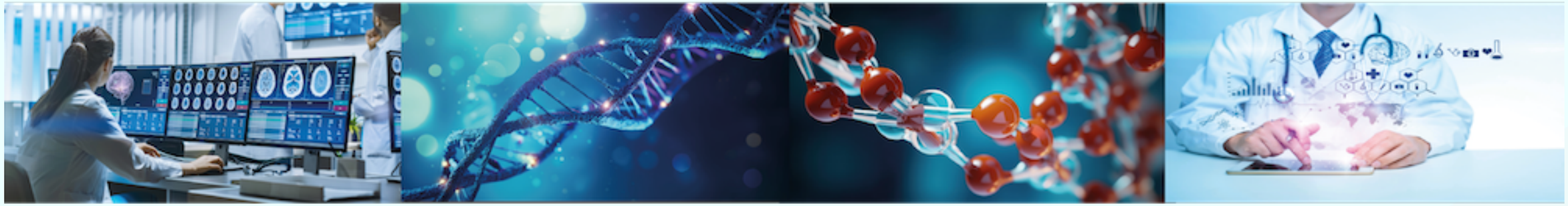


Programming for Biomedical Informatics

Lecture 17 “Working with Multiple Data Modalities”

https://github.com/tisimpson/pbi

Ian Simpson
ian.simpson@ed.ac.uk

THE UNIVERSITY of EDINBURGH
informatics

Opportunities & Challenges in Biomedical Informatics

Opportunities

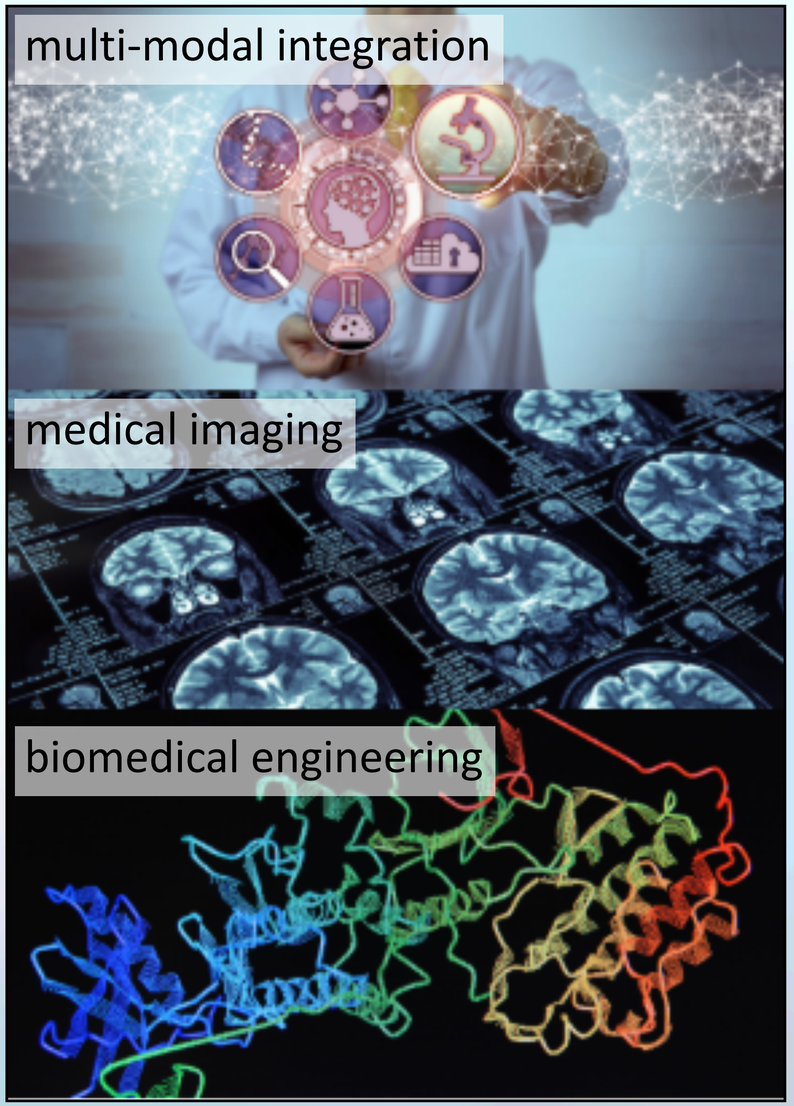
Clinical & Health
- Administration Support
- Decision Support
- Patient Engagement
- Synthetic Data Generation
- Clinical Trial Design & Monitoring
- Population Level Modelling
- Professional Education

Biomedical Science
- Drug Discovery and Design
- Protein Structure Prediction
- Biomedical Image Synthesis
- Patient Data Generation
- Drug Response Prediction
- Biological Sequence Generation
- Medical Text Generation
- Biomedical Signal Generation
- Disease Progression Modeling

Challenges

Technical Challenges
- Unlabelled & Unstructured Data
- Missing Values
- Model & Data Bias
- Poor Longitudinal Coverage
- Scaling Problems
- Lack of Realistic Evaluation Benchmarks
- Explainability
- Data Availability & Inter-Operability

Societal & Health Systems
- Clincal safety, Efficacy, & Reliability
- Evaluation, Regulation, & Certification
- Privacy
- Copyright & Ownership
- Implementation & Adoption



A Cornucopia of Biomedical Data

Clinical Data
Electronic Health Records (EHRs): Patient-level data including medical history, diagnoses, treatments, medications, lab results, and imaging.
Claims Data: Billing and insurance data capturing diagnoses, procedures, and healthcare costs.
Reported Outcomes (PROs): Information provided directly by patients about their health conditions, symptoms, or quality of life.
Clinical Trials Data: Structured data from clinical studies, including intervention details, outcomes, and adverse

Omics Data
Genomics: DNA sequence data (e.g., whole genome or exome sequencing) capturing genetic variations.
Transcriptomics: Gene expression profiles from RNA sequencing or microarrays.
Proteomics: Data on protein expression and post-translational modifications.
Metabolomics: Small molecule metabolites involved in biochemical processes.
Epigenomics: Data on DNA methylation, histone modifications, or chromatin accessibility.

Imaging Data
Radiology Images: X-rays, CT scans, MRIs, PET scans used for diagnostics.
Pathology Images: Digital scans of tissue samples from biopsies or surgeries.
Microscopy Images: Cellular and sub-cellular imaging for research purposes (e.g., fluorescence microscopy).

Epidemiological Data
Population Health Surveys: large cohorts demographics, health behaviours, and disease prevalence.
Disease Registries: registries tracking specific conditions (e.g., cancer registries).
Public Health Surveillance Data: disease outbreaks, vaccination rates, and other health indicators.

Wearable and Sensor Data
Activity Monitors: Data from fitness trackers or smartwatches on physical activity, heart rate, and sleep patterns.
Continuous Monitoring Devices: Blood glucose monitors, ECGs, and other wearable medical devices.

Environmental and Social Determinants Data
Environmental Exposure Data: Air quality, pollution, climate data linked to health outcomes.
Socioeconomic Data: Income, education, and other social factors influencing health.

Biomedical Literature and Knowledge Graphs
Text Data: PubMed abstracts, clinical guidelines, and research articles.
Ontologies and Knowledge Graphs: Structured representations of biomedical knowledge (e.g., UMLS, SNOMED-CT, and GO).

Behavioural and Lifestyle Data
Survey Data: Self-reported behaviours such as smoking, diet, or exercise.
Digital Health Data: Data from apps tracking behaviours like calorie intake or mental health.

Biobanks and Bio-specimen Data
Sample Metadata: Information about biological samples (e.g., blood, saliva, tissues) stored in biobanks.
Linked Clinical Data: Associated patient characteristics and health outcomes.

Pharmacological Data
Drug Databases: Information about drugs, their mechanisms, side effects, and interactions (e.g., DrugBank).
Prescription Data: Patterns of medication use across populations.

Synthetic and Simulated Data
Synthetic Datasets: Artificially generated data mimicking real-world datasets for training models without compromising privacy.
Modeling and Simulation Data: In silico experiments to model biological systems or predict outcomes.

Why Integrate Data?

Advantages

Holistic View of Biological Systems
Combines diverse data types (e.g., genomics, proteomics, and clinical data) to provide a comprehensive understanding of complex biological processes.

Improved Disease Mechanism Discovery
Links molecular-level data (e.g., gene expression) with phenotypic data (e.g., disease symptoms) to uncover disease pathways and biomarkers.

Enhanced Predictive Models
Increases the accuracy and robustness of predictive models by incorporating multi-dimensional data.

Cross-Validation of Findings
Enables verification of results across different datasets, increasing confidence in findings.

Identification of Novel Patterns
Facilitates the discovery of relationships and patterns not evident within individual datasets.

Personalised Medicine
Supports tailored healthcare approaches by integrating genomic, clinical, and lifestyle data for individualized treatment.

Efficient Use of Resources
Leverages existing datasets, reducing the need for redundant experiments.

Support for Multi-Scale Analysis
Bridges scales of biology, from molecular interactions to organ-level and population-wide studies.

Disadvantages

Data Heterogeneity
- Different data types (e.g., genomic vs. clinical) may have varying formats, scales, and resolutions, making integration complex.
- Integrated data may contain errors or missing values from individual sources, compounding the problem in the combined dataset.
- Lack of universal standards for data collection and annotation complicates integration across studies or institutions.
- Integrating data may introduce artificial patterns or average out real biological signal.

Complexity of Interpretation
Multi-modal data can produce results that are difficult to interpret, requiring domain expertise across multiple fields.

Data Availability and Access
Limited access to proprietary or sensitive datasets can restrict integration efforts.

Computational Challenges
Requires significant computational power and advanced algorithms to handle large, multi-dimensional datasets.

Bias
Integration may amplify biases inherent in individual datasets (e.g., population-specific biases).

Privacy and Ethical Concerns
Combining sensitive data (e.g., genomic and clinical information) increases the risk of privacy breaches and ethical issues.

