Knowledge-enhanced biomarker discovery

Trifels Spring School 2025: AI in Bioinformatics

David Selby
DFKI

24th March 2025

Agenda

1 Introduction & Motivation
2 Prior knowledge
3 Multi-omics integration
4 Visible neural networks
5 Hands-on
6 Discussion

Preamble

David Selby

Senior Researcher
Data Science & its Applications
david.selby@dfki.de

dfki ai
Deutsches Forschungszentrum für Künstliche Intelligenz
German Research Center for Artificial Intelligence

RPTU
Rheinland-Pfälzische Technische Universität
Kaiserslautern
Landau

Workshop objectives

By the end of this session, we aim to:

- **Understand** role of prior knowledge in biomarker discovery
- **Learn** how to integrate biological context into workflows
- **Explore** tools for knowledge-guided analysis
- **Discuss** challenges in knowledge-guided AI for biomedicine

1 Introduction & Motivation

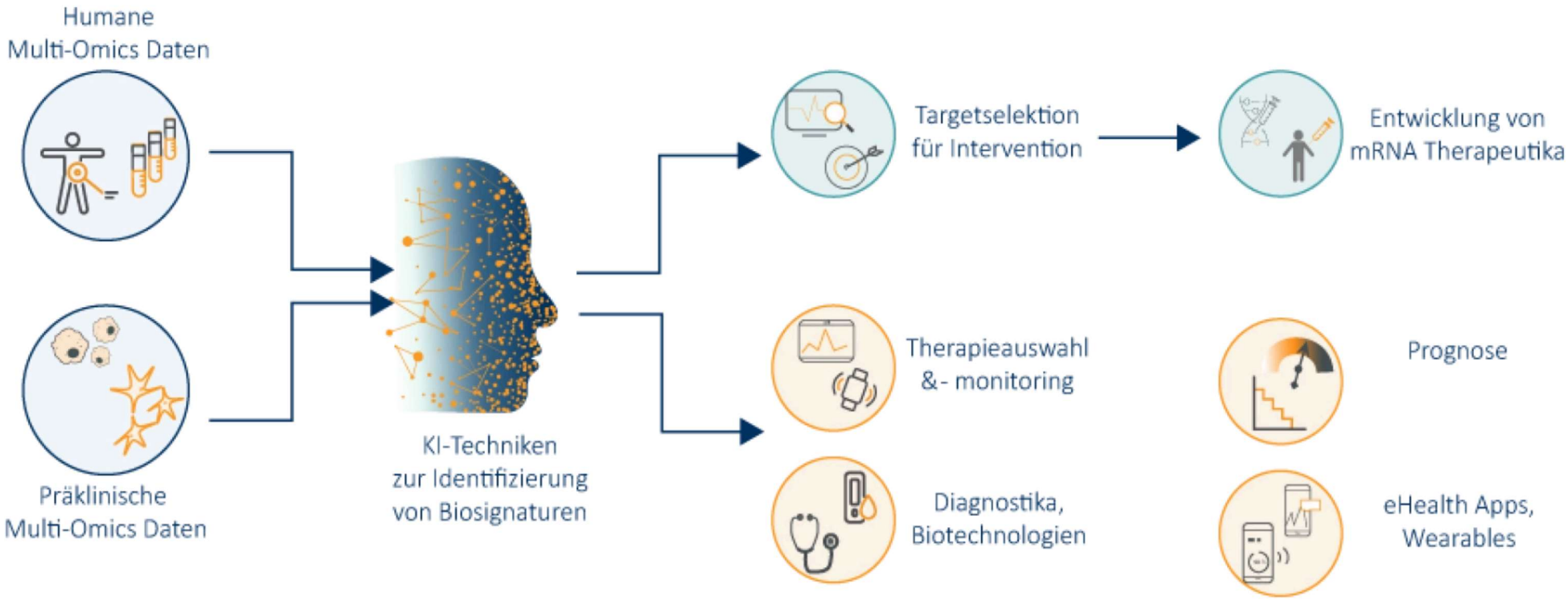
Knowledge enhanced (multi-omics) **biomarker discovery**

Why? Where? How?

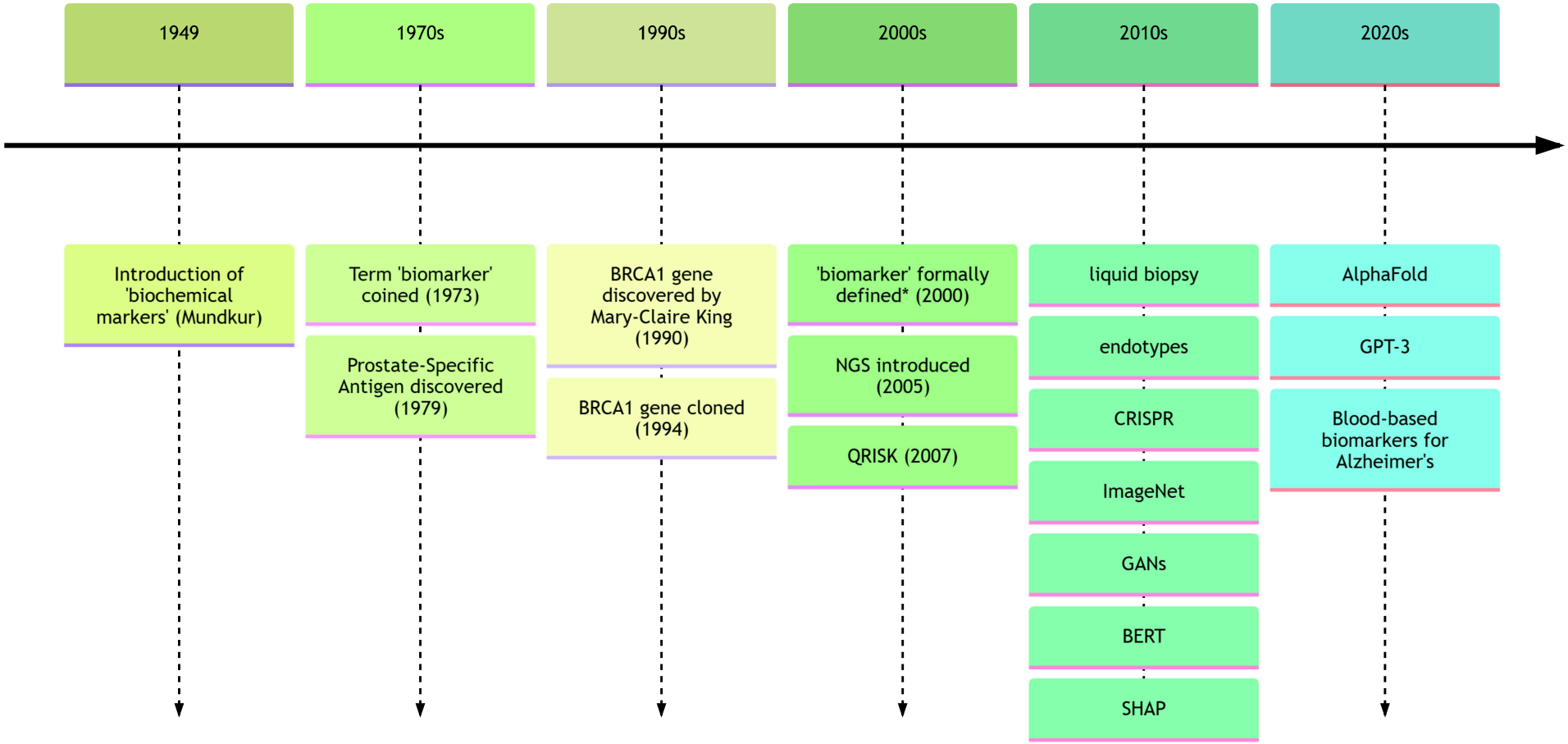
What makes it hard?

What are we looking for?

curATime


Humane
Multi-Omics Daten
Präklinische
Multi-Omics Daten
KI-Techniken
zur Identifizierung
von Biosignaturen
Targetselektion
für Intervention
Entwicklung von
mRNA Therapeutika
Therapieauswahl
&- monitoring
Prognose
Diagnostika,
Biotechnologien
eHealth Apps,
Wearables

A short history of biomarker discovery



*__Definition__: “indicators of biologic/pathogenic processes/responses … frequently measured and evaluated”

Key challenges

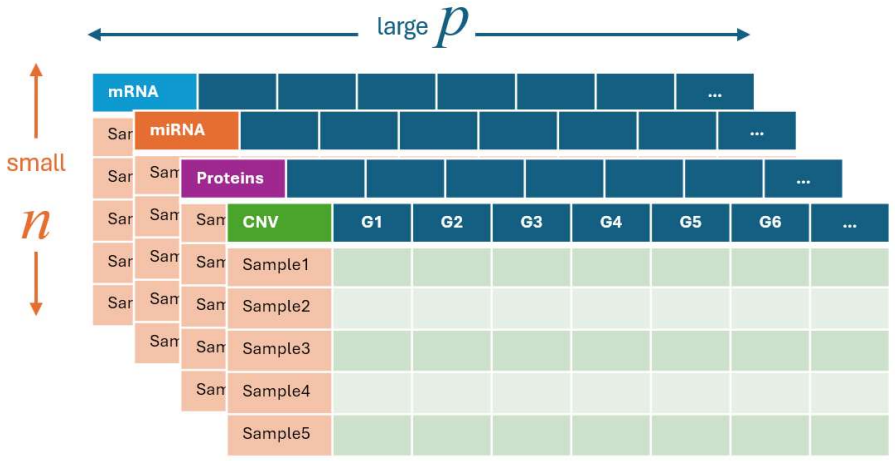
- High dimensionality & small sample sizes ($p \gg n$)
- Heterogeneous data modalities
- Complexity–interpretability tradeoff
- Validation in diverse cohorts
- FAIRness of data, methods and tools

Signatures or biomarkers

(Sets of) features predictive of a biological outcome

Multi-omics biomarker discovery

- Contrasts with **single-omics**
- Combines
  - genomics, transcriptomics, proteomics, metabolomics, …
  - cliniomics, radiomics, …
- Identify robust **signatures** for disease diagnosis, prognosis or treatment



Knowledge-intensive machine learning

- **Prior knowledge** can guide model training
- **Interpretability** is crucial for clinical adoption
- Easy to **overfit** in high-dimensional space
- Biomedical analytical insights not easy to **reproduce**
  - data wrangling
  - domain expertise
  - model interpretations

2 Prior knowledge

What is prior knowledge?

