Multi-Omic Graph Diagnosis (MOGDx) : A data integration tool to perform classification tasks for heterogeneous diseases

Barry Ryan[1], **Riccardo E. Marioni**[2], **and T. Ian Simpson**[1]

[1]School of Informatics, University of Edinburgh, 10 Crichton Street, Edinburgh EH8 9AB, UK
[2]Centre for Genomic and Experimental Medicine, Institute of Genetics and Cancer, University of Edinburgh, Edinburgh, EH4 2XU, UK

ABSTRACT

Heterogeneity in human diseases presents challenges in diagnosis and treatments due to the broad range of manifestations and symptoms. With the rapid development of labelled multi-omic data, integrative machine learning methods have achieved breakthroughs in treatments by redefining these diseases at a more granular level. These approaches often have limitations in scalability, oversimplification, and handling of missing data. In this study, we introduce Multi-Omic Graph Diagnosis (MOGDx), a flexible command line tool for the integration of multi-omic data to perform classification tasks for heterogeneous diseases. MOGDx is a network integrative method that combines patient similarity networks with a reduced vector representation of genomic data. The reduced vector is derived from the latent embeddings of an auto-encoder and the combined network is fed into a graph convolutional network for classification. MOGDx was evaluated on three datasets from the cancer genome atlas for breast invasive carcinoma, kidney cancer, and low grade glioma. MOGDx demonstrated state-of-the-art performance and an ability to identify relevant multi-omic markers in each task. It did so while integrating more genomic measures with greater patient coverage compared to other network integrative methods. MOGDx is available to download from https://github.com/biomedicalinformaticsgroup/MOGDx. Overall, MOGDx is a promising tool for integrating multi-omic data, classifying heterogeneous diseases, and interpreting genomic markers.

Introduction

Heterogeneity in human diseases is a pertinent yet difficult issue that can confound the analysis of clinical trials, genetic association testing, drug responses, and intervention strategies. Heterogeneous diseases encompass any single disease with a broad range of manifestations or symptoms. Redefining such diseases through sub-type classification, symptomatic grading or similar has the potential to uncover new treatments, re-purpose old treatments or identify intervention strategies. This approach has already been shown to improve patient outcomes in a number of diseases[1,2]. Performing classification tasks with heterogeneous diseases is a complex problem often requiring analysis of multiple types of data of varying scale and complexity, as such it needs analytic frameworks that are flexible and scalable. The use of Artificial Intelligence (AI) has emerged as a popular method to solve this problem and has been facilitated by the development of high-throughput sequencing technologies. Such technologies have made various types of biological data, coined 'omic' data, available. Individual omics provide 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. An analytical tool which can integrate multiple omic measures and perform heterogeneous disease classifications could significantly improve patient outcomes in this area.

There is an increasing number of methods which integrate multi-omic data in both the supervised and unsupervised classification space. There exists two main taxonomies for data integration, which can be broadly categorised as input data-fusion and output data-fusion, although no gold standard method exists. Input data-fusion methods combine data sources into a single dataset prior to analysis, while output data-fusion methods analyse each dataset separately and combine the results. Input data-fusion methods estimate an embedding which projects datasets into a shared latent space which minimizes variance between datasets while maximising individual variability within each data set[3]. For example, Lock et al.[4] achieved state-of-the-art performance on characterization of tumour types from the Breast Invasive Carcinoma (BRCA) dataset in The Cancer Genome Atlas (TCGA)(https://www.cancer.gov/tcga). Their analysis was able to effectively uncover both individual and joint data structures, resulting in better interpretation while also improving unsupervised classification results on the BRCA dataset. In its simplest form, output data-fusion resembles ensemble methods, whereby an independent analysis of each dataset is performed, and the results combined using an aggregation technique. An example of this is presented by Phan et al.[5] who use a stacked-generalisation model on an ovarian dataset in TCGA.

These methods show that classification performance is increased when multiple modalities are considered, however they often scale poorly, are overly simplistic or do not take into account the cross correlation between modalities[3,6]. The use of a network taxonomy for multi-omic data integration has risen in popularity recently. The advantage of networks is that they are easily integrated and can readily handle missing data. Gliozzo et al.[7] and Li et al.[8] show that representing data as a Patient Similarity Network (PSN) can retain information and have superior or competitive performance compared to standard Euclidean methods for a single modality. netDx, developed by Pai et al.[9], uses ridge regression and label propagation algorithms to integrate and perform ranked classifications on PSN's. Wang et al.[6] define each modality as a single PSN, perform classification using a Graph Convolutional Network (GCN) and concatenate predictions into a cross-omic correlation tensor before making final label predictions. Li et al.[10] perform classifications using a GCN on integrated PSN's. These methods are novel strategies for the integration of network data at the input and output space, however, they don't leverage the full advantages of representing data as a network. Current network methods cannot handle patients missing one or more omic measures, and methods can only handle a fixed number of modalities.

Hence, we introduce Multi-Omic Graph Diagnosis (MOGDx), a flexible tool for the integration of multi-omic data to perform classification tasks for heterogeneous diseases. The MOGDx pipeline integrates omic data into a single PSN before performing classification using an Graph Neural Network (GNN) algorithm. Each omic measure undergoes pre-processing steps to extract the most informative features. These features are used to inform the PSN for each individual omic measure. Similarity Network Fusion (SNF)[11] is performed to integrate the PSN's into a single network. In parallel, each unprocessed omic measure is passed through an Autoencoder (AE) for dimensionality reduction. The latent embedding of each modality is concatenated into a single vector. Each vector corresponds to a single patient node, and the two are combined into a single GNN model, namely GCN in this instance. The performance of MOGDx is benchmarked on the BRCA, Low-Grade Glioma (LGG) and Pan Kidney Cohort (KIPAN) datasets, with state-of-the-art performance demonstrated. MOGDx is the first tool of its kind which can handle missing patient data as well as any number of data modalities or omic measures. We show the benefit of integrating multiple modalities and the importance of representing data as a PSN. We also demonstrate that MOGDx can identify important omic markers relating to the targeted biomedical problem.

