Gene-Disease network analysis

Vincenzo Altavilla
v.altavilla1@studenti.unipi.it
Student ID: 690642

Chiara Capodagli
c.capodagli@studenti.unipi.it
Student ID: 693737

Alice Malfatti
a.malfatti3@studenti.unipi.it
Student ID: 586284

Abstract

This study investigates the structure of the human gene–disease network to reveal patterns of gene interactions and disease associations. We applied community detection to identify both large hubs and small gene modules and performed enrichment analyses to understand their biological relevance. Additionally, we explored indirect genetic relationships to uncover connections between diseases across different biomedical categories. Our findings highlight specialized gene communities, central hub genes and cross-disease associations, offering insights into the organization and functional relationships within the gene–disease network.
[1]

Keywords

Social Network Analysis, Biological networks, Community detection, Network enrichment, Gene–disease associations.

ACM Reference Format:
Vincenzo Altavilla, Chiara Capodagli, and Alice Malfatti. 2019. Gene-Disease network analysis. In *Social Network Analysis.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction

Understanding how genes relate to human diseases is essential for uncovering biological mechanisms and identifying potential therapeutic targets. Networks that map gene–disease associations can reveal modular structures, where groups of genes are functionally related, as well as hub genes that influence multiple conditions. Beyond direct associations, indirect genetic relationships can connect diseases across different physiological systems, providing a broader view of disease biology. In this work, we examine both the modular organization of gene communities and higher-order connections between diseases, aiming to highlight biologically meaningful patterns in the human gene–disease network.

[1]**Project Repositories**
Data Collection: https://github.com/sna-unipi/sna-final-project-2025-2025_altavilla_capodagli_malfatti/tree/main/data_collection
Analytical Tasks: https://github.com/sna-unipi/sna-final-project-2025-2025_altavilla_capodagli_malfatti/tree/main/network_analysis
Open Problem: https://github.com/sna-unipi/sna-final-project-2025-2025_altavilla_capodagli_malfatti/tree/main/open_problem
Report: https://github.com/sna-unipi/sna-final-project-2025-2025_altavilla_capodagli_malfatti/tree/main/report

SNA, University of Pisa, Italy

ACM ISBN 978-x-xxxx-xxxx-x/YYYY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

2 Data Collection

This section describes the procedures adopted to construct the dataset used in our analyses. We first identify the data source, then outline the crawling methodology and finally state the assumptions made during the collection process.

2.1 Selected Data Sources

For the purpose of this project we selected the **Harmonizome** platform as our data source. Harmonizome is an online platform that integrates information from multiple biological resources and provides programmatic access to genes and their functional terms, which are extracted and organized from over a hundred publicly available resources [1]. Specifically, we focused on the dataset *"DISEASES Experimental Gene-Disease Association Evidence Scores 2025"*, which contains curated associations between human genes and diseases, each characterized by a standardized evidence score[2]. These scores are derived from the confidence scoring scheme of the DISEASES database and integrate multiple sources of information, like namely curated databases, genome-wide association studies (GWAS) and automatic text mining of biomedical literature [2]. While DISEASES internally reports star-based confidence levels (1-5), Harmonizome exposes a continuous standardized score which preserves the relative strength of associations across evidence types. This dataset was chosen because it provides a large-scale bipartite structure of gene-disease relationships and represents one of the few sources that could be effectively extracted given both computational feasibility and the need to balance dataset size.

2.1.1 ***Crawling Methodology.*** The collection process of the dataset was performed through the official Harmonizome REST API. We developed a crawler to retrieve and integrate three complementary layers of information:

(1) **Gene-Disease Associations:** For each disease in the selected dataset we retrieved all associated genes, storing their identifiers, hyperlinks and standardized evidence scores. The resulting table contains 171,854 associations between 10,636 unique genes and 557 distinct diseases.
(2) **Gene Metadata:** For each unique gene we collected additional attributes such as synonyms, full name, description, NCBI identifiers, links to protein products and membership to HGNC gene families. This step produced information for 10,634 genes. Two genes (*HSPA1A* and *HSPA1B*) could not be retrieved because the API repeatedly failed to provide a valid response for them, resulting in missing metadata.
(3) **Disease Metadata:** For each disease we retrieved naming authority information, external identifiers, textual descriptions and connections to other gene sets where available. This step successfully returned information for all 557 diseases.

[2]URL: https://maayanlab.cloud/Harmonizome/dataset/DISEASES%2BExperimental%2BGene-Disease%2BAssociation%2BEvidence%2BScores%2B2025

The resulting data was consolidated into a single tabular structure, producing a unified dataset that links associations to enriched gene and disease attributes. This final table contains 171,708 rows and 25 features.

2.1.2 ***Assumptions and Graph Construction.*** Prior to graph construction, several preprocessing steps were applied to the raw dataset to ensure the robustness and interpretability of the resulting network. Starting from the unified table described above, we retained, temporarily, only the columns corresponding to gene identifiers and disease names, as these represent the two partitions of our bipartite system. A preliminary inspection of the raw graph revealed an unexpectedly **high density (**0.98**)**, suggesting substantial redundancy in the disease representation. Upon closer examination, we found that many disease nodes referred to the same pathological condition under different names or identifiers, leading to multiple overlapping entries connected to largely identical gene sets. To address this issue, we first applied a consolidation step by collapsing equivalent disease entities. Specifically, we employed the `collapse_node` function provided by the *HyperNetX* library[3], which merges nodes belonging to the same equivalence class into a single representative node while preserving the union of their incident edges. This procedure grouped together all diseases describing the same condition and connected all associated genes to the resulting collapsed disease node. As a result, redundancy was reduced and the network topology became more representative of true biological relationships, without loss of connectivity information.

Subsequently, we performed an additional consolidation step: we noticed that several disease entries represented extremely broad categories, such as *"disease"* or *"disease of anatomical entity"*, which were linked to almost every gene in the dataset. These overly generic terms acted as artificial hubs, obscuring the underlying structure of the network. To mitigate this effect, we excluded all diseases whose names contained the word *"disease"*. Since these entries typically corresponded to unspecific or umbrella categories overlapping with more specific pathological entities, their removal effectively improved the specificity and interpretability of the final network. After these processing steps, the curated dataset comprised 9,334 unique genes and 250 distinct diseases.

