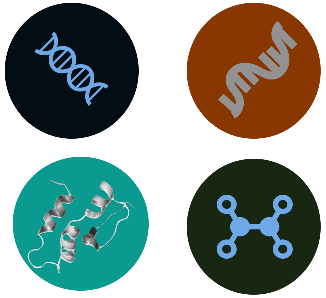
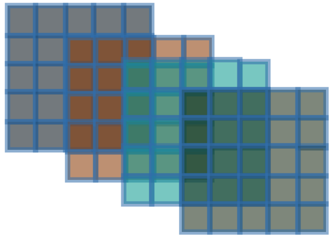
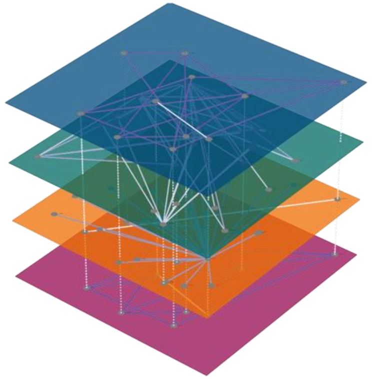
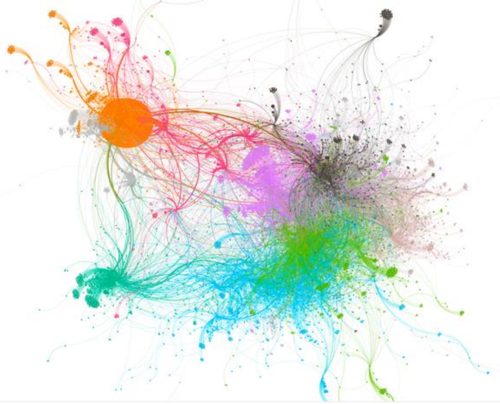








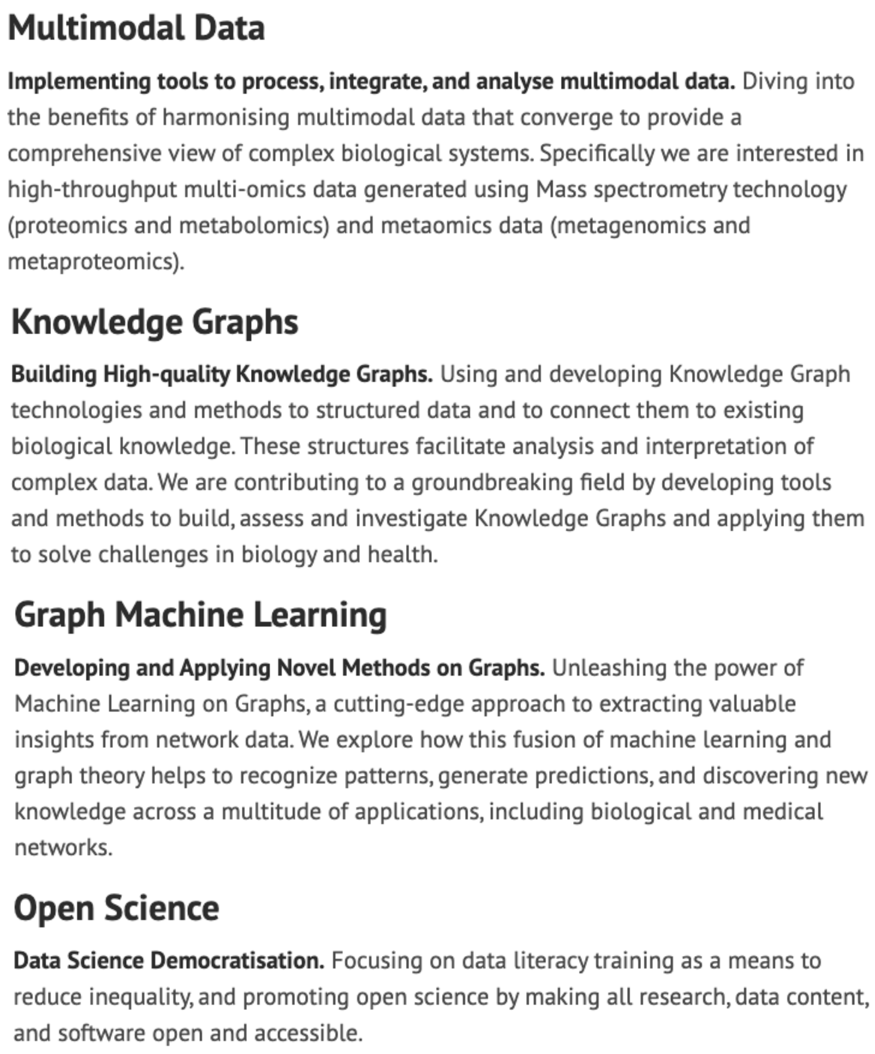
Basics of Downstream Proteomics Analysis

14th May 2025

Alberto Santos

The Novo Nordisk Foundation
Center for Biosustainability

Multi-omics Network Analytics (MoNA)



Multimodal Data

Implementing tools to process, integrate, and analyse multimodal data. Diving into the benefits of harmonising multimodal data that converge to provide a comprehensive view of complex biological systems. Specifically we are interested in high-throughput multi-omics data generated using Mass spectrometry technology (proteomics and metabolomics) and metaomics data (metagenomics and metaproteomics).

Knowledge Graphs

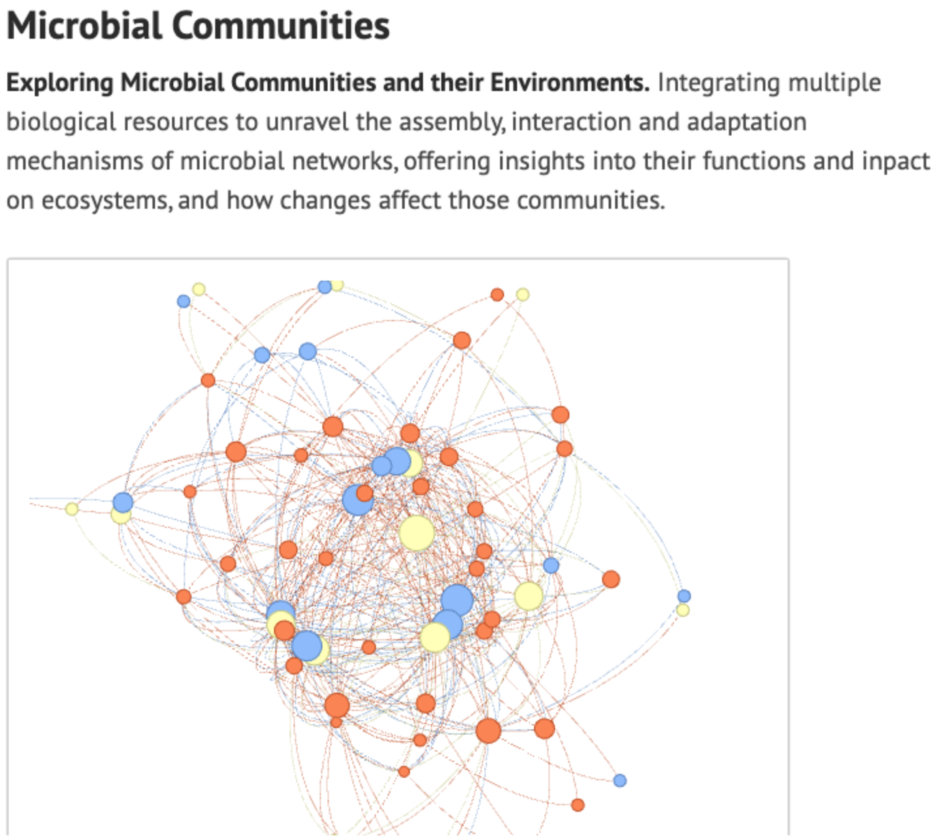
Building High-quality Knowledge Graphs. Using and developing Knowledge Graph technologies and methods to structured data and to connect them to existing biological knowledge. These structures facilitate analysis and interpretation of complex data. We are contributing to a groundbreaking field by developing tools and methods to build, assess and investigate Knowledge Graphs and applying them to solve challenges in biology and health.

Graph Machine Learning

Developing and Applying Novel Methods on Graphs. Unleashing the power of Machine Learning on Graphs, a cutting-edge approach to extracting valuable insights from network data. We explore how this fusion of machine learning and graph theory helps to recognize patterns, generate predictions, and discovering new knowledge across a multitude of applications, including biological and medical networks.

Open Science

Data Science Democratisation. Focusing on data literacy training as a means to reduce inequality, and promoting open science by making all research, data content, and software open and accessible.



Microbial Communities

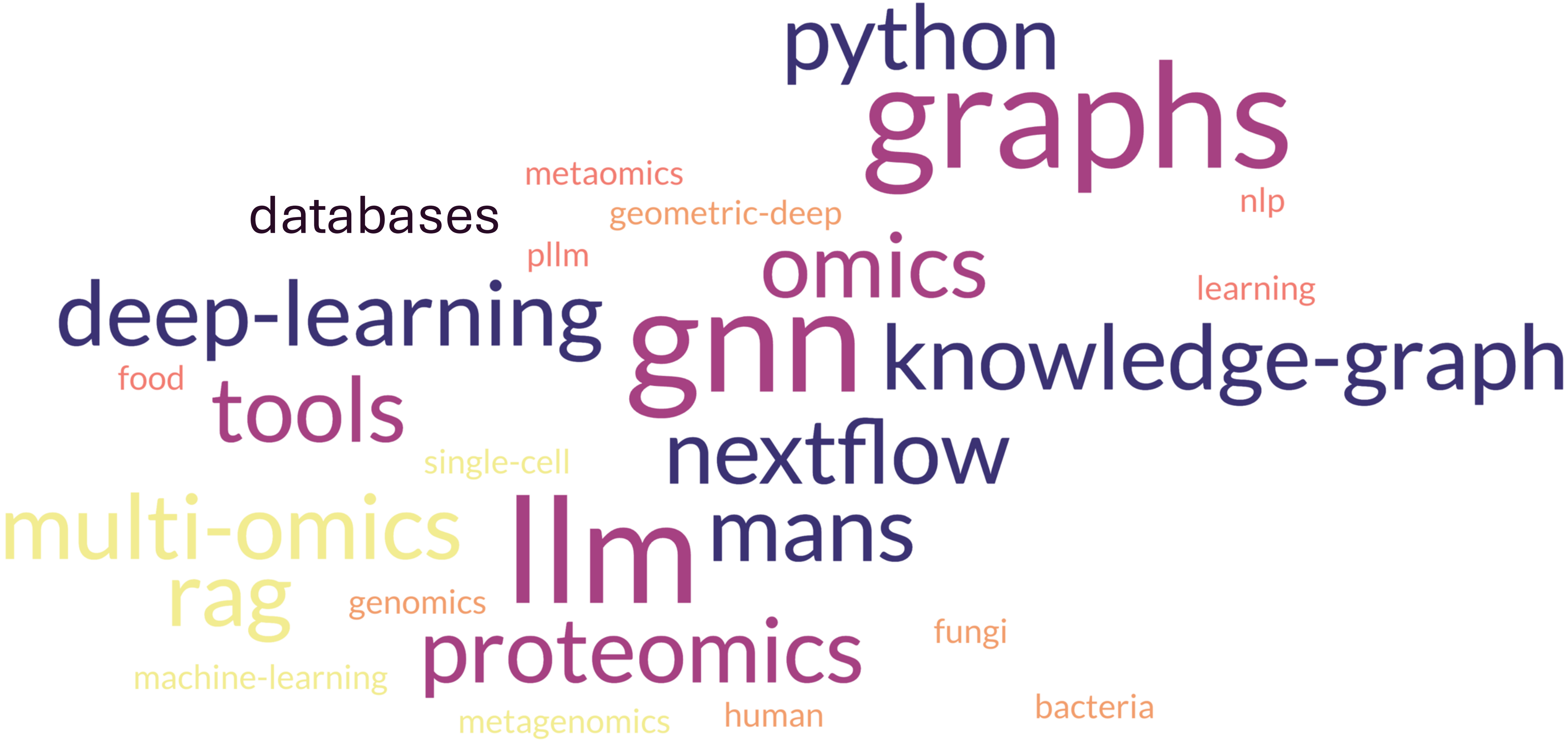
Exploring Microbial Communities and their Environments. Integrating multiple biological resources to unravel the assembly, interaction and adaptation mechanisms of microbial networks, offering insights into their functions and inpact on ecosystems, and how changes affect those communities.

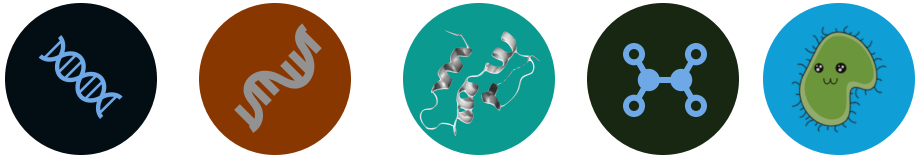
Clinical Computational Omics

Providing tools for the analysis and interpretation of clinical omics data. Integration of high-throughput omics data with computational and bioinformatics approaches to advance precision medicine and disease research. These projects aim to identify biomarkers, uncover disease mechanisms, and tailor treatments based on individual molecular profiles.

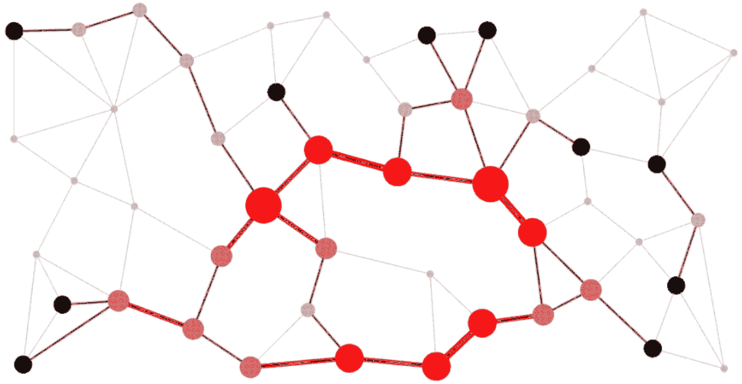
https://multiomics-analytics-group.github.io/

Research


python
graphs
metaomics
databases
geometric-deep
nlp
pllm
omics
learning
deep-learning
gnn
knowledge-graph
food
tools
nextflow
single-cell
multi-omics
llm
mans
rag
genomics
fungi
proteomics
machine-learning
bacteria
metagenomics
human



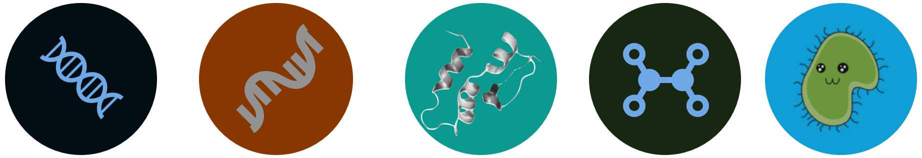
Graphs



What is Omics?

