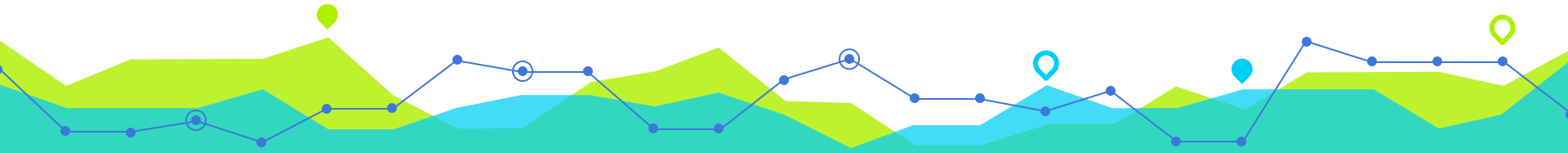


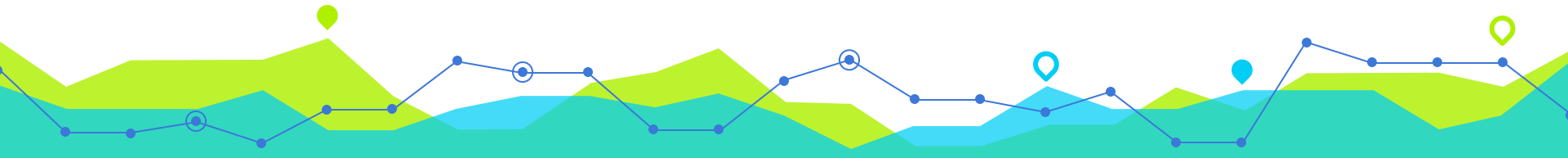
Network Analysis: Cardiovascular Disease Subtype Association and Ion Channel Proteins

Irsyad Adam, Seung Chung, Andy Goh, and Ethan Young

Introduction

For this project, we analyzed the strengths of relationships between proteins and different types of cardiovascular diseases by modeling the information as a network.

We first studied the network on a more global scale, then analyzed the network on a more local level using two different approaches: community detection and clustering through dimension reduction.



Motivation: Why Cardiovascular Disease (CVD)?

Construction of the Network

176 Cardiovascular MeSH (Medical Subject Heading) Terms from MeSH Database

- “Cardiomyopathies”, “Myocarditis” , “Endomyocardial Fibrosis" …

13495 PubMed Documents corresponding to MeSH Terms

- “Influence of echocardiographic measurements and renal impairments on the prognosis of fulminant myocarditis.” …

424 Ion Channel Proteins from UniProtKB

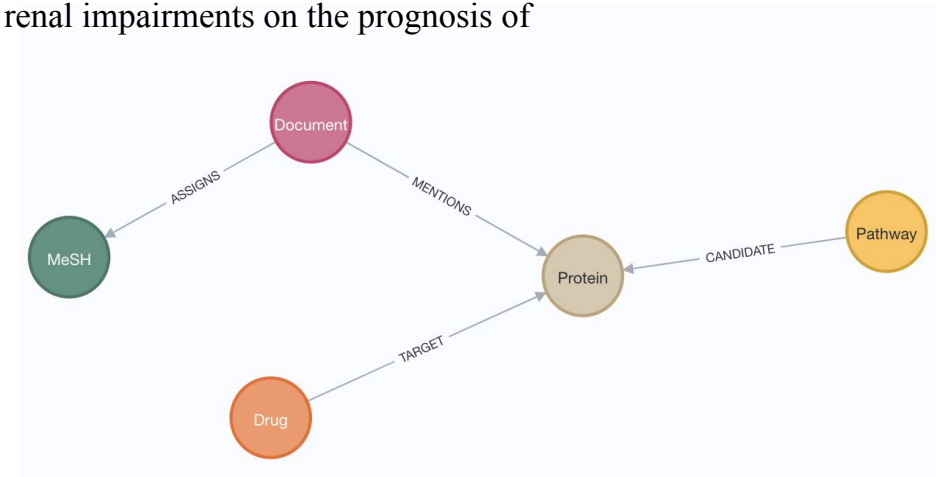
- “Amyloid beta A4 protein”, “Caveolin-1” …

142 Cardiovascular Drugs from DrugBank

- “Heparin”, “Thrombolytics”, “Tirofiban” …

535 Biological Pathways from Reactome

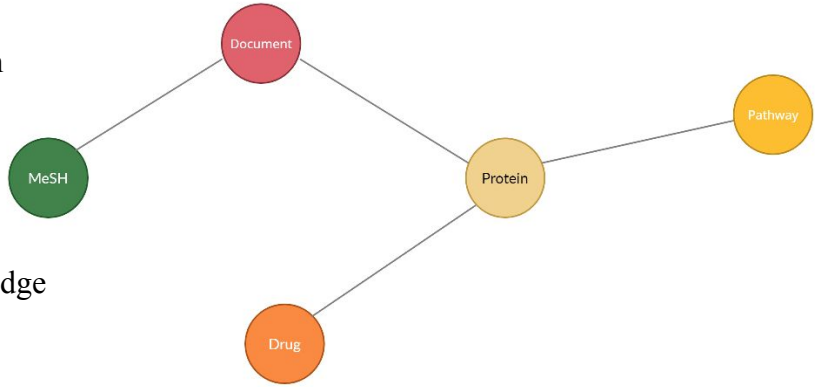
- “Platelet homeostasis”, “Ion transport by P-type ATPases” …



Significance of the Network (What Information is Captured)

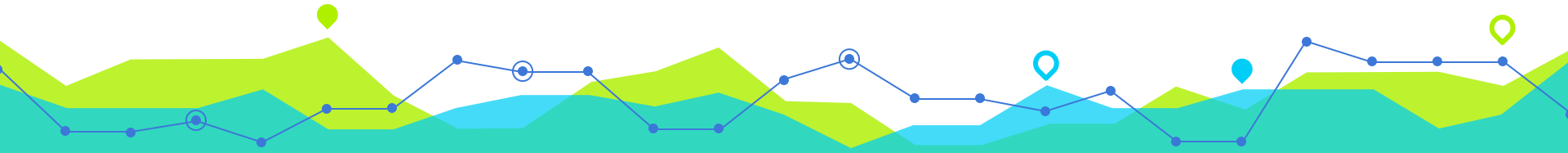
Using the property of networks on biological data and bioinformatics, we are able to see which ion channel proteins in general relate to which subtype of cardiovascular disease without doing wet lab research.

Because of this, we implement various algorithms to quantify the correlation between ion channel proteins and different CVD sub-types.



