09-02-2023

Application of integrated network analyses in disease characterization

Rui Benfeitas

NBIS - National Bioinformatics Infrastructure Sweden

Science for Life Laboratory, Stockholm

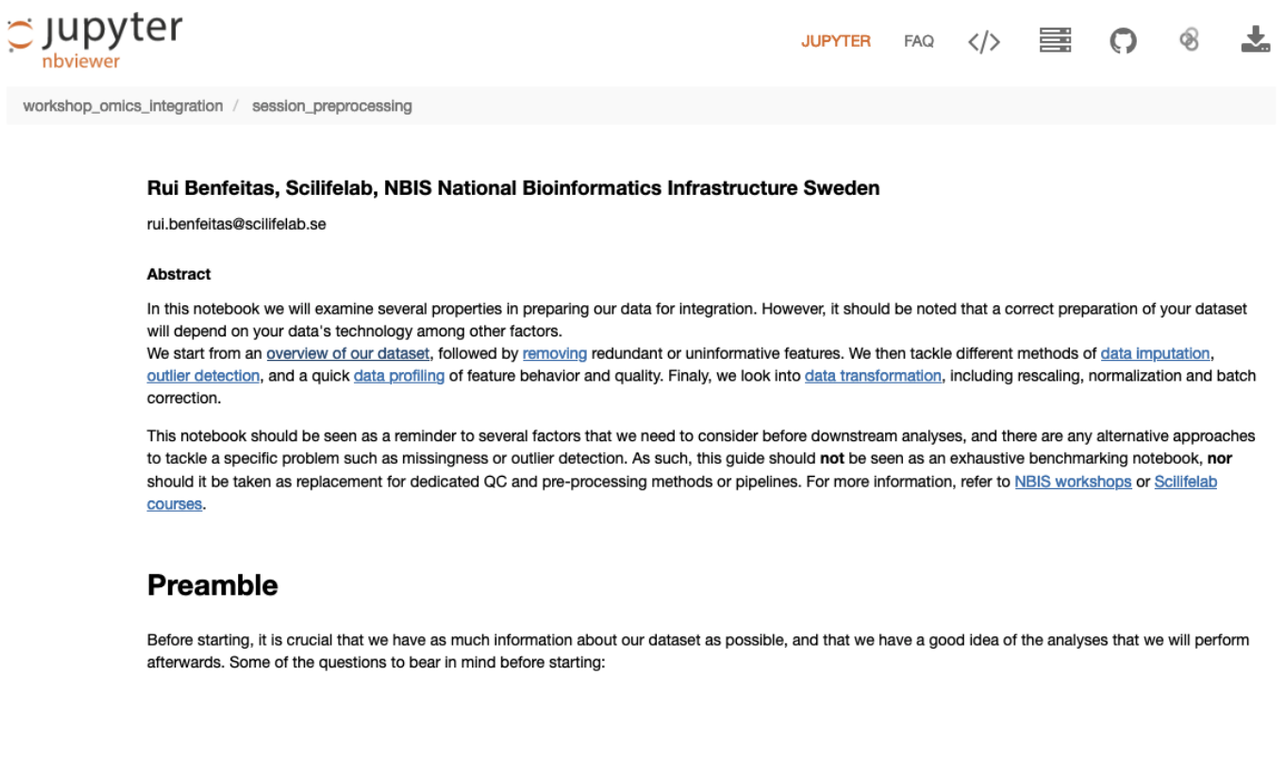
rui.benfeitas@scilifelab.se

Companion resources

Pre-course information and installation instructions

Data pre-processing notebook

Notebooks


jupyter nbviewer

JUPYTER FAQ

workshop_omics_integration / session_preprocessing

Rui Benfeitas, Scilifelab, NBIS National Bioinformatics Infrastructure Sweden

rui.benfeitas@scilifelab.se

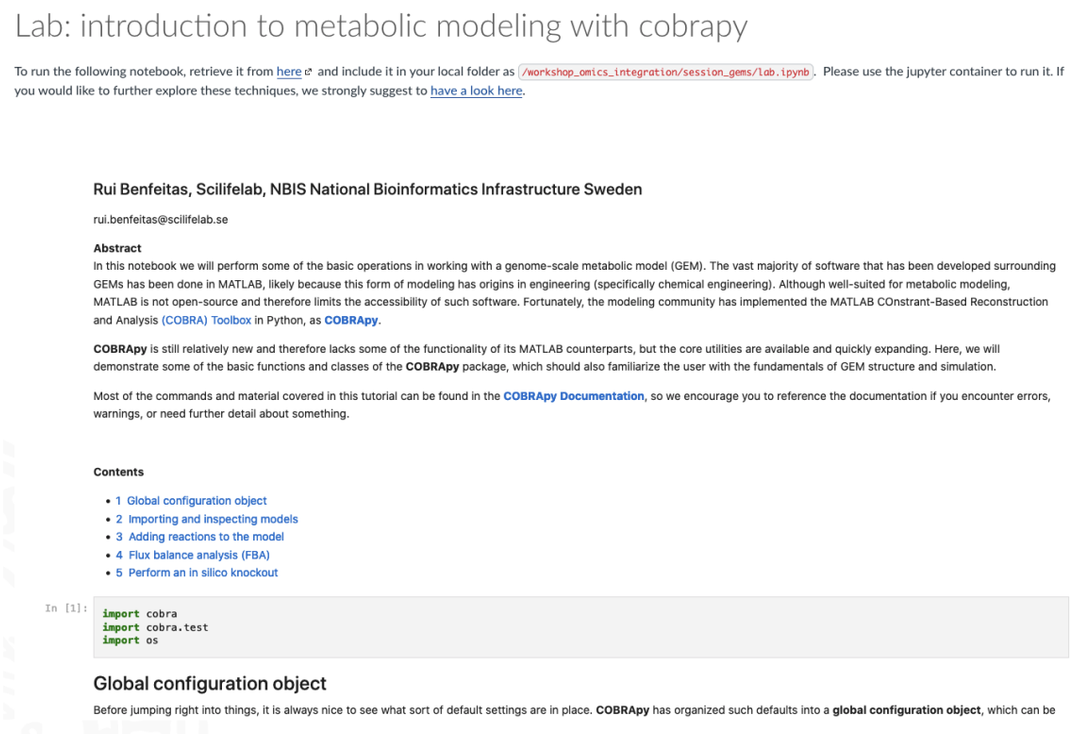
Abstract

In this notebook we will examine several properties in preparing our data for integration. However, it should be noted that a correct preparation of your dataset will depend on your data's technology among other factors.
We start from an overview of our dataset, followed by removing redundant or uninformative features. We then tackle different methods of data imputation, outlier detection, and a quick data profiling of feature behavior and quality. Finaly, we look into data transformation, including rescaling, normalization and batch correction.

This notebook should be seen as a reminder to several factors that we need to consider before downstream analyses, and there are any alternative approaches to tackle a specific problem such as missingness or outlier detection. As such, this guide should **not** be seen as an exhaustive benchmarking notebook, **nor** should it be taken as replacement for dedicated QC and pre-processing methods or pipelines. For more information, refer to NBIS workshops or Scilifelab courses.

Preamble

Before starting, it is crucial that we have as much information about our dataset as possible, and that we have a good idea of the analyses that we will perform afterwards. Some of the questions to bear in mind before starting:


Lab: introduction to metabolic modeling with cobrapy

To run the following notebook, retrieve it from here and include it in your local folder as `/workshop_omics_integration/session_gems/lab.ipynb`. Please use the jupyter container to run it. If you would like to further explore these techniques, we strongly suggest to have a look here.

Rui Benfeitas, Scilifelab, NBIS National Bioinformatics Infrastructure Sweden

rui.benfeitas@scilifelab.se

Abstract

In this notebook we will perform some of the basic operations in working with a genome-scale metabolic model (GEM). The vast majority of software that has been developed surrounding GEMs has been done in MATLAB, likely because this form of modeling has origins in engineering (specifically chemical engineering). Although well-suited for metabolic modeling, MATLAB is not open-source and therefore limits the accessibility of such software. Fortunately, the modeling community has implemented the MATLAB COnstrant-Based Reconstruction and Analysis (COBRA) Toolbox in Python, as **COBRApy**.

COBRApy is still relatively new and therefore lacks some of the functionality of its MATLAB counterparts, but the core utilities are available and quickly expanding. Here, we will demonstrate some of the basic functions and classes of the **COBRApy** package, which should also familiarize the user with the fundamentals of GEM structure and simulation.

Most of the commands and material covered in this tutorial can be found in the **COBRApy Documentation**, so we encourage you to reference the documentation if you encounter errors, warnings, or need further detail about something.

Contents

- 1 Global configuration object
- 2 Importing and inspecting models
- 3 Adding reactions to the model
- 4 Flux balance analysis (FBA)
- 5 Perform an in silico knockout

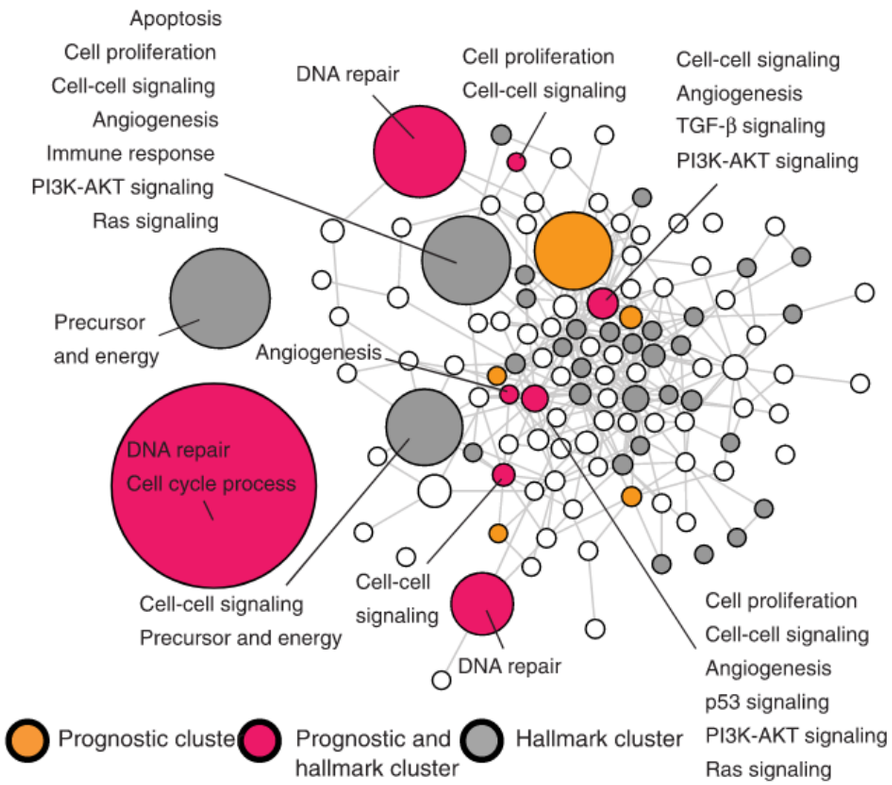
In [1]:
```
import cobra
import cobra.test
import os
```

Global configuration object

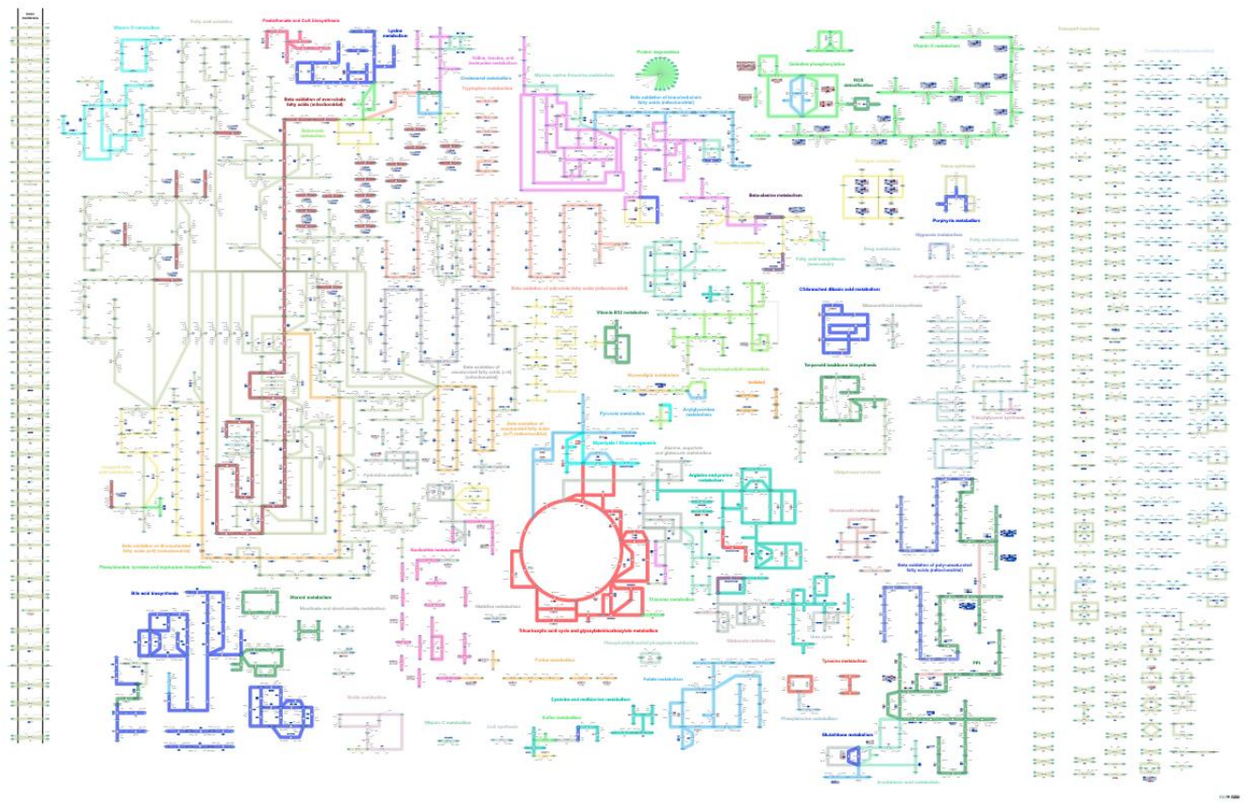
Before jumping right into things, it is always nice to see what sort of default settings are in place. **COBRApy** has organized such defaults into a **global configuration object**, which can be

Frameworks for biological network analysis in health and disease

Introduction to application of graph analysis in disease



Genome-scale metabolic modeling for data integration and simulation



Uhlen 2017
https://metabolicatlas.org/

Outline of part I: graph analysis

1. Introduction to network analysis

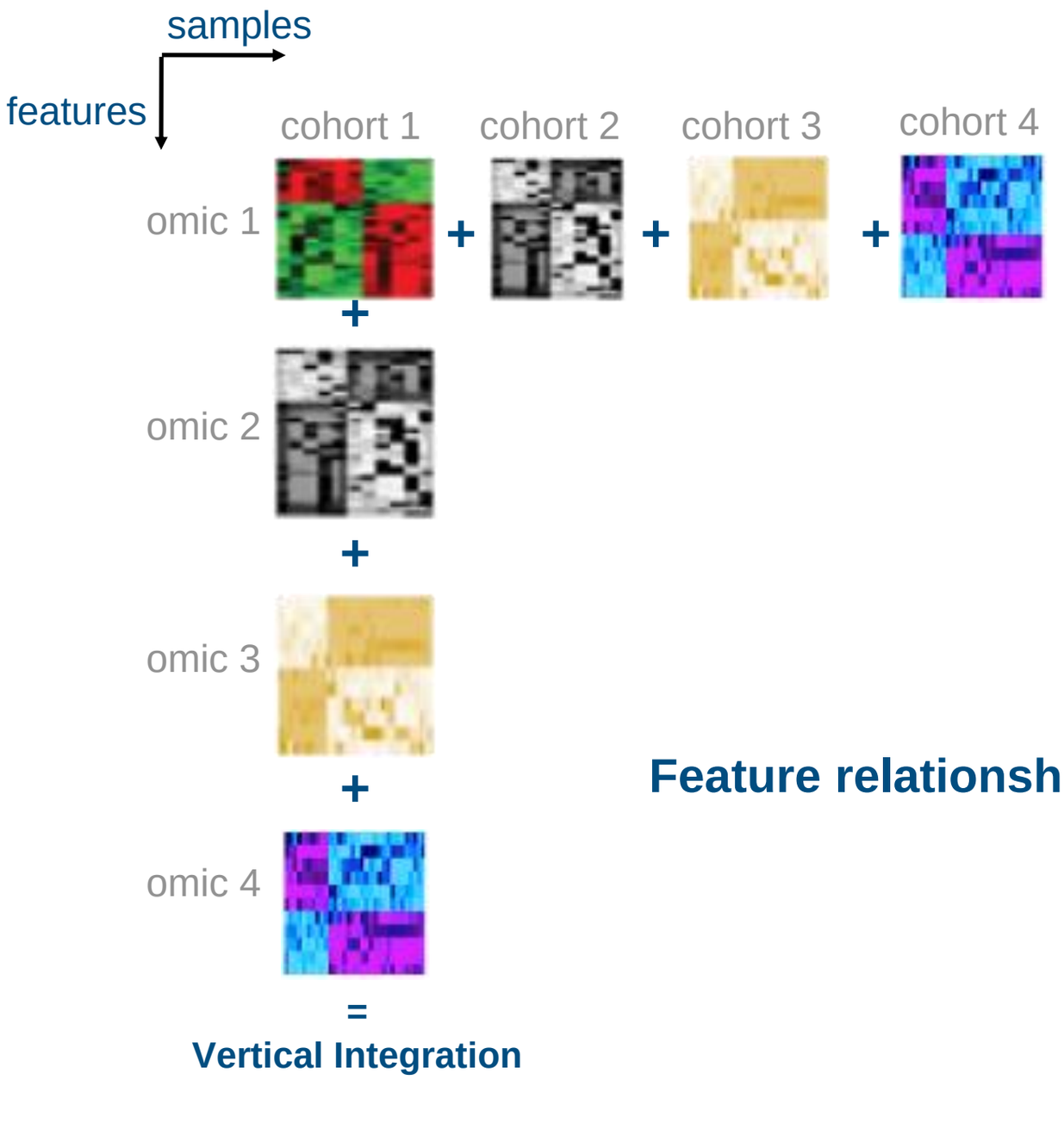
2. Network inference

3. Analysis of key properties

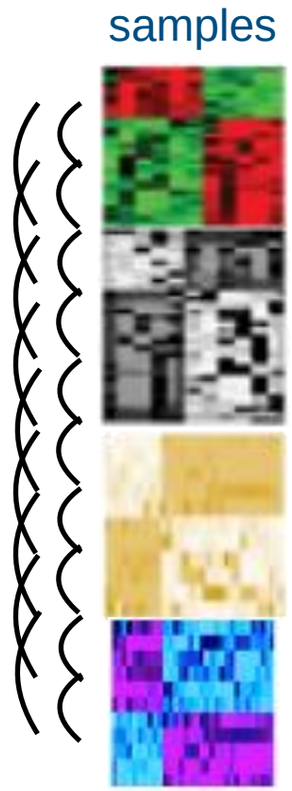
4. Communities and functional analysis

Original sources of images provided as reference and hyperlinks where applicable.