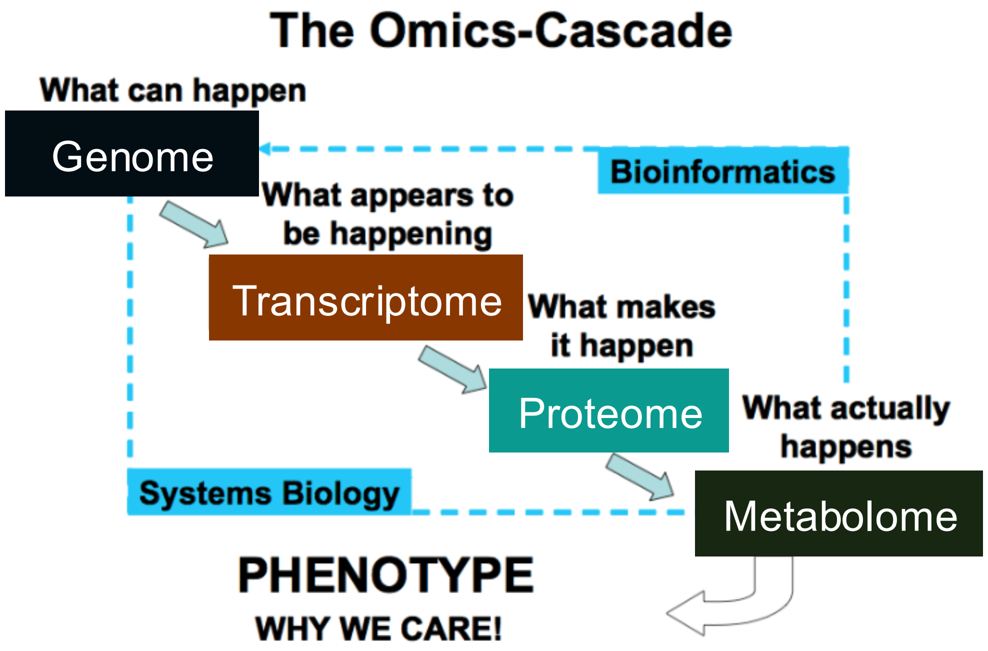
Understanding biology on a large scale

- Fields of study that aim to **map**, **quantify**, and **understand** sets of biological molecules within an organism or system— genes, proteins, metabolites, and more
- Provide:
  - **Holistic View** beyond single-gene or single-protein studies, providing a comprehensive view of biological processes
  - **High-Resolution Data** generated with high-throughput technologies



Types of Omics

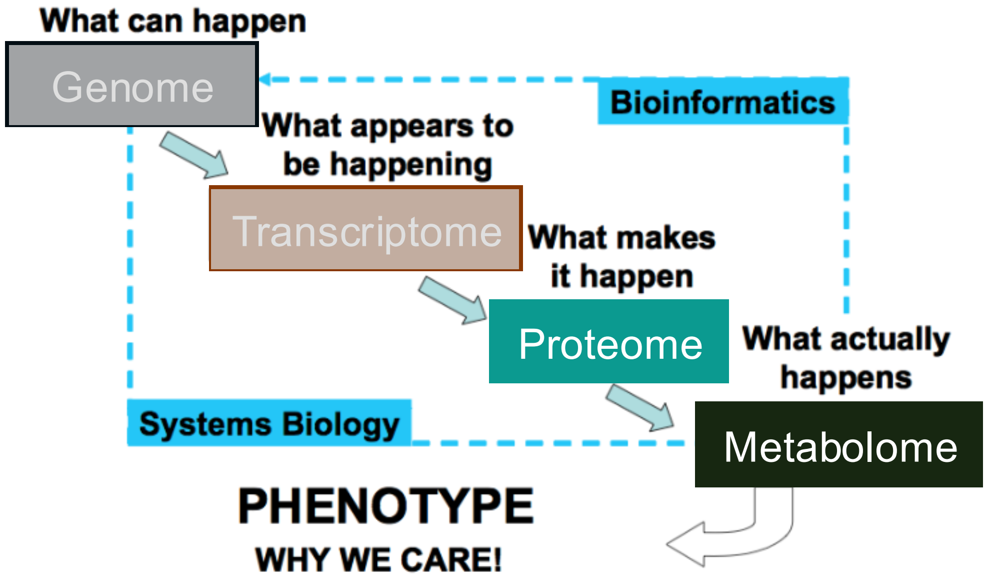
- Genomics Study of the genome, which includes all DNA within an organism
  - Sequence, structure, and function of genes
  - Key technology — Next-generation sequencing (NGS)
- Transcriptomics Study of the transcriptome, which is the complete set of RNA transcripts
  - Gene expression and regulation
  - Key technology — RNA sequencing (RNA-seq)
- Proteomics Study of the proteome, or the complete set of proteins in a cell or organism
  - Protein structure, function, interactions, and modifications
  - Key technology — Mass spectrometry (MS)
- Metabolomics Study of the metabolome, which includes all small-molecule metabolites in a cell or biological system
  - Cellular processes and metabolic pathways
  - Key technology — Mass spectrometry (MS)
- Metaomics Studies the collective genetic material, proteins, metabolites, and other molecular components from entire communities of organisms in a specific environment, without needing to isolate or culture individual species.



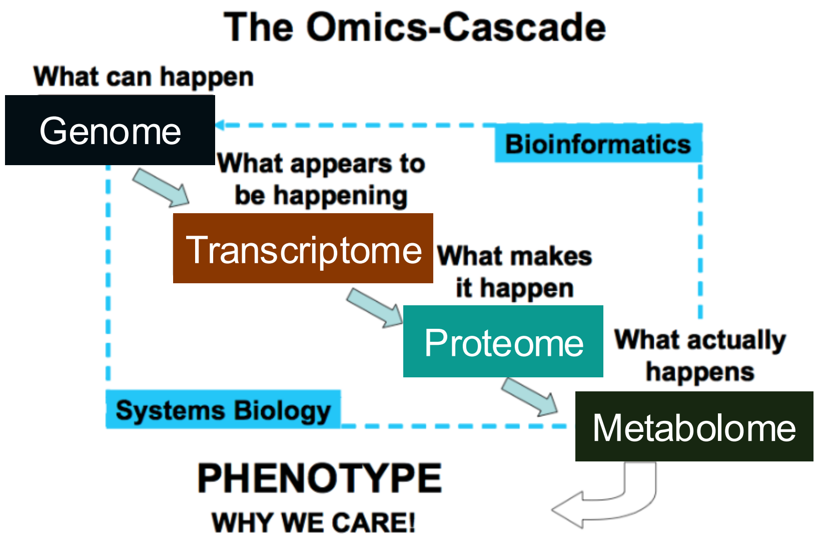
Types of Omics

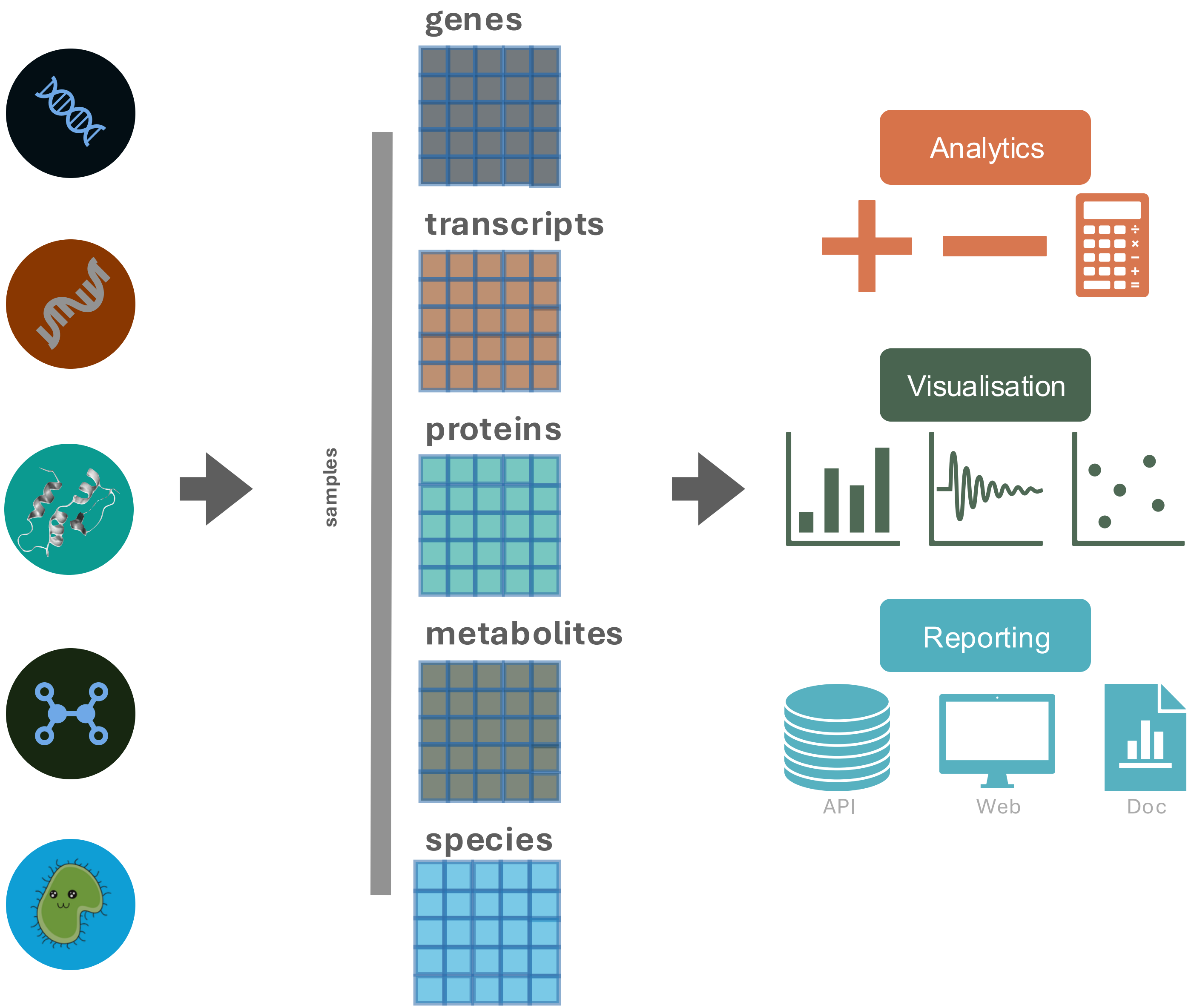
- Genomics Study of the genome, which includes all DNA within an organism
  - Sequence, structure, and function of genes
  - Key technology — Next-generation sequencing (NGS)
- Transcriptomics Study of the transcriptome, which is the complete set of RNA transcripts
  - Gene expression and regulation
  - Key technology — RNA sequencing (RNA-seq)
- Proteomics Study of the proteome, or the complete set of proteins in a cell or organism
  - Protein structure, function, interactions, and modifications
  - Key technology — Mass spectrometry (MS)
- Metabolomics Study of the metabolome, which includes all small-molecule metabolites in a cell or biological system
  - Cellular processes and metabolic pathways
  - Key technology — Mass spectrometry (MS)
  - Metaomics Studies the collective genetic material, proteins, metabolites, and other molecular components from entire communities of organisms in a specific environment, without needing to isolate or culture individual species.

The Omics-Cascade

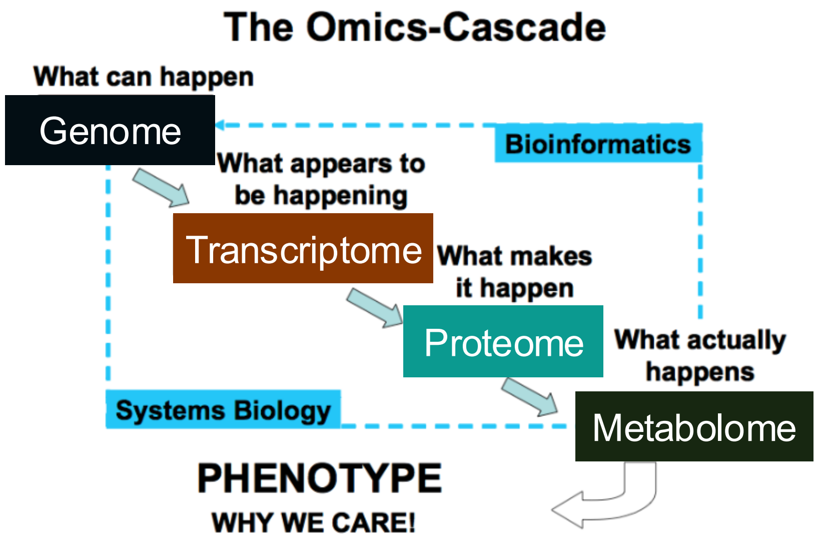


Downstream Analysis





Downstream Omics Analysis



The Omics-Cascade

What can happen

Genome

What appears to be happening

Transcriptome

What makes it happen

Proteome

What actually happens

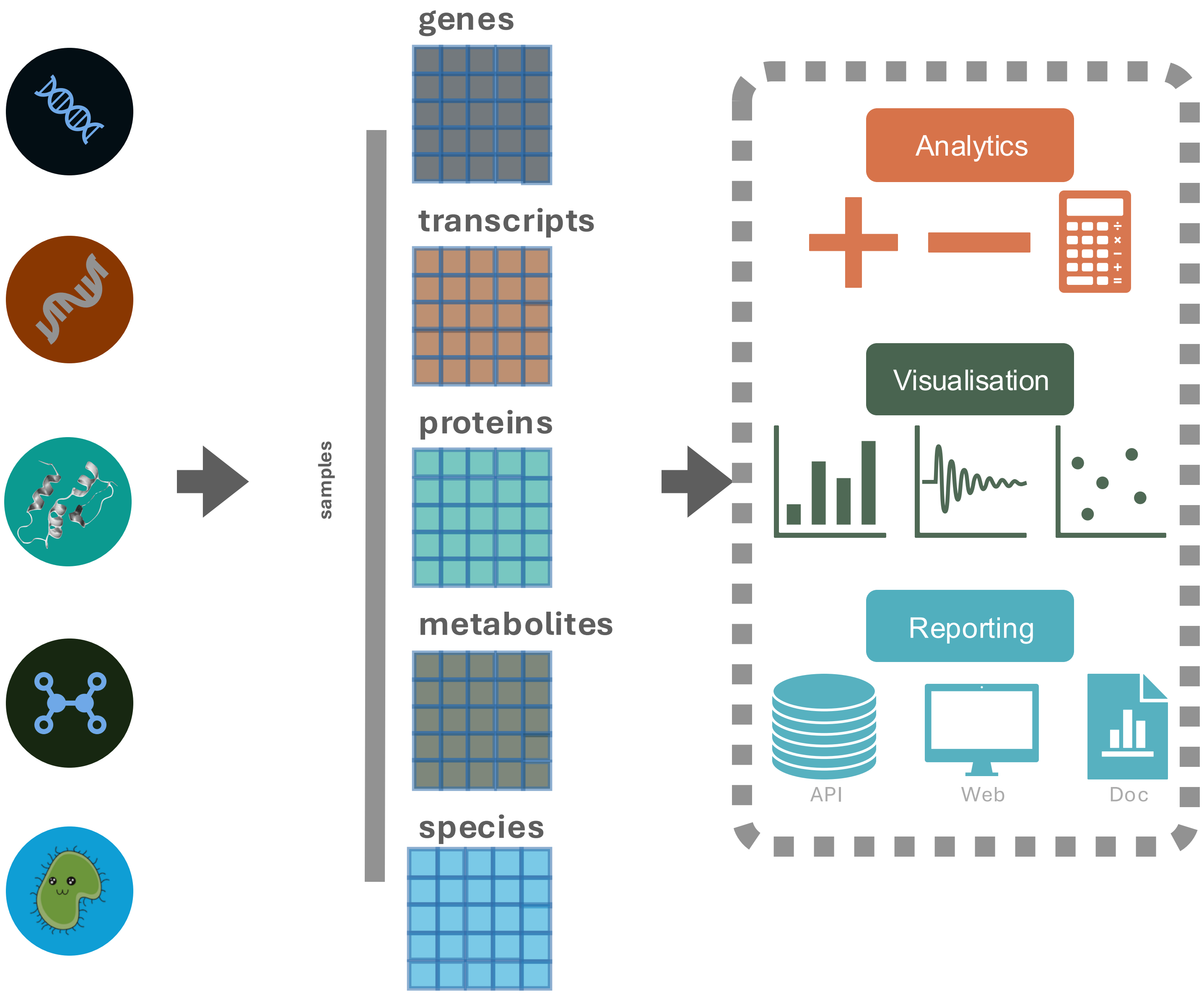
Metabolome

Bioinformatics

Systems Biology

PHENOTYPE

WHY WE CARE!



genes

transcripts

proteins

metabolites

species

samples

Analytics

Visualisation

Reporting

API

Web

Doc

Downstream Proteomics Analysis

Challenges

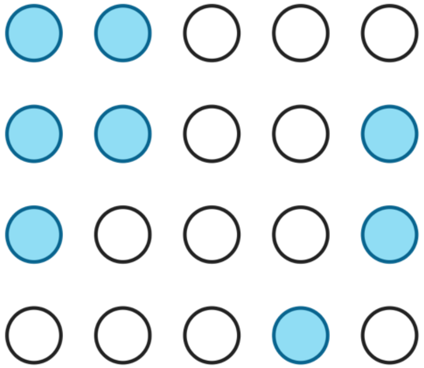
Main goals

- Identify **significant changes**
- Infer **biological meaning**
- **Integrate** with other omics data