- scientific publications in literature
- open datasets (e.g. TCGA, OpenML, UCI)
- domain-specific databases (e.g. KEGG, Reactome, GO)
- networks data (e.g. protein-protein interactions)
- ontologies
- expert knowledge (Bayesian decision-making)

How can prior knowledge be encoded in a **transparent, reproducible** way?

Approaches

1. Regularization
2. Knowledge graphs
3. Biologically-informed architectures

Information theory

Entropy $H(X)$ quantifies uncertainty in a random variable X

- Higher entropy → more uncertainty
- Lower entropy → more certainty

Prior knowledge reduces entropy:

- Narrows the hypothesis space
- Focuses on biologically plausible solutions

Example: biomarker discovery

Without prior knowledge, search space includes **all possible gene combinations** (high entropy, expensive).

With prior knowledge, focus on **pathways, interactions, or known gene sets** (low entropy, faster convergence).

Mutual information $I(X;Y)$ measures shared information between X (data) and Y (prior knowledge).

Higher $I(X;Y)$ → more effective integration of prior knowledge

Mutual information can be expressed

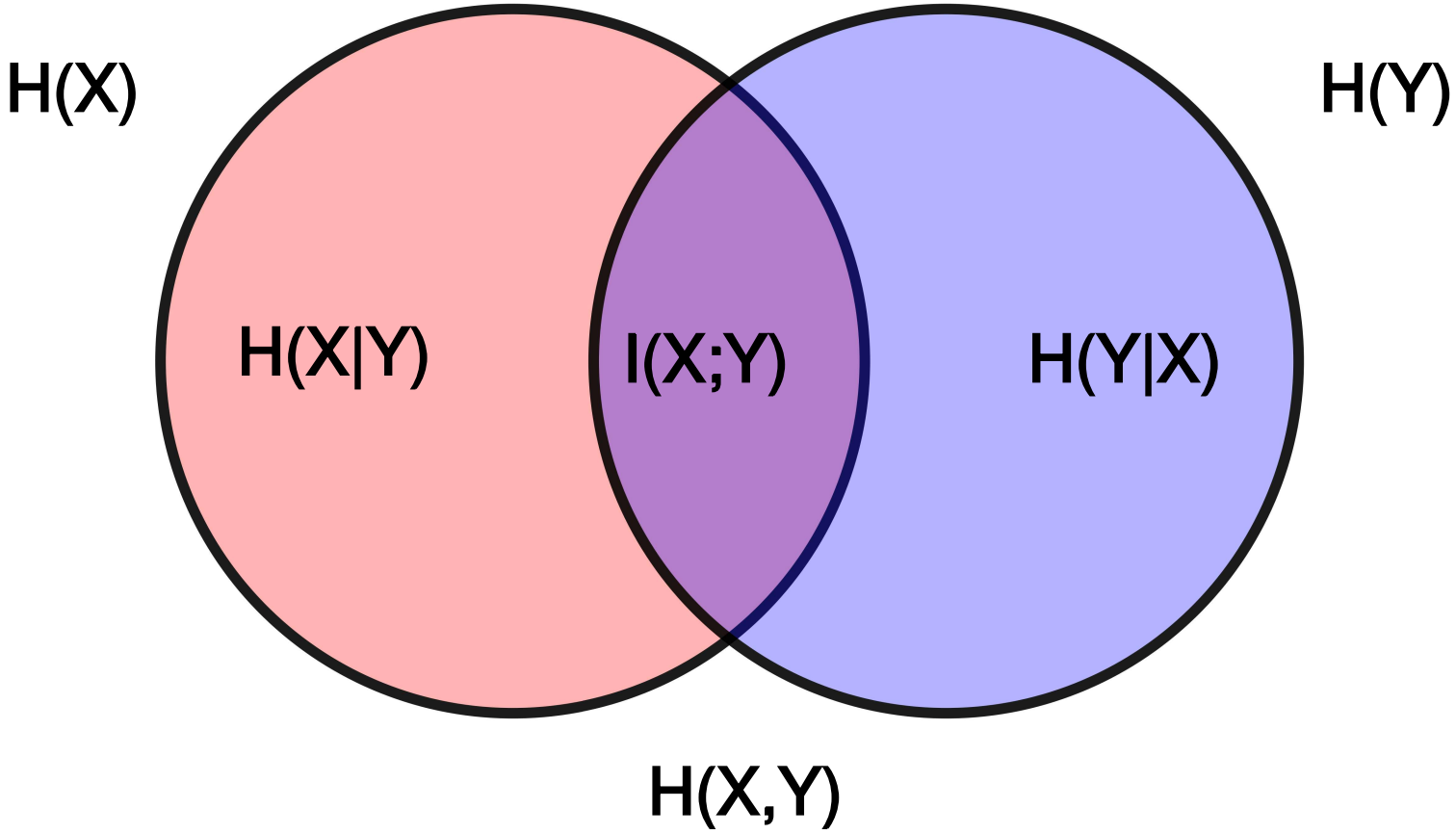
$$I(X;K) = H(X) - H(X|K),$$

where

- $H(X)$ is the entropy (uncertainty) of the data X
- $H(X|K)$ is the conditional entropy of X given prior knowledge K.

Reduction in entropy is equivalent to the mutual information: $\Delta H = I(X;K)$.

> **Tip**
>
> The greater the mutual information $I(X;K)$, the more effective the prior knowledge K is in narrowing the hypothesis space.



Entropy $H(\cdot)$, conditional entropy $H(\cdot \mid \cdot)$ and mutual information $I(\cdot;\cdot)$

Example of entropy in biomarker discovery

Without Prior Knowledge:

- Searching among **10,000 genes**.
- Entropy: $H(X) = \log_2(10,000) \approx 13.29 \text{ bits}.$

With Prior Knowledge:

- Prior knowledge narrows the search to **100 candidate genes**.
- Entropy: $H(X|K) = \log_2(100) \approx 6.64 \text{ bits}.$

Mutual Information

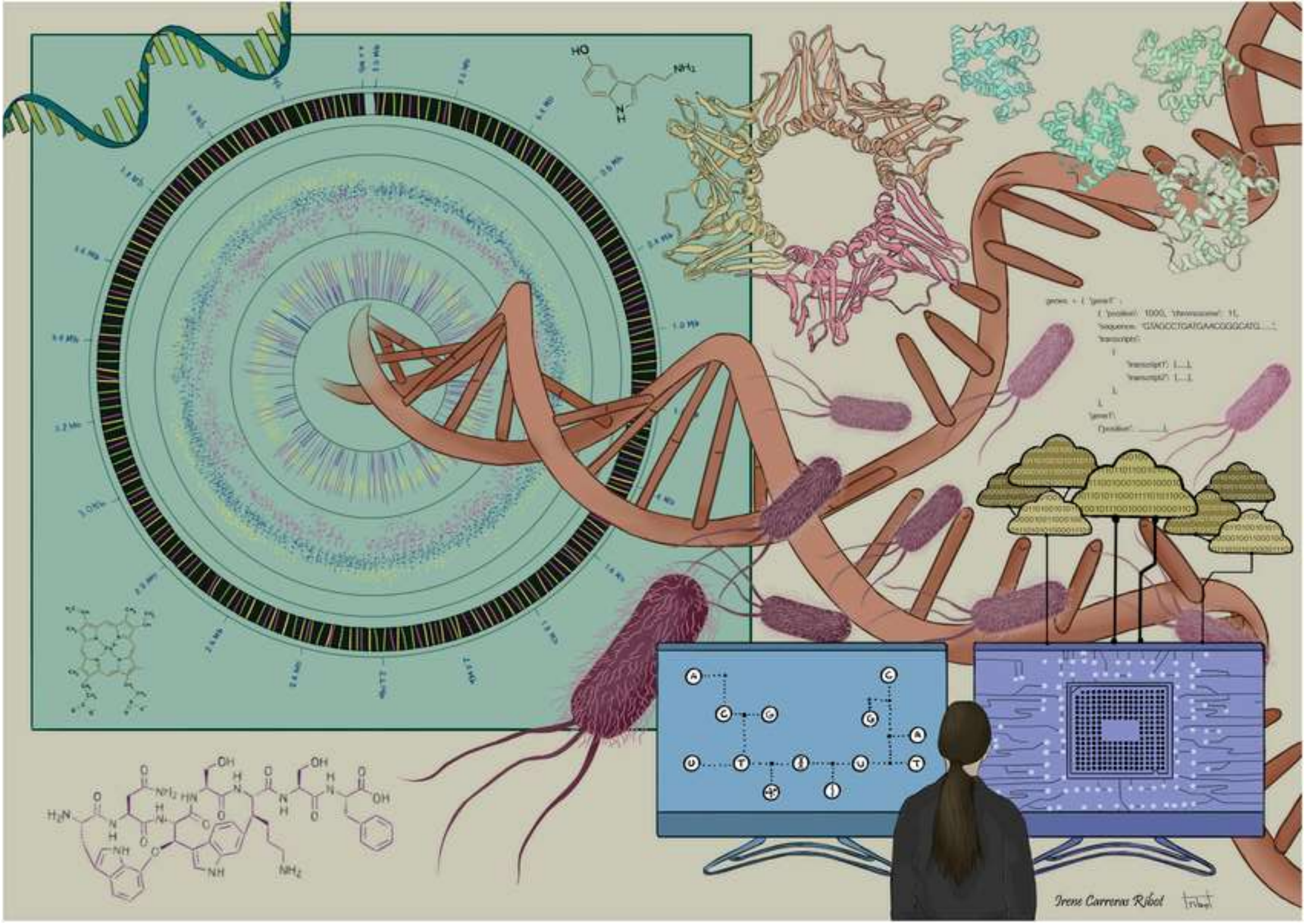
- Reduction in entropy:
  $I(X;K) = H(X) - H(X|K) = 13.29 - 6.64 = 6.65 \text{ bits}.$

⚠ Lower entropy ≠ correctness

Prior knowledge (and data) can still be **wrong/biassed**!

