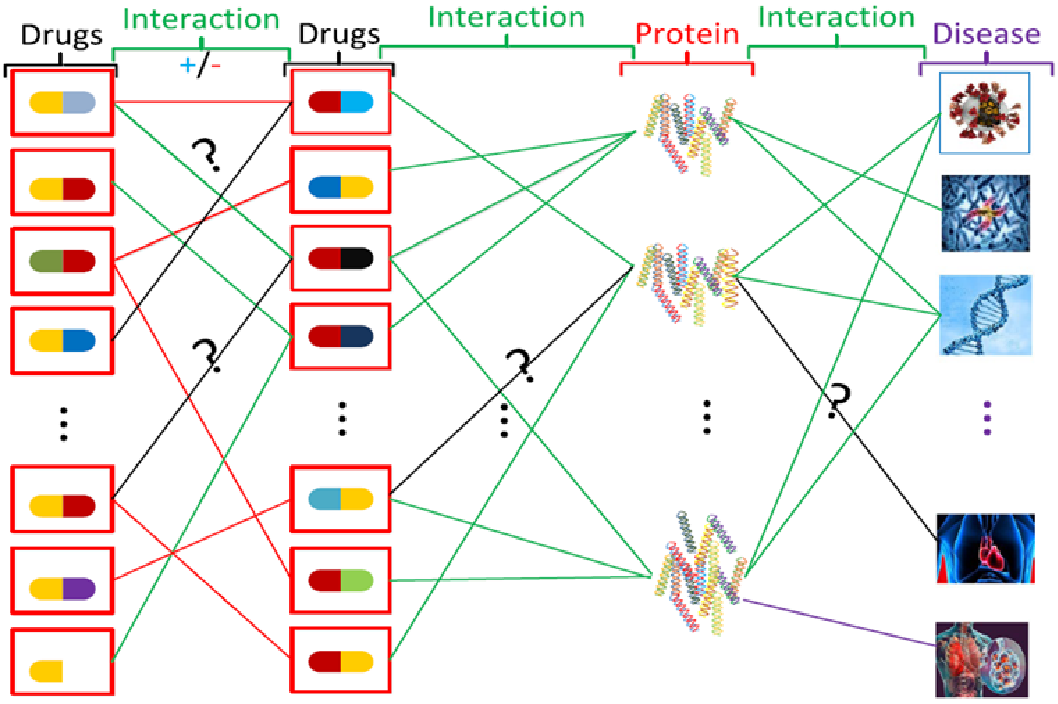
## CSE 412 Introduction to Bioinformatics & CSE 5012 Bioinformatics Project Topics

1. **Graduade** students graduate students must do 2 projects : DTI predictions (%70) and Mechanism of Action (%30)

   **Undergraduade** students must do only DTI predictions (%100)

### a-) DTI predictions (both undergraduade (%100) and graduade (%70) students are responsible)

Graphs (networks) are commonly used to represent a broad range of interactions and associations (as edges) among biomedical entities (as nodes). Developing computational methods to analyze and understand these networks is one of the major research challenges in bioinformatics. A common problem that arises in the analysis of biomedical networks is the prediction of new associations or interactions using existing information on the networks.  This problem is often abstracted in the form of "link prediction", a commonly studied problem in data mining and machine learning. In the context of biomedical networks, link prediction is useful in discovering previously unknown associations or interactions, as well as identifying missing or spurious interactions. Link prediction problems on biological networks include  prediction of drug-target associations (DTI).