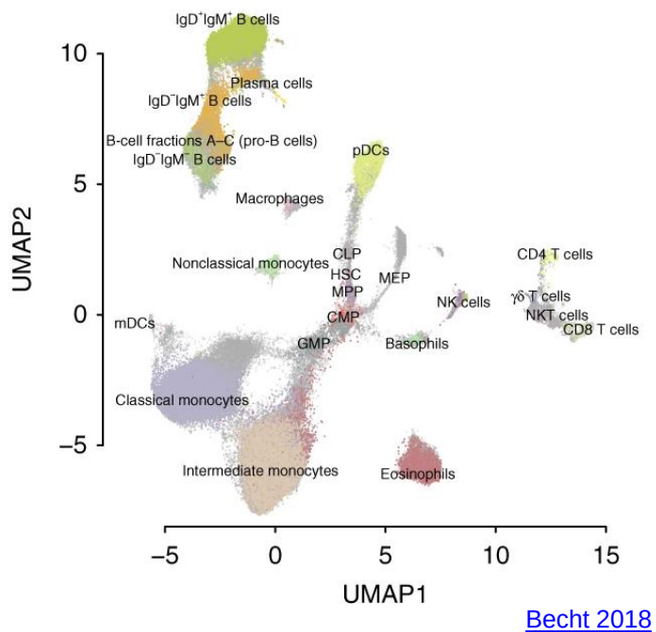
Techniques may be employed at feature and sample levels



Feature relationships



Application at sample and cell level



Becht 2018

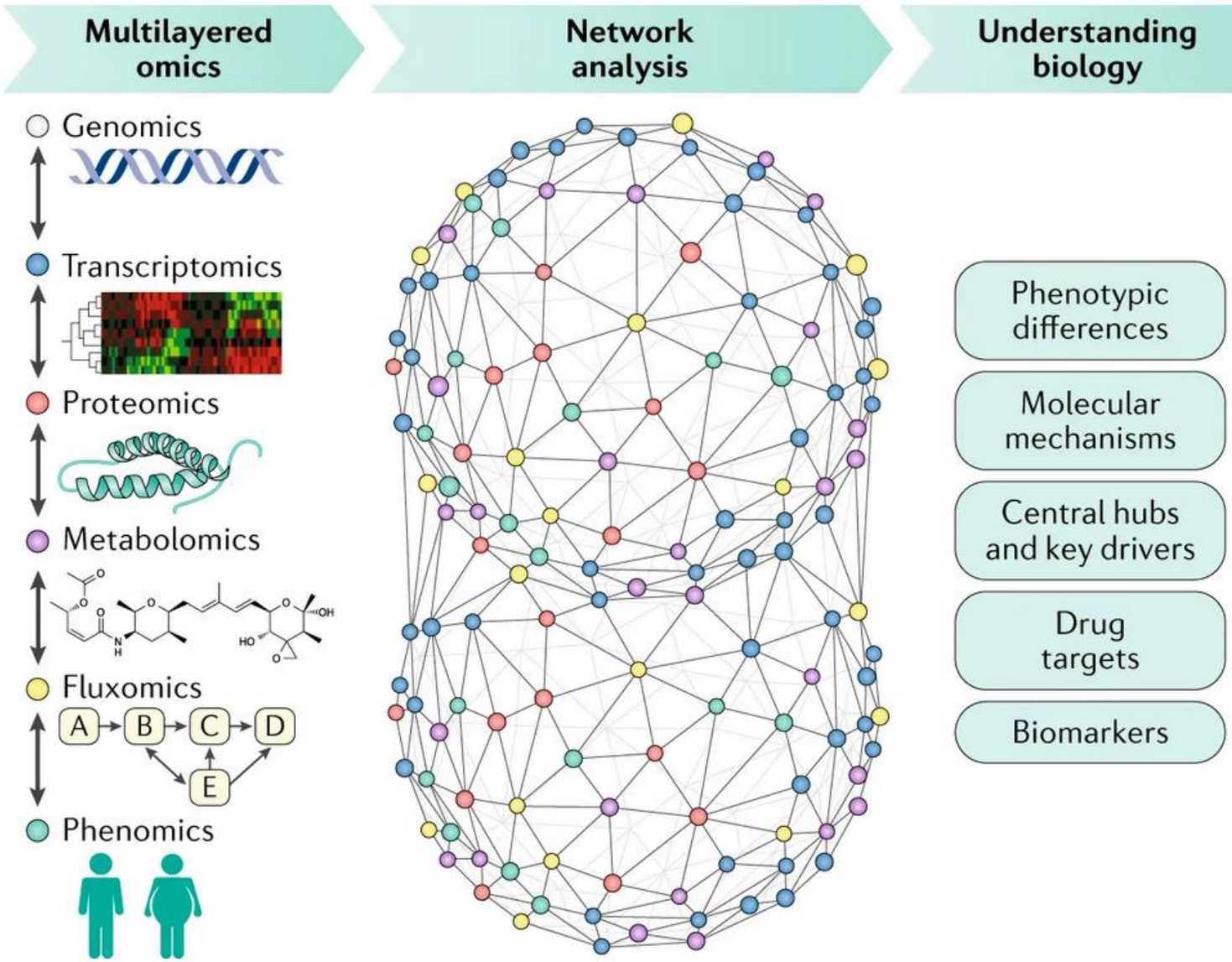
Network formalisms balance parameter number and size

	Pros	Cons
Kinetic models	Detailed Quantitative Dynamic / Steady state	Small Requires detailed parameterization
Stoichiometric GEMs	Large Semi-quantitative Steady state	Static
Topological Graphs	Comprehensive No extensive parameterisation required	Static

Size

Adapted from Hartmann et al.

Network as integrative frameworks



Systematically examine coordinated patterns

Phenotypic subgroup specificity

Feature prioritisation and target prediction

Pathway dysregulation

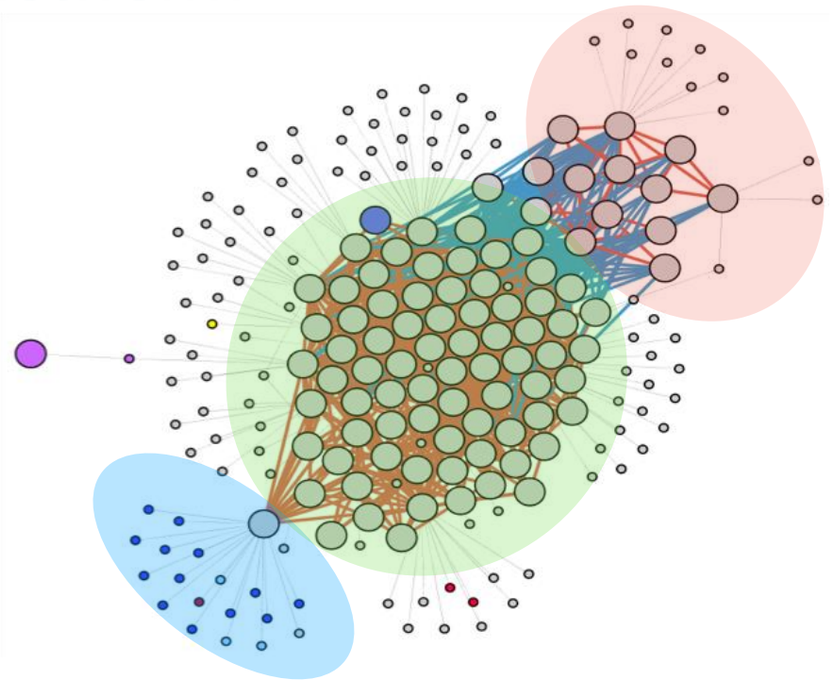
In silico simulation

Mardinoglu 2018

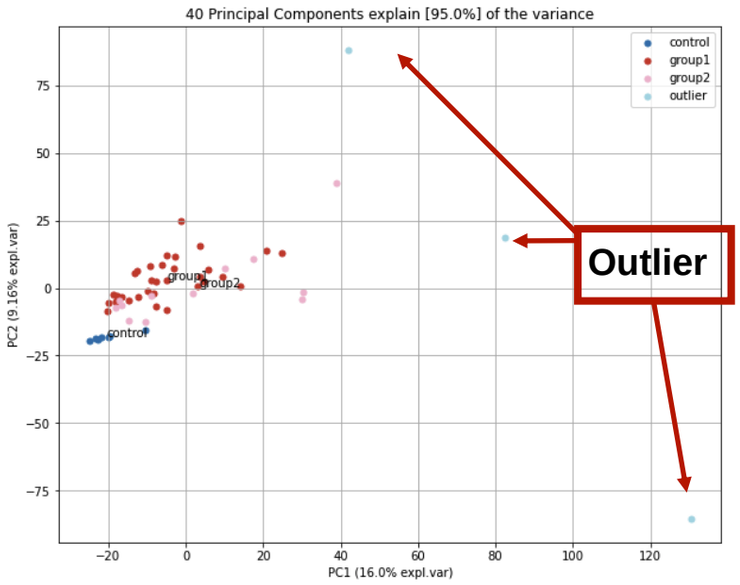
Analysis of network topology for disease characterization

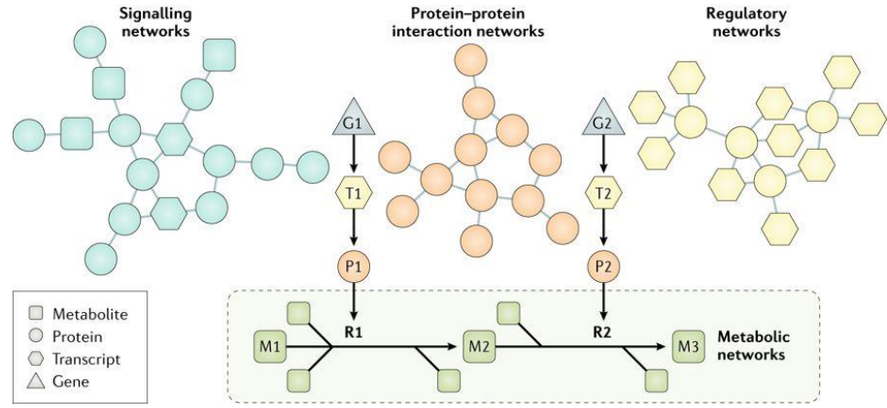
Examples:

- What features are associated with my gene of interest?
- What are the feature communities in my network?
- What possible functional relationships do they share?
- What are the key elements in a community?
- How connected is a feature to the rest of the network?

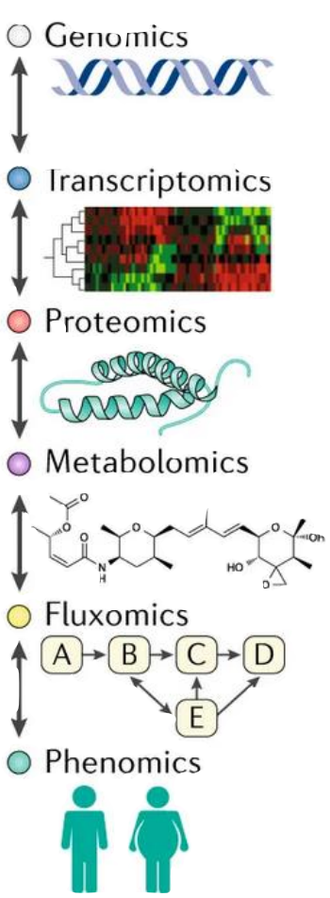


Network inference and analysis workflow





Raw → Pre-processing → **Distance calculation** → Graph analysis



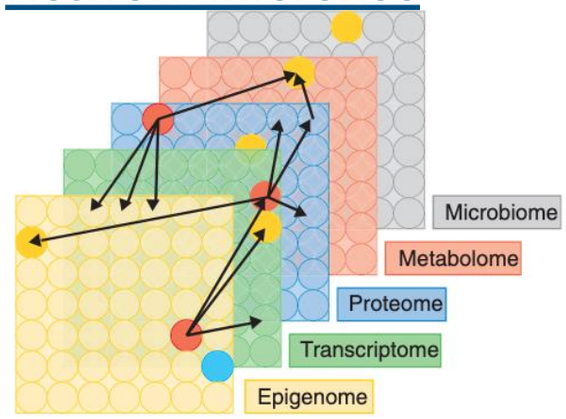
Batch effect correction

Feature selection

Anomaly detection

…

Network inference



Intra-/Interomic

Piening 2018
Hasin 2017

Mardinoglu 2018

Different approaches for network inference

1. Feature association (e.g. correlation)
2. Genome-scale metabolic models

No prior graph structure

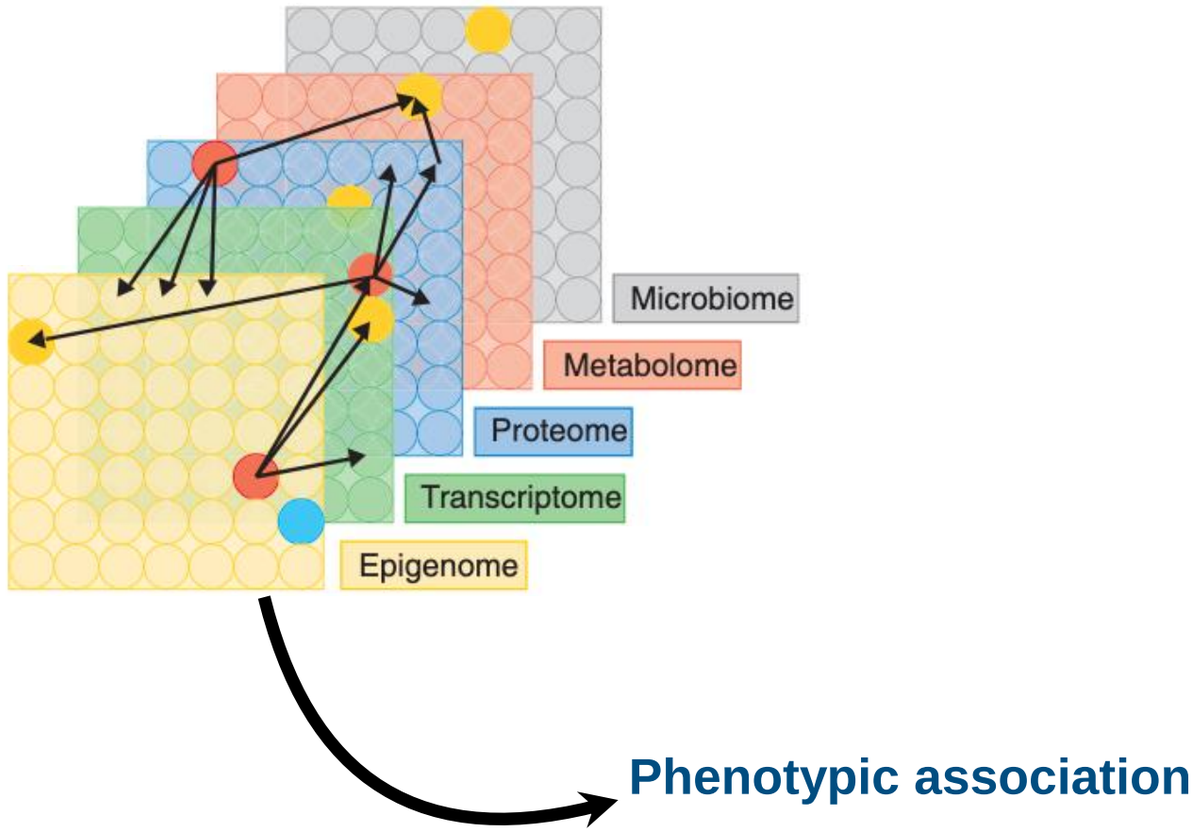
Based on prior knowledge

Association analysis

Common approach: compute correlations between features

- Spearman
- Pearson
- WGCNA

Extend known associations



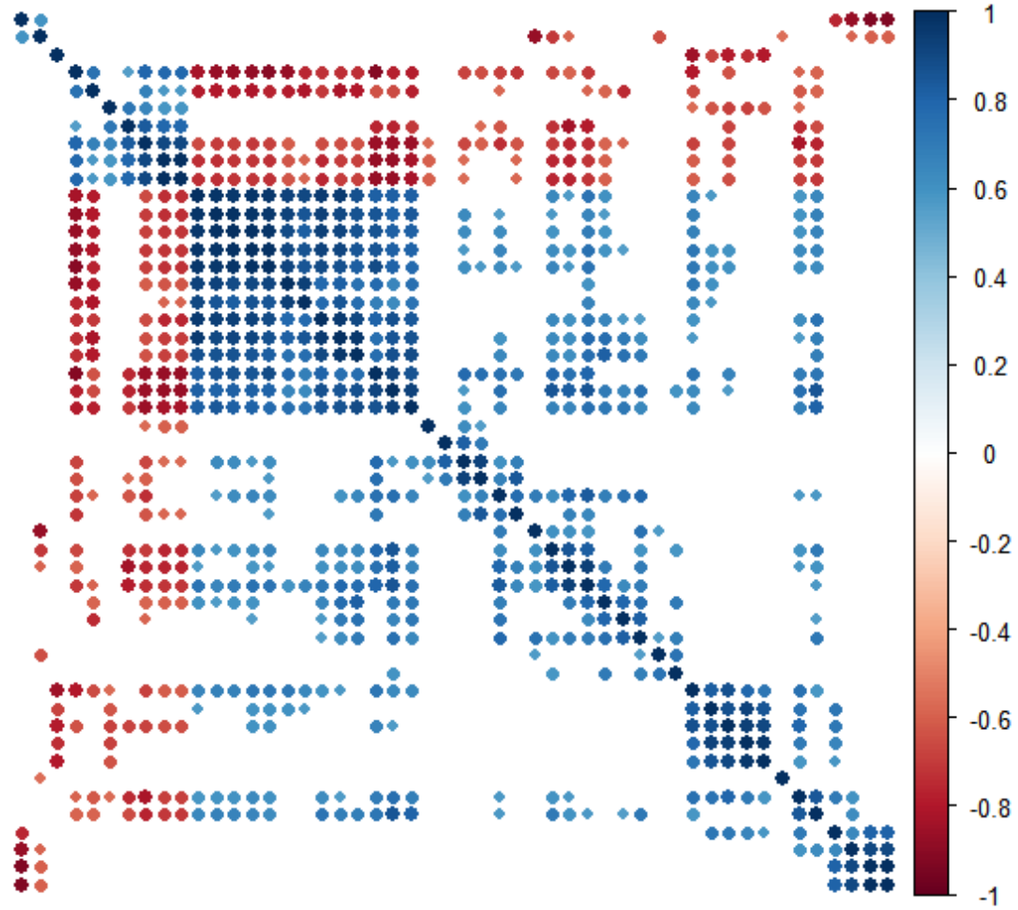
Adapted from Piening 2018

Association analysis

Correlations are easy to interpret

Unweighted vs weighted ($-1 \leq \rho \leq 1$)

