

Latent Feature Models for Multiomic Data Analysis



Aleksandra Ziubroniewicz

Kellogg College, University of Oxford

Supervisors: Prof. Dan Woodcock and Prof. Calliope Dendrou

A thesis submitted for the degree of

Doctor of Philosophy

October 2024

Abstract

Every year, cancer kills millions of people around the world. Treatment efficacy, including surgery, radiotherapy, and chemotherapy, varies considerably across tumour types, and growing evidence shows that the molecular subtype of the disease can be linked to clinical outcomes and hence inform clinical decision making. While DNA mutations are pivotal to cancer development, other factors such as methylation, RNA, and proteins also play critical roles, requiring a comprehensive, multimodal analysis for a holistic understanding of the disease. This has spurred significant research into multiomic data integration and analysis, which can reveal meaningful subtypes and guide treatment decisions. However, integrative analysis of multiomic datasets is challenging, and current methods fail to sufficiently address the complexity, dimensionality, prevalence of missing data, and heterogeneities that characterise different omics outputs.

To address these limitations, this thesis focuses on the development of a novel latent feature model tailored for the integrative analysis of multiomic datasets with missing modalities. We introduce iCS-GAN (integrative Cancer Subtyping with Generative Adversarial Networks) - a method that leverages adversarially learned inference to extract clustering-relevant binary

latent features from multiomic data. The proposed approach employs a combination of shared and modality-specific layers, layer-wise pre-training, robust imputation techniques, and adversarial loss functions, to consistently integrate heterogeneous data, even in the presence of incomplete datasets. Non-negativity constraints ensure that the latent variables remain fully interpretable and any results are amendable for translation for clinical use. Furthermore, clustering and survival penalties guide the latent encodings and subsequent analysis towards clinically-relevant disease subtypes.

We demonstrate the utility of iCS-GAN through a comprehensive analysis of the PanProstate Cancer Group multiomic prostate cancer dataset. Our study identifies three distinct multiomic prostate cancer subtypes, including a novel aggressive subtype characterized by low expression levels of the *ERG* and *TFF3* genes. To facilitate clinical translation, we develop a highly accurate predictive test, capable of classifying patients into these subtypes using only 24 RNA gene expression levels. Upon external validation, this test could support low-cost, clinically viable patient stratification, paving the way for improved cancer outcomes and personalized care.

Acknowledgements

First and foremost, I would like to express my sincere gratitude to my supervisors, Dan Woodcock and Calliope Dendrou, for their invaluable support, guidance, and feedback throughout my DPhil. My gratefulness extends to EPSRC Centre for Doctoral Training in Health Data Science for the training, resources and funding provided.

I would also like to thank all members of the PanProstate Cancer Group and the Prostate Cancer Research charity for their support and stimulating discussions, with special acknowledgments to Daniel Brewer and David Wedge, with whom I had the pleasure of working over the past few years.

I would like to extend my deepest thanks to my husband, Damian, for his unwavering love, encouragement, and for always ensuring our home office remained a joyful space. I am also grateful for his time spent proofreading this work.

My sincere appreciation goes to my parents, Izabela and Karol, and my brother Lukasz, for their continuous support and encouragement throughout my academic journey.

Finally, thank you to you, the reader, for taking the time to read my thesis.

Aleksandra Ziubroniewicz

Preface

This DPhil thesis represents an exploration into latent feature methods suitable for multiomic data analysis, and the application of such methods for prostate cancer subtyping, undertaken under the supervision of Dan Woodcock and Calliope Dendrou at the University of Oxford, with support from the PanProstate Cancer Group and the Prostate Cancer Research charity.

This DPhil has been funded by a EPSRC Centre for Doctoral Training in Health Data Science (EP/S02428X/1) in partnership with Kellogg College, University of Oxford.

The research conducted is original and builds on the foundations established by numerous previous studies in the field, with proper acknowledgment of these works through appropriate citations. ChatGPT was utilized solely for refining the writing. No part of this thesis has been previously submitted for a degree. This research represent primarily my independent work, under the guidance of Dan Woodcock and Calliope Dendrou, with the following exceptions:

- The primary dataset explored in this DPhil, namely the PanProstate Cancer Group prostate cancer dataset, was generated, curated and pre-processed by the members of the PanProstate Cancer Group consortium.

- The publicly available datasets used for validation of the method proposed in this DPhil, were generated, curated and pre-processed by The Cancer Genome Atlas Program, representing joint effort between the National Cancer Institute and the National Human Genome Research Institute.
- The subtype assignments for established prostate cancer classification frameworks used for comparison with the subtyping schema proposed in this DPhil in Chapter 5 were generated by other members of the PanProstate Cancer Group: Sergio Llaneza Lago (DESNT classification), Valeriia Haberland (You et al. classification) and Emre Esenturk (Evotypes classification).

Throughout my degree, I have additionally contributed to the following publication, which I summarize and refer to for comparative purposes in this work:

- Woodcock, D.J., Sahli, A., Teslo, R., Bhandari, V., Gruber, A.J., **Ziubroniewicz**, A., Gundem, G., Xu, Y., Butler, A., Anokian, E., Pope, B.J., Jung, C-H., Tarabichi, M., Dentre, S.C., Farmery, J.H.R., CRUK ICGC Prostate Group, Loo, P.V., Warren, A.Y., Gnanapragasam, V., Hamdy, F.C., Bova, G.S., Foster, C.S., Neal, D.E., Lu, Y-J., Kote-Jarai, Z., Fraser, M., Bristow, R.G., Boutros, P.C., Costello, A.J., Corcoran, N.M., Hovens, C.M., Massie, C.E., Lynch, A.G., Brewer, D.S., Eeles, R.A., Cooper, C.S., Wedge, D.C. “Genomic evolution shapes prostate cancer disease type”. In: Cell Genomics 4.3 (2024).

Table of Contents

List of Figures	xi
List of Tables	xv
List of Abbreviations	xvii
Introduction	1
1 Data	10
1.1 Terminology	11
1.1.1 DNA and the Genome	11
1.1.2 RNA and the Transcriptome	16
1.1.3 Methylation and the Methylome	18
1.1.4 Other Omics Sources	19
1.1.5 Targeted Therapies	19
1.2 PPCG Dataset	20
1.2.1 Sample Selection	22
1.2.2 Clinical Measurements	24
1.2.3 DNA Data	26
1.2.4 RNA Data	33
1.2.5 The Final Dataset	35

1.2.6	Prostate Cancer Subtyping Frameworks	36
1.3	TCGA Datasets	38
1.4	Synthetic Datasets	43
2	Methodology: Literature Review and Background	46
2.1	Multiomic Data Integration	47
2.1.1	Early Integration	48
2.1.2	Mixed Integration	48
2.1.3	Intermediate Integration	48
2.1.4	Late Integration	49
2.1.5	Hierarchical Integration	49
2.1.6	Choosing an Integration Strategy	50
2.2	Latent Feature Models	51
2.2.1	Statistical Methods	53
2.2.2	Machine Learning Methods	57
2.3	Generative Adversarial Networks	64
2.3.1	Basic Definition	64
2.3.2	Wasserstein GANs	67
2.3.3	Inference Networks in GANs	69
2.3.4	GANs and Multiomic Data	73
2.4	Data Types and Distributions	74
2.5	Missing Modalities	77
2.6	Interpretability	81
2.7	Subtyping and Survival Analysis	83
2.7.1	Subtyping (Clustering)	83
2.7.2	Survival Analysis	88

3	Proposed Method	90
3.1	Single-Modality Model	92
3.1.1	Wasserstein Training Procedure	93
3.1.2	Interpretability	94
3.1.3	Support for Mixed Data Types and Distributions	96
3.1.4	Clustering Regularization	99
3.1.5	Survival Regularization	101
3.2	Multiomic Data Integration Model	102
3.2.1	Shared and Independent Layers	103
3.2.2	Layer-Wise Pre-Training	103
3.2.3	Indian Buffet Process Prior	106
3.3	Multiomic Data Integration with Missing Modalities	109
4	Testing and Validation	112
4.1	Evaluation Metrics	113
4.2	Experimental Design	114
4.3	Synthetic Datasets: Testing	118
4.4	TCGA: Single-Modality Validation	121
4.5	TCGA: Multiomic Integration Validation	128
4.6	TCGA: Gold Standard Comparison	133
4.6.1	BRCA - ER, PR and HER2 Subtypes	133
4.6.2	COAD - CMS Subtypes	136
4.6.3	HNSC - HPV Subtypes	138
4.6.4	Remarks	140
4.7	TCGA: Survival Experiments	141
4.8	TCGA: Experiments with Missing Modalities	144

5	Results: Prostate Cancer Subtyping	146
5.1	Single Modality Subtyping	147
5.1.1	Subtyping with Summary Measurements	148
5.1.2	Subtyping with Drivers	153
5.1.3	Subtyping with CNAs	157
5.1.4	Subtyping with RNA	162
5.2	Multiomic Subtyping	166
5.2.1	Subtype Characteristics	169
5.2.2	Latent Variables	174
5.2.3	Clinical Characteristics	178
5.2.4	Comparison with Single-Modality Subtypes	179
5.2.5	Predictive Tests	182
5.3	Comparison with Previously Established Subtypes	185
5.3.1	Evotypes and Multiomic Subtypes	187
5.3.2	Risk Stratification Utility Comparison	195
6	Discussion	196
6.1	Contributions to Methods for Multiomic Data Analysis	197
6.1.1	Adversarially Learned Inference Models are an Effective Tool for Multiomic Data Integration	197
6.1.2	Survival Regularization Can Guide the Discovery of Clinically-Relevant Disease Subtypes	198
6.1.3	Interpretable Models Extract Domain-Relevant Latent Variables	199
6.2	Contributions to Prostate Cancer Subtyping	200
6.2.1	PGA, Kataegis, Chromothripsis and Ploidy Predict Negative Prostate Cancer Outcomes	200

6.2.2	ETS Status and ERG Fusions Are Not Indicators of Prostate Cancer Patient Prognosis	201
6.2.3	CNA Burden Predicts Prostate Cancer Relapse	202
6.2.4	Multomic Analysis Reveals 3 Distinct Prostate Cancer Subtypes	202
6.2.5	Prostate Cancer Subtypes Can Be Accurately Predicted from 24 RNA Expressions	204
6.2.6	Multomic Analysis Provides Additional Insights into Prostate Cancer Evotypes	204
6.3	Limitations	206
6.4	Future Work	207
6.5	Ethical Considerations	208
6.5.1	PPCG Dataset: Consent, Privacy and Security	208
6.5.2	PPCG Dataset: Sample Bias	209
6.5.3	Environmental Impact of GANs	210
6.5.4	ML Models: Transparency and Explainability	210
6.5.5	Commercialization vs Improving Patient Care	211
6.6	Final Remarks	211
	References	212
	A Implementation Details and Experimental Setup	237
A.1	iCS-GAN - Implementation and Default Hyperparameters	237
A.2	AE - Implementation and Hyperparameter Search	243
A.3	nnAE - Implementation and Hyperparameter Search	244
A.4	VAE - Implementation and Hyperparameter Search	246
A.5	PPCG - Experimental Setup	248
	B Supplementary Results	251

List of Figures

1.1	PPCG - Demographic and Clinical Measurements	25
1.2	PPCG - Summary Measurements	28
1.3	PPCG - Mutational Landscape	31
1.4	PPCG - CNA LOH Landscape	32
1.5	PPCG - CNA Gains Landscape	32
1.6	PPCG - CNA HD Landscape	33
1.7	PPCG - RNA Data	34
1.8	PPCG - Venn Diagram	36
1.9	TCGA - Survival Summary	41
1.10	TCGA - Frequency Histograms	42
1.11	Synthetic Datasets - Frequency Histograms	44
1.12	Synthetic Datasets - UMAP Visualisations	45
2.1	Multiomic Data Integration Strategies	50
2.2	IBP Matrices	57
2.3	Autoencoder	58
2.4	Variational Autoencoder	61
2.5	Restricted Boltzmann Machine	63
2.6	Generative Adversarial Network	66
2.7	Adversarially Learned Inference	70
2.8	Adversarially Learned Inference with Conditional Entropy	72

3.1	iCS-GAN Multiomic Integration Generator	104
3.2	iCS-GAN Layer-Wise Pre-Training	107
4.1	Testing - Synthetic Datasets Reconstructions	119
4.2	Testing - Synthetic Datasets Clustering UMAPs	120
4.3	Single-Modality Validation - BRCA UMAPs	123
4.4	Single-Modality Validation - KIRC UMAPs	124
4.5	Single-Modality Validation - BLCA UMAPs	124
4.6	Single-Modality Validation - COAD UMAPs	125
4.7	Single-Modality Validation - HNSC UMAPs	125
4.8	Single-Modality Validation - TCGA Reconstructions	126
4.9	Single-Modality Validation - TCGA Interpretability	127
4.10	Single-Modality Validation - TCGA (V)AE Interpretability	127
4.11	Multiomic Integration Validation - BRCA UMAP	129
4.12	Multiomic Integration Validation - KIRC UMAP	130
4.13	Multiomic Integration Validation - BLCA UMAP	130
4.14	Multiomic Integration Validation - COAD UMAP	130
4.15	Multiomic Integration Validation - HNSC UMAP	131
4.16	Multiomic Integration Validation - TCGA Reconstructions	131
4.17	Multiomic Integration Validation - TCGA Encodings	132
4.18	Multiomic Integration Validation - TCGA Interpretability	132
4.19	Gold Standard Subtype Comparison - BRCA	135
4.20	Gold Standard Subtype Comparison - COAD	138
4.21	Gold Standard Subtype Comparison - HNSC	139
4.22	Gold Standard Subtype Comparison - HNSC (mRNA)	140
4.23	Survival Regularization Validation - KIRC UMAPs I	143
4.24	Survival Regularization Validation - KIRC UMAPs II	143

5.1	PPCG SM Results - KM Plots	149
5.2	PPCG SM Results - Subtype Characteristics	150
5.3	PPCG SM Results - Encoding Weights	151
5.4	PPCG SM Results - SHAP LV Plots	152
5.5	PPCG Drivers Results - KM Plots	153
5.6	PPCG Drivers Results - Subtype Characteristics	154
5.7	PPCG Drivers Results - Encoding Weights	155
5.8	PPCG Drivers Results - SHAP LV Plots	156
5.9	PPCG CNA Results - KM Plots	158
5.10	PPCG CNA Results - Subtype Characteristics	159
5.11	PPCG CNA Results - Encoding Weights	160
5.12	PPCG CNA Results - SHAP LV Plots	161
5.13	PPCG RNA Results - KM Plots	162
5.14	PPCG RNA Results - Subtype Characteristics	164
5.15	PPCG RNA Results - SHAP LV Plots	165
5.16	PPCG Multiomic Results - UMAP (Complete Data)	167
5.17	PPCG Multiomic Results - UMAP (Entire Dataset)	167
5.18	PPCG Multiomic Results - KM Plots	168
5.19	PPCG Multiomic Results - Subtype Characteristics I	170
5.20	PPCG Multiomic Results - Subtype Characteristics II	171
5.21	PPCG Multiomic Results - Subtype Characteristics III	172
5.22	PPCG Multiomic Results - Subtype Characteristics IV	173
5.23	PPCG Multiomic Results - SHAP LV Plots I	174
5.24	PPCG Multiomic Results - SHAP LV Plots II	175
5.25	PPCG Multiomic Results - SHAP LV Plots III	176
5.26	PPCG Multiomic Results - Latent Space Encodings	177
5.27	PPCG Multiomic Results - Clinical Characteristics	178

5.28	PPCG Multiomic vs Summary Measurement Subtypes	180
5.29	PPCG Multiomic vs Driver Genes Subtypes	180
5.30	PPCG Multiomic vs CNA Subtypes	181
5.31	PPCG Multiomic vs RNA Subtypes	181
5.32	PPCG Multiomic Results - Predictive Test SHAP	183
5.33	PPCG Multiomic Results - Predictive Test KM Plots	184
5.34	PPCG Comparison Results - DESNT	185
5.35	PPCG Comparison Results - You et al.	186
5.36	PPCG Comparison Results - Evotypes	186
5.37	PPCG Evotypes vs Multiomic Subtypes - UMAP	188
5.38	PPCG Multiomic-Evotypes Results - KM Plots	189
5.39	PPCG Multiomic-Evotypes Results - Characteristics I	191
5.40	PPCG Multiomic-Evotypes Results - Characteristics II	192
5.41	PPCG Multiomic-Evotypes Results - Characteristics III	193
5.42	PPCG Multiomic-Evotypes Results - Characteristics IV	194
5.43	PPCG Subtyping Schemas Comparison - KM Plots	195

List of Tables

1.1	TCGA Datasets - Summary	40
1.2	Synthetic Datasets - Summary	44
4.1	Testing - Synthetic Datasets Results	118
4.2	Single-Modality Validation - TCGA Results	122
4.3	Single-Modality Validation - (V)AE Comparison	122
4.4	Single-Modality Validation - TCGA Ablation Analysis	123
4.5	Multiomic Integration Validation - TCGA Results	128
4.6	Multiomic Integration Validation - TCGA Ablation Analysis	129
4.7	Gold Standard Subtype Comparison - BRCA	135
4.8	Gold Standard Subtype Comparison - COAD	138
4.9	Gold Standard Subtype Comparison - HNSC	140
4.10	Survival Regularization Validation - TCGA Concordance	142
4.11	Survival Regularization Validation - KIRC Results	142
4.12	Missing Modalities Validation - TCGA Missing Data	145
4.13	Missing Modalities Validation - TCGA Results	145
4.14	Missing Modalities Validation - TCGA Comparison Results	145
B.1	Single-Modality Validation - TCGA AE Results	251
B.2	Single-Modality Validation - TCGA mAE Results	252
B.3	Single-Modality Validation - TCGA VAE Results	252

B.4	Single-Modality Validation - TCGA Results W/O WD	253
B.5	Single-Modality Validation - TCGA Results W/O INT	253
B.6	Single-Modality Validation - TCGA Results W/O BIN	254
B.7	Single-Modality Validation - TCGA Results W/O CL	254
B.8	Multiomic-Integration Validation - TCGA Results W/O IND .	254
B.9	Multiomic-Integration Validation - TCGA Results W/O PT .	255
B.10	Multiomic Integration Validation - TCGA Results W/O CL .	255
B.11	Multiomic Integration Validation - TCGA Results W/O CLI .	255
B.12	Multiomic Integration Validation - TCGA Results W/O IBP .	255

List of Abbreviations

AE	Autoencoder
ALI	Adversarially Learned Inference
ALICE	Adversarially Learned Inference with Conditional Entropy
AR	Androgen Receptor
ARI	Adjusted Rand Index
BCR	Biochemical Recurrence
BLCA	Bladder Urothelial Carcinoma
BRCA	Breast Invasive Carcinoma
CNA	Copy Number Alteration
COAD	Colorectal Adenocarcinoma
CQRNN	Censored Quantile Regression Neural Network
DCEC	Deep Convolutional Embedded Clustering
DNA	Deoxyribonucleic Acid

GAIN	Copy Number Gain
GAN	Generative Adversarial Network
HD	Homozygous Deletion
HNSC	Head and Neck Squamous Cell Carcinoma
IBP	Indian Buffet Process
KIRC	Kidney Renal Clear Cell Carcinoma
KNN	K-Nearest-Neighbors
LOH	Loss of Heterozygosity
LVs	Latent Variables
MFS	Metastasis-Free Survival
mRNA	Messenger RNA
NMF	Non-Negative Matrix Factorization
PCA	Principal Component Analysis
PFS	Progression-Free Survival
PPCG	Pan Prostate Cancer Group
PSA	Prostate-Specific Antigen
RBM	Restricted Boltzmann Machine
RFC	Random Forest Classifier
RFS	Relapse-Free Survival

RI	Rand Index
RNA	Ribonucleic Acid
RNA-seq	RNA Sequencing
rRNA	Ribosomal RNA
SHAP	SHapley Additive exPlanations
SNV	Single Nucleotide Variant
SV	Structural Variant
TCGA	The Cancer Genome Atlas
tRNA	Transfer RNA
VAE	Variational Autoencoder
WGAN	Wasserstein Generative Adversarial Network
WGAN-GP	Wasserstein Generative Adversarial Network with Gradient Penalty
WGS	Whole Genome DNA Sequence

Introduction

Every year, cancer kills millions of people around the world [20]. In the UK specifically, there are around 167,000 cancer deaths per annum, which translates to nearly 460 cancer deaths every day [21]. A person gets diagnosed with cancer every two minutes and it is estimated that one in two people will get some form of cancer in their lifetime [21].

The term cancer itself refers to a group of diseases characterised by an uncontrolled growth and proliferation of abnormal cells. The human body is composed of many different cell types, each with a defined functionality. Cell behaviour is governed by many different interacting processes, from inter- and intra-cellular signalling and communication, to the intricate mechanisms of the cell cycle and apoptosis (programmed cell death). In healthy cells, these processes ensure that cells grow, divide, and die at the right time, maintaining tissue integrity and function. In cancerous cells, these processes are disrupted, leading to a variety of pathological outcomes, most notably the uncontrolled growth and division of cells, evasion of programmed cell death, formation of new blood vessels (angiogenesis) and the potential to invade adjacent tissues and spread to distant sites in the body (metastasis). These are sometimes referred to as the “Hallmarks of Cancer” [75].

Cancer begins in normal cells via changes to their DNA sequence - a process known as carcinogenesis. These changes, often known as mutations,

can occur naturally during DNA replication in cell division, and can also be induced through other mechanisms including environmental exposures, such as smoking, and viruses, such as the human papillomavirus (HPV). While mutations occur at the genomic or DNA level, their effects mainly manifest at the gene level. Genes are segments of DNA that are transcribed into RNA in a process known as gene expression. A type of RNA called messenger RNA (mRNA) is then translated to proteins, which carry out most of the cellular processes [39]. Therefore, while DNA mutations can be thought to directly affect the behaviour of genes, these can have wide-ranging effects, from altering individual protein functions to affecting entire cellular pathways and ultimately the fate of the organism. Furthermore, we are also expanding our understanding of how epigenetic factors, which are modifications of the genome that do not affect the DNA code itself, can also influence gene expression in cancer. In particular, DNA methylation has been shown to repress gene expression, which can invoke carcinogenic behaviour when genes that prevent the formation of tumours (tumour suppressors) are hypermethylated.

Once these early carcinogenic mutations occur, they create a cellular state that favours the occurrence of further genetic alterations. This sequential accumulation of mutations drives cancer progression from a precancerous lesion to a full-blown malignancy. Some mutations confer survival advantages, for instance the ability to evade the immune system, and over time, cells with these mutations will outcompete normal cells and less well-adapted malignant cells. Cancer progression is therefore an evolutionary process that shapes tumours to be increasingly aggressive and eventually gain the capability to invade adjacent tissue and metastasise to distant sites. These advanced tumours, if left untreated, will ultimately result in the death of the host.

Treatments such as surgery, radiotherapy, and chemotherapy or other drugs are available to remove or eradicate the tumour, or alternatively, to manage incurable disease. However, treatment efficacy varies considerably across tumour types. We now know that subtypes of many cancer types exist, and that patients with different subtypes should be managed and treated differently [e.g. 63, 184]. For example, we know that certain melanoma drugs such as BRAF and MEK inhibitors are effective only if the *BRAF* gene mutation is present [24], and conversely, colorectal cancer drugs such as EGFR inhibitors are prescribed only for patients without the *KRAS* mutation [117]. Given that DNA mutations are fundamental to cancer development, their utility in determining clinically-relevant cancer subtypes that inform the subsequent treatment selection and response is not surprising.

However, the enormous complexity of cancer cannot be encapsulated by DNA mutations alone. Numerous other factors feature in cancer development at all levels, and subtypes have been proposed based on methylation, RNA, proteins, and even evolutionary trajectories [e.g. 19, 167, 175, 213]. While each of these subtyping schemas may contribute significant insights into the composition of cancers across the population, it is likely that a fully comprehensive description requires complementary information obtained from several modalities. Integrative analysis of these data sources is therefore an active and extensive area of research often referred to as multiomic data analysis. The term ‘multiomics’ itself broadly describes a biological analysis approach focused on tabular data sourced from several ‘omes’ that each pertain to a specific biological component, e.g. genome, proteome, transcriptome, microbiome, etc. Unsurprisingly, joint analysis of data generated via quite different processes that reflect distinct biological phenomena across multiple scales is non-trivial. Nonetheless, many consider multiomic data and

its analysis to be the future of precision medicine and personalized cancer treatments [e.g. 80, 123, 161].

An important application of multiomic data analysis is the identification of molecular subtypes within a specific cancer type. Molecular subtypes are smaller, more specific classifications of a cancer type, defined by distinct genetic alterations, biomarkers, epigenetic modifications, or gene expression profiles, among other factors. These subtypes serve three major purposes: mechanistic, aetiological, and clinical. Mechanistic uses focus on understanding the molecular mechanisms underlying the disease, offering insights into its biology that may lead to the discovery of new therapeutic targets or biomarkers [e.g. 44, 88]. Aetiological uses aim to identify causes and risk factors, such as environmental or lifestyle factors, that contribute to the development of specific subtypes, enabling the design of preventative strategies and exploration of population-level differences in disease incidence [e.g. 59, 163]. Finally, clinical uses leverage molecular subtypes to predict disease outcomes, treatment responses, and resistance mechanisms, allowing for personalized treatment strategies and improved patient stratification [e.g. 45, 46]. It is therefore evident that the identification of molecular subtypes plays a crucial role in advancing cancer research and treatment by enabling deeper biological insights, preventive measures, and personalized therapeutic approaches.

Of specific interest to us in this DPhil is prostate cancer, the most common cancer and the second most common cause of cancer deaths in men in the United Kingdom [22]. It is expected that 1 in 8 men in the UK will get prostate cancer in their lifetime. Currently, the disease is often curable, however, it appears to be more aggressive for some patients, for whom the treatment ultimately fails. Growing evidence shows that the molecular

subtype of prostate cancer can be linked to its aggressiveness and inform clinical decision-making [e.g. 10, 99, 120, 222]. Despite this, clinical decision-making in prostate cancer treatment is currently based on tumour staging, histopathological grading and biochemical markers only. Histopathological evaluation is performed by expert pathologists using the Gleason grading system, which assesses the aggressiveness of the cancer based on the microscopic appearance of prostate cancer cells. Biochemical markers are specific molecules produced during the disease process used to assess its presence or severity. For example, elevated levels of the Prostate-Specific Antigen (PSA) biomarker, a protein exclusively produced by prostate cells, can indicate the presence of prostate cancer. The fact that prostate cancer decision-making is limited to tumour staging and histopathological and biomarker-based assessments is largely caused by the heterogeneous nature of this disease, which renders our ability to identify clinically-relevant subtypes, or signatures, of aggressive prostate cancer challenging [e.g. 35, 99]. Currently known prostate cancer subtypes based on gene translocations [154, 192], gene expressions [106, 174], mutations [13, 14, 67, 182] or oncogenic signatures [133, 194] are fragmented and inconsistent. Identifying clinically-relevant subtypes is particularly problematic as disease recurrence generally only occurs many years after initial treatment, necessitating long follow-up of patients. Therefore, while subtypes that are informative of prognosis and treatment response have been proposed [e.g. 100, 227, 237], these are often weakly evidenced and are not used in the clinic. As such, there is a clear unmet need for a comprehensive, clinically-relevant subtyping schema.

