upstream of or within	5	240	0
causes	1	9	0
colocalizes with	0	315	0
contributes to	9	73	9
enables	11,048	20,553	10,931
expressed in	12,993	21,085	4,627
gene associated with condition	3	18	0
has mode of inheritance	3	13	2
has phenotype	13,824	27,811	8,040
interacts with	1,318,479	1,328,063	1,316,254
located in	0	31876	0
orthologous to	11	75	0
part of	0	613	0
participates in	0	682	0
related to	6	80	2
subclass of	27	189	7
treats or applied or studied to treat	11	19	0
no interaction	-	1,328,063	-
Total	1,356,448	2,778,749	1,339,872

Table 3.5: Relationships in Huntington's and Alzheimer's graphs. Highlighted in light gray we have the most shared and popular types of relationships. In darker gray the relationship that was added to characterize unconnected pairs.

Phenotypes

Using the knowledge graph structure, it was also possible to find particular coincidences in both graphs. For instance, we checked which phenotypes were directly attached to Huntington's and Alzheimer's. This means there existed an edge with the label *has phenotype* from the disease node to a phenotype.

We found 59 phenotypes for Huntington's (figure 3.3) and 47 for all the different types of Alzheimer's (figure 3.4). Out of them, only 11 were in common: *Seizure, Agitation, Gait disturbance, Dystonia, Babinski sign,* ***Memory impairment****, Myoclonus, Dementia, Hallucinations,* ***Disinhibition*** and *Personality changes.* As expected they

were all behavioral and psychiatric disturbances, as motor dysfunction symptoms are not very common in Alzheimer's patients.

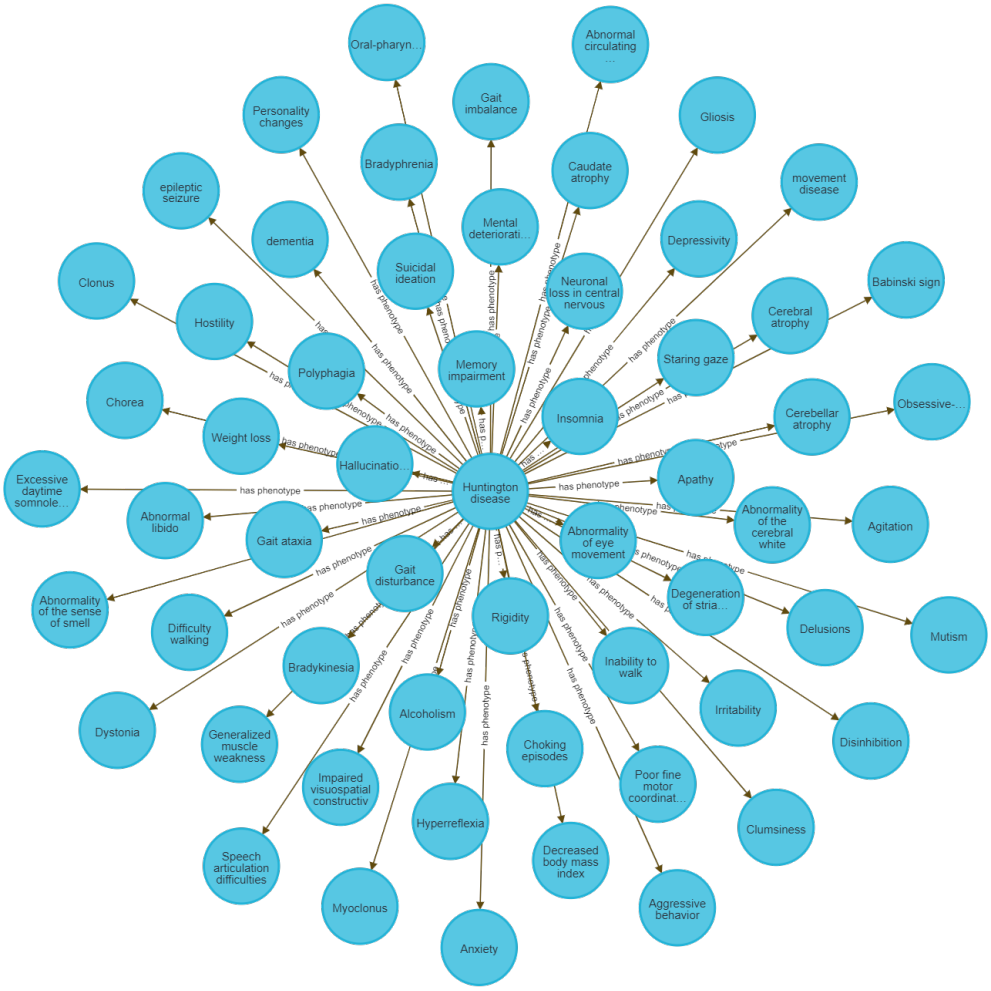


Figure 3.3: Huntington's disease phenotypes. In the center we have the Huntington's disease node, as the subject, and all its 59 phenotypes as the objects.

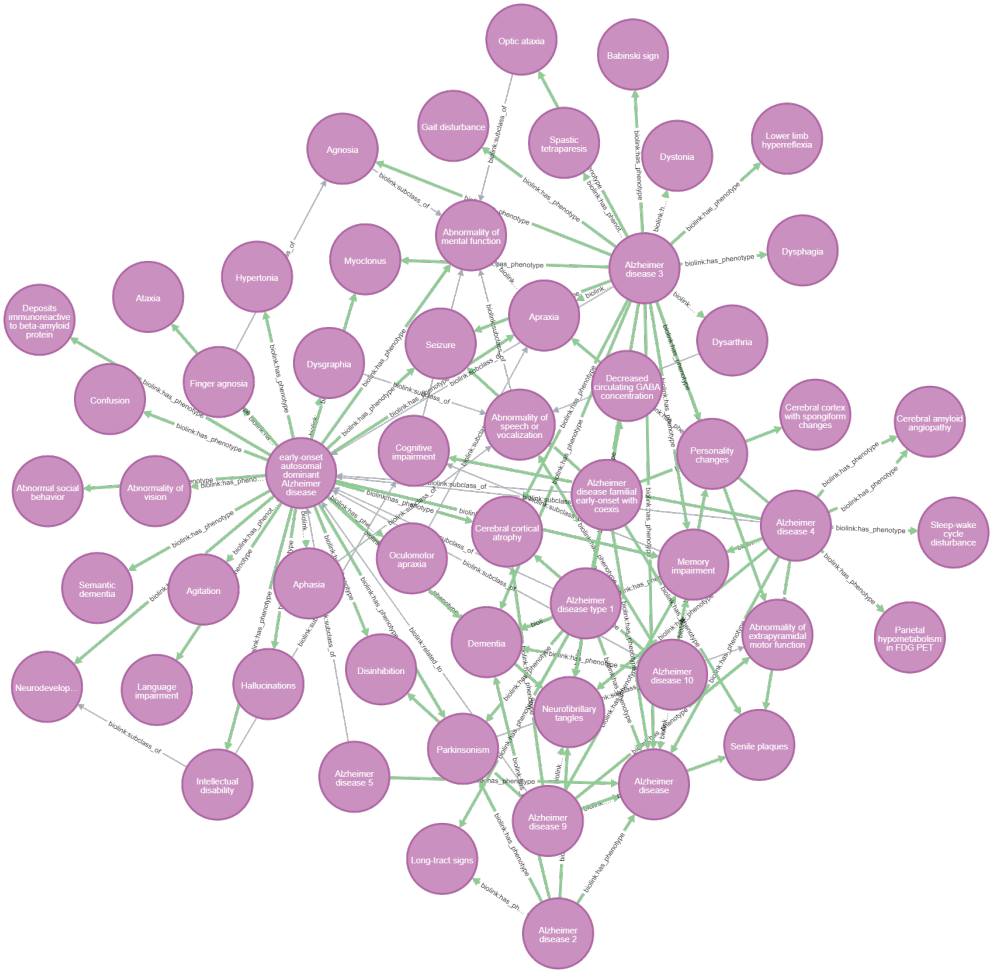


Figure 3.4: Alzheimer's disease phenotypes. Represented in the graph are the phenotypes connected to the different kinds of Alzheimer's disease. The multiple types of the disease cause the subgraph to be a bit more complex than in the Huntington's case.

Therapeutic treatments

Another interesting relationship is *treats or applied or studied to treat*. In figure 3.5 we visualize all the edges with this type of relationship. As observed in table 3.5, the Huntington's graph (in green in figure 3.5) contained this edge 11 times, 7 of which had as object the node *Huntington disease*. Thus, the Huntington's graph included 7 drugs used to mitigate the effects of the disease. In the case of Alzheimer's (in pink in figure 3.5), 16 out of the 19 edges of this kind were connected to the Alzheimer's disease node. Only 2 drugs were shared between the diseases: *minocycline* (CHEBI:50694) and *resveratrol* (CHEBI:27881). Moreover, the other 14 drugs directly connected to Alzheimer's were not present on the Huntington's graph. These are the nodes used for the drug repurposing through the BioBERT model.

```
$ MATCH p=()-[r:`biolink:treats_or_applied_or_studied_to_treat`]→() RETURN p
```

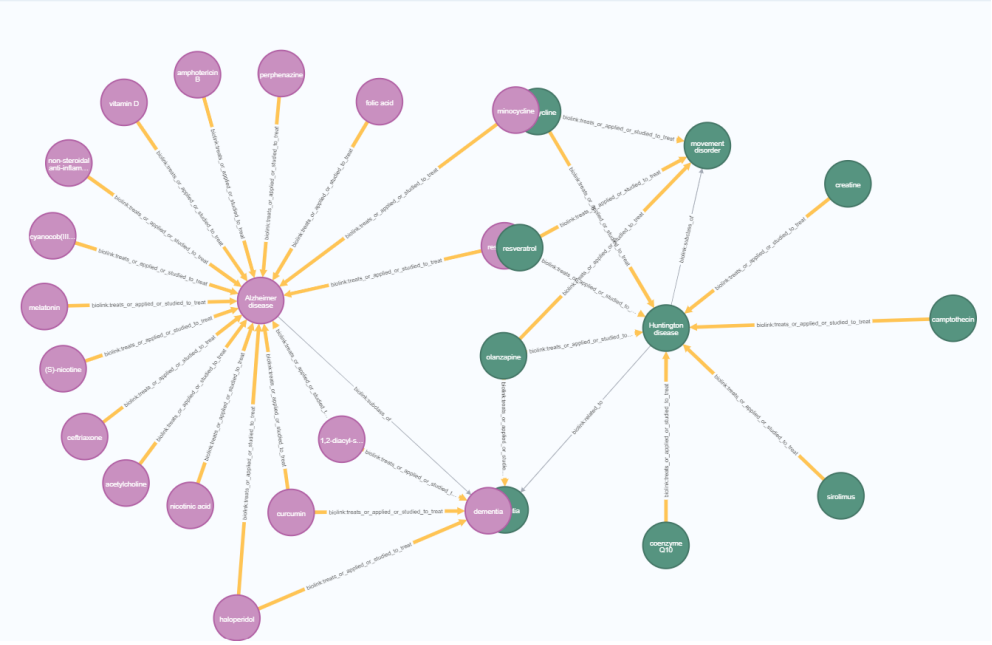


Figure 3.5: Drug subgraphs. Retrieved from Neo4j. In yellow we have the edges connecting drug to disease (*treats or applied or studied to treat*). HD subgraph (in green): 7 drugs are pointing to *Huntington's disease*. AD subgraph (in pink): 16 drugs showed. Only 2 are in common between the 2 subgraphs.