Cost and Resource Intensive
Collecting, processing, and integrating diverse datasets often requires significant financial and human resources.

Gene Correlation Network

BMC Bioinformatics

BioMed Central

Software

Open Access

WGCNA: an R package for weighted correlation network analysis

Peter Langfelder[1] and Steve Horvath*[2]

Address: [1]Department of Human Genetics, University of California, Los Angeles, CA 90095, USA and [2]Department of Human Genetics and Department of Biostatistics, University of California, Los Angeles, CA 90095, USA

Email: Peter Langfelder - Peter.Langfelder@gmail.com; Steve Horvath* - shorvath@mednet.ucla.edu

* Corresponding author

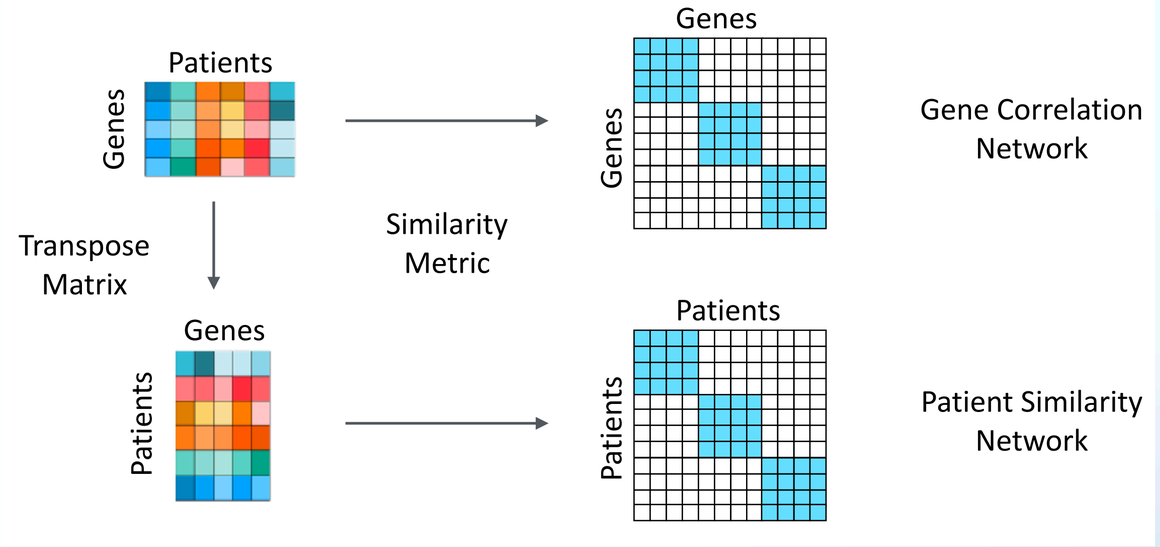
Published: 29 December 2008

BMC Bioinformatics 2008, 9:559 doi:10.1186/1471-2105-9-559

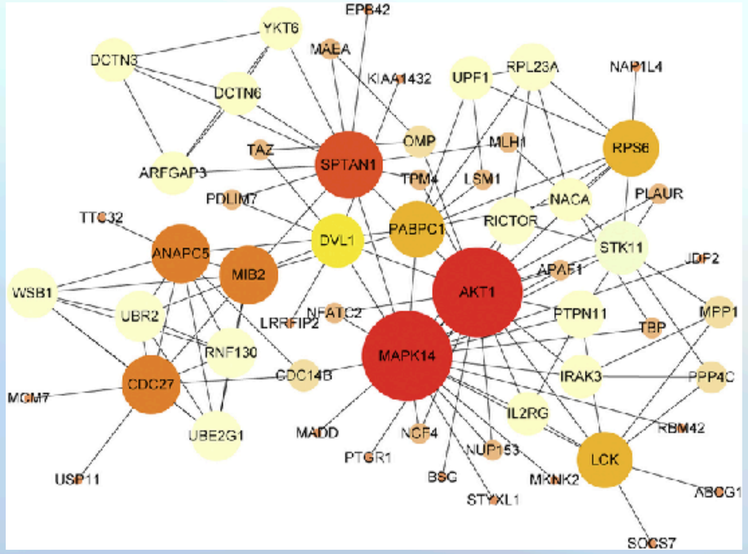
Received: 24 July 2008

Accepted: 29 December 2008

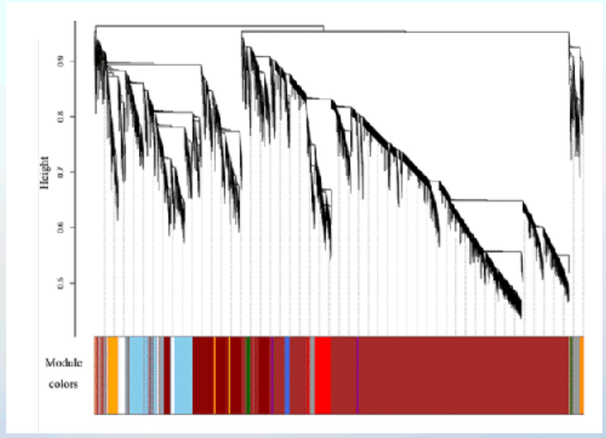
This article is available from: http://www.biomedcentral.com/1471-2105/9/559



