Introduction to Systems Biology: Network models inferring insights from -omic and imaging analysis.

4-hrs crash-course

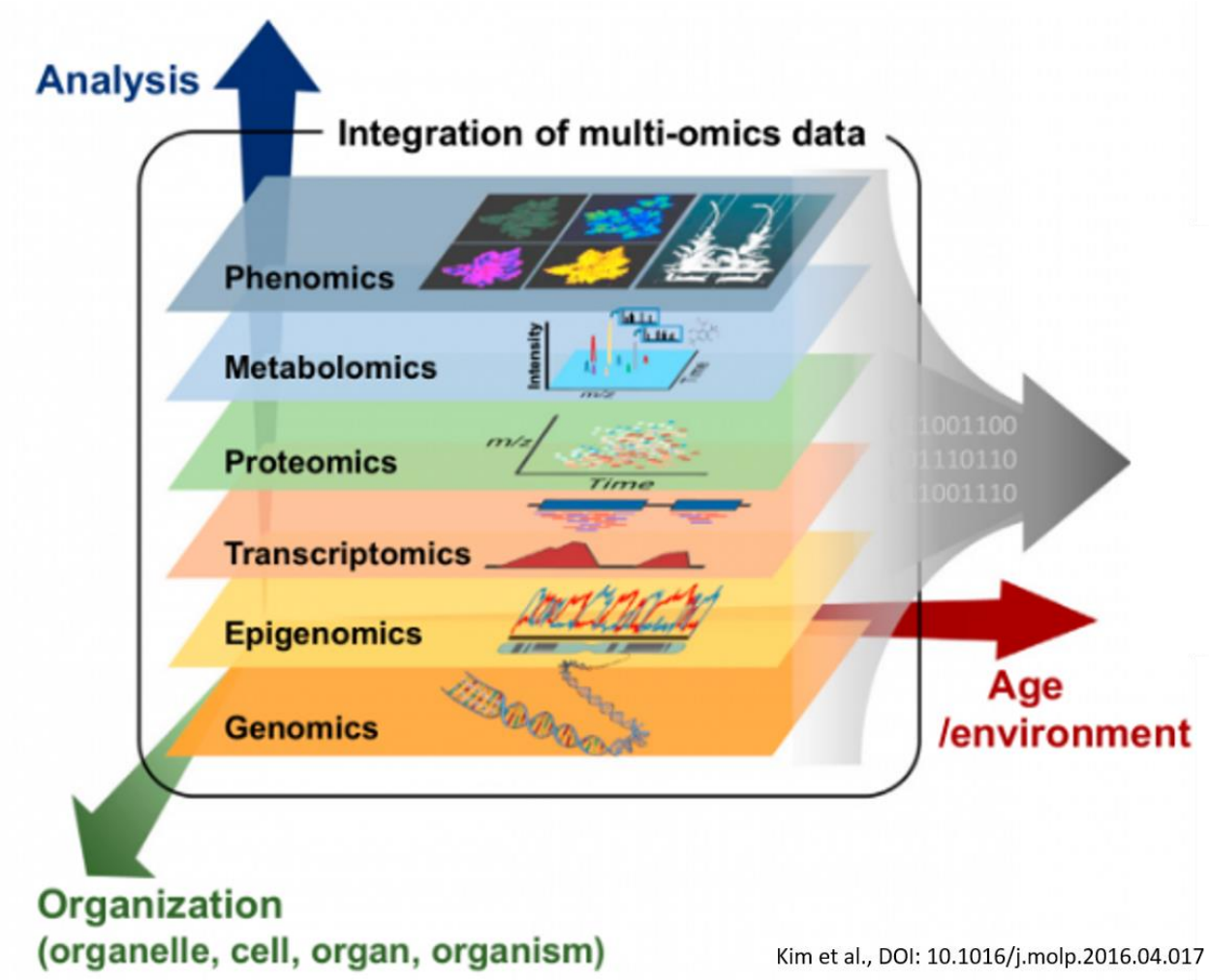
George Gavriilidis MSc, PhD

Post-doctoral researcher in Systems Pharmacology and single-cell omics

Systems Biology – What is the “Systems”?

- What is Systems Biology?
- Biology itself is an inclusive term… in the context of this course, biology encompasses
- Molecular → Cellular → Tissue Organ → Physiological Function
- ***Systems biology***

how molecules interact and come together to give rise to subcellular machinery that forms the functional units capable of operations that are needed for cell, tissue/organ level physiological functions

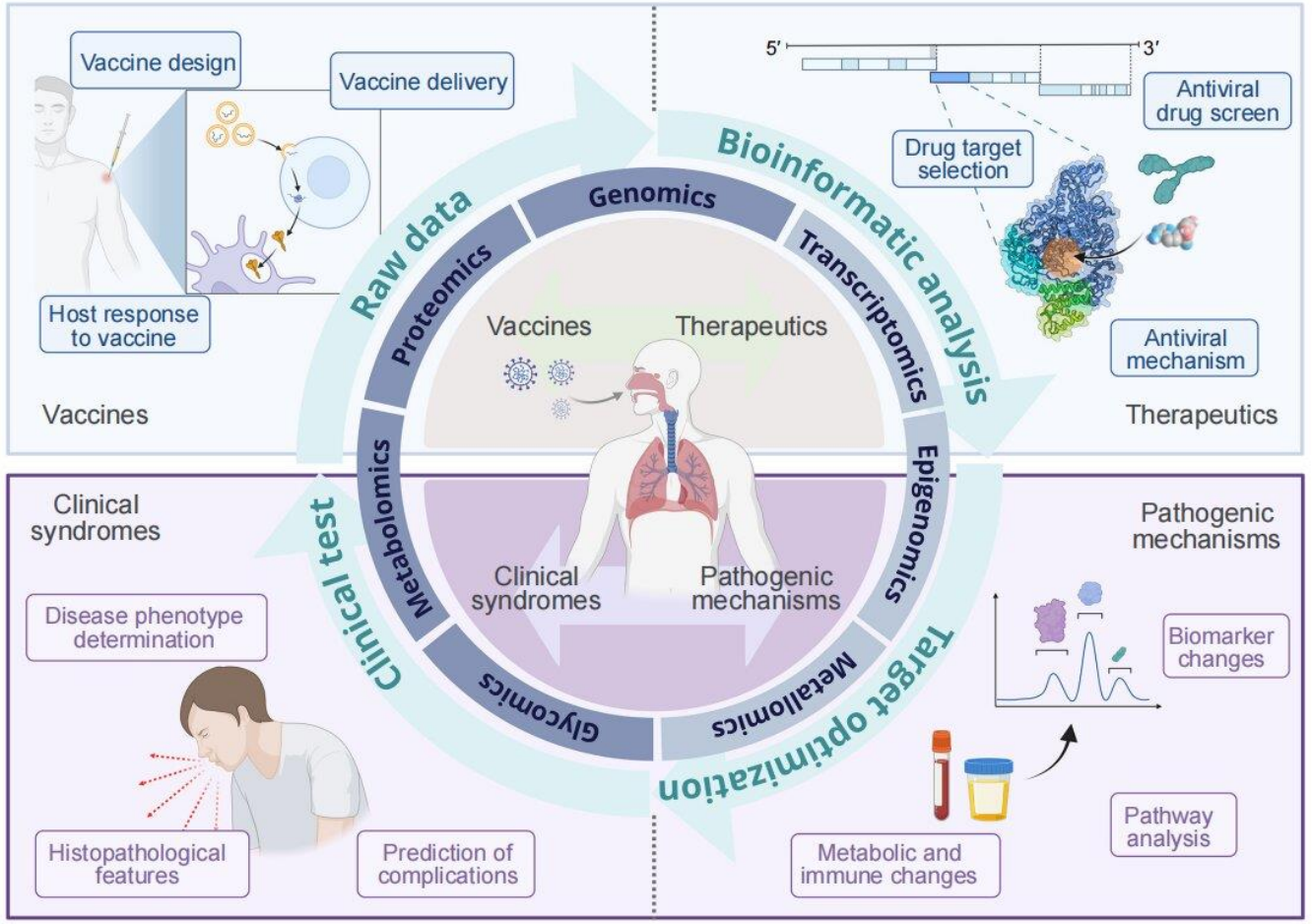


Kim et al., DOI: 10.1016/j.molp.2016.04.017

MORPHOPHEN

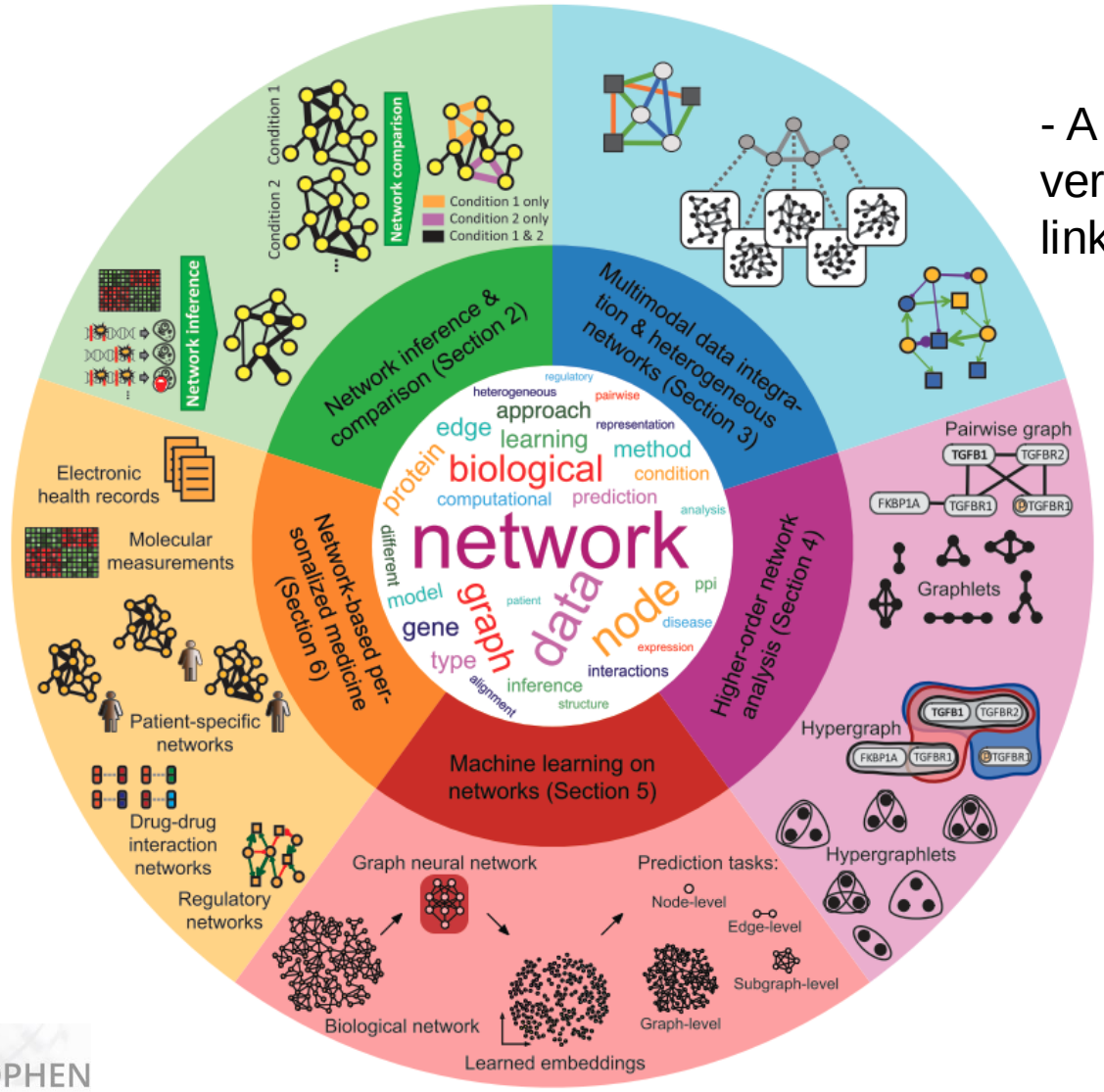
Why do we need Systems Biology?

- All changes in cell state both normal and disease related are at least part due to changes in gene expression, proteins, metabolites… – hence ***concurrent surveying of these omic modalities is crucial for therapeutic results (e.g., COVID-19 vaccines)***



https://medicalxpress.com/news/2023-07-multi-omics-generation-covid-therapeutics-vaccines.html

Network Biology - an overview



- A network (or graph) is comprised of a set of nodes (or vertices) that are connected by a set of edges (or links)

- Networks allow us to study the properties of a complex system that emerge from interactions between its components.

- The focus is on **biological networks** (genes, proteins, metabolites, patients, drugs, diseases)

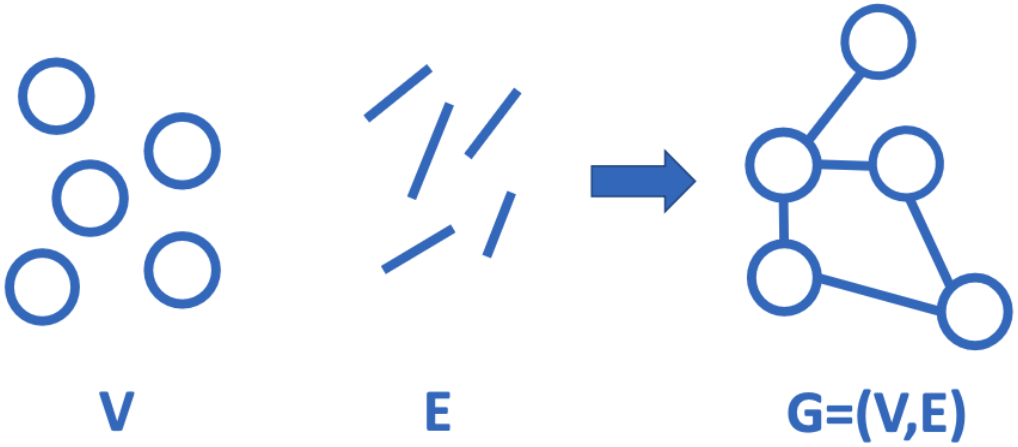
- **Network biology** encompasses computational (e.g., algorithms, graph theory, network science, data mining, and machine learning) and biological sciences.

