Benchmarking Multi-Omics Latent Factor Methods to Predict Anticancer Drug Response Using Baseline Cancer Cell Line Data

by

Shannon Brown

Thesis submitted to the University of Nottingham in partial fulfilment of the requirements for the degree of MRes in Bioinformatics

September 2022

Acknowledgements

I would like to express my sincere gratitude to my supervisor Dr Stephan Gade for providing his invaluable knowledge, support and patience throughout the course of the project. Additionally, many thanks to Dr Matthew Heatley for his advice and feedback, which has greatly improved my academic writing. Special thanks to Dr Emma Laing for her feedback, but most of all, her moral support and encouragement that helped to build my self-confidence and kept me motivated during the project. Thanks also to all my colleagues at GSK for always being welcoming, supportive and willing to share their knowledge. Lastly, I'd like to recognise my partner for his understanding and belief in me, and my dog for the entertainment and emotional support.

It’s the little details that are vital. Little things make big things happen.

– John Wooden

Abstract

Cancer is a complex, heterogeneous disease that arises from genomic instability causing molecular alterations at multiple levels, making it notoriously difficult to treat. Computational methods are being applied to discover predictive biomarkers of drug response, which aim to resolve these challenges by focusing on the genotype, rather than the phenotype of tumours. Latent factor methods enable simultaneous analysis of multiple omics datasets and so hold unprecedented opportunity to understand the relationship between molecular layers. These methods have been successfully applied in biomarker discovery, however, there are many different methods available. In this study two latent factor methods, multi-omics factor analysis (MOFA) and multiple co-inertia analysis (MCIA), are benchmarked using baseline cancer cell line data. Performance is evaluated using three objectives. The first investigates quality control of multi-omics data processing, the second compares the latent representation of data, and the final evaluates and compares method ability to predict anticancer drug response. Our analysis shows different underlying statistical frameworks result in contrasting model sensitivity to noise and bias, and as a result, produce different low-dimensional representations of data. Both methods were equally able to explain variance in drug response across cell lines, however, not enough to be able to predict response. We conclude that further studies are required to determine whether this result is due to inadequately processed data, or due to true inability of either method to predict drug response.

Table of Contents

Abbreviations

ATAC-seq	Assay for transposase-accessible chromatin sequencing
AUC	Area under the dose response curve
CCA	Canonical correlation analysis
CCLE	Cancer Cell Line Encyclopaedia
CIA	Co-inertia analysis
CLL	Chronic lymphocytic leukaemia
COA	Correspondence analysis
CTRP	Cancer Therapeutics Response Portal
GDSC	Genomics of drug sensitivity in cancer
GFA	Group factor analysis
GWAS	Genome wide association study
HVF	Highly variable feature
intNMF	Integrative non-negative matrix factorisation
JIVE	Joint and individual variance explained
LC-MS/MS	Liquid chromatography-tandem mass spectrometry
MCIA	Multiple co-inertia analysis
MOFA	Multi-omics factor analysis
MS	Mass spectrometry
MSFA	Multi-study factor analysis
NCBI	National Center for Biotechnology Information
NCI60	National Cancer Institute 60
PCA	Principal components analysis
PLS	Partial least squares
QC	Quality control
RGCCA	Regularised generalised canonical correlation analysis
RNA-seq	RNA sequencing
SD	Standard deviation
TCGA	The Cancer Genome Atlas
tICA	Tensorial independent component analysis
TMT	Tandem mass tags
TPM	Transcripts count per million
VSN	Variance stabilising normalisation
VST	Variance stabilising transformation

List of Tables

List of Figures

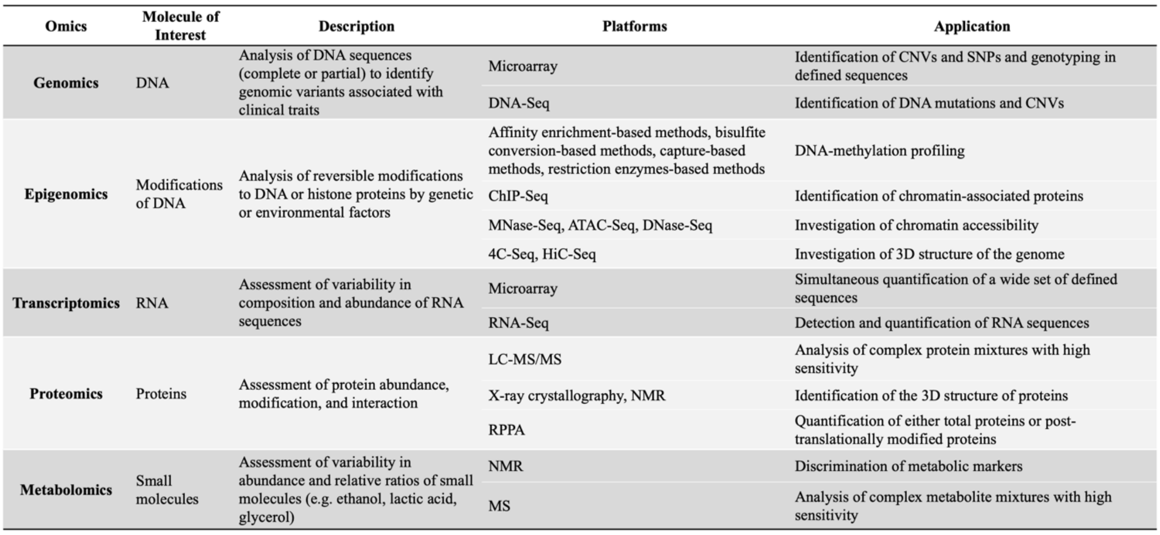
Chapter 1. Introduction

Cancer is one of the leading causes of death worldwide making it a major area of focus in research and the pharmaceutical industry. Due to the complexity and heterogeneity of tumours, there is no one-size-fits-all therapy, making cancer challenging to treat. There has been a shift in current approaches to therapy as a result of genome-wide screening. This has enabled patient stratification using molecular biomarkers to help inform on diagnosis and treatment. By targeting specific proteins, drugs are delivered with higher specificity and lower toxicity, resulting in improved patient outcomes. However, there are limitations to targeted therapies. A key example being, despite patients having presence of a specific target, drugs remain ineffective or only partially effective [1]. This can be related to the complex mutational landscape and molecular cross-talk driving cancer progression [1, 2].

Molecular alterations must occur at many levels (e.g. genome, transcriptome, proteome) for cells to undergo malignant transformations [3]. Thus, looking at single-level omics data in isolation is not enough to establish causal relationships between phenotype and molecular alterations. To fully understand the development of cancer, the complexity of interactions that take place between dynamic molecular layers and the influence of environmental factors must be deciphered. This encompasses a systems biology approach, where integrating multidisciplinary data can capture different aspects of cellular function to understand biological interactions holistically and systematically [3].

1.1. Multi-Omics in Oncology

Over the past decade there have been significant advances in cost-effective high-throughput omics technologies, enabling large-scale data generation in each field. Omics technology refers to biochemical assays that measure biological molecules of the same type. Each omics offers a different view of biological function and organisation of molecular systems [4]. The development of these technologies has followed the central dogma, starting by capturing alterations of the genome (genomics), differentially expressed genes driving disease (transcriptomics) and protein expression profiling and post-translational modifications (proteomics). Beyond this, there has been expansion to investigating modifications in epigenetic regulation of the genome (epigenomics) and metabolic regulation of the cell (metabolomics) (see Table 1.1) [5, 6]. The number of omics fields has continued to grow over time, to areas such as lipidomics (analysis of lipids), glycomics (analysis of sugars) and microbiomics (analysis of microbiota), to name a few [7]. These more recent areas of research are gaining traction in oncology, however, fall out of scope for this project. Common assays used for omics analysis are RNA sequencing (RNA-seq) for transcriptome profiling, assay for transposase-accessible chromatin sequencing (ATAC-seq) for epigenome profiling and mass spectrometry (MS) for metabolome and proteome profiling. RNA-seq provides fast, precise quantification of transcripts and their isoforms by converting long RNAs to cDNA fragments and harnessing high-throughput sequencing technology to output sequence reads that are aligned to a reference genome [8]. ATAC-seq is one of many methods used to investigate the epigenome, specifically it assesses chromatin accessibility. Epigenomic assays vary in specificity, alternatives assess other forms of regulation such as DNA modifications or histone modifications, which impact chromatin dynamics and structure. Chromatin can exist in several different states which associate with particular patterns of gene regulation. ATAC-seq is a simple, scalable technique that uses *in vitro* transposition of sequencing adapters into chromatin to create sequenceable DNA fragments. Following fragment alignment, peaks of accessible chromatin are identified by enrichment of transposition events in genomic regions [9]. Lastly, MS provides diverse utility in metabolomics, proteomics, phosphoproteomics and the emerging field of lipidomics. The biomolecule of interest is separated, ionised and vaporised to form gas phase ions, which are input into the mass spectrometer. Ions are sorted according to their mass-to-charge ratio to produce a mass spectrum of ion abundance that can be mapped back to peptides/metabolites based on mass [10].



Omics	Molecule of Interest	Description	Platforms	Application
Genomics	DNA	Analysis of DNA sequences (complete or partial) to identify genomic variants associated with clinical traits	Microarray	Identification of CNVs and SNPs and genotyping in defined sequences
			DNA-Seq	Identification of DNA mutations and CNVs
Epigenomics	Modifications of DNA	Analysis of reversible modifications to DNA or histone proteins by genetic or environmental factors	Affinity enrichment-based methods, bisulfite conversion-based methods, capture-based methods, restriction enzymes-based methods	DNA-methylation profiling
			ChIP-Seq	Identification of chromatin-associated proteins
			MNase-Seq, ATAC-Seq, DNase-Seq	Investigation of chromatin accessibility
			4C-Seq, HiC-Seq	Investigation of 3D structure of the genome
Transcriptomics	RNA	Assessment of variability in composition and abundance of RNA sequences	Microarray	Simultaneous quantification of a wide set of defined sequences
			RNA-Seq	Detection and quantification of RNA sequences
Proteomics	Proteins	Assessment of protein abundance, modification, and interaction	LC-MS/MS	Analysis of complex protein mixtures with high sensitivity
			X-ray crystallography, NMR	Identification of the 3D structure of proteins
			RPPA	Quantification of either total proteins or post-translationally modified proteins
Metabolomics	Small molecules	Assessment of variability in abundance and relative ratios of small molecules (e.g. ethanol, lactic acid, glycerol)	NMR	Discrimination of metabolic markers
			MS	Analysis of complex metabolite mixtures with high sensitivity

Table 1.1. Overview of omics data types and technologies applied in cancer research.

Table adapted from Gallo Cantafio, M.E., et al. 2018 and Hasin,Y., et al. 2017 [11, 12].

CNV = copy number variation; SNP = single-nucleotide polymorphism; MNase = micrococcal nuclease; ATAC = assay for transposase-accessible chromatin; 4C = chromosome conformation capture-on-chip; HiC = high-throughput chromosome conformation capture; LC-MS/MS = liquid chromatography-tandem mass spectrometry; NMR = nuclear magnetic resonance; RPPA = reverse phase protein array

As previously mentioned, the synergistic interactions and complementary effects between omics layers cannot be assessed by the reductionist approach of single-omics analysis. Multi-omics offers an integrated approach to understand the relationship between molecules and the flow of information in dynamic multi-dimensional biological networks [13]. This approach holds promise to bridge the gap from genotype to phenotype [3]. A major challenge in cancer is distinguishing the small number of driver mutations that provide selective advantage to tumourigenic cells from the vast number of passenger mutations that do not alter the phenotype. Integration of omics data can intensify relevant signals underlying disease mechanism and cancer progression to uncover driver somatic mutations and thus enabling dissection of the heterogeneity of cancer cells. This in turn contributes to revealing cancer subtypes, finding reliable biomarkers and discovering potential drug targets (see Figure 1.1) [3, 14]. *In silico* techniques, such as this, are key in modern drug development to help prioritise new targets and stratify patients using biomarkers in clinical decision support [3, 15]. Biomarkers assist earlier diagnosis to prevent cancer-related deaths, improve the prognostic and predicative accuracy of disease progression and clinical outcomes and lastly, advance clinical subtyping. It is common within cancer subtypes for patients to have varying degrees of responsiveness to therapies, thus better patient stratification utilising preserved clinical and molecular biomarkers are required to predict suitable interventions for patient groups. This in turn can help to improve patient outcomes, increase understanding of drug mode of action and prevent development of drug resistance [2, 3, 14-16]. An additional interesting application of multi-omics in the pharmaceutical industry is in drug repurposing (also known as drug repositioning). This strategy aims to expand opportunities for approved drugs outside of the original medical use, providing potential benefits of lower development costs and shorter development times by using already de-risked compounds. Computational approaches like multi-omics

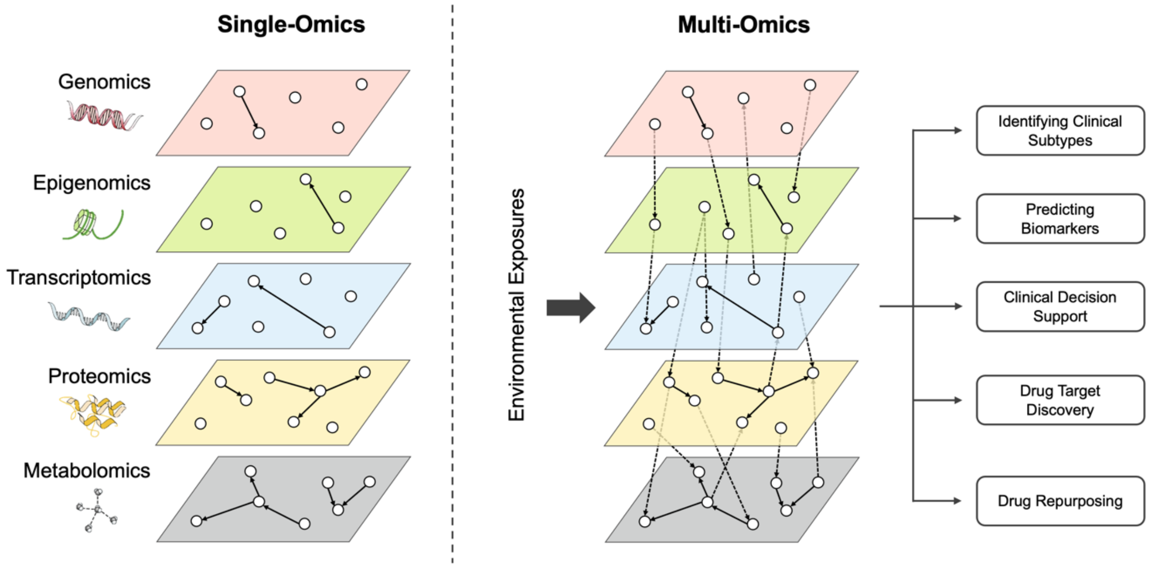


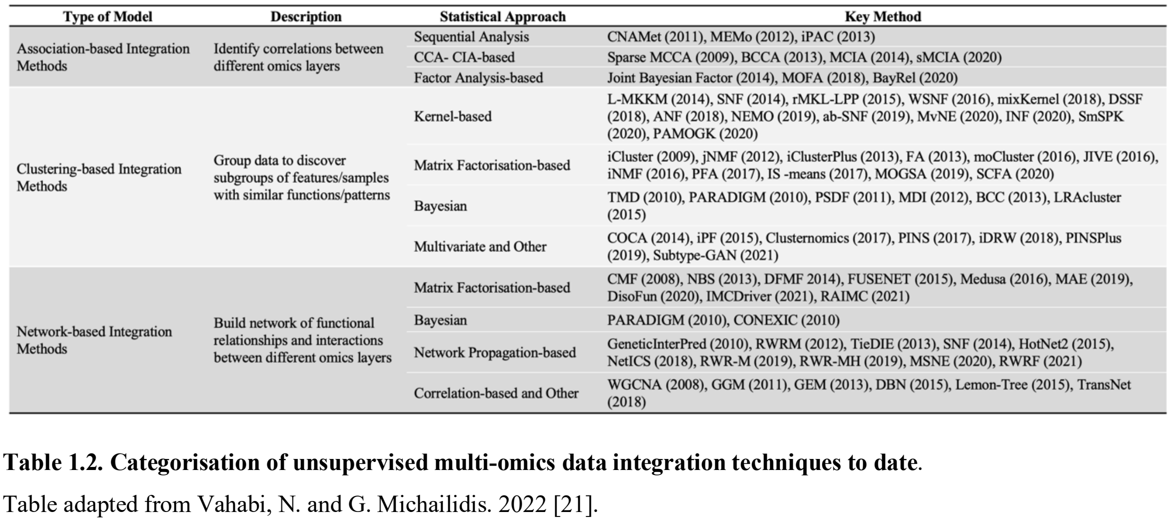
Figure 1.1. Conceptual model of single-omics analysis compared to multi-omics analysis and its applications in oncology.

Each coloured rhomboid represents a molecular layer (genomics, epigenomics, transcriptomics, proteomics and metabolomics). Solid arrows depict interactions of features (white dots) within the same molecular layer and dashed arrows depict interactions between molecular layers. Figure adapted from Hasin,Y., et al. 2017 and Yugi, K., et al. 2016 [12,19].

can help in formulating drug repurposing hypotheses to speed-up shortlisting candidates for assessment in preclinical models [17]. This is of particular interest in oncology, where drug attrition rates are higher relative to other therapeutic areas [18].

1.2. Multi-Omics Data Integration Approaches

Depending on the focus of the investigation, integrative analysis can be approached in two different ways: bottom-up or top-down. A bottom-up strategy can also be thought of as phenotype-first, where the investigation is centred on a given disease and seeks to understand the pathways associated, rather than focusing on a particular locus. Alternatively, a top-down strategy tries to determine how a GWAS locus of interest contributes to disease, hence this is also referred to as a genome-first approach [3, 12, 20]. There have been an enormous variety of integration techniques developed over the years, that can be categorised in a multitude of ways. Here, integration techniques are categorised in three tiers: 1) type of machine learning, 2) type of model and 3) statistical approach. Additional documented ways of categorising are, using biological objective (e.g. biomarker prediction, disease subtyping or disease insights) or method of data ensemble (i.e. whether integration occurred before or after data analysis) [15, 21]. There are two main types of machine learning, supervised and unsupervised. For supervised data integration, the model is trained with data labelled with known outcome variables for prediction. In contrast, unsupervised data integration aims to draw inferences and find patterns in unlabelled data, where the outcomes are unknown. Currently in the field, the number of unsupervised methods outweighs the number of supervised, with the majority showcasing methods for disease classification or biomarker discovery [21-23]. The next level of categorisation looks at the type of model, which refers to how the method derives actionable insight from data. Categories for unsupervised methods include association-based, clustering-based and networks-based (see Table 1.2). Association-based methods look for correlations between different molecular assays, while clustering-based methods group data to



Type of Model	Description	Statistical Approach	Key Method
Association-based Integration Methods	Identify correlations between different omics layers	Sequential Analysis	CNAMet (2011), MEMo (2012), iPAC (2013)
		CCA- CIA-based	Sparse MCCA (2009), BCCA (2013), MCIA (2014), sMCIA (2020)
		Factor Analysis-based	Joint Bayesian Factor (2014), MOFA (2018), BayRel (2020)
Clustering-based Integration Methods	Group data to discover subgroups of features/samples with similar functions/patterns	Kernel-based	L-MKKM (2014), SNF (2014), rMKL-LPP (2015), WSNF (2016), mixKernel (2018), DSSF (2018), ANF (2018), NEMO (2019), ab-SNF (2019), MvNE (2020), INF (2020), SmSPK (2020), PAMOGK (2020)
		Matrix Factorisation-based	iCluster (2009), jNMF (2012), iClusterPlus (2013), FA (2013), moCluster (2016), JIVE (2016), iNMF (2016), PFA (2017), IS -means (2017), MOGSA (2019), SCFA (2020)
		Bayesian	TMD (2010), PARADIGM (2010), PSDF (2011), MDI (2012), BCC (2013), LRAcluster (2015)
		Multivariate and Other	COCA (2014), iPF (2015), Clusternomics (2017), PINS (2017), iDRW (2018), PINSPlus (2019), Subtype-GAN (2021)
Network-based Integration Methods	Build network of functional relationships and interactions between different omics layers	Matrix Factorisation-based	CMF (2008), NBS (2013), DFMF 2014), FUSENET (2015), Medusa (2016), MAE (2019), DisoFun (2020), IMCDriver (2021), RAIMC (2021)
		Bayesian	PARADIGM (2010), CONEXIC (2010)
		Network Propagation-based	GeneticInterPred (2010), RWRM (2012), TieDIE (2013), SNF (2014), HotNet2 (2015), NetICS (2018), RWR-M (2019), RWR-MH (2019), MSNE (2020), RWRF (2021)
		Correlation-based and Other	WGCNA (2008), GGM (2011), GEM (2013), DBN (2015), Lemon-Tree (2015), TransNet (2018)

Table 1.2. Categorisation of unsupervised multi-omics data integration techniques to date.

