DISEASE PROTEIN PREDICTION

AN ANALYSIS OF MACHINE LEARNING APPROACHES

Sabri Eyuboglu and Pierce Freeman

The Disease Protein Prediction Problem

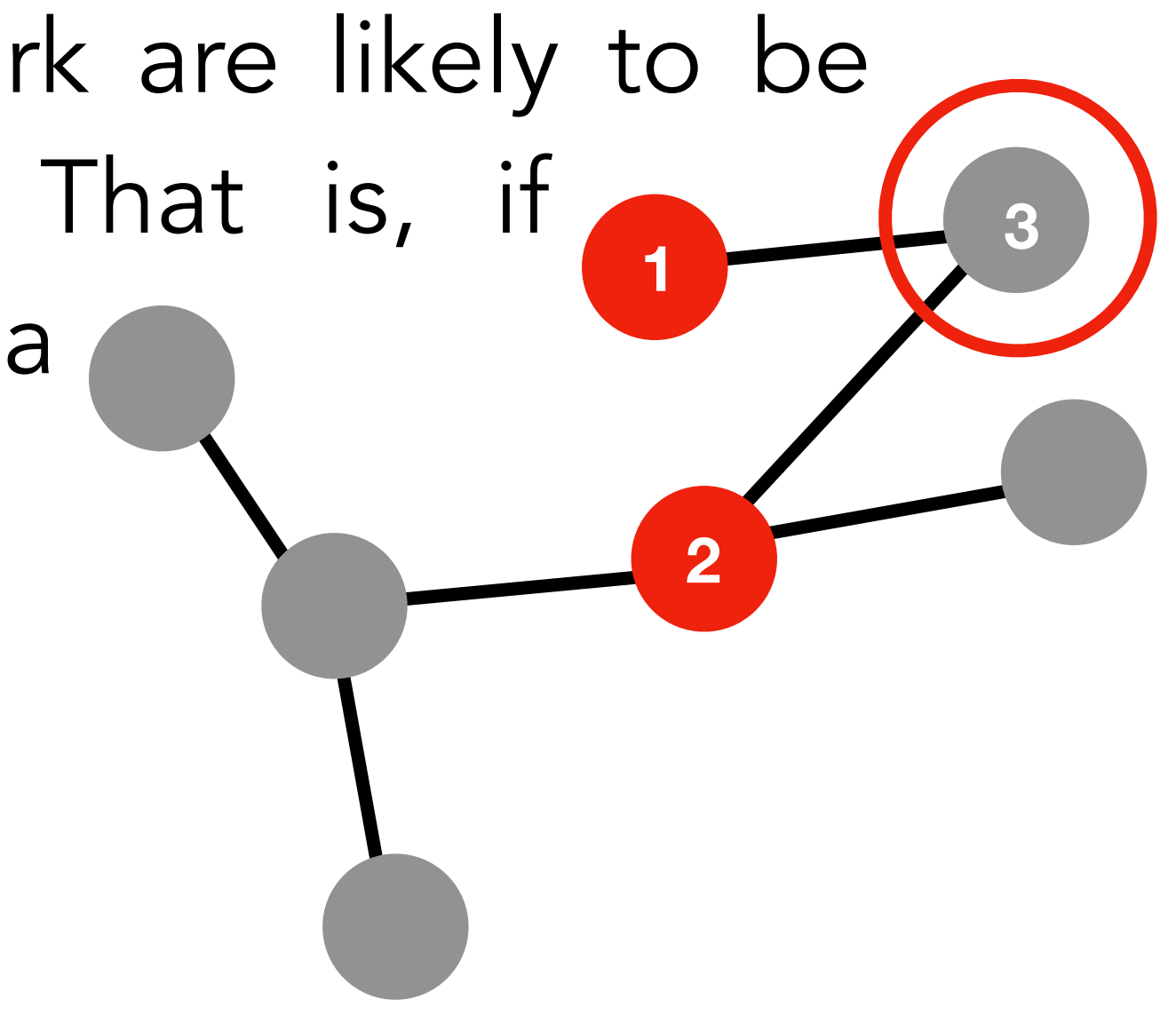
In human cells, **proteins** and their interactions can be disturbed by viruses, genetic mutations, or other processes in and out of the cell. Behind most human diseases are a set of proteins that when disturbed are responsible for the manifestation of the disorder. For medical researchers and scientists, understanding the set of proteins involved in disease is invaluable when studying the disease and designing new treatments. As it turns out, we can frame the problem of predicting which proteins are involved in a disease very naturally as a supervised machine learning problem. In this work, we use known disease proteins as a training-set, use data on proteins for features, and predict with **Logistic Regression** and **Support Vector Machines** the proteins that are most likely to be involved in the disease. We