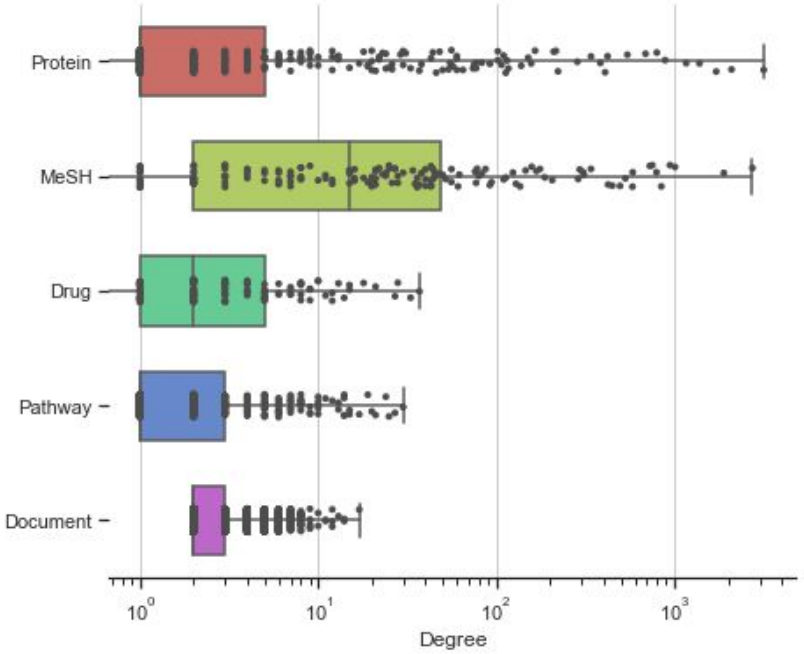
Data Cleaning:

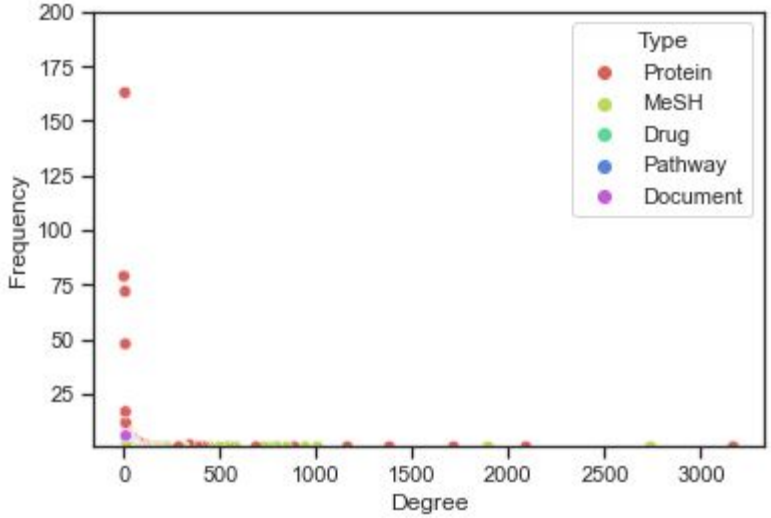
- Removal of degree-0 nodes due to coding errors
- Changing relationships to undirected and treating them as 1 type of edge as direction is mainly for clarification