Then *smaller* hypothesis space → barking up the *wrong* tree.

3 Multi-omics integration


Irene Carreras Ribot

Single-omics methods

Task	Tools
Differential expression analysis	DESeq2, edgeR, limma
Clustering samples/features	k-means, hierarchical clustering, t-SNE, UMAP
Pathway enrichment analysis	GSEA, Reactome
Predictive modelling	GLMs, random forests, SVM, XGBoost, NNs

Single-omics challenges

Challenge	Methods	Tools/Packages
$p \gg n$	Dimensionality reduction	PCA, t-SNE, UMAP, LASSO, ElasticNet
Batch effects	Batch effect correction	ComBat (sva), limma, Harmony
Missing data	Imputation	MICE, KNN imputation, MissForest

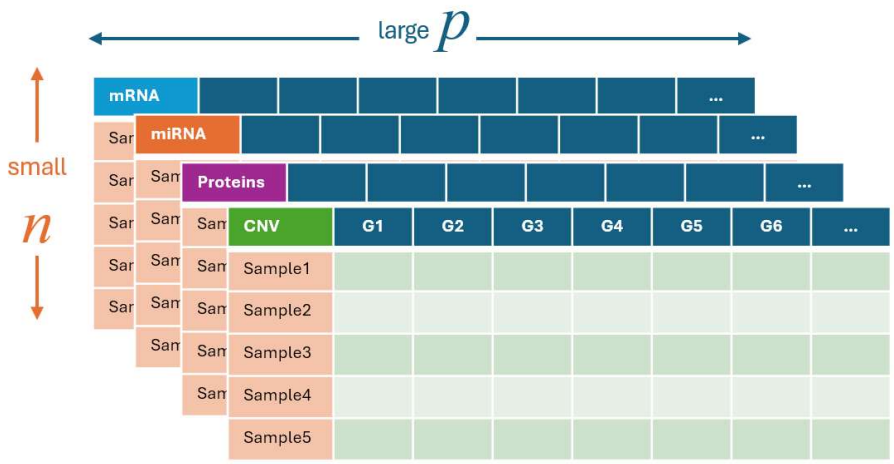
Tip

Single-omics data is **tabular**, for which **tree-based models** (e.g. random forests) can outperform deep learning.

⚠ Multi-omics data bring extra challenges

- **Integration** of different data types
- **Interpretation** of complex interactions

Multi-omics datasets



- Tabular data
- High-dimensional
- Small samples
- **Multimodal structure**

Why deep learning?

Why not just use classical methods?

Classical approaches

Multi-omics factor analysis (MOFA)

Unsupervised, generalization of PCA

Canonical correlation analysis (CCA)

Find linear combinations of features that are maximally correlated

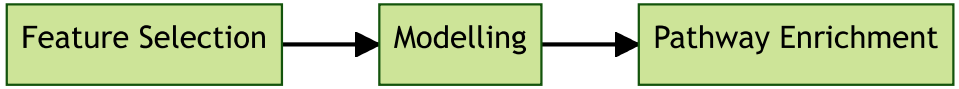
Grouped LASSO

(Linearly) penalize groups of features together

Why deep learning?

Why not just use classical methods? Isn’t it tabular data?

- **Non-linear** relationships
- **Interactions** between features
- **Representation learning** from raw data
- **End-to-end** learning



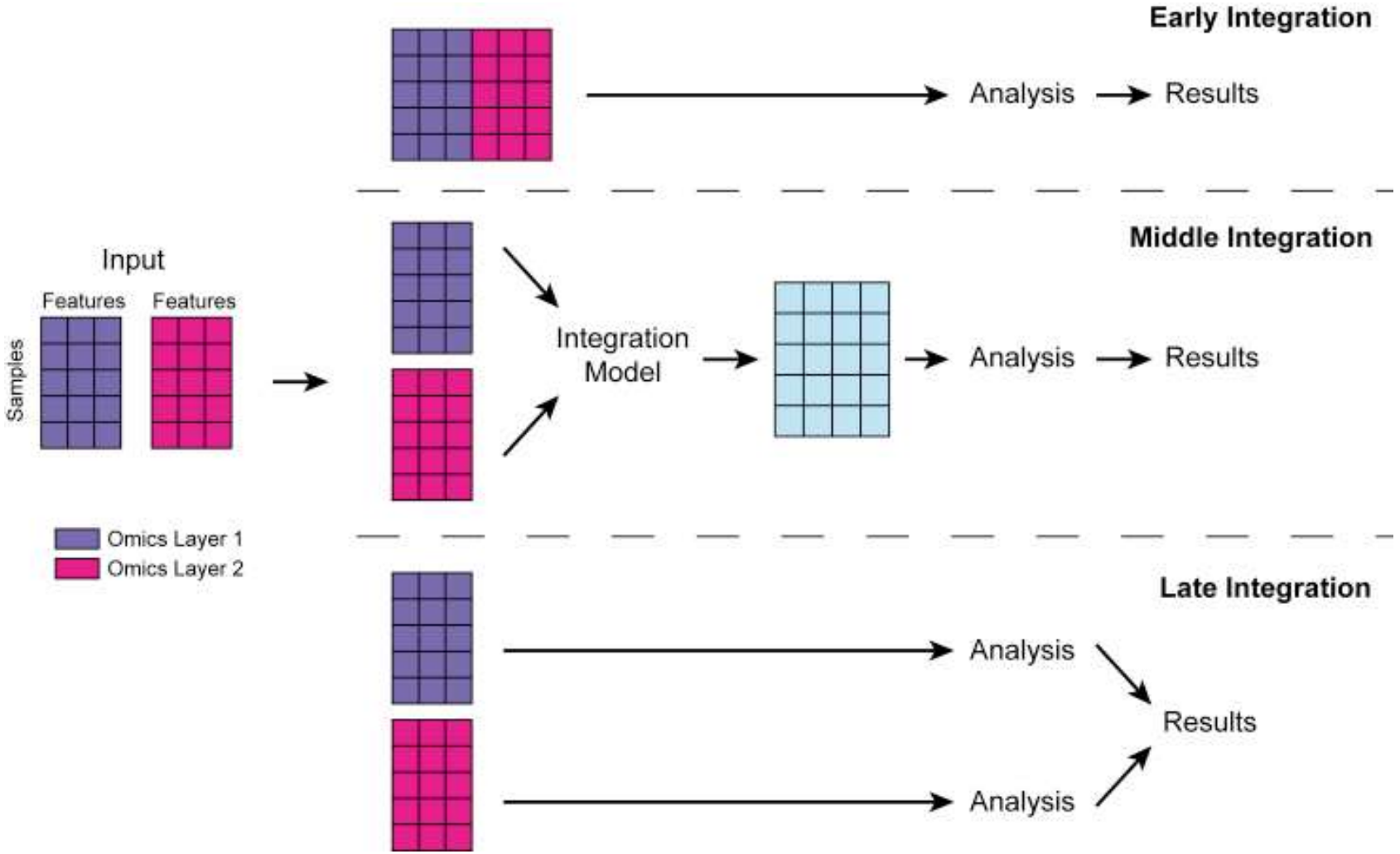
Deep learning architectures

Question

Which neural network architectures are suitable for omics?

Multimodal fusion

When should we combine omics layers?



Multimodal fusion

When should we combine omics layers?