Results

Pipeline of MOGDx

We present MOGDx, a pipeline for the supervised classifications of patients with heterogenous diseases (Figure 1). MOGDx takes as input any number of omic measures such as genomic, transcriptomic and proteomic datasets. The raw data is processed into omic measure matrices, with each row corresponding to a patient and each column a feature of that measure. First, depending on the omic dataset (see Methods), either differential expression or penalised logistic regression is performed to identify important features of that dataset.

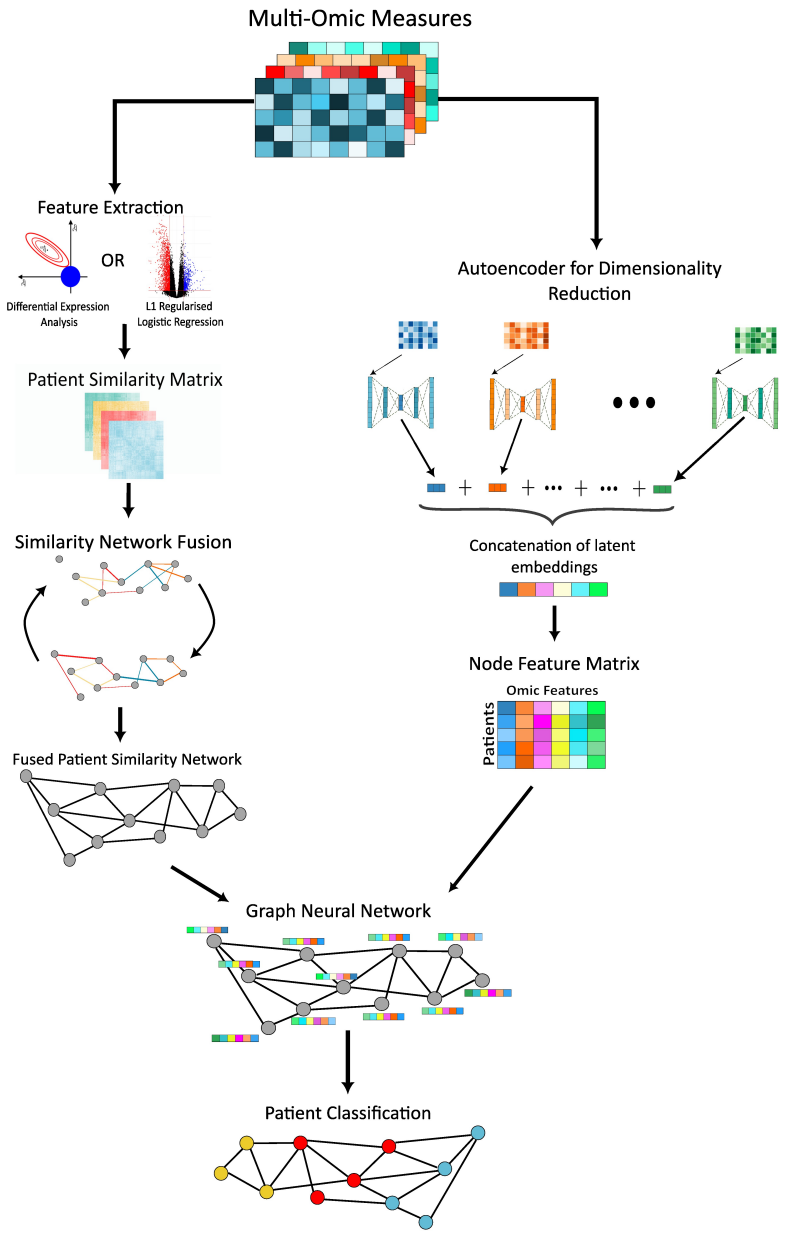

Figure 1. Pipeline of MOGDx | MOGDx takes any number of omic measures as input. Feature extraction is performed to maximise similarities between patients. Each patient similarity matrix is converted to a network and these patient similarity networks are fused using SNF. In parallel, an AE is trained 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.

These features are used to inform a weighted similarity matrix calculated using Pearson correlation. A PSN is then constructed using the K-Nearest Neighbours (KNN) algorithm. SNF is performed to fuse the PSN's into a single network.

One AE per omic measure is applied for dimensionality reduction. The reduced latent embeddings of each AE are concatenated, forming the node feature matrix. The node feature matrix and fused PSN are combined and input into a GNN for supervised classification. For simplicity, a GCN architecture was implemented using the Stellargraph[12] library (version 2.2.1) in python. This methodology integrates the predictive power of PSN's with a reduced representation of classical omic characteristics. MOGDx obtains state-of-the-art predictive performance on the integration of network and vector characteristics, exhibiting the benefit of including both. MOGDx is a command line tool for the supervised classification of heterogeneous diseases, which can be used for a wide range of biomedical applications.

Table 1. Summary of TCGA datasets

Dataset	Categories		Modalities			
				Raw Feature Count	Counts After Processing	PSN Extracted Feature Counts
BRCA	HER2	82	mRNA	60666	29995	1657
	Basal	190	miRNA	1882	423	465
	Luminal A	562	DNAm	485577	293649	191
	Luminal B	209	RPPA	488	464	111
	Normal-like	40	CNV	60624	60265	341
LGG	Grade 2	215	mRNA	60660	22185	488
	Grade 3	229	miRNA	1881	345	200
			DNAm	485577	321999	318
			RPPA	487	457	65
			CNV	60623	60274	181
KIPAN	KICH	66	mRNA	60660	28212	1200
	KIRP	284	miRNA	1881	1556	352
	KIRC	514	DNAm	485577	310045	167
			RPPA	487	469	48
			CNV	60623	60274	157

mRNA refers to mRNA gene expression data, miRNA refers to micro RNA gene expression data, DNAm refers to DNA methylation data, RPPA refers to reverse phase protein array data and CNV refers to Copy Number Variation data. The Breast Invasive Carcinoma (BRCA) dataset is for PAM50 sub-type classification consisting of 5 classes; HER2, Basal-like, Luminal A, Luminal B and Normal-like. The LGG dataset is a grade classification task for Low-Grade Glioma. The KIPAN dataset is sub-type classification task consisting of 3 classes; KICH, KIRP, KIRC.