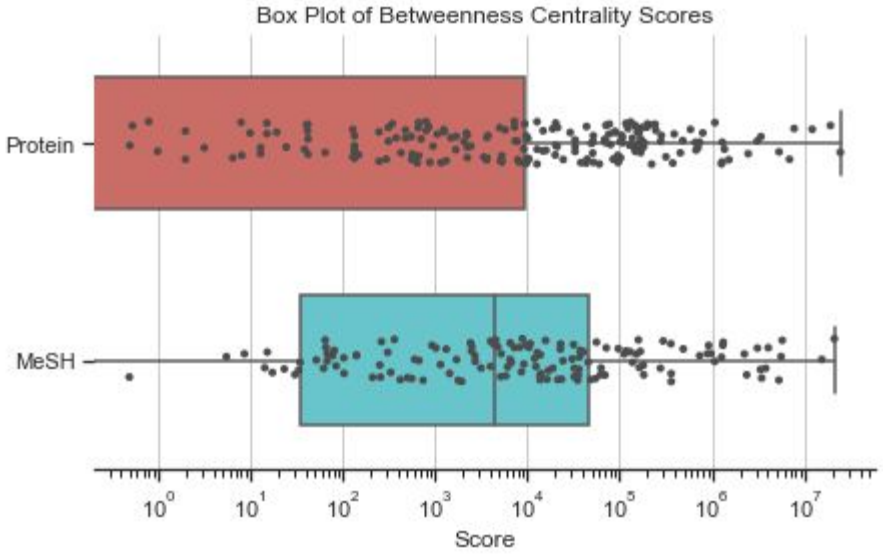
I. Global Network Analysis

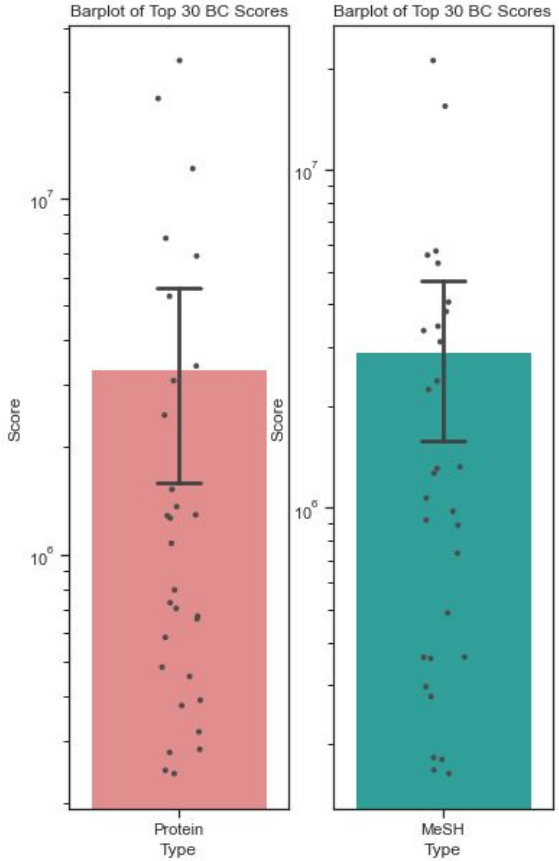
Degree Distribution



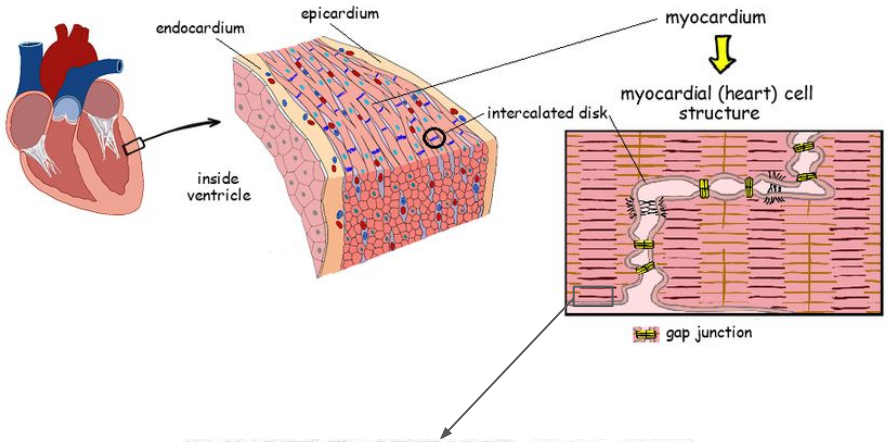


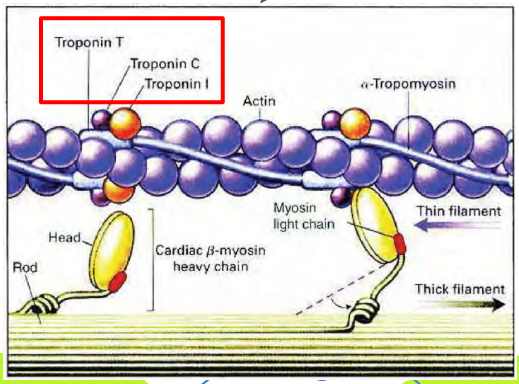
Betweenness Centrality





Biological Implications





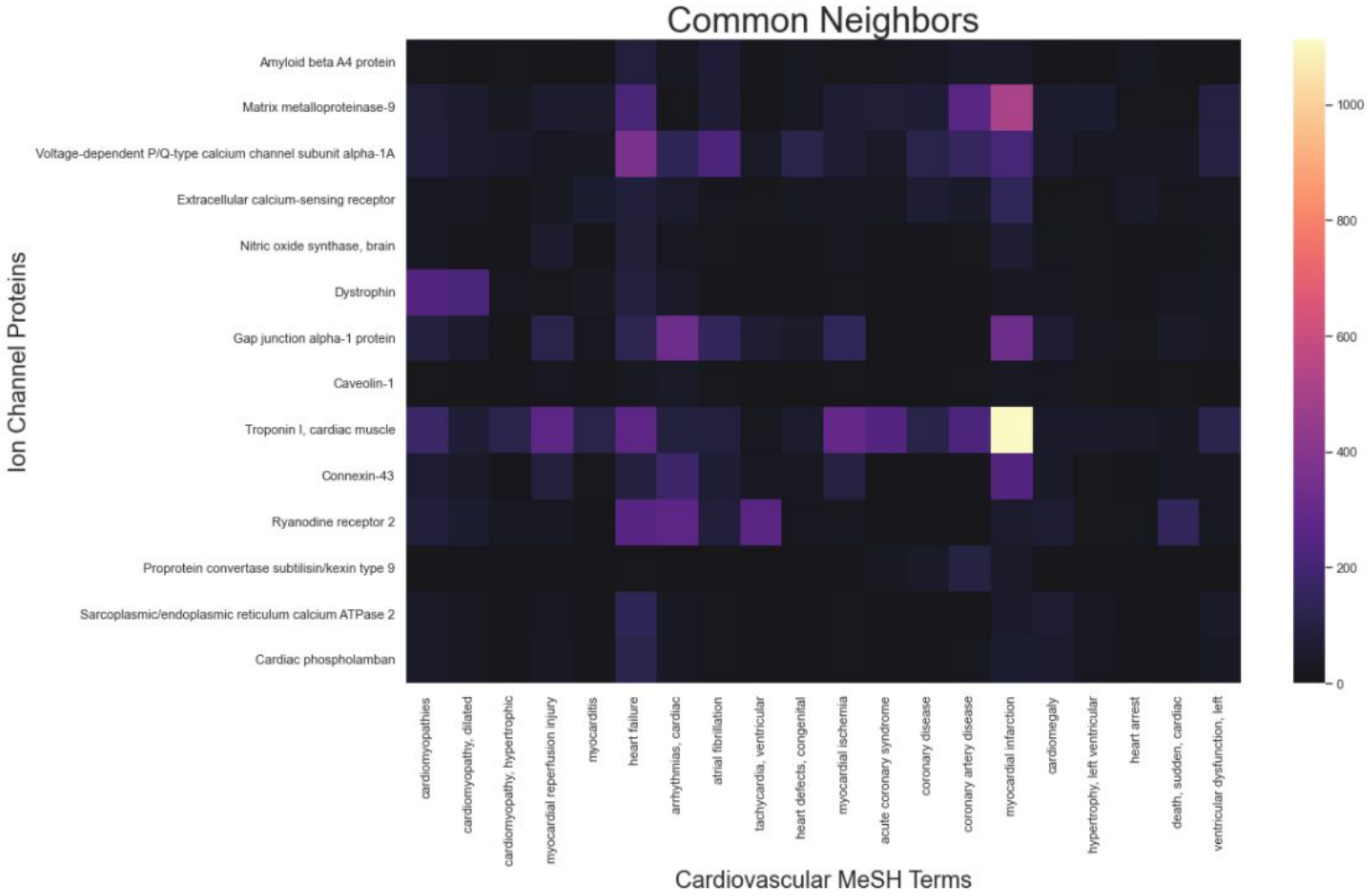
[Figures]. https://mysciencesquad.weebly.com/d4-the-heart.html

Protein

Name	Betweenness Centrality Score
Troponin-I	2.445644e+07
Voltage-dependent P/Q-type calcium channel	1.911941e+07
Matrix Metalloproteinase-9	1.214294e+07
Gap Junction Alpha-1 Protein	7.737463e+06
Ryanodine Receptor 2	6.897274e+06
Extracellular Calcium-Sensing Receptor	5.314467e+06
Amyloid beta A4 Protein	3.384747e+06
Dystrophin	3.077908e+06
Connexin-43	2.462704e+06

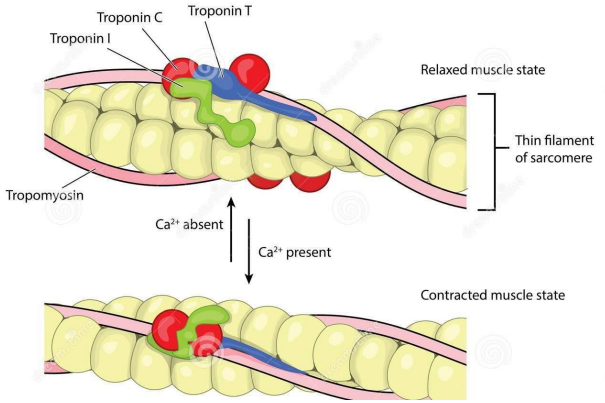
MeSH

Name	Betweenness Centrality Score
Myocardial Infarction	2.106038e+07
Heart Failure	1.543012e+07
Arrhythmias	5.752145e+06
Coronary Artery Disease	5.594684e+06
Cardiomyopathies	5.294915e+06
Myocardial Ischemia	4.063651e+06
Atrial Fibrillation	3.808096e+06
Myocardial Reperfusion Injury	3.446566e+06
Cardiomegaly	3.345379e+06