Early

easier, loss of information, worse performance*

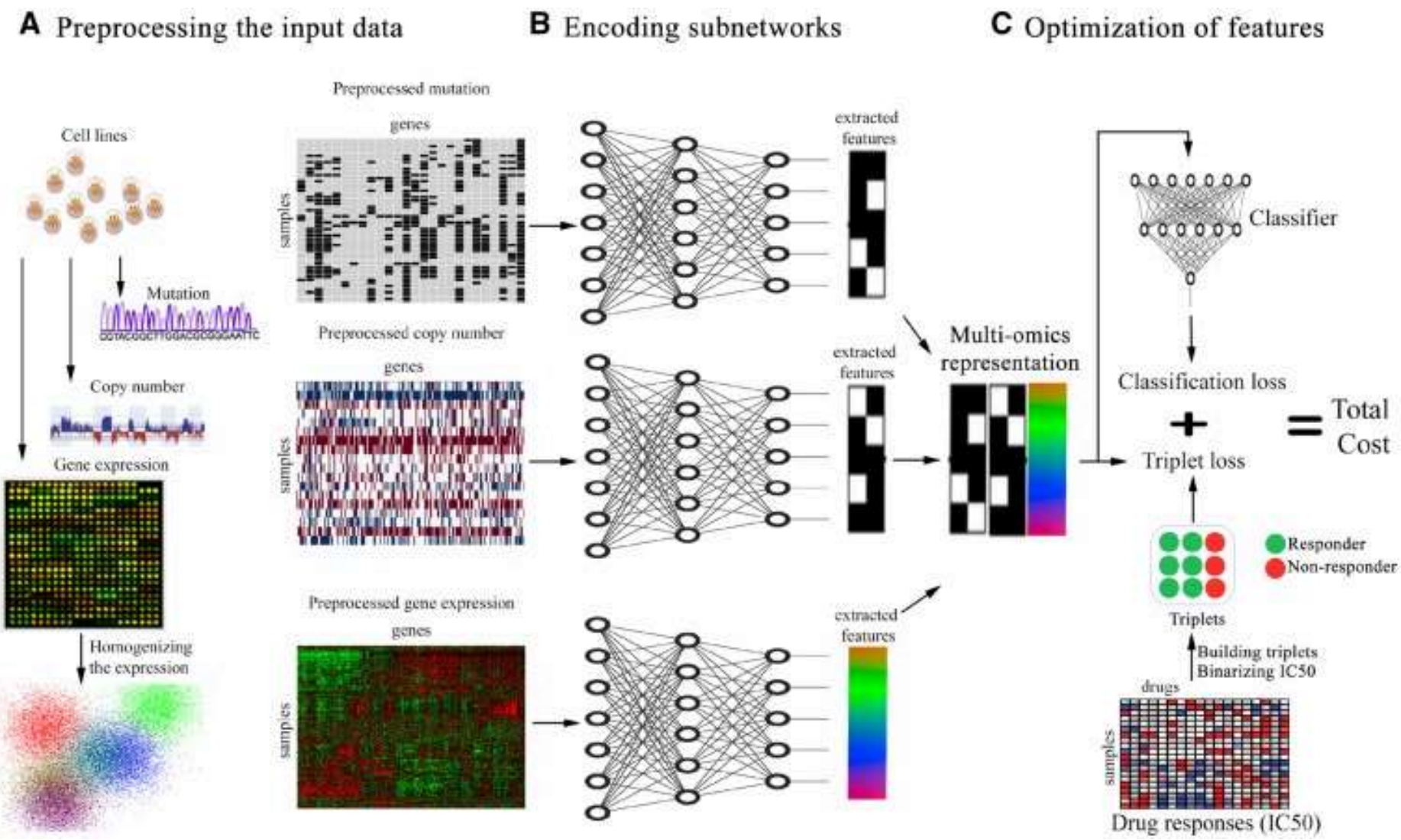
Intermediate (mixed, joint)

modality-specific layers, but harder to train

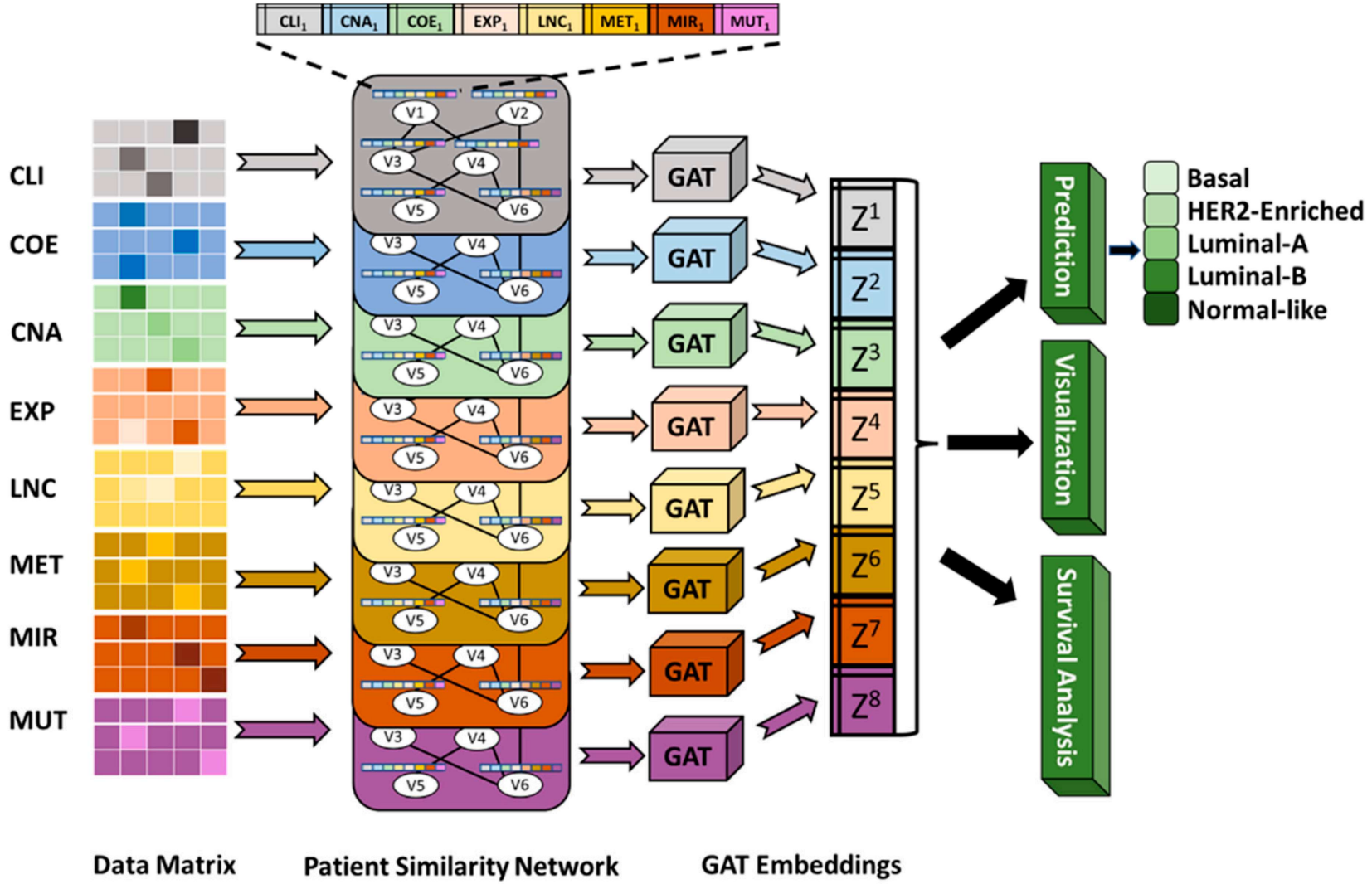
Late

may not capture interactions

Late fusion (MOLI)

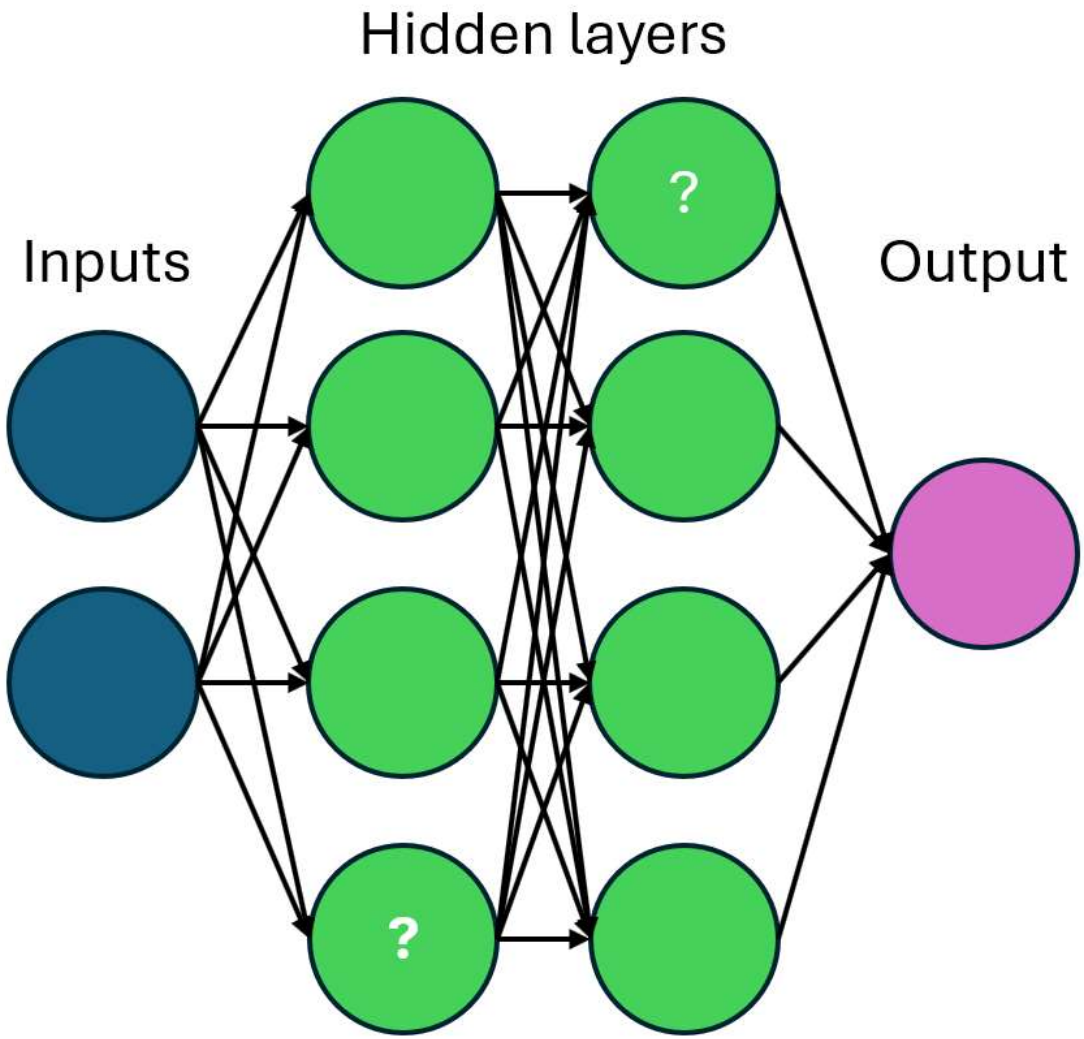


Graph neural networks (GNNs)



MOGAT

Model explanations



Can we call hidden nodes in neural networks ‘biomarkers’?

