

Integrative Rare Disease Biomedical Profile based Network Supporting Drug Repurposing, a case study of Glioblastoma

Erin McGowan

NCATS: National Center for Advancing Translational Sciences

Jaleal Sanjak

NCATS: National Center for Advancing Translational Sciences

Ewy A. Mathé

NCATS: National Center for Advancing Translational Sciences

Qian Zhu (✉ qian.zhu@nih.gov)

National Institute of Health <https://orcid.org/0000-0002-4858-6333>

Research Article

Keywords: Data integration, Network analysis, Drug repurposing, Glioblastoma

Posted Date: April 18th, 2023

DOI: <https://doi.org/10.21203/rs.3.rs-2809689/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License. [Read Full License](#)

Abstract

Background

Glioblastoma (GBM) is the most aggressive and common malignant primary brain tumor; however, treatment remains a significant challenge. This study aims to identify drug repurposing candidates for GBM by developing an integrative rare disease profile network containing heterogeneous types of biomedical data.

Methods

We developed a Glioblastoma-based Biomedical Profile Network (GBPN) by extracting and integrating biomedical information pertinent to GBM-related diseases from the NCATS GARD Knowledge Graph (NGKG). We further clustered the GBPN based on modularity classes which resulted in multiple focused subgraphs, named mc_GBPN. We then identified high-influence nodes by performing network analysis over the mc_GBPN and validated those nodes that could be potential drug repositioning candidates for GBM.

Results

We developed the GBPN with 1,466 nodes and 107,423 edges and consequently the mc_GBPN with forty-one modularity classes. A list of the ten most influential nodes were identified from the mc_GBPN. These notably include Riluzole, stem cell therapy, cannabidiol, and VK-0214, with proven evidence for treating GBM.

Conclusion

Our GBM-targeted network analysis allowed us to effectively identify potential candidates for drug repurposing. This could lead to less invasive treatments for glioblastoma while significantly reducing research costs by shortening the drug development timeline. Furthermore, this workflow can be extended to other disease areas.

Background

Glioblastoma (GBM) is a rare, malignant variety of brain tumor that develops from astrocyte and oligodendrocyte cells.[1] GBM is both the most aggressive and most common malignant primary brain tumor, making up 54% of all gliomas and 16% of all primary brain tumors.[2] The incidence of GBM ranges from 0.59 to 5 per 100,000 persons, and this number is rising in many countries.[3] GBM patients have a median survival of only 15 months, and less than 5% of patients survive 5 years following diagnosis. There is currently no cure for GBM.[2] The standard of care (SOC) for grades 3 or 4 high-grade gliomas (HGGs) including GBM, a grade 4 glioma, currently relies on maximally safe surgical resection followed by concurrent radiation therapy and temozolomide (TMZ).[4] While this SOC has increased the median survival time of GBM patients,[2] Stupp et al.[5] has shown that the two year survival rate of patients who undergo radiotherapy plus TMZ treatment is only 26.5 percent. Moreover, Stupp et al.[5] found that concomitant treatment with radiotherapy plus TMZ resulted in grade 3 or 4 hematologic toxic effects in 7 percent of patients.

By April 2022, four drugs besides TMZ and one device were approved by the FDA for treating HGGs: lomustine,[6] intravenous carmustine,[7] carmustine implants,[8] bevacizumab,[9] and tumor treatment fields.[10] Only TMZ, carmustine implants, and tumor treatment fields are FDA-approved for new diagnoses (the rest are approved for recurrent HGGs).[4] Carmustine wafer implants are expensive and have a high complication rate (42.7%), whereas tumor treatment fields are expensive, inconvenient for patients, and yield marginal survival benefits.[4] Toxicity is also a common issue with all current therapies.[4] Thus, there is an urgent need for therapy discovery for GBM patients that are both effective and less invasive than the current SOC.

Drug repurposing is the practice of repurposing an active pharmaceutical ingredient already approved for use in the treatment of one condition for the treatment of another.[11] This approach reduces research costs and allows treatments to reach patients more quickly. Repurposed drugs seeking approval are 150% more likely to be introduced on the market than novel drugs.[12] The exponential growth of large-scale, publicly-available biomedical and pharmaceutical data combined with advancements in high-performance computing have enabled the development of various computational drug repurposing approaches including data mining, machine learning, and network analysis.[13] These in silico strategies, along with disease molecular profiles (e.g. associated genes, biomarkers, signaling pathways, environmental factors, etc.), empower researchers to determine the degree of similarity between diseases by their molecular features.[11] Network analysis in particular has been used extensively in computational drug repurposing, as networks provide an intuitive method of modeling biological and biomedical entities and their interactions and relationships to each other.[13] Centrality measures play a vital role in network analysis, allowing researchers to identify important nodes within a network from a structural perspective.[14] Though frequently used in social network analysis, centrality measures have been adapted as a metric for biological studies since as early as 2001.[15] A previous drug repurposing study ranked drugs by their centrality scores within networks composed of drugs connected based on their side effects and interactions.[16] Another study suggests that the centralities of drugs in a network of drugs connected based on their side-effect similarities may have significant implication in drug repurposing.[14] Most of those published applications mainly leveraged one aspect of drugs, such as side effects or interactions; thus nodes in their established network were specifically associated with drugs (as opposed to other data types such as diseases, phenotypes, proteins, etc). Inspired by these studies, we proposed to generate integrative rare disease biomedical profiles with heterogeneous types of data from our previously developed NCATS Genetic and Rare Diseases (GARD) Knowledge Graph (NGKG)[17] with forty-three rare disease related data resources. Furthermore, instead of mining the entire NGKG, only GBM-associated subgraphs based on pre-calculated disease clusters were derived, and multiple network analysis techniques, such as centrality measures and community detection, were combined and applied to generate GBM-focused graphs for identifying high-influence nodes, which might be potential drug repurposing candidates.

Materials And Methods

In this study, to uncover significant associations relevant to GBM for drug repurposing, we performed network analysis in three steps: 1) we developed a GBM-based Biomedical Profile Network (GBPN) by obtaining the GBM-related biomedical data extracted from the NGKG,[17] 2) we clustered the GBPN into a modularity classes-based network (mc_GBPN) by performing community detection, and 3) we identified high-influence nodes as potential candidates for drug repurposing for GBM from mc_GBPN via various centrality measures. Figure 1 shows the study workflow.

A. NCATS GARD Knowledge Graph (NGKG)[17]

The GARD Information Center was managed by the NCATS to provide freely accessible consumer health information on over 6500 genetic and rare diseases. To expand the use of information from GARD for biomedical research in rare diseases, we previously developed the NGKG,[17] a knowledge graph that integrated data from GARD and other well-known rare disease related resources including Orphanet,[18] OMIM,[19] MONDO,[20] and curated mappings between FDA orphan designations to GARD, and information on FDA approval status and drug indications from Inxight Drugs,[17] using our stitcher[21] software. Stitcher defines edges to link equivalent/relevant concepts from different resources; for instance, “N_Name” denotes linked concepts with the same concept names, while “L_CODE” denotes linked concepts sharing the same external reference. In addition, stitcher adopts predicates from original resources, such as “R_equivalentClass” from MONDO. More examples are shown in Fig. 2.

