Bioinformatics and Network Medicine Project

Aur Marina Iuliana 1809715
Petrucci Ilaria 1732987

Group 10 - Fibrosing Alveolitis C4721507

15 January 2024

Contents

1 Abstract

Fibrosing Alveolitis, a chronic and progressive interstitial lung disease, predominantly affects older adults and is characterized by a rapid decline in lung function and a poor prognosis. The pathogenesis of IPF involves repeated alveolar injuries in genetically susceptible individuals, leading to fibroblast activation and excessive collagen accumulation, which impairs lung function. The only effective solution for this disease is lung transplant, while more easily available pharmacological treatments offer limited efficacy, emphasizing the need for novel therapeutic strategies. Our study explores computational approaches to identify new disease genes as potential drug targets for IPF.

Starting from known disease genes, we applied various algorithms on the *Human Protein-Protein Interaction Network*, based on different approaches, to discover putative disease genes; we then investigated at the molecular level the processes in which they were mainly involved and attempted to identify effective drugs targeting them via database explorations. Additionally, we examined the presence of active clinical trials focusing on the identified drugs to better characterize the treatment options.

2 Introduction

Fibrosing Alveolitis, better known as Idiopathic Pulmonary Fibrosis (IPF) consists of a a chronic, progressive interstitial lung disease (ILD) that leads to the progressive worsening of dyspnea and lung function in the affected patients. It is associated to aging, since is prevalent in older adults and has a poor prognosis with, if untreated, an average life expectancy of 3–5 years after diagnosis.

The pathogenesis of the disease is not completely understood; however, it is established that it often occurs in genetically susceptible alveolar epitheliums and involves a pattern of repeated alveolar injuries followed by an impaired ability of the epithelium to undergo proper re-epithelialization and repair; other risk factors besides genetic predisposition are cigarette smoking and viral infections. At tissue level it is known that alveolar cells are activated and secrete various cytokines and growth factors able to attract and stimulate lung fibroblasts to transform into myofibroblasts; this conversion results in an accumulation of collagen, that alters the phisiology of lungs, causing scarring and an irreversible decline in their function.

Nowadays, the only real cure for the disease is lung transplant which is known to be a procedure subject to the availability of organs and compatibility; moreover an eventual transplant for IPF patients is made more difficult by the fact that there are often co-morbidities and the age of patients is usually high.

There exists some drug treatments available, mainly based on two compounds, namely *pirfenidone* and *nintedanib*. The first is capable to inhibit collagen production and fibroblast proliferation regulating tumor necrosis factor (TNF) pathways, while the latter is a tyrosine kinase inhibitor (TKI) that binds to a family of growth factor receptors and prevents the proliferation of fibroblasts. Nevertheless these drugs are not able to heavily slow down the disease course and improve in a significant way the life quality and expectancy of patients; thus the identification of new compounds capable of treating the disease is crucial.

A valid approach is based on the identification of new disease genes, which may turn out to be potential targets towards which a drug can act and allow disease remission. From a computational point of view, it is possible to start from known disease genes in order to use them to identify other putative disease genes that can be evaluated as potential targets on which to act.

For this purpose, we validated different approaches, experimenting with the application of various algorithms in order to identify the best performing one and obtain a list of new putative disease genes, with the aim to better characterize at a molecular level the pathogenesis of the disease. Following this, we turned our attention to the identification of drugs that could target the identified genes and finally assessed the presence of active clinical trials on them.

3 Materials and Methods

3.1 Data Gathering and Networks Construction

First, we downloaded human protein-protein interaction data from **BioGRID 4.4.228** database and, working in `Python`, we built the interactome with `networkX` library, considering only physical interactions and as organism type *Homo sapiens* (ID: 9606). Then, we removed self and redundant loops, isolating the *Largest Connected Component* of the network, which we then used for the entire study.

To identify known fibrosing alveolitis disease associated genes, we collected them from **DisGeNET**, and created a network by isolating a subgraph from the interactome. Also for this resulting network, whose features can be observed in Table 1, we isolated the *Largest Connected Component* and noted that it was the unique connected component of the network, with all other genes being disconnected nodes (see Figure 1). Following this, we calculated different centrality measures for the seed genes in the obtained LCC, as shown in Table 2.

3.2 Algorithms Comparison

To assess how effectively the chosen algorithms could identify putative disease genes throughout the available data, we implemented a 5-fold cross-validation approach to evaluate their performance. We began by dividing our set of disease genes, denoted as $S0$, into five distinct subsets. For each iteration of the validation, we designated one subset as the probe set PS, which served as our test data, while the remaining four subsets were combined to form the training set TS, used to run the algorithm.

We computed the *Precision* metric, which, in our context, represents the fraction of probe set genes successfully retrieved among all the genes returned by the algorithms (that we chose to be 100).

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}}$$

Then we computed the *Recall* metric, which measures the fraction of probe set genes successfully retrieved over the number of genes in the probe set PS.

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}}$$

Finally, we proceeded to calculate the *F1-score* metric, which is the harmonic mean between precision and recall, representing a balanced trade-off between the two metrics.

$$\text{F1-score} = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$$

By analyzing the algorithm's metrics at different cutoffs, we aimed to gain a comprehensive understanding of its effectiveness in predicting disease-associated genes. More precisely, we assessed the algorithm's performance at different cutoff points: the top 50 genes and the top n genes, where n represented various fractions of the total number of known Gene-Disease Associations (GDAs). These fractions included 1/10, 1/4, 1/2, and the entirety of the known GDAs.

After validating the algorithms, we identified the one exhibiting the best performance and replicated the analysis using the complete set of seed genes, denoted as $S0$. Additionally, we specifically considered the first 100 genes, on which we focused for subsequent analysis.

3.2.1 DIseAse MOdule Detection (DIAMOnD)

The first algorithm we considered was *DIseAse MOdule Detection* (DIAMOnD), which is based on the hypothesis that locally dense neighbourhoods of the interactome (i.e., topological modules) overlap with functional modules but are not able to capture disease modules. Thus, the rationale is to evaluate the significance of protein connections instead of their density, identifying proteins with more connections to seed proteins than expected in a random scenario.

To achieve this, the algorithm computes the probability that a protein with a total of k links has exactly ks links to $s0$ seed proteins using the following formula:

$$p_{k,k_s,k_{s_0}} = \frac{\binom{s_0}{k_s}\binom{N-s_0}{k-k_s}}{\binom{N}{k}}$$

Subsequently, to determine whether a specific protein exhibits a higher connectivity to seed proteins than expected by chance, the connectivity p-value is computed to assess the statistical significance:

$$p\text{-value}(k, k_s) = \sum_{k_i=k_s}^{k} p(k, k_i)$$

3.2.2 DIAMOnD Background Local Expansion (DiaBLE)

Next, we assessed a variant of *DIAMOnD*, still based on the connectivity significance evaluation but introducing an adaptive gene universe (N). In fact, while in the latter approach the size of the universe remained fixed during the execution of the hypergeometric test and corresponded to the entire number of genes present in the interactome, in *DiaBLE* (Diamond Background Local Expansion) it is updated at each iteration to reflect the smaller expansion of the current seed set.

In particular, in the version we implemented by modifying *DIAMOnD* code, we included, with the function `calculate_universe_size_diable`, in the universe N not only seed genes but also the candidate genes identified during the iteration and their first neighbors.

3.2.3 Network Diffusion

Afterward, we tested *network diffusion*, also known as network propagation, using `Cytoscape` software. This method aims to identify network neighborhoods within a larger network that are relevant to the given set of seed nodes, relying on the network interactions.

Heat diffusion can be considered similar to a random walk, where the spread from the initial distribution is influenced by the time parameter t, without being affected by the restart probability.

In this regard, we compared the performance of the algorithm at different t times: in particular at $t = 0.002$, $t = 0.005$ and $t = 0.01$. To perform the five fold cross-validation for this specific case, we manually used the `Cytoscape` software by selecting each time one of the 5 traning sets TS as seed genes.

3.2.4 Community Detection Algorithms

Finally, we explored a different approach for the identification of putative disease genes, based on the disease module hypothesis, i.e. the assumption that topological communities overlap with functional and disease modules and that the disease phenotype can be described by their breakdown. In this context, we compared the outcomes of two algorithms that align with this assumption: *Markov Clustering (MCL)* and *Louvain Algorithm*.

The first relies on the detection of densely connected regions based on random flow, while the latter maximizies a global modularity function. To assess modules enriched in seed genes, we performed a hypergeometric test and identified communities enriched in disease genes by establishing a threshold on the p-value ≤ 0.05, adjusting it with Bonferroni correction for multiplicity, as suggested in Jafari et al. (2019), in order to reduce the overall false positive rate.

3.3 Enrichment Analysis

After acquiring the putative disease genes (see Table 4 and Figure 8), to understand their functions at a biological and molecular level we performed an enrichment analysis and compared results with known disease genes.

As a reference database, we used *GO Biological Processes* to understand the general process in which they were involved; then we moved on to *GO Molecular Function* and *GO Cellular Component*, which provide information respectively on the type of activity and location of the gene product and finally, *KEGG* that informs about pathways in which they are involved.

3.4 Drug Repurposing

Then we delved into exploring their potential as drug targets based on the principle of drug repurposing. This approach aims to utilize drugs that already exist and have therefore already undergone the extensive clinical trial phase, which involves both in vitro and in vivo studies to rule out drug toxicity and determine their efficacy, for the treatment of a different disease than the one for which they were discovered.

3.4.1 Drug Identification

The first step was to identify drugs for which the obtained genes represent targets. Specifically, of the 100 putative genes obtained, we took only the first 20 and accessing manually the database **DGIdb** we compiled a ranking of the approved drugs associated with them, based on the number of targets.

3.4.2 Clinical Trial Validation

In order to better understand whether there were already ongoing testings on the drugs identified relating to the disease of interest and thus validate whether any of them might emerge as suitable for treating it with the right tradeoff of cost benefit in terms of side effects, we investigated the presence of active clinical trials.

In particular we wanted to assess whether on the obtained drugs in the ranking there were active clinical trials related to fibrosing alveolitis. To do so we realized a database search on **ClinicalTrials.gov** of the *National Insitute of Health* (NIH).

4 Results

4.1 Disease Network LCC

From Table 2 we can make some statements about the nodes role in the disease network. We see that the nodes possessing a higher eigenvector centrality, namely `CCL5`, `CCL2` and `CCL11` are chemokines, i.e. polypeptides capable of mediating the inflammatory response. This suggests that in LCC disease they are likely to play an important regulatory role on the other nodes, in particular `CCL5` which has also the higher degree.

