Multi-Omic Graph Diagnosis (MOGDx) :
A tool for the integration and classification of heterogeneous diseases

Barry Ryan, Riccardo Marioni & T. Ian Simpson

https://homepages.inf.ed.ac.uk/tsimpson/ barry.ryan@ed.ac.uk github.com/Barry8197/MOGDx

MOGDx is a command line tool designed to **integrate multi-omic data** for heterogenous disease classification. It extracts important features, integrates modalities using similarity network fusion, imputes missing samples and utilises a **graph neural network** to perform accurate classification on a **patient similarity network**.

1. Motivation & Aims

Heterogeneity in human diseases confounds everything; clinical trials, genetic association testing, drug response and intervention strategies to name a few. Redefining such diseases through subtype classification, symptomatic grading or similar has the potential to uncover new treatments, repurpose old treatments or identify intervention strategies.

An individual omic measure provides a single measure of biological complexity however, the integration of multiple omic types could combine multiple measures of biological complexity, mirroring the heterogeneity in these diseases.

The use of a network taxonomy for multi-omic data integration has risen in popularity recently[1,2,3]. Networks are easily integrated, can readily handle missing data, and have been used in a wide variety of biomedical applications in the unsupervised setting[4].

Graph Neural Networks (GNN) have shown powerful classification performance on several benchmark network datasets[5]. The use of GNN's in a supervised setting for disease classification is a promising avenue to redefine heterogenous diseases.

Our aims are to develop MOGDx, a tool which:

- Performs accurate classification tasks for heterogenous diseases
- Yields interpretable results
- Is reproducible, can be downloaded and run on the command line

2. Methods

- **Patient Similarity Network (PSN)**

A PSN is a method of classifying patients based on a similarity measure in various features. We calculated patient similarity using Pearson correlation on extracted features from each modality and utilised a K-Nearest Neighbour algorithm to keep the strongest connections.

- **Similarity Network Fusion (SNF)**

SNF is a computational method to integrate multiple data types when represented as a PSN. It is effective in capturing the full underlying spectrum of the data, as well as inferring missing connections. We used SNF to integrate our PSN's as per Figure 1.

- **Autoencoder (AE)**

An AE is a particular type of neural network which is trained to copy its input to its output. We utilised this architecture to extract the hidden dimension, which is a reduced representation of its input.

- **Graph Neural Network (GNN)**

A GNN is a class of neural networks which learn from network structure and embeddings. We utilised a graph convolutional network to learn from the fused PSN and an embedded vector. The embedded vector was created by concatenating the reduced representation from each modality

3. Work in Progress

- **Developing a better method of dimensionality reduction for node embeddings or improving the performance of the AE**

As can be seen in Figure 3 (B), the AE is significantly less informative than the PSN. We are working on improving this performance by calculating a joined loss of all omics as opposed to the current parallel implementation.

- **Graph convolutional networks require a fully connected network during training**

Graph convolutional network is a transductive algorithm, meaning all patient samples have to be present during training. It requires a full re-training of the algorithm when new samples are collected. We are working on moving to the inductive setting and using an algorithm such as GraphSage.

- **Early predictive power, longitudinal analysis and novel datasets**

MOGDx demonstrated some early predictive power when classifying on the BRCA dataset. We are working on extending this analysis to a novel dataset. We are also beginning to work on a longitudinal aspect to MOGDx which can predict if a patient classification label will change.

4. How you can help

If you use, have experience with or interest in multi-modal network integration or heterogenous disease classification, we would love to talk with you about your research and expectations. We are particularly keen to discuss methods or datasets which could further develop or improve future network analyses.