Table adapted from Vahabi, N. and G. Michailidis. 2022 [21].

discover biologically relevant subgroups of features or samples. Lastly, network-based methods seek to build networks of functional relationships between features from different modalities i.e. data types with different formation methods and internal structures [21]. The plethora of methods spanning various statistical approaches presents both an opportunity and a challenge. It is recommended to benchmark data integration methods, doing a rigorous performance comparison using the same multi-omics datasets. This highlights the strengths of each method and potential advantages for usage when investigating a particular biological objective [15, 24].

Regardless of multi-omics strategy or integration technique, the field still presents many challenges across data collection and data integration. Non-uniform missing data is a common challenge during data collection. Missing values can arise from features failing to be measured or whole sample measurements being unavailable either as a result of quality control (QC) procedures or potential unbalanced study design. It is possible for feature values to be imputed or re-measured using alternative technologies, while missing samples are more disruptive. Another challenge is heterogeneity in signal-to-noise ratio between assays. Differing precision levels between assays could potentially lead to false conclusions, as a weak association between molecules could be due to a true lack of relationship or as a result of poor detection. Lastly, inefficient computation and storage is creating a bottleneck in analysis, as the cost per unit measurement is reducing, large volumes of data are being generated with expensive long-term storage requirements. Cloud computing infrastructures have multicore central processing units available for parallel computing, however implementing this efficiently on high-dimensional data is problematic [25, 26].

1.2.1. Latent Factor Methods

This project focuses on unsupervised latent factor methods of integration, these can also be referred to as dimensionality reduction methods. Dimension reduction aims to map data to a lower dimensional space that is represented by a set of new variables, which aim to explain that the majority of variance present across the data. These new variables are referred to as latent factors, as they are linear combinations of the original variables that are not directly observable in the data. Latent factor methods involve using matrix decomposition, reducing the data into a small number of latent factors, which represent the underlying biological processes that are being measured, and a set of loadings which represent the weights of features in the model, thus indicating the relative importance of each feature in explaining the variation in the data. The factors and loadings are chosen to minimise the reconstruction error, which is the distance between the original data point and its projection in the lower-dimensional space [27].

Each omics dataset in an analysis can be viewed as a matrix ($\mathbf{X}^{\mathbf{i}}$) of dimension $n \times p$ with p_i features and n samples, where p ranges from thousands to millions thus presenting a large dimensional space. Matrices are decomposed into the product of weight matrices ($\mathbf{A}^{\mathbf{i}}$) of dimensions $p_i \times k$, and a factor matrix (**F**) of dimensions $k \times n$, where k represents the number of latent factors (see Figure 1.2). Latent factors in each derived matrix ($\mathbf{A}^{\mathbf{i}}$ or **F**) represent projections of biological signals on different spaces and so can extract different insights. Factors from weight matrices represent projections on the feature space and so top-ranked features can inform on markers or pathways associated with the variance explained. Meanwhile, the factor matrix represents the relationships defined by all omics datasets i.e. projections on the sample space, where factors can be interpreted similarly to principal components for sample clustering [28, 29]. Dimension reduction is advantageous as it offers more robust and sensitive conclusions that are less likely to reflect technical or batch effects [28].

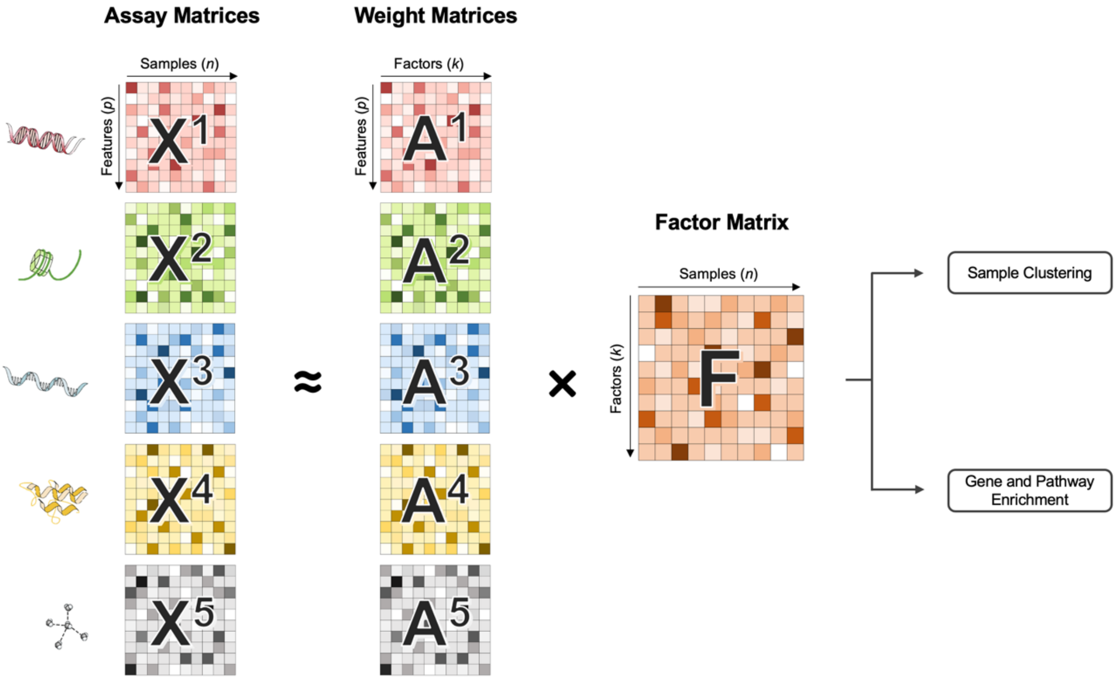
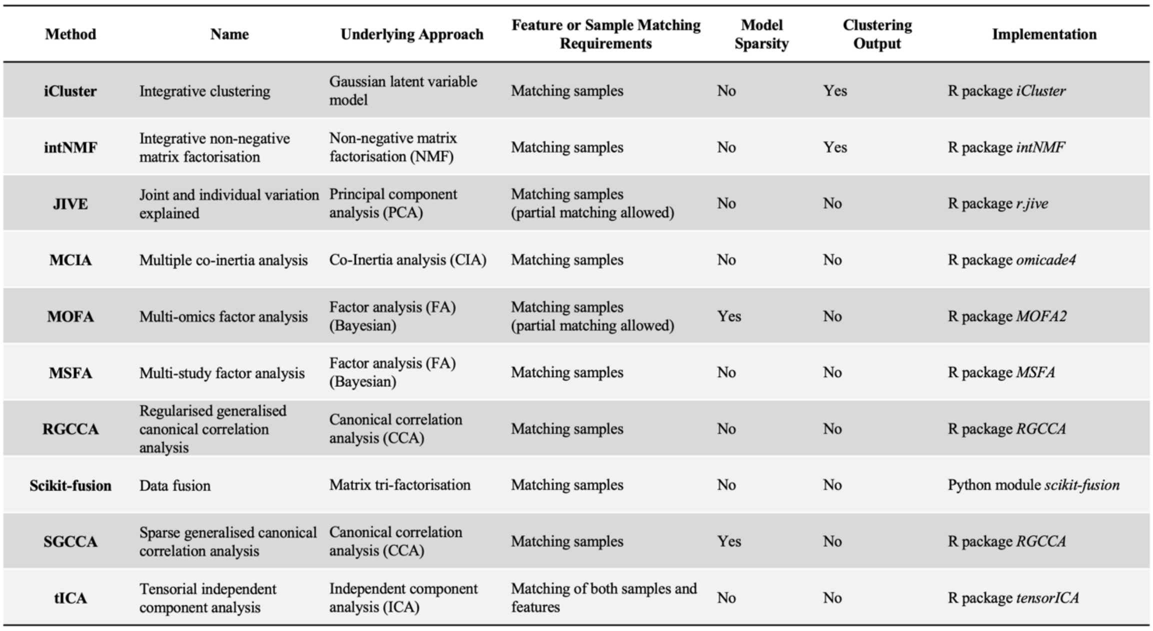


Figure 1.2. Latent factor model overview.

Multi-omics are measured from a matched set of samples. Each omics corresponds to a matrix $\mathbf{X}^{\mathbf{i}}$, which is factorised into the product of weight matrices $\mathbf{A}^{\mathbf{i}}$ and factor matrix **F**. These products can then be used for sample clustering and gene/pathway enrichment to identify molecular processes. Figure adapted from Cantini, L., et al. 2021 [29].

Two key considerations when inputting data to latent factor methods are feature variance and feature filtering (also known as feature selection). Number of variables and count scales vary between different omics data which results in different variance. This can potentially cause latent factors to be dominated by more variable datasets. Hence, pre-processing is required to centre and normalise features prior to integration to prevent overlooking small sources of biological variance [28]. Only a small set of features contribute to major biological processes and tend to be correlated across modalities [30]. Additionally, modalities with larger numbers of features are often over-represented in latent factors [29]. Together this highlights the requirement of feature filtering during data processing, ideally based on variability, to make feature number and variance comparable and reduce noise entering the model [21, 29, 31].

Table 1.2 highlighted the vast range of software available to execute unsupervised multi-omics analysis. Within the subset of unsupervised latent factor methods there are different aspects that differentiate them beyond statistical approach (see Table 1.3). For example, ability to integrate datasets with mis-matched features and samples. It is common for multi-omics datasets to have overlapping samples but varying numbers of unmatched features, making integration more complex. Datasets with matched features are rare but can be created by converting all features to the same molecular level (e.g. genes), however this is not always feasible. For example, miRNAs can't be converted to gene symbols. The vast majority of methods are able to cope with unmatched features, however very few can handle unmatched samples i.e. samples must be profiled for all omics [28]. Model sparsity is another aspect that differentiates methods. Sparsity essentially follows the concept of "less is more" and aims to have a small number of non-zero parameters or weights in the model by implementing an implicit feature selection. This can be use useful for high-dimensional datasets, such as multi-omics, where it can help to reduce overfitting, increase efficiency and improve the overall interpretability of the model. There are several techniques that can be used to encourage sparsity in a model, such as regularisation methods, that penalise the model for having a large number of non-zero parameters. These methods can be used



Method	Name	Underlying Approach	Feature or Sample Matching Requirements	Model Sparsity	Clustering Output	Implementation
iCluster	Integrative clustering	Gaussian latent variable model	Matching samples	No	Yes	R package *iCluster*
intNMF	Integrative non-negative matrix factorisation	Non-negative matrix factorisation (NMF)	Matching samples	No	Yes	R package *intNMF*
JIVE	Joint and individual variation explained	Principal component analysis (PCA)	Matching samples (partial matching allowed)	No	No	R package *r.jive*
MCIA	Multiple co-inertia analysis	Co-Inertia analysis (CIA)	Matching samples	No	No	R package *omicade4*
MOFA	Multi-omics factor analysis	Factor analysis (FA) (Bayesian)	Matching samples (partial matching allowed)	Yes	No	R package *MOFA2*
MSFA	Multi-study factor analysis	Factor analysis (FA) (Bayesian)	Matching samples	No	No	R package *MSFA*
RGCCA	Regularised generalised canonical correlation analysis	Canonical correlation analysis (CCA)	Matching samples	No	No	R package *RGCCA*
Scikit-fusion	Data fusion	Matrix tri-factorisation	Matching samples	No	No	Python module *scikit-fusion*
SGCCA	Sparse generalised canonical correlation analysis	Canonical correlation analysis (CCA)	Matching samples	Yes	No	R package *RGCCA*
tICA	Tensorial independent component analysis	Independent component analysis (ICA)	Matching of both samples and features	No	No	R package *tensorICA*

Table 1.3. Subset of unsupervised latent factor methods for multi-omics data integration.

Table adapted from Cantini, L., et al. 2021 [29] to automatically select a small number of important features from the data and set the weights for other features to zero [32]. This investigation will focus on benchmarking MOFA and MCIA, two high performing but statistically divergent techniques.

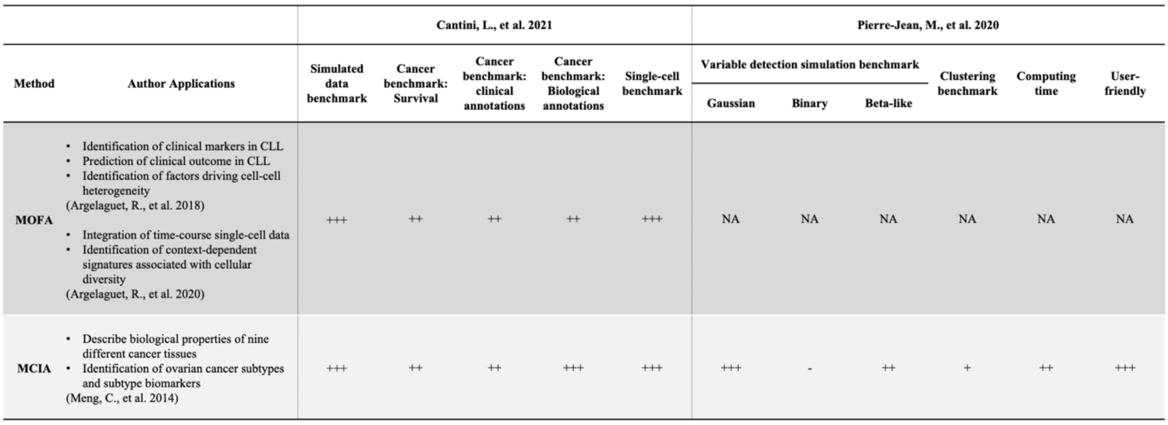
MOFA [33] is simultaneously an extension of factor analysis and a generalisation of PCA, that captures major sources of variation across omics datasets. MOFA differs from other methods by having a noise model and a sparsity constraint. The noise model aims to understand and reduce the impact of random error, i.e. noise, on the accuracy and precision of estimates. While the sparsity constraint is a type of regularisation added to the loss function of the model, where the loss function measures the error between the predicted output and the true output. It encourages the model to have fewer non-zero parameters, by penalising the loss function proportional to the number of non-zero parameters. During training, the model will try to minimise the loss function, to prune away less important parameters and become more sparse [32]. This means many of the loadings for latent factors will be zero or close to zero, which in turn means latent factors only strongly associate with a small number of omics data types, rather than all of them. Another important characteristic of MOFA is that latent factors can be correlated and are not constrained to orthogonality. Imposing orthogonality would infer omics are independent, which may not be a realistic assumption on the data. Factor correlations are a disadvantage of having a sparse model, therefore it is important to select an optimal number of latent factors whereby they capture independent sources of variation [33]. The MOFA R package has built-in functionalities for downstream analysis, such visualisation, annotation of latent factors and imputation of missing values. MOFA has been successfully applied to a wide range of biological questions. For example, in the study of chronic lymphocytic leukaemia (CLL) MOFA was able to extract already known clinical markers in addition to novel biomarkers, of which some were found to be predictive of clinical outcome [33]. In microbiology MOFA was able to derive molecular signatures partitioning the microbiome compositions of critically ill patients, healthy patients and patients on antibiotic treatment [34] and lastly in systems toxicology MOFA robustly extracted molecular mechanisms activated by cigarette smoking in mouse lungs [35]. In 2020 Argelaguet et al published MOFA+ [36], a new implementation of MOFA with all the same features but with a new multi-group framework and incresed computation speed. Analysis aims to find factors shared across groups or explanatory of single groups, importantly, factors do not separate groups. Furthermore, the ability to investigate spatial and temporal relationship has been added to MOFA+ since release [37]. For clarity, MOFA and MOFA+ will be referred to synonymously as MOFA.

MCIA [38] is an extension of co-inertia analysis (CIA), which was originally applied in environmental and ecological studies. It aims to find latent factors, also called co-inertia axes, by finding the linear combinations of variables in the different datasets that maximise the co-inertia, which is the measure of similarity between datasets. The resulting linear combinations are the latent factors, while the correlation between the variables and latent factors form the loadings. A key difference between MCIA and MOFA is that MCIA has no sparsity constraint in the model, so is not required to have many zero or near-zero loadings, thus reducing interpretability. Another key difference is that latent factors in MCIA are typically constrained to orthogonality, which means they are uncorrelated and independent of one another [32, 38]. MCIA has proven success in biomarker prediction and disease subtyping, being shown to discriminate four previously described subtypes of high-grade serous ovarian cancer and uncover robust subtype biomarkers. This investigation additionally demonstrated that integration using MCIA improved knowledge of pathways in leukaemia over analysis of gene expression alone [38]. MCIA has also been applied outside of bioinformatics. Afshari et al showed MCIA was able to detect and characterise relationships between microbiome and metabolome that were explanatory of quality attributes of cheese. Though this sounds trivial, the signatures found could be used to monitor cheese quality and product authenticity, which highlights the potential versatility of method applications [39].

In 2021, Cantini et al benchmarked nine multi-omics dimensionality reduction techniques representative of the most prevalent underlying mathematical frameworks. Three complementary benchmarks were used to evaluate methods: 1) sample clustering on simulated multi-omics data, 2) association of latent factors with survival, clinical annotations and biological annotations using real cancer datasets, and 3) ability to integrate single-cell datasets. The first benchmark found that the two methods designed for clustering, integrative non-negative matrix factorisation (intNMF) and iCluster, expectedly performed the best at clustering simulated datasets. Of the seven remaining methods, MCIA and MOFA were the top performing across simulated scenarios, where a k-means consensus clustering was applied to the factor matrix. The second benchmark used three Cancer Genome Atlas (TCGA) omics datasets for ten different cancer types. MCIA performed very well in this benchmark, finding factors significantly associated with survival in 70% of cancers, scored amongst the top 3 methods for associating with clinical annotations in 50% of cancer and performed well in finding associations with MsigDB hallmarks annotation and gene ontology (GO) annotations in 80% of cancers. MOFA also performed well, finding factors predictive of survival in 50% of cancers, also scored amongst the top 3 methods for associating with clinical annotations in 50% of cancer, and found metagene associations with biological annotations in 50% of cancers. Interestingly, this analysis showed the number of factors associated with survival was more dependent on cancer type than method. In addition, during evaluation of associations with biological processes and pathways, methods ranked variably depending on the biological annotation database being used. All methods in this study, except MOFA, were designed to be applied to bulk multi-omics datasets, therefore the authors believed investigating integration of single-cell data would be a valuable benchmark. All nine methods performed extremely well in this benchmark, with MCIA ranking third and MOFA sixth. Methods were also compared to single-cell analysis methods, which revealed all methods to perform equally well or better than Seurat or LIGER. Overall, this study showed both MCIA and MOFA performed well across all three interdependent benchmarks, particularly when assessing factor-level information i.e. survival and clinical annotation associations. It also highlighted MCIA to be a slightly more versatile method, performing better across multiple applications and data types [29].

A similar comparison to Cantini et al was performed by Pierre-Jean et al in 2020, but with a panel of 13 unsupervised multi-omics data integration methods, including MCIA and MOFA. This investigation had a more exhaustive list of evaluation metrics, covering computation time, subgroup clustering performance and feature importance evaluation (i.e. how much the model uses that feature to make accurate predictions) for both real and simulated data. Contrastingly to previous studies, the authors found it difficult to calibrate MOFA parameters and were unable to get model convergence on any simulation benchmarks. As a result MOFA was discarded from the study and not included in the results. This study found MCIA to have low computation time and comparatively poor clustering performance, where the method appeared to have consistent stability issues with varying subgroup number and composition. For investigation of method ability to select important features driving clusters, three likelihood models were simulated. MCIA performed very well with gaussian and beta-like datasets but failed to recover relevant variables on binary data. Therefore, MCIAs feature selection performance appears to depend on the heterogeneity of the data. Lastly, the authors summarised the user-friendliness of methods, where MCIA was deemed one of the most user friendly [40].

In summary, both methods of interest have varying underlying statistical frameworks, but overall perform very similarly (see Table 1.4). Literature review found there are limited benchmarking studies including both MOFA and MCIA to be able to directly compare methods. From the two direct comparisons available, MCIA appears to be the best allrounder and is easily implemented. While



		Cantini, L., et al. 2021					Pierre-Jean, M., et al. 2020					
Method	Author Applications	Simulated data benchmark	Cancer benchmark: Survival	Cancer benchmark: clinical annotations	Cancer benchmark: Biological annotations	Single-cell benchmark	Variable detection simulation benchmark			Clustering benchmark	Computing time	User-friendly
							Gaussian	Binary	Beta-like			
MOFA	• Identification of clinical markers in CLL • Prediction of clinical outcome in CLL • Identification of factors driving cell-cell heterogeneity (Argelaguet, R., et al. 2018) • Integration of time-course single-cell data • Identification of context-dependent signatures associated with cellular diversity (Argelaguet, R., et al. 2020)	+++	++	++	++	+++	NA	NA	NA	NA	NA	NA
MCIA	• Describe biological properties of nine different cancer tissues • Identification of ovarian cancer subtypes and subtype biomarkers (Meng, C., et al. 2014)	+++	++	++	+++	+++	+++	-	++	+	++	+++

Table 1.4. Summary of MOFA and MCIA applications and performance within previous studies.