Current and future directions in network biology, Zitnik M et al. 2023

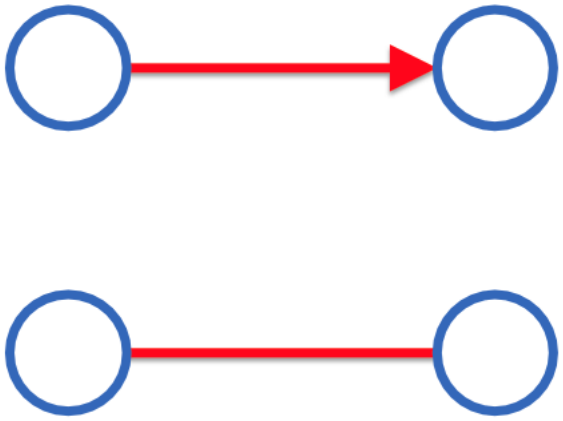
MORPHOPHEN

Network models – The basics

V: A nonempty finite set of vertices (nodes)
E: A nonempty finite set of pairs of vertices (edges)



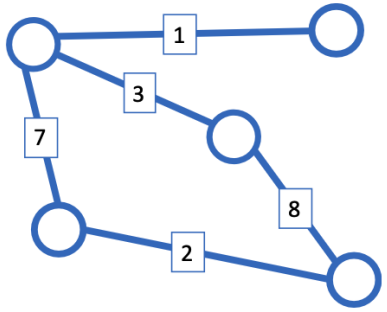
Edges can be
- Directed (a,b) ≠ (b,a) (e.g., metabolic networks)
- Undirected (a,b) = (b,a) (e.g., PPI networks)



In a biological network, nodes may represent:
- genes
- proteins
-drugs
-biological pathways
-diseases

Edges represent relationships between pair of entities

Edges can be weighted!



MORPHOPHEN

Monopartite vs Bipartite graph

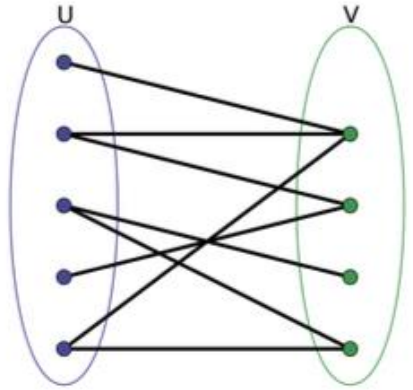
In a biological network, **<u>nodes</u> may represent different types of elements**.. The most common is that they include two types of elements:

-genes
-proteins

Versus

-Regulators
-Drugs
-Biological functions
-Diseases
-etc

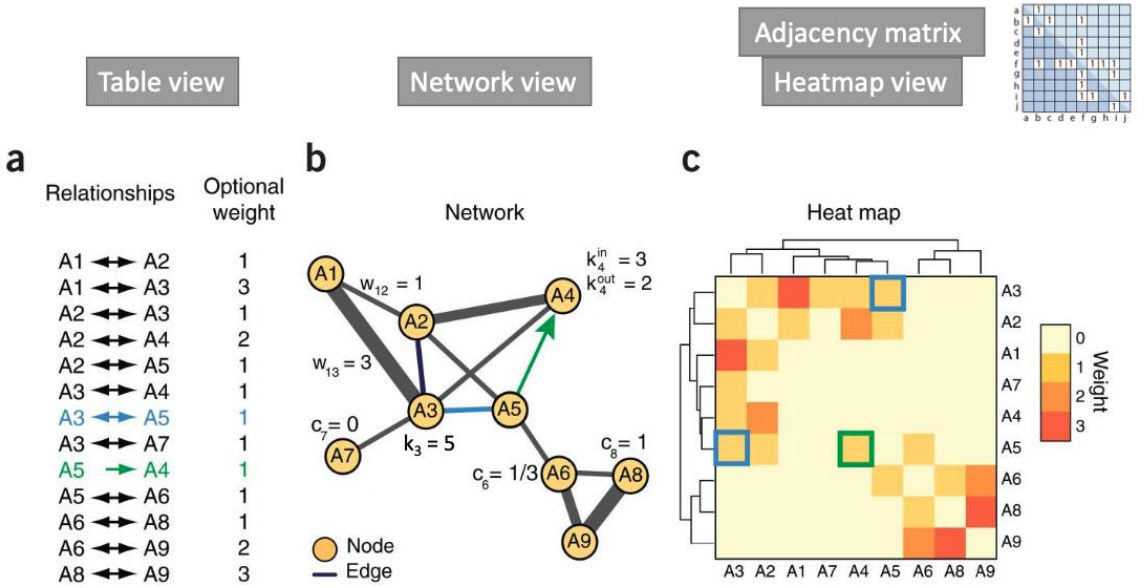
<u>Edges</u> represent relationships between the pairs of entities and are generally directed

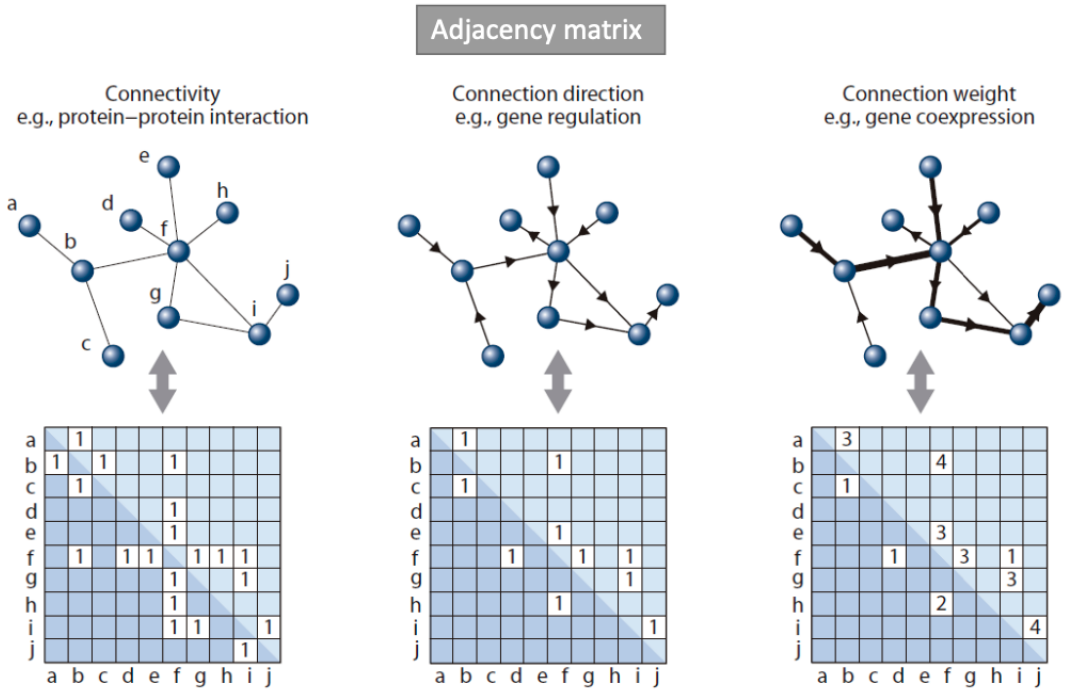


Nodes can represent two different entities in the same graph (**bipartite** graphs)

MORPHOPHEN

Network Adjacency matrix

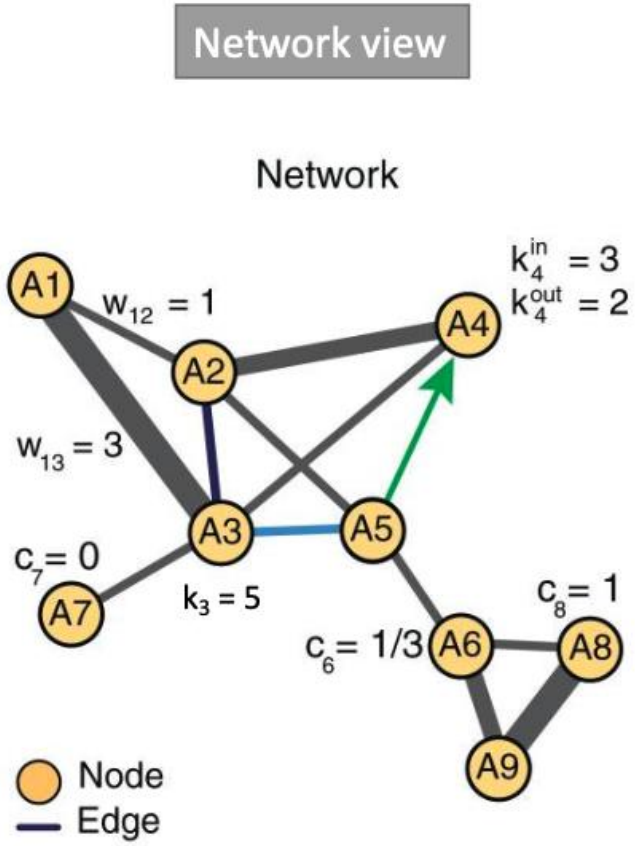




Gerstein et al. (2018) *Annu Rev Biomed Data Sci*

Merico, Gfeller & Bader (2010) *Nature Biotechnology*

Node and edge attributes



We can add **visual information** (like edge width representing the weight)

We can measure some properties, like:

- ***ki***: node degree, the number of edges attached to a node. If edges have directions, we can distinguish between the in-degree and the out-degree

- ***cj***: clustering coefficient, which counts the number of edges among the neighbors of a node, divided by the maximal possible number of such edges

Graph theory concepts in biological networks

- **Network topology** plays a vital role in understanding network architecture and performance.

- Several of the most important and commonly used topological parameters include:

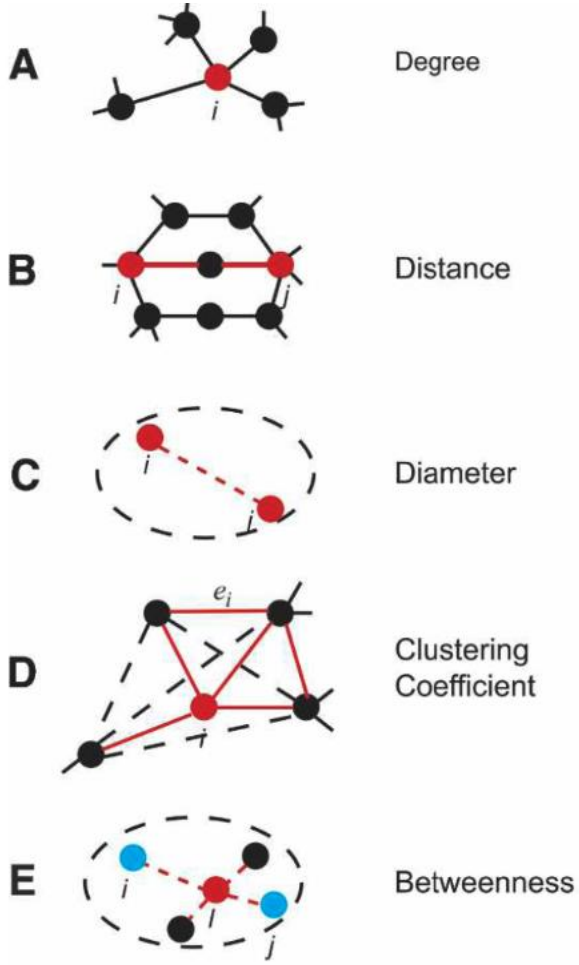
degree number of links connected to 1 vertex

distance shortest path length

diameter maximum distance between any two nodes

clustering coefficient number of links between the nodes within its neighbourhood divided by the number of possible links between them

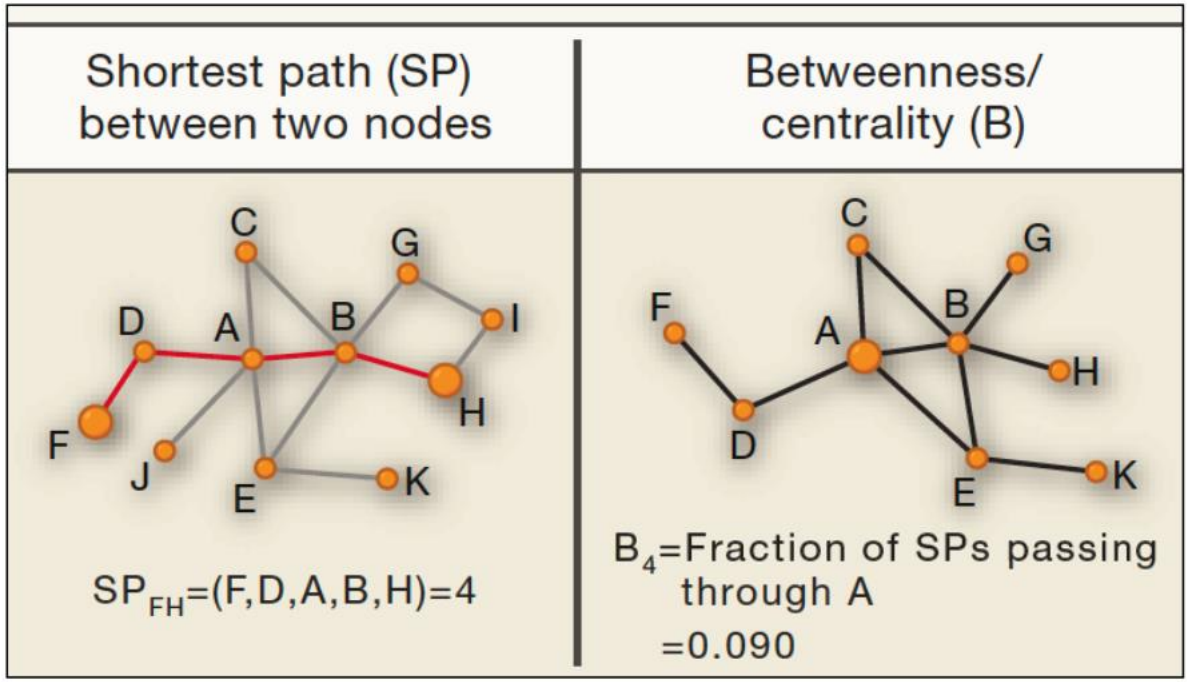
betweenness fraction of the shortest paths between all pairs of vertices that pass through one vertex or link



Zhu et al. (2007) Genes Dev.