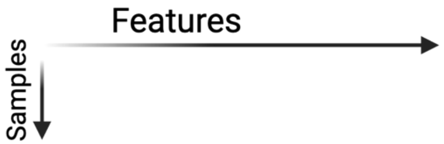
Challenges

- High dimensionality, small sample sizes
- Missing values and batch effects
- Interpretation bias in functional analysis
- Reproducibility

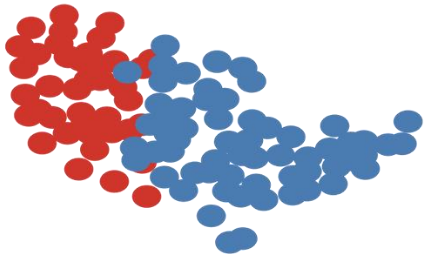
Missing Values



High-dimensionality



Missing Values



Proteomics Data

SDRF file

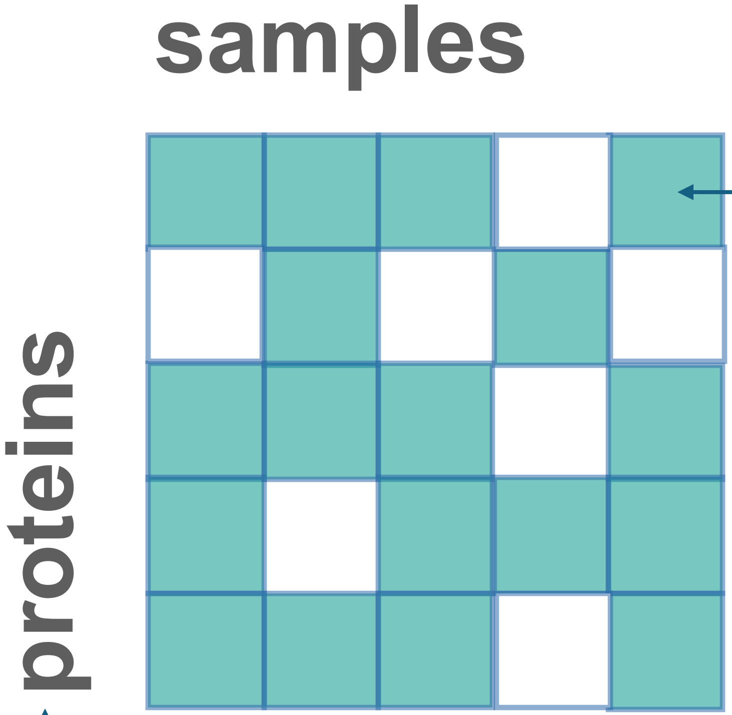
Stardised Experimental Metatada

nf-core

quantms: Introduction

Quantitative mass spectrometry workflow.

Pipeline



LFQ intensities

Normalised intensities – normalisation is crucial for ensuring reliable comparison of protein levels across biological samples

Protein groups

ProteinA;ProteinB;ProteinC
Peptides can match multiple proteins; protein groups handles redundancy in the matching of peptides to protein hits.
Razor protein -> first reported protein

Proteomics Data

SDRF file

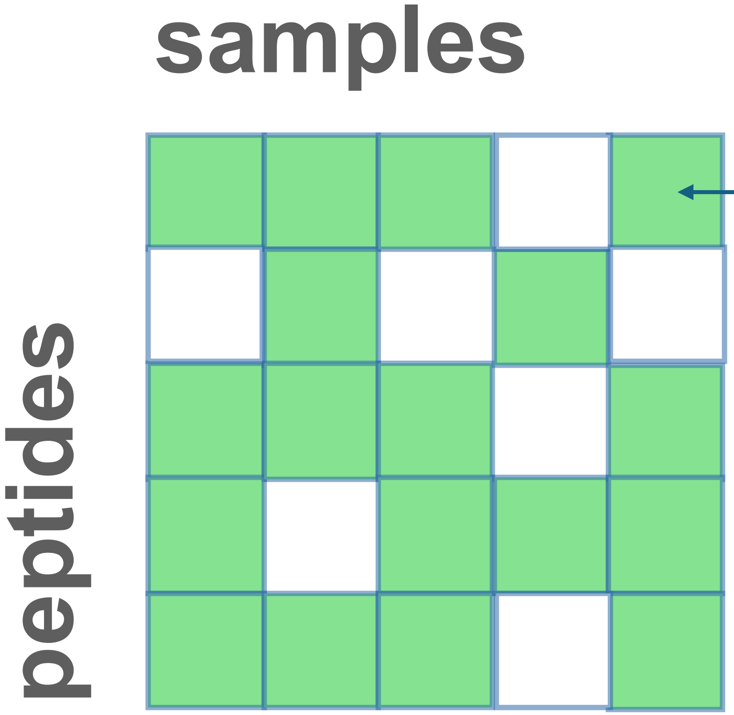
Stardised Experimental Metatada

nf-core

quantms: Introduction

Quantitative mass spectrometry workflow.

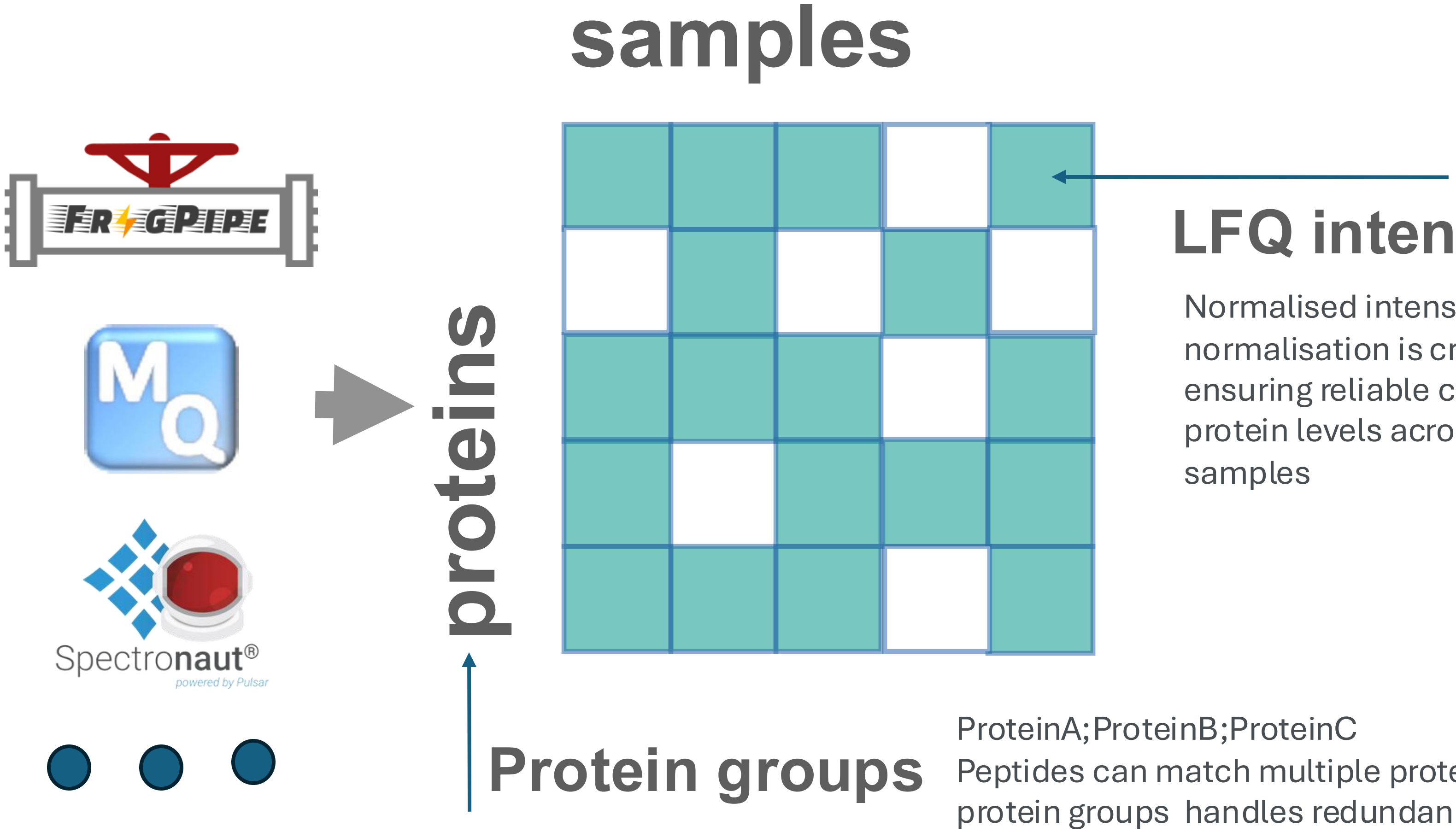
Pipeline



LFQ intensities

Normalised intensities – normalisation is crucial for ensuring reliable comparison of protein levels across biological samples

Proteomics Data



samples

proteins

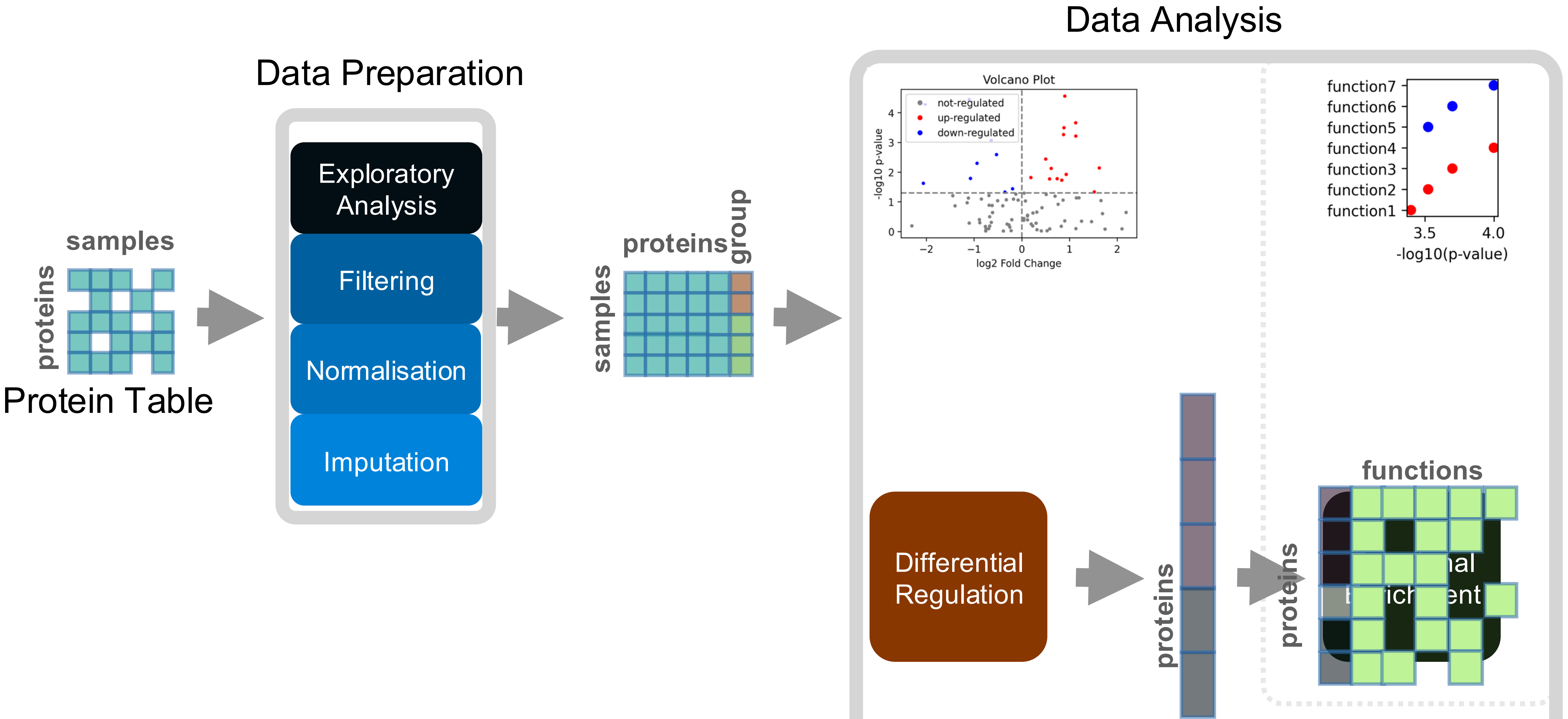
LFQ intensities

Normalised intensities – normalisation is crucial for ensuring reliable comparison of protein levels across biological samples

Protein groups

ProteinA;ProteinB;ProteinC
Peptides can match multiple proteins; protein groups handles redundancy in the matching of peptides to protein hits.
Razor protein -> first reported protein

Analytical Workflow



Data Preparation

Exploratory Analysis

Filtering

Normalisation

Imputation

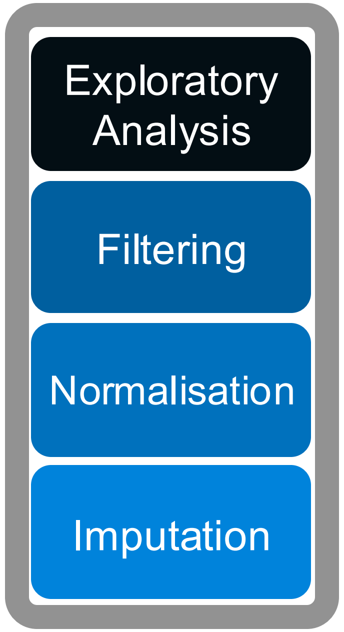
Protein Table

Data Analysis

Differential Regulation

Functional Enrichment

Data Preparation



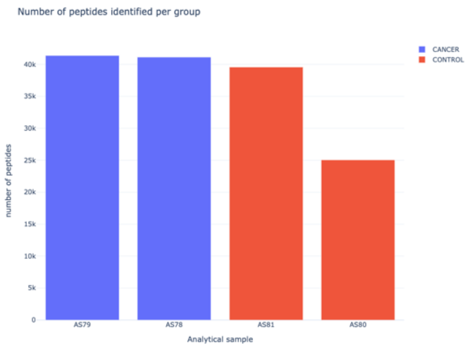
Exploratory Analysis

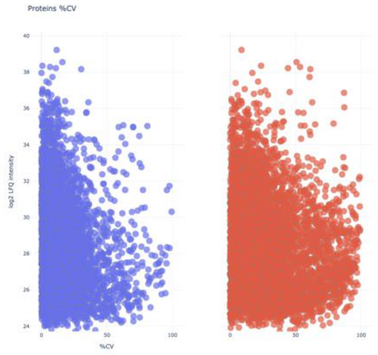
Filtering

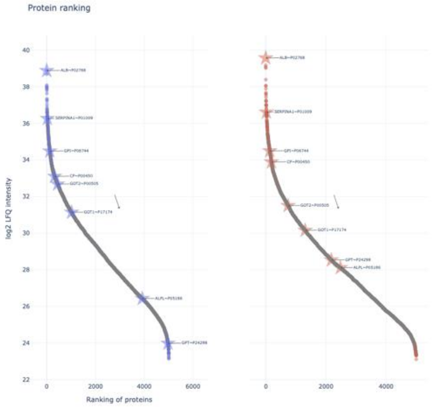
Normalisation

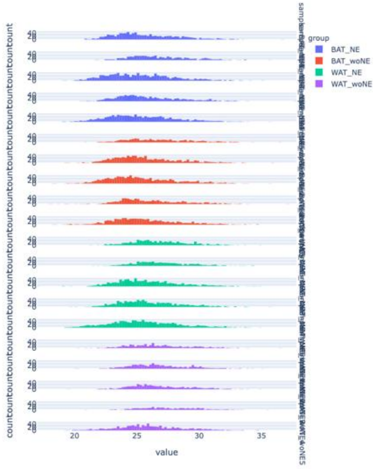
Imputation

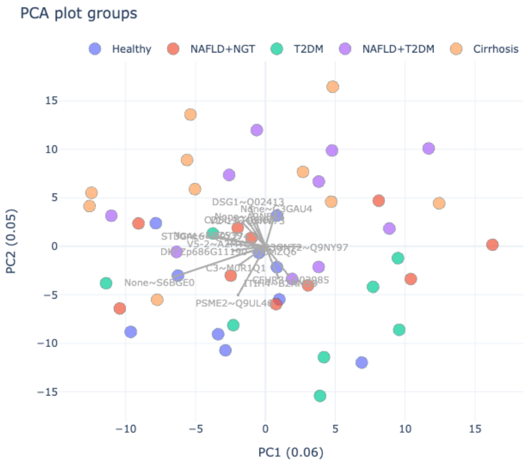
- **Exploratory Analysis**: understand the structure and quality of the data
- **QC/Filtering**: remove proteins and samples that do not meet quality criteria (e.g., missing too many values) – Boxplots, PCA, heatmaps
- **Normalisation**: correct for systematic biases (e.g., sample or instrument variation) — log2, median, z-score, quantile normalisation, etc.
- **Imputation**: handle missing data (Missing Not at Random: below detection limit / Missing at Random: instrument errors, fragmentation efficiency, etc.) — low-intensity imputation, KNN-imputation