Good performances are denoted by +++ and bad performances by -. NA means not available. Table adapted from Cantini, L., et al. 2021 and Pierre-Jean, M., et al. 2020 [29, 40]

CLL = chronic lymphocytic leukaemia

MOFA performed well in factor-based analysis in the Cantini et al study, it had to be excluded in in the Pierre-Jean et al study due to inability to converge and the authors noted difficulty tuning parameters [29, 40]. In literature, both methods are strongly documented for application in predictive biomarker discovery and improving understanding of disease, but only MCIA was found to have successful usage in disease subtyping [38]. Despite MOFA performing poorer in benchmarking studies to date, its statistical framework is the most interpretable and has attractive built-in analysis capabilities, which MCIA is comparatively lacking [33].

1.3. Predicting Anticancer Drug Response

Discovery of predictive biomarkers for drug response is gaining a lot of traction in research and computational tool development. Such biomarkers are driving the field of precision oncology, which aims to understand molecular mechanisms of response to ultimately be able to account for patient genotype when making treatment decision [41, 42]. Additionally, this opens up opportunity for recommendation on early-phase clinical trial design and the repurposing of existing drugs for different cancers [43]. However, there are currently very few established biomarkers for anticancer drugs and use of genomic status of drug target as a therapeutic indicator is not always effective for molecular targeted therapies [44]. Machine learning methods have been used to build drug response prediction models from multi-omics data. Each omics layer brings different value to anticancer drug response prediction as a result of molecular alterations having varying impacts on drug response. For example, metabolic re-wiring has been shown to influence drug response to chemotherapy in several cancers, for instance lung and ovarian cancer in response to Cisplatin treatment. Therefore, metabolomics can inform on alterations in cellular metabolism that are essential to sustain tumour cell growth and proliferation [45]. Meanwhile, mass spectrometry-based proteomic and phosphoproteomic profiling has shown capability to improve drug sensitivity predictions through yielding proteome-wide cancer cell signalling activity [46]. In addition, an investigation of gene expression, DNA methylation, somatic mutations and copy number variations in 11,289 tumours across 29 tissues was able to be mapped to 1,001 human cancer cell lines and correlated with response to 265 anticancer compounds. This investigation went on to explore the relative importance of data types in predicting drug response, where gene expression was found to have the best predictive power and 85% of multi-input predictive models performed better than the best single-predictor model [43]. Therefore, by combining these data in multi-omics analyses more stable and reliable predictions can be made compared to analysing datasets in isolation. Large volumes of cell model data are available in public databases, offering the ability to test multiple drugs and combinations in parallel [42]. Inevitably, as greater volumes of multi-dimensional data become available, so will the demand for new, more sophisticated bioinformatics tools [47].

Typically, computational approaches to drug response prediction have three key steps. Firstly, datasets are obtained from public data resources, such as the Cancer Cell Line Encyclopaedia (CCLE) or the Genomics of Drug Sensitivity in Cancer (GDSC) project. Data is then normalised, and features are selected to filter out noisy or irrelevant data. It is possible for feature selection/dimension reduction to be embedded into the model training [47]. This is a crucial step, as a high feature to sample ratio can lead to model overfitting, whereby the model will perform well on the training data but have poor generalisability to the evaluation data [42]. Secondly, the model is trained to create a mathematical representation of the relationship between features and drug response, and the final step comprises evaluation of the selected model on the new data [47]. As inferred from the steps above, supervised learning techniques are most widely used for building models for drug response prediction. Neighbourhood component analysis [48], deep neural networks [49-51] and random forests [52] are all examples of supervised learning techniques used in literature. It is possible for unsupervised learning techniques to provide the basis for generation of predictive models. For example, Cai et al [53] showed multiple multi-omics data integration techniques are able to accurately predict drug response assisted by random forest.

1.4. Research Aims

This work seeks to execute an in-depth comparison of multi-omics latent factor methods in the context of cancer data analysis. MOFA and MCIA have proven their ability to find predictive biomarkers and molecular mechanisms from complex omics datasets in a variety of different contexts [33-35, 38, 39]. Cantini et al showed MCIA to perform the most consistently and effectively across three complementary benchmarks, highlighting strong potential for application in research with diverse and open biological questions [29]. While MOFA performed less consistently, it possesses a more stable and interpretable framework for uncovering insights into drivers of variation [33]. Thus, these methods show great promise to find predictive biomarkers of response to anticancer drugs. Here, a neutral stance is taken to benchmark approaches against three objectives, using omics data from baseline cancer cell lines. The first objective aims to quality control multi-omics data processing. This will enable better interpretation of the second objective, comparing and evaluating the statistical frameworks of methods. In the literature, no benchmarking reviews discussed the impact of different statistical approaches on variance decomposition or the implementation of methods, which proved a gap in the field. The final objective evaluates the ability of methods to recover and explain responder and non-responder cell lines to anticancer therapies, without assistance from supervised learning.

Chapter 2. Materials & Methods

2.1. Omics Datasets

Methods were tested on five omics datasets covering, transcriptomics, epigenomics, metabolomics, expression- and phosphoproteomics. These datasets were available for 46 untreated, unstimulated (baseline) cancer cell lines spanning lung, breast and ovarian cancer (refer to Supplementary Table S1 for the list of cell lines). Omics datasets were prepared, processed and quality controlled by GSK, prior to data retrieval for use in this investigation.

Cell cultures were divided into three sub-cultures to produce three biological replicates, which were each aliquoted for omics measurement. The transcriptome was investigated using bulk RNA sequencing (RNA-Seq), epigenome using Assay of Transposase Accessible Chromatin sequencing (ATAC-Seq), while mass spectrometry (MS) was used to analyse metabolomics, expression- and phosphoproteomics. In instances when there was a low cell count, specific omics measurements were prioritised over others. All raw data was processed using internally standardised computational pipelines, followed by QC, normalisation and batch-correction. Datasets underwent variance stabilising normalisation (VSN) using the *vsn* R package [54], other than transcriptomics which was normalised and transformed by calculating the binary logarithm of the transcript count per million (TPM). Next, datasets were batch corrected using the *limma* R package [55], however, for epigenomics data, no batch correction could be performed due to confounding with the cancer type. Lastly, replicate measurements for each cell line were aggregated by taking the mean, except epigenomics where the median was taken. Due to the proprietary nature of this data, it is unable to be shared.

Additional filtering was applied following data retrieval to handle missing values across datasets. This involved removing cell lines missing one or more omics measurements and within each omics dataset, removing features with missing measurements for one or more cell lines. Details on data pre-processing and additional processing steps are outlined below.

2.1.1. Transcriptomics

Bulk RNA-seq reads were mapped to the full human genome, GRCh38.p13 from the National Center for Biotechnology Information (NCBI) [56], using *STAR* [57]. Mapped reads were counted at the gene-level using *featureCounts* [58] and an in-house workflow management tool applet calculated gene TPM values using the *featureCounts* output and total mapped reads from the *samtools flagstat* [59] output.

Prior to data retrieval, genes with zero variance or missing counts were removed and lowly expressed genes were removed, where lowly expressed genes were defined as having a mean log2(TPM+1) less than 1 for all samples.

2.1.2. Epigenomics

ATAC-seq reads were mapped to the full human genome, GRCh38.p13 from the NCBI [56], using *Bowtie2* [60]. Genomic regions of open chromatin, or peaks, were called using *Genrich* [61] in ATAC-seq mode. This tool was configured to use an interval length of 100bp, a minimum AUC for a peak of 20 and a maximum FDR-adjusted p-value of 0.05 (per sample). Peaks were quantified using *featureCounts* [58], which used an internally generated universal chromatin accessibility reference. This reference is a comprehensive feature space of accessible regions across multiple biological contexts. Following sample QC, *bedtools* [62] was used to naïvely merge common peaks across samples where there was a 1bp overlap, resulting in larger, more diverse peak widths. The new BAM files output were then used to re-quantify peaks and peaks with missing values in at least one cell line were removed.

To reduce noise and improve the biological interpretability of epigenomics data, promoter peaks were filtered for inclusion in the analysis. Since it can be assumed that an open promoter corresponds to a higher expression of the gene, peaks were filtered for those in a promoter region. This was done using the *ChIPseeker* package in R [63], where the promoter region was defined as ±200 base pairs from the transcription start site with a flank distance of 200 base pairs. The *TxDb.Hsapiens.UCSC.hg38.knownGene* transcript annotation object [64] and *org.Hs.eg.db* annotation R package [65] were used to define promoters. Peaks in the same promoter region were not combined.

2.1.3. Expression- and Phospho-proteomics

MS based expression- and phoshoproteomics data were processed as described previously [66]. Briefly, proteins were digested and resulting peptides were subjected to tandem mass tags (TMT) enabling relative quantification of up to ten conditions in one MS run. Labelled samples were measured using liquid chromatography-tandem mass spectrometry (LC-MS/MS) on an Orbitrap Fusion Lumos and a Q Exactive (Thermo Fisher Scientific). Mascot 2.5 was used for protein identification using a customised version of the SwissProt protein database (https://www.uniprot.org/) [67], from December 2018. Protein quantification values were calculated from individual spectra matching unique peptides using sum-based bootstrap algorithm. Phosphoproteomics employed an additional phospho-enrichment step before TMT labelling to enrich for phosphorylated peptides. Spectra quantification values were combined to phosphosites using the median yielding quantification values per sample and phosphosite in a specific protein.

2.1.4. Metabolomics

Untargeted metabolomics were performed as described previously [68]. Samples were measured on an LC-MS platform using a Q Exactive. The resulting raw data were processed using an in-house built R pipeline. Detected ions were tentatively matched to metabolites using the Human Metabolome Database (HMDB) (https://hmdb.ca/) [69], where annotation was solely based on accurate mass, therefore isomers or other metabolites within a given tolerance cannot be distinguished. Log_{10}-transformed ion intensities were used as measure of quantification.

2.1.5. Feature Filtering

For each dataset, a feature selection step was performed to remove features (i.e. genes, genomic regions, proteins or metabolites) with low variance from input to methods [36]. Thresholds were set arbitrarily based off the assumption that a higher number of total features in a dataset requires a higher number of selected features to capture a similar minimum variance (see Table 2.1). Therefore, varying numbers of features in datasets will result in different numbers of selected features. Datasets with less than 1,000 features are presumed to capture a small amount of variance in too few features to be included in the model.

Complete (Number of Features)	Selected (Number of Features)
<1,000	0
1,000 – 2,500	1,000
2,500 – 7,500	2,000
7,500 – 15,000	3,000
15,000 – 25,000	4,000
25,000 – 40,000	5,000

Table 2.1. Thresholds for feature filtering

2.2. Drug Response Data

Drug response data was accessed (August 2022) from the Genomics of Drug Sensitivity in Cancer (GDSC) database (www.cancerRxgene.org/) [70], where the GDSC2 dataset (release 8.4) was downloaded for analysis. This dataset contains data on 288 anticancer drugs for a total of 969 cell lines. The authors use a non-linear mixed effect model to fit dose-response curves of all available cell line/drug combinations to obtain area under the dose response curve (AUC) estimates, which are utilised for prediction of drug response (see section 2.5).

2.3. Software and Packages

All analyses were performed using R statistical software (v4.0.2) [71]. See R Session Information in the supplementary information for details of all software packages used during this investigation. Source code available at https://github.com/shannonkatrina/benchmarking-mofa-mcia.

2.4. Data Integration Approaches

Detailed below are the two unsupervised latent factor methods benchmarked in this investigation, MOFA and MCIA. Default parameters were selected for each approach. Although each method can optimise the number of latent factors detected, for the sake of comparison, the same number of factors were imposed on both methods. Datasets were decomposed into ten factors as a starting point, based on recommendations by Argelague et al [36].

2.4.1. Multi-Omics Factor Analysis

MOFA [33, 36], can be viewed as a generalisation of (sparse) PCA applied to multiple omics datasets. Although, technically it is an extension of Bayesian group factor analysis, meaning Bayesian inference and probabilistic models are used to estimate the latent factors and their loadings [72]. The matrix factorization framework in MOFA+ can be described as:

$$Y_{gm} = Z_g W_m + \varepsilon_{gm} \tag{1}$$

Y_{gm} = matrix of observations for the *m*th modality and the gth group
Z_g = factor matrix for the gth group
W_m = weight matrix for the *m*th modality
ε_{gm} = residual noise for the *m*th modality and the gth group

The model is able to efficiently handle missing values and can flexibly combine different likelihood models for different data modalities. The noise term (ε_{gm}) contains unexplained variance for each feature in each modality and varies depending on the types of data input into the method. A combination of different noise models are supported to integrate different data types, such as binary (Bernoulli), discrete (Poisson) and continuous data (Gaussian). Following a Bayesian framework, a prior distribution is assigned to the factor matrix, weight matrix and parameters of the noise term. MOFA applies a two-step symmetric regularisation of the weights and factors to account for structure in both the sample and feature space. It first boosts view- and factor-wise sparsity to enable distinction of active factors in omics datasets using an automatic relevance determination (ARD) prior. A spike-and-slab prior is then applied to induce feature-wise sparsity to highlight small sets of features with active weights.

The core of MOFA+ is implemented in Python package *mofapy2*, while the R package *MOFA2* is recommended for use as an interface for model training and downstream analysis. The code to run MOFA+ is available at https://github.com/bioFAM/MOFA2.

2.4.2. Multiple Co-Inertia Analysis

MCIA [38], is an extension of co-inertia analysis (CIA) [73] that enables analysis of two or more omics datasets. It requires a set of matrices where either samples or features are matched with equal weights and can accommodate both discrete and continuous data. MCIA factorises omics data into latent factors in two steps. Firstly, an ordination technique, such as PCA or correspondence analysis (COA), is applied to each matrix (**M$^{\mathbf{i}}$**) separately, transforming data into new, comparable, lower dimensional datasets (**X$^{\mathbf{i}}$**). In this analysis PCA was used.

$$x_{ij} = \frac{p_{ij}}{r_i} - c_j \tag{2}$$

x_{ij} = relative abundance of element to the measurement's weight
p_{ij} = single element contribution to the total variance in matrix (M^{i})
r_i = relative contribution of row *i* over the total variance in matrix (M^{i})
c_j = relative contribution of column *j* over the total variance in matrix (M^{i})

The second step is derived from CIA, which aims to maximise the sum of squared co-variance, i.e. the co-inertia, between scores of each matrix (**X$^{\mathbf{i}}$**). This maximisation means the norm constraint ($\|u\|=\|v\|=1$) is applied on orthogonal directions (*u* and *v*). For each latent factor the problem is defined as:

$$argmax_{q_1^1 \ldots q_p^1} \sum_{k=1}^{P} cov^2(X_k^i q_k^i, X^i q^i\ ) \tag{3}$$

q^i = global PCA projections

Features and samples with similar trends are closely projected in the latent space. MCIA is implemented in the R package *omicade4* (https://bioconductor.org/packages/release/bioc/html/omicade4.html).

2.5. Prediction Evaluation

Cell lines were categorised into drug response categories using the AUC normalised to zero mean and unit variance across the available data for the 46 cell lines (z-score) for easier interpretation of results. As AUC is dependent on the range of tested drug concentrations, in the event a drug was tested across multiple concentration ranges, these were treated as separate drug entities for z-score calculation [74]. As described previously [75], cell lines with a z-score less than 0.8 standard deviations (SD) away from the mean were defined as a responder whereas cell lines more than 0.8 SDs away from the mean were classified as a non-responder. Cell lines that fell between these thresholds were classified as intermediate. To evaluate the ability of latent factor approaches to predict drug response the coefficient of determination (R^2) was calculated for every latent factor/drug combination within each method (see equation 4). It is inferred that ability to explain the variance driving differences in drug response enables prediction of response using unsupervised learning.

$$R^2 = \ 1 - \frac{SS_{RES}}{SS_{TOT}} \tag{4}$$

R^2 = coefficient of determination
SS_{RES} = sum of squares of residuals
SS_{TOT} = total sum of squares

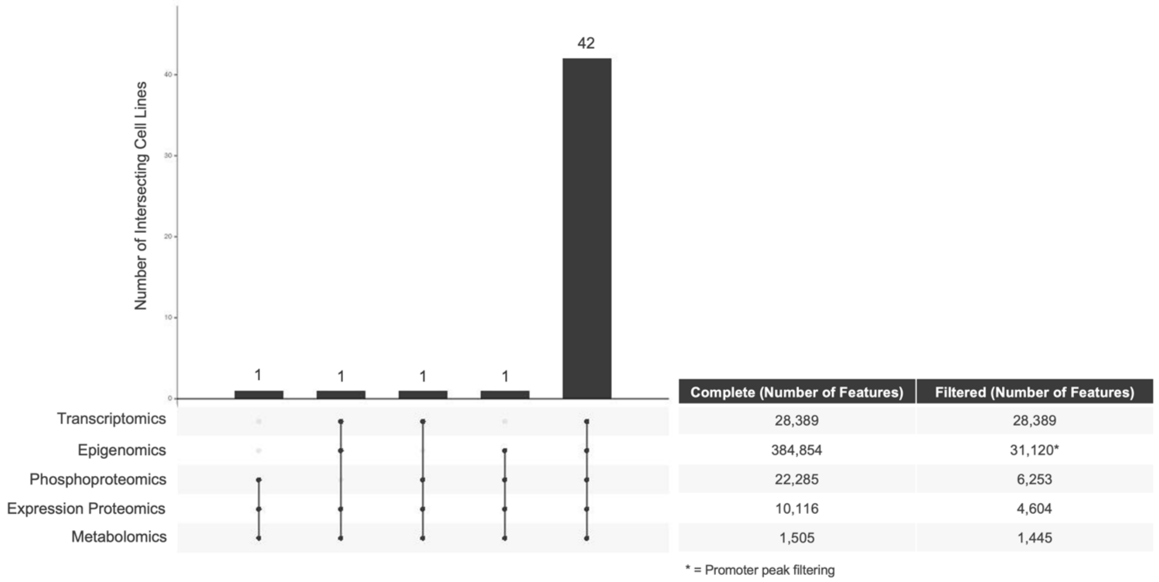
P-values were derived using an F-test from the linear fit used to calculate the R^2. The p-values were adjusted for multiple testing using the method of Benjamini and Hochberg [76].

Chapter 3. Results

In this project, baseline transcriptomics, epigenomics, expression- and phosphoproteomics and metabolomics data from 46 cancer cell lines were used to benchmark MOFA and MCIA, two multi-omics latent factor methods. The aim was to quality control multi-omics data processing followed by evaluation and comparison of statistical frameworks of the methods, particularly the similarities and differences in captured variance. Lastly method ability to predict anticancer drug response using the AUCs from 288 anticancer drugs was evaluated and compared. This chapter summarises the analyses performed to answer the above outlined objectives.

3.1. Data Processing

At the point of data retrieval, data had already been quality controlled, normalised, filtered and, where applicable, batch corrected (refer to section 2.1 for more information). A total of 447,149 features were measured across all assays, of which there were 28,389 genes, 383,854 genomic regions, 22,285 phosphorylated proteins, 10,116 proteins and 1,505 metabolites. This showed epigenomics to be over-represented and metabolomics to be under-represented, though this was expected given ATAC-seq



Complete (Number of Features)	Filtered (Number of Features)
28,389	28,389
384,854	31,120*
22,285	6,253
10,116	4,604
1,505	1,445

* = Promoter peak filtering

Figure 3.1. Overview of multi-omics datasets following processing.

Upset plot depicting the overlap in cell lines assessed by each omics technology. Rows correspond sets of omics data and columns correspond to possible intersections of cell lines. Filled cells represent datasets composing an intersection and the bar chart above shows the number of cell lines in an intersection. The table on the right shows the number of features (i.e. genes, genomic regions, proteins or metabolites) in each omics dataset before (complete) and after (filtered) additional processing. processing.

identifies multiple genomic regions per gene and there are vastly less metabolites relative to other biomolecules. Additional data processing was required prior to input into the models (see Figure 3.1). This consisted of removing features with missing values, which affected 16,032 phosphorylated proteins, 5,512 proteins and 60 metabolites, whereas genes and genomics regions were already filtered during pre-processing. Secondly, epigenomics data was filtered to retain only promoter regions, which removed 353,734 genomic regions (see section 2.1.2 for more details). Lastly, only cell lines with measurements from all omics layers were kept resulting in a total of 42 out of 46 remaining for data integration. The resulting data were five complete data matrices with no requirement for imputation.