B. GBM-based Biomedical Profile Network (GBPN)

Development. To construct the GBPN with GBM-relevant information, we generated a disease cluster pertinent to GBM. This cluster containing GBM and 91 other GBM-related rare diseases was generated through a modified version of DL2Vec[22] applied to data obtained from the NGKG and enriched with additional data sources. Specifically, a focused subgraph of the NGKG was extracted containing disease, genes and phenotypes. The subgraph was annotated with Gene Ontology[23] and Human Phenotype Ontology[24] and then enriched with small molecule and pathway data from Pharos[25] and The Pathway Commons,[26] respectively. Random walks emanating from each rare disease were used to generate a corpus from which disease node embeddings were created. The disease node embeddings were clustered using the k-means algorithm. Detailed description of the disease clustering procedure has been described in a separate submission.[27]

We extracted 92 subgraphs from the NGKG, each an ego graph[28] of radius of 3 centered on a node containing one of those 92 GBM-related rare diseases. Figure 2 shows one subgraph that is centered on the node of Familial Alzheimer Disease, one disease from the GBM-related disease cluster. We then merged the union of these subgraphs to create the GBPN.

Optimization. The NGKG maintains connections among equivalent or relevant concepts from different resources via pre-defined edges, e.g. “N_Name” and “L_CODE” or those adopted predicates, e.g. “R_equivalentClass” and “R_exactMatch”. We optimized the GBPN by merging associated diseases, genes, treatments, etc. with those aforementioned edges into singular nodes, yielding a more condensed graph of nodes with enriched biomedical information for efficient network analysis. Specifically, we optimized the GBPN via these rules: 1) the attributes of merged nodes were concatenated; 2) edges were removed if the connected nodes were merged (i.e. if nodes A and B merged, all edges between A and B would be removed); 3) edges were maintained between unmerged and newly-merged nodes (i.e. if node A and B merged into node AB, an edge from A to node C would be reassigned as an edge from AB to C). The code used to implement rules 1–3 is in the supplemental materials. Synonyms were subsequently filtered out of name labels within newly merged nodes. For instance, if the nodes “Addison’s Disease” and “Adrenal aplasia” were merged, both of these labels (which denote the same disease) would be concatenated within the newly merged node. In this case, we would verify that “Adrenal aplasia” is a synonym of “Addison’s Disease” by querying the NGKG for the “synonyms” attribute of the “Addison’s Disease” node and would subsequently remove “Adrenal aplasia” from the newly merged node’s name label in the GBPN. This process was automated and applied to each newly merged node; some other complementary resources, including the NORD Rare Diseases database,[29] GeneCards,[30] the National Library of Medicine’s MedlinePlus,[31] PubChem,[32] and the National Cancer Institute’s List of Cancer Drugs,[33] were applied for this process as well. Figure 3 illustrates one merging example.

C. mc_GBPN Development

To group the GBPN into focused subgraphs, we clustered the GBPN into modularity classes (mc_GBPN) using community detection[34] available from Gephi 0.9.6.[35] Gephi is an open source tool for creating and exploring interactive network visualizations that includes functions for network analysis. Gephi uses the Louvain modularity algorithm[36] for community detection, which maximizes a modularity score for each community and is well-suited to large networks. [37] We set randomize to “On” and the resolution to 1.0. Smaller resolution values recover more communities (each containing fewer nodes), while larger resolution values recover fewer communities (each containing more nodes).[38] While larger resolution values may fail to separate distinct communities,[39] smaller resolution values may produce communities that are too small to perform meaningful network analysis on. In the case of the GBPN, resolution values less than 1.0 translated to over half of the communities being too small (three nodes or fewer) to analyze. We prioritized the mc_GBPN by modularity score and the top ten mc_GBPN were applied for further investigation. Specifically, we sorted mc_GBPN with more than three nodes in descending order by modularity score. The modularity score of a mc_GBPN is defined as $\frac{L_c}{m} - \gamma \left(\frac{k_c}{2m} \right)^2$ where L_c is the number of intra-community edges for mc_GBPN, k_c is the sum of degrees of the nodes in mc_GBPN, m is the total number of edges across all mc_GBPN, and γ is the resolution parameter (in this case, 1.0).[40–41] A mc_GBPN with a higher modularity score contains more internal-connections and less external-connections, which results in a large number of “hub nodes” with high centrality scores and therefore is of interest to our investigation for drug repurposing. Thus, we sought out mc_GBPN with a high modularity score.

The mc_GBPN were then reviewed and assigned a class label based on parent-child relationships denoted in the NGKG and Disease Ontology.[42] For example, one mc_GBPN containing disease nodes of “Tumor Grade 1,” “Intracranial Cystic Lesion,” “Hemangioblastoma,” “Benign Neoplasm,” etc. was assigned the class label “Abnormal Brain Growths,” as the majority of its nodes are associated with abnormal growths in the brain.

D. Drug repurposing candidate identification

High-influence node identification. We calculated the degree, closeness, betweenness, eigenvector, and PageRank centrality for each node within their respective mc_GBPN. Each centrality measure detects the amount of influence a given node has over the flow of information in the mc_GBPN. Specifically, the degree centrality of a node is the number of edges connected to it.[43] Closeness centrality measures the average distance between a node and all other nodes in its mc_GBPN.[44] Betweenness centrality of a node is the percentage of shortest paths between any other pair of nodes in the graph which include the given node.[45–46] Eigenvector centrality measures the transitive influence of nodes; edges originating from a node with a high eigenvector centrality score contribute more to the score of the node they target than edges originating from a node with a lower eigenvector centrality score. Thus, if a node has a high eigenvector centrality score, it is connected to many other nodes with high eigenvector centrality scores.[47] We used 100 iterations in our eigenvector centrality calculations[48] (though we note that after experimenting with values ranging from 50–200 iterations, number of iterations had a negligible impact on the calculation and particularly did not affect the order of nodes from highest to lowest eigenvector centrality score). Finally, PageRank centrality is a subtype of eigenvector centrality that uses indegree rather than total degree.[49] We used the default probability setting in Gephi of 0.85 and the default epsilon setting 0.001 in our PageRank centrality calculations.[50] Note that all centrality scores will be greater than zero, and that closeness, eigenvector, and PageRank centrality must all be within the range of zero to one.[43–45, 48–49] In general across all metrics, higher centrality scores indicate a node is connected to a greater number of other nodes and/or is more centrally located within the network.

Drug repurposing candidate identification. We ranked the five most influential nodes for each top ranked mc_GBPN by the five aforementioned centrality measures. We manually reviewed and selected the most interesting nodes from prioritized mc_GBPN based on their influence as potential candidates for drug repurposing for GBM.

Results

A. Results of the GBPN

The NGKG contains 3,819,623 nodes and 84,223,681 edges from forty-three different biomedical data resources. Of these, 4,789 nodes and 177,106 edges were extracted and applied to generate the GBPN. After optimization, the GBPN contained 1,466 nodes (538 of which contained the merged information of two or more pre-optimization nodes) and 107,423 edges with average degree 73.276, defined as the total number of edges divided by the total number of nodes. Additional network properties can be found in Table 1.

Table 1
Network properties of the GBPN.

