Universiteit Leiden

Master Computer Science

Structured review on Huntington's disease iron hypothesis

Name:	Karolis Cremers
Student ID:	s2682095
Date:	05/07/2022
Specialisation:	Bioinformatics
1st supervisors:	Núria Queralt-Rosinach and Eleni Mina
2nd supervisor:	Katy Wolstencroft

Master's Thesis in Computer Science

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

1 Abstract

Structured Reviews (SRs) organize and semantically represent the current knowledge around a research hypothesis in a structured manner, enabling semantic querying and data mining [1]. In this work we present the application of a SR to explore the relationship of iron with Huntington's Disease (HD). HD (OMIM:143100) is a heritable rare neurodegenerative disease caused by an elongated CAG repeat within the huntingtin (HTT, HGNC:4851) gene. The exact mechanisms that lead to disease pathogenesis remain unclear, however one of the current hypotheses implicates the accumulation of iron in HD brain. Abnormal accumulation of iron in the brain has been associated with several other neurodegenerative diseases. Therefore, current therapies often include iron chelators to combat iron build up. Our SR is a knowledge graph that includes information surrounding the iron hypothesis in HD. We constructed a HD knowledge graph that integrates genes, anatomy, genotypes, variants, physiology and disorders as concepts and the relationships between these concepts. In order to encourage the use of the SR throughout the research cycle it is hosted in parallel on a Neo4j instance and a Wikibase server, where Wikipedia style pages represent node information in the KG. This allows users to review, edit and update the graph with knowledge involving the hypothesis. We use the Graph Data Science Library (GDSL) and LibKGE to apply relationship prediction algorithms to provide potential research hypotheses. In addition to the relationship predictions, we improve the semantic richness of the SR by using NEOsemantics to integrate the Gene Ontology and measure the effects of this added information on our predictions.

Contents

2 Introduction

2.1 Huntington's Disease and iron

HD (OMIM:143100) is a hereditary neurodegenerative disease caused by an enlarged CAG repeat within the huntingtin (HTT, HGNC:4851) gene[2]. The size of the CAG repeat, which translates to the PolyQ region of the huntingtin protein, correlates with the age of onset and progression of the disease. CAG repeats smaller than 36 are considered to be non-symptomatic, repeats larger than 39 consistently cause symptoms and early onset (including juvenile HD) occurring with repeats larger than 70 [3, 4]. The mutant HTT protein (mHTT) has a wide range of effects within the cell, as shown in Figure 1. This includes transcriptional deregulation and mitochondrial toxicity. Focusing on potential iron-related processes in brain cells that are disrupted by mHTT, Muller et al. provide a good overview as seen in figure 2. Muller et al. [5] discuss interactions that mHTT has within the medium spiny neurons: with mHTT causing increased calcium levels, which promotes the activation of Rhes, which binds to mHTT (and HTT) together with Dexras1 and ACBD3 to activate DMT1, divalent metal transporter 1, which transports iron into the neuron. Furthermore, mHTT interferes with mitochondrial ion transport, leading to increased production of Reactive Oxygen Species (ROS) molecules. ROS molecules interact with Fe^3 to form Fe^2 and various radicals. These interactions are enhanced by mHTT causing an increased influx of Fe^2 through calcium activated nNOS proteins which are part of the DMT1 activation complex. This increase in radicals and iron ions causes the lipid layers that hold the cell together to peroxidise, opening up holes in the cell [5]. In addition to cell membrane damage, ROS molecules and radicals are a major source of DNA damage, further reducing cell survival [6].

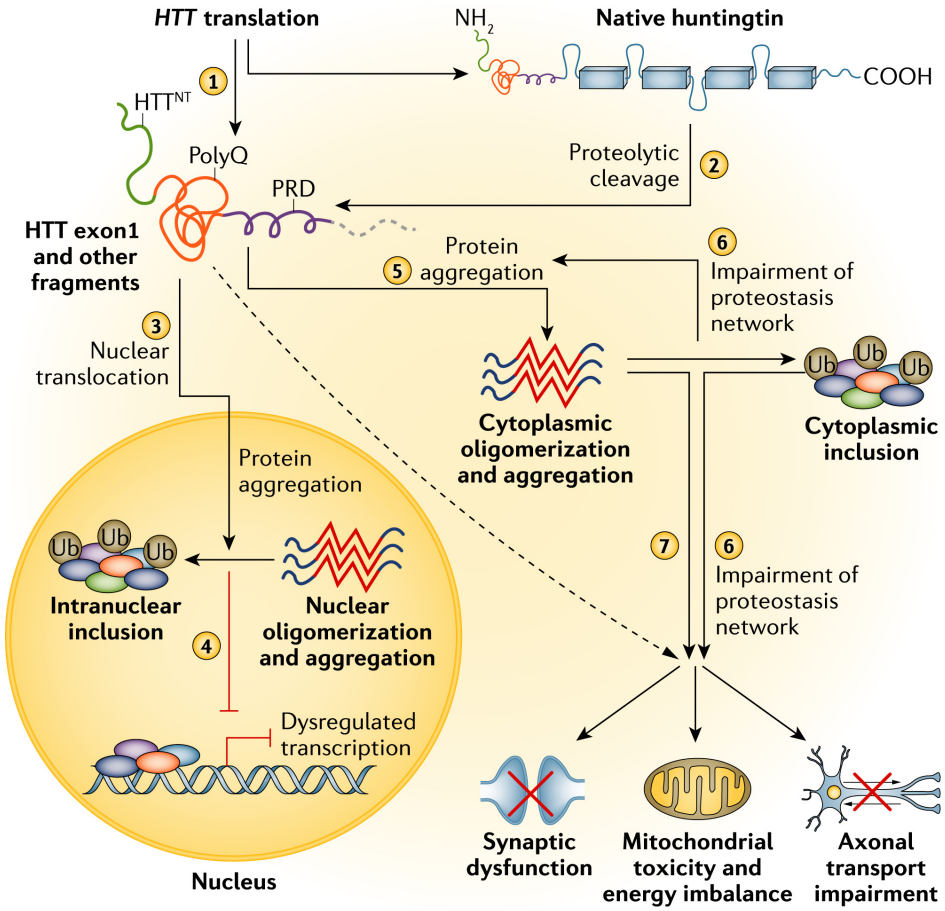

Figure 1: The role of mutant *HTT* in Huntington's Disease. Figure source: Mccolgan, et al. [7]

2.2 Structured Review

When exploring a possible hypotheses, a literature search has to be done. To help researchers with this task other researchers produce review articles summarize the most up-to-date knowledge on a certain topic. As with any

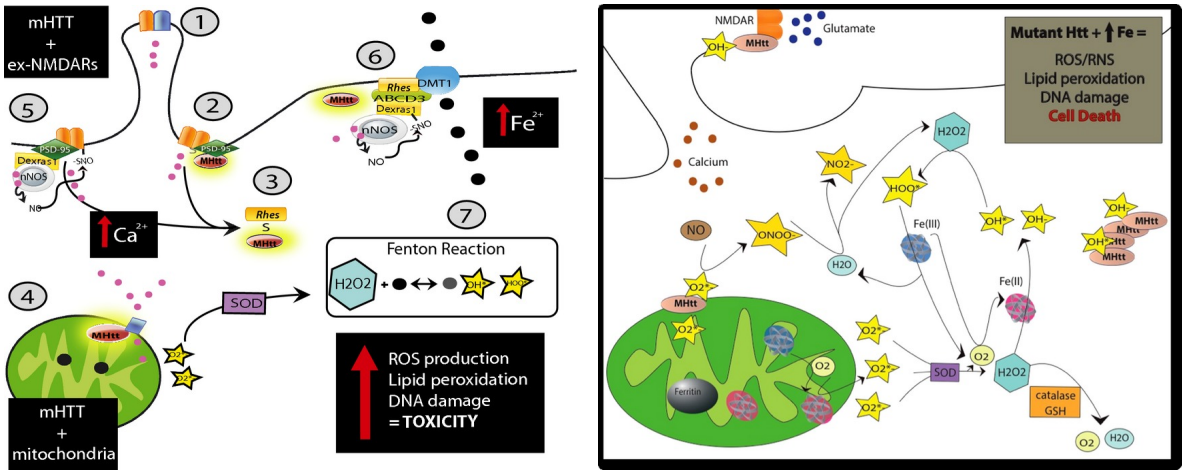

Figure 2: The role of mutant *HTT* in Huntington's Disease in association with iron. With cell toxicity on the figure left and cell death on the right. Figure source: Muller, et al. [5]

labour intensive tasks, automation of the literature search has been an active field of development.

2.2.1 Automated knowledge capturing

There are two tactics available to collect relevant information in an automated way: Directly capturing information using text mining or obtaining relevant information from available databases. Papers such as KnIT [8] and "A Text Mining Model for Hypothesis Generation" [9] apply machine learning on full texts to summarize knowledge contained within. These methods allow non structured knowledge to be included. Whereas tools like BionKnowledge-Reviewer (BKR) [1], BioGraph [10], and CROssBAR [11] use curated database information to capture relevant curated information. Both methods store their information in a machine-readable format in the form of a knowledge graph. The Knowledge Graph (KG) is a collection of interconnected facts that are represented as a graph network.[12] Nodes represent concepts and edges their relations. In contrast to just listing related facts to a certain subject, a knowledge graph also shows how these facts are interconnected [12]. A powerful aspect of the KG is that any similarly structured information such as databases already in a KG format can be incorporated into the original graph. This allows for the use of multiple data sources to create one network containing information from a plethora of fields of biological sciences. When querying this graph on the relationship between two concepts results may include one or more of these different data sources automatically connecting the different fields and provide novel information to the user. Allowing for a more complete representation of current knowledge. A KG on a specific topic is called a Structured Review (SR)[1]. In this work we chose to use the BKR library and adapt it to our subject of interest. The BKR was chosen as it's original implementation was made to capture information surrounding NGLY1 deficiency and so has been proven to work in the context of rare diseases. We are also interested in its ability to collect data from databases and from differential gene expression experiment results, as these results can provide a unique perspective on potential genes of interest within HD affected tissues.

2.3 Project goal definition

The aim of this work is to expand upon the BKR to produce a SR and to explore the role of iron in Huntington's disease: We will implement the Wikibase interface part of the BioKnowledge-Reviewer workflow paper, extend the BKR library by building a module for the integration of Gene Ontology information into the Structured Review (SR), and apply link prediction algorithms to the SR to exploit it's machine-readable structure. Using these improvements we want to explore high potential research directions and hypothesis surrounding iron in HD within our SR.

3 Original BioKnowledge-Reviewer implementation

The BKR library structure is depicted in Figure 3. The BKR uses four data sources which are either already in a graph based format or are converted into one [1]. First of which is the curated dataset. This data is a csv file

filled with facts manually curated by topic experts.

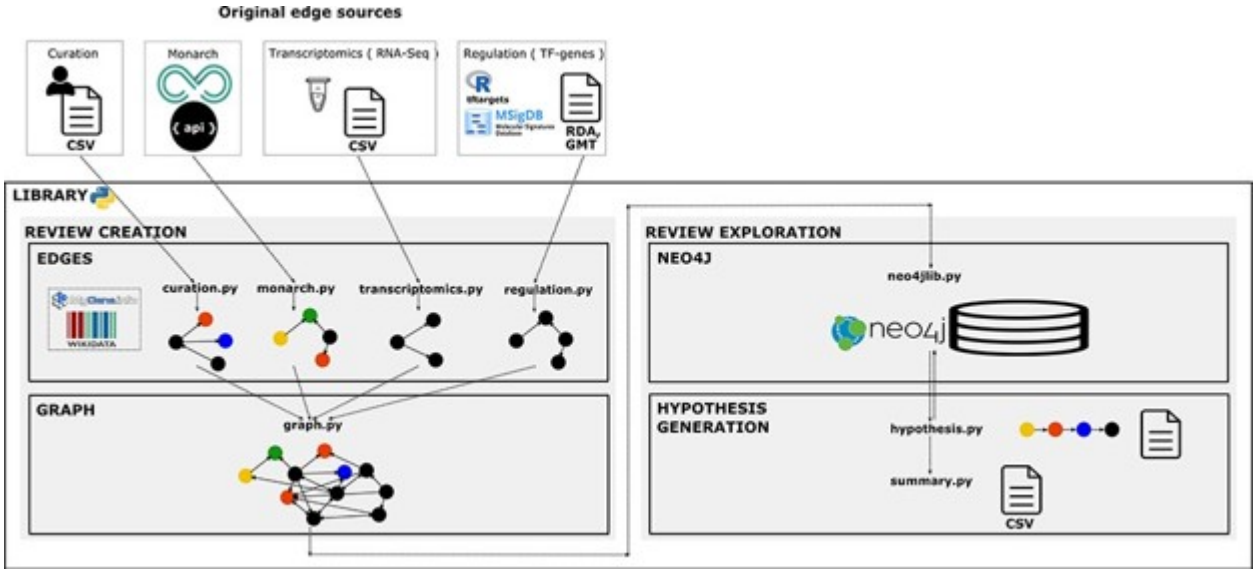

Figure 3: The original library structure of the BKR. Figure source: [1]

3.1 Monarch Initiative

The Monarch dataset used in a BKR run is a subset of the Monarch Initiative database obtained through a series of API calls. The Monarch Initiative is a database that integrates a large amount of cross-species genetic information into a central KG [13]. This includes phenotypes, gene variants, research organisms, diseases, and genes. The Monarch Initiative provides download options for use on its KG. The download options provided allow the extraction of direct and specific node type neighbors from the KG. This is used by the BKR to specifically include human gene orthologs in the KG. If a certain gene-gene interaction is found in a model organism, it may be that this interaction has not yet been found in humans, and the orthologous interaction can be used to justify a new hypothesis. As mentioned by Mungall et al. (The authors of the Monarch Initiative), including orthologs also improves the coverage of gene-phenotype associations [13].