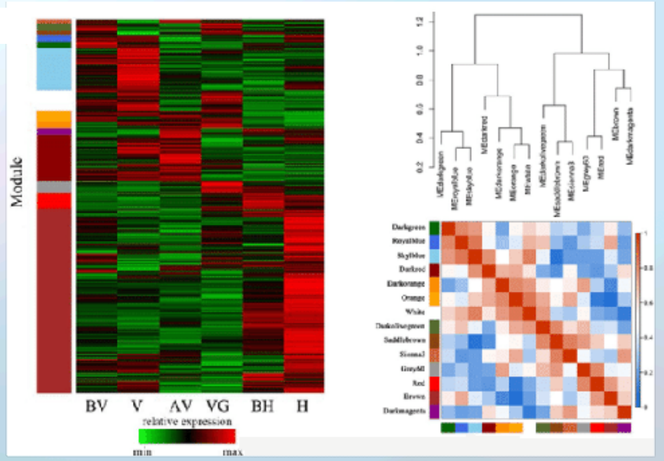
gene correlation network



network clustering

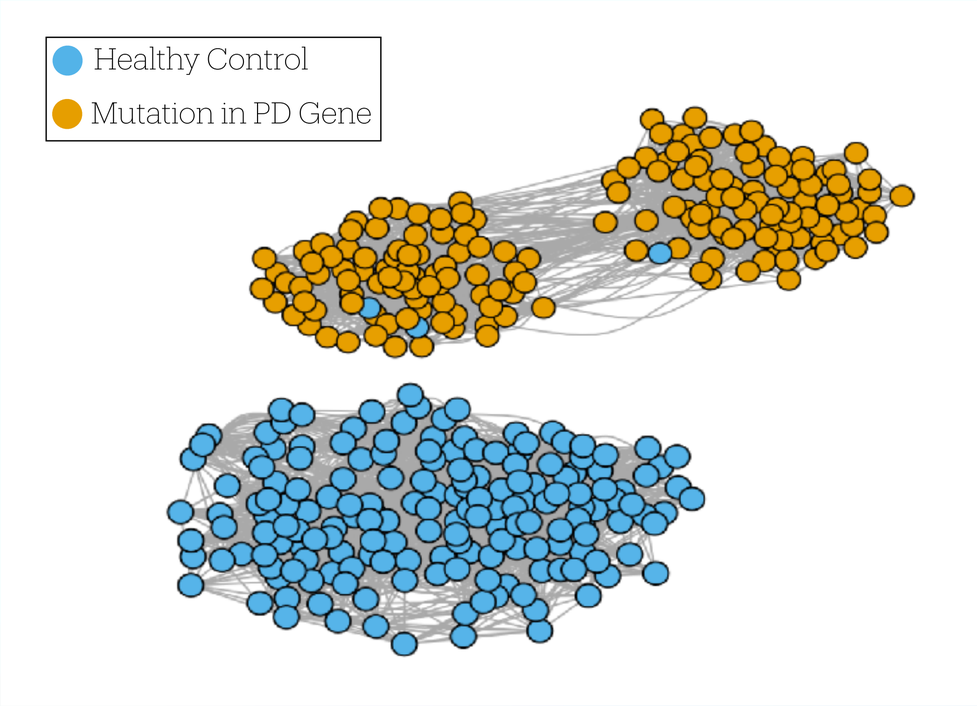


module profiling



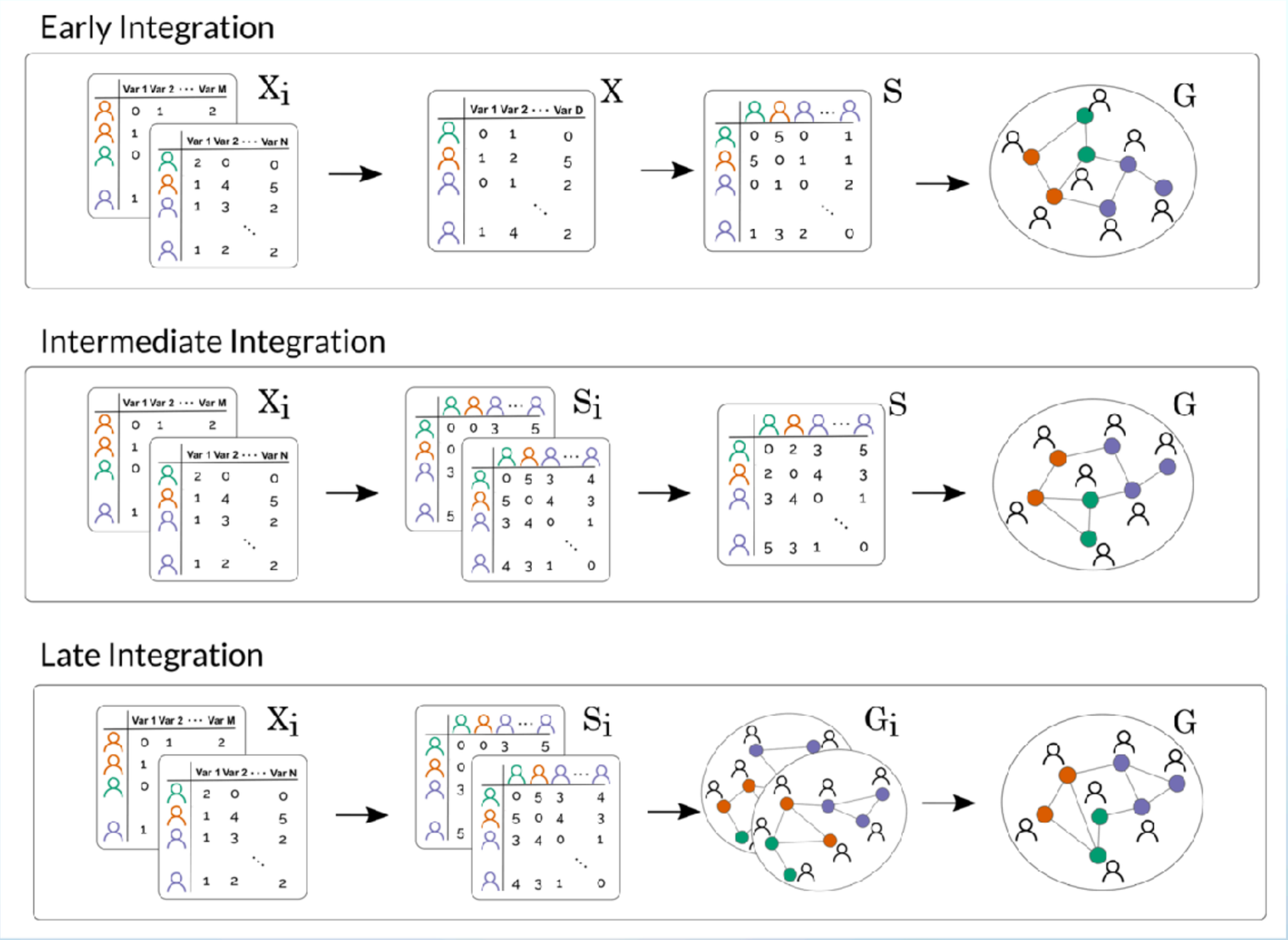
Patient Similarity Networks (PSNs)

- Networks where participants are represented as nodes connected by edges where associated features share similarities
- Useful way to represent commonly unstructured data with common reference point
- Networks are constructed in a variety of ways depending on the underlying data modality, type, and distribution
- Motivation for patient similarity
  - Modelling at the level of an individual
  - Mimics comparative practice within and between populations
  - Scalable
  - Strong predictive power
  - Interpretable



Similarity Network of participants in the PPMI (Parkinson's Disease) study using selected SNPs. Groups segregate by causative alleles using SNPs alone.

Strategies for Multi-Modal Network Integration



Algorithmic Approaches to Network Integration

- **Similarity Network Fusion (SNF)** — For each pairwise modality distance $S^{(K)}$, the pairwise value between a node *i* with *NaN* in X_k and any node *j* is set to max distance/dissimilarity for that modality. SNF is then computed as normal with max dissimilarity included.

- **NEighborhood Based Multi-Omic Clustering (NEMO)** — NEMO was developed to analyse partial data. The mean relative similarity for any pair of nodes *i* and *j* is computed over the modalities where both nodes have recorded data.

- **Concatenated** X_i — Feature mean value imputation in X_k for all nodes with *NaN* values. Then distance/similarity calculated as normal.

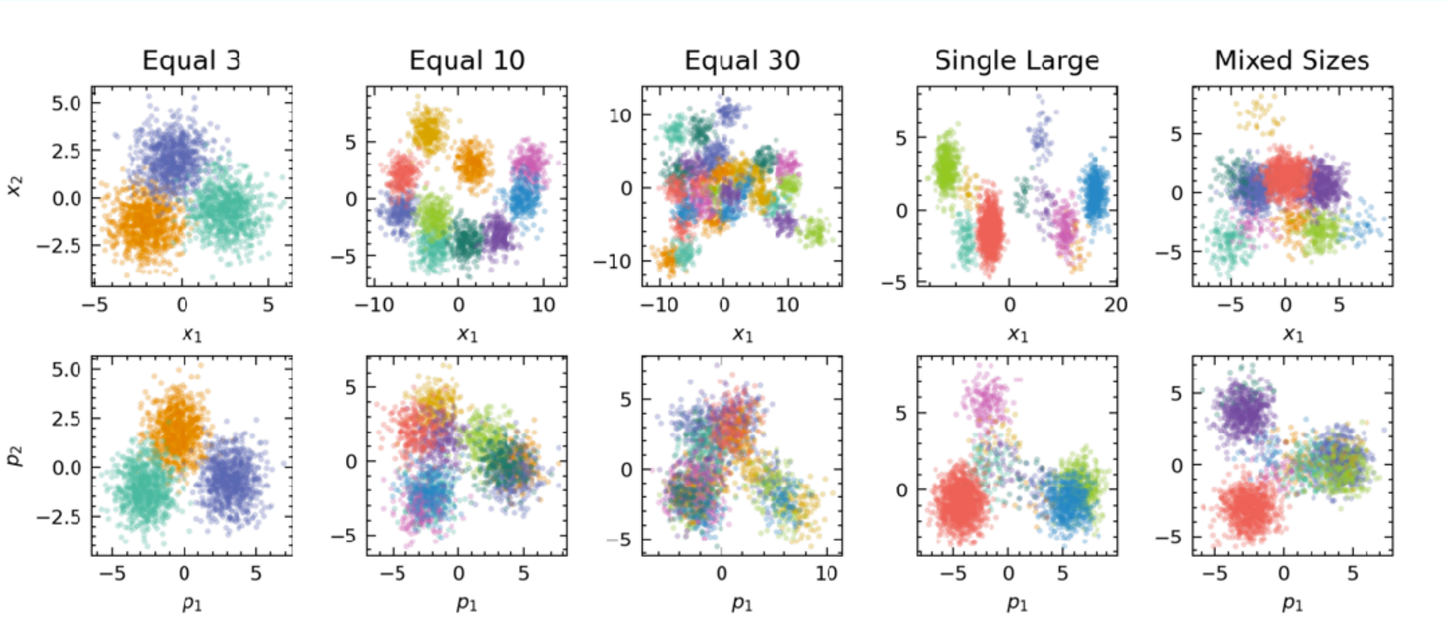
- **Mean** S_i **imputing *Max*** — For each pairwise modality distance $S^{(K)}$, the pairwise value between a node *i* with *NaN* in X_k and any node *j* is set to max distance/dissimilarity for that modality. Mean similarity then computed between a pair of nodes *i* and *j* across all modalities.

- **Mean** S_i **ignoring** *NaN* — The mean similarity for any pair of nodes *i* and *j* is computed over the modalities where both nodes have recorded data.

- **Extreme Mean** — Thresholding is performed on the pairwise similarity between nodes with recorded values in the modality. The mean similarity for any pair of nodes *i* and *j* is computed over the modalities where both nodes have recorded data. If all values between *i* and *j* are *NaN* after thresholding (including *NaN* for where *i* has no recorded data in a modality) then the dissimilarity is set to max.
-

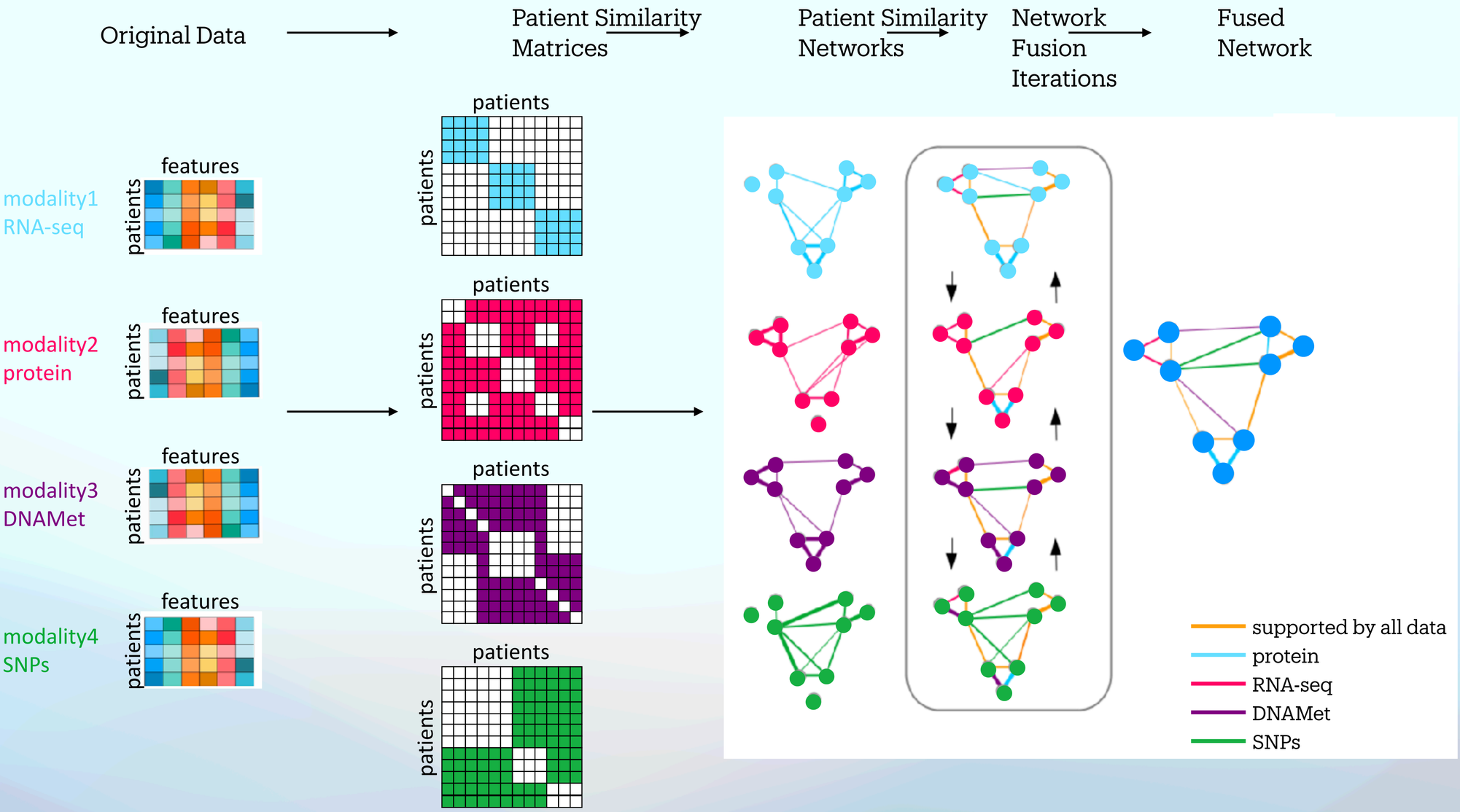
Fusion Method Performance on Clustering of Ground Truth Networks