Data-driven drug discovery approach that models drug-target interactions (DTI) as networks between two sets of nodes: the drug candidates, and the entities affected by the drugs (i.e. diseases, genes, and other drugs), which are referred to as targets. In this projet your aim is to predict the missing nodes (i.e. drugs and targets) and links between them. For example, you attempt to predict which candidate drugs might treat a list of diseases. There is a large amount of data identifying which drugs treats which diseases, but some diseases have very few drugs available. Thus, discovering which existing drugs can treat them is of great importance. Further, it is critical to determine which drugs have side effects in the presence of other drugs as interactions among drugs may be harmful or lethal to patients. Therefore, drug interaction networks are considered to predict what is the likelihood of reactions between combinations of drugs in a patient's body. Likewise, you can formulate a drug-target interaction network to predict missing links between drugs and target diseases. When considering n drugs, then there will be $n * (n - 1)/2$ combinations of drug–drug relationships for trials. Because a patient could be taking more than two medicines together, the resulting combinations are of an even higher-order and are not feasible to test via experiments. Thus, link prediction offers an important solution. Besides, it allows to find additional uses of existing drugs, with 30% of 84 drugs introduced in 2013 being reused. Many drugs affect more than one particular protein or gene, and some medical conditions involve multiple genes and proteins. Modeling such situations as network interactions and formulating a link prediction problem enables drug-target gene prediction.

Traditional machine learning approaches applied to the drug–target interaction (DTI) problem have many constraints, including dimensionality (for complex and large pharmacological datasets) and incompleteness, sparsity, and heterogeneity (mainly in biological datasets). For instance, logistic regression and support vector machines suffer from the high dimensionality and numerous implicit relationships in the data. These are the result of many factors including measurement technologies and bias problems during the recording of the data. Besides, the spreading speed of diseases or other causes of infection such as viruses evolve quickly is not considered in traditional machine learning methods. In addition, the hierarchical nature of biological data (connections among genes, proteins, and so on) cannot be easily modeled by traditional machine learning approaches. Therefore, there is a need for methods and models capable of addressing these problems.

Network-based approaches are gaining attention because of their simplicity (node and edge representation) which effectively considers high dimensionality and heterogeneity as well as implicit relationships. For drug discovery, these relationships include sharing a common chemical formula and structure or affecting the same protein. Such ability supports reusing existing drugs in new ways, as with a recent breast cancer treatment. As noted previously, this accelerates the drug discovery process, saving time and expense. As an example, other medications such as Duloxetine, used for treating depression, have been found powerful in treating urine leakage issues. Thus, we can consider drug discovery as a missing link problem between chemicals and proteins as shown in Figure.

In this project, your proposed work exploits network-based link prediction models for solving the following pharmaceutical problems:

- **Drug–target interaction prediction**: This task is to predict which drug will affect which protein this is one of the application in drug repurposing.

- **Drug–drug side effect prediction**: From existing drug–drug side effect data, you can create a network, in which a link reflects the two drugs (nodes) has shown some side effects. So in this task you will predict which new drug combinations can cause side effect.

- **Disease–gene association prediction**: Some disease affects the genes which is more lethal as it can transfer to the next generations. Therefore in this task you will aim to predict which new disease can affect which particular gene.

- **Disease–drug association prediction**: Some drugs might not be pharmaceutical chemicals such as arsenic. So in this problem, you will aim to predict which drug is associated to which disease.

To carry out your tasks, you first build the networks  bipartite, and ensured they were undirected then apply the link prediction models to solve the following drug discovery problems:

1. **Drug–Target prediction**: The aim is to predict which drug will affect which unknown proteins, considering the bipartite networks of drugs and their target proteins.

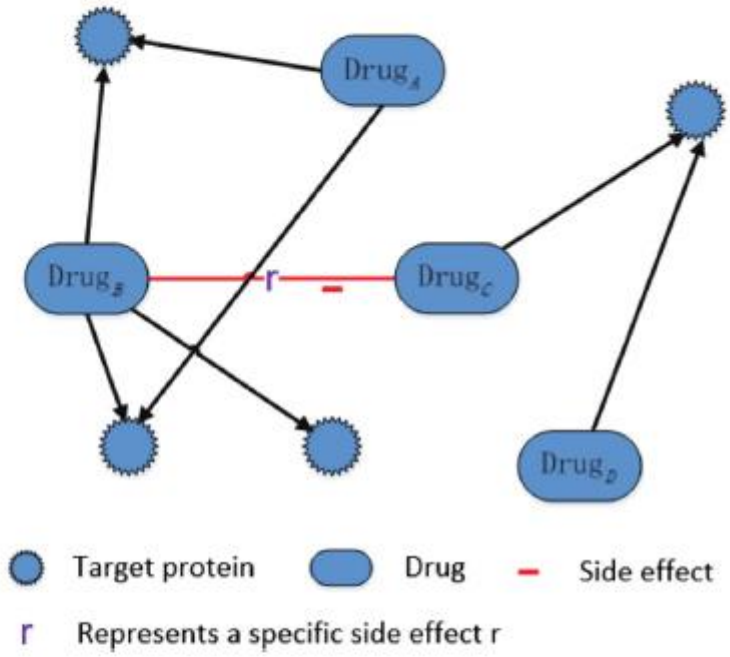
2. **Drug–Disease prediction**: The aim is to study drug chemical structures and target proteins (as the disease and drugs both affects proteins) to find similarities between drug structures. It has been found that similar drug structures affect similar proteins. You represent each chemical structure as a network. Once a similar drug is found, it can be used to target similar proteins. There is currently a lack of systematic research in this area.