3.2 Transcriptomics

Data from a differential gene expression (transcription) experiment can provide specific tissue related information involving the topic of interest in the SR. The original BKR paper used an RNA-seq experiment on Drosophila model of NGLY1 deficiency [1]. This was done by adding an "association" relationship between the gene of interest (NGLY1) of the original application of

3.3 Transcription factors

Transcription factors are proteins (gene products) that bind to DNA in a sequence specific manner and regulate transcription of genes based on this binding [14]. These interactions are specifically of interest to us, due to mHTT's effects on Calcium, Iron and ROS levels and their effects on transcription factor activation [15, 16, 17].

3.4 Curated information

The BKR also allows users to input their own curated subgraph into the SR. This is often done for information a domain expert already knows about but has not yet published, or if the information is not yet available in the data sources used by the BKR.

3.5 Graph merging

When all subgraphs are constructed they are merged based on their ID. This necessitates that the subgraphs are using the same ontology for concepts or reference ID's for genes.

3.6 Neo4J

Neo4j is a graph database product that includes a user interface that displays nodes and relationships based on Cypher queries. Cypher is a language that simplifies SQL queries by symbolic representation of nodes and relationships within the syntax. Cypher represents a node by closed brackets; (), while a relationship is represented by two dashes with closed square brackets: -[]-. Users can filter nodes and edges based on the content of their brackets such as node or edge attributes. In the following example, we query Neo4j on a shortest path containing any direct relationship between HD and HTT. Neo4j can filter any node or edge attribute that exists in the graph, the query uses the preferential label (display name) for HD and the HTT gene id as filters.

```
MATCH Path = (node1 {preflabel: \Huntington's Disease"})-[]-(node2 {id: HGNC:4851})
RETURN Path
```

The results of this query are shown in figure 4. A powerful aspect of using Neo4j is the fact that the query results displayed are interactive: One can select edges and nodes to display their attributes and can show neighboring nodes of already displayed nodes by selecting the expand button on a selected node. This allows users to visualize the context to the concept displayed within the shortest path in an iterative method.

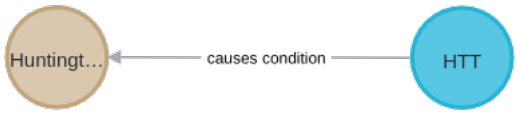

Figure 4: A Neo4j visualization of the relationship between Huntington's Disease and HTT

3.7 Hypothesis generation

The original BKR library contains two scripts that allow the user of the SR to use common queries without having to learn Cypher. The hypothesis script allows queries involving different filters for finding the shortest path between two concepts. While the summary script produces a tabular overview of the type of paths, e.g. what node types or relationships that are used to connect two concepts.

3.8 Wikibase

Wikibase is a software system used to run the Wikidata knowledge graph [18]. Wikibase has several interfaces, including Mediawiki, the software that runs the Wikipedia website [18]. Mediawiki is extended by the Wikibase repository to allow structured data storage and the Wikibase client to interface with the Wikibase repository extension. Wikibase has been used as an extension to the NGLY1 SR by the original authors of the library to improve the usability of their SR [1]. Specifically, the Wikibase runs a parallel version of the SR that can be manually edited by users without knowledge of the Neo4j Cypher language. The Mediawiki software allows the full functionality of Wikipedia: tracking page changes and users, and setting up bots with specific permissions for automated editing. The Neo4j graph is updated with new information stored in the Wikidata database whenever a synchronization between the two databases is initiated. The Wikibase also enforces the internal structure of the KG. To add a new node or relationship type to the SR, one first has to generate a separate page describing its properties. This process ensures that the SR has clear definitions and references on new information and information types.

4 Expansion and adaptions on the original BioKnowledge-Reviewer framework

Our work is inspired by several technologies associated with graph data science. First we improve upon the NGLY1 BKR implementation the following expansion and adaptations.

4.1 Neosemantics and Gene Ontology

We expand upon the Neo4j implementation by using two available plugins for Neo4j databases. First, we use Neosemantics to import the Gene Ontology into the existing SR. Second, we use the Graph Data Science library to perform a series of link predictions. The Gene Ontology (GO) is a structured and logical description of biological functions of genes. The GO is a project that provides structured, controlled vocabularies and classifications for genes, gene products and sequences. The GO uses a directed acyclic graph structure to describe individual gene products and consists of three name spaces: Biological Process (BP), Cellular Component (CC) and Molecular Function (MF). The GO is a directed acyclic graph structure where relationships between terms only direct to more generic concepts.
We decided to import the most recent GO [1] into the SR to improve its semantic richness, query ability, context to existing knowledge and the predictive power of link prediction algorithms. Due to the structure of the GO, one can make more generic searches on the SR, for example: instead of searching for genes related to each-other through the concept of "Golgi apparatus" we use the parent term "Cytoplasm" to capture genes that are expressed in it's surrounding area. We hypothesise that adding the GO will help improve the quality of predictions for our link prediction, as genes with similar BP, CC or MF parent terms have a higher chance to be related in some fashion.

4.2 Link prediction

One of the simplest ways of link prediction is based on common neighbor nodes between concepts node pairs. The likelihood of two nodes in a graph having a relationship is higher when they have a lot of overlapping neighbors [19].
We perform multiple common neighbor algorithms predictions on the Neo4j graph. First we have a direct measurement of common neighbors as seen in equation 1. This is one of the most basic measurements, but has been shown to be one of the most accurate [20].

$$Similarity(x, y) = N(x) \cap N(y) \tag{1}$$

Second, we score node pairs based on the Adamic Adar [19] algorithm, which weights common neighbor nodes by their own neighbor count. The algorithm gives a higher weight to common neighbors that are less connected other nodes besides evaluation pair. This is done by taking the inverse log of the connectivity of the common neighbor as seen in equation 2. Unique connections between concepts are often a justification for hypotheses. Applying this algorithm will hopefully provide users with recommendations that includes these unique connections.

$$Similarity(x, y) = \sum_{s \in N(x) \cap N(y)} \frac{1}{\log(N(s))} \tag{2}$$

Third, we apply the resource allocation algorithm [20], which includes also the connectivity of the common neighbors within the calculation as seen equation 3. As mentioned by Zhou, et al., their scoring system is highly similar to that of Adamic Adar [20]. However, in their evaluation of their algorithm they measured a higher accuracy in networks with a higher average connectivity. As such, we include it in our measurements for comparison and posterity of the BKR workflow.

$$Similarity(x, y) = \sum_{s \in N(x) \cap N(y)} \frac{1}{N(s)} \tag{3}$$

The last of the common neighbor algorithms we apply is preferential attachment [21], which applies the logic of scale-free networks that states that high connectivity nodes are more likely to receive another connection. Therefore the algorithm gives high scores the higher the connectivity of both nodes:

$$Similarity(x, y) = |N(x)| * |N(y)| \tag{4}$$

[1] http://purl.obolibrary.org/obo/go.owl

The recommendations given by this scoring algorithm are good starting points to get familiar with highly connected concepts within the SR and provides a good starting point for literature research on the topic of the SR.

4.2.1 Embeddings

Besides measuring the likelihood of two nodes based on their common neighbors, we can also compare their neighborhoods. This can be done by representing attributes of a concept, such as neighborhood, as a point in (multi)dimensional space. The most basic embedding is a list of a single value $[X]$. This value can be interpreted as the tip of a vector starting from the center of the space. A one-dimensional vector works best when concept attributes are correlated. For example, we could summarize surface area and total edge length with the radius of a circle. However, when embedding concepts more complex than a circle, a better representation is often obtained by using multiple dimensions: $[X, Y, etc.]$.

4.2.2 FastRP

FastRP [22] is a method to rapidly build node embeddings of a given graph. It relies on the Johnson-Lindenstrass lemma, which states that points in sufficiently high dimensional Euclidean space can be represented in lower dimension without significant error[23]. FastRP takes advantage of this by representing each node in a graph with a random vector. This vector is then multiplied by the vectors representing nodes in its neighborhood. This multiplication against the neighboring node vectors is weighted by a random walk where the chance of moving to a node further away from the node of interest is cut in half each step. This has the effect that the node vector in the embedding space will move closer to the nodes in their neighborhood, and the distances between the points can be used to represent similarity.

4.2.3 Cosine similarity

A simple technique for comparing embeddings is to use cosine similarity. This method measures the angle between two vectors (cosine). Since both vectors are in the same vector space, a small angle represents a high degree of similarity in the features of the nodes, while a large angle indicates that the features of the nodes are dissimilar.

4.2.4 RESCAL

In addition to the common neighbor algorithms, we applied the RESCAL embedding model. This model decomposes a relationship into two embeddings representing the subject and object, and a matrix representation of the interactions the relationship type has with node features. This model is bilinear, for given relationship (subject, predicate, object) it can predict scores based on the direction of the relationship ((A, is a, B) or (B, is a, A) and allows for directionality of the edges in predictions. When performing RESCAL training, the original graph is represented by a three dimensional matrix χ with cells χ_{ijk} where i and j are nodes in the network and k the relationship type. The cell has either a 1 for an existing relationship and a 0 for no relationship. This matrix representations of the graph is decomposed into three matrices per relationship: $\chi_k \approx A\mathrm{R}_k A.T$. Matrix A contains the embeddings of nodes, is universal across relationship types and is size nxr. R is a matrix with size of r x r containing relationship interactions related to the embedded neighborhood of the node pair. R is unique for every relationship type. Both A and R are initialized with random values.
The goal of training the model is to reach a point where the matrix multiplication of the decomposition results in a correct representation of the original network the model has been trained on. First, we separate our data into training examples, Training RESCAL is done by calculating the error between a given example of an existing relationship and the current output of the matrix multiplication. Based on the error we edit the values a tiny amount so that for this example the model will be more correct in the future. We use small steps per example to ensure that the model predictions become better in general. The size of these training steps (learning rate) is usually controlled by an optimization algorithm. These algorithms ensure that we do not miss a optimal state of the model during training.
After completing these steps for all examples in the training data, we use the test dataset to test the general accuracy of the model on relationships that exist, but it has no knowledge of. This test provides a more realistic accuracy score compared to the training set and is used as an indication of how well the model has learned from the data and alter the training step following the optimization algorithm. The score is also used to prevent overfitting: When a model is being trained on a specific dataset giving too many of the same examples may cause the model to learn patterns in that dataset instead of general patterns. A human example of this is someone learning to solve a puzzle using muscle memory and brute force instead of looking at the information available on

the puzzle and applying the appropriate steps to get to a solved state. Another example can be the calculation of $1 + 1 = 2$, where a model doesn't actually learn that both added numbers have the same value but solves any calculation involving the number 1 with the number 2, E.G: $1 + 6 = 2$. We prevent this behavior by applying early-stopping: when the accuracy score from the test set stagnates, we stop training the model.
Of interest is that RESCAL constructs decompositions per relationship type, and can have different accuracy for each relationship type. This can give insight into the data quality stored within the SR. For example, if a given relationship is rare within the SR, but is easy to predict by the model this may indicate that the context on this relationship is captured very well by the SR. Encountering the opposite; a relationship that has a high frequency but not predicted well, may indicate that the relationship may not be defined correctly and contains subgroups or may not behave consistently based on node context.