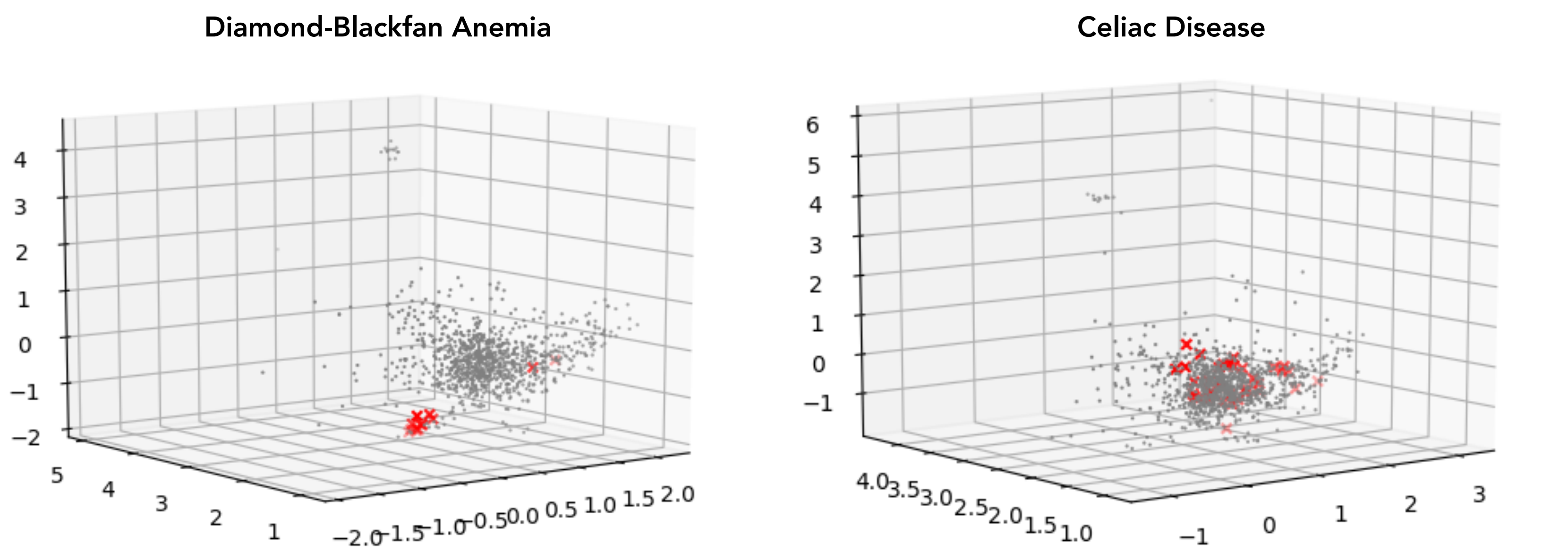
Mining the PPI Network for Features

The **Protein-Protein Interaction Network** is a graph that encodes protein interactions in the cell: each node is a protein and an edge defines an interaction between them.

IDEA Proteins that have similar structural roles or network neighborhoods in the PPI network are likely to be involved in the same diseases. That is, if proteins 1 and 2 are involved in a disease, protein 3 likely is too.