The development of comprehensive subtyping schemas in diseases as heterogeneous as prostate cancer requires large datasets sourced from multiple sites. Such resources do exist and could facilitate thorough subtyping.

One such dataset is the Prostate Adenocarcinoma dataset from The Cancer Genome Atlas program (TCGA) [189]. While the TCGA datasets are an invaluable resource, and have commonly been used for cancer subtyping [e.g. 29, 84, 208, 221], they remain limited in some crucial aspects. For instance, although the number of samples (494) in the prostate cancer dataset is large compared to most cancer studies, the data was mainly generated from patients in the US, the vast majority of samples came from patients with indolent prostate cancer, the clinical and demographic data is largely incomplete, and the patients have short follow-up times. As such, the TCGA data is more often used for method development or validation rather than exploratory analysis. Recently, an international consortium known as the Pan-Prostate Cancer Group (PPCG) was set up to compile and curate an extensive multiomic dataset that addresses these issues. The data consists of whole genome sequencing (WGS), transcriptome and methylome data from a global cohort of approximately 1600 men with prostate cancer. This dataset therefore provides an unprecedented basis on which to derive multiomic subtypes. Unfortunately, current methodologies for analysing multiomic datasets like this exhibit numerous limitations, including, but not limited to, the lack of interpretability and the restricted support for missing data.

Integrative analysis of multiomic data is challenging due to their dimensionality, noisiness, complexity and heterogeneity among different omics outputs [156]. Existing methods therefore often rely on analysing each omics separately and consolidating the results [e.g. 188], likely leading to inconsistent conclusions [201]. Other approaches are based on identifying cluster-relevant features via integrative clustering [137, 138, 172] and commonly produce results that are not amenable for further analysis. Alternatively, methods based on machine learning involve the extraction of latent features

from the data, followed by subsequent analysis conducted on the compressed latent representations [e.g. 29, 84, 208, 213, 221, 230, 234]. The latent features themselves represent underlying variables that can be inferred, but not directly observed, from data. As such, latent features capture the hidden signals and regularities in data, which, together with their intrinsic dimensionality reduction, can prove invaluable for subtyping, i.e. clustering. While this appears promising, deep learning techniques such as Autoencoders [165] or their variational extensions [103] are most commonly used for the extraction of latent features, resulting in a complex latent space that is difficult to interpret. Moreover, current methodologies, machine-learning based or not, frequently prove inadequate for the analysis of multi-modal tabular data wherein entire modalities are absent for certain patients - a frequent occurrence within real-world cancer datasets. The above issues restrict the immense potential of using datasets like PPCG for comprehensive cancer subtyping and justify the need for the development of approaches capable of overcoming the aforementioned limitations.

In this thesis, we propose a bespoke, fully interpretable latent feature model designed specifically for multiomic data analysis. More precisely, we investigate how generative adversarial learning can be applied for the extraction of latent features effective for cancer subtyping. Adversarial learning techniques, in particular Generative Adversarial Networks (GANs) [65], have been impressively successful in generating realistic images [e.g. 89, 102, 231], suggesting the ability to learn complex patterns, regularities, and latent structures in the data. In spite of this, the application of such methods as latent feature models for multi-modal tabular data remains largely unexplored. Here, we hypothesise that such techniques can enable the extraction of meaningful latent features capturing low-level synergies and subtle rela-

tionships in multiomic data. Furthermore, we anticipate that the generative capabilities inherent in GANs will prove invaluable in addressing challenges posed by absent modalities and the potential need for data imputation.

There are also notable challenges arising from the data itself that our proposed model will need to address. We therefore identified several requirements that we have deemed necessary for its application domain. First, the model should be capable of integrating data from multiple sources (modalities), characterized by different biases, scales and data types. The latent features extracted from the model should encapsulate relationships both within and between data from different sources. Second, the model should demonstrate proficiency in addressing scenarios wherein large amounts of data are missing, for example, when not all data sources are available for all patients. Furthermore, the model should preserve interpretable links to the underlying biology so that the rationale behind its outputs can be understood by clinicians and patients. Finally, the model should have the ability to guide subtype discovery toward those that are clinically-relevant.

Further to devising a methodology that fulfils the aforementioned criteria, we will use the proposed method to provide broader insights into prostate cancer subtypes. As such, once developed and validated, the proposed method will be applied to extract latent features from the PPCG prostate cancer dataset. Patient data represented in the form of latent features will then be used for comprehensive downstream analysis, including subtyping and survival modelling. With our focus on interpretability, any findings should be amenable for translation to clinical use, meaning potential improvements in prognostic accuracy, more informed clinical decision making and possible identification of therapeutic targets.

The structure of this thesis is as follows. In Chapter 1, we introduce

the datasets used in the DPhil, explain their format, and provide a comprehensive summary of the biological terminology used throughout this thesis. Chapter 2 serves as a review of the literature and methods most relevant to multiomic data integration, latent feature extraction, subtyping and Generative Adversarial Networks. In Chapter 3, we introduce a novel adversarially learned latent feature extraction model suitable for the analysis of multiomic cancer datasets with missing modalities, a method we will refer to as iCS-GAN (integrative Cancer Subtyping with Generative Adversarial Networks). Then, in Chapter 4, we thoroughly test and validate our proposed method on a number of synthetic and real-life datasets, other than the PPCG resource. In Chapter 5, we apply the now validated iCS-GAN to the PPCG dataset and present our results. Specifically, we uncover and characterize three multiomic prostate cancer subtypes, including an aggressive subtype characterized by the downregulation of the *ERG* and *TFF3* genes, which, to the best of our knowledge, has not been described in the literature before. Finally, in Chapter 6, we discuss and critically analyse the significance, implications and limitations of the deliverables of this DPhil, summarize our contributions to the field of multiomic data integration, and suggest areas for further research.

Chapter 1

Data

In this chapter, we describe the multiomic cancer datasets used in this thesis. Throughout this DPhil, we assume that multiomic cancer datasets follow a single uniform format which our proposed latent feature extraction model will support. Specifically, we assume that each multiomic dataset can be represented as a collection of two or more tables, with each table reflecting data sourced from a single ome (e.g. genome or methylome), or some summary multiomic measurements. In each tabular dataset, i.e. in each table, rows represent patient samples and columns represent specific features (i.e. observations of interest) describing these samples. Certain observations may be absent in certain rows. Furthermore, one-to-one index mapping between tables in a given dataset cannot be guaranteed, i.e. all considered omics sources may not be available for all patients. Finally, we assume that certain clinical variables and pertinent outcomes, such as survival variables and treatment response indicators, are available in addition to the aforementioned tabular data.

We begin the data chapter by providing, in Section 1.1, an overview of the biological concepts and terminology relevant to cancer. Subsequently,

we introduce the primary dataset studied in this thesis, the PanProstate Cancer Group (PPCG) prostate cancer dataset, in Section 1.2. However, to avoid pitfalls associated with generalisation, such as over-fitting and bias, we did not use the PPCG dataset during the model development phase. Instead, alternative multiomic cancer datasets were utilized for this specific purpose. In particular, we selected five multiomic cancer datasets from The Cancer Genome Atlas Program (TCGA), which we review in Section 1.3. Furthermore, multiple synthetic datasets, with distributions and data types resembling those in the PPCG dataset, were generated and utilized for additional testing of the proposed model. We introduce these synthetic datasets in Section 1.4.

1.1 Terminology

In this thesis we propose a method suitable for the integrative analysis of multiomic data, to allow for an improved identification of molecular subtypes and signatures of cancer. The term ‘multiomic data’ describes tabular data sourced from multiple ‘omes’, for example, genome, transcriptome, or methylome. In this section, we concisely explain those terms and their relevance to cancer. The central dogma of molecular biology [39] states that DNA is transcribed into RNA, which in turn is translated into chains of amino acids that form a protein; we follow this hierarchy as we explain the various biological concepts and their relevance to the data used in this thesis.

1.1.1 DNA and the Genome

DNA, or deoxyribonucleic acid, is a hereditary material that holds the genetic instructions for building and maintaining an organism. DNA itself consists of

four chemical bases — adenine (A), cytosine (C), guanine (G), and thymine (T). These base sequences are organized into two adjacent strands where each base on one strand pairs with a complementary base on the opposite strand: adenine with thymine and cytosine with guanine. The DNA is packaged into discrete structures known as chromosomes. The human DNA sequence (*genome*) contains approximately 3 billion base pairs arranged into 23 pairs of chromosomes. It is the sequence of these bases that encodes the information needed to build and maintain an organism.

DNA has the crucial ability to be replicated. Replication happens when cells divide, so that new cells will contain copies of the DNA present in the previous generation of cells. However, errors during cell division can occur, leading to changes in the DNA sequence, often referred to as *mutations* or *genetic alterations*. These genetic alterations can also be caused by environmental and lifestyle factors, e.g. smoking or exposure to radiation, or as a result of viral infections. Genetic alterations can take many forms depending on the specific change to the DNA, including single nucleotide variants (SNVs), insertions or deletions (indels), copy number alterations (CNAs), structural variants (SVs) or gene fusions. Mutations are fundamental in the development of cancer as they can alter cellular function through their impact on gene expression and corresponding protein behaviour [5].

Single nucleotide variants (SNVs) are the simplest form of mutation, as they involve a single base change (e.g. from A to T). This creates a *mismatch* with the base on the complementary strand, and so the SNV is usually corrected via the *mismatch repair* machinery of the cell. Sometimes this machinery fails to correct the SNV before cell division occurs and so DNA replication copies this error into the next generation of cells. It should also be noted that some cancers are actually *mismatch repair deficient* (MMRd)

[e.g. 110], meaning that the machinery itself is dysfunctional, which results in large numbers of SNVs. How SNVs affect the function of the gene depends on where they occur. For instance, *nonsense* mutations will stop the gene from being transcribed into RNA, resulting in loss-of-function for that gene, while *missense* mutations actually lead to a structural change in the resulting protein, which may bestow additional functionality and change the behaviour of that gene. Finally, *silent* mutations do not affect the resulting protein and so cell function is usually maintained. There can also be SNVs that affect DNA that lies outside the region that becomes transcribed into RNA (the *reading frame*) - these are called *non-coding* mutations. Most of these do not effect the gene expression process, but some can affect how the gene is transcribed into RNA.

Insertions/deletions (indels) are similar to SNVs but involve the addition (insertion) or loss (deletion) of one or more base(s) in the DNA sequence of a gene. As bases in RNA are translated to amino acids in sequence in groups of three (triplets), indels generally still enable the production of RNA but can greatly affect the resulting protein. If the indel base length is a multiple of three, then the protein will usually still be produced as before but including extra or missing amino acids. These are called *in-frame mutations*. Conversely, if the indel base length is not a multiple of three then this leads to a *frameshift mutation* in which the amino acids from the position of the indel are translated to a completely random amino acid sequence that usually leads to a non-functional protein. Most indels are of this type. Similarly to SNVs, indels can occur outside of the reading frame, in which case they usually do not have an effect.

Copy number alterations (CNAs) are similar to indels in that they describe addition or loss of a number of bases in the DNA sequence. However,

the size of the affected region is much larger in CNAs, consisting of DNA segments spanning multiple genes up to entire chromosomes. In healthy *diploid* cells, each chromosome pair contains two copies of DNA covering the same genes, with each copy inherited from one parent. As each copy is different, this state is referred to as *heterozygous*. There are three main types of CNAs:

1. Loss of heterozygosity (LOH), in which one copy of a DNA segment is lost, leaving one copy remaining.
2. Homozygous deletion (HD), in which both copies of a DNA segment are lost so there are no copies remaining.
3. Gains (GAIN), where extra copies of the DNA segment are present. In this thesis we do not distinguish whether one or more copies are gained.

In this thesis we describe each CNA by the type followed by the chromosome and location (in bases) of the region affected. For instance LOH.17.7.571.739-590.808, abbreviated as LOH.17.571-590, describes a LOH event on chromosome 17 covering the *TP53* gene¹.

Structural variants (SVs), involve large genomic rearrangements such as deletions, duplications, inversions, and translocations. Deletions result in the loss of a segment of a DNA, duplications create multiple copies of a given DNA segment, inversions cause a segment of a chromosome to be inverted end-to-end and translocations involve the exchange of a DNA segment between chromosomes. In addition, cells can undergo what are called *complex SVs*, such as *chromoplexy* and *chromothripsis*. Chromoplexy is a series of interdependent rearrangements that occur in one event, typically a series of deletions and translocations across several chromosomes. Chromothripsis occurs when a section of a chromosome shatters into fragments that are then

¹in coordinates relative to the hg19 reference genome

reassembled in a haphazard fashion, leading to SVs of all types affecting the chromosomal region.

Gene fusions are a type of mutation in which two genes join together to create a single hybrid gene. For instance, in prostate cancer, there is a frequent gene fusion of the *TMPRSS2* and *ERG* genes [193]. These genes are adjacent on chromosome 21 and this gene fusion usually results from an LOH affecting most of the *TMPRSS2* gene and the intragenic region up to the start of the *ERG* gene. This results in the promoter (a region of DNA preceding the transcription start site (TSS) that regulates the expression of a gene) of the *TMPRSS2* gene becoming attached to the coding sequence of the *ERG* gene. As a result, the *ERG* gene becomes under the control of the *TMPRSS2* promoter, which is responsive to Androgen Receptor (AR), the protein that enacts a response to male hormones (androgens). This leads to the overexpression of *ERG* in the cancer cells, which is a key driver of prostate cancer [2]. Gene fusions can also occur between distant genes, when they are more likely the result of a translocation SV.

Mutations can be broadly split into two categories:

1. Driver mutations: these directly contribute to cancer development and progression. They occur in *driver genes* that when affected confer some advantage to the growth and survival of the cancer cells. Driver genes themselves can be put into two categories: *tumour suppressors*, which generally regulate cellular processes and ensure homeostasis, and *oncogenes* that control cell growth and division. In cancer evolution it is common for cells to first undergo loss-of-function events affecting tumour suppressors, followed by gain-of-function events affecting oncogenes.
2. Passenger mutations: these do not contribute directly to cancer devel-

opment. They usually occur in intergenic regions or affect genes that aren't relevant to the increased proliferation of the cell. However, identifying passenger mutations is still useful as they provide information on the mutational processes that are in effect.

It is common to refer to driver and passenger mutations in the context of SNVs and indels, since identifying the drivers in larger genetic alterations is difficult to define as they may affect multiple genes with synergistic effects. However, it is standard to mention some known driver genes in a CNA region to aid interpretation of the effect of the CNA.

Mutations can be detected via DNA sequencing [169], a method that determines the exact order of bases in a sample. Often a reference genome is used for comparison, usually provided by non-tumour cells from the same person. As there are so many mutations that occur concurrently in cancer cells, it is common to interpret tumour status through a number of summary measurements, obtained with advanced bioinformatics algorithms and pipelines. For example, percentage genome altered (PGA) can be calculated to describe the proportion of the total genome affected by genetic alterations, providing insights into the complexity and aggressiveness of a tumour [81].

1.1.2 RNA and the Transcriptome

RNA, or ribonucleic acid, acts as the intermediary in the conversion from DNA instructions to functional proteins, which are essential for various cellular functions. There are three main types of RNA: messenger RNA (mRNA), ribosomal RNA (rRNA) and transfer RNA (tRNA). mRNA carries the information from DNA to ribosomes, where it acts as the template for proteins to be made. rRNA uses the instructions from mRNA and builds proteins by linking together the amino acids transported to the ribosomes by the tRNA.

For the purposes of this thesis, we restrict our focus to mRNA, as these molecules encode the proteins that drive the cancerous behaviour of tumours. While measuring protein levels directly would provide the most detailed insights into the functional aspects of the cellular processes, this is incredibly challenging in practice and so it is common to use mRNA as a proxy for protein behaviour. However, mRNA and protein levels are not always strongly associated [150] so the results should be interpreted under this caveat.

RNA expression is controlled by a number of factors that regulate transcription from DNA. The most important regulators are *transcription factors*, which are proteins that bind to specific DNA sequences in or around the gene and affect how it is expressed. It is sometimes useful to categorise transcription factors based on the origin of their activation:

- Intracellular transcription factors originate from within the cell and are often themselves created as part of a wider gene network with multiple component genes.
- Externally-activated transcription factors respond to signals from outside the cell. For example, the androgen receptor (AR) is a cell surface protein that, upon binding to an androgen hormone, is internalized into the cell cytoplasm. It then translocates to the nucleus, where it functions as a transcription factor.

As a result, DNA mutations can have wide-ranging effects on RNA expression. Mutations in intracellular transcription factors can disrupt entire gene networks within individual cells, leading to changes in the expression of multiple genes. Similarly, mutations in externally-activated transcription factors can cause uncoordinated gene expression patterns across different cells within a tissue or organ. Therefore, identifying *signatures* of co-expressed genes that

share common regulatory mechanisms is more effective than monitoring the expression levels of individual genes. Determining which RNA signatures correspond to specific DNA mutations remains a significant challenge in cancer research.

To study the levels of RNA in cancer tissue samples, it is common to use RNA sequencing (RNA-seq) [207]. This is a technology that can be used to study the entire set of RNA molecules, referred to as the *transcriptome*. This enables an estimate of gene expression levels, determining which genes are active, and how active they are, by estimating the number of RNA transcripts. As we can measure the entire transcriptome simultaneously, this technology enables the extraction of signatures that reflect disrupted biological processes and can therefore be used to define cancer subtypes.

1.1.3 Methylation and the Methylome

Methylation is a biochemical process in which a methyl group is added to the DNA molecule, typically on cytosine bases that are followed by guanine (often called the *CpG* context). This process regulates gene expression levels to allow for turning genes on and off without affecting the actual DNA sequence.

The *methylome* represents the set of DNA methylation events across the entire genome, and can be measured with various technologies, such as sequencing and micro-arrays [15, 158]. Abnormal methylation events can affect RNA expression in a similar way to DNA mutations, and can therefore also contribute to the cancer development process. For example, hypermethylation (increased methylation) can disable tumour suppressor genes, and conversely, hypomethylation (decreased methylation) can contribute to the activation of oncogenes. Notably, methylation is a reversible process. This means that specific enzymes can be utilized to remove methyl groups from

DNA, paving the way for targeted therapies to potentially reactivate silenced genes.

1.1.4 Other Omics Sources

The majority of the data considered in this thesis was sourced from either the genome, the transcriptome, the methylome, or any combination of the three. However, other omics sources exist, and could be integrated into our analysis, subject to data availability. For example, we could integrate the data from the *proteome*, which describes the complete set of proteins expressed by a cell, tissue, or organism, the *microbiome* which considers all the microorganisms, such as bacteria or viruses, living in and on the human body, or the *glycome* which focuses on sugar molecules attached to proteins and lipids. As these omic sources are not relevant to this thesis, we will not review them further.

1.1.5 Targeted Therapies

Understanding which alterations characterize a given cancer, or its subtype, is essential for the development of targeted therapies, that is, tailored treatments that attempt to modify the activity of specific molecules. For example, non-small cell lung cancer characterized by the *EGFR* driver gene mutation can be treated with drugs inhibiting the activity of the EGFR protein [159], which reduces tumour growth. Similarly, drugs like Herceptin [61] can target the RNA overexpressed HER2 protein, commonly occurring in the HER2-positive breast cancer, to slow down the growth of cancer cells. Finally, demethylating agents, that is drugs that reverse hypermethylation, can lead to the reactivation of silenced tumour suppressor genes, contributing to the death of cancerous cells.

1.2 PPCG Dataset

In this section, we introduce the main dataset of interest in this DPhil, namely the PanProstate Cancer Group prostate cancer dataset, and provide a brief overview of prostate cancer for context and to facilitate understanding of the data.

Prostate cancer is the most common cancer and the second most common cause of cancer deaths in men in the United Kingdom [22]. It is estimated that 1 in 8 men in the UK will get prostate cancer in their lifetime. Most prostate cancers are known as prostate adenocarcinomas and develop in the cells that line the glands within the prostate. The disease is more common in older men and other risk factors include ethnicity (prostate cancer affects Black men more often than White or Asian men), family history, obesity, and inherited faults or mutations in the *BRCA1* and *BRCA2* genes. Currently, prostate cancer is often curable, although it appears to be more aggressive for some patients, for whom the treatment ultimately fails. Distinguishing aggressive disease from more indolent types remains a significant challenge in the management of prostate cancer.

Several tests to detect prostate cancer are available at various stages of the diagnostic pathway, including digital rectal exams, PSA (Prostate-Specific Antigen) biomarker blood tests, and prostate biopsies. Treatment options depend on the severity of the disease. For localized (confined to the prostate) cancer, options may include active surveillance or radical (curative) treatments such as prostatectomy (surgical removal of the prostate) or radiotherapy. Chemotherapy and hormone therapy are typically reserved for very high-risk tumours or those that have already spread beyond the prostate, although these treatments are increasingly being used as neoadjuvant therapies (to enhance the effectiveness of the primary treatment approach). Clinical

decision-making is usually based on a small number of factors:

- TNM staging that describes the size and spread of the cancer:
 - tumour stage T1-T4 indicates the extent of the primary tumour, from very small (T1) to extensive spread beyond the prostate (T4),
 - node stage N0-N1 indicates whether cancer has spread to nearby lymph nodes (N1) or not (N0),
 - metastasis stage M0-M1 indicates whether cancer has metastasized to distant parts of the body (M1) or not (M0).
- The Gleason grade: cancerous glandular morphologies are assigned a score from 1-5 that describes the level of differentiation of cancerous glands in the prostate. The Gleason score is made up of two of values (X+Y), the first (X) describing the most prevalent (predominant) morphology and the second (Y) the less prevalent morphology. Broadly speaking,
 - scores 3+3 and lower indicate low risk disease,
 - scores 3+4 and 4+3 indicate intermediate risk disease,
 - scores 4+4 and higher indicate high risk disease.
- PSA tests measure the concentration of prostate specific antigen in the blood. A PSA level of less than 4ng/ml is considered normal; higher values indicate that further tests should be carried out. This threshold is somewhat arbitrary as baseline PSA levels differ between men.

After treatment or being assigned to active surveillance, patients are often monitored with PSA tests. If the prostate is removed then PSA usually drops to 0ng/ml; if PSA rises past 2ng/ml after this it is referred to as a

biochemical recurrence (BCR) and normally triggers an evaluation on progression to the next treatment stage, which might include hormone therapy or chemotherapy. However, BCR is considered to be an imperfect metric for clinically-significant disease as many men can experience BCR that would not go on to develop metastasis, leading to over-treatment. Confirmed metastatic disease is considered to be incurable, although advances in treatments often allow patients to survive for many years with good quality of life.

Although growing evidence shows that the molecular subtype of prostate cancer can be linked to its aggressiveness [e.g. 10, 99, 120, 222], to the best of our knowledge, a comprehensive and integrative multiomic prostate cancer subtyping schema is yet to be developed and incorporated in the clinical decision making procedures.

The PanProstate Cancer Group consortium was set up to enable comprehensive multiomic investigations into prostate cancer and advance the treatment of the disease. The consortium's goal is to harmonise the Whole Genome Sequence (WGS) data generated around the world with the associated transcriptome and methylome data. We have privileged access to the full unpublished PPCG dataset, although quality control issues prevented the methylation data from being suitable for analysis. For the remainder of this section, we review and visualise the available data, explain the sample and feature selection processes applied, and review some of the well-established prostate cancer subtyping schemas.

1.2.1 Sample Selection

The PPCG dataset contained 959 unique DNA samples and 1757 unique RNA samples from 1798 patients. For the DNA data, each of the 959 samples represented a different patient, and for RNA, some samples were taken

from the same patient at different points in time. Additionally, the RNA dataset contained benign samples obtained from matching controls. Both data sources included samples with potential contamination and samples with a large number of missing measurements, necessitating the need for a careful sample selection process. Furthermore, as we wanted to focus our analysis on subtypes of localised disease to ultimately inform treatment decisions at the curative stage, we decided to omit any patients that presented with disease that had already metastasised at first diagnosis. With the help of consortium members involved with this dataset, we created an inclusion-exclusion list based on the following criteria:

- samples marked as ‘blacklist’, either due to sample contamination, or sample mismatch, were excluded from the dataset,
- samples initially presenting with metastatic diseases and samples pre-treated before sample collection were excluded from the dataset,
- only non-metastatic tumour samples were used in further analysis,
- samples with 30% or more missing values for a given modality were excluded for that modality only - the RNA data had no missing values, and as such this criterion was applied solely to the DNA dataset, resulting in the removal of fewer than 4% of all samples, with the threshold of 30% chosen to ensure both the reliability of the analysis and the minimal loss of valuable data,
- samples not present in the reference sheet, a PPCG file used for sample identification, were excluded as their origin could not have been verified,
- where multiple RNA samples were available for a given patient, only the first record was included in further analysis - as there was no systematic

ordering applied to sample labels, this is equivalent to selecting a sample at random; as samples from the same patient are biologically similar, selecting one record suffices to represent the genomic alterations present and avoids biasing the analysis with multiple measurements from the same individual,

- outliers in the RNA dataset were excluded from further analysis.

With these criteria, samples from 1627 prostate cancer patients were selected for further analysis. Of these, 857 had DNA data available, and 1235 had RNA data available. A complete set of observations, i.e. both DNA and RNA measurements, was available for 465 patients only.

1.2.2 Clinical Measurements

The PPCG dataset contains a range of clinical and demographic measurements, including observations such as age, country, Gleason grade, PSA at tumour collection and survival indicators. For the survival indicators, two variables are of specific interest to us: Relapse-Free Survival and Metastasis-Free Survival. Relapse-Free Survival (RFS) measures the length of time after primary treatment during which a patient remains free from either BCR or metastatic disease. Metastasis-Free Survival (MFS) measures the length of time after primary treatment during which a patient remains free from the development of metastatic disease only. Complete RFS and MFS data was available for 961 and 749 of the retained PPCG samples, respectively. Selected clinical and demographic measurements, as well as the RFS and MFS survival outcomes are visualised in Figure 1.1.