3.2 Knowledge graph completion

To perform Knowledge graph completion a BioBERT model was constructed and fine-tuned. Followed are shown the results while constructing the model and the final knowledge predicted.

3.2.1 Hyperparameter tuning

An overview of the results of the hyperparameter tuning can be visualized in figure 3.6. The vertical plot displays for each hyperparameter combination the accuracy obtained. For more detail see table 3.6, which gives an overview of the metrics obtained for each experiment. The best combination (learning rate: 1e-5, weight decay: 0.01, maximum sequence length: 128 and batch size: 8) is highlighted in green. This combination was utilized to create the final model that was then used for the following steps.

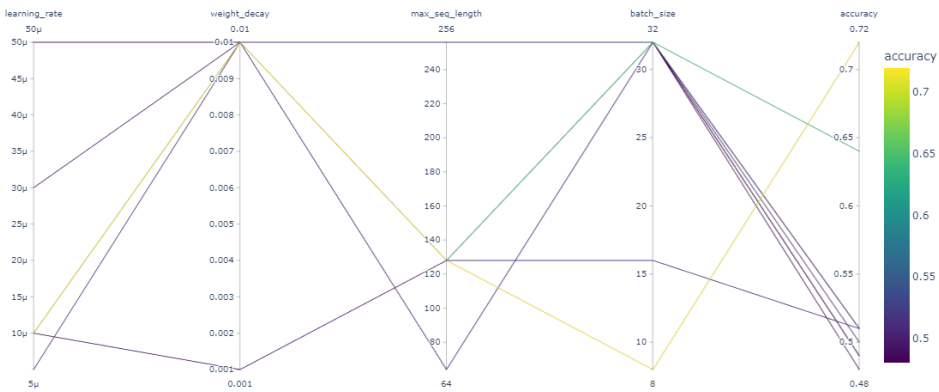


Figure 3.6: Validation results Each vertical line contains a range of values for the hyperparameters. Each line that goes through them represents a combination of hyperparameter values and the accuracy obtained. The color palette corresponds to the accuracy score. The yellow line gives the best accuracy and consists of the best combination of hyperparameters (learning rate: 1e-5, weight decay: 0.01, maximum sequence length: 128 and batch size: 8).

Parameter altered	value	precision	w	recall	w	f1	w	accuracy
learning rate	5e-5	0.02	0.23	0.04	0.48	0.03	0.31	0.48
	3e-5	0.03	0.24	0.08	0.49	0.05	0.32	0.45
batch size	8	0.24	0.71	0.3	0.72	0.26	0.71	0.72
	16	0.21	0.27	0.28	0.51	0.15	0.35	0.51
maximum sequence length	64	0.23	0.27	0.26	0.5	0.21	0.34	0.5
	256	0.24	0.73	0.43	0.51	0.21	0.35	0.51
weight decay	0.001	0.15	0.25	0.17	0.49	0.14	0.33	0.49
baseline		0.12	0.69	0.13	0.69	0.12	0.61	0.64

Table 3.6: Hyperparameter results. The table contains the values obtained for each metric in every experimental setting during the hyperparameter tuning. Every metric was calculated for each relationship type and then averaged, with the general mean and with weights (column w). The results are worse in comparison to the baseline with the exception of the downsizing of the batch size (highlighted in green).

3.2.2 Prediction results

Having established our final model, we predicted on unseen data. For each candidate pair (subject-object pair) we got a probability score for each possible relationship type.

Drug repurposing

The first area of interest consisted of the drugs used to treat Alzheimer's disease. In this case, we got no significant results. The appendix includes figure .1, where we show the distribution of the predicted classes for each drug. The highest probability is *no interaction* in all cases. Thus, the model concluded that the drug nodes should not be connected to the Huntington's disease node through any relationship. In the discussion section, we delve deeper into why this may have happened.

Iron subgraph

As stated, the model was also used to predict on a relevant iron subgraph from the HD graph that consisted of 484 nodes and **56,288 possible relationships** (entity pairs). Given that we were looking for new information the first step consisted of filtering out the triplets predicted as negative. That is the pairs of nodes where the maximum probability class was *no interaction*. The distribution of classes for the remaining **19,735 relationships** is shown in figure 3.7. Most samples were predicted as *interacts with*, but 4 other classes were also predicted: *located in*, *actively involved in*, *expressed in* and *enables*.

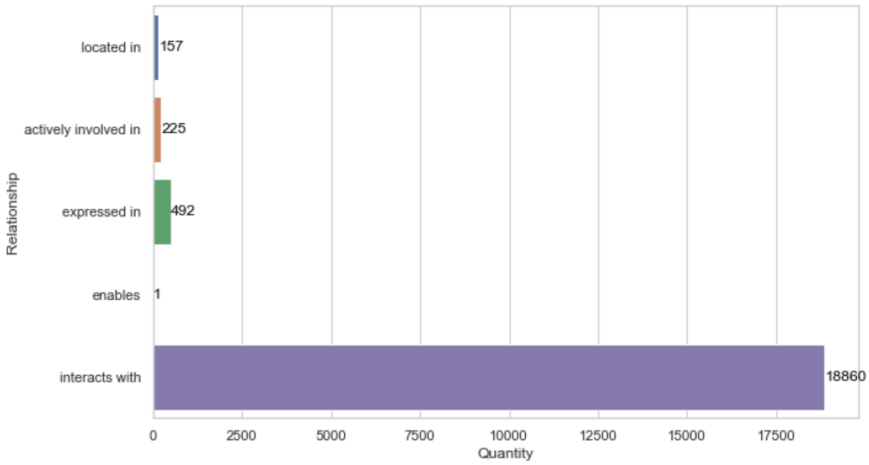


Figure 3.7: Prediction results. Showed are the positive relationships predicted for the iron dataset. Out of the 24 possibilities we see 5: *located in*, *actively involved in*, *expressed in*, *enables* and *interacts with*. The number next to each bar indicates the number of triplets predicted to be connected through that relationship.

A threshold on the probability scores was set to filter out distinctive results and perform quality control. Figure 3.8 illustrates the density plot of the probability scores. Based on the observed distribution, the threshold was set to 0.85, retaining only the more probable predictions. Furthermore, note that since the highest probability out of all classes was selected for each sample, the left tail of the distribution was already cut off, already eliminating considerably low probabilities. Consequently, in all cases, the class that was selected by the model with higher probability had significant value in comparison to all other classes. This indicated that the model is both distinctive and deterministic. The updated bar plot with the predicted labels after this filtering

is shown in figure 3.9. The relationship types remained unchanged but most of the *interacts with* predictions got filtered out.

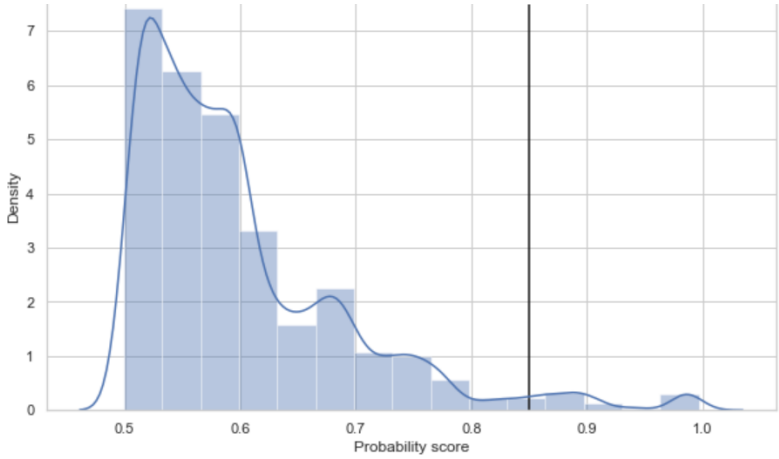


Figure 3.8: Density plot of probability scores The density plot accumulates the probability scores showing which were most popular. A vertical line in black is shown to visualize the threshold selected.

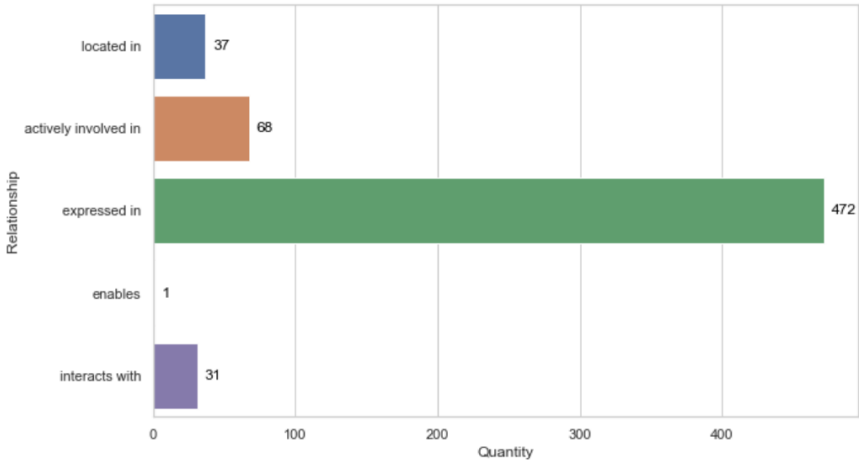


Figure 3.9: Prediction results with probability > 0.85 . Showed are the positive relationships predicted for the iron dataset after filtering out by probability score. The same 5 classes remain, with different proportions of samples (depicted by the number next to each bar).

Finally, to further refine the selection of interesting predictions, we filtered by semantic group, only keeping the groups *genes* and *diseases*. The latter englobes diseases, and most importantly *phenotypes*. Thus, the nodes that represented anatomical structures or physical objects were excluded. Figure 3.10 illustrates the distribution of the remaining **105 predictions**. Out of the three remaining classes, we deemed most interesting the *actively involved in* and *interacts with*. *Actively involved in* is defined in Biolink as: "holds between a continuant and a process or function, where the continuant actively contributes to part or all of the process or function it realizes" [48]. The predictions for this relationship linked several genes to three phenotypes: *decreased body mass index*, *disinhibition* and *memory impairment*. Note that the last two phenotypes were shared between the two diseases, while the first was only directly connected to Huntington's disease on our knowledge graphs. The relationship *interacts with* is defined as: "holds between any two entities that directly or indirectly interact with each other" [48]. In our predictions, it related genes to other genes. The entire list of predictions for this subset is provided in a supplementary table in the appendix.