“Paths” in networks

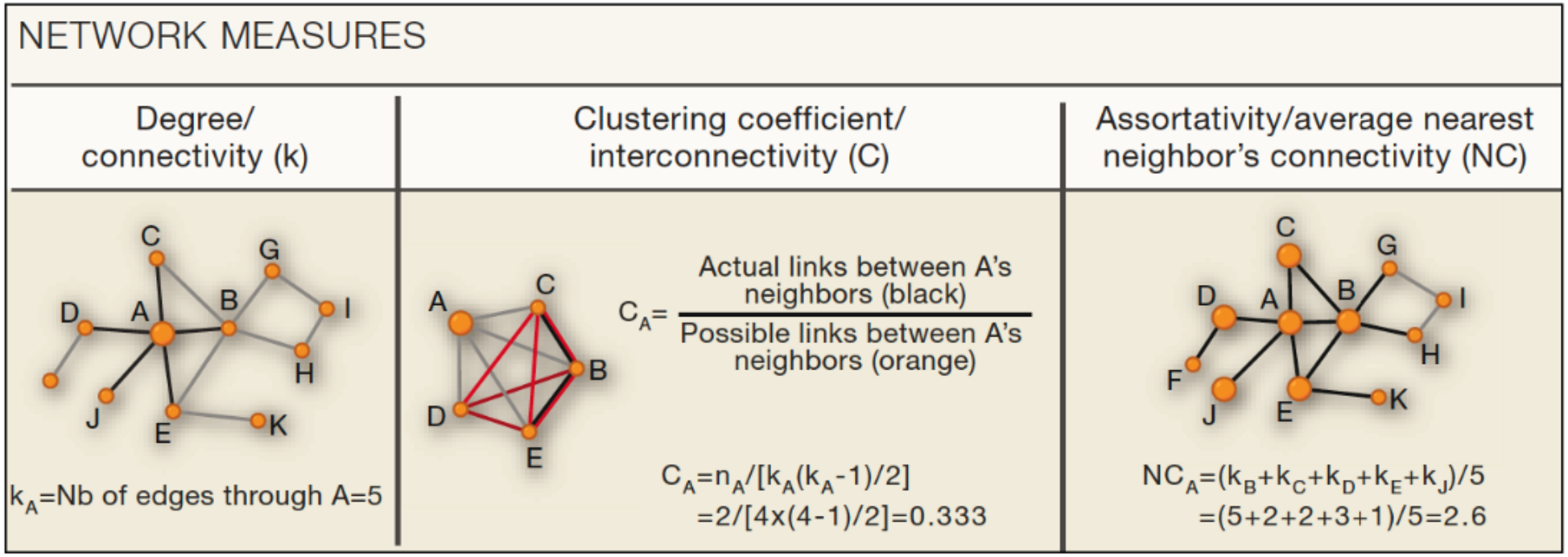
Network measures related to "number of ways" (path-ways): – **shortest path**
– **betweenness** = centrality



Seebacher & Gavin (2011) Cell

Graph theory concepts in biological networks

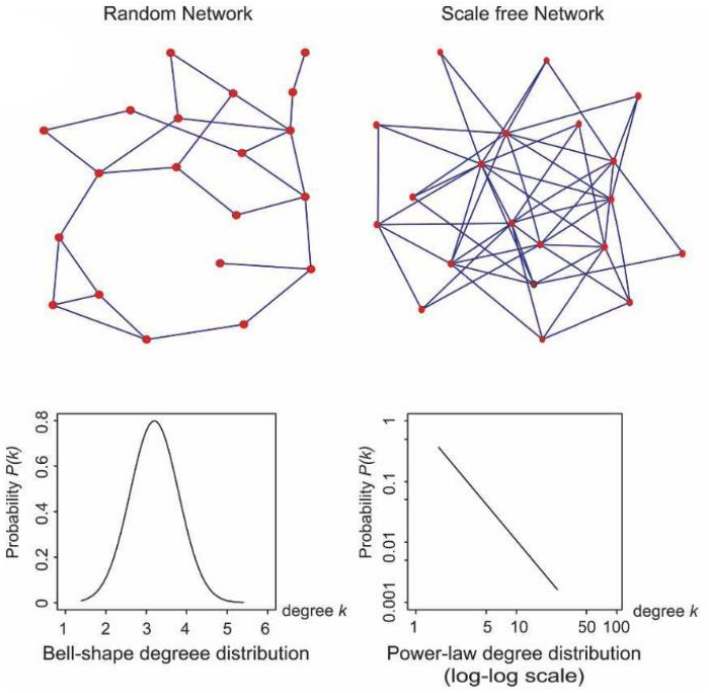
Network measures related to "number of friends" (connectivity):
– **degree** = connectivity
– **clustering coefficient** = inter-connectivity
– **assortativity** = average nearest neigbor's connectivity

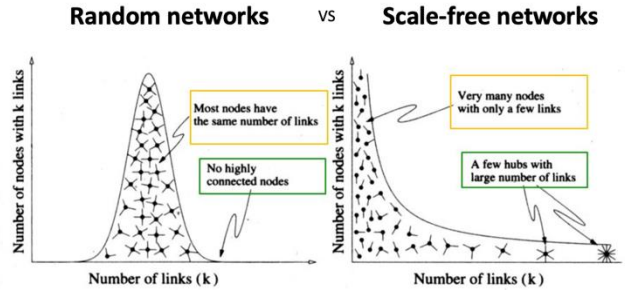


MORPHOPHEN

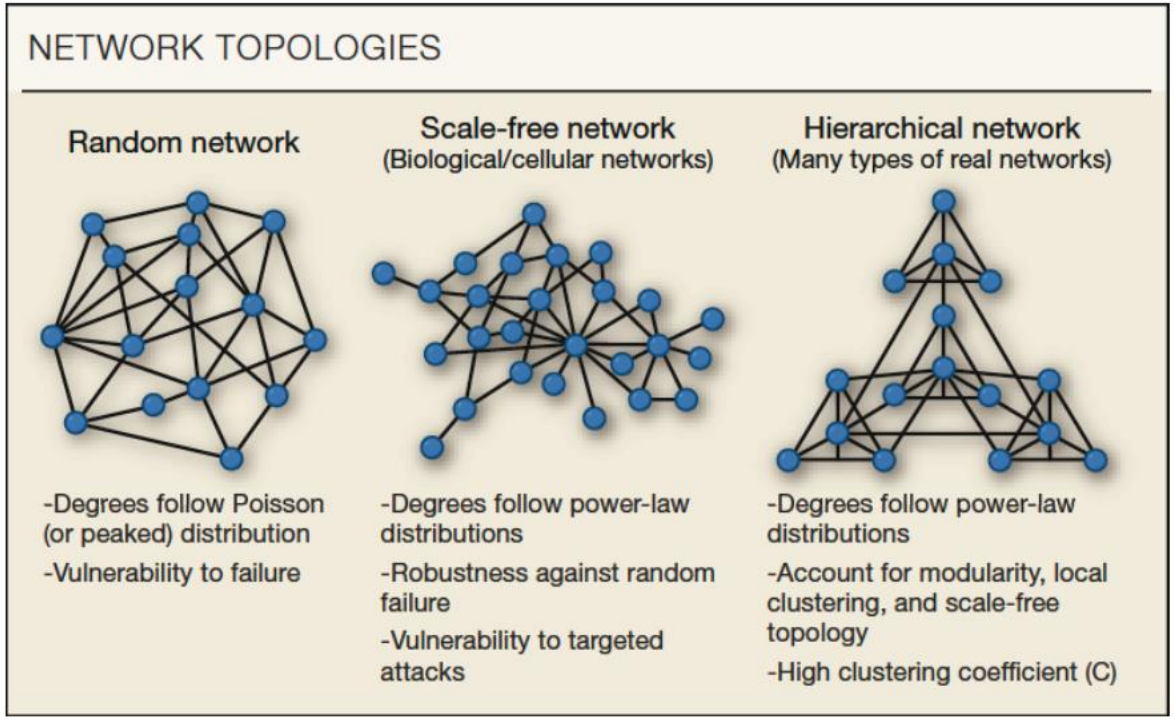
Seebacher & Gavin (2011) Cell

Random networks vs Scale free Network





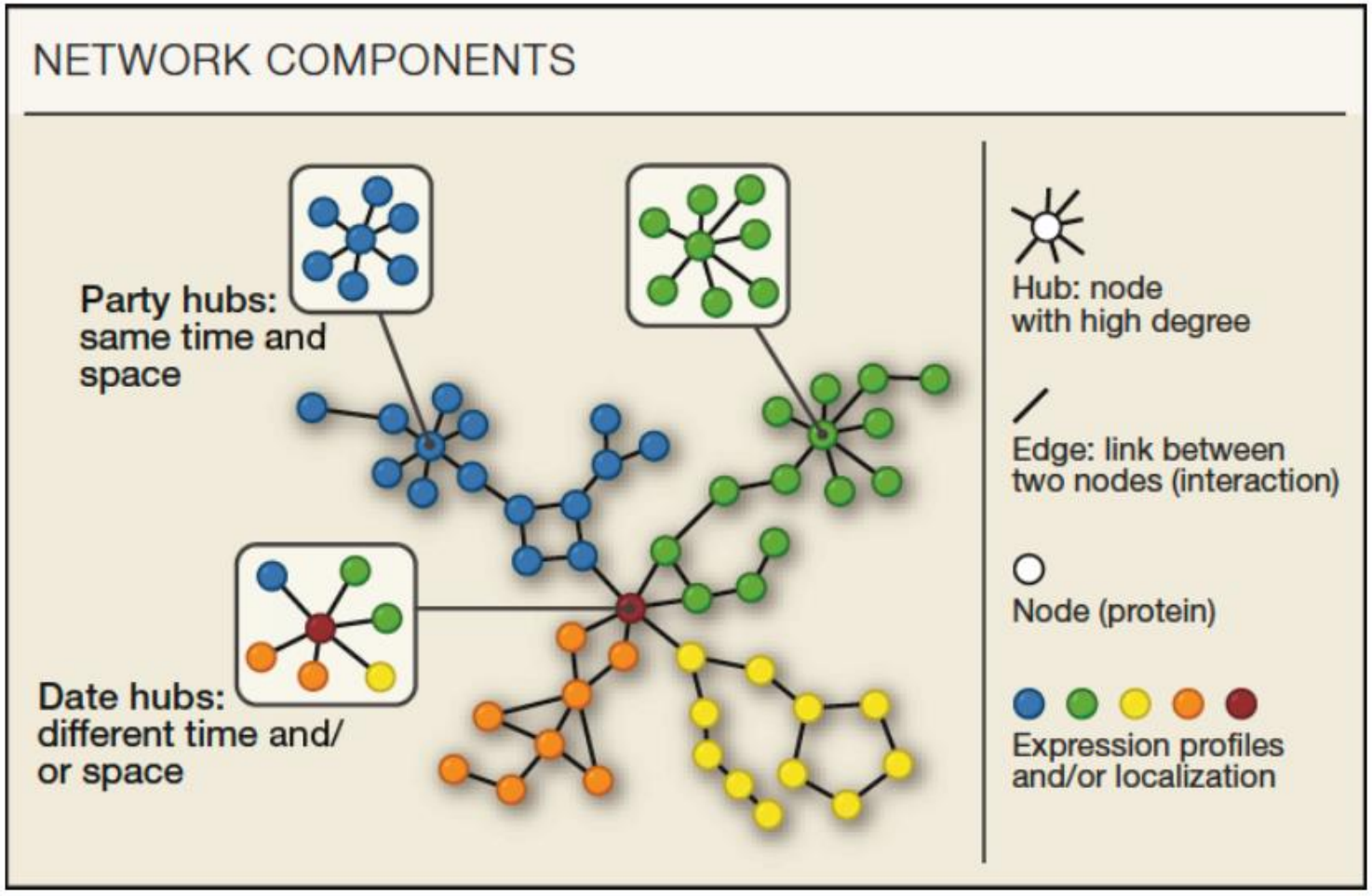
Biomolecular networks **are scale-free**: A scale-free network has more high-degree nodes and a power-law distribution, which leads to a straight line when plotting the total number of nodes with a particular degree versus that degree in log-log scales



Zhu et al. (2007) Genes Dev.
Seebacher & Gavin (2011) Cell
Linked: The New Science of Networks" by Albert-László Barabási

MORPHOPHEN

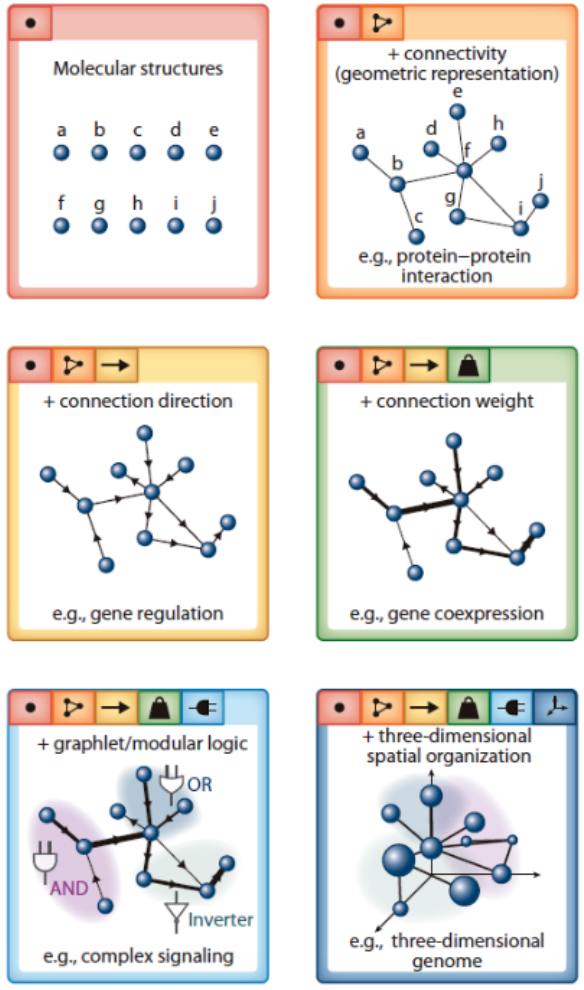
Networks like to party and mingle!