Datasets

The performance of MOGDx is benchmarked on three different TCGA datasets, BRCA PAM50 sub-type classification, grade classification in LGG and KIPAN for kidney type classification. Data was downloaded using the TCGABiolinks[13] Bioconductor package (version 2.28.3). All modalities available in the TCGA database were included, resulting in 5 types of omics data used for classification. The omic data types available are mRNA expression (mRNA) data, micro RNA expression (miRNA) data, DNA methylation (DNAm) data, Reverse Phase Protein Array (RPPA) data and Copy Number Variation (CNV) data. All patient samples were utilised irrespective if they were available in only one or all datasets, with specific details reported in Table 1. Raw feature count is the total number of features available per modality. Processing was performed to remove features which were mostly missing or had a standard deviation of 0. These features were directly inputted into the AE. Further processing was performed to extract the most predictive features, which were used to construct the PSN's.

BRCA PAM50 is a 50-gene signature used to sub-type breast cancer into 5 classifications; Normal-like, Basal-like, HER2-enriched, Luminal A and Luminal B[14,15]. Patients included in this dataset have a mutation to their BRCA gene and therefore have a larger risk of developing breast cancer. Sub-typing by gene expression separates the carcinomas by varying biological properties and prognoses. For example, Luminal A has the best prognosis, while HER2 and Basal are considered more aggressive forms of cancer[15]. The KIPAN dataset consists of three categories separated by chromosomal differences[16]. Clear Renal Cell Carcinoma (KIRC) is characterised by loss of chromosome 3p, Papillary Renal Cell Carcinoma (KIRP) is characterised by loss of chromosome 9p and Chromophobe Renal Cell Carcinoma (KICH) is characterised by loss of multiple other chromosomes. The LGG dataset consists of grade 2 and grade 3 which are characterised by the World Health Organisation based on their histopathologic characteristics[17]. All of these datasets categorise a heterogeneous disease by a genetic association, making them suitable tasks for classification. They were chosen to demonstrate the generalisability of MOGDx to different diseases, as well as to benchmark the performance of MOGDx against other integrative methods[6,10,18].

Performance & Evaluation

The performance of MOGDx was compared to existing network integrative methods that perform heterogeneous disease classification. Ablation experiments were also performed to understand the importance of different components of MOGDx. The performance metrics used to compare the classification performance of MOGDx were accuracy and F1-score (F1). The F1 score was calculated by the mean F1 score of each class, weighted by the size of that class. k-fold cross validation was performed with 5 randomly generated splits to obtain the mean and standard deviation metrics reported. Within each split, the training set was randomly split into training and validation sets to produce an overall train/validation/test split of 65%/15%/20% respectively.

MOGDx achieves superior performance when integrating a variable number of modalities while including a larger number of samples

The classification performance of MOGDx was compared to similar PSN multi-omic integrative methods as well as benchmark classification algorithms namely; Support Vector Machine (SVM), L1 regularized linear regression (Lasso) and gradient tree boosting (XGBoost). Table 2 shows MOGDx outperforms all benchmark classification algorithms when trained on a single best modality or low dimensional representation of multiple modalities. This demonstrates the predictive power of the PSN and GCN learning architecture.

Table 2. Summary of Performance

Method	Dataset	Number of Modalities	Number of Samples	Number of Classes	Accuracy	F1
MOGDx	BRCA	5	1083	5	0.866 ± 0.007	0.827 ± 0.010
	BRCA	5	1043	4	0.890 ± 0.013	0.857 ± 0.017
	LGG	1	457	2	0.875 ± 0.032	0.851 ± 0.044
	KIPAN	5	888	3	0.949 ± 0.013	0.938 ± 0.017
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. All 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[6] was achieved using mRNA, miRNA and DNAm. The performance reported by MoGCN[10] 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.

MOGDx outperforms the other comparative integrative methods, MOGONET[6] and MoGCN[10]. On the BRCA dataset, MOGDx performs better in both accuracy and F1 metrics compared to MOGONET. MoGDx achieves comparable performance to MoGCn when trained on four classes and crucially retains double the number of samples. In supplementary figure S1, it can be seen that the Normal-like class is a difficult sub-type to classify. This is due to the small number of samples and the likelihood of these samples to go on to develop into one of the other sub-types. MoGCN does not include this sub-type in their classification, simplifying the task, resulting in higher accuracy. Interestingly, MOGDx strongly associates some of the Normal-like samples with other sub-types. This could suggest predictive power of this method to identify early signatures of BRCA. MOGDx identified a single omic measure, DNAm, which achieved optimal performance on the LGG dataset. All labelled samples were available in this single omic measure. While MOGDx did significantly outperform MOGONET on this dataset, there is a relatively large difference in number of samples. MOGONET obtained their data from Broad GDAC Firehose which stores TCGA data version from 2016 which could explain this discrepancy. Finally, MOGDx achieves slightly lower metrics on the KIPAN dataset compared to MOGONET and MoGCN. Once again, the number of samples differ due to differences between methods and differences in data collection. The lower performance of MOGDx in this dataset could be due to the imputation methods applied to account for missing samples in one or more data modality.