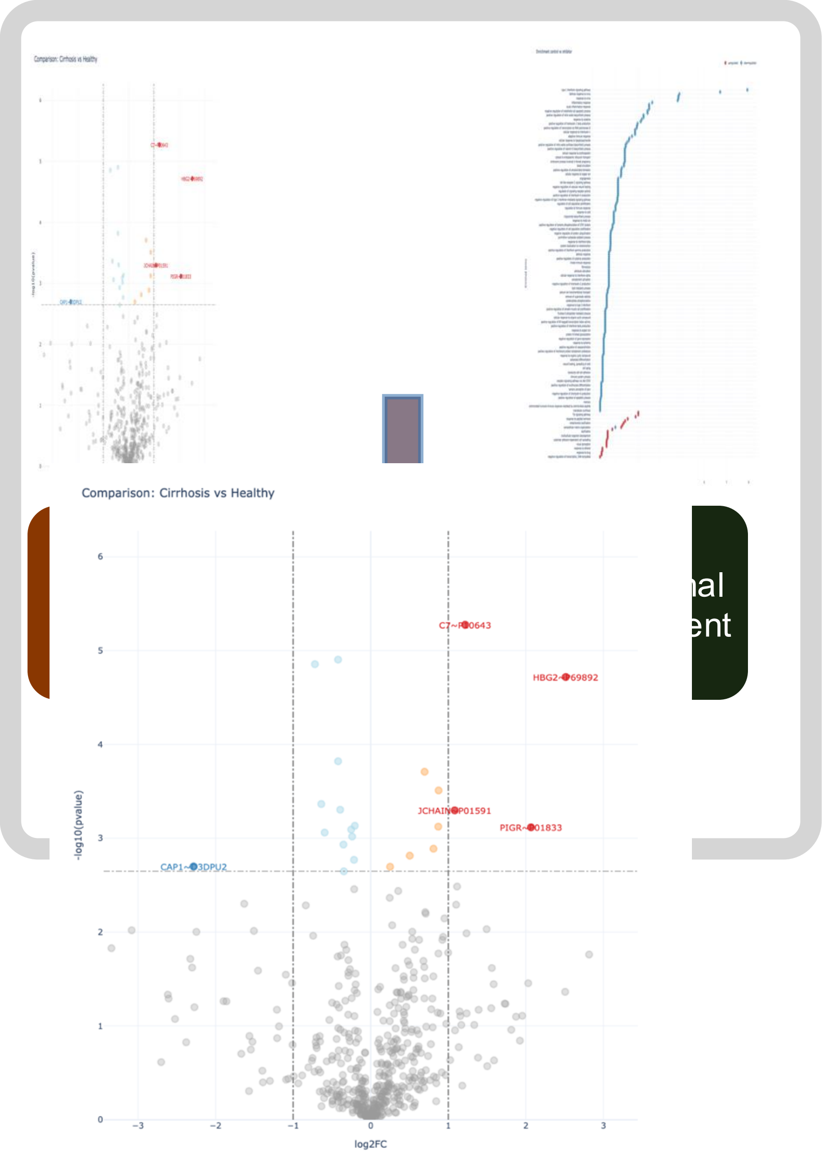








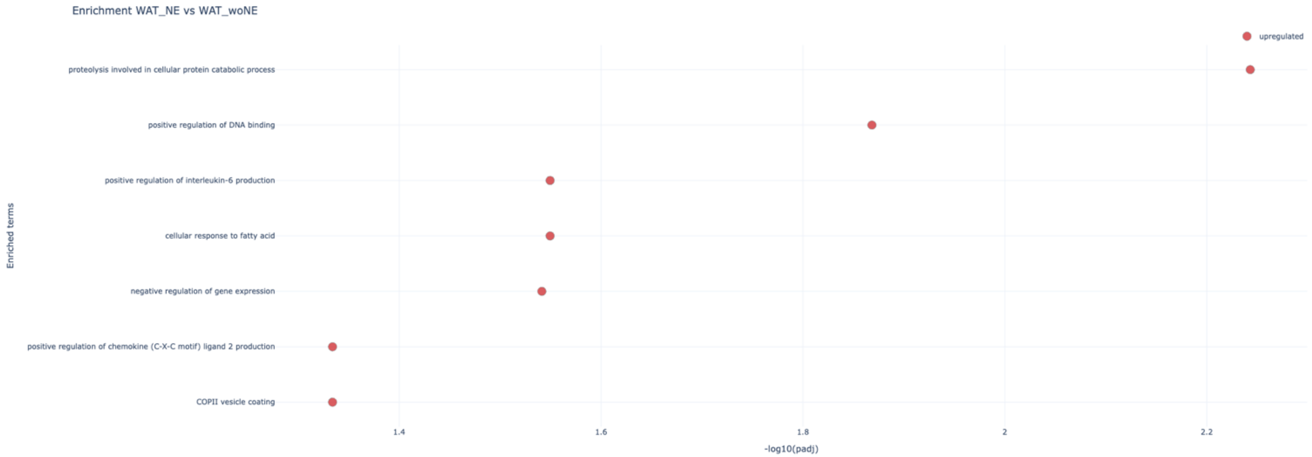
Data Analysis



Differential Regulation: Apply appropriate statistical tests to compare protein intensities between groups — T-test, ANOVA

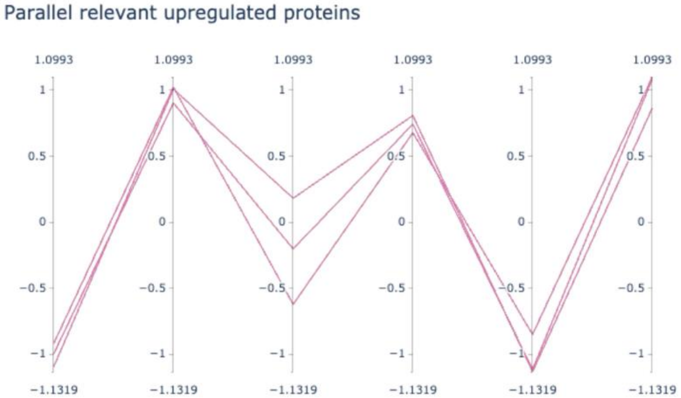
Multiple test correction (e.g., Benjamini-Hochberg False Discovery Rate (FDR))

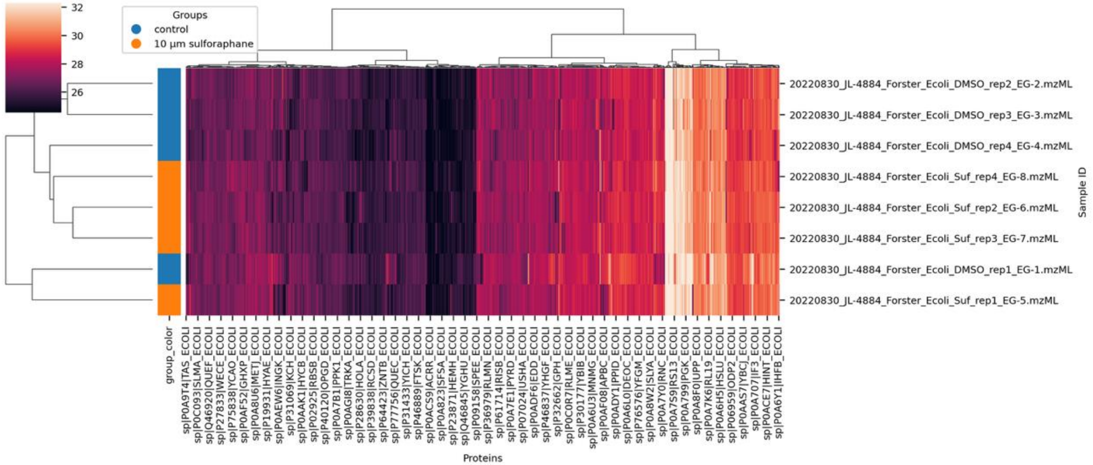
Functional Enrichment: Identify the biological functions, pathways, or processes associated with the differentially regulated proteins — Fisher exact test, Gene Set Enrichment Analysis (GSEA)



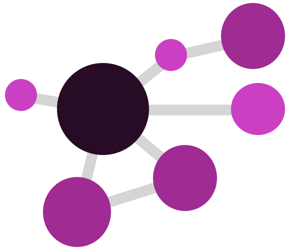
Other Potential Analysis

Clustering/Pattern Discovery hierarchical clustering of samples and proteins – heatmap, profile plots, correlation analysis.

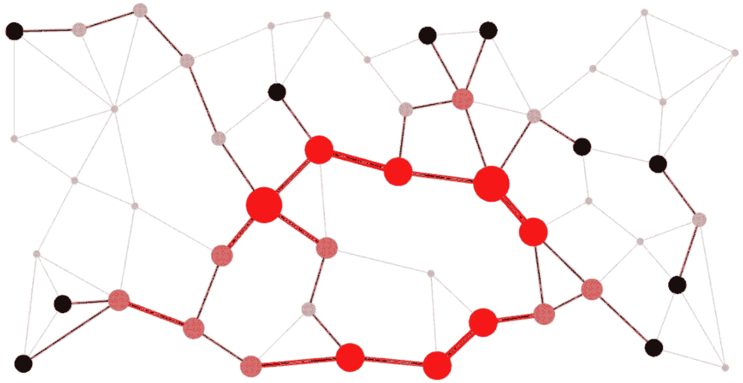




Graph Analysis: Build a protein graph/network and use the structure of nodes and relationships to find relevant patterns.



Graphs

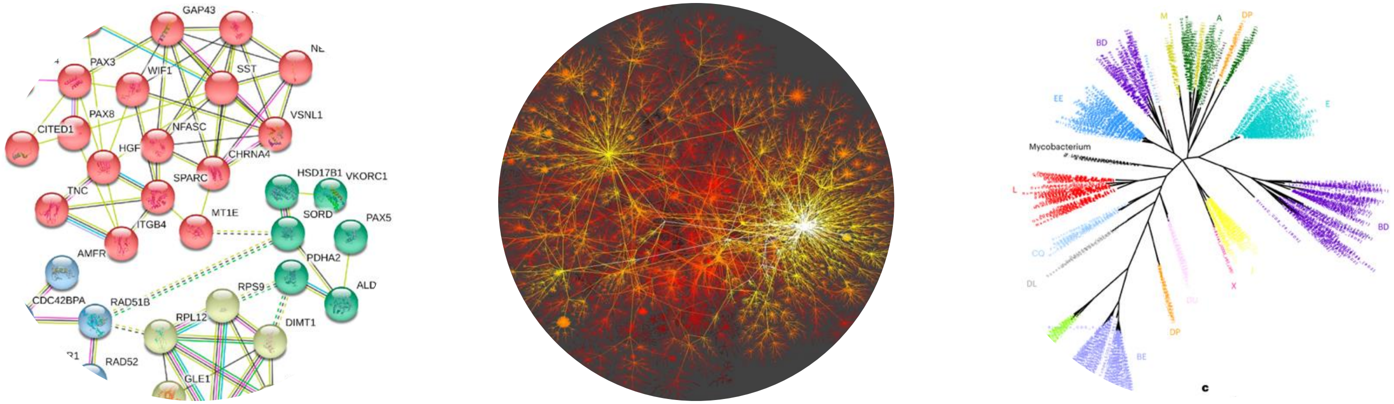


What is a Graph/Network?

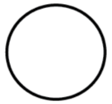
- Data structures of **components (nodes)** connected by **relationships (edges)**

Social networks

Biological networks

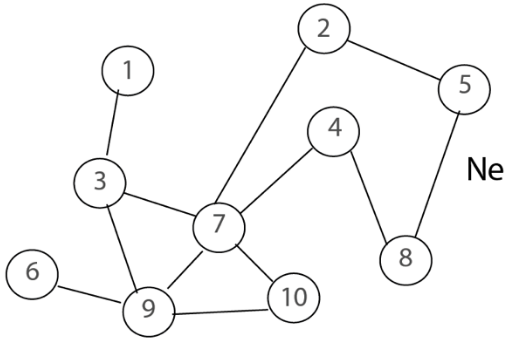


Graphs



Node

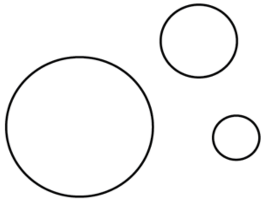
Edge



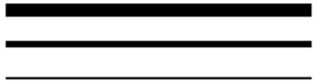
Network

$$\begin{bmatrix} 0&0&1&0&0&0&0&0&0&0 \\ 0&0&0&0&1&0&1&0&0&0 \\ 1&0&0&0&0&0&1&0&1&0 \\ 0&0&0&0&0&0&1&1&0&0 \\ 0&1&0&0&0&0&0&1&0&0 \\ 0&0&0&0&0&0&0&0&1&0 \\ 0&1&1&1&0&0&0&0&1&1 \\ 0&0&0&1&1&0&0&0&0&0 \\ 0&0&1&0&0&1&1&0&0&1 \\ 0&0&0&0&0&0&1&0&1&0 \end{bmatrix}$$

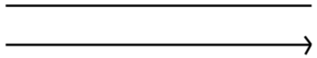
Adjacency matrix



weighted nodes (size)

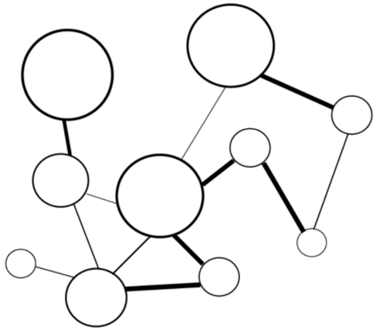


weighted edges (thickness)



undirected edge

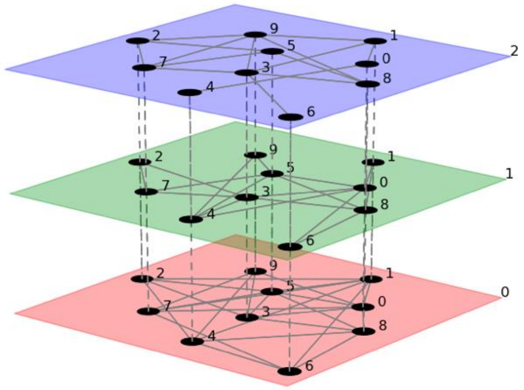
directed edge

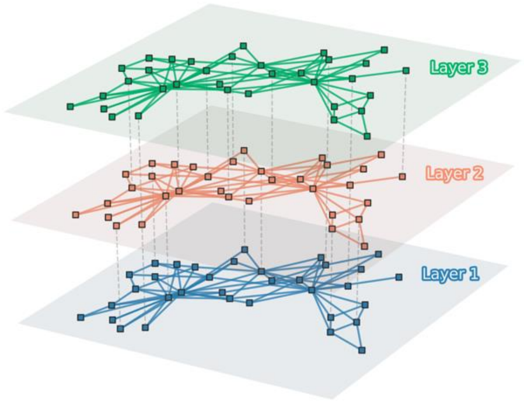


weighted undirected network (thickness)

$$\begin{bmatrix} 0&0&W_{1,3}&0&0&0&0&0&0&0 \\ 0&0&0&0&W_{2,5}&0&W_{2,7}&0&0&0 \\ W_{3,1}&0&0&0&0&0&W_{3,7}&0&W_{3,9}&0 \\ 0&0&0&0&0&0&W_{4,7}&W_{4,8}&0&0 \\ 0&W_{5,2}&0&0&0&0&0&W_{5,8}&0&0 \\ 0&0&0&0&0&0&0&0&W_{6,9}&0 \\ 0&W_{7,2}&W_{7,3}&W_{7,4}&0&0&0&0&W_{7,9}&W_{7,10} \\ 0&0&0&W_{8,4}&W_{8,5}&0&0&0&0&0 \\ 0&0&W_{9,3}&0&0&W_{9,6}&W_{9,7}&0&0&W_{9,10} \\ 0&0&0&0&0&0&W_{10,7}&0&W_{10,9}&0 \end{bmatrix}$$