Example Ground Truth Distributions



Modality Problem	Easy		Single Merged		Merged		Split		1Rand		Mixed 1Rand		Mixed Noisy	
	Max	Mean	Max	Mean	Max	Mean	Max	Mean	Max	Mean	Max	Mean	Max	Mean
Graph														
SNF	**0.998**	0.968	0.997	0.906	0.941	0.696	**0.998**	**0.936**	0.987	**0.935**	**0.837**	0.720	0.674	0.546
NEMO	**0.999**	**0.981**	0.985	0.818	0.921	0.651	**0.998**	**0.937**	0.972	0.911	0.794	**0.724**	0.644	0.557
Mean S_i	**1.000**	0.976	**1.000**	**0.937**	**0.983**	**0.778**	0.982	0.688	**0.995**	0.755	0.825	0.571	**0.797**	**0.562**
Concatenated X_i	**1.000**	0.975	**1.000**	**0.938**	**0.982**	0.772	0.958	0.666	0.993	0.750	0.813	0.564	0.647	0.471
Extreme Mean	0.896	0.660	0.781	0.658	0.717	0.556	0.906	0.677	0.749	0.542	0.603	0.463	0.572	0.470

Multi-Modal Integration Through Similarity Network Fusion

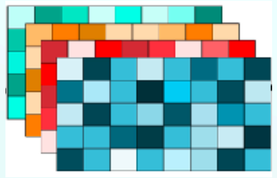


after Wang, B. et al. Similarity network fusion for aggregating data types on a genomic scale. Nat Methods 11, 333–337 (2014).

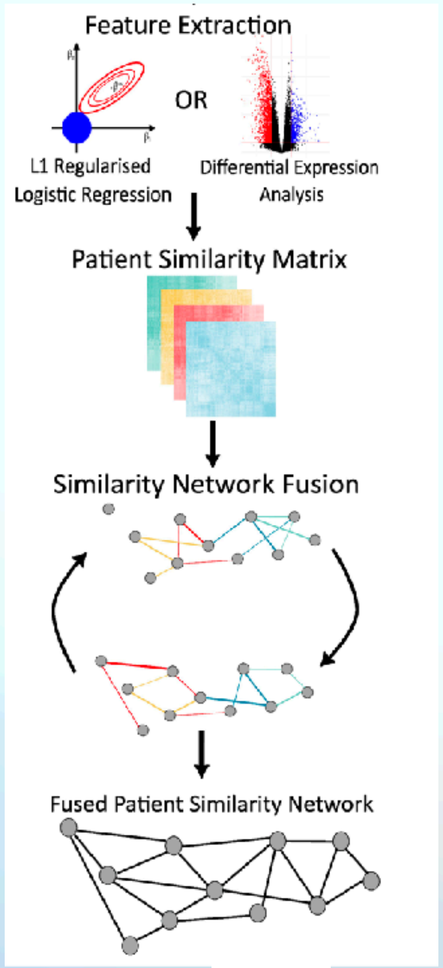
Cohort Stratification Using Multi-Modal Fused Patient Similarity Networks

generate fused patient network

multi-omic data

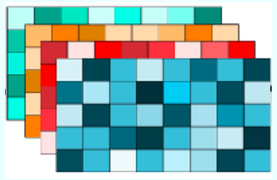


mRNA
miRNA
DNA-Met
CNV
RPPA

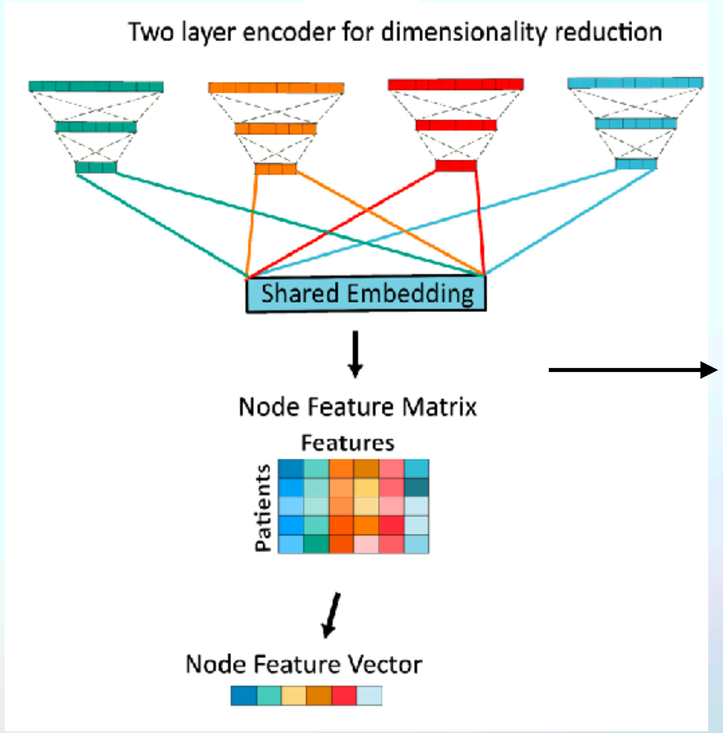


feature embedding

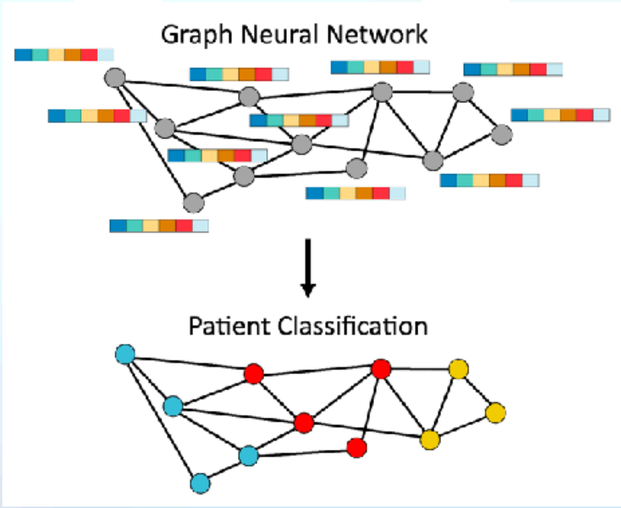
multi-omic data



mRNA
miRNA
DNA-Met
CNV
RPPA



GCN training & classification



Barry Ryan, Riccardo E Marioni, T Ian Simpson, Multi-Omic Graph Diagnosis (MOGDx): a data integration tool to perform classification tasks for heterogeneous diseases. Bioinformatics, Volume 40, Issue 9, September 2024, btae523, https://doi.org/10.1093/bioinformatics/btae523

Cancer Classification from Multi-Modal Fused Patient Networks

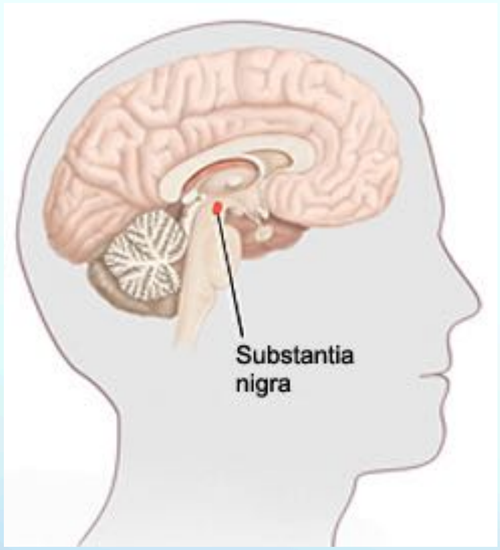
Method	Dataset	Number of Modalities	Number of Samples	Number of Classes	Accuracy	F1
MOGDx	BRCA	4	1083	5	0.893 ± 0.014	0.874 ± 0.012
	BRCA	4	1043	4	0.904 ± 0.014	0.887 ± 0.016
	LGG	1	457	2	0.899 ± 0.016	0.881 ± 0.019
	KIPAN	4	888	3	0.958 ± 0.003	0.948 ± 0.004
MOGONET	BRCA	3	875	5	0.829 ± 0.018	0.825 ± 0.016
	LGG	3	510	2	0.816 ± 0.016	0.814 ± 0.014
	KIPAN	3	658	3	0.999 ± 0.002	0.999 ± 0.002
MoGCN	BRCA	3	511	4	0.898 ± 0.025	0.902 ± 0.024
	KIPAN	3	698	3	0.977 ± 0.017	0.977 ± 0.017
SVM	BRCA	1	869	5	0.782 ± 0.033	0.721 ± 0.030
Lasso	BRCA	1	1047	5	0.829 ± 0.014	0.771 ± 0.012
XGBoost	BRCA	1	1047	5	0.762 ± 0.036	0.692 ± 0.033

The optimal MOGDx performance is shown for each dataset. 4 of 5 available modalities were used for both BRCA and KIPAN. Only DNAm was used on the LGG dataset as it achieved the best accuracy while still including maximum number of samples. The performance reported by MOGONET(Wang et al., 2021) was achieved using mRNA, miRNA and DNAm. The performance reported by MoGCN(Li et al., 2022) was achieved using CNV, mRNA and RPPA. The performances reported on SVM, Lasso and XGBoost methods were achieved using the omic measure which gave the highest accuracy.