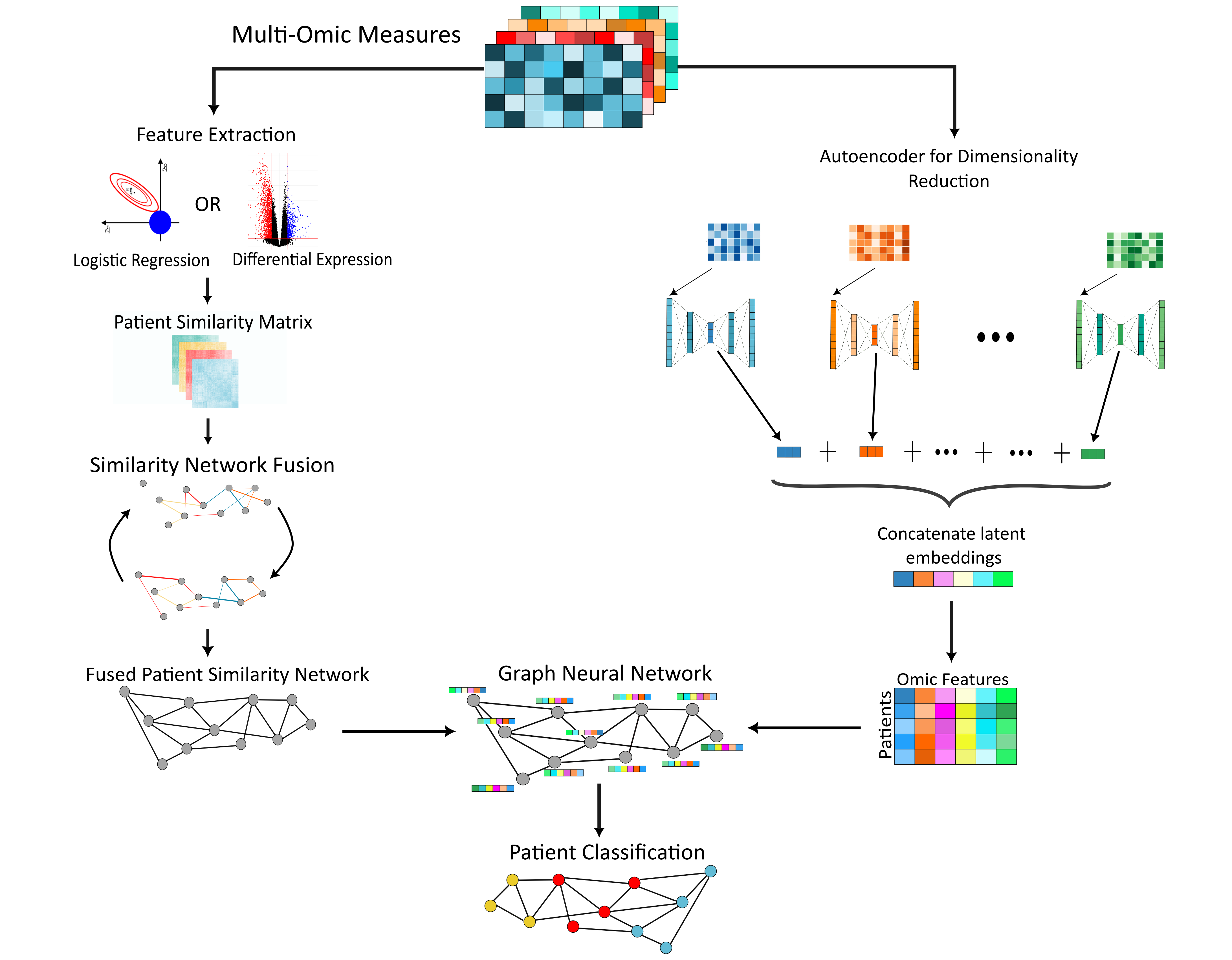
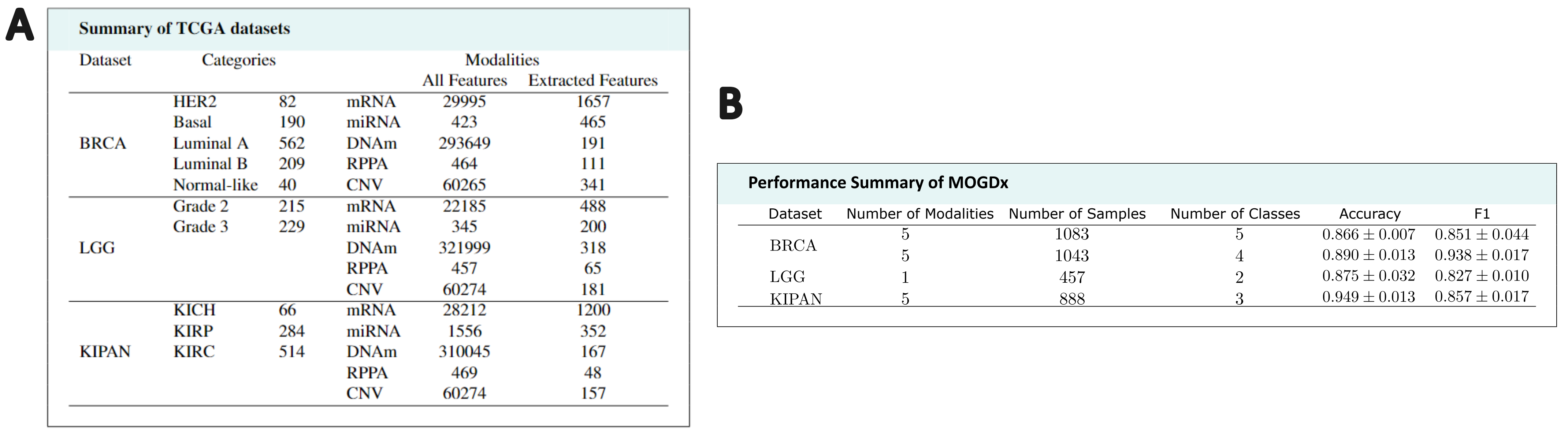