Prone to type I errors



Filtering

- FDR / Bonferroni
- Correlation coefficient cutoff

Need adjustment to possible confounding factors

Wu 2019
Uhlen 2017

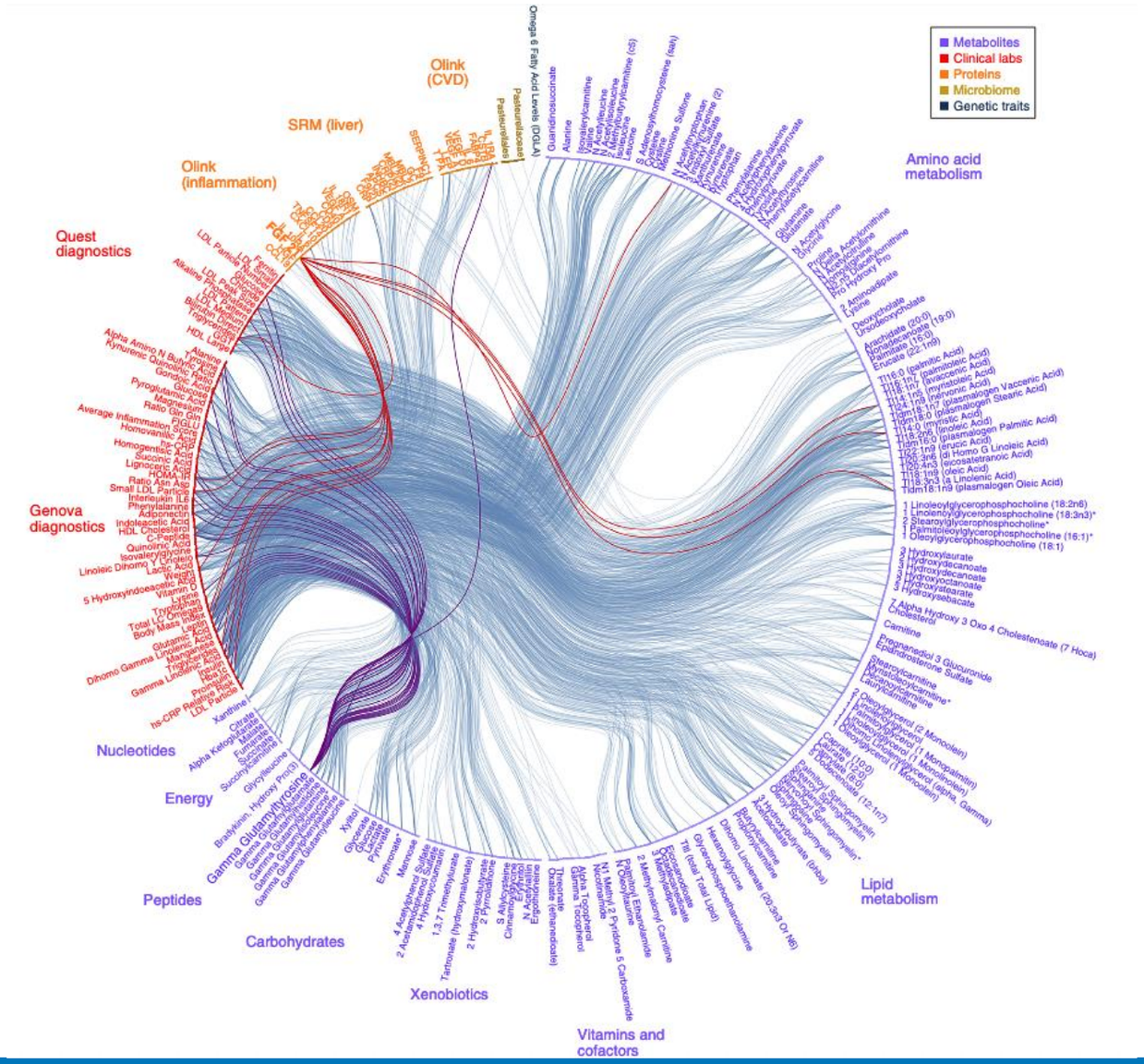
Association analysis

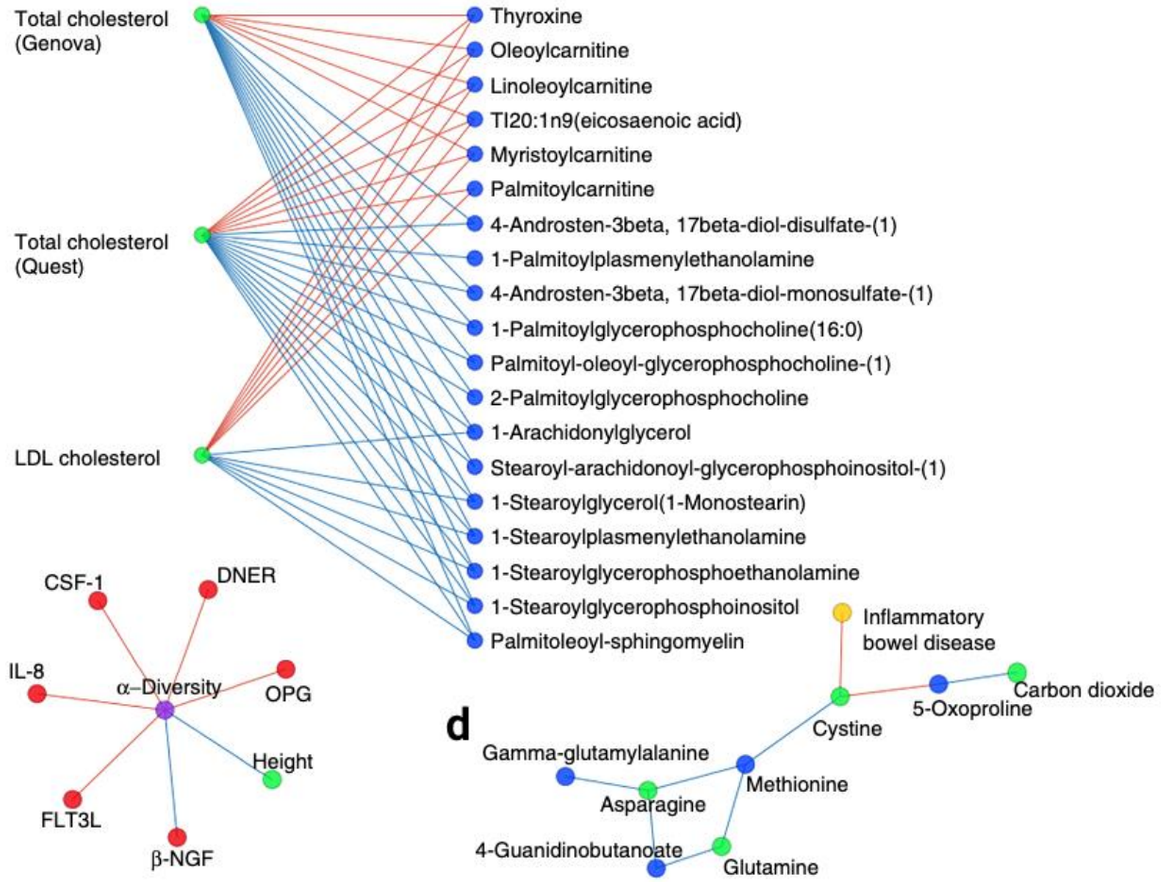
Adjusting for confounding factors

Below:

- gender and age are known confounding factors

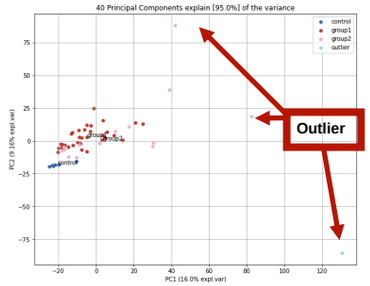
- feature regression on confounding factors, followed by correlation on the residuals of each model

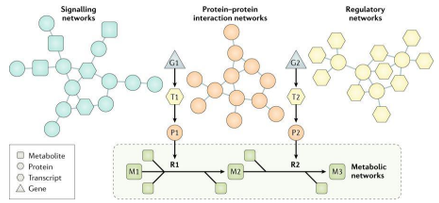




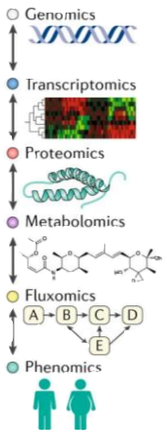
Price 2016

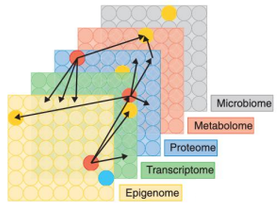
Network inference and analysis workflow





Raw → Pre-processing → Distance → **Graph analysis**





- **Computing global properties (e.g. centrality, …)**
- **Comparison against null models**
- **Identification of key nodes**
- Module assessment
- Phenotypic interrogation
- Experimental validation of key drivers

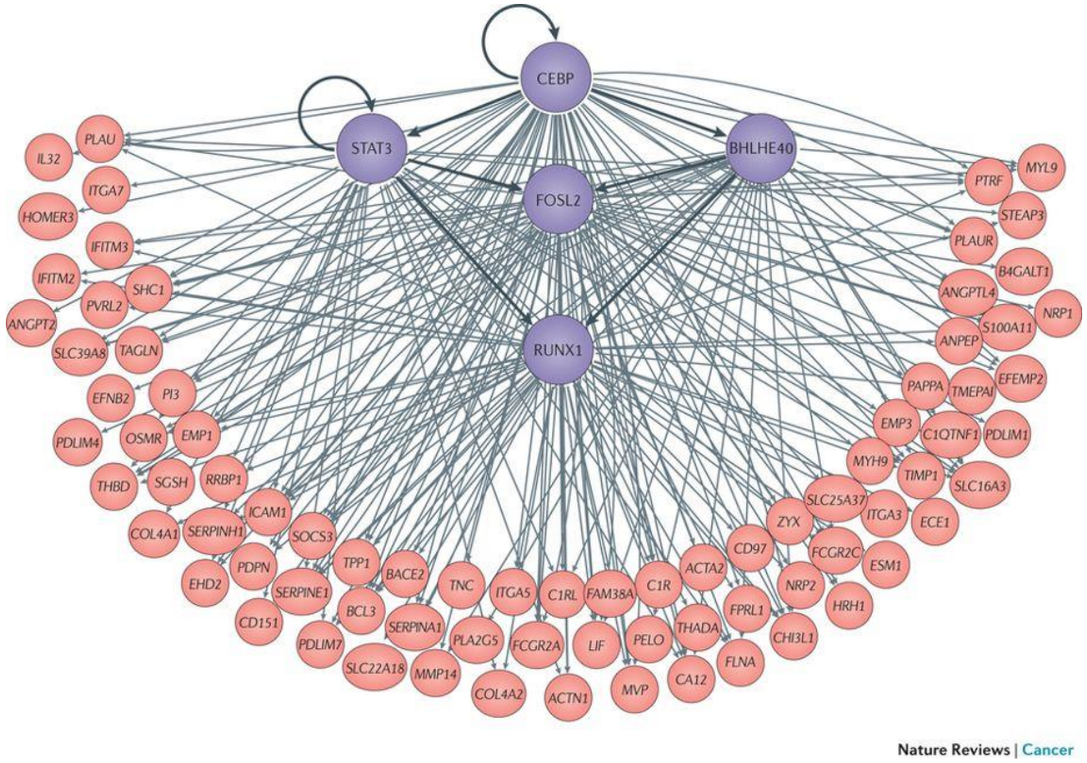
Why centrality is important

Indicate the most central nodes in a network

Why look at the central nodes?

Hubs

Example: Transcription Factor Master Regulators



Tyagarajan 2014

Centrality metrics

Indicate the most central nodes in a network

Hypothesis: central nodes are important in the network

There are many different measures of centrality:

- Eccentricity

- Degree

- Betweenness

- Closeness

- Eigenvector

- PageRank

- Katz

- Percolation

- Cross-clique

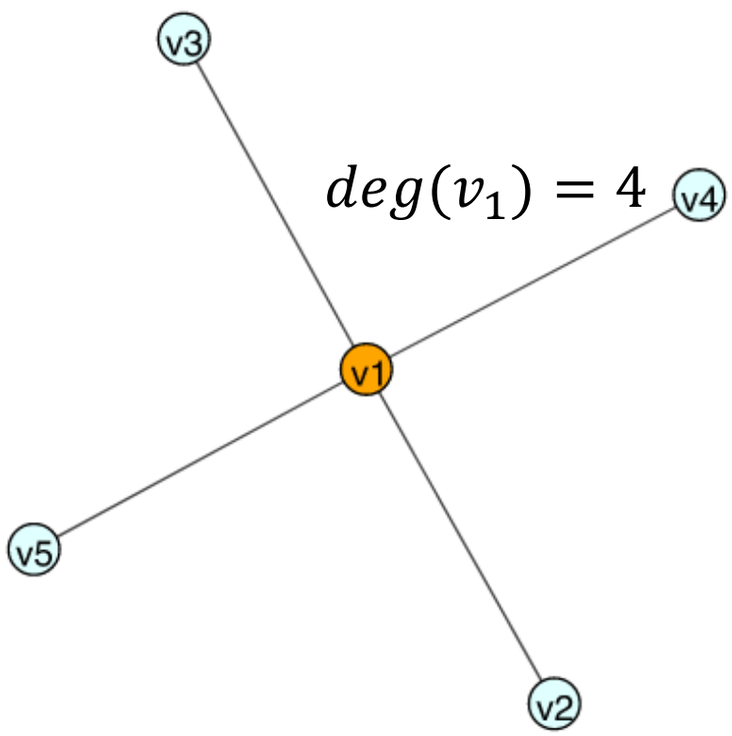
…

Degree centrality

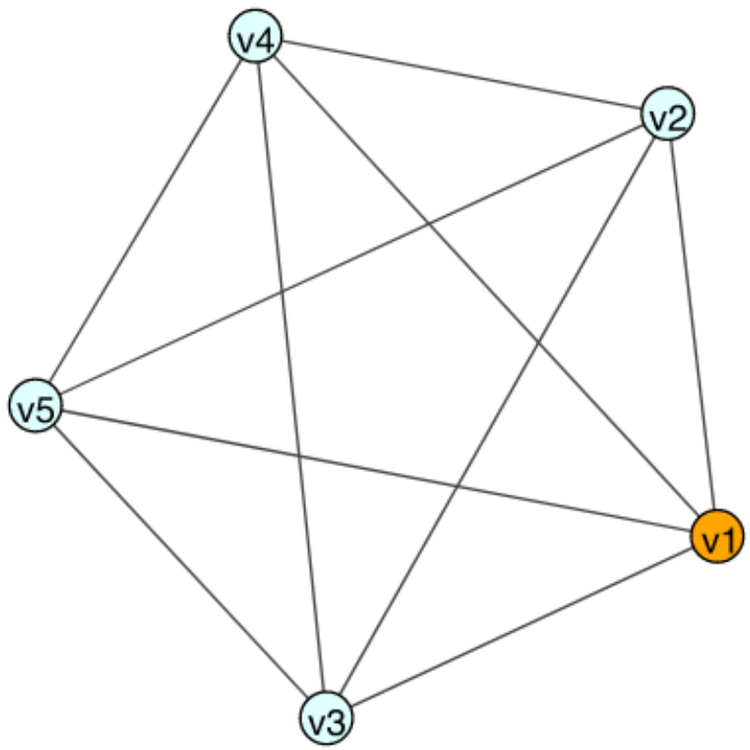
Degree indicates the number of connections with a node

$$d(v) = |N(i)|$$

where *N(i)* is the number of 1st neighbours of a node.



$deg(v_1) = 4$

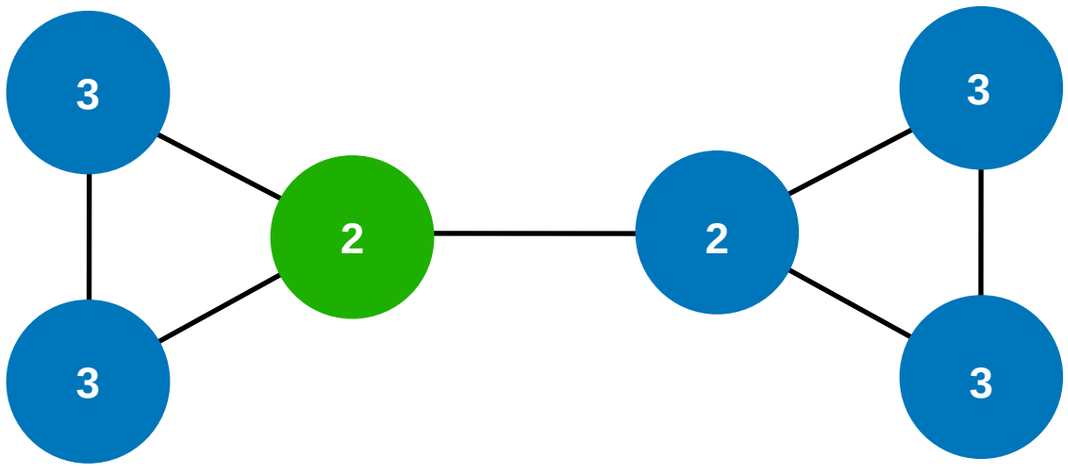


$deg(v_1) = 4$

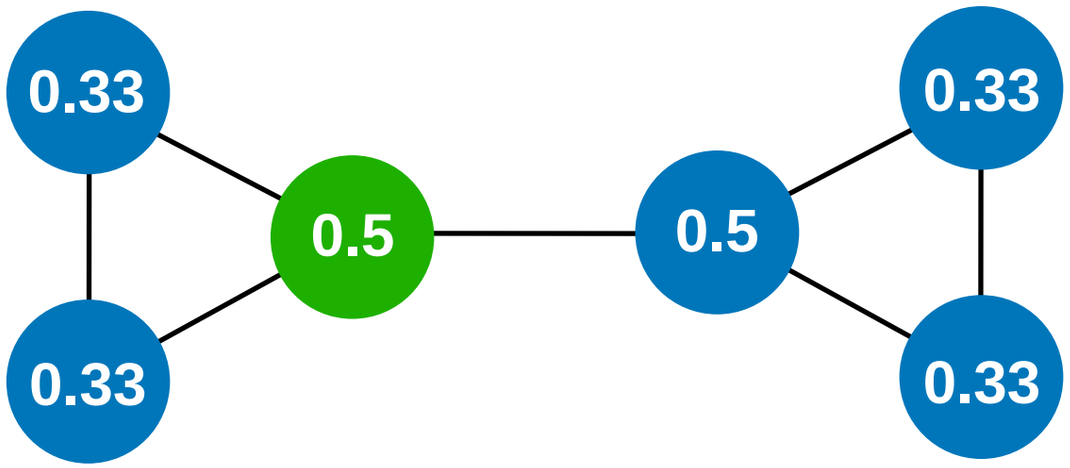
Eccentricity centrality

Eccentricity considers the maximum shortest path passing through a node

[max shortest path]



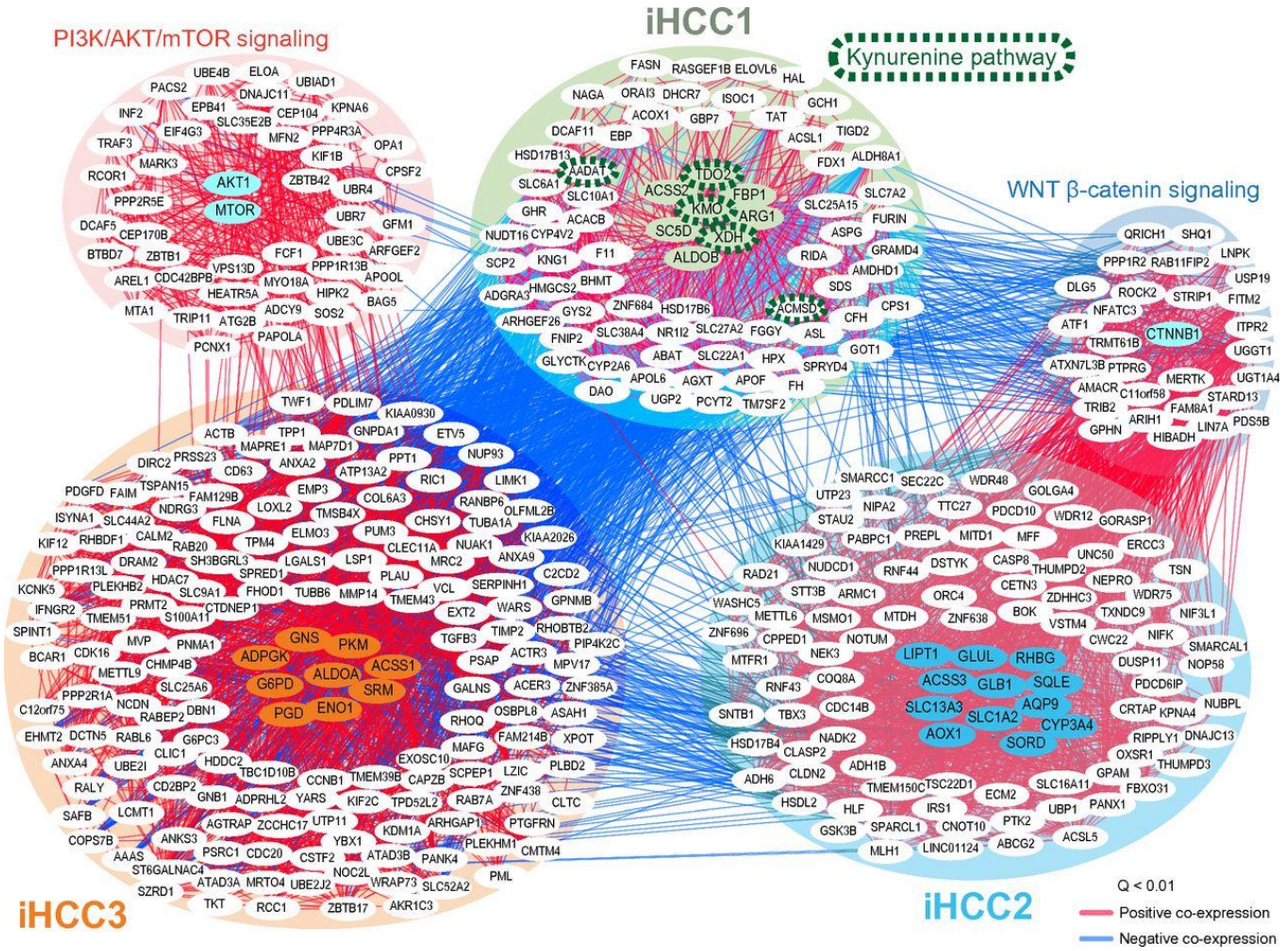
$$\frac{1}{[max\ shortest\ path]}$$

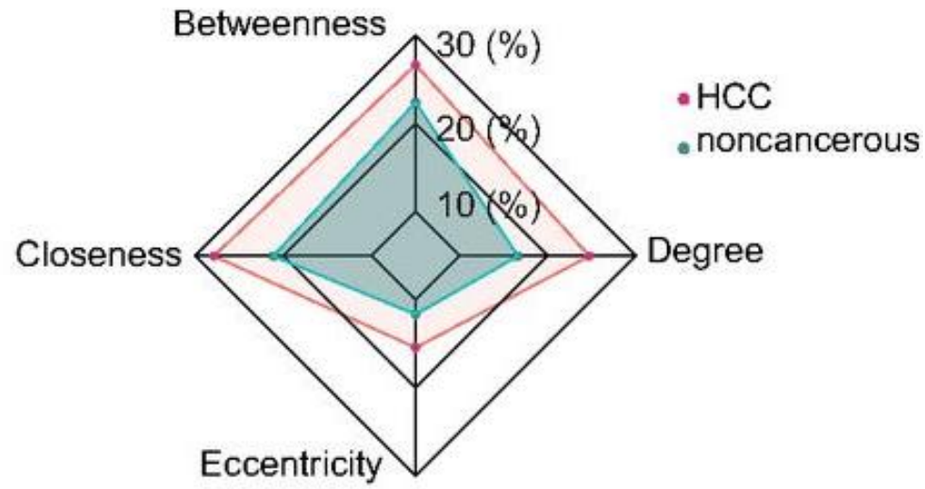


Hage 1995

Centrality: good practices

1. Compute multiple metrics and understand their differences
2. Find nodes with largest ranked centrality

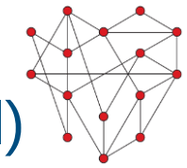
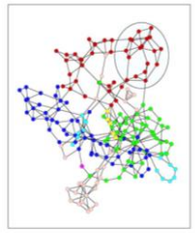




3. Compute node **influence,** modifications of centrality
Measure **information transmission** rather than *connectiveness*

Bidkhori 2018

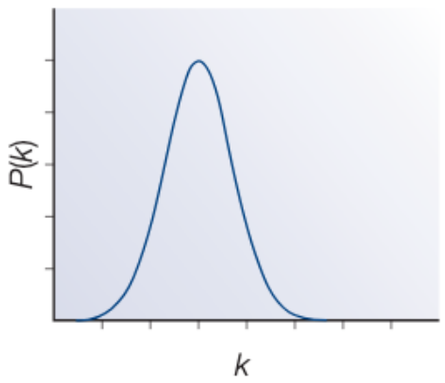
What distinguishes biological networks from random?