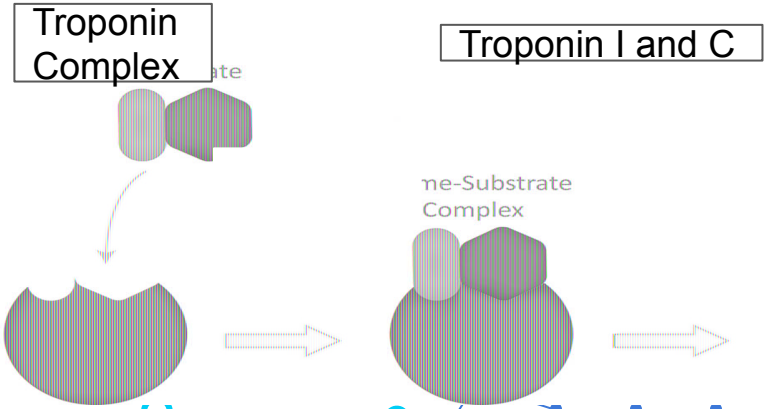

Common Neighbors
Ion Channel Proteins
Amyloid beta A4 protein
Matrix metalloproteinase-9
Voltage-dependent P/Q-type calcium channel subunit alpha-1A
Extracellular calcium-sensing receptor
Nitric oxide synthase, brain
Dystrophin
Gap junction alpha-1 protein
Caveolin-1
Troponin I, cardiac muscle
Connexin-43
Ryanodine receptor 2
Proprotein convertase subtilisin/kexin type 9
Sarcoplasmic/endoplasmic reticulum calcium ATPase 2
Cardiac phospholamban
cardiomyopathies
cardiomyopathy, dilated
cardiomyopathy, hypertrophic
myocardial reperfusion injury
myocarditis
heart failure
arrhythmias, cardiac
atrial fibrillation
tachycardia, ventricular
heart defects, congenital
myocardial ischemia
acute coronary syndrome
coronary disease
coronary artery disease
myocardial infarction
cardiomegaly
hypertrophy, left ventricular
heart arrest
death, sudden, cardiac
ventricular dysfunction, left
Cardiovascular MeSH Terms
1000
800
600
400
200
0

Biological Interpretation and Potential Implications

Let X be the set of all cardiovascular-related proteins and Y be the set of all CVD. Define X~Y as relation between X and Y such that protein $x \in X$ is the most prominent in disease $y \in Y$. Then by common neighbors, we claim that:

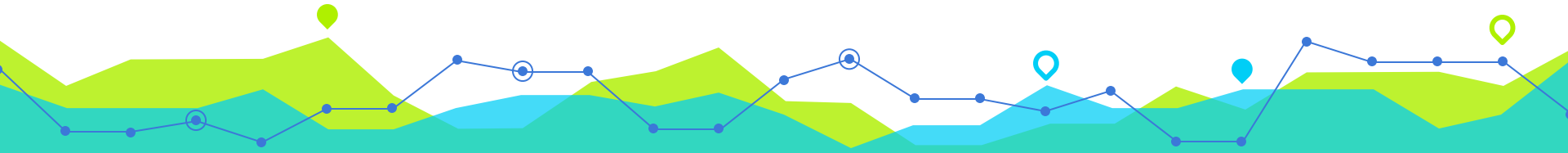


- **Troponin-I ~ Myocardial Infarction**



- **Matrix Metalloproteinase-9 ~ Myocardial Infarction**

[Images]. https://www.dreamstime.com/muscle-contraction-relaxation-mechanism-which-contracts-relaxes-due-to-action-calcium-ions-troponin-complex-image121391732]



II. Community Detection

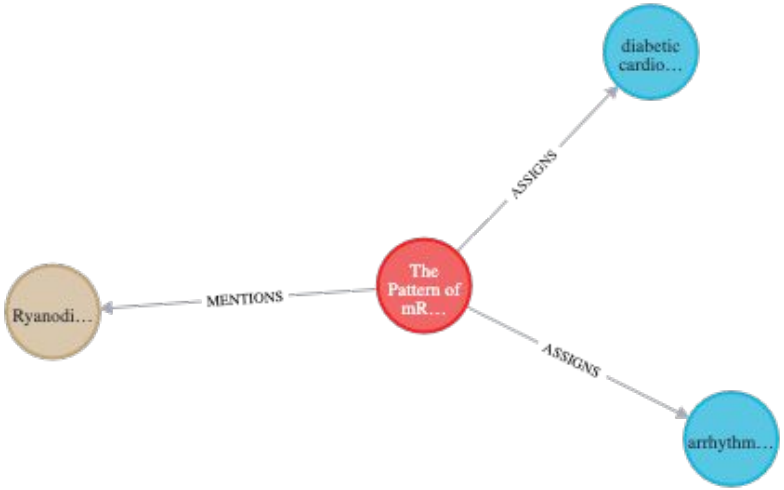
Motivation

Documents
MeSH
Proteins

Betweenness and common neighbor methods don't consider groupings.

CVD is the malfunction of multiple proteins, which suggests looking at groupings is necessary.

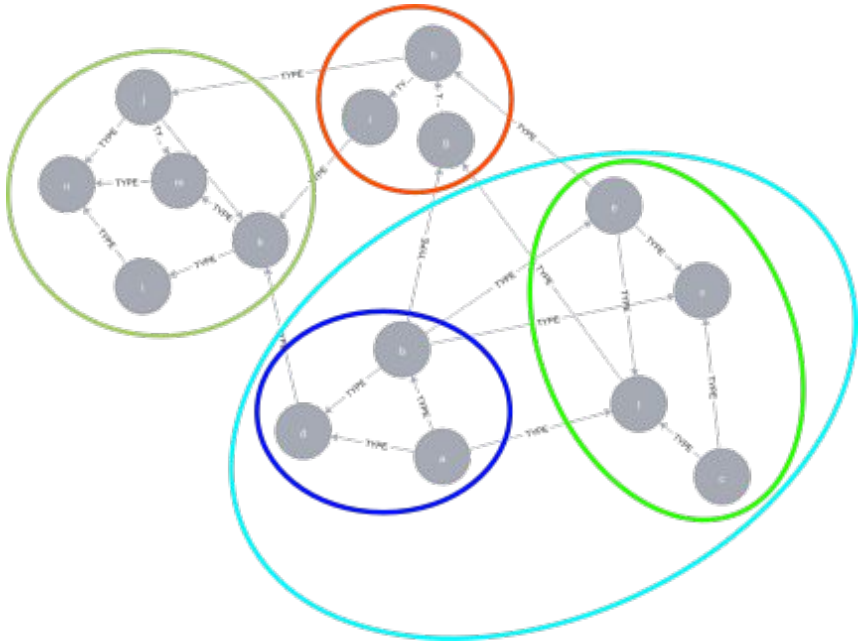
Intuitively, we would expect there to be clusters within this network around proteins, documents, and MeSH terms that are **related to similar diseases.**



Community Detection: Algorithm and Assumptions

The Louvain method is a greedy algorithm that maximizes a modularity score for each community.

