Network Analysis: Cardiovascular Disease Subtype Association and Ion Channel Proteins

Irsyad Adam[a], **Seung Chung**[a], **Andy Goh**[b], **and Ethan Young**[a, c]

[a]Department of Mathematics, University of California, Los Angeles, CA 90095; [b]Department of Computer Science, University of California, Los Angeles, CA 90095; [c]Department of Statistics, University of California, Los Angeles, CA 90095

This manuscript was compiled on February 6, 2023

Many scientific literatures from cross-sectional studies to clinical research attempt to unravel the causes of CVD, including the ion channel proteins that are involved in such diseases. However, there seems to be a lack of ways to quantify and formalize the strength of correlations between these proteins and the affected CVDs. One approach we attempt to formalize this measure is to model proteins, research papers, and CVDs using a network in order to compute various network measures that may encapture such correlations. Namely, we explored common neighbor scores, community-wise betweenness centralities, and the distance-from-centroid in k-means clustering. Our results suggest that the community-wise betweenness centralities align most appropriately with previous research on the importance and prominence of certain proteins in CVDs.

cardiovascular disease | networks | community detection | graph embedding

To correlate proteins with CVDs, we first study the general types of heart disease: valve disease, aneurysm, arrhythmia, cardiomyopathy, pericarditis, heart failure, and coronary artery disease. These can be categorized into two pathological issues: structural changes and signaling error. When a small quantitative changes take place within a protein, they can be qualitatively visualized within the structure, which has a strong relationship with the function. Within the set of proteins, ion channel proteins can be considered the most essential to humans since they allow parts of the body to communicate with others. Namely, cardiac muscles require each cell to communicate precisely to create a uniform beat that starts from one cell and pulsates throughout the heart. Moreover, Abriel et al. (1) conducted a review of research in the last 20 years showing that cardiac ion channel may function as part of larger macromolecular complexes that play a role in gene transcription, amino acid translation, and oligomerization.

Muscle contraction is undoubtedly the most essential function of the heart. For contraction to occur, the myosin heads must be attached to the actin through a middle man: troponin complex. Then, the head can pull the actin through the bond energy in Adenosine Triphosphate (ATP). This is a highly ordered, multi-step mechanism that is sensitive to the external conditions; thus, it is justified to state that the troponin complex is of high importance when studying CVDs. Furthermore, Ruan et al. (2) conducted research on CVD and associated risk factors among older adults in low to middle income countries, and discovered that ischemic heart disease (a subset of coronary heart disease) and stroke were the top two leading causes of CVD health lost in each world region. As previous research has suggested the importance of the troponin complex and its correlation to coronary heart disease, we use this relation as one of the baselines in analyzing the validity of our network measures.

Network Background. The network created for this study consists of 14772 nodes of five types: 176 nodes that contain Medical Subject Headings (MeSH), 142 nodes that contain unique Cardiovascular drug information, 13495 nodes that contain PubMed document metadata, 424 nodes that contain data for unique ion channel proteins, and 535 nodes that contain biological pathway information. To first assemble the network, all MeSH terms that pertain to subtypes of cardiovascular diseases are scraped from the MeSH tree and deployed in the network (3). These MeSH terms are what allows us to integrate CVD subtypes into our network. Next, all PubMed documents that correspond to these subject headings are pulled (4), their metadata collected and deployed as nodes, and then connected to their respective MeSH term using the relationship "ASSIGNS." After these two types of nodes are connected, all of the ion channel proteins were extracted from the UniProtKB knowledge base (5), deployed as nodes, and connected to the PubMed document nodes that mentioned that specific protein using the relationship "MENTIONS." Next, drugs were collected from the DrugBank database (6) and filtered by their category, leaving only cardiovascular drugs. These drugs were then deployed along with their metadata as nodes and connected to the ion channel proteins that they target using the relationship "TARGET". However, there are some drug target proteins that are not ion channel proteins. These specific target proteins are then deployed in the graph if they share a biological pathway with an ion channel protein. Otherwise, these drug targets are left out. Finally, biological pathways of all of the proteins in the network, whether they are ion

Significance Statement

Cardiovascular disease (CVD) is a global epidemic that affects the lives of millions. Moreover, there has recently been multiple cases of heart inflammation associated with COVID vaccines. Medical research has studied the possible causes of this inflammation, but the novelty of the vaccine limits our knowledge about this CVD side effect. Our project analyzed the strength of the relationships between proteins and different types of CVD's by representing the information as a network. These correlation scores can be the starting point of further research into high-impact proteins in CVDs.