Metabolomics, expression proteomics and phosphoproteomics were successfully normalised but some systematic mean-SD bias remained in transcriptomics and epigenomics

Following processing, the quality of data normalisation was assessed by plotting SD as a function of rank mean for each feature in their respective datasets (see Figure 3.2). Well normalised data should

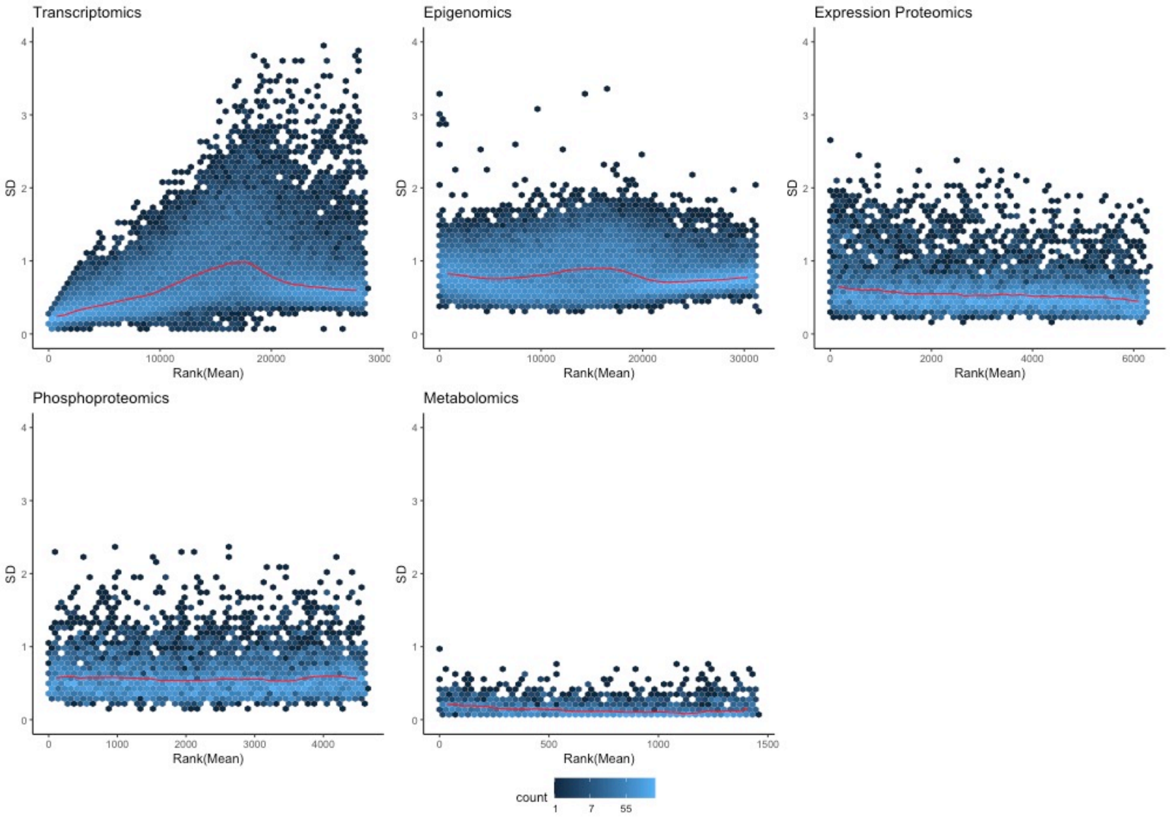


Figure 3.2. Assessing systematic bias within omics datasets using standard deviation as a function of rank mean. Each point represents a feature, and the colouring scale depicts the count (abundance). The x-axis is rank of the mean, therefore high abundance features are on the left and low abundance features on the right. The red line represents the running median estimator of variance. If there is no variance-mean dependence, the line should be approximately horizontal.

contain no systematic mean-SD bias; the variance should not be dependent on feature abundance. This can be seen for expression proteomics, phosphoproteomics and metabolomics data, exhibited by the flat running median estimator of variance. Contrastingly, the median estimator of variance deviated from straight for transcriptomics and mildly for epigenomics. This showed some systematic bias remained in these data, which could introduce unwanted variance into data integration models. Ideally, data normalisation would have been redone, however, raw counts data was unavailable. Thus, transcriptomics and epigenomics datasets were unaltered and retained in the analysis, considering small, estimated impacts on model learning and downstream analysis.

Selected highly variable features had discordant minimum variance across omics datasets

When running unsupervised data integration methods it is strongly recommended to filter highly variable features (HVFs) to remove uninformative features and reduce imbalances between modalities [36]. Arbitrary thresholds were used to determine the number of selected features based off the total number of features in the dataset (outlined in section 2.1.2). As modalities varied in size, the number of selected features varied (transcriptomics = 5,000; epigenomics = 5,000; expression proteomics = 2,000; phosphoproteomics = 2,000; metabolomics = 1,000) and the resulting minimum variance of feature subsets also differed (transcriptomics = 1.11; epigenomics = 1.19; expression proteomics = 0.49; phosphoproteomics = 0.37; metabolomics = 0.01) (see Figure 3.3). Details of the impact removing features and cell lines with missing data had on HVFs selected for each assay can be found in the supplementary information (Supplementary Figure S1). The vastly lower feature variance in the metabolomics data poses challenges of being under-represented during data integration. Despite this, data was retained to see how well methods cope with disparity in variance between modalities.

3.2. Method Evaluation and Comparison

After checking normalisation and selecting the HVFs for each of the five datasets, the data were ready to be stacked into a multi-assay experiment for input into data integration methods. The explained variance across modalities was investigated prior to evaluating the methods. Both MOFA and MCIA were run with the same input data, respective default model parameters and learned ten latent factors. The cumulative proportion of variance explained across the ten factors by each assay, and proportion of total variance explained by individual factors for each assay was then explored (see Figure 3.4). The total variance explained by all factors gives a good indication of the model fit to the data. It is desirable for each assay to have a cumulative variance explained greater than 10% and, for the purposes of identifying functions across modalities and data integration, learnt factors to have a proportion of total variance explained by two assays or more. Otherwise, it is likely the factor is capturing technical variance [36].

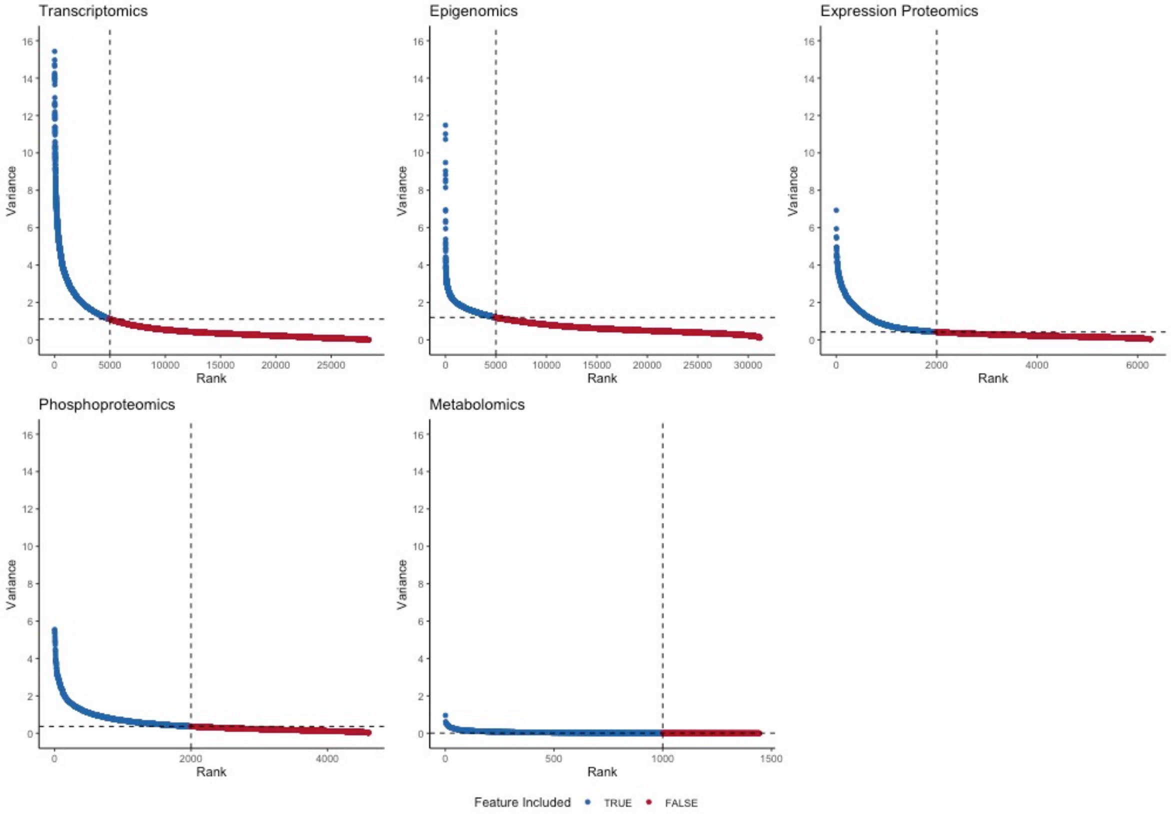


Figure 3.3. Filtering of highly variable features across each of the five omics datasets.

Features were ranked in descending order based on their variance estimate for the 42 cell lines. The vertical dashed line represents the feature rank threshold used for filtering highly variable features and the horizontal dashed line represents the minimum variance of selected features. The colour corresponds to whether a feature is included as input for latent factor models (blue) or removed (red).

Metabolomics data introduced noise to latent factor models

Overall the variance decomposition looked quite different between the two methods. MCIA had a higher cumulative variance explained for all assays compared to MOFA, meaning MCIAs latent representation explained more cell line heterogeneity per assay (see Figure 3.4A). Cumulative variance explained varied more between modalities for MOFA than MCIA (R^2 range: MOFA = 16.3% to 50.2% ; MCIA = 33.0% to 48.4%). Metabolomics had the smallest representation in the latent space for both methods, which could mean the dataset was introducing noise to the model, particularly in MOFA due to only explaining 16.3% of the variance. To investigate this further, the proportion of total variance explained by individual factors was explored (see Figure 3.4B). All MCIA factors explained a proportion of variance for all assays, whereas each MOFA factor explained variance for a fraction of assays i.e. factors explained near zero variance for some assays. For MOFA, metabolomics was the only assay to be solely explained by a single factor (factor 4, R^2 = 15.5%), while all other assays shared factors. This gave a strong indication that the data was introducing technical variance and would likely require removal. For MCIA the opposite occurred, where metabolomics had a consistently low variance explained by all factors, ranging from 2.7% to 4.5%. Altogether this inferred that metabolomics was not integrating well with other datasets and was contributing little biologically relevant variance in either method. As a result metabolomics data was removed for subsequent analyses evaluating methods.

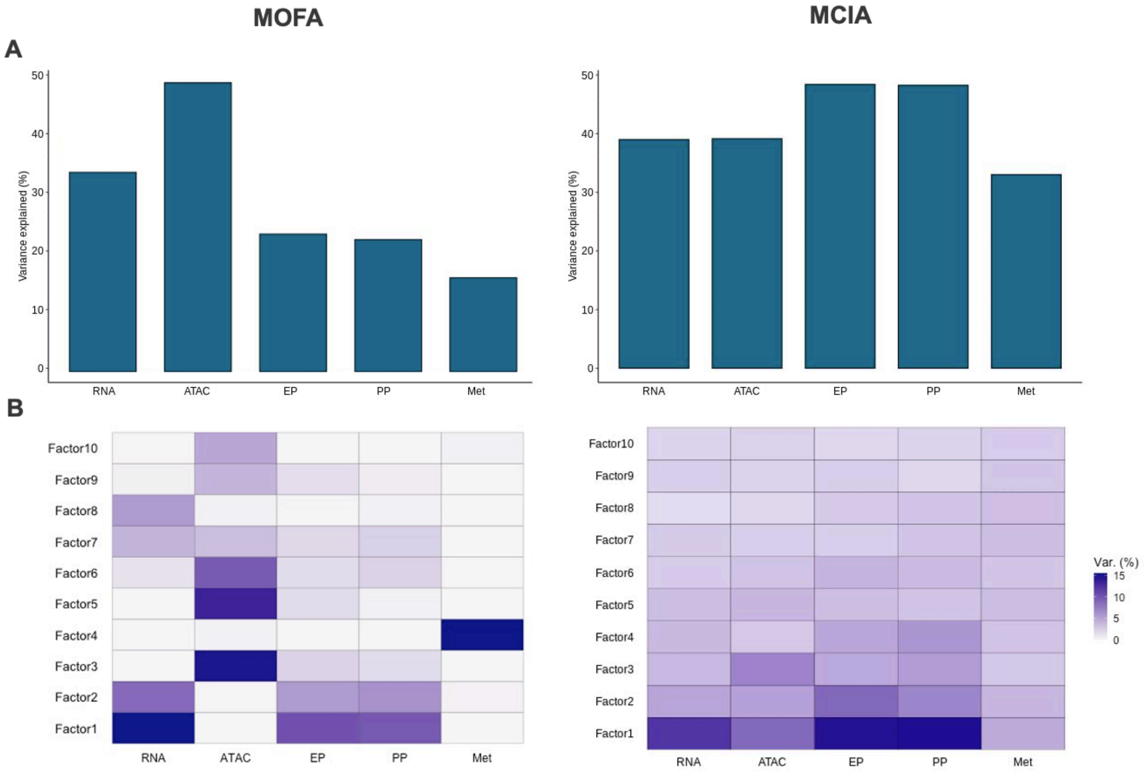


Figure 3.4. Evaluation and comparison of total variance explained by MOFA (left) and MCIA (right) models. **A** Cumulative proportion of total variance explained by each assay. **B** Proportion of total variance explained by individual factors for each assay. Colour density represents the proportion of the variance explained.

RNA = Transcriptomics; ATAC = Epigenomics; EP = Expression proteomics; PP = Phosphoproteomics; Met = Metabolomics

Variance decomposition by view and by factor presented differently in each method

After metabolomics was dropped, both methods were re-run with respective default model parameters and learned ten latent factors from the remaining four omics datasets. During this analysis the correlation of factors was explored (Figure 3.5A). This is another recommended assessment for MOFA model fitting, where many correlations between factors indicates poor model fit [36]. MOFA had a moderate correlation between factor 1 and factor 4 (r = 0.43), but otherwise all Pearson correlation coefficients were below 0.21. When numerous correlations are found, it is advised to reduce the number of factors or check data normalisation to reduce correlations between learnt factors. As normalisation

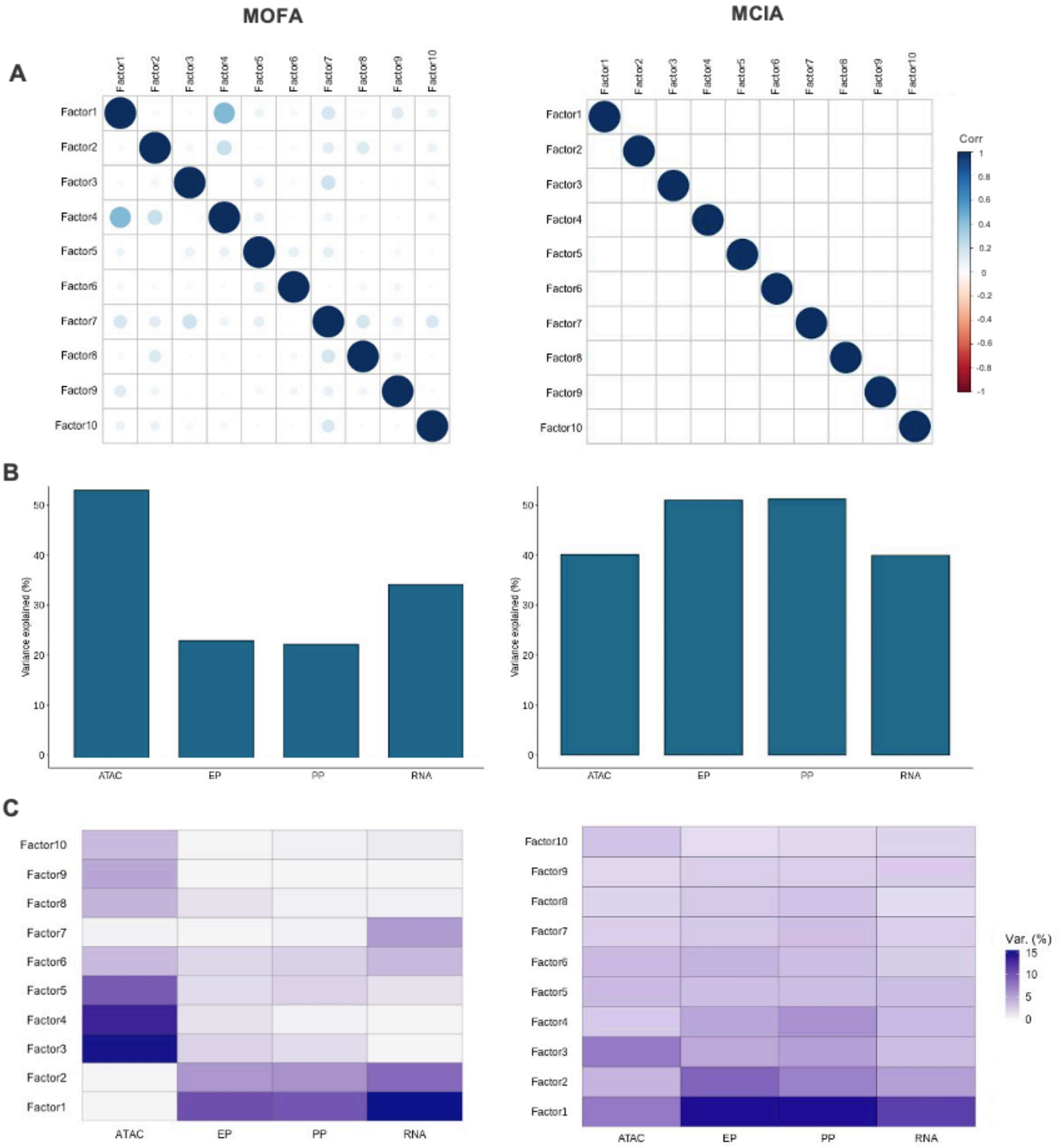


Figure 3.5. Evaluation and comparison of latent factor correlations and variance explained by MOFA (left) and MCIA (right) models using refined inputs.

A Plot of correlation matrix for latent factors. The areas of circles show the absolute value of corresponding Pearson correlation coefficients. Colour depicts the direction (positive correlations are displayed in blue and negative correlations in red) and size (low correlations have low colour intensity and high correlations have high colour intensity) of correlation coefficients. **B** Cumulative proportion of total variance explained by each assay. **C** Proportion of total variance explained by individual factors for each assay. Colour density represents the proportion of the variance explained.

RNA = Transcriptomics; ATAC = Epigenomics; EP = Expression proteomics; PP = Phosphoproteomics had already been considered, a reduction in factors was investigated but did not impact the correlation between factors 1 and 4 (see Supplementary Figure S2). Given the size and number of correlations, the MOFA model fit was suboptimal but adequate to continue. As expected, MCIA factors showed no correlation.

Similar to the previous analysis, the cumulative proportion of variance explained across the ten factors by each assay, and proportion of total variance explained by individual factors for each assay was explored (Figure 3.5B and Figure 3.5C, respectively). The removal of metabolomics data affected the total variance explained per assay in both methods by only a small amount, where the largest effects were seen in epigenomics for MOFA (+3.5%) and phosphoproteomics and expression proteomics for MCIA (+3.1% and +2.7%, respectively). Thus, this provided further evidence that metabolomics had not integrated well, and it made reasonable to remove the dataset from analysis for model evaluation and comparison. Each method emphasised variance of different assays, shown by different cumulative explained variance. Epigenomics had the greatest variance explained across MOFAs ten learnt factors (53.7%), followed by transcriptomics (34.9%). Whereas phosphoproteomics and expression proteomics jointly had the greatest cumulative variance explained in MCIA (51.3% and 51.1%, respectively). Interestingly, for both MCIA and MOFA, expression proteomics and phosphoproteomics had almost identical cumulative variance explained within each method, differing by 0.2% in MCIA and 0.7% in MOFA, but between methods total variance differed by approximately 30%. Inspecting more closely at a factor level, similarly to in the previous analysis, all MCIA factors explained a proportion of variance for all assays, but all MOFA factors, except factor 6, explained almost zero variance for at least one assay. The average range of variance explained by factors was 8.0% for MOFA and 2.3% for MCIA, emphasising that the proportion of variance explained by each factor in MCIA was less divergent across assays compared to MOFA. The first factor of MOFA captured a strong source of variability present across transcriptomics (15.5%), phosphoproteomics (9.9%) and expression proteomics (10.2%). Correspondingly, factor 1 of MCIA captured a strong signal from all data modalities (transcriptomics = 11.4% ; epigenomics = 7.7% ; expression proteomics = 14.3% ; phosphoproteomics = 14.2%). Factors that capture strong variance across multiple modalities, such as these, are likely to be important sources of variability in the data.