We ignored node types (document, protein, etc.) when running community detection

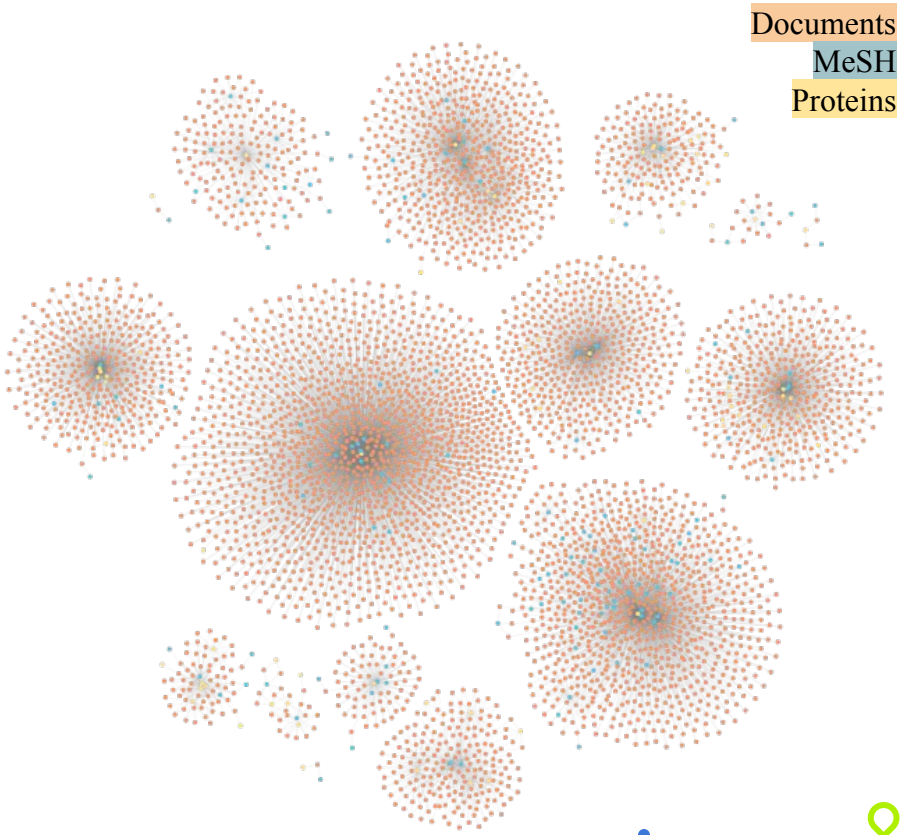


[Figure]. https://neo4j.com/docs/graph-data-science/current/algorithms/louvain/#algorithms-louvain-examples-stream-intermediate

Community Detection: Result

Communities are mostly determined by **Documents**, as that is the link between proteins and MeSH terms, and they are most abundant

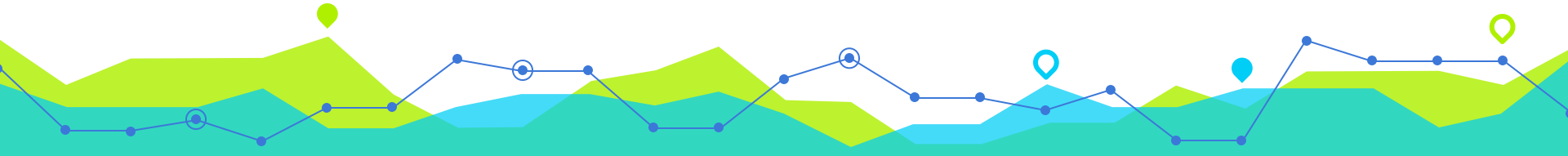
Main focus on which **proteins** and **MeSH** terms are in which community



What do the communities represent?

MeSH	myocardial reperfusion injury
Protein	Nitric oxide synthase, brain
Protein	78 kDa glucose-regulated protein
Protein	Sodium/hydrogen exchanger 1
Protein	Voltage-dependent anion-selective channel protein 1
Protein	Alpha-1-syntrophin
Protein	Beta-2-syntrophin

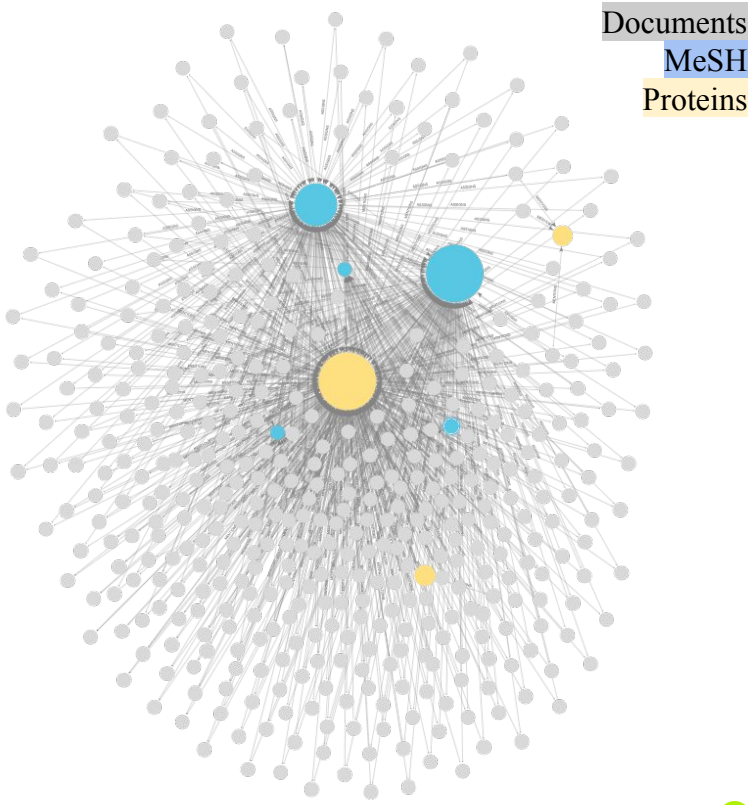
MeSH	cardiomyopathies
Protein	**Dystrophin**
MeSH	cardiomyopathy, dilated
MeSH	barth syndrome
Protein	Selenoprotein N
MeSH	sarcoglycanopathies
MeSH	glycogen storage disease type iib
Protein	Chloride intracellular channel protein 4

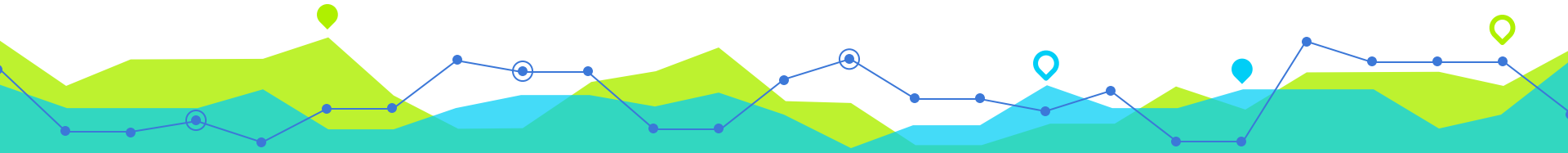


Community-wise Betweenness

Within each community, which proteins and MeSH terms are the most "important"?

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
Protein	Chloride intracellular channel protein 4	106.8





III. Graph Embeddings and K-Means

Graph Embeddings

To further explore any potential structure in the network using K-means clustering, we employ graph embedding techniques in order to reduce the dimension of our data, while preserving as much information (e.g., distance) as possible. The reason we do this is to use distances between data points as a potential metric for correlation when clustering.

One method of growing popularity is random-walk-based. Here, the use of random projections (RP) assign a weight to each dimension of the embedding by using the probability of reaching node *j* from *i* for varying numbers of random walks, as seen in [6].

