Knowledge-enhanced biomarker discovery

Trifels Spring School 2025: AI in Bioinformatics

David Selby

24th March 2025

DFKI

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

Introduction & Motivation

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

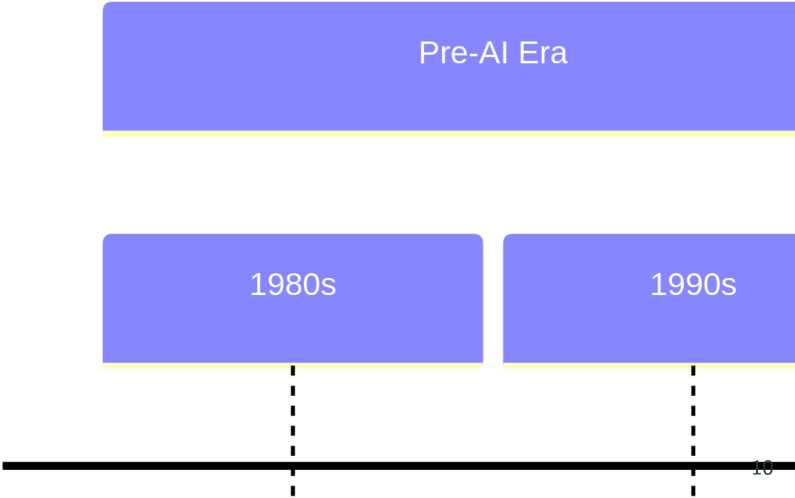
Knowledge enhanced (multi-omics) **biomarker discovery**

Why? Where? How?

What makes it hard?

What are we looking for?

Krassowski at al. (2020). State of the Field in Multi-Omics Research. *Frontiers in Genetics.* doi:10.3389/fgene.2020.610798



Pre-AI Era

1980s

1990s

What is prior knowledge?

- scientific publications in literature

What is prior knowledge?

- scientific publications in literature
- open datasets (e.g. TCGA, OpenML, UCI)

What is prior knowledge?

- scientific publications in literature
- open datasets (e.g. TCGA, OpenML, UCI)
- domain-specific databases (e.g. KEGG, Reactome, GO)

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)

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

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)

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)

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?

What is multi-omics biomarker discovery?

- Combines genomics, transcriptomics, proteomics, metabolomics, ...
- Identify robust signatures for disease diagnosis, prognosis or treatment

Sign

curATime project

A bit of motivation

What is multi-omics?

Definition Integration of genomics, transcriptomics, proteomics, metabolomics, epigenomics, ...

Rationale Capture complementary processes to improve biomarker robustness

Challenges Data heterogeneity, modality-specific noise; batch effects and small sample sizes; alignment of feature spaces

Opportunities Advanced techniques to enhance interpretability, reproducibility

Prior knowledge

Approaches

1. Knowledge graphs
2. Regularization
3. Biologically-informed neural networks

- **Knowledge Graphs:**
  - Encode relationships between genes, pathways, and diseases
  - Use graph convolutional networks for structured representation
- **Regularization Strategies:**
  - Incorporate pathway-level priors in loss functions
  - Penalize biologically implausible connections in high-dimensional space
- **Biologically-Informed Neural Networks:**
  - Architectures that enforce modularity reflecting known biology
  - Example: Visible neural networks where hidden nodes map to biological entities
- **Benchmarking:**
  - Compare performance on independent cohorts

Multi-omics integration

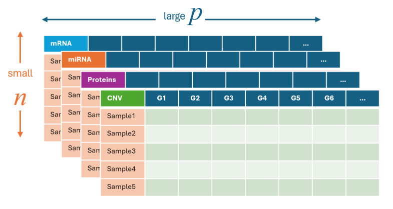
Single-omics

Classical techniques

Regularized regression

Batch effect correction

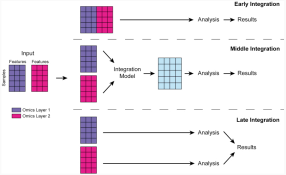
Multi-omics datasets



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

Multimodal fusion

When should we combine omics layers?