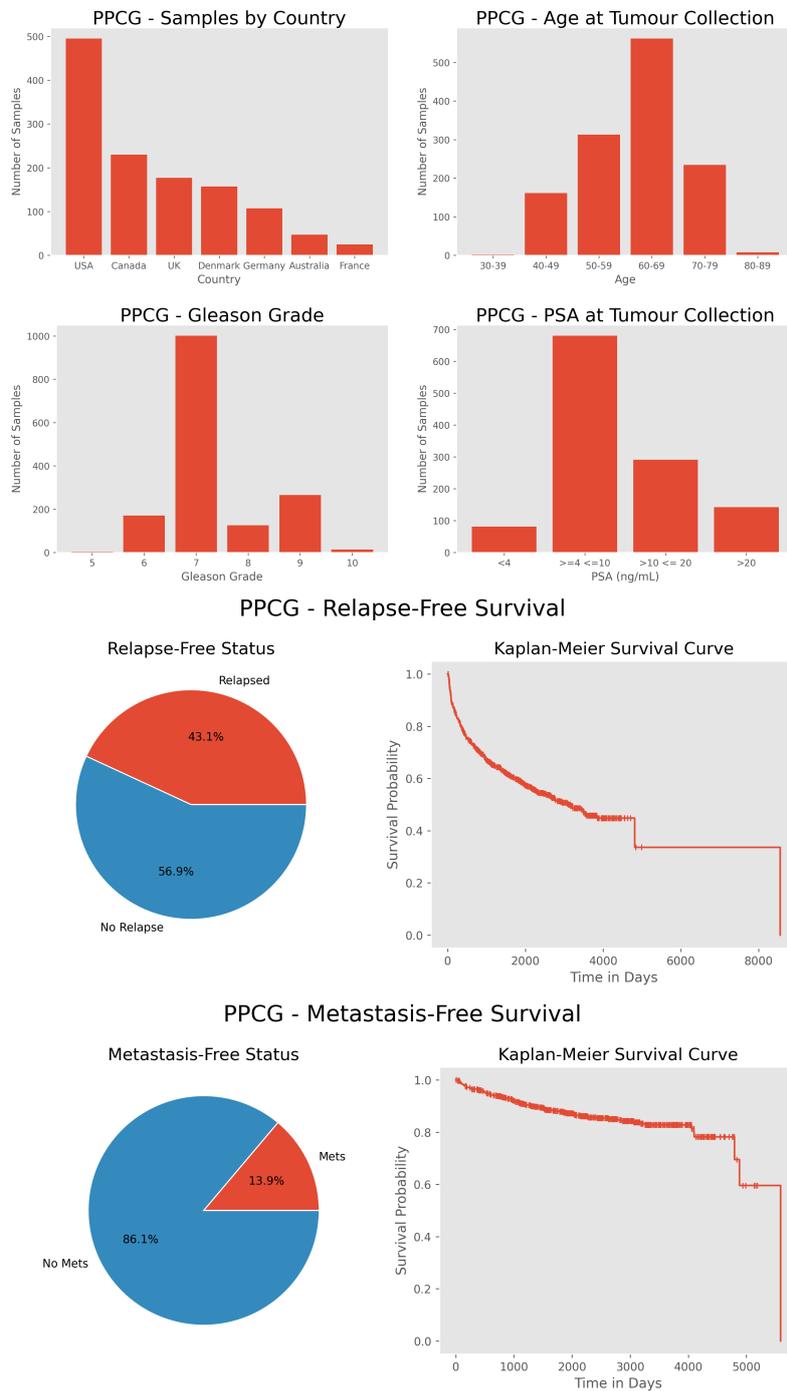


Figure 1.1: Bar plots visualising the number of PPCG samples by country, age, Gleason grade and PSA at tumour collection, summary of the relapse-free and metastasis-free survival status with the associated Kaplan-Meier survival curves (left to right, top to bottom). Note that the visualised measurements were not available for all PPCG samples.

1.2.3 DNA Data

As the raw WGS output is not amenable to direct analysis, the PPCG consortium applied a number of bioinformatics tools to generate 119 measurements that related to phenomena observed in the WGS, similar to the approach developed in our previous publication [213] that was based on a part of the UK subset of the dataset. These 119 features can be viewed as belonging to one of 3 distinct classes of measurements: summary variables, driver gene mutations and copy number alterations. As these each have their own analytical challenges, we treat them as separate ‘modalities’. Next, we provide an overview of these, with some context on how they relate to prostate cancer.

Summary Measurements

The DNA dataset contained 34 summary measurements generated from the WGS data using a number of previously published algorithms. The summary features consisted of multiple data types including continuous, count and binary measurements. Figure 1.2 visualises the summary data. Below we describe the specific summary measurements in the PPCG DNA dataset, providing feature names as used in Figure 1.2 in brackets:

- **Numbers of SNVs, Indels and Structural Variants (10 Features)** - the measurements represent the total numbers of SNVs (SNVs), indels (Indels) and rearrangements (Rearrangements) per sample, the total number numbers of the three types of indels (insertions (Indels: Insertions), deletions (Indels: Deletions) and complex (Indels: Complex)) and the total numbers of the 4 types of SVs (duplications (SVs: Duplications), deletions (SVs: Deletions), inversions (SVs: Inversions) and translocations (SVs: Translocations)). The measurements were

created using the Cancer Genome Project Wellcome Trust Sanger Institute pipeline described in [38]. Higher SV counts have previously been associated with biochemical recurrence of prostate cancer [213].

- **Percentage Genome Altered (3 Features)** - the measurements represent the percent total of the genome affected by CNAs (PGA (Total)), including the percentage affected by the clonal (PGA (Clonal)) and subclonal (PGA (Subclonal)) CNAs [81]. CNA burden across the genome was previously associated with prostate cancer biochemical recurrence and metastasis [81].
- **Ploidy (1 Feature)** - the measurement (Ploidy) identifies, using the procedure described in [210], samples with average ploidy (number of sets of chromosomes) greater than 3. The combination of DNA ploidy status and the *PTEN/6q15* deletions has been shown to be a predictor of poor patient outcomes [115].
- **Kataegis (1 Feature)** - kataegis, a term used to describe the presence of small DNA regions with a large number of highly patterned base pair mutations, was identified using <https://github.com/cran/SeqKat> (Kataegis). In African men, kataegis-associated mutational processes have been linked to adverse prostate cancer presentation [79].
- **ETS Status (1 Feature)** - the measurement (ETS Status) identifies an in-frame ETS fusion. The ETS fusion describes a DNA breakpoint involving *ERG*, *ETV1*, *ETV3*, *ETV4*, *ETV5*, *ETV6*, *ELK4*, or *FLI1* gene and partner DNA sequences. The relationship between ETS fusions and prostate cancer prognosis is still not fully understood with studies showing contradictory results [51, 66, 142, 154, 202].

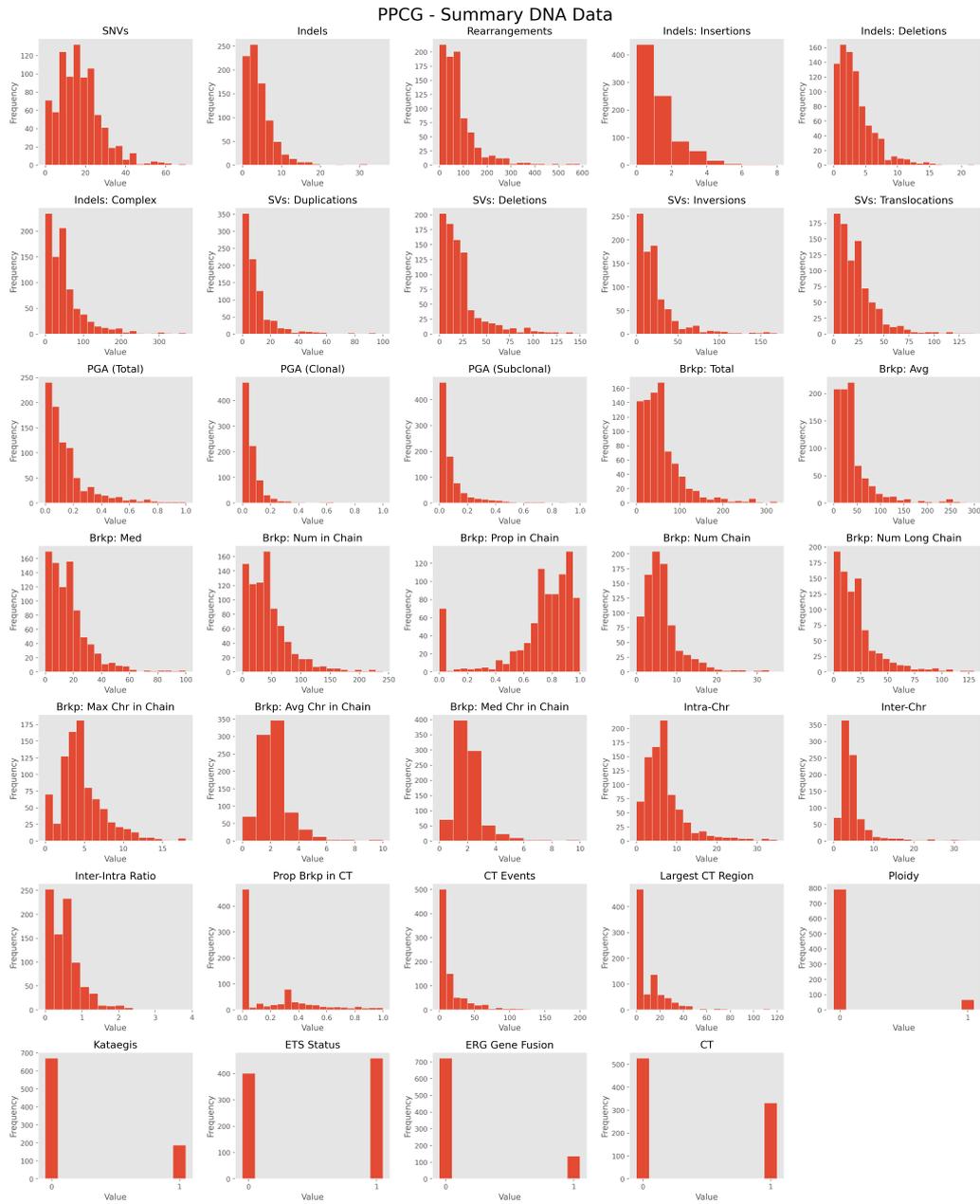


Figure 1.2: Frequency histograms of the summary measurements available in the PPCG DNA dataset. The features include heterogeneous data types, such as continuous variables (e.g. PGA), count data (e.g. SNVs), and binary measurements (e.g. Ploidy). Additionally, heavily skewed distributions can be observed. Potential outliers were removed for the clarity of this visualisation.

- **ERG Gene Fusion (1 Feature)** - the measurement (ERG Gene Fusion) identifies an in-frame gene fusion affecting the *TMPRSS2 / ERG* genes. The *ERG* gene is the part of the ETS family and the *TMPRSS2 / ERG* gene fusions are most commonly considered in the studies concerning ETS and prostate cancer [51, 66, 142, 154, 202].
- **Breakpoints (13 Features)** - the term DNA breakpoint refers to locations in the DNA sequence in which a double strand break (DSB) has occurred, resulting in a SV, CNA or gene fusion. For the PPCG dataset, breakpoints were identified with the ChainFinder algorithm [13] that can detect chains of linked SVs that occur in interdependent events (chromoplexy) and therefore the breakpoints associated with these. The specific measurements describe the total (Brkp: Total), average (Brkp: Avg) and median (Brkp: Med) numbers of breakpoints, the total number (Brkp: Num in Chain) and proportion (Brkp: Prop in Chain) of breakpoints in chains, the number of chains (Brkp: Num Chain), the number of breakpoints in the longest chain (Brkp: Num Long Chain), the maximum (Brkp: Max Chr in Chain), average (Brkp: Avg Chr in Chain) and median (Brkp: Med Chr in Chain) numbers of chromosomes involved in a chain, the number of intra- (Intra-Chr) and inter-chromosomal (Inter-Chr) events and the inter-chromosomal to intra-chromosomal ratio (Inter-Intra Ratio). DNA breakpoint burden has previously been associated with the biochemical recurrence of prostate cancer [213].
- **Chromothripsis (4 Features)** - chromothripsis describes the shattering and uncoordinated reassembly of a region of DNA in a chromosome. In the PPCG dataset, a chromothripsis region was defined as a

high density breakpoint region with more than 15 copy number breakpoints. The specific measurements describe the presence or absence of chromothripsis (CT), the proportion of all breakpoints in chromothripsis events (Prop Brkp in CT), the number of chromothripsis events in each sample (CT Events) and the size of the largest chromothripsis region (Largest CT Region). Chromothripsis has previously been associated with the biochemical recurrence of prostate cancer [213] and found to occur early in tumour progression [105, 171].

The summary dataset contained missing values, which we mean imputed after the train-test split, rounding to the nearest integer value for non-continuous data types. The only exception was the binary chromothripsis feature, imputed as 0 if missing, to indicate that chromothripsis was not identified.

Driver Genes

In addition to the summary measurements described previously, the PPCG DNA dataset contained 25 binary features representing mutations in driver genes, identified in Wedge et al. [210]. Figure 1.3 visualises the mutational landscape of the PPCG dataset. Notable genes include the tumour suppressor *TP53* gene, which was initially associated with metastatic prostate cancer and castration resistance, with more recent studies detecting it in primary prostate cancer and correlating with worse patient outcomes (see the review in Chapter 8 of [186]). Other common mutations involve the *SPOP* gene, known to be associated with best prostate cancer outcomes if no other mutations are present [141], the *PTEN* gene correlated with adverse oncological outcomes [93], or the *FOXA1* gene, thought to be a predictor of prostate cancer recurrence and a potential therapeutic target [180]. The drivers data contained no missing values.

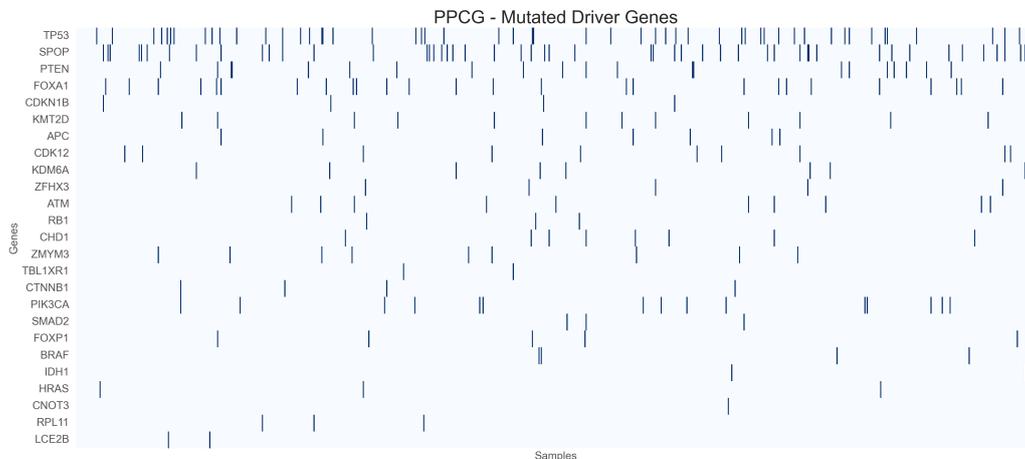


Figure 1.3: Mutational landscape of the PPCG dataset. Heatmap rows represent driver genes, columns represent patient samples. Dark mark indicates that a given driver gene is mutated for a given patient. The driver gene matrix consists of relatively sparse binary variables, with each mutation affecting between 0.12% and 6.07% of the patient population.

Copy Number Alterations

The CNA part of the PPCG DNA dataset contained 60 binary features identifying consistently aberrant regions, separately for CNA losses of heterozygosity (37 fields), gains (11 fields) and homozygous deletions (12 fields) [210]. Figures 1.4 - 1.6 visualise the CNA landscapes of the PPCG dataset. Copy number alterations play a vital role in prostate cancer disease evolution. Studies have shown their association with metastatic-lethal disease progression [206] or disease recurrence and tumour aggressiveness [81, 168].

Within the PPCG CNA dataset, 36 samples failed quality checks. As such, we have decided to remove these samples from the CNA dataset only (i.e. treat them as missing modalities) but keep the corresponding samples in the summary and drivers datasets. This resulted in the CNA dataset containing 821 instead of 857 samples. As obtained, the CNA dataset contained no missing values.

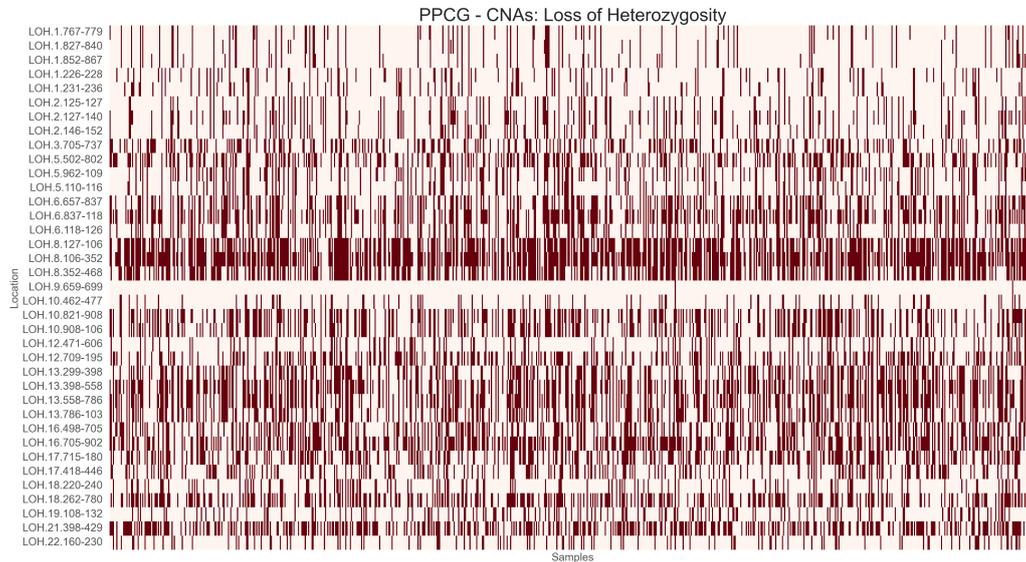


Figure 1.4: CNA LOH landscape of the PPCG dataset. Heatmap rows represent regions, columns represent patient samples. Dark mark indicates that a given region is affected by an LOH event for a given patient. The LOH matrix comprises binary variables with widely varying activation probabilities, with each LOH event affecting between 0.24% and 70.52% of the patient population.

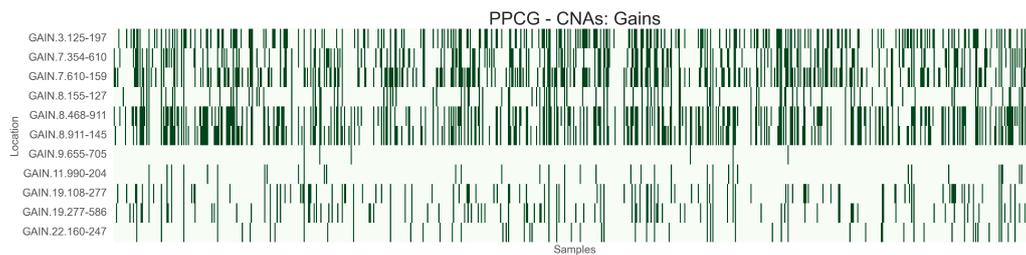


Figure 1.5: CNA gains landscape of the PPCG dataset. Heatmap rows represent regions, columns represent patient samples. Dark mark indicates that a given region is affected by a gain event for a given patient. The gains matrix comprises binary variables with widely varying activation probabilities, with each gain event affecting between 0.73% and 37.39% of the patient population.

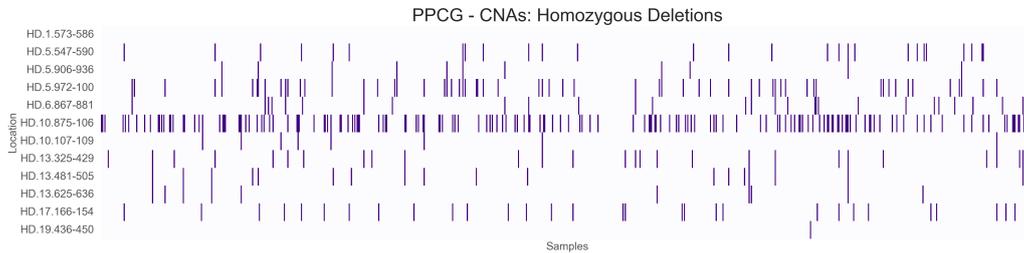


Figure 1.6: CNA HD landscape of the PPCG dataset. Heatmap rows represent regions, columns represent patient samples. Dark mark indicates that a given region is affected by an HD event for a given patient. The HD matrix consists of relatively sparse binary variables, with each event affecting at most 19.73% of the patient population.

1.2.4 RNA Data

Upon sample selection, the PPCG RNA-seq dataset contained 1235 patients represented by 34265 features, each describing gene-level RNA expression. Upon the removal of measurements representing mitochondrial genes, 34228 features remained. The RNA dataset as obtained was already normalized, using the RUV-III method [139], with no missing values present. In line with standard RNA-seq analysis approaches, we performed feature selection to retain only 1000 features with highest median absolute deviation, as measured on the training set after the train-test split. Median absolute deviation was used instead of the more common mean absolute deviation due to its robustness to outliers. Figure 1.7 shows frequency histograms of selected RNA features. A number of links between RNA expressions and prostate cancer has been described in the past. For example, *TFF3* and *PCAT5* genes have been identified as potential prostate cancer therapeutic targets [121, 225].

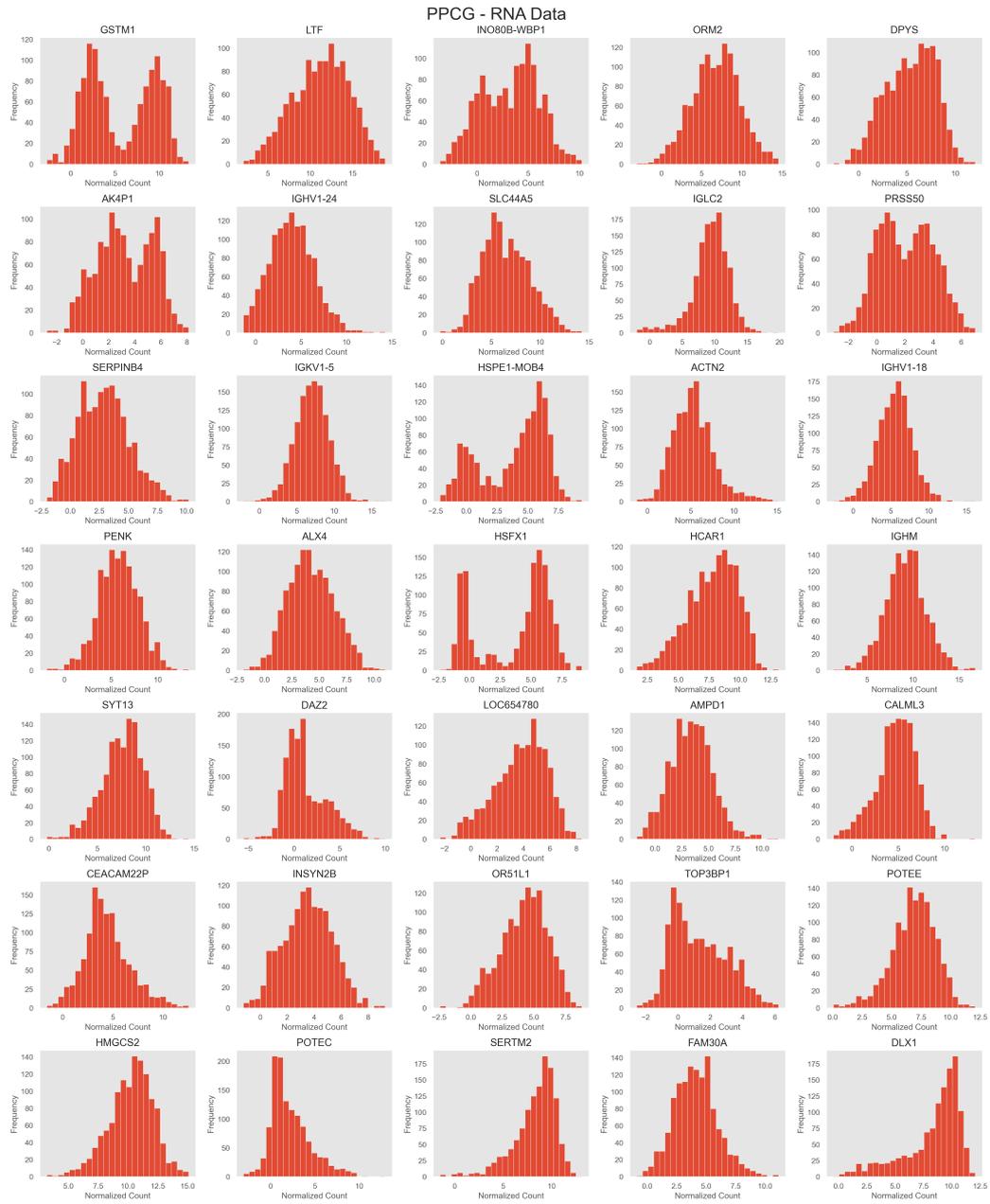


Figure 1.7: Frequency histograms of selected RNA-seq measurements available in the PPCG RNA dataset. In contrast to the distributions of the summary measurement features, the RNA distributions generally exhibit an approximately normal shape but show signs of bimodality and skewness, suggesting the presence of underlying subpopulations and asymmetry.

1.2.5 The Final Dataset

Upon the completion of the sample and feature selection processes, and the removal of samples from the CNA dataset that failed quality control, the finalized PPCG dataset contained the following 4 sub-datasets, treated as separate modalities:

- **DNA summary data** - 857 patients, 34 mixed-type measurements,
- **DNA driver data** - 857 patients, 25 binary measurements,
- **DNA CNA data** - 821 patients, 60 binary measurements,
- **RNA data** - 1235 patients, 1000 continuous features,

together with the associated RFS and MFS variables, available for 961 and 749 patients, respectively. The 4 sub-datasets accounted for total number of 1627 unique patient samples. Of these samples, 453 patients had all modalities available, 368 samples had complete DNA and no RNA data, 24 samples had summary and driver data only, 12 had only CNA data missing and 770 samples had only RNA data available. A Venn diagram illustrating the numbers of samples in the PPCG dataset by available modality is given in Figure 1.8.

For further analysis, following standard machine learning practices, we split the entire dataset into independent training (80% of all samples) and test (20% of all samples) sets, stratifying the split by available modalities, sample origin (country) and the RFS and MFS survival indicators. The purpose of performing the train-test split was to ensure that any biological findings discovered during our analysis would remain valid when applied to independent, previously unseen data. The non-binary features were then normalized and min-max scaled to the $[0, 1]$ range.

PPCG Samples - Venn Diagram

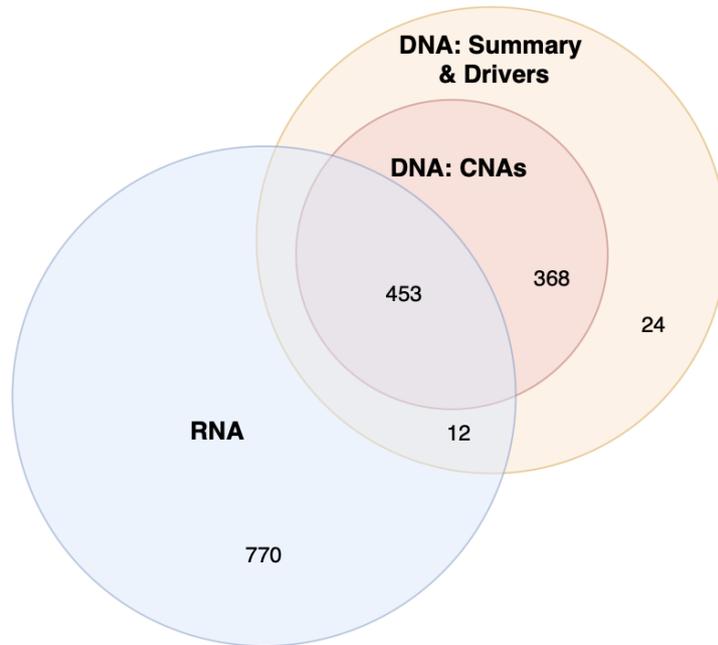


Figure 1.8: Venn diagram visualising the numbers of samples in the PPCG dataset by available modality.

1.2.6 Prostate Cancer Subtyping Frameworks

Multiple prostate cancer subtyping frameworks have been proposed, however none have shown sufficient clinical utility to be adopted into the current care pathway. We review the most notable ones below.