Here, we use a graph embedding algorithm in Neo4J based on an RP method proposed by H. Chen (2019), called FastRP. Figure 3.1 shows an example computation of the embedding.

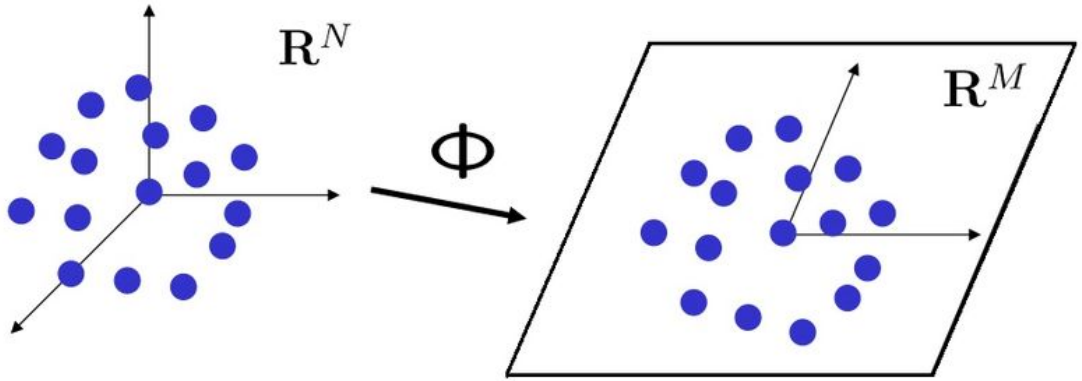
$$N = (\alpha_1 \tilde{A} + \alpha_2 \tilde{A} + \alpha_3 \tilde{A} + \alpha_4 \tilde{A}) \cdot R$$

[Figure 3.1].
https://towardsdatascience.com/behind-the-scenes-on-the-fast-random-projection-algorithm-for-generating-graph-embeddings-efb1db0895

Motivating FastRP: The Johnson-Lindenstrauss Lemma

This lemma states that a set of points in N-dimensional space can be projected into a M-dimensional space, where $M \ll N$, such that the distances between the points are approximately preserved.

Figure 3.2 shows the geometric interpretation of this lemma.



[Figure 3.2]. https://www.researchgate.net/figure/Johnson-Lindenstrauss-lemmas-illustration_fig3_341576089

The FastRP Algorithm

Goal: we want $N = M \cdot R$, where $N \in \mathbb{R}^{nxd}$, $M \in \mathbb{R}^{nxm}$, $R \in \mathbb{R}^{mxd}$, and $d \ll m$.

Here, N is the embedding matrix, M is the feature matrix, and R is a random projection matrix where its entries are sampled i.i.d. (independently and identically distributed) with zero mean from some distribution (e.g., Gaussian) and satisfies the J-L lemma.

Exploiting the associativity of matrix multiplication, we have that:

$$N_k = (A \cdot \ldots \cdot)\,(A \cdot L \cdot R) = \tilde{A}^k \cdot L \cdot R,$$

where A is the random-walk Laplacian and L is a normalization matrix that reduces the influence of high-degree nodes.

The FastRP Algorithm

Goal: we want $N = M \cdot R$, where $N \in \mathbb{R}^{nxd}$, $M \in \mathbb{R}^{nxm}$, $R \in \mathbb{R}^{mxd}$, and $d \ll m$.

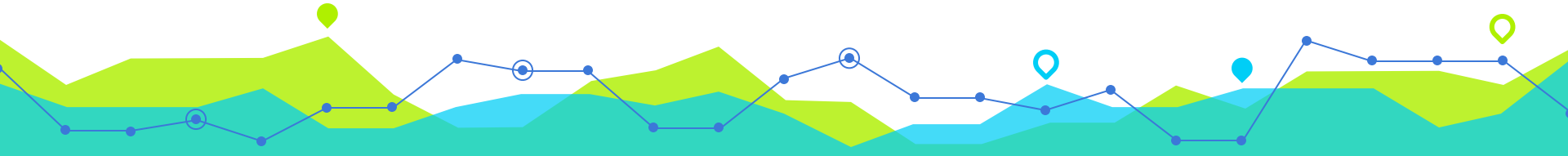
We further consider a weighted sum, so our final computation of embeddings is:

$$N_k = (A \cdot \ldots \cdot)\,(A \cdot L \cdot R) = \tilde{A}^k \cdot R$$

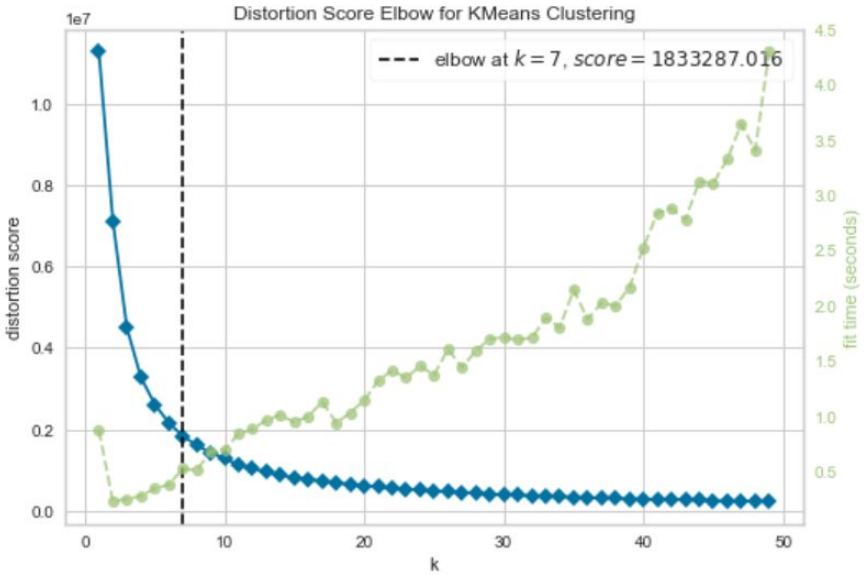
$$N = \boldsymbol{\alpha}_1 N_1 + \boldsymbol{\alpha}_2 N_2 + \ldots + \boldsymbol{\alpha}_k N_k,$$

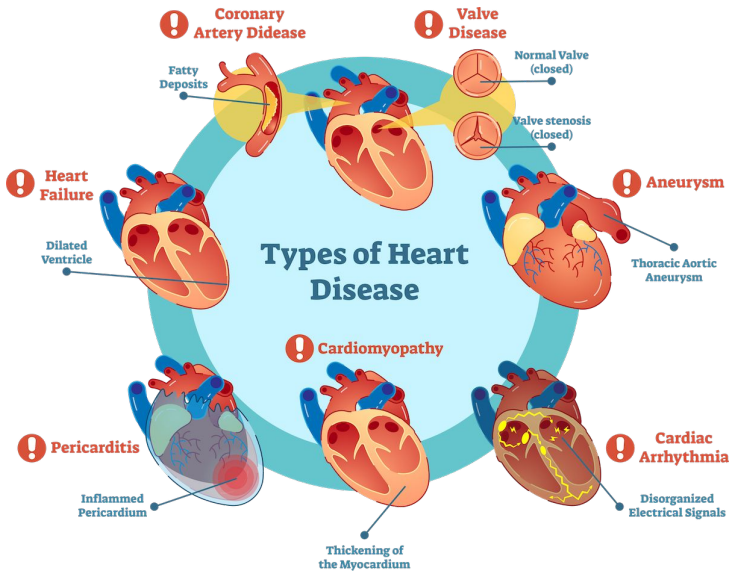
where $\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \ldots, \boldsymbol{\alpha}_k$ are weights that tell us how much influence the neighbors k-steps away of a node has.