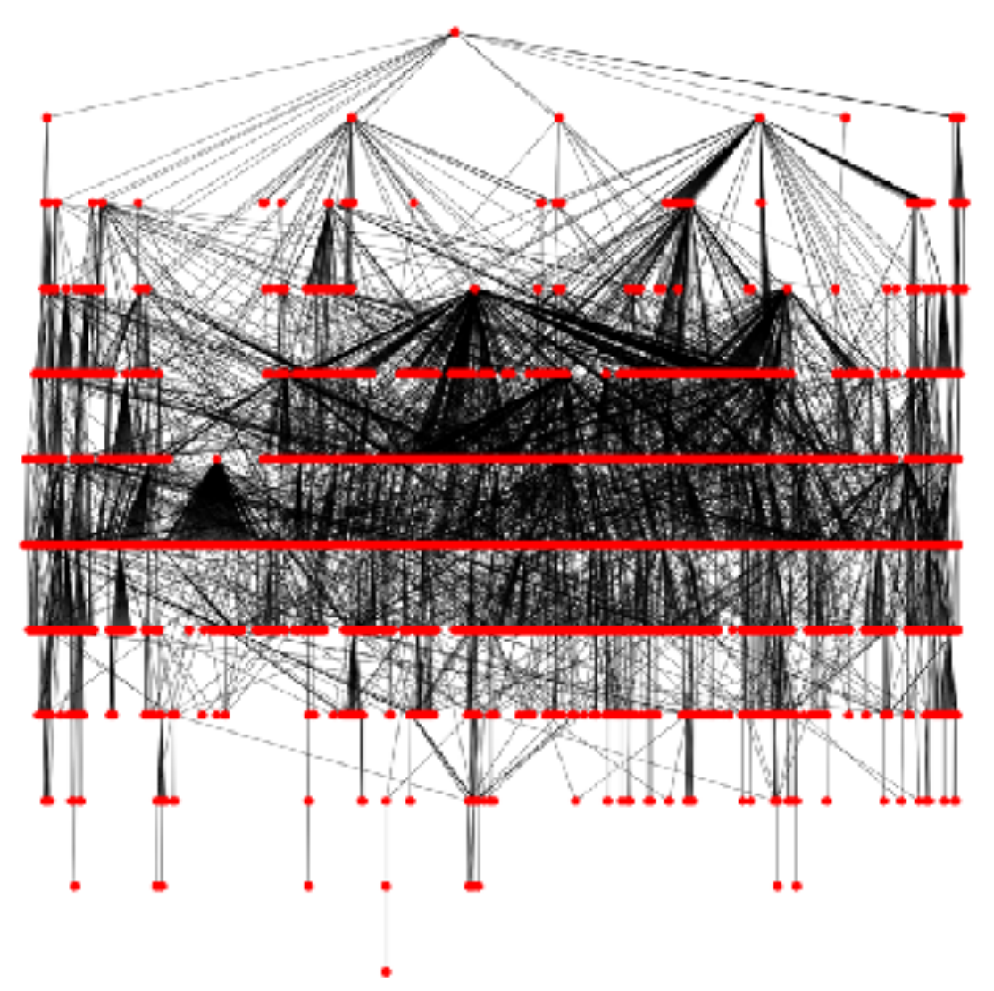
node2vec Let's build protein feature vectors using node2vec, an algorithm that generates a neighbor-hood preserving d-dimensional embedding of a network using maximum likelihood estimation.