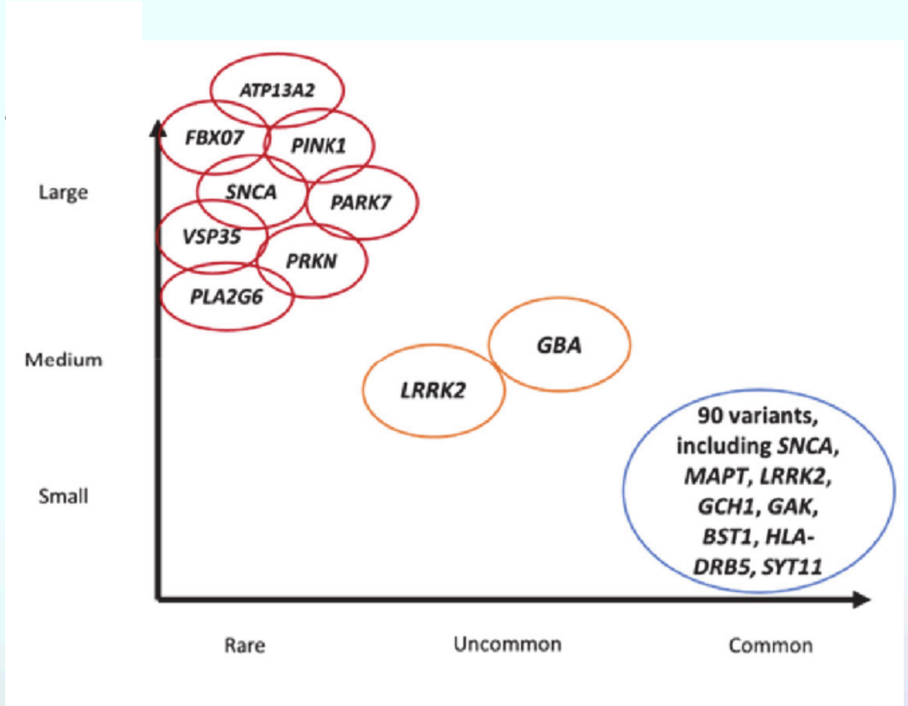
Ryan B, Marioni RE, Simpson TI. Multi-Omic Graph Diagnosis (MOGDx) : A data integration tool to perform classification tasks for heterogeneous diseases. medRxiv 2024. https://doi.org/10.1101/2023.07.09.23292410v2

Temporal Classification for Parkinson's Disease

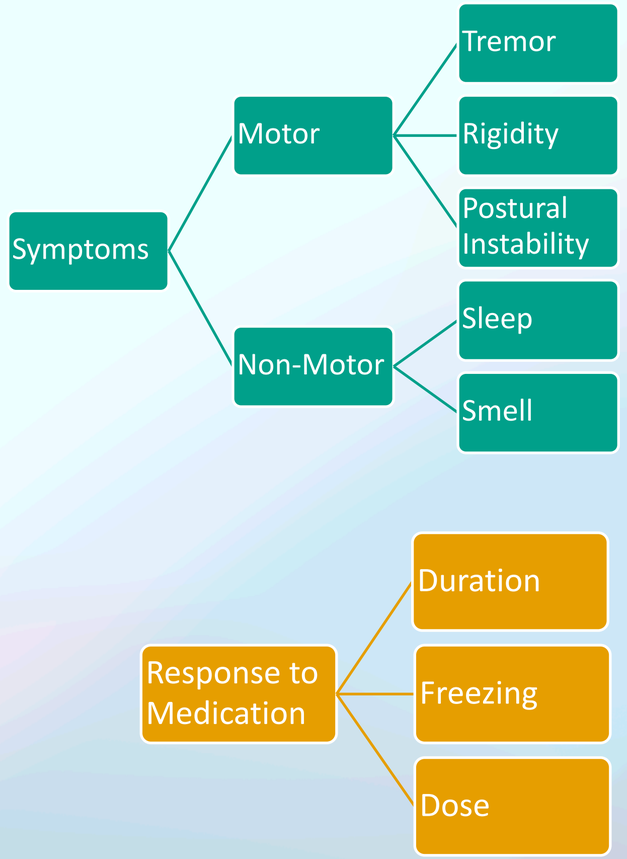
Death of dopamine producing cells in Substantia Nigra



Genetic Association

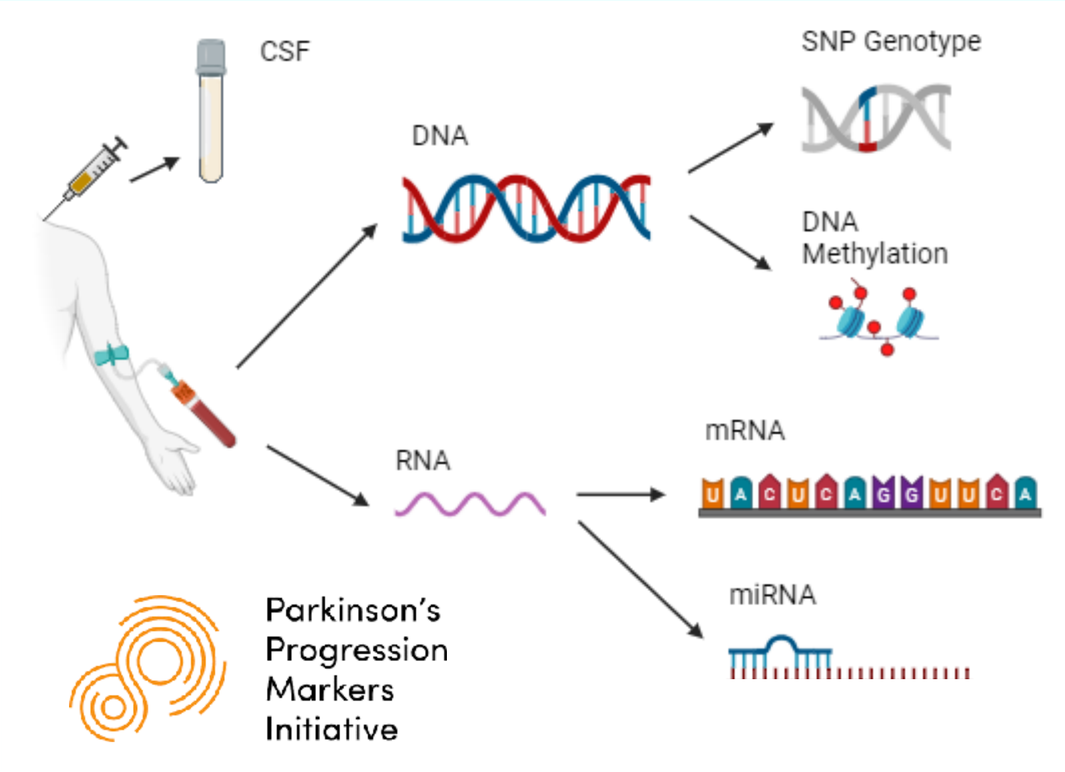


Variability Between Patients



1. Klein, C. & Westenberger, A. Genetics of Parkinson's Disease. Cold Spring Harb. Perspectives Medicine 2, a008888
2. Severson, K. A. et al. Discovery of Parkinson's disease states and disease progression modelling: a longitudinal data study using machine learning. The Lancet Digit.Heal. 3 (2021)

Multi-modality Genomic Data from PPMI



- The Parkinson's Progression Marker Initiative (PPMI) is a longitudinal dataset with over 900 Parkinson's Disease (PD) patients
- Whole-blood samples taken over a 3-year period
- Genomic measures are interrelated, but no single measure comprehensively captures the entire human biological system

Marek et al. The Parkinson Progression Marker Initiative (PPMI). Progress in Neurobiology, 95:629–635, 12 2011