Model explanations

Input-level explanations:

- p-values, features importance
- DeepLIFT
- SHAP
- LIME

→ *post-hoc* gene-set enrichment analysis (GSEA) or “pathway analysis”

Gene set enrichment analysis

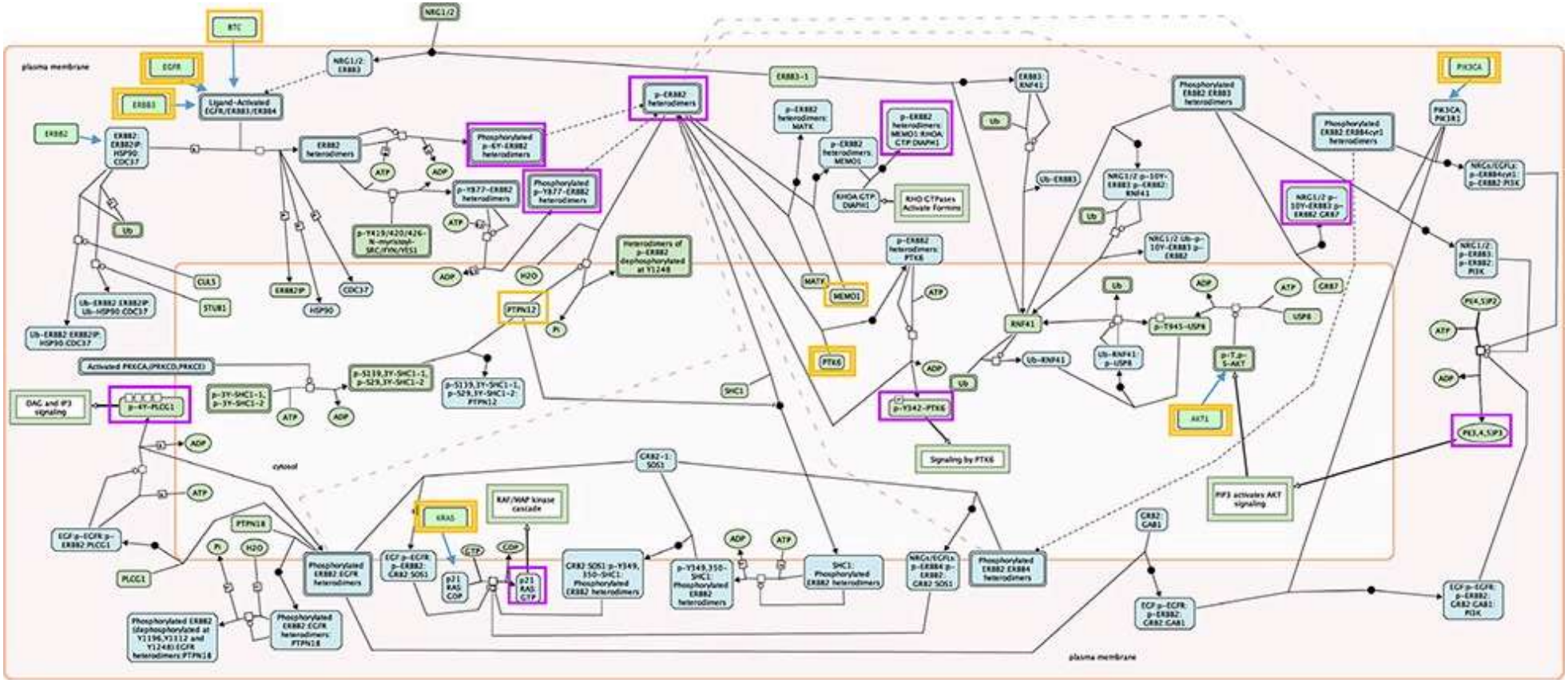
1. Set of genes $G = \{g_1, g_2, \ldots, g_N\}$. Order by ranking metric $S(g_i)$ (e.g. *t*-statistic)
2. Compute **enrichment score** using running sum statistics, or **overrepresentation score** with hypergeometric test:

$$P(X = x) = \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}}$$

with p-value

$$p = \sum_{i=x}^{\min(M,K)} P(X = i).$$

What is a pathway (gene set)?



Reactome pathways

A pathway is a set of genes that are known to interact in a biological process.

Pathway databases

reactome

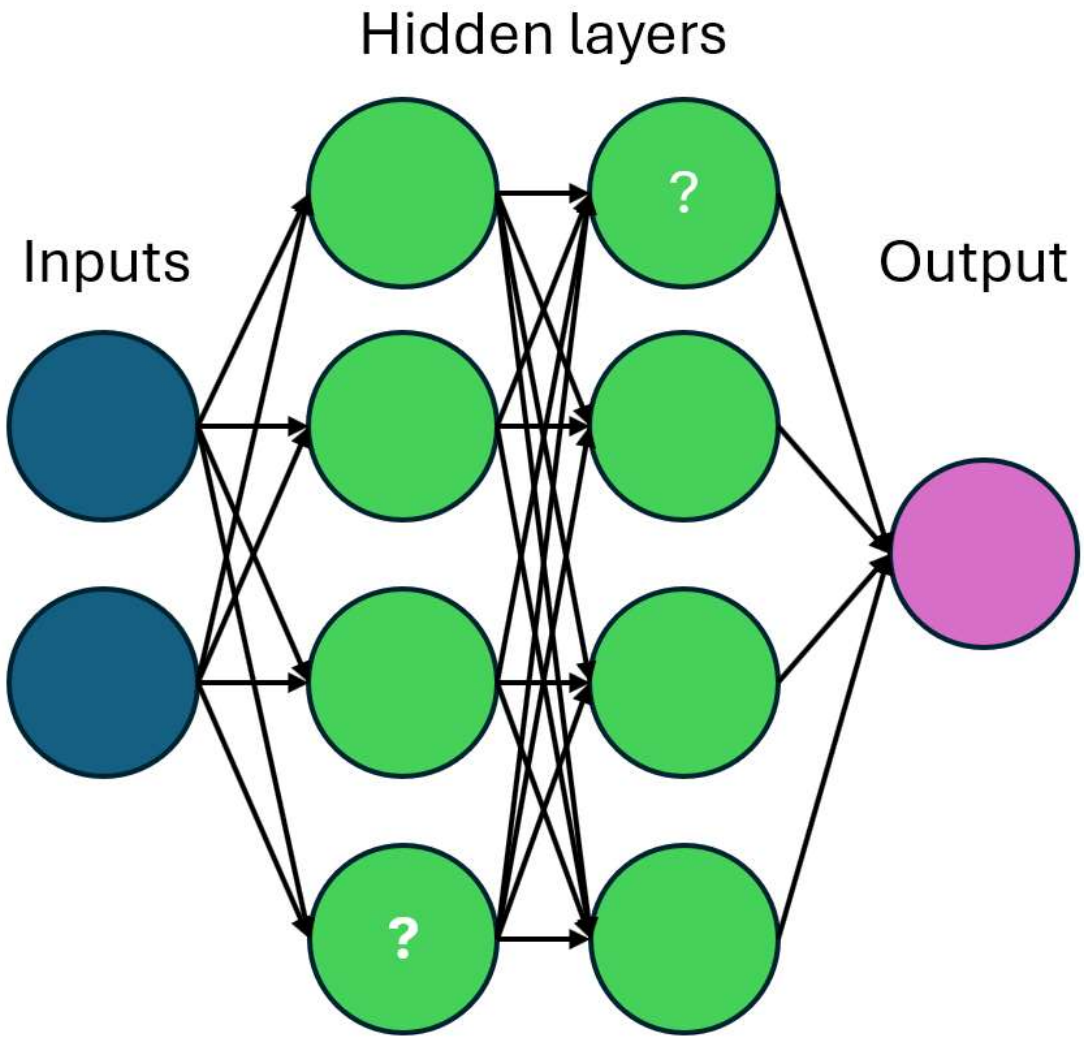
GENEONTOLOGY
Unifying Biology

Pathway enrichment analysis

What are the disadvantages of using *post-hoc* GSEA or GSOA?