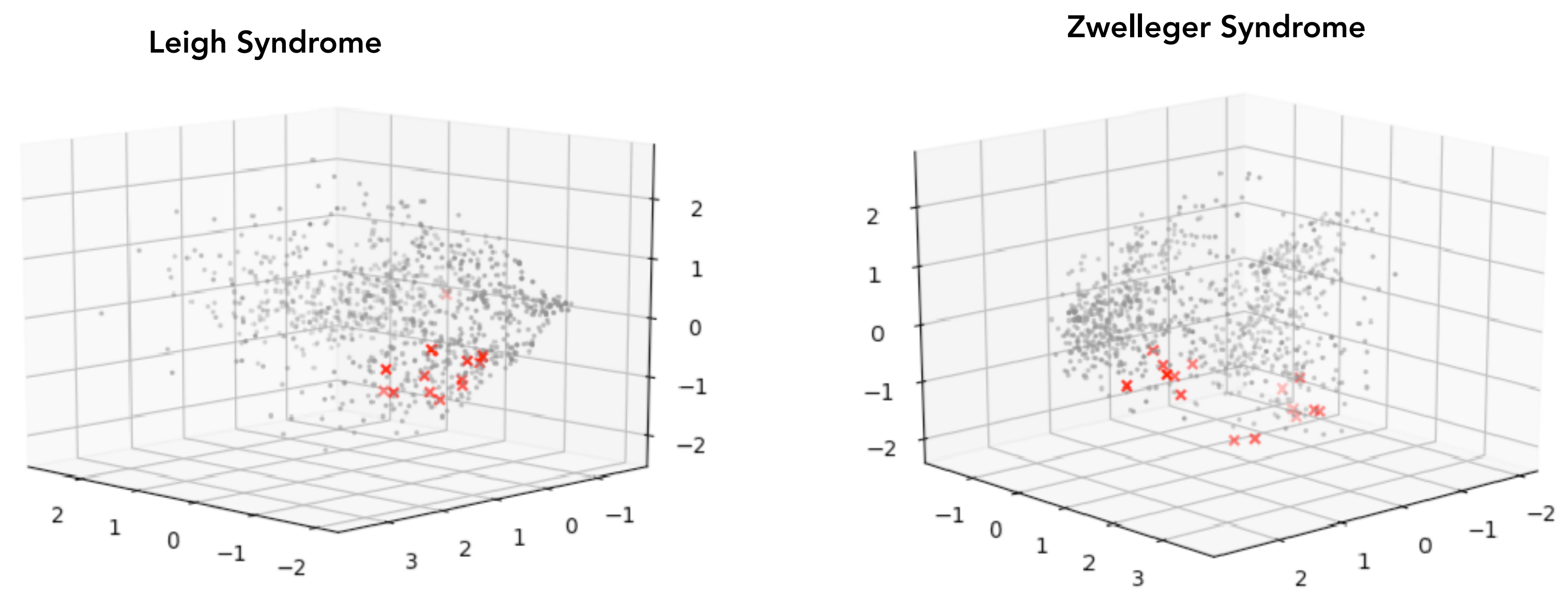




Traversing the Gene Ontology for Features

We use the **Gene Ontology** for descriptive cellular features of each protein. The dataset contains a directed ontology tree where the nodes describe different general organism properties and the edges define connections between them. These terms range from broad (human cell) to specific (organelle envelope lumen). Each protein is labeled with a few of these terms, allowing us to uncover their cellular function. We created an algorithm that crawls this term network and allows us to specify the desired level of specificity.





PCA Embedding of Gene Ontology Feature Space

Disease Protein Classification Models

We leverage **2** models to make predictions on proteins.

Logistic Regression Maximizes log-likelihood of training disease protein classifications given observed features.

$$\ell(\theta) = \sum_{i=1}^{m} \log p(y^{(i)} \mid x^{(i)}; \theta) \qquad p(y^{(i)} = 1 \mid x^{(i)}; \theta) = \frac{1}{1 + e^{-\theta^T x^{(i)}}}$$

Support Vector Machines Minimizes hinge-loss of disease protein classifications with the use of kernels to capture non-linearities. Kernels used: Linear, Radial Basis Function