5 Materials and methods

Code and data are available on github[2].

5.1 Materials

Study data used to construct the SR was obtained from several sources based on the original version of the BKR. Data from differential gene expression and the Monarch Initiative KG have been adapted to suite the context of iron in HD.

5.1.1 Transcriptomoics data

To get a representation of differentially expressed genes in HD Brains we used results obtained from the Gene Expression Omnibus (GEO) GSE64810 produced by Adam Labadorf, Andrew G. Hoss, et al.[24]. This data was particularly of interest because they used tissue samples from the prefrontal cortex, a region that is affected in HD patients but experiences a lower amount of neuronal cell death compared to the most affected regions. This data is extremely valuable as it provides a proxy representation of HD patient brain transcriptomic status during early disease onset [24]. Data from this source is represented by a "association" relationship to the HD node within the SR. We used the differential gene expression results presented in the original publication and filtered with an adjusted p-value threshold of 0.05. This resulted in 3.209 relationships to the HD node.

5.1.2 Monarch Initiative querying

A subsection of the Monarch Initiative KG was obtained by a series of API calls. Starting out with the ID's for Huntingtons disease (MONDO:0007739), HTT (HGNC:182293), Rhes or RASD2 (HGNC:182293) and the iron uptake pathway (REACT:R-HSA-917937) as seed nodes we query the database for their neighboring concepts using the BKR settings. This obtain a sub-graph containing 13.394 nodes and 207.836 edges. We chose to include RASD2 as it has been detailed in [5] as part of protein complex that is involved with the homeostasis of the neuron.

5.1.3 Transcription factors

Transcription factor data has been sourced from the TFtargets R package [3] results of the original NGLY1 version of the BKR pipeline. This package combines data from TRED, ITFP, ENCODE, TRRUST, Neph et al. and Marbach et al. [25, 26, 27, 28, 29, 30]. As data sources used by TFtargets such as TRED, ITFP, Neph et al., Marbach et al. were unavailable for access either due to lack of funding or technical difficulties.

5.2 Methods

5.2.1 Wikibase

Modified versions of the Wikibase server and the Krusty import scripts are adapted from `https://github.com/wmde/wikibase-docker` and `https://github.com/SuLab/Krusty` to handle duplicate gene symbols created as artifacts from converting the transcriptomics module to allow mapping of gene ID's to HUGO Gene Nomenclature

[2] https://github.com/KarolisCremers/bioknowledge-reviewer
[3] https://github.com/slowkow/tftargets

Committee (HGNC) gene identifiers. This includes improved API error handling, allowing for a continous upload to Wikibase.

5.2.2 Neo4j

To apply both Neosemantics and the GDS libraries we updated the Neo4j implementation to version 5.1 which is compatible with both plugins. We adapted the neo4j implementation file at: neo4jlib.py which now automatically downloads necessary plugins and updates the Neo4j configuration file to install them.

5.2.3 Neosemantics

We used the Neosemantics plugin to load the Gene Ontology into Neo4J and merge nodes using a series of Cypher commands.

5.2.4 Common Neighbors algorithms

The common neighbors algorithms applied to the SR is based on the Graph Data Science (GDS) plugin of Neo4j. This plugin allows for the application of the common neighbors and FastRP algorithms discussed earlier. We constructed a script to implement the complete set of nearest neighbor algorithms using the Neo4j API this script can be found at: recommender.py.

5.2.5 Cosine similarity

Cosine similarity measurement of target nodes from the common neighbors script are performed using a separate script

5.2.6 RESCAL

RESCAL has been implemented using the LibKGE library [31] on both the original SR and the GO expanded SR. Scripts used to convert the original SR and the GO expanded SR to LibKGE standards are found in the RESCAL model data folder. This folder also includes the configuration files used to perform the RESCAL hyperparameter search on the original SR and training on the original SR and GO expaneded version.

5.2.7 Cypher guide

A Cypher guide for exploring the SR has been written, this guide is made interactive by hosting it on a separate server that the Neo4j server loads on startup. This uses the code from: `https://github.com/neo4j-contrib/neo4j-guides` as provided by Neo4j[4].

6 Results

Here we highlight the characteristics of the SR, examples of information of interest in the SR and the results of the link prediction experiments.

6.1 Structured Review

Figure 5 shows the SR's schema, a schema is a representation of what type of nodes exists in the graph and how these nodes are interconnected.
The node types in the SR are GENE; a gene, ANAT; an anatomical location of gene expression, PHYS; a physiology, cell biological function or location, DISO; Disorder or phenotype, GENO; Genotype, VARI; Gene variant, NA; placeholder node type for concepts which could not be identified. The SR without the Gene Ontology contains a total of $31,266$ nodes and $392,872$ edges. The total counts of nodes and relationship per type can be found in Appendix B
(Section 11.1)

[4] https://neo4j.com/developer/guide-create-neo4j-browser-guide/

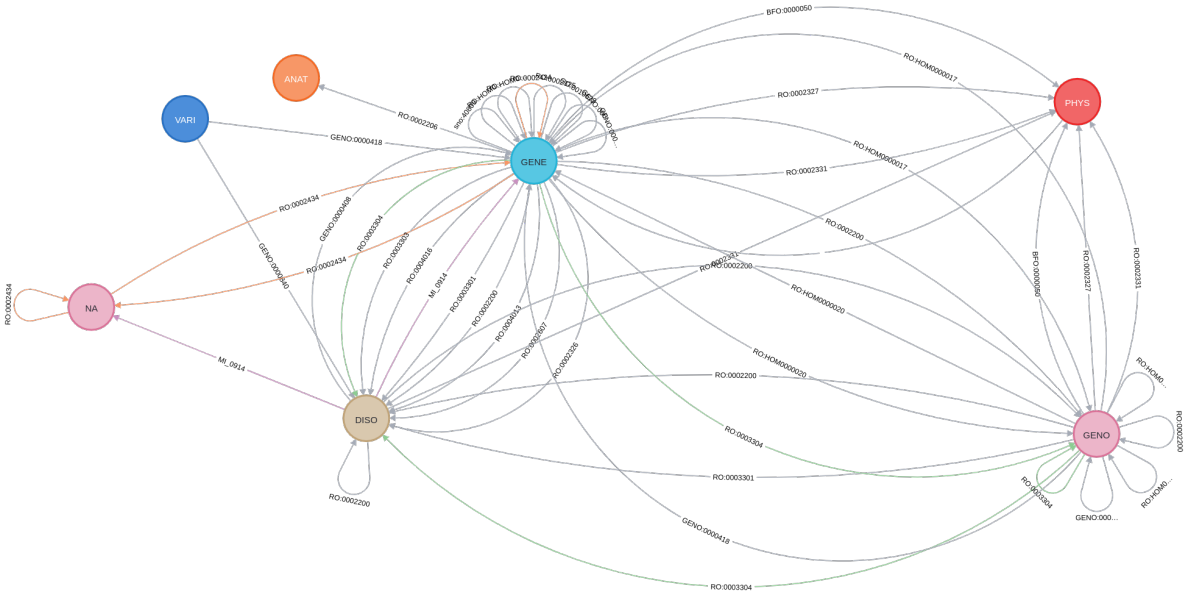

Figure 5: The schema of the SR: the node types within the SR and what relationship types connect which node types.

Table 1: Graph count and edge count in the original SR and when the Gene Ontology is added to the network

Attributes	Original	Gene Ontology
Node count	31,544	82,692
Edge count	408,288	495,090

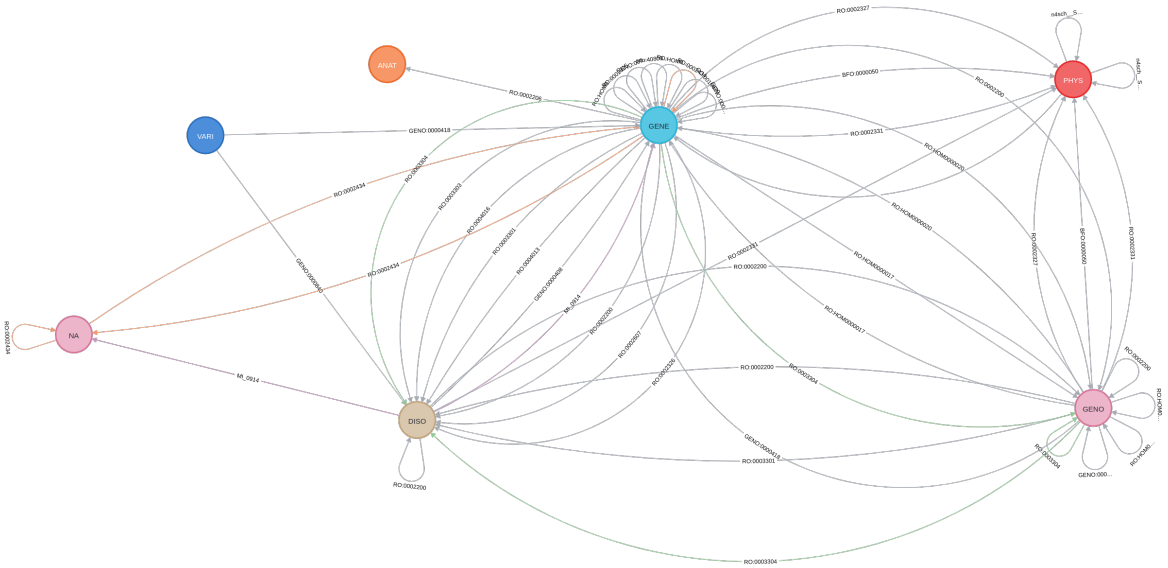

Figure 6: The schema of the SR when including the Gene Ontology: Two relationships between Physiology nodes are added to the general pattern of the SR (Top right).

Merging the SR with the Neosemantics's Gene Ontology increases the total amount of nodes of $51,148$ and edges of $86,802$ as seen in table 1. As seen in figure 6, the merger of the graphs added two relationship types between physiology nodes: "n4sch__SCO" and "n4sch__SCO_RESTRICTION". "n4sch__SCO" represents the "is a" relationship of ontology terms and "n4sch__SCO_RESTRICTION" contains all other relationships available in the GO (part of, negatively regulates, has part, etc.).

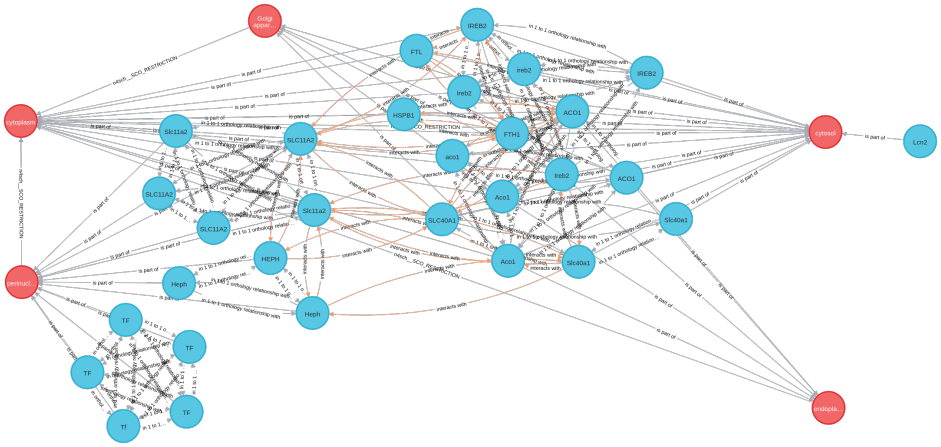

Figure 7: An example of added relationships through incorporation of the Gene Ontology via Neosemantics. "n4sch__SCO_restriction" relationships indicate that the Golgi apparatus and the perinuclear region of cytoplasm are part of the cytoplasm with certain restrictions.

6.2 Huntington and Iron

Here we show an example of exploring the relationships between HD and iron using the Neo4j interface as an average user would. To find genes of interest, we query Neo4j for neighbouring genes with "iron" in their description. As seen in figure 8 18 genes were found. These genes are related to HD through the "association" relationship, the identifier for differential expression from our experimental data.