Author contributions: I.A., S.C., A.G., and E.Y. wrote the paper and the slide deck; I.A., S.C., A.G., and E.Y. analyzed results; A.G. and E.Y. examined limitations of the methods; I.A. compiled the data, constructed the network, and implemented common neighbors, FastRP, t-SNE, and k-means; S.C. conducted a literature search and researched biological implications of the results; A.G. implemented and analyzed community structure and within-community betweenness centrality; E.Y. implemented and analyzed degree distribution and betweenness centrality, helped I.A. with the details of FastRP and t-SNE and with interpretation of the results.

The authors declare no conflict of interest.

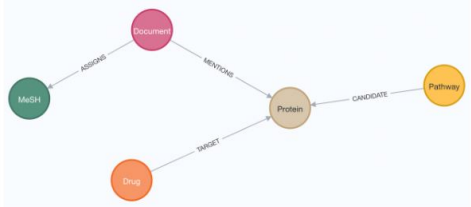

Fig. 1. Initial Directed Schema

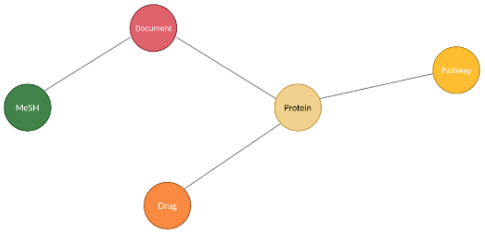

Fig. 2. Schema after Editing Edges

channel proteins or not, are extracted from Reactome (7) and deployed as nodes using their Reactome ID, and linked to the proteins that are contained in them using the "CANDIDATE" relationship. Thus, a connected graph is formed with MeSH terms, PubMed documents, ion channel proteins, drugs, and biological pathways. The overall schema in the network is seen in Fig. 1.

After more consideration, we decided to ignore the directions and labels of the relationships connecting each node, leading to an undirected network with a single edge type. This is due to the fact that the types of relationships or directions of these relationships had no biological significance and was only a placeholder for clarification. Thus, it was more logical to include a single undirected edge type (Fig. 2), instead of introducing biases into network analysis.

Methods

From the created network, we first looked at the degree distribution and betweenness centralities for the entire network. Then, we studied three potential metrics: common neighbor score, community-wise betweenness, and distance-from-centroid in a K-means clustering. Neo4j, a graph database system, was used to build this network, run analytics, and visualize results.

Degree Distribution. We begin our analysis by looking at the degree distribution of our network. Specifically, we look at the degree distributions of each node type and compare their variance as well as largest and smallest degrees.

Betweenness Centrality. To formulate some metric to quantify the importance of protein and MeSH nodes, we elected to use betweenness centrality, given by Eq. 1. The betweenness centrality of node i is defined as

$$x_i = \sum_{st} \frac{n_{st}^i}{g_{st}}, \qquad [1]$$

where n_{st}^i is the number of shortest paths from s to t that include node i, and g_{st} is the total number of shortest paths from s to t. By convention, n_{st}^i/g_{st} is zero if both are zero (8). We chose this measure because we thought that out of all the centrality measures learned in class (e.g., Katz, PageRank, eigenvector), betweenness centrality fit the best as seeing which proteins are the most influential in connecting MeSH terms to certain pathways and drugs that target that protein.

Common Neighbors. To formulate a simple measure of the correlation between proteins and MeSH terms, we focus on the proteins and MeSH terms with the highest betweenness centrality scores and count the number of common neighbors between each protein-Mesh term pair. The number of common neighbors of nodes i and j is given by

$$n_{ij} = \sum_k A_{ik} A_{kj}, \qquad [2]$$

which is the same as the ijth element of A^2 (8).