Multi-Omic Measures
Feature Extraction
OR
Logistic Regression
Differential Expression
Patient Similarity Matrix
Similarity Network Fusion
Fused Patient Similarity Network
Autoencoder for Dimensionality Reduction
Concatenate latent embeddings
Omic Features
Patients
Graph Neural Network
Patient Classification

Figure 1. Pipeline of MOGDx | *MOGDx takes any number of omic measures as input. Feature extraction is performed to maximise similarities between patients. Each patient similarity matrix is converted to a network, and these patient similarity networks are fused using SNF. In parallel, an AE is trained for dimensionality reduction. The reduced latent embeddings are concatenated and added to the fused network as node features. A graph neural network is trained and patient classification performed.*



A

Summary of TCGA datasets

Dataset	Categories		Modalities	All Features	Extracted Features
BRCA	HER2	82	mRNA	29995	1657
	Basal	190	miRNA	423	465
	Luminal A	562	DNAm	293649	191
	Luminal B	209	RPPA	464	111
	Normal-like	40	CNV	60265	341
LGG	Grade 2	215	mRNA	22185	488
	Grade 3	229	miRNA	345	200
			DNAm	321999	318
			RPPA	457	65
			CNV	60274	181
KIPAN	KICH	66	mRNA	28212	1200
	KIRP	284	miRNA	1556	352
	KIRC	514	DNAm	310045	167
			RPPA	469	48
			CNV	60274	157

B

Performance Summary of MOGDx

Dataset	Number of Modalities	Number of Samples	Number of Classes	Accuracy	F1
BRCA	5	1083	5	0.866 ± 0.007	0.851 ± 0.044
	5	1043	4	0.890 ± 0.013	0.938 ± 0.017
LGG	1	457	2	0.875 ± 0.032	0.827 ± 0.010
KIPAN	5	888	3	0.949 ± 0.013	0.857 ± 0.017

Figure 2. (A) Summary of Datasets | The BRCA dataset is for PAM50 subtype classification of Breast Invasive Carcinoma, consisting of 5 classes. The LGG dataset is a grade classification task for Low Grade Glioma. The KIPAN dataset is a subtype classification task consisting of 3 classes. (B) Summary of Performance | *MOGDx demonstrates state-of-the-art classification accuracy in a variety of tasks. All available modalities were used for both BRCA and KIPAN. Performance is shown for BRCA with and without the Normal-like class. Only DNAm was used on the LGG dataset, as it achieved the best accuracy while still including the maximum number of samples.*