DESNT

The DESNT [125] (Desmoplastic-like, Epithelial, and Stromal, with Neuroendocrine and Tumour cell plasticity features) subtyping schema identified, based on transcriptome profiling, a poor prognosis prostate cancer category characterized by distinct gene expression patterns, elevated levels of neuroendocrine differentiation-related genes, epithelial-to-mesenchymal transition (EMT) and stromal remodelling. Neuroendocrine-differentiated can-

cer cells exhibit characteristics of hormone-secreting neuroendocrine cells resulting in higher likelihood of metastasis and resistance to therapies. EMT transforms epithelial cells into mesenchymal ones, which are more invasive and resistant to apoptosis, contributing to cancer metastasis. Stromal remodelling involves changes that occur in the tissues surrounding a tumour, contributing to its progression. DESNT prostate cancers tend to be more aggressive and exhibit poor outcomes relative to other patients [125].

You et al.

You et al. [227] classified prostate cancers, based on transcriptome profiling, into 3 distinct subtypes PCS1, PCS2 and PCS3: subtypes PCS1 and PCS2 represent luminal subtypes, and PCS3 reflects a basal subtype. The terms ‘luminal’ and ‘basal’ refer to two types of epithelial cells in the prostate. Luminal cells line the inside of the prostate glands, facing the lumen (hollow space in the gland). Basal cells are located in a layer underneath the luminal cells, closest to the other prostate tissue. The authors designed a clinically-relevant 37-gene panel to accurately classify tumours to one of the three PCS subtypes [227]. PCS1 tumours have been shown to progress more rapidly to metastatic disease than PCS2 or PCS3 tumours [227].

Evotypes

The Evotypes [213] prostate cancer classification schema revealed the existence of two distinct prostate cancer subtypes, Canonical and Alternative, arising from divergent evolutionary trajectories. The divergence is related to the dysregulation of androgen receptor (AR) signalling, which itself can be caused by a number of factors. Therefore, classifying by the evolutionary trajectory of the disease rather than the presence and absence of specific

genetic anomalies is called the evolutionary subtype, or evotype. Each evotype is characterized by a different propensity of certain genomic aberrations, with no single aberration being necessary or sufficient for assignment to either evotype [213]. Specifically, the Canonical evotype exhibits greater likelihood of LOH:17p, LOH:19p (*TP53*), LOH:21q (*ERG*), *ETS* gene mutations, and higher inter- to intra-chromosomal breakpoint ratio. The Alternative evotype is characterized by higher likelihood of LOH:1q, LOH:2.q, LOH:5q (*IL6ST*, *PDE4D*, *CHD1*), LOH:6q (*MAP3K7*, *ZNF292*), LOH13q (*BRCA2*, *RB1*, *EDNRB*), GAIN:3q, GAIN:7, GAIN:8 (*MYC*), SPOP mutations, kataegis, ploidy, chromothripsis and higher percentage genome altered by clonal CNAs [213]. The Alternative evotype was associated with worse patient prognosis with respect to biochemical disease recurrence [213].

1.3 TCGA Datasets

In this section, we introduce datasets from The Cancer Genome Atlas [189] utilized for the validation of our proposed latent feature model. While our main objective was to analyse the PPCG dataset introduced in the previous section, to avoid pitfalls associated with generalisation, over-fitting and bias, we purposefully did not use this dataset during the model development phase and instead utilized other similar multiomic cancer datasets, specifically:

- Breast Invasive Carcinoma (BRCA),
- Kidney Renal Clear Cell Carcinoma (KIRC),
- Bladder Urothelial Carcinoma (BLCA),
- Colorectal Adenocarcinoma (COAD),

- Head and Neck Squamous Cell Carcinoma (HNSC).

TCGA datasets are used throughout this thesis for the purpose of model validation and ablation analysis, rather than to discover new subtypes of the breast, kidney, bladder, colorectal or head and neck cancers. Therefore, we review them in lesser detail than the PPCG dataset.

All TCGA datasets used for evaluating our model were downloaded from the cBioPortal for Cancer Genomics [25], and the PanCancer Atlas version of each dataset was used. For each dataset, we selected 3 modalities for integration: methylation, mRNA and CNA. For methylation, we used the methylation between platforms (hm27 and hm450) normalization values. For mRNA, we used expression z-scores of tumour samples compared to the expression distribution of all log-transformed mRNA expression of adjacent normal samples in the cohort. Finally, for CNAs, we used the putative arm-level copy-number from GISTIC 2.0. We selected the ‘Progression Free Status’ and ‘Progress Free Survival (Months)’ clinical variables as our survival target variable. Progression-Free Survival (PFS) measures the length of time a patient lives with cancer without the disease getting more severe or progressing.

For each dataset, data pre-processing involved excluding all samples for which the selected clinical variables were unavailable, or where the survival time was given as 0. For mRNA data, we dropped all genes with missing ‘Hugo_Symbol’ (standardized nomenclature for human genes) and removed duplicates by keeping only the first record with a given ‘Hugo_Symbol’. Additionally, for each modality, we removed all measurements with more than 20% missing values. For further analysis, we selected only the samples for which complete data (i.e. all 3 modalities) was available, to later allow for tests involving ‘masked’ modalities. Additionally, a number of outlier sam-

ples was removed from the KIRC dataset.

Each dataset was then split into independent training (80%) and test (20%) sets. For the continuous modalities (methylation and mRNA) missing values were mean imputed and feature selection that involved selecting the 250 most variable genes, as measured by median absolute deviation, was applied. For each CNA measurement, given as either ‘Gain’, ‘Loss’ or ‘Unchanged’ values for each chromosome arm, we created two binary features, the first one representing gains (1 - gain, 0 - otherwise) and the second one representing losses (1 - loss, 0 - otherwise). We then imputed the missing CNA values with the most common value for each feature and dropped all measurements with less than 5% aberration rate. Finally, we normalized and min-max scaled to the $[0, 1]$ range all continuous features.

Table 1.1 summarises the number of samples and features in each dataset, after pre-processing. Figure 1.9 visualises the progression-free survival for each dataset. Distributions of selected normalized observed features are shown in Figure 1.10.

Dataset	#Samples	#Methylation Features	#mRNA Features	#CNA Features
BRCA	1050	250	250	43
KIRC	478	250	250	37
BLCA	406	250	250	35
COAD	561	250	250	59
HNSC	514	250	250	50

Table 1.1: Summary of the TCGA datasets.

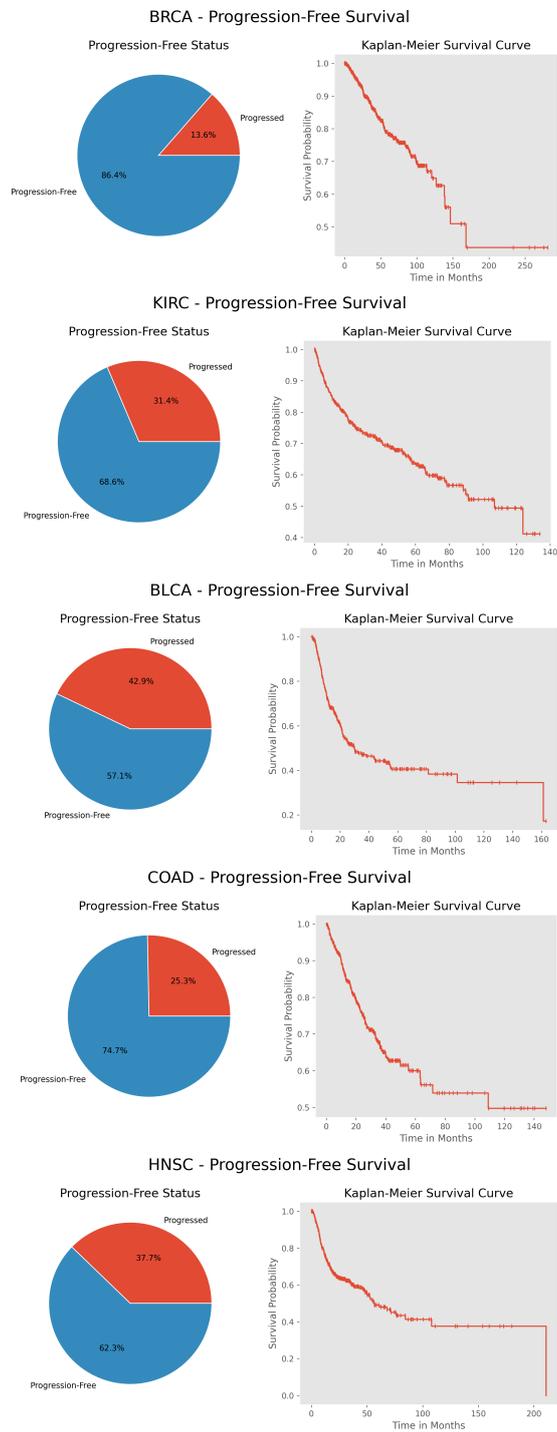


Figure 1.9: Summary of the progression-free survival status (left) and the associated Kaplan-Meier survival curves (right) for BRCA, KIRC, BLCA, COAD and HNSC datasets (top to bottom).

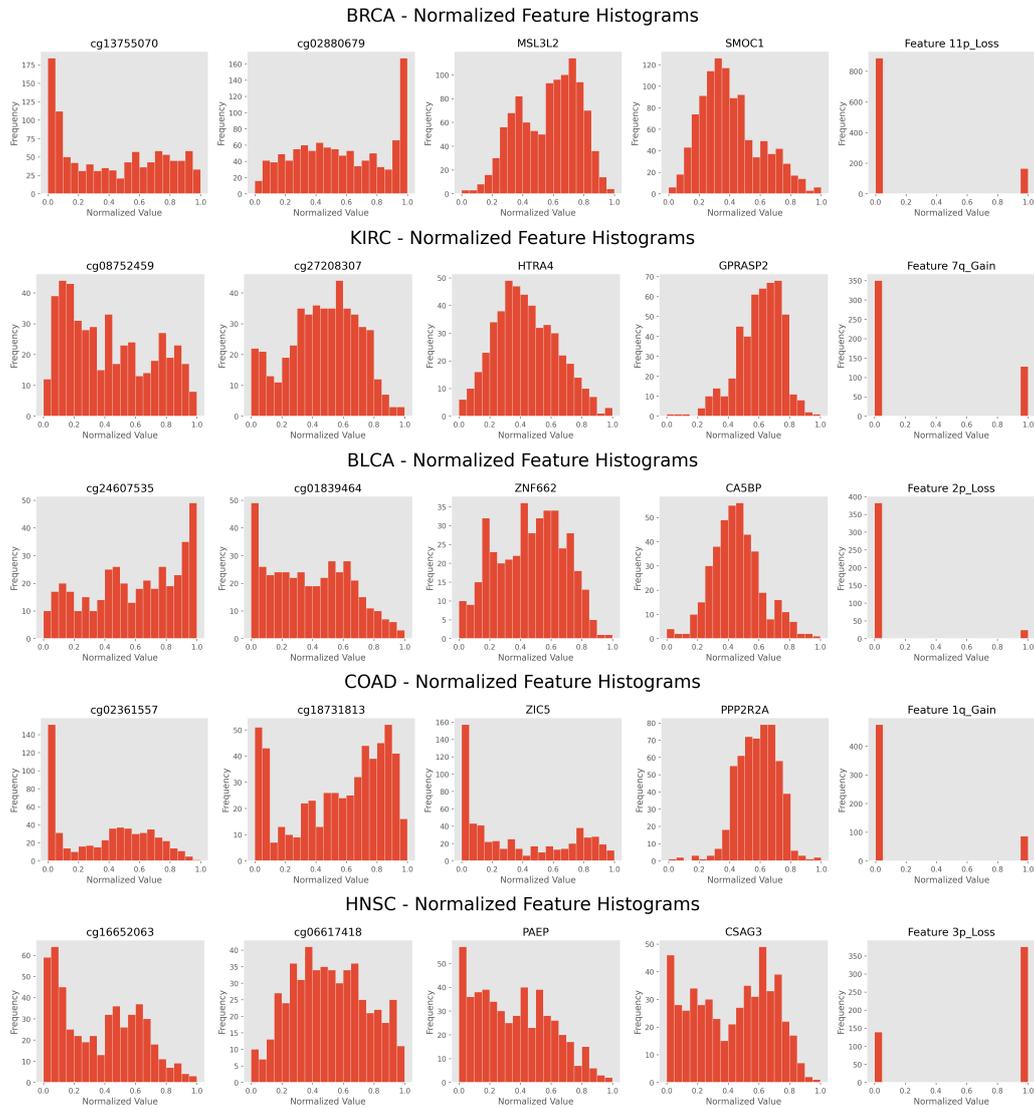


Figure 1.10: Frequency histograms for selected normalized methylation (columns 1 and 2), mRNA (columns 3 and 4) and CNA (column 5) features from BRCA, KIRC, BLCA, COAD and HNSC datasets (top to bottom).

1.4 Synthetic Datasets

In addition to the 5 validation datasets described in the previous section, we generated 5 simple synthetic datasets, for model testing. The main purpose of using the synthetic datasets was to check how well can the model recover known clusters (subtypes) present in the data, and how well it handles different data types and distributions.

Table 1.2 summarizes the synthetic datasets. Datasets S1-S4 are single modality (3 with only continuous features, 1 with only binary features), and dataset S5 is bimodal, with both continuous and binary modalities. All datasets were synthesized according to the following procedure.

1. Isotropic Gaussian distributions with predefined numbers of clusters and numbers of samples in each cluster were generated to mimic the unobserved latent space.
2. Observed features were generated as linear combinations of latent variables and uniform noise.
3. Thresholding or activation similar to the leaky rectified linear unit function were applied for some of the observed features, to introduce binary variables, non-linearity and skewed distributions.

In dataset S5, each of the two modalities was generated from the same latent space, however, selected latent variables were used for the creation of only one of the two modalities. Selected normalized and re-scaled to the $[0, 1]$ range (after 80-20 train-test split) features are visualised in Figure 1.11. UMAPs of latent and observed spaces of each dataset are visualised in Figure 1.12.

Each of the datasets was generated for a specific purpose. Dataset S1 is the simplest, containing mostly normally distributed features and clusters of

equal sizes. Features in dataset S2 are highly skewed, to help us verify how well our model works in the presence of non-Gaussian distributions. With dataset S3, which contains one cluster much smaller than the other two, we want to test how sensitive our model is to small subpopulations. Dataset S4 can be used to verify the method’s suitability for sparse binary features. Finally, dataset S5 is bimodal, with different data types for both modalities, and as such can be used for multimodal data integration method testing.

Dataset	#Latent Features	#Observed Features	#Clusters	Cluster Sizes
S1	20	150 continuous	5	200, 200, 200, 200, 200
S2	20	150 continuous	3	450, 350, 200
S3	15	150 continuous	3	500, 400, 100
S4	15	50 binary	2	600, 400
S5	20	150 continuous, 50 binary	3	400, 350, 250

Table 1.2: Summary of the synthetic datasets.

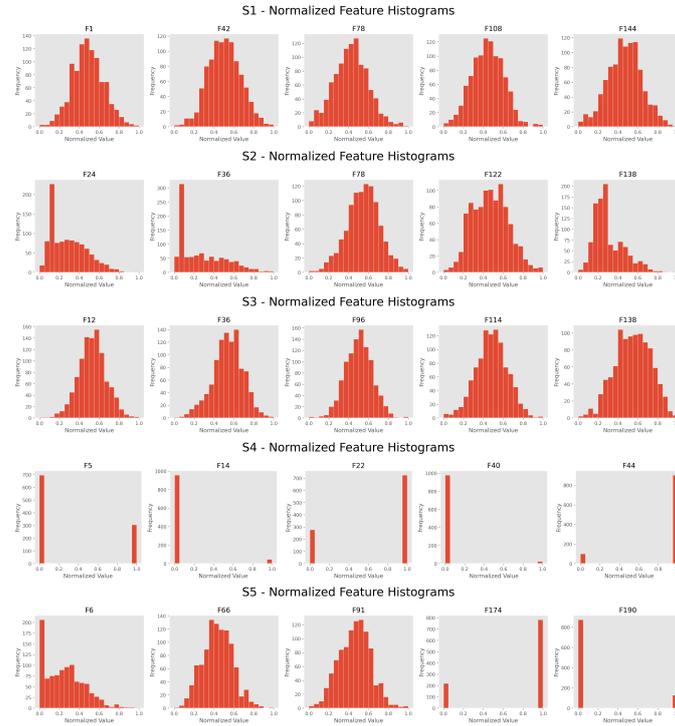


Figure 1.11: Frequency histograms for selected normalized features from synthetic datasets S1, S2, S3, S4 and S5 (top to bottom).

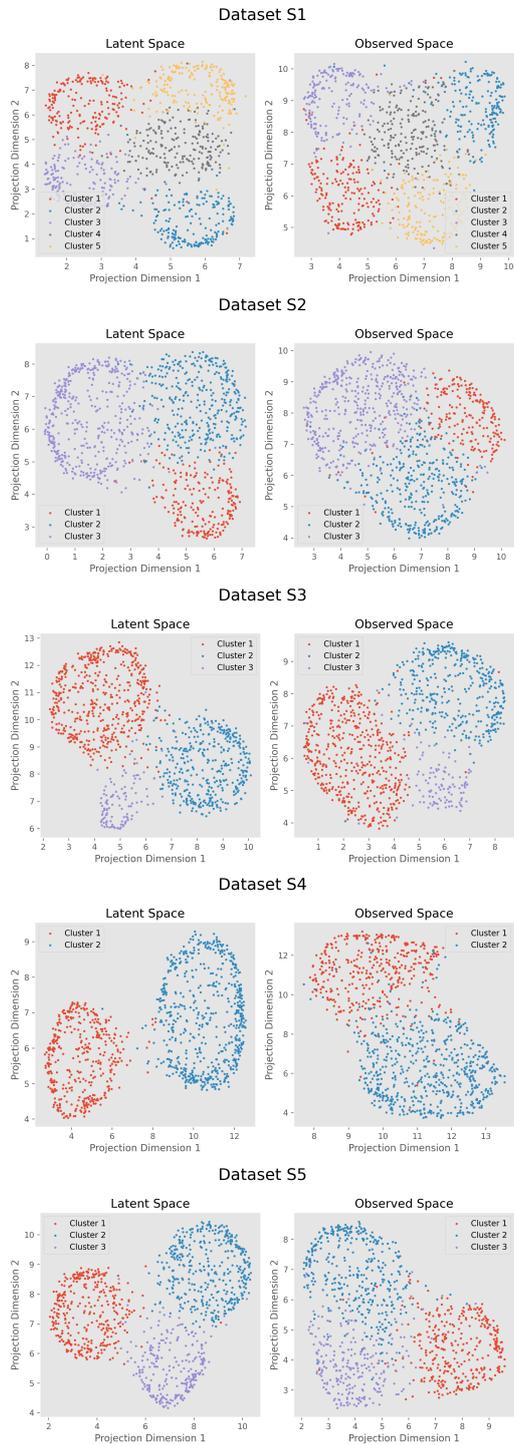


Figure 1.12: UMAP visualisations of the latent (left) and observed (right) spaces of synthetic datasets S1-S5 (top to bottom).

Chapter 2

Methodology: Literature Review and Background

This chapter constitutes a review of the literature and methods most relevant to the methodology proposed in this thesis. We begin Section 2.1 by explaining the potential and challenges associated with multiomic data analysis and by reviewing the most common multiomic data integration strategies. In Section 2.2, we introduce the concept of latent features and review the methods most commonly used for their extraction from multiomic data. We additionally explain the limitations associated with each of the presented approaches. Then, in Section 2.3, we formally introduce the concept of Generative Adversarial Networks [65] and explain how they could be repurposed as latent feature extraction models. In Sections 2.4 - 2.6, we review the relevant solutions to other challenges associated with multiomic data analysis, including mixed-type datasets with diverse distributions, missing modalities and interpretability. Finally, in Section 2.7, we explain the methods commonly used for downstream analysis on the integrated multiomic data, including selected approaches to subtyping and survival modelling.

2.1 Multiomic Data Integration

As explained in Chapter 1, the term ‘multiomics’ broadly refers to tabular data sourced from multiple ‘omes’, or modalities, e.g. genome, proteome, transcriptome, and the analysis thereof. Studies using multiomic data have shown numerous advantages compared to single-omics analysis, including the ability to provide better explanations for complex phenotypes, patterns or molecular pathways, or to improve the performance of classifiers, subtype or drug response predictors [e.g. 4, 160, 170, 191, 220, 228]. As such, many consider multiomic data and its analysis to be the future of precision medicine and personalized cancer treatments [e.g. 80, 123, 161].

The relatively recent decrease in costs of whole genome sequencing and other related screening technologies contributed to the generation of a large number of well-curated multiomic cancer datasets, available for researchers to analyse [e.g. 189]. We reviewed some of these datasets in Chapter 1. However, multiomic data analysis requires careful data integration, which often proves difficult for the currently existing methodologies. As pointed out in the review by Picard et al. [156], and evident from the datasets descriptions in Chapter 1, multiple challenges may arise when integrating multiomic data. These include the noise and complexity inherent in biological data, the risk of overfitting due to a large number of features for the relatively small number of patient samples, the heterogeneity in data types among different omics sources, and the high prevalence of missing values, often involving entire omics representations being unavailable for many patients. Several integration strategies attempt to address the aforementioned challenges, each exhibiting varying degrees of efficacy. Picard et al. [156] comprehensively characterized five main multiomic data integration strategies: early, mixed, intermediate, late, and hierarchical. Here we provide a short overview of each

of the aforementioned strategies.

2.1.1 Early Integration

Early integration (Figure 2.1-A) involves concatenating all available omics datasets into a single matrix prior to downstream analysis. While early integration is a simple and straightforward solution that naturally allows ML models to infer interactions between different omics, data distributions specific to each modality are ignored, potentially leading to irrelevant findings that reflect only the features' membership to a given omics source [176]. Furthermore, size differences between modalities can lead to an imbalanced training process in which omics with the smallest number of features are overlooked [1, 176].

2.1.2 Mixed Integration

The mixed integration strategy (Figure 2.1-B) addresses the heterogeneities present among different omics sources by transforming each omics dataset independently into a simpler latent format before joint analysis on the combined representations [e.g. 32, 90, 94, 130, 218, 221, 230]. The intermediate representations can be obtained using kernel based methods [e.g. 64], graph based methods [e.g. 130, 201, 212], or with neural networks capable of extracting latent representations (e.g. Autoencoders [165], Restricted Boltzmann Machines [83] or Variational Autoencoders [103]).

2.1.3 Intermediate Integration

With the intermediate integration strategy (Figure 2.1-C), data from multiple omics sources can be jointly integrated without requiring prior trans-

formations or data concatenation. The output of this approach is usually a single latent representation shared by all modalities, and multiple omics-specific independent representations. Intermediate integration strategies are capable of discovering joint inter-omics structures, while also highlighting the complementary intra-modality information. Example methods used for intermediate integration include the extensions of the Non-Negative Matrix Factorization [27, 114, 223, 232], Canonical Correlation Analysis [129, 185], Co-Inertia Analysis [136], and integrative clustering approaches [137, 138, 172].

2.1.4 Late Integration

Late integration strategy (Figure 2.1-D) involves analysing each omics dataset separately and consolidating the respective results. In the late integration, an ML model is normally applied to each modality separately, and the predictions are combined with some chosen aggregation function [e.g. 179, 204]. The major disadvantage of such approach is that interactions between different omics sources cannot be captured and the use of their carried complementary information is rather limited. Additionally, late integration often leads to inconsistent conclusions [201].

2.1.5 Hierarchical Integration

Hierarchical integration involves the inclusion of prior knowledge, for example from interaction databases or scientific literature, into the development of the integration method [156]. Specific approaches include Bayesian analysis of genomics data (iBAG) [205], linear regulatory modules (LRMs) [240], Assisted Robust Marker Identification (ARMI) [26] and Robust Network [214].

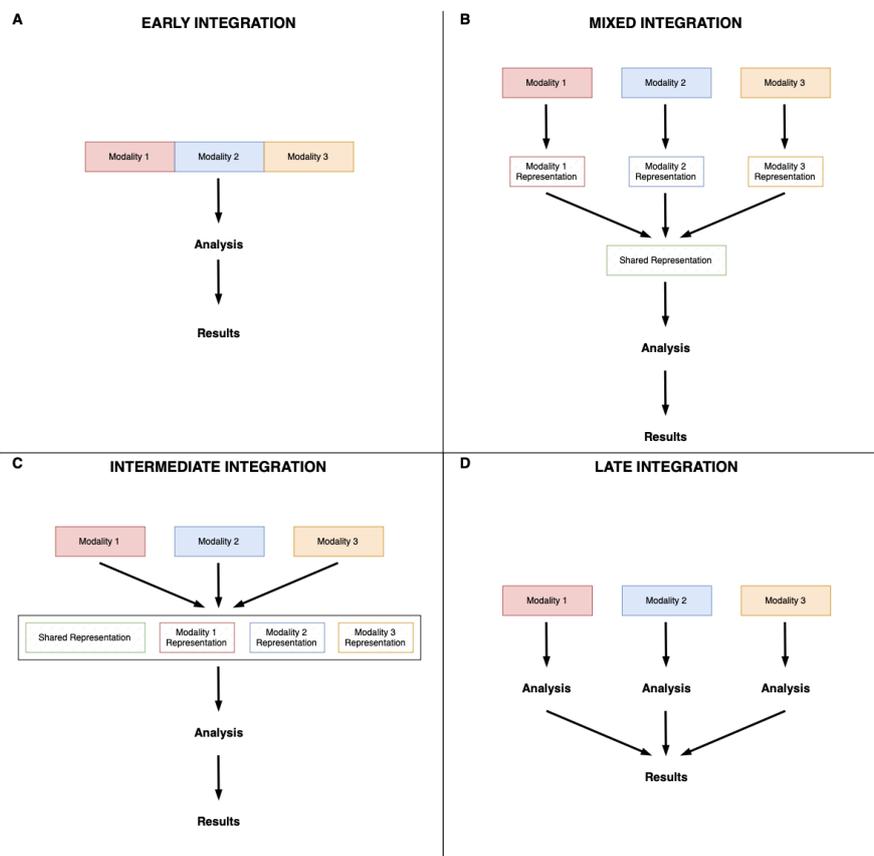


Figure 2.1: Four main multiomic data integration strategies: early (A), mixed (B), intermediate (C) and late (D). Early integration involves concatenating all data sources into a single matrix before downstream analysis. With the mixed integration strategy, each omics source is first transformed into a simpler modality-specific representation, after which a joint modality-shared representation used for downstream analysis is inferred. With the intermediate integration strategy, modality-shared and modality-specific representations used for downstream analysis are inferred at the same time. Late integration strategy involves separate analysis of each modality and consolidation of the respective results.

2.1.6 Choosing an Integration Strategy

Each of the reviewed integration strategies has its own advantages and disadvantages. Which approach should be chosen depends on many factors, including the goals of the analysis, the familiarity of the researcher with various machine learning models and biological concepts, the specific omics datasets available, and the desired level of interpretability. Considering all

aspects, mixed and intermediate integration strategies are perhaps the most reasonable approaches for many datasets, given that they do not require the advanced domain knowledge, that is a deep understanding of the underlying biological networks and concepts, necessary for hierarchical integration, and that they overcome many of the challenges associated with the early and late integration solutions. Additionally, some models could potentially be used for exploring multiple integration strategies for a given dataset. For example, with appropriate modifications, an Autoencoder [165] could be used for early [e.g. 29] or mixed [e.g. 218] integration. An Autoencoder is just one example of model capable of extracting latent representations from the data, a crucial step within many of the multiomic data integration strategies and a concept which we will now delve into further.