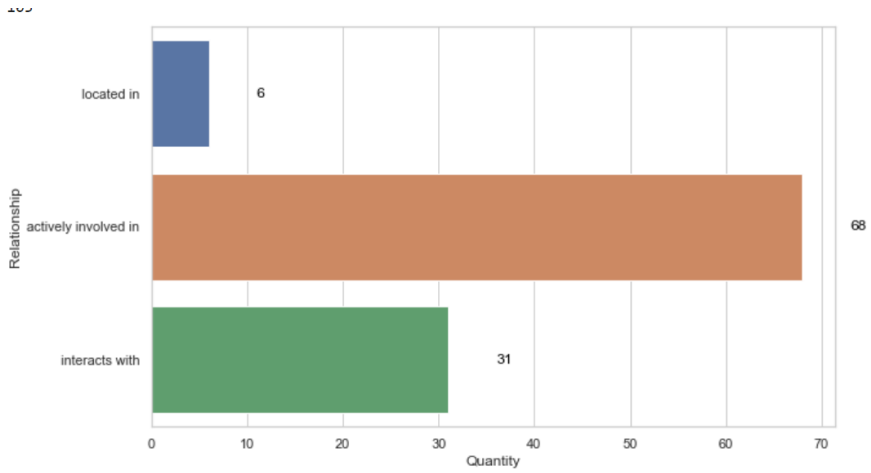


Figure 3.10: Prediction results with probability 0.85 and semantic groups gene and disease. Showed are the positive relationships predicted for the iron dataset after filtering out by probability score and semantic group. There are 5 classes: *located in*, *actively involved in*, *expressed in*, *enables* and *interacts with*. The number next to each bar indicates the number of triplets predicted to be connected through that relationship.

3.3 Analysis of the predictions

In this last step, we analyzed some of the predictions out of the subset of the most relevant results. For each of the 2 aforementioned relationships, *interacts with* and *actively involved in*, we selected a prediction to investigate.

The investigation consisted of an initial exploration of both entities in their previous context in the knowledge graph. Followed by a deep investigation of relevant literature regarding the predicted relationship.

3.3.1 Gene interaction prediction: FTL-TP53

In the case of the *interacts with* relationship, which connects 2 entities that interact with each other, we got 31 gene interactions. These represent 31 pairs of genes that are predicted to interact with each other. The complete table with the probabilities scores can be found in the appendix. Following the systematic approach detailed in the methods section (Chapter 2), we reviewed literature and consulted external databases to choose a relevant prediction. The selected interaction was **FTL - interacts with - TP53**. We will first provide a thorough explanation of the selection process for this particular gene pair. Then, we will discuss the pertinent analysis for the prediction.

Using the STRING database, we identified that 4 out of the 31 predicted gene interactions were already known. Although these interactions were not present in our knowledge graphs and were predicted during knowledge graph completion, this information was not novel. Figure 3.11 shows the result of the STRING investigation with the interactions marked with a black box (*DDR2*, *MAF1*, *SDHB*, *LTF* interact with *TP53*).

Regarding the 27 other interactions, 23 of them relate genes to gene *TP53* (HGNC:11998). This gene encodes tumor protein p53 which is a transcription factor involved in mechanisms such as cell proliferation, apoptosis and ferroptosis. There is on-going research relating *TP53* with iron homeostasis and how it could affect diseases associated with abnormal iron levels [66]. *TP53* has been shown to be increased in Huntington's disease and there have been several studies investigating how its regulation can decrease *HTT* aggregation [67,68]. Its context within the Huntington's disease knowledge graph is shown in figure 3.12. The most relevant connections are with *HTT* and a group of common phenotypes with the disease.

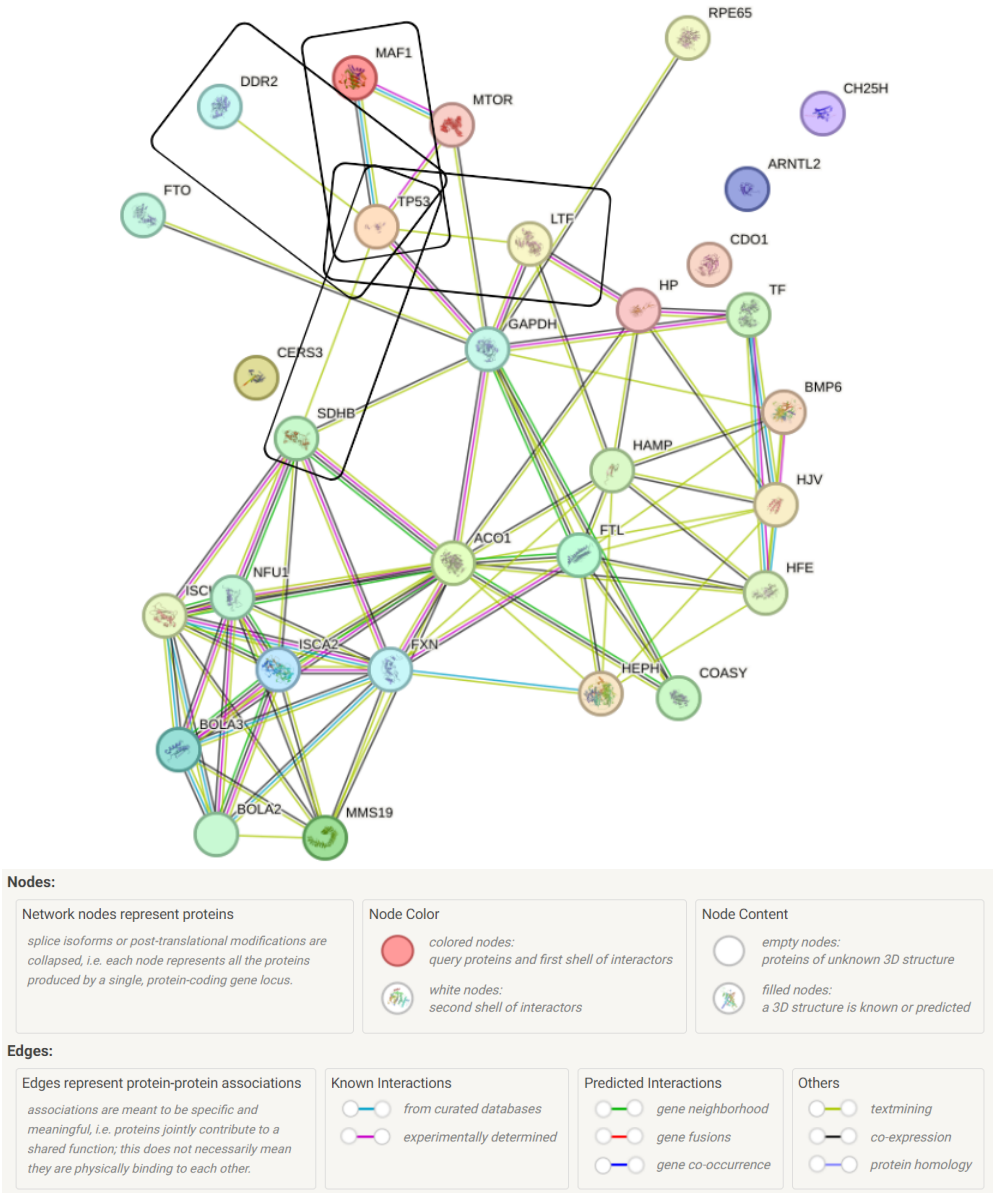


Figure 3.11: Known interactions from STRING. Every edge represents an interaction found by the STRING database, from diverse sources as depicted in the legend. The relationships marked with a black box were not reported in our knowledge graph and were predicted during the knowledge graph completion.

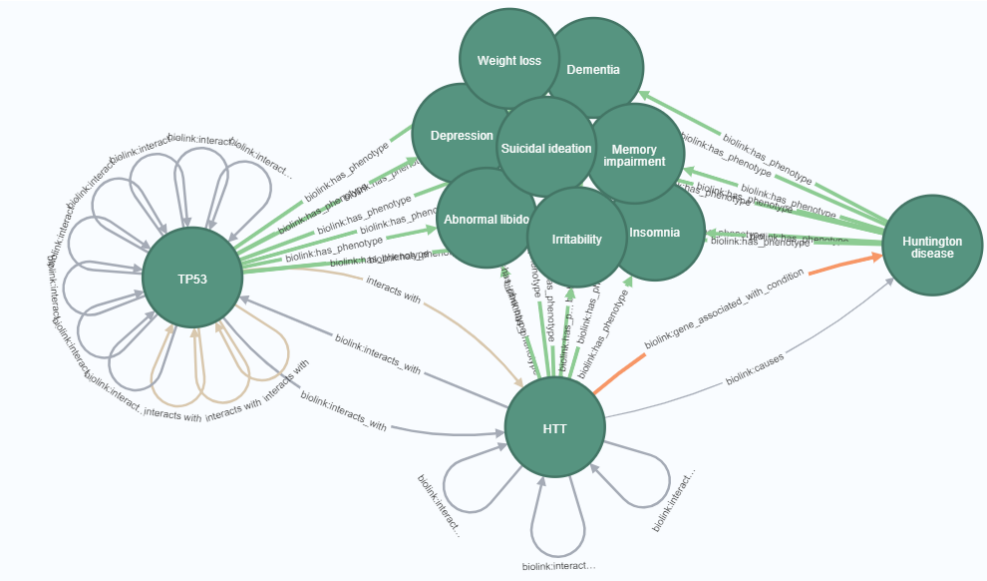


Figure 3.12: Relevant nodes between *Huntington's Disease* and *TP53* (HD graph). The gene is directly involved to several relevant phenotypes of the disease, as well as interacting with *HTT* gene.

Out of this set, **FTL - TP53** was chosen as an interesting prediction to explore. The gene *FTL* (HGNC:3999) encodes the light subunit of the ferritin protein, the major intracellular iron storage protein [69]. It is the cause of *Neuroferritinopathy*, a Huntington's disease-like disorder that is also characterized by chorea, dystonia and speech and swallowing deficits [70]. This prediction was selected because there is a direct connection of *FTL* with *HTT* in the HD graph (Shown in the iron subgraph Figure 3.2). We found a relationship of *interacts with* that was retrieved from Monarch. It is based on an experimental study in a yeast-two hybrid system, where a physical association was found between *HTT* and *FTL* [71]. This relationship is also displayed in figure 3.13 with other relevant nodes from the HD graph, such as common phenotypes between *FTL* and Huntington's disease.

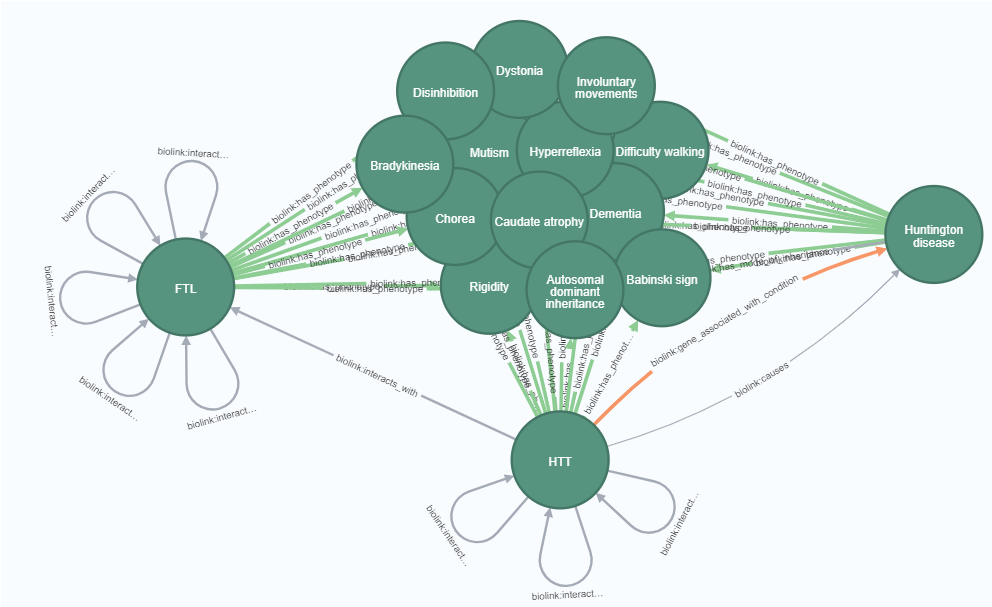


Figure 3.13: Relevant nodes between *FTL* and *Huntington's Disease* (HD graph). Shown are the experimental interaction with HTT and several common phenotypes.