Informative Modalities Vary During Disease Progression

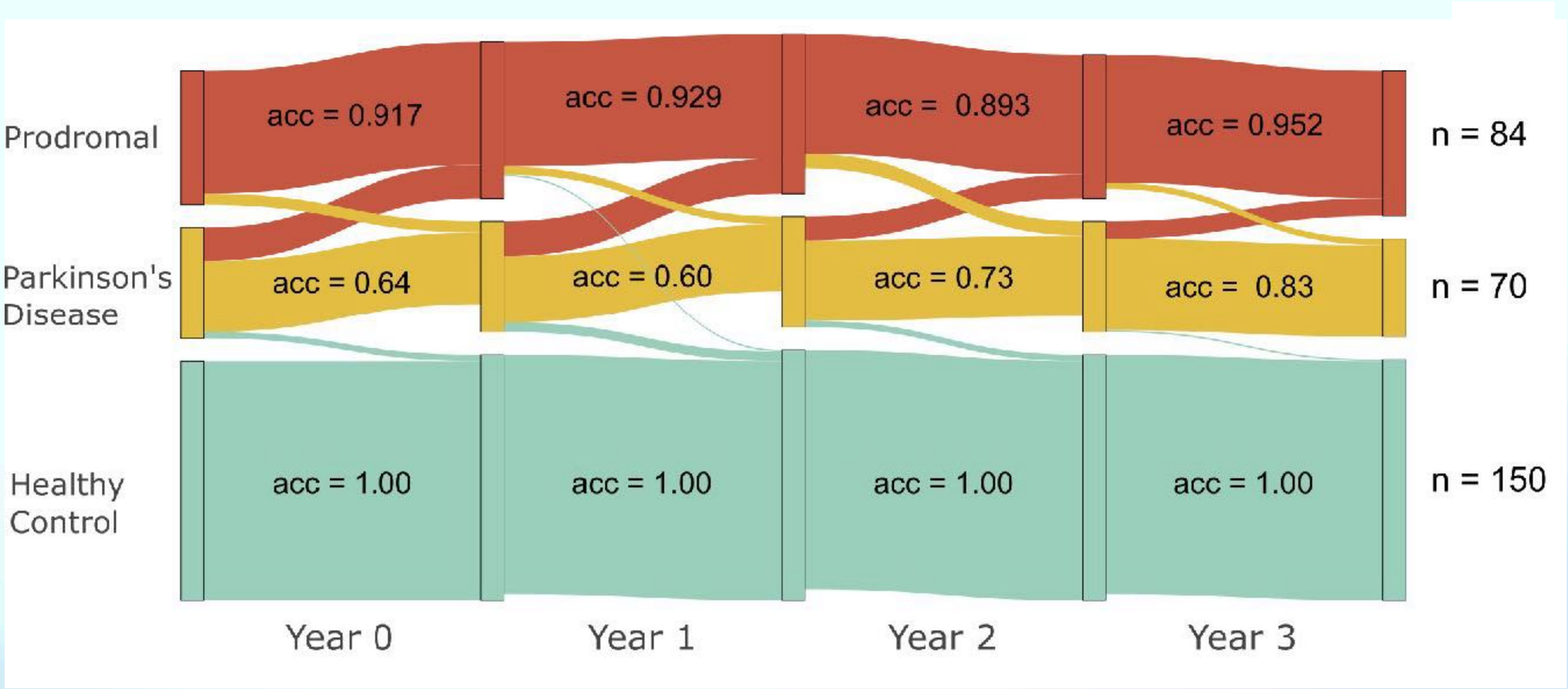
Cross-Sectional performance of MOGDx when stratifying participants into:

- Parkinson's Disease
- Prodromal (early indicators of disease but no clinical diagnosis)
- Healthy Control

		Modalities	Number of Participants	Accuracy	F1 score	Improvement in Accuracy
Genetic + Idiopathic (All)	Year 0	DNAm + SNP + mRNA + miRNA	1515	0.630 ± 0.019	0.665 ± 0.017	0.110 ± 0.018
	Year 1	DNAm	548	0.624 ± 0.020	0.667 ± 0.032	0.111 ± 0.02
	Year 2	Clinical + DNAm	542	0.694 ± 0.037	0.717 ± 0.034	0.166 ± 0.037
	Year 3	DNAm	493	0.712 ± 0.018	0.699 ± 0.048	0.146 ± 0.018
Genetic	Year 0	DNAm + SNP	489	0.789 ± 0.036	0.753 ± 0.04	0.419 ± 0.036
	Year 1	DNAm + SNP	443	0.867 ± 0.018	0.835 ± 0.02	0.472 ± 0.018
	Year 2	DNAm + SNP	432	0.866 ± 0.031	0.837 ± 0.032	0.477 ± 0.031
	Year 3	DNAm + SNP	365	0.841 ± 0.034	0.811 ± 0.038	0.403 ± 0.034
Idiopathic	Year 0	SNP + miRNA	667	0.681 ± 0.031	0.752 ± 0.008	0.069 ± 0.031
	Year 1	CSF + DNAm + SNP	582	0.720 ± 0.039	0.776 ± 0.035	0.122 ± 0.039
	Year 2	CSF + Clinical + DNAm	399	0.805 ± 0.022	0.770 ± 0.022	0.246 ± 0.022
	Year 3	CSF + DNAm	360	0.764 ± 0.022	0.721 ± 0.021	0.183 ± 0.022

Ryan B, Marioni RE, Simpson TI. An Integrative Network Approach for Longitudinal Stratification in Parkinson's Di

Models Trained at Latest Time-point are Useful for Early Detection

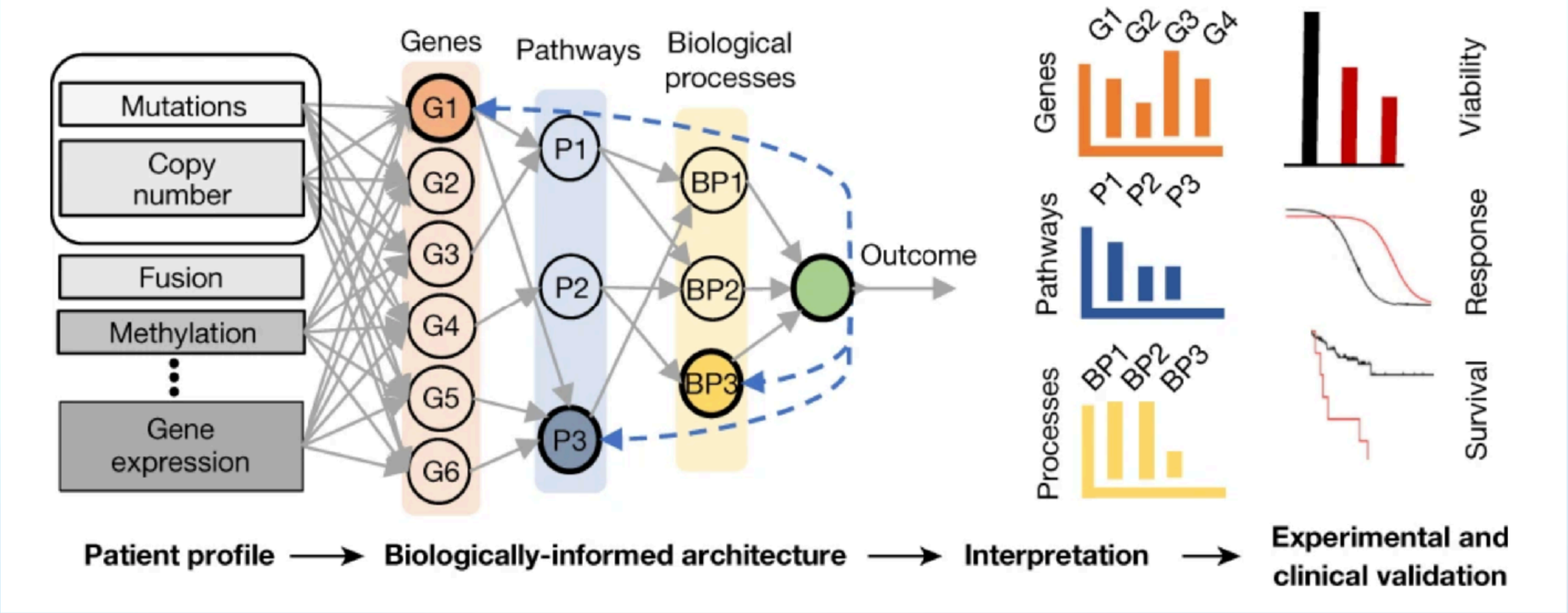


- Strong disease signature found in integration of SNP and DNAm modalities for individuals with a genetic association
- Models trained later in the disease course are more accurate
- Epigenetic modifications are informative throughout the disease course

Ryan B, Marioni RE, Simpson TI. An Integrative Network Approach for Longitudinal Stratification in Parkinson's Disease. medRxiv 2024. https://doi.org/10.1101/2024.01.25.24301595

Biologically Informed Deep Neural Networks - Prostate Cancer

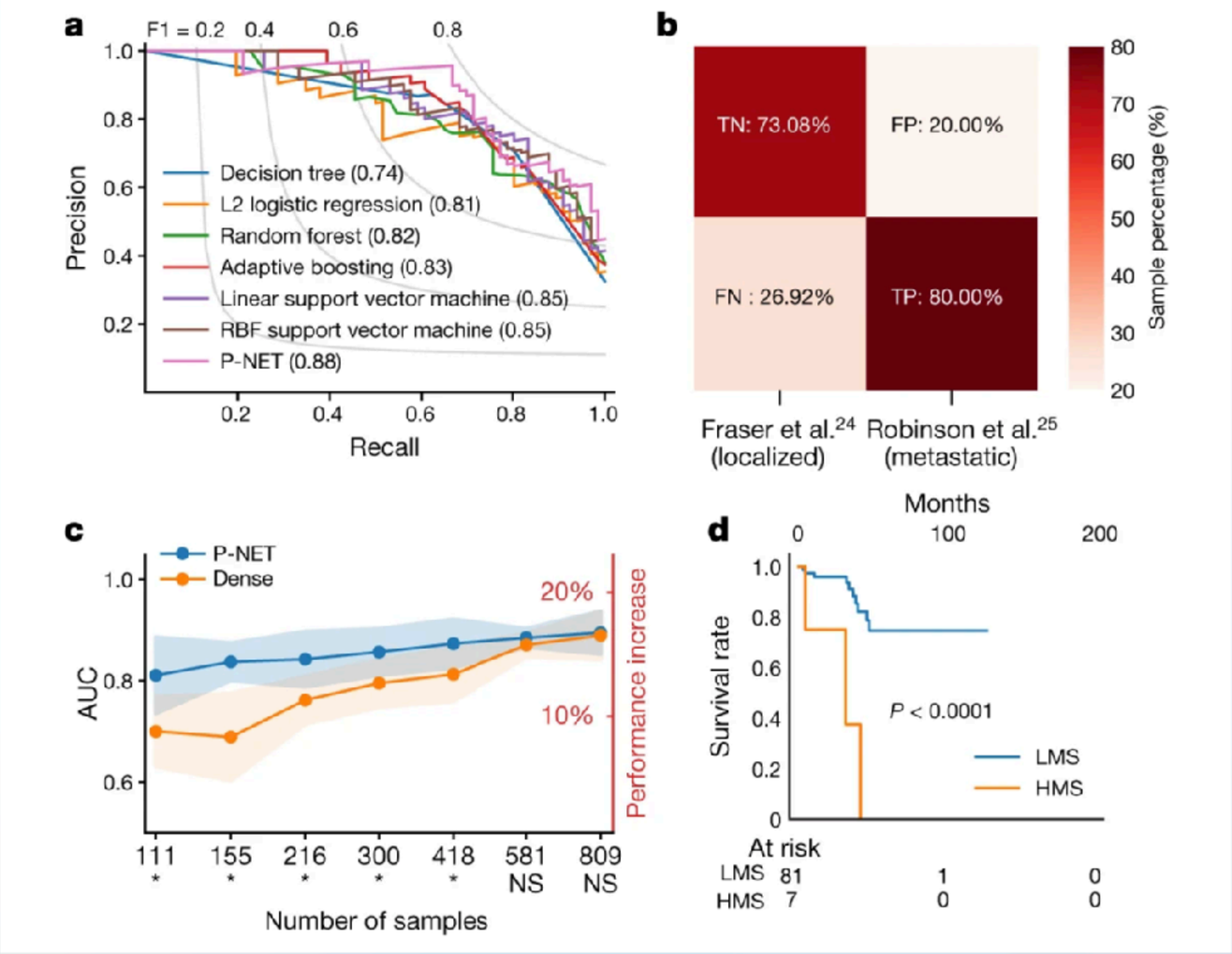
- domain knowledge used to generate customised DNN architectures (P-NET)
- trained P-NET provides relative ranking of nodes in each layer