From this preprocessed data, we constructed a bipartite graph $G = (V_g, V_d, E)$, where V_g and V_d denote the sets of genes and diseases, respectively, and each edge $(g_i, d_j) \in E$ indicates an association between gene g_i and disease d_j. We then derived the **gene–gene projection** of this bipartite network, where nodes represent genes and an undirected edge between two genes indicates that they are both associated with at least one common disease. This projection enables the identification of gene communities and allows us to investigate the structural properties and connectivity patterns emerging from shared pathological associations.

3 Network Characterization

The gene-gene network consists of 9,334 nodes and 10,146,630 edges, with a density of 0.2329. The minimum degree is 1, the maximum is 7,452 and the average degree is 2,174. This indicates a heterogeneous topology with a few highly connected *hub* genes and many nodes with lower connectivity, a pattern typical of scale-free and complex biological networks where hubs often correspond to pleiotropic genes[3] implicated in multiple diseases.

The network comprises three connected components. The first component links the genes *UBE2U* and *ROR1* via *bacterial meningitis*[4]. The second component includes *CST3*, *CST4*, *CST9* and *CST9L*, forming a 4–clique connected through DOID:3394 (*myocardial ischemia*)[5]. The remaining component is a giant connected component (GCC) encompassing most nodes.

The GCC has a diameter of 5, an average shortest path length of 1.789 and a global clustering coefficient of 0.85. These measures reveal a small-world structure with high local clustering and efficient connectivity. This implies that genes are organized in tightly connected clusters, yet a few hub genes provide shortcuts across the network, facilitating rapid spread of disease associations. Such a topology suggests both modularity and hub dominance, highlighting genes that may play central roles in multiple disease pathways.

3.1 Comparision with ER

We compared the gene-gene network with an Erdős–Rényi (ER) random graph having the same number of nodes and edges.

Several key differences are evident:

- **Degree distribution:** The ER graph has a very narrow degree range (min 2,025, max 2,328), indicative of a homogeneous connectivity pattern. In contrast, the gene-gene network is highly heterogeneous, with hubs reaching degree 7,452 and many low-degree nodes, highlighting its scale-free nature.
- **Clustering:** The ER graph has a low global clustering coefficient (0.2329), roughly equal to its density, whereas the GCC of the real network shows a high clustering (0.85), demonstrating the presence of tightly connected local modules, a feature absent in the ER model.
- **Small-world structure:** Both networks have similar average shortest path lengths (real: 1.789, ER: 1.767) and small diameters, indicating that short paths are not unusual in either case. However, in the real network, short paths coexist with high clustering and hubs, characteristic of a small-world topology, whereas in the ER network, short paths arise from its homogeneous randomness and may do not imply modularity.

In summary, while both networks are well connected and allow short paths, the real gene-gene network differs strongly from a random graph by exhibiting high clustering and heterogeneous degree distribution. These features underline the biological relevance of hubs and tightly connected gene modules in disease associations.

[3] **Gene pleiotropy** refers to the phenomenon whereby a single gene influences multiple distinct phenotypic traits or disease processes, reflecting the gene's involvement in diverse biological pathways.

[4] Bacterial meningitis is a severe infectious disease characterized by inflammation of the meninges, most commonly caused by *Neisseria meningitidis* or *Streptococcus pneumoniae* [4].

[5] Myocardial ischemia (DOID:3394) is a cardiovascular disease caused by reduced blood flow to the heart muscle due to narrowed coronary arteries [5].

3.2 Comparision with BA

We further compared the gene-gene network with a Barabási–Albert (BA) scale-free graph having the same number of nodes and edges.

The comparison reveals different observations:

- **Degree distribution:** Similar to the real network, the BA graph shows a heterogeneous degree distribution with a few hubs and many low-degree nodes. However, the real network exhibits even higher-degree hubs (max 7,452) and a wider range, highlighting stronger hub dominance.
- **Clustering:** The BA graph has a moderate clustering coefficient (0.2975), higher than its density (0.2058), indicating some local cohesion. Nevertheless, the real network's clustering in the GCC (0.85) is substantially higher, reflecting stronger modular organization and tightly connected gene communities.
- **Small-world structure:** Both networks have very short average path lengths (real: 1.789, BA: 1.794) and small diameters, consistent with small-world behavior.

Overall, the comparison shows that while the BA model captures the heterogeneous degree distribution of the real network, it underestimates clustering and modularity. The real gene-gene network thus exhibits a combination of scale-free topology, strong small-world characteristics, and modular organization, all of which are biologically meaningful for disease association patterns.
Therefore, the real gene-gene network can be characterized as a *scale-free and small-world* network. Its structure is dominated by high-degree hub genes that connect dense local modules, suggesting biologically meaningful organization in terms of disease associations. Both ER and BA models capture some aspects of connectivity, but only partially reproduce the clustering and hub prominence observed in the empirical network. In Table 1, a summary of the results.

Table 1: Network metrics for gene-disease (GG), Erdős–Rényi (ER) and Barabási–Albert (BA) graphs.

Metric	GG	ER	BA
Nodes	9 334	9 334	9 334
Edges	10 146 630	10 146 630	10 146 630
Density	0.233	0.233	0.206
Max deg	7 452	2 328	5 118
Min deg	1	2 025	1 087
Avg deg	2 174	2 173	1 921
# Components	3	1	1
Diameter (GCC)	5	2	2
Glob. Clust. Coeff. (GCC)	0.85	0.233	0.298
Avg Shortest Path (GCC)	1.789	1.767	1.794

4 Task 1: Community detection

Community detection identifies densely connected subgroups within networks, revealing modular structures that reflect functional organization. In gene-gene networks, communities cluster genes with similar disease association patterns, potentially indicating shared biological pathways or coordinated roles in pathogenesis. This approach reduces network complexity and facilitates the identification of functionally coherent gene modules for biological interpretation.