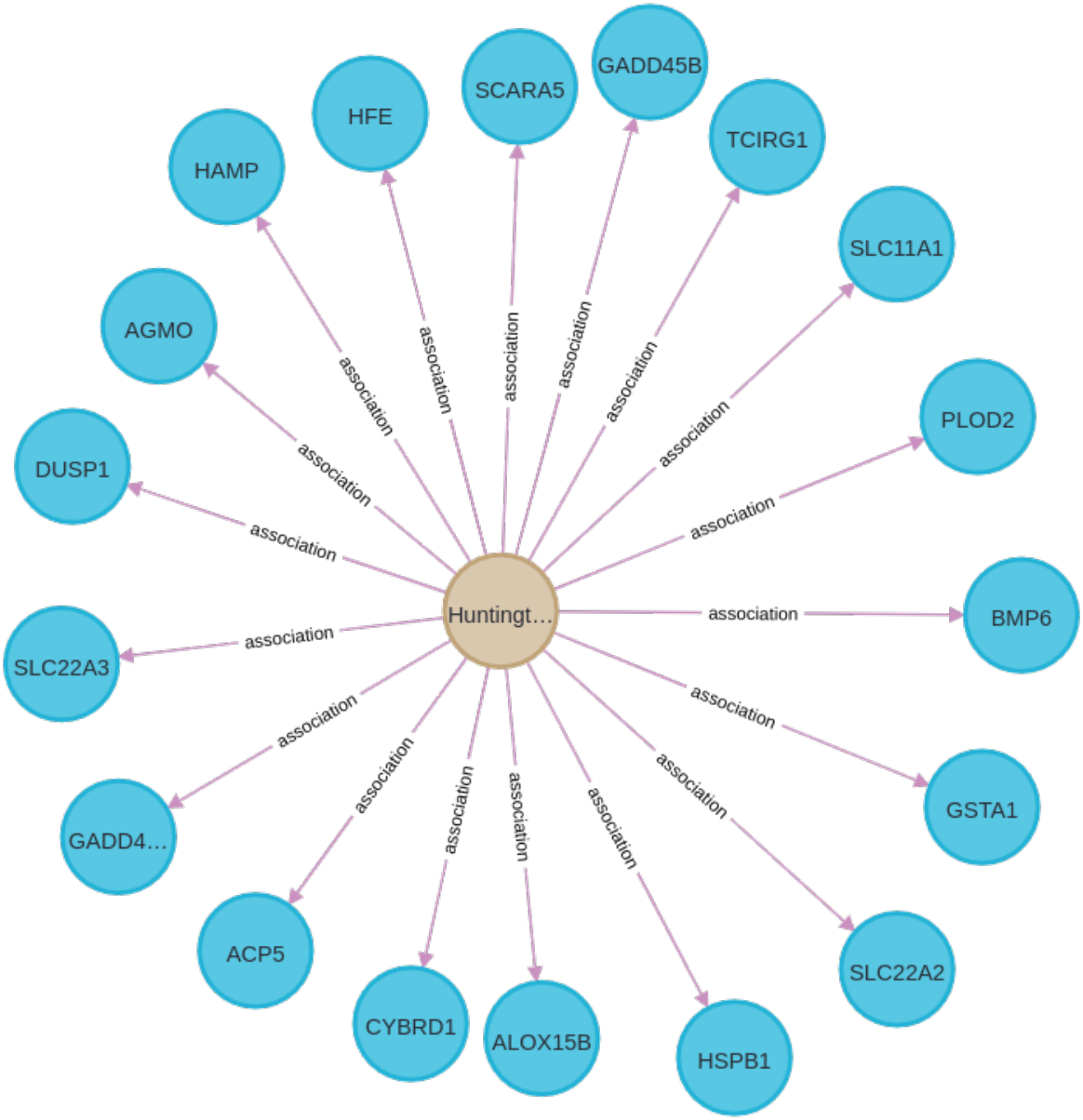

Figure 8: Results from a Neo4j query on HTT related genes with "iron" in their description. All genes displayed have been asocciated to iron through differential expression.

Based on the previous results we queried for anatomical information available on the genes of interest. As seen in figure 9, there are two genes of which the anatomical expression is known within the SR. HSPB1 (heat shock protein family B (small) member 1) is expressed in both the forebrain and the neocortex. CYBRD1 (Cytochrome B Reductase 1) is expressed within the eye, forebrain, retina, skeletal muscle, neocortex and "material anatomical entity". The "material anatomical entity" node represents the Uber-anatomy ontology definition UBERON:0000465, described as "Anatomical entity that has mass.".

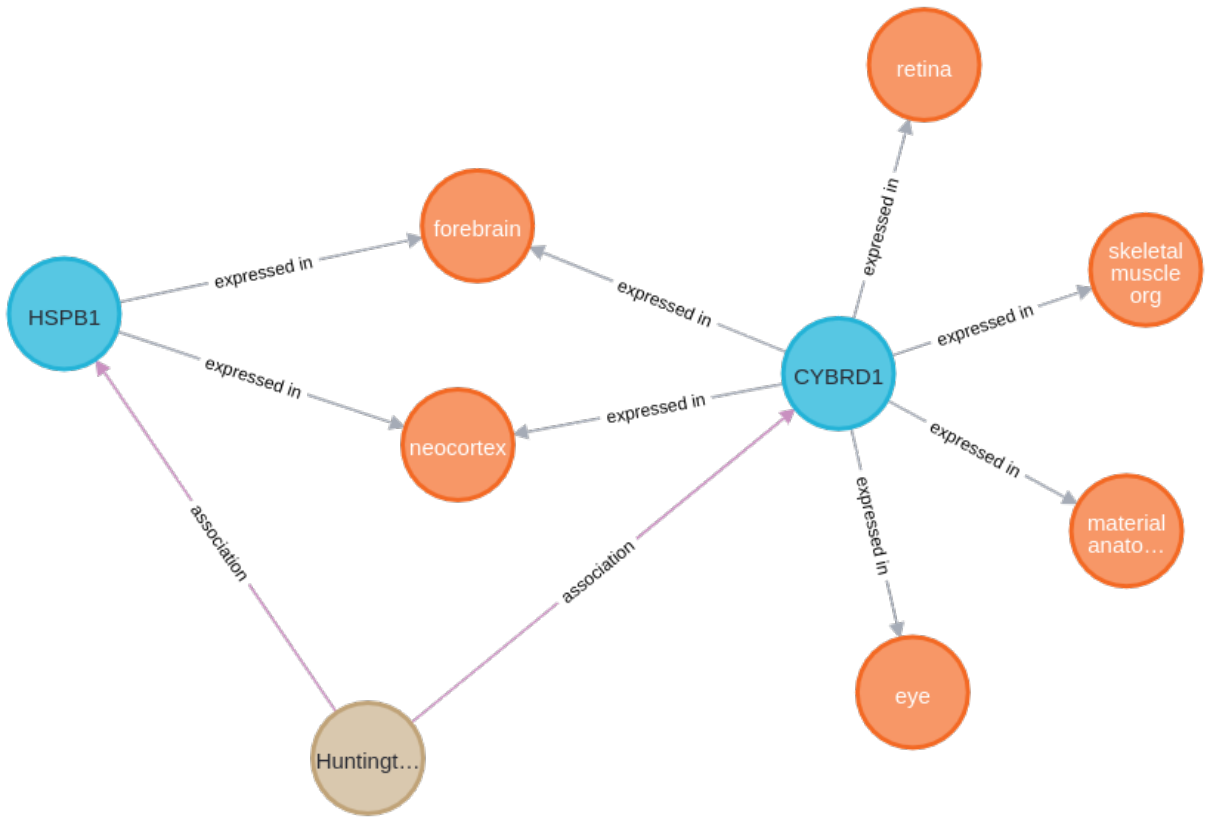

Figure 9: Results from a Neo4j query on anatomical expression of HD and genes with "iron" within their description.

Next we queried the SR on disorder type nodes with relationships to the genes captured in the first query. As seen in Figure 10, 11 genes have disorder nodes related to them. Disorders related to Iron include "Iron accumulation" associated to CP, "High platelet count" to MID1 and "Hemolytic anemia" to HMOX1. Disorder nodes related to the cardio-vascular system are seen to be related to MID1, HMOX1, BAG3 and EGFR. Other HD disorders such as cognitive impairment and movement dysfunction related nodes are associated with CP, RNASET2 and TUBA1A.

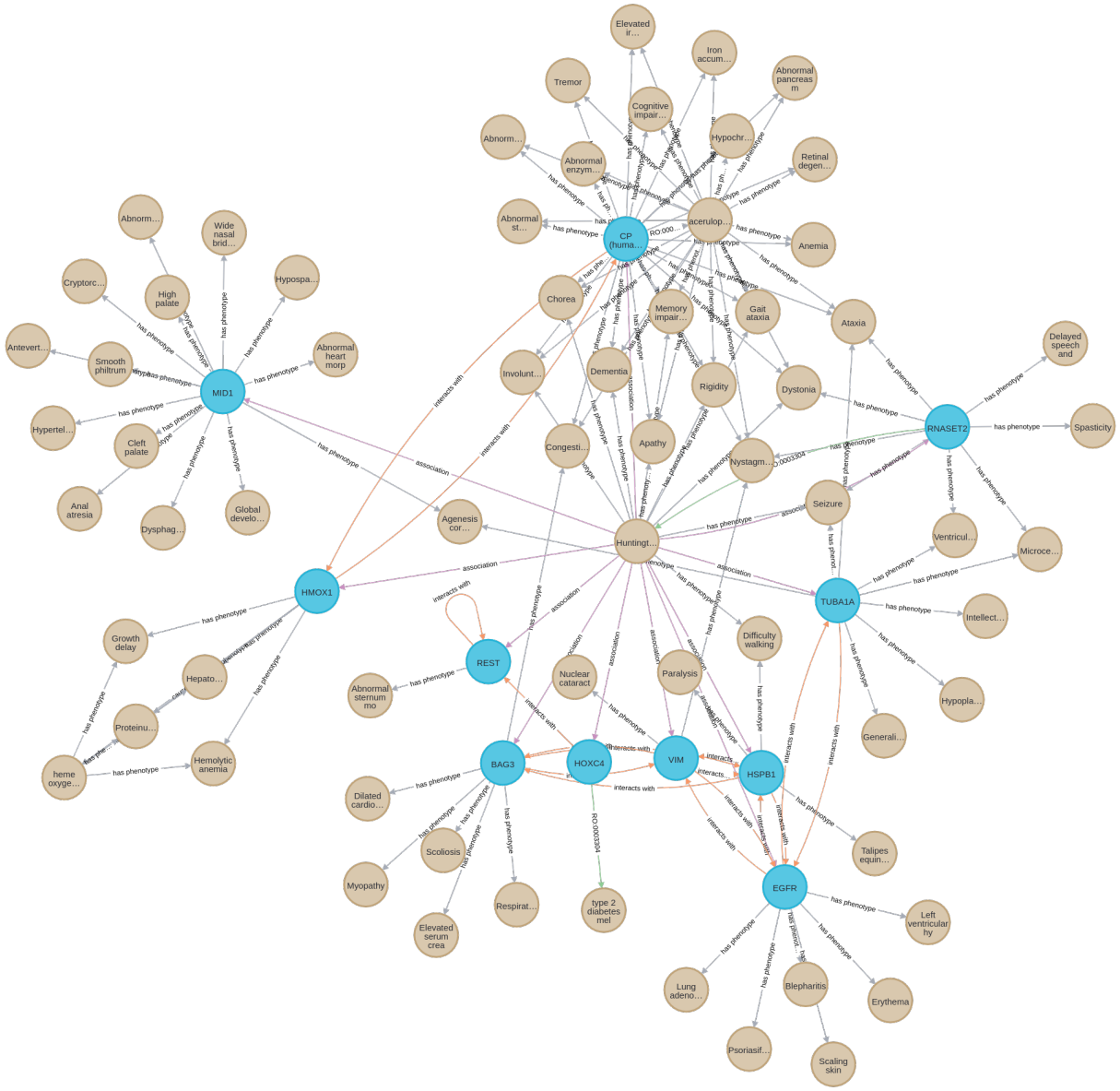

Figure 10: Disorder nodes related with differentially expressed genes within HD patient brain tissue, containing iron in their description. Purple edges indicate a gene being differentially expressed in HD, while orange edges indicate interactions between gene nodes.