Network Property	Results
Nodes	1,466
Edges	107,423
Average Degree	73.276
Net Diameter ^a	10
Average Path Length ^b	3.751
Graph Density ^c	0.05
^a The net diameter is defined as the maximal distance between all pairs of nodes.[51] ^b The average path length is defined as the average distance between all pairs of nodes.[52] ^c The graph density is a measure of how close the graph is to being complete, and a complete graph contains all possible edges between any two nodes and has a density of 1.[53]	

B. Results of the mc_GBPN

We performed community detection by Louvain modularity[36] on the GBPN, obtaining forty-one mc_GBPN. Brief descriptions including class labels, number of nodes/edges and modularity scores for the ten mc_GBPN with the highest modularity scores are in Table 2. A full list of forty-one mc_GBPN is in the supplemental materials.

Table 2
Descriptions of the ten ranked mc_GBPB with the highest modularity score.

mc_GBPB Index	Class labels	Nodes	Edges	Modularity Score
0	White Matter-Related Conditions	140	14,176	0.100
12	Movement Disorders	209	12,519	0.083
38	Seizures and Epilepsies	132	7,390	0.060
27	Amyotrophic Lateral Sclerosis and Related Conditions	156	5,984	0.052
21	Parkinsonism, Progressive Supranuclear Palsy, Dementia, and Related Conditions	94	5,936	0.048
14	Sensory and Motor Conditions	58	4,887	0.039
22	Neurodegeneration with Brain Iron Accumulation and Dystonias	65	3,904	0.033
16	Developmental Disorders	61	3,430	0.031
2	Dystonias and Mental Health Conditions	52	2,920	0.025
17	Cerebrovascular Conditions	75	2,735	0.024

We identified the five most influential nodes from each of the ten mc_GBPB (Table 2) by each centrality measure. The identified high-influence nodes from the mc_GBPB with an index of 0 are shown in Fig. 4. Centrality scores were normalized to a 0–1 range using the scikit-learn MinMaxScalar preprocessing function fit_transform method.[54] The full list of the five most influential nodes by each centrality measure within these ten mc_GBPB is in the supplemental materials.

Table 3
The top five nodes in mc_GBPB #0 ranked by five centralities.

Degree	Closeness	Betweenness	Eigenvector	PageRank
1 X-linked Adreno-leukodystrophy; ABCD1 WT; VK-0214	Abnormal Eye	X-linked Adreno-leukodystrophy; ABCD1 WT; VK-0214	X-linked Adreno-leukodystrophy; ABCD1 WT; VK-0214	X-linked Adreno-leukodystrophy; ABCD1 WT; VK-0214
2 Autosomal Dominant Leukodystrophy with Autonomic Disease; LMNB1	X-linked Adreno-leukodystrophy; ABCD1 WT; VK-0214	Autosomal Dominant Leukodystrophy with Autonomic Disease; LMNB1	Attention Deficit Hyperactivity Disorder (ADHD)	Paraparesis
3 Muscle Spasticity	Autosomal Dominant Leukodystrophy with Autonomic Disease; LMNB1	Paraparesis	Vision impairment or loss	Facial Abnormality
4 Paraparesis	Progressive Disorder	Muscle Spasticity	Hearing Impairment	Autosomal Dominant Leukodystrophy with Autonomic Disease; LMNB1
5 Cognitive Decline/Impairment	Muscle Spasticity	Cerebro-medullospinal Disconnection	Dementias, Amentia	ABCD1 Gene

C. Drug Repurposing Candidate Identification

We examined the five most influential nodes from the top ten mc_GBPB (Table 2) by their centrality scores as potential candidates for drug repurposing for GBM. We first normalized the centrality scores of the top five nodes by each centrality measure to a 0–1 range using the fit_transform method of the scikit-learn MinMaxScalar preprocessing function.[54] We then calculated a total normalized centrality score (TNCS) for each distinct node. The TCNS of a node is defined as the sum of its normalized centrality scores across degree, closeness, betweenness, eigenvector, and PageRank centralities. The TCNS of a node may range from 0–5, as there are five centrality measures. The nodes with the highest TCNS in each mc_GBPB listed in Table 2 are identified in Table 4. Of the nodes in Table 4, six had the highest centrality scores across all five centrality measures within their respective mc_GBPB.

Table 4
The most influential nodes in the mc_GBPN, selected by their TNCS across all centrality measures.

mc_GBPN Index	Node Description	TNCS
0	● X-linked Adrenoleukodystrophy (X-ALD) ● ABCD1 WT Allele, ABC42, AMN, ACOX1, ECK2921, ECK1408 genes ● VK-0214	4.407
12	● Ataxias, Chorea ● SETX, WASF1, GCH1 genes	5
38	● Absence Epilepsy ● CSTB, GABRA1, EC1 genes	3.986
27	● Amyotrophic Lateral Sclerosis (ALS) ● ALS1, ALS2, IGFALS genes ● Riluzole, Cannabidiol, Stem Cell Therapy	5
21	● Parkinson's Disease ● FBXO7, DCTN1, GBA, PARK1, PARK2, PARK5, PARK6, PARK7, PARK8 genes	4.413
14	● Spastic Paraplegia 10 (SPG10) ● KIF5A gene	3.625
22	● Neurodegeneration with Brain Iron Accumulation 3 (NBIA3) ● FTL gene	5
16	● Rett Syndrome ● MECP2 gene	5
2	● Myoclonus Dystonia (M-D) ● SGCE gene	5
17	● CADASIL and CARASIL Syndromes ● NOTCH3, COL4A1, HTRA1 genes ● Cerebroslysin, Palm Tocotrienol Complex	5

We observed that the most influential nodes in the mc_GBPN are associated with central nervous system conditions,[42] the main disease category GBM belongs to. Many are also genetic disorders and x-linked (e.g. x-linked adrenoleukodystrophy, Rett syndrome,[55] and some forms of Parkinson's disease,[56] amyotrophic lateral sclerosis,[57] chorea,[58] and ataxia).[59]

The high-influence nodes in Table 4 shed light on drug repurposing. For instance, a novel COL4A1 gene variant associated with CADASIL syndrome was recently found to be associated with GBM.[60] Moreover, the NOTCH3 gene (also associated with CADASIL syndrome) is a prognostic factor that promotes glioma cell proliferation, migration, and invasion.[61] Several drugs were identified as potential candidates for GBM, although they have not been clinically administered for GBM. Riluzole, a treatment for amyotrophic lateral sclerosis (ALS), has been shown to be an effective pretreatment that sensitizes glioma to radiation therapy. It also has synergistic effects in combination with select other drugs when used to treat GBM.[62] Cannabidiol, another ALS treatment, sensitizes GBM to TMZ in multiple orthotopic tumor models.[63] Inhalant cannabidiol has also been shown to inhibit the progression of GBM through regulation of the tumor environment.[64] Finally, stem cell therapy has shown potential for treating neuron and glial cell damage in the brain or spinal cord that results from neurological conditions such as GBM.[65] Interestingly, VK-0214 is currently being tested in a clinical trial as a treatment for x-linked adrenoleukodystrophy.[66] VK-0214 is a thyroid beta receptor agonist[67] which induces the ABCD2 gene by binding to and activating the thyroid beta receptor.[68] In *ABCD1* knockout mice, overexpression of ABCD2 via thyroid receptor activation has been shown to decrease the accumulation of very long chain fatty acids (VLCFA).[68] Based on these findings, selective thyroid receptor agonists are being evaluated as a novel treatment for X-ALD, which is characterized by the accumulation of VLCFA.[68] However, inhibition of fatty acid accumulation and oxidation has been shown to reduce GBM proliferation,[69] growth,[70] and survival[71] as well. The fatty acid accumulation-inhibiting effect of VK-0214 may be beneficial in the treatment of GBM. We will perform additional experimental validation. The full list of associations we examined between the nodes in Table 4 and GBM is in the supplemental materials.