Community Detection. As opposed to common neighbors that can only consider correlations between pairs of proteins and MeSH terms, looking at the communities within the network allows for the groupings of multiple proteins and MeSH terms. This is important as CVDs are often the culminations of malfunctions in multiple proteins together, and not the result of any one, singular protein.

For this project, the Louvain algorithm was used to detect communities, a greedy algorithm that recursively builds up communities while maximizing modularity over the partitions of the network (8). In particular, this algorithm was chosen as its main advantage is its relatively short computation time. Comparing against different algorithms provided by Neo4j (such as Label Propagation), it was also experimentally determined that Louvain ran the quickest while producing similar results.

After the communities are formed by the Louvain algorithm, these communities were partitioned into disjoint subgraphs. Then, within each subgraph, the betweenness centrality for the proteins and MeSH terms were computed. Note that because the graph was partitioned into communities, the betweenness centrality calculation (Eq. 1) will have the restriction that s, t only include nodes in the same community as i.

Graph Embeddings and K-means Clustering. To further explore any potential structure in the network, we employ graph embeddings and K-means clustering. We use graph embeddings to reduce the dimension of our data, while preserving as much information (e.g., distance) as possible. Our primary goal here is to see if distances between data points (nodes) as a potential metric to measure correlation.

To create graph embeddings of our network, we implement an algorithm in Neo4j called FastRP, or Fast Random Projection. This algorithm is a random-walk-based-method of growing popularity proposed by H. Chen et al. and has strong theoretical guarantees (9) from the Johnson-Lindenstrauss lemma, which states that a set of points in n-dimensional space can be projected into a d-dimensional space, where $d \ll n$, such that the distances between the points are approximately preserved (10). This result forms the basis of the FastRP algorithm as well as many other dimension reduction techniques.

In summary, this algorithm first computes a series of intermediate embeddings N_k in Eq. 3, where A is the random-walk

Laplacian, L a normalization matrix that reduces the influence of high-degree nodes, and $R \in \mathbb{R}^{nxd}$ a random projection matrix where its entries are sampled i.i.d. (independently and identically distributed) with 0 mean from some distribution (e.g., Gaussian), $\tilde{A}^k = (A \cdot ... \cdot A \cdot L)$, n is the number of rows (nodes) in the network, and d is the number of embedding dimensions.

$$N_k = (A \cdot ....)(A \cdot L \cdot R) = \tilde{A}^k \cdot R \quad [3]$$

Eq. 4 shows the computation of the final embedding matrix $N \in \mathbb{R}^{nxd}$.

$$N = \alpha_1 N_1 + \alpha_2 N_2 + ... + \alpha_k N_k \quad [4]$$

Here, the constants α_1, α_2, ..., α_k are weights that tell us how much influence the neighbors k-steps away of a node has. Each row of N is the node where the columns are the approximate distances preserved). Furthermore, k tells us how much of the neighborhood around a node is included in the embeddings (e.g., if $k = 3$, then we cannot randomly move to a node that is 4 or more edges away). For a detailed proof, see (9).

For our purposes, we chose our embedding dimension to be 256, as suggested in the Neo4j documentation. This choice is a balance between preserving higher-order information and computational tractability. We then visualize our dimension-reduced data in two dimensions using t-distributed stochastic neighbor embeddings, a method originally developed by Hinton and Rowels (11). The choice to use this method is arbitrary, as another dimension reduction technique such as principal component analysis (PCA) could have been used instead.

We employ graph embeddings to use K-means and attempt to cluster different MeSH terms and proteins together, quantifying correlation with distance. To do this, the initialization of multiple clusters is needed. The number of clusters was initially based of the number of cardiovascular subtypes. However, due to this number being only an intuition, we used a heuristic known as a distortion score elbow method to find the optimal number of centroids. This heuristic helps find the number of clusters that neither overfits nor underfits the data. Thus, in Fig 8, the distortion score, or the total sum of the squared distance between the points in a cluster and the centroid, is plotted against the number of clusters to see the optimal number of clusters. Using the elbow method, we noticed that the "elbow" is roughly 7 clusters.