6.3 Editing the Structured Review with Wikibase

Here we show the effects of implementing the wikibase on our SR. This allows us to edit the graph with new knowledge without completely reconstructing the SR. For example, the current SR does not contain an iron build-up related relationship involving the ACBD3 gene from Muller and Leavitt [5]. ACBD3, or "Acyl-coenzyme A binding domain containing 3 protein", binds to Divalent Metal Transporter (DMT1, SLC11A2, HGNC:10908), RASD1 (Dexras1) and RASD2 (Rhes) causing iron uptake in the cell [32].

```
MATCH p=(m:GENE{id:"HGNC:18229"})-[*1..3]-(n:GENE {id:"HGNC:10908"})
RETURN p
```

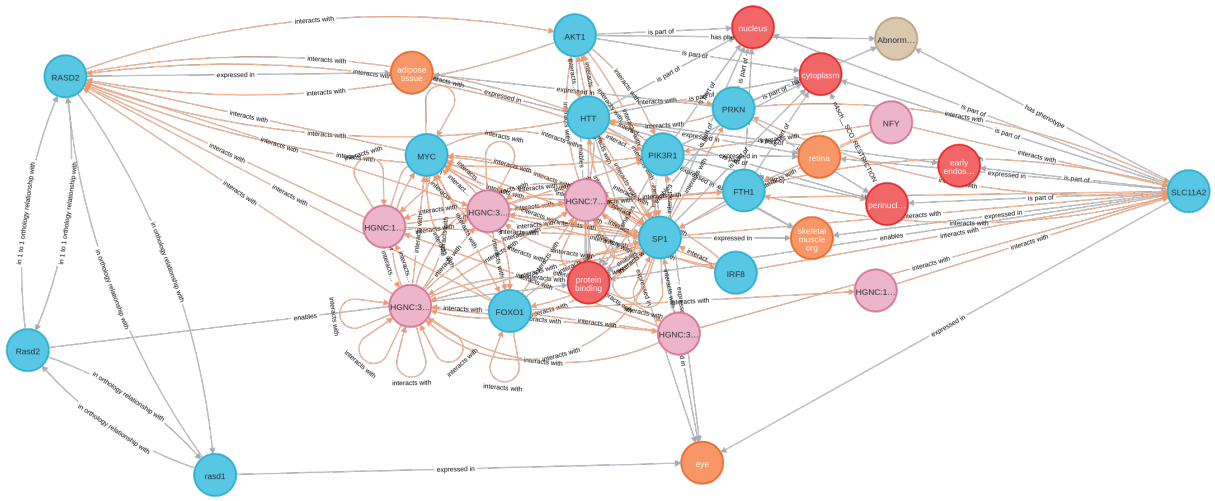

Figure 11: The results of querying the SR for the closest connection (within 3 steps) between RASD2 (Rhes) and SLC11A2 (DMT1)

As an example, the steps to add this gene to the SR through the use of the Wikibase is shown in this section: First, check if the ACBD3 gene is already in the SR by looking up it's HGNC id (HGNC:15453) 12. As seen in figure 13, there already exists a ACBD3 gene within the SR, however, the node does not contain any gene information (figure 13. Basic gene information and the ACBD3 connection to DMT1 is added by editing the page, as shown in figure 14. Adding a new relationship to a gene is done by pressing the "add statement" text and naming the edge type in the title. When filling in the title, the Wikibase gives auto-complete options on the relationship name and more importantly restricts the page editor to already existing relationships.

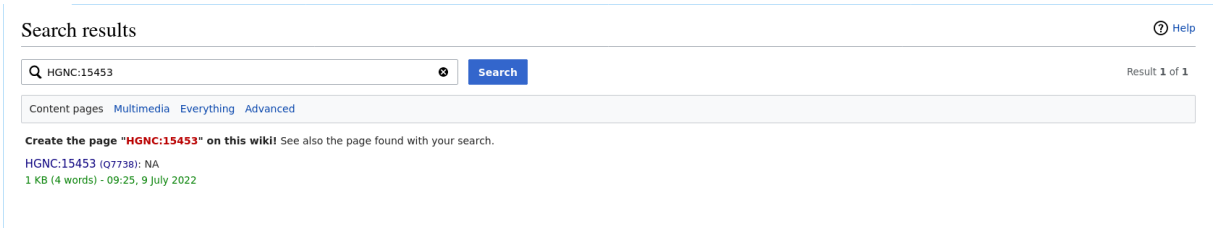

Figure 12: The result of querying the Wikibase on the ACB3 HGNC id.

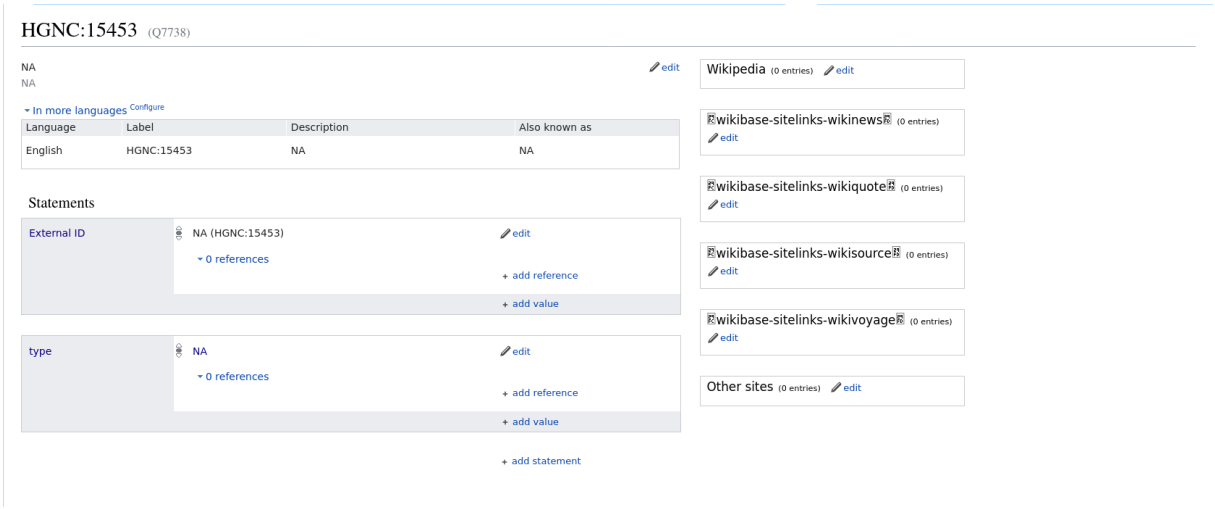

Figure 13: The ACBD3 gene information within the SR.

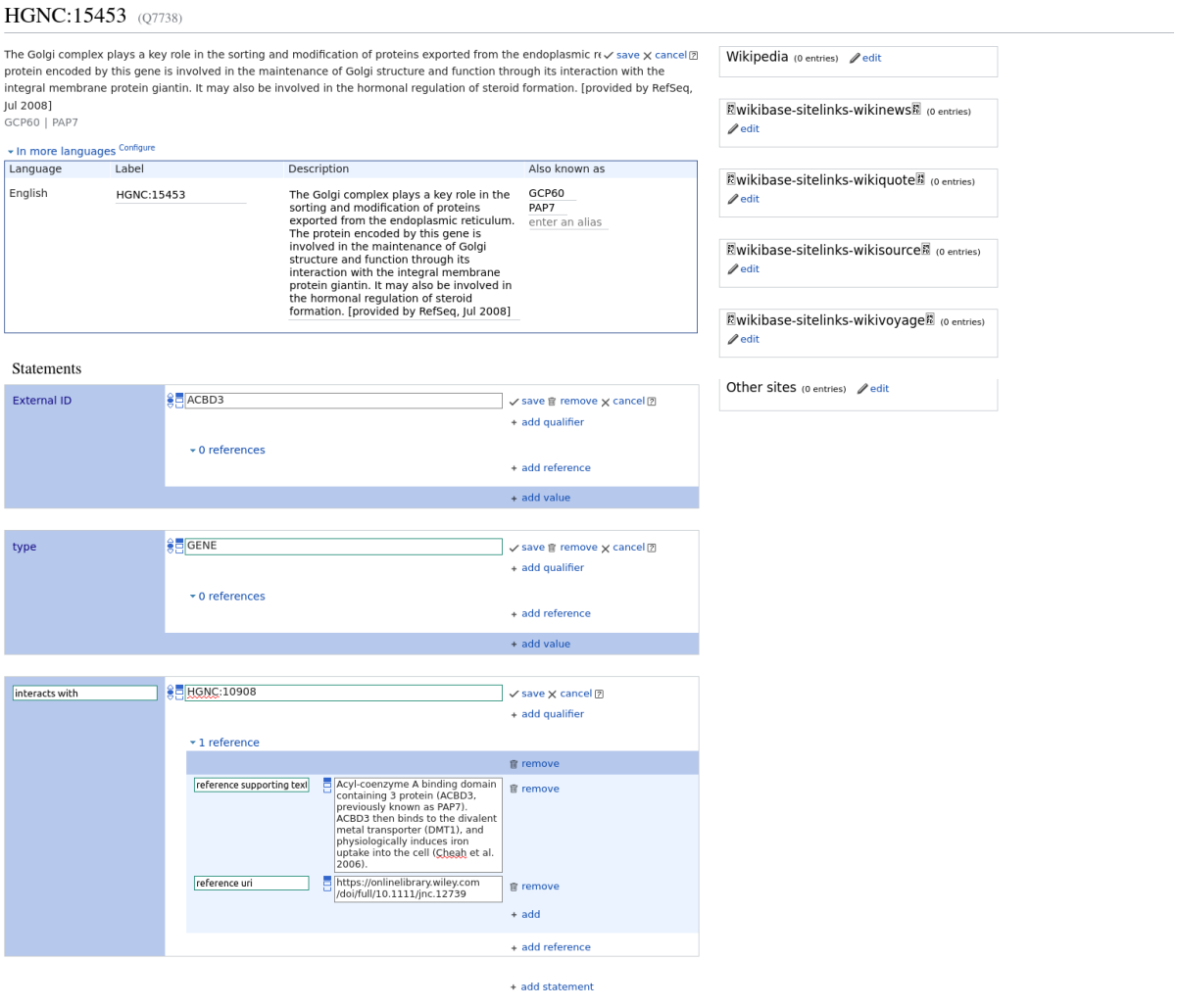

Figure 14: Filled in spaces adding relevant information on the ACB3 gene to the SR.

An "interacts with" relationships from RASD2 to ACBD3 is added and the Krusty library was used to update the Neo4j database. Using the same query as above we can see that the ACBD3 gene information and relationships are added to the results.

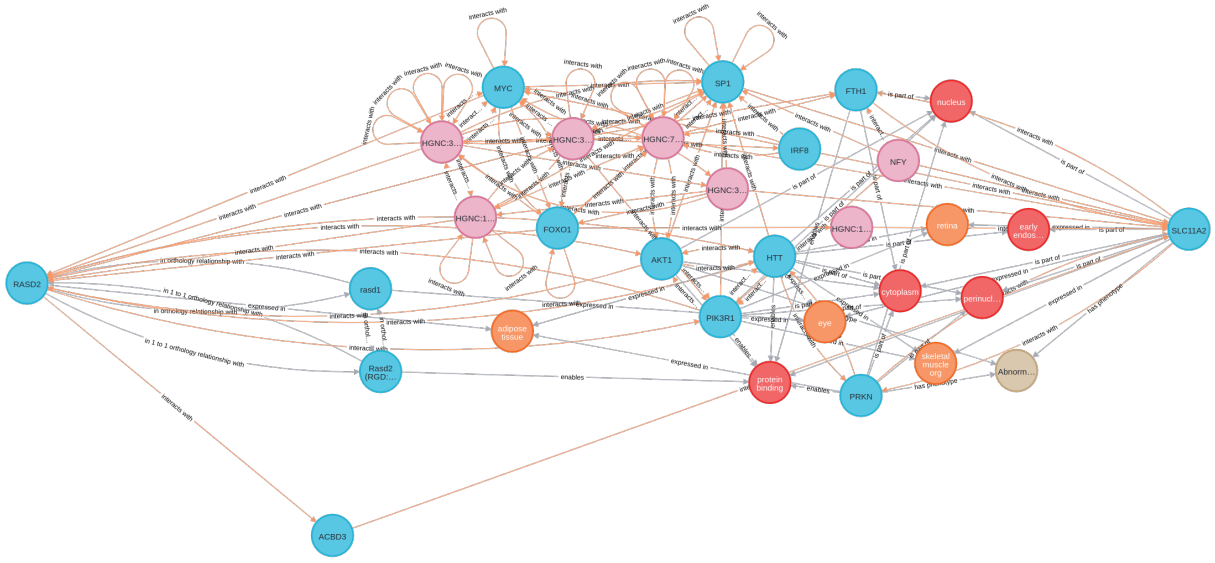

Figure 15: This figure shows the new information on ACBD3 added to the SR (Bottom) through the Wikibase interface.