4.1 Application on the Gene–Gene Network

To identify clusters of functionally related genes in the gene–gene association network, multiple community detection algorithms were implemented and evaluated. The bipartite network, composed of 9,584 nodes and 52,769 edges, was projected onto the gene space, resulting in a unipartite network of 9,334 genes connected by 10,146,630 edges. Three algorithms, *Label Propagation*, *Infomap* and *Weighted Louvain*, were applied to this projected gene network using the `cdlib.algorithms` and `cdlib.evaluation` [6] modules in Python to identify gene communities based on shared disease associations.

4.1.1 Label Propagation Algorithm. Applied to the projected gene network, LPA detected 10 communities. The results show an extremely high coverage (0.9999), meaning that almost all edges are contained within communities and a high average internal density (0.8944). However, the modularity value (0.00018) indicates that the division of the network into communities provides minimal improvement over a random partition. This suggests that while LPA produced dense clusters, it may have merged large regions of the graph into few broad groups, limiting its capacity to uncover modular biological structures.

4.1.2 Infomap Algorithm. When applied to the same network, Infomap identified 20 communities, twice as many as LPA. The modularity (0.00053), though still low, was slightly higher than that of LPA, while the coverage (0.9993) remained very high and the average internal density (0.9476) indicated compact community structures. These results imply that Infomap succeeded in creating more internally cohesive clusters compared to LPA, but with a much greater computational cost. Nonetheless, the extremely low modularity once again reflects the highly interconnected nature of the gene–gene network, where dense associations reduce the visibility of distinct modular boundaries.

4.1.3 Weighted Louvain Algorithm. The Weighted Louvain method is a modularity-based hierarchical algorithm designed to optimize community structure iteratively while accounting for edge weights. The weighted projection of the gene network contained the same 9,334 nodes but included weighted edges that reflected the strength of co-occurrence between genes across shared diseases. The Louvain algorithm detected a larger number of communities with improved modular characteristics. The computed modularity (0.2603) indicates a substantially stronger community structure compared to both LPA and Infomap, revealing meaningful modular patterns in the weighted network. The coverage (0.6822) decreased relative to the unweighted methods, as expected, since stronger weighting emphasizes well-defined intra-community connectivity while excluding weaker inter-community links. Moreover, the average internal density (0.8172) suggests compact and internally cohesive clusters. The average edge weight across communities confirmed that intra-community links tended to represent more substantial gene–disease associations. This indicates that incorporating edge weights into the Louvain framework significantly enhanced the

algorithm's sensitivity to the intensity of gene co-occurrence, yielding more biologically interpretable results.

While Label Propagation and Infomap achieved nearly complete edge coverage, this came at the expense of modular resolution, as most edges fell within very large communities. Louvain's lower coverage but higher modularity represents a better balance between cohesion and separation, typical of biologically relevant modular structures. The results of the comparison are shown in Table 2.

Table 2: Comparison of community detection algorithms

Algorithm	Communities	Modularity	Coverage	Avg. Density	Runtime (s)
Label Propagation	10	0.00018	0.9999	0.8944	8.06
Infomap	20	0.00053	0.9993	0.9476	510.84
Louvain (weighted)	9	0.2603	0.6822	0.8172	341.89

In conclusion, the Weighted Louvain algorithm outperformed the other methods in revealing meaningful modular organization in the gene–gene network.

4.2 Community Structure

Given the superior performance of the Weighted Louvain algorithm in terms of modularity and structural coherence, subsequent analyses focused on the communities detected by this method. The Louvain algorithm revealed nine gene communities of varying sizes, ranging from large clusters of several thousand genes to small, tightly connected groups. The largest communities (1–3) contained the majority of the network's nodes and exhibited progressively higher average edge weights, indicating stronger intra-community associations and potentially greater biological relevance. In contrast, the smaller communities (4–9) consisted of very few genes, each connected with low-weight edges (average ≈ 1.0), suggesting specialized or isolated gene modules. The overall distribution of the communities is summarized in Table 3.

Table 3: Structural metrics for the communities detected by the Weighted Louvain algorithm

Community	Genes	Edges	Avg. Degree	Conductance	Internal Density
1	3,464	793,368	784.45	0.6088	0.1323
2	3,114	2,524,685	3424.76	0.3312	0.5209
3	2,690	3,603,385	10617.04	0.3023	0.9963
4	29	406	28.00	0.2632	1.0000
5	20	134	13.50	0.0625	0.7053
6	7	21	6.00	0.1429	1.0000
7	4	6	3.00	0.5000	1.0000
8	4	6	3.00	0.0000	1.0000
9	2	1	1.00	0.0000	1.0000

The three largest communities encompass more than 95% of the network's nodes, indicating the presence of extensive gene modules linked by multiple shared diseases. From a structural perspective, Communities 1–3 display high average degrees and relatively low conductance values, suggesting strong internal connectivity and limited interaction with other parts of the network. Among them, Community 3 shows an exceptionally high average degree (10,617.04) and an internal density close to 1, denoting an almost fully connected subnetwork. The smaller communities (4–9) show very low conductance and internal densities close to 1, reflecting small, self-contained gene modules likely representing highly specific functional processes.