Concerning betweeness centrality we notice that `TGF-1` has the highest and seems to have a bridge role between chemokines and the other portion of the network; in this sense, this makes it an optimal target to disrupt the disease network communication and indeed, its role has in fact been extensively investigated as evidenced in the review of Ye et al. (2021).

Observing the scatterplot in Figure 2 there is no clear increasing trend, which implies that the most connected nodes are not always the most central when it comes to controlling the flow of information across the network. Some nodes with a lower degree have a relatively high betweenness centrality, indicating that they may have fewer connections but are still critical in connecting different parts of the network,

4.2 Algorithms Computational Validation

In Table 3, we present a performance evaluation of *DIAMOnD*, *DiaBLE* and *Network Diffusion* algorithms using Precision, Recall and F1-score metrics (average ± SD). The table highlights the comparable performance of *DIAMOnD* and *DiaBLE*, attributed to the specific characteristics of our network, which significantly influence the potential improvements introduced by the *DiaBLE* algorithm over *DIAMOnD* one. Upon exploring our network, we observed that:

- The seed genes in the network have an eccentricity of 4-5, suggesting a form of centrality. This indicates that these nodes can easily reach other nodes in the network with a limited number of steps.
- Conducting a Breadth-First Search (BFS) in the network, starting from our seed genes, we found that nodes at a distance greater than two are only 484.

Based on these considerations, it is reasonable to observe that *DIAMOnD* and *DiaBLE* have demonstrated similar performances in computational validation. The optimization introduced by the *DiaBLE* algorithm appears to have a limited impact on our network, because the majority of nodes are located within a distance equal to or less than 2 (first and second-degree neighbors). Consequently, from the initial iterations, *DIAMOnD* and *DiaBLE* exhibit very similar universe sizes and overall algorithmic performance remains consistent.

To determine the best algorithm in our scenario, we decided to consider the average F1-score metric as a crucial determinant, because it represents an optimal trade-off between precision and recall and consequently, also a good balance between false negative and false positive rates.

Both *DIAMOnD* and *DiaBLE* algorithms demonstrated higher average F1-scores compared to *Network Diffusion.* Additionally, these two algorithms also exhibited better execution times than *Network Diffusion*, which is also an important factor in scenarios involving large data sizes as networks.

Although the two algorithms demonstrated comparable performances in computational validation, we decided to choose *DiaBLE* as the best algorithm because it operates on a slightly smaller universe, making it computationally more efficient. Moreover the p-values calculated in the hypergeometric tests slightly differ, resulting in distinct ranking positions. As highlighted in Petti et al. (2019), this difference might lead *DiaBLE* to produce better results compared to *DIAMOnD* at the biological validation level.

As it possible to observe in the error plots of metrics at different cutoffs, both the *DIAMOnD* (see Figure 3) and *DiaBLE* (see Figure 4) algorithms consistently exhibit better performances and a more stable behavior than Network Diffusion (see Figures 5, 6, 7).[1]

It is important to note that when the analysis is limited to only the top $n/10$ and $n/4$ positions in the output, the outcomes become unstable across all three algorithm. This instability arises from the challenges these algorithms encounter with small test sets and outputs as in our case.

Starting from the $n/2$ cutoff and beyond, the model exhibits significant improvements, achieving best performance at n, where n represents the quantity of known *GDAs.* This suggests that our algorithms require, starting from our data a sufficiently large output to achieve the best metrics.

4.3 Enrichment Results

The enrichment analysis carried out on the known disease genes revealed from a process perspective a variety of them all aimed at a promotion of different cellular activities such as metabolism, intracellular signal transduction, gene expression and replication, as well as processes concerning cell activation in an inflammatory sense, mediated by chemokines. This is confirmed from a molecular function point of view as we see enriched functions of growth factors and their receptors as well as cytokines/chemokines with corresponding receptors.

The molecules that perform these activities are typical markers of the inflammatory and tissue repair process and through the known disease genes it is already possible to guess that the pathogenesis of the disease depends precisely on this inflammatory aspect. For example, we observe in Figure 10 enriched the activation function of tyrosine kinases, for which one of the previously mentioned drugs *nintedanib* is an inhibitor, confirming the importance of the activity of these proteins in the disease pathogenesis.

With seed genes alone, some of the enriched *KEGG* pathways refer to other pathological conditions, different from IPF, but underlying their similarity or involvement in some cases. In fact, for example as studied in Diesler et al. (2022) rheumatoid arthritis, which we see in the Figure 12, and which is the most common inflammatory autoimmune disease, often involves the lungs causing a condition of interstitial lung disease (ILD), very similar to that caused by IPF. This suggests that a combined study of the diseases could benefit the treatment of both. On the other hand, with the putative disease genes, this pathway was found to be no more enriched, as maybe genes more specific to IPF than to RA were probably captured.

Performing the enrichment on putative disease genes obtained with *DiaBLE*, however, we see that the results reinforce those that partially emerged with seed genes alone, in the sense that most of the molecular processes and functions concern the action of cytokines and chemokines and the subsequent recruitment of cells of the immune system, both innate (e.g. macrophages, granulocytes) and adaptive (lymphocytes).

This clearly confirms what has been described in Liu et al. (2023), in which the role of inflammatory mediators appears to be of primary importance in the disease. Indeed, it seems that some of those chemokines also play specialized roles in the remodeling processes associated with the impaired tissue repair, acting also

[1] When interpreting error bars, if it assumes negative values, this does not represent negative metrics but is due to instability, meaning that the mean is lower than the standard deviation at low cutoffs as explained in the main text.

on fibroblast; they are in fact able to determine their migration, proliferation, and activation. As such, they probably represents the critical link connecting ongoing inflammation with the atypical tissue repair observed in the development of IPF.

The enrichment with *GO Cellular Component* in 15 confirms this aspect providing information on the localization of those main actors; as well as the remodelling process itself that we see in the Figure 13, the collagen-containing extracellular matrix is of particular importance, with a much lower p-value than in the enrichment of seed genes alone. This implies that the identified putative disease genes contribute to the composition and remodelling of the extracellular matrix of the lung tissue. Among the genes identified there are are structural collagen chains (eg. `COL18A`, `COL2A1` ,`COL2A1`), as well as the other regulatory factors.

4.4 Community Detection Algorithms

The *Louvain* and *MCL* community detection algorithms have demonstrated a remarkable similarity in their results, both identifying only a single community as significantly enriched for the specified seed genes. Specifically, the *Louvain* algorithm identified a community comprising 38 genes, including 6 seed genes, while the *MCL* algorithm detected a marginally smaller community of 32 genes, with 5 of those being seed genes. Notably, there was a considerable overlap between the two communities, with 27 genes shared between them, indicating a high degree of similarity in the communities they identified, albeit with some minor differences. Furthermore, when assessed biologically, the enrichment results against all databases were strikingly similar, suggesting that both algorithms were comparably effective in discerning the modular structure of the network.

Comparing with *DiaBLE* in terms of biological validation, the results showed substantial effectiveness of both approaches in detecting valid disease genes, with some subtle differences. In particular, the community detection algorithms did not detect the process concerning the organisation of the extracellular matrix in the first 10 terms, whereas in the KEGG pathways were revealed two fundamental ones in inflammatory processes, namely TNF, which is a target of the drug *pirfenidone*, and of `Nf-kb`, which is a key transcriptional factor able to determine fibroblast-to-myofibroblast (FMT) transition. the latter in particular was considered in Jaffar et al. (2021) as a potential target to attack in IPF and *ACT-001* compound was tested in vitro ; this suggests the possibility of further investigating this transcription factor and its role in determining the disease.

4.5 Drug Repurposing

4.5.1 Drug Identification

For what concerns the drugs ranking based on the number of genes associated with them, it was not possible with the approved drugs to identify any associated with more than one target. Mainly we found drugs associated with two molecules of the family of chemokines, which play a pivotal role in infectious and inflammatory processes as they are signal molecules capable of recruiting and regulating the activity of immunity cells.

These two molecules are `CXCL12` which is a fundamental granulocyte neutrophil chemoattractant and `CXCL10` that exercise its chemoattractive function with many different immune system cells as monocytes/macrophages, T cells, NK cells, and dendritic cells.

4.5.2 Clinical Trial Validation

Since we did not have a real ranking of the drugs, we searched for clinical trials on all the drugs that emerged from the analysis. Specifically with regard to *Rituximab*, *Alemtuzumab* and *Fludarabine*, we have identified a clinical trial involving all of them for fibrosing alveolitis but which is not aimed at evaluating its efficacy directly, but rather at assessing whether bone marrow transplantation can resolve the disease; the role of *Alemtuzumab* together with other drugs would be to bring about immunosuppression to ensure the success of the transplantation. The first two are monoclonal antibodies, *Rituximab* directed against the `CD20` antigen found on the surface of normal and malignant B lymphocytes, while *Alemtuzumab* is is directed against `CD52` of B lymphocytes. Both are able to induce depletion of B lymphocytes, mediating in this way a suppression of inflammation.

Rituximab on the other hand, has been found to be directly involved in five other direct studies on the disease, two of which were in combination with steroids (a group of molecules including testosterone). We then identified no less than 12 studies involving *Prednisone* and 9 involving *Methylprednisolone* which are two corticosteroids with a similar structure used to treat inflammation or immune reactions (see Figures 25, 26). In practically all the clinical trials investigated, however, they were administered in combination with other drugs probably because as described in the recent literature e.g. Raghu et. al (2022) these corticosteroids alone, although powerful anti-inflammatory agents, are not adequate to slow disease progression.

5 Conclusions

In conclusion, we can assert that both algorithms based on connectivity significance and those on the disease module hypothesis have emphasized the role of inflammatory and fibrosis mediators in the pathogenesis of the disease as described in Liu et. al (2023). The analysis of drugs and related clinical trials has primarily focused on chemokines as targets and has shown how, some clinical trials are underway on identified drugs, such as Rituximab and corsticosteroids, while it has been observed that the latters are not effective in treatment; as suggested in Liu et al. (2023), this is probably due to the fact that the role of the relationship between inflammation and abnormal repair needs to be further clarified, specifically in terms of timing.

Given all the obtained results is suggested that probably inflammation starts in an early phase of the disease and therefore it could be appropriate to investigate not only its mediators but evaluate the ones that are relevant to fibrosis and are able to not only recruit inflammation factors but also to induce fibroblasts activation, that in turn determine extracellular matrix remodelling and abnormal collagen deposition.

6 Authors Contribution

The work described above was conducted collaboratively throughout its duration, with ongoing communication to ensure alignment of our decisions and integration of each other's contributions as needed. Specifically, M.I.A. focused more on the performance evaluation of various algorithms and enrichment analysis, while I.P. worked more on the data gathering and network construction, Network Diffusion algorithm using Cytoscape and the drug repurposing and clinical trials part. Concerning the optional Clustering part, M.I.A. mainly implemented the MCL algorithm, while I.P. focused more on the Louvain algorithm. Both of us were involved in writing, reviewing and editing the report and in all the non mentioned tasks.