6.4 Link Prediction Algorithms

As we are interested relationships between HD and iron, we focus the exploration of our prediction algorithm results on predicted relationships with the HD node in the SR and closely related nodes associated with iron.

6.4.1 Common neighbours

Table 2 shows the predictions given by our common neighbours algorithms, before and after the inclusion of the GO to the SR. Prediction scores for a relationship based on common neigbor nodes is calculated for all second degree neigbhors of the HD node. A single table is used as the scores between the two versions of the SR have not changed.

6.4.2 Cosine similarity

As seen in table 3, adding the GO to the SR has caused a slight increase in cosine similarity of the top 10 highest scoring predicted relationships. With retinal dysplasia similarity being impacted the most.

Table 2: Top 10 relationships to HD predictions based on average rank of common neighbor algorithms (recommmendationscore) before and after addition of the GO to the SR

Node2.preflabel	Node2.id	adamicAdar	commonNeighbors	preferentialAttachment	resourceAllocation	totalNeighbors	recommendationscore
rs72715653-	dbSNP:rs72715653	2.88539008177793	2	10752	1	3584	0.666759672619048
PTDSS1	HGNC:9587	1.442695040888896	1	326144	0.5	3648	0.504231770833333
OR10A2	HGNC:8161	1.442695040888896	1	215040	0.5	3613	0.502790178571429
EFR3A	HGNC:28970	1.442695040888896	1	182784	0.5	3611	0.502371651785714
ADGB	HGNC:21212	1.442695040888896	1	103936	0.5	3598	0.501348586309524
AC0046912	ENSEMBL:ENSG00000250424	1.442695040888896	1	10752	0.5	3585	0.500139508928571
LINC01337	HGNC:50546	1.442695040888896	1	10752	0.5	3585	0.500139508928571
rs8031584-	dbSNP:rs8031584	1.442695040888896	1	7168	0.5	3584	0.500093005952381
ENSEMBL:ENSG00000259720	ENSEMBL:ENSG00000259720	1.442695040888896	1	7168	0.5	3584	0.500093005952381
46XX47XX1247XX12{{}13{}}46XX{{}37{}} {{}GM02166{}}	dbSNPIndividual:21199	1.442695040888896	1	7168	0.5	3584	0.500093005952381

Table 3: Top 10 cosine similarty between HD and the same targets from the common neighbor algorithms. When the GO is added we can see that the cosine similarity between has increased with Retinal dysplasia having the highest increase in similarity.

Predicted Node label	Node Id	cosinesim	cosinesim GO
abnormal photoreceptor outer segment morphology	MP:0001004	1	1
abnormal retinal photoreceptor morphology	MP:0003729	1	1
abnormal retinal outer plexiform layer morphology	MP:0003732	0.961377385660785	0.961514573097162
abnormal photoreceptor inner segment morphology	MP:0003730	0.947036400267797	0.947001602953837
retinal photoreceptor degeneration	MP:0008450	0.944392617828452	0.944595181779083
Tg(HDexon1)62Gpb [involves: C57BL6 CBAJ]	MGI:2653604	0.831962588254057	0.831842131268968
Tg(HDexon1)61Gpb [involves: C57BL6 CBAJ]	MGI:2653603	0.829836387473637	0.829809666862312
abnormal eye electrophysiology	MP:0005551	0.818747370027381	0.818818300467252
Retinal dysplasia	HP:0007973	0.746549219573545	0.748002648616711
Tg(HDexon1)61Gpb	MGI:2389466	0.667962569821915	0.668252875288466

6.4.3 RESCAL

For a given query, the RESCAL model returns a ranked list of concepts that are most likely to match the query pattern. The RESCAL models' accuracy scores on the validation dataset are shown in table 4. The frequency of the target relationship being accuratly predicted is slightly higher when only looking at the highest scoring and lower when comparing top 10 and 100 highest scoring relationships.

Table 4: The accuracy score of the RESCAL algorithm when requiring a given predicted relationship to be in the top # 1, 10 or 100 of the ranked scores of given relationship query.

	hits at 1:	hits at 10:	hits at 100:
Standard	0.03410094714629239	0.24765667173774356	0.5673798890906414
Ontology	0.03448127791137066	0.24736222211316680	0.5672142611768171

We explore a selection of the predictions given by the two models for the query: "Huntington's Disease, contributes to condition, ?". The highest ranked recommendation is that HD contributes to HD. The second highest ranked concept is "type 2 diabetes melitus" (MONDO:0005148). The relationship between these two concepts is shown in figure 16. The genes KIF9, PTPRM, MSH3, PTDSS1 contribute to HD. Of this gene HOXC4 is differentially expressed in HD (association relationship) and contributes to type 2 diabetes and interacts with HTT. HTT causes HD, contributes to type 2 diabetes and interacts with HOXC4. This knowledge could be start of a future hypothesis on HOXC4's involvement in HD.

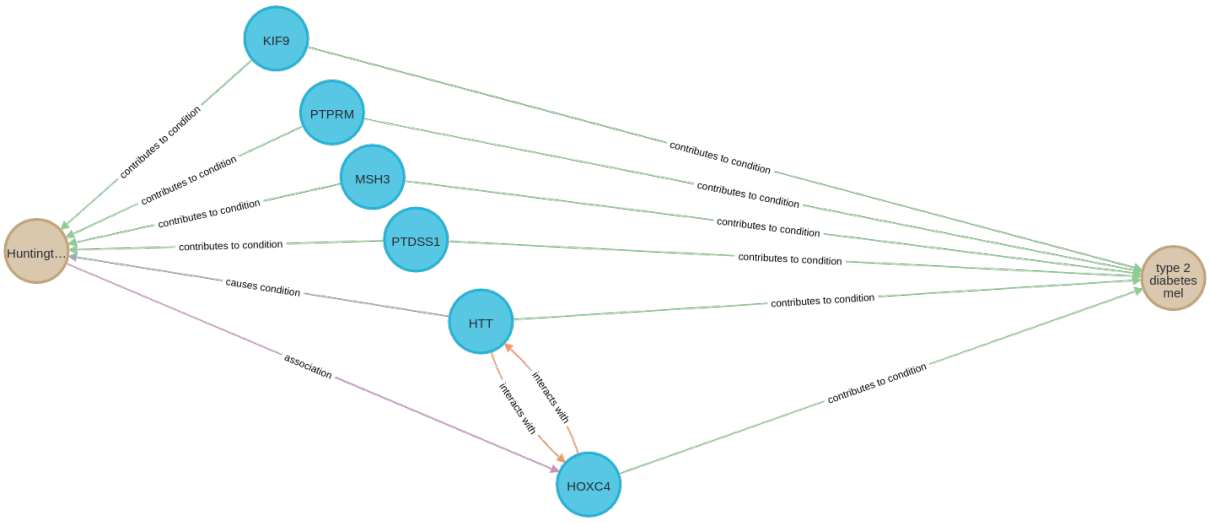

Figure 16: Shortest paths between Huntington's Disease and Type 2 Diabetes concept, ranked 2nd place in predicting the "Huntington's Disease, contributes to condition, ?" using the RESCAL model

The 3rd ranked concept related to HD with the "contributes to condition" relationship is platelet count. As shown in figure 17 HD and platelet count are most closely related to one another through 5 genes. BRF1 and KIF9 contribute to both concepts, with BRF1 interacting with REST. REST is differentially expressed in HD, as is EGFR, and these genes are contributors to platelet count and interact with HTT. As in the previous prediction, HTT causes HD and contributes to platelet count.

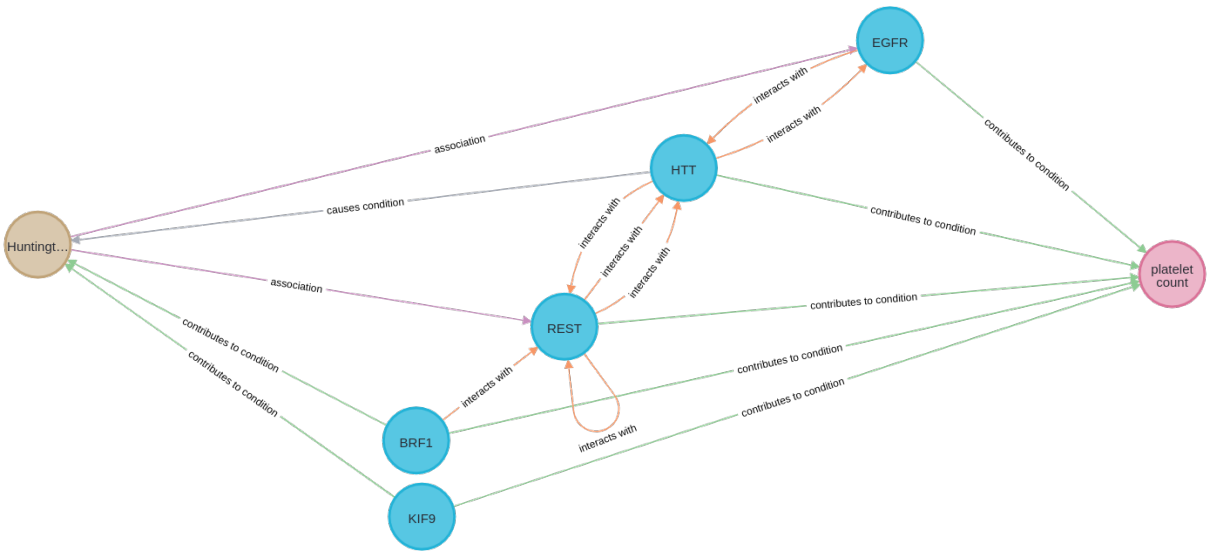

Figure 17: Shortest paths between Huntington's Disease and Platelet count concept, ranked 3th (or 6th including ontology terms) place in predicting the "Huntington's Disease, contributes to condition, ?" using the RESCAL model

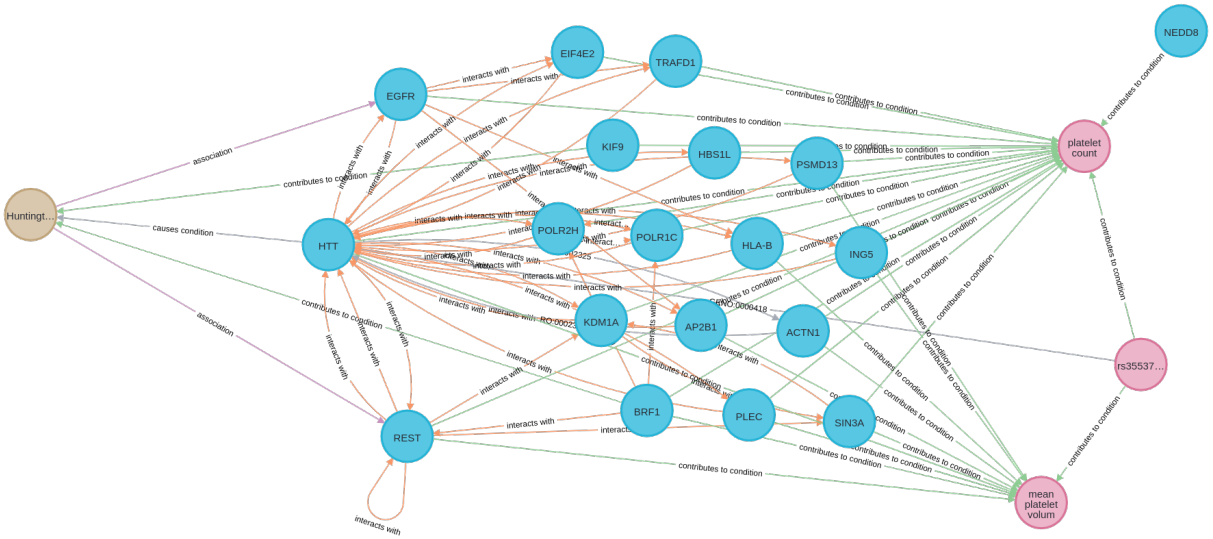

Figure 18: An expansion on figure 17, here we can see the addition of the SNP rs35537543-G and the mean platelet volume genotype nodes.

The model based on the SR with the integration of GO structure provides some differing predictions. Of interest is "neutrophil count" which now appears in the top 10 highest ranked predictions. The results of the Neo4j shortest path algorithm result of HD to "neutrophil count" is shown in figure 19. To provide more context to the "neutrophil count" within HD, we expand the relationships to the node (figure 20). This new figure shows a large amount of genes that interact with HTT that are associated to "neutrophil count" through a "contributes to condition" relationship. These genes have all been associated with neutrophil count through the GWAS study of Hensman Moss and Davina J[33].