Latent factor combinations were unable to cluster by cancer type in either method

The relationship between factors was investigated by visualising pairwise combinations of factors and evaluating is clusters align with cancer type. Factors 1 and 2 are shown as an example (see Figure 3.6), but the full set of pairwise combinations can be found in the supplementary information (see Supplementary Figure 3). All factor combinations for both methods poorly formed clusters, indicating neither method was able to identify groupings of similar cell lines. As a result, both methods were unable to cluster cell lines by cancer type.

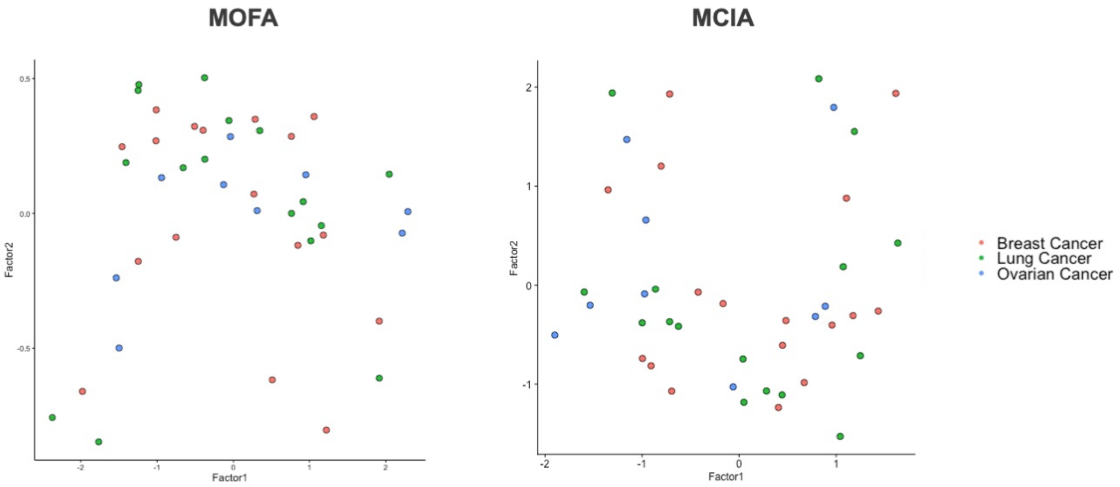


Figure 3.6. Visualisation of cancer cell lines using MOFA (left) and MCIA (right) latent factors 1 and 2. Each point represents a cancer cell line and colour denotes the cancer type.

Strongly correlated latent factors could be identified between the two methods

To investigate the relationship between the two models further, pairwise Pearson correlation coefficients between the latent factors from the two models were calculated and plotted (see Figure 3.6). In general, there was little to no correlation between each methods factors, with some exceptions. Strong negative correlations were found between factor 1 (r = -0.89) and factor 2 (r = -0.78) with their relative factor. Whereas, strong positive correlations were found between MOFA factors 5 and 6 and MCIA factors 3 and 5, respectively (r = 0.79; r = 0.87). It is assumed that strong absolute correlation (r > 0.8) between method factors showed both methods were able to find factors explanatory of the same variance.

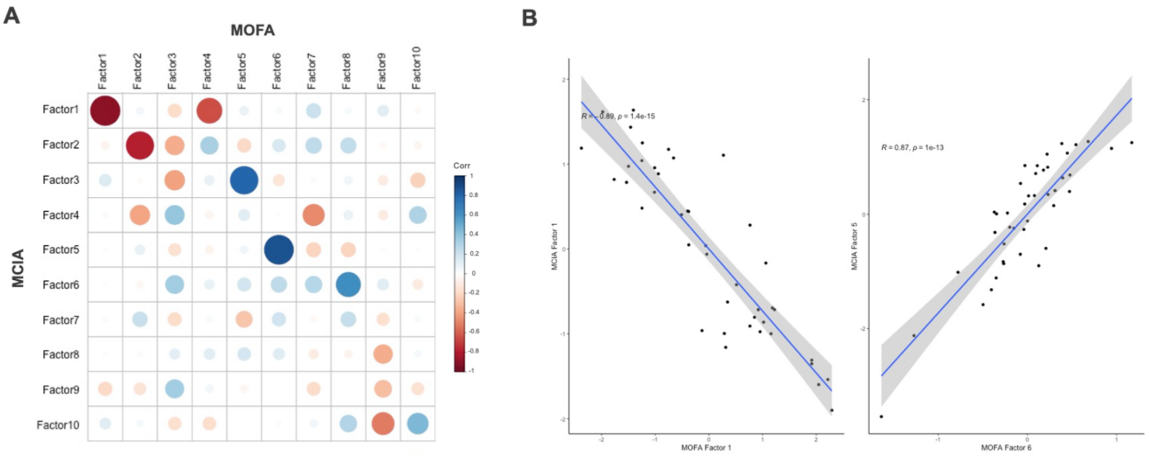


Figure 3.7. Investigation of correlation between latent factors of MOFA and MCIA.

A correlation plot for MOFA (top) and MCIA (left) latent factors. The size of the dots indicates the absolute value of the corresponding Pearson correlation coefficients. Colour depicts the direction (positive correlations are displayed in blue and negative correlations in red) and size (low correlations have low colour intensity and high correlations have high colour intensity) of correlation coefficients. **B** Relationship between cell line factor values for highly correlated latent factors shown in A. MOFA factor values are on the x-axis and MCIA factor values on the y-axis. Left shows the relationship between MOFA factor 1 and MCIA factor 1 and on the right shows the relationship between MOFA factor 6 and MCIA factor 5.

3.3. Evaluation of Drug Response Prediction

Model factors were unable to confidently predict anticancer drug response

Next, the ability of both methods to predict anticancer drug response in the 42 cell lines was assessed. Using trained model outputs, the relationship between latent factors and drug response was explored by calculating the proportion of variance for drug response that is explained by factor values (R^2). Drug response values range between zero and one, where one shows a low responsiveness to a drug and zero shows high responsiveness to a drug. The GDSC2 dataset contained data for 288 anticancer drugs across 38 of the 42 cell lines, though the data were incomplete as some drugs had data on less than 38 cell lines. To ensure there were sufficient data for drug response prediction, only drugs with data for at least 20 cell lines were considered for the analysis. In addition, some drugs appeared multiple times in the dataset with different dose response concentrations. In these instances the drugs were treated separately, which resulted in a total of 285 compounds being analysed. This resulted in a total of 5,700 comparisons made.

MOFA and MCIA performed very similarly, as neither of the model factors were able to explain more than approximately 30% variance in drug response across cell lines and the majority of drugs had less than approximately 10% variance explained (see Figure 3.8). In this instance, outliers are a positive result as they indicate drugs whose response could be better predicted using the latent factors than the majority of the drugs. MOFA was able to mildly predict response ($R^2 > 0.3$) of two drugs, Redmodelin

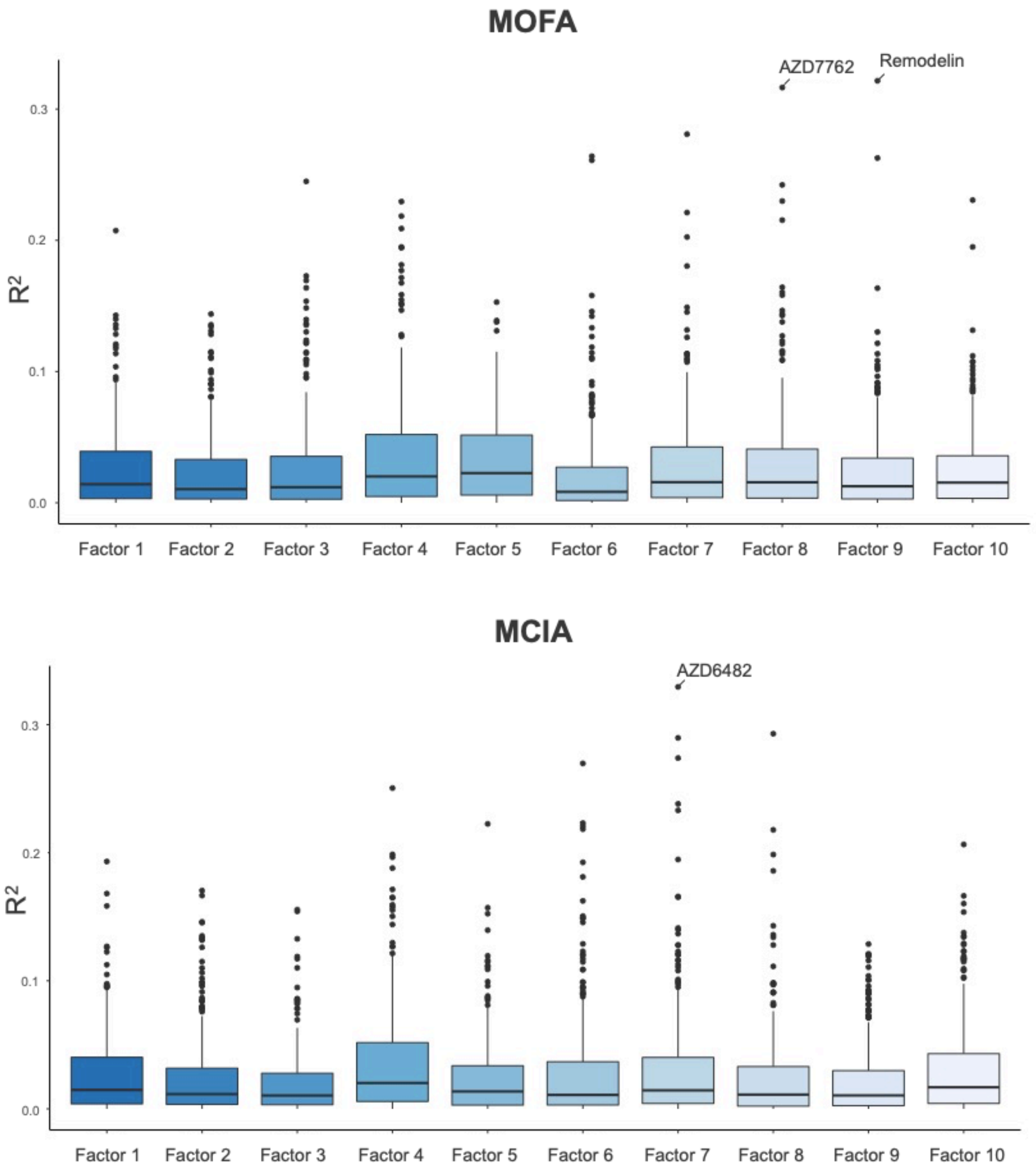
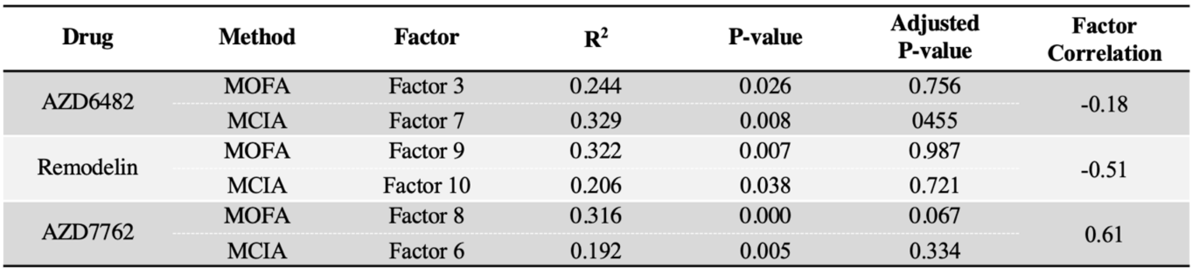


Figure 3.8. Evaluation of MOFA (top) and MCIA (bottom) latent factor ability to explain variance in anticancer drug response.

Boxplots depict the spread of explained variance (R^2) between factor value and drug response. Boxes represent the interquartile range (IQR) and whiskers represent the minimum (Q1-1.5*IQR) and maximum (Q3+1.5*IQR). Dots represent outliers that fall beyond the minimum or maximum. Drugs with an R^2 greater than 0.3 are labelled.

($R^2 = 0.32$) and AZD7762 ($R^2 = 0.32$), while MCIA was only able to predict one, AZD6482 ($R^2 = 0.33$). These three drugs were inspected more closely for significance using p-values derived from the linear regression model and compared to the best prediction (highest R^2) in the respective other method (see Table 3.1). Between methods there was a difference of approximately 10% of variance explained for each drug and interestingly, moderate correlations were found between best performing factors for Remodelin (r = -0.51) and AZD7762 (r = 0.61). However, no factors predictions were significant following Benjamini & Hochberg correction [76]. Although not significant, MOFA factor 8 was characterised in relation to AZD7762 drug response for demonstrative purposes (see Figure 3.8). AZD7762 was chosen over Remodelin and AZD6482 due having the lowest adjusted p-value whilst having a similar R^2 value. From this point on, factor 8 will refer to MOFA factor 8 and drug response will refer to AZD7762 drug response.



Drug	Method	Factor	R^2	P-value	Adjusted P-value	Factor Correlation
AZD6482	MOFA	Factor 3	0.244	0.026	0.756	-0.18
	MCIA	Factor 7	0.329	0.008	0455	
Remodelin	MOFA	Factor 9	0.322	0.007	0.987	-0.51
	MCIA	Factor 10	0.206	0.038	0.721	
AZD7762	MOFA	Factor 8	0.316	0.000	0.067	0.61
	MCIA	Factor 6	0.192	0.005	0.334	

Table 3.1. Summary of best performing drug response predictions by model factors.

P-values were derived using an F-test from the linear fit used to calculate the R^2 and adjusted for multiple comparisons within methods and factors using the Benjamini & Hochberg correction [76].

Epigenomic signatures of MOFA factor 8 were unable to explain AZD7762 drug response

Cell lines were categorised into responder and non-responder groups for easier interpretation of results (see section 2.5 for more information). Out of the 38 cell lines there was data available for, eight were classified as responder ($\bar{x}$ - 0.8SD) and ten as non-responder ($\bar{x}$ + 0.8SD). The remaining 20 cell lines were classified as intermediate, meaning drug response did not deviate far enough from the mean to be classified in either group (see Figure 3.9A). Factor 8 and drug response had a moderate positive correlation (r = 0.56), however factor value was unable to separate the three drug response categories (see Figure 3.9B). The top ten weighted epigenomic features in factor 8 were then explored (see Figure 3.9C). Data showed that factor 8 explained the most variance in the epigenomics dataset (3.8%) and so this assay had the greatest likelihood of finding biologically relevant signal to drug response within this factor. Weights indicate how much a single feature contributes to the latent factor, enabling biological interpretation of factors. Weights range from zero to one, where zero indicates no association to a factor and one shows high association. The direction of effect is represented by the sign of the weight, positive

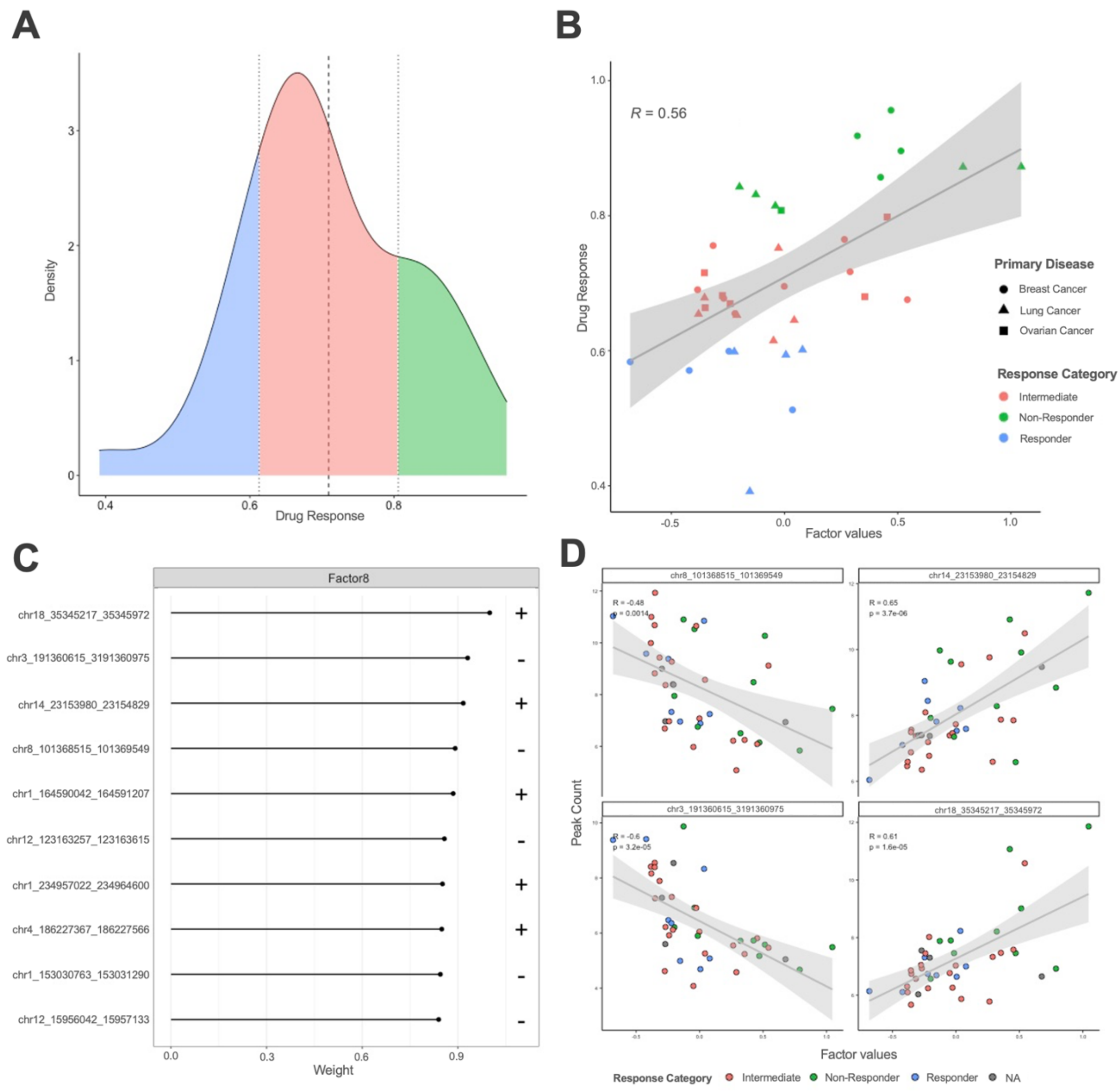


Figure 3.9. Epigenomic characterisation of factor 8 in relation to AZD7762 response using MOFA.

A Density of AZD7762 drug response across 38 cell lines. Drug response values range between 0 and 1, where 1 shows a low responsiveness to a drug and 0 shows high responsiveness to a drug. The vertical dashed line represents the mean drug response for AZD7762 and dotted lines represent ±0.8SD from the mean. Coloured areas correspond responders (blue), non-responders (green) and intermediate (red). **B** Relationship between factor 8 values (x-axis) and AZD7762 drug response (y-axis). Colour represents drug response category and shape represents primary disease. **C** Absolute weight of top 10 features of MOFA factor 8 in the epigenomics data. The corresponding weight sign is depicted on the right, where positive signs depict higher levels of abundance in cell lines, and vice-versa. **D** Relationship between MOFA factor 8 values (x-axis) and peak count (y-axis) for the top 4 peaks with largest absolute weight. Cell lines are coloured by response category.

indicates the feature has higher abundance in cell lines with positive factor values, and negative indicates the feature has lower abundance in cell lines with positive factor values. The top ten features had weights over 0.8, showing high association to factor 8. Lastly, the relationship of the top four features with factor values and response categories of cell lines was investigated (see Figure 3.9D). All features had a moderate correlation with factor values (abs(r) = 0.48 to 0.65), however, showed no relationship with drug response category, indicated by lack of separation of categories.