7 Figures

7.1 Network Exploration

Disease name	UMLS ID	MeSH	Total Genes	Genes in interactome	LCC size
Alveolitis, Fibrosing	C4721507	D011658	83	77	41

Table 1: Disease network info

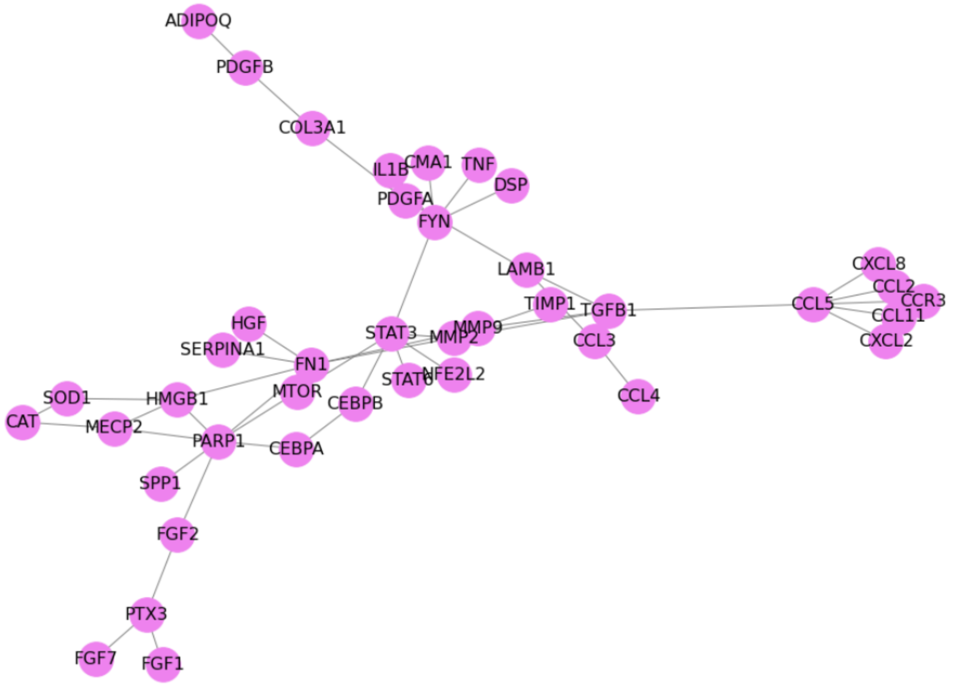

Figure 1: Largest Connected Component of the Disease Network

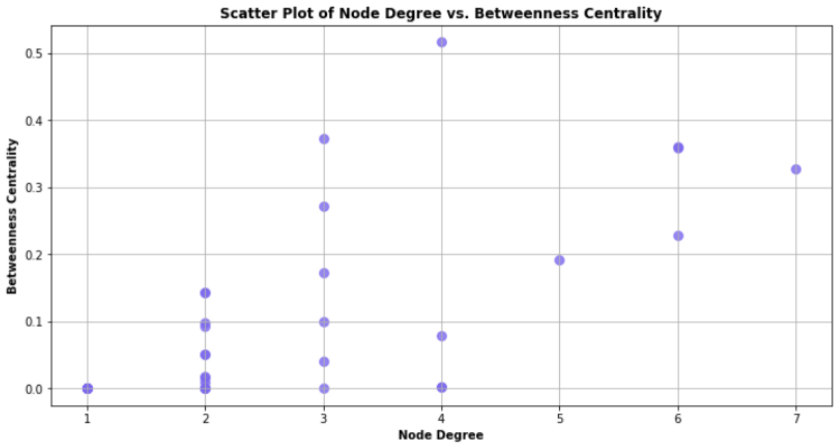

Figure 2: Scatterplot of Degree vs Betweeness Centrality

Ranking	Gene	Degree	Betweenness	Closeness	Eigen	Bet/Deg
1	PARP1	7	0.326496	0.303030	0.141850	0.046642
2	FN1	6	0.357906	0.325203	0.151309	0.059651
3	CCL5	6	0.228205	0.261438	0.505746	0.038034
4	STAT3	6	0.360256	0.317460	0.077344	0.060043
5	FYN	5	0.191026	0.254777	0.031927	0.038205
6	CCL2	4	0.001282	0.211640	0.407804	0.000321
7	CCL11	4	0.001282	0.211640	0.407804	0.000321
8	TGF-1	4	0.516239	0.322581	0.215820	0.129060
9	HMGB1	4	0.077991	0.272109	0.110197	0.019498
10	CCR3	3	0.000000	0.210526	0.355656	0.000000
11	PTX3	3	0.098718	0.200000	0.013277	0.032906
12	MECP2	3	0.039530	0.239521	0.076501	0.013177
13	LAMB1	3	0.271795	0.264901	0.068951	0.090598
14	MMP2	3	0.373291	0.347826	0.119872	0.124430
15	MMP9	3	0.172650	0.303030	0.106751	0.057550
16	PDGFB	2	0.050000	0.156250	0.001711	0.025000
17	PDGFA	2	0.142308	0.217391	0.020158	0.071154
18	MTOR	2	0.092308	0.287770	0.059221	0.046154
19	CXCL2	2	0.000000	0.209424	0.245894	0.000000
20	CCL3	2	0.050000	0.212766	0.020024	0.025000
21	COL3A1	2	0.097436	0.182648	0.005890	0.048718
22	CEBPB	2	0.016880	0.261438	0.033708	0.008440
23	CAT	2	0.000641	0.196078	0.030998	0.000321
24	CEBPA	2	0.015598	0.248447	0.047445	0.007799
25	CMA1	2	0.000000	0.205128	0.011816	0.000000
26	FGF2	2	0.142308	0.242424	0.041928	0.071154
27	IL1B	2	0.000000	0.205128	0.011816	0.000000
28	CXCL8	2	0.000000	0.209424	0.245894	0.000000
29	SOD1	2	0.009188	0.217391	0.038163	0.004594
30	ADIPOQ	1	0.000000	0.135593	0.000461	0.000000
31	SPP1	1	0.000000	0.233918	0.038338	0.000000
32	FGF7	1	0.000000	0.167364	0.003590	0.000000
33	SERPINA1	1	0.000000	0.246914	0.040861	0.000000
34	DSP	1	0.000000	0.204082	0.008623	0.000000
35	FGF1	1	0.000000	0.167364	0.003590	0.000000
36	STAT6	1	0.000000	0.242424	0.020883	0.000000
37	HGF	1	0.000000	0.246914	0.040861	0.000000
38	CCL4	1	0.000000	0.176211	0.005393	0.000000
39	NFE2L2	1	0.000000	0.242424	0.020883	0.000000
40	TNF	1	0.000000	0.204082	0.008623	0.000000
41	TIMP1	1	0.000000	0.233918	0.028786	0.000000

Table 2: Disease LCC genes and their metrics

7.2 Algorithms Performance Evaluation

Algorithms	AVG Precision	Precision SD	AVG Recall	Recall SD	AVG F1	F1 SD
DIAMOnD	0.024000	0.013565	0.154167	0.085554	**0.041529**	0.023410
DiaBLE	0.024000	0.013565	0.154167	0.085554	**0.041529**	0.023410
Diffusion t = 0.002	0.012000	0.007483	0.077500	0.048419	0.020780	0.012959
Diffusion t = 0.005	0.012000	0.007483	0.077500	0.048419	0.020780	0.012959
Diffusion t = 0.001	0.016000	0.010198	0.102500	0.063988	0.027676	0.017589

Table 3: Algorithms Performance Evaluation

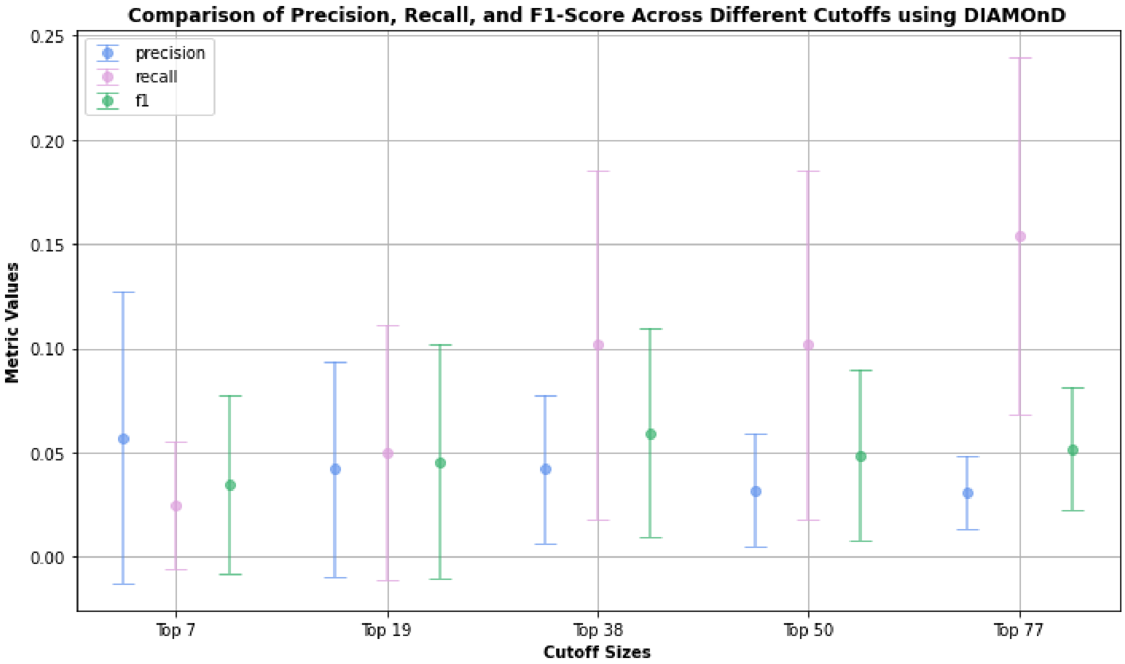

Figure 3: Error plot of Precision, Recall, and F1-Score Across Different Cutoffs using DIAMOnD algorithm