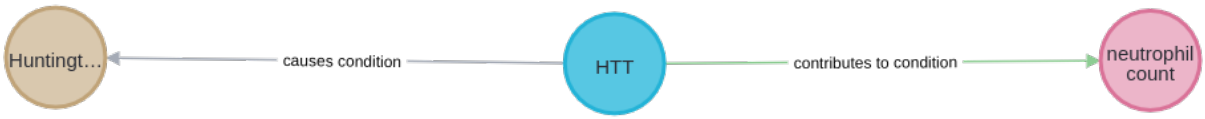

Figure 19: Shortest paths between Huntington's Disease and neutrophil count concept.

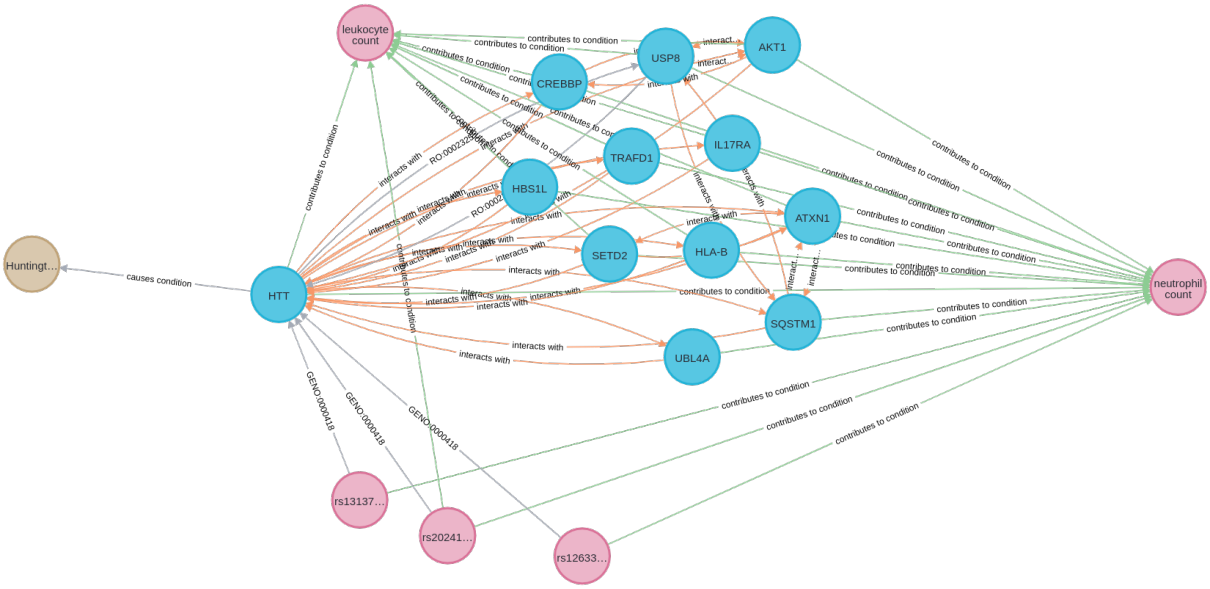

Figure 20: An expansion of the relationships of the "neutrophil count" concept in figure 19. Here we can see the addition of HTT SNP: rs1263309-T, rs1313766-C, rs2024115-G, and the "leukocyte count" genotype nodes.

7 Conclusion

We successfully achieved our first goal in this project: we implemented the Wikibase, extended the SR with Gene Ontology information and applied link prediction algorithms to the SR to exploit it's machine-readable structure. The RESCAL algorithm in particular may not have very high prediction metrics, but has shown to capture current topics of interest in HD research related to iron.

8 Discussion

The results of this project show that there is a bright future ahead for applying link prediction on SRs in research. However, there As this project has a wide array of results, we will tackle each results section one by one.

8.1 Structured Review

Section 6.1 shows that we have captured related information on HD and Iron. However, when one considers the fact that including the GO more than doubles the size of the SR, inclusion of a larger section of the Monarch Initiative may be justified as a majority of the information has become GO structure. A feature of the SR is that it is context specific across information types [1] and expanding the SR using the Monarch Initiative has to carefully balanced to retain the specificity of the SR. A solution to this would be to increase the amount of data sources the BKR uses to construct the SR or to expand the extraction

8.2 Gene Ontology

Based on the amount of nodes added into the SR we would have expected a larger impact on the predictions made by both the common neighbor and RESCAL algorithms. The main reason why the common neighbor algorithms have not changed their scores is because the addition of the GO does not add common neighbors between HD and it's second degree neighbors. Even for the Resource Allocation algorithm the impact on the connectivity of common neighbors is not enough to make significant changes in the rankings of scores. However, the algorithms using node embeddings are affected by the inclusion of the GO, as seen in the cosine similarity scores and the RESCAL prediction accuracy. There are multiple methods that may improve the impact of the GO in future iterations of the SR. First, we could, as with gene homologs, add a step to the Monarch Initiative extraction to specifically obtain the GO terms associated to concepts. This will increase the frequency of GO terms and increase the contribution of adding GO parent-child relationships into the SR. However, this will not make the addition of the GO through neosemantics improve the nearest neighbor scoring. Because of this, we recommend prioritizing the use of algorithms that use embeddings to maximize information value extraction. As the RESCAL model generates a separate matrix for every relationship type in a graph, merging relationship across the GO and the SR or nuanced SR relationship edges into more generic relationships may improve the model accuracy. This does brings in the question if using more generic relationships reduce the interpretability of RESCAL predictions, as some implications that specific relationship types can contain are lost.

8.3 Common neighbors

There are multiple ways of interpreting the common neighbor algorithm scores. We chose to prioritize our analysis based on average rank across all measurements for the sake of time. With average rank scores close to 0.5, which represent middle-of-the-pack average scoring, we concentrate on rs72715653. This is a variant within dbSNP in CPSF2, this gene produces a RNA binding protein that may affect the stability of mRNA isoforms [34]. Extensive analysis based on an per algorithm approach may provide a result. A pattern found in the common neighbor results is that genes have a higher score on average, this is expected as genes are the node type with the highest connectivity in the graph.

8.4 Cosine Similarity

Compared to other results the cosine similarity measurement is focused on a seemingly different topic. But these results still help with highlighting possible information of interest on iron in HD. Retinal dysplasia has been found in both diabetes and neurodegenerative disorders [35].

8.5 RESCAL predictions

The predictions from RESCAL show a lot of promise for the future of SRs. The top three ranked predictions can be justified through the SR and literature. For the first, it makes sense that the model ranks HD as contributing to itself. HD is one of the core concepts of the SR and a majority of it's neighbouring concepts are contributing to HD. With interactions of concepts with themselves allowed in the graph it is within reason that the model suggest that HD is contributing to itself.

A link with HD and type 2 diabetes melitus has been established in literature with increased prevalence of diabetes and drugs that target diabetes being suggested as potential HD therapies [36, 37, 38]. The fact that both RESCAl and cosine similarity point to diabetes relates concepts but from different perspectives highlights the added value of using multiple methods of analyzing the SR giving researchers a stronger justification for investigating the relationship in the future.

The platelet count prediction is of special interest to our goal of exploring the role of iron in HD. Literature on blood platelets in HD has seen a increase in publications in recent years and has been the primairy subject of research

into HD before HTT was identified as the causal gene. As blood platelets are derived from megakaryocytes. The development of megakaryocyte cells is partially based on iron levels [39] and use iron in their mitochondria production.
KIF9 and BRF1 have been associated to HD and platelet count through multiple GWAS studies [33, 40, 41]. Represented by the "contributes to condition" relationships to HD and platelet count. KIF9 is a Kinesin (microtubule transport protein). However, specific interactions within megakaryocytes of these genes is still a largely unexplored subject within literature. Of interest is the fact that an association has been made with KIF9 to type 2 diabetes in literature [42]. Linking both predictions in a manner beyond the knowledge stored within the SR. These results shows the power of structured data and the application of knowledge graph embedding techniques to extract the maximum value of current knowledge.
The final relationship prediction we highlighted, HD and neutrophil count, is another concept that has been found in literature. Neutrophil production has been associated to plasma iron [43] relevant genes are seen in figure 7, aided by the new integrated GO relationships.

8.6 Graph size

The size of our base SR can be a major topic of discussion. The size of our SR was chosen to be appropriate at the time of construction as a relatively small graph has its own advantages: First of, small graph require less resources to store and transfer across the Neo4j and Wikibase services. This smaller storage size can promote sharing of SR data between researchers or provide an incentive to construct SR's "off the cuff" as BKR run time is reduced. Smaller graphs also require a smaller amount of time investment for our recommendation algorithms. Even in our current results, we specifically calculated the common neighborhood and cosine similarity scores surrounding the HD node as a time-saving measure as calculating these scores is a $NLOG(N)$ O problem as it is a pair-wise action. For users without prior knowledge on the subject who would perform these measurements at a larger scale, a restriction of the graph size can help focus on core concepts within the context of a topic that the SR tries to capture. However, the purpose of a SR is to capture the context surrounding a certain topic. We argue that the size of the SR was insufficient to fully capture the context of iron in HD. This is reflected in the query results figure 8, the genes shown in this graph do not have any relationships wiht GO terms. This is most likely due to a to early stop in the subgraph extraction from the Monarch Initiative data source, as these genes are annotated when visited on their web-interface. An issue is caused by this is that users are not able to associate genes to the biological process, molecular functions and cellular components defined in the GO at a glance. Furthermore, this early-stop have decreased the quality of training ratios for our RESCAL model. Whereby the model may not have gotten enough examples of physiology nodes being related to genes. This is due to the fact that every unique relationship type has a separate interaction matrix, a low population of specific relationship type can cause these matrices to be unable to be generalized for multiple SR's and reduce overall accuracy scores. We recommend future implementations of the SR's to include multiple intermediate exploratory checks beyond the graph schema, pathfinding between seed nodes and checking the presence of nodes from the different data-sources. Specifically, we suggest performing similar user-story like queries on the graph as in our current result section. Equally important is that now that link prediction can be applied to reduce graph exploration time, we recommend extracting more information from the Monarch Initiative to make sure that every concept is adequately annotated by Physiology concept nodes. Beyond that, a more thorough exploration of node and relationship type frequency in the graph may give an indication of blind spots in the SR.
Similar to any review article where there are limits to the amount of knowledge that can be gained from a single publication. A SR is best used as a tool within a literature study, as it is not (yet) possible to fully automate this important step within the scientific process. The results of this project suggest that the value of the SR as a hypothesis-generating and collaborative tool will increase over time, as the quality of BKR workflow and other tools improve through additional source integration, predictive models and manual refinement.

9 Future work

There are multiple ways to expand upon the current SR:

9.1 More experimental data.

The more experimental data we incorporate, the more likely we are to find a relationship between iron or HTT that wasn't readily understood by including a single experiment.

9.2 Neo4j Bloom

For researchers less familiar with programming languages, adding a higher level interface to the SR may improve the querying capabilities of the SR. One possible option is Neo4j Bloom[5]. This allows a more natural language to be used to query the graph.

9.3 Combining recommendation systems

Currently, the recommendation scores are stored independently of the Neo4j or Wikidata user interfaces. Displaying the recommended relationships in these interfaces may provide a more user-friendly method of exploring the results of these predictions.

10 Acknowledgements

Many thanks to my supervisors Eleni Mina, Núria Queralt-Rosinach and Katy Wolstencroft, this has been an enormous undertaking which I would not be able to perform without your continuous support and advice. Also thanks to Elsa Kuijper for her expert opinion and advice when testing and exploring the SR together. Finally, I thank the Biosemantics LUMC group and my family for their support.
This thesis uses and displays data and algorithms from the Monarch Initiative. The Monarch Initiative (https://monarchinitiative.org) makes biomedical knowledge exploration more efficient and effective by providing tools for genotype-phenotype analysis, genomic diagnostics, and precision medicine across broad areas of disease.

[5] https://neo4j.com/product/bloom/

References

[1] N. Queralt-Rosinach, G. S. Stupp, T. S. Li, M. Mayers, M. E. Hoatlin, M. Might, B. M. Good, and A. I. Su, "Structured reviews for data and knowledge-driven research," *Database*, vol. 2020, 04 2020. baaa015.

[2] F. O. Walker, "Huntington's disease," *The Lancet*, vol. 369, no. 9557, pp. 218–228, 2007.