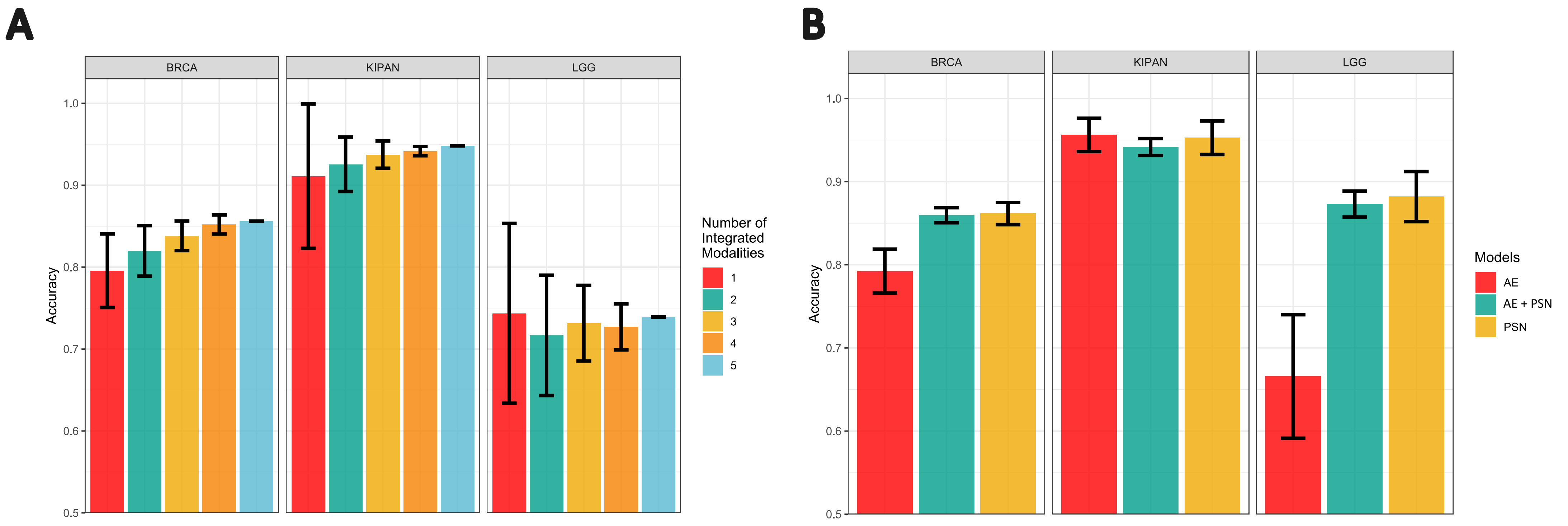


Figure 3. (A) Modality Integration Importance | Some omic measures are more predictive than others depending on the classification task and should only be integrated if they improve classification performance or improve patient sample coverage. (B) PSN Importance | Combining AE and PSN reduces variance in train and test splits while maintaining optimal accuracy.

Acknowledgements

This work was supported by the United Kingdom Research and Innovation [grant EP/S02431X/1], UKRI Centre for Doctoral Training in Biomedical AI at the University of Edinburgh, School of Informatics.

The results published here are in whole or part based upon data generated by the TCGA Research Network: https://www.cancer.gov/tcga

With thanks to my supervisors Ian Simpson and Riccardo Marioni for their advice and help throughout.

References

1.Gliozzo, J. et al. Network modeling of patients' biomolecular profiles for clinical phenotype/outcome prediction. Sci. Reports 10, 3612, DOI: 10.1038/s41598-020-60235-8 (2020). Number: 1 Publisher: Nature Publishing Group
2. Li, L. et al. Identification of type 2 diabetes subgroups through topological analysis of patient similarity. Sci. Transl. Medicine 7, 311ra174–311ra174, DOI:10.1126/scitranslmed.aaa9364 (2015). Publisher: American Association for the Advancement of Science.
3. Pai, S. et al. netDx: interpretable patient classification using integrated patient similarity networks. Mol. Syst.Biol. 15, e8497, DOI: 10.15252/msb.20188497 (2019). Publisher: John Wiley & Sons, Ltd.
4. Wang, T. et al. MOGONET integrates multi-omics data using graph convolutional networks allowing patient classification and biomarker identification. Nat. Commun. 12, 3445, DOI: 10.1038/s41467-021-23774-w (2021). Number: 1 Publisher: Nature Publishing Group.
5. Jie Zhou et al. "Graph neural networks: A review of methods and applications". en. In: AI Open 1 (Jan. 2020), pp. 57–81. ISSN: 2666-6510. DOI: 10 . 1016 / j . aiopen . 2021 . 01 . 001. URL: https : / / www .sciencedirect.com/ science/article/pii/S2666651021000012 (visited on 10/12/2022).