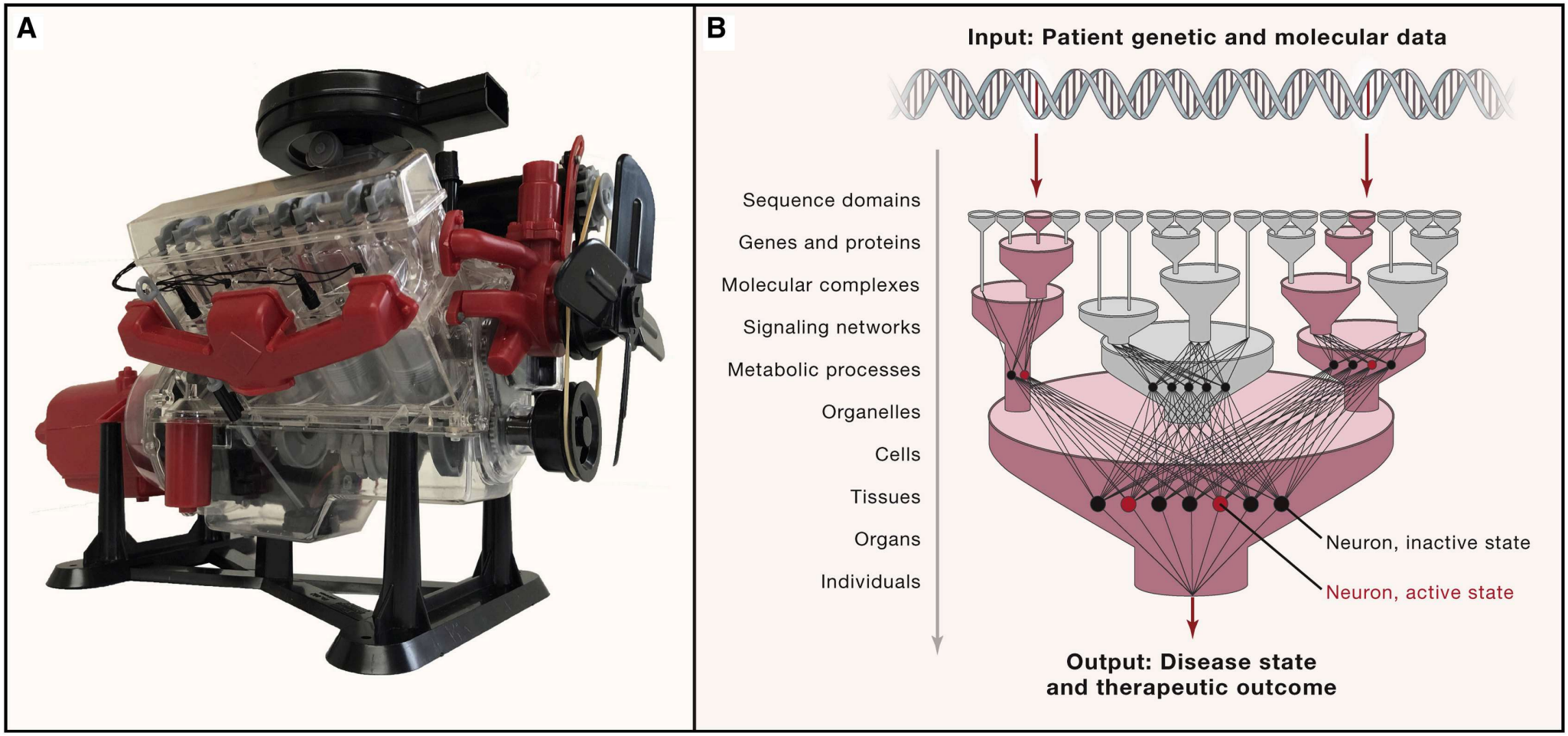
4 Visible neural networks

Feedforward neural network



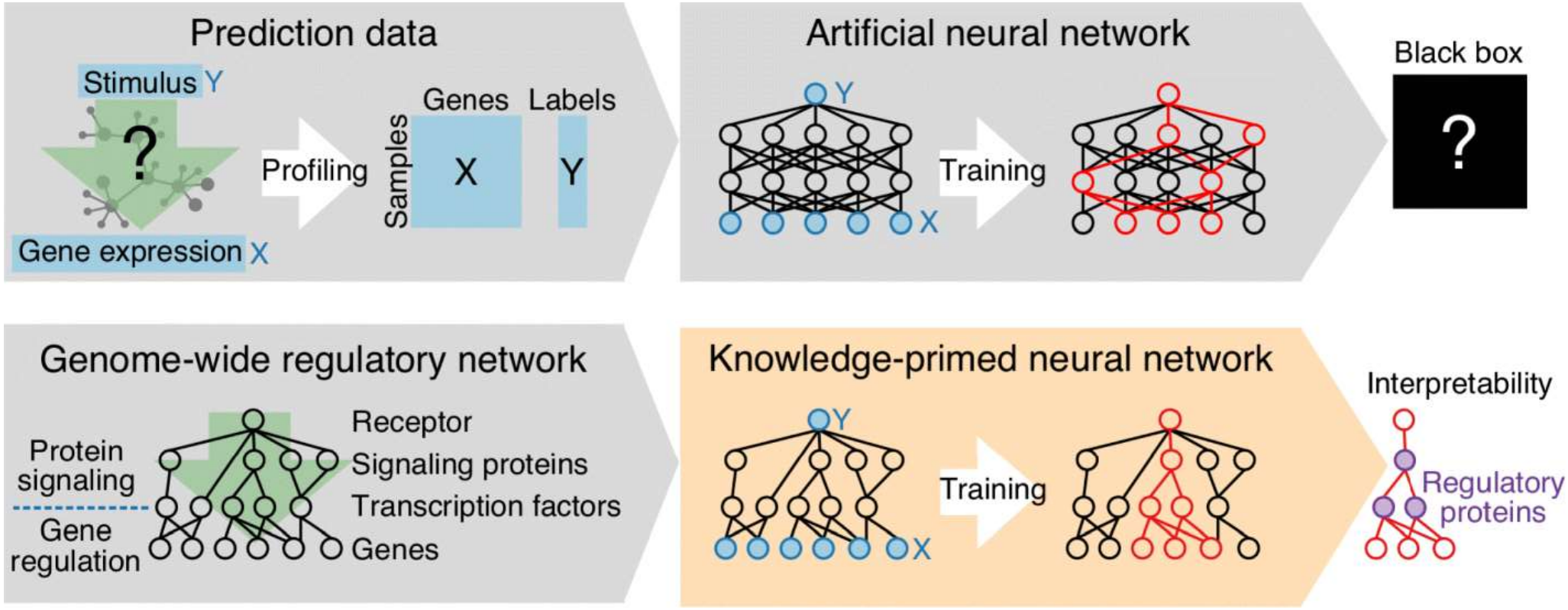
Where are the biomarkers?

Visible neural network (VNN)