The context of the 2 genes together in the knowledge graphs is shown in figure 3.14. The nodes are only one node away. Some relevant neighbors are the iron genes *TFRC* (HGNC:11763), which encodes a receptor required for iron uptake and neurologic development, *FTH1* (HGNC:3976), which encodes a heavy subunit of ferritin, as well as the aforementioned *HTT*. Moreover, both genes contribute to the phenotype *Dementia* (common phenotype in Alzheimer's and Huntington's).

Through the literature research, we could not find any study directly relating *FTL* and *TP53* genes in the field of neurodegeneration. However, there is a recent study suggesting *FTL* plays an essential part in liver hepatocellular carcinoma (LIHC). Li et al. showed that *FTL* expression in LIHC, promotes mutation and expression of *TP53* [72]. And on a pancreatic cancer study, it is demonstrated how *TP53* is highly associated with *FTL* (and *FTH1*) based on the role of *TP53* in ferroptosis. The authors go on to suggest that these genes could modulate pancreatic carcinoma, but that more research is required, specially in the role of *TP53* in ferroptosis [73].

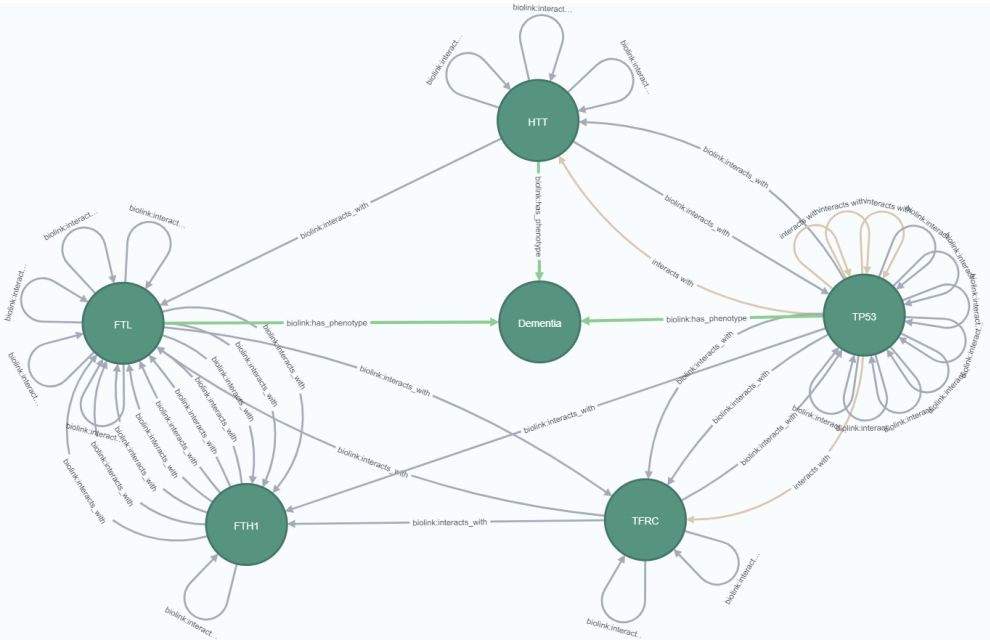


Figure 3.14: Shortest path between *FTL* and *TP53* (HD graph). Their common neighbors include 2 iron genes (*TFRC* and *FTH1*), *HTT* and the phenotype Dementia.

Ferroptosis is a form of regulated cell death due to iron accumulation and lipid peroxidation. The *FTL* protein plays a regulator role in the mechanism through its protein (in charge of iron storage). Activated *TP53* can both promote and inhibit ferroptosis, and as stated it has been mainly studied in regards to cancer [74]. In the ferroptosis pathway (shown in figure 3.15), both *FTL* and *TP53* are present but considerably far away. An alteration on either of them could potentially have an effect on the other, but there is no clear path showing it explicitly.

Ferroptosis is one of the pivotal focuses of Alzheimer's disease. Thus, research on this pathway, and specifically on this gene interaction, would potentially benefit Huntington's research [76]. Other interesting findings connecting both genes individually to our research area further prove the transfer learning of this prediction. In Alzheimer's disease, following the hypothesis that iron accumulation can cause the disease to progress, it was found that the protein ferritin light chain (*FTL*) can accumulate in microglia [77]. Furthermore, upregulation of p53 (protein generated by *TP53*) has been connected with pathogenesis. It is involved in many cell stress control mechanisms relevant to the disease such as the control of Aβ peptides [78, 79].

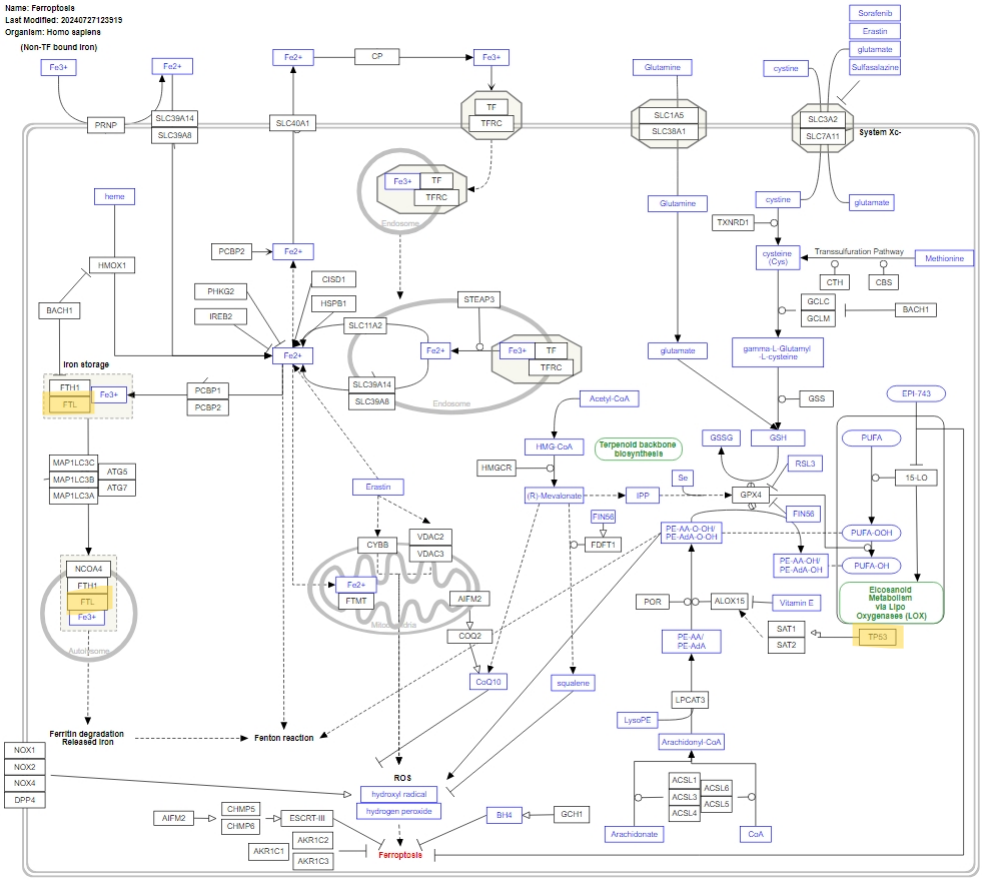


Figure 3.15: Ferroptosis pathway [75]. Highlighted are genes *FTL* and *TP53*. They appear to be rather separated in the mechanism.

All the presented findings are summarized in table 3.7 for easier understanding. Based on the gathered information, we hypothesize that altering the expression of *TP53*, and therefore the production of p53, could change the expression of *FTL*, potentially alleviating or regulating symptoms of Huntington's disease patients and reducing the iron accumulation in the brain. This hypothesis becomes relevant in the Huntington's disease context given the interaction of both *FTL* and *TP53* with *HTT*.

	Iron	Alzheimer's	Huntington's	Other
TP53	dysregulation of iron homeostasis	upregulation of p53 causing Aβ plaques	p53 increased in HD, regulation can decrease *HTT* aggregation	
FTL	encodes light subunit of ferritin	*FTL* protein accumulation in microglia	experimental association of *HTT* and *FTL*	causes Neuroferritinopathy (HD-like disorder)
both genes	involved in ferroptosis	have phenotype dementia	have phenotype dementia	interaction in cancer studies

Table 3.7: Summary of *FTL-TP53* analysis

3.3.2 Gene - phenotype prediction: BMP6 - memory impairment

The second kind of relevant predictions generated were relationships between genes and phenotypes, linked by the relationship *actively involved in*. As with the gene interactions, we selected a case to investigate thoroughly. There are 47 genes that were predicted to be related to a phenotype. All of them were predicted to be involved with *memory impairment*, 20 only to *decreased body mass index* and 1 to *disinhibition*. This last one was with gene *BMP6*, which was also predicted to be related to *memory impairment*. The prediction we decided to explore is **BMP6 - actively involved in - memory disinhibition**.

BMP6 (HGNC:1073) is a bone morphogenetic protein (BMP) gene, which is part of the TGF-β (transforming growth factor-beta) superfamily. This gene encodes a secreted ligand of the TGF-β superfamily of proteins, which regulates processes such as iron homeostasis. *BMP6* is associated with iron overload [80]. The TGF-β superfamily has been shown to be involved in the pathogenesis of neurological disorders [81].

The most relevant study on Huntington's disease consisted of a *Drosophila* HD model. The authors demonstrated that reducing BMP signaling activity could reduce the severity of the disease. They found that pathogenic *HTT* can accumulate in nerve terminals, interfering with the signaling mechanism and causing aberrant activation of BMP signaling, which leads to neuronal dysfunction [82]. The specific role of gene *BMP6* in Huntington's disease has yet to be studied. Soldati et al. performed a study on the role of *REST* (HGNC:9966) in neural stem/progenitor cells, where they identified *BMP6* as a candidate regulator of adult neuronal stem cell (aNSC) function.

Given that *REST* binding is implicated in the aberrant transcription of neuronal genes in Huntington's disease, the authors discussed that exploring *BMP6* regulation, in the context of altered REST, could provide valuable insight into the disease. [83]

The graph exploration for *BMP6* in Huntington's is shown in figure 3.16. It contributes to two common phenotypes: apathy and weight loss. Through the genes *NOG* (HGNC:7866), known to inactivate *BMP4* (also shown) and other TGF-β family members [84], and *SMAD9* (HGNC:6774), which transduces signals from TGF-beta family members and is activated by BMP and SMAD4 (also shown in the graph) [85], we see its interaction with *REST*. This last one interacts with *HTT*.