Cai, Poulos, Liu & Zhong. Machine learning for multi-omics data integration in cancer. *iScience.* (2022). doi:10.1016/j.isci.2022.103798

Multimodal fusion

When should we combine omics layers?

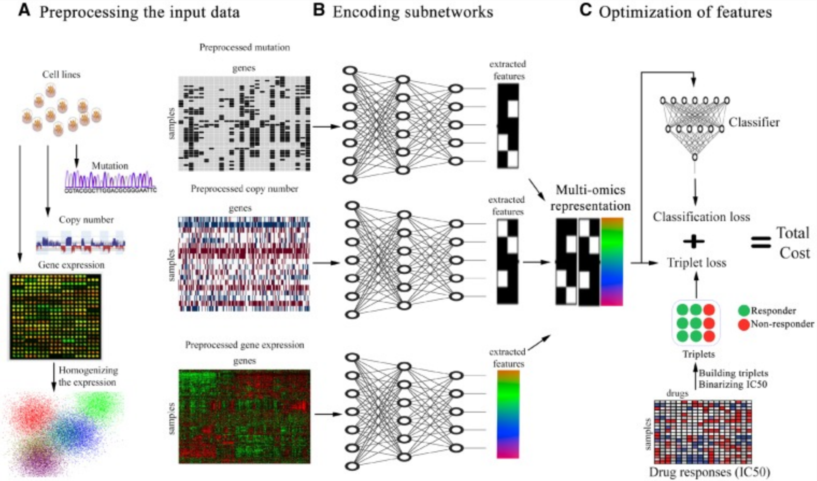
Early easier, loss of information, worse performance*

Intermediate (mixed, joint) modality-specific layers, but harder to train

Late may not capture interactions

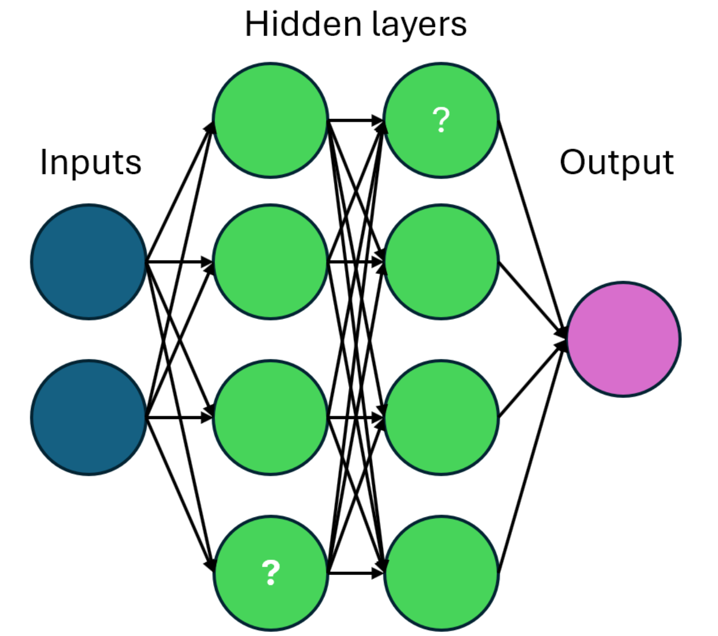
*Hauptmann, T., Kramer, S. A fair experimental comparison of neural network architectures for latent representations of multi-omics for drug response prediction. *BMC Bioinformatics* 24, 45 (2023). doi:10.1186/s12859-023-05166-7

Late fusion (MOLI)



Sharifi-Noghabi et al. *Bioinformatics*. 2019. doi: 10.1093/bioinformatics/btz318

Model explanations



Model explanations

Input-level explanations:

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

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

reactome

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:

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

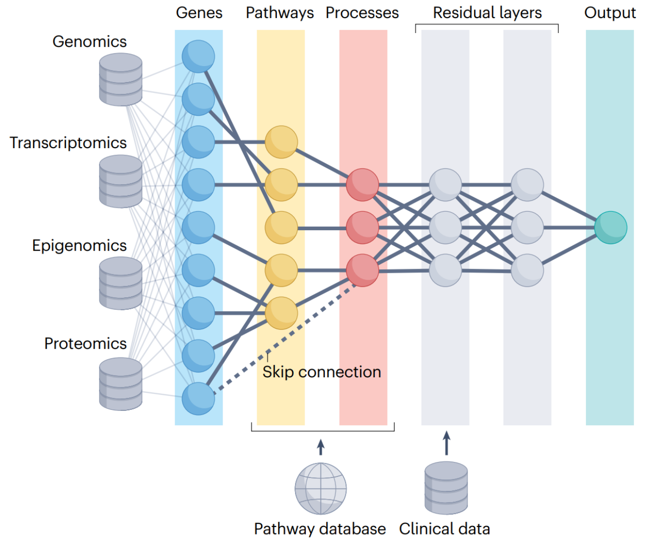
with p-value

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

Multi-omics integration

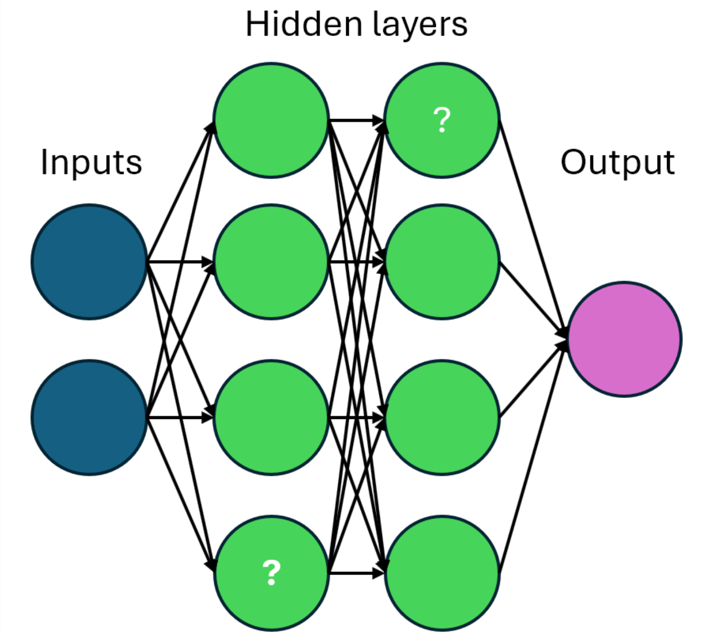
Multi-omics integration

Visible neural networks



Selby, D.A. et al. *Nat Rev Genet* (2025). doi:10.1038/s41576-025-00826-1

Feedforward neural network



Hands-on session

Worksheet

- **Workflow Overview:**
  - Data preprocessing and integration in a reproducible Colab notebook
  - Training of biologically-informed neural networks with clear hyperparameter tuning
  - Post-hoc interpretation using DeepLIFT, SHAP, and GSEA
- **Code Walkthrough:**
  - Annotated code snippets emphasizing reproducible pipelines
  - Discussion on containerization and version control for reproducible research
- **Live Demo:**
  - Execute a minimal example on multi-omics data to illustrate integration and interpretation steps

The `binn` package

Discussion

Things we didn’t cover today

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

Challenges & Future Directions

- **Limitations:**
  - Residual uncertainty in integrating diverse modalities
  - Interpretability challenges in highly complex models
  - Potential biases in available biological knowledge bases
- **Future Directions:**
  - Integration of causal inference techniques
  - Dynamic updating of knowledge graphs as new data emerges
  - Scaling to larger, more diverse cohorts to validate reproducibility
- **Open Questions for Debate:**
  - How to balance model complexity with biological interpretability?
  - What standards ensure FAIRness in rapidly evolving multi-omics workflows?

Thank you!

Thank you!

github.com/datasciapps/trifels2025

Contact: david.selby@dfki.de

Further reading:

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