After specifying 7 clusters and running the algorithm, the MeSH terms for each cluster were carefully collected to analyze the type of cardiovascular disease for each cluster. Using the cardiovascular MeSH terms corresponding to each cluster, the 7 sub-types of cardiovascular disease were assigned to the 7 clusters as seen in 9. Since the purpose of this clustering was to find a correlation between different types of cardiovascular disease and ion channel proteins, the centroids of each CVD cluster were plotted, along with the MeSH terms and ion channel proteins, as seen in Fig. 10.

Results

Degree Distribution. As we can see in Fig. 3, the degree distributions of the "Protein" and "MeSH" nodes share the highest degree variance and also share very similar maximum degrees. For our project, we focus on these nodes to try to gain some insights into their importance as well as the structure of the network as a whole.

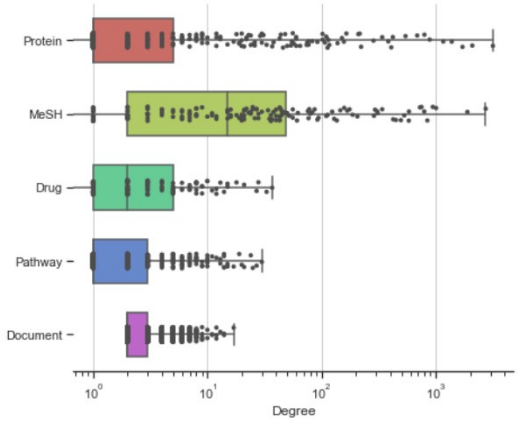

Fig. 3. Degree Distribution of each Node Type

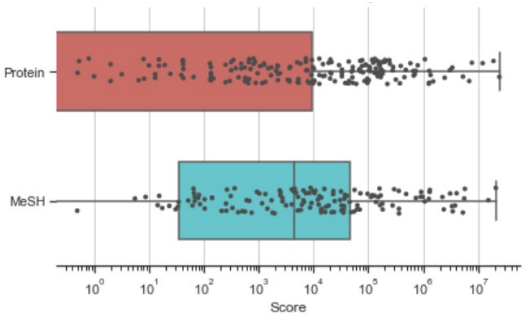

Fig. 4. Box Plot of Betweenness Centrality Scores

Betweenness Centrality. We focused on the nodes with the highest betweenness centrality scores for common neighbors, as seen in Fig. 4. We see that due to the structure of our network, many protein and MeSH nodes have very high betweenness centrality scores. Also of interest are the protein and MeSH terms with the highest betweenness centrality scores, which suggests their prevalence in scientific studies.

Common Neighbors. We visualize the common neighbors score between proteins and MeSH terms with a heatmap, as seen in Fig. 5. For the purposes of visual clarity, we focus on the top 14 proteins and the top 20 MeSH terms. We see a clear outlier between the protein Troponin I and the CVD myocardial infarction. Their large number of common numbers suggests that both feature prominently in scientific studies. Apart from that value, we see that the row for Troponin I has generally has higher numbers of common neighbors than other proteins, which also suggests its importance as a protein of interest in the study of cardiovascular diseases.

Community Detection. The Louvain algorithm outputted around 100 communities, but we decided to filter out the communities with sizes less than 10 for ease of analysis, as these small communities do not give much information about the correlations. This left 21 communities with sizes greater than 10. An example of one such community (with only its proteins and MeSH terms shown) and its community-wise betweenness centralities are shown in Table 1 and visualized in Fig. 6.