Chapter 4. Discussion

In this study, where transcriptomics, epigenomics, metabolomics, expression- and phosphoproteomics datasets from 46 baseline cancer cell lines were used to benchmark latent factor methods, MOFA and MCIA, three objectives were completed. The first objective comprised quality control of multi-omics data processing prior to model input, such as normalisation and highly variable feature filtering. The second aimed to compare the low dimensional representations of the five omics datasets produced by MOFA and MCIA, in addition to the computational implementation of methods. Lastly, the third objective was dedicated to evaluating the ability of model factors to predict anticancer drug response, and if applicable, whether molecular signatures associated with drug response could be found.

Several insights were gained from this study, as follows. Prior to model fitting, selecting the number of highly variable features based on total number of features is not sufficient to capture a consistent minimum variance across datasets. Following model training, MOFA and MCIA derive different amounts of signal from each dataset and so produce contrasting latent representations of the same data. Despite these differences, both found the metabolomics dataset to be noisy, though portrayed in different ways, and were able to find latent factors that explained the same variance. When investigating the variance explained by factors, neither of the model factors were able to confidently predict cell line response to any anticancer drug and no molecular signatures of response could be found in the single example explored.

4.1. Model Inputs

Below, limitations of data processing impacting results are reviewed. As in most multi-omics datasets, the five omics datasets retrieved for analysis had both unmatched features and unmatched samples. To enable a fair comparison of both MCIA and MOFA, only cell lines with measurements across all omics were kept and features with any missing measurements across samples were removed. By filtering the datasets to complete matrices, this may have limited the ability to see strengths of MOFA over MCIA, in terms of integration and imputation capability. Unlike MCIA, MOFA can handle unmatched, incomplete datasets as input i.e. samples do not need to be matched across datasets and missing measurements are tolerated within datasets. Therefore, a limitation of this study was that there was unnecessary data loss to the MOFA model by removing cell lines and features with missing measurements, which impacted the subset of HVFs selected for all omics datasets (Supplementary Figure S1). As a result, it is plausible that different sources of biological variation across modalities could have been discovered and differentiated performance of MOFA from MCIA, for better or worse.

If time permitted, a comparison of the impact of removing missing data on the latent space and drug response predictions would have been very interesting. It should be noted that when missing data is present, MOFA simply ignores the missing values from the likelihood estimations and there is no hidden imputation step. Rather, as part of the downstream analysis values can be imputed for biological interpretation, which has been shown to be more accurate than other established imputation strategies [33]. To avoid data loss in the first instance, methods are available that handle missing rows in multiple omics datasets through exploiting the correlation structure across datasets [77]. However, these methods risk reducing the variability in the low-dimensional representation and breach modality independence assumptions required by underlying statistical approaches of many methods. Thus, although imputation can be performed as part of data pre-processing, it is best avoided given the unknown impact on downstream analyses [25].

During data processing it was noted that the transcriptomics data had a deviated running median estimator and so there was some systematic bias present (Figure 3.2). It is probable that this is the result of normalisation using the TPM method opposed to the variance stabilising transformation (VST) method. Argelaguet et al [36] state that appropriate normalisation during data processing is critical for MOFA and recommend data should be normalised according to the likelihood model used. For example, counts-based data should undergo size factor normalisation followed by VST to fulfil a Gaussian distribution [36]. Meng et al do not make any explicit recommendations on data normalisation for MCIA, but note from their analysis that the variance in transcriptomics data was sensitive to pre-processing [38]. It is likely that any bias remaining after normalisation would be captured strongly in an early factor and downweight other sources of variation. Interestingly, transcriptomics contributes the most variance explained in MOFA factor 1, whereas in MCIA factor 1, transcriptomics only captures a small amount of signal relative to expression- and phosphoproteomics. This finding could suggest that MOFA is more sensitive to data normalisation, which would be reasonable given the noise term that relies on likelihood models. This hypothesis could be tested by investigating the relationship between raw RNA counts and gene rank in top loadings for factor 1 in each model, where a high correlation would suggest inadequate normalisation for the given model.

It is advised to subset HVFs during data processing to decrease large imbalances in size between modalities, simply interpretation and speed up model training [36]. In this analysis, arbitrary thresholds were chosen to select HVFs. Thresholds for transcriptomics and epigenomics were set at 5,000 features, in line with implementation by Argelaguet et al [33], and altered accordingly to the reduced size of other omics datasets. This selection method successfully reduced the imbalance of size between modalities, however, the minimum variance of feature subsets differed (Figure 3.3). This potentially means that modalities with higher minimum variance (transcriptomics and epigenomics) are over-represented in factors, causing smaller sources of biological variation from other modalities (metabolomics, expression- and phosphoproteomics) to be missed. If this were the case, the total variance explained by each assay would be expected to follow the same pattern as the minimum variance of feature sets, i.e. low minimum variance results in low total variance explained and vice-versa. However, this does not occur for either method. This means that either the minimum variance does not impact the total variance explained of modalities, or other sources of technical variance are having a greater effect. Arguably, more sophisticated methods for feature selection could have been used, based on the variance of each dataset. A heuristic approach would be to use the elbow method, a technique often used to determine the optimal number of clusters in a dataset during clustering analysis. The method consists of plotting sum squared error (SSE) as a function of the number of clusters and using the "elbow" on the curve to decide number of clusters to use [78]. This can be translated to this analysis by using the "elbow" on the feature variance/rank curves (Figure 3.3) to determine the optimal number of features, which would produce omics-specific thresholds and likely reduce the range of minimum variance between selected feature sets. However, this method could be problematic due to the high number of features causing difficulty in unambiguously identify at which feature rank the "elbow" falls. An alternative method could be defining a minimum variance threshold to apply across all omics and determine the number of kept features for each dataset by looking at which rank the threshold line crosses the variance/rank curve (Figure 3.3). Although these methods have stronger influence over the minimum feature variance across modalities, there is no control over the number of features selected. Therefore, this could result in extremely small datasets or even removal of whole datasets, which would be the case for the metabolomics data in this investigation. It is plausible that having too few features in a dataset, despite having high variance, could impact representation in factors and ability to extract functional insights. Altogether, this raises the question of where the balance should be struck between quality and quantity of features for optimal data integration and downstream analysis. If time permitted, it would be interesting to look at the impact of different feature selection methods on model outputs.

4.2. Implementation

MCIA and MOFA are implemented and available in the R/Bioconductor packages *omicade4* and *MOFA2*, respectively. MCIA was first released in 2014, being built on top of the *ade4* R package. Whereas MOFA was first released more recently in 2018 and has since been re-released under the name MOFA+ (R package *MOFA2*) in 2020. In terms of documentation, MOFA benefits from having a GitHub Pages website, covering installation, troubleshooting, tutorials and much more. In addition, there is a MOFA community Slack group for quick, personalised help. Contrary to this, MCIA has limited documentation and user support. Running both models was very simple in this investigation as default parameters were used and the number of factors were pre-defined. MOFA has a multitude of parameters for defining data, model and training options, whereas MCIA has very few. Therefore, MOFA is better equipped to be fine-tuned for a given analysis, but greater complexity and risk of overfitting are a consequence. Though, MOFA has been shown to lack overfitting in large-sample settings, however stability reduces with smaller sample size [79]. For downstream interpretation, MCIA has limited built-in capabilities compared to MOFA. The package produces a 4-panel figure that summarises the sample space, feature space and factor values, but can only view the relationship between two latent factors at a time. Thus, custom libraries need to be built to produce more customisable outputs. MCIA imposes no sparsity on results, so interpretation requires additional methods, such as enrichment analysis, to be able to reveal functional insights. In comparison, MOFA has functions available for data extraction, plotting and enrichment analysis, plus the sparsity constraint means that results are more interpretable. To note, the additional capabilities of MOFA such as multi-group analysis, spatio-temporal relationships and GPU acceleration were not applicable for investigation during this project.

4.3. Performance of Data Integration

At present, latent factor method benchmarking studies have evaluated methods by investigating clustering, outcome prediction, classification into sub-groups and ability to find relevant variables in data types [29, 40]. However, performance of data integration has not been benchmarked and is rarely included in published multi-omics studies. It is important to assess the variance decomposition of factors to understand how well factors are capturing variance across modalities, and so, how well the data is integrating and whether technical variance is being introduced. In the first integration of all five omics datasets, metabolomics data was considered to integrate poorly and introduce noise to the model (Figure 3.4). This conclusion was primarily driven by the presentation of the metabolomics data in MOFA, where the dataset was explained by a single, unshared factor with high variance explained. Whereas in MCIA metabolomics had a consistently low variance explained across all factors. Argelaguet et al noted that MOFA factors capture variance of multiple modalities, which can help to mitigate assay noise. Therefore it can be inferred that a factor driven by a single modality will be noisy [33]. In addition, the tutorial for MOFA analysis of CLL cohort data highlights noisy datasets with strong non-linearities will have low total variance explained, which the is exhibited by metabolomics data [80]. Alternatively, it is possible that metabolomics doesn't introduce noise and rather biologically doesn't co-vary with other omics datasets. However, this is unlikely given multiple studies have successfully integrated metabolomics with other abundance-based modalities in multi-omics analyses [81-83]. It is important to account for differing numbers and weights of features, which can introduce bias towards the modalities with a large number of variable features. Without proper scaling issues can arise with datasets such as metabolomics, due to inherently having a reduced number of features relative to other omics such as transcriptomics and proteomics [84]. Processing of datasets showed the HVFs selected for metabolomics had vastly lower variance than selected features from other modalities, making feature selection and scaling the most plausible reasons why the data failed to integrate. Moreover, metabolomics differs to other omics technologies as cell culture media can heavily influence cell metabolism, particularly when cells are in a steady state [85]. Therefore, metabolite measurements may have been more representative of differences in growth media opposed to biologically meaningful differences. This could have been another contributing factor to the poor integration. The differing presentation of suspected noisy data between methods is likely due to the sparsity constraint imposed in MOFA, which reduces non-zero loadings in latent factors. This in turn causes MOFA factors to strongly associate with a small number of modalities [32]. It is possible that the sparse model considered the majority of metabolomic loadings to be unimportant due to having comparatively low variance to other omics and so were pruned away. This would result in a small number of features representative of the modality, which are less likely to explain overall variance or co-vary with other modalities, thus resulting in a single latent factor strongly associated with metabolomics. Whereas in MCIA, loadings were not pruned so the model was able to find a higher total variance explained across all modalities. Variance in the metabolomics data was able to be related to variance in other datasets and the imposed orthogonality of factors meant that a small amount of variance was distributed across all latent factors.

Given the above observations, data integration was rerun with the metabolomics data removed, which formed the basis for evaluation and comparison of the low dimensional representations produced by the methods. Removal of metabolomics appeared to have no impact on the variance decomposition of other modalities when compared to the first set of results, which is further evidence that this dataset was likely an outlier. Furthermore, it could be argued that this shows both MOFA and MCIA are relatively stable in the presence of noisy data, however, this may only be the case as metabolomics had such low variance. The first comparison made between the two methods was investigating the intra-model factor correlations. Expectedly, MCIA factors had no correlation to one-another, while some MOFA factors were mildly correlated due to no orthogonality constraints (Figure 3.5A). As it is possible for MOFA factors to be correlated, it is important to check factor correlations as part of model QC. Correlation between two factors indicates that they are partially representative of the same variability in the data, and so the model is not optimally fitted. This can be a result of poor normalisation causing factors to share systematic bias or due to too many factors being learnt causing overlap in captured variability [33]. In this investigation, factors 1 and 4 of MOFA were correlated and a reduction in number of factors did not reduce factor correlations (Supplementary Figure S2). Therefore, this leaves inadequate normalisation to be the most likely cause. Looking closely at these factors' variance decomposition, factor 1 is most active in the transcriptomics, while factor 4 is most active in the epigenomics. Both transcriptomics and epigenomics displayed a degree of systematic bias, so poor normalisation is a plausible explanation for this correlation. It would have been preferable to redo the data normalisation during data processing to attempt to correct this, however the raw counts data was not available for either dataset.

Further differences can be seen in the cumulative proportion of total variance explained by each assay between methods. MOFA found the most signal in the epigenomics data, followed by transcriptomics and lastly expression- and phosphoproteomics. Meanwhile, MCIA found the most variation equally in the expression- and phosphoproteomics data, followed by epigenomics and transcriptomics (Figure 3.5B). It was previously discussed that MOFA appears sensitive to technical variance. The emphasis of variation in the epigenomics and transcriptomics data could be another representation of this. Both datasets had technical variance remaining after normalisation, in addition, no batch correction was performed on the epigenomics dataset, which likely explains why it has the highest captured variance. Epigenomics is most strongly explained in factor 3 and factor 4, which notably capture little variance in other modalities. After investigating factor correlations it was hypothesised that factor 1 and factor 4 are representative of systematic bias in the transcriptomics and epigenomics, respectively. Therefore, it is possible that factor 3 is explanatory of technical batch effects present in the epigenomics. Interestingly, MOFA factor 3 was found to weakly correlate with all MCIA factors, so it possible that the noise from the batches has been spread across all factors at a low level of variance. This is a similar to how the metabolomics data presented in the first data integration, so would have been interesting to investigate in more depth to assess whether this dataset should have been removed from analysis. As batch and cancer type are confounded, the three cancer types could be analysed separately to see whether there is an increase in shared variance of epigenomics with other modalities in factors, which would suggest a reduction in technical variation. Alternatively, to look more specifically at whether MOFA factor 3 was representative of batch effect, the association between top epigenomics feature loadings of factor 3 and batch/cancer type could be investigated. However, this may not be possible with this data given latent factors in neither model were able to cluster by cancer type (Supplementary Figure S3). There is increasing evidence that cancers are comprised of many subpopulations of cells, termed cancer heterogeneity, which means even cell lines of the same cancer type are biologically divergent [86]. This is a possible reason for failure of cell types to cluster in this analysis, however there is no clustering of cell lines in any pairwise factor plots. Heterogeneity exists across multiple omics layer, thus data integration offers potential to decipher subpopulations of cancer cells or highlight similarities in subpopulations of different cancer types [87]. Meng et al demonstrated capability of MCIA to cluster nine different cancer types using transcriptomic and proteomic datasets [38], however Argelaguet et al did not investigate this capability [33, 36]. This raises minor concern with this analysis and again highlights the possibility of noise or technical variation being introduced into the models' preventing factors from explaining biologically relevant variation. Figure 3.5B also exhibited that MCIA captures over double the amount of variance present in expression- and phosphoproteomics compared to MOFA, despite selected HVFs having lower minimum variance than transcriptomics and epigenomics. Phosphoproteomics constitutes changes in both protein abundance and phosphorylation, therefore it can be described as dependent on expression proteomics [88]. This relationship means that expression- and phosphoproteomics violate MCIAs assumption of independence and are highly likely to explain the same variance. This is a possible explanation for why MCIA appears to upweight the variance explained in these two modalities. To test this theory, it would have been interesting to assess if correcting for protein abundance in the phosphoproteomics data lowered expression- and phosphoproteomics total variance explained in MCIA and if this de-coupled the two datasets in the variance decomposition of both methods. Overall, these findings suggest that MOFA is more sensitive to noise and the benefits of sparsity are reliant on proper removal of technical variance. Whereas MCIA appears more sensitive to fulfilment on underlying statistical assumptions. However, these are only preliminary hypotheses that require further investigation to fully understand the impact on downstream analyses.

Although MOFA and MCIA appear very different, some factors in MOFA and MCIA were highly correlated (Figure 3.7). This indicates that despite statistical differences, both methods are able to capture very similar variance. It is important to note that the factor values produced during dimension reduction are note directly interpretable, rather they should be interpreted analogously to principal components, where the relative positioning is most important. Therefore, when interpretating factor-factor correlations, it is the absolute correlation that should be focused on and not the direction. High absolute correlations were found between factors 1 and 2 to their respective factor in the other model, showing they are representative of extremely similar variance. Finding correlated factors increases the likelihood of finding biologically relevant data, however this is heavily reliant on data processing removing all sources of technical variance.

4.4. Performance of Drug Response Prediction

The final objective of this investigation was to evaluate the ability of MOFA and MCIA to predict anticancer drug response in cancer cell lines. AUC was used as the measure of drug response as it combines information on efficacy and potency and can be calculated for any dose-response curve, therefore there are never missing values. Also it has been shown to be robust when making comparisons across cell lines [89]. A disadvantage of this metric is that AUC is dependent on the range of drug concentrations tested, which can vary between experiments and studies. In this investigation, this resulted in some drugs having multiple predictions. R^2 values were calculated for all drug and factor combinations, which were used to assess the performance of prediction (Figure 3.8). Neither model factors were able confidently predict response for any drug, as only a total of three drugs had an R^2

greater than 0.3, which was considered a threshold for weak predictions. Comparison of method performance was difficult as there were no strong predictions ($R^2 > 0.8$). It would have been preferable to compare the total number of strong predictions, assess the overlap of drugs predicted and how well highly predictive factors of drugs correlate from each method. However, as this was not possible, the difference in R^2 and correlation between MOFA and MCIAs best predictive factor of each of the three drugs was investigated (Table 3.1). This showed the respective methods to differ by approximately 10% of variance explained, but two out of the three of the predictive factor pairs were moderately correlated. This showed that despite varying ability to predict, both methods extracted similar biological variance to explain differences in drug response. However, this small subset of examples not sufficient to make a robust comparison between methods. It is possible that benchmarking drug response prediction using an alternative method might have been more successful. Due to latent factor methods being unsupervised, common performance evaluation metrics such as precision, recall, F1-measure and balanced accuracy could not be used [90]. Cox proportional-hazards regression model is a possible alternative method to test the association between factors and drug response. This technique was successfully applied in multi-omics analysis by Cantini et al to benchmark survival prediction, which found both MOFA and MCIA could find predictive latent factors. Although, none of the nine models investigated were unable to find predictive factors for lung or ovarian cancer [29]. This is evidence that MOFA and MCIA are capable of prediction, but the cancers being investigated might be more difficult to predict, possibly due to being particularly heterogeneous [91].

In this analysis model factors were only explored in isolation for drug response prediction, opposed to additionally investigating how combinations of factors can explain data. Looking at pairs of latent factors in multi-omics data is similar to looking at PCA plots, where samples will cluster based on their similarity [92]. These visualisations are useful to uncover major axes of heterogeneity and align them to biological variables. As the axes represent variation across multiple modalities, multiple sets of loadings can be investigated to build a more holistic and reliable view of drivers of biological variation [93]. For example, Meng et al utilised factors 1 and 2 generated by MCIA to distinguish four subtypes of ovarian cancer using transcriptional data from multiple platforms. This enabled consensus genes commons across platforms to be identified and associated to ovarian subtypes, forming robust biomarkers [38]. Argelaguet et al also looked at the relationship between factors during MOFA analysis of CLL data, which found factors 1 and 2 distinguished samples based on somatic mutation status of the immunoglobulin heavy-chain variable region gene and chromosome 12 trisomy status. Investigation of factor loadings of transcriptomics data found gene associations consistent with literature, reinforcing these as important clinical markers in CLL [33]. Although this type of investigation would be very valuable for evaluation of drug response prediction, it would be impractical to execute due to the number of drugs included in this study. As this type of analysis is unsupervised it requires manual inspection for agreement of clusters with biological variables, which in this study would result in a multitude of plots to evaluate. Cantini et al and Pierre-Jean et al evaluated the clustering of methods using Jaccard index and adjusted rand index on simulated and real datasets [29, 40]. Therefore, a way to partially circumvent the requirement to evaluate all individual plots could be to use a quantitative measure such as Jaccard index, adjusted rand index or silhouette score [94]. These coefficients provide a measure for how well data is clustered, and so could be applied to subset high scoring factor combinations for manual evaluation of whether clusters align with response categories for any anticancer compounds.

For demonstrative purposes, the top epigenomics loadings of MOFA factor 8 was investigated in relation to AZD7762 response (Figure 3.9). Unsurprisingly, given the small R^2, this investigation was unable to find separation of drug response categories in any of the top four feature loadings. This demonstrated that none of these four peaks alone, would be able to explain variance in drug response, despite being highly associated with factor 8. Should a factor-drug combination have had a high R^2, it would have been interesting to do gene set enrichment analysis and pathway enrichment analysis on the genes associated with the most weighted epigenetic features to improve understanding of molecular signatures and pathways associated with drug response [95, 96].