Note: 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).

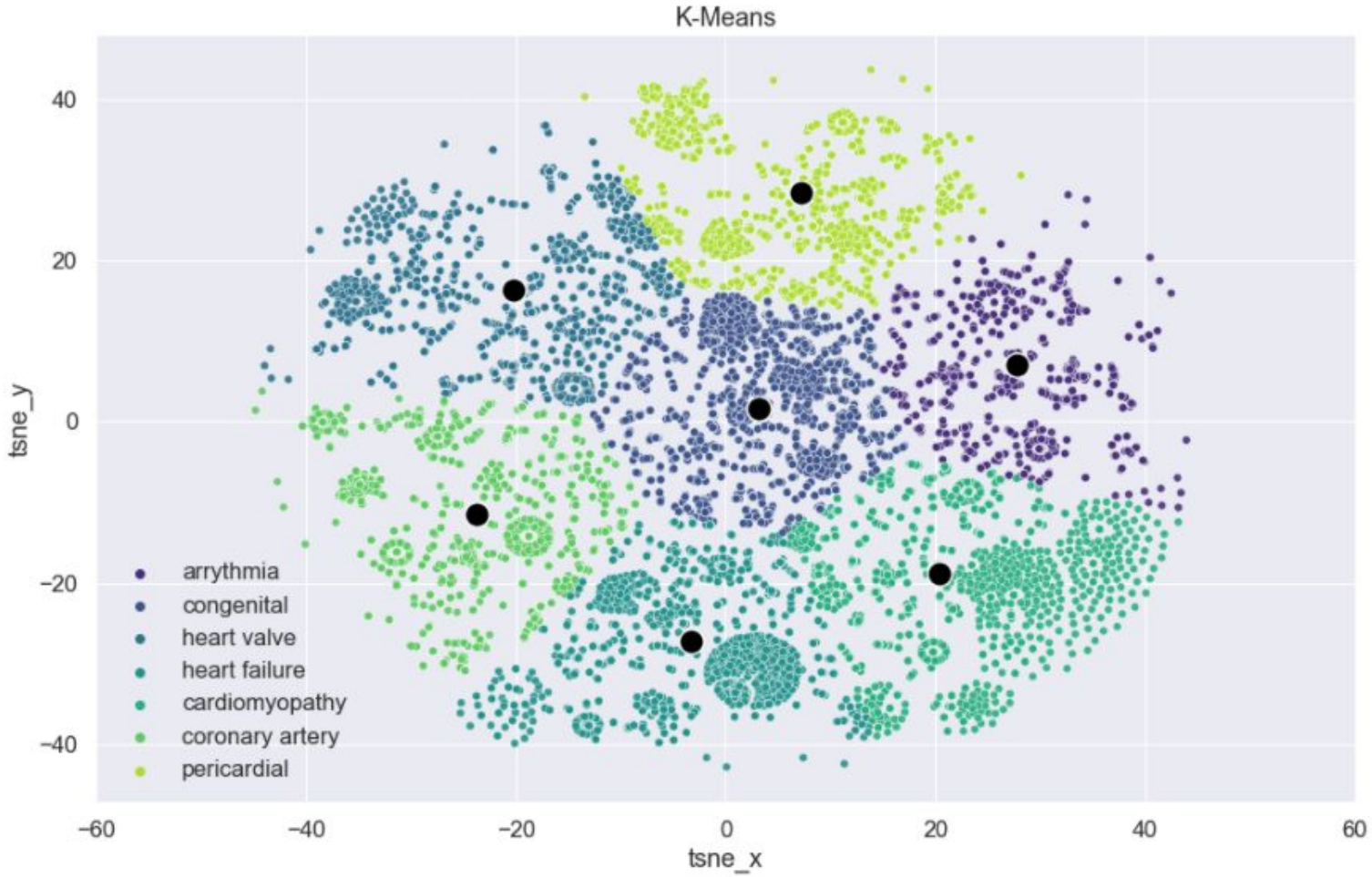


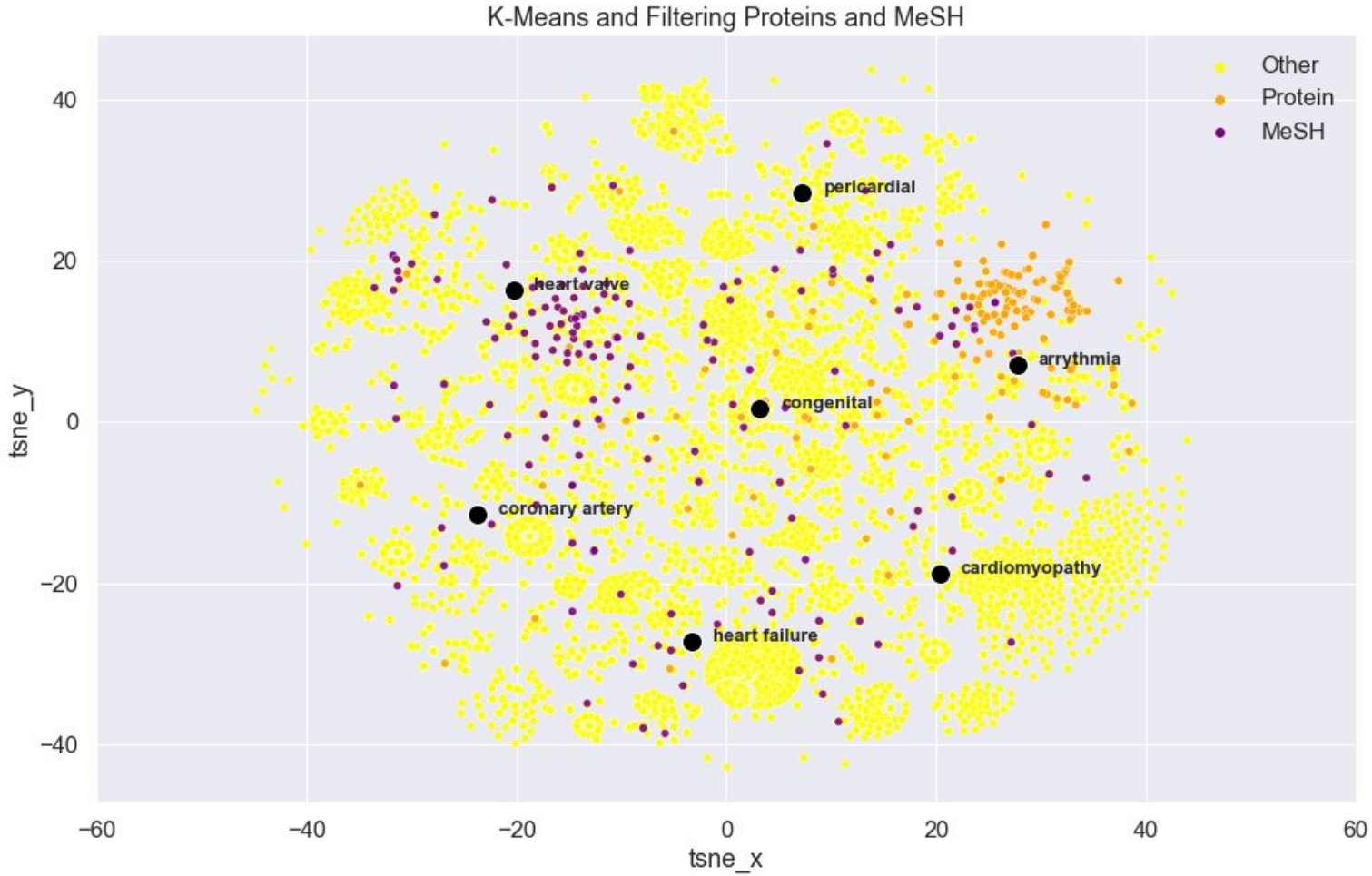
For Fast RP, we used an embedding dimension of 256, as suggested by this article.