Discussion

In this study, we introduced an integrative GBM-based Biomedical Profile Network (GBPN) by integrating heterogeneous types of data, including disease, gene, drug, etc. based on their shared concept characteristics. To further construct focused subgraphs from the GBPN for supporting high-influence node identification for drug repurposing, we derived modularity class-based subnetworks (mc_GBPN) by leveraging community detection, a form of graph clustering algorithm. Through implementing multiple network analysis techniques over the mc_GBPN, we successfully identified multiple high-influence nodes as potential drug repurposing candidates for GBM. This presented framework sheds light on supporting drug repurposing in a more effective manner. While

integrating more data to expand the search space, we organized the data in a more manageable scale with consideration of their relevance from the network view.

A. Observations and Findings

We applied a rare disease cluster consisting of 92 GBM-related diseases to construct the GBPN by exploring data from the NGKG. We optimized the GBPN for integrative rare disease profile generation by merging associated diseases, genes, treatments, etc. into singular nodes based on their shared concept names or external references. This approach allowed us to explore a large scale of GBM-relevant data in a concentrated and scalable form, which effectively supports drug repurposing with lower computational burden as demonstrated in the Results section. As shown in Fig. 3, some level of inference was introduced during the optimization. When we merged Lafora disease, EPM2A, EPM2B and Metformin, we declared the new connections between Metformin and EPM2A and EPM2B based on inference, since there are no existing connections among them obtained from the NGKG. Since PME2 shares different degrees of associations (different numbers of edges) with EPM2A, EPM2B and Metformin, we inferred these four concepts are potentially associated with each other, leading to node merging. The findings from Bisulli et al.[72] proved the inference introduced for this particular case. In the future study, we will attach relevant references gathered from the previously developed scientific annotation knowledge graph,[73] to the merged nodes, as scientific evidence enrichment.

After GBPN optimization, we generated focused subgraphs of the GBPN by performing community detection as a graph clustering algorithm, resulting in a network partitioned into modularity classes (mc_GBPN). mc_GBPN as a set of subgraphs (i.e., clusters) derived from the GBPN were ranked by their modularity scores, which allowed programmatically upgrade those top prioritized clusters for further investigation and downgrade those with lower priority. Our experiments showed that such a strategy did not lose any important information compared to the GBPN, instead more high-influence nodes were exposed in the top ranked clusters for easy extraction. For instance, nine distinct top high-influence nodes derived from the GBPN appear in the top five most influential node lists from their respective mc_GBPN. We calculated the five most influential nodes by each centrality measure in the GBPN and found that seven of the ten most-influential nodes (see Table 3) were included in the resulting list. The remaining nodes (i.e., Spastic Paraplegia 10, Rett syndrome, Myoclonus Dystonia) were present exclusively in the lists of high-influence nodes derived from the mc_GBPN. The complete lists of the five most influential nodes by each centrality measure in the GBPN and in each modularity class of the mc_GBPN are in the supplemental materials.

B. Limitations of This Study

Due to lack of standardization across the biomedical resources that the NGKG sources data from, integrating information from different resources with a high level of precision proved to be a significant challenge. While we optimized the GBPN by merging nodes with closely associated information into a singular node, we were not able to fully automate this process because the data was not represented in a standard form and the nature of the NGKG that does not contain predefined data models, instead a rule-based semi-automatic approach. A more sophisticated harmonization process will be proposed when we obtain data to build the GBPN. For instance, rare diseases from different resources will be harmonized and standardized by using GARD ID, genes with HGNC ID, etc. Additionally, during the step of high-influence node identification, we manually searched for scientific evidence to support our findings. In the future study, we will programmatically query the rare disease-based scientific annotation knowledge graph[73] for evidence collection. In the future study, we will adopt/extend the strategy of network optimization to apply on the datasets with well-defined data models underneath, then we will be able to generate highly condensed graphs by merging nodes/relationships by different concept types.

C. Future Directions

We presented a preliminary analysis of GBM-related data that allowed us to identify potential candidates for drug repurposing to treat the condition. Although scientific evidence has been identified to support our initial findings, experimental validation is necessary to determine whether these candidates would be effective in treating GBM patients in practice. Clinical observations/efficacy regarding those candidates administered for patients with GBM, derived from Electronic Medical Records (EMR) can serve as another layer of validation. We propose to mine clinical data from National COVID Cohort Collaborative (N3C) and the Biomedical Translational Research Information System (BTRIS) at NIH for clinical evidence identification. Our pipeline is modularized as shown in Fig. 1, thus we propose to extend the use of each module. We will expand to other disease areas by starting with other disease clusters and generating corresponding GBPN. We also propose to explore other clustering algorithms besides community detection for focused subgraph generation. Besides the application of drug repurposing we started with, we believe mc_GBPN as a collection of rare disease profiles providing a complete picture of direct and indirect associations to the target disease can be a valuable source to help us understand the etiology of rare diseases.

Conclusion

In this study we presented a preliminary network analysis-based approach to drug repurposing for GBM. We successfully identified several potential candidates

via centrality and community detection calculations, and substantiated the connections between these candidates and GBM. We reinforced the findings of emerging studies into some treatments and also identified a new candidate, VK-0214, that could be potentially repurposed to treat GBM. These findings can guide future experimental validation, which could lead to new, more effective treatments that extend the lifespan of patients living with GBM.

Declarations

Ethical Approval and Consent to participate

None declared

Consent for publication

None declared

Availability of data and materials

The full documentation, codes, and other supplemental data is found on GitHub (https://github.com/ncats/drug_rep/tree/main/Glioblastoma_Subgraph).

Competing interests

None declared

Funding

None declared

Authors' contributions

EM: performed the experiments and wrote the manuscript. JS: generated the GBM disease cluster and edited the manuscript. EAM: participated in the discussion and edited the manuscript; QZ: conceived, designed, and supervised this study and wrote the manuscript. All authors reviewed and approved the manuscript.