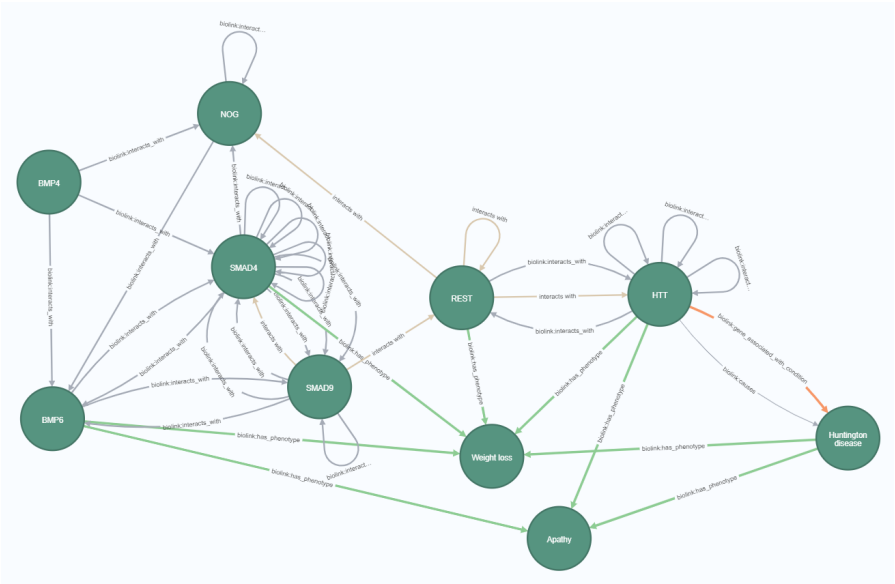


Figure 3.16: BMP6 in HD graph Shown are some relevant interactions with other genes of the TGF-β family (BMP4, SMAD4 and SMAD9), genes NOG, REST and HTT, and common phenotypes with Huntington's disease apathy and weight loss.

In the field of Alzheimer's, there is numerous research on *BMP6*. It is proposed that members of the TGF-β superfamily, TGF-βs and BMPs, are involved in its pathogenesis. TGF-βs have been observed to contribute to the β amyloid (Aβ) accumulation, microglia activation and neurodegeneration. For *BMP6* it was shown it augmented in the hippocampus of AD patients, potentially altering neurogenesis. Nevertheless, the research on the role of the TGF-β/BMP signaling in AD is still on-going. [81]

In figure 3.17 are the relationships of BMP6 in the Alzheimer's graph. We show more information to understand how this prediction came to be, which knowledge may have been leveraged to pair *BMP6* with the phenotype *memory impairment*. Some relevant nodes include another BMP gene (*BMP4*), BMP receptors (which encode protein receptors that bind to BMP to transmit signal to the cells) [86], some phenotypes (associated to either *BMP6* or to its interacting genes and Alzheimer's), *HFE* (HGNC:4886) a gene that encodes an iron regulator protein [87], and relevant genes to Alzheimer's disease (*APOE*, *APP*, *ADAM10*).

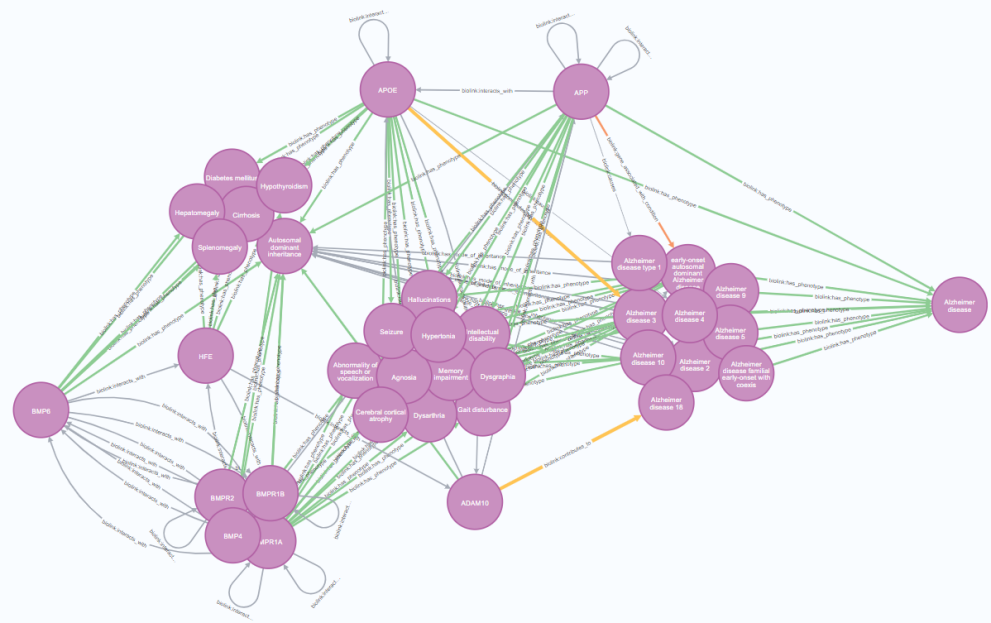


Figure 3.17: BMP6 in AD graph. Relevant phenotypes connected to *BMP6* and interactors of *BMP6*. Several of its interactors are from the TFG-β family, relevant AD genes and iron node HFE.

The phenotype *memory impairment* (HP:0002354) is defined by the Human Phenotype Ontology (HPO) as "An impairment of memory as manifested by a reduced ability to remember things such as dates and names, and increased forgetfulness." It is associated with Huntington's and several types of Alzheimer's [88]. In HD it is one of the most prominent deficits in the earliest stages of the disease [89]. In AD, as the cause for the disease is not clear, it is often used to diagnose individuals, particularly in the early stages of the disease, through Neuropsychological assessment techniques [90]. This cognitive impairment has been shown to be caused by iron overload due to oxidative

damage in the brain. [91]

The relationship of *memory impairment* with the gene *BMP6* in the context of Huntington's is illustrated in figure 3.18. It shows that the interaction could be through the transcription factor *NF1* and receptor *BMPR1A* (HGNC:1076), which interacts with multiple members of the TGF-β family such as *BMP4*. [92]

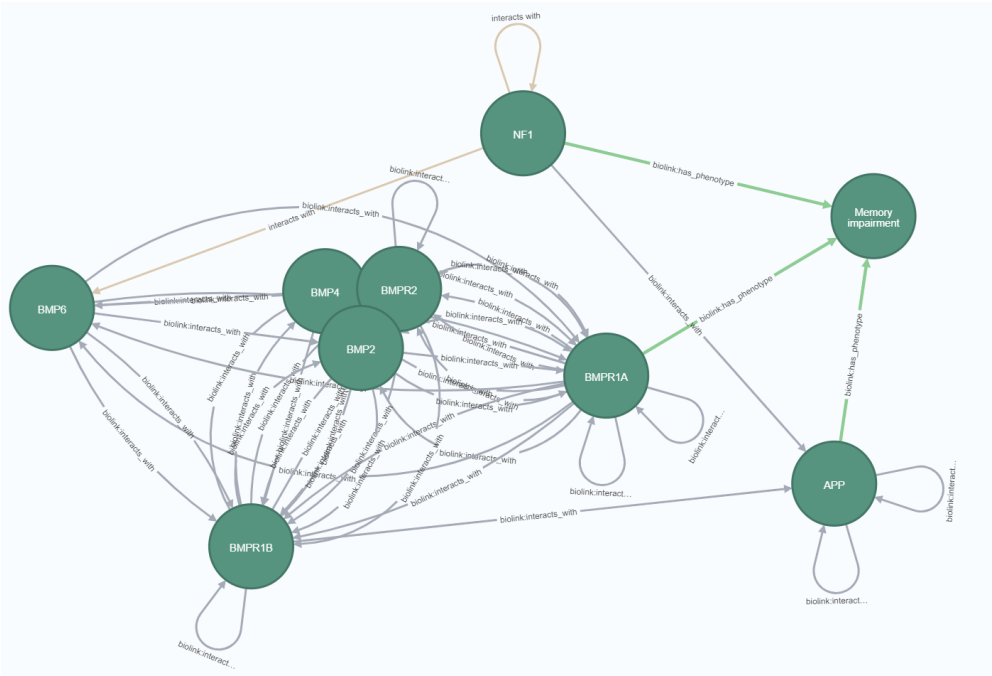


Figure 3.18: BMP6 - memory impairment in AD graph. Linked through *BMPR1A* gene and transcription factor *NF1*

In literature, we found no mention directly linking *BMP6* with *memory impairment*. There is a study relating the upregulation of *BMP4* (HGNC:1071) with memory deficits in Alzheimer's. The authors found *BMP4* inhibited nerves in the hippocampus causing *BMP4* transgenic mice to exhibit memory impairment. As there are numerous types of BMP in the hippocampus, the authors go on to suggest that BMP factors could regulate the activity [93]. Recall that in Alzheimer's disease *BMP6* is shown to be augmented in the hippocampus. Furthermore, through the graph exploration, we saw *BMP4* interacts with *BMP6*.

A summary of the information relating to this prediction can be found in table 3.8.

Looking at the entire picture, we hypothesize that exploring the activity and interactions of *BMP6* in the hippocampus could provide valuable insight into the mechanisms of an important phenotype such as *memory impairment*, and therefore into the disease. Analyzing the effect of this iron-related gene, which is key in the phenotype, could provide a new understanding of therapeutic strategies.

	Iron	**Alzheimer's**	**Huntington's**
BMP	iron homeostasis regulated by multiple members of BMP	involved in pathogenesis	HTT accumulation causes aberrant activity by BMP leading to neuronal dysfunction
BMP6	causes iron overload	augmented in the hippocampus altering neurogenesis	discussed regulation of BMP6 when REST is altered to reduce pathology
Mem. Imp.	potential cause for the phenotype (oxidation in the brain)	important phenotype used for diagnosis, linked to *BMP4*	early stage phenotype

Table 3.8: Summary of *BMP6-memory impairment* analysis

Chapter 4

Discussion

The primary research question of this thesis consisted of assessing if new data on a rare disease can be gained using information on a common disease, stored in a knowledge graph, using Large Language Models. We managed to create a a proof-of-concept for knowledge transfer between diseases. Its relevancy is not only in the computational applications but in the advantage it can offer researchers. Performing an experiment in a wet lab can be significantly expensive and should be guided by strong evidence. Rare diseases suffer or a lack of funding, and this approach can aid researchers in prioritizing their experimental investigations. The tool developed has the potential to become a key resource for identifying candidate hypotheses for further exploration.

4.1 Relevant findings

One of the goals we strove to investigate was if we could uncover new knowledge on current research areas of the disease. Our computational predictions enable the generation of interesting new mechanistic hypotheses partially supported by evidence found in the graph in the literature. Significant efforts ([8–10,18,19]) have been made to characterize the mechanisms underlying iron accumulation in the brain and its impact on neurodegenerative diseases. In Huntigton's disease, altered levels in the basal ganglia have been shown to affect HTT accumulation, potentially accelerating the disease progression [8–10]. We thoroughly examined two specific predictions that could offer new directions for the disease's research. On the one hand, the potential interaction of *FTL* and *TP53* may help reduce iron levels in the brain, consequently alleviating some of the disease's phenotypes. This hypothesis is supported by pre-existing interactions

between the two genes in other domains [71–73]. Recent cancer studies ([72,73]) have demonstrated that the two genes can influence each other's expression and even promote mutations due to their common involvement in the ferroptosis pathway [74]. Additionally, both genes have been implicated in the pathogenesis of Huntington's disease [67,68]. On the other hand, we hypothesized that regulating *BMP6* could be key in preventing *memory impairment*, one of the earliest signs of Huntington's disease [89]. The phenotype's cause remains unclear, although several studies point to iron overload [91]. The prediction contributes to the on-going investigation of the BMP family's role in regulating iron homeostasis in the brain and its involvement in the pathogenesis of neurological disorders [80,81].