[Figure]. https://www.udmi.net/cardiovascular-disease-risk/

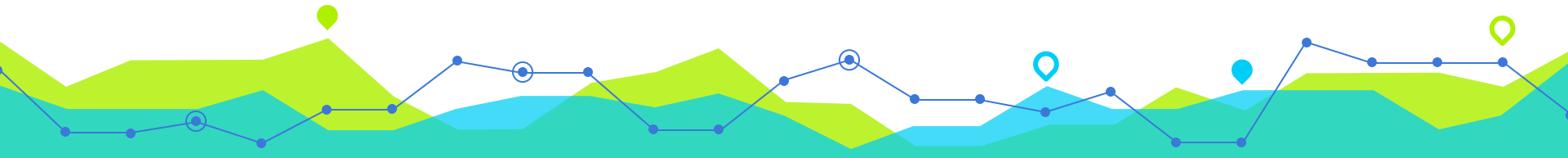

K-Means
40
20
0
−20
−40
tsne_y
−60
−40
−20
0
20
40
60
tsne_x
arrythmia
congenital
heart valve
heart failure
cardiomyopathy
coronary artery
pericardial


K-Means and Filtering Proteins and MeSH
Other
Protein
MeSH
pericardial
heart valve
arrythmia
congenital
coronary artery
cardiomyopathy
heart failure
tsne_y
tsne_x
40
20
0
−20
−40
−60
−40
−20
0
20
40
60

Limitations & Further Study

Mathematical and computational constraints and errors:

- Making the network undirected changes the mathematics done on it.
  - Given that this network structure is rather unconventional, it’s unclear what effects it has on the outputs of algorithms run on it.
- Errors in our graph construction may have also led to errors such as mislabeled nodes and missing edges.



Limitations & Further Study

Biological limitations:

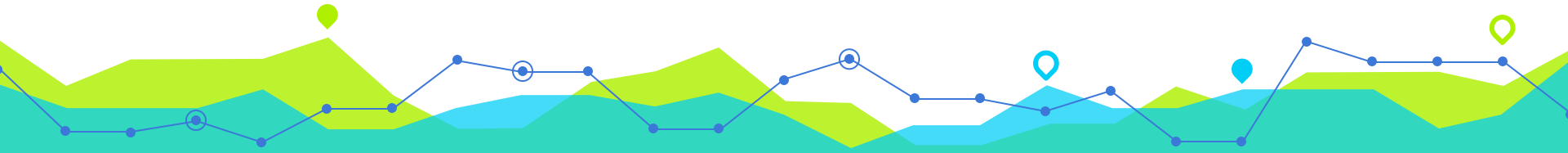
- Some proteins in the network have different names (e.g., Gap Junction Alpha-1 and Connexin-43), but have a same UniProtID and are associated with the same function in the body
- Other entities have similar names, but are associated with different parts of the body (Nitric Oxide synthase and Nitric Oxide, Brain)
- CVD is usually a combination of different proteins malfunctioning rather than one specific protein, so it is difficult to attribute CVD subtypes to a single protein

Protein	
Name	Betweenness Centrality Score
Troponin-I	2.445644e+07
Voltage-dependent P/Q-type calcium channel	1.911941e+07
Matrix Metalloproteinase-9	1.214294e+07
Gap Junction Alpha-1 Protein	7.737463e+06
Ryanodine Receptor 2	6.897274e+06
Extracellular Calcium-Sensing Receptor	5.314467e+06
Amyloid beta A4 Protein	3.384747e+06
Dystrophin	3.077908e+06
Connexin-43	2.462704e+06

Limitations & Further Study

Future Work:

- Given more time, we would have focused on extending the analysis of the dimension-reduced data and train machine learning models on it.
    - Some tasks of interest are node inference using a graph convolutional network.
- Particular attention should also be devoted to better understanding network structure and any communities or clusters produced by algorithms.
- We use several methods with metrics dependent on distance between data points; given the nature of this data set, more study is needed to evaluate if distance is a measure that gives useful information.



Thank you!

Any questions?

References

(1) CDC – National Center for Health Statistics on Heart Disease – Homepage. https://www.cdc.gov/heartdisease/facts.htm March 14, 2022

(2) Islami, F., Mańczuk, M., Vedanthan, R., Vatten, L., Polewczyk, A., Fuster, V., Boffetta, P., & Zatoński, W. A. (2011). A cross-sectional study of cardiovascular disease and associated factors. *Annals of agricultural and environmental medicine : AAEM*, *18*(2), 255–259.

(3) Flora, G. D., & Nayak, M. K. (2019). A Brief Review of Cardiovascular Diseases, Associated Risk Factors and Current Treatment Regimes. *Current pharmaceutical design*, *25*(38), 4063–4084. https://doi.org/10.2174/1381612825666190925163827

(4) Lee, Y. B., Han, K., Kim, B., Lee, S. E., Jun, J. E., Ahn, J., Kim, G., Jin, S. M., & Kim, J. H. (2019). Risk of early mortality and cardiovascular disease in type 1 diabetes: a comparison with type 2 diabetes, a nationwide study. *Cardiovascular diabetology*, *18*(1), 157. https://doi.org/10.1186/s12933-019-0953-7

(5) Parente, J. M., Blascke de Mello, M. M., Silva, P., Omoto, A., Pernomian, L., Oliveira, I. S., Mahmud, Z., Fazan, R., Jr, Arantes, E. C., Schulz, R., & Castro, M. M. (2021). MMP inhibition attenuates hypertensive eccentric cardiac hypertrophy and dysfunction by preserving troponin I and dystrophin. *Biochemical pharmacology*, *193*, 114744. https://doi.org/10.1016/j.bcp.2021.114744

(6) Chen, H., Sultan, S. F., Tian, Y., Chen, M., Skiena, S. (2019). Fast and Accurate Network Embeddings via Very Sparse Random Projection. *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, 399-408. https://doi.org/10.1145/3357384.3357879