	Random network (e.g. Erdös-Rényi model) 		**Metabolic network** (hierarchical organization) 
Node number	N	=	N
Edge number	E	=	E
Density	D	=	D
Average shortest path	L	<	$\boldsymbol{L}$

Node failure easily propagates

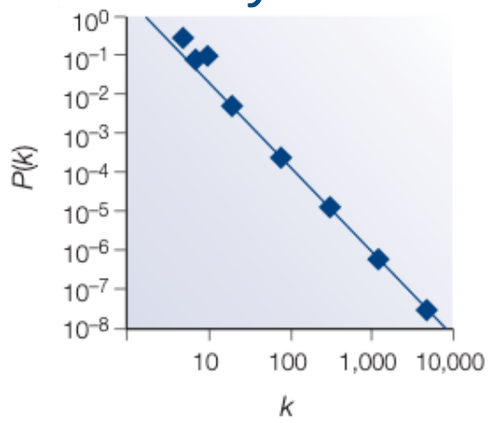
Degree distribution

no highly connected nodes



Most nodes have ~ <*k*>

many with low degrees
few very central

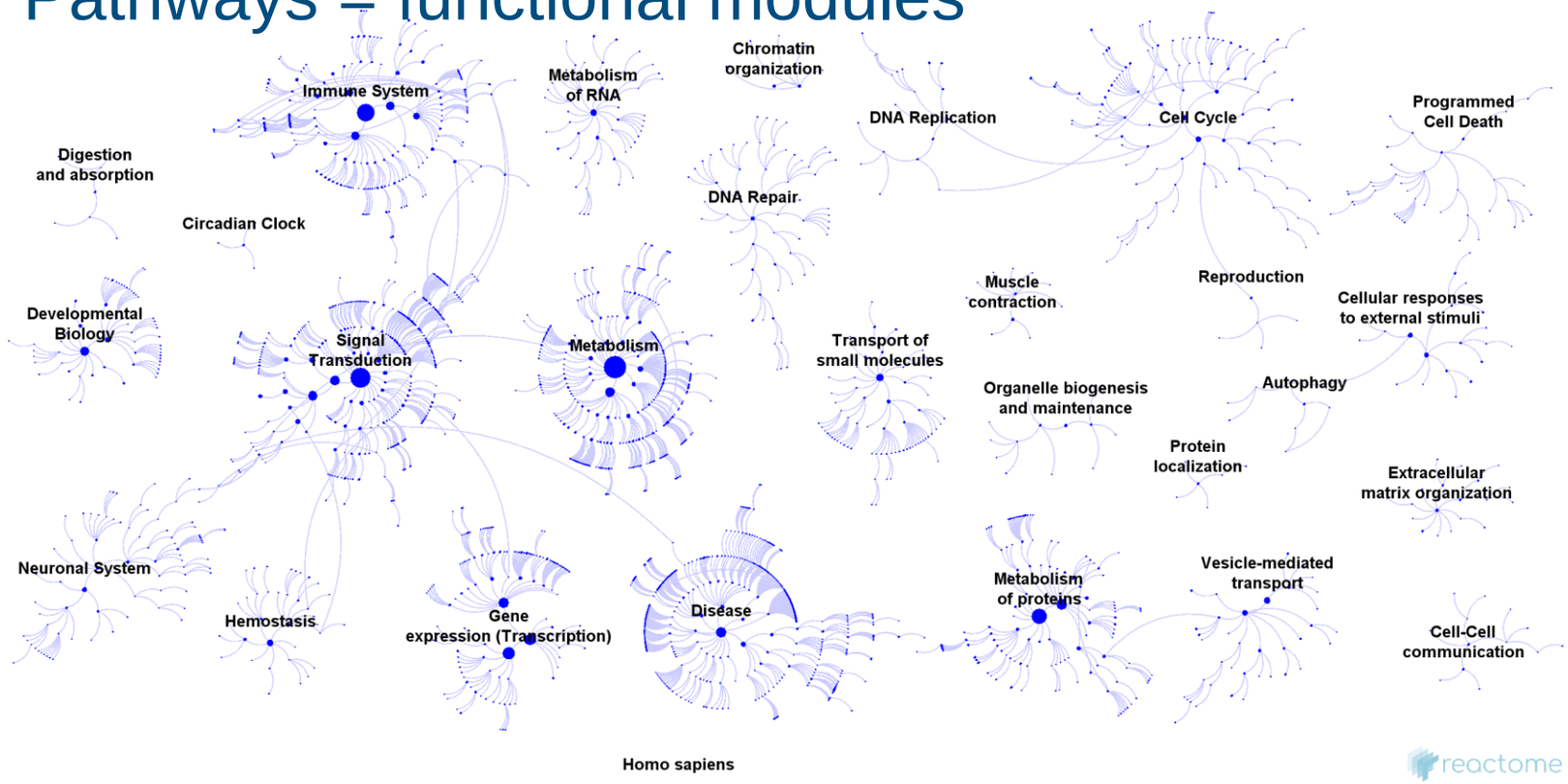


Barabasi 2004
Jeong 2000
Ravasz 2002

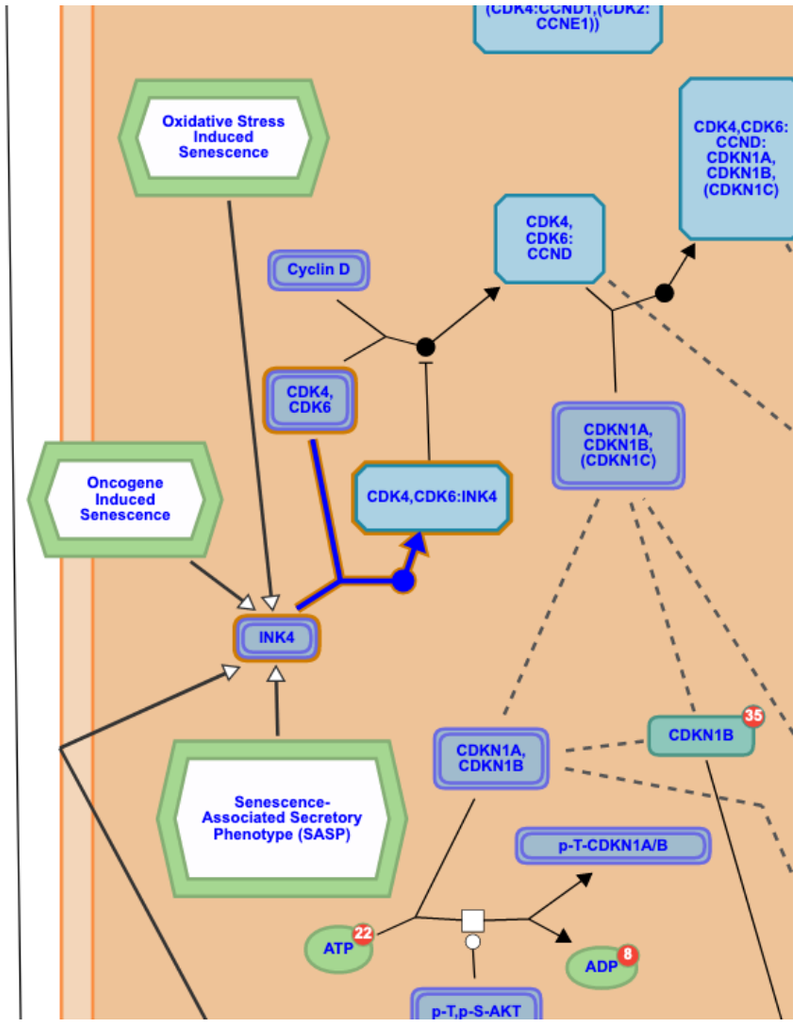
What are modules?

Modules (=communities / clusters) are physically or functionally associated nodes that work together to achieve a given function

Pathways = functional modules



Protein complexes = physical modules



reactome.org

What are modules?

In addition to physical or functional modules, one may identify other types of modules

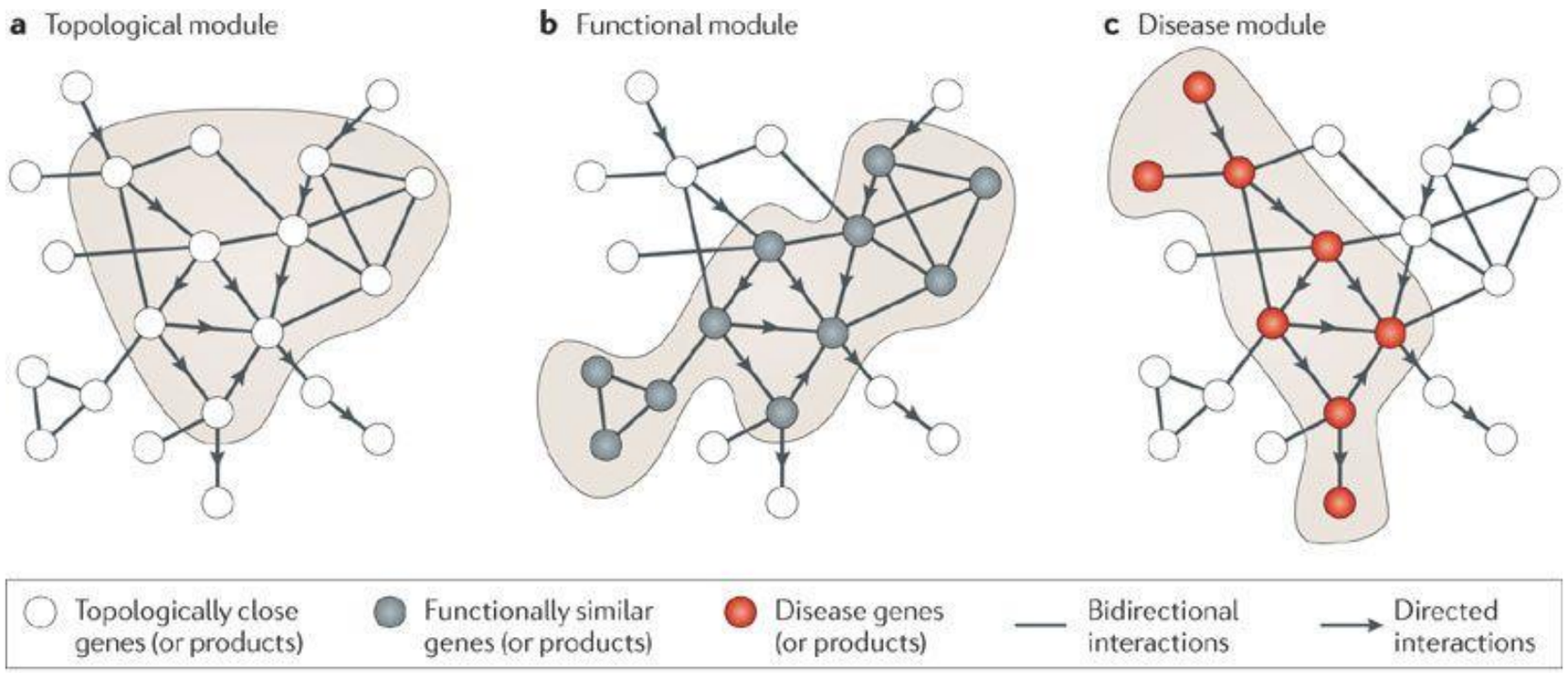
Topological: derived from their high within-module degree

Disease/drug: highly interconnected nodes associated with a disease response or drug

Subgroup: highly interconnected nodes associated with a sample subgroup (e.g. cancer subtype)

Tissue-, cell-type-specific: highly interconnected nodes associated with a specific tissue or cell type

Highly interlinked local regions of a network



Barabasi 2011

Module detection: Louvain algorithm

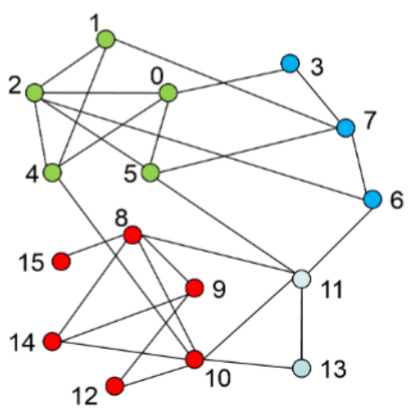
Phase 1: greedy modularity optimisation

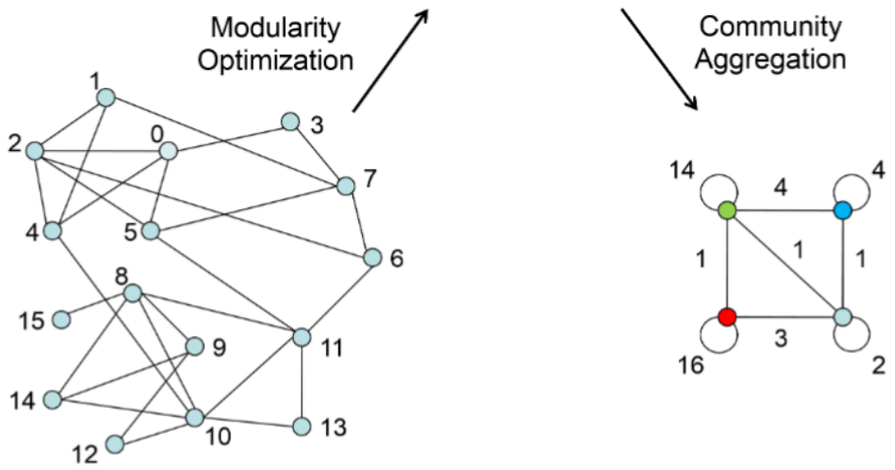
1. Start with 1n/community
2. Compute Q by moving *i* to the community of *j*
3. If ΔQ>1, node is placed in community
4. Repeat 1-3 until no improvement is found. Ties solved arbitrarily

Phase 2: coarse grained community aggregation

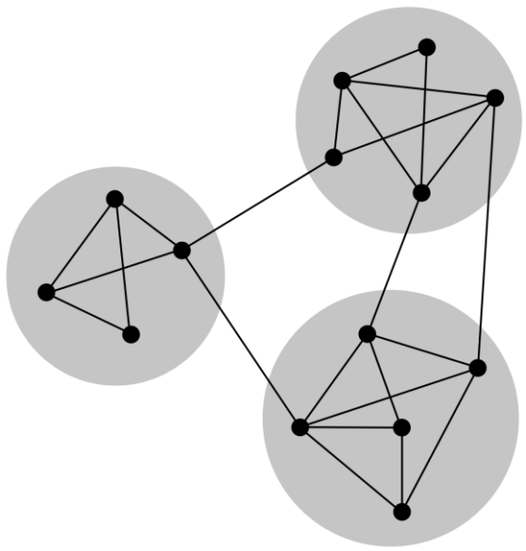
5. Link nodes in a community into single node.
6. Self loops show intra-community associations
7. Inter-community weights kept

Second pass: repeat phase 1 on the new network





Other methods:
Walktrap
Label propagation
…
(benchmarking)

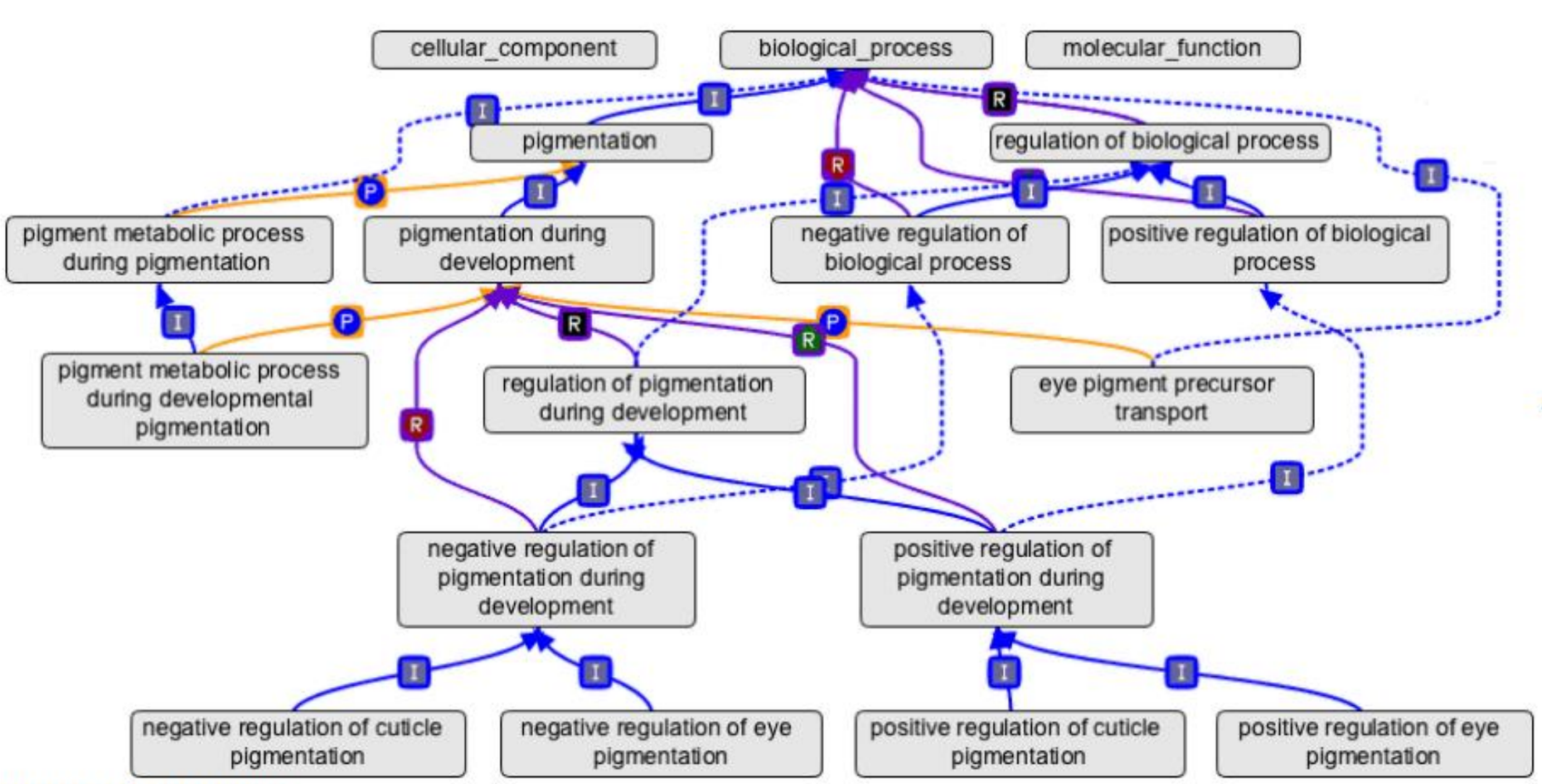


Campigotto 2014
Traag 2019

Community characterisation

Hypothesis: community-associated features show coordinated changes related with common biological processes or phenotypes