To perform the predictions, our aim was to employ knowledge graphs assuming that the semantic structures, which include detailed relationships between biological entities [26], are crucial for understanding disease mechanisms and discovering new knowledge. Knowledge graphs have been used for knowledge discovery in previous research, such as in drug repurposing for COVID-19 [40]. To fully leverage the potential of knowledge graphs, we performed knowledge graph completion with LLMs. Knowledge graph completion consists of exploiting the data to infer new information to be added to the graph [41]. Although several tools exist for this purpose ([42,43]), we utilized Large Language Models. These models have emerged as powerful tools for knowledge discovery [32] and are well-suited for applications involving diverse data types, including knowledge graphs [34]. Additionally, LLMs have demonstrated effectiveness in biomedical research [46], but to the best of our knowledge not to perform knowledge graph completion in a biomedicine task. The use of a Large Language Model allowed a straight forward adaptation, as these computational models are capable of creating embeddings that handle different data structures, have powerful semantic comprehension, and can integrate novel data [94]. Upon reviewing the model's predictions, we were able to validate several of them. Finding some of the predictions in the STRING database [62] provided confidence in the model's outputs. Importantly, the limited number of matches suggests that training the model with our knowledge graphs does not overfit it, indicating that our approach represents a source of novel insights. Although further research is needed, our framework demonstrates its potential to make informed inferences based on existing data. Moreover, the combination of the knowledge graph semantic structures with the capabilities of Large Language Models in biomedicine is an emerging approach that has not been widely explored to date. Our study shows it has the potential to become a powerful tool, facilitating the discovery

of new knowledge and enhancing our understanding of complex diseases.

Moreover, the fine-tuning of the LLM facilitated the implementation of the knowledge transfer from Alzheimer's to Huntington's disease. The model successfully learned from the Alzheimer's graph structures and applied them to Huntington's, as in the case of the BMPs that have been widely studied in the context of AD [80,81,93] By comparing the Knowledge graphs as a whole, we were able to see that most of the processes associated with the diseases overlapped. This finding further supported the hypothesis that the two neurodegenerative diseases are similar, suggesting insights gained from one may enhance our understanding of the other. Further investigation into the specific entities offered a deeper understanding of the processes involved and how they could be related. Although LLMs are often considered black boxes, the literature found helped us make the predictions more understandable, supporting its plausability for future experimental testing. For both of the reviewed predictions, we were able to find literature validating their relevancy and previous research in Alzheimer's disease. Firstly, *FTL* was shown to accumulate in the brain of Alzheimer's disease patients [77] and protein p53, encoded by gene *TP53*, is involved in the formation of the Aβ plaques, a known cause for the disease [78,79]. Secondly, *memory impairment* is a phenotype in common of Huntington's and Alzheimer's [89,90], that in AD has been shown to be exhibited by *BMP4* transgenic mice [93]. Notably, *BMP6* interacts with *BMP4* [80]. Furthermore, in AD *BMP6* has been shown to be augmented in the hippocampus [81], where it interacts with other BMP genes [93].

4.2 Limitations

Throughout the study, several limitations had important effects on the methodology.

The data used to create the knowledge graphs came from the BioKnowledge Reviewer [26], which had to be updated with the new Monarch [27] version from this same year, 2024. It is notable to mention that the adaptation was a considerable setback, mainly due to the limited documentation. Furthermore, a large proportion of the graph's entities were genes. Although this was not unexpected given that the data gathering revolved around genes, it resulted in a highly imbalanced dataset. In computational models, data imbalance can become very problematic as it can introduce unwanted biases and affect the training process. To mitigate it, we adapted the loss function and applied class weights, in addition to creating the negative sample space [53]. Nevertheless, this issue can not be considered to be fully resolved. Addi-

tionally, the limited amount of drug-related data from the Monarch database [27], with only 16 drugs related to Alzheimer's and 7 to Huntington's, constrained our ability to derive meaningful insights about drug interactions. Even though the inner workings of LLMs are not fully explainable [31], the model was fine-tuned with mainly gene data [27, 47], primarily learning how genes relate to other entities within the graph. We forced the model to learn how the structures in Alzheimer's worked and in doing so it mostly learned how genes relate. As the data concerning drugs and treatments was significantly scarce, the model lacked sufficient information to derive meaningful connections involving drugs and treatments, rendering the drug-related predictions non-significant.

Upon filtering the results, a notable proportion of our gene interaction predictions were associated with the *TP53* gene. This outcome was not entirely unexpected, as there is extensive on-going research surrounding this gene and its effect on several diseases [66–68]. Its role in multiple cell functions leads to interactions with numerous genes, making the discovery of new interactions plausible. Nevertheless, it is a concern that the model may have identified the relationships trivially, simply because *TP53* is known to interact with a large number of genes.

The computational power required posed significant challenges. Although this was dealt with using the GPU from the Snellius supercomputer [61], its capacity is not limitless. This called for consideration of resource usage, including tasks such as code optimization and reduction of the sample space.

One relevant limitation in our literature analysis arose with the phenotype identifications. While our knowledge graphs associate phenotypes with HPO IDs (ensuring traceability and hierarchical organization, for instance within cognitive impairment phenotypes) [88], the literature often lacked consistent use of identifiers, making it challenging to determine if researchers were referring to the same mechanisms. This issue underscores the importance of using FAIR (Findability, Accessibility, Interoperability, and Reusability) data principles [95] to enhance the accuracy and interoperability of information in transfer learning projects.

Despite conducting extensive literature research to support the results, it is important to highlight caution when using LLMs. It is well reported that they can generate misleading or false content, namely hallucinations [31]. Consequently, the findings from this study be regarded as strong indications warranting further investigation.

4.3 Future work

The use of the BioKnowledge Reviewer [26] facilitated the creation of more comprehensive knowledge graphs, with a broader range of entities, including therapeutic agents. To enhance the model further, incorporating additional sources of therapeutic data would be advantageous. Additionally, integrating experimental transcriptomics data could significantly improve the validation of gene interactions, and provide a better understanding of gene relationships. It would also be addressing one of the main limitations, the data imbalance. More diverse data would likely create a sturdier model that would produce superior results.

One of our aims was to harness the semantics encoded in a knowledge graph. Although this was achieved using the LLM, we acknowledge there is still potential for improvements. Our methodology involved utilizing the semantic information to construct sentences by concatenating the graph's triplets, which were then employed as input. While this approach provided richer information compared to a basic network structure [24, 94], the graph could be further exploited by incorporating additional data on the relationships For instance, integrating temporal data (such as that found in Temporal Knowledge Graphs, which capture dynamic relationship [96]) or conducting preliminary semantic analysis to augment the attributes of entities, could offer significant improvements.

In terms of the actual prediction model, it could be improved to do also entity predictions. To infer which nodes should be connected to already existing nodes on the graph. Computationally this would make the model significantly more complex as the possibilities would increase exponentially [41]. Thus, it should be done with some criteria of which nodes could be added to the graph based on some hard evidence.

Our research focused on two main areas: drug repurposing and the iron hypothesis. Nevertheless, the framework was constructed to be adaptable to other research areas within the field. Given the various dimensions of research on Huntington's disease, it would be valuable to explore whether this framework could yield significant insights in other areas of study.

To further refine the investigation it would be advantageous to get expert opinions. From which areas should be considered of interest to how the candidate predictions should be prioritized. The aim of the tool is to facilitate

Finally, while the literature review provided validation for the predictions made in

this project, the results require further investigation. The in-silico predictions offered promising insight, but they should be corroborated by experimental evidence to confirm their validity and relevance.

4.4 Conclusions

This research demonstrated that similarities between neurodegenerative diseases, specifically Alzheimer's and Huntington's, can be leveraged to gain new insights. By employing knowledge graphs and LLMs, we identified relevant suggestions for new research directions in Huntington's disease. Our predictions suggest precise research directions by identifying specific genes related to iron accumulation in the brain that should be investigated to potentially alter disease progression.

The results obtained support the importance of exploring the regulation of iron. The potential interactions and the comprehensive literature review to back them up reinforce the significance of this area for investigating the disease's pathology. The developed framework effectively utilized the semantic structures of knowledge graphs through state-of-the-art Large Language Models, allowing the leveraging of the semantic information in the graphs to enhance the model's performance [31]. The tool also allowed a seamless transfer learning by applying insights into Alzheimer's disease, a common neurodegenerative disease, to infer new knowledge about Huntington's disease, a rare disease. This was further validated during the literature review, as several studies on Alzheimer's disease corroborated the rationale for the proposed relationships [77–79, 81, 90, 93].

In conclusion, our framework demonstrates significant potential for generating relevant predictions that can lead to the discovery of new knowledge. By integrating state-of-the-art methods, the structured knowledge graphs and the powerful LLMs, it performs transfer learning to make informed predictions. Employing this tool could significantly aid researchers overcome the challenges associated with data scarcity and limited funding by focusing on research with a strong evidence base.

Bibliography

[1] EURORDIS. What is a rare disease?, February 28 2024.

[2] Orphanet. Quality charterabout rare diseases, 2024.

[3] pharmaphorum. Why is rare disease funding rare?, 2013.

[4] Dagmar E. Ehrnhoefer, Yee Wong, and Michael R. Hayden. Convergent pathogenic pathways in alzheimer's and huntington's diseases: shared targets for drug development. *Nature Reviews Drug Discovery*, 10(11):853–867, 2011.

[5] National Institute of Neurological Disorders and Stroke. Huntington's disease, 2023.

[6] Christopher A Ross and Sarah J Tabrizi. Huntington's disease: from molecular pathogenesis to clinical treatment. *The Lancet Neurology*, 10:83–98, 2011.

[7] Nancy S. Wexler. Venezuelan kindreds reveal that genetic and environmental factors modulate Huntington's disease age of onset. *Proceedings of the National Academy of Sciences*, 101(10):3498–3503, 2004.

[8] Carolina Sánchez-Castañeda, Ferdinando Squitieri, Margherita Di Paola, Marco Dayan, Marco Petrollini, and Ugo Sabatini. The role of iron in gray matter degeneration in huntington's disease: A magnetic resonance imaging study. *Human Brain Mapping*, 36(1):50–66, 2014.

[9] J. Chen, E. Marks, B. Lai, Z. Zhang, J. A. Duce, L. Q. Lam, and J. H. Fox. Iron accumulates in huntington's disease neurons: Protection by deferoxamine. *PloS One*, 8(10):e77023, 2013.

[10] M. Muller and B. R. Leavitt. Iron dysregulation in huntington's disease. *Journal of Neurochemistry*, 130(3):328–350, 2014.

[11] National Institute on Aging. Alzheimer's disease fact sheet, 2023.

[12] Johns Hopkins Medicine. Alzheimer's disease, 2024.

[13] Alzheimer's Association. 2023 alzheimer's disease facts and figures. *Alzheimer's & Dementia*, 19(4):1598–1695, 2023.