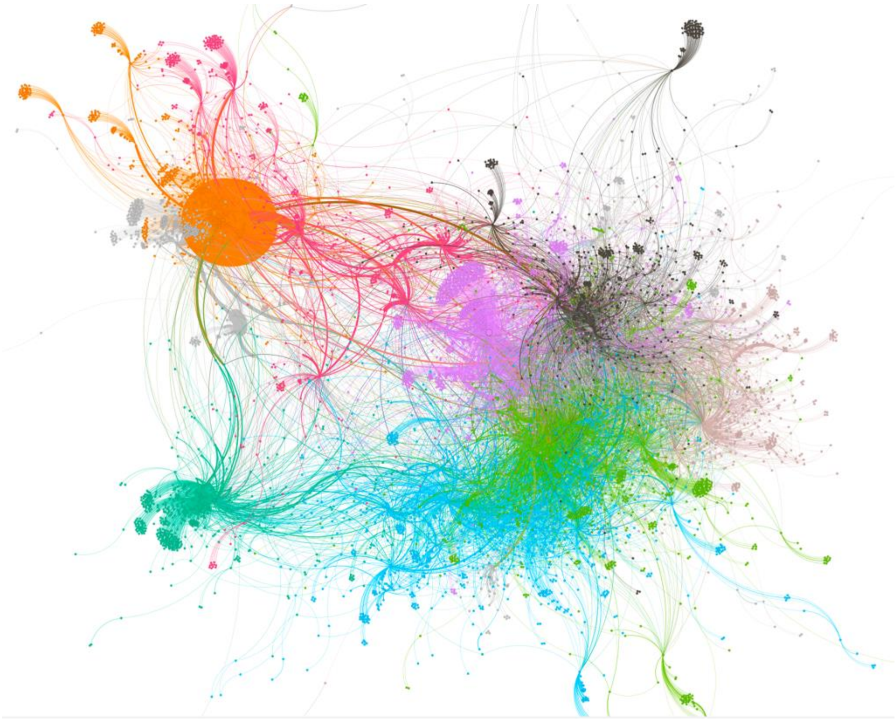
Weighted adjacency matrix





Why Graphs?

- These structures allow:
  - Quick **integration** of **heterogeneous data** based on relationships
  - **Graph theory** methods can be used to **analyse** and **interpret** data, e.g., topological properties can be used to explain:
    - The possible **role** of specific components
    - The **flow** of information
    - The **robustness** of the system
- **Visualize** data



How to Analyse Graph Structures

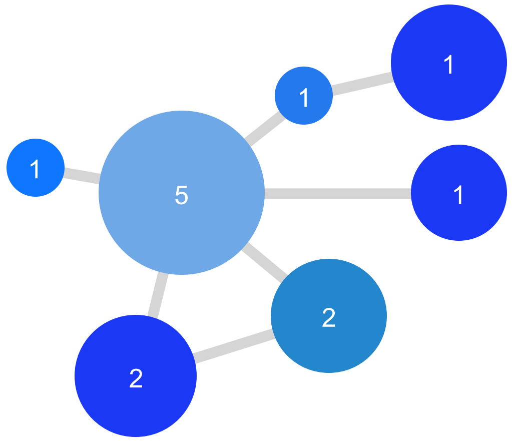
- **Graph Theory**: algorithms that allow you to extract relevant information from the topology of the graph.
  - **Topological Features:** Centrality, degree, clustering, etc.

- **Graph Machine Learning**:
  - Embeddings
  - Graph Neural Networks

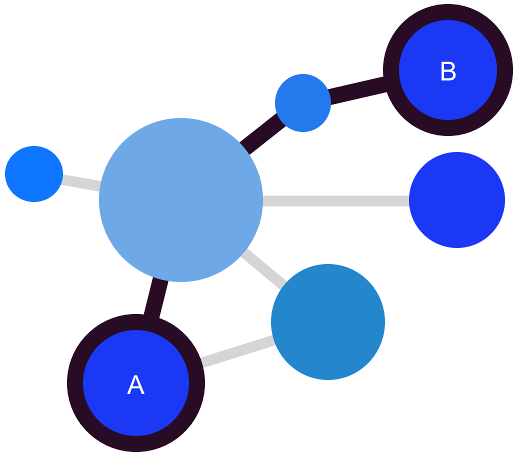
https://huggingface.co/blog/intro-graphml

Some Topological Features

Topological properties can help extract meaningful information and identify relevant structures within the network

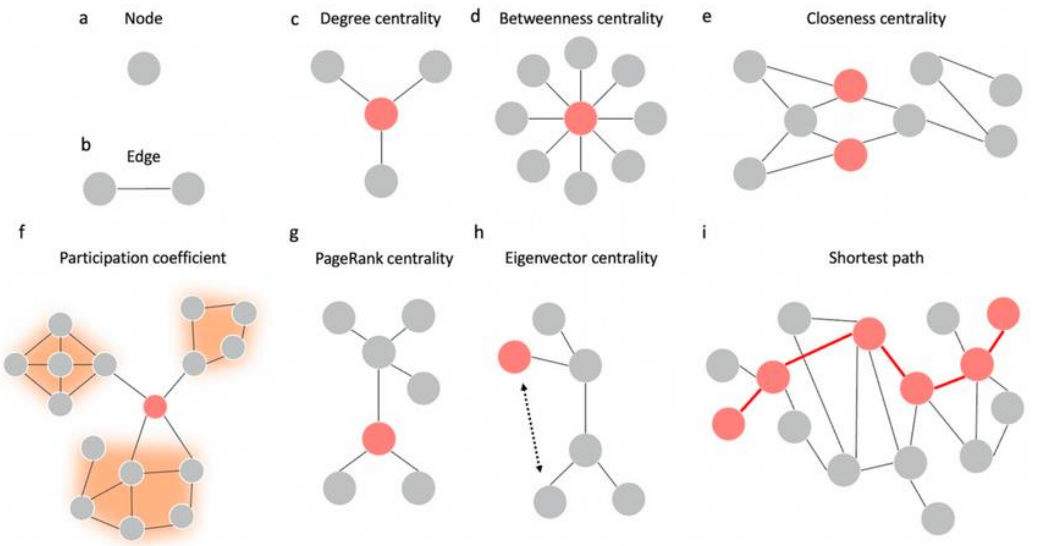


Degree

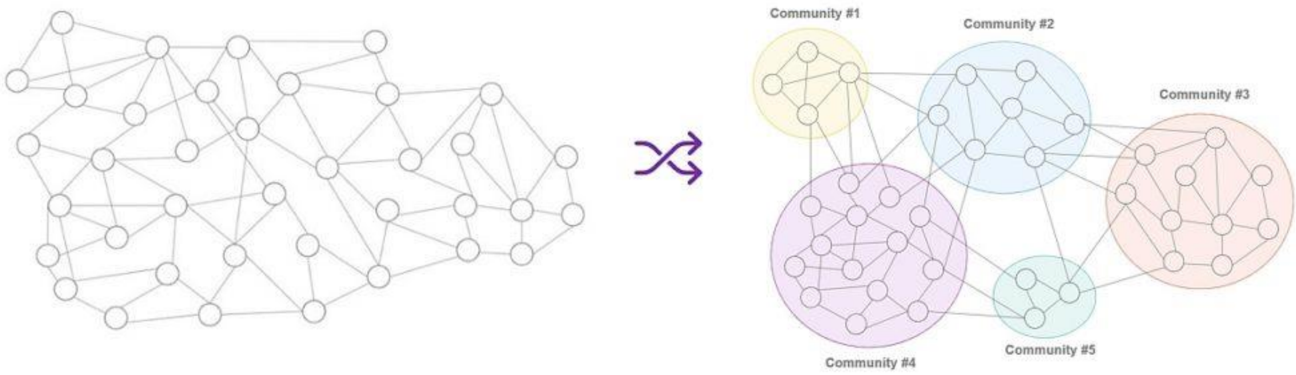


Shortest path

Centrality



Clustering



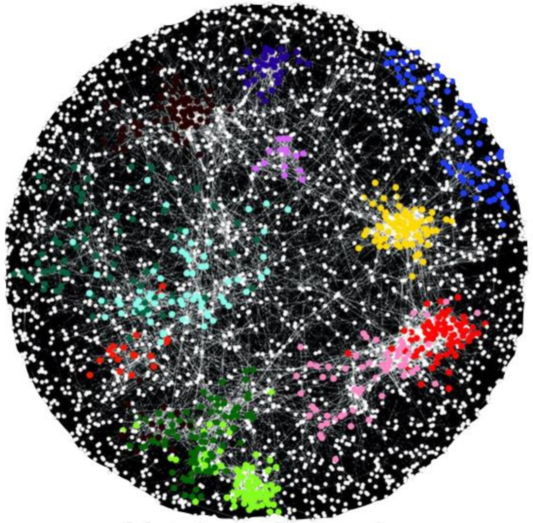
https://timbr.ai/community-detection-algorithm/

Graph Theory Measures and Their Application to Neurosurgical Eloquence. Cancers 2023

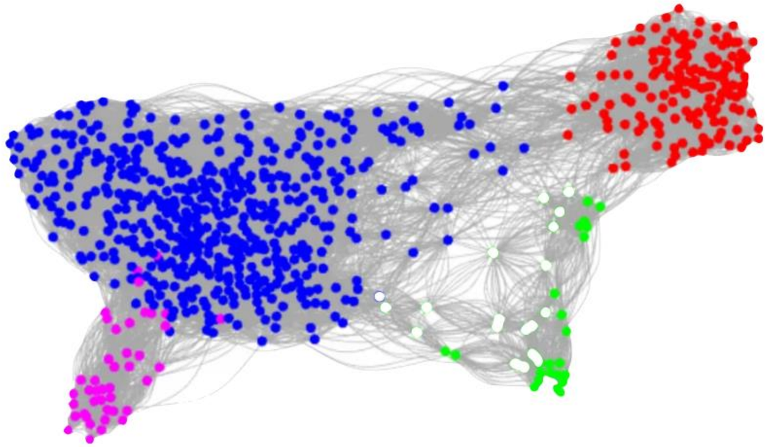
Graphs in Biology

https://towardsdatascience.com/umap-for-data-integration-50b5cfa4cdcd
http://snap.stanford.edu/deepnetbio-ismb/ipynb/Human+Disease+Network.html
https://cytoscape.org/cytoscape-tutorials/presentations/ppi-tools1-2017-mpi.html#/
https://en.wikipedia.org/wiki/Metabolic_network
https://www.scienceandfood.org/the-flavor-network/

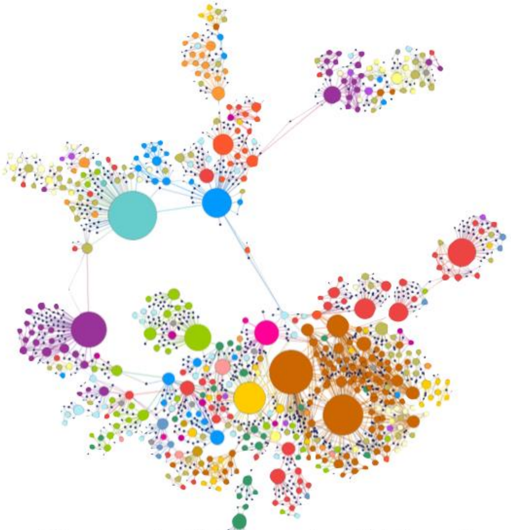
Protein-protein Interaction Networks



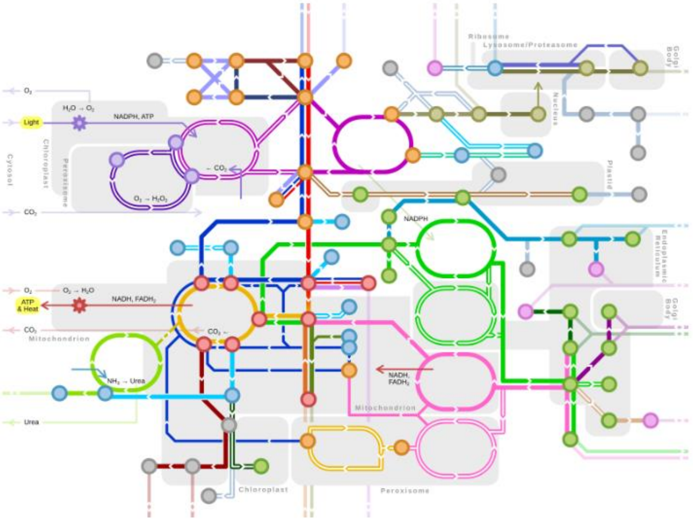
Single cell Networks



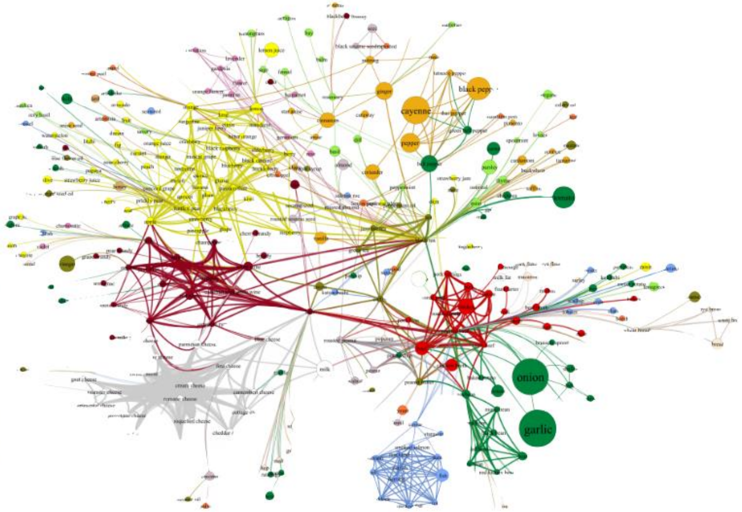
Disease Networks



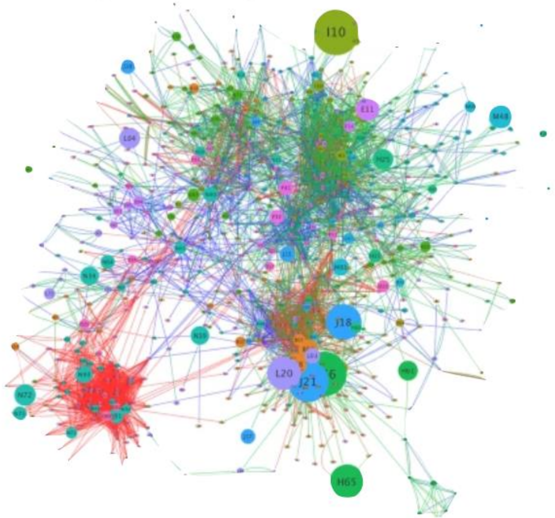
Metabolic Networks



Food Networks

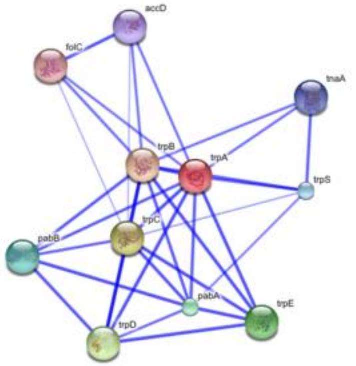


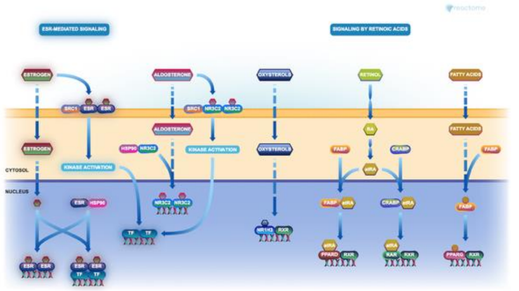
Diagnosis Progression Networks



How to Build a Network

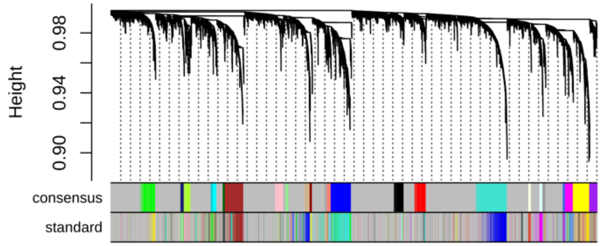
Data to Graph

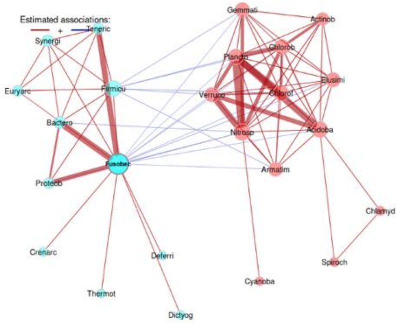




Data sources

- STRING — https://string-db.org/
- BioGRID — https://thebiogrid.org/
- IntAct — https://www.ebi.ac.uk/intact
- REACTOME — https://reactome.org/
- KEGG — https://www.genome.jp/kegg/
- MINT — https://mint.bio.uniroma2.it/