These findings altogether show that although some variance in cell line drug response can be explained by methods, neither MOFA nor MCIA are able to predict drug response using baseline omics data. However, it is difficult to tell if the models truly can't predict drug response or whether the data input was not good enough to be able to extract biologically relevant variance. Literature shows success of a variety of algorithms, such as deep learning [97], random forests [98], naïve bayes classifier [52] and manifold learning [99], to predict drug response using publicly available omics and drug response data. Although manifold learning is the only example of an unsupervised learning technique, this points towards the data being the reason for poor drug response predictions. Common data types present in these studies are mutation and/or copy number variation, which aren't present in this analysis. One element contributing to low predictive power could be a lack of genomic profile data, such as these. Additionally, these studies use a range of 3 to 4 modalities. Thus it is possible that reducing the data input to models, for example only including transcriptomic and proteomic data, could improve ability to predict. In future work it would be interesting to investigate combinations of different modalities and how this alters predictive power in each model. Another impacting factor could be the small number of cell lines included in this study. A pharmacogenomic study showed that downsampling subsets of cell lines rapidly reduced the number statistically significant associations between cancer functional events and drug sensitivity [43]. This highlights the benefit of a large collection of cell lines to increase statistical power.

This analysis has presented that there was likely to be remaining systematic bias and technical variance in the data following data processing, as well as probable low statistical power due to only having complete data on 42 cell lines. Furthermore, feature selection prior to model training resulted in unfair feature subsets with differing variance. Consequently, these factors together plausibly interfered with both models' ability to find smaller sources of more relevant variation to derive functional insights. If time permitted, it would have been interesting to see the impact of the following on drug response prediction performance: 1) revised normalisation for transcriptomics and epigenomics data (pending raw counts data retrieval), 2) protein abundance correction of phosphoproteomics data and 3) revised feature selection using a minimum variance threshold. Improvements in experimental design could have also been made, such as analysing each cancer type separately. This would not only have removed noise introduced by epigenomic batches, but also potentially improved predictions. This is evidenced by Cantini et al finding that the number of latent factors associated with survival was driven more by cancer type than data integration method during their benchmarking study [29]. Additionally, this would have doubled up as a good use case to test the multi-group function of MOFA. The drug response data is another variable that might have impacted the ability of methods to predict response. This analysis used data from the public database GDSC. Instantly, a potential issue using public data in combination with in-house data is that it isn't guaranteed cell lines used for drug response analysis come from the same source as those used for omics analysis. Different sources can be considered as different strains of the same cell line, where vast differences in gene expression, morphology and, importantly, drug response can be observed [100]. Therefore, it is possible that the drug response values used during this analysis might differ to the true drug response of cells analysed, thus impacting the accuracy of predictions. Secondly, using a single drug response dataset may reduce the robustness of predictions. There are multiple public datasets available that use different viability assays for response profiling and contain varying cell lines and drugs, therefore ability to predict may vary depending on the dataset used. Examples of other datasets available include the National Cancer Institute 60 (NCI60), Cancer Cell Line Encyclopaedia (CCLE) and Cancer Therapeutics Response Portal (CTRP) [101]. Multiple data harmonisation approaches exist that could be implemented on public datasets to provide a more accurate and robust measure of drug response for use in this analysis [74, 102-104].

Chapter 5. Conclusion

This research aimed to give an in-depth comparison of MOFA and MCIA, divided into three objectives. The first objective comprised investigating quality control of multi-omics data processing, for example data normalisation to remove systematic bias and feature selection to reduce imbalances between modalities. This complemented the second objective of evaluating and comparing the similarities and differences in captured variance by factors and assays. This research highlighted adequate data processing to be very important prior to training models, though the extent of impact appeared to depend on the statistical framework of methods. Differing underlying assumptions meant each method was more sensitive to different parts of data processing, and so decomposed variance very differently. For instance, MOFA was found to be more sensitive to bias and noise introduced through poor normalisation and batch correction, which related to the sparsity constraint of the method. This showed the benefits of sparsity require proper data processing. On the other hand, MCIA was found to be more sensitive to the relationship between modalities, due to the assumption of independence between datasets when maximising covariance. Considering this, it was difficult to detect the impact of inconsistent minimum variance between modalities. Nonetheless, additional investigation into alternative feature selection methods could bring value to understanding the impact on data integration once data processing concerns have been rectified. To summarise, this showed the importance of good data processing and that underlying statistics govern how variance is captured in the latent representation of data.

The final objective aimed to assess and compare the ability of trained models to predict anticancer drug response, unassisted by supervised learning. This analysis found both methods derived latent factors that were able to explain variance in drug response. The distribution of variance explained by factors was very similar between the two methods, showing both performed equally. However, neither model sufficiently explained variance at a level to be able to predict response. Given the results from the first objectives, further research is required to ascertain if this result occurred due to lack of data and/or inadequate data. Improvements in data normalisation, batch correction, feature selection and experimental design should be able to give a better determination if the models truly can’t predict drug response.

References

1. Lee, Y.T., Y.J. Tan, and C.E. Oon, *Molecular targeted therapy: Treating cancer with specificity.* European Journal of Pharmacology, 2018. **834**: p. 188-196.
2. Finotello, F., et al., *Editorial: Multi-omic Data Integration in Oncology.* Front Oncol, 2020. **10**: p. 1768.
3. Menyhárt, O. and B. Győrffy, *Multi-omics approaches in cancer research with applications in tumor subtyping, prognosis, and diagnosis.* Comput Struct Biotechnol J, 2021. **19**: p. 949-960.
4. Conesa, A. and S. Beck, *Making multi-omics data accessible to researchers.* Scientific Data, 2019. **6**(1): p. 251.
5. Dai, X. and L. Shen, *Advances and Trends in Omics Technology Development.* Front Med (Lausanne), 2022. **9**: p. 911861.
6. Lu, M. and X. Zhan, *The crucial role of multiomic approach in cancer research and clinically relevant outcomes.* Epma j, 2018. **9**(1): p. 77-102.
7. Olivier, M., et al., *The Need for Multi-Omics Biomarker Signatures in Precision Medicine.* Int J Mol Sci, 2019. **20**(19).
8. Wang, Z., M. Gerstein, and M. Snyder, *RNA-Seq: a revolutionary tool for transcriptomics.* Nat Rev Genet, 2009. **10**(1): p. 57-63.
9. Grandi, F.C., et al., *Chromatin accessibility profiling by ATAC-seq.* Nat Protoc, 2022. **17**(6): p. 1518-1552.
10. Griffiths, W.J. and Y. Wang, *Mass spectrometry: from proteomics to metabolomics and lipidomics.* Chem Soc Rev, 2009. **38**(7): p. 1882-96.
11. Gallo Cantafio, M.E., et al., *From Single Level Analysis to Multi-Omics Integrative Approaches: A Powerful Strategy towards the Precision Oncology.* High Throughput, 2018. **7**(4).
12. Hasin, Y., M. Seldin, and A. Lusis, *Multi-omics approaches to disease.* Genome Biology, 2017. **18**(1): p. 83.
13. Sun, Y.V. and Y.J. Hu, *Integrative Analysis of Multi-omics Data for Discovery and Functional Studies of Complex Human Diseases.* Adv Genet, 2016. **93**: p. 147-90.
14. Heo, Y.J., et al., *Integrative Multi-Omics Approaches in Cancer Research: From Biological Networks to Clinical Subtypes.* Mol Cells, 2021. **44**(7): p. 433-443.
15. Subramanian, I., et al., *Multi-omics Data Integration, Interpretation, and Its Application.* Bioinform Biol Insights, 2020. **14**: p. 1177932219899051.
16. Paananen, J. and V. Fortino, *An omics perspective on drug target discovery platforms.* Briefings in Bioinformatics, 2020. **21**(6): p. 1937-1953.
17. Pushpakom, S., et al., *Drug repurposing: progress, challenges and recommendations.* Nature Reviews Drug Discovery, 2019. **18**(1): p. 41-58.
18. Hutchinson, L. and R. Kirk, *High drug attrition rates—where are we going wrong?* Nature Reviews Clinical Oncology, 2011. **8**(4): p. 189-190.
19. Yugi, K., et al., *Trans-Omics: How To Reconstruct Biochemical Networks Across Multiple 'Omic' Layers.* Trends Biotechnol, 2016. **34**(4): p. 276-290.
20. Yu, X.-T. and T. Zeng, *Integrative Analysis of Omics Big Data*, in *Computational Systems Biology: Methods and Protocols*, T. Huang, Editor. 2018, Springer New York: New York, NY. p. 109-135.
21. Vahabi, N. and G. Michailidis, *Unsupervised Multi-Omics Data Integration Methods: A Comprehensive Review.* Front Genet, 2022. **13**: p. 854752.
22. Huang, S., K. Chaudhary, and L.X. Garmire, *More Is Better: Recent Progress in Multi-Omics Data Integration Methods.* Front Genet, 2017. **8**: p. 84.
23. Richardson, S., G.C. Tseng, and W. Sun, *Statistical Methods in Integrative Genomics.* Annu Rev Stat Appl, 2016. **3**: p. 181-209.
24. Weber, L.M., et al., *Essential guidelines for computational method benchmarking.* Genome Biology, 2019. **20**(1): p. 125.

25. Tarazona, S., A. Arzalluz-Luque, and A. Conesa, *Undisclosed, unmet and neglected challenges in multi-omics studies.* Nature Computational Science, 2021. **1**(6): p. 395-402.
26. Palsson, B. and K. Zengler, *The challenges of integrating multi-omic data sets.* Nature Chemical Biology, 2010. **6**(11): p. 787-789.
27. Bartholomew, D.J., M. Knott, and I. Moustaki, *Latent variable models and factor analysis: A unified approach*. 2011: John Wiley & Sons.
28. Meng, C., et al., *Dimension reduction techniques for the integrative analysis of multi-omics data.* Briefings in Bioinformatics, 2016. **17**(4): p. 628-641.
29. Cantini, L., et al., *Benchmarking joint multi-omics dimensionality reduction approaches for the study of cancer.* Nature Communications, 2021. **12**(1): p. 124.
30. Wang, B., et al., *Similarity network fusion for aggregating data types on a genomic scale.* Nature Methods, 2014. **11**(3): p. 333-337.
31. Yu, L. and H. Liu. *Feature selection for high-dimensional data: A fast correlation-based filter solution*. in *Proceedings of the 20th international conference on machine learning (ICML-03)*. 2003.
32. Hastie, T., R. Tibshirani, and M. Wainwright, *Statistical learning with sparsity.* Monographs on statistics and applied probability, 2015. **143**: p. 143.
33. Argelaguet, R., et al., *Multi-Omics Factor Analysis-a framework for unsupervised integration of multi-omics data sets.* Mol Syst Biol, 2018. **14**(6): p. e8124.
34. Haak, B.W., et al., *Integrative Transkingdom Analysis of the Gut Microbiome in Antibiotic Perturbation and Critical Illness.* mSystems, 2021. **6**(2).
35. Titz, B., et al., *Multi-omics systems toxicology study of mouse lung assessing the effects of aerosols from two heat-not-burn tobacco products and cigarette smoke.* Comput Struct Biotechnol J, 2020. **18**: p. 1056-1073.
36. Argelaguet, R., et al., *MOFA+: a statistical framework for comprehensive integration of multi-modal single-cell data.* Genome Biology, 2020. **21**(1): p. 111.
37. Velten, B., et al., *Identifying temporal and spatial patterns of variation from multimodal data using MEFISTO.* Nature Methods, 2022. **19**(2): p. 179-186.
38. Meng, C., et al., *A multivariate approach to the integration of multi-omics datasets.* BMC Bioinformatics, 2014. **15**(1): p. 162.
39. Afshari, R., et al., *New insights into cheddar cheese microbiota-metabolome relationships revealed by integrative analysis of multi-omics data.* Scientific Reports, 2020. **10**(1): p. 3164.
40. Pierre-Jean, M., et al., *Clustering and variable selection evaluation of 13 unsupervised methods for multi-omics data integration.* Briefings in Bioinformatics, 2020. **21**(6): p. 2011-2030.
41. Garraway, L.A., J. Verweij, and K.V. Ballman, *Precision oncology: an overview.* J Clin Oncol, 2013. **31**(15): p. 1803-5.
42. Adam, G., et al., *Machine learning approaches to drug response prediction: challenges and recent progress.* npj Precision Oncology, 2020. **4**(1): p. 19.
43. Iorio, F., et al., *A Landscape of Pharmacogenomic Interactions in Cancer.* Cell, 2016. **166**(3): p. 740-754.
44. Roy, R., et al., *Expression Levels of Therapeutic Targets as Indicators of Sensitivity to Targeted Therapeutics.* Mol Cancer Ther, 2019. **18**(12): p. 2480-2489.
45. Zaal, E.A. and C.R. Berkers, *The Influence of Metabolism on Drug Response in Cancer.* Front Oncol, 2018. **8**: p. 500.
46. Ali, M., et al., *Global proteomics profiling improves drug sensitivity prediction: results from a multi-omics, pan-cancer modeling approach.* Bioinformatics, 2018. **34**(8): p. 1353-1362.
47. Azuaje, F., *Computational models for predicting drug responses in cancer research.* Brief Bioinform, 2017. **18**(5): p. 820-829.
48. Malik, V., Y. Kalakoti, and D. Sundar, *Deep learning assisted multi-omics integration for survival and drug-response prediction in breast cancer.* BMC Genomics, 2021. **22**(1): p. 214.
49. Li, M., et al., *DeepDSC: A Deep Learning Method to Predict Drug Sensitivity of Cancer Cell Lines.* IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2021. **18**(2): p. 575-582.
50. Sakellaropoulos, T., et al., *A Deep Learning Framework for Predicting Response to Therapy in Cancer.* Cell Rep, 2019. **29**(11): p. 3367-3373.e4.

51. Zhang, H., Y. Chen, and F. Li, *Predicting Anticancer Drug Response With Deep Learning Constrained by Signaling Pathways.* Frontiers in Bioinformatics, 2021. **1**.
52. De Niz, C., et al., *Algorithms for Drug Sensitivity Prediction.* Algorithms, 2016. **9**(4).
53. Cai, Z., et al., *Machine learning for multi-omics data integration in cancer.* iScience, 2022. **25**(2): p. 103798.
54. Huber, W., et al., *Variance Stabilization Applied to Microarray Data Calibration and to the Quantification of Differential Expression*. 2002: Bioinformatics. p. S96-S104.
55. Ritchie, M., et al., *limma powers differential expression analyses for {RNA}-sequencing and microarray studies*. 2015: Nucleic Acids Research. p. e47.
56. NCBI, *Genome Reference Consortium Human Build 38 patch release 13 (GRCh38.p13)*. 2019.
57. Dobin, A., et al., *STAR: ultrafast universal RNA-seq aligner.* Bioinformatics, 2013. **29**(1): p. 15-21.
58. Liao, Y., G.K. Smyth, and W. Shi, *featureCounts: an efficient general purpose program for assigning sequence reads to genomic features.* Bioinformatics, 2014. **30**(7): p. 923-930.
59. Li, H., et al., *The Sequence Alignment/Map format and SAMtools.* Bioinformatics, 2009. **25**(16): p. 2078-2079.
60. Langmead, B. and S.L. Salzberg, *Fast gapped-read alignment with Bowtie 2.* Nature Methods, 2012. **9**(4): p. 357-359.
61. Gaspar, J., *Genrich: detecting sites of genomic enrichment*. 2018.
62. Quinlan, A.R. and I.M. Hall, *BEDTools: a flexible suite of utilities for comparing genomic features.* Bioinformatics, 2010. **26**(6): p. 841-842.
63. Yu, G., L.-G. Wang, and Q.-Y. He, *ChIPseeker: an R/Bioconductor package for ChIP peak annotation, comparison and visualization.* Bioinformatics, 2015. **31**(14): p. 2382-2383.
64. BC, T. and M. BP, *TxDb.Hsapiens.UCSC.hg38.knownGene: Annotation package for TxDb object(s)*. 2019.
65. Carlson, M., *org.Hs.eg.db: Genome wide annotation for Human*. 2019.
66. Zinn, N., et al., *Improved Proteomics-Based Drug Mechanism-of-Action Studies Using 16-Plex Isobaric Mass Tags.* J Proteome Res, 2021. **20**(3): p. 1792-1801.
67. Bairoch, A. and B. Boeckmann, *The SWISS-PROT protein sequence data bank.* Nucleic Acids Res, 1991. **19 Suppl**(Suppl): p. 2247-9.
68. Perrin, J., et al., *Identifying drug targets in tissues and whole blood with thermal-shift profiling.* Nature Biotechnology, 2020. **38**(3): p. 303-308.
69. Wishart, D.S., et al., *HMDB: the Human Metabolome Database.* Nucleic Acids Research, 2007. **35**(suppl_1): p. D521-D526.
70. Yang, W., et al., *Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells.* Nucleic Acids Research, 2013. **41**(D1): p. D955-D961.
71. R Core Team, *R: A Language and Environment for Statistical Computing*. 2020, R Foundation for Statistical Computing: Vienna, Austria.
72. Virtanen, S., et al. *Bayesian group factor analysis*. in *Artificial Intelligence and Statistics*. 2012. PMLR.
73. Dray, S., D. Chessel, and J. Thioulouse, *Co-Inertia analysis and the linking of ecological data tables.* Ecology, 2003. **84**: p. 3078-3089.
74. Pozdeyev, N., et al., *Integrating heterogeneous drug sensitivity data from cancer pharmacogenomic studies.* Oncotarget, 2016. **7**(32): p. 51619-51625.
75. Dong, Z., et al., *Anticancer drug sensitivity prediction in cell lines from baseline gene expression through recursive feature selection.* BMC Cancer, 2015. **15**(1): p. 489.
76. Benjamini, Y. and Y. Hochberg, *Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing.* Journal of the Royal Statistical Society. Series B (Methodological), 1995. **57**(1): p. 289-300.
77. Voillet, V., et al., *Handling missing rows in multi-omics data integration: multiple imputation in multiple factor analysis framework.* BMC Bioinformatics, 2016. **17**(1): p. 402.
78. Kodinariya, T.M. and P.R. Makwana, *Review on determining number of Cluster in K-Means Clustering.* International Journal, 2013. **1**(6): p. 90-95.
79. McCabe, S.D., D.-Y. Lin, and M.I. Love, *Consistency and overfitting of multi-omics methods on experimental data.* Briefings in Bioinformatics, 2020. **21**(4): p. 1277-1284.

80. Velten, B. and R. Argelaguet. *Applying MOFA+ to the Chronic Lymphocytic Leukaemia cohort.* 2020; Available from: https://raw.githack.com/bioFAM/MOFA2_tutorials/master/R_tutorials/CLL.html.

81. Eicher, T., et al., *Metabolomics and Multi-Omics Integration: A Survey of Computational Methods and Resources.* Metabolites, 2020. **10**(5).

82. Wörheide, M.A., et al., *Multi-omics integration in biomedical research – A metabolomics-centric review.* Analytica Chimica Acta, 2021. **1141**: p. 144-162.

83. Dutta, N.K., et al., *Integration of metabolomics and transcriptomics reveals novel biomarkers in the blood for tuberculosis diagnosis in children.* Scientific Reports, 2020. **10**(1): p. 19527.

84. Spicker, J.S., et al., *Integration of clinical chemistry, expression, and metabolite data leads to better toxicological class separation.* Toxicol Sci, 2008. **102**(2): p. 444-54.

85. Daskalaki, E., et al., *The influence of culture media upon observed cell secretome metabolite profiles: The balance between cell viability and data interpretability.* Anal Chim Acta, 2018. **1037**: p. 338-350.

86. Fisher, R., L. Pusztai, and C. Swanton, *Cancer heterogeneity: implications for targeted therapeutics.* Br J Cancer, 2013. **108**(3): p. 479-85.

87. Lee, D., Y. Park, and S. Kim, *Towards multi-omics characterization of tumor heterogeneity: a comprehensive review of statistical and machine learning approaches.* Briefings in Bioinformatics, 2021. **22**(3): p. bbaa188.

88. Wu, R., et al., *Correct interpretation of comprehensive phosphorylation dynamics requires normalization by protein expression changes.* Mol Cell Proteomics, 2011. **10**(8): p. M111.009654.

89. Fallahi-Sichani, M., et al., *Metrics other than potency reveal systematic variation in responses to cancer drugs.* Nature Chemical Biology, 2013. **9**(11): p. 708-714.

90. Qureshi, R., et al., *Machine learning based personalized drug response prediction for lung cancer patients.* Scientific Reports, 2022. **12**(1): p. 18935.

91. Dagogo-Jack, I. and A.T. Shaw, *Tumour heterogeneity and resistance to cancer therapies.* Nature Reviews Clinical Oncology, 2018. **15**(2): p. 81-94.

92. Ringnér, M., *What is principal component analysis?* Nature Biotechnology, 2008. **26**(3): p. 303-304.

93. Canzler, S., et al., *Prospects and challenges of multi-omics data integration in toxicology.* Archives of Toxicology, 2020. **94**(2): p. 371-388.

94. Lovmar, L., et al., *Silhouette scores for assessment of SNP genotype clusters.* BMC Genomics, 2005. **6**(1): p. 35.