MORPHOPHEN

Seebacher & Gavin (2011) Cell

From abstract graph theory to Network Biology



Network inference

Biological DATA

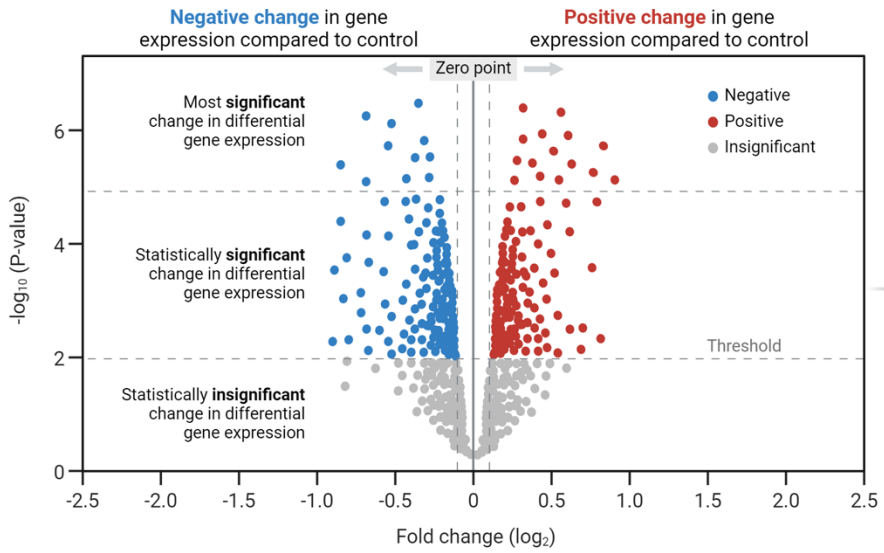
nodes + edges

"Network representations can be built through a progressive layering of information and logic"

Biological NETWORKS

Gerstein et al. (2018) *Annu Rev Biomed Data Sci*

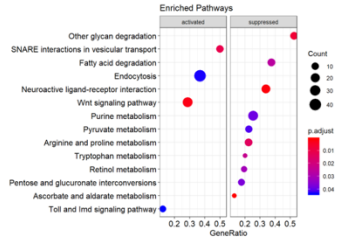
Implementing network analysis in omic studies

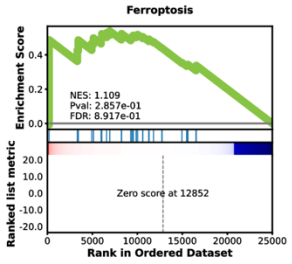


ORA

Pathway enrichment

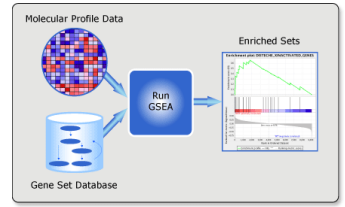
GSEA



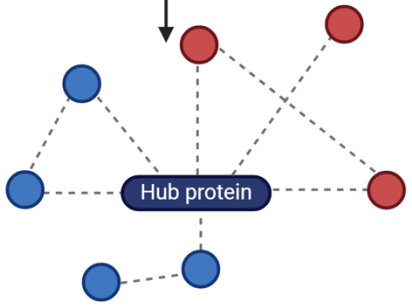


Over-representation analysis (ORA): are genes from pre-defined sets (ex: GO term or KEGG pathway) more ***than would be expected*** (over-represented) in a subset of your data?
(blue and red calculated <u>separately</u>)

Gene Set Enrichment Analysis (GSEA) is a computational method that determines whether an *a priori* defined set of genes shows statistically significant, concordant differences between two biological states (e.g. phenotypes)
(blue and red are <u>co-calculated</u>)

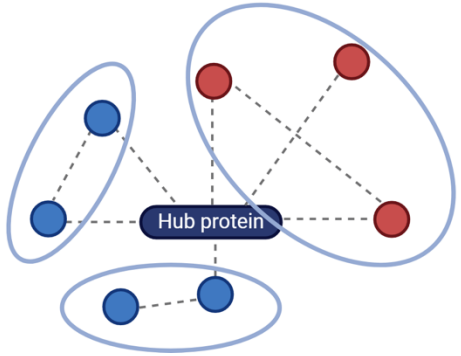


Network view of perturbations



Association PPI network

Prioritizing network elements



Clustering

Further analysis...

+ Biological plausibility based on curated datasets

- Minimal description of experimental-ground-truth (what is happening in our samples????) – <u>only differential expression can be mapped…</u>

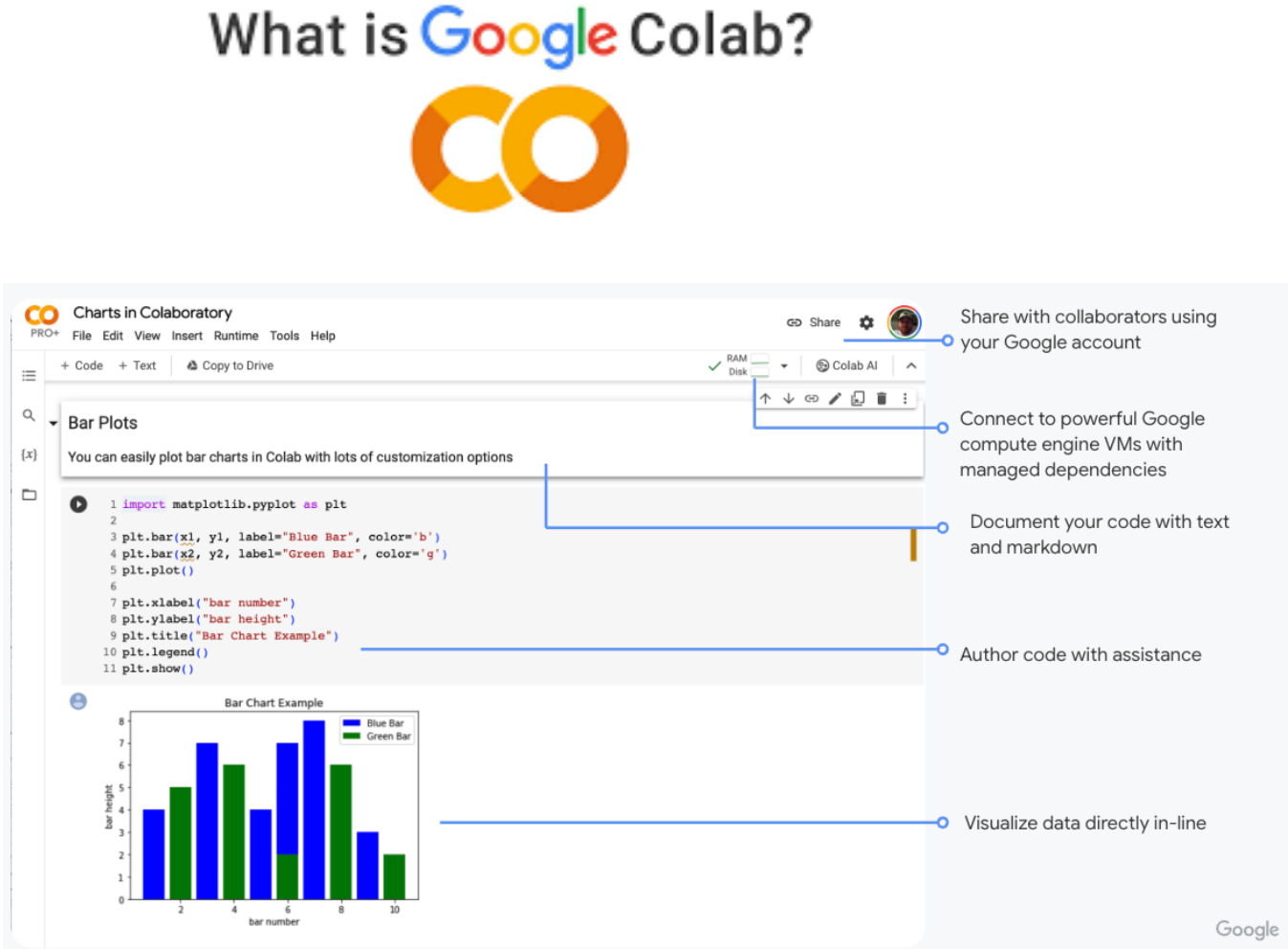
MORPHOPHEN

Recap no1!

- Systems Biology aspires to provide a holistic approach to how cells, tissues and organs operate by analyzing high-throughput biological data (“omics”: mRNA, proteins, metabolites, epigenetic changes etc..)
- Networks are a mathematically driven approach (*graph theory*) to represent and analyze omic data – The field is called Network Biology
- Network topology, pathways and distribution can reveal underlying biological motifs

Hands-on practical #0: Google Colab

- Go to https://colab.research.google.com/github/geogav/morphophen_networks_spatial/blob/main/notebooks/MorphoPhen_Networks_Module1.ipynb **-> File -> Save a Copy in Drive**
- **We run Python on the Google Cloud**
- **Cloud compute requires installation of packages and has by default a temporary working directory**



Hands-on practical #1

🧬 Dataset Overview: RNA-Seq Analysis of iPS-Derived Neuronal Samples

We will be working with a publicly available dataset from the **Gene Expression Omnibus (GEO)**, which focuses on RNA-Seq analysis of purified **induced pluripotent stem (iPS) cell-derived neuronal samples**. This study investigates gene expression differences between **midbrain dopaminergic (mDA) neurons** derived from **Parkinson's disease (PD) patients** and **healthy controls**.

📂 Dataset Information

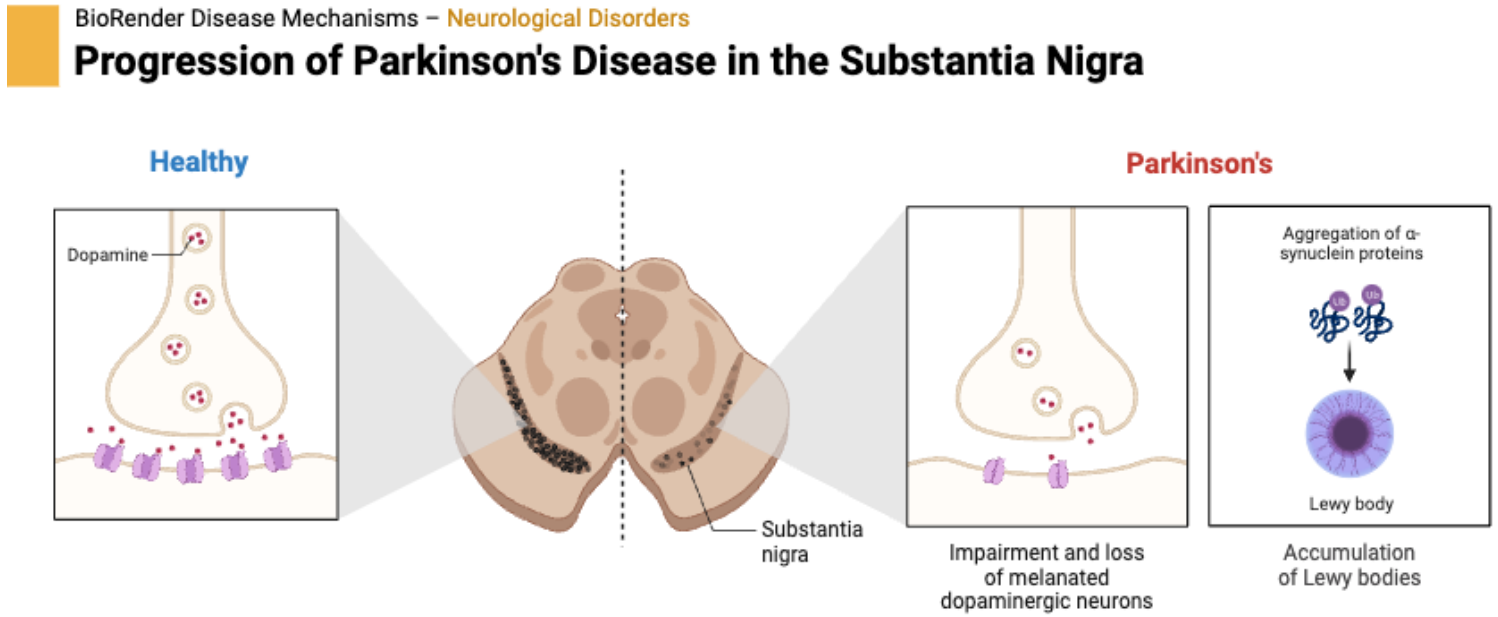
Attribute	Details
🔗 GEO Accession	GSE62642
📊 Number of GEO Samples	14
🗃️ Number of SRA Runs	14
🧬 Species	*Homo sapiens*
📝 Title	RNA-Seq Analysis in Purified iPS Cell-Derived Neuronal Samples

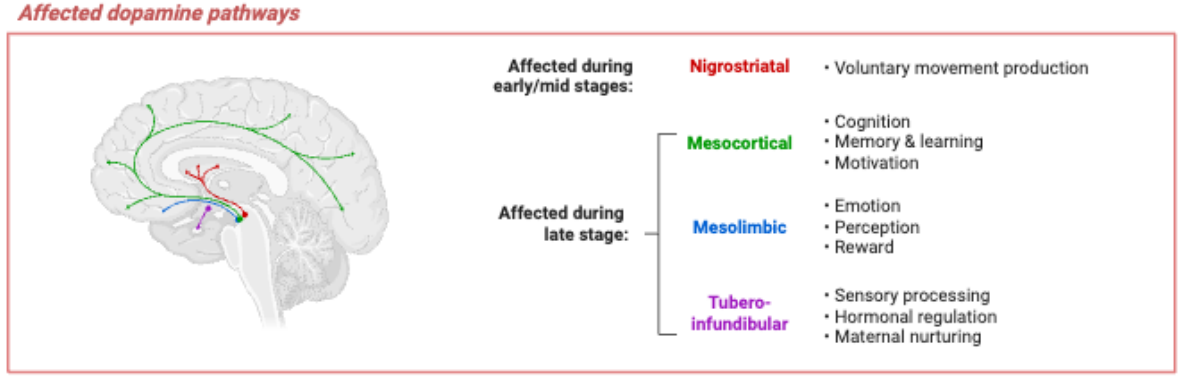
🧩 Study Summary

This dataset characterizes **gene expression differences** in **midbrain dopaminergic (mDA) neurons** from:

- 🧪 **6 Parkinson's disease (PD) cases**
- 🧬 **8 healthy controls**

Hands-on practical #1

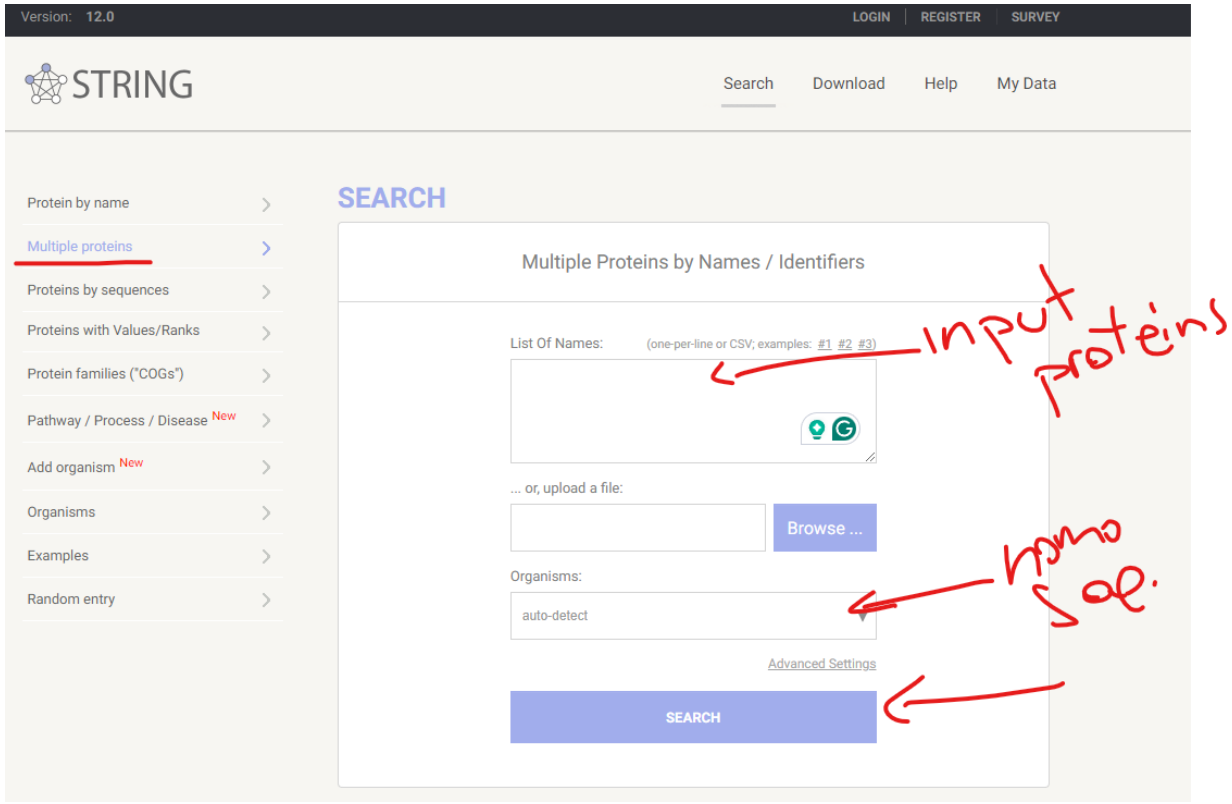


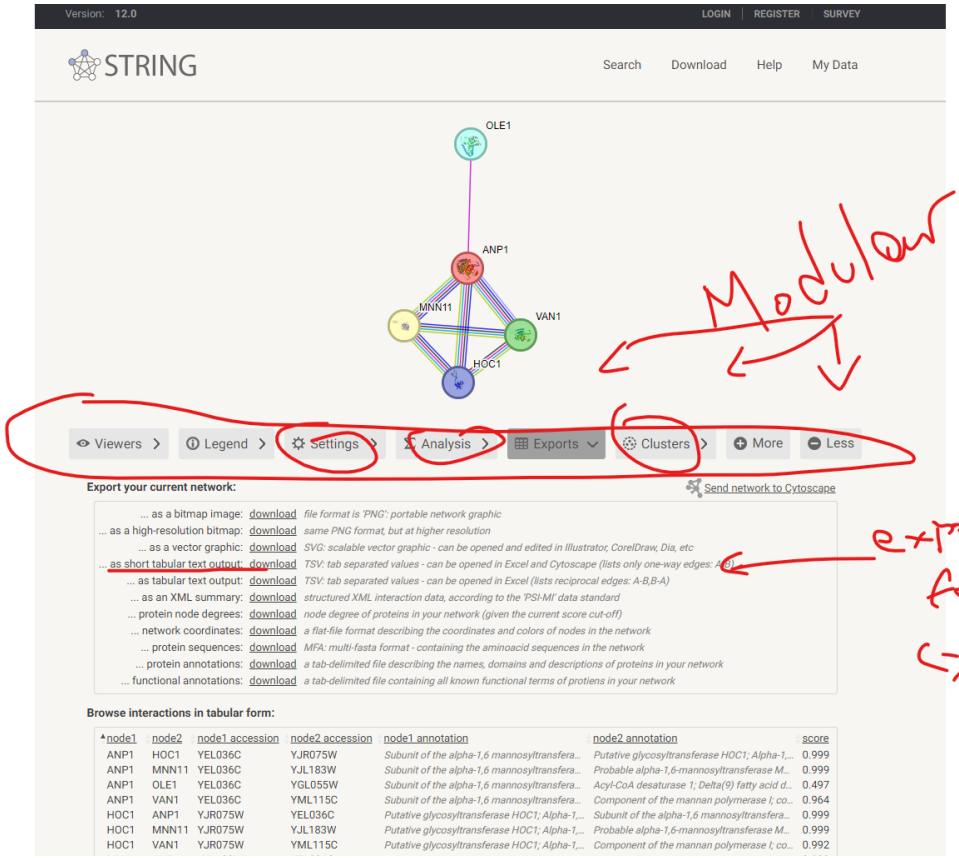


Hands-on practical #1 – Working on GREIN platform to analyse RNA-seq

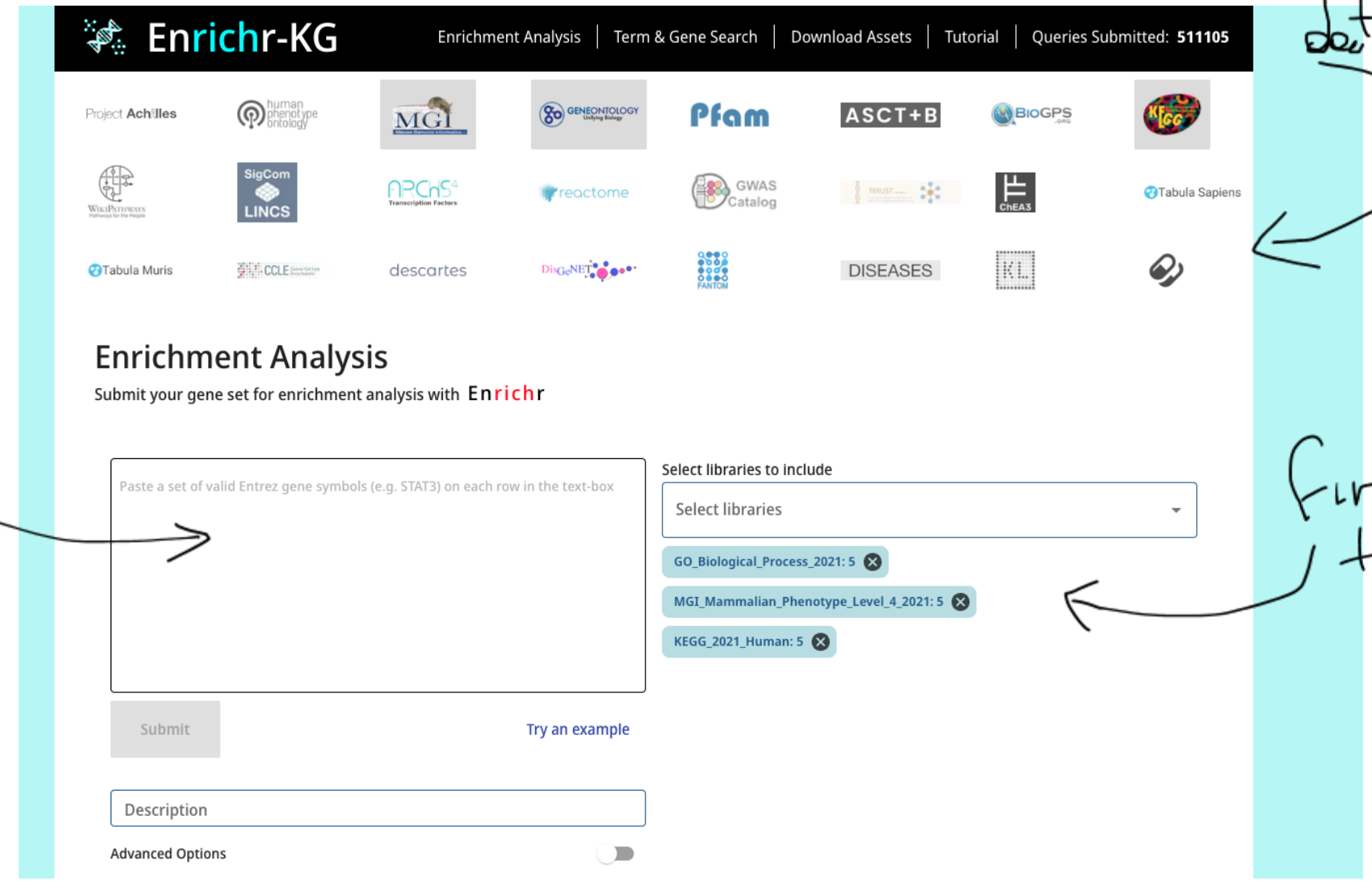
- https://www.ilincs.org/apps/grein/?gse=
- Insert in “Search for GEO series (GSE) accession”: **GSE62642**
- **Press the Blue tab that writes GSE62642**
- Go through Description, Metadata, Counts table, QC report, Visualisation (**Draw heatmap**) **– what does it tell you?**
- **Go to ”Analyze dataset”, select “Disease” in the Factor of Interest and set up:**
  - **Experimental group = *Parkinson’s Disease***
  - **Control group = *Healthy***
- **Generate Signature!**
- **Signature visualization!**
- **Explore visualisation – Static Heatmap the most informative?**
- **What do the column names mean in the Signature table? Discuss…**
- **Open the .csv file created in the Colab, and ”click and drag” *Gene_symbol* column**