3. **Drug–Drug reaction prediction**: This problem examines the search for combinations of drugs for conditions that require targeting more than one protein, as with degenerative neurological conditions, such as Alzeheimer's and Parkinson's. You will incorporate known combinations that cause adverse side effects (headaches, vomiting, rashes, etc.) to predict which additional combinations might cause reactions in patients.

4. **Disease–Gene association prediction**: Thanks to high-throughput screening technologies, you have large volumes of genomic data. Yet, there are many diseases for which a genomic basis is unknown. Genomic alleles and malignant mutations are continuously sequenced, which is why most of them are identified or annotated. Traditionally, linkage analysis has been done to find non-experimental disease-gene associations, and it has been based on the likelihood of observing alleles. However, this kind

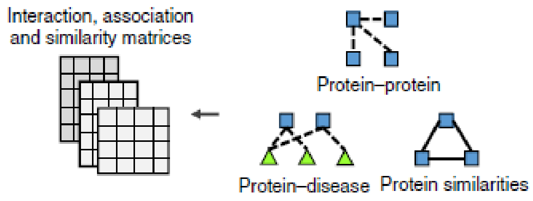
of analysis fails for a multifactorial and heterogeneous diseases. Considering genomic data association is a newer approach to solve this problem, but with the downside of producing hundreds of candidates for complex diseases, which hinders experimental validation. Therefore, you will present a network approach for genomic data analysis.

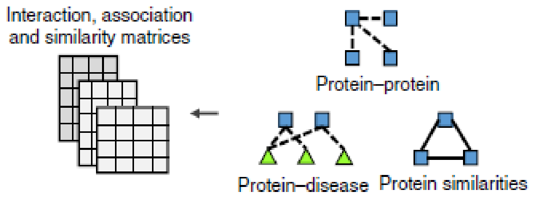
In addition to existing network approaches, such as common neighbour, path-based and random walk-based methods, recent developments in network-based learning technologies, such as geometric deep learning, offer the prospect of using genomic data to find gene-disease associations that are still unknown.



Example Drug heterogeneity network. The figure is a heterogeneous DDI network with one kind of side effect

**Summary of the Project :**





In your Project you first integrate a variety of drug-related information sources to construct a heterogeneous network and applies a compact feature learning algorithm to obtain a low-dimensional vector representation of the features describing the topological properties for each node.

1. You need to develop a computational pipeline to predict novel drug–target interactions from a constructed heterogeneous network, which integrates diverse drug-related information.
2. Your proposed model focuses on learning a low-dimensional vector representation of features, which accurately explains the topological properties of individual nodes in the heterogeneous network, and then makes prediction based on these representations via a vector space projection scheme.
3. Moreover, you experimentally validate the novel interactions between drugs and targets, and demonstrate the new potential applications of your results. You must use the **evaluations metrics AUROC and AUPR**
4. In order to integrate the heterogeneous information to predict new drug–target interactions , you can propose an end–to–end network model that predicts DTIs from low level representations by using GCN to achieve the task of link prediction.

**Hint about Project:**

Link prediction is an important and well-studied problem in network biology. Recently, graph representation learning methods, including Graph Convolutional Network (GCN)-based node embedding have drawn increasing attention in link prediction. Graph Representation Learning (GRL) can effectively combine feature information and structural information to obtain low-dimensional embedding and Graph Convolutional Network (GCN) model.