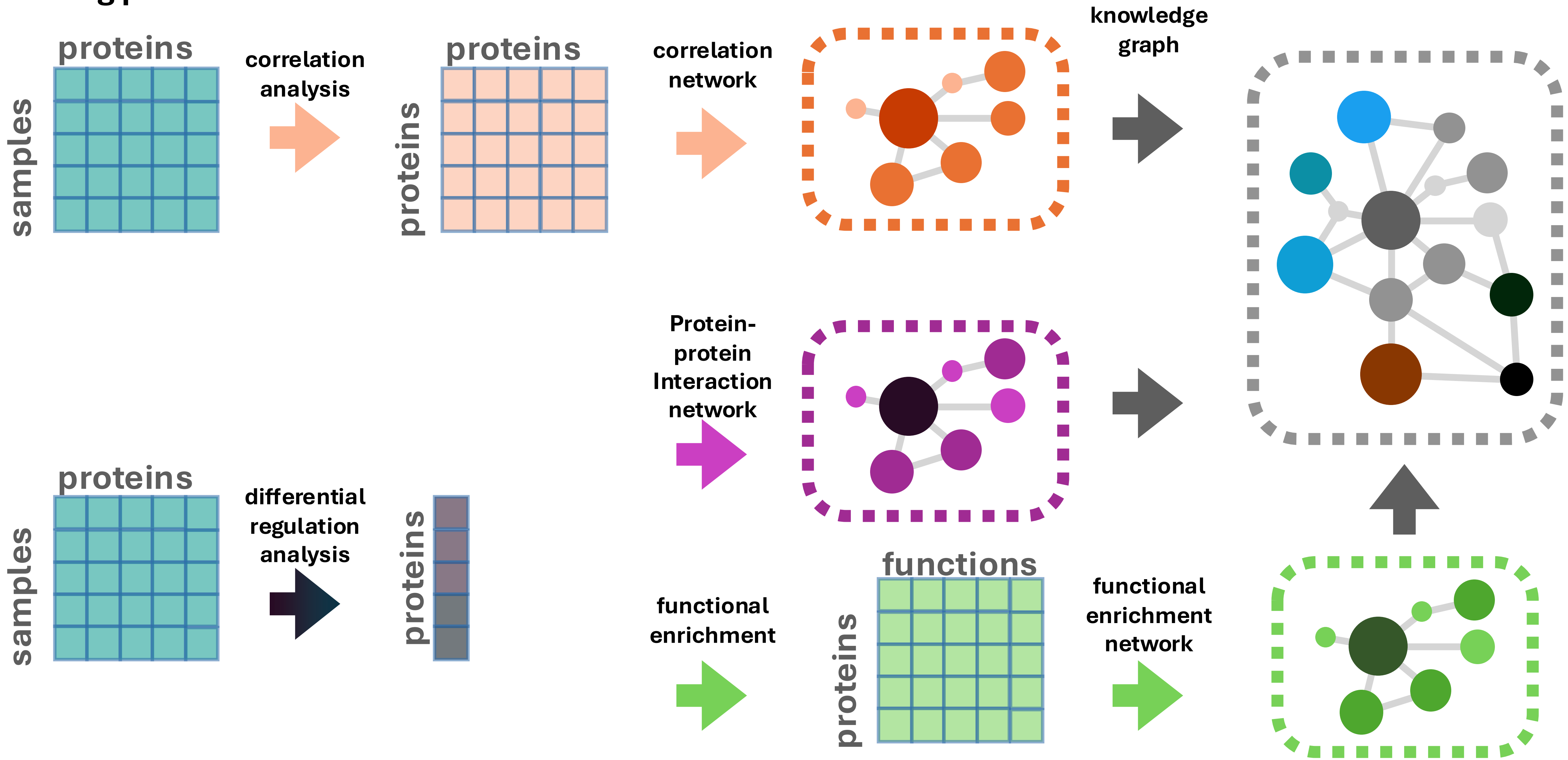


Correlation-based networks — constructed by calculating pairwise correlations between entities based on their expression profiles across multiple conditions, time points, or samples (Weighted gene co-expression network analysis (WGCNA), co-abundance networks)

Knowledge-base approaches — also called knowledge graphs and built by integrating heterogeneous data from multiple sources —> Knowledge Graphs

How to Build a Network

Starting point



Knowledge Graphs

Knowledge Graphs

- A way to organise **knowledge/information** by defining associations or relationships
- These relationships facilitate **integration**, **management** and **enrichment** of data
- The objective when setting up a KG:

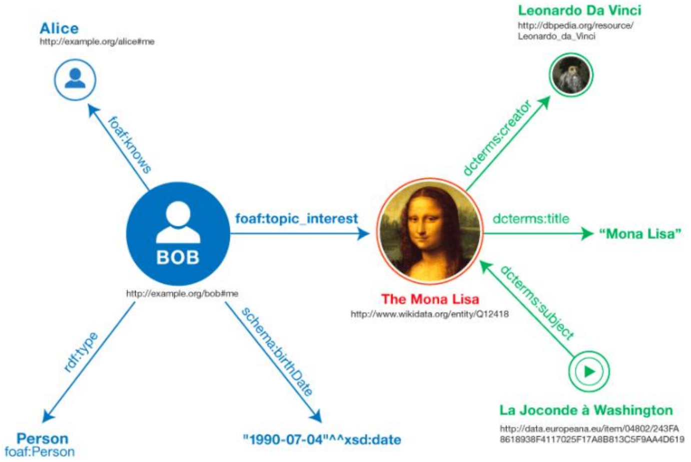
Standardisation / FAIRification

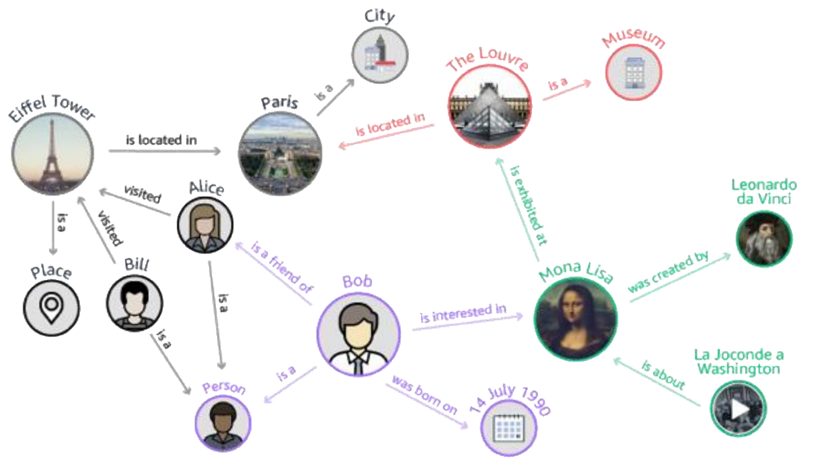
Reusability

Interpretability

Automation

Representation/Visualisation


Alice
http://example.org/alice#me
foaf:knows
BOB
http://example.org/bob#me
foaf:topic_interest
rdf:type
Person
foaf:Person
schema:birthDate
"1990-07-04"^^xsd:date
Leonardo Da Vinci
http://dbpedia.org/resource/Leonardo_da_Vinci
dcterms:creator
The Mona Lisa
http://www.wikidata.org/entity/Q12418
dcterms:title
"Mona Lisa"
dcterms:subject
La Joconde à Washington
http://data.europeana.eu/item/04802/243FA8618938F4117025F17A8B813C5F9AA4D619

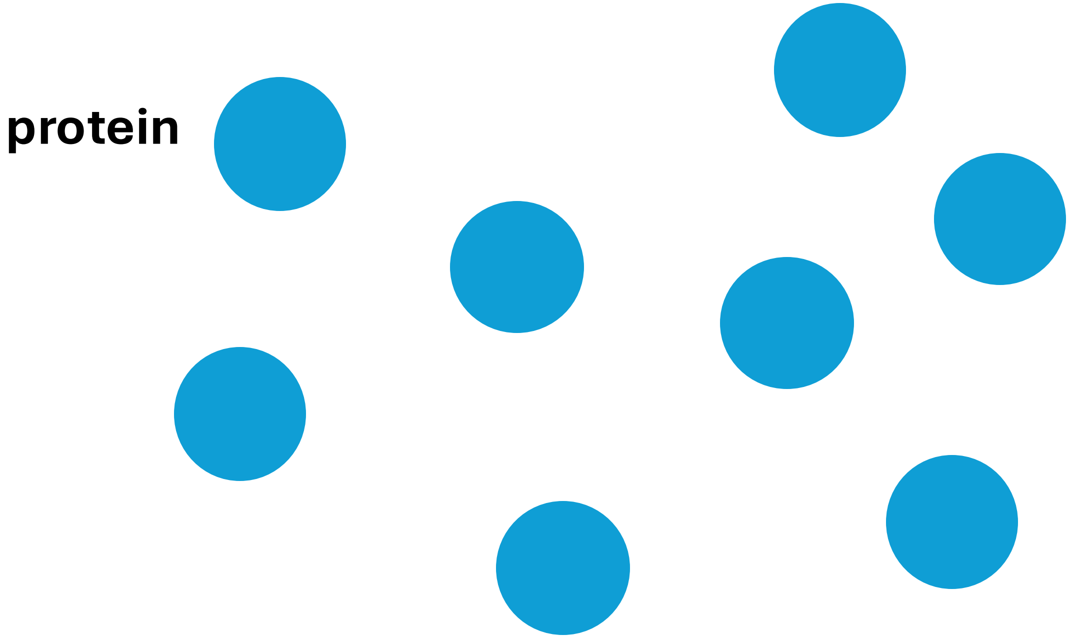


The Knowledge Graph Cookbook. Andreas Blumauer and Helmut Nagy. 2020

Knowledge Graphs

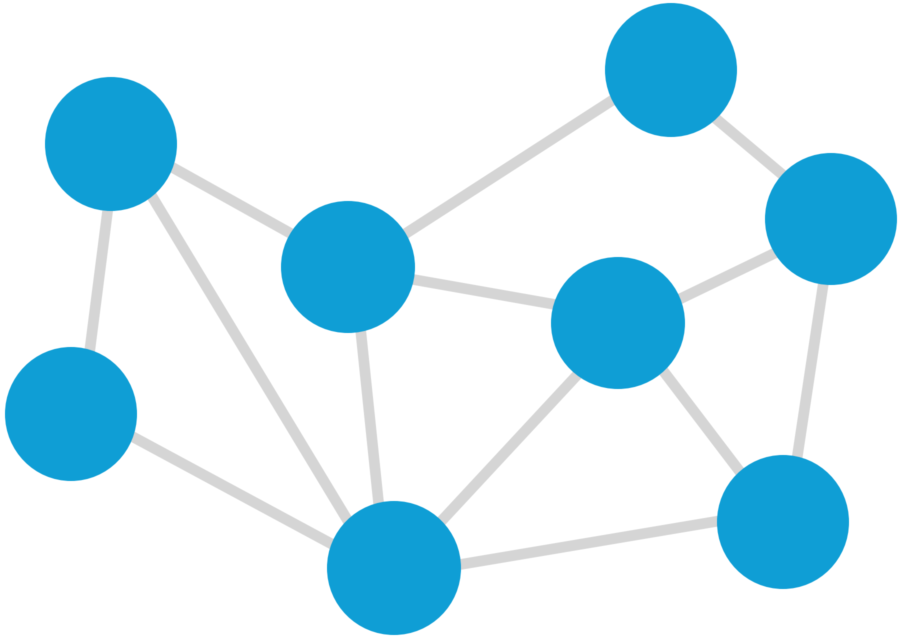
How Does It Work?

How Does It Work?



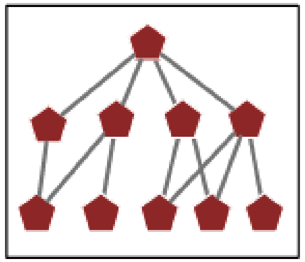
protein

How Does It Work?

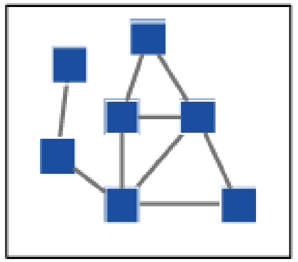


How Does It Work?