Hands-on practical #1 - STRINGdb

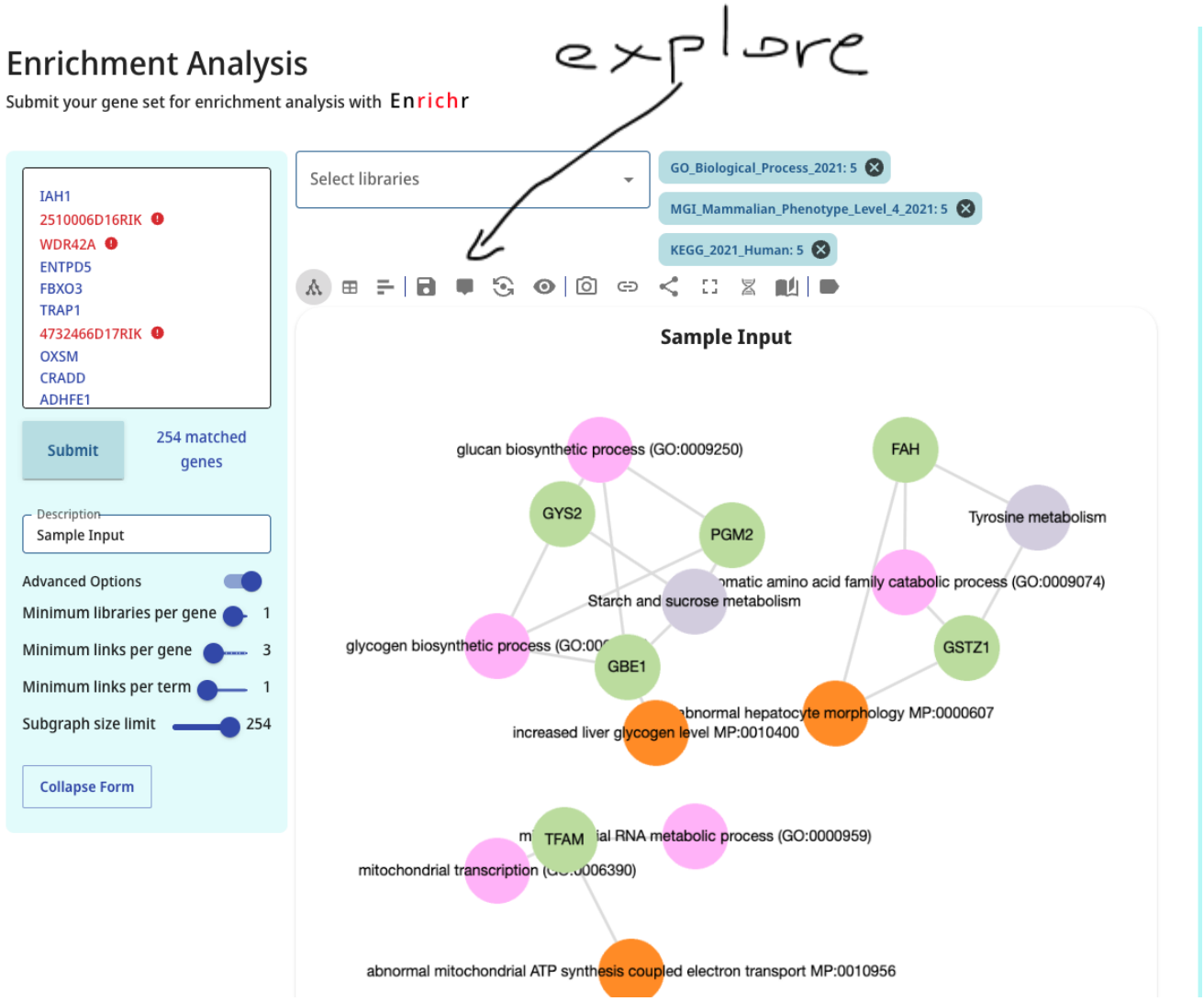




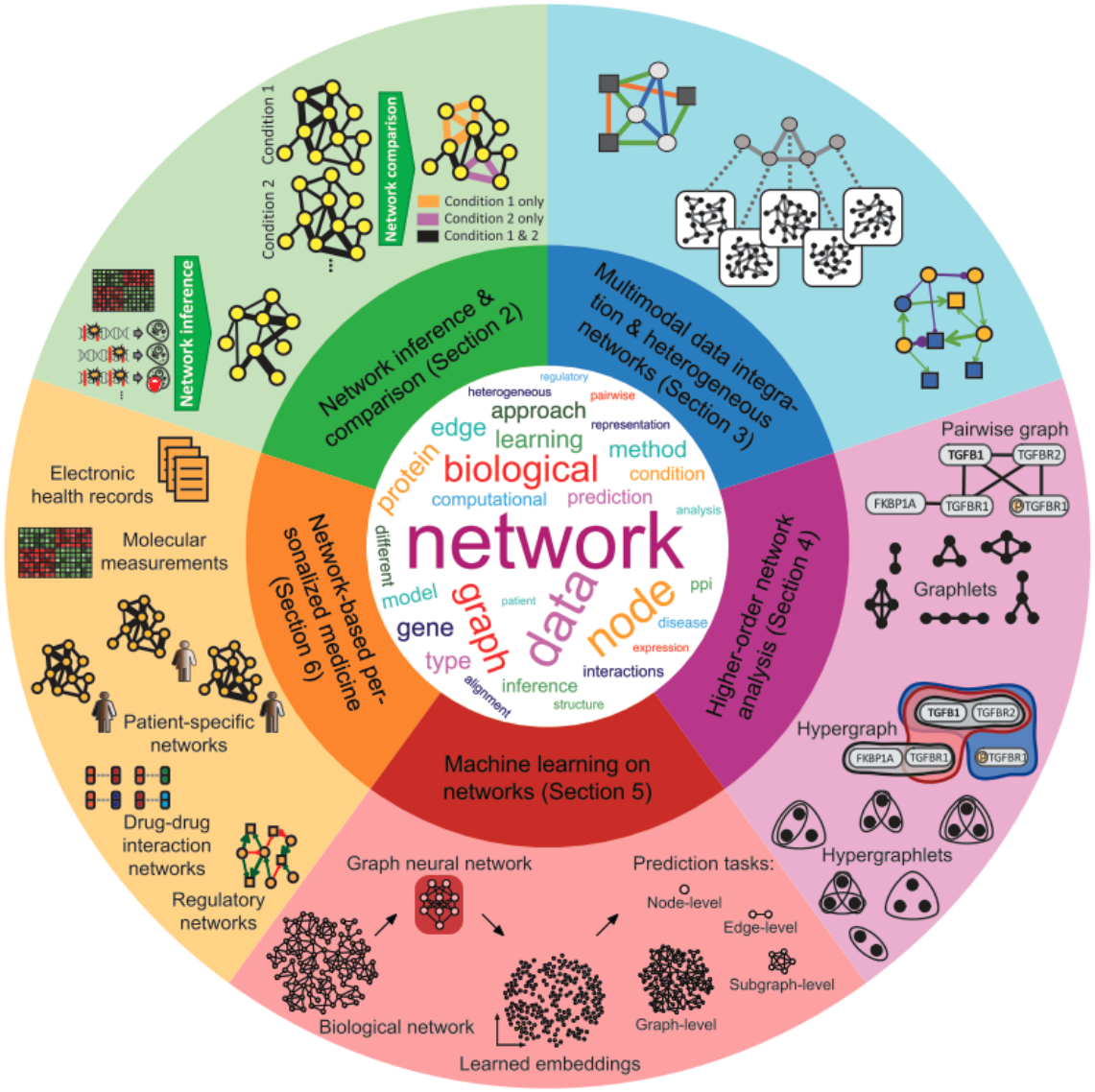
Hands-on practical #1 – Enrichr KG



Hands-on practical #1



Let's revisit the beginning…



Protein-protein interaction (PPI) network (physical interactions among proteins)

Association PPI network (physical + literature mining + genetic interactions + 3D structural similarities + …)

Correlation network (omic measurements from multiple samples; undirected; shows correlated entities like genes *that may* participate in a common pathway)

Regulatory network (directed regulatory connections of transcription factors or signal drivers with their targets)

Biomedical knowledge graphs (semantic relationships between diverse biomedical entities like genes, diseases, and patients, as well as measurements associated with them)

Cluster or community (a set of topologically related nodes, typically nodes that are densely connected and loosely connected to nodes in other clusters)

MORPHOPHEN

Current and future directions in network biology, Zitnik M et al. 2023

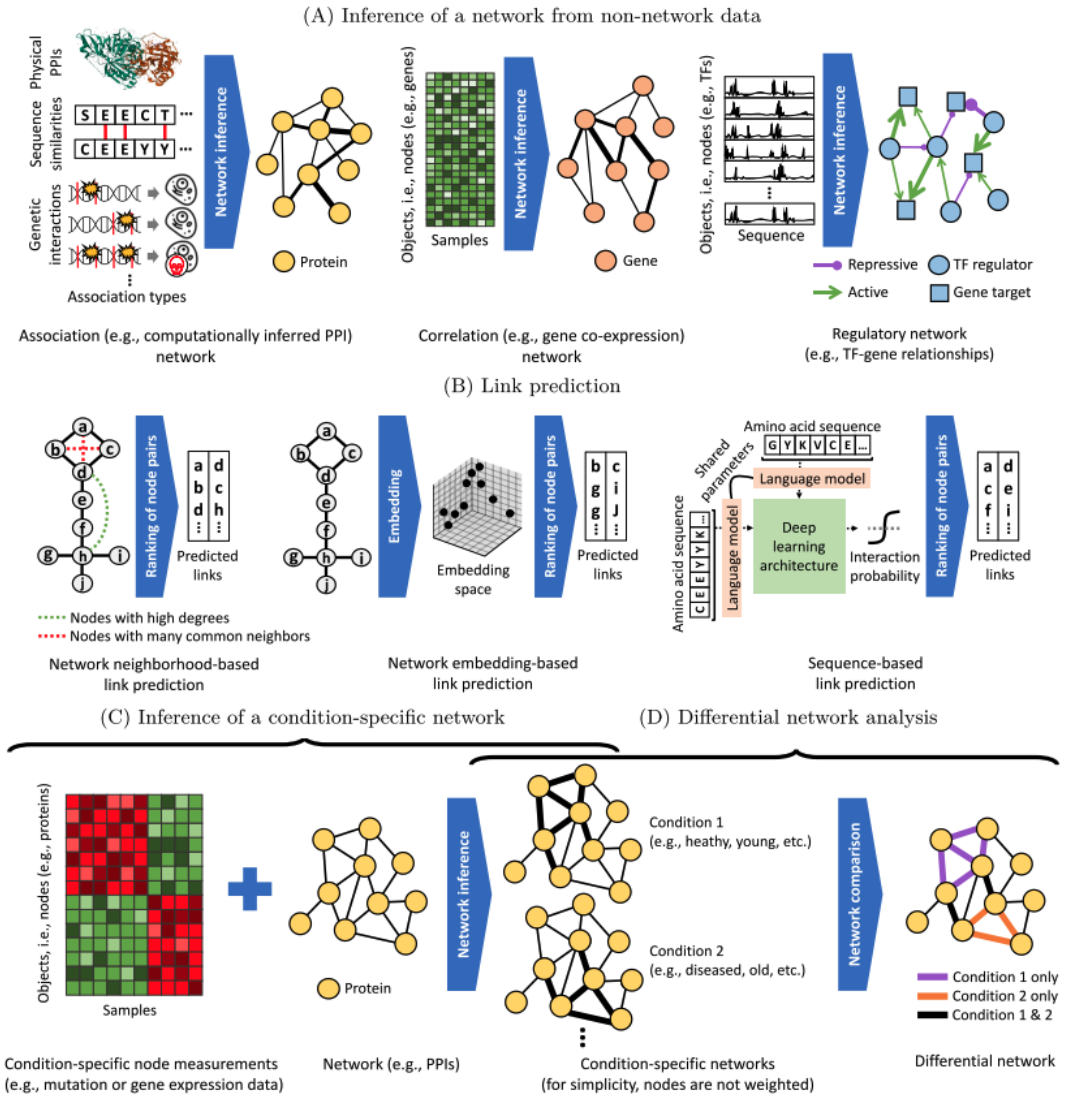
Inference and comparison of biological networks

Network inference from non-network data (A)
-- PPI + Associative networks
-- Correlation (ARACNe, CLR, WGCNA…)
-- Regulation (Inferelator, GENIE3…)
--- Genie3: ML using TF-target mapping

Link Prediction (B)
link nodes that have high degrees, that have many common interacting partners – or neighbors that share many paths, or that are topologically similar

Condition-specific network (C)
mutations, gene expression, in combination with PPI or regulatory or correlation networks – multiple samples, multiple combinations

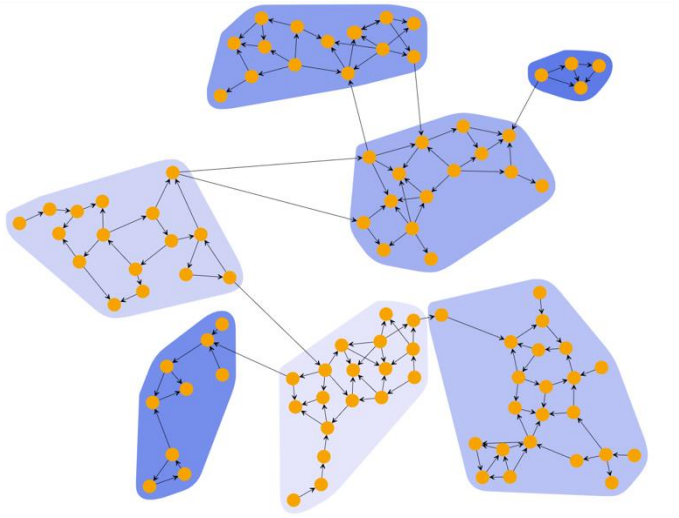
Differential network analysis (D)
differential network analysis, topological differences (a central node in a healthy network can less important in a disease network), difference in edges…



Current and future directions in network biology, Zitnik M et al. 2023

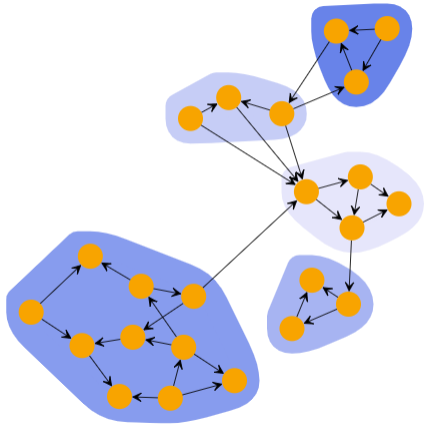
Why clustering a biological network?

- To extract biologically meaningful structures from data.
- To understand the relationship among these structures.
- To reduce data to an analyzable size.
- To visualize data of high-dimension and large-size



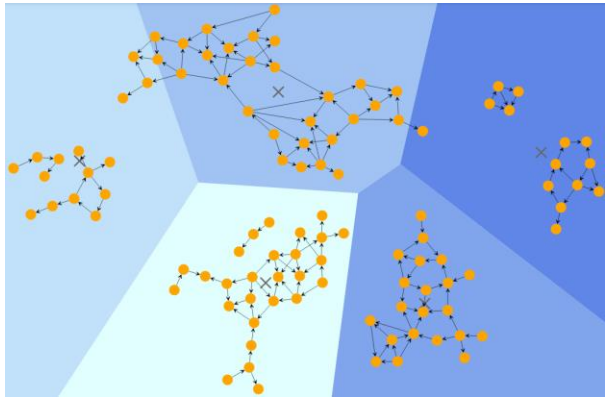
Edge Betweenness Clustering

Measures how often a node/edge lies on the shortest path between each pair of nodes in the diagram



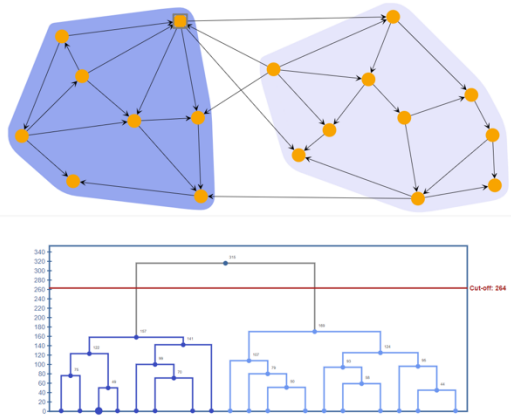
Biconnected Components Clustering

A biconnected component is a connected component with the property that the removal of any node keeps the component connected.



k-means Clustering

Partitions the graph into k clusters based on the location of the nodes such that their distance from the cluster's mean (centroid) is minimum



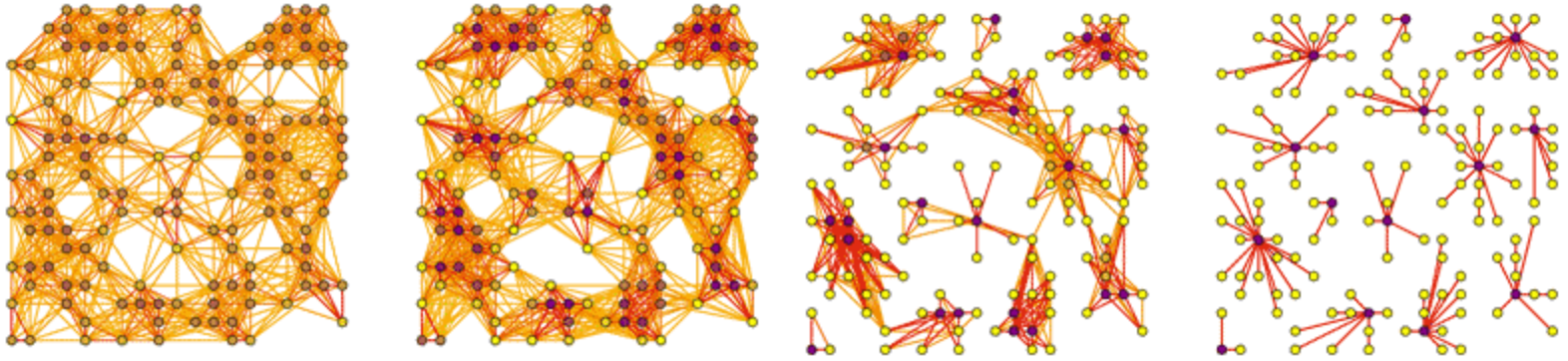
Hierarchical Clustering

Partitions the graph into a hierarchy of clusters. There exist two different strategies for hierarchical clustering, namely the agglomerative and the divisive.

https://www.yworks.com/pages/clustering-graphs-and-networks

MORPHOPHEN

The famous Markov Clustering algorithm



- **MCL** helps detecting communities, amongst the nodes present and also since it is un-supervised, it won't require any input form clusters unlike K-means algorithms
- Useful on weighted graphs
- A markov chain is a system in which the next state is dependent upon the current state based on some probability or rule.
- *The travelling salesmen problem*: being at one city, the probability of Sam visiting a nearby city is more than visiting a city very far!
- **Inflation** (*strong neighbor values are strengthened and large neighbor values are demoted*) & **Expansion (**making the farther nodes or neighbors reachable**)** adjustment ----------→ Clusters!