[14] GeneCards Database. Clu gene - genecards — clus protein — clus antibody, 2024.

[15] NCBI. Mapt microtubule associated protein tau [homo sapiens (human)] - gene - ncbi, 2024.

[16] NCBI. Bace1 beta-secretase 1 [homo sapiens (human)] - gene - ncbi, 2024.

[17] L. C. Walker. Ab plaques. *PubMed*, 1:1, 2020.

[18] Jin-Lian Liu, Yan-Guo Fan, Zhao-Sheng Yang, Zuo-Yi Wang, and Chuang Guo. Iron and alzheimer's disease: From pathogenesis to therapeutic implications. *Frontiers in Neuroscience*, 12:632, 2018.

[19] F. Wang, J. Wang, Y. Shen, H. Li, W.-D. Rausch, and X. Huang. Iron dyshomeostasis and ferroptosis: A new alzheimer's disease hypothesis? *Frontiers in Aging Neuroscience*, 14, 2022.

[20] R. J. Ward, F. A. Zucca, J. H. Duyn, R. R. Crichton, and L. Zecca. The role of iron in brain ageing and neurodegenerative disorders. *The Lancet Neurology*, 13(10):1045–1060, 2014.

[21] S. Levi, M. Ripamonti, A. S. Moro, and A. Cozzi. Iron imbalance in neurodegeneration. *Molecular Psychiatry*, 29(4):1139–1152, 2024.

[22] J. L. Beard, J. R. Connor, and B. C. Jones. Iron in the brain. *Nutrition Reviews*, 51(6):157–170, 2009.

[23] Katrin Hänsel, S. N. Dudgeon, K.-H. Cheung, Thomas, and W. L. Schulz. From data to wisdom: Biomedical knowledge graphs for real-world data insights. *Journal of Medical Systems*, 47(1):1, 2023.

[24] Caohongliu. The advantages of using the knowledge graph part i, 2021.

[25] Jesús Barrasa. What is a knowledge graph?, 2023.

[26] Núria Queralt-Rosinach, Gregory S. Stupp, Tong Shu Li, Michael Mayers, Maureen E. Hoatlin, Matthew Might, and Andrew I. Su. Structured reviews for data and knowledge-driven research. *Database*, 2020.

[27] Monarch Initiative. Monarch initiative, 2024.

[28] Neo4j. What is a graph database? - getting started, 2024.

[29] GeeksforGeeks. Neo4j query cypher language, 2019.

[30] NVIDIA Glossary. What are large language models?, 2017.

[31] Muhammad Usman Hadi, Qasem Al-Tashi, R. Qureshi, and Seyedali Mirjalili. Large language models: A comprehensive survey of its applications, challenges, limitations, and future prospects, July 10 2023.

[32] Bernardino Romera-Paredes, Mohammadamin Barekatain, A. Novikov, M. Balog, M. Pawan Kumar, E. Dupont, and A. Fawzi. Mathematical discoveries from program search with large language models. *Nature*, 625(7995):468–475, 2023.

[33] Turing.com. Fine-tuning llms: Overview, methods & best practices, 2024.

[34] Lin Yao, Chengsheng Mao, and Yuan Luo. KG-BERT: Bert for knowledge graph completion, 2019.

[35] Sachin Mathur and Deendayal Dinakarpandian. Finding disease similarity based on implicit semantic similarity. *Journal of Biomedical Informatics*, 45:363–371, 2012.

[36] Liang Cheng, Hengqiang Zhao, Pingping Wang, Wenyang Zhou, Meng Luo, Tianxin Li, Junwei Han, Shulin Liu, and Qinghua Jiang. Computational methods for identifying similar diseases. *Molecular therapy. Nucleic acids*, 18:590–604, 2019.

[37] Shuhui Su, Lei Zhang, and Jian Liu. An effective method to measure disease similarity using gene and phenotype associations. *Frontiers in Genetics*, 10:466, 2019. Published: 21 May 2019.

[38] H. Hassan, Z. Ren, C. Zhou, M. A. Khan, Y. Pan, J. Zhao, and B. Huang. Supervised and weakly supervised deep learning models for covid-19 ct diagnosis: A

systematic review. *Computer Methods and Programs in Biomedicine*, 218:106731–106731, 2022.

[39] H. E. Kim, A. Cosa-Linan, N. Santhanam, M. Jannesari, M. E. Maros, and T. Ganslandt. Transfer learning for medical image classification: a literature review. *BMC Medical Imaging*, 22(1):1–14, 2022.

[40] R. Zhang, D. Hristovski, D. Schutte, A. Kastrin, M. Fiszman, and H. Kilicoglu. Drug repurposing for covid-19 via knowledge graph completion. *Journal of Biomedical Informatics*, 115:103696, 2021.

[41] Paperswithcode.com. Papers with code - knowledge graph completion, 2022.

[42] Amine Dadoun. Knowledge graph embeddings 101, 2023.

[43] X. Ge, Y.-C. Wang, B. Wang, and Jay. Knowledge graph embedding: An overview, 2023.

[44] Hanwen Zha, Zhiyu Chen, and Xifeng Yan. Inductive relation prediction by bert. In *Proceedings of the Thirty-Sixth AAAI Conference on Artificial Intelligence (AAAI-22)*, University of California, Santa Barbara, 2022. AAAI.

[45] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding, 2018.

[46] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 2019.

[47] Gsea — msigdb. https://www.gsea-msigdb.org/gsea/msigdb/, 2024.

[48] Biolink Model Documentation. Biolink model documentation, 2022.

[49] Nature. Transcription factor, 2014.

[50] S. Robertson. Role of transcription factors, 2018.

[51] Project Jupyter. Project jupyter, 2024.

[52] Helmholtz-muenchen.de. The negatome database 2.0., 2014.

[53] Ryutaro Ichise Tiroshan Madushanka. Negative sampling in knowledge graph representation learning: A review. 2024.

[54] scikit-learn. train_test_split, 2024.

[55] dmis-lab. dmis-lab/biobert-v1.1 · hugging face. https://huggingface.co/dmis-lab/biobert-v1.1, 2024.

[56] K. Team. Keras documentation: Adamw, 2019.

[57] Vincent Jung and Lonneke van der Plas. Understanding the effects of language-specific class imbalance in multilingual fine-tuning. *Idiap Publications*, 2024.

[58] GeeksforGeeks. What is cross-entropy loss function?, 2024.

[59] X. Jiang and C. Xu. Deep learning and machine learning with grid search to predict later occurrence of breast cancer metastasis using clinical data. *Journal of Clinical Medicine*, 11(19):5772, 2022.

[60] PyTorch. Softmax — pytorch 2.4 documentation, 2023.

[61] Snellius - surf user knowledge base. https://servicedesk.surf.nl/wiki/display/WIKI/Snellius, 2023.

[62] STRING protein-protein interaction networks. String protein-protein interaction networks. https://string-db.org/cgi/input?sessionId=bne8Tq7zQBLf&input_page_show_search=on, 2015.

[63] NetworkX. degree_centrality — networkx 3.3 documentation, 2024.

[64] G. Database. Trim67 gene - genecards — tri67 protein — tri67 antibody. https://www.genecards.org/cgi-bin/carddisp.pl?gene=TRIM67&keywords=trim67, 2024.

[65] NetworkX. closeness_centrality — networkx 3.3 documentation, 2024.

[66] Jing Zhang and Xiaoyuan Chen. p53 tumor suppressor and iron homeostasis. *FEBS Journal*, 286(4):620–629, 2018.

[67] J. S. Steffan, A. Kazantsev, O. Spasic-Boskovic, M. Greenwald, Y.-Z. Zhu, H. Gohler, and L. M. Thompson. The huntington's disease protein interacts with p53 and creb-binding protein and represses transcription. *Proceedings of the National Academy of Sciences*, 97(12):6763–6768, 2000.

[68] R. H. Mansky, E. A. Greguske, D. Yu, N. Zarate, T. A. Intihar, W. Tsai, and R. Gomez-Pastor. Tumor suppressor p53 regulates heat shock factor 1 protein degradation in huntington's disease. *Cell Reports*, 42(3):112198, 2023.

[69] GeneCards. Ftl gene - genecards — fril protein — fril antibody, 2024.

[70] P. F. Chinnery. Neuroferritinopathy, 2022.

[71] B. Suter, J.-F. Fontaine, R. Yildirimman, T. Raskó, M. H. Schaefer, A. Rasche, and E. E. Wanker. Development and application of a dna microarray-based yeast two-hybrid system. *Nucleic Acids Research*, 41(3):1496–1507, 2012.

[72] A. Li, Y. Li, X. Li, C. Tang, Y. Yang, N. Li, and Y. Jin. Ferritin light chain as a potential biomarker for the prognosis of liver hepatocellular carcinoma. *Heliyon*, 10(16):e36040–e36040, 2024.

[73] Ji Min Park, Chen Zou Mau, Yang Ching Chen, Yen Hao Su, Hsin An Chen, Shih Yi Huang, ..., and Ching Feng Chiu. A case–control study in taiwanese cohort and meta-analysis of serum ferritin in pancreatic cancer. *Scientific Reports*, 11(1), 2021.

[74] Chao Dai, Xiaoyu Chen, Jie Li, Patrick Comish, Rui Kang, and Daolin Tang. Transcription factors in ferroptotic cell death. *Cancer Gene Therapy*, 27(9):645–656, 2020.

[75] WikiPathways. Ferroptosis. https://www.wikipathways.org/pathways/WP4313.html, July 28 2024.

[76] Q. Han, L. Sun, and K. Xiang. Research progress of ferroptosis in alzheimer disease: A review. *Medicine*, 102(36):e35142–e35142, 2023.

[77] B. Kenkhuis, A. Somarakis, L. de Haan, O. Dzyubachyk, M. E. IJsselsteijn, Noel, and Louise. Iron loading is a prominent feature of activated microglia in alzheimer's disease patients. *Acta Neuropathologica Communications*, 9(1):185, 2021.

[78] J. S. Clark, R. Kayed, G. Abate, D. Uberti, P. Kinnon, and S. Piccirella. Post-translational modifications of the p53 protein and the impact in alzheimer's disease: A review of the literature. *Frontiers in Aging Neuroscience*, 14:835288, 2022.

[79] P. Wolfrum, A. Fietz, S. Schnichels, and J. Hurst. The function of p53 and its role in alzheimer's and parkinson's disease compared to age-related macular degeneration. *Frontiers in Neuroscience*, 16:1029473, 2022.

[80] G. Database. BMP6 Gene - GeneCards — BMP6 Protein — BMP6 Antibody. https://www.genecards.org/cgi-bin/carddisp.pl?gene=BMP6&keywords=bmp6, 2024.

[81] R. Kashima and A. Hata. The role of tgf- superfamily signaling in neurological disorders. *Acta Biochimica et Biophysica Sinica*, 50(1):106–120, 2018.

[82] Yulia Akbergenova and J. Troy Littleton. Pathogenic huntington alters bmp signaling and synaptic growth through local disruptions of endosomal compartments. *Journal of Neuroscience*, 37(12):3425–3439, 2017.

[83] C. Soldati, P. Caramanica, M. J. Burney, C. Toselli, A. Bithell, G. Augusti-Tocco, and E. Cacci. Re1 silencing transcription factor/neuron-restrictive silencing factor regulates expansion of adult mouse subventricular zone-derived neural stem/progenitor cells in vitro. *Journal of Neuroscience Research*, 93(8):1203–1214, 2015.