Elmarakeby, Haitham A., Justin Hwang, Rand Arafeh, Jett Crowdis, Sydney Gang, David Liu, Saud H. AlDubayan, et al. 'Biologically Informed Deep Neural Network for Prostate Cancer Discovery' https://doi.org/10.1038/s41586-021-03922-4

Biologically Informed Deep Neural Networks - Prostate Cancer

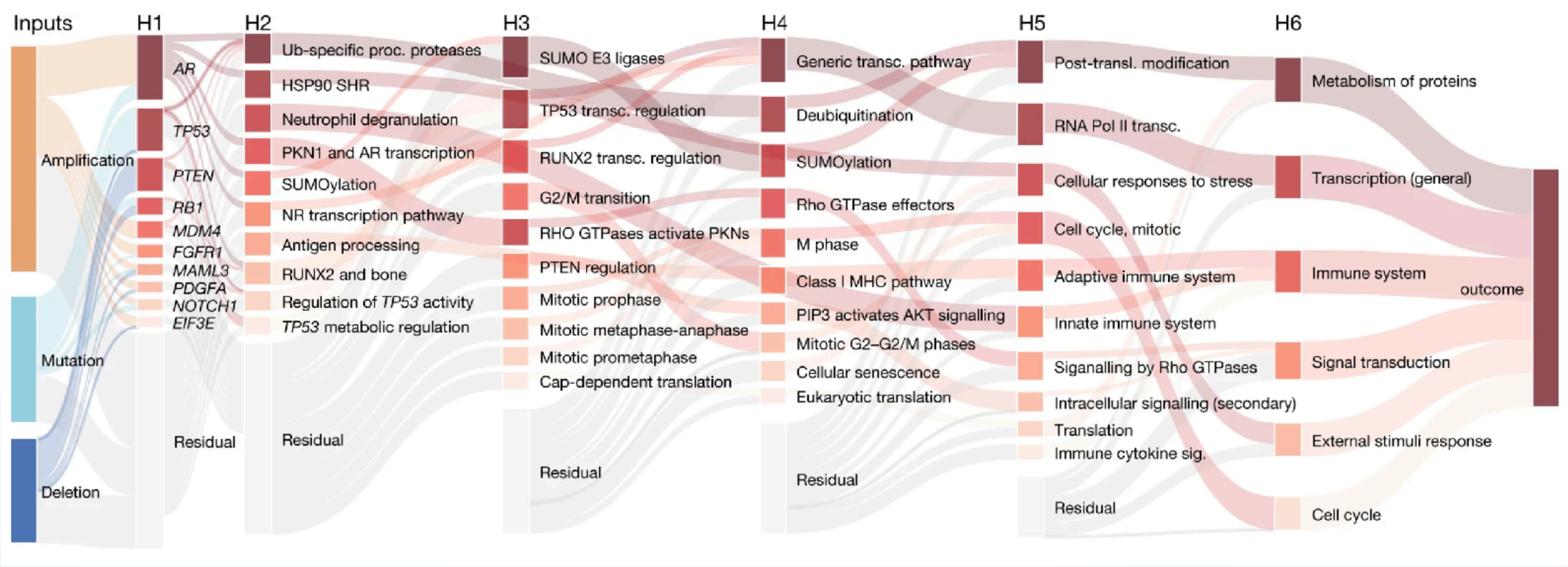
- trained and tested P-NET with 1,013 prostate cancers (333 CRPCs and 680 primary cancers - 80% training, 10% validation and 10% testing)
- predict disease state (primary or metastatic disease) using somatic mutation and copy number data
- evaluated how sparse model compared to fully connected DNN
- validated model with 2 external datasets
- used *Deep-LIFT* attribution to derive gene importance scores [evaluates contribution of input features to the output prediction]
- flexible model for integrating different data modalities
- used to predict aggressive cancers
- allowed molecular stratification



Elmarakeby, Haitham A., Justin Hwang, Rand Arafeh, Jett Crowdis, Sydney Gang, David Liu, Saud H. AlDubayan, et al. 'Biologically Informed Deep Neural Network for Prostate Cancer Discovery' https://doi.org/10.1038/s41586-021-03922-4

Biologically Informed Deep Neural Networks - Prostate Cancer

- inputs, genes, and pathways hierarchically linked - intuitive
- genes/terms ranked by importance - explainable

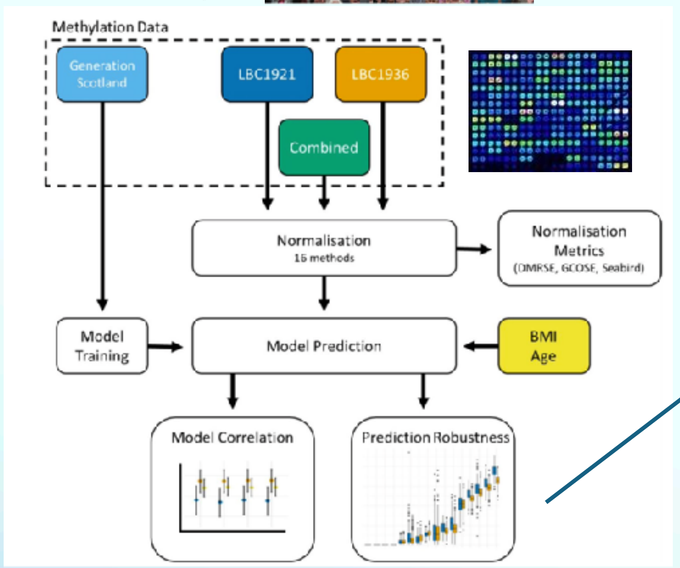


Elmarakeby, Haitham A., Justin Hwang, Rand Arafeh, Jett Crowdis, Sydney Gang, David Liu, Saud H. AlDubayan, et al. 'Biologically Informed Deep Neural Network for Prostate Cancer Discovery' https://doi.org/10.1038/s41586-021-03922-4

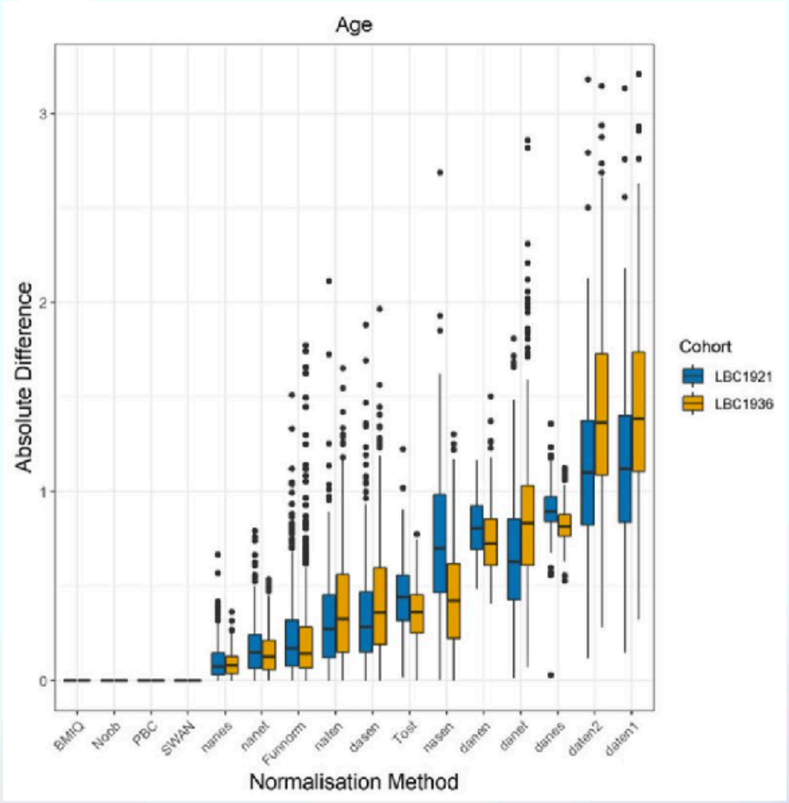
Integration of datasets for individual prediction of DNA methylation-based biomarkers

DNA Methylation Data

1921 Lothian Birth Cohort 1936



Robustness of Predictions



- Projecting data from new datasets onto existing reference data is challenging
- Identification of technical vs biological variation
- Normalisation can help to resolve this but approach is critical.
- Epigenetic measures, such as EpiScore can be used to aid retention of biological variation during normalisation