Acknowledgements

This project was supported by the intramural program (ZIA TR000410-03) at NCATS, and was conducted while one of the authors, Erin McGowan, was a Data Science Fellow through Coding it Forward. Special thanks to Iyanuoluwa Odebo for valuable suggestions on network analysis

References

1. Glioblastoma. <https://rarediseases.info.nih.gov/diseases/2491/glioblastoma>. Accessed September 11, 2022.
2. Tamimi AF, Juweid M. Epidemiology and Outcome of Glioblastoma. In: De Vleeschouwer S, editor. Glioblastoma [Internet]. Brisbane (AU): Codon Publications; 2017 Sep 27. Chapter 8. Available from: <https://www.ncbi.nlm.nih.gov/books/NBK470003/> doi: 10.15586/codon.glioblastoma.2017.ch8
3. Grech N, Dalli T, Mizzi S, et al. Rising Incidence of Glioblastoma Multiforme in a Well-Defined Population. *Cureus* 2020;12(5):e8195. <https://pubmed.ncbi.nlm.nih.gov/32572354/> (accessed 11 September 2022).
4. Fisher JP, Adamson DC. Current FDA-Approved Therapies for High-Grade Malignant Gliomas. *Biomedicines* 2021;9(3):324. <https://doi.org/10.3390/biomedicines9030324> (accessed 11 September 2022).
5. Stupp R, Mason WP, van den Bent MJ, et al. Radiotherapy plus concomitant and adjuvant temozolomide for glioblastoma. *The New England journal of medicine* 2005;352(10):987–996. <https://doi.org/10.1056/NEJMoa043330> (accessed 11 September 2022).
6. Lomustine. <https://www.cancer.gov/about-cancer/treatment/drugs/lomustine>. Accessed September 11, 2022.
7. Carmustine. <https://www.cancer.gov/about-cancer/treatment/drugs/carmustine>. Accessed September 11, 2022.
8. Carmustine Implant. <https://www.cancer.gov/about-cancer/treatment/drugs/carmustineimplant>. Accessed September 11, 2022.
9. Bevacizumab. <https://www.cancer.gov/about-cancer/treatment/drugs/bevacizumab>. Accessed September 11, 2022.
10. Treatment Options. <https://braintumor.org/brain-tumors/diagnosis-treatment/diagnosis/treatment-options/>. Accessed September 11, 2022.
11. Jourdan JP, Bureau R, Rochais C, et al. Drug repositioning: a brief overview. *The Journal of pharmacy and pharmacology* 2020;72(9):1145–1151. <https://doi.org/10.1111/jphp.13273> (accessed 11 September 2022).
12. Thayer AM. Drug Repurposing. <https://cen.acs.org/articles/90/i40/Drug-Repurposing.html>. Accessed September 11, 2022.
13. Jarada TN, Rokne JG, Alhajj R. A review of computational drug repositioning: strategies, approaches, opportunities, challenges, and directions. *J Cheminform* 2020;12(46). <https://doi.org/10.1186/s13321-020-00450-7> (accessed 11 September 2022).
14. Keng YY, Kwa KH, Ratnavelu K. Centrality analysis in a drug network and its application to drug repositioning. *Applied Mathematics and Computation* 2021;395. <https://doi.org/10.1016/j.amc.2020.125870> (accessed 11 September 2022).
15. Jalili M, Salehzadeh-Yazdi A, Gupta S, et al. Evolution of Centrality Measurements for the Detection of Essential Proteins in Biological Networks. *Frontiers in physiology* 2016;7:375. <https://doi.org/10.3389/fphys.2016.00375> (accessed 11 September 2022).
16. Mohd Ali YE, Kwa KH, Ratnavelu K. Predicting new drug indications from network analysis [abstract]. *International Journal of Modern Physics C* 2017;28(9). <https://ui.adsabs.harvard.edu/abs/2017IJMPC..2850118M/abstract> (accessed 11 September 2022).
17. Zhu Q, Nguyen DT, Grishagin I, et al. An integrative knowledge graph for rare diseases, derived from the Genetic and Rare Diseases Information Center (GARD). *Journal of biomedical semantics* 2020;11(1):13. <https://doi.org/10.1186/s13326-020-00232-y> (accessed 11 September 2022).
18. Orphanet: an online rare disease and orphan drug data base. Copyright, INSERM 1999. Available on <http://www.orpha.net> (accessed 13 September 2022).
19. Online Mendelian Inheritance in Man, OMIM®. McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University (Baltimore, MD). <https://omim.org/> (accessed 13 September 2022).
20. Mungall CJ, McMurtry JA, Köhler S, et al. The Monarch Initiative: An Integrative Data and Analytic Platform Connecting Phenotypes to Genotypes across Species. *Nucleic Acids Research* 2017;45(D1):D712–22. <https://academic.oup.com/nar/article/45/D1/D712/2605791> (accessed 11 September 2022).