5 Enrichment Analysis

To investigate the biological significance of the detected communities, an enrichment analysis was conducted on the gene sets identified by the *Weighted Louvain algorithm*. The analysis aimed to identify over-represented pathways, biological processes and disease associations within each community.
The enrichment procedure was implemented in Python using the *gseapy* package [7], through the `gp.enrichr` function. Each gene list was tested against three reference libraries: **KEGG_2021_Human** [8], **GO_Biological_Process_2021** [9] and **DisGeNET** [10]. The main steps of the procedure were as follows:

(1) **Community selection and pre-filtering.** Genes were grouped according to the communities returned by the Weighted Louvain algorithm. Communities containing fewer than three genes were excluded from enrichment to avoid unstable statistics.
(2) **Statistical testing.** For each community and each reference library, enrichment was computed using Enrichr's implementation of a Fisher's exact hypergeometric test [11]. This evaluates whether the overlap between the community gene list and a reference term is greater than expected by chance.
(3) **Overlap representation.** Results are expressed as k/M, where k is the number of community genes present in a given term and M is the total number of genes annotated to that term.
(4) **Multiple testing correction.** Raw p-values were adjusted using the Benjamini–Hochberg procedure [12] to control the False Discovery Rate (FDR). Terms with an adjusted p-value < 0.05 were considered statistically significant.
(5) **Score transformation.** To facilitate visualization, adjusted p-values were converted into significance scores using:

$$\text{score} = -\log_{10}(\text{adjusted p-value}),$$

such that higher scores indicate stronger evidence of enrichment (note that an adjusted p-value of 0.05 corresponds approximately to a score of 1.30).
(6) **Aggregation and visualization.** Significant terms were aggregated into a Community × Term matrix, where each cell contains the maximum score observed for that community-term pair. This matrix was visualized as a heatmap, with color intensity proportional to $-\log_{10}(\text{adj p-value})$.

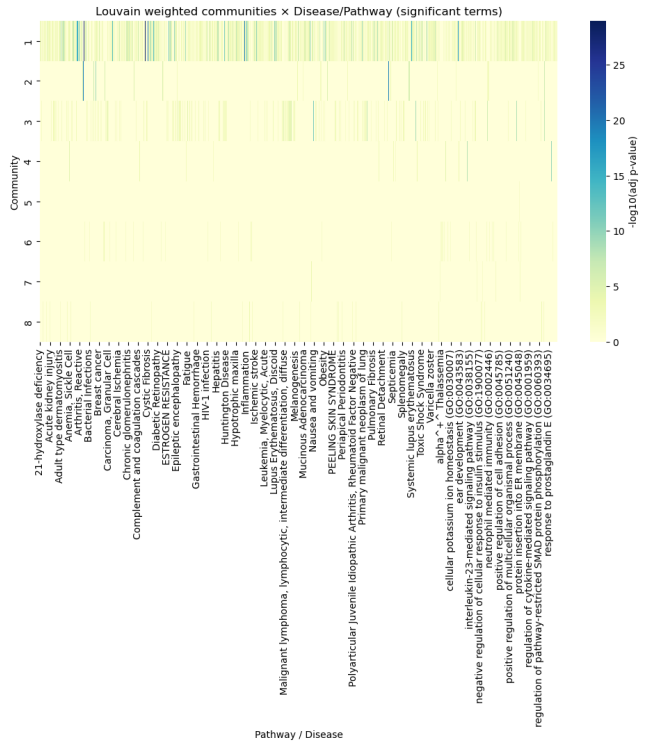


Figure 1: Heatmap of significant enrichment terms across gene communities. Color intensity represents $-\log_{10}$(adjusted p-value). Only significant terms (adjusted p-value < 0.05) are shown.

The global heatmap representation of the Community × Term matrix (constructed from $-\log_{10}$(adj p-value) values) 1 allows rapid visualization of shared versus community-specific enrichment patterns whereas Table 4 reports the most significant enriched term for each weighted Louvain community, together with the source database, the adjusted p-value and the transformed significance score.

Table 4: Top enriched term per weighted Louvain community (terms with adj. p-value < 0.05).
Score **$= -\log_{10}$(adjusted p-value).**

Community	Top enriched term	Source	Adj. p-value	Score
1	Lupus Erythematosus, Systemic	DisGeNET	1.05×10^{-29}	28.98
2	Intelligence	DisGeNET	4.25×10^{-26}	25.37
3	Protein C antigen measurement	DisGeNET	2.36×10^{-16}	15.63
4	Olfactory transduction	KEGG_2021_Human	1.68×10^{-18}	17.77
5	RNA splicing (GO:0008380)	GO_Biological_Process_2021	1.54×10^{-2}	1.81
6	CYSTIC FIBROSIS MODIFIER 1	DisGeNET	2.41×10^{-4}	3.62
7	Severe myopia	DisGeNET	2.75×10^{-5}	4.56
8	Salivary secretion	KEGG_2021_Human	1.28×10^{-4}	3.89

Communities 1–3 exhibited highly significant enrichments (scores > 15), corresponding mainly to immune, neurological and coagulation-related terms. These results are consistent with the large size and high internal connectivity of these communities. Smaller communities (4–8) showed more specific functional signals, such as olfactory transduction, RNA splicing, cystic fibrosis modifiers and myopia. These results suggest that even the smallest clusters correspond to biologically coherent submodules.

5.0.1 Interpretation and results . The enrichment profiles revealed distinct biological and pathological themes across the weighted Louvain communities.

- **Community 1** displayed strong enrichment for autoimmune disorders such as *Systemic Lupus Erythematosus*, *Psoriasis*, *Rheumatoid Arthritis* and *Crohn's Disease*, with extremely low adjusted p-values (adj. $p < 10^{-26}$). This suggests that the corresponding gene cluster comprises immune-regulatory genes and inflammatory mediators commonly involved in systemic autoimmune and inflammatory pathways.
- **Community 2** was primarily associated with neuropsychiatric and cognitive traits, including *Intelligence*, *Schizophrenia* and *Autism Spectrum Disorders*, as well as body mass index–related traits. These results indicate that this gene module may reflect neuronal development, brain function and metabolic regulation mechanisms.
- **Community 3** showed significant associations with cancer-related and coagulation-related phenotypes, such as *Protein C measurement*, *Nasopharyngeal Neoplasms*, *Testicular Germ Cell Tumors* and *Breast Carcinoma*. The enrichment pattern suggests involvement of genes linked to tumorigenesis, vascular biology and hemostatic balance.
- **Community 4** was dominated by sensory and oxidative processes, with enrichment in *Olfactory transduction* and *hydrogen peroxide metabolic and catabolic processes*. This cluster likely represents genes involved in chemosensory perception and oxidative stress responses in sensory tissues.
- **Community 5** presented moderate enrichment for RNA-processing mechanisms, particularly *RNA splicing* and *protein insertion into the endoplasmic reticulum membrane*, pointing to genes that play structural and regulatory roles in post-transcriptional modification and protein biogenesis.
- **Community 6** was enriched for cystic fibrosis–related and intestinal transport processes (*CYSTIC FIBROSIS MODIFIER 1*, *Meconium ileus*, *Intestinal obstruction*), consistent with genes associated with ion transport, mucus production and epithelial function.
- **Community 7** included genes linked to ocular morphology and chromosome organization, with enrichment in *Severe myopia* and *protein localization to centromeric and kinetochore regions*, suggesting a relationship between cell division mechanisms and degenerative eye disorders.
- **Community 8** was associated with *Salivary secretion*, *Intersex conditions* and regulation of proteolytic and protein metabolic processes, suggesting functional involvement of genes in epithelial secretion and hormonal differentiation pathways.