Enrichment analysis and Phenotypic association

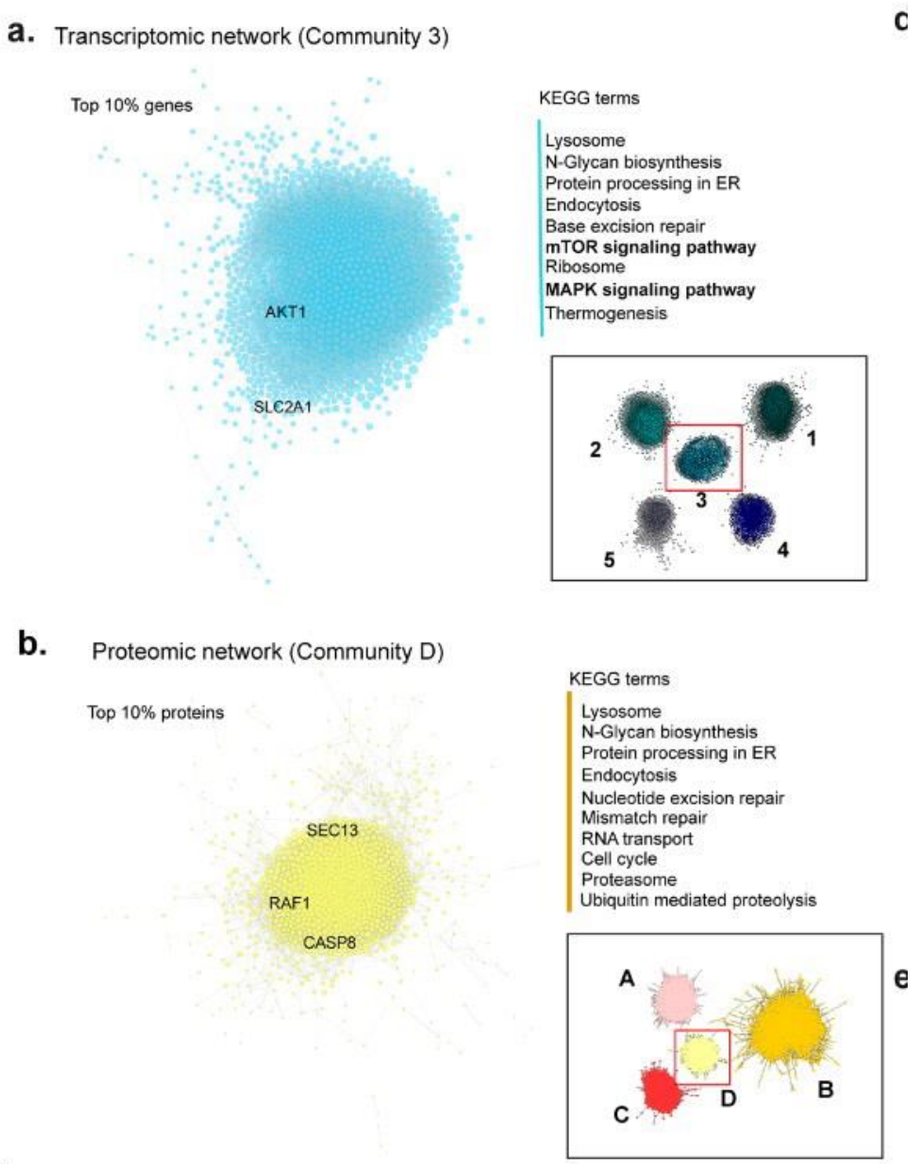


http://geneontology.org/

Multi-tissue network analysis of proteo-transcriptomics data in response to Covid-19 infection

Graph analysis of proteo-transcriptomic data

Centrality and Community characterization



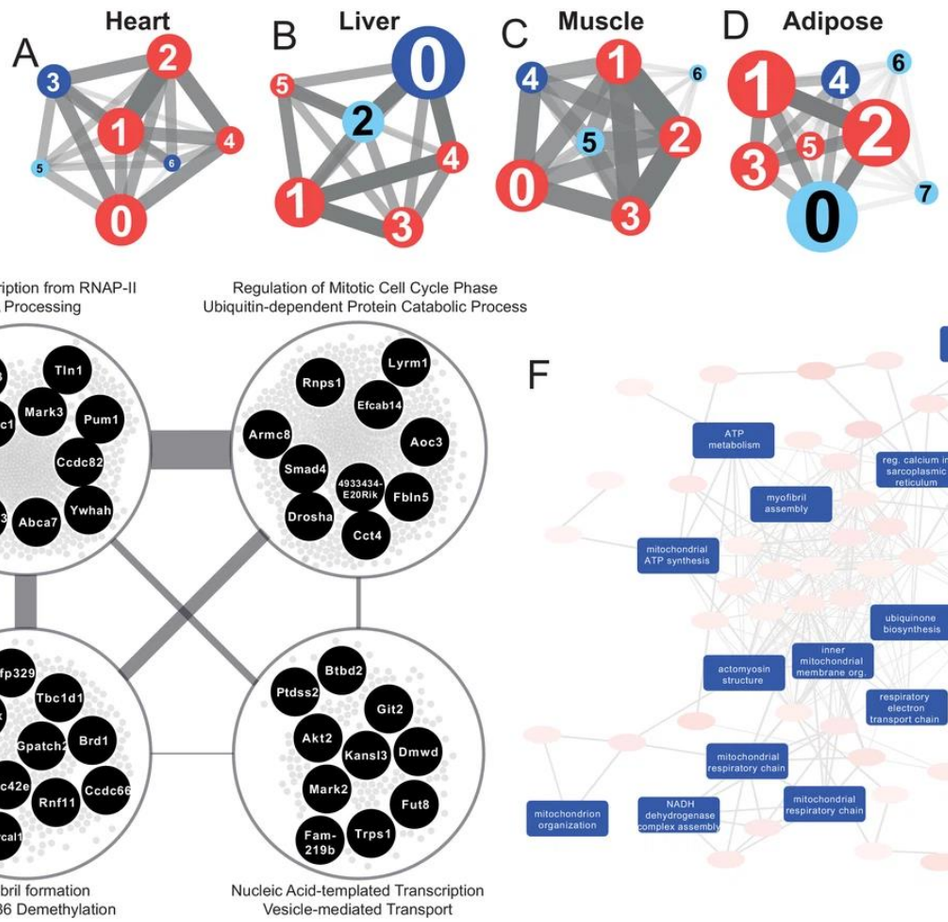
Appelberg 2020

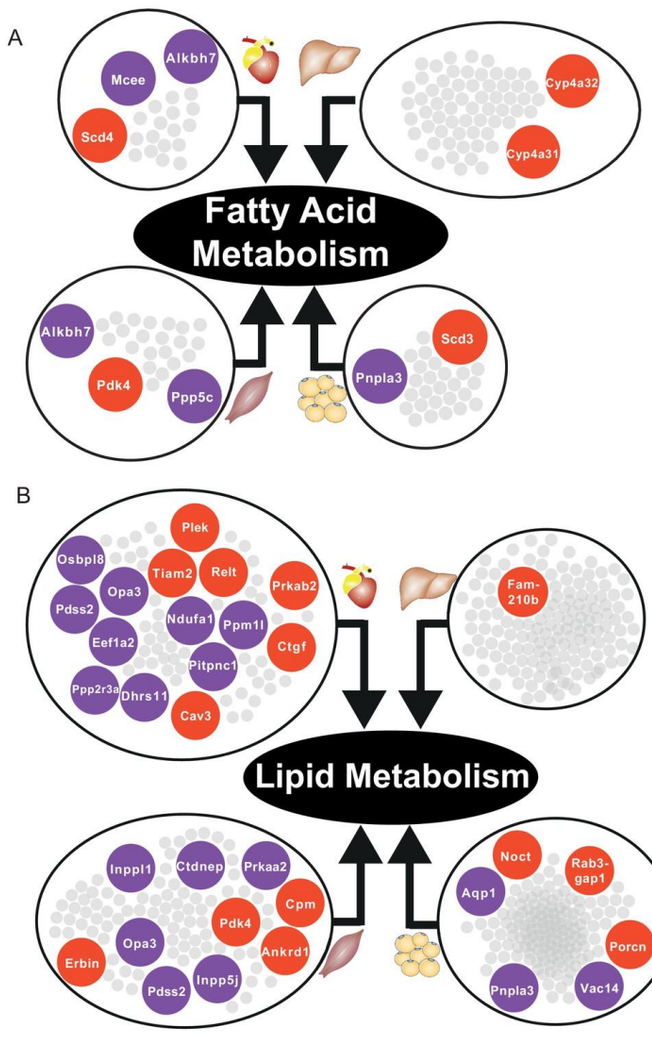
Multi-tissue network analysis of RNAseq data in CVD

Graph analysis (centrality, modularity, cluster coefficients, …)

Community identification and characterization

Tissue-specific and shared responses to myocardial infraction



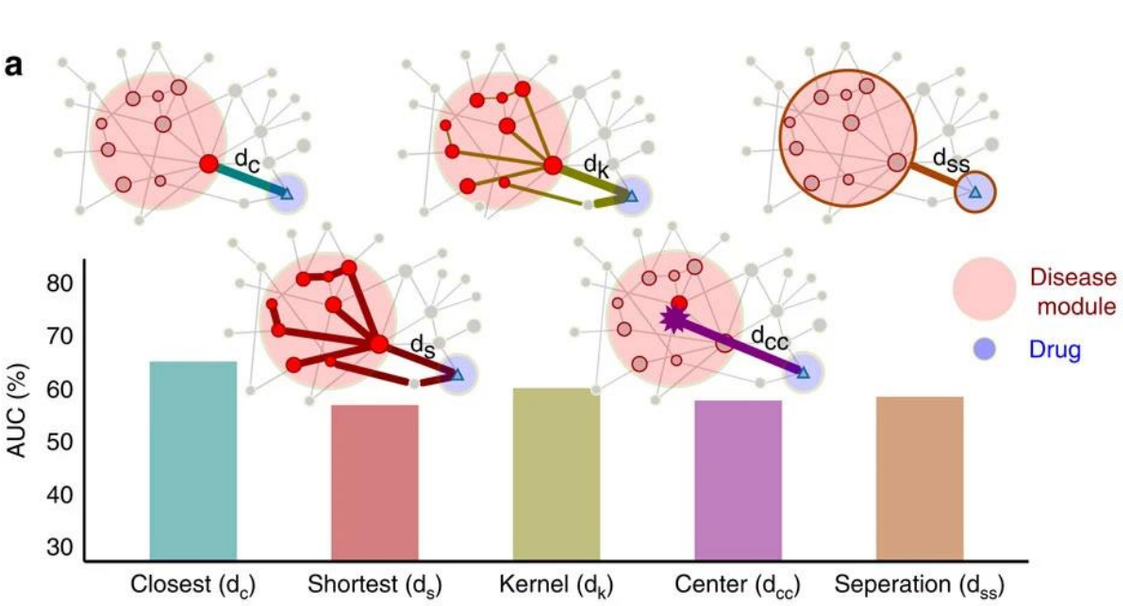


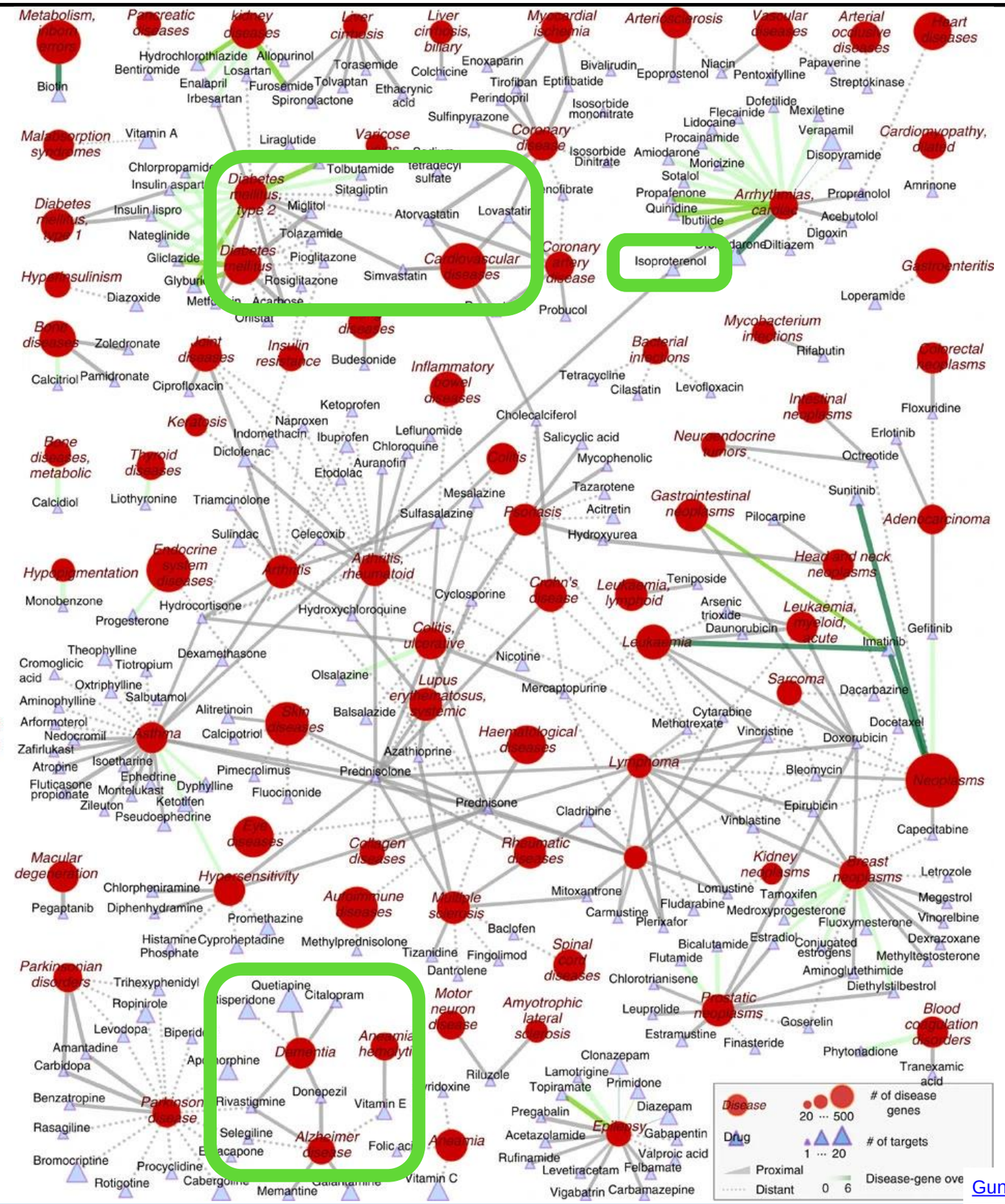
Arif 2021

Modules / centrality applied to drug repositioning

Centrality used to prioritise drug associations

Identified suitable candidates for repositioning





Guney 2016

Recap: graph analysis

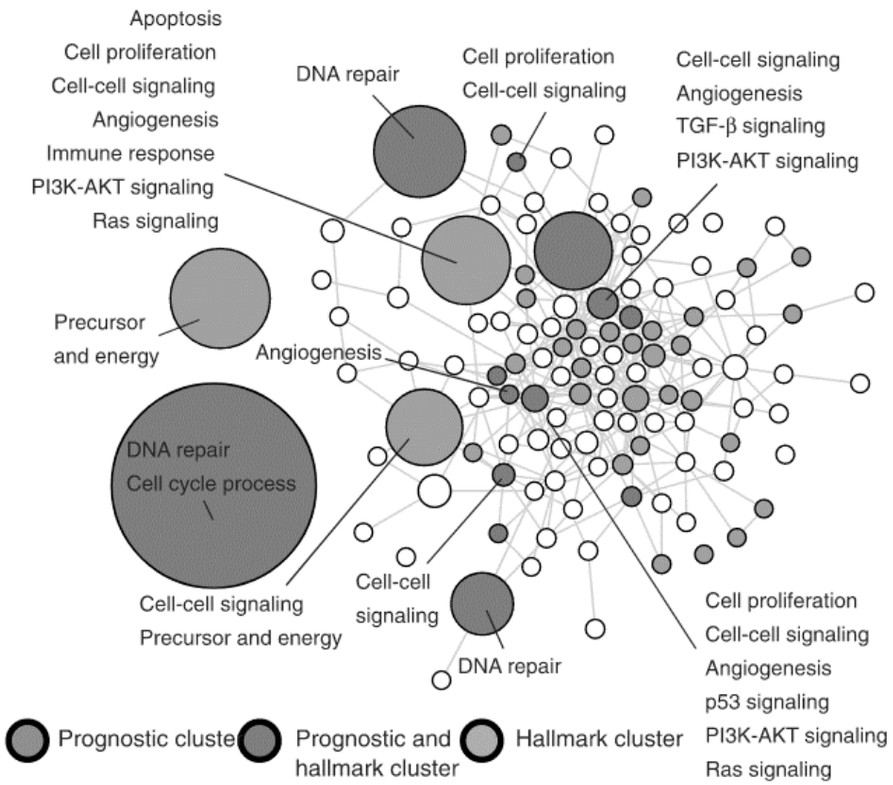
Graph analysis can be used to examine coordinated patterns of variation at feature or sample levels

Technical biases need to be tackled, and graphs should be compared against matching null models

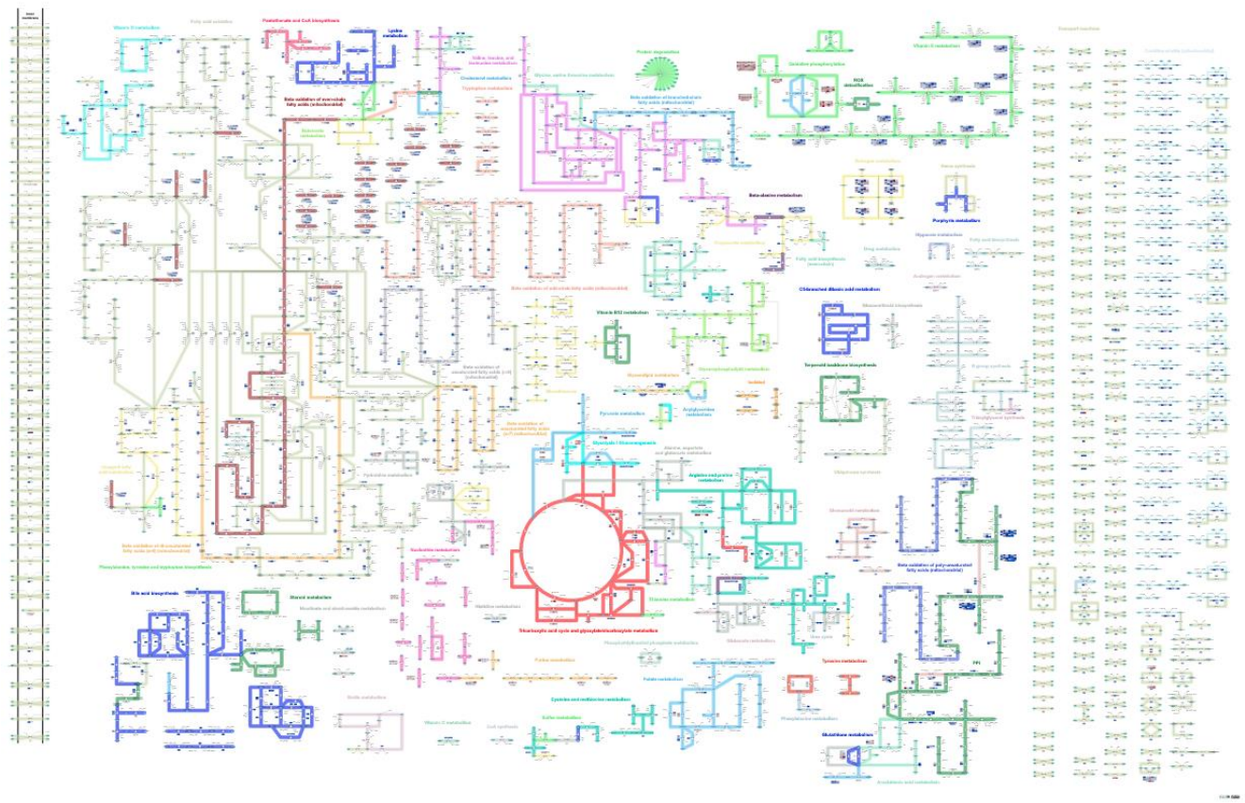
Centrality and module assessment may be used to characterise and prioritise features or their groups