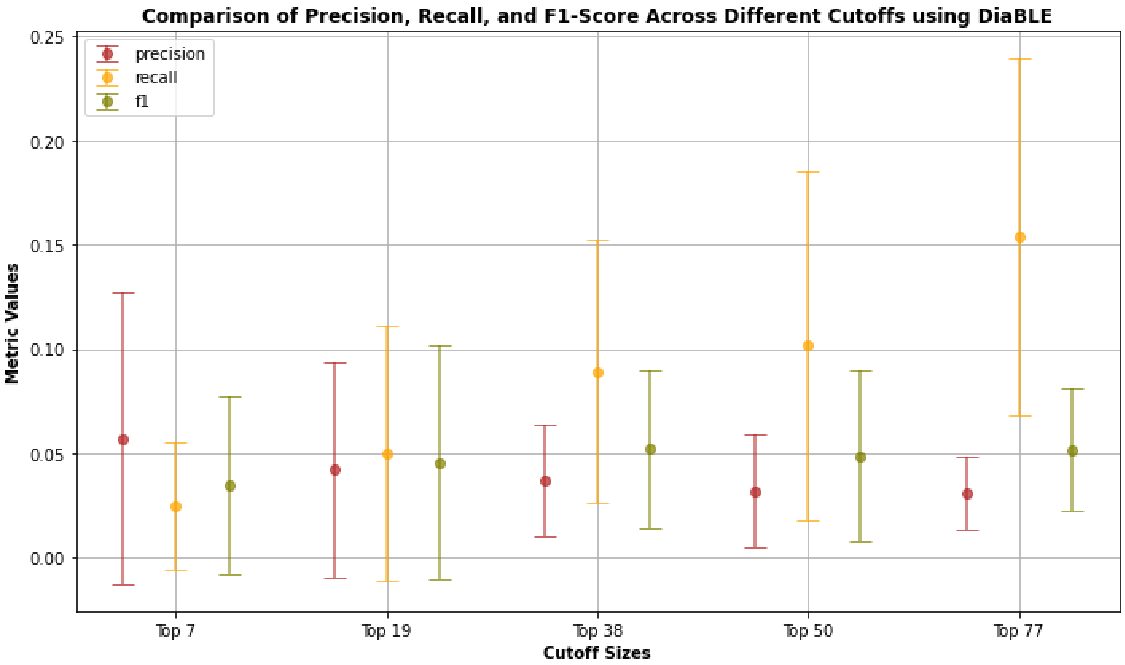

Figure 4: Error plot of Precision, Recall, and F1-Score Across Different Cutoffs using DiaBLE algorithm

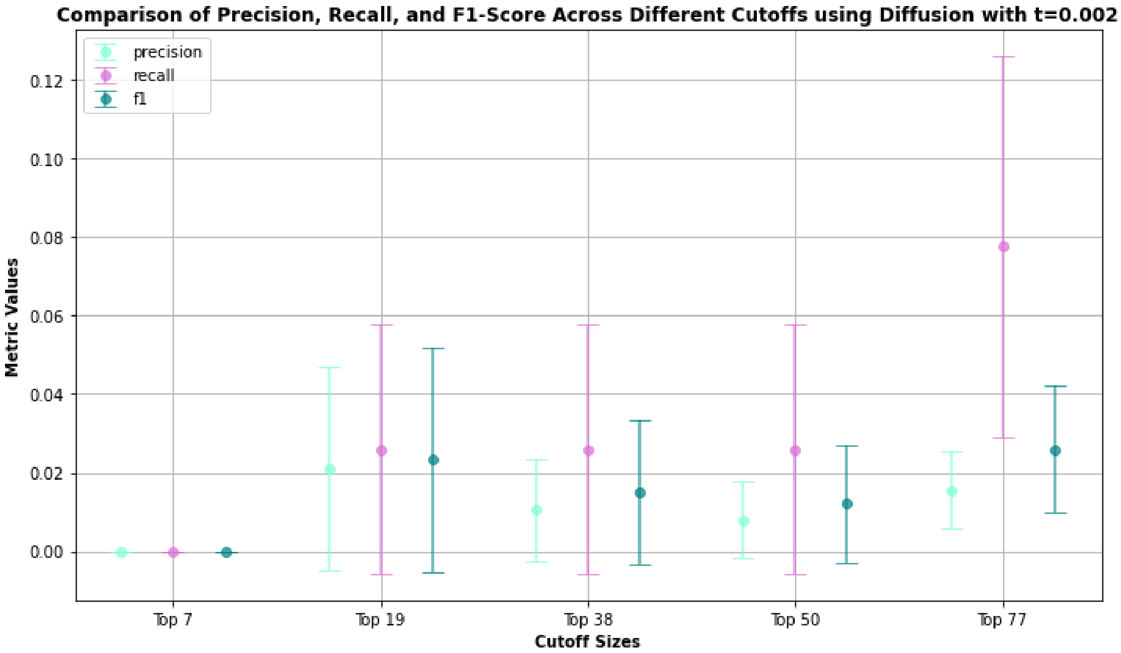

Figure 5: Error plot of Precision, Recall, and F1-Score Across Different Cutoffs using Diffusion algorithm with t = 0.002

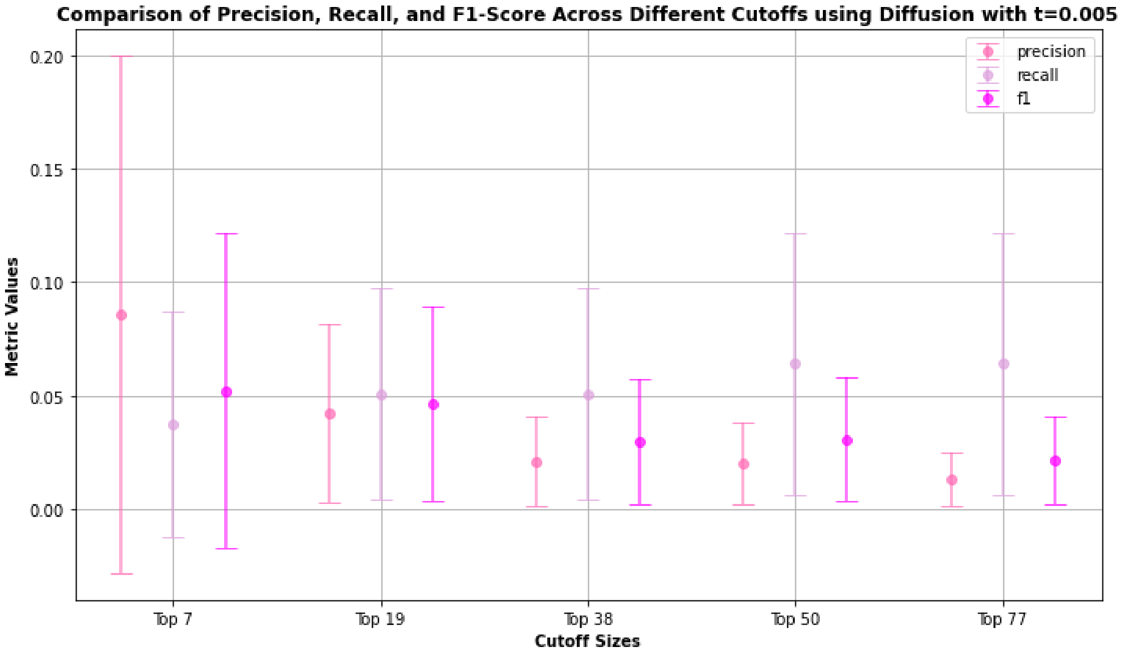

Figure 6: Error plot of Precision, Recall, and F1-Score Across Different Cutoffs using Diffusion algorithm with t = 0.005

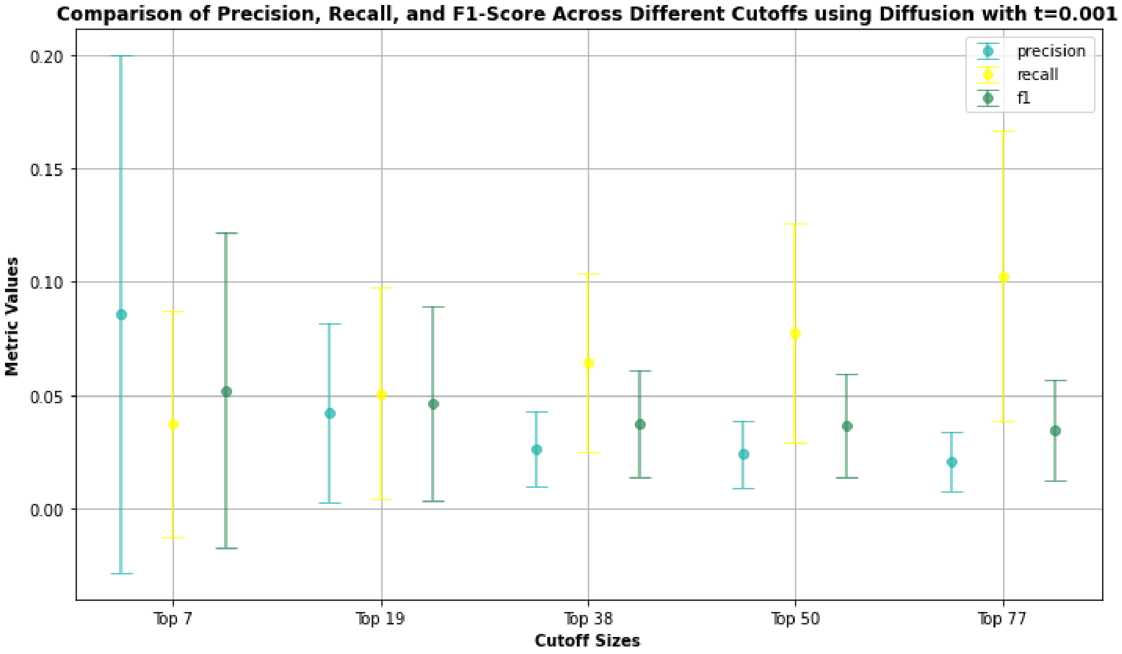

Figure 7: Error plot of Precision, Recall, and F1-Score Across Different Cutoffs using Diffusion algorithm with t = 0.01