21. Stitcher. <https://github.com/ncats/stitcher>. Accessed September 11, 2022.
22. Chen J, Althagafi A, Hoehndorf R. Predicting candidate genes from phenotypes, functions and anatomical site of expression. *Bioinformatics (Oxford, England)* 2021;37(6),853–860. <https://doi.org/10.1093/bioinformatics/btaa879> (accessed 11 September 2022).
23. Ashburner M, Ball CA, Blake JA, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature genetics* 2000;25(1),25–29. <https://doi.org/10.1038/75556> (accessed 11 September 2022).
24. Robinson PN, Köhler S, Bauer S, et al. The Human Phenotype Ontology: a tool for annotating and analyzing human hereditary disease. *American journal of human genetics* 2008;83(5),610–615. <https://doi.org/10.1016/j.ajhg.2008.09.017>. (accessed 11 September 2022).
25. Nguyen DT, Mathias S, Bologa C, et al. Pharos: Collating protein information to shed light on the druggable genome. *Nucleic Acids Res.* 2017;45(D1):D995-D1002. doi:10.1093/nar/gkw1072. (accessed 27 September 2022).
26. Cerami EG, Gross BE, Demir E, et al. Pathway Commons, a web resource for biological pathway data. *Nucleic Acids Res.* 2011;39(Database issue):D685-D690. doi:10.1093/nar/gkq1039. (accessed 27 September 2022).
27. J. Sanjak, E. Mathe, and Q. Zhu, "Clustering rare diseases within an ontology-enriched knowledge graph," *bioRxiv*, p. 2023.02. 15.528673, 2023.
28. Ego Graph. https://networkx.org/documentation/stable/reference/generated/networkx.generators.ego.ego_graph.html. Accessed September 11, 2022.
29. Rare Disease Database. <https://rarediseases.org/for-patients-and-families/information-resources/rare-disease-information/>. Accessed September 11, 2022.
30. GeneCards: The Human Gene Database. <https://www.genecards.org/>. Accessed September 11, 2022.
31. MedlinePlus. <https://medlineplus.gov/>. Accessed September 11, 2022.
32. PubChem. <https://pubchem.ncbi.nlm.nih.gov/>. Accessed September 11, 2022.
33. A to Z List of Cancer Drugs. <https://www.cancer.gov/about-cancer/treatment/drugs>. Accessed September 11, 2022.
34. Khan BS, Niazi MA. Network Community Detection: A Review and Visual Survey. 2017. <https://arxiv.org/pdf/1708.00977.pdf> (accessed 11 September 2022).
35. Gephi 0.9.6. <https://github.com/gephi/gephi/releases/tag/v0.9.6>. Accessed September 11, 2022.
36. Blondel VD, Guillaume JL, Lambiotte R, et al. Fast unfolding of communities in large networks. *J. Stat. Mech.* 2008;P10008. <https://iopscience.iop.org/article/10.1088/1742-5468/2008/10/P10008> (accessed 11 September 2022).
37. Louvain. <https://neo4j.com/docs/graph-data-science/current/algorithms/louvain/>. Accessed September 11, 2022.
38. Louvain Clustering. <https://orange3.readthedocs.io/projects/orange-visual-programming/en/latest/widgets/unsupervised/louvainclustering.html>. Accessed September 11, 2022.
39. Fortunato S, Barthélemy M. Resolution limit in community detection. *Proceedings of the National Academy of Sciences* 2007;104(1)36-41. <https://doi.org/10.1073/pnas.0605965104> (accessed 15 September 2022).
40. Modularity. <https://networkx.org/documentation/stable/reference/algorithms/generated/networkx.algorithms.community.quality.modularity.html#networkx.algorithms.community.quality.modularity>. Accessed September 12, 2022.
41. Clauset A, Newman MEJ, Moore C. Finding community structure in very large networks. *Phys. Rev. E* 2004;70,066111. <https://doi.org/10.1103/PhysRevE.70.066111> (accessed 12 September 2022).
42. Disease Ontology. <https://disease-ontology.org/>. Accessed September 12, 2022.
43. Degree. <https://github.com/gephi/gephi/wiki/Degree>. Accessed September 12, 2022.
44. Closeness Centrality. <https://github.com/gephi/gephi/wiki/Closeness-Centrality>. Accessed September 12, 2022.
45. Betweenness Centrality. <https://github.com/gephi/gephi/wiki/Betweenness-Centrality>. Accessed September 12, 2022.
46. Brandes U. A Faster Algorithm for Betweenness Centrality. *Journal of Mathematical Sociology* 2001;25(2):163-177. <http://snap.stanford.edu/class/cs224w-readings/brandes01centrality.pdf> (accessed 12 September 2022).
47. Eigenvector Centrality. <https://neo4j.com/docs/graph-data-science/current/algorithms/eigenvector-centrality/#:~:text=Eigenvector%20Centrality%20is%20an%20algorithm,who%20themselves%20have%20high%20scores>. Accessed September 12, 2022.
48. Eigenvector Centrality. <https://github.com/gephi/gephi/wiki/Eigenvector-Centrality>. Accessed September 12, 2022.
49. Page L, Brin S, Motwani R, et al. The PageRank Citation Ranking: Bringing Order to the Web. *Technical Report, Stanford InfoLab* 1999. <http://ilpubs.stanford.edu:8090/422/> (accessed 12 September 2022).
50. PageRank. <https://github.com/gephi/gephi/wiki/PageRank>. Accessed September 12, 2022.
51. Diameter. <https://github.com/gephi/gephi/wiki/Diameter>. Accessed September 12, 2022.
52. Average Path Length. <https://github.com/gephi/gephi/wiki/Average-Path-Length>. Accessed September 12, 2022.
53. Graph Density. <https://github.com/gephi/gephi/wiki/Graph-Density>. Accessed September 12, 2022.
54. MinMaxScalar. <https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.MinMaxScaler.html>. Accessed September 12, 2022.
55. Rett syndrome. <https://rarediseases.info.nih.gov/diseases/5696/rett-syndrome>. Accessed September 12, 2022.
56. Evidente VGH. X-Linked Dystonia-Parkinsonism. 2005 Dec 13 [Updated 2018 Feb 15]. In: Adam MP, Everman DB, Mirzaa GM, et al., editors. GeneReviews® [Internet]. Seattle (WA): University of Washington, Seattle; 1993-2022. Available from: <https://www.ncbi.nlm.nih.gov/books/NBK1489/> (accessed 13 September 2022).

57. Chen S, Sayana P, Zhang X, et al. Genetics of amyotrophic lateral sclerosis: an update. *Mol Neurodegener* 2013;8:28. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3766231/> (accessed 12 September 2022).
58. Martinez-Ramirez D, Walker RH, Rodriguez-Violante M, et al. Review of Hereditary and Acquired Rare Chorea. *Tremor and Other Hyperkinetic Movements* 2020;10:24. <http://doi.org/10.5334/tohm.548> (accessed 13 September 2022).
59. Zanni G, Bertini E. X-linked ataxias. *Handb Clin Neurol*. 2018;155:175-189. doi:10.1016/B978-0-444-64189-2.00011-1 (accessed 13 September 2022).
60. Muto K, Miyamoto R, Terasawa Y, et al. A novel *COL4A1* variant associated with recurrent epistaxis and glioblastoma. *Hum Genome Var* 2021;8:18. <https://doi.org/10.1038/s41439-021-00150-0> (accessed 13 September 2022).
61. Alqudah MAY, Agarwal S, Al-Keilani MS, et al. NOTCH3 Is a Prognostic Factor That Promotes Glioma Cell Proliferation, Migration and Invasion via Activation of CCND1 and EGFR. *PLoS ONE* 2013;8(10):e77299. <https://doi.org/10.1371/journal.pone.0077299> (accessed 13 September 2022).
62. Blyufer A, Lhamo S, Tam C, et al. Riluzole: A neuroprotective drug with potential as a novel anti-cancer agent (Review). *International Journal of Oncology* 2021;59:95. <https://doi.org/10.3892/ijo.2021.5275> (accessed 13 September 2022).
63. Soroceanu L, Singer E, Dighe P, et al. Cannabidiol inhibits RAD51 and sensitizes glioblastoma to temozolomide in multiple orthotopic tumor models. *Neurooncol Adv*. 2022;4(1):vdac019. <https://doi.org/10.1093/noajnl/vdac019> (accessed 13 September 2022).
64. Khodadadi H, Salles ÉL, Alptekin A, et al. Inhalant Cannabidiol Inhibits Glioblastoma Progression Through Regulation of Tumor Microenvironment [published online ahead of print, 2021 Dec 16]. *Cannabis Cannabinoid Res*. 2021;10.1089/can.2021.0098. <https://doi.org/10.1089/can.2021.0098> (accessed 13 September 2022).
65. Ebrahimi T, Abasi M, Seifar F, et al. Transplantation of Stem Cells as a Potential Therapeutic Strategy in Neurodegenerative Disorders. *Current Stem Cell Research & Therapy* 2021;16(2). <https://dx.doi.org/10.2174/1574888X15666200628141314> (accessed 13 September 2022).
66. A Study to Assess the Pharmacodynamics of VK0214 in Male Subjects With AMN. <https://clinicaltrials.gov/ct2/show/NCT04973657>. Accessed September 13, 2022.
67. Pipeline Overview. <http://www.vikingtherapeutics.com/pipeline/overview/>. Accessed September 13, 2022.
68. Jones RS, Chen EC. Transporters as therapeutic targets in human diseases. In: You G, Morris ME, eds. *Drug Transporters: Molecular Characterization and Role in Drug Disposition*, 3rd Edition. Wiley 2022:497-514.
69. Taïb B, Aboussalah AM, Moniruzzaman M, et al. Lipid accumulation and oxidation in glioblastoma multiforme. *Sci Rep*. 2019;9(1):19593. Published 2019 Dec 20. doi:10.1038/s41598-019-55985-z (accessed 13 September 2022).
70. Sperry J, Condro MC, Guo L, et al. Glioblastoma Utilizes Fatty Acids and Ketone Bodies for Growth Allowing Progression during Ketogenic Diet Therapy. *iScience* 2020;23,101453. [https://www.cell.com/iscience/pdf/S2589-0042\(20\)30645-3.pdf](https://www.cell.com/iscience/pdf/S2589-0042(20)30645-3.pdf) (accessed 13 September 2022).
71. Juraszek B, Czamecka-Herok J, Nałęcz KA. Glioma cells survival depends both on fatty acid oxidation and on functional carnitine transport by SLC22A5. *J. Neurochem*. 2021;156:642– 657. <https://doi.org/10.1111/jnc.15124> (accessed 13 September 2022).
72. Bisulli F, Muccioli L, d'Orsi G, et al. Treatment with metformin in twelve patients with Lafora disease. *Orphanet journal of rare diseases* 2019;14(1),149. <https://doi.org/10.1186/s13023-019-1132-3> (accessed 15 September 2022).
73. Zhu Q, Qu C, Liu R, et al. Rare Disease based Scientific Annotation Knowledge Graph. *Front. Artif. Intell*. 2022;5:932665. doi: 10.3389/frai.2022.932665 (accessed 13 September 2022).