Overall, these enrichment profiles confirm that the Louvain-derived communities capture biologically coherent modules, each corresponding to distinct physiological systems or disease domains — immune, neurological, oncogenic, sensory or epithelial — thereby

validating the structural and functional consistency of the detected network communities.

6 Task 2: Higher-order Network Analysis

In this task, we extend our analysis by exploring the structure of the system from a higher-order perspective. Starting from the gene–disease associations previously collected, we model the data as a hypergraph in which each node corresponds to a disease and each hyperedge represents a gene linking all the diseases it is associated with. This formulation allows us to analyze the network from a complementary point of view: instead of focusing on how genes share diseases (as in the bipartite or projected graph), we study how diseases are co-associated through shared genes. In this way, we can capture collective patterns of co-occurrence that are not visible in a pairwise representation and identify diseases that frequently appear together across multiple genes.

Using the *HyperNetX* Python library, we constructed the hypergraph directly from the gene–disease dataset, defining diseases as nodes and genes as hyperedges. The hypergraph displayed three connected components: two small components with 2 and 4 hyperedges, respectively, and a large giant component containing all remaining nodes and hyperedges. The resulting structure is characterized by a very low *aspect ratio* ($r = 0.027$), the ratio between the number of nodes and the number of hyperedges and a density of 0.023. These low values indicate a sparse system in which many nodes (diseases) participate in numerous distinct hyperedges.

To reduce redundancy and obtain a more compact representation, we applied the same collapsing function used for diseases to merge genes with functional or semantic overlap. This process yielded a collapsed hypergraph composed of 250 nodes and 2,203 hyperedges. A comparative summary of the main structural parameters between the original and collapsed hypergraphs is reported in Table 5.

Table 5: Comparison between the original and collapsed hypergraphs.

Metric	Original	Collapsed
Nodes (Diseases)		
Average degree	211.0	82.4
Minimum degree	1	1
Maximum degree	3168	1086
Hyperedges (Genes)		
Average size	5.7	9.4
Minimum size	1	1
Maximum size	81	81
Global Metrics		
Aspect ratio (r)	0.027	0.113
Density	0.023	0.037

The comparison between the original and collapsed hypergraphs highlights the effect of the collapsing procedure on network structure. The average degree of diseases and the maximum degree both decrease markedly, reflecting the removal of redundant or overlapping disease entities. Conversely, the average hyperedge size increases, indicating that the remaining genes are associated with a broader set of distinct diseases. At the global level, both the aspect ratio (r) and density increase slightly, which is expected after the elimination of duplicated nodes and edges. Overall, the collapsed representation yields a denser yet more interpretable network, offering a cleaner and more biologically meaningful basis for higher-order analysis.

6.1 Top 20 hub diseases by number of associated genes

To better understand the topological impact of the collapsing procedure, we compared the most connected disease nodes (i.e., hub diseases) in both the original and collapsed hypergraph representations.

Table 6 reports the top 20 hub diseases, showing the number of associated genes in the non-collapsed and collapsed hypergraphs, respectively. These results highlight biologically meaningful hubs emerging from aggregated associations.

Table 6: Top 20 hub diseases by number of associated genes, before and after the collapsing procedure.

Disease	Non-collapsed	Collapsed
cancer	3168	1086
organ system cancer	3143	1079
cell type cancer	2292	798
carcinoma	2173	758
developmental disorder of mental health	1768	572
reproductive organ cancer	1556	492
sleep disorder	1823	490
arthritis	1020	445
cognitive disorder	1134	415
specific developmental disorder	1311	415
male reproductive organ cancer	1353	399
attention deficit hyperactivity disorder	1239	387
mood disorder	1039	374
prostate cancer	1216	359
Depressive disorder	1001	357
glaucoma	797	286
open-angle glaucoma	780	280
autism spectrum disorder	670	269
rheumatoid arthritis	574	259
colitis	577	247

As shown in Table 6, the collapsing procedure substantially reduced the absolute number of gene connections per disease, as expected from the merging of semantically equivalent gene entries. Nevertheless, the ranking of the most connected diseases remains highly consistent between the two representations. Generic oncological categories, such as *cancer*, *organ system cancer*, *cell type cancer* and *carcinoma*, dominate both lists, confirming that cancer-related processes are characterized by extensive genetic pleiotropy.

The persistence of these categories across both scales indicates that oncological mechanisms remain central even after normalization and redundancy reduction.
Beyond cancer, several neurodevelopmental and psychiatric conditions, such as *developmental disorder of mental health*, *autism spectrum disorder* and *attention deficit hyperactivity disorder*, retain relatively high connectivity, suggesting shared genetic factors across different cognitive and behavioral disorders.
Overall, the predominance of generic oncology-related terms suggests that tumorigenic processes are widely shared across the collapsed gene set, although other categories such as cognitive and developmental disorders remain highly central within the network, highlighting the multifactorial nature of gene involvement across diverse disease classes.