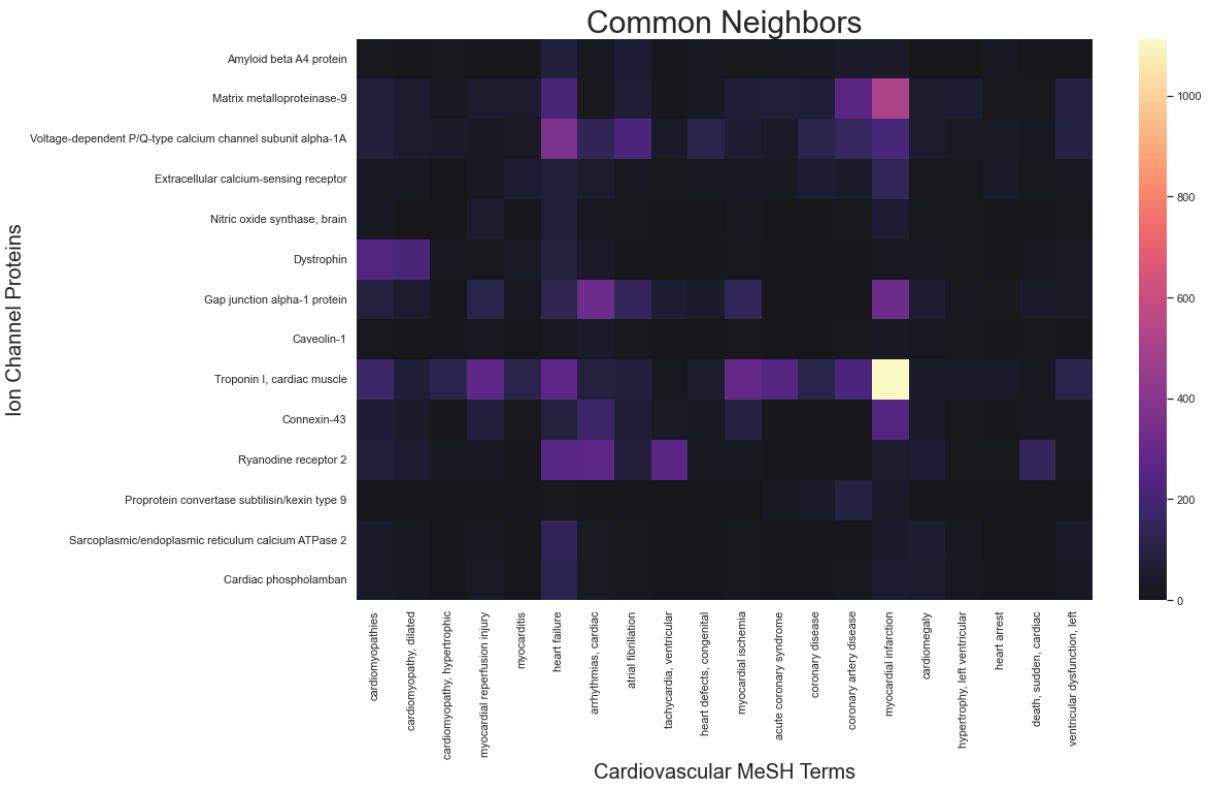

Fig. 5. Heatmap of Common Neighbor Score

Table 1. Community-wise betweenness score for the cardiomyopathy community

Type	Name	Betweenness
MeSH	cardiomyopathies	341929.8
Protein	Dystrophin	320870.7
MeSH	cardiomyopathy, dilated	79902.9
MeSH	barth syndrome	4787.9
Protein	Selenoprotein N	4735.0
MeSH	sarcoglycanopathies	1893.6
MeSH	glycogen storage disease type iib	1623.0
MeSH	chloride intracellular channel protein 4	106.8

Clustering with K-Means. In Fig. 7, the graph embeddings are visualized to help show the structure of the network in 2D space. Here, the points are also color coded by type for clarity. In Fig. 9, K-Means was done on the t-SNE data and then plotted to show the varying clusters. The clusters are also labeled with cardiovascular disease sub-type and their centroids are plotted, showing how different MeSH, protein, and pathway points correspond to each sub-type by varying distances. Finally, in Fig. 10, the centroids for each cluster are plotted and labeled with their corresponding CVD subtype. In addition, the protein and MeSH term nodes are highlighted to easier understand the spread of the nodes in regards to the centroids of the sub-types of CVD.

Discussion

Biological Interpretation.