Frameworks for biological network analysis in health and disease

Introduction to application of graph analysis in disease



Genome-scale metabolic modeling for data integration and simulation



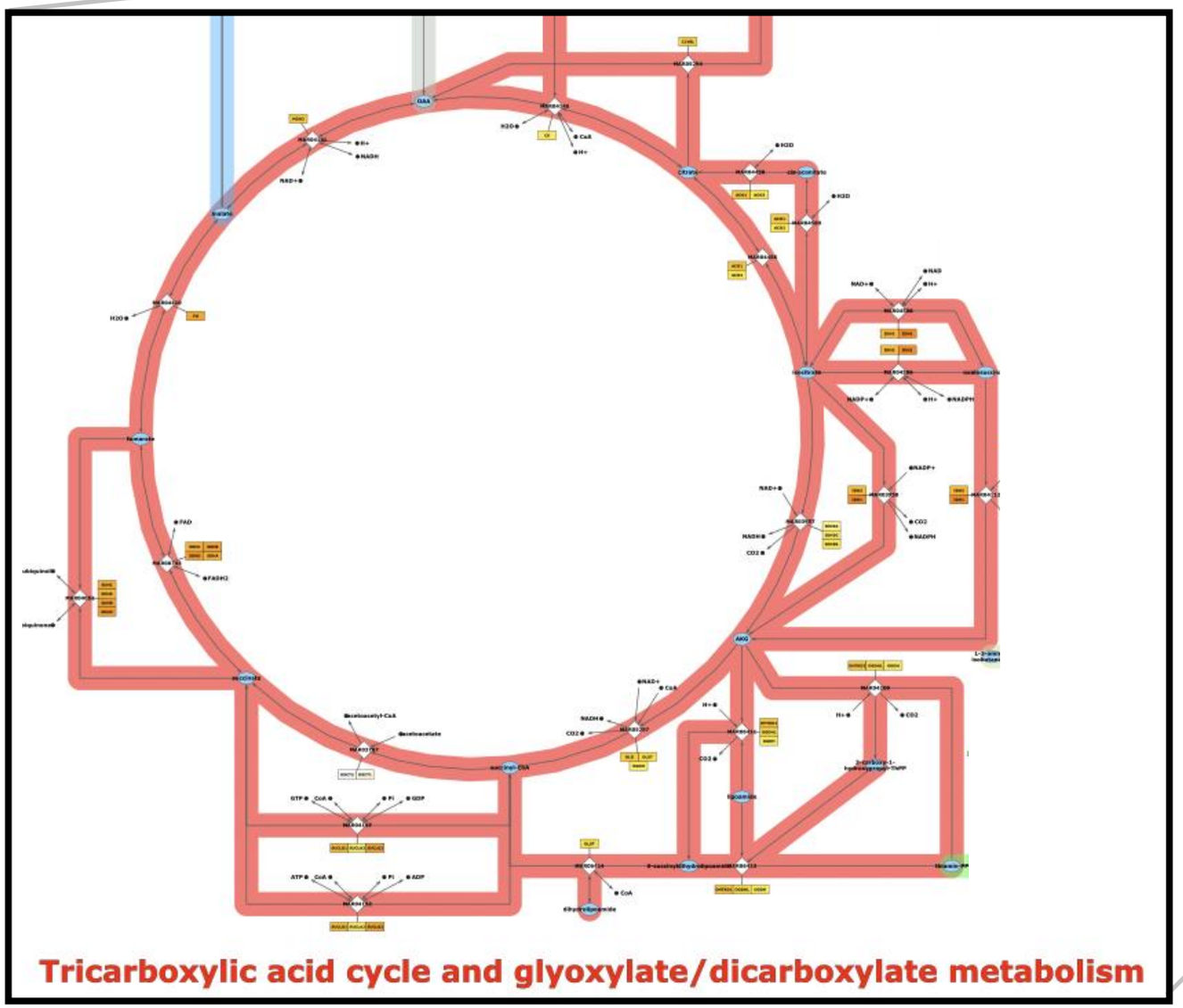
Uhlen 2017
https://metabolicatlas.org/

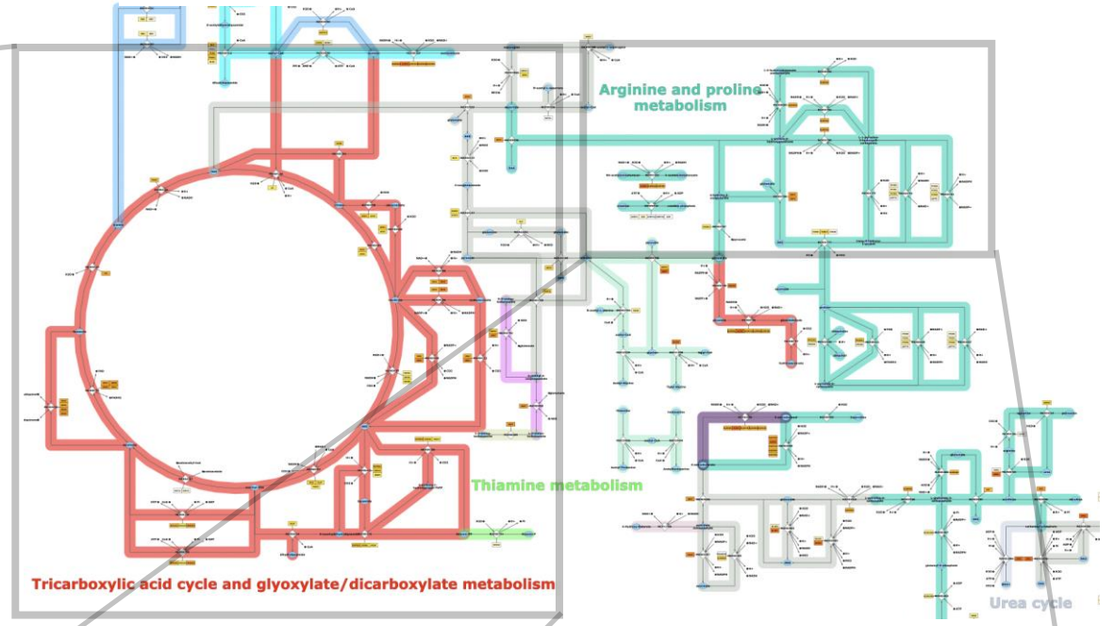
Outline of part II: genome-scale metabolic models

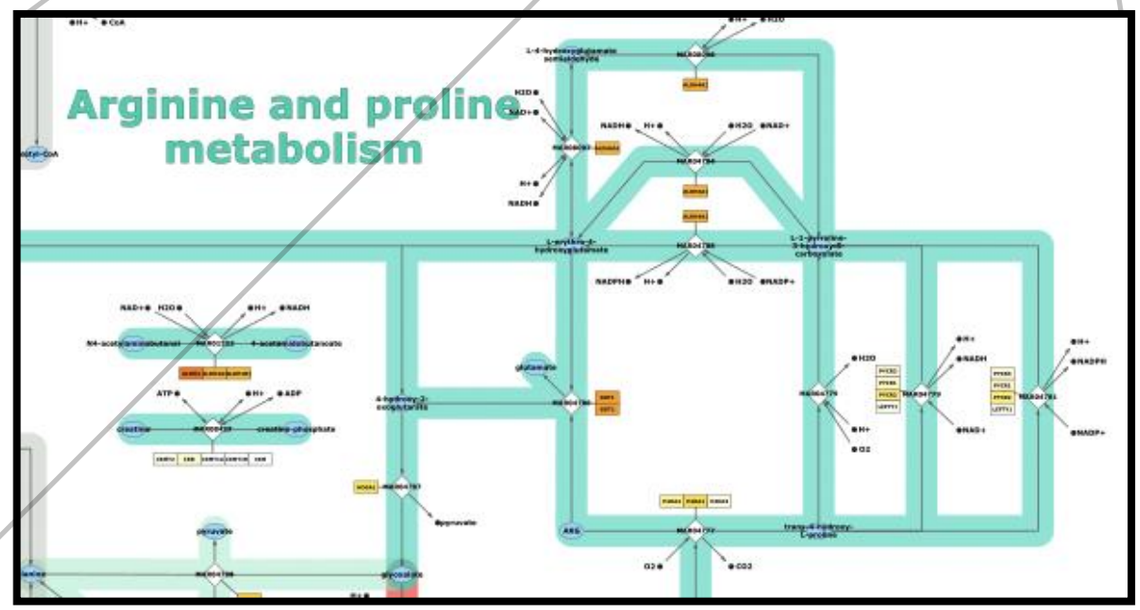
1. Challenge in systematically characterising metabolic disruption
2. Introduction to metabolic modelling and FBA
3. Downstream analysis and combination with topology analysis

Characterizing metabolic disruption in disease is hindered by biological complexity

Human mitochondrial reactions



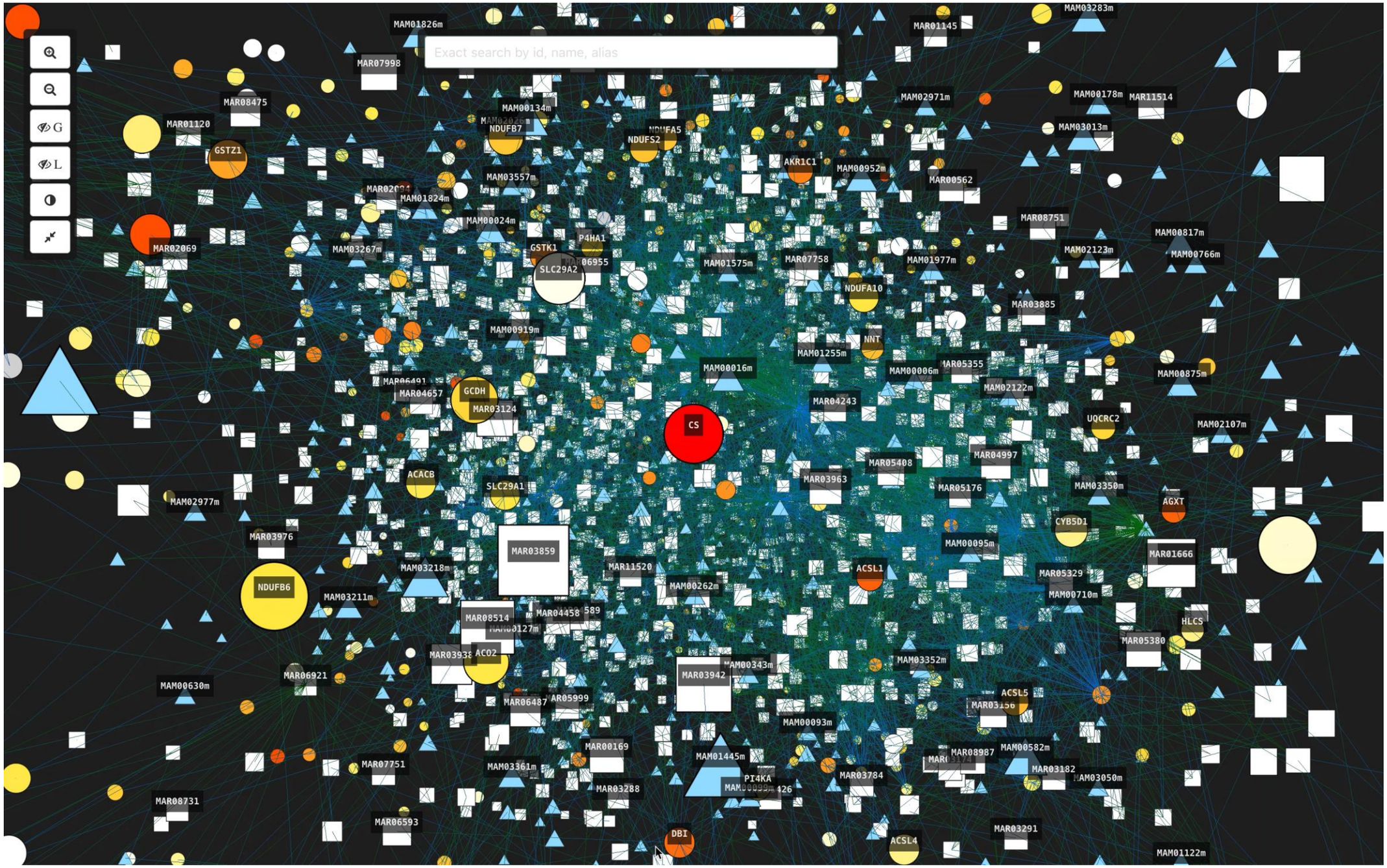




metabolic ATLAS

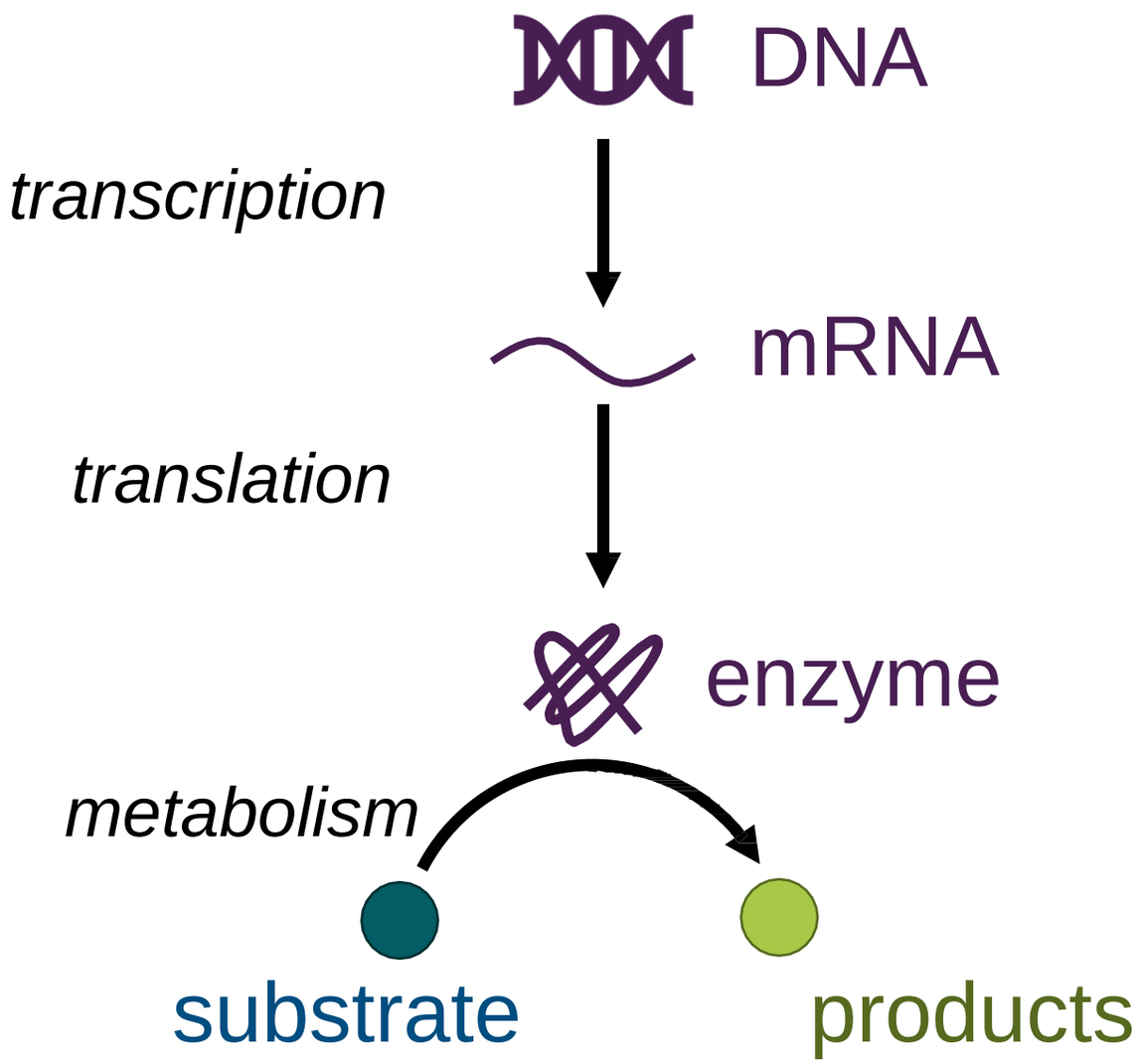
Characterizing metabolic disruption in disease is hindered by biological complexity

3D map of mitochondrial reactions in Human



metabolic ATLAS

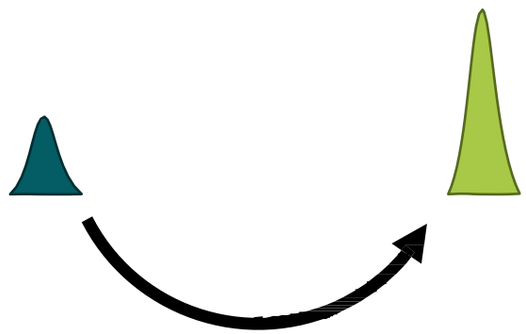
Moving from genetic to metabolic characterisations



Metabolism provides the **energy** and **building blocks** necessary to sustain life.

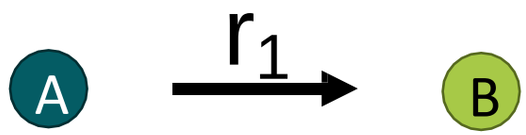
Quantifying fluxes

We can generally measure metabolite concentrations



…but what is often important is the flow or **flux** of metabolites through the reactions.

Enzyme kinetics require knowledge of many kinetic parameters

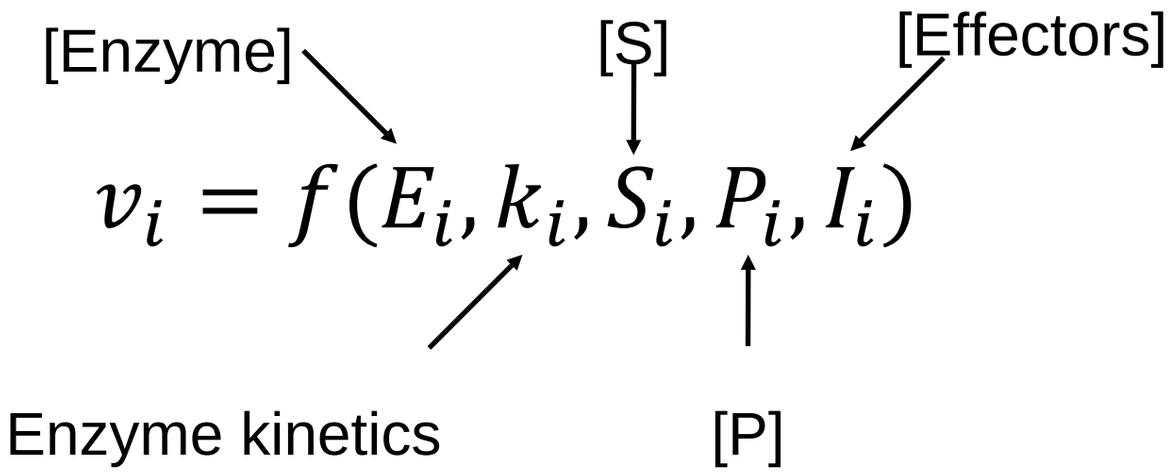


flux = v_1

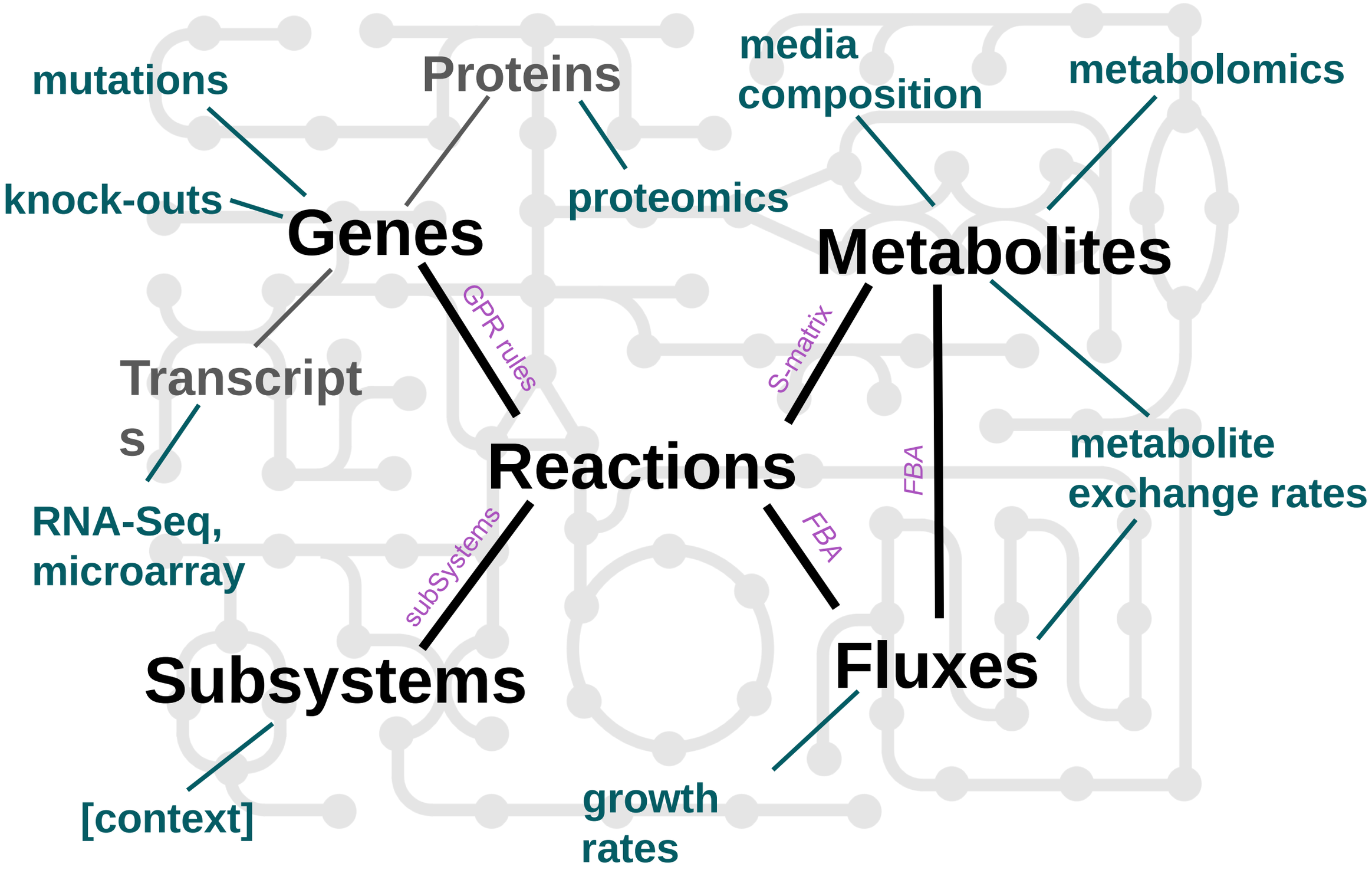
Estimated experimentally

$$\frac{d[A]}{dt} = -v_1 = k_1 \times [A]$$

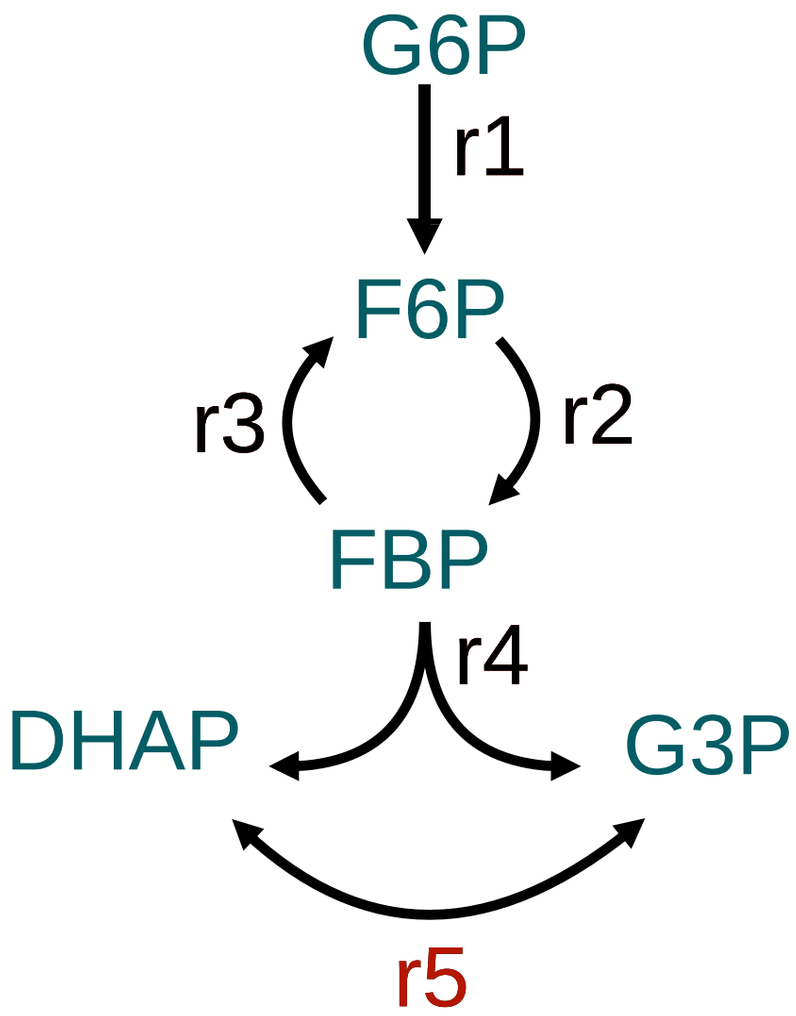
$$\frac{d[A]}{dt} = -v_1 = \frac{V_{max} \times [A]}{K_M + [A]}$$