MORPHOPHEN

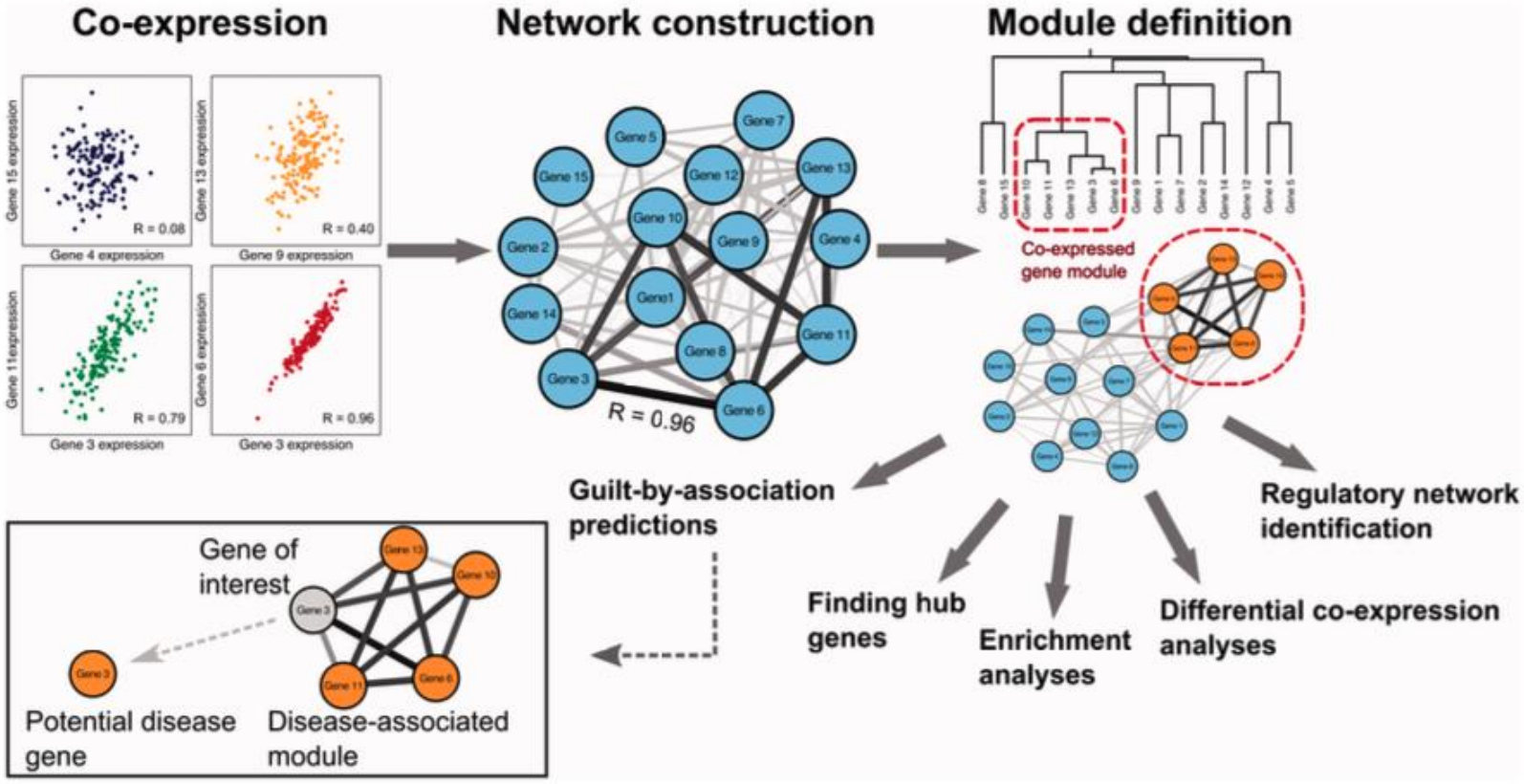
https://github.com/micans/mcl

Co-expression reveals related co-perturbed genes

Methods like WGCNA use Pearson correlation (assuming linear relationships)…
Un-directional networks, spurious results on regulation! (many false positives)

Can we use global human gene expression data (i.e. transcriptomic genome- wide data) to derive **gene coexpression networks** ? (Closer to ground-truth)

• Is it a **reliable** way to find coexpression (knowing the noise and background in genome-wide expression and the bad effect of outliers on correlation) (Linearity vs non-linearity) ?



Van Dam et al. (2018) *Briefings in Bioinformatics*

Inferring Gene Regulatory Networks (GRNs) beyond correlation

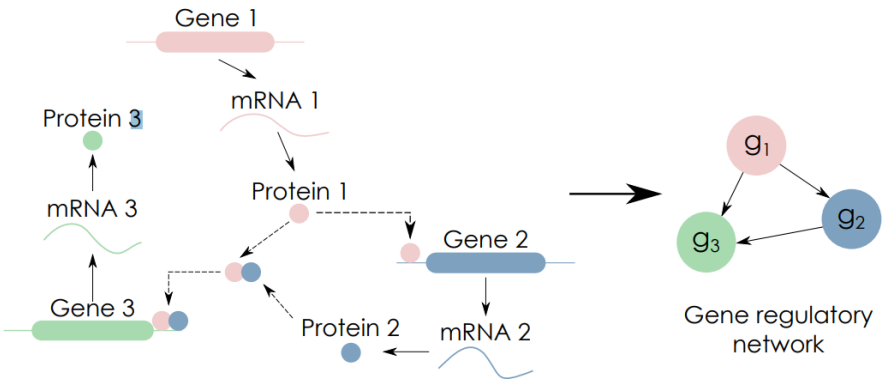
- ***Information theoretic scores*** (Mutual Information/MI to see how dependent two variables are, MI=0 means independence, 1 more dependence, empirical distributions (estimated from the samples) of gene expression levels for each pair of genes, **ARACNe**, CLR, MRNET, good scaling, noise-sensitive when samples are few..)

- ***Regression-based methods*** (predict one variable from the other, GENIE3 which uses non-parametric regression is well-studied algorithm, captures higher-order relationships, computationally intensive!, good for time-series data…)

- ***Probabilistic methods*** (probabilistic model of the data, using global measures of fit - joint likelihood- or Bayesian approaches with the latter having the advantage of which prior information can be encoded, and in the way the intrinsic uncertainty in the system is represented

- ***Dynamic methods for time-series data*** (dynamic Bayesian or dif.equations, the former are the most popular)

- ***Machine Learning/Deep Learning*** !



MORPHOPHEN

Vân Anh Huynh-Thu, Guido Sanguinetti, arXiv:1801.04087

Recap no2!

- Biological networks are highly modular depicting different types of data (mono-partite, multi-partite, from PPIs to Knowledge Graphs)
- Networks combine well with pathway enrichments to reveal signaling profiles changing upon perturbations
- Gene regulatory networks are highly informative into revealing the regulatory capacity of transcription factors towards their target-genes but they can become highly complex
- Extracting information from complex networks is usually mediated by clustering

Hands-on practical #2

- **Run basic Network Analysis of the STRINGdb network in Colab**
- **Run Clustering**
- **Run Pathway Enrichment**

MORPHOPHEN

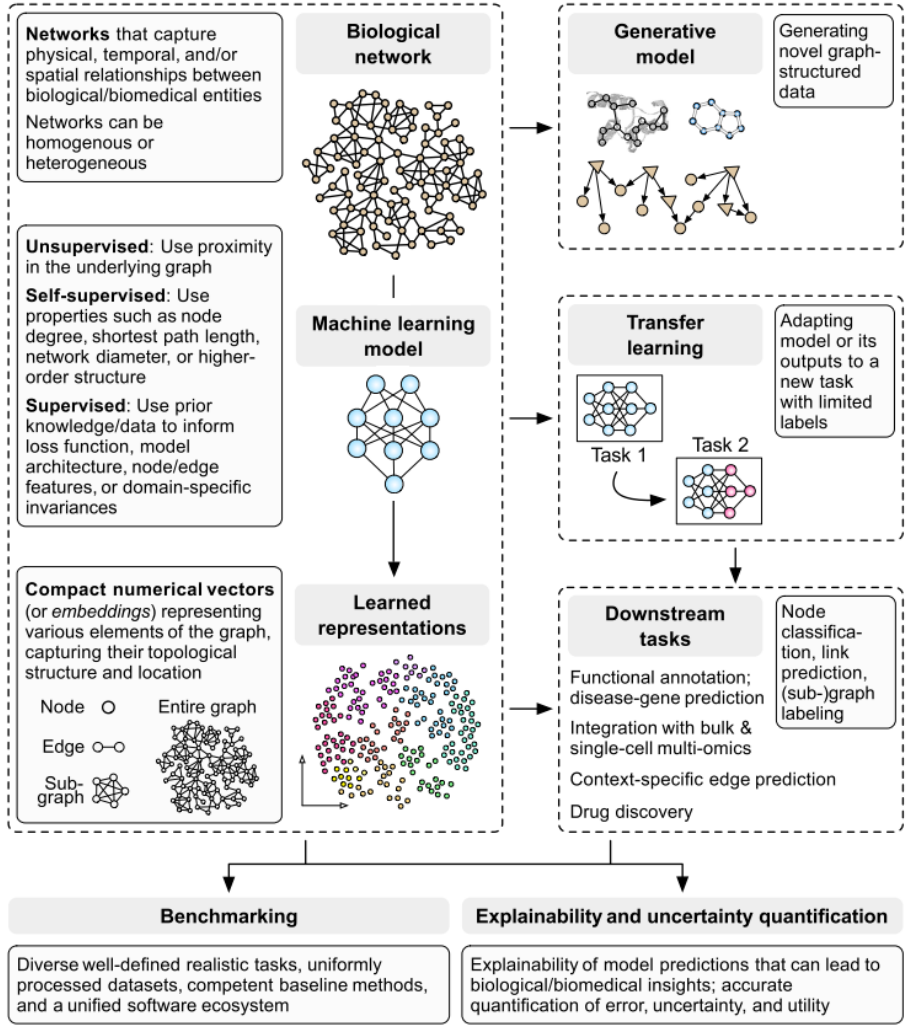
Machine learning on networks

- A machine learning model, typically a neural network, that takes one or more biological networks as input and learns ***representations (embeddings)*** of various graph elements in an *unsupervised*, *self-supervised*, or *supervised manner*.

- These representations can be used for ***exploratory analysis*** or as ***input to train a new machine learning*** model to perform a downstream task.

- Models can also be trained for one task with abundant labels and transferred (modified and fine-tuned) to a new related task with limited labels (***Transfer Learning***)

- **Downstream tasks!**



Current and future directions in network biology, Zitnik M et al. 2023

Network-based precision medicine

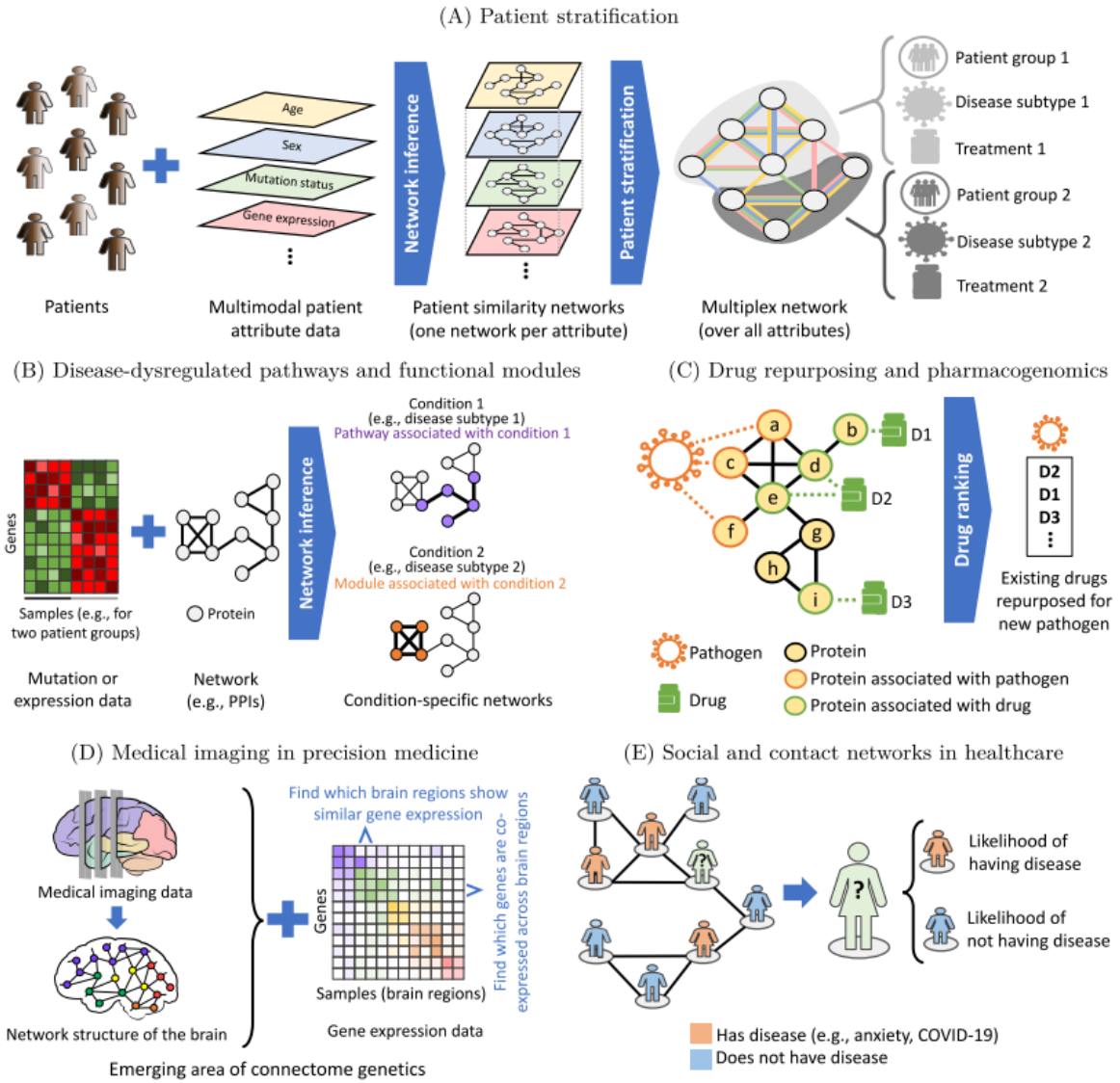
Patient stratification (A)
Groups of patients that correspond to their communities (clusters) in a patient similarity network –> distinct disease subtypes