MOGDx can incorporate greater number of samples and modalities in its methodology. MOGONET and MoGCN are limited to the intersection of samples, which reduce the number of samples included in their analysis when more modalities are included. This is evident as both Lasso and XGBoost have greater number of samples available when trained on mRNA only. Conversely, MOGDx can incorporate all available samples due to imputation methods employed without a significant degradation in performance. Moreover, MOGONET and MoGCN are fixed to the integration of three modalities. The flexibility of MOGDx allows any number of modalities to be included, resulting in significantly improved performance on the LGG dataset as per Table 2.

The performance of MOGDx varies under different omic data types for different classification tasks

Figure 2.A shows the performance of MOGDx varies significantly when different omic measures are integrated. There exists a trade-off between modality integration and performance. As can be seen in supplementary figure S2, typically three to four modalities are required to ensure full sample coverage. Figure 2.A shows that some omic measures are significantly more predictive than others. The standard error bars when only one omic measure is considered are large, meaning there is a spread in accuracy. Omic measures should be included if they improve performance or contribute a significant number of samples not contained in other measures. In order to test this, all combinations of modalities were trained using MOGDx. The modality or combination of modalities which achieved the best classification performance and including all samples were reported in Table 2 with the performance of all other combinations reported in supplementary tables S2-S4. For the BRCA and KIPAN datasets, integration of all 5 modalities resulted in optimal performance. Whilst including all modalities was not necessary to produce optimal accuracy it did reduce the standard deviation of accuracy estimates from cross-validation. Conversely, training MOGDx on only DNAm resulted in the best performance on the LGG dataset. The DNAm dataset had all samples present, meaning there was full sample coverage. In this case, it was clear that DNAm was the only informative modality for tumour grade in the LGG dataset. This demonstrates that flexibility to train on any number and/or combination of modalities is an important requirement for integrative network approaches.

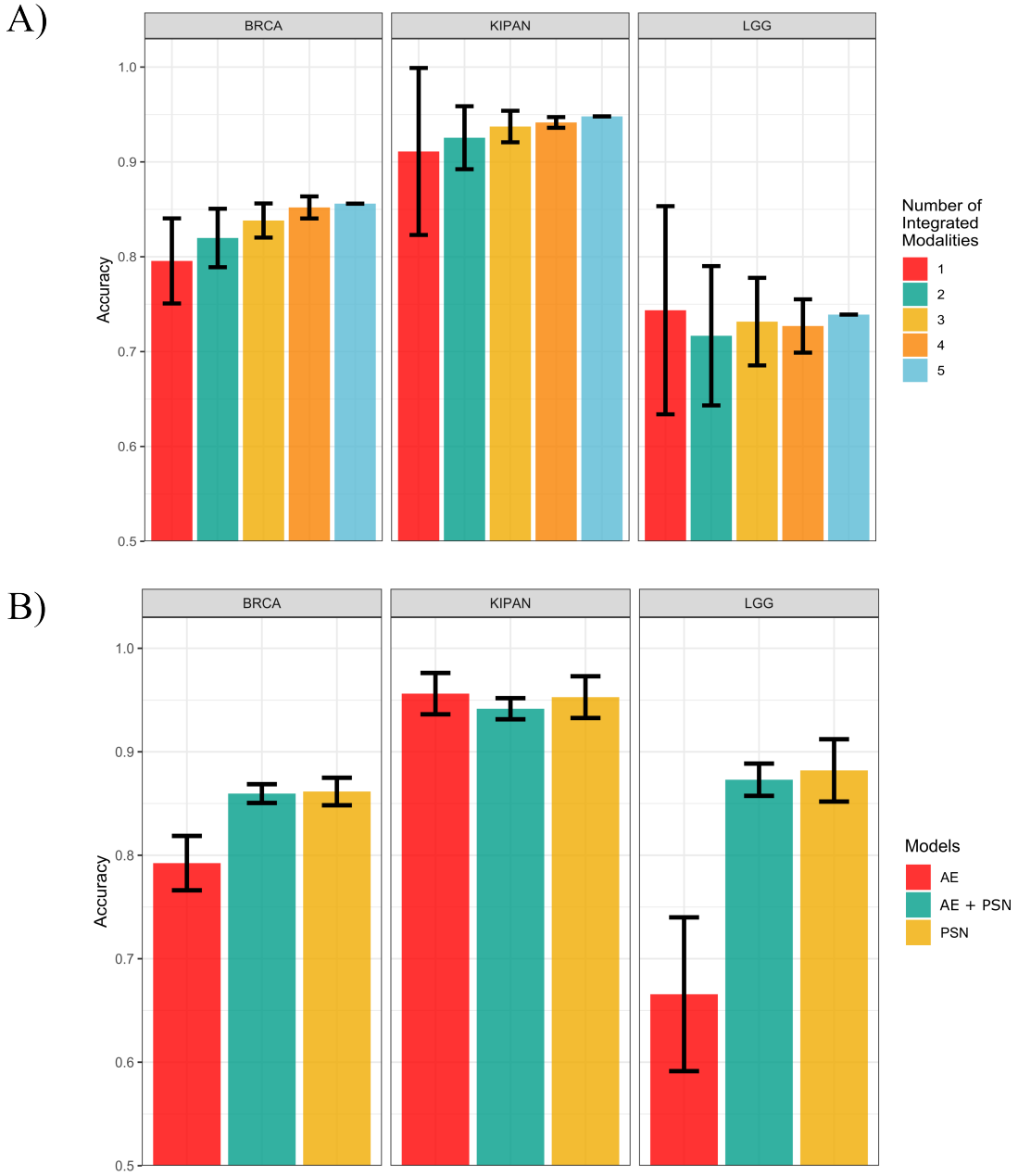

Figure 2. A Modality Importance | Mean accuracy per number of integrated modalities across the three datasets. The error bars represent the standard deviation of accuracy across all combinations of integrated modalities, e.g. The standard deviation of all 5 modalities is zero as there is only one combination. **B PSN Importance |** Performance of MOGDx on the best model when trained on AE only, PSN only and when trained on AE + PSN (MOGDx). The AE model has been trained on the AE only by removing all edges from the PSN. Similarly, the PSN model was trained by one hot encoding each node feature, thus only allowing the GCN to learn from the structure of the PSN. MOGDx learns from both AE node features and PSN.