7.3 Putative Disease Genes

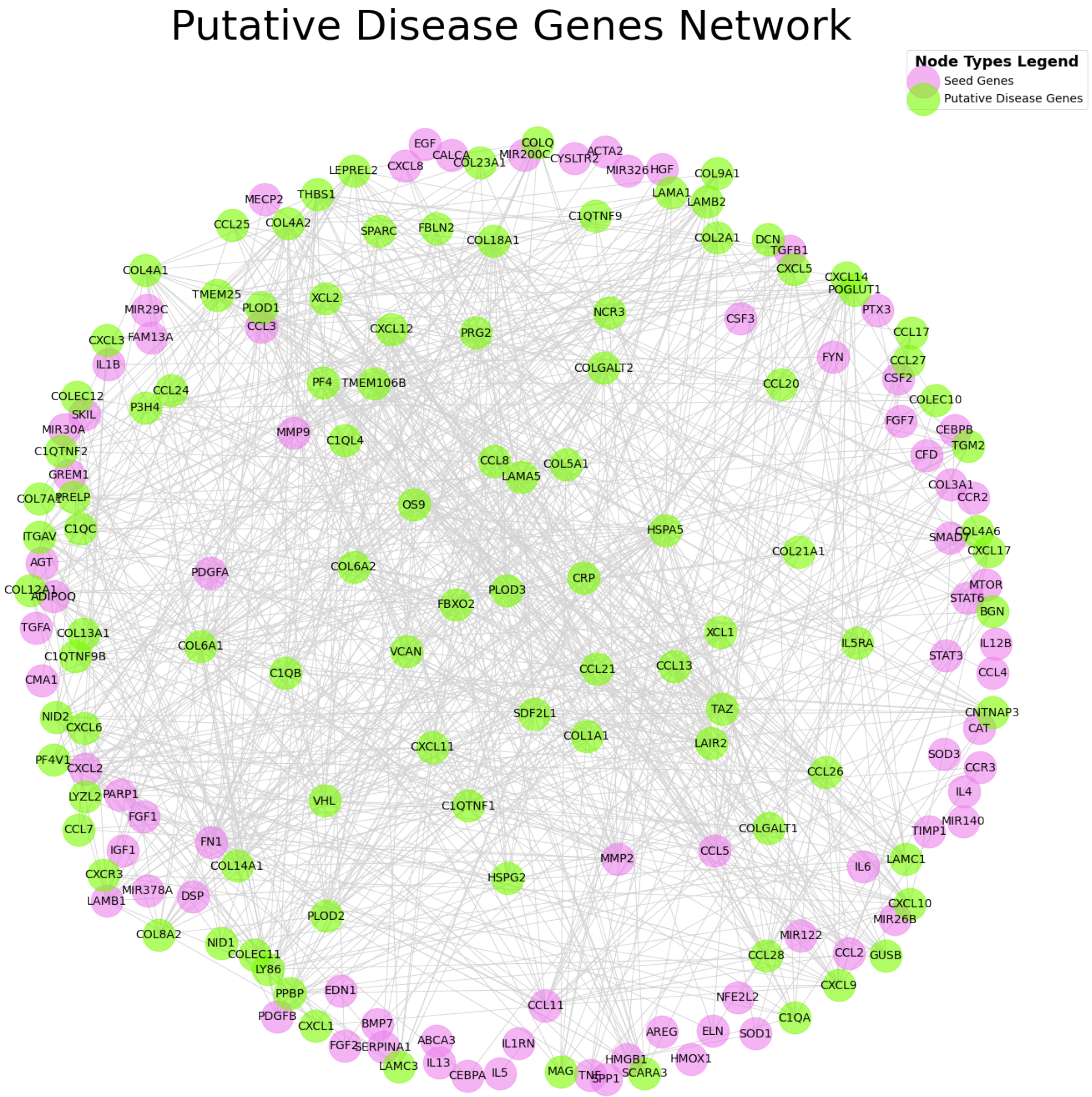

Figure 8: Network of seed genes (purple nodes) and putative disease genes (green nodes) identified with DiaBLE (best algorithm)

Putative Disease Genes Ranking			
1	PF4	51	COLQ
2	CXCL12	52	COL21A1
3	CXCL6	53	PLOD1
4	CCL8	54	COL4A6
5	CXCL10	55	COL1A1
6	CCL13	56	COLEC12
7	CXCL11	57	COL13A1
8	CCL26	58	DCN
9	XCL1	59	LEPREL2
10	CXCL17	60	C1QB
11	CCL28	61	PLOD2
12	CXCL14	62	COLEC11
13	CCL21	63	C1QA
14	CXCL9	64	COLEC10
15	PPBP	65	VHL
16	XCL2	66	P3H4
17	PF4V1	67	SPARC
18	CCL25	68	COLGALT1
19	CCL24	69	TGM2
20	CCL20	70	ITGAV
21	VCAN	71	COL23A1
22	CCL17	72	COL7A1
23	CXCL5	73	OS9
24	CXCL3	74	PRELP
25	CXCL1	75	FBXO2
26	CCL27	76	SCARA3
27	CXCR3	77	TMEM106B
28	CCL7	78	HSPG2
29	THBS1	79	IL5RA
30	COL2A1	80	LAMA5
31	COL4A1	81	NID2
32	BGN	82	TMEM25
33	COL18A1	83	LAMA1
34	COL6A1	84	FBLN2
35	MAG	85	NID1
36	COL5A1	86	HSPA5
37	C1QTNF1	87	LAMC1
38	LAIR2	88	CRP
39	COLGALT2	89	LY86
40	COL4A2	90	LAMB2
41	COL9A1	91	NCR3
42	COL8A2	92	PRG2
43	COL12A1	93	LAMC3
44	C1QC	94	GUSB
45	COL6A2	95	SDF2L1
46	C1QTNF9	96	C1QL4
47	COL14A1	97	TAZ
48	C1QTNF9B	98	POGLUT1
49	C1QTNF2	99	LYZL2
50	PLOD3	100	CNTNAP3

Table 4: Ranked Putative Disease Genes obtained with DiaBLE

7.4 Enrichment Analysis

7.4.1 Enrichment Analysis of Known disease genes

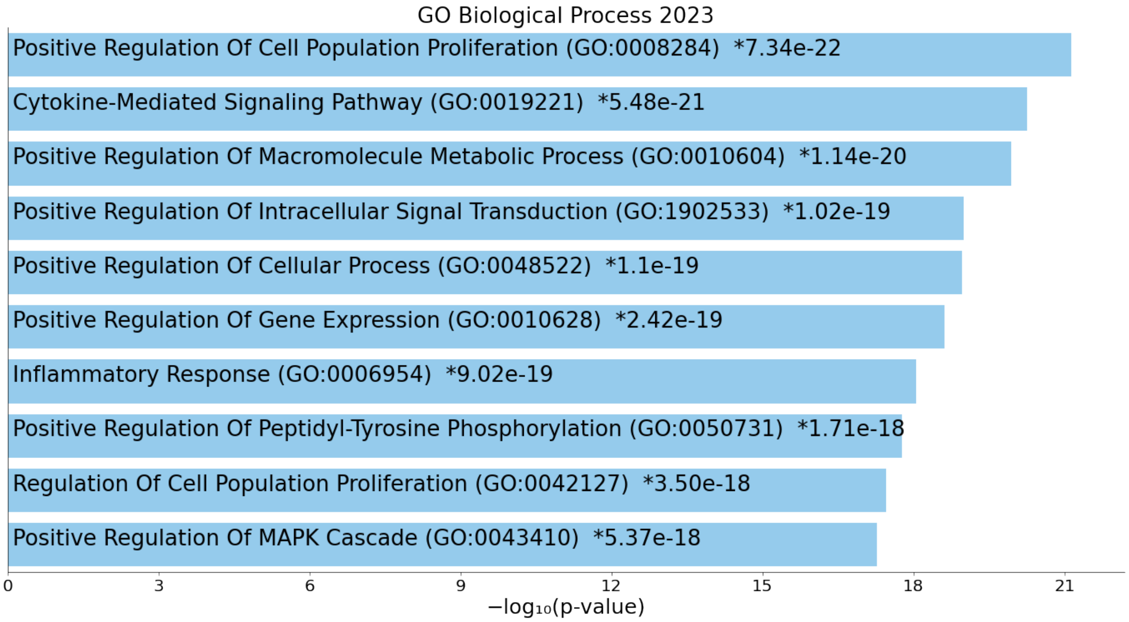

Figure 9: Barchart of the top enriched terms with Gene Ontology Biological Process 2023 of known disease genes present in the interactome

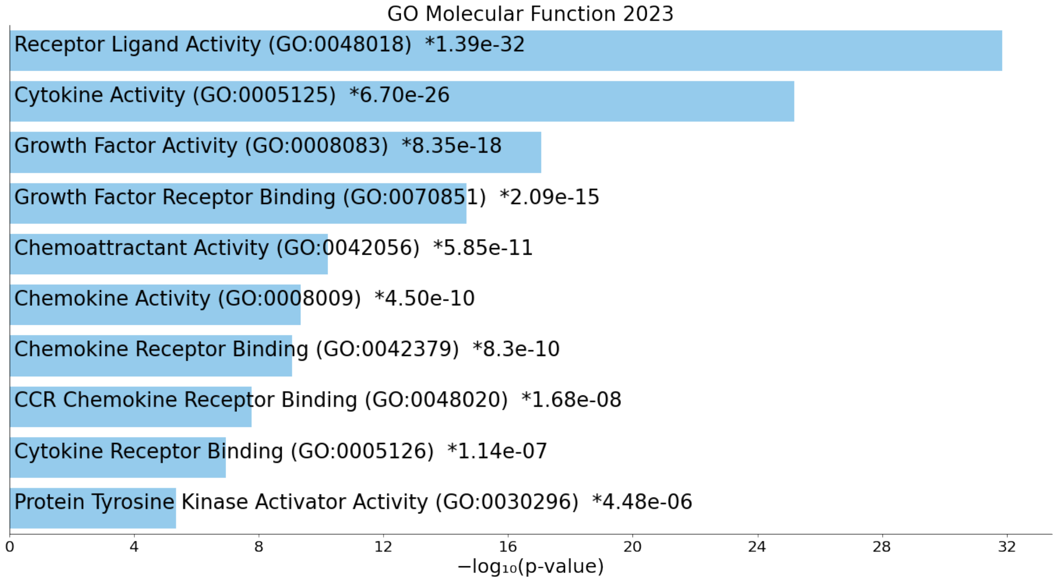

Figure 10: Barchart of the top enriched terms with Gene Ontology Molecular Function 2023 of known disease genes present in the interactome