Betweenness Centrality. Troponin-I received the highest betweenness centrality score within the set of all cardiovascular proteins. Indeed, Troponin-I, or Troponin Inhibitor, is a subunit of the troponin complex within the cardiac cell. It binds to the actin to hold the actin-troponin complex together. When calcium binds to the troponin complex, it causes a conformation change to the actin-troponin complex which pushes Troponin-I away such that the myosin heads can bind to the actin for contraction. Hence, a loss- or gain-of-function by Troponin-I can bring about life threatening conditions because heart would not be able to contract. Thus, betweenness centrality aligned with the knowledge from previous research that proteins from the Troponin complex (Troponin-I, in this case) are essential to heart functions.

Common Neighbor. The network's common neighbor depicts that Troponin-I is strongly associated with myocardial infarction (MI). John Hopkins Medicine defines MI as a disease that occurs when one or more areas of the heart muscle don't get enough oxygen. Coronary artery disease (CAD) happens when the arteries of the heart becomes obstructed by fatty deposits such that blood has difficulty passing through. Since blood is the cell's source of oxygen, there exists a high correlation between CAD and MI. In fact, Ojha et al. (12) reviewed the basic pathophysiology of myocardial infarction and found that most MIs are due to underlying CAD.

This implies that Troponin-I must have a strong correlation with CAD, since MI is associated with CAD. The heat map (Fig. 5) also shows that Troponin-I has a direct correlation with both CAD and MI, but the correlation to CAD is not as prominent as to MI. As MI is the more specific term and more closely related to Troponin-I, this difference has biological significance.

Community Detection. Our communities should be categorized into different types of cardiovascular disease by grouping similar proteins and MeSH terms. One of our communities contains

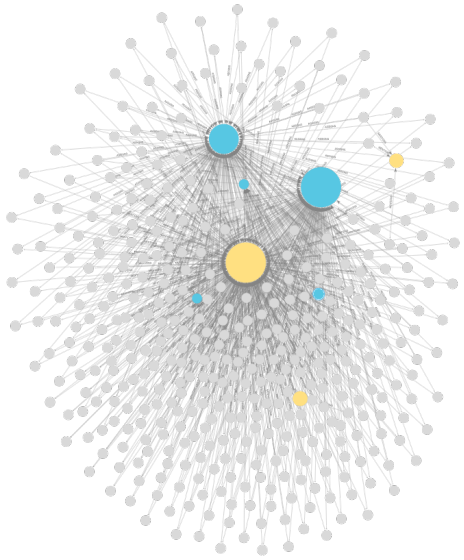

Fig. 6. Visualization of proteins (yellow) and MeSH terms (blue) in the cardiomyopathy community with documents greyed out, where nodes are sized by betweenness centrality