Optimal performance is achieved when MOGDx is trained on fused PSN and node features

Figure 2.B demonstrates the predictive power of a GCN trained on a PSN. It is clear that the main predictive power comes from learning the relationships between patient samples. The PSN achieves the best mean accuracy in each dataset when k-fold cross validation is performed. Integrating the PSN with an AE reduces the variation in accuracy of the GCN to splits in the data. It is known that GNN's are sensitive to train/test and validation splits. Schur et al. showed

that classification metrics are susceptible to inflated results when models are trained on the same splits[19]. To overcome this limitation of GNN's MOGDx was trained on shuffled splits of the data. Embedding a reduced representation of the modality as node features also allows the GNN to learn from more than just the network structure, reducing this variability furthermore, as can be seen in Figure 2. In this manner, MOGDx achieves a balanced trade-off between accuracy and robustness by integrating the AE with the PSN.

MOGDx can identify relevant biological markers of heterogeneous diseases

Ablation experiments were performed to determine which omic measures were most predictive of each classification task. It was determined that mRNA and DNAm were most predictive of BRCA, all modalities except CNV were predictive of KIPAN and only DNAm was predictive of LGG. The ten most predictive features per omic measure, as determined by adjusted p-value, are shown in Table 3.

Table 3. Top 10 Important Biomarkers per Modality Identified by MOGDx

	Modalities	Markers				
BRCA	mRNA	FOXC1	FAM83D	GINS1	ESR1	XBP1
		ABCC2	GATA3	CA12	SPAG5	AURKA
	DNAm	cg25941751	cg08651590	cg02218932	cg04016621	cg02212575
		cg00011460	cg08836954	cg21861233	cg01065161	cg07951083
KIPAN	mRNA	COL23A1	MEF2C	ANGPT2	FLT1	NOVA2
		NPTX2	CA9	ABCC3	VEGFA	CDH6
	miRNA	hsa-mir-651	hsa-mir-221	hsa-mir-874	hsa-mir-222	hsa-mir-532
		hsa-mir-660	hsa-mir-188	hsa-mir-29a	hsa-mir-96	hsa-mir-766
	DNAm	cg25941751	cg08651590	cg02218932	cg04016621	cg02212575
		cg00011460	cg08836954	cg21861233	cg01065161	cg07951083
	RPPA	ASNS	ATR	CKIT	CYCLINE1	DRP1
		Grp75	ERCC1	CD31	PKA-a	Stat3
LGG	DNAm	cg25941751	cg08651590	cg02218932	cg04016621	cg02212575
		cg00011460	cg08836954	cg21861233	cg01065161	cg07951083

Further enrichment analysis was performed on all extracted features in the mRNA and DNAm modalities. Supplementary figure S3 and tables S5-S7 show the functional processes and genes associated with mRNA and DNAm respectively for; BRCA PAM50 sub-type classification, KIPAN sub-type classification and LGG grade classification. The functional processes found to be enriched in the BRCA mRNA dataset are related to the regulation of cellular and nucleic metabolic processes, which corroborates what is already known in literature[20,21]. The KIPAN mRNA dataset was enriched in processes relating to myeloid and leukocyte cells. These cells have known associations with carcinoma tumours and have a known association with survivability of these cancers[22,23]. Similarly, the DNAm enrichment identified known gene associations in all three datasets. The genes *CD44*, *CRIM1* and *USP1* were all enriched in the CpG sites extracted by MOGDx. These genes have been strongly associated with breast cancer and its prognosis[24–26]. Extracted CpG sites were also enriched in genes pertaining to the target dataset in KIPAN and LGG. For example, upregulation of *KRT8* is predictive of a poor prognosis in KIRC and *WIPF1* is indicative of a favourable prognosis in LGG[27,28].

t-SNE plots, shown in supplementary figure S1, show the grouping of clusters in each task. These figures show the groups that MOGDx has learnt in training. MOGDx finds good separation between all classes in the KIPAN and LGG datasets, however, there is no cluster for the 'Normal-like' sub-type in the BRCA dataset. There is also signification overlap between the "Luminal A" and "Luminal B" sub-types, which is likely due to the similarity of the two sub-types.

Discussion

Disease heterogeneity has moved medical research from a population-based perspective towards a personalised approach where diagnosis, prognosis and treatments are selected based on biomedical characteristics. Driving this movement is the development of large, diverse omic technologies and studies which provide labelled biomedical data at unprecedented levels. The integration of these omic measures offer the opportunity to build quantitative models, which can aid the understanding of heterogeneous disease architectures and inform clinical guidance. Therefore, a tool which can flexibly incorporate omic measures and identify specific biomedical characteristics based on these labels has the ability to redefine heterogeneous diseases.

We propose MOGDx, a network integrative architecture for the classification of heterogeneous diseases. What separates MOGDx from its competitors is the flexibility in its integration of omic measures, its inclusion of all available patient samples and its leverage of the predictive power of PSN's. MOGDx includes omic measures, which either improves predictive performance or include patients who may only have samples in one omic measure. This allows users to fine tune to the most predictive modalities while incorporating the maximum number of patient samples in an analysis. In this analysis, we maximised data usage while maintaining competitive or state-of-the-art performance on a variety of classification tasks. Fundamental to the predictive performance of MOGDx is the integration of PSN's. In this analysis, we have shown that patient similarity is a very effective determinant of heterogeneous disease sub-typing and grading. The use of PSN's is analogous to clinical diagnosis, where a diagnostician will compare a new patient to a database of similar cases. Similarly, MOGDx captures the variability in similarity and uses this to perform accurate sub-type classification and grading.