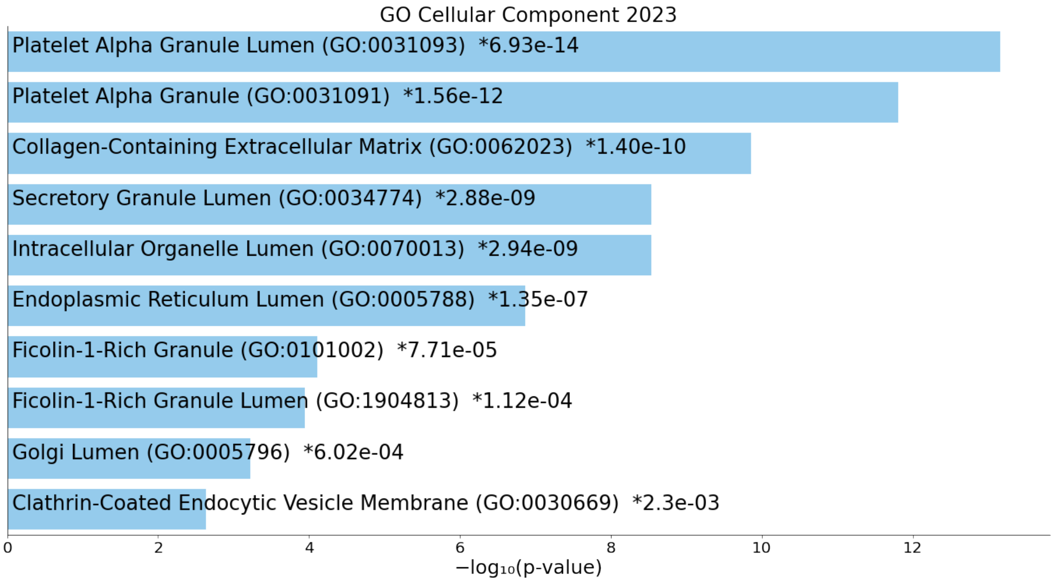

Figure 11: Barchart of the top enriched terms with Gene Ontology Cellular Component 2023 of known disease genes present in the interactome

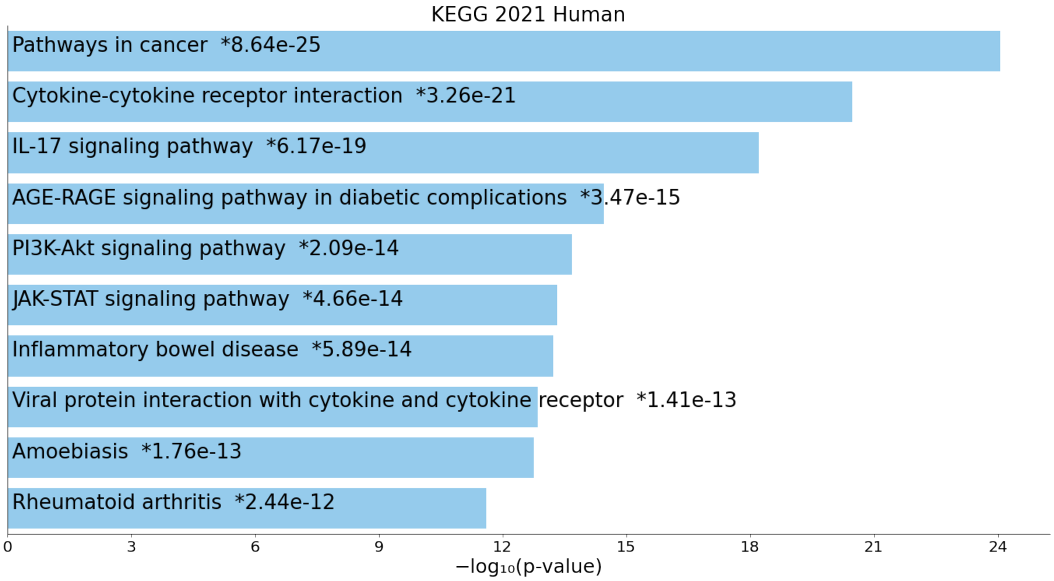

Figure 12: Barchart of the top enriched terms with KEGG 2021 Human of known disease genes present in the interactome

7.4.2 Enrichment Analsysis of Putative disease genes

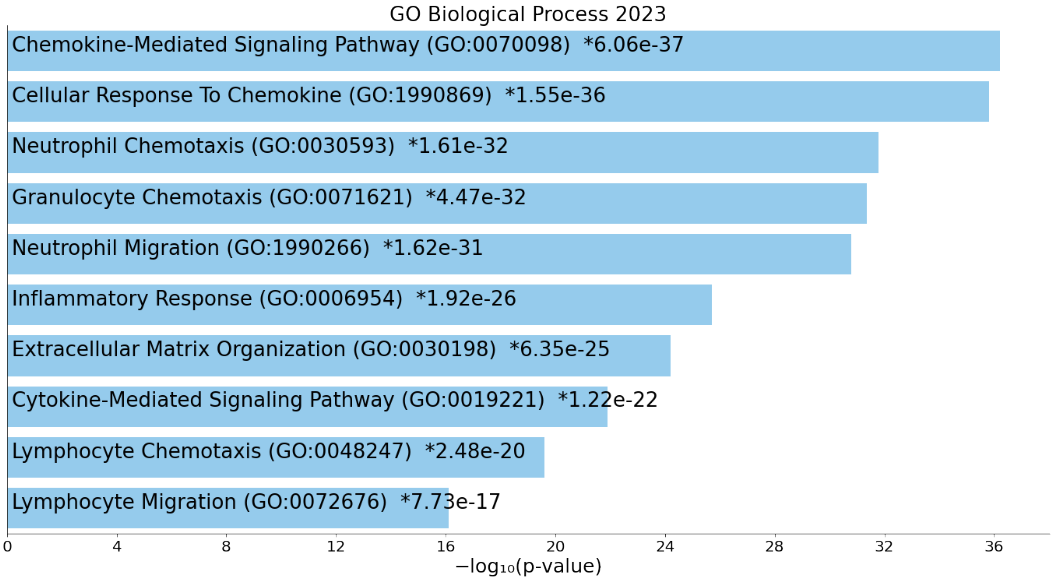

Figure 13: Barchart of the top enriched terms with Gene Ontology Biological Process 2023 of Putative disease genes in green identified with DiaBLE algorithm

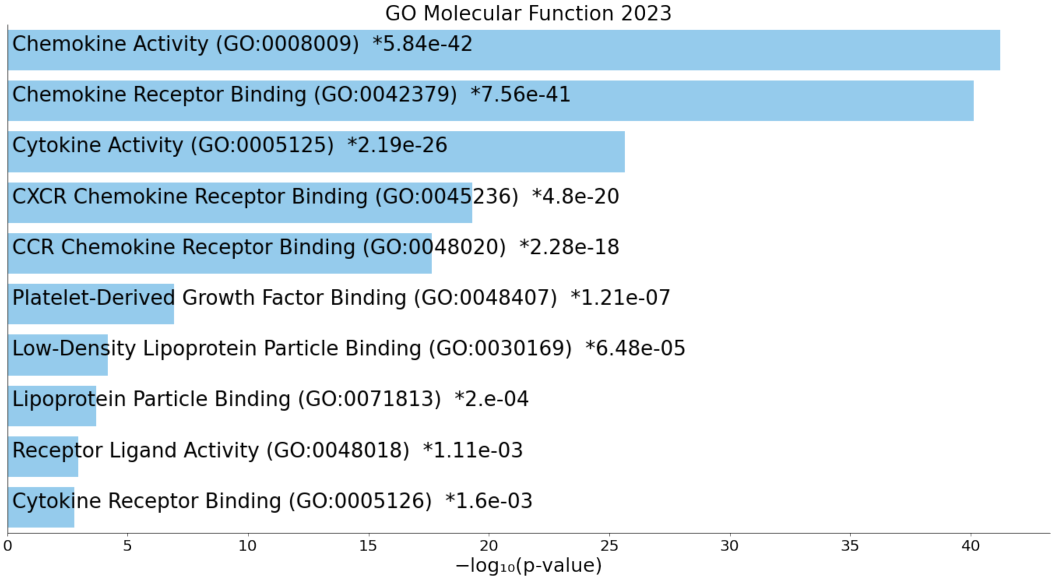

Figure 14: Barchart of the top enriched terms with Gene Ontology Molecular Function 2023 of Putative disease genes in green identified with DiaBLE algorithm

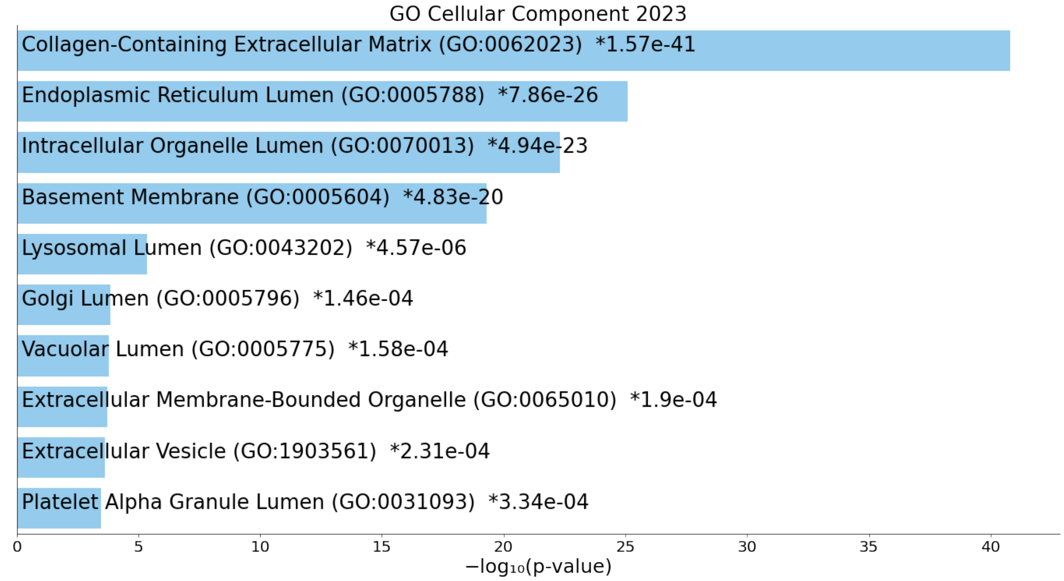

Figure 15: Barchart of the top enriched terms with Gene Ontology Cellular Component 2023 of Putative disease genes in green identified with DiaBLE algorithm

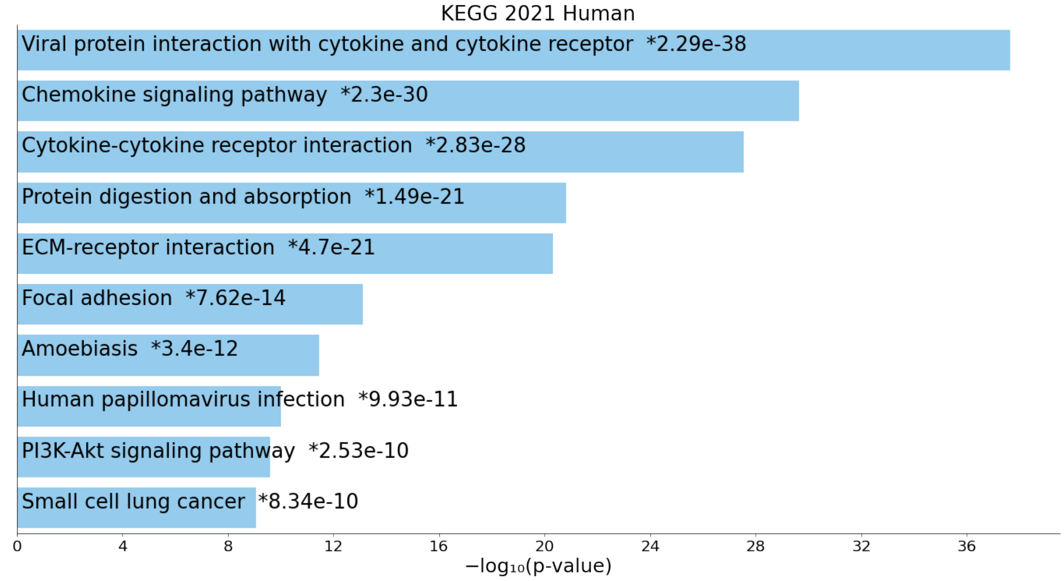

Figure 16: Barchart of the top enriched terms with KEGG 2021 Human of Putative disease genes in green identified with DiaBLE algorithm