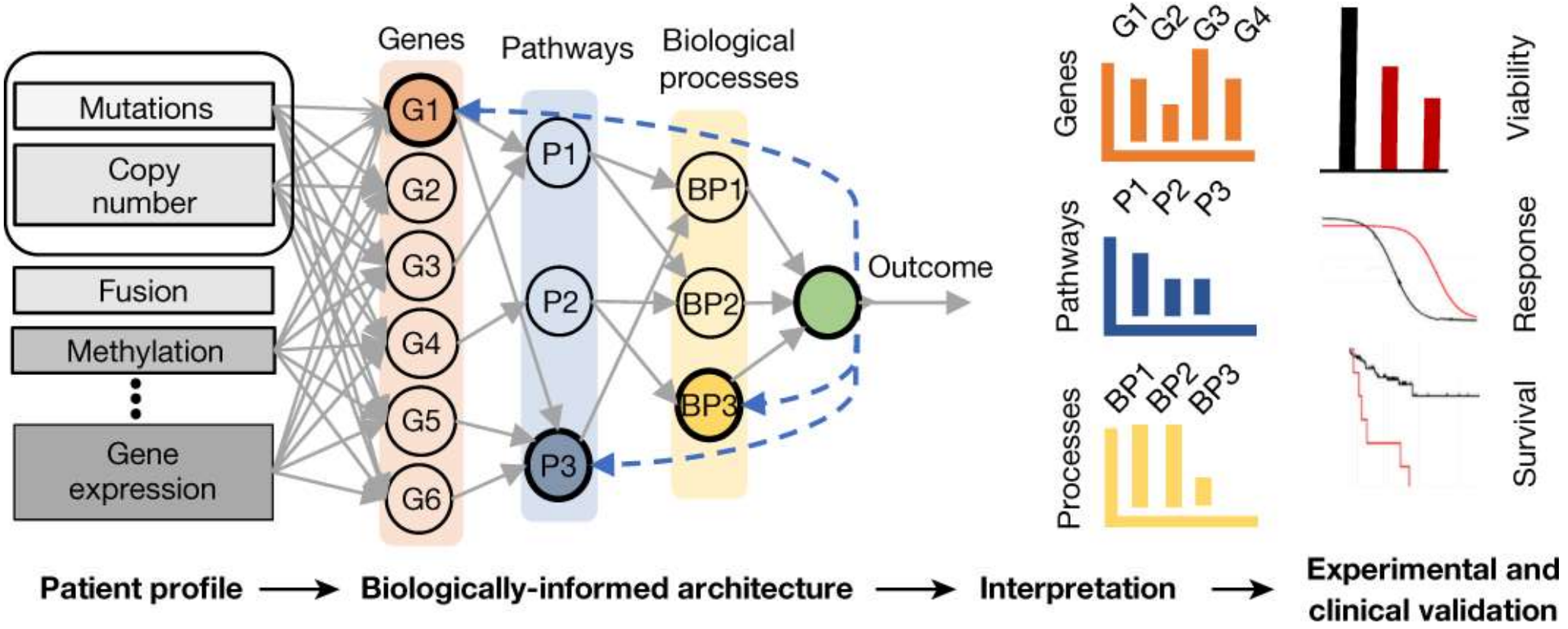


DCell

Knowledge-primed neural network

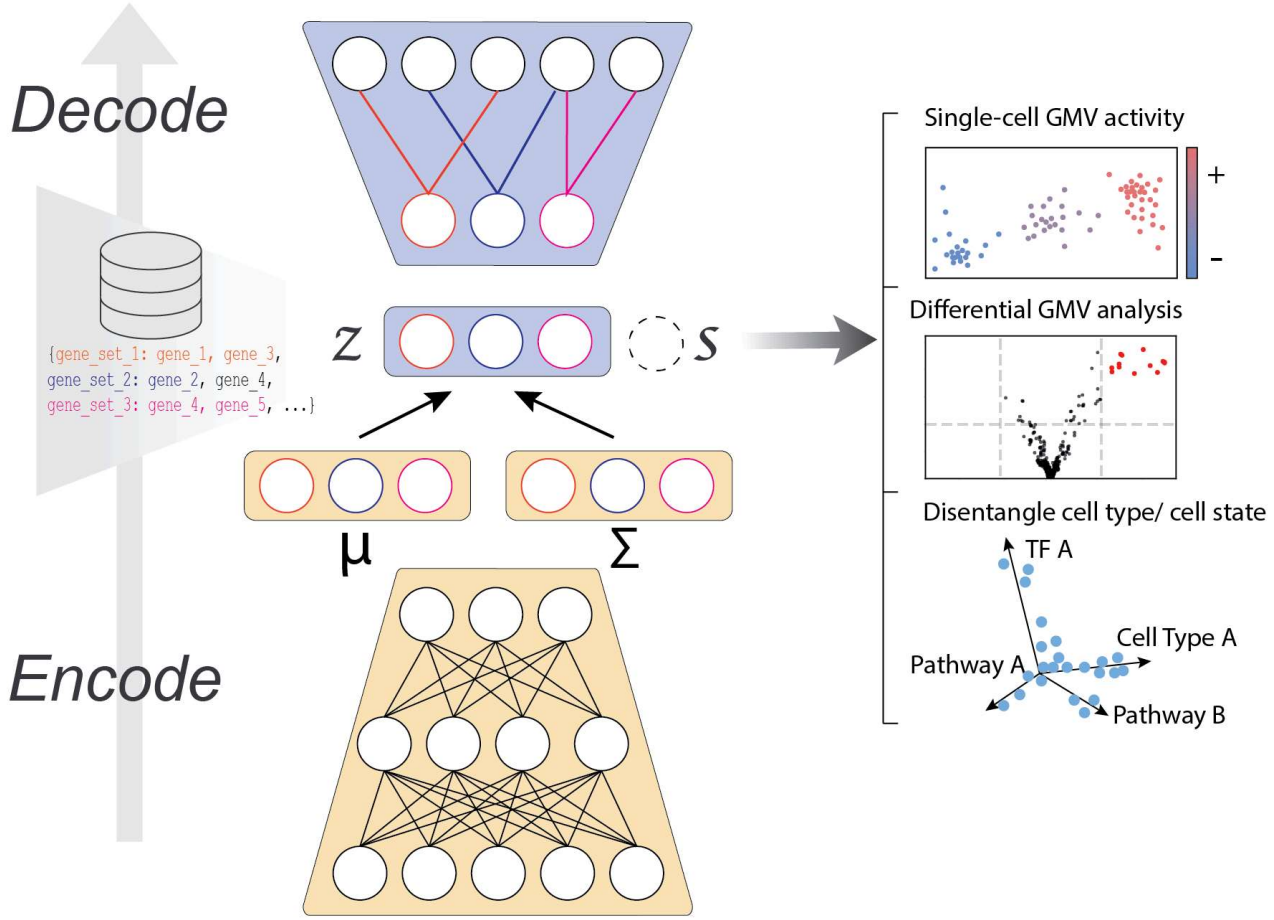


Biologically informed neural network (BINN)



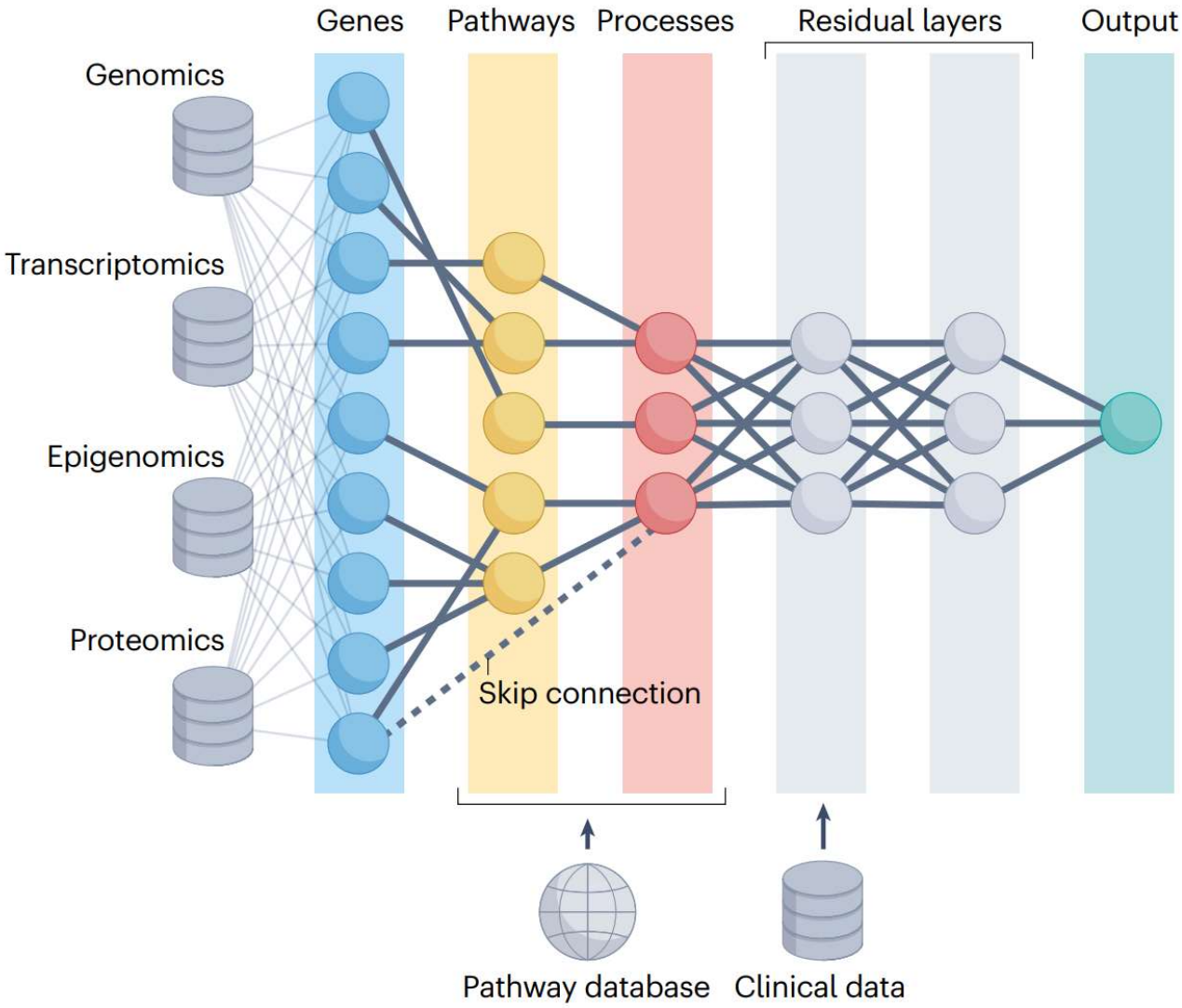
P-Net

VAE enhanced by gene annotations (VEGA)



VEGA

Anatomy of a BINN/VNN

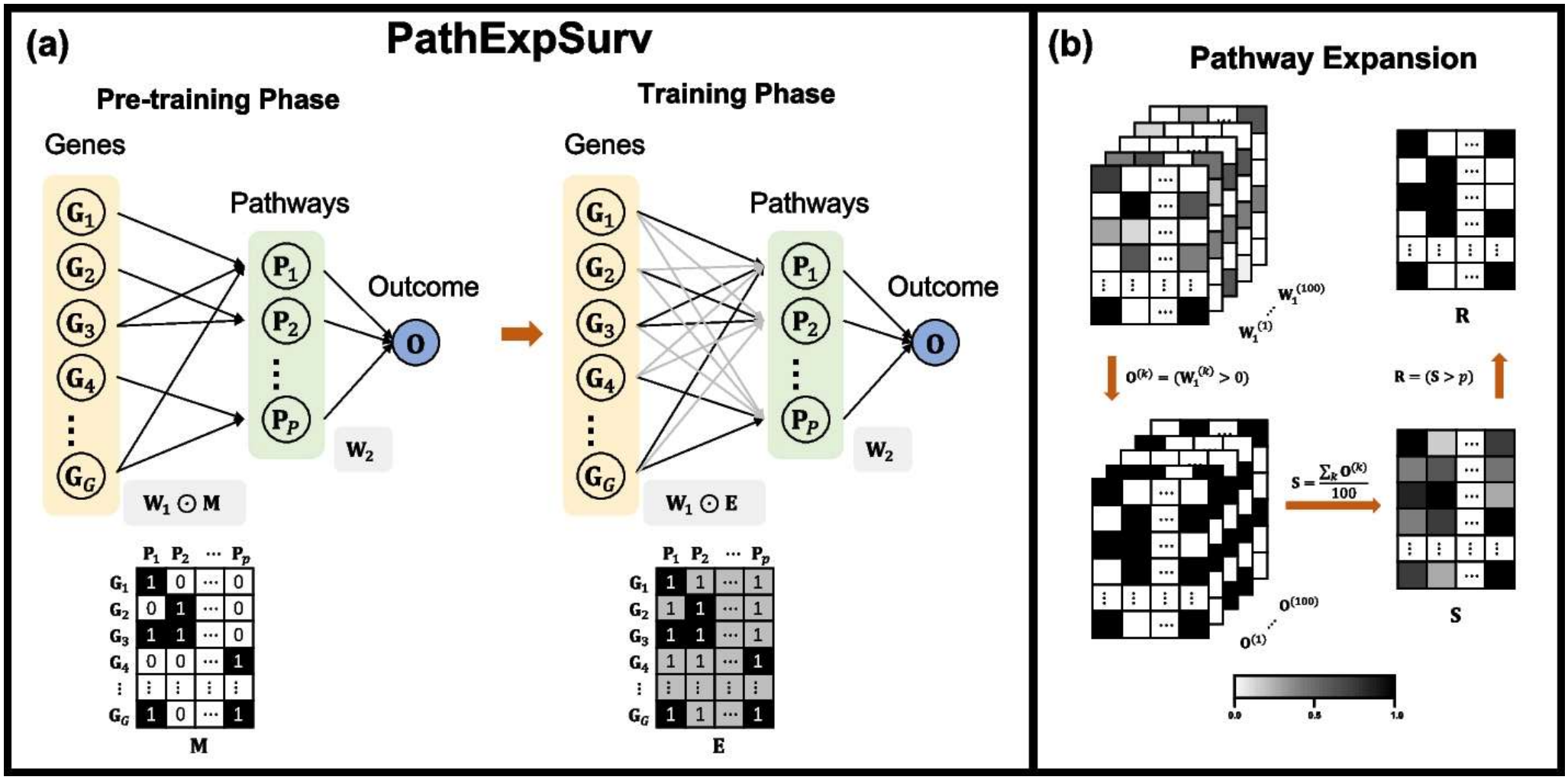


How to build a VNN

1. Start with dense sequential neural network
2. Use adjacency matrix of inputs → pathways as a **masking matrix**
3. (Optional) for next layer, mask = mapping of pathways → higher pathways
4. Omit all masked weights from backpropagation

> ⚠ **Warning**
>
> What if the pathways aren't all the same depth?