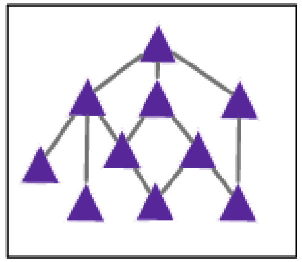
DISEASES



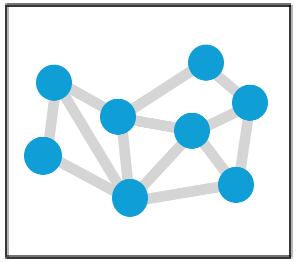
DRUGS



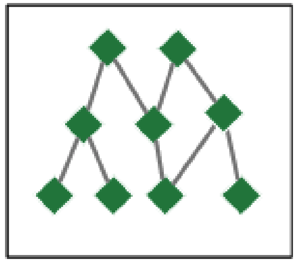
PATHWAYS



PROTEINS



SIDE EFFECTS

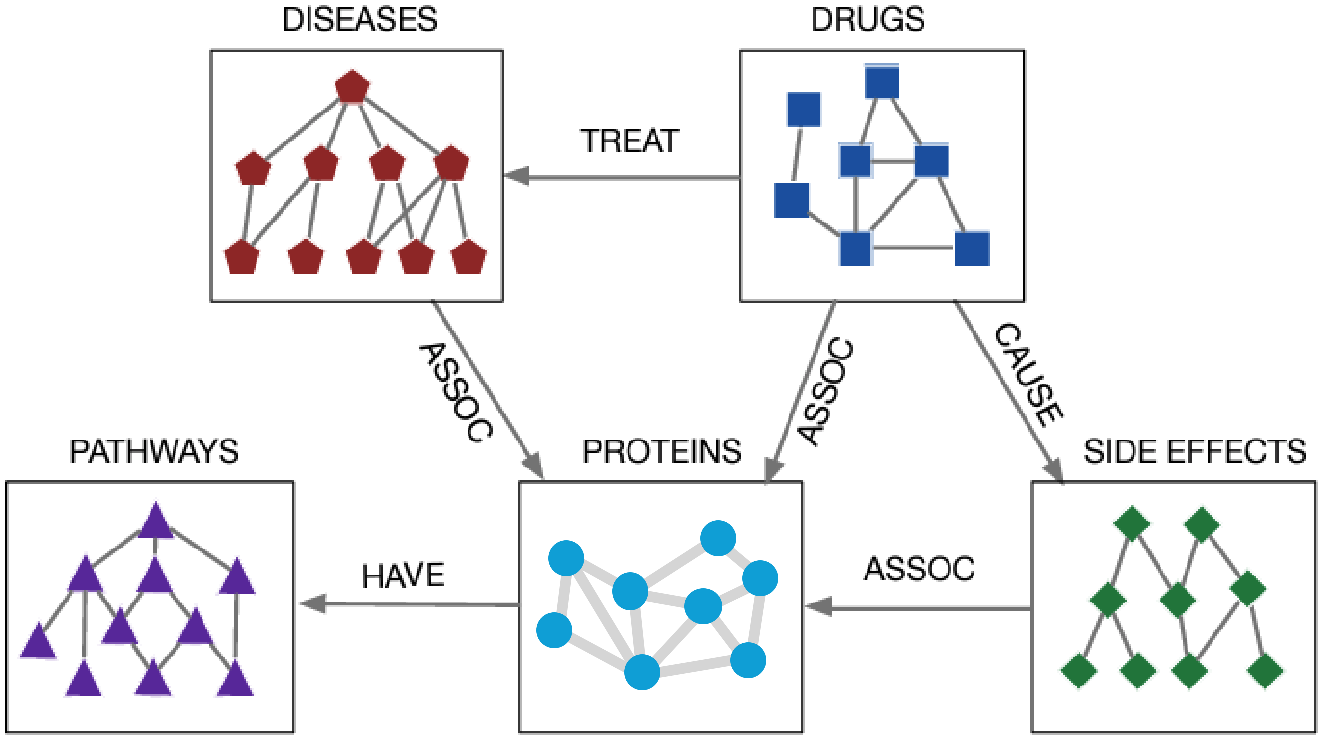


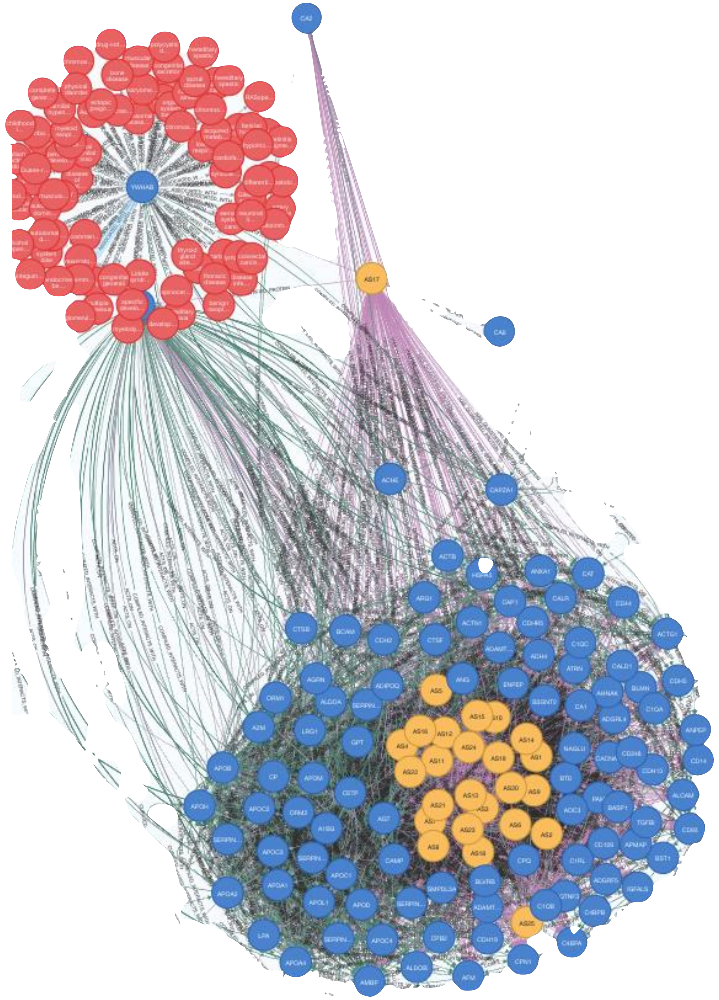
https://snap.stanford.edu/graphlearning-workshop/slides/stanford_graph_learning_Biomedicine.pdf

https://snap.stanford.edu/graphlearning-workshop/slides/stanford_graph_learning_Biomedicine.pdf

Knowledge Graph

Focus on data integration to represent complex biological systems and be able to reason over them





https://snap.stanford.edu/graphlearning-workshop/slides/stanford_graph_learning_Biomedicine.pdf

https://snap.stanford.edu/graphlearning-workshop/slides/stanford_graph_learning_Biomedicine.pdf

Building a Knowledge Graph

1. Define the **questions** you want to answer
2. Define **what data** can be used to answer these questions and **how it is linked** — Data model
3. Find **where to get these data**
4. Get the data, **standardise** it and **format** it
5. Generate the **graph**
6. **Query the graph** to answer the questions

Building a Knowledge Graph

Building a Knowledge Graph

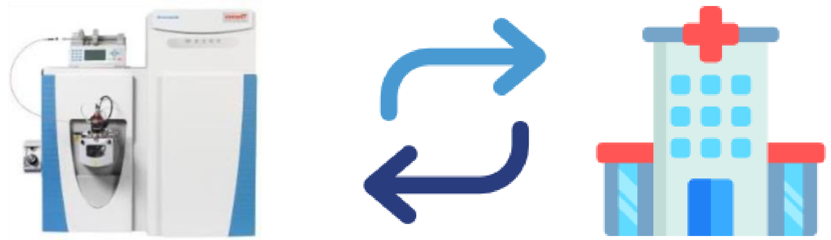
Exercise

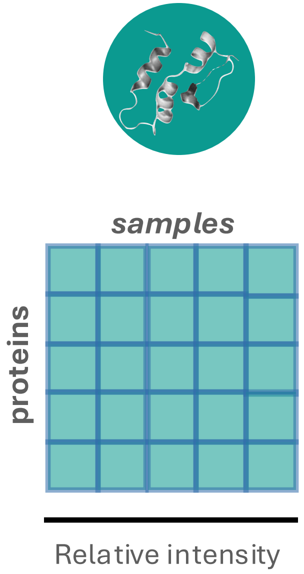
Create a data model that allows us to answer the question:

What drugs related to our disease of interest target some of the proteins identified in our experiment or relevant protein complexes and pathways?

Application

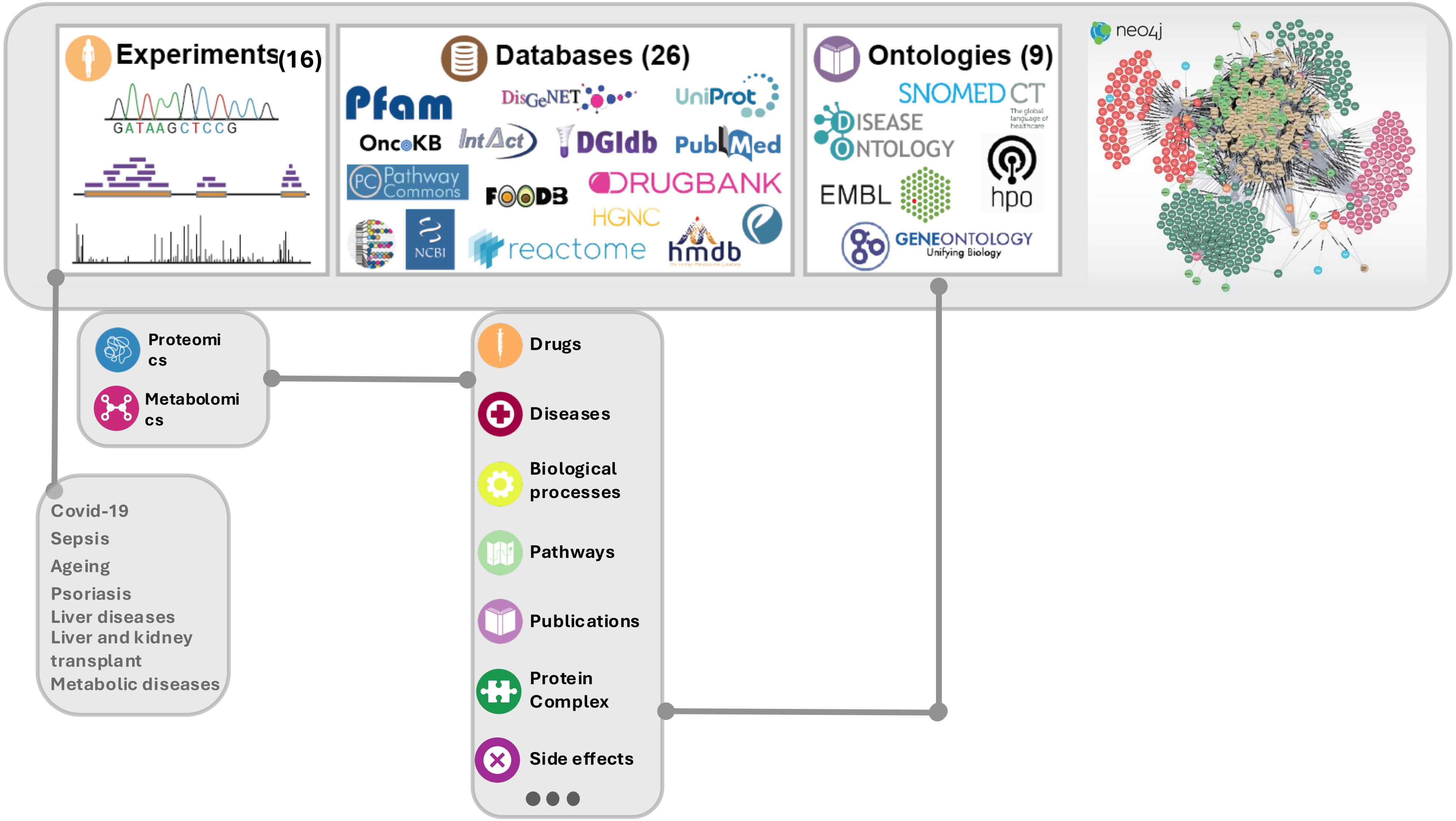
Clinical Knowledge Graph





https://www.nature.com/articles/s41587-021-01145-6

Clinical Knowledge Graph – CKG



Experiments(16)

Databases (26)

Ontologies (9)

Proteomics

Metabolomics

Covid-19
Sepsis
Ageing
Psoriasis
Liver diseases
Liver and kidney transplant
Metabolic diseases