Figures

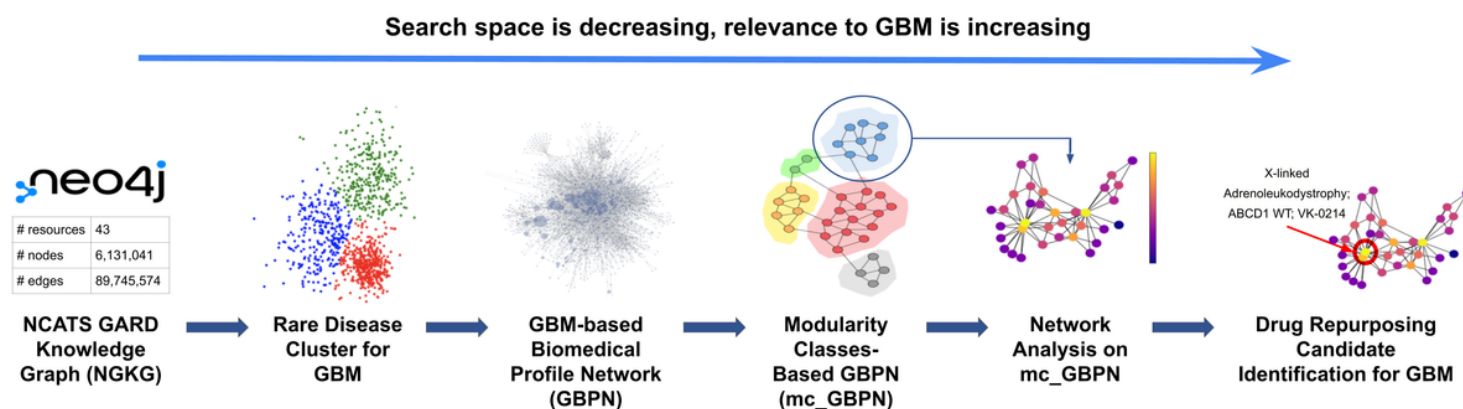


Figure 1

Workflow for identifying drug repurposing candidates for GBM.

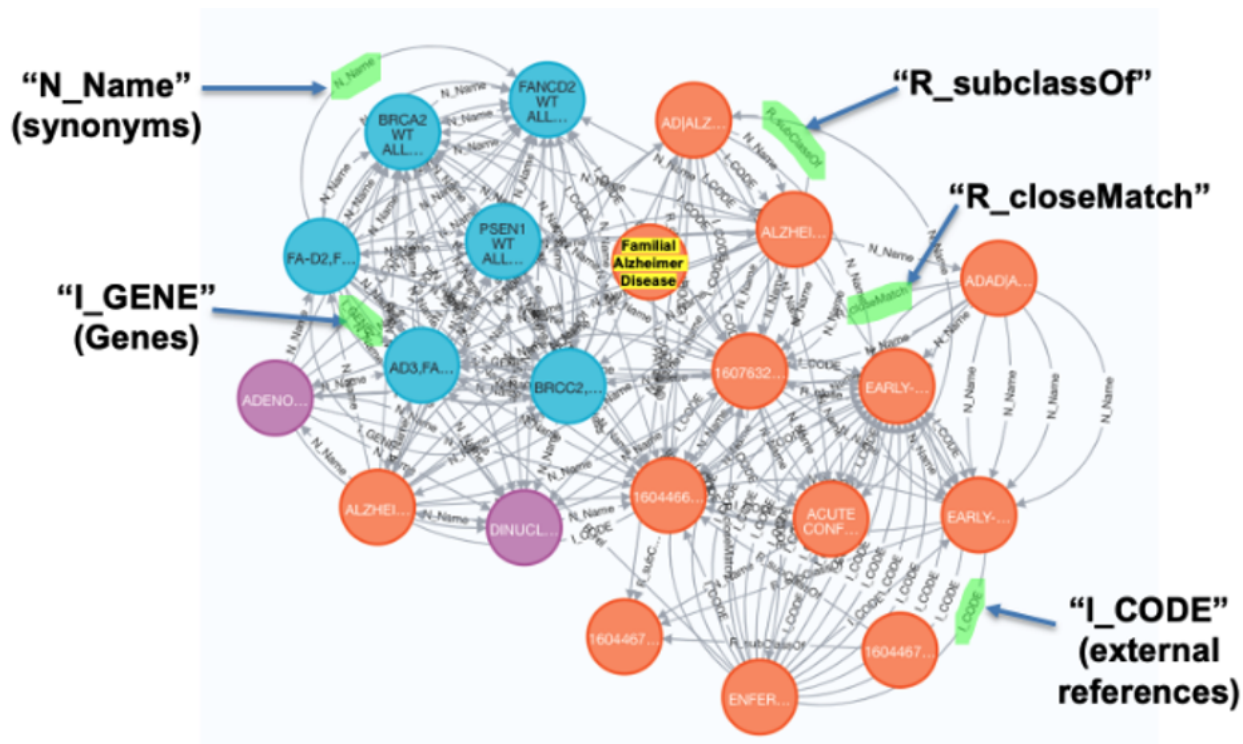


Figure 2
 Familial Alzheimer Disease-based subgraph derived from the NGKG. Orange nodes denote diseases, blue nodes denote genes, and purple nodes denote drugs. Familial Alzheimer Disease is highlighted in yellow.