In otherwords,you can propose machine learning model using node similarity matrices (computed using local measures of node similarity) as convolution matrices for GCNs that are used to compute node embeddings for link prediction. In the context of various machine learning tasks, GCNs facilitate the use of network topology in computing latent features from input features associated with network nodes. GCNs are also used to compute node embeddings, i.e., features that represent network topology, by setting the loss function appropriately to capture the correspondence between the embeddings and network topology.

**Using GCN is not mandatory, just recommended**. All groups can design their own original models.

**Grading about Project :**

a.Designing Original Graph Based Machine Learning Model (% 40)

b.Comparison of your results with baselines and experimental results (% 30)
**Evaluations metrics:** AUROC, AUPR

c.Report (% 30)

**Datasets:** https://github.com/luoyunan/DTINet/tree/master/data

**References :** Abbas, K., Abbasi, A., Dong, S. et al. Application of network link prediction in drug discovery. BMC Bioinformatics 22, 187 (2021). https://doi.org/10.1186/s12859-021-04082-y

Luo, Y., Zhao, X., Zhou, J. et al. A network integration approach for drug-target interaction prediction and computational drug repositioning from heterogeneous information. Nat Commun 8, 573 (2017). https://doi.org/10.1038/s41467-017-00680-8

Cheng S, Zhang L, Jin B, Zhang Q, Lu X. Drug Target Prediction Using Graph Representation Learning via Substructures Contrast. Preprints.org; 2021. DOI: 10.20944/preprints202103.0337.v1.

**b-) Mechanism of Action (%30 only for graduade students are responsible)**

What is the Mechanism of Action (MoA) of a drug? And why is it important?

In the past, scientists derived drugs from natural products or were inspired by traditional remedies. Very common drugs, such as paracetamol, known in the US as acetaminophen, were put into clinical use decades before the biological mechanisms driving their pharmacological activities were understood. Today, with the advent of more powerful technologies, drug discovery has changed from the serendipitous approaches of the past to a more targeted model based on an understanding of the underlying biological mechanism of a disease. In this new framework, scientists seek to identify a protein target associated with a disease and develop a molecule that can modulate that protein target. As a shorthand to describe the biological activity of a given molecule, scientists assign a label referred to as mechanism-of-action or MoA for short.

How do we determine the MoAs of a new drug?

One approach is to treat a sample of human cells with the drug and then analyze the cellular responses with algorithms that search for similarity to known patterns in large genomic databases, such as libraries of gene expression or cell viability patterns of drugs with known MoAs.

In this competition, you will have access to a unique dataset that combines gene expression and cell viability data. The data is based on a new technology that measures simultaneously (within the same samples) human cells' responses to drugs in a pool of 100 different cell types (thus solving the problem of identifying ex-ante, which cell types are better suited for a given drug). In addition, you will have access to MoA annotations for more than 5,000 drugs in this dataset.

As is customary, the dataset has been split into testing and training subsets. Hence, your task is to use the training dataset to develop an algorithm that automatically labels each case in the test set as one or more MoA classes. Note that since drugs can have multiple MoA annotations, the task is formally a multi-label classification problem.

How to evaluate the accuracy of a solution?
Based on the MoA annotations, the accuracy of solutions will be evaluated on the average value of the logarithmic loss function applied to each drug-MoA annotation pair.

If successful, you'll help to develop an algorithm to predict a compound's MoA given its cellular signature, thus helping scientists advance the drug discovery process.

**Datasets :** https://www.kaggle.com/c/lish-moa/data

2. **Report format:** The paper formats below are prepared according to A4-size paper format as presented in IEEE website

   https://www.ieee.org/conferences/publishing/templates.html

   All reports should be written with the headings below.

   Abstract (summary of your contributions and what challenges you have solved)
   1.Introduction
   2.Related Work (Literature Review)
   3.Method
   4.Experimental Result
   5. Conclusion

**3. All the Python codes and report will be checked with The Copyleaks code plagiarism checker and CodeGrade. Reports and codes must be original**