95. Canzler, S. and J. Hackermüller, *multiGSEA: a GSEA-based pathway enrichment analysis for multi-omics data.* BMC Bioinformatics, 2020. **21**(1): p. 561.

96. Paczkowska, M., et al., *Integrative pathway enrichment analysis of multivariate omics data.* Nature Communications, 2020. **11**(1): p. 735.

97. Safikhani, Z., et al., *Gene isoforms as expression-based biomarkers predictive of drug response in vitro.* Nature Communications, 2017. **8**(1): p. 1126.

98. Costello, J.C., et al., *A community effort to assess and improve drug sensitivity prediction algorithms.* Nature Biotechnology, 2014. **32**(12): p. 1202-1212.

99. Ahmadi Moughari, F. and C. Eslahchi, *ADRML: anticancer drug response prediction using manifold learning.* Scientific Reports, 2020. **10**(1): p. 14245.

100. Ben-David, U., et al., *Genetic and transcriptional evolution alters cancer cell line drug response.* Nature, 2018. **560**(7718): p. 325-330.

101. Xia, F., et al., *A cross-study analysis of drug response prediction in cancer cell lines.* Briefings in Bioinformatics, 2022. **23**(1): p. bbab356.

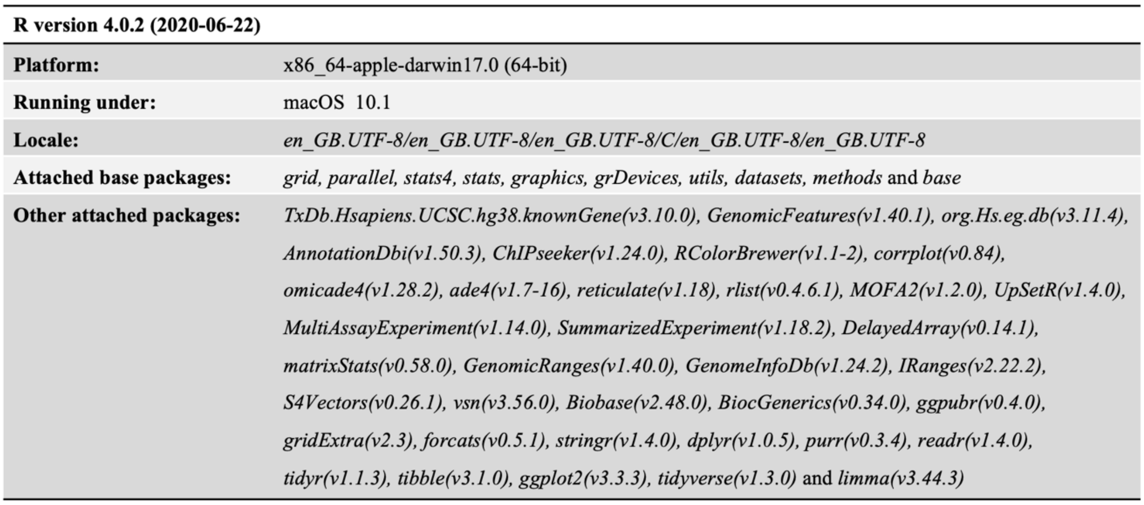
102. Rahman, R., et al., *Evaluating the consistency of large-scale pharmacogenomic studies.* Briefings in Bioinformatics, 2019. **20**(5): p. 1734-1753.

103. Gupta, A., et al., *A normalized drug response metric improves accuracy and consistency of anticancer drug sensitivity quantification in cell-based screening.* Communications Biology, 2020. **3**(1): p. 42.

104. Smirnov, P., et al., *PharmacoDB: an integrative database for mining in vitro anticancer drug screening studies.* Nucleic Acids Research, 2018. **46**(D1): p. D994-D1002.

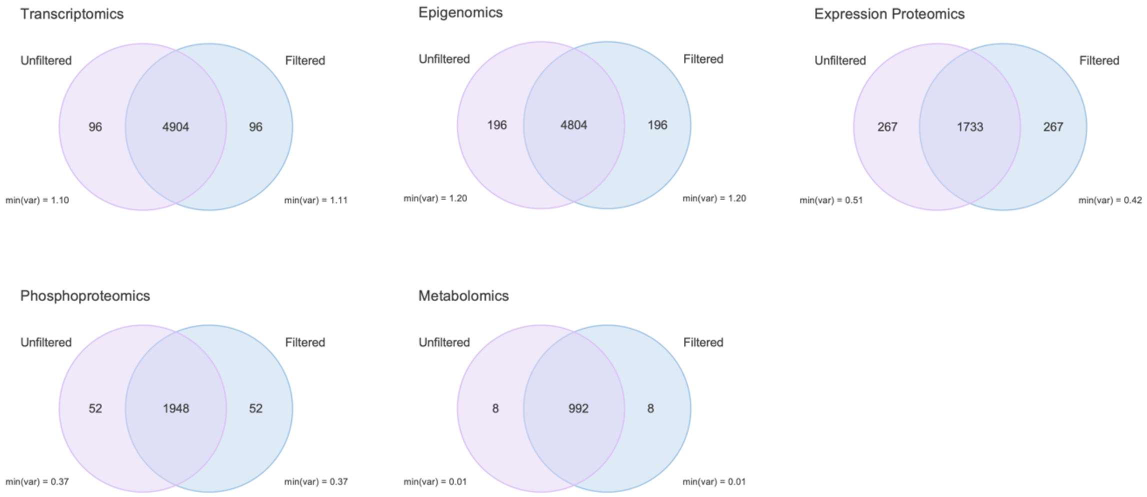
Supplementary Information

R Session Information



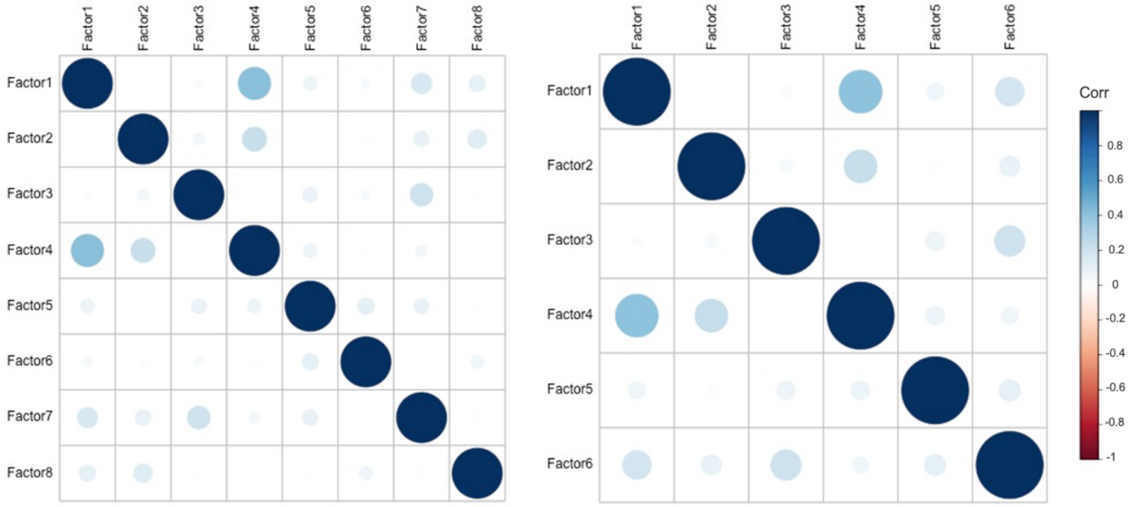
R version 4.0.2 (2020-06-22)	
Platform:	x86_64-apple-darwin17.0 (64-bit)
Running under:	macOS 10.1
Locale:	*en_GB.UTF-8/en_GB.UTF-8/en_GB.UTF-8/C/en_GB.UTF-8/en_GB.UTF-8*
Attached base packages:	*grid, parallel, stats4, stats, graphics, grDevices, utils, datasets, methods* and *base*
Other attached packages:	*TxDb.Hsapiens.UCSC.hg38.knownGene(v3.10.0), GenomicFeatures(v1.40.1), org.Hs.eg.db(v3.11.4), AnnotationDbi(v1.50.3), ChIPseeker(v1.24.0), RColorBrewer(v1.1-2), corrplot(v0.84), omicade4(v1.28.2), ade4(v1.7-16), reticulate(v1.18), rlist(v0.4.6.1), MOFA2(v1.2.0), UpSetR(v1.4.0), MultiAssayExperiment(v1.14.0), SummarizedExperiment(v1.18.2), DelayedArray(v0.14.1), matrixStats(v0.58.0), GenomicRanges(v1.40.0), GenomeInfoDb(v1.24.2), IRanges(v2.22.2), S4Vectors(v0.26.1), vsn(v3.56.0), Biobase(v2.48.0), BiocGenerics(v0.34.0), ggpubr(v0.4.0), gridExtra(v2.3), forcats(v0.5.1), stringr(v1.4.0), dplyr(v1.0.5), purr(v0.3.4), readr(v1.4.0), tidyr(v1.1.3), tibble(v3.1.0), ggplot2(v3.3.3), tidyverse(v1.3.0)* and *limma(v3.44.3)*

Supplementary Figures



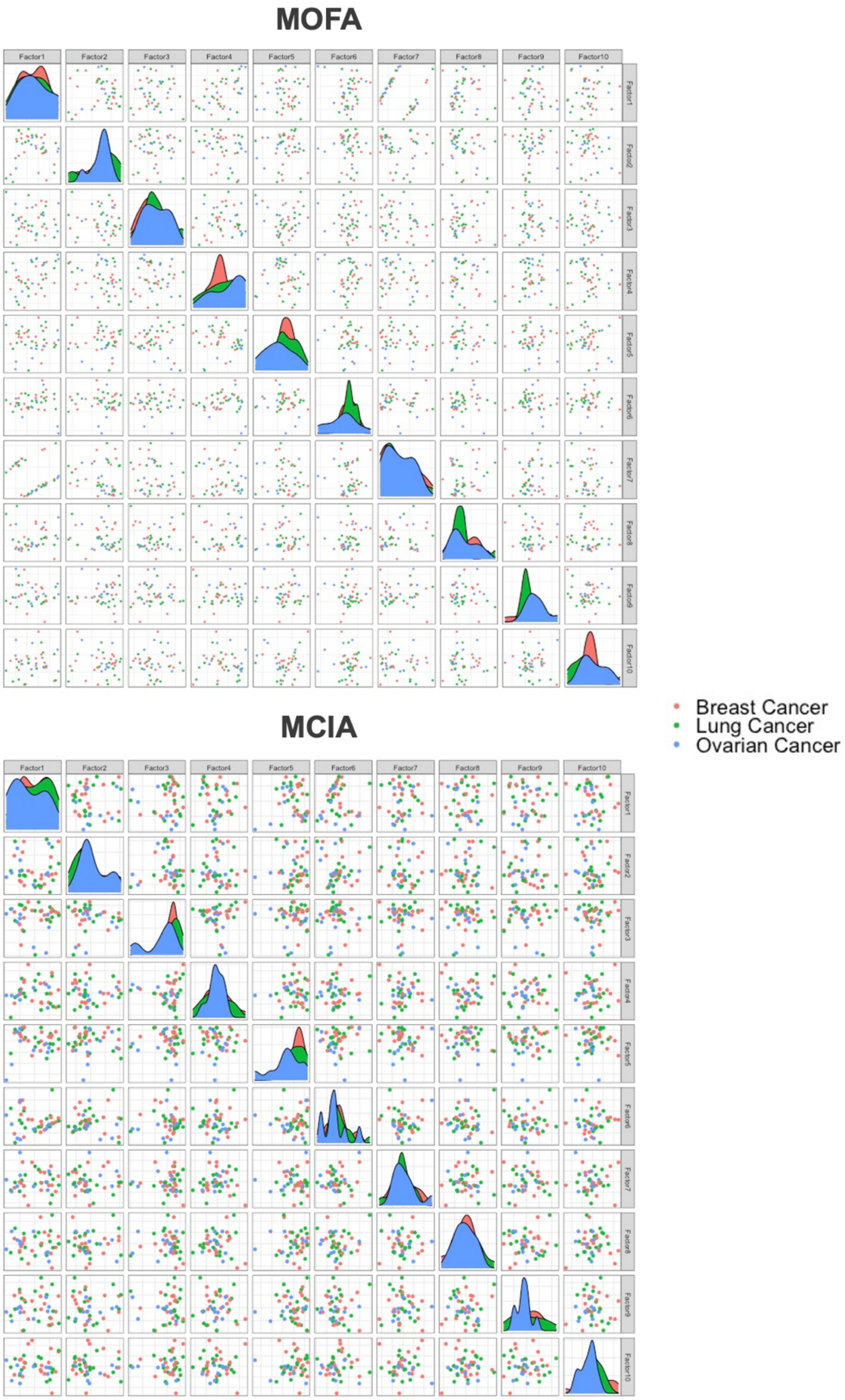
Supplementary Figure S1. Overlap of selected highly variable features before and after filtering features and cell lines missing data for the five omics datasets.

min(var) = minimum variance of the feature subset.



Supplementary Figure S2. Evaluation of reducing number of learnt features on MOFA factor correlations.

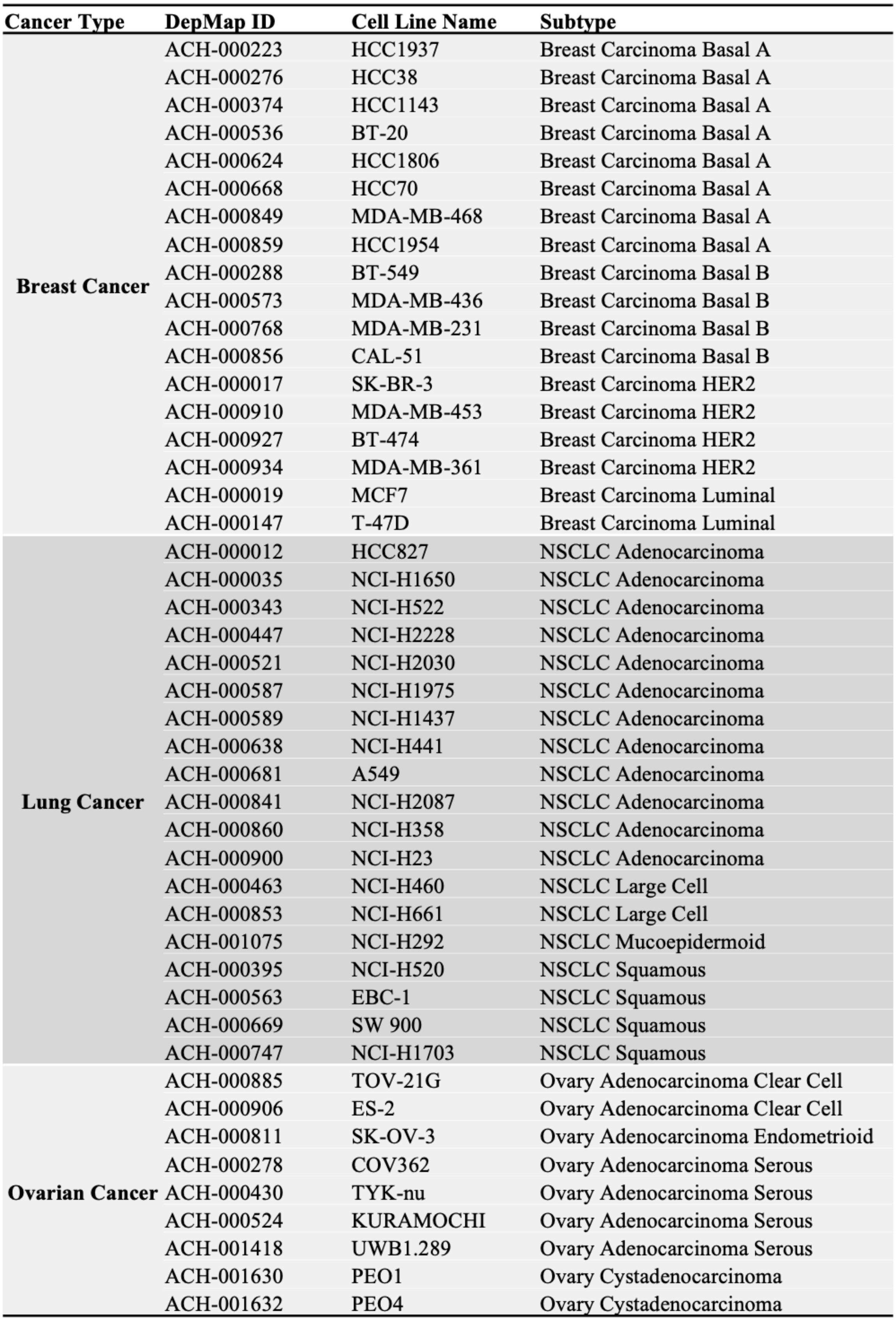
Plots of correlation matrix for latent factors (left = 8 learnt factors, right = 6 learnt factors). The areas of circles show the absolute value of corresponding Pearson correlation coefficients. Colour depicts the direction (positive correlations are displayed in blue and negative correlations in red) and size (low correlations have low colour intensity and high correlations have high colour intensity) of correlation coefficients.



Supplementary Figure S3. Visualisation of cancer cell lines using all pairwise comparisons of MOFA (top) and MCIA (bottom) latent factors

Each point represents a cancer cell line and colour denotes the cancer type.

Supplementary Tables



Cancer Type	DepMap ID	Cell Line Name	Subtype
Breast Cancer	ACH-000223	HCC1937	Breast Carcinoma Basal A
	ACH-000276	HCC38	Breast Carcinoma Basal A
	ACH-000374	HCC1143	Breast Carcinoma Basal A
	ACH-000536	BT-20	Breast Carcinoma Basal A
	ACH-000624	HCC1806	Breast Carcinoma Basal A
	ACH-000668	HCC70	Breast Carcinoma Basal A
	ACH-000849	MDA-MB-468	Breast Carcinoma Basal A
	ACH-000859	HCC1954	Breast Carcinoma Basal A
	ACH-000288	BT-549	Breast Carcinoma Basal B
	ACH-000573	MDA-MB-436	Breast Carcinoma Basal B
	ACH-000768	MDA-MB-231	Breast Carcinoma Basal B
	ACH-000856	CAL-51	Breast Carcinoma Basal B
	ACH-000017	SK-BR-3	Breast Carcinoma HER2
	ACH-000910	MDA-MB-453	Breast Carcinoma HER2
	ACH-000927	BT-474	Breast Carcinoma HER2
	ACH-000934	MDA-MB-361	Breast Carcinoma HER2
	ACH-000019	MCF7	Breast Carcinoma Luminal
	ACH-000147	T-47D	Breast Carcinoma Luminal
Lung Cancer	ACH-000012	HCC827	NSCLC Adenocarcinoma
	ACH-000035	NCI-H1650	NSCLC Adenocarcinoma
	ACH-000343	NCI-H522	NSCLC Adenocarcinoma
	ACH-000447	NCI-H2228	NSCLC Adenocarcinoma
	ACH-000521	NCI-H2030	NSCLC Adenocarcinoma
	ACH-000587	NCI-H1975	NSCLC Adenocarcinoma
	ACH-000589	NCI-H1437	NSCLC Adenocarcinoma
	ACH-000638	NCI-H441	NSCLC Adenocarcinoma
	ACH-000681	A549	NSCLC Adenocarcinoma
	ACH-000841	NCI-H2087	NSCLC Adenocarcinoma
	ACH-000860	NCI-H358	NSCLC Adenocarcinoma
	ACH-000900	NCI-H23	NSCLC Adenocarcinoma
	ACH-000463	NCI-H460	NSCLC Large Cell
	ACH-000853	NCI-H661	NSCLC Large Cell
	ACH-001075	NCI-H292	NSCLC Mucoepidermoid
	ACH-000395	NCI-H520	NSCLC Squamous
	ACH-000563	EBC-1	NSCLC Squamous
	ACH-000669	SW 900	NSCLC Squamous
	ACH-000747	NCI-H1703	NSCLC Squamous
Ovarian Cancer	ACH-000885	TOV-21G	Ovary Adenocarcinoma Clear Cell
	ACH-000906	ES-2	Ovary Adenocarcinoma Clear Cell
	ACH-000811	SK-OV-3	Ovary Adenocarcinoma Endometrioid
	ACH-000278	COV362	Ovary Adenocarcinoma Serous
	ACH-000430	TYK-nu	Ovary Adenocarcinoma Serous
	ACH-000524	KURAMOCHI	Ovary Adenocarcinoma Serous
	ACH-001418	UWB1.289	Ovary Adenocarcinoma Serous
	ACH-001630	PEO1	Ovary Cystadenocarcinoma
	ACH-001632	PEO4	Ovary Cystadenocarcinoma

Supplementary Table S1. List of 46 cell lines used in analysis