Drugs

Diseases

Biological processes

Pathways

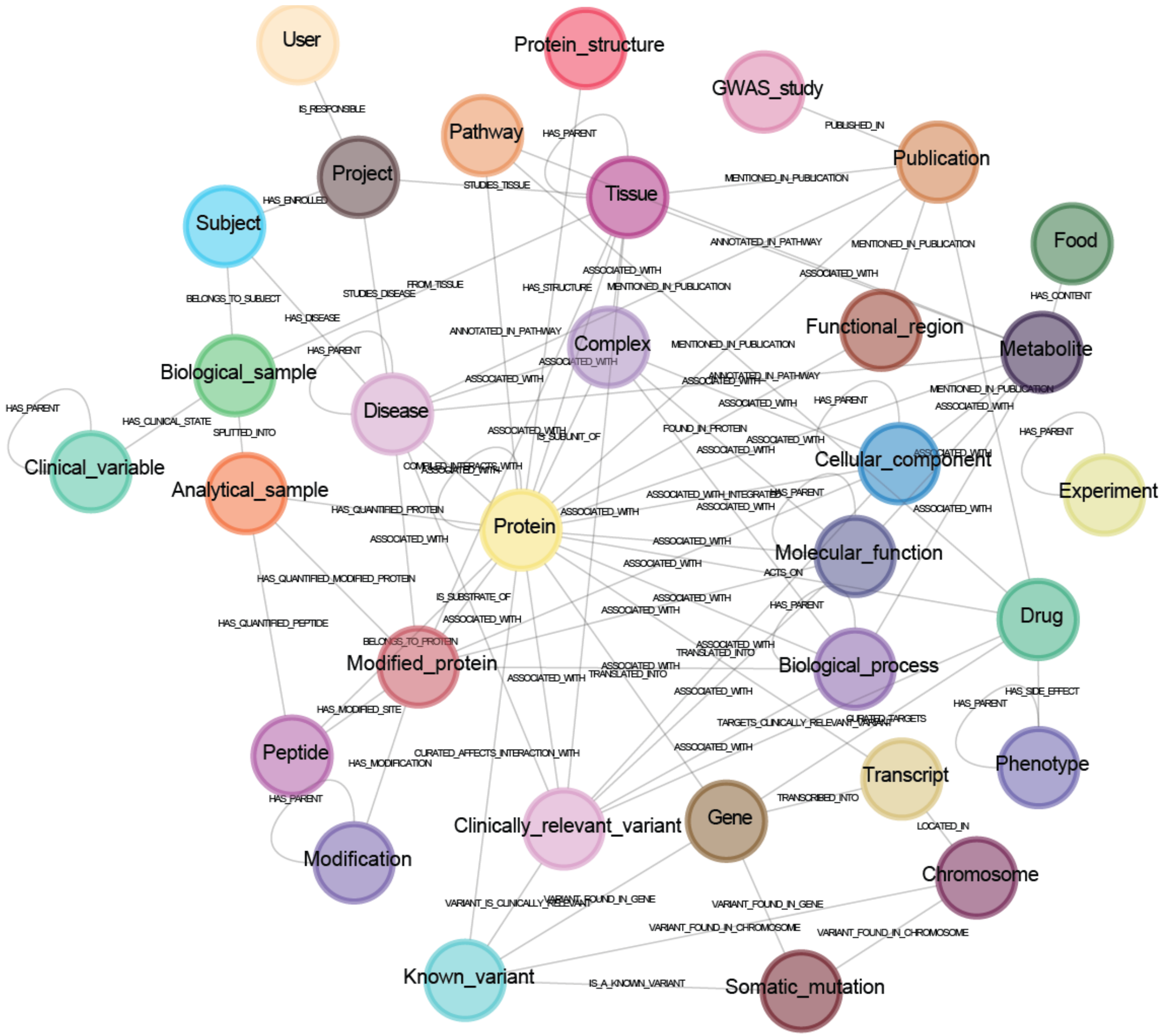
Publications

Protein Complex

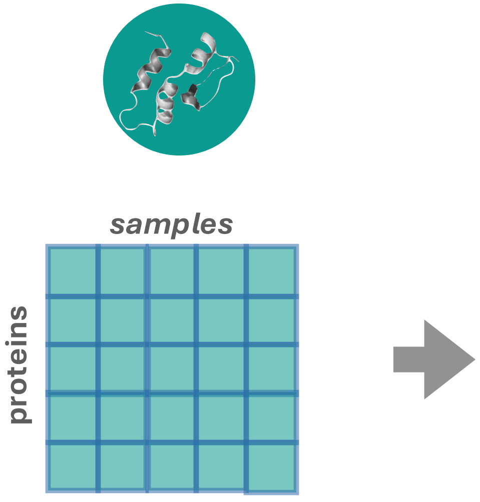
Side effects

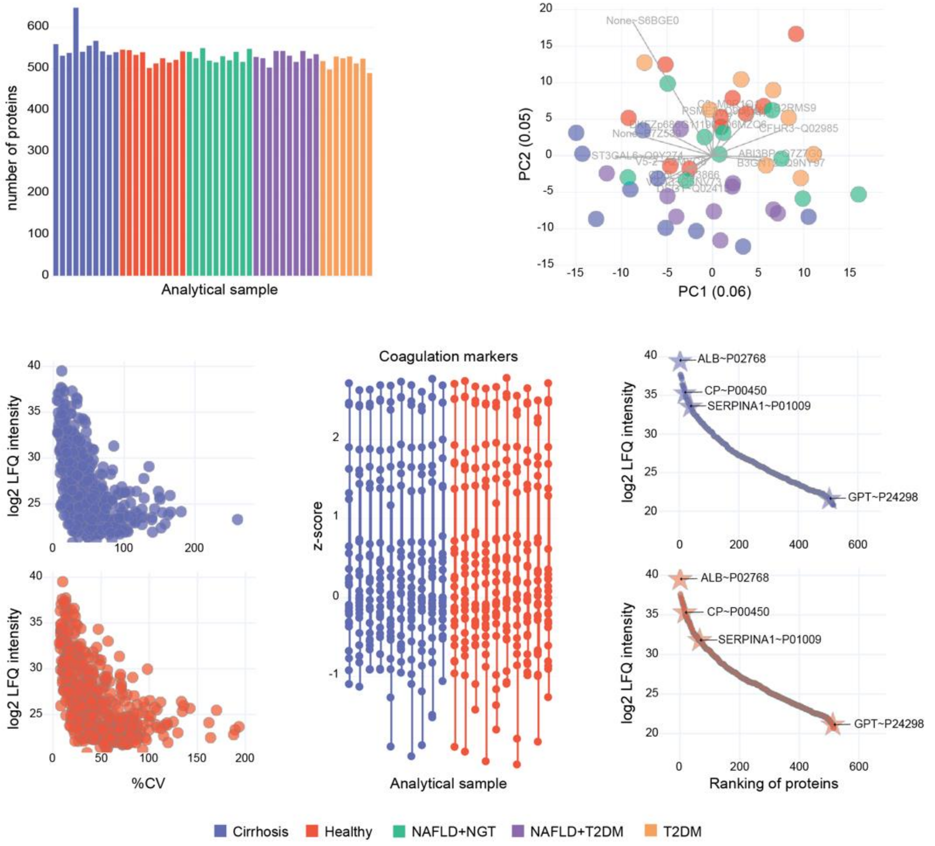
●●●

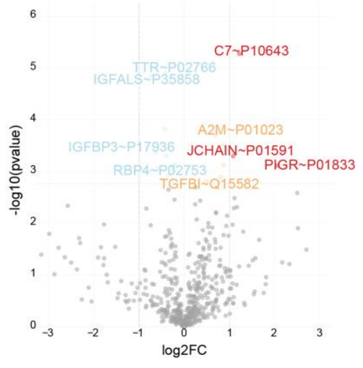
Clinical Knowledge Graph Data Model

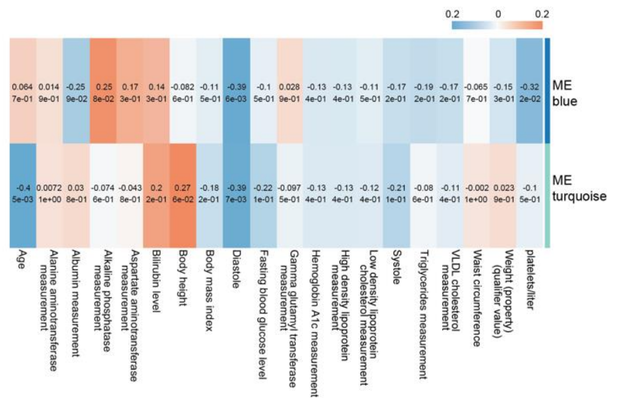


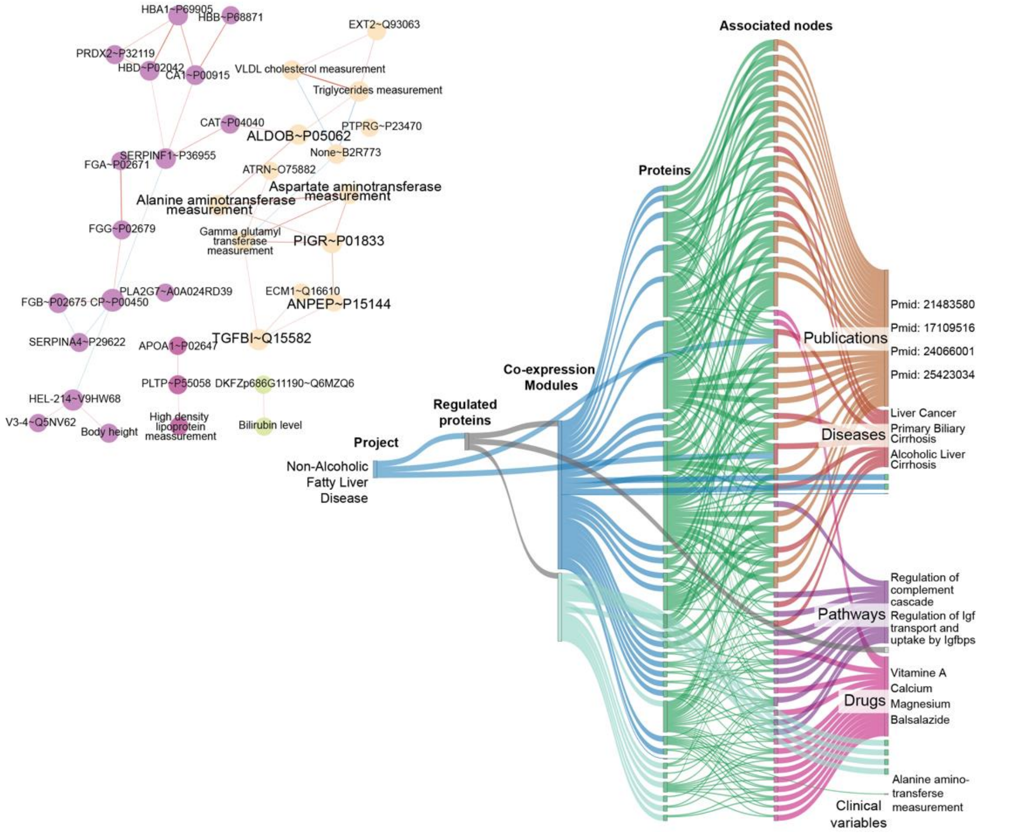
Clinical Knowledge Graph – CKG





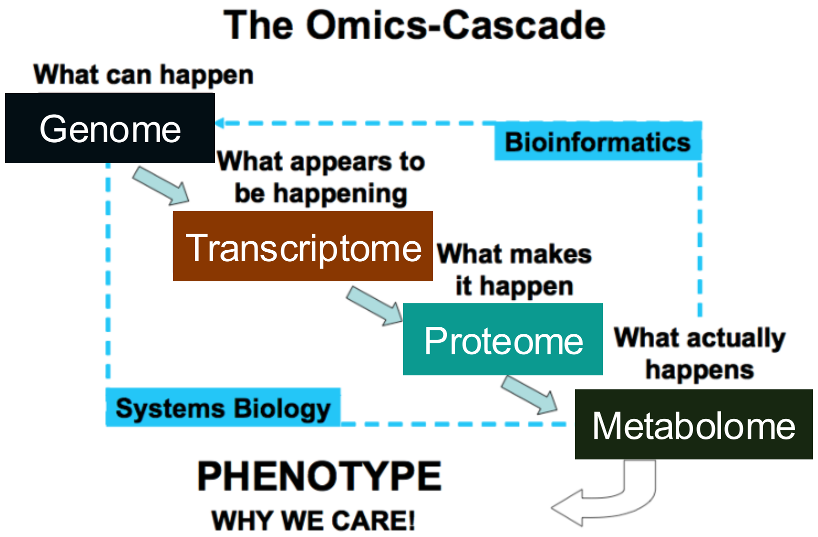






Open Source Tools

Omics Data



The Omics-Cascade

What can happen

Genome

What appears to be happening

Bioinformatics

Transcriptome

What makes it happen

Proteome

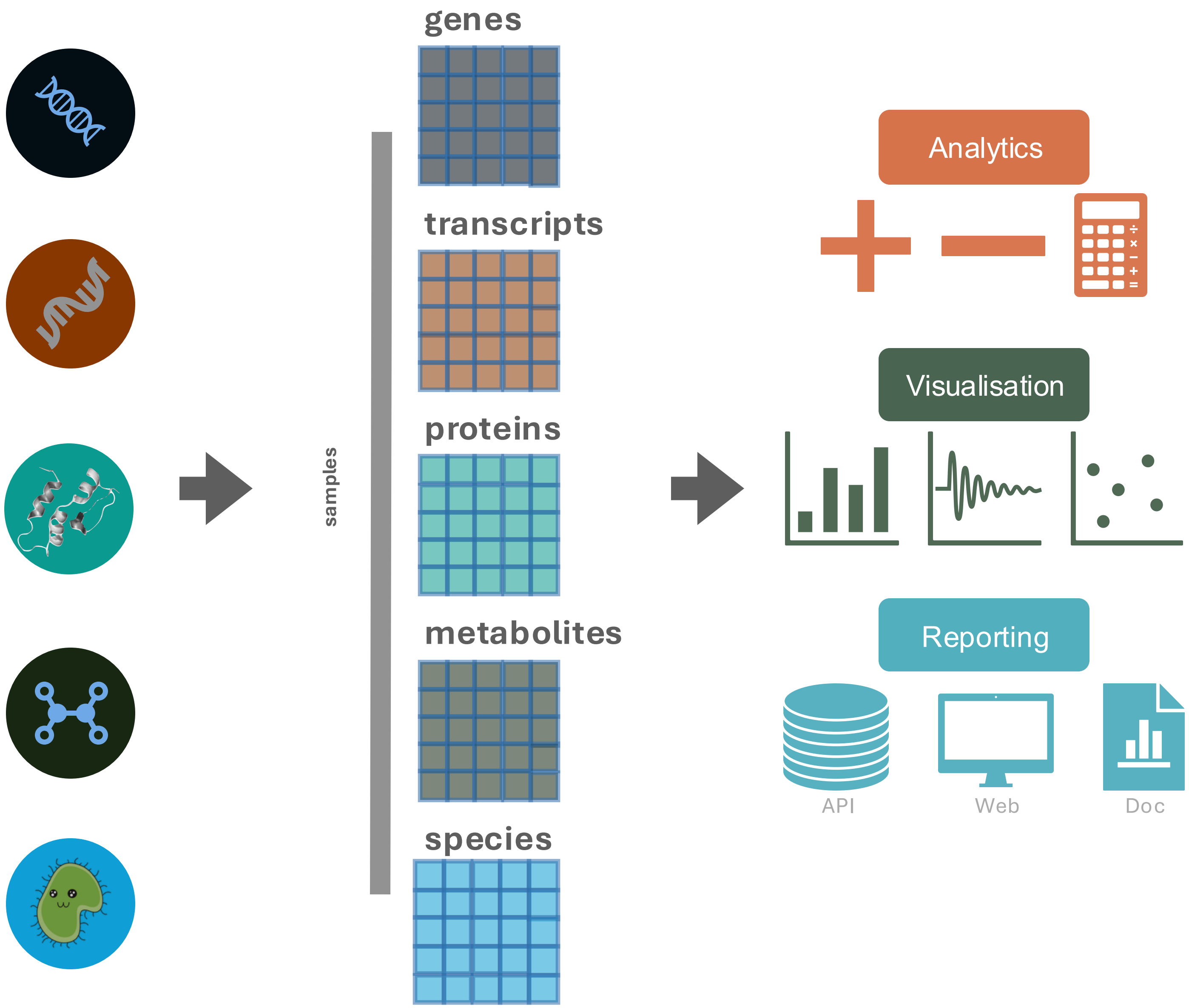
What actually happens

Systems Biology

Metabolome

PHENOTYPE

WHY WE CARE!



genes

transcripts

proteins

metabolites

species

samples

Analytics

Visualisation

Reporting

API

Web

Doc

MoNA Open Source Tools

Analytics

Visualisation

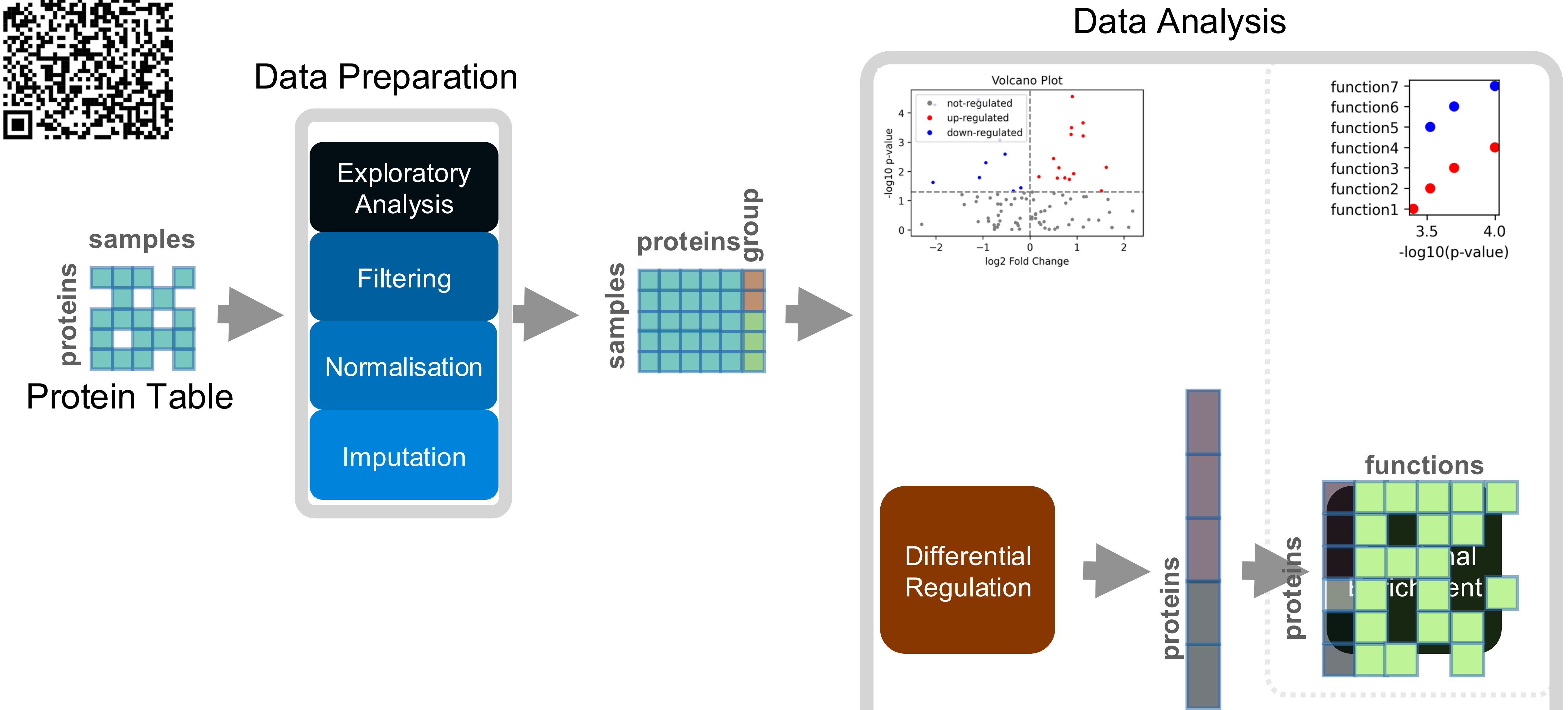
Reporting

Analytics core library — analytics-core.readthedocs.io/

Visualization core library — github.com/Multiomics-Analytics-Group/vuecore

Automated Reporting library and cli — github.com/Multiomics-Analytics-Group/vuegen

Acore – Analytical core – workflow example



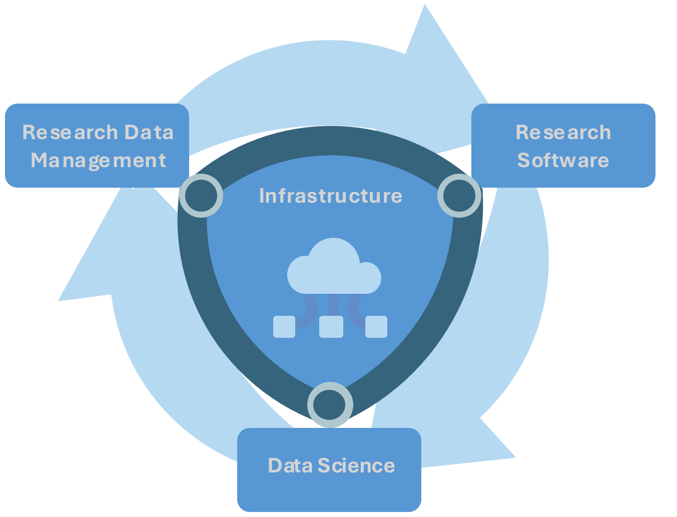
Other Tools for Downstream Proteomics Analysis

Category	Tool	
Statistical Analysis	Perseus, limma, MSstats, AlphaStats	GUI and R/Python-based options
Functional Enrichment	Enrichr	Web tool
Pathway Analysis	Reactome, IPA (Qiagen)	Curated databases
Network Analysis	STRING, Cytoscape, Gephi	Visual and analytical network tools
Integrated Platforms	CKG, Proteome Discoverer, AlphaPept	Combine multiple steps

Acknowledgements

Multi-omics Network Analytics Research Group

Informatics Platform



The Novo Nordisk Foundation
Center for Biosustainability

novo nordisk **fonden**

PIONEER CENTRE FOR ARTIFICIAL INTELLIGENCE

✉ albsad@dtu.dk

https://github.com/Multiomics-Analytics-Group

https://multiomics-analytics-group.github.io/