2.2 Latent Feature Models

Latent features, or latent variables (LVs), can be defined as underlying variables within a dataset that are not directly observed but can be inferred from data. They can reveal hidden patterns or relationships underlying the observed measurements. As such, latent feature models are methods for extracting said features from the data. The extraction of latent features is related to the concept of dimensionality reduction, which involves generating low-rank representations of high-dimensionality datasets. Both approaches aim to produce compressed representations of the data while preserving as much relevant information as possible. However, the main difference between the two concepts is that latent feature models additionally focus on describing the relationships between latent and observed variables and providing insights into the underlying factors and patterns. Nevertheless, the

two terms are often used interchangeably.

As pointed out by Picard et al. [156], the extraction of latent features is commonly used as a pre-processing step during multiomic data analysis, irrespective of the specific integration strategy chosen. This is often done in conjunction with feature selection, which involves reducing the dimensionality of the dataset by selecting only specific features of interest. For example, we may choose to focus our analysis on 1000 genes with most variable RNA expressions, or limit mutational data used to established driver genes only. Furthermore, latent feature models can be used as an integrative part of the mixed or intermediate integration strategies, for the extraction of the independent or shared representations.

Various latent feature models / dimensionality reduction methods exist, each with its own advantages and disadvantages. Which latent feature model should be selected for a given problem depends on a number of factors including the complexity of the data and the goals of the analysis. For multiomic data analysis specifically, there are several desirable properties that the selected latent feature model should possess. One such property would be the possibility of enforcing a specific prior (e.g. the normal distribution) over the latent space. Such property would provide a substantial advantage when integrating data using the mixed integration strategy. Specifically, before final integration, all considered modalities - irrespective of their dimensionalities, data types or feature distributions - could be first projected onto latent spaces with certain pre-defined sizes and properties, thus alleviating the majority of issues arising from the heterogeneities among different omics sources. Another desirable characteristic for a latent feature model is interpretability. Should the model be capable of preserving interpretable links between latent and observed features, it will be easier to explain the ratio-

nale behind the findings, an invaluable property in biomedical applications. Furthermore, the method’s ability to support a high prevalence of missing modalities common in multiomic datasets must be taken into account when choosing or developing a latent feature model. Finally, as we are primarily interested in deriving cancer subtypes, the structure of the latent space should be suitable for downstream analysis that enables this.

In the subsequent sections, we review the latent feature models most commonly used in single- or multi- omic data analysis. We divide the latent feature extraction approaches into two main categories: statistical (Section 2.2.1) and machine learning based (Section 2.2.2). It is important to note that, while some of the methods reviewed in later sections allow for a flexible prior choice, none of the approaches can support missing data, or multiomic data with missing modalities. However, for many of these methods, certain adjustments can be made to introduce such functionality. As these adjustments are often method independent, we review them separately in Section 2.5.

2.2.1 Statistical Methods

The most recognised statistical method related to latent feature models is Principal Component Analysis (PCA) [85, 98, 153], in which the original variables are transformed into a set of new variables, referred to as principal components (PCs), that preserve as much variance present in the original data as possible. PCA is calculated by performing the *eigenvector decomposition* of the covariance matrix, after which the data can be transformed by projection onto the eigenvectors. This means that the PCs have two highly desirable properties: (1) they are *orthogonal* to each other, that is to say the information contained in each PC is independent, and (2) the variance

is proportional to the *eigenvalue* associated with each eigenvector. Therefore the most informative PCs are those with the greatest eigenvalues, and dimensionality reduction is generally performed by only projecting onto a subset of the eigenvectors with the greatest eigenvalues. Although widely used and easy to deploy, PCA has numerous limitations. Most notably, in PCA, the principal components are linear combinations of the observed features, which means that the method fails to recover non-linear relationships and patterns present in the data. Moreover, what stems from linearity is that a specific prior cannot be explicitly enforced over the latent space. Furthermore, although interpretable variants such as Sparse or Non-Negative PCA exist [229, 241], the interpretability of PCA in its standard form is rather limited with each latent variable normally being a dense combination of all observed variables. Additionally, many claim that PCA is unsuitable for use with non real-valued (e.g. binary) data due to its implicit minimization of a squared loss function [e.g. 36].

In the context of statistical latent feature models, Factor Analysis (FA) is commonly recognized as the foundational approach. The term itself broadly refers to a group of statistical methods aimed at transforming a large number of observations into a smaller set of *factors*, which are the statistical version of latent features. The fundamental setup is that the data originates from a linear combination of some unknown factors plus some noise terms, and then statistical inference methods are employed to estimate the factors from the observed variables.

Multiple Factor Analysis (MFA) extends the concept of Factor Analysis to datasets with multiple sets of variables available for the same set of observations, e.g. multiple omics sources available for the same set of patients. The technique involves performing Factor Analysis on each set of

variables and integrating the results to capture factors shared across sources. The method shares the limitations of FA, including the assumption of linearity. Additionally, MFA requires complete data matrices and as such cannot handle datasets with missing modalities.

Another example of a statistical latent feature model is Independent Component Analysis (ICA) [37] which involves separating a multivariate signal into a smaller set of additive, statistically independent components capable of representing meaningful patterns within the data. Akin to PCA, ICA is based on linear transformations and as such is unable to capture non-linear signals in the data and does not allow for the selection of a specific latent space prior.

Non-Negative Matrix Factorization (NMF) [114] is a dimensionality reduction technique in which a dataset, here in the form of a non-negative matrix, is decomposed into two lower-rank, also non-negative, matrices, one representing the set of the inferred latent variables, the other one containing coefficients that correspond to the contribution of each of the latent features to the observed variables. In NMF, the non-negativity constraint improves the method's interpretability. However, similarly to PCA and ICA, NMF is limited by it being a linear dimensionality reduction technique.

Other statistical methods, such as Gaussian Mixture Models (GMMs), do not explicitly extract latent features from data, but focus on representing a given dataset as a mixture of latent components, with each data point assigned to one such component. GMMs specifically assume that each data sample belongs to one of the underlying Gaussian distributions, here called mixture components. The mixture components themselves can be viewed as clusters, or latent features present in the data. Similarly, Dirichlet Processes (DPs) [54] provide distributions over probability distributions, i.e. model

datasets based on the belonging of data points to latent clusters, or components. Although the aforementioned approaches are commonly used in statistical modelling, they are characterized by several limitations including sensitivity to initialization, computationally expensive inference and restricted scalability to complex datasets with high dimensionality.

Finally, Indian Buffet Processes (IBPs) [68] can be used to model latent features present in data as distributions over sparse binary matrices with finite numbers of rows and infinite numbers of columns. Formally, an IBP places the following prior on an $N \times K$ binary matrix Z indicating the presence or absence of each latent feature for each sample:

$$p(Z) = \frac{\alpha^{K^+}}{\prod_{i=1}^N K_1^{(i)}!} \exp\{-\alpha H_N\} \prod_{k=1}^{K^+} \frac{(N - m_k)!(m_k - 1)!}{N!}. \quad (2.1)$$

IBP is an analogy to a food buffet in which customers (rows) sample some of the possibly infinite number of dishes (columns). Here, K^+ is the number of non-zero columns in Z , i.e. the number of dishes that have already been sampled, m_k indicates the number of times a dish k has been sampled, H_N is the N^{th} harmonic number, $K_1^{(i)}$ is the number of new dishes sampled by customer i , and α controls the expected number of dishes sampled by each customer, i.e. the expected number of features present in each observation. The i^{th} customer takes each previously sampled dish with probability $\frac{m_k}{i}$ and additionally samples $Poisson(\alpha/i)$ new dishes. Finite approximations of the IBP are commonly derived using the stick-breaking construction process [183] and have been applied as latent space priors, for example in Variational Autoencoders [73]. Assuming a finite number of columns K , the stick-breaking process first samples the feature activation probability μ_k independently for $k = 1, 2, \dots, K$ (for each column of Z) as $\mu_k \sim Beta(\frac{\alpha}{K}, 1)$. Each entry (i, k) of Z , denoted as z_{ik} is then independently sampled as $z_{ik} | \mu_k \sim Bernoulli(\mu_k)$.

IBPs are commonly represented as left-ordered matrices (Figure 2.2). In models concerned with multiomic data analysis, and Indian Buffet Process prior would allow us to, for example, interpret the extracted latent features as binary indicators signifying the presence or absence of the underlying biological processes connected to various patterns of genetic alterations visible in the observed multiomic measurements.

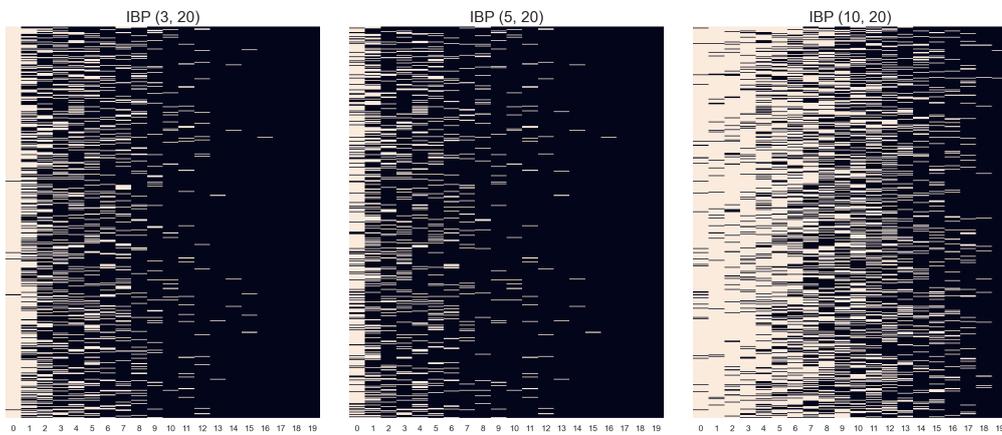


Figure 2.2: Heatmap visualisations of example Indian Buffet Process matrices generated using the stick-breaking construction process for $K = 10$ and $\alpha = 3$ (left), $\alpha = 5$ (middle) and $\alpha = 10$ (right). Heatmap columns represent latent variables (dishes), rows represent data samples (customers). A white mark indicates that a given latent feature is active for a given data sample (a given dish has been sampled by a given customer).

2.2.2 Machine Learning Methods

Machine Learning based latent feature models offer a flexible and versatile alternative to traditional statistical approaches. Autoencoders (AEs) [165] are perhaps the most common neural networks used for latent feature extraction. A basic Autoencoder has two components, encoder E and decoder D . The role of E is to compress (or encode) each data sample \mathbf{x} into its low-dimensional latent representation $\mathbf{z} = E(\mathbf{x})$. The role of D , in turn, is to

reconstruct \mathbf{x} from its latent representation \mathbf{z} , i.e. to produce $\hat{\mathbf{x}} = D(E(\mathbf{x}))$ as similar to \mathbf{x} as possible. The entire model is trained to minimize the reconstruction error between the input data and the obtained reconstructions. A usual choice for the reconstruction error metric is the Mean Squared Error (MSE), in this context defined as:

$$MSE = \frac{1}{N} \sum_{i=1}^N (\mathbf{x}_i - D(E(\mathbf{x}_i)))^2, \quad (2.2)$$

where N is the total number of samples in the dataset. Figure 2.3 visualises a simple Autoencoder.

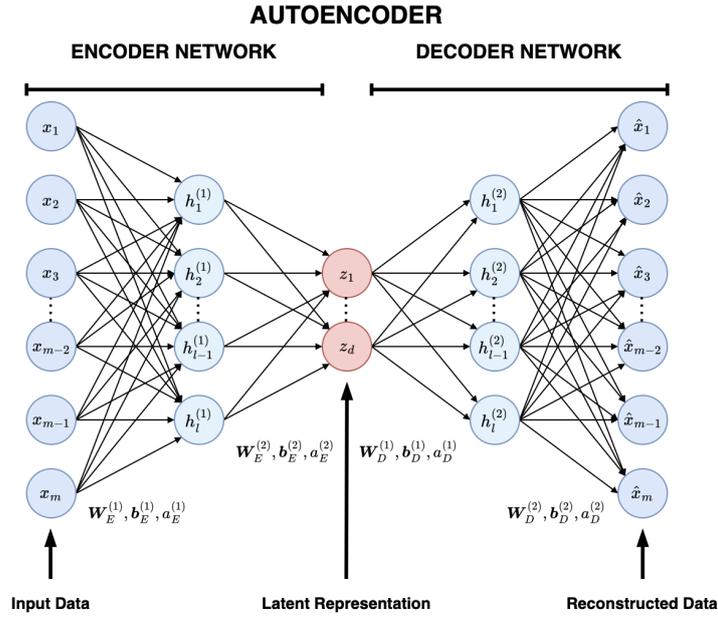


Figure 2.3: A simple Autoencoder model. The encoder network takes a data point \mathbf{x} as input and produces its compressed latent space representation \mathbf{z} , calculated as $\mathbf{z} = a_E^{(2)}(\mathbf{W}_E^{(2)}(a_E^{(1)}(\mathbf{W}_E^{(1)}\mathbf{x} + \mathbf{b}_E^{(1)})) + \mathbf{b}_E^{(2)})$, using the respective encoding weights \mathbf{W}_E , biases \mathbf{b}_E and activation functions a_E . The decoder network then produces the reconstruction $\hat{\mathbf{x}}$ of \mathbf{x} , calculated as $\hat{\mathbf{x}} = a_D^{(2)}(\mathbf{W}_D^{(2)}(a_D^{(1)}(\mathbf{W}_D^{(1)}\mathbf{z} + \mathbf{b}_D^{(1)})) + \mathbf{b}_D^{(2)})$, using the respective decoding weights \mathbf{W}_D , biases \mathbf{b}_D and activation functions a_D . The entire network is trained to minimize the MSE loss function between \mathbf{x} and $\hat{\mathbf{x}}$, as defined in Equation 2.2. Commonly, Autoencoder's weights are tied, which in this case could be enforced by the setting following constraints: $\mathbf{W}_E^{(1)} = \mathbf{W}_D^{(2)\top}$ and $\mathbf{W}_E^{(2)} = \mathbf{W}_D^{(1)\top}$.

Autoencoders are flexible and easy to train models that allow for non-linear dimensionality reduction. Although largely non-interpretable in their standard forms, AEs can be made more explainable by enforcing sparsity or non-negativity constraints on the model weights [e.g. 12, 40]. Furthermore, traditionally trained in an unsupervised fashion, Autoencoders can be extended with supervised classification or clustering losses [e.g. 72, 111, 217], to encourage encodings more suitable for the downstream tasks chosen. While commonly used, Autoencoders have several limitations, many of which stem from the use of the Mean Squared Error (MSE) loss function. Specifically, MSE is sensitive to outliers due to the squaring of error terms [6, 92]. Additionally, MSE assumes that the error terms follow Gaussian distributions, which may limit the suitability of Autoencoders for data where this does not apply [235]. Finally, standard Autoencoders do not offer a flexible prior choice.

Another common network-based latent feature model - Variational Autoencoder (VAE) [103] - extends the deterministic Autoencoder into a probabilistic model capable of synthetic data generation. In standard AEs, the deterministic space is usually non-continuous and does not allow for easy interpolations. As such, sampling from the latent space for the purpose of generating new samples is generally not feasible. VAEs address this limitation by leveraging amortized variational inference in a Bayesian latent variable model, effectively acting as ‘probabilistic autoencoders’.

A Bayesian latent variable model attempts to infer the posterior distribution $p_{\theta}(\mathbf{z}|\mathbf{x})$ of the unobserved latent variables $\mathbf{z} \sim p(\mathbf{z})$, given observed data \mathbf{x} . The observed data is assumed to be generated via $\mathbf{x} \sim p_{\theta}(\mathbf{x}|\mathbf{z})$, where $p_{\theta}(\mathbf{x}|\mathbf{z})$ is a likelihood function parametrized by θ , often represented as a neural network. Inferring the posterior directly is however intractable due to the

marginal likelihood computation: $p_\theta(\mathbf{x}) = \int p_\theta(\mathbf{x}|\mathbf{z})p(\mathbf{z})d\mathbf{z}$. To address this, variational inference approximates the intractable posterior $p_\theta(\mathbf{z}|\mathbf{x})$ with a simpler distribution $q_\phi(\mathbf{z}|\mathbf{x})$, parametrised by ϕ . This approximation is optimized by minimizing the Kullback–Leibler (KL) divergence between $q_\phi(\mathbf{z}|\mathbf{x})$ and $p_\theta(\mathbf{z}|\mathbf{x})$, which is equivalent to maximizing the Evidence Lower Bound (ELBO):

$$\mathcal{L}(\theta, \phi) = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z})] - \text{KL}(q_\phi(\mathbf{z}|\mathbf{x})||p(\mathbf{z})). \quad (2.3)$$

Rather than learning a separate variational distribution $q_\phi(\mathbf{z}|\mathbf{x})$ for each data point, VAEs utilize amortized variational inference, where q_ϕ is parametrized by a shared neural network, an encoder, that maps each data point \mathbf{x} to a distribution over \mathbf{z} .

In simple terms, a VAE, similarly to an Autoencoder, consists of two components: (1) an encoder that maps \mathbf{x} to a latent representation \mathbf{z} , and (2) a decoder that reconstructs \mathbf{x} from \mathbf{z} . However, unlike in a deterministic Autoencoder, the encoder in a VAE outputs a distribution $q_\phi(\mathbf{z}|\mathbf{x})$ instead of a single point, and the decoder models the likelihood $p_\theta(\mathbf{x}|\mathbf{z})$, capturing uncertainty in the reconstruction. This design allows us to interpret VAEs as probabilistic Autoencoders. In VAEs each latent feature is usually represented as a probability distribution with mean μ and standard deviation σ . Therefore, the latent space in VAEs is continuous and can be sampled from. The continuity is enforced by adding a regularization loss, usually the Kullback–Leibler (KL) Divergence, to the standard Autoencoder’s loss. Formally, VAE’s loss \mathcal{L}_{VAE} can be defined as:

$$\mathcal{L}_{\text{VAE}} = \mathcal{L}_{\text{rec}} + \mathcal{L}_{\text{KL}}, \quad (2.4)$$

where \mathcal{L}_{rec} denotes the reconstruction loss as used in an Autoencoder, e.g. the MSE loss, and \mathcal{L}_{KL} is the KL Divergence between the learned distributions in

the latent space, and some predefined target distribution, usually a standard Gaussian, in which case:

$$\mathcal{L}_{\text{KL}} = -\frac{1}{2} \sum_{d=1}^D (1 + \log \sigma_d^2 - \mu_d^2 - \sigma_d^2), \quad (2.5)$$

where D denotes the dimensionality of the latent space. Figure 2.4 visualises a simple VAE.

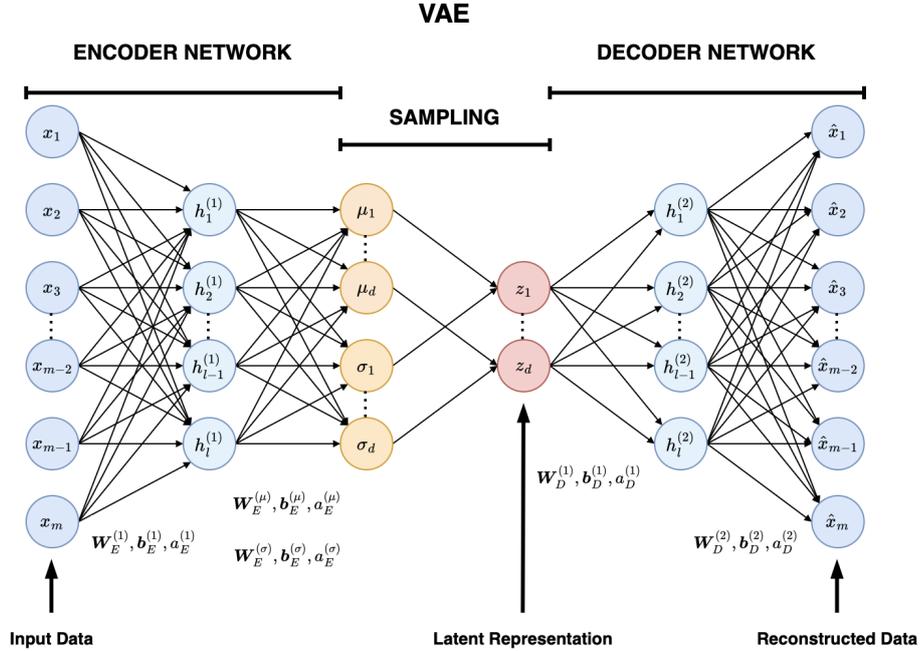


Figure 2.4: A simple VAE model. The encoder network takes a data point \mathbf{x} as input and produces its probabilistic representation characterized by $\boldsymbol{\mu} = a_E^{(\mu)}(\mathbf{W}_E^{(\mu)}(a_E^{(1)}(\mathbf{W}_E^{(1)}\mathbf{x} + \mathbf{b}_E^{(1)})) + \mathbf{b}_E^{(\mu)})$ and $\boldsymbol{\sigma} = a_E^{(\sigma)}(\mathbf{W}_E^{(\sigma)}(a_E^{(1)}(\mathbf{W}_E^{(1)}\mathbf{x} + \mathbf{b}_E^{(1)})) + \mathbf{b}_E^{(\sigma)})$, using the respective encoding weights \mathbf{W}_E , biases \mathbf{b}_E and activation functions a_E . The latent space representation \mathbf{z} of \mathbf{x} is then sampled as $\mathbf{z} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\sigma})$. The decoder network then produces the reconstruction $\hat{\mathbf{x}}$ of \mathbf{x} , calculated as $\hat{\mathbf{x}} = a_D^{(2)}(\mathbf{W}_D^{(2)}(a_D^{(1)}(\mathbf{W}_D^{(1)}\mathbf{z} + \mathbf{b}_D^{(1)})) + \mathbf{b}_D^{(2)}$, using the respective decoding weights \mathbf{W}_D , biases \mathbf{b}_D and activation functions a_D . The entire network is trained to minimize the loss function defined in Equation 2.4, that is the reconstruction loss between \mathbf{x} and $\hat{\mathbf{x}}$ and the KL Divergence between $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ and a predefined target distribution.

Although VAEs resolve some of the issues present in Autoencoders, they are still restricted by the limitations stemming from the use of the MSE reconstruction loss. Furthermore, VAEs are known to produce blurry reconstructions with a severely limited attention to detail [e.g. 107]. Finally, while enforcing a latent space prior different than the Gaussian distribution in VAEs is possible, complex modifications are often required [e.g. 58].

Restricted Boltzmann Machines (RBMs) [83] can also be used as latent feature models and offer an interesting alternative to the two architectures reviewed above. In contrast to the input, hidden and reconstruction layers present in Autoencoders and VAEs, standard RBMs have binary hidden and visible layers only. Instead of attempting to reconstruct the input data, an RBM aims to learn its probability distribution. Formally, the training of an RBM attempts to maximize the joint probability $P(\mathbf{v}, \mathbf{h})$ of a configuration of the visible and hidden units, \mathbf{v} and \mathbf{h} respectively, given as:

$$P(\mathbf{v}, \mathbf{h}) = \frac{e^{-E(\mathbf{v}, \mathbf{h})}}{Z}, \quad (2.6)$$

with Z being a normalization constant and $E(\mathbf{v}, \mathbf{h})$ denoting the energy function defined as:

$$E(\mathbf{v}, \mathbf{h}) = -\mathbf{a}^\top \mathbf{v} - \mathbf{b}^\top \mathbf{h} - \mathbf{v}^\top \mathbf{W} \mathbf{h}, \quad (2.7)$$

where \mathbf{a} and \mathbf{b} are the biases of the visible and hidden layers respectively, and \mathbf{W} is a matrix of weights. Figure 2.5 visualises a simple RBM.

While RBMs can allow for the extraction of easily interpretable binary latent variables, especially when used in conjunction with non-negativity weight constraints [213], the main limitation of RBMs lies in their restricted ability to work with datasets containing continuous features. As RBMs assume visible units to be binary, continuous variables need to be discretized, potentially leading to a significant information loss.

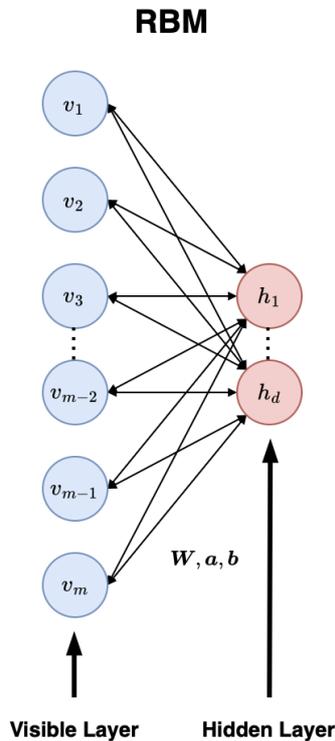


Figure 2.5: A simple RBM model. The network is trained to maximize the probability $P(\mathbf{v}, \mathbf{h})$ of the configuration of the visible units (data samples, \mathbf{v}) and hidden (encoded representations, \mathbf{h}) units, defined as $P(\mathbf{v}, \mathbf{h}) = \frac{e^{-E(\mathbf{v}, \mathbf{h})}}{Z}$, where Z is a normalization constant and $E(\mathbf{v}, \mathbf{h}) = -\mathbf{a}^\top \mathbf{v} - \mathbf{b}^\top \mathbf{h} - \mathbf{v}^\top \mathbf{W} \mathbf{h}$. Here, \mathbf{W} is the weight matrix, and \mathbf{a} and \mathbf{b} are the biases of the visible and hidden layers, respectively.

Despite their limitations, ML models based on Autoencoders, VAEs and RBMs are the most common neural networks used for unsupervised single- or multi- omics data analysis [e.g. 29, 84, 208, 213, 230, 234]. In addition to the aforementioned models, Generative Adversarial Networks (GANs) [65] have more recently emerged as a notable alternative to these network architectures. Although GANs have been successfully used in the image and video domain [e.g. 89, 102, 231], and recently more so for tabular data generation [e.g. 33, 149, 219, 236], their application for multiomic data analysis remains largely unexplored.

2.3 Generative Adversarial Networks

Generative Adversarial Networks [65], or GANs, are generative models mainly used for synthetic data generation. GANs are extremely popular in the image, audio and video domain, finding applications in photo-realistic image generation [e.g. 112], image-to-image translation [e.g. 91] or cross-modality translation [e.g. 173], to name a few. Additionally, GANs can be used for tabular data generation [e.g. 219, 236] and missing data [e.g. 226] or missing modality [e.g. 148] imputation. Nevertheless, despite their great potential, except for few studies [3, 28, 62, 134, 148, 199, 221], GANs are not usually applied for multiomic data analysis.

As GANs are a major focus of this thesis, we explore them in detail in this section. We begin by reviewing some of the relevant GAN concepts and explaining how can we repurpose GANs for latent feature extraction. We then summarize the relevant studies applying GANs for multiomic data analysis and point out the gap in the research.