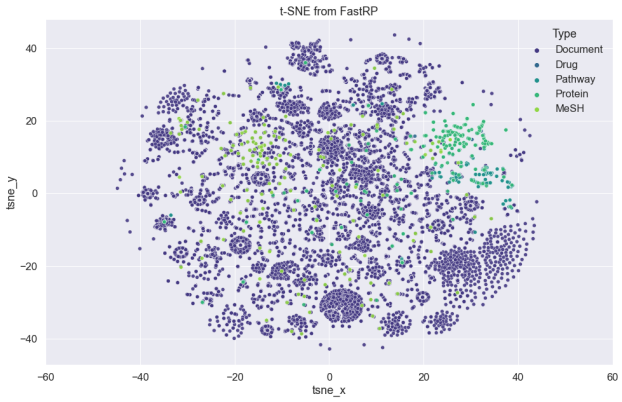

Fig. 7. t-SNE of Embedded Nodes

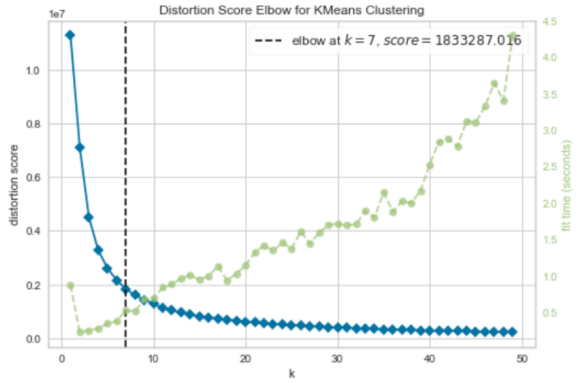

Fig. 8. Distortion Score against Clusters

Fig. 9. K-Means on t-SNE

the proteins Troponin-I and Protein Kinase C (PKC) along with various MeSH terms, including myocardial infarction and angina pectoris. Steinberg (13) discovered that PKC activation leads to rapid changes in contractile performance. Namely, PKC increases the myosin head's force of contraction by increasing the calcium sensitivity of the troponin complex. Hypertension, or high blood pressure, is associated with an increase risk of MI especially in patients with acute coronary artery syndromes. Thus, PKC has some correlation with CAD. Angina pectoris is also a symptom of MI whereby the lack oxygen to the heart causes chest pains. Thus, it suggests that this community is for CAD.

Another small community included the proteins, dystrophin, Selenoprotein N, and chloride intracellular channel protein 4, along with the MeSH terms, cardiomyopathy, sacroglycanopathies, and glycogen storage disease type iib (Table 1, Fig. 6). In a previous study, Duan (14) found that chloride channels have been found to play a role in myocardial hypertrophy, which is a subset of cardiomyopathy. A gain- or loss-of-function of selenoprotein N may cause hypertension, and chronic hypertension has been associated with myocardial hypertrophy (15). Finally, dystrophin bridges the cardiac tissue to the inner wall of our chest. A disease associated with dystrophin will cause the cells to rub against the inner wall during contraction, which will destroy the epithelial tissues and eventually cause cardiomyopathy. Hence, this community seems to be well defined for cardiomyopathy.

In addition, the betweenness centralities seem to align with the prominence of each protein and MeSH term within their respective communities. Consider the cardiomyopathy community, where the MeSH term "barth syndrome", which has a betweenness centrality of 4787. Compared to the centrality of 79902 for dilated cardiomyopathy in the same community, implying that "barth syndrome" is significantly less important. This is well-justified as barth syndrome is a rare form of dilated cardiomyopathy, meaning that it will not be as strongly related to other proteins and MeSH terms that may be more generally applicable (16).

Although our communities are categorized, they are not necessarily partitioned. For example, the coronary artery disease community also included the MeSH term, cardiomyopathy. These are distinct CVDs but a protein like Troponin-I is involved in many different function other than just contraction. Thus, there exists some overlap of MeSH terms and protein between each communities.

K-Means Clustering. The result had seven clusters(K=7) correlating to the seven types of CVDs, which was due to either structural changes of the heart or signaling error between

Fig. 10. K-Means on t-SNE

the cells. When we filtered the protein and MeSH nodes, we observed that most of the protein nodes are densely packed within the arrhythmia cluster. Out of all the CVD, only arrhythmia is related to signaling error; the rest originate from changes in muscle structure. Pathways and proteins pertaining to cardiac muscle are typically limited to the cardiovascular system, but arrhythmia involves the cardiovascular, endocrine, nervous, and urinary system. So, all the proteins in the cardiac muscle are already contained in the set of proteins involving arrhythmia. Thus, this may speculate that some proteins that are involved in multiple organ system may be scattered further from the center of the arrhythmia cluster, while proteins that are solely involved in arrhythmia are clustered together. Furthermore, there may be more associations contained within this data, but limited time and resource deterred further investigations.

Conclusion

Our study attempted to determine the most prominent protein in our network and quantify the strength of correlation between the proteins and CVD to further understand their role in a macromolecular perspective. When the analysis correlated a single protein with one MeSH term, we discovered significance of Troponin-I in cardiac function and its association with myocardial infarction. Computing the betweenness centralities within each community in the network gave intuitive results that lined up with previous research. Finally, we found that translating the network using graph embedding and using k-means clustering did not translate well into an effective metric. This is because the notion of distance that results from graph embedding does not have a clear biological interpretation for this network.

However, community-wise betweenness is not without faults. While we grouped multiple proteins with many different MeSH terms, community detection did not allow for one node to be in multiple communities. This is not realistic as a protein may have effects on various different CVDs, but only the CVD with the highest correlation was chosen.