The application of MOGDx has been benchmarked on three cancer datasets from the TCGA, namely; BRCA, LGG and KIPAN. Cancer is widely regarded as a highly heterogeneous disease however, MOGDx was able to accurately classify breast cancer sub-types, kidney cancer sub-types and brain tumour grades from integrated omic data. MOGDx identified the optimal combination of modalities which resulted in greater patient coverage while maintaining a state-of-the-art classification performance compared to its competitors, as per Table 2.

Interpretability is an important aspect to consider for biomedical applications in order to transform research into novel diagnoses, grades or treatments. We have demonstrated the interpretability of MOGDx in several ways. Firstly,

through leave one out experiments we have identified the modalities which are most predictive of the classification task and their most important features, as summarised in Table 3. We have also performed enrichment analyses on the extracted features in the mRNA and DNAm modalities to identify the main drivers of their variability. Finally, we have plotted the latent embeddings from the GCN model as a t-SNE which shows the clusters the model has learnt in supplementary figure S1.

The use of different omics modalities allows us to assay different parts of the biological systems involved in disease mechanism, their integration can help reduce biological noise improving signal and allowing for the identification of previously undetectable informative features. Understanding which omic modalities are most predictive for a given disease can allow us to design more efficient and informative experiments, minimising impacts on patients and reducing costs. Further, because different omics modalities capture different components of the genetic and environmental contributions to disease their integration can help us to gain a more complete picture of disease. We performed enrichment analysis on mRNA and DNAm modalities if they were shown to be predictive for that disease. MOGDx was able to identify features enriched in processes and genes relating to the pathology and prognosis of the disease. These findings were supported by similar findings in the literature demonstrating MOGDx's ability to identify important omic markers. We have demonstrated in this work that the MOGDx architecture can successfully produce interpretable and reproducible insights into multiple heterogeneous diseases.

The cluster plots, shown in supplementary figure S1, show what clusters the GCN algorithm has learnt. The clusters in the KIPAN end LGG datasets show good separation between the class, with each class having their own distinct cluster. There is no distinct cluster for the 'Normal-like' subtype in the BRCA dataset. This could be an artefact of the small sample size, however small sample size was not an issue for the smallest class of the KIPAN dataset. As the BRCA dataset consists of individuals with a mutation in their BRCA gene, they are predisposed to developing the disease at some stage. As a result, it is possible that MOGDx is catching early signs of the different sub-types of the disease. This is an important finding as it could inform treatment strategies. For example, if someone was classified as 'Normal-like' but it was very likely they were to develop a more aggressive sub-type such as Basal, they should be treated differently. Further longitudinal studies to identify the predictive performance of MOGDx on the BRCA dataset would be an interesting avenue of further research.

The main drawback of MOGDx is that the GNN algorithm employed in this analysis, is a transductive algorithm. Transductive algorithms require the entire network to be available during training. In a clinical setting, this is not possible as it will be required to perform predictions on unseen patients. Hence, it will be required to extend MOGDx to an inductive algorithm which will not require the entire network to be available during training and can make predictions on unseen patients. In summary, MOGDx is a flexible and accurate classification tool which can be applied to a broad range of heterogeneous diseases.

Methods

Framework of MOGDx

The framework for MOGDx can be summarised into four main components; 1) Pre-processing and feature extraction, 2) Graph generation and SNF, 3) Dimensionality reduction and node feature augmentation and 4) GNN training and classification. Before integration, each modality is treated individually. An individual modality will undergo processing steps where an expression matrix and meta file for each modality is created. Feature extraction will be performed on this expression matrix and a PSN generated on the most important features. This PSN will then be used to create a network based on the KNN algorithm. The expression matrix will be inputted into a de-noising AE for dimensionality reduction. Integration is performed using SNF on the networks and simple concatenation on the reduced latent dimensions of the expression matrices. The nodes of the fused PSN are augmented with a vector representing the concatenated embedding from the AE. A GNN is trained on the combination of PSN and AE to perform heterogeneous disease classifications. MOGDx is a command line tool which can integrate a variable number of omic measures. Specific details of each component are described in the following sections.

Pre-processing and Feature Extraction

Pre-processing is performed to remove unwanted noise and variations in the data due to experimental or technical effects. For mRNA expression (mRNA) and micro RNA expression (miRNA) all genes which had either zero expression or zero variance in all samples were removed. Next, any samples which were more than 2 standard deviations from the mean node connectivity distance were removed. Differential expression was performed using a one-vs-the-rest methodology, and the most significantly differentially expressed genes were extracted.

The DNA methylation (DNAm) data downloaded from TCGA-Biolinks used multiple generations of Illumina Infinium DNA methylation arrays, however, they have been corrected and standardised using the SeSAMe[29] pipeline. Further steps were taken to remove any CpG sites which contained missing values. Important CpG sites were identified by performing a penalised Logistic Regression algorithm and keeping any CpG sites which had a non-zero weight.

To overcome significant missingness in the Copy Number Variation (CNV) and Reverse Phase Protein Array (RPPA) datasets, sites which contained more than 50% missingness were removed, and mean imputation was performed. The CNV data was log transformed to give it a close to normal distribution, and penalised Logistic Regression was applied to

both. The CNV and RPPA sites which had a non-zero weight were extracted.

Graph Generation and Similarity Network Fusion

The modalities were represented as graphs and Similarity Network Fusion (SNF) was performed to integrate the modalities. A patient similarity matrix was first created for each modality. The Pearson correlation coefficient (Eq. 1) between the extracted features was used as a measure of similarity.