2.3.1 Basic Definition

Formally, Generative Adversarial Networks (GANs) [65] are a framework for inferring generative models via adversarial processes. In the basic GAN approach, two neural networks, generator G and discriminator D compete against each other in a two-player minimax game. Given a random vector \mathbf{z} sampled from a prior, usually Gaussian, distribution $p_{\mathbf{z}}(\mathbf{z})$, G generates a fake data sample $G(\mathbf{z})$. Given a real data sample \mathbf{x} from the data distribution $p_{\text{data}}(\mathbf{x})$, or a fake sample generated by G , D outputs a single scalar representing the probability that the sample came from the data rather than from G . D 's training objective is to maximize the probability of correctly

classifying each sample as either real or fake, while G is trained to maximize the probability of D making a mistake. Both models are trained simultaneously, which can be represented as a two-player minimax game with the following cross-entropy based objective [65]:

$$\min_G \max_D \mathbb{E}_{\mathbf{x} \sim p_{\text{data}}(\mathbf{x})} [\log (D(\mathbf{x}))] + \mathbb{E}_{\mathbf{z} \sim p_{\mathbf{z}}(\mathbf{z})} [\log (1 - D(G(\mathbf{z})))] \quad (2.8)$$

Early in the training, samples from G are of a rather poor quality, and as such D rejects them with high confidence, which in turn leads to the saturation of $\log(D(G(\mathbf{z})))$ and disrupts G 's learning process [65]. To alleviate this issue, the objective from Equation 2.8 is most commonly used in its non-saturating version, which prevents the discriminator from ‘getting too far ahead’ of the generator [65]:

$$\begin{aligned} \max_D \mathbb{E}_{\mathbf{x} \sim p_{\text{data}}(\mathbf{x})} [\log (D(\mathbf{x}))] + \mathbb{E}_{\mathbf{z} \sim p_{\mathbf{z}}(\mathbf{z})} [\log (1 - D(G(\mathbf{z})))] \\ \max_G \mathbb{E}_{\mathbf{z} \sim p_{\mathbf{z}}(\mathbf{z})} [\log (D(G(\mathbf{z})))] \end{aligned} \quad (2.9)$$

In this above equation, G is trained to maximize $\log(D(G(\mathbf{z})))$, instead of minimizing $\log(1 - D(G(\mathbf{z})))$, a modification that provides stronger gradients early in the training [65]. Figure 2.6 visualises a basic GAN model.

Although GANs are capable of synthesising high-quality data [e.g. 89, 102, 231], they are characterized by multiple well-known failure modes, including training instability and sensitivity to hyperparameters. Specifically, reaching training convergence, or achieving Nash equilibrium, i.e. a point at which G produces fake data indistinguishable from the real data and D cannot differentiate between the real and fake samples, is challenging and often even impossible in practice. A further issue with a standard GAN involves the so-called mode collapse, which occurs when the generator learns to produce only a small set of highly plausible outputs, and so generates only these, ignoring all other structures in the data. As such, mode collapse leads to

the lack of diversity, poor generalization and the loss of information with important features and signals in the data remaining uncaptured.

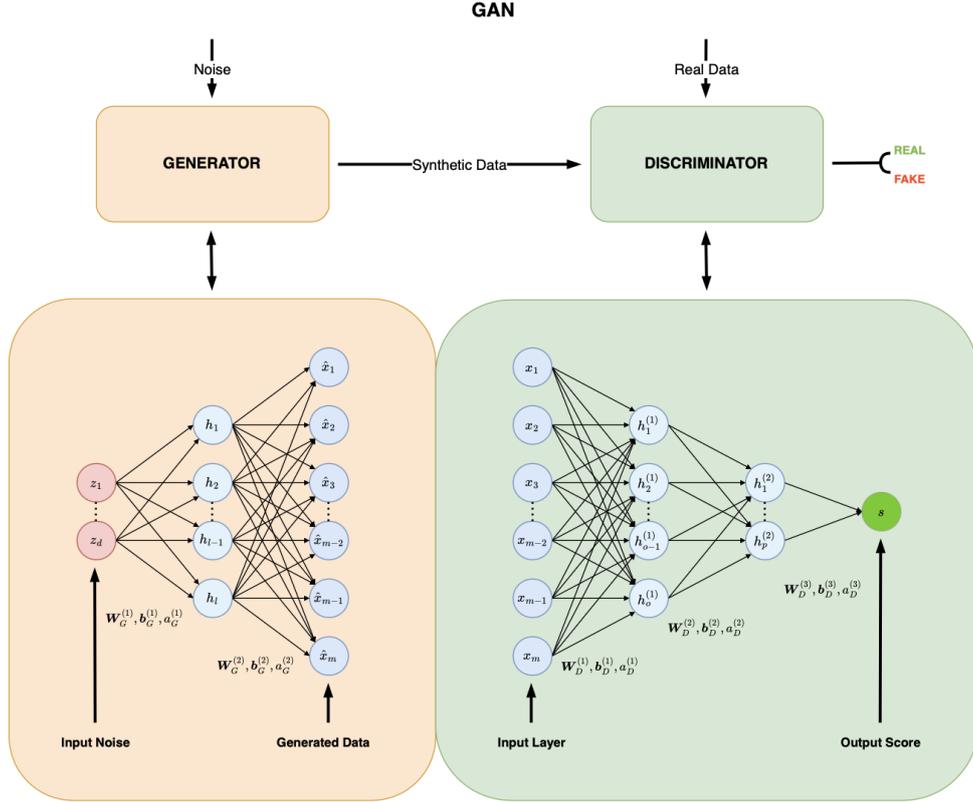


Figure 2.6: A simple GAN model. Given a noise sample \mathbf{z} , commonly obtained as $\mathbf{z} \sim \mathcal{N}(0, 1)$, the generator G produces a synthetic data sample $G(\mathbf{z}) = a_G^{(2)}(\mathbf{W}_G^{(2)}(a_G^{(1)}(\mathbf{W}_G^{(1)}\mathbf{z} + \mathbf{b}_G^{(1)}) + \mathbf{b}_G^{(2)})$, using the respective generating weights \mathbf{W}_G , biases \mathbf{b}_G and activation functions a_G . Given a real data sample \mathbf{x} , or a synthetic data sample $G(\mathbf{z})$, the discriminator D produces a its real/fake score calculated as $D(\mathbf{x}) = a_D^{(3)}(\mathbf{W}_D^{(3)}(a_D^{(2)}(\mathbf{W}_D^{(2)}(a_D^{(1)}(\mathbf{W}_D^{(1)}\mathbf{x} + \mathbf{b}_D^{(1)}) + \mathbf{b}_D^{(2)}) + \mathbf{b}_D^{(3)})$ or $D(G(\mathbf{z})) = a_D^{(3)}(\mathbf{W}_D^{(3)}(a_D^{(2)}(\mathbf{W}_D^{(2)}(a_D^{(1)}(\mathbf{W}_D^{(1)}G(\mathbf{z}) + \mathbf{b}_D^{(1)}) + \mathbf{b}_D^{(2)}) + \mathbf{b}_D^{(3)})$, using the respective discriminator's weights \mathbf{W}_D , biases \mathbf{b}_D and activation functions a_D . The entire model is usually trained to optimize the objective from Equation 2.9.

2.3.2 Wasserstein GANs

In an attempt to rectify such failure modes, Arjovsky et al. [9] proposed a modification to the standard GAN, called Wasserstein GAN (WGAN). WGAN adjusts the GAN’s training procedure by replacing the discriminator with a critic, which, instead of classifying samples as either real or fake, outputs scores quantifying their ‘realness’ or ‘fakeness’. Such scores are then used to calculate the modified loss function based on the approximation of the Wasserstein distance between the distributions of the real and generated samples. The Wasserstein distance, also called the Earth-Mover distance, between two distributions \mathbb{P}_r and \mathbb{P}_g is defined in [9] as:

$$W(\mathbb{P}_r, \mathbb{P}_g) = \inf_{\gamma \in \Pi(\mathbb{P}_r, \mathbb{P}_g)} \mathbb{E}_{(x,y) \sim \gamma} [\|x - y\|], \quad (2.10)$$

where $\Pi(\mathbb{P}_r, \mathbb{P}_g)$ denotes the set of all joint distributions $\gamma(x, y)$ with marginals \mathbb{P}_r and \mathbb{P}_g , respectively. The distance describes the ‘cost’ of an optimal transport needed to transform the distribution \mathbb{P}_r into \mathbb{P}_g . Direct computation of the Wasserstein distance as defined in Equation 2.10 is however challenging, due to the highly intractable infimum. Instead, the Kantorovich-Rubinstein duality [198] is commonly used to approximate it:

$$W(\mathbb{P}_r, \mathbb{P}_g) = \sup_{f \in \mathcal{D}} \mathbb{E}_{x \sim \mathbb{P}_r} [f(x)] - \mathbb{E}_{y \sim \mathbb{P}_g} [f(y)], \quad (2.11)$$

where \mathcal{D} is a set of 1-Lipschitz continuous functions. A function $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is said to be 1-Lipschitz continuous if the following inequality holds for any two points x and y in its domain:

$$|f(x) - f(y)| \leq \|x - y\|. \quad (2.12)$$

In practical scenarios, the dual formulation of the Wasserstein distance can be approximated by parametrizing f with a neural network f_θ and enforcing

the 1-Lipschitz constraint via techniques such as weight clipping [9] or gradient penalty terms [71]. During training, gradients of the loss function with respect to θ can be computed with the use of standard back-propagation algorithms [9]. The complexity of such computation is linear in the numbers of network parameters and samples considered (batch size).

The creators of WGAN argued the superior suitability of the Wasserstein metric for GAN training and demonstrated how their proposed modification improves training stability, alleviates the mode collapse issue and provides a meaningful loss metric indicative of the training’s progression [9]. Formally, the WGAN’s training objective [9] (based on the Kantorovich-Rubinstein duality [198]) is defined as:

$$\min_G \max_{D \in \mathcal{D}} \mathbb{E}_{\mathbf{x} \sim p_{\text{data}}(\mathbf{x})} [D(\mathbf{x})] + \mathbb{E}_{\mathbf{z} \sim p_{\mathbf{z}}} [D(G(\mathbf{z}))], \quad (2.13)$$

The authors of WGAN proposed to achieve the 1-Lipschitz continuity by clipping the weights of the critic to a compact space (e.g. $[-0.01, 0.01]$) after each gradient update [9].

Gulrajani et al. [71] showed how weight clipping in Wasserstein GANs can lead to multiple optimization difficulties, including capacity underuse and exploding and vanishing gradients, and proposed to solve them by using gradient penalties to enforce the 1-Lipschitz constraint instead. In their model, a gradient penalty term encourages the norm of the critic’s gradient towards 1. The training objective of their model, often referred to as WGAN-GP (Wasserstein GAN with Gradient Penalty) [71] now becomes:

$$\min_G \max_D \mathbb{E}_{\mathbf{x} \sim p_{\text{data}}(\mathbf{x})} [D(\mathbf{x})] - \mathbb{E}_{\mathbf{z} \sim p_{\mathbf{z}}} [D(G(\mathbf{z}))] - \lambda \mathbb{E}_{\bar{\mathbf{x}} \sim p_{\bar{\mathbf{x}}}} [(\|\nabla_{\bar{\mathbf{x}}} D(\bar{\mathbf{x}})\|_2 - 1)^2], \quad (2.14)$$

where $\bar{\mathbf{x}}$ is a sample from $p_{\bar{\mathbf{x}}}$, a distribution defined as uniform sampling along straight interpolations between samples from the data and the genera-

tor distribution. λ is a parameter defining the gradient penalty regularization strength.

2.3.3 Inference Networks in GANs

All of the previously introduced GAN variants are generation networks that, in their standard forms, do not allow for the extraction of latent features. Damoulin et al. [48] extended the concept of GANs to Adversarially Learned Inference (ALI), where both the generation network and an inference networks are learned, still, following the two-player minimax game training paradigm. In an ALI model, visualised in Figure 2.7, the generator has two components - encoder $G_z(\mathbf{x})$ and decoder $G_x(\mathbf{z})$. $G_z(\mathbf{x})$ is trained to map data samples \mathbf{x} from the data distribution $p_{\text{data}}(\mathbf{x})$ to the latent space of \mathbf{z} . $G_x(\mathbf{z})$ is trained to map samples \mathbf{z} from the prior distribution $p_z(\mathbf{z})$ to the data space. The discriminator is trained to distinguish between joint samples $(\mathbf{x}, G_z(\mathbf{x}))$ and $(G_x(\mathbf{z}), \mathbf{z})$. The ALI [48] learning objective is defined as:

$$\min_G \max_D \mathbb{E}_{\mathbf{x} \sim p_{\text{data}}(\mathbf{x})} [\log (D(\mathbf{x}, G_z(\mathbf{x})))] + \mathbb{E}_{\mathbf{z} \sim p_z(\mathbf{z})} [\log(1 - D(G_x(\mathbf{z}), \mathbf{z}))], \quad (2.15)$$

or, in a non-saturating version:

$$\begin{aligned} & \max_D \mathbb{E}_{\mathbf{x} \sim p_{\text{data}}(\mathbf{x})} [\log (D(\mathbf{x}, G_z(\mathbf{x})))] + \mathbb{E}_{\mathbf{z} \sim p_z(\mathbf{z})} [\log(1 - D(G_x(\mathbf{z}), \mathbf{z}))] \\ & \max_G \mathbb{E}_{\mathbf{x} \sim p_{\text{data}}(\mathbf{x})} [\log (1 - (D(\mathbf{x}, G_z(\mathbf{x}))))] + \mathbb{E}_{\mathbf{z} \sim p_z(\mathbf{z})} [\log(D(G_x(\mathbf{z}), \mathbf{z}))]. \end{aligned} \quad (2.16)$$

It is worth mentioning that the ALI [48] model has also been independently proposed as Bidirectional GAN (BiGAN) by Donahue et al. [47].

Li et al. [116] pointed out that the ALI framework tends to produce unfaithful reconstructions of inputs and proposed to solve this problem with the use of the Conditional Entropy framework. In their model, ALICE (Ad-

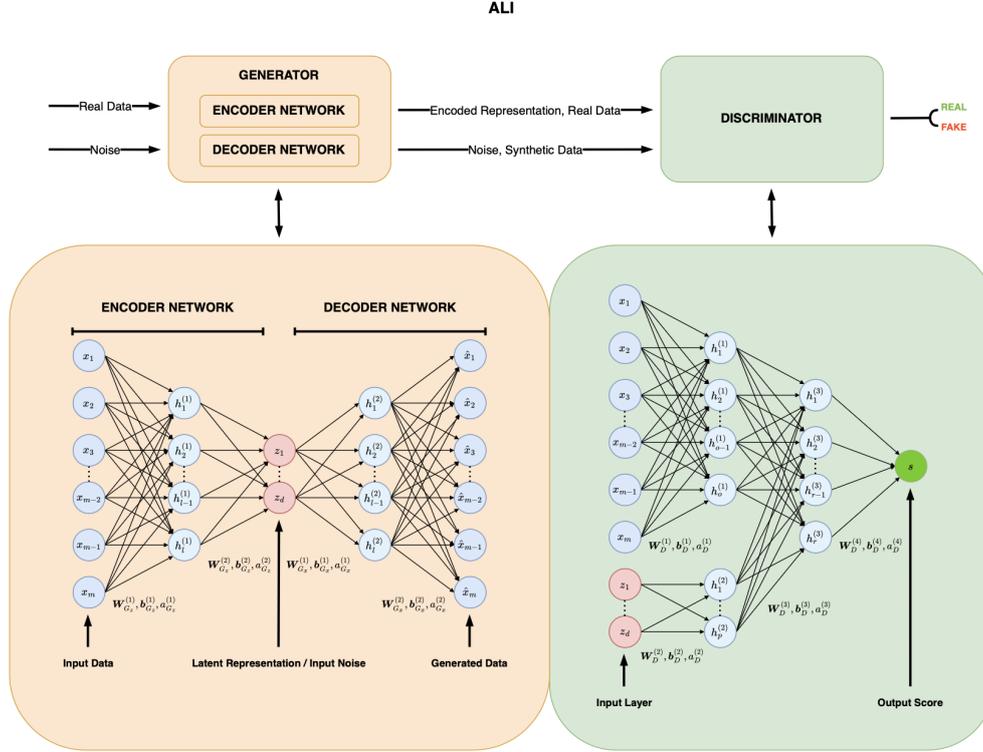


Figure 2.7: A simple ALI model. Given a noise sample \mathbf{z} , commonly obtained as $\mathbf{z} \sim \mathcal{N}(0, 1)$, the generator $G_{\mathbf{x}}(\mathbf{z})$ produces a synthetic data sample $G_{\mathbf{x}}(\mathbf{z}) = a_{G_{\mathbf{x}}}^{(2)}(\mathbf{W}_{G_{\mathbf{x}}}^{(2)}(a_{G_{\mathbf{x}}}^{(1)}(\mathbf{W}_{G_{\mathbf{x}}}^{(1)}\mathbf{z} + \mathbf{b}_{G_{\mathbf{x}}}^{(1)})) + \mathbf{b}_{G_{\mathbf{x}}}^{(2)})$, using the respective generating weights $\mathbf{W}_{G_{\mathbf{x}}}$, biases $\mathbf{b}_{G_{\mathbf{x}}}$ and activation functions $a_{G_{\mathbf{x}}}$. Given a real data sample \mathbf{x} , the generator $G_{\mathbf{z}}(\mathbf{x})$ produces its encoded representation $G_{\mathbf{z}}(\mathbf{x}) = a_{G_{\mathbf{z}}}^{(2)}(\mathbf{W}_{G_{\mathbf{z}}}^{(2)}(a_{G_{\mathbf{z}}}^{(1)}(\mathbf{W}_{G_{\mathbf{z}}}^{(1)}\mathbf{x} + \mathbf{b}_{G_{\mathbf{z}}}^{(1)})) + \mathbf{b}_{G_{\mathbf{z}}}^{(2)})$, using the respective encoding weights $\mathbf{W}_{G_{\mathbf{z}}}$, biases $\mathbf{b}_{G_{\mathbf{z}}}$ and activation functions $a_{G_{\mathbf{z}}}$. Given a (real data, encoding) pair $(\mathbf{x}, G_{\mathbf{z}}(\mathbf{x}))$ or a (synthetic data, noise) pair $(G_{\mathbf{x}}(\mathbf{z}), \mathbf{z})$ the discriminator D generates their respective scores $D(\mathbf{x}, G_{\mathbf{z}}(\mathbf{x})) = a_D^{(4)}(\mathbf{W}_D^{(4)}(a_D^{(3)}(\mathbf{W}_D^{(3)}((a_D^{(1)}(\mathbf{W}_D^{(1)}\mathbf{x} + \mathbf{b}_D^{(1)})) \oplus (a_D^{(2)}(\mathbf{W}_D^{(2)}G_{\mathbf{z}}(\mathbf{x}) + \mathbf{b}_D^{(2)}))) + \mathbf{b}_D^{(3)})) + \mathbf{b}_D^{(4)}$ and $D(G_{\mathbf{x}}(\mathbf{z}), \mathbf{z}) = a_D^{(4)}(\mathbf{W}_D^{(4)}(a_D^{(3)}(\mathbf{W}_D^{(3)}((a_D^{(1)}(\mathbf{W}_D^{(1)}G_{\mathbf{x}}(\mathbf{z}) + \mathbf{b}_D^{(1)})) \oplus (a_D^{(2)}(\mathbf{W}_D^{(2)}\mathbf{z} + \mathbf{b}_D^{(2)}))) + \mathbf{b}_D^{(3)})) + \mathbf{b}_D^{(4)}$ using the discriminator's weights \mathbf{W}_D , biases \mathbf{b}_D and activation functions a_D . The entire model is usually trained to optimize the objective from Equation 2.16.

verserially Learned Inference with Conditional Entropy), visualised in Figure 2.8, the Conditional Entropy is bound using the cycle-consistency criterion [239], enforced via a fully adversarial training. In image-generating GANs, the cycle consistency property requires that when translating an image from one domain to another, and then translating the image back to its original domain, we should be able to retrieve the original image with only minimal

changes. In ALICE, the property requires that the reconstruction of each real data point, obtained by decoding (generating) from its corresponding latent space encoding, is highly similar to the original data point. This can be achieved by adversarially training an additional reconstruction discriminator D_r to distinguish between real data samples \mathbf{x} and their reconstructions $\hat{\mathbf{x}} = G_{\mathbf{x}}(G_{\mathbf{z}}(\mathbf{x}))$. The ALICE [116] objective extends the Equation 2.16 to:

$$\begin{aligned}
& \max_D \mathbb{E}_{\mathbf{x} \sim p_{\text{data}}(\mathbf{x})} [\log(D(\mathbf{x}, G_{\mathbf{z}}(\mathbf{x})))] + \mathbb{E}_{\mathbf{z} \sim p_{\mathbf{z}}(\mathbf{z})} [\log(1 - D(G_{\mathbf{x}}(\mathbf{z}), \mathbf{z}))] \\
& \max_{D_r} \mathbb{E}_{\mathbf{x} \sim p_{\text{data}}(\mathbf{x})} [\log(D_r(\mathbf{x}, \mathbf{x}))] + \mathbb{E}_{\mathbf{x} \sim p_{\text{data}}(\mathbf{x})} [\log(1 - D_r(\mathbf{x}, G_{\mathbf{x}}(G_{\mathbf{z}}(\mathbf{x}))))] \\
& \max_G \mathbb{E}_{\mathbf{x} \sim p_{\text{data}}(\mathbf{x})} [\log(1 - (D(\mathbf{x}, G_{\mathbf{z}}(\mathbf{x}))))] + \mathbb{E}_{\mathbf{z} \sim p_{\mathbf{z}}(\mathbf{z})} [\log(D(G_{\mathbf{x}}(\mathbf{z}), \mathbf{z}))] + \\
& \quad \mathbb{E}_{\mathbf{x} \sim p_{\text{data}}(\mathbf{x})} [\log(D_r(\mathbf{x}, G_{\mathbf{x}}(G_{\mathbf{z}}(\mathbf{x})))].
\end{aligned} \tag{2.17}$$

As pointed out by Li et al. [116], an alternative to the adversarial approach for enforcing cycle-consistency would be the use of the ℓ_k losses, as previously applied in CycleGAN [239], DiscoGAN [181] and DualGAN [224]. Such solution could however result in blurry reconstructions characteristic for AEs and VAEs [107]. Furthermore, alternative approaches to the ALI(CE) learning of an inference network for a GAN also exist. As suggested by Damoulin et al. [48], one could train an encoder to reconstruct \mathbf{z} so that $\mathbb{E}_{\mathbf{z} \sim p_{\mathbf{z}}(\mathbf{z})} [\|\mathbf{z} - G_{\mathbf{z}}(G_{\mathbf{x}}(\mathbf{z}))\|] \approx 0$, which would again be limited by the blurriness resulting from the MSE loss. ALI could also be trained in two phases, training a standard GAN first, and then fitting an encoder using the adversarial ALI procedure, with a frozen pre-trained decoder (i.e. GAN generator) [48]. Such post-hoc learned inference would however limit the ability to model the interactions between the encoder and the decoder [48]. Yet another alternative would be the use of the InfoGAN model [30]. InfoGAN attempts to maximize the mutual information between data \mathbf{x} and a sub-

set \mathbf{c} of the latent code, thus allowing for a partial inference on \mathbf{z} . In an InfoGAN, however, \mathbf{z} cannot be fully inferred and the latent code \mathbf{c} is used mostly as a style manipulator, controlling e.g. image rotation or width [30]. Finally, one could resort to the use of hybrid approaches such as Adversarial Autoencoders [132] or VAE/GANs [107], which, as pointed out by [48], often exhibit the weaknesses of both (Variational) Autoencoders and GANs.

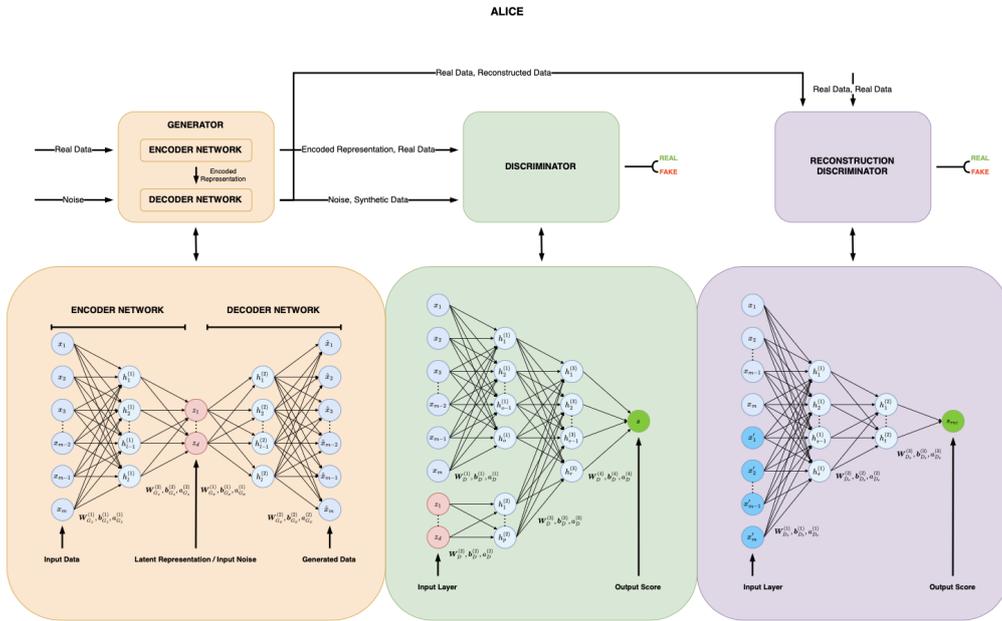


Figure 2.8: A simple ALICE model. ALICE extends ALI (Figure 2.7) with the following modifications: given an encoded representation $G_{\mathbf{z}}(\mathbf{x})$ of a real data sample \mathbf{x} , the generator $G_{\mathbf{x}}(\mathbf{z})$ produces the reconstruction of \mathbf{x} calculated as $G_{\mathbf{x}}(G_{\mathbf{z}}(\mathbf{x})) = a_{G_{\mathbf{x}}}^{(2)}(a_{G_{\mathbf{x}}}^{(1)}(\mathbf{W}_{G_{\mathbf{x}}}^{(1)} G_{\mathbf{z}}(\mathbf{x}) + \mathbf{b}_{G_{\mathbf{x}}}^{(1)}) + \mathbf{b}_{G_{\mathbf{x}}}^{(2)})$, using the respective generating weights $\mathbf{W}_{G_{\mathbf{x}}}$, biases $\mathbf{b}_{G_{\mathbf{x}}}$ and activation functions $a_{G_{\mathbf{x}}}$. And additional reconstruction discriminator D_r then takes a (real data, real data) pair (\mathbf{x}, \mathbf{x}) or a (real data, reconstruction) pair $(\mathbf{x}, G_{\mathbf{x}}(G_{\mathbf{z}}(\mathbf{x})))$ and produces their respective scores $D_r(\mathbf{x}, \mathbf{x}) = a_{D_r}^{(3)}(\mathbf{W}_{D_r}^{(3)}(a_{D_r}^{(2)}(\mathbf{W}_{D_r}^{(2)}(a_{D_r}^{(1)}(\mathbf{W}_{D_r}^{(1)}(\mathbf{x} \oplus \mathbf{x}) + \mathbf{b}_{D_r}^{(1)}) + \mathbf{b}_{D_r}^{(2)})) + \mathbf{b}_{D_r}^{(3)})$ and $D_r(\mathbf{x}, G_{\mathbf{x}}(G_{\mathbf{z}}(\mathbf{x}))) = a_{D_r}^{(3)}(\mathbf{W}_{D_r}^{(3)}(a_{D_r}^{(2)}(\mathbf{W}_{D_r}^{(2)}(a_{D_r}^{(1)}(\mathbf{W}_{D_r}^{(1)}(\mathbf{x} \oplus G_{\mathbf{x}}(G_{\mathbf{z}}(\mathbf{x}))) + \mathbf{b}_{D_r}^{(1)}) + \mathbf{b}_{D_r}^{(2)})) + \mathbf{b}_{D_r}^{(3)})$, using the respective reconstruction discriminator's weights \mathbf{W}_{D_r} , biases \mathbf{b}_{D_r} and activation functions a_{D_r} . The entire model is usually trained to optimize the objective from Equation 2.17.