Disease pathways (B)
inference of a condition-specific network (subgraphs…)

Drug repurposing (C)
evaluation of the fit of existing drugs to new diseases based on network "relatedness"

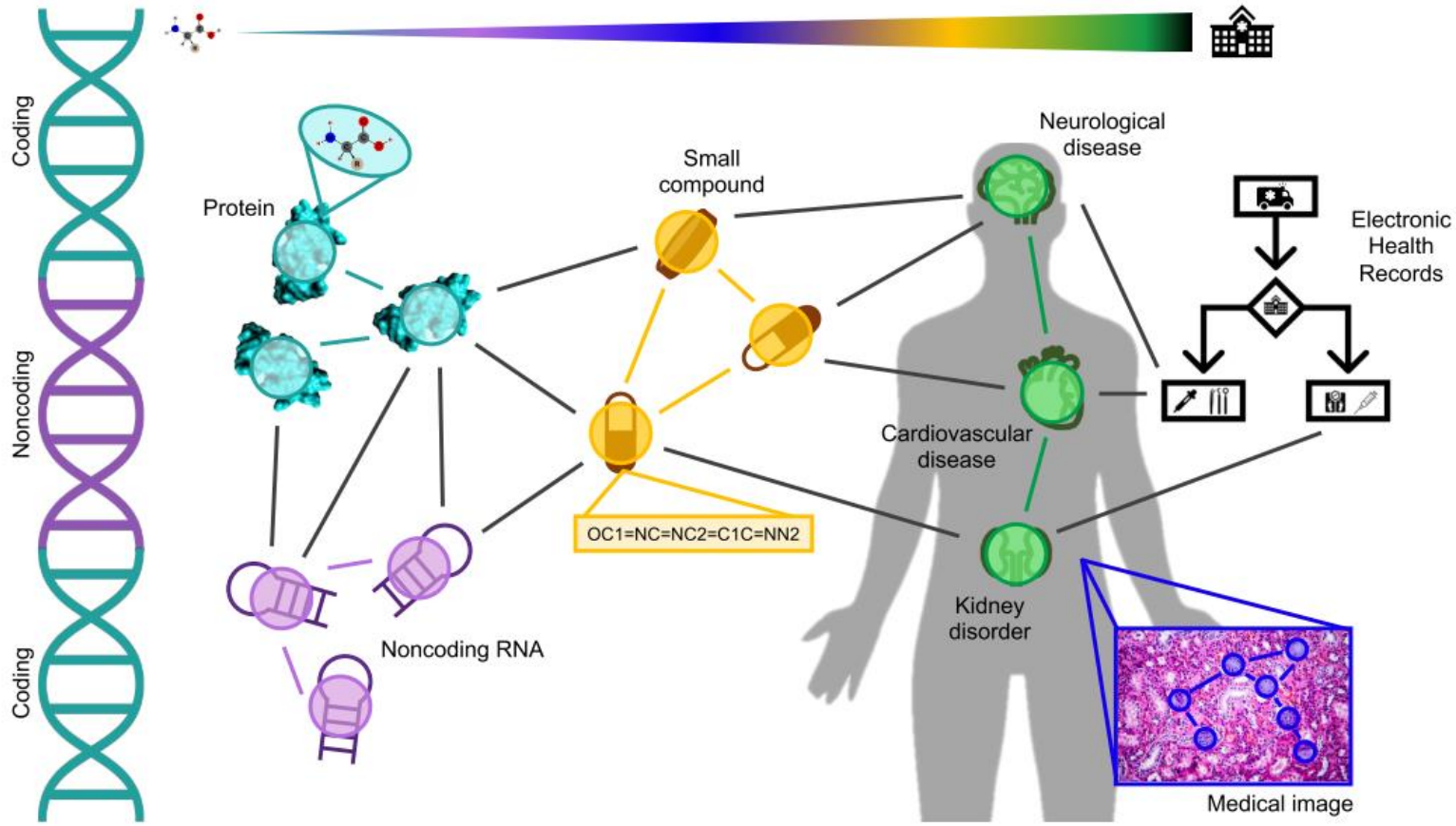
Medical imaging data (D)
connectome genetics, network structure of the brain meets -omics data

Societal and contact networks (E)
Social position, demographics, geography, mental health…



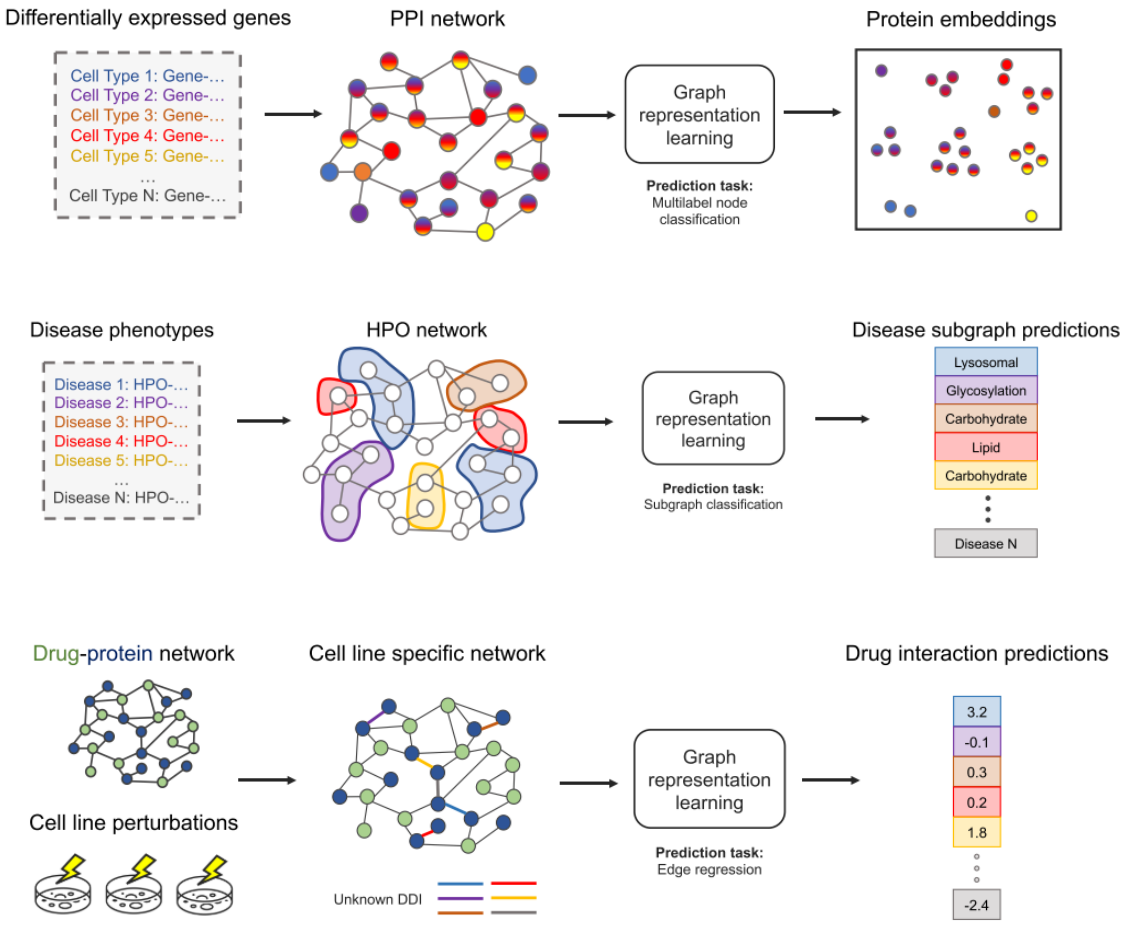
Current and future directions in network biology, Zitnik M et al. 2023

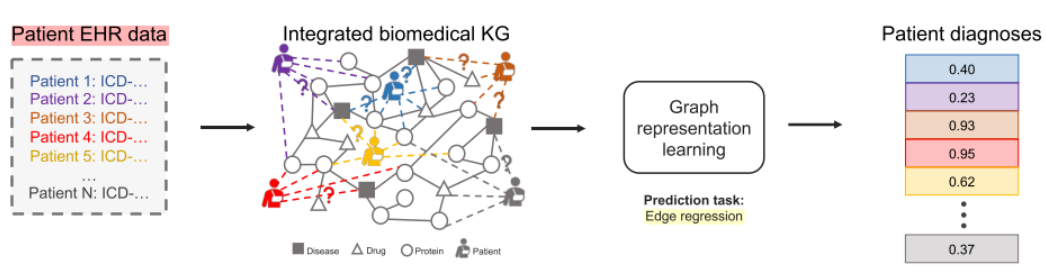
The overall Network Biology vision



MORPHOPHEN

Graph representation learning





Node classification/regression

Edge classification/regression

Module detection

Clustering or community detection

Latent graph learning ….

(can you guess what it is all about??)

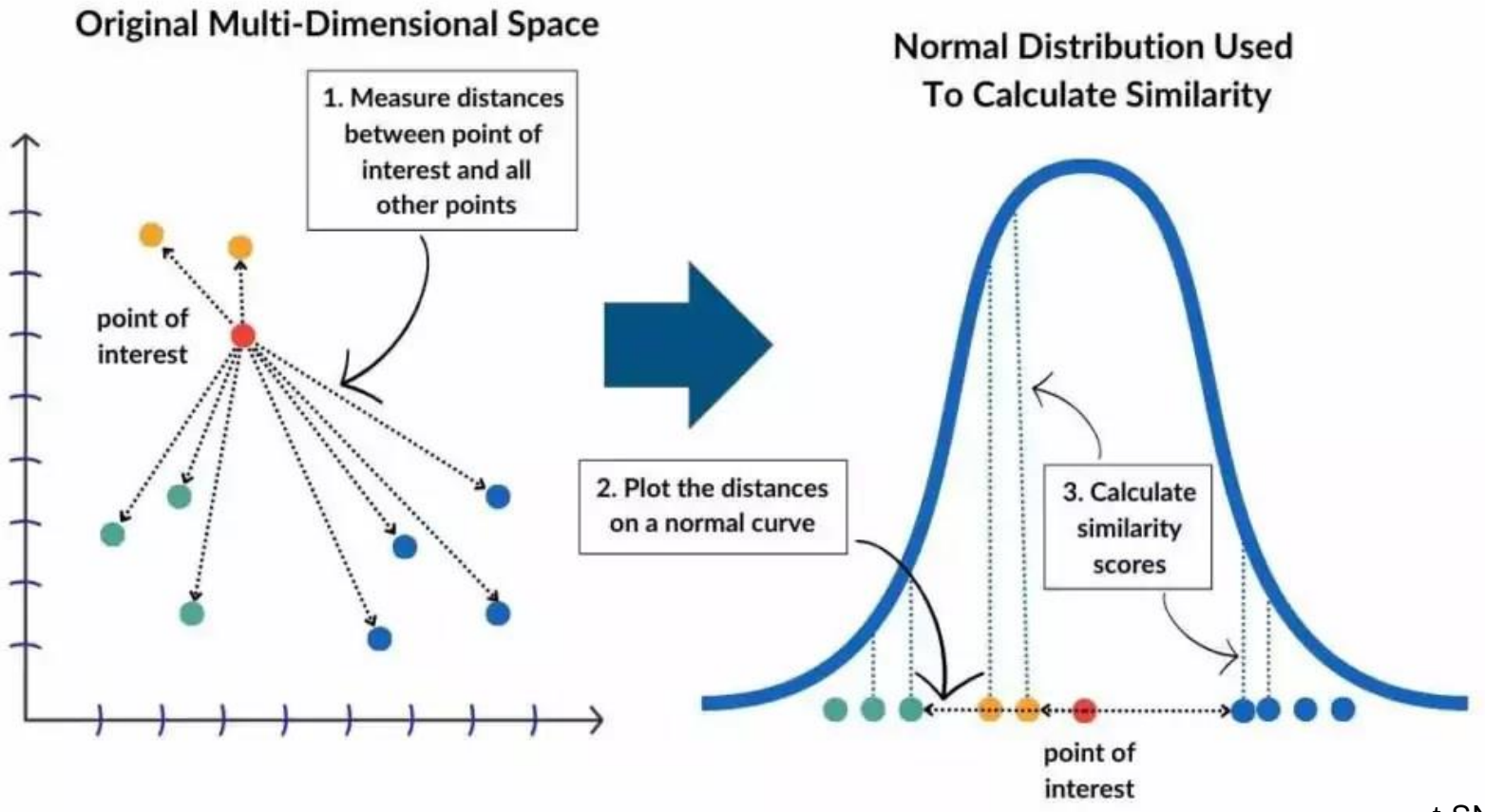
MORPHOPHEN

Hands-on practical #3

- Run Node2Vec in the network
- Wrap up

Hands-on practical #3 – t-SNE

t-distributed Stochastic Neighbor Embedding (t-SNE)



t-SNE is the “cartographer” using the scale to preserve an accurate representation of neighborhoods on miniaturized map

https://spotintelligence.com/2023/12/22/practical-guide-t-sne/

MORPHOPHEN

Recap no3!

- Machine Learning and Deep Learning is taking the network biology field by storm! (non-linearity, biological priors, complex relationships)
- Network-based precision medicine can be applied to omics, patients, imaging, diseases, drugs offering a unified paradigm for tailor-made therapies to each patient
- Representation learning can augment networks of biological entities leading to the assembly of predictive models which provide actionable insights

Before wrap up!– some key resources

- https://cytoscape.org/cytoscape-tutorials/contents/index.html#/
- https://snap.stanford.edu/deepnetbio-ismb/
- https://www.youtube.com/watch?v=lH75WJgLeoo
- https://graphery.reedcompbio.org/
- https://igraph.org/
- https://networkx.org/

Thank you for your attention!

Next episode: Single-cell omics…and spatiality???