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{(x_i - \bar{x})^2 (y_i - \bar{y})^2}} \tag{1}$$

The K-Nearest Neighbours (KNN) algorithm was used to build the graph with edges created between the 15 nearest neighbours. SNF[11] was applied to fuse the graphs into a single network representing the full spectrum of the underlying data. SNF allows complimentary information to be shared between modalities, and also is effective in identifying novel relationships between patients. It also integrates missing patient samples inherently by complimenting a missing edge in one modality with the same relationship from others.

Dimensionality Reduction and Node Feature Augmentation

One Autoencoder (AE) was constructed for dimensionality reduction of each modality. AE's consist of an encoder and decoder. The encoder (f) maps the original domain to a reduced latent space and the decoder (g) maps the latent space back to the original domain. The encoder consists of a single linear layer, which is batch normalised and passed through a sigmoid function. The aim of the AE is to minimise noise in the reconstruction of the latent space according to Eq. 2 where the loss is calculated as the mean square error (MSE) loss.

$$E = argmin_{f,g}\left(MSE_{Loss}(x, f(g(x)))\right) \tag{2}$$

The size of the latent space is arbitrary and can be tuned to each modality. If a patient is missing a sample from an omic measure, the latent embedding from this modality is imputed with the median value derived from all other patients. Each node in the fused network is augmented with a vector formed by concatenating the latent spaces of each modality.

GNN Training and Classification

GNN's are a powerful architecture for the learning of graph structure and information in a supervised setting. We implemented a GCN model from the StellarGraph[12] library in Python. The differentiation between GCN and neural network architectures is their ability to learn from the local neighbourhood as opposed to handcrafted features. The performance of GCN and other GNN architectures has been demonstrated on a variety of benchmark tasks, hence extending their application to a biomedical setting is an exciting avenue with great potential.

GCN requires two inputs. A network, consisting of nodes and edges, and a vector of features for each node. For MOGDx, the network created was a PSN and the vector of features was a reduced feature representation from the AE. Formulating the GCN algorithm as a network represented by an adjacency matrix $A \in R^{nxn}$ and a feature matrix $X \in R^{nxd}$ where n is the number of patients and d is the latent dimension selected for the AE. The GCN then consists of stacked convolutional layers defined by Eq. 3[30].

$$H^{l+l} = \sigma\left(LH^{(l)}W^{(l)}\right) \tag{3}$$

Where $L = \tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}$ is the normalised graph Laplacian; $\tilde{A} = A + I$ is the adjacency matrix; $\tilde{D}$ is the degree matrix of $\tilde{A}$; W is the weight matrix learned during training; σ is the non-linear activation function, ELU activation in this case, $H^{(l)}$ is the input to each layer and $H^{(0)}$ corresponds to X the node feature matrix.

Interpretability of MOGDx

Interpretability in biomedical applications is important to understand how specific features contribute to prediction so that therapeutic interventions or novel diagnoses can be well understood. MOGDx shows interpretability in a number of ways. Firstly, through ablation experiments, we can identify which omic measures are most predictive of the targeted outcome. Ablation experiments are widely adopted for feature importance and ranking in neural networks[31]. Similarly, we can treat entire modalities as features in MOGDx to identify the most informative. Enrichment analysis is a well understood methodology to map selected genes to their biological and molecular pathways. Functional enrichment analysis was carried out using the clusterProfiler[32] algorithm in R on the extracted features from the mRNA datasets. Similarly, enrichment analysis was carried out using the mCSEA[33] algorithm in R on the extracted CpG sites from the DNAm datasets. Results from these analyses were further compared to existing literature. TSNE is a statistical method to represent high dimensional data in an x-y plane. In this case, the GCN algorithm was able to map each point to a coordinate with close points having the same classification outcome. Through this visualisation, we can see which classes are most difficult to predict and which classes overlap in predictions.

References

1. Brodlie, M., Haq, I. J., Roberts, K. & Elborn, J. S. Targeted therapies to improve CFTR function in cystic fibrosis. *Genome Medicine* **7**, 101, DOI: 10.1186/s13073-015-0223-6 (2015).

2. Sosman, J. A. *et al.* Survival in BRAF V600-mutant advanced melanoma treated with vemurafenib. *The New Engl. J. Medicine* **366**, 707–714, DOI: 10.1056/NEJMoa1112302 (2012).

3. Gliozzo, J. *et al.* Heterogeneous data integration methods for patient similarity networks. *Briefings Bioinforma.* **23**, bbac207, DOI: 10.1093/bib/bbac207 (2022).

4. Lock, E. F., Hoadley, K. A., Marron, J. & Nobel, A. B. Joint and Individual variation explained (JIVE) for integrated analysis of multiple data types. *The annals applied statistics* **7**, 523–542, DOI: 10.1214/12-AOAS597 (2013).

5. Phan, J. H., Hoffman, R., Kothari, S., Wu, P.-Y. & Wang, M. D. Integration of Multi-Modal Biomedical Data to Predict Cancer Grade and Patient Survival. ... *IEEE-EMBS Int. Conf. on Biomed. Heal. Informatics. IEEE-EMBS Int. Conf. on Biomed. Heal. Informatics* **2016**, 577, DOI: 10.1109/BHI.2016.7455963 (2016). Publisher: NIH Public Access.

6. 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.

7. 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.

8. 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.

9. 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.

10. Li, X. *et al.* MoGCN: A Multi-Omics Integration Method Based on Graph Convolutional Network for Cancer Subtype Analysis. *Front. Genet.* **13**, 806842, DOI: 10.3389/fgene.2022.806842 (2022).

11. Wang, B. *et al.* Similarity network fusion for aggregating data types on a genomic scale. *Nat. Methods* **11**, 333–337, DOI: 10.1038/nmeth.2810 (2014). Number: 3 Publisher: Nature Publishing Group.