2.3.4 GANs and Multiomic Data

Despite the overall popularity of GANs, very few studies apply them for (multi)omic data analysis. Ahmed et al. [3] proposed a framework in which a Wasserstein GAN is used to integrate two highly related omics types - mRNA and microRNA - for the purpose of generating synthetic data with improved predictive signals. Park, Mu and Jiao et al. [148] used a GAN as an imputation technique for missing multiomic data. Viñas et al. [199] applied GANs to generate gene expression data, Marouf and Machart et al. [134] used them to synthesize and augment single-cell RNA-seq data and Chaudhari et al. [28] to augment gene expression data. More recently, Giansanti et al. [62] used a Wasserstein GAN to generate synthetic multiomic datasets for information transfer among measured assays.

All the aforementioned methods focus on data generation, augmentation or imputation. To the best of our knowledge, the only work that has attempted to use a GAN as a latent feature model for multiomic data is that of Yang et al. [221]. Their proposed method, Subtype-GAN, employs a mixture of independent and shared layers to adversarially integrate copy number, mRNA, miRNA, DNA methylation and other omics data into a single shared latent space, that can later be used for clustering analysis. However, in Subtype-GAN the inference network is trained with the standard MSE loss and the discriminator is used only to enforce a specific prior over the latent space, making the method more of an Adversarial Autoencoder [132], than an adversarially learned inference network, thus exhibiting the limitations stemming from the MSE based training. Additionally, Subtype-GAN provides no support for the integration of mutation data [221], lacks interpretability, and does not support datasets with missing modalities.

All of the aforementioned points to a significant research gap related to

adversarially learned latent feature models for multiomic data analysis, a gap we attempt to fill in this thesis. Before we can introduce our proposed method, however, we review other challenges related to multiomic data analysis, including mixed data types and heterogeneous distributions, missing modalities, interpretability, and approaches to downstream analysis, which will be the focus of the remainder of this chapter.

2.4 Data Types and Distributions

Multiomic data is naturally heterogeneous with datasets sourced from different omics characterized by diverse distributions and data types. In the TCGA datasets, for example, continuous methylation and mRNA features are often skewed and bi-modal and CNA features are represented with binary values (recall Figure 1.10). Irrespectively of the latent feature model chosen, such differences in scales, distributions and data types need to be addressed properly, to ensure a balanced training process.

In machine learning, the most common strategy for addressing diverse continuous distributions is data standardization, which rescales each feature to have a zero mean and a unit variance. This is often followed by min-max scaling, i.e. the scaling of values to a desired range, often $[0, 1]$ or $[-1, 1]$, to ensure all features are on the same scale and carry a similar importance. Binary features are usually left unchanged, and categorical variables are often one-hot encoded as binary values. Standardization is sensitive to outliers and does not affect distribution shapes, which means that, for example, bi-modal data remains bi-modal. This may pose a significant challenge for MSE-trained models, but could, potentially, be an appropriate pre-processing choice for adversarially trained architectures. Additionally, the use of non-

Gaussian data in neural networks can potentially cause the vanishing gradient problem, which can slow down the training or prevent the network from converging to an optimal solution.

Log transformations, i.e. transformations based on replacing each feature value with its logarithm, are commonly applied to reduce the skewness of variables and make the data more Gaussian, often before feature standardization [e.g. 221]. However, as shown by Feng et al. [53], log transformations reduce skewness only if the original data roughly follows a log-normal distribution, which is often not the case. As such, in many situations, using log-transformations can actually exacerbate the skewness of the transformed features, contrary to the intent of the transformation [53]. Furthermore, when standard statistical tests are performed on the log-transformed data, their results are often no longer relevant for the original data [53].

As an alternative to standardization and log transformations, some choose to transform continuous features to follow the uniform or normal distributions using quantile or rank-based scaling [e.g. 213], again leaving binary variables unchanged and potentially min-max scaling the data. Ranking involves replacing each feature value with its rank, i.e. its relative position within a sorted dataset. Quantile transformers estimate the cumulative distribution function of a given feature and use it to map the original values to a uniform distribution. While robust to outliers, rank and quantile transformers involve non-linear transformations and can potentially distort the important correlations between variables.

Instead of data pre-processing, certain latent variable models can inherently handle heterogeneous data types by employing appropriate data modelling techniques. For instance, Nabney et al. [140] extended the Hierarchical Generative Topographic Mapping (HGTM) [135] framework by incorporating

the Latent Trait Model (LTM) [16], enabling the analysis and visualization of both discrete and continuous data. LTMs [16] define a class of latent variable models where noise models are drawn from the exponential family of distributions. Generative Topographic Mapping (GTM) [200] itself is a non-linear latent variable model in which a grid of points in the latent space is mapped onto a non-linear manifold in the data space, generated by a radial basis function neural network. Hierarchical GTMs (HGTMs) [135] build upon GTMs by introducing hierarchies of these manifolds, making them highly effective for both data visualization and clustering. This hierarchical structure enables them to capture both broad patterns and fine details within complex datasets. By incorporating LTMs, i.e. supporting noise models beyond the standard Gaussian - such as Student's t, Bernoulli, or multinomial distributions - HGTMs further enhance their ability to visualize and cluster mixed-type data in a principled and flexible way [140]. Furthermore, unlike standard hierarchical clustering algorithms, HGTMs are not constrained by heuristic distance measures and are more robust to noise [140].

Lastly, in the context of GANs specifically, an interesting method for dealing with non-Gaussian, multi-modal distributions was introduced by Xu et al. [219]. Their proposed approach, mode-specific normalization, encodes each continuous value as a one-hot vector indicating the mode, and a scalar indicating the value within the mode [219], in practice representing each multi-modal distribution as a combination of several single-mode Gaussian-like distributions. While useful for synthetic data generation, the proposed approach is not suitable for latent feature models. With mode-specific normalization, each variable is re-coded as multiple features (6 in [219]), not only contradicting the goals of dimensionality reduction but also rendering the interpretability of the method more challenging.

In addition to appropriate data pre-processing, in neural networks, adjustments to activation functions can be introduced to allow for the simultaneous reconstruction (or generation) of both discrete and continuous variables [e.g. 219]. For example, sigmoid or tanh activations are normally used for continuous features, with softmax and softsign being common choices for discrete data.

2.5 Missing Modalities

The presence of missing data is a frequent occurrence in multiomic datasets. Multiomic measurements can be missing due to a number of reasons including the errors during data acquisition or processing, the low quality, insufficient quantity or contamination of samples, the removal of outliers or low confidence data points, or simply the differences in data collection processes applied across various study centres.

In datasets composed of multiple omics sources specifically, we can distinguish between two types of missing data. First, data can be missing at feature level, i.e. values of some singular measurements may not be available for some patients. Second, entire modalities can be missing for some patients, for example RNA expressions may not be available for a fraction of samples in the dataset, as can DNA mutations.

The first scenario, common in machine learning, can be easily dealt with. Missing values can be simply imputed, using e.g. the mean imputation technique where each missing value is replaced with the variable's overall mean. Additionally, features with high percentage of missing values can be removed from the dataset, as they are not useful for analysis due to the limited amount of information they carry. This imputation or removal strategy is typically

effective when the proportion of missing values is relatively small. In the PPCG dataset specifically, the driver genes, CNA, and RNA data sources had no missing values at the feature level. Imputation at the feature level was required only for the summary measurements modality. Within this modality, no feature had more than 7.4% missing values (after feature and sample selection), making the mean imputation approach reasonable. Overall, we imputed approximately 5.02% of all summary measurement values (1,462 out of 29,138 entries), which accounted for 0.11% of all values in the entire PPCG dataset (1,462 out of 1,334,823 entries).

Handling of the second scenario, where entire modalities are missing, is more complicated and requires careful consideration. As such, we will now review the most relevant approaches to handling missing modalities in multiomic studies, largely basing our review on the summaries and references provided by Flores et al. [56].

When attempting to develop a solution for any problem involving missing values, one needs to take into account the mechanism that generated them. A commonly used classification system for such mechanisms, developed by Donald J. Rubin [164], assumes that missing data belongs to one of the three classes: missing not at random (MNAR), missing at random (MAR) or missing completely at random (MCAR). A value is MAR if the probability of it being missing depends on the observed data, but not on the missing measurement itself [60, 119]. MCAR is a special case of MAR and describes values whose missingness can be considered purely stochastic [211]. MNAR, on the other hand, describes any situation in which MAR does not apply, i.e. when the value's missingness depends on the value itself, or on some other unobserved variables. An example of an MNAR variable would be a feature that is missing due to its value being under the limit of detection or

quantification [56]. In multiomic studies, missing data is usually MNAR or MAR [56].

Perhaps the most common approach to handling missing modalities in multiomic data analysis is to perform the so-called complete case analysis, i.e. to limit the study subjects to those for whom all measured omics sources are completely observed [56]. This simple approach not only decreases the sample size and thus the power of the analysis, but can additionally bias the results if the MCAR assumptions are not met [52, 77].

Missing modality imputation offers an alternative to complete case studies, circumventing some of the issues arising from limiting the analysis to parts of the dataset only. Naive approaches, such as the mean, zero-value, or limit of detection imputation techniques can however lead to biased parameter and variability estimates [18, 109, 124, 178, 209]. Alternatively, more complex approaches such as the K-Nearest-Neighbors (KNN) [55], random forest [147], or expectation maximization [43] based imputation frameworks could potentially be applied, although these techniques are usually limited to single-modality datasets only. Yet, several approaches designed for multiomic data specifically have been proposed. For example, Eltager et al. [50] used a variant of the KNN framework to impute single-cell transposase accessibility chromatin data based on the corresponding nearest neighbours (samples with the smallest distance to a given sample) identified with single-cell transcriptomics data. Howey et al. [86] proposed yet another version of nearest neighbour imputation using Bayesian networks to decide on the features used for selecting the nearest neighbour for each sample. Methods based on the KNN framework are sensitive to noisy data, outliers and parameters (e.g. the choice of the number of nearest neighbours) and are known to perform best when the number of features is small [97], which is not usually the case

in multiomic datasets. As an alternative to KNN, transfer learning-based approaches, such as the one proposed by Zhou et al. [238], can be used, for example to impute RNA-seq data from the available DNA methylation features. Such techniques are limited in that they usually do not work both ways, and with the aforementioned approach, for example, RNA could not have been imputed using methylation [238].

Finally, as an alternative to complete case analysis or missing data imputation, several approaches robust to missing modalities have been developed. For the mixed or intermediate, latent feature based, multiomic data integration specifically, these approaches can be divided into two classes: joint-imputation methods capable of imputing the missing data during the main analysis and optimization-masking approaches that mask the missing components of the partially observed samples during optimization, still allowing them to contribute to the overall model estimates [56]. Example joint-imputation approaches include extensions of PCA for multiomic data such as MOFA (Multi-Omics Factor Analysis) [8], or its successor MOFA+ [7]. Parameters of these methods are updated using an expectation maximization based algorithm that allows for the incorporation of samples with missing modalities. Optimization-masking approaches such as MVAE (Multimodal Variational Autoencoder) [215] or DeepIMV [113] often base their architectures on the product of experts (PoE) [82] framework. The PoE technique models a probability distribution as the output of several simpler distributions and allows to make decisions without access to the full dimensionality of a problem, making it robust to missing components.

2.6 Interpretability

Method interpretability is an essential property for latent feature models when applied for multiomic data analysis. Here specifically, we define interpretability as our ability to understand, or explain, the latent features extracted from the data, as well as the patterns and relationships associated with them. For example, we would like to be able to explain how latent and observed features are correlated, and what exactly does each latent feature represent. With our proposed method, we want to model a scenario where latent variables signify the presence or absence of underlying biological processes, and observed variables display the patterns of genetic alterations associated with these processes. As such, interpretable latent-to-observed links tell us which alterations are caused by which processes. What would follow is that the model’s output and the results of any downstream analysis would be easy to explain to clinicians to amend for translation to clinical use.

Neural networks are considered to be ‘black box’ approaches, in their standard forms offering little to no interpretability. This is because they are normally composed of many hidden layers, with each unit in each hidden layer being a dense combination of all units from the previous layer. Several modifications can be introduced to improve the interpretability of the connections in neural networks. Shallow networks, i.e. those with a limited number of hidden layers, are generally easier to investigate than deeper networks. In some applications, however, shallow networks may not perform as well as deeper architectures with larger capacities. Alternatively, or additionally, to shallow architectures, penalties, such as L1 and L2 norms, can also be used to encourage sparse weight matrices in neural networks, again leading to improvements in interpretability. Possibly the most interpretable architectures are those with non-negative weights only [e.g. 12, 195, 213]. Non-negative

network-based latent feature models can be viewed as non-linear extensions of the Non-Negative Matrix Factorization [114] method and usually lead to sparse solutions in which each input feature is represented with only a small number of latent variables [213].

Two main approaches to encouraging weight non-negativity exist. First, the quadratic barrier function [144], as used for example by Nguyen et al. [143] or Woodcock et al. [213], can be applied. Formally, for a weight matrix \mathbf{W} , the quadratic barrier based regularizer that can be added to the model’s overall loss function is defined as:

$$-\frac{\alpha}{2} \sum_i \sum_j f(W_{i,j}), \quad (2.18)$$

where α is the regularisation strength and

$$f(x) = \begin{cases} x^2 & \text{if } x < 0, \\ 0 & \text{otherwise.} \end{cases} \quad (2.19)$$

As with the majority of regularizers, the use of the above requires a careful selection of the strength parameter α . Furthermore, using a quadratic barrier does not ensure that all model weights will be exclusively positive. In practice, many weights may, and most likely will, still exhibit small negative values. An alternative, parameter-free solution with full non-negativity guarantees, would involve clipping all weights in a given matrix to the $[0, +\infty)$ range in each forward optimization pass [34].

In shallow architectures, the interpretable weight matrices extracted from non-negative models could be easily investigated, for example, by using heatmaps to visualise the latent to observed connections and their respective strengths. For deeper models, invaluable for visualising such connections would be the tool known as SHapley Additive exPlanations [127], or SHAP,

for short. Based on a game theoretic approach, the SHAP framework assigns each feature an importance value for a particular prediction. A positive value indicates that the feature contributed towards increasing the prediction, and conversely, a negative value indicates that the feature contributed to decreasing the prediction. As such, SHAP values could be used to visualise how exactly each observed variable is linked to each latent variable, for each sample in the dataset. SHAP values could of course also be plotted for standard models without any weight constraints. With dense connections, such plots would however be majorly uninformative.

2.7 Subtyping and Survival Analysis

Latent feature extraction for multiomic data integration is only the first step in multiomic data studies, usually succeeded by downstream analysis. In the majority of cases, downstream analysis focuses on the identification of subtypes, or signatures, of aggressive diseases. Identifying molecular cancer subtypes, i.e. groups of patients with the same cancer type and for whom the disease manifests itself and progresses similarly, can prove invaluable for predicting therapy response and matching patients with best possible drugs and treatment options. In machine learning, subtype identification corresponds to the unsupervised analysis approach called clustering.

2.7.1 Subtyping (Clustering)

In machine learning terms, subtypes are usually referred to as clusters, and techniques used to define these clusters are called clustering algorithms. Clustering is based on partitioning a set of data points into subsets, such that the data points within the same subset are more similar to each other than

to those in other subsets, according to some chosen similarity metric. Many clustering algorithms exist, and the most common ones include K-means, hierarchical or agglomerative clustering, Gaussian Mixture Models, or DBSCAN (Density-Based Spatial Clustering of Applications with Noise). K-means clustering is specifically of interest to us for several reasons:

1. Unlike methods such as agglomerative clustering or DBSCAN, K-means supports predictions on new samples, making it straightforward to evaluate results on a holdout test set.
2. Our preliminary results obtained using the synthetic datasets showed that K-means significantly outperformed agglomerative clustering and DBSCAN in recovering true clusters. Notably, DBSCAN was highly sensitive to hyperparameters and classified a large proportion of samples as outliers.
3. KMeans clustering can be easily integrated into the latent feature extraction process [72, 217] since, unlike DBSCAN or agglomerative clustering, it is defined by a clear optimization objective rather than heuristic rules.

As such, we will now briefly summarize the algorithm and elaborate on its integration into latent feature models.

In K-means clustering, the pre-selected number of clusters K also refers to the number of centroids in the dataset, i.e. the number of locations representing cluster centres. The goal of the K-means algorithm is to assign each sample in the dataset to the cluster defined by its closest centroid, while keeping the centroids as small as possible. Formally, the algorithm can be summarized into the following four steps:

1. Choose the number of clusters K .

2. Randomly initialize the cluster centroids μ_1, \dots, μ_K : pick K random points in the dataset and use them as the initial cluster centres.
3. Assign each data point \mathbf{x}_i to its nearest cluster C_j : $j = \underset{j}{\operatorname{argmin}} \|\mathbf{x}_i - \mu_j\|^2$.
4. Update cluster centroids: set each cluster centroid to the mean of all data points assigned to said cluster $\mu_j = \frac{1}{|C_j|} \sum_{\mathbf{x}_i \in C_j} \mathbf{x}_i$

Steps 3 and 4 are then repeated until the algorithm converges, i.e. until the cluster centroids stop changing.

As initially suggested by Xie et al. [217], and later extended by Guo, et al. [72], K-means clustering can be integrated into the latent feature extraction process of neural networks, here Autoencoders, to better preserve the local structures of data generating distributions in the latent spaces. In Deep Convolutional Embedded Clustering (DCEC) [72], a clustering layer and a clustering loss are introduced to fine-tune a pre-trained Autoencoder. Formally, the clustering layer maintains cluster centres $\{\mu_j\}_1^K$ as trainable weights. Each embedded point \mathbf{z}_i (latent space representation of a data point \mathbf{x}_i) is mapped into a probabilistic cluster assignment q_i using the Student's t-distribution:

$$q_{ij} = \frac{(1 + \|\mathbf{z}_i - \mu_j\|^2)^{-1}}{\sum_j (1 + \|\mathbf{z}_i - \mu_j\|^2)^{-1}}. \quad (2.20)$$

The clustering loss, appropriately weighted, added to the overall Autoencoder's loss, is then defined as:

$$\mathcal{L}_c = \operatorname{KL}(P||Q) = \sum_i \sum_j p_{ij} \log \frac{p_{ij}}{q_{ij}}, \quad (2.21)$$

where $\operatorname{KL}(P||Q)$ denotes the KL-divergence between the distribution Q of the probabilistic cluster assignments, and a pre-defined target distribution

P . P is defined as:

$$p_{ij} = \frac{q_{ij}^2 / \sum_i q_{ij}}{\sum_j (q_{ij}^2 / \sum_i q_{ij})} \quad (2.22)$$

so that it (1) strengthens predictions, (2) gives higher importance to data points assigned with high confidence and (3) normalizes the loss contribution of each centroid so that large clusters do not distort the latent space [217].

Given a pre-trained Autoencoder, the cluster centres can be initialized by performing K-means clustering on the latent space embeddings of all data points. The clustering fine-tuning process involves updates to both the Autoencoder’s weights and the cluster centres at each optimization iteration. The target distribution P is updated every T iterations, to avoid instability [72]. The proposed regularization could be easily introduced into other network-based latent feature models.

The main challenge with K-means, and with many other clustering algorithms, is the selection of the number of clusters K . In certain applications, this can be done using domain knowledge, for example, should we want our algorithm to recover the known prostate cancer Evotypes [213] subtypes, we would naturally set K to 2. In other applications, it may be beneficial to explore and compare the results of the algorithm at different numbers of K , for example for any K between 2 and 10, and select the final result as one that provides the most useful, clinically or otherwise, partition, or one that maximizes some chosen clustering evaluation metric.

Evaluating clustering results, be it according to evaluation metrics or for their clinical relevance, is yet another challenge. Clustering is an exploratory analysis, and unlike for classification or regression problems, a single correct answer usually does not exist. The most commonly used evaluation metric is the silhouette score which measures how similar a sample is to other samples in its own cluster, as compared to other clusters. Formally, the silhouette

score $s(\mathbf{x}_i)$ of a data point \mathbf{x}_i can be calculated as:

$$s(\mathbf{x}_i) = \frac{b(\mathbf{x}_i) - a(\mathbf{x}_i)}{\max(a(\mathbf{x}_i), b(\mathbf{x}_i))}, \quad (2.23)$$

where $a(\mathbf{x}_i)$ is the average distance from \mathbf{x}_i to all other points in its own cluster, and $b(\mathbf{x}_i)$ is the smallest average distance from \mathbf{x}_i to all points in any other cluster. The overall silhouette score is obtained by averaging the individual silhouette scores for all points in the dataset. The value of the silhouette score ranges from 1 (best) to -1 (worst), with values near 0 indicating overlapping clusters.

As pointed out in the review by Ullmann et al. [196], we should also be evaluating clustering results for their reproducibility, i.e. the ability to achieve similar results, or labellings, when the algorithm is re-run with a different seed, or on different partitions of the original dataset. If subtyping results cannot be re-produced, they are likely meaningless. A metric useful for quantifying the similarity of two clustering results is the Adjusted Rand Index (ARI). For any two clustering results on the same data, ARI considers all pairs of samples and counts pairs that are assigned to the same or different clusters, adjusting the results for chance. Formally,

$$\text{ARI} = \frac{\text{RI} - \mathbb{E}(\text{RI})}{\max(\text{RI}) - \mathbb{E}(\text{RI})}, \quad (2.24)$$

where the Rand Index (RI) is defined as:

$$\text{RI} = \frac{\text{number of agreeing pairs}}{\text{number of pairs}}. \quad (2.25)$$

ARI values close to 0.0 indicate random labelling, while a value of 1.0 signifies a perfect agreement. The stability and robustness of clustering results can be improved by consensus clustering [177] which involves combining the solutions obtained from multiple clustering algorithms or differently parametrized runs of the same algorithm. Consensus clustering generally involves building

a consensus matrix in which each entry represents the proportion of time two data points are clustered together in different solutions, followed by standard clustering on the consensus matrix.

Yet another evaluation criterion, also pointed by Ullmann et al. [196], concerns itself with the associations of clusters with external variables, broadly encapsulating the usefulness of the obtained partitions in its particular domain. For example, a relevant and clinically actionable clustering result in the cancer domain would return clusters of patients with different mean survival rates.

2.7.2 Survival Analysis

Survival rates themselves lead us to the topic of survival analysis, which frequently features in multiomic data studies. The main goal of survival analysis is to analyse the time until an event occurs, for example the time until the disease becomes more aggressive, or the time before the patient dies from the disease. With survival analysis, we usually want to plot the survival curves, i.e. visualisations of survival functions describing the probability of survival beyond a certain time. For subtyping specifically, we would like the identified subtypes to have largely different survival curves.

An important distinguishment between survival analysis and simple regression tasks is the censoring of variables present in survival problems. Censoring occurs when we know that an event (e.g. death) did not happen for a certain sample until some specific point in time, however, we do not know what happened after said time point. As censored outcomes represent incomplete information, they need to be properly accounted for in survival analysis.

Survival analysis is normally carried out using linear statistical methods

such as the Kaplan-Meier [101] or the Portnoy [157] estimators. For this thesis, of specific interest is the more recently proposed neural network based approach for predicting survival by performing quantile regression on (possibly censored) data. The method, called censored quantile regression neural network (or CQRNN, for short), proposed by Pearce et al. [152] optimizes a grid of quantile outputs through an efficient implementation of the Portnoy [157] estimator in a neural-network based setting. The objective of the Portnoy estimator [157], as defined in the CQRNN [152] paper for their network, is to minimize, for a target quantile τ , the following loss:

$$\begin{aligned} \mathcal{L}_{Port.}(\theta, \mathcal{D}, \tau, \mathbf{w}, y^*) = & \sum_{i \in S_{observed}} \rho_{\tau}(y_i, \hat{y}_{i,\tau}) + \\ & \sum_{j \in S_{censored}} w_j \rho_{\tau}(y_j, \hat{y}_{j,\tau}) + (1 - w_j) \rho_{\tau}(y^*, \hat{y}_{j,\tau}) \end{aligned} \quad (2.26)$$

The loss is defined over two disjoint sets, $S_{observed}$ and $S_{censored}$, containing non-censored and censored samples only, respectively. θ are model parameters and \mathcal{D} is a dataset defined as $\mathcal{D} = \{\{\mathbf{x}_1, y_1, \Delta_1\}, \dots, \{\mathbf{x}_N, y_N, \Delta_N\}\}$, where $\mathbf{x}_i \in \mathbb{R}^M$ are the features, $y_i \in \mathbb{R}$ is the possibly censored time-to-event, and Δ_i is the event/censorship indicator. Furthermore, given a model prediction $\hat{y}_{i,\tau} = \psi_{\tau}(\mathbf{x}_i, \theta)$, the loss $\rho_{\tau}(y_i, \hat{y}_{i,\tau})$ is defined as $\rho_{\tau}(y_i, \hat{y}_{i,\tau}) = (y_i - \hat{y}_{i,\tau})(\tau - \mathbb{I}[\hat{y}_{i,\tau} > y_i])$, with $\mathbb{I}[\cdot]$ denoting the indicator function. Finally, $y^* \gg \max_i y_i$ is some large value, and $w_j = (\tau - q_j)/(1 - q_j)$ is a weighting multiplier, with q_j denoting the quantile at which the data point j was censored. In the CQRNN model, the loss from Equation 2.26 is simultaneously optimised over a grid of M increasing, evenly spaced, quantiles $\tau \in \text{grid}_{\tau}$.

In the subsequent chapters, we will explore how to utilize the CQRNN [152] network as a latent space regularizer, to encourage latent space encodings more indicative of patient survival.

Chapter 3

Proposed Method

In Chapter 2 we reviewed the common approaches to multiomic data integration and the challenges associated with this process. Motivated by the previously highlighted research gap related to the application of GANs for such purpose, in this chapter, we formally introduce iCS-GAN (integrative Cancer Subtyping with Generative Adversarial Networks) - a novel, fully interpretable, adversarially learned inference model suitable for the integration of multiomic data with missing modalities. There are multiple reasons as to why we believe in the suitability of GANs for said task, summarized below.