6.2 s-Subgraph Decomposition of the Collapsed Hypergraph

To investigate the structure of the collapsed hypergraph at different scales, we constructed s-subgraphs, retaining only genes that connect at least s disease nodes. This approach allows us to filter out low-degree associations and focus on stronger and more informative subnetworks. As s increases, smaller and more interpretable components emerge, while the initially large, highly connected component gradually fragments, revealing core patterns of gene–disease connectivity. Table 7 presents the number of components for different s values.

Table 7: Overview of s-subgraph components in the collapsed hypergraph. For each threshold s, the total number of components.

s	N. Components
10	13
20	9
30	8
40	4
50	2
60	1
70	1
80	0

For low thresholds ($s < 30$), a giant component dominates, containing a large fraction of nodes and hyperedges. This indicates that many genes are connected to multiple diseases, forming a highly interconnected core of shared associations. Multiple smaller components coexist, representing more specialized sets of diseases connected by fewer genes.
From $s = 30$ onward, the giant component fragments and the remaining components result smaller, more focused and easier to analyze. These subnetworks likely highlight strong, functionally coherent gene-disease associations, since only genes linked to many diseases are retained.
At very high thresholds ($s \geq 60$), only one or two compact components remain, representing the most central and highly pleiotropic genes and their key disease connections.

Taken together, the s-subgraph analysis shows that increasing the threshold gradually simplifies the network, reducing broad overlaps and revealing smaller, more coherent groups of genes and diseases. This helps to identify the core structure of the network and the most relevant gene–disease connections.

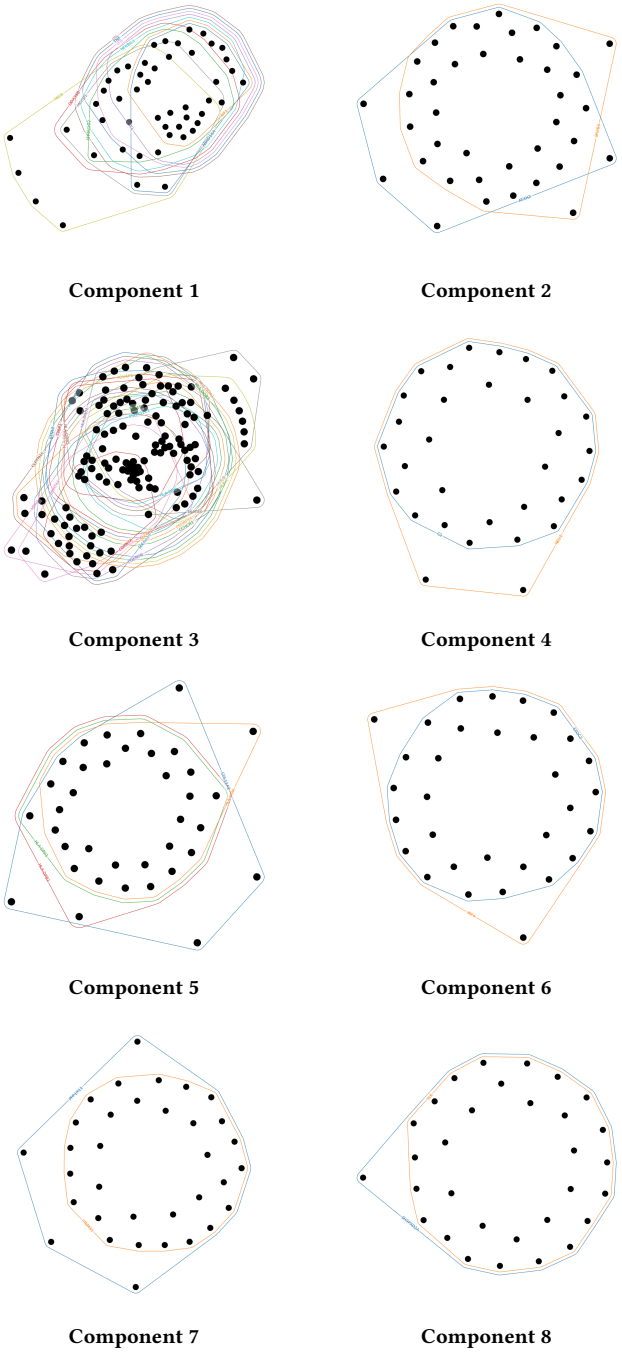


Figure 2: Visualization of the 8 s-subgraph components for $s = 30$, arranged in 2 images per row within a single column.

Figure 2 shows the 8 s-subgraph components for $s = 30$. To improve readability, node labels are omitted, allowing the structural patterns to be more clearly observed.
Components 1 and 3 are the largest, involving many genes and diseases. These components correspond to highly pleiotropic genes connected to multiple disease classes, forming dense substructures that reflect widespread shared genetic associations.
The remaining components are smaller, involving 2 to 4 genes and represent more specific gene–disease modules. These subnetworks highlight focused interactions where a few genes are strongly associated with a limited set of diseases, offering interpretable and biologically coherent patterns.

Now, let us focus on Component 3 of the s-subgraph with threshold $s = 30$, the largest and most connected subnetwork. It features many highly connected genes from the HLA family, which encode major histocompatibility complex (MHC) proteins crucial for immune system recognition [13]. These genes act as central hubs, linking a wide variety of conditions. The associated diseases include cancers such as *cervical cancer*, *lung cancer* and *B-cell lymphoma*, autoimmune disorders like *vitiligo*, *autoimmune thyroiditis* and *rheumatoid arthritis*, and neurological or developmental disorders including *autism spectrum disorder* and *cognitive disorder*. Some genes, like `BTNL2` and `C6ORF15`, are highly pleiotropic, connecting to many diseases, while others show more specific associations, reflecting a hierarchical pattern of genetic influence.