7.4.3 Enrichment Analysis of the Disease Module identified with the Louvain algorithm

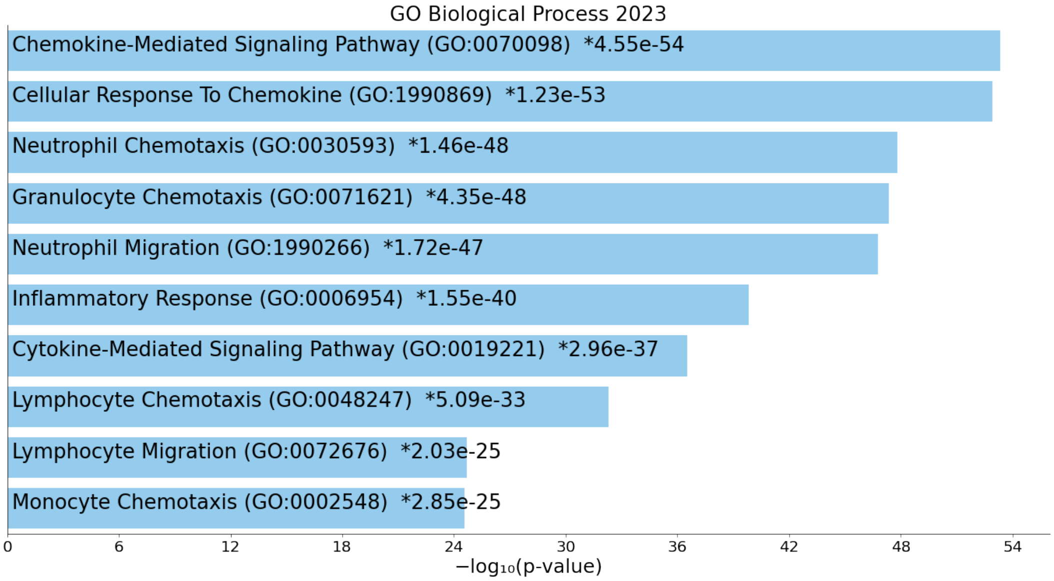

Figure 17: Barchart of the top enriched terms with GO Biological Process 2023 of the disease module identified with the Louvain algorithm

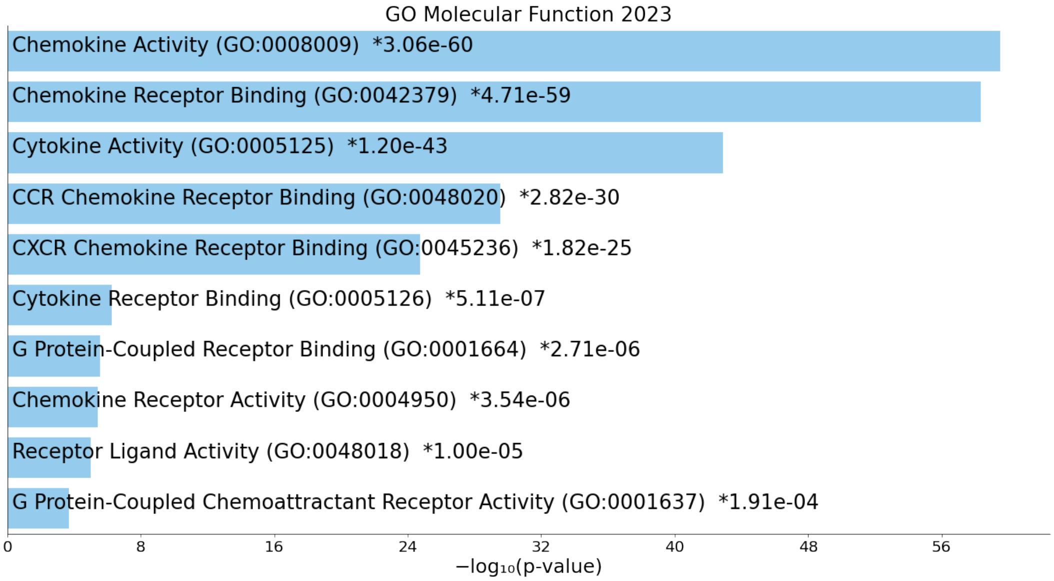

Figure 18: Barchart of the top enriched terms with GO Molecular Function 2023 of the disease module identified with the Louvain algorithm

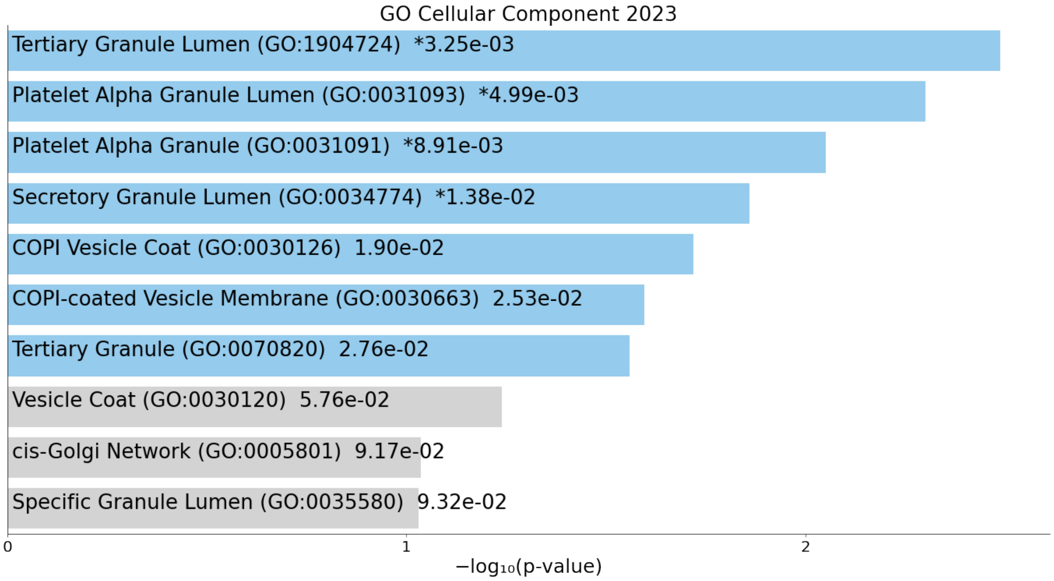

Figure 19: Barchart of the top enriched terms with GO Cellular Component 2023 of the disease module identified with the Louvain algorithm

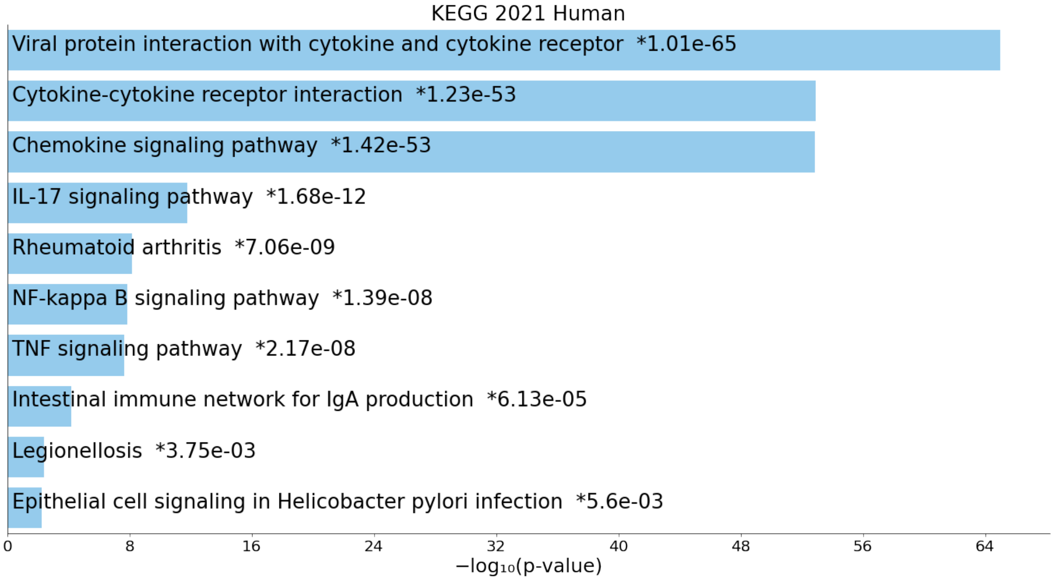

Figure 20: Barchart of the top enriched terms with KEGG 2021 Human of the disease module identified with the Louvain algorithm

7.4.4 Enrichment Analysis of the Disease Module identified with the MCL algorithm

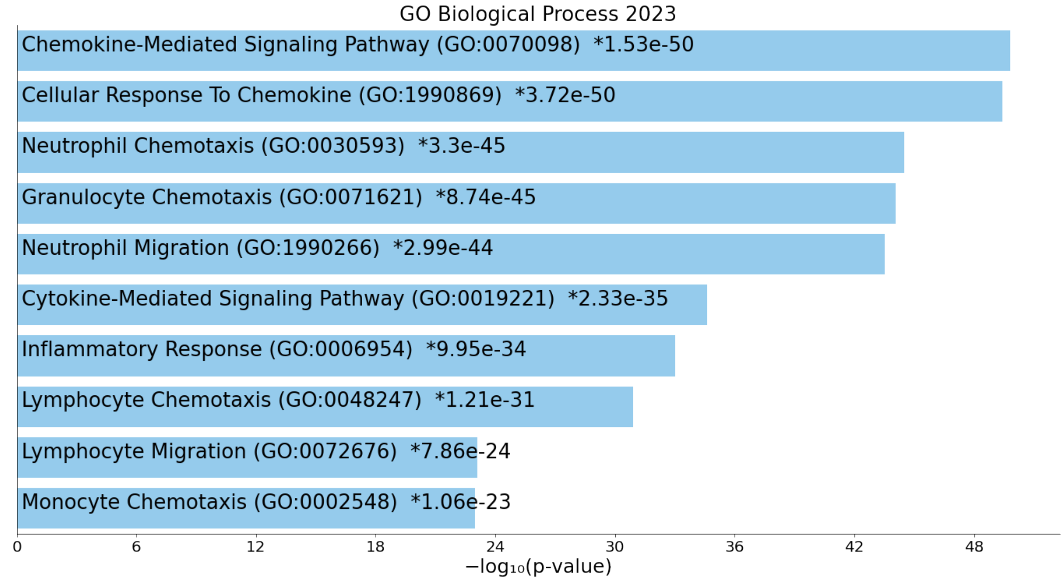

Figure 21: Barchart of the top enriched terms with GO Molecular Function 2023 of the disease module identified with the MCL algorithm