Pathway discovery with BINNs



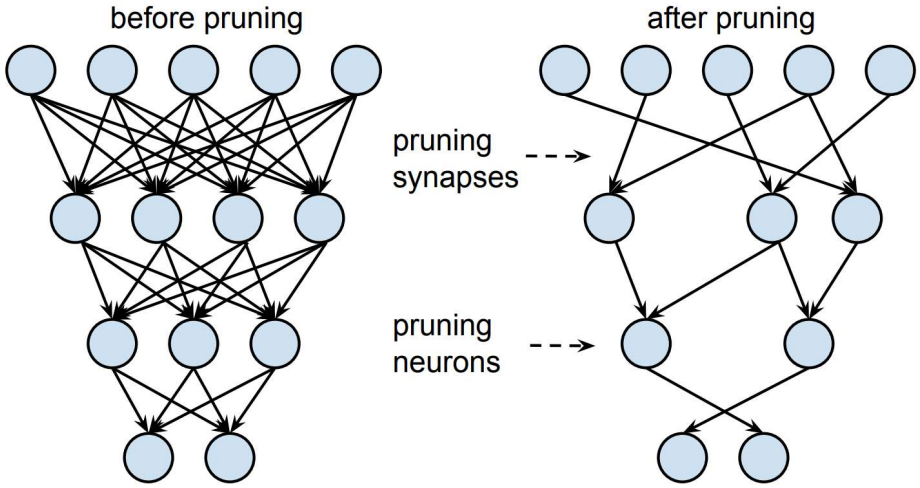
PathExpSurv

Interpretable AutoML discovery

Open question

If (1) VNNs perform better, and (2) Pathway DBs are incomplete; then

Can neural architecture search over VNNs discover new pathways?



5 Hands-on

Accessible tools for BINNs

…would be a very good idea

BINN

- `pip install binn`
- supervised learning
- proteomics or single-omics
- omics data: as `pandas` or csv
- pathways: ‘Reactome’ or own csv (`parent,child`)

VEGA

- `pip install scvega`
- unsupervised learning
- transcriptomics
- omics data: as Anndata
- pathways: GMT from MSigDB

…and various other paper repos you can `git clone`.

Nothing for R… yet!

⚠ What's missing?

- multi-omics
- integration with clinical data
- survival analysis (censored time-to-event)
- R + BioConductor

The binn package

Installation

```
pip install binn==0.1.1          # 0.1.0 has a bug
```

Usage

```
from binn import BINN
binn = BINN(
    data_matrix=input_data,
    network_source="reactome",
    input_source="uniprot",
    n_layers=4,
    dropout=0.2
)
```

Documentation

https://infectionmedicineproteomics.github.io/BINN/

The binn package - inputs

What's required:

`data_matrix`

- $n \times p$ matrix of omics data

`mapping`

- $p \times 2$ 2 column matrix of features to pathways
- or `'uniprot'`

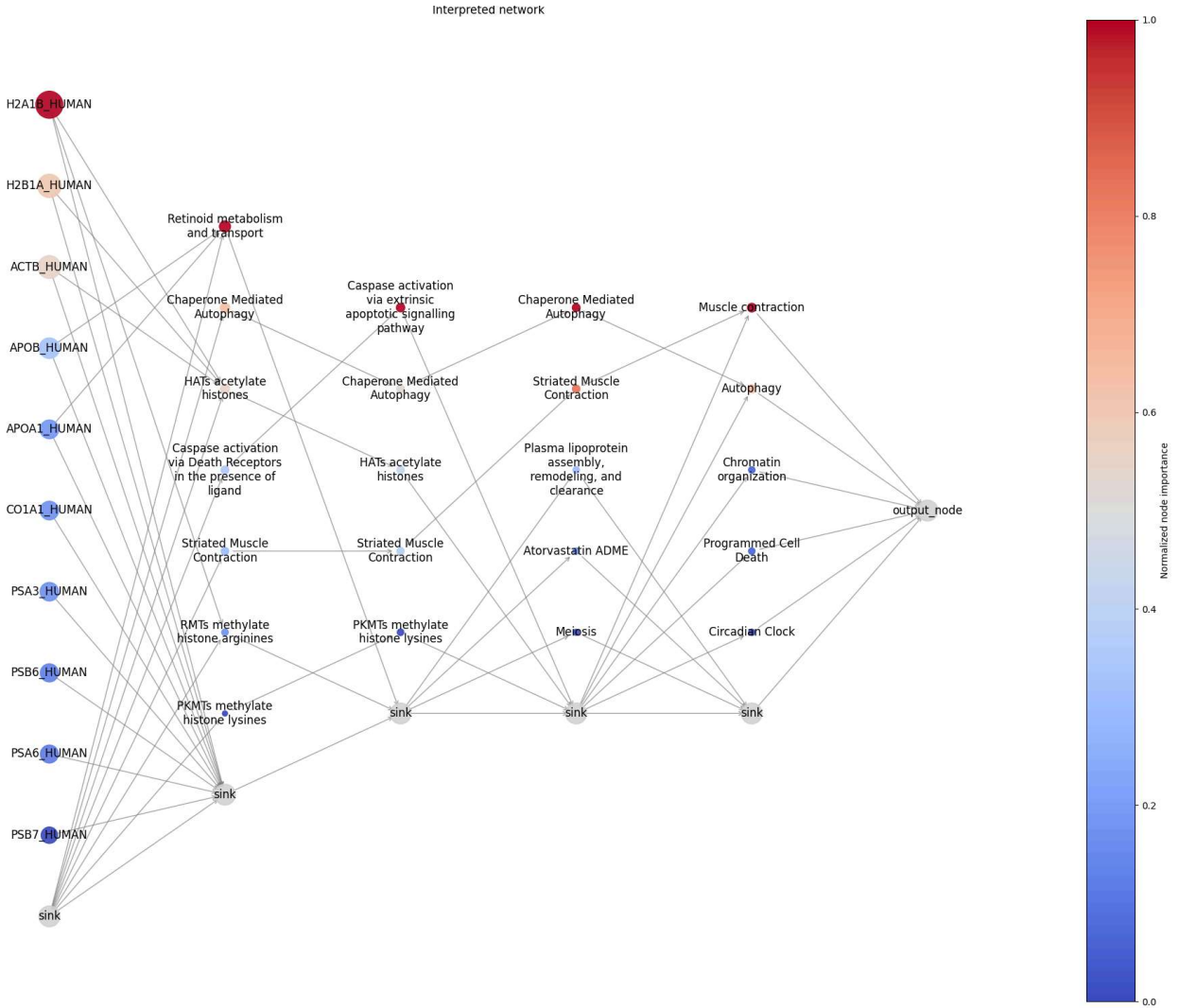
`pathways`

- $? \times 2$ matrix: parent → child pathway
- or `'reactome'`

`design_matrix`

- matrix $n \times 2$ of targets `sample,group`
- `group` = output class

The binn package - visualization



Node importance plot

Worksheet

Colab notebook and slides:
github.com/datasciapps/trifels2025

6 Discussion

Challenges

- **Lack of Robust Tools**: Few standardized frameworks, poor interpretability and limited benchmarks.
- **Alternative Models**:
  - **Classical methods**: faster, more interpretable, statistically robust
  - **GNNs**: Better structural representation but need high-quality pathway graphs.
  - **Transformers/LLMs**: Strong sequence modeling but lack explicit pathway structure.
  - **Hybrid Approaches**: Combining VNNs with GNNs or transformers may improve learning.
- **Pathway Databases Are Incomplete**: KEGG, Reactome, and GO suffer from curation bias, lack context specificity, and may be outdated.

Open questions

Biological fidelity

- Is a BINN really a digital twin of a cell?
- Biological sparsity or just noisy sparsity?

Model validation

- Synthetic and experimental data for validation
- Uncertainty-aware architectures to quantify confidence in biological predictions.

Performance vs interpretability

- Mechanistic insight vs. predictive accuracy: which should take priority?
- Do black-box models merely reinforce existing biases in pathway databases?

Things we didn’t cover today

- Causal inference
- Dynamic updating of knowledge graphs
- Bayesian prior elicitation
- GenAI: LLM agents & retrieval-augmented generation

Thank you!

github.com/datasciapps/trifels2025
david.selby@dfki.de

Further reading

Selby, D.A. et al. Beyond the black box with biologically informed neural networks. *Nature Reviews Genetics* (2025). doi:10.1038/s41576-025-00826-1