Charlotte Merzbacher , **Barry Ryan** , **Thibaut Goldsborough**, Robert F Hillary , Archie Campbell , Lee Murphy , Andrew M McIntosh, David Liewald, Sarah E Harris, Allan F McRae , Simon R Cox , Timothy I Cannings , Catalina A Vallejos, Daniel L McCartney, Riccardo E Marioni (2023) Integration of DNA methylation datasets for individual prediction of DNA methylation-based biomarkers. Genome Biology. https://doi.org/10.1186/s13059-023-03114-5

Automated Mood Disorder Symptoms Monitoring From Multivariate Time-Series Sensory Data: Getting the Full Picture Beyond a Single Number

Psychiatric assessments are scheduled infrequently and rely on self-reported experiences

Wearables can enable near-continuous remote monitoring leveraging passively collected physiological data

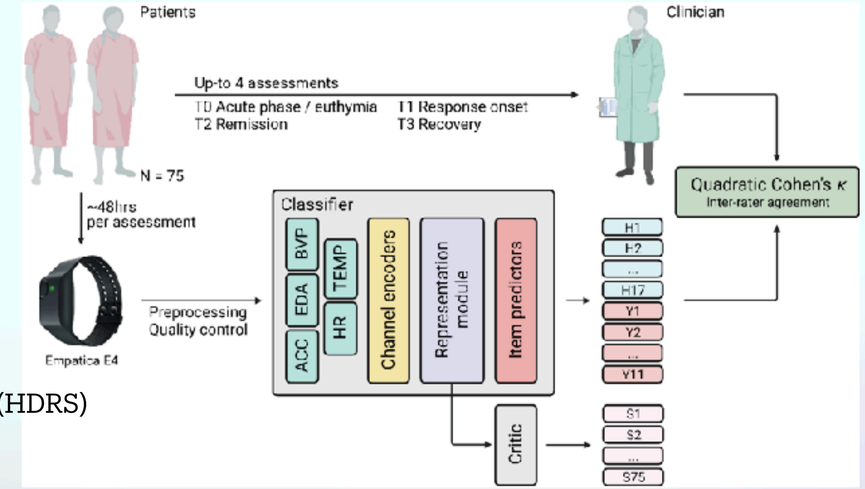
Timely Interventions
Better Outcomes

Reducing monitoring to acute episode detection misses on actionable clinical information

Novel Task: Inferring all items from YMRS and HDRS (two popular standardised psychometric scales)

Hamilton Depression Rating Scale (HDRS)

Young Mania Rating Scale (HDRS)

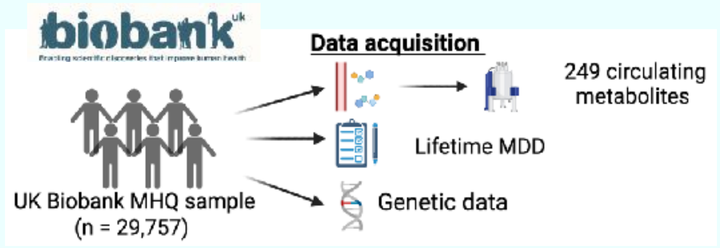


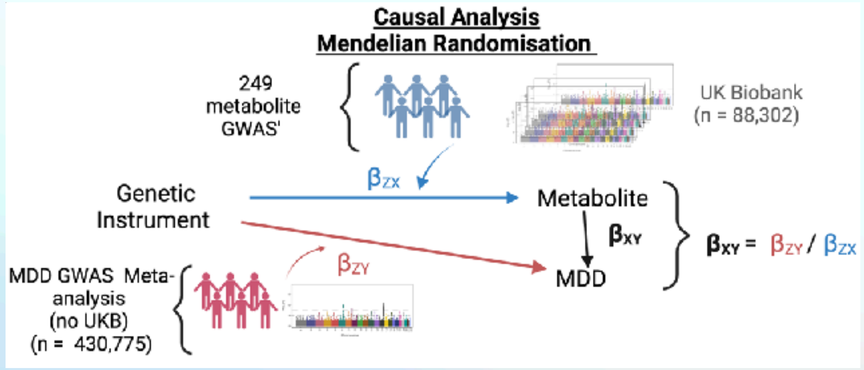
Model		ACC		F_1score		AUROC	
		segment	subject	segment	subject	segment	subject
SL	ENET	66.38	71.88	66.54	70	72.24	82.25
	KNN	70.37	82.81	71.34	80.6	73.27	83.26
	SVM	71.25	81.25	71.63	78.81	73.44	83.21
	XGBoost	72.02	82.81	71.72	82.03	72.44	83.17
	E4mer	75.35	81.25	74.39	81.33	75.68	82.22
SSL	MP (LR)	77.53	87.5	77.87	88.3	78.02	89.2
	MP (FT)	**81.23**	**90.63**	**81.45**	**91.47**	**82.02**	**93.11**
	TP (LR)	71.16	81.25	72.06	82.37	71.89	84.12
	TP (FT)	75.69	84.38	75.1	83	75.21	84.23

Filippo Corponi, **Bryan M. Li**, Gerard Anmella, Ariadna Mas, Isabella Pacchiarotti, Marc Valentí, Iria Grande, Antoni Benabarre,Marina Garriga, Eduard Vieta, Stephen M. Lawrie, Heather C. Whalley, Diego Hidalgo-Mazzei, Antonio Vergari (2024) Automated mood disorder symptoms monitoring from multivariate time-series sensory data: getting the full picture beyond a single number. Nature: Translational Psychiatry. https://doi.org/10.1038/s41398-024-02876-1

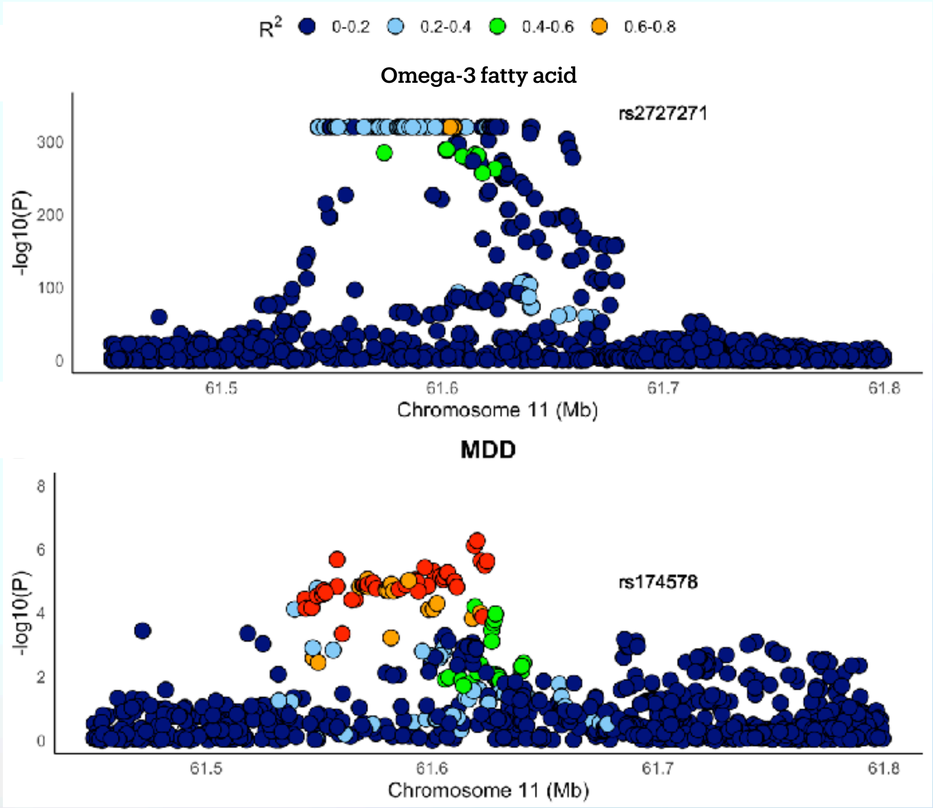
Metabolites and Major Depressive Disorder in the UK Biobank

Most metabolites were significantly associated with depression.



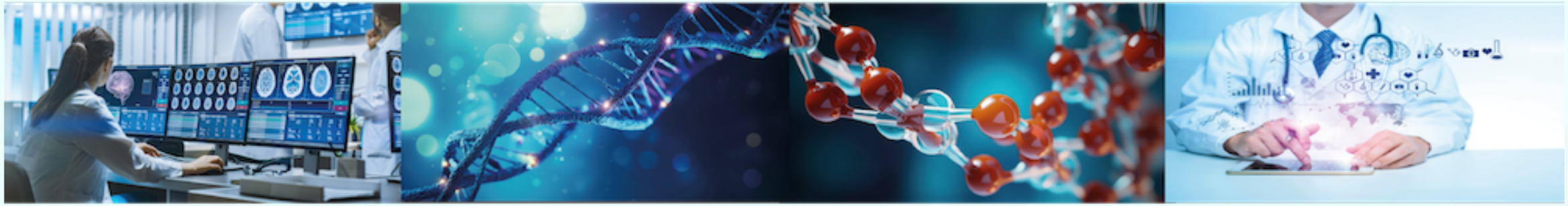


Evidence of causality between lowered omega-3 and higher omega-6:omega-3 fatty acid ratio with depression.



Colocalisation of MDD and Omega-3 fatty acid loci in the FADs cluster (PP.H4 > 0.90)

Eleanor Davyson, Xueyi Shen, Danni A. Gadd, Elena Bernabeu, Robert F. Hillary, Daniel L. McCartney, Mark Adams, Riccardo Marioni, and Andrew M. McIntosh (2023) Metabolomic Investigation of Major Depressive Disorder Identifies a Potentially Causal Association With Polyunsaturated Fatty Acids. Society of Biological Psychiatry. https://doi.org/10.1016/j.biopsych.2023.01.027



Programming for Biomedical Informatics

Revision Session 1 this Thursday

"Course Lecture Material Overview & Short Answer Questions"

Ask Questions on the EdStem Discussion Board

https://github.com/tisimpson/pbi