Here is some more information for the project.

1) Useful resources of experimental data:
   a) NDeX has many types of networks. One useful paper that categorizes >20 popular networks derived from all types of experiments, are entered into NDeX.

   b) Brain-related networks:
   -PsychEncode Consortium
   -Human Brain Atlas
   -Allen Brain Atlas

   c) Hi-C datasets

   d) GEO (https://www.ncbi.nlm.nih.gov/gds), a repository for many gene expression datasets. If you see a paper that did some Next-Generation-Sequencing, check to see if there is a Sequence Read Archive (SRA) or GEO entry.

   e) ENCODE Many kinds of epigenetic data, one of which is ChIP-Seq , which can be seen as a protein-DNA interaction

2) Community and ontology building methods:
   a) Course lectures
   b) Current network community detection procedures: https://www.sciencedirect.com/science/article/pii/S0370157316302964

3) Performance:
   a) Ability to recover unknown synapse related genes. This will be given after training with a set of unknown genes, some related to the synapse and others not.
   b) Ability for the modules to enrich for disease genes. This will be implemented using an enrichment analysis, described below.

4) Output format:
   a) There are two examples of what the ontology format should be. The ID names can be completely arbitrary, or you can name them into known biological parts (e.g. chromatin) if you wanted to use it for some downstream process.

5) Enrichment analysis:
   a) Enrichment analysis starts with a certain set of genes of interest and looks to see if they are associated with certain ontology modules or communities. We will ultimately be testing the ontology on a set of genes related to the synapse and see if certain modules in your ontologies enrich for these genes. While there are many different approaches to test for enrichment in structured tests, we use a

commonly employed hypergeometric test with a p-value correction to account for the multiple hypothesis tested for.

b) The test we will use is implemented in ddot, and a script that will run this enrichment is attached in the Final Project folder (Network_Biology_Enrichment.py).

c) The input will be the ontology network (described in #4), the list of genes of interest, and the number of genes within the the ontology.

d) Performance will be based on the ability to recover the most number of unique test genes found in significant modules relative to the number of total genes contained within the significant modules.

6) Training suggestions:

   a) When building your ontology, you may be facing certain parameter decisions. One way to optimize these is to use enrichment tests. You may download other disease genes that are related to the synapse, or see if specific synaptic components from GO is enriched in your ontology (i.e. are you recovering certain parts of the synapse components in GO).

   b) Additionally, you may use the CliXo algorithm, (also wrapped within the ddot package) to align your synaptic genes to the gene ontology and see if it performs well.

   c) Not overfitting to one specific list tends to help when clustering unseen data. Using negative samples with your ontology may highlight the ability to detect synaptic genes. Additionally, using subsets of synaptic genes may optimize for parameters that can handle different areas of the sampling space.