On the other hand, Component 1 is another highly connected subnetwork but with a different profile. It also links cancers (*lung*, *cervical*, *uterine*), autoimmune and inflammatory diseases (*rheumatoid arthritis*, *autoimmune thyroiditis*, *vasculitis*), and neurodevelopmental conditions (*autism spectrum disorder*, *developmental disorders of mental health*). Highly pleiotropic genes such as `ABHD16A`, `AIF1`, and `TNF` form central hubs, similar to the HLA genes in Component 3. In addition, Component 1 includes metabolic, autoimmune, and sensory conditions (*age-related macular degeneration*, *sleep disorder*, *abdominal obesity-metabolic syndrome*), which are less represented in Component 3, suggesting that Component 1 complements Component 3 by covering additional physiological systems and disease types.

Overall, comparing Components 1 and 3 shows that the collapsed hypergraph organizes genes into overlapping but distinct pleiotropic cores. Both highlight genes central to multiple diseases, but each captures different aspects of gene–disease connectivity, reinforcing that complex disorders arise from interconnected genetic pathways.

In conclusion, the s-subgraph analysis shows that increasing the threshold simplifies the network, highlighting core, pleiotropic genes in large components and more specific gene–disease associations in smaller subnetworks.

7 Task 3: Open Question: which diseases are most influenced by indirect genetic relationships?

In this section, we aim to identify the diseases that are most influenced by **indirect genetic relationships**. Specifically, we focus on second order connections among diseases, defined as paths of the form:

$$\text{Disease}_1 \rightarrow \text{Gene}_A \rightarrow \text{Disease}_{\text{bridge}} \rightarrow \text{Gene}_B \rightarrow \text{Disease}_2 \quad (1)$$

These correspond to situations in which two diseases do not share any gene directly, but are connected through pairs of genes that interact across diseases. In other words, for a valid second order path between Disease_1 and Disease_2, there must exist at least one **bridge disease** connected to both via distinct genes, while no direct gene α exists such that the path $\text{Disease}_1 - \text{Gene}_\alpha - \text{Disease}_2$ exists.
This concept allows us to explore how diseases can be connected through multiple genetic steps, revealing indirect similarities that are not visible when considering only direct gene–disease associations.
We further refined our dataset by removing general disease categories, like *cancer*, *organ system cancer*, *cell type cancer*, *carcinoma* and *syndrome*. These generic labels would increase the network connectivity without providing meaningful insights. After this filtering step, only well-defined pathological entities were retained, resulting in a final dataset composed of 9,334 genes, 245 diseases and 41,517 associations.
We built the bipartite network and derived the disease–disease projection, where two diseases are connected if they share at least one gene. The edge weight between two diseases d_i, d_j is computed as the average of their gene–disease association scores.

$$w(d_i, d_j) = \frac{1}{|G_{ij}|} \sum_{g \in G_{ij}} \frac{w(d_i, g) + w(d_j, g)}{2}$$

where G_{ij} is the set of genes shared by diseases d_i and d_j and $w(d_i, g)$ represents the score of the association between disease d_i and gene g.
Starting from this projected graph, we searched for all valid second–order paths of the form (1).
For each pair $(\text{Disease}_1, \text{Disease}_2)$, we excluded cases where the two diseases shared any gene directly. Only paths mediated by a distinct bridge disease were retained.
For each valid triplet $(\text{Disease}_1, \text{Disease}_{bridge}, \text{Disease}_2)$, we computed the average weight of the two gene–disease associations, denoted as $\overline{w}$ and the total number of possible gene–gene pairs along the path, n_{pairs}. Subsequently, all triplets were grouped by disease pair $(\text{Disease}_1, \text{Disease}_2)$. Within each group, we calculated the mean of the average weights across all distinct bridge diseases and summed the corresponding gene–gene pairs. In addition, we counted the number of unique bridges n_{bridges} connecting each disease pair. This aggregation step allowed us to capture, for every disease pair, both the biological strength and the topological richness of their indirect genetic relationships. To combine these aspects into a unique balanced score, we defined a composite relevance metric, defined as:

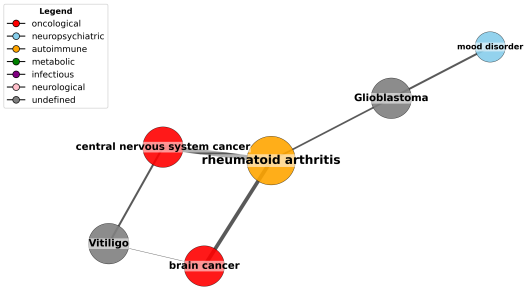
$$\text{Score} = (\overline{w}^{\alpha}) \cdot \left(\log_{10}(1 + n_{\text{pairs}})^{\beta}\right) \cdot \left(n_{\text{bridges}}^{\gamma}\right)$$

where α, β and γ determine the relative importance of the three components: the average strength of gene–disease associations ($\overline{w}$), the number of gene–gene connections (n_{pairs}) and the number of bridge diseases (n_{bridges}). In our analysis, we set $\alpha = 1$, $\beta = 1$ and $\gamma = 0.5$, giving equal weight to biological strength and gene connectivity, while slightly down-weighting the influence of multiple bridge diseases. All disease pairs were finally ranked according to this score to identify those most affected by indirect genetic relationships. We focused on the top 50.

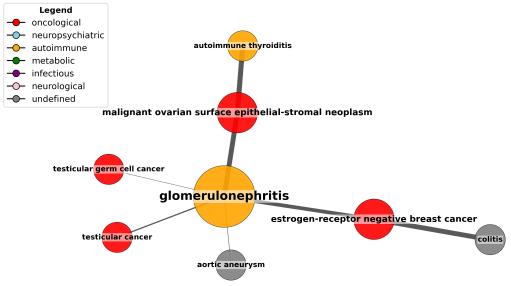
For each disease in these top connections, we retrieved its general biomedical category by querying the Wikidata knowledge base using SPARQL[6]. The query extracted the superclass of each disease, which was then processed to assign each disease to one of the following broad categories: *oncological*, *neuropsychiatric*, *autoimmune*, *metabolic*, *infectious*, *neurological* or, lastly, *undefined*.