GEMs as an integrative tool



Using reaction stoichiometry to describe metabolism



Reactions

Metabolites

	r1	r2	r3	r4	r5
G6P	**−1**	0	0	0	0
F6P	**1**	**-1**	**1**	0	0
FBP	0	**1**	**-1**	**-1**	0
DHAP	0	0	0	**1**	**-1**
G3P	0	0	0	**1**	**1**

Genome-scale model (GEM)

	Genes (symbol)	**Proteins** (UniProt)	**Transcript IDs**
	GPI	P06744	**GO Terms**
	n/a		**Orthologs**
	FBP1, FBP2	P09467, O00757	…
	ALDOA, …	P04075, …	
	TPI1	P60174	

	r1	r2	r3	r4	r5
G6P	**−1**	0	0	0	0
F6P	**1**	**-1**	**1**	0	0
FBP	0	**1**	**-1**	**-1**	0
DHAP	0	0	0	**1**	**-1**
G3P	0	0	0	**1**	**1**

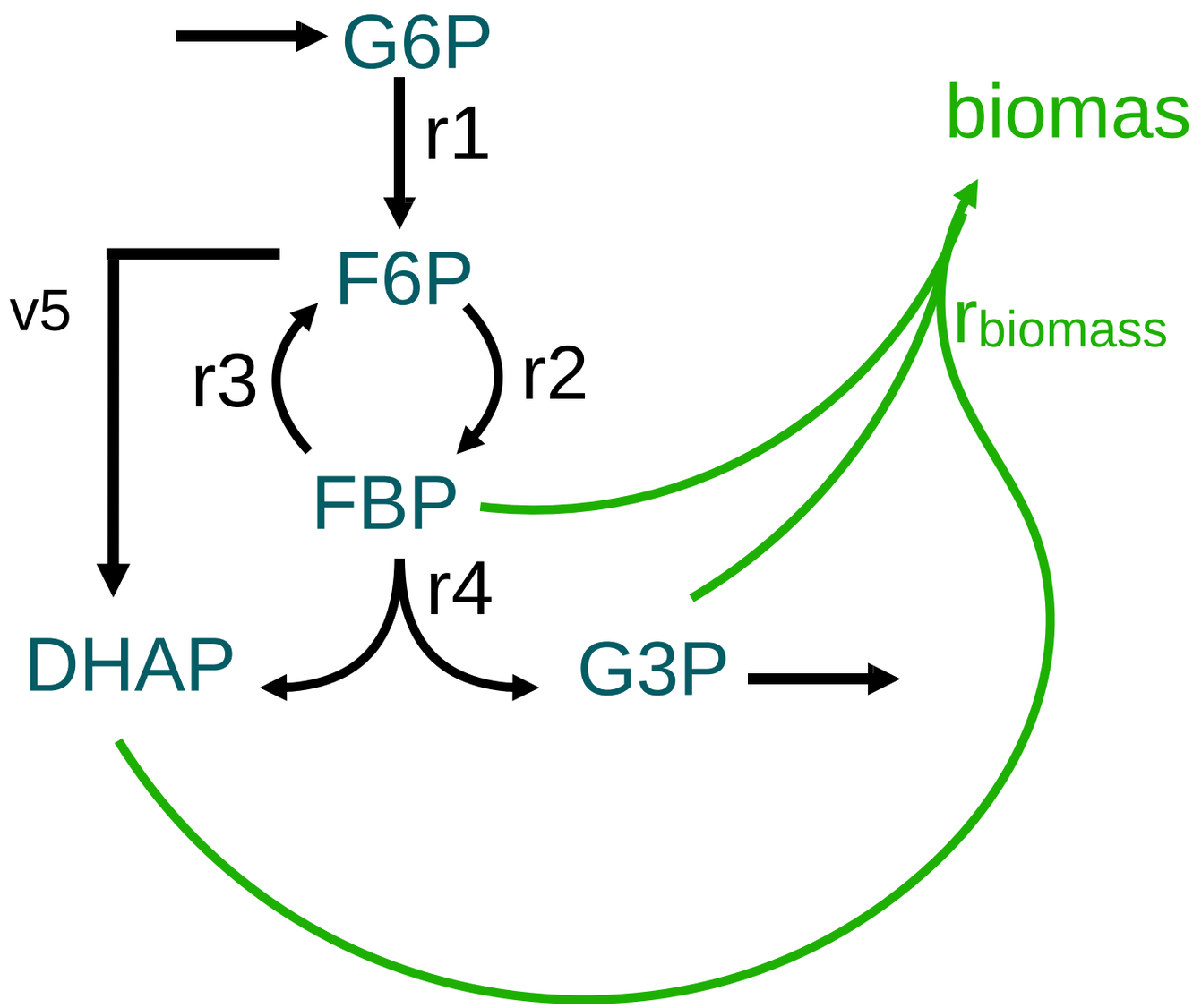
S

Reactions linked to genes that encode the enzymes that catalyze the reaction

“gene-protein rules” (GPR rules)

Flux Balance Analysis (FBA)

Objective function (i.e. optimisation objective) is often:
maximise an artificial **“biomass” reaction** or **ATP production**



Flux Balance Analysis (FBA)

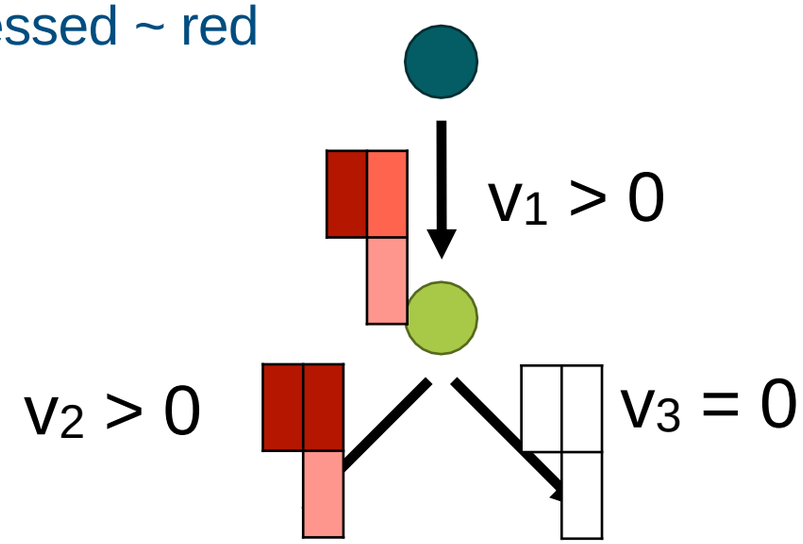
We can further constrain the solution space by limiting reaction fluxes based on their reversibility:

Irreversible reactions  $0 \leq v \leq$ upper bound

Reversible reactions  lower bound $\leq v \leq$ upper bound

Gene expression:

Expressed ~ red



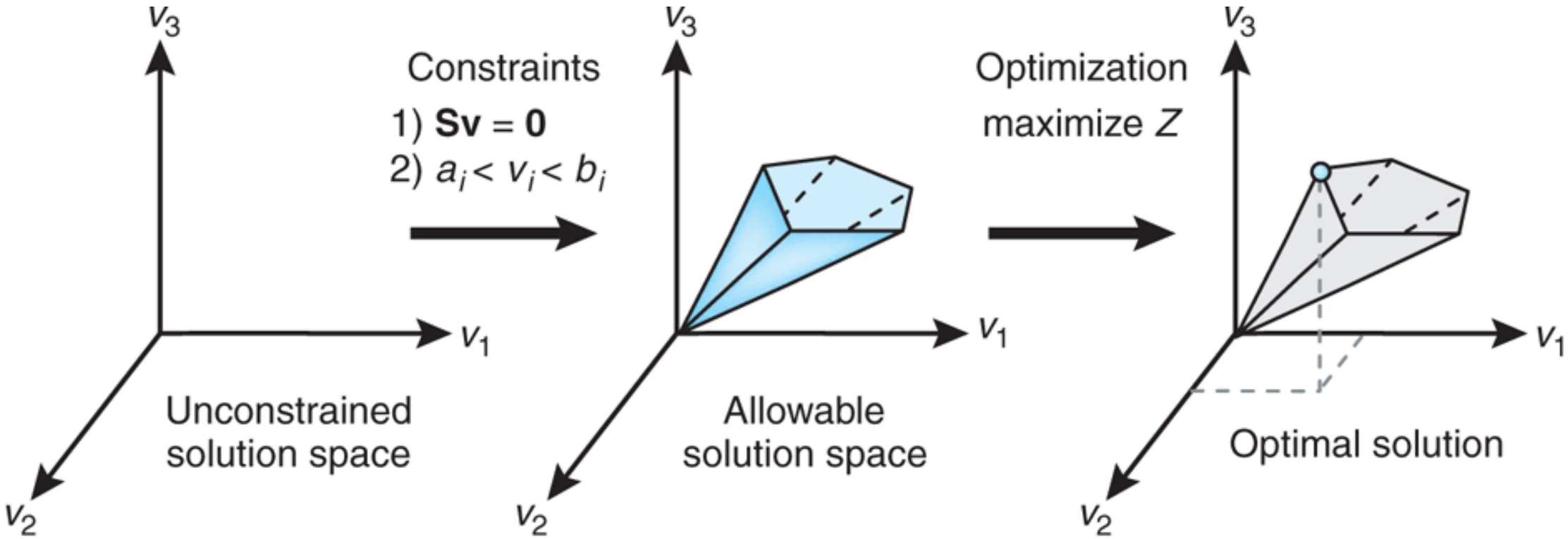
Others:

- Enzyme kinetics
- Thermodynamic constraints

Metabolic tasks
(~biological feasibility)

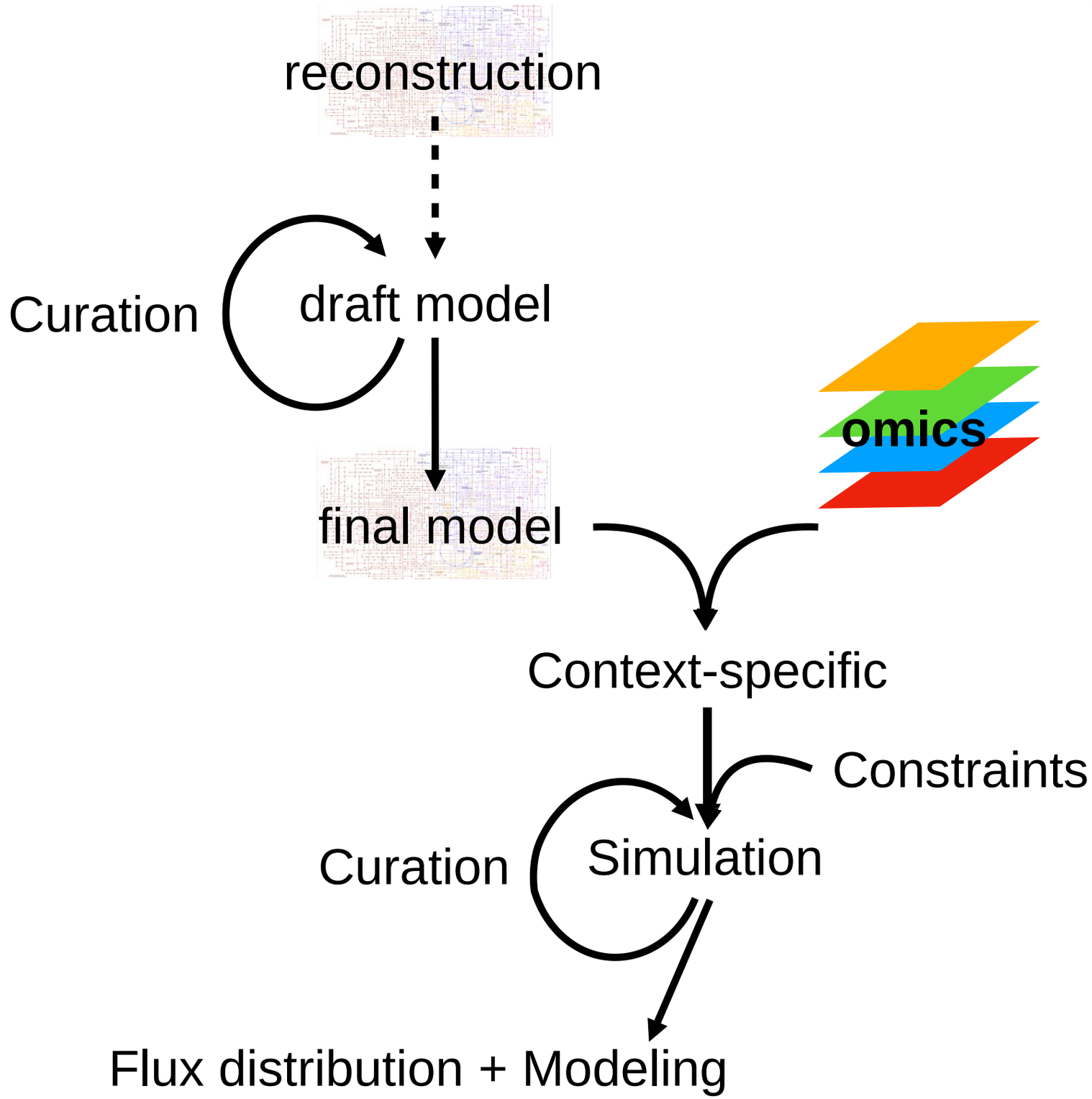
Flux Balance Analysis (FBA)

Since the problem is still **under-defined**, FBA uses linear **optimization** to identify a solution that maximizes (or minimizes) some **objective (Z)**



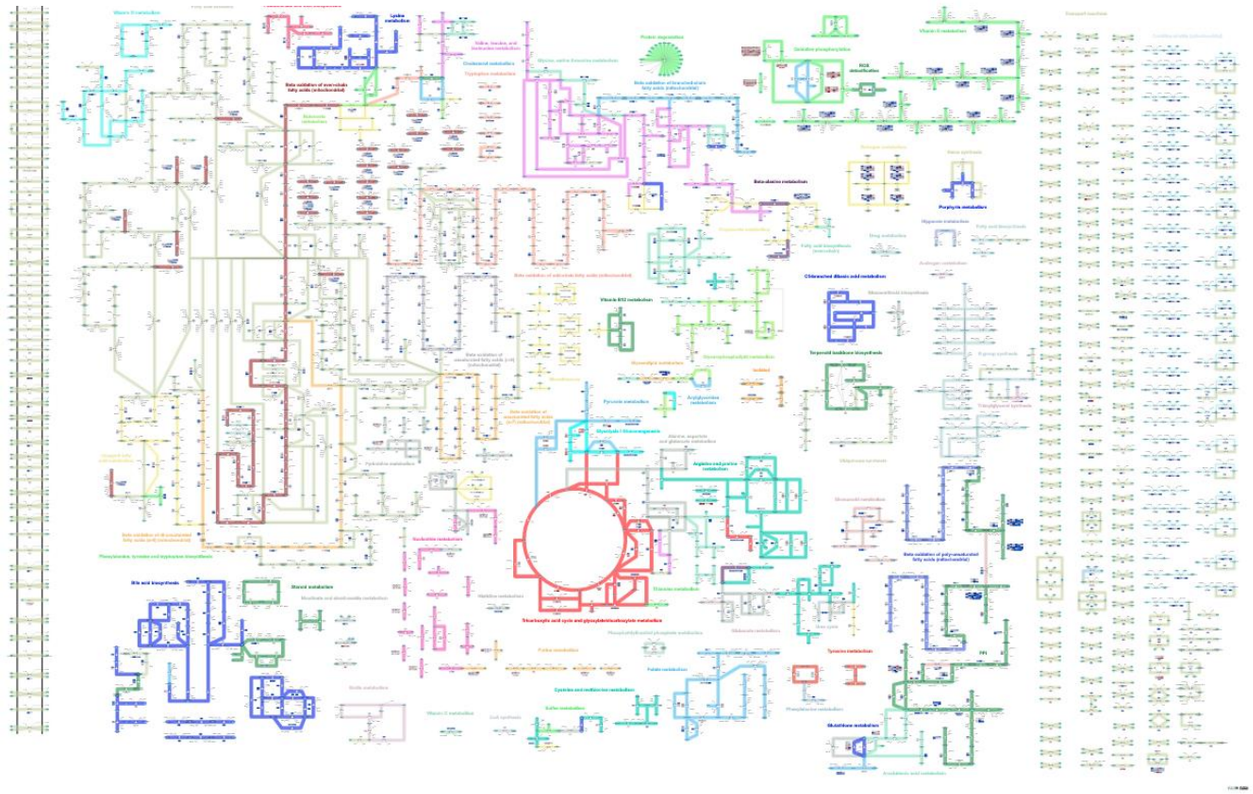
Orth, J., Thiele, I. & Palsson, B. *Nat Biotechnol* (2010).