[84] G. Database. Nog gene - genecards — nogg protein — nogg antibody. https://www.genecards.org/cgi-bin/carddisp.pl?gene=NOG&keywords=nog, 2024.

[85] G. Database. Smad9 gene - genecards — smad9 protein — smad9 antibody. https://www.genecards.org/cgi-bin/carddisp.pl?gene=SMAD9&keywords=smad9, 2024.

[86] G. Sanchez-Duffhues, E. Williams, Marie-Jose Goumans, Carl-Henrik Heldin, and Peter ten Dijke. Bone morphogenetic protein receptors: Structure, function and targeting by selective small molecule kinase inhibitors. *Bone*, 138:115472, 2020.

[87] G. Database. HFE Gene - GeneCards — HFE Protein — HFE Antibody. https://www.genecards.org/cgi-bin/carddisp.pl?gene=HFE&keywords=hfe, 2024.

[88] Human Phenotype Ontology. Memory impairment (hp:0002354). https://hpo.jax.org/app/browse/term/HP:0002354, 2024.

[89] A. Montoya, M. Pelletier, M. Menear, E. Duplessis, F. Richer, and M. Lepage. Episodic memory impairment in huntington's disease: A meta-analysis. *Neuropsychologia*, 44(10):1984–1994, 2006.

[90] G. Lowndes and G. Savage. Early detection of memory impairment in alzheimer's disease: A neurocognitive perspective on assessment. *Neuropsychology Review*, 17(3):193–202, 2007.

[91] M. Noemia, M. Polydoro, D. C. Laranja, F. Bonatto, E. Bromberg, F. Moreira, and N. Schröder. Recognition memory impairment and brain oxidative stress induced by postnatal iron administration. *European Journal of Neuroscience*, 21(9):2521–2528, 2005.

[92] GeneCards Database. Bmpr1a gene - genecards — bmr1a protein — bmr1a antibody. https://www.genecards.org/cgi-bin/carddisp.pl?gene=BMPR1A&keywords=bmpr1a, 2024.

[93] X. Zhang, J. Li, L. Ma, H. Xu, Y. Cao, W. Liang, and Y. Li. Bmp4 overexpression induces the upregulation of app/tau and memory deficits in alzheimer's disease. *Cell Death Discovery*, 7(1), 2021.

[94] Z. Chen, H. Mao, H. Li, W. Jin, H. Wen, X. Wei, and J. Tang. Exploring the potential of large language models (llms) in learning on graphs. *SIGKDD Explorations*, 25(2):42–61, 2024.

[95] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, et al. The fair guiding principles for scientific data management and stewardship. *Scientific Data*, 3(1):160018, 2016.

[96] Temporal knowledge graph reasoning based on evolutional representation learning. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, New York, NY, USA, 2021. ACM.

Appendix

Subject	Object	Probability score
BMP6	Disinhibition	0.866613

Table .1

Subject	Object	Probability score
GSTM5	Decreased body mass index	0.872353
OGFOD1	Decreased body mass index	0.872353
NDUFS3	Decreased body mass index	0.872353
GSTM1	Decreased body mass index	0.872353
MAP3K4	Decreased body mass index	0.872353
PLOD3	Decreased body mass index	0.872353
NUBP1	Decreased body mass index	0.872353
MAP2K3	Decreased body mass index	0.872353
SLC22A4	Decreased body mass index	0.860373
CD274	Decreased body mass index	0.872353
LYRM4	Decreased body mass index	0.872353
UQCRFS1	Decreased body mass index	0.860373
LCN2	Decreased body mass index	0.872353
PLOD2	Decreased body mass index	0.872353
SLC39A4	Decreased body mass index	0.860373
SLC22A2	Decreased body mass index	0.860373
NUBPL	Decreased body mass index	0.872353
SLC22A3	Decreased body mass index	0.860373
FDX1	Decreased body mass index	0.872353
SLC46A1	Decreased body mass index	0.860373

Table .2

Subject	Object	Probability score
SLC46A1	Memory impairment	0.861569
DDR2	Memory impairment	0.872740
UBE2D1	Memory impairment	0.855968
ALOX15B	Memory impairment	0.863384
OGFOD1	Memory impairment	0.863384
MAP2K3	Memory impairment	0.863384
TFRC	Memory impairment	0.872740
GADD45G	Memory impairment	0.855968
BOLA2	Memory impairment	0.872740
LCN2	Memory impairment	0.863384
SLC22A2	Memory impairment	0.861569
PLOD2	Memory impairment	0.863384
CD274	Memory impairment	0.863384
BMP6	Memory impairment	0.872740
GSTM1	Memory impairment	0.863384
ARVCF	Memory impairment	0.855968
EDDM3A	Memory impairment	0.855968
PLOD3	Memory impairment	0.863384
MAF1	Memory impairment	0.872740
HEPH	Memory impairment	0.876864
ISCU	Memory impairment	0.876864
TKT	Memory impairment	0.876864
FTH1	Memory impairment	0.872740
NDUFS3	Memory impairment	0.863384
GSTM5	Memory impairment	0.863384
FDX1	Memory impairment	0.863384
PCSK7	Memory impairment	0.863384
CISD2	Memory impairment	0.863384
SLC22A4	Memory impairment	0.861569
NUBPL	Memory impairment	0.863384
SLC22A3	Memory impairment	0.861569
NFU1	Memory impairment	0.872740
MAP3K4	Memory impairment	0.863384
CDO1	Memory impairment	0.872740
AIFM3	Memory impairment	0.863384
CIAO2B	Memory impairment	0.855968
FTO	Memory impairment	0.876864
ACO1	Memory impairment	0.872740
ABCB7	Memory impairment	0.855968
SLC39A4	Memory impairment	0.861569
OSR1	Memory impairment	0.863384
UQCRFS1	Memory impairment	0.861569
ISCA2	Memory impairment	0.872740
NUBP1	Memory impairment	0.863384
SDHB	Memory impairment	0.872740
LYRM4	Memory impairment	0.863384
HSCB	Memory impairment	0.863384

Table .3

Subject	Object	Probability score
COASY	TP53	0.859514
ISCU	MTOR	0.873008
LTF	TP53	0.859514
ISCA2	TP53	0.859514
HJV	TP53	0.859514
ACO1	TP53	0.859514
HP	MTOR	0.895397
CH25H	TP53	0.859514
DDR2	TP53	0.859514
HP	TP53	0.908900
FXN	TP53	0.859514
CERS3	TP53	0.859514
BOLA2	TP53	0.859514
ISCU	TP53	0.855486
HEPH	TP53	0.855486
FTO	TP53	0.855486
BMP6	TP53	0.859514
BMAL2	TP53	0.859514
HP	GAPDH	0.863198
CDO1	TP53	0.859514
FTL	TP53	0.855486
HEPH	MTOR	0.873008
RPE65	TP53	0.859514
HFE	TP53	0.855486
HAMP	TP53	0.855486
BOLA3	TP53	0.859514
NFU1	TP53	0.859514
MMS19	TP53	0.859514
MAF1	TP53	0.859514
TF	TP53	0.855486
SDHB	TP53	0.859514

Table .4: Gene interaction predictions. Marked in yellow the predictions found also in STRING. In orange the prediction explored in this project.

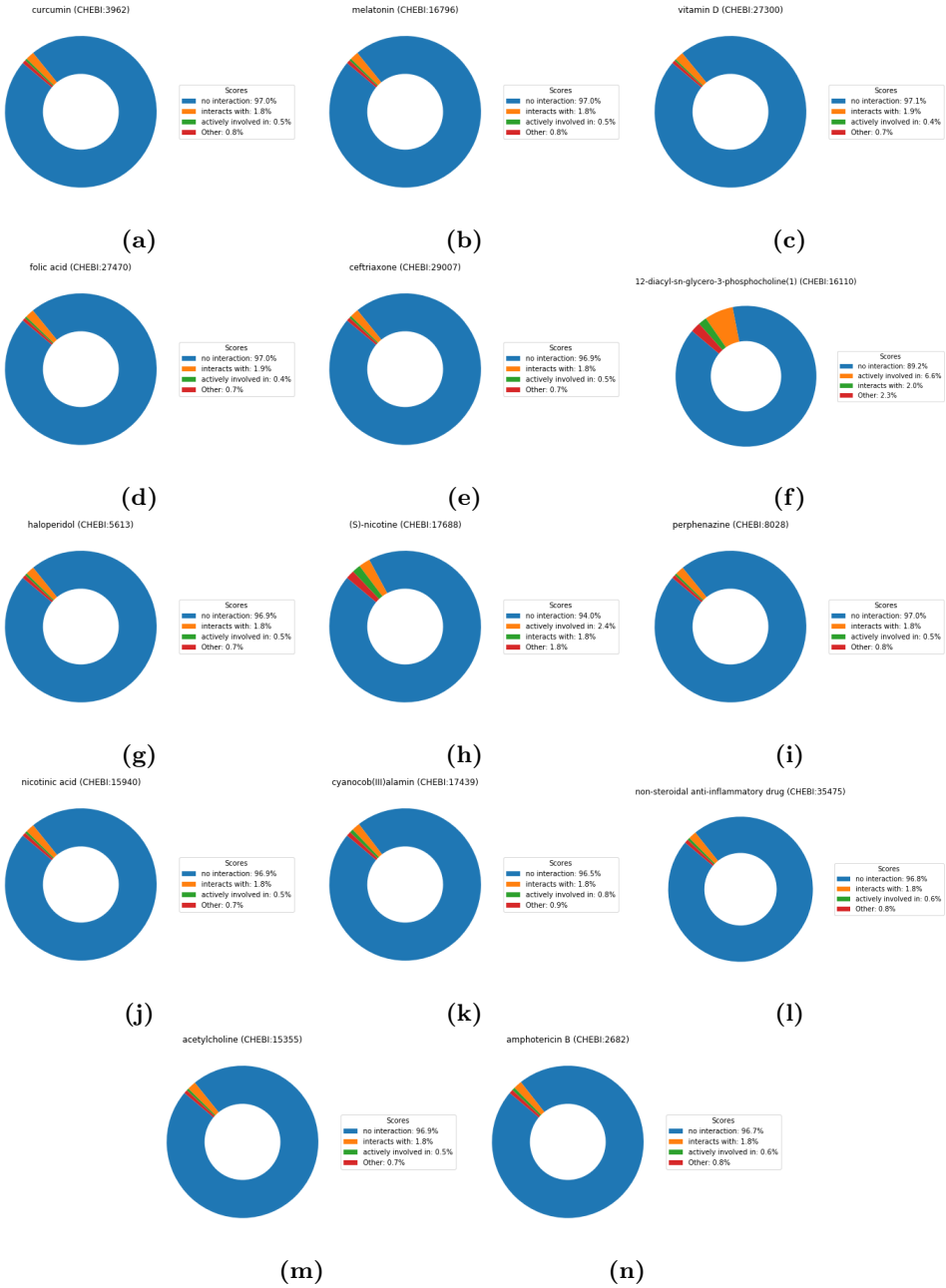


Figure .1: Results of drug-HD predictions. The pie charts depict the distribution of the probability score for each drug. In all cases *no interaction* is the most significant relationship predicted.