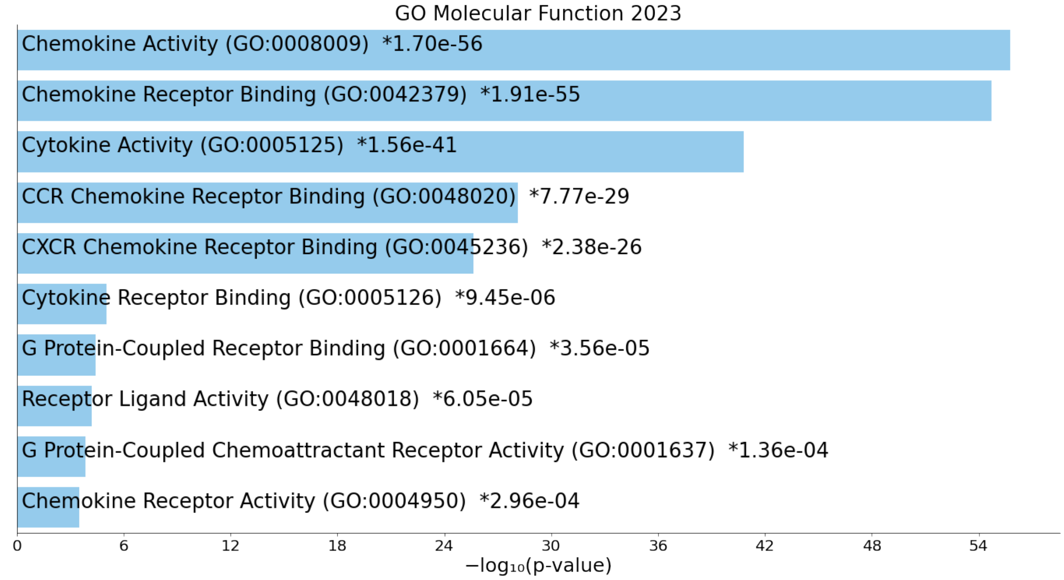

Figure 22: Barchart of the top enriched terms with GO Molecular Function 2023 of the disease module identified with the MCL algorithm

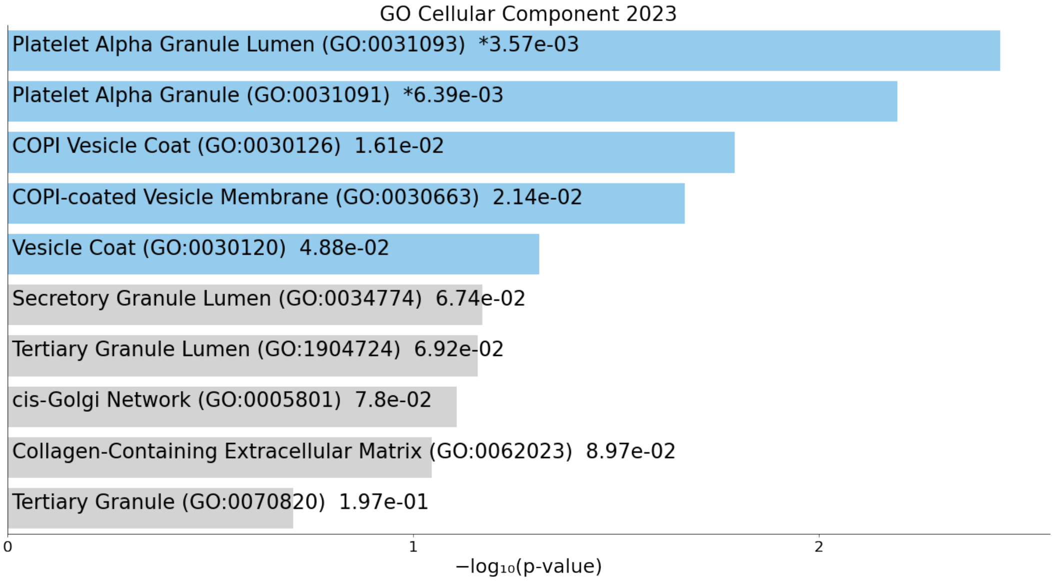

Figure 23: Barchart of the top enriched terms with GO Cellular Component 2023 of the disease module identified with the MCL algorithm

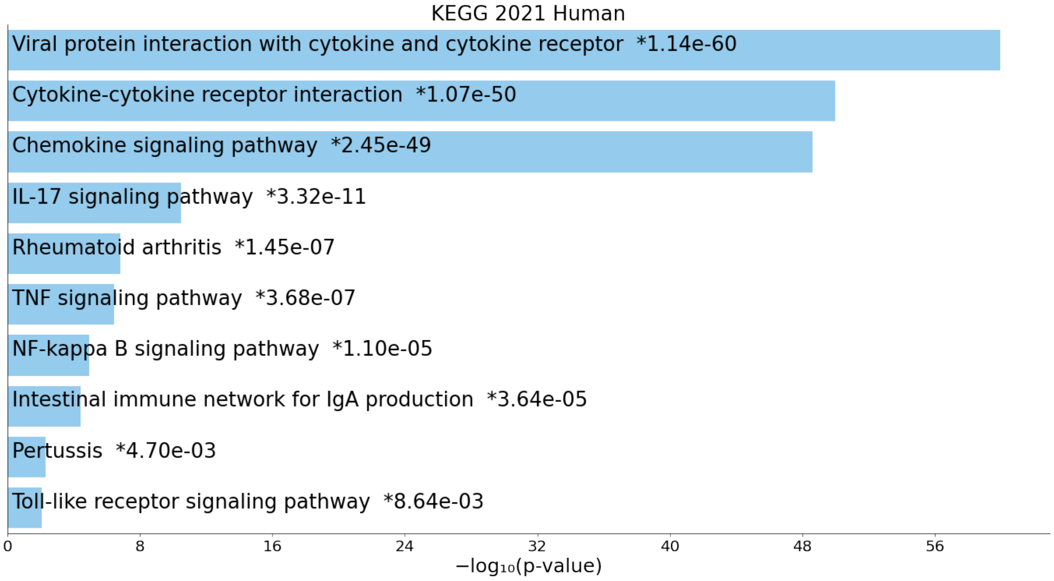

Figure 24: Barchart of the top enriched terms with KEGG Human 2021 of the disease module identified with the MCL algorithm

7.5 Drug repurposing

Ranking	Drug	Nr. of Genes	Genes
1	ALEMTUZUMAB	1	CXCL12
2	PEGINTERFERON ALFA-2B	1	CXCL10
3	VINCRISTINE	1	CXCL12
4	TINZAPARIN SODIUM	1	CCL21
5	TESTOSTERONE	1	CXCL10
6	STAVUDINE	1	CXCL10
7	RITUXIMAB	1	CXCL12
8	RITONAVIR	1	CXCL10
9	PREDNISONE	1	CXCL12
10	PEGINTERFERON ALFA-2A	1	CXCL10
11	ATORVASTATIN CALCIUM TRIHYDRATE	1	CXCL10
12	OXALIPLATIN	1	CXCL10
13	METHYLPREDNISOLONE	1	CXCL10
14	HUMAN CHORIONIC GONADOTROPIN	1	CXCL10
15	FLUDARABINE	1	CXCL12
16	CYCLOPHOSPHAMIDE ANHYDROUS	1	CXCL12
17	CHLORAMBUCIL	1	CXCL12
18	ATROPINE	1	CXCL10
19	ZIDOVUDINE	1	CXCL10

Table 5: Drugs ranking based on the association with disease genes

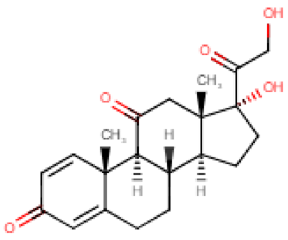

Figure 25: Prednisone structure from DRUGBANK database

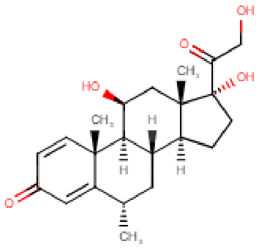

Figure 26: Methylprednisolone structure from DRUGBANK database

8 References

Diesler, Remi, and Vincent Cottin. ”Pulmonary fibrosis associated with rheumatoid arthritis: from pathophysiology to treatment strategies.” Expert Review of Respiratory Medicine 16.5 (2022): 541-553.

Jafari, Mohieddin, and Naser Ansari-Pour. “Why, when and how to adjust your P values?.” Cell Journal (Yakhteh) 20.4 (2019): 604.

Jaffar, Jade, et al. ”Inhibition of NF-B by ACT001 reduces fibroblast activity in idiopathic pulmonary fibrosis.” Biomedicine Pharmacotherapy 138 (2021): 111471.

Liu, Shanshan, et al. “CC chemokines in idiopathic pulmonary fibrosis: pathogenic role and therapeutic potential.” Biomolecules 13.2 (2023): 333.

Petti, Manuela, et al. “Connectivity significance for disease gene prioritization in an expanding universe.” IEEE/ACM transactions on computational biology and bioinformatics 17.6 (2019): 2155-2161.

Raghu, Ganesh, et al. “Idiopathic pulmonary fibrosis (an update) and progressive pulmonary fibrosis in adults: an official ATS/ERS/JRS/ALAT clinical practice guideline.” American Journal of Respiratory and Critical Care Medicine 205.9 (2022): e18-e47.

Spagnolo, Paolo, et al. “Idiopathic pulmonary fibrosis: Disease mechanisms and drug development.” *Pharmacology & Therapeutics*, 222 (2021): 107798.

Ye, Zhimin, and Yongbin Hu. “TGF-1: Gentlemanly orchestrator in idiopathic pulmonary fibrosis.” International Journal of Molecular Medicine 48.1 (2021): 1-14.