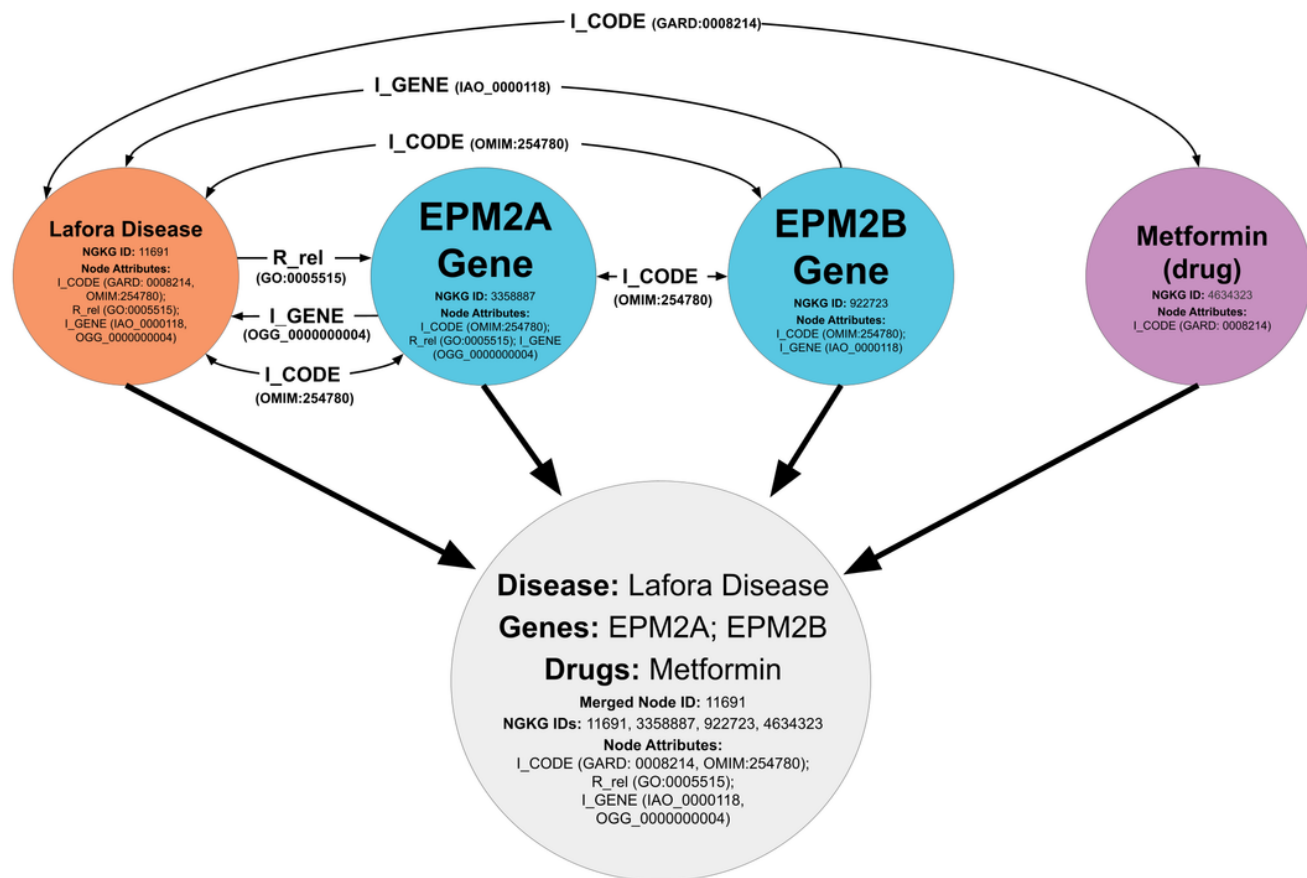


Figure 3

A node containing Lafora disease is merged with nodes connected to it by an edge label of "L_CODE": two Lafora disease related genes, the EPM2A gene, the EPM2B gene, and Metformin, a treatment that has been used for Lafora disease. The gray node is one of the merged nodes in the GBPN.

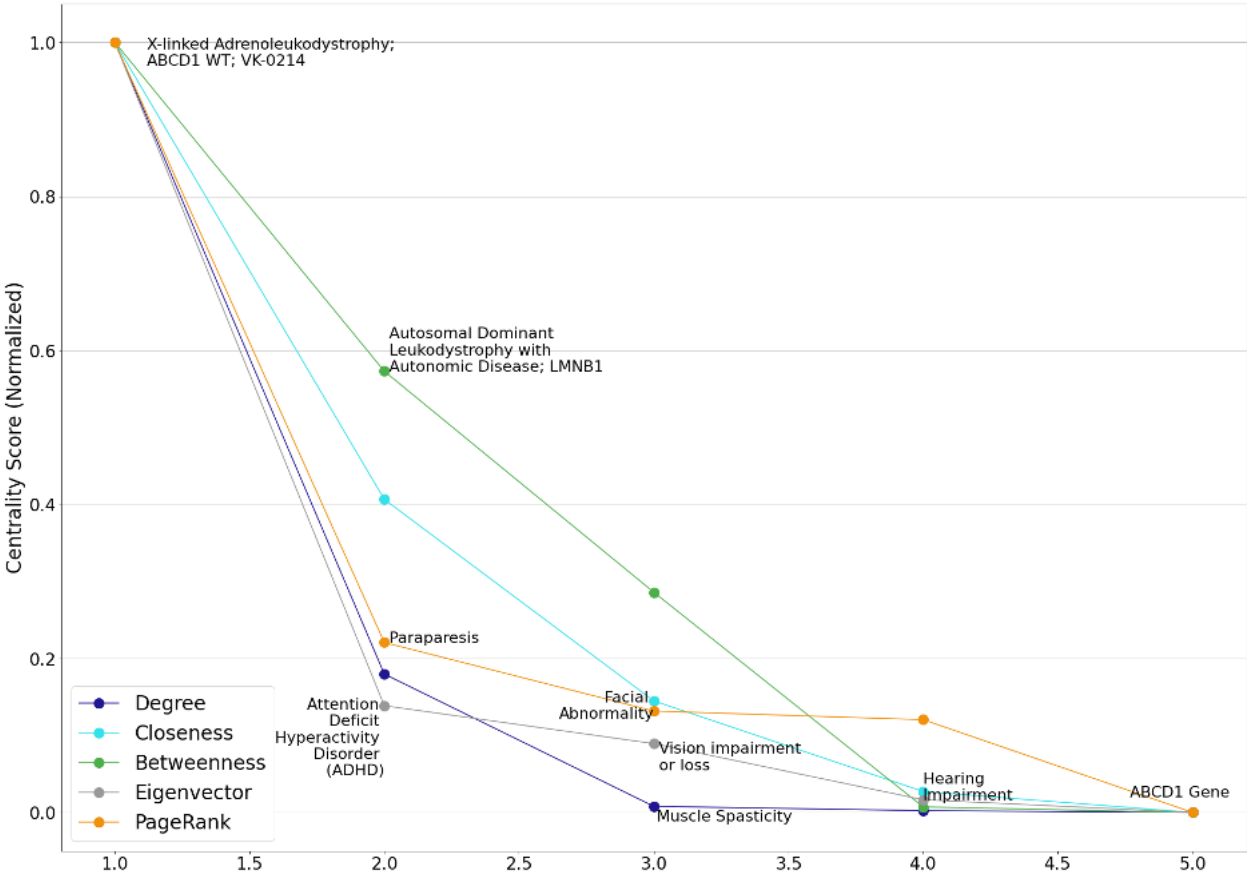


Figure 4

High-influence nodes identified by degree, closeness, betweenness, eigenvector, and PageRank centrality in mc_GBPN with an index of 0. The nodes displayed have a strong relationship to white matter-related conditions (as does GBM). Note that several nodes have high centrality scores across multiple measures; these nodes have a higher potential for drug repurposing.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [TopCandidateReferenceList.xlsx](#)
- [mcGBPNModularityClassDescriptions.pdf](#)