[3] M. Duyao, C. Ambrose, R. Myers, A. Novelletto, F. Persichetti, M. Frontali, S. Folstein, C. Ross, M. Franz, M. Abbott, *et al.*, "Trinucleotide repeat length instability and age of onset in huntington's disease," *Nature genetics*, vol. 4, no. 4, pp. 387–392, 1993.

[4] S. Chen, F. A. Ferrone, and R. Wetzel, "Huntington's disease age-of-onset linked to polyglutamine aggregation nucleation," *Proceedings of the National Academy of sciences*, vol. 99, no. 18, pp. 11884–11889, 2002.

[5] M. Muller and B. R. Leavitt, "Iron dysregulation in huntington's disease," *Journal of neurochemistry*, vol. 130, no. 3, pp. 328–350, 2014.

[6] U. S. Srinivas, B. W. Tan, B. A. Vellayappan, and A. D. Jeyasekharan, "Ros and the dna damage response in cancer," *Redox biology*, vol. 25, p. 101084, 2019.

[7] P. McColgan and S. J. Tabrizi, "Huntington's disease: a clinical review," *European journal of neurology*, vol. 25, no. 1, pp. 24–34, 2018.

[8] S. Spangler, A. D. Wilkins, B. J. Bachman, M. Nagarajan, T. Dayaram, P. Haas, S. Regenbogen, C. R. Pickering, A. Comer, J. N. Myers, *et al.*, "Automated hypothesis generation based on mining scientific literature," in *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 1877–1886, 2014.

[9] W. Jin, R. K. Srihari, and H. H. Ho, "A text mining model for hypothesis generation," in *19th IEEE International Conference on Tools with Artificial Intelligence (ICTAI 2007)*, vol. 2, pp. 156–162, IEEE, 2007.

[10] A. M. Liekens, J. De Knijf, W. Daelemans, B. Goethals, P. De Rijk, and J. Del-Favero, "Biograph: unsupervised biomedical knowledge discovery via automated hypothesis generation," *Genome biology*, vol. 12, no. 6, pp. 1–12, 2011.

[11] T. Doğan, H. Atas, V. Joshi, A. Atakan, A. S. Rifaioglu, E. Nalbat, A. Nightingale, R. Saidi, V. Volynkin, H. Zellner, *et al.*, "Crossbar: comprehensive resource of biomedical relations with knowledge graph representations," *Nucleic Acids Research*, vol. 49, no. 16, pp. e96–e96, 2021.

[12] X. Chen, S. Jia, and Y. Xiang, "A review: Knowledge reasoning over knowledge graph," *Expert Systems with Applications*, vol. 141, p. 112948, 2020.

[13] K. A. Shefchek, N. L. Harris, M. Gargano, N. Matentzoglu, D. Unni, M. Brush, D. Keith, T. Conlin, N. Vasilevsky, X. A. Zhang, *et al.*, "The monarch initiative in 2019: an integrative data and analytic platform connecting phenotypes to genotypes across species," *Nucleic acids research*, vol. 48, no. D1, pp. D704–D715, 2020.

[14] S. A. Lambert, A. Jolma, L. F. Campitelli, P. K. Das, Y. Yin, M. Albu, X. Chen, J. Taipale, T. R. Hughes, and M. T. Weirauch, "The human transcription factors," *Cell*, vol. 172, no. 4, pp. 650–665, 2018.

[15] R. E. Dolmetsch, R. S. Lewis, C. C. Goodnow, and J. I. Healy, "Differential activation of transcription factors induced by ca2+ response amplitude and duration," *Nature*, vol. 386, no. 6627, pp. 855–858, 1997.

[16] A. S. Fleischhacker and P. J. Kiley, "Iron-containing transcription factors and their roles as sensors," *Current Opinion in Chemical Biology*, vol. 15, no. 2, pp. 335–341, 2011. Biocatalysis and Biotransformation/Bioinorganic Chemistry.

[17] A. Gupta, S. F. Rosenberger, and G. T. Bowden, "Increased ROS levels contribute to elevated transcription factor and MAP kinase activities in malignantly progressed mouse keratinocyte cell lines ," *Carcinogenesis*, vol. 20, pp. 2063–2073, 11 1999.

[18] D. Diefenbach, M. D. Wilde, and S. Alipio, "Wikibase as an infrastructure for knowledge graphs: The eu knowledge graph," in *The Semantic Web – ISWC 2021* (A. Hotho, E. Blomqvist, S. Dietze, A. Fokoue, Y. Ding, P. Barnaghi, A. Haller, M. Dragoni, and H. Alani, eds.), (Cham), pp. 631–647, Springer International Publishing, 2021.

[19] L. A. Adamic and E. Adar, "Friends and neighbors on the web," *Social Networks*, vol. 25, no. 3, pp. 211–230, 2003.

[20] T. Zhou, L. Lü, and Y.-C. Zhang, "Predicting missing links via local information," *The European Physical Journal B*, vol. 71, pp. 623–630, 2009.

[21] H. Jeong, Z. Néda, and A.-L. Barabási, "Measuring preferential attachment in evolving networks," *Europhysics letters*, vol. 61, no. 4, p. 567, 2003.

[22] H. Chen, S. F. Sultan, Y. Tian, M. Chen, and S. Skiena, "Fast and accurate network embeddings via very sparse random projection," in *Proceedings of the 28th ACM international conference on information and knowledge management*, pp. 399–408, 2019.

[23] S. Dasgupta and A. Gupta, "An elementary proof of a theorem of johnson and lindenstrauss," *Random Structures & Algorithms*, vol. 22, no. 1, pp. 60–65, 2003.

[24] A. Labadorf, A. G. Hoss, V. Lagomarsino, J. C. Latourelle, T. C. Hadzi, J. Bregu, M. E. MacDonald, J. F. Gusella, J.-F. Chen, S. Akbarian, *et al.*, "Rna sequence analysis of human huntington disease brain reveals an extensive increase in inflammatory and developmental gene expression," *PloS one*, vol. 10, no. 12, p. e0143563, 2015.

[25] C. Jiang, Z. Xuan, F. Zhao, and M. Q. Zhang, "Tred: a transcriptional regulatory element database, new entries and other development," *Nucleic acids research*, vol. 35, no. suppl_1, pp. D137–D140, 2007.

[26] G. Zheng, K. Tu, Q. Yang, Y. Xiong, C. Wei, L. Xie, Y. Zhu, and Y. Li, "Itfp: an integrated platform of mammalian transcription factors," *Bioinformatics*, vol. 24, no. 20, pp. 2416–2417, 2008.

[27] E. P. Consortium *et al.*, "An integrated encyclopedia of dna elements in the human genome," *Nature*, vol. 489, no. 7414, p. 57, 2012.

[28] S. Neph, A. B. Stergachis, A. Reynolds, R. Sandstrom, E. Borenstein, and J. A. Stamatoyannopoulos, "Circuitry and dynamics of human transcription factor regulatory networks," *Cell*, vol. 150, no. 6, pp. 1274–1286, 2012.

[29] H. Han, H. Shim, D. Shin, J. E. Shim, Y. Ko, J. Shin, H. Kim, A. Cho, E. Kim, T. Lee, *et al.*, "Trrust: a reference database of human transcriptional regulatory interactions," *Scientific reports*, vol. 5, no. 1, pp. 1–11, 2015.

[30] D. Marbach, D. Lamparter, G. Quon, M. Kellis, Z. Kutalik, and S. Bergmann, "Tissue-specific regulatory circuits reveal variable modular perturbations across complex diseases," *Nature methods*, vol. 13, no. 4, pp. 366–370, 2016.

[31] S. Broscheit, D. Ruffinelli, A. Kochsiek, P. Betz, and R. Gemulla, "LibKGE - A knowledge graph embedding library for reproducible research," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 165–174, 2020.

[32] B.-R. Choi, S. Bang, Y. Chen, J. H. Cheah, and S. F. Kim, "Pka modulates iron trafficking in the striatum via small gtpase, rhes," *Neuroscience*, vol. 253, pp. 214–220, 2013.

[33] D. J. Hensman Moss, *Identification of genetic factors underpinning phenotypic heterogeneity in Huntington's disease and other neurodegenerative disorders*. PhD thesis, UCL (University College London), 2020.

[34] L. Romo, A. Ashar-Patel, E. Pfister, and N. Aronin, "Alterations in mrna 3 utr isoform abundance accompany gene expression changes in human huntington's disease brains," *Cell reports*, vol. 20, no. 13, pp. 3057–3070, 2017.

[35] J. M. Sundstrom, C. Hernández, S. R. Weber, Y. Zhao, M. Dunklebarger, N. Tiberti, T. Laremore, O. Simó-Servat, M. Garcia-Ramirez, A. J. Barber, *et al.*, "Proteomic analysis of early diabetic retinopathy reveals mediators of neurodegenerative brain diseases," *Investigative ophthalmology & visual science*, vol. 59, no. 6, pp. 2264–2274, 2018.

[36] M. T. Montojo, M. Aganzo, and N. González, "Huntington's disease and diabetes: Chronological sequence of its association," *Journal of Huntington's disease*, vol. 6, no. 3, pp. 179–188, 2017.

[37] S. J. Schönberger, D. Jezdic, R. L. Faull, and G. J. Cooper, "Proteomic analysis of the human brain in huntington's disease indicates pathogenesis by molecular processes linked to other neurodegenerative diseases and to type-2 diabetes," *Journal of Huntington's disease*, vol. 2, no. 1, pp. 89–99, 2013.

[38] D. Hervas, V. Fornés-Ferrer, A. P. Gómez-Escribano, M. D. Sequedo, C. Peiro, J. M. Millán, and R. P. Vazquez-Manrique, "Metformin intake associates with better cognitive function in patients with huntington's disease," *PloS one*, vol. 12, no. 6, p. e0179283, 2017.

[39] E. Brissot, M.-B. Troadec, O. Loréal, and P. Brissot, "Iron and platelets: a subtle, under-recognized relationship," *American Journal of Hematology*, vol. 96, no. 8, pp. 1008–1016, 2021.

[40] D. J. H. Moss, A. F. Pardiñas, D. Langbehn, K. Lo, B. R. Leavitt, R. Roos, A. Durr, S. Mead, A. Coleman, R. D. Santos, *et al.*, "Identification of genetic variants associated with huntington's disease progression: a genome-wide association study," *The Lancet Neurology*, vol. 16, no. 9, pp. 701–711, 2017.

[41] M.-H. Chen, L. M. Raffield, A. Mousas, S. Sakaue, J. E. Huffman, A. Moscati, B. Trivedi, T. Jiang, P. Akbari, D. Vuckovic, *et al.*, "Trans-ethnic and ancestry-specific blood-cell genetics in 746,667 individuals from 5 global populations," *Cell*, vol. 182, no. 5, pp. 1198–1213, 2020.

[42] A. Khamis, M. Canouil, A. Siddiq, H. Crouch, M. Falchi, M. von Bulow, F. Ehehalt, L. Marselli, M. Distler, D. Richter, J. Weitz, K. Bokvist, I. Xenarios, B. Thorens, A. M. Schulte, M. Ibberson, A. Bonnefond, P. Marchetti, M. Solimena, and P. Froguel, "Laser capture microdissection of human pancreatic islets reveals novel eqtls associated with type 2 diabetes," *Molecular Metabolism*, vol. 24, pp. 98–107, 2019.

[43] J. N. Frost, S. K. Wideman, A. E. Preston, M. R. Teh, Z. Ai, L. Wang, A. Cross, N. White, Y. Yazicioglu, M. Bonadonna, *et al.*, "Plasma iron controls neutrophil production and function," *Science Advances*, vol. 8, no. 40, p. eabq5384, 2022.

11 Appendix

11.1 Total node and edge counts of the SR

Table 5

Node type	Count
"DISO":	5317,
"NA":	15841,
"GENE":	9018,
"PHYS":	48,
"VARI":	6,
"GENO":	1016,
"ANAT":	20

Table 6

Edge type	Count
"BFO:0000050"	5327,
"RO:0002326"	5,
"RO:0002327"	1424,
"RO:0002206"	2796,
"sno:408094002"	159,
"GENO:0000840"	6,
"RO:HOM0000017"	76332,
"RO:0002607"	1,
"RO:0004013"	2,
"RO:0002331"	459,
"RO:0002200"	22656,
"RO:0003301"	288,
"RO:HOM0000020"	65772,
"MI_0914"	1356,
"RO:0004016"	1,
"GENO:0000222"	9,
"RO:0002434"	215255,
"RO:0003303"	12,
"RO:0003304"	363,
"RO:0002325"	562,
"GENO:0000418"	50,
"GENO:0000408"	14,
"dc:source"	23