Limitations. The structure of our network yields deep mathematical and computational implications. Of particular note is making the network undirected. This decision changes the mathematics done on it; however, it is also unclear what effects this has on the outputs of the algorithm outputs and their interpretation. Furthermore, errors in graph construction may have led to mislabeled nodes, nodes with no edges, and missing edges. While we have done our best in the time allotted to correct these errors, we are unsure if our network is completely free of these errors. Finally, many algorithms we chose to use treated all the nodes as the same type (e.g., the network is homogenous). Again, we are unsure of its implications in the analysis of our network.

A biological process may not output the same value even if two input values were identical, so it is difficult to replicate the results. The mechanisms are only a simplified model that merely speculates the implications, and new theories are continuously being developed at a rapid rate. Furthermore, some molecules have the same name but play a multi-purpose role in various parts of the body in different pathways. For example, nitric oxide plays a role in muscle cells by increasing glucose uptake for ATP production, but neurons use NO as a neurotransmitter. In our protein betweenness centrality, we had gap junction alpha-1 protein and connexin-43, but both referred to the same protein. This case may extend to other MeSH terms across our network. Finally, health issues such as CVD come about from a combination of different proteins complexes rather than one target protein subunit. Thus, even though we correlated one protein with one disease, it does not mimic the real world application.

Further Study. Future studies may involve link inferences between nodes that may provide new data regarding the mechanisms of various CVD's that were never investigated before. A deeper dive can be done relating Troponin-I with arrhythmia pathways since many previous studies only associates Troponin-I with contraction force rather than rate. New research constantly changes the landscape of science, so these studies allow us to better address the growing problem of CVD's.

Code and Data. The code and dataset used in this paper can be found here.

ACKNOWLEDGMENTS. We acknowledge Professor Mason Porter's advice on analyzing community structure and understanding the consequences of making the network undirected.

1. H Abriel, JS Rougier, J Jalife, Ion channel macromolecular complexes in cardiomyocytes: roles in sudden cardiac death. *Cirulation Res.* **116**, 1971–1988 (2015).
2. Y Ruan, et al., Cardiovascular disease (cvd) and associated risk factors among older adults in six low-and middle-income countires: results from sage wave 1. *BMC Public Heal.* **18**, 778 (2018).
3. U.S. National Library of Medicine, Medical subject headings (2022).
4. U.S. National Library of Medicine, Pubmed (2022).
5. The UniProt Consortium, Uniprot: Universal protein knowledgebase (2022).
6. D Wishart, et al., Drugbank: a comprehensive resource for in silico drug discovery and exploration (2022).
7. L Stein, P D'Eustachio, H Hermjakob, G Wu, Reactome pathway browser (2022).
8. M Newman, *Networks.* (Oxford University Press) No. 2, (2018).
9. H Chen, SF Sultan, Y Tian, M Chen, S Skiena, Fast and accurate network embeddings via very sparse random projection. *Proc. 28th ACM Int. Conf. on Inf. Knowl. Manag.* p. 399–408 (2019).
10. K Makarychev, Y Makarychev, I Razenshteyn, Performance of johnson–lindenstrauss transform for k-means and k-medians clustering. *SIAM J. on Comput.* **0**, STOC19–269–STOC19–297 (2022).
11. GE Hinton, S Roweis, Stochastic neighbor embedding. *Adv. Neural Inf. Process. Syst.* **15** (2002).
12. N Ojha, A Dhamoon, Myocardial infarction. *NCBI* (2021).
13. SF Steinberg, Cardiac action of protein kinase c isoforms. *Physiol. (Bethesda)* **27**, 130–139 (2012).
14. D Duan, Phenomics of cardiac chlordie channels. *Compr. Physiol.* **3**, 667–692 (2013).
15. AC Egbe, et al., Persistent hypertension and left ventricular hypertrophy after repair of native coartation of aorta in adults. *Hypertension* **78**, 672–680 (2021).
16. Barth syndrome: Medlineplus genetics (2021).