This classification allowed us to identify cases in which two diseases connected through an indirect genetic relationship belong to different domains (e.g., an autoimmune disease linked to a neurological one). Thus, finally, we filtered only those disease pairs with differing categories, obtaining the most biologically heterogeneous indirect associations.

The resulting graph revealed four connected components, each corresponding to a distinct cluster of cross–category interactions. These components are shown in Figure 3.

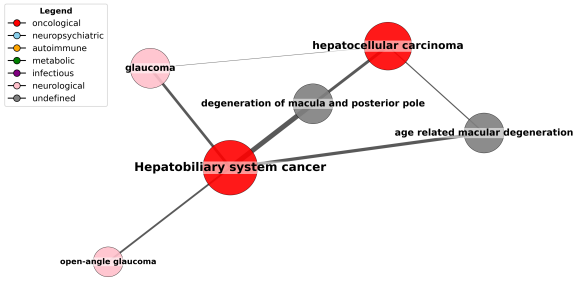


Component 1

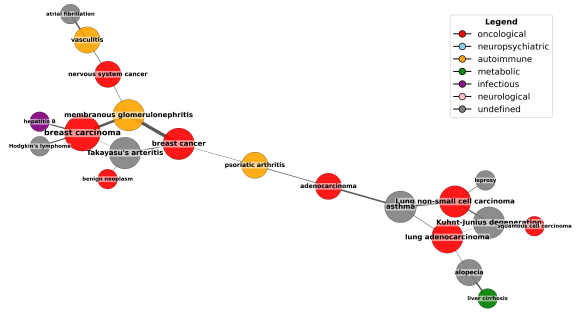


Component 2

[6]https://query.wikidata.org/



Component 3



Component 4

Figure 3: Visualization of the 4 components of the cross-category interactions among diseases.

8 Discussion

The four connected components highlight different structural patterns in the network of indirect genetic relationships among diseases. **Component 1** is centered around *rheumatoid arthritis*, which acts as a highly connected node bridging diseases from oncological and neuropsychiatric domains. **Component 2**, dominated by *glomerulonephritis*, shows strong connections between autoimmune and oncological nodes. **Component 3** presents a structure which links hepatic cancers to disorders categorized as neurological or undefined, suggesting a cross–domain interaction pattern. Finally, **Component 4** contains the largest and most heterogeneous group, where autoimmune, oncological and infectious diseases are mixed within the same network region, indicating a complex multi–domain structure.

Overall, the most frequent cross category associations involve *autoimmune* and *oncological* diseases. Both categories tend to include diseases with a large number of associated genes. From a network perspective, these types of disease may reflect the presence of complex gene interaction patterns and overlapping biological processes that make them central in the overall structure of the graph. From a biomedical point of view, both categories are closely linked to immune system regulation; for this reason, they may share several genes related to immune signaling and cell regulation [14], which explains their frequent co-occurrence as connected nodes in the network.

Among all disease nodes, the most frequent bridge diseases (those

appearing most often as intermediaries between other pairs) include *sleep disorder*, *developmental disorder of mental health*, *autism spectrum disorder*, *cognitive disorder* and several forms of cancer such as *gastrointestinal system cancer*, *prostate cancer* and *basal cell carcinoma*. These diseases tend to link conditions from distinct biomedical domains. At the gene level, the most frequent gene pairs contributing to indirect connections are `(DMRTA1, HLA-DRB1)` and `(DMRTA1, HLA-DQA1)`, suggesting that these three genes play a key role in bridging distant disease categories.

References

[1] I. Diamant, D. J. B. Clarke, J. E. Evangelista, N. Lingam, and A. Ma'ayan. Harmonizome 3.0: integrated knowledge about genes and proteins from diverse multi-omics resources. *Nucleic Acids Research*, 53(1):D1016–D1028, 2024.

[2] Dhouha Grissa, Alexander Junge, Tudor I. Oprea, and Lars Juhl Jensen. Diseases 2.0: a weekly updated database of disease–gene associations from text mining and data integration. *Database*, 2022:baac019, 2022.

[3] David Smith, Austin Benson, Bruce Hendrickson, Robert Leland, and et al. Hypernetx: A python library for analysis of hypergraphs. *Journal of Open Source Software*, 5(52):2174, 2020.

[4] World Health Organization. Meningitis. *WHO Fact Sheets*, 2024.

[5] Mayo Clinic Staff. Myocardial ischemia - symptoms and causes. *Mayo Clinic*, 2024.

[6] Giulio Rossetti, Letizia Milli, Salvatore Rinzivillo, Dino Pedreschi, Fosca Giannotti, and Roberto Interdonato. Cdlib: a python library to extract, compare and evaluate communities from complex networks. *Applied Network Science*, 4(1):52, 2019.

[7] Zhuoqing Fang, Xinyuan Liu, and Gary Peltz. Gseapy: a comprehensive package for performing gene set enrichment analysis in python. *Bioinformatics*, 2022. btac757.

[8] Minoru Kanehisa, Masahiro Furumichi, Yoko Sato, Michihiro Ishiguro-Watanabe, and Masumi Tanabe. Kegg: integrating viruses and cellular organisms. *Nucleic Acids Research*, 2021.

[9] The Gene Ontology Consortium. Gene ontology resource: enriching a gold mine. *Nucleic Acids Research*, 49(D1):D325–D334, 2021.

[10] Jordi Piñero, J. Marc Ramírez-Anguita, Joaquí Saüch, Félix Ronzano, Elisa Centeno, Ferran Sanz, and Laura I. Furlong. Disgenet: a comprehensive platform integrating information on human disease-associated genes and variants. *Nucleic Acids Research*, 45(D1):D833–D839, 2016.

[11] Ronald A. Fisher. On the interpretation of χ^2 from contingency tables, and the calculation of p. *Journal of the Royal Statistical Society*, 85(1):87–94, 1922.

[12] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300, 1995.

[13] Peter Parham. *The Immune System*. Garland Science, New York, 3rd edition, 2005.

[14] D. Bernal-Bello et al. Cancer risk in autoimmune and immune-mediated diseases. *Journal of Clinical Medicine*, 2025.