12. Data61, C. StellarGraph machine learning library (2018).

13. Colaprico, A. *et al.* TCGAbiolinks: TCGAbiolinks: An R/Bioconductor package for integrative analysis with GDC data, DOI: 10.18129/B9.bioc.TCGAbiolinks (2023).

14. Parker, J. S. *et al.* Supervised risk predictor of breast cancer based on intrinsic subtypes. *J. Clin. Oncol. Off. J. Am. Soc. Clin. Oncol.* **27**, 1160–1167, DOI: 10.1200/JCO.2008.18.1370 (2009).

15. Kensler, K. H. *et al.* PAM50 Molecular Intrinsic Subtypes in the Nurses' Health Study Cohorts. *Cancer epidemiology, biomarkers & prevention : a publication Am. Assoc. for Cancer Res. cosponsored by Am. Soc. Prev. Oncol.* **28**, 798–806, DOI: 10.1158/1055-9965.EPI-18-0863 (2019).

16. Tabibu, S., Vinod, P. K. & Jawahar, C. V. Pan-Renal Cell Carcinoma classification and survival prediction from histopathology images using deep learning. *Sci. Reports* **9**, 10509, DOI: 10.1038/s41598-019-46718-3 (2019).

17. Forst, D. A., Nahed, B. V., Loeffler, J. S. & Batchelor, T. T. Low-Grade Gliomas. *The Oncol.* **19**, 403–413, DOI: 10.1634/theoncologist.2013-0345 (2014).

18. Zhang, G. *et al.* A novel liver cancer diagnosis method based on patient similarity network and DenseGCN. *Sci. Reports* **12**, 6797, DOI: 10.1038/s41598-022-10441-3 (2022).

19. Shchur, O., Mumme, M., Bojchevski, A. & Günnemann, S. Pitfalls of Graph Neural Network Evaluation (2019). ArXiv:1811.05868 [cs, stat].

20. Werner, H. BRCA1: An Endocrine and Metabolic Regulator. *Front. Endocrinol.* **13** (2022).

21. Privat, M. *et al.* BRCA1 Induces Major Energetic Metabolism Reprogramming in Breast Cancer Cells. *PLoS ONE* **9**, e102438, DOI: 10.1371/journal.pone.0102438 (2014).

22. Schmid, M. C. & Varner, J. A. Myeloid Cells in the Tumor Microenvironment: Modulation of Tumor Angiogenesis and Tumor Inflammation. *J. Oncol.* **2010**, 201026, DOI: 10.1155/2010/201026 (2010).

23. Danaher, P. *et al.* Gene expression markers of Tumor Infiltrating Leukocytes. *J. for ImmunoTherapy Cancer* **5**, 18, DOI: 10.1186/s40425-017-0215-8 (2017).

24. Wright, M. H. *et al.* Brca1 breast tumors contain distinct CD44+/CD24- and CD133+ cells with cancer stem cell characteristics. *Breast cancer research: BCR* **10**, R10, DOI: 10.1186/bcr1855 (2008).

25. Wen, W. *et al.* Low CRIM1 Levels Predict Poor Prognosis in Breast Cancer Patients. *Front. Oncol.* **12**, 882328, DOI: 10.3389/fonc.2022.882328 (2022).

26. Suan Lim, K. *et al.* USP1 is Required for Replication Fork Protection in BRCA1-Deficient Tumors. *Mol. cell* **72**, 925–941.e4, DOI: 10.1016/j.molcel.2018.10.045 (2018).

27. Tan, H.-S. *et al.* KRT8 upregulation promotes tumor metastasis and is predictive of a poor prognosis in clear cell renal cell carcinoma. *Oncotarget* **8**, 76189–76203, DOI: 10.18632/oncotarget.19198 (2017).

28. Staub, E. *et al.* An expression module of WIPF1-coexpressed genes identifies patients with favorable prognosis in three tumor types. *J. Mol. Medicine (Berlin, Ger.* **87**, 633–644, DOI: 10.1007/s00109-009-0467-y (2009).

29. Zhou, W., Triche, T. J., Laird, P. W. & Shen, H. SeSAMe: reducing artifactual detection of DNA methylation by Infinium BeadChips in genomic deletions. *Nucleic Acids Res.* **46**, e123, DOI: 10.1093/nar/gky691 (2018).

30. Kipf, T. N. & Welling, M. Semi-Supervised Classification with Graph Convolutional Networks (2017). ArXiv:1609.02907 [cs, stat].

31. Setiono, R. & Liu, H. Neural-network feature selector. *IEEE Transactions on Neural Networks* **8**, 654–662, DOI: 10.1109/72.572104 (1997). Conference Name: IEEE Transactions on Neural Networks.

32. Wu, T. *et al.* clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. *The Innov.* **2**, 100141, DOI: 10.1016/j.xinn.2021.100141 (2021).

33. Martorell-Marugán, J., González-Rumayor, V. & Carmona-Sáez, P. mCSEA: detecting subtle differentially methylated regions. *Bioinformatics* **35**, 3257–3262, DOI: 10.1093/bioinformatics/btz096 (2019).

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. For the purpose of open access, the author has applied a creative commons' attribution [CC BY] licence to any author accepted manuscript version arising.

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

Author contributions statement

BR gathered all data, performed analysis, designed the study, conducted experiments and drafted the manuscript. TIS contributed to analysis, results and discussions. TIS and REM supervised the study, revised the manuscript and approved the final version of the manuscript.

Competing Interests

REM is a scientific advisor to Optima Partners and the Epigenetic Clock Development Foundation.

Additional information

Data Availability

All data is available to download from The Cancer Genome Atlas (TCGA)(https://www.cancer.gov/tcga). A script to download and obtain data exactly as it is presented is available from https://github.com/Barry8197/MOGDx.

Supplementary Material

Supplementary Material is available in the accompanying files S1FiguresTablesMOGDx.pdf and S2ExtractedFeaturesMOGDx.pdf