Analysis workflow



Genome-scale metabolic models as integrative networks

metabolic ATLAS



Simulate flux distributions and pathway activity

Dysregulated pathways

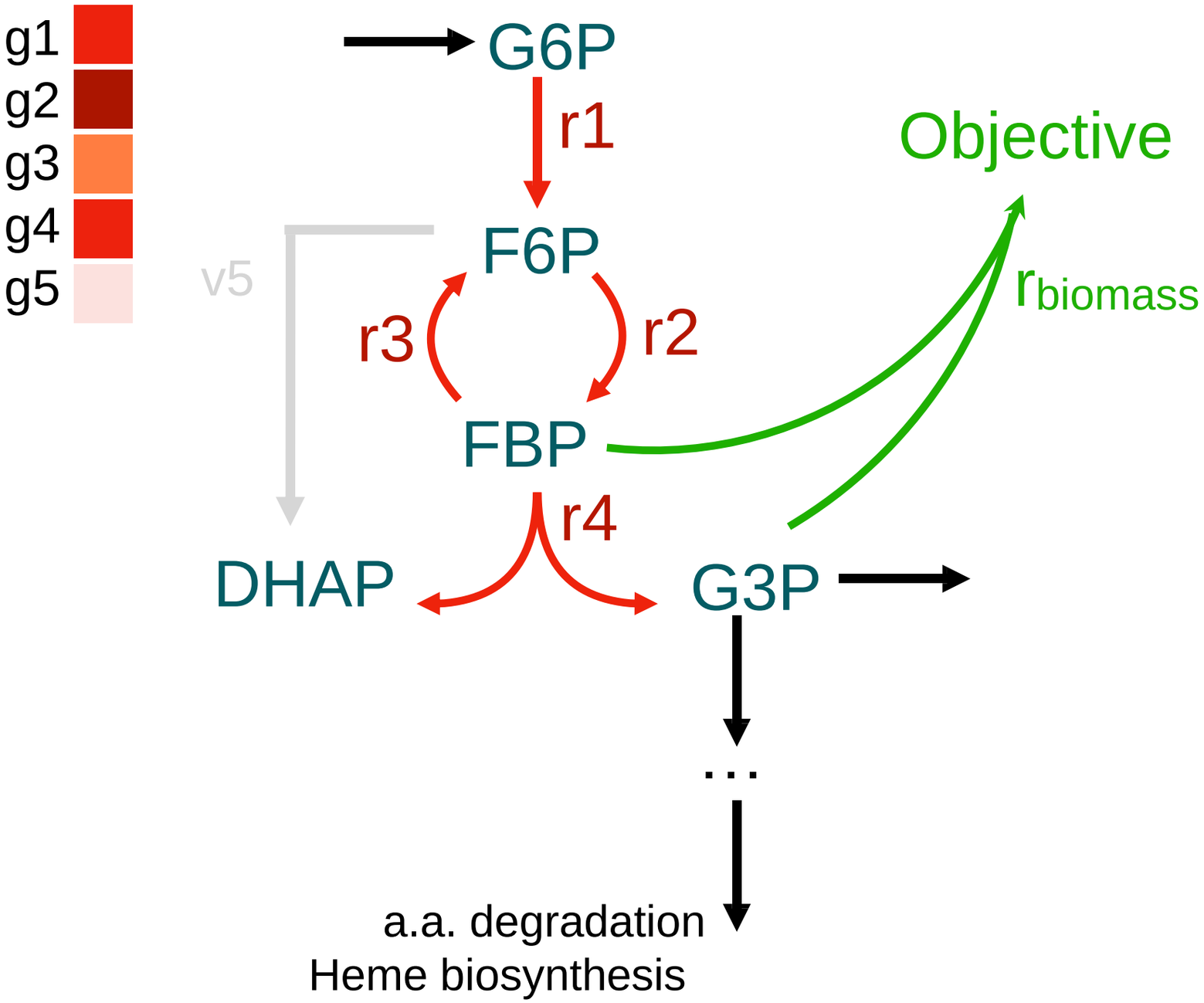
Essential genes & metabolites

...

metabolicatlas.org/

Essential genes, metabolites and metabolic tasks

Essential metabolites and genes can be identified based on fulfilment of certain biological tasks



Agren 2014

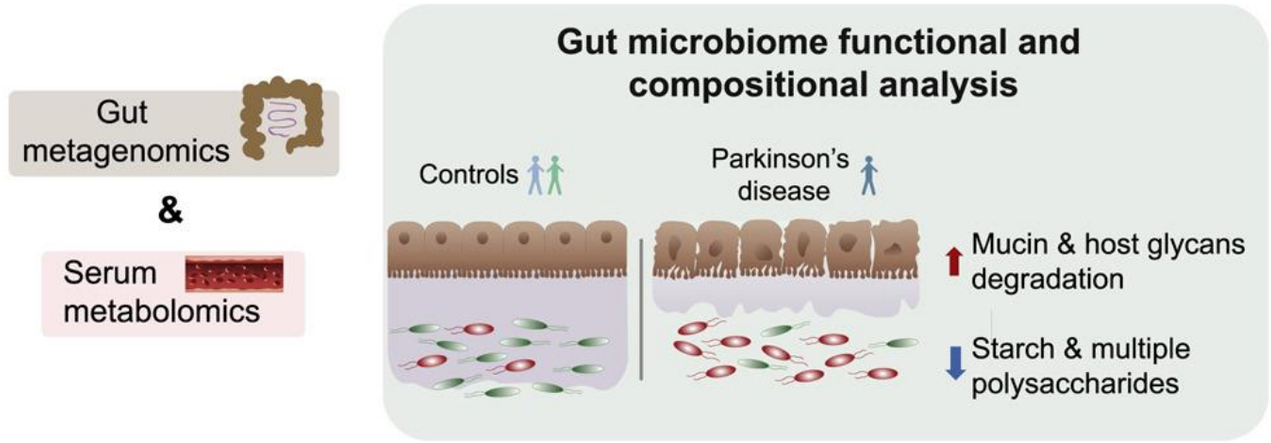
Integrative host + gut GEMs in Parkinson's disease

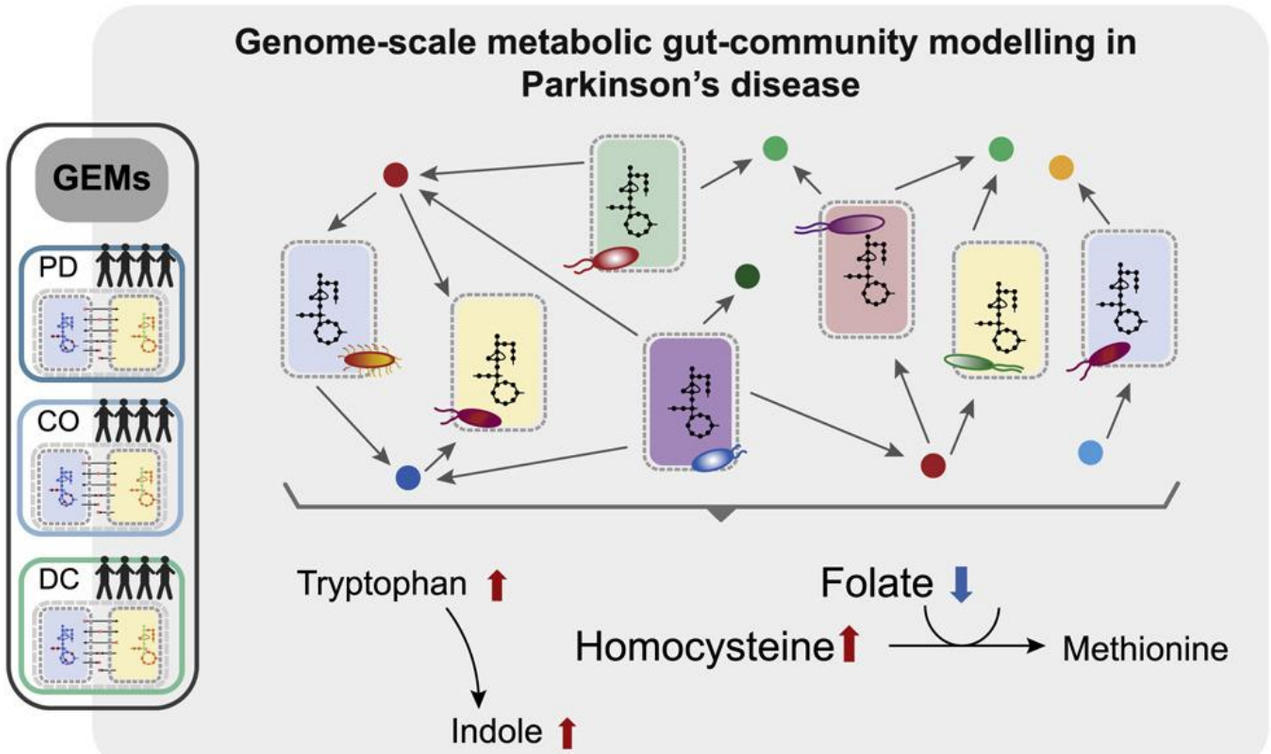
Cohort with PD, healthy controls (CO) and diseased controls (DC)

Microbial GEMs + Intestinal lumen

Personalised community GEMs are modelled and analysed

Substantial imbalances in bacterial mucin and host catabolic enzymes associated with PD

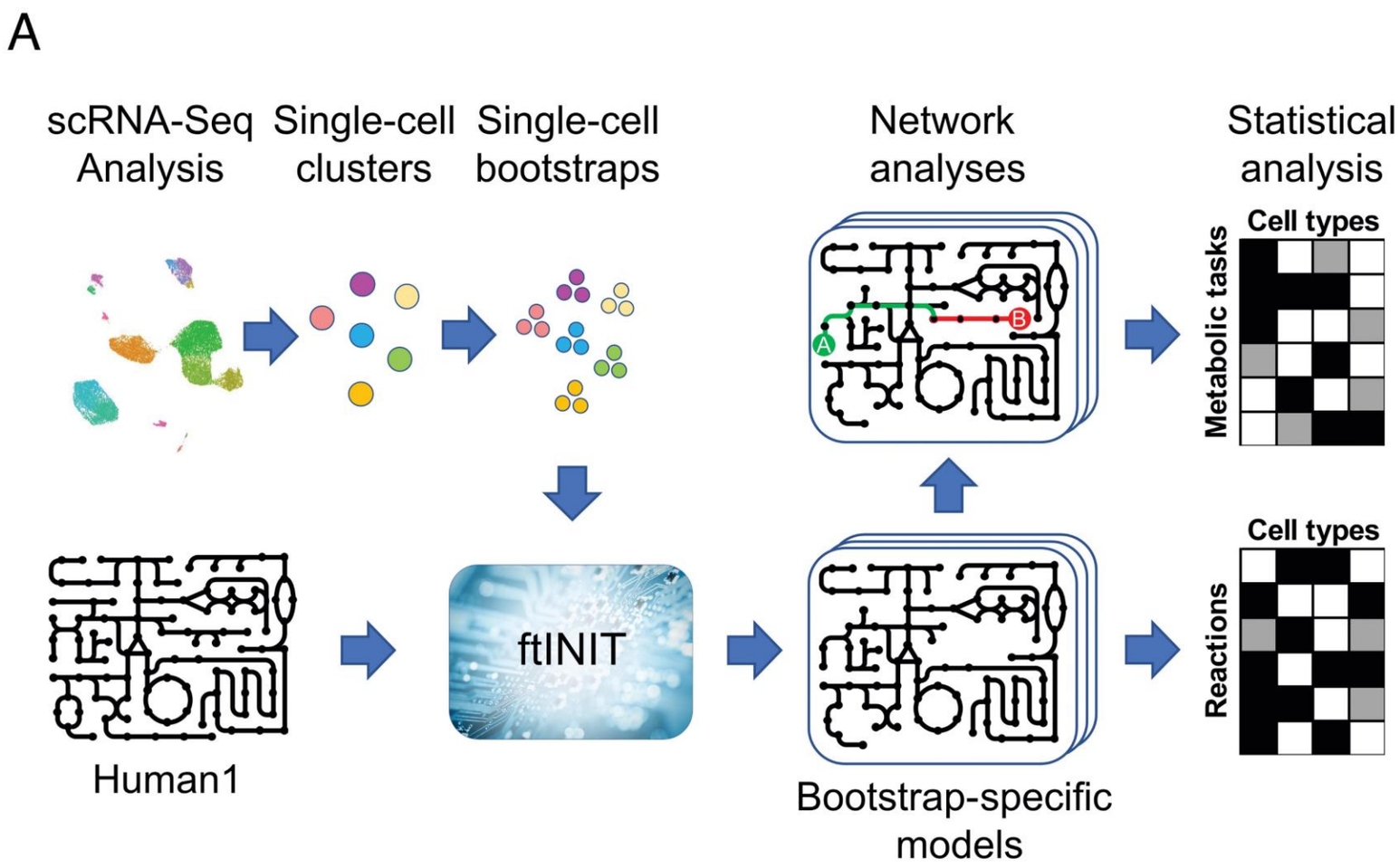




Rosario 2021

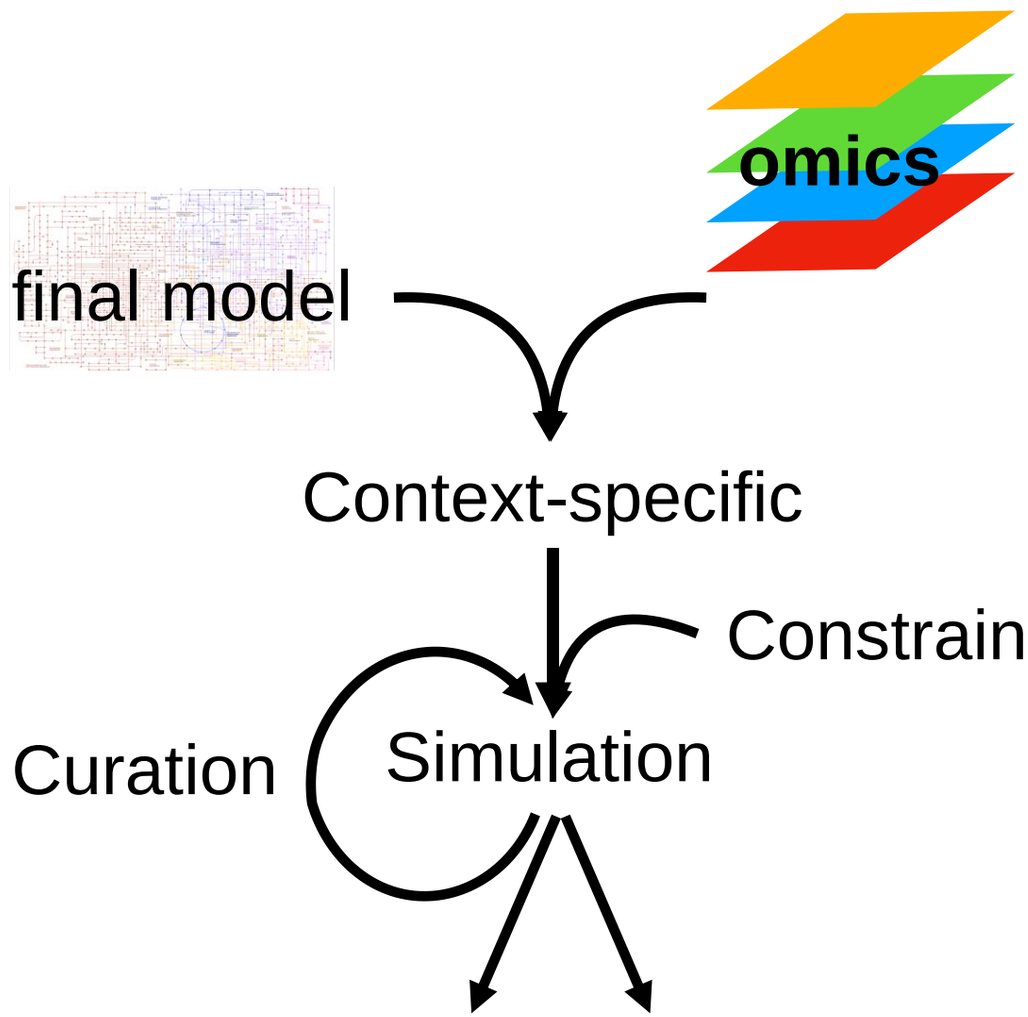
GEMs in single-cell data

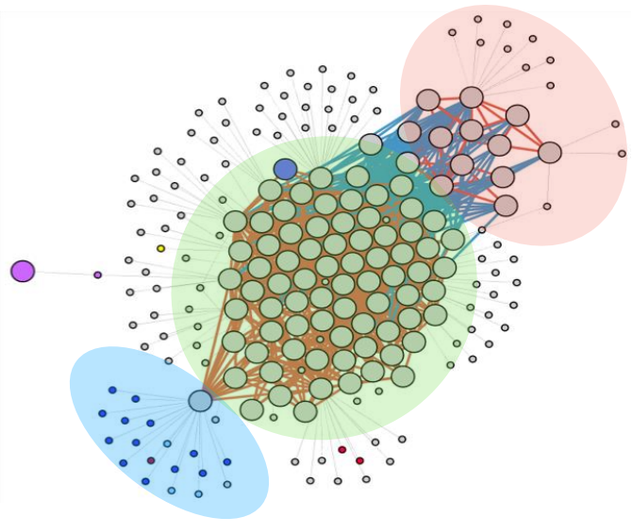
- GEMs show limited applications in single cell data due to sparsity
- Novel bootstrapping method, coupled with improvements in data integration, proposes (cluster) cell-specific GEMs



Gustafsson 2023

Approach for analysis





Modeling | **Topological analysis**

Personalised GEMs + Topology analysis highlight disruption of central metabolism associated with SARS-CoV-2 severity

RNAseq + Metabolomics +
…

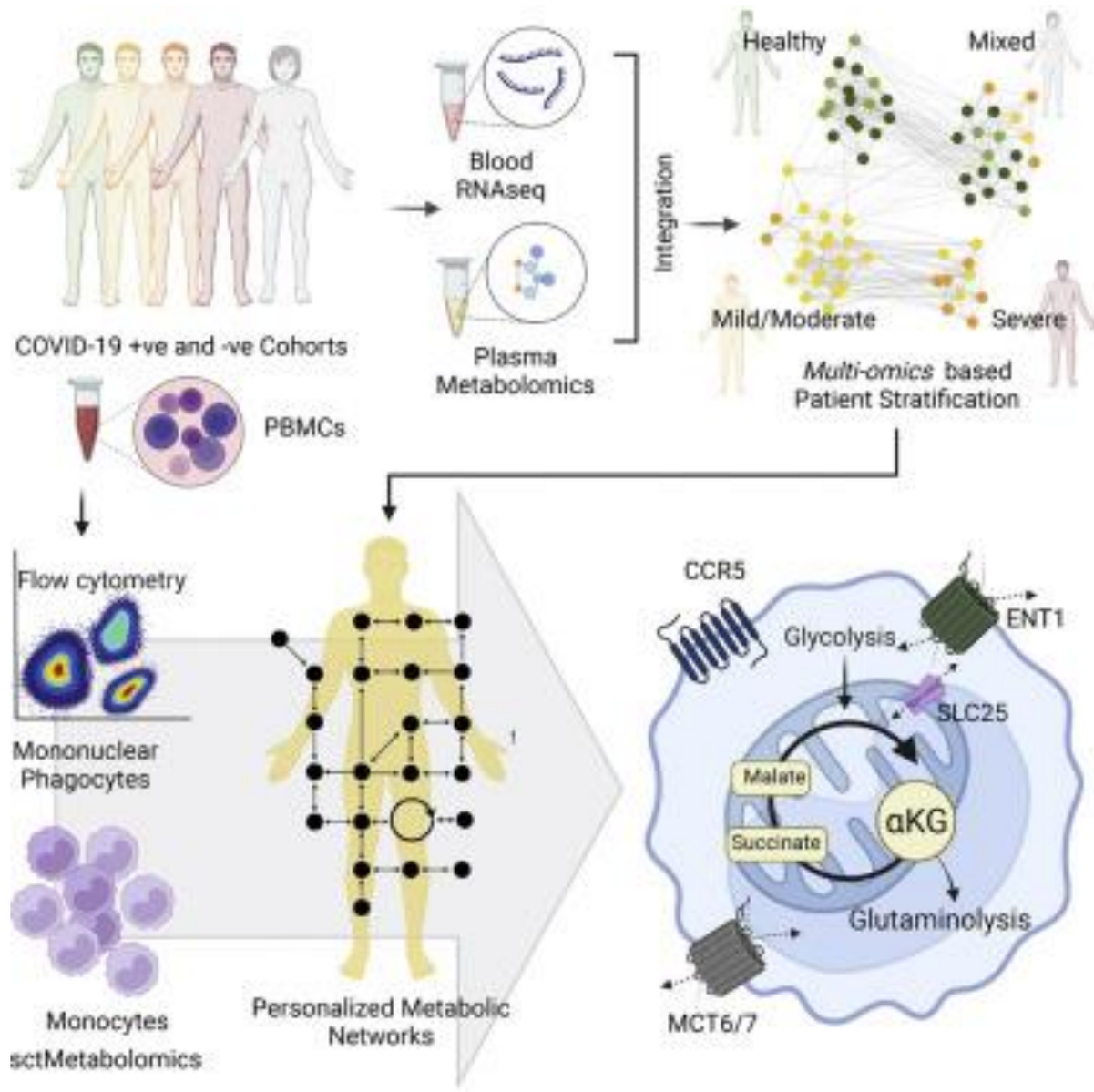
Healthy + Mild + Severe +
Convalescent subjects

Data-driven patient stratification (SFN)

Integrated network analysis for characterisation and feature prioritisation

Human+viral personalised and subgroup-specific GEMs

GEM -> Topology analysis show altered metabolism in key transporters and central metabolites



Ambikan 2022

Conclusions

Genome-scale metabolic models are solid frameworks for integration of multi-omic data, for single individuals or multiple species

GEMs may be specifically constructed for patient subgroups or personalised levels

Elucidate mechanistic alterations and key metabolic effectors associated with pathology

A combination of metabolic modelling and topology analysis provides systematic clues not easily elucidated by the individual approaches

Special thanks to Jonathan Robinson for some of the slides