$$J(w, b) = \sum_{i=1}^{m} \max[0, 1 - y^{(i)}(wx^{(i)} - b)]$$

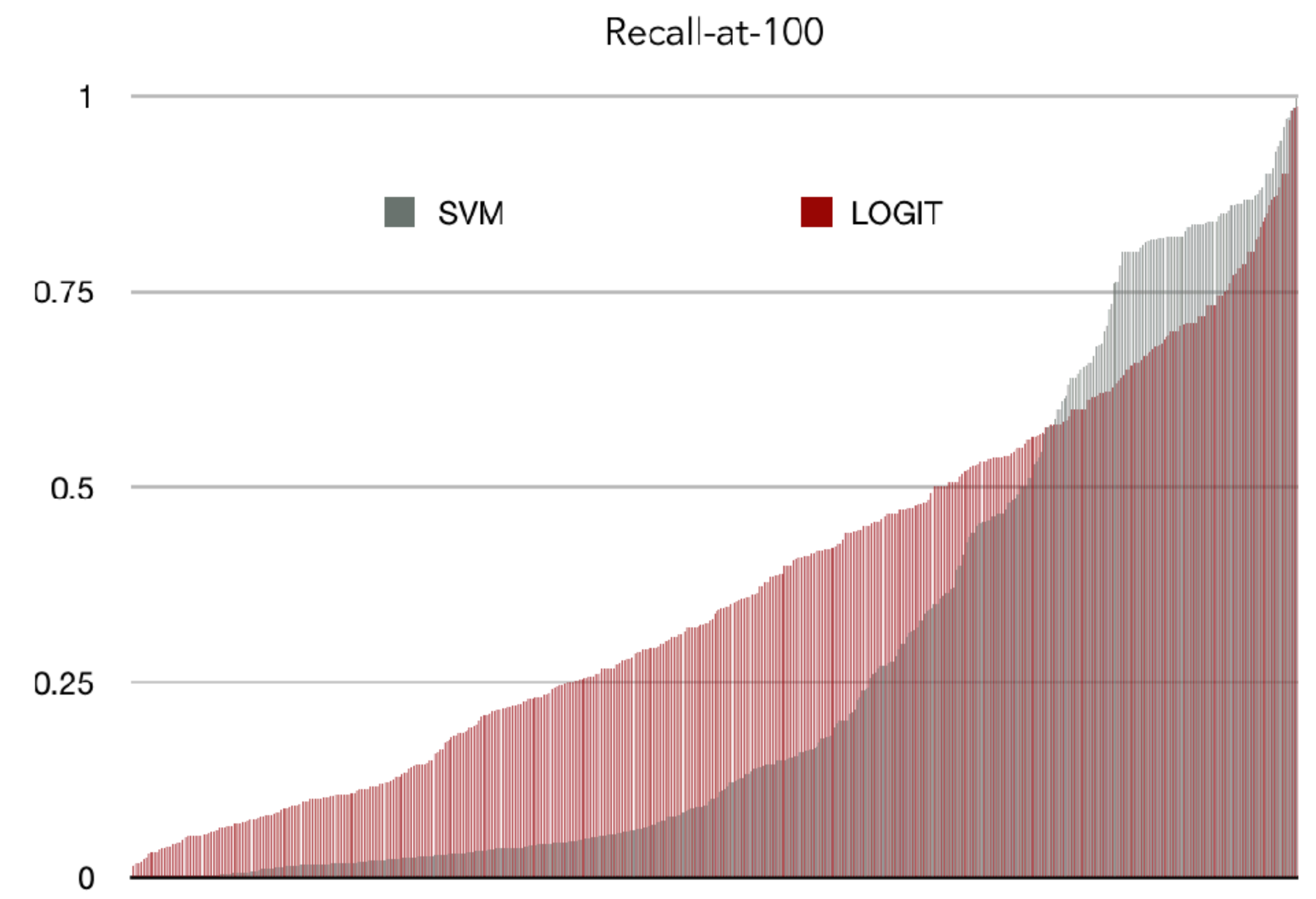
Experiments

For each of the 519 diseases in our dataset, we performed **5-fold cross validation**. On each iteration of the cross validation , we split disease's known protein set into a training and development set. We then trained the model on the train sets and computed accuracy and recall.

Results

PPI Network Embedding Features

	Logistic Regression			Linear SVM		
Metric	Train Accuracy	Test Accuracy	Recall at 100	Train Accuracy	Test Accuracy	Recall at 100
P = 1 Q = 1	0.93	0.79	0.366	0.97	0.77	0.268
P = 2 Q = 0.5	0.92	0.79	0.365	0.98	0.77	0.280
P = 0.5 Q = 4	0.94	0.79	0.386	0.98	0.76	0.270


Recall-at-100
SVM
LOGIT