1. The success of GANs in the image and video domains, specifically their ability to generate synthetic media good enough to trick even human observers into believing their reality, points towards their intrinsic capabilities to discern patterns and regularities in data. As such, we believe that GANs will be an invaluable tool for extraction of meaningful latent features capturing low-level synergies and subtle relationships in multiomic data.
2. The adversarial loss used in GAN-based models should be suitable for

overcoming the limitations stemming from the use of the MSE loss in Autoencoders and VAEs. Specifically, GAN-based models should be able to better capture the non-Gaussian, possibly multi-modal, distributions of multiomic data. Furthermore, a well-trained discriminator with sufficient capacity is likely to encourage the generator to synthesize highly variable data, covering the whole range of values each given feature can take. This, in turn, should encourage the associated inference network to be sensitive to small subpopulations present in the data, for example, small subsets of patients with highly aggressive subtypes of the disease.

3. Adversarially learned inference models naturally allow for the selection of a latent space prior, an invaluable property when integrating multiomics data with the commonly chosen mixed integration strategy. While the normal or uniform distribution is usually selected as a latent space prior in GANs, we will show, that other, more suitable for biological data, priors can also be easily introduced.
4. GANs are generative by nature, and as such, with appropriate modifications, they should be highly suitable for performing inference in the presence of missing modalities, and even for imputing said modalities.

This chapter is dedicated to the introduction of our proposed GAN-based latent feature model, iCS-GAN. Specifically, in Section 3.1, basing our work on the concept of ALICE [116], we define the main single-modality building block of iCS-GAN. We additionally introduce a number of modifications designed to improve training stability and interpretability, provide support for different data types and distributions, and regularize the latent space so that the extracted latent features are more suitable for downstream analysis, in

this case disease subtyping and survival analysis. In Section 3.2, we show how to extend our model for multiomic data integration and introduce an Indian Buffet Process [68] latent space prior, so that patient samples can be represented with easily interpretable binary features that signify the presence or absence of the underlying biological processes. Finally, in Section 3.3, we extend our method to support multiomic data with missing modalities.

3.1 Single-Modality Model

In this section, we introduce the single-modality building block of iCS-GAN. As outlined previously, we base our model on the concept of ALICE (Adversarially Learned Inference with Conditional Entropy) [116], described earlier in Section 2.3.3. To recap, ALI [48], and its extension ALICE [116], augment a standard GAN [65] with an inference network to allow for the extraction of latent variables from the purely generative GAN architecture. In an ALI model, the generator has two components - encoder $G_z(\mathbf{x})$ and decoder $G_x(\mathbf{z})$. $G_z(\mathbf{x})$ is trained to map data samples \mathbf{x} from the data distribution $p_{\text{data}}(\mathbf{x})$ to the latent space of \mathbf{z} . $G_x(\mathbf{z})$ is trained to map samples \mathbf{z} from the prior distribution $p_z(\mathbf{z})$ to the data space. The discriminator is trained to distinguish between joint samples $(\mathbf{x}, G_z(\mathbf{x}))$ and $(G_x(\mathbf{z}), \mathbf{z})$. In ALICE, to allow for faithful reconstructions of encoded inputs, a further reconstruction discriminator D_r is trained to distinguish between real data samples \mathbf{x} and their reconstructions $\hat{\mathbf{x}} = G_x(G_z(\mathbf{x}))$. The overall model is trained to optimize:

$$\begin{aligned} & \max_D \mathbb{E}_{\mathbf{x} \sim p_{\text{data}}(\mathbf{x})} [\log(D(\mathbf{x}, G_z(\mathbf{x})))] + \mathbb{E}_{\mathbf{z} \sim p_z(\mathbf{z})} [\log(1 - D(G_x(\mathbf{z}), \mathbf{z}))] \\ & \max_{D_r} \mathbb{E}_{\mathbf{x} \sim p_{\text{data}}(\mathbf{x})} [\log(D_r(\mathbf{x}, \mathbf{x}))] + \mathbb{E}_{\mathbf{x} \sim p_{\text{data}}(\mathbf{x})} [\log(1 - D_r(\mathbf{x}, G_x(G_z(\mathbf{x}))))] \end{aligned}$$

$$\begin{aligned} \max_G \mathbb{E}_{\mathbf{x} \sim p_{\text{data}}(\mathbf{x})} [\log(1 - (D(\mathbf{x}, G_{\mathbf{z}}(\mathbf{x}))))] + \mathbb{E}_{\mathbf{z} \sim p_{\mathbf{z}}(\mathbf{z})} [\log(D(G_{\mathbf{x}}(\mathbf{z}), \mathbf{z}))] \\ + \mathbb{E}_{\mathbf{x} \sim p_{\text{data}}(\mathbf{x})} [\log(D_r(\mathbf{x}, G_{\mathbf{x}}(G_{\mathbf{z}}(\mathbf{x})))], \end{aligned}$$

which defines the ALICE training objective outlined earlier in Equation 2.17.

The suitability of the ALICE model for the extraction of interpretable latent features effective for disease subtyping, and from data characterized by different data types and distributions, is rather limited. As such, we propose a number of modifications to the ALICE model, designed to support the Wasserstein training procedure, interpretability, binary variables and clustering and survival regularization, reviewed below.

3.1.1 Wasserstein Training Procedure

The ALICE [116] model, as evident from Equation 2.17, is normally trained with an extension of the standard GAN [65] objective (Equation 2.8), or its non-saturating version (Equation 2.9). As Wasserstein GANs [9, 71] are generally considered superior GAN architectures, we modify the ALICE training objective to follow that of a WGAN-GP [71] from Equation 2.14. In line with this modification, we also replace both ALICE discriminators with WGAN-style critics. The new training objective can be written as:

$$\begin{aligned} \max_D \mathbb{E}_{\mathbf{x} \sim p_{\text{data}}(\mathbf{x})} [D(\mathbf{x}, G_{\mathbf{z}}(\mathbf{x}))] - \mathbb{E}_{\mathbf{z} \sim p_{\mathbf{z}}(\mathbf{z})} [D(G_{\mathbf{x}}(\mathbf{z}), \mathbf{z})] - \\ \lambda \mathbb{E}_{\bar{\mathbf{x}} \sim p_{\bar{\mathbf{x}}}(\bar{\mathbf{x}})} [(\|\nabla_{\bar{\mathbf{x}}} D(\bar{\mathbf{x}})\|_2 - 1)^2] \\ \max_{D_r} \mathbb{E}_{\mathbf{x} \sim p_{\text{data}}(\mathbf{x})} [D_r(\mathbf{x}, \mathbf{x})] - \mathbb{E}_{\mathbf{x} \sim p_{\text{data}}(\mathbf{x})} [D_r(\mathbf{x}, \hat{\mathbf{x}})] - \\ \lambda_r \mathbb{E}_{\bar{\mathbf{x}}_r \sim p_{\bar{\mathbf{x}}_r}(\bar{\mathbf{x}}_r)} [(\|\nabla_{\bar{\mathbf{x}}_r} D(\bar{\mathbf{x}}_r)\|_2 - 1)^2] \\ \min_G \mathbb{E}_{\mathbf{x} \sim p_{\text{data}}(\mathbf{x})} [D(\mathbf{x}, G_{\mathbf{z}}(\mathbf{x}))] - \mathbb{E}_{\mathbf{z} \sim p_{\mathbf{z}}(\mathbf{z})} [D(G_{\mathbf{x}}(\mathbf{z}), \mathbf{z})] + \\ \mathbb{E}_{\mathbf{x} \sim p_{\text{data}}(\mathbf{x})} [D_r(\mathbf{x}, \mathbf{x})] - \mathbb{E}_{\mathbf{x} \sim p_{\text{data}}(\mathbf{x})} [D_r(\mathbf{x}, \hat{\mathbf{x}})], \end{aligned} \tag{3.1}$$

where $\hat{\mathbf{x}}$ is the reconstruction of \mathbf{x} obtained as $\hat{\mathbf{x}} = G_{\mathbf{x}}(G_{\mathbf{z}}(\mathbf{x}))$, the terms $\lambda \mathbb{E}_{\bar{\mathbf{x}} \sim p_{\bar{\mathbf{x}}}(\bar{\mathbf{x}})}[(\|\nabla_{\bar{\mathbf{x}}} D(\bar{\mathbf{x}})\|_2 - 1)^2]$ and $\lambda_r \mathbb{E}_{\bar{\mathbf{x}}_r \sim p_{\bar{\mathbf{x}}_r}(\bar{\mathbf{x}}_r)}[(\|\nabla_{\bar{\mathbf{x}}_r} D(\bar{\mathbf{x}}_r)\|_2 - 1)^2]$ define gradient penalties for the standard and reconstruction critics, respectively, λ and λ_r are the respective gradient penalty strengths and $\bar{\mathbf{x}}$ and $\bar{\mathbf{x}}_r$ are samples from $p_{\bar{\mathbf{x}}}(\bar{\mathbf{x}})$ and $p_{\bar{\mathbf{x}}_r}(\bar{\mathbf{x}}_r)$, distributions defined as uniform sampling along straight interpolations between samples from the data and samples generated, or reconstructed, by the generator, respectively.

We decided to modify the ALICE training procedure to follow that of a WGAN-GP because of the improved training stability and other desirable properties of Wasserstein GANs, discussed earlier in Section 2.3.2. The ALICE [116] and Wasserstein GAN [9] papers were both published in 2017, which likely explains why the Wasserstein training procedure was not originally used in ALICE. Wasserstein GANs with inference components, without additional reconstruction critics, have already been studied, for example by Chen et al. [31], who demonstrated their superior performance for capturing the latent space distributions as compared to e.g. ALI [48].

3.1.2 Interpretability

To achieve interpretability with iCS-GAN, that is to be able to identify clear links between the latent and observed variables and to generally have the ability to understand the latent features extracted from the data, we introduce the following three constraints into the generator network.

1. We force all weights in the generator network to be non-negative, which is known to induce sparseness and thus boost the interpretability of the latent-to-observed feature maps [12, 195, 213]. We achieve non-negativity by clipping all weights in the generator to the $[0, +\infty)$ range in each forward pass [34]. While weight clipping can potentially cause

training instability, such as exploding or vanishing gradients [71], we did not observe such issues in the case of iCS-GAN. Furthermore, unlike the use of the quadratic barrier function [143, 144], weight clipping guarantees non-negativity and does not introduce any additional parameters. With our preliminary experiments, we observed that with the use of the quadratic barrier function, many weights still exhibited small negative values, which is consistent with previous observations [e.g. 12, 213]. Additionally, we found that selecting the appropriate regularization strength for the non-negativity penalty in the context of the unbounded GAN-based loss function was particularly challenging, leading us to choose weight clipping as our method. For further interpretability adjustments, in each forward pass of our model, the weights of the encoder are additionally re-scaled so that the sum of weights used for the calculation of each hidden unit is 1, to ensure that the magnitudes of connections are comparable between units.

2. We tie the encoder-decoder weights of G_x and G_z , which ensures that the latent-to-observed and observed-to-latent feature maps are identical. Tied weights are a common architectural choice in Autoencoders, and have also been used in the original ALI [48] model.
3. Where feasible, we keep the networks as shallow as possible.

With the above modifications, connections between the latent and observed variables can be easily explored via heatmap visualisations, or with the use of the SHAP plots [127].

The primary intent behind making our model interpretable was to ensure that its outputs and rationale can be effectively communicated to clinicians, biologists, and patients, that is, people who may not have a background in

machine learning. While weight heatmaps and SHAP scores are standard tools for ML practitioners and computer scientists, it is important to assess their usefulness for these stakeholders. Through consultations, we found that heatmaps offer an intuitive way to visualize results: deeper colours naturally indicate stronger connections between latent and observed variables, highlighting the importance of specific features. SHAP plots, on the other hand, were initially less intuitive for stakeholders and required additional explanation. By providing background on SHAP by explaining how these plots illustrate the direction and magnitude of each observed variable’s influence on a latent variable for individual patient samples, we were able to make them more accessible and useful. In practice, heatmaps are often the preferred visualization method due to their simplicity. However, for deeper models where heatmaps may not be sufficient, SHAP plots offer a more detailed breakdown of variable importance.

3.1.3 Support for Mixed Data Types and Distributions

Measurements characterised by heterogeneous distributions, data types, scales and biases are a natural occurrence in multiomic datasets (for examples, see Figures 1.2 or 1.10). While skewed and multimodal distributions with potential outliers can pose a significant challenge for models optimized with MSE-based loss functions [6, 92, 235], we hypothesised that the adversarial loss characteristic to GAN-based models should be naturally suitable for such an application. Following from this assumption, to handle different data types, scales and distributions in our model, we propose, depending on the specific data type, the use of the pre-processing techniques and activation functions outlined next.

- For continuous (or count) features we apply feature normalization (i.e.

re-scaling to zero mean and unit variance) and min-max scale each variable to the $[0, 1]$ range. We use sigmoid activation in the output layer of the generator for such features.

- For categorical variables, we use one-hot encoding and proceed as with binary features.
- For binary variables, we do not apply any pre-processing, however, we do assume that binary values are encoded as $\{0, 1\}$, unlike in some machine learning applications where $\{-1, 1\}$ are used. For a direct generation of almost discrete values, to prevent the critics from easily rejecting the generated samples based on the data-type mismatch, we propose to use the modified softsign activation function. The softsign function, previously used in GANs for example by Cao et al. [23], is defined as:

$$\text{softsign}(h) = \frac{h}{|h| + \epsilon}, \quad (3.2)$$

where ϵ is a smoothing constant. As such, softsign is a smooth, easy to optimize over, approximation of the sign function:

$$\text{sign}(h) = \frac{h}{|h|}. \quad (3.3)$$

In this case, $\text{sign}(h) = 1$ if $h > 0$ and $\text{sign}(h) = -1$ if $h < 0$. Similarly, for softsign, assuming $\epsilon > 0$, we have $\text{softsign}(h) \approx 1$ if $h \geq 0$ and $\text{softsign}(h) \approx -1$ if $h < 0$. For our specific application, to allow for the generation of binary values in the $\{0, 1\}$ range, and with different activation probabilities, we replace the standard softsign function with its modified version:

$$\text{mod_softsign}(h) = \left(1 + \frac{h - \beta}{|h - \beta| + \epsilon}\right) / 2, \quad (3.4)$$

where β is a thresholding parameters. Now, assuming $\epsilon > 0$, we have $mod_softsign(h) \approx 1$ if $h \geq \beta$ and $mod_softsign(h) \approx 0$ if $h < \beta$.

With the above modifications, iCS-GAN should be able to handle heterogeneous multiomic cancer data in a consistent fashion.

As our pre-processing approach transforms all data into a non-negative format, it is important to consider the effects of this transformation, particularly when combined with the previously introduced non-negativity constraints. While non-negativity enhances the overall interpretability of the method, one could argue that this may hinder the ability to capture inhibition or down-regulation [e.g. 203]. Non-negative weights enforce positivity in learned transformations, meaning that connections between nodes in the network can only contribute additive effects to the activations. However, biological inhibition or down-regulation typically involves suppressive effects. For instance, a gene mutation might silence a signalling pathway [e.g. 104, 162], and a transcription factor could inhibit the expression of certain genes [e.g. 108, 128]. Although this inhibition-related limitation that stems from our focus on interpretability is apparent, it should be considered with two important points in mind: (1) non-linear activation functions (e.g. sigmoid or ReLU) combined with batch normalization can help approximate inhibitory effects by introducing thresholding and saturation, and (2) in biological systems, the down-regulation of one pathway usually co-occurs the up-regulation of another. Therefore, while specific inhibitory effects may not be captured, the primary cause can likely still be identified, even with non-negative weights and data representations.

3.1.4 Clustering Regularization

To encourage latent space encodings that are more suitable for clustering analysis, we follow the DCEC clustering regularization approach [72, 217] reviewed earlier in Section 2.7.1. Similarly to Guo et al. [72], we first pre-train iCS-GAN as described so far, and then fine-tune it with an additional clustering layer and the associated clustering loss. As a remainder, the clustering layer maintains cluster centres $\{\mu_j\}_1^K$ as trainable weights and each embedded point $\mathbf{z}_i = G_{\mathbf{x}}(\mathbf{x}_i)$ is mapped into a probabilistic cluster assignment q_i using the Student’s t-distribution (Equation 2.20):

$$q_{ij} = \frac{(1 + \|\mathbf{z}_i - \mu_j\|^2)^{-1}}{\sum_j (1 + \|\mathbf{z}_i - \mu_j\|^2)^{-1}}.$$

Given (Equation 2.22):

$$p_{ij} = \frac{q_{ij}^2 / \sum_i q_{ij}}{\sum_j (q_{ij}^2 / \sum_i q_{ij})},$$

the appropriately weighted clustering loss (Equation 2.21):

$$\mathcal{L}_c = \text{KL}(P||Q) = \sum_i \sum_j p_{ij} \log \frac{p_{ij}}{q_{ij}},$$

can be added to the overall loss of iCS-GAN based on Equation 3.1, to fine tune the model.

Guo et al. [72] empirically set the clustering loss weight to 0.1 to prevent excessive disruption of the latent space, which could lead to high correlations among the extracted latent variables. For iCS-GAN, due to the unbounded nature of the adversarial loss as compared to the loss of a standard Autoencoder, we recommend testing different regularization strengths. The optimal choice should yield the best clustering performance while minimizing the increase in correlation between the latent variables.

Furthermore, Guo et al. [72], pre-trained their model for a set number of 200 epochs before clustering fine-tuning. The fine-tuning process ended when the cluster assignments changed for less than 0.1% of all samples between two successive updates of the target distribution P . Since GAN-based methods typically require many optimization steps and cancer datasets often contain a limited number of samples, we recommend pre-training iCS-GAN until the initial loss converges and fine-tuning until the clustering loss stabilizes.

In summary, to introduce the clustering loss into iCS-GAN, we recommend the following training pipeline:

1. **Initial Training:** Train iCS-GAN without the clustering loss until the model converges.
2. **Determine the Optimal Number of Clusters (K):**
 - Apply KMeans clustering to the embeddings generated by iCS-GAN for various values of K , $K \in \{2, 3, 4, 5\}$.
 - Select the optimal K based on a chosen evaluation metric. In our case, we select K that maximizes clustering stability across multiple runs with different seeds, measured using the Adjusted Rand Index (ARI).
3. **Fine-Tuning with Clustering Loss:** Fine-tune iCS-GAN with clustering loss using only the optimal K identified in step 2. Use the corresponding embedding to initialize cluster centres. Experiment with different regularization strengths ($\{0, 5, 10, 15, 20, 25\}$).
4. **Final Model Selection:** Choose the regularization strength that produces the best results according to the chosen evaluation metric. Once

again, we use the ARI to select the regularization strength that maximizes clustering stability across multiple runs.

3.1.5 Survival Regularization

To encourage the latent space encodings of iCS-GAN to be more indicative of patient survival, and to potentially direct the subsequent subtyping analysis towards subtypes with significant survival differences, we propose a survival regularization technique based on the CQRNN [152] network described earlier in Section 2.7.2.

Recall the CQRNN loss from Equation 2.26:

$$\begin{aligned} \mathcal{L}_{Port.}(\theta, \mathcal{D}, \tau, \mathbf{w}, y^*) &= \sum_{i \in \mathcal{S}_{observed}} \rho_{\tau}(y_i, \hat{y}_{i,\tau}) + \\ &\sum_{j \in \mathcal{S}_{censored}} w_j \rho_{\tau}(y_j, \hat{y}_{j,\tau}) + (1 - w_j) \rho_{\tau}(y^*, \hat{y}_{j,\tau}). \end{aligned}$$

Formally, given a training dataset $\mathcal{D} = \{\{\mathbf{x}_1, y_1, \Delta_1\}, \dots, \{\mathbf{x}_N, y_N, \Delta_N\}\}$, where $\mathbf{x}_i \in \mathbb{R}^M$ are the features, $y_i \in \mathbb{R}$ is the possibly censored time-to-event, and Δ_i is the event/censorship indicator, and a frozen CQRNN network M_{CQRNN} pre-trained on \mathcal{D} , we suggest the following regularization component to be included in the overall iCS-GAN’s training objective:

$$\mathcal{L}_{surv}(\alpha_{surv}, \hat{\mathcal{D}}, M_{CQRNN}) = \alpha_{surv} \cdot \mathcal{L}_{Port.}(\theta, \hat{\mathcal{D}}, \tau, \mathbf{w}, y^*), \quad (3.5)$$

where $\theta, \tau \in \text{grid}_{\tau}$, y^* and \mathbf{w} are the parameters, quantiles, large value y^* and weighting multipliers, respectively, of the network M_{CQRNN} trained on \mathcal{D} , α_{surv} is the survival penalty strength and, with G denoting the generator of iCS-GAN, $\hat{\mathcal{D}} = \{\{G_{\mathbf{x}}(G_{\mathbf{z}}(\mathbf{x}_1)), y_1, \Delta_1\}, \dots, \{G_{\mathbf{x}}(G_{\mathbf{z}}(\mathbf{x}_N)), y_N, \Delta_N\}\}$. If survival outcomes are not available for all samples in the dataset, the penalty can be calculated using only the samples with available data.

The regularizer defined above uses a survival network trained on the real data to penalize iCS-GAN for generating data reconstructions from which patient survival cannot be predicted accurately. The reconstruction of each data point is generated directly from its latent space encoding, hence this should, in turn, encourage the latent space encodings to be more indicative of patient survival.

An alternative approach worth considering would involve connecting the CQRNN network directly to the embedding layer of iCS-GAN and optimizing both models together. However, we found this solution impractical due to the CQRNN network’s sensitivity to parameter choices. Ensuring optimal performance would demand an extensive hyper-parameter search, which, given the computational expensiveness of GAN-based models, would be challenging in practice.

We recommend treating the survival regularizer as an optional component of the iCS-GAN framework, as its effectiveness is closely tied to the performance of the CQRNN network itself. If the survival network’s predictive accuracy is poor, the regularizer is unlikely to have the intended effect.

3.2 Multiomic Data Integration Model

As defined so far, iCS-GAN is not explicitly suitable for deployment on datasets with multiple modalities, that is, for multiomic data integration. However, early integration could be performed by concatenating all available omics sources into a single matrix that would later serve as an input to our model. In Section 2.1, we reviewed the pitfalls associated with the early integration strategy. To avoid them, in this section, we propose two modifications designed to extend our single-modality model with support for the

mixed multiomic data integration strategy: the use of shared and independent embedding layers and layer-wise pre-training. Additionally, we show how our method can support an Indian Buffet Process [68] binary latent space prior for the final shared embedding space.

3.2.1 Shared and Independent Layers

Rather than concatenating all omics sources into a single matrix used as an input to iCS-GAN, similarly to the approach chosen by Yang et al. in subtypeGAN [221], we propose to use a mixture of independent and shared layers in both the encoder and decoder networks of the generator, thus introducing the support for the mixed multiomic data integration strategy. With this modification, we first generate an independent latent space for each modality, and then use the latent modality-specific encodings to infer the final latent space shared amongst multiple modalities. As it is possible to specify a pre-selected dimensionality for each independent layer, our approach should allow us to easily remove the integration challenge associated with size differences between modalities. Figure 3.1 visualises an example iCS-GAN’s multiomic data integration generator with shared and independent layers.

3.2.2 Layer-Wise Pre-Training

With the mixed integration strategy, supported by the use of the shared and independent layers in our model, we can additionally largely avoid the pitfalls stemming from various data types and distributions present in different modalities by applying the layer-wise pre-training strategy, common in Autoencoders. By pre-training each modality-specific layer independently as a stand-alone iCS-GAN model, before shared latent space training and final tuning, we can enforce a pre-defined prior over the independent latent

iCS-GAN: MULTIMODAL GENERATOR

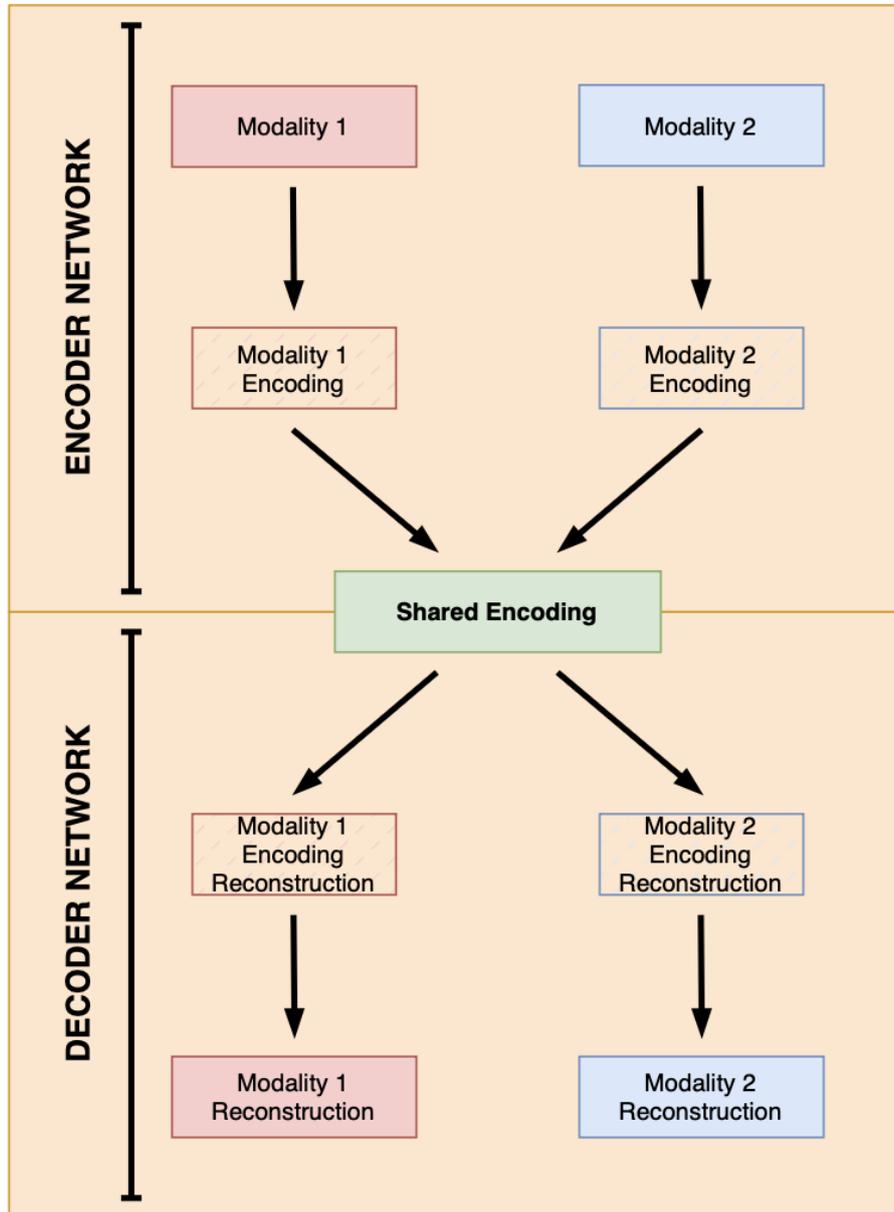


Figure 3.1: Schematic representation of an example multiomic generator for the integration of 2 modalities with iCS-GAN.

spaces, thus alleviating the issues arising from the heterogeneities present among different omics sources.

The layer-wise pre-training strategy for iCS-GAN (Figure 3.2) can be described as follows. Suppose we are given a dataset \mathbf{X} that can be described as a multiomic collection of N_M single-omic datasets $\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_{N_M}$, i.e. $\mathbf{X} = [\mathbf{X}_1 \mathbf{X}_2 \dots \mathbf{X}_{N_M}]$. For now, we can assume that \mathbf{X} is complete, that is, all of the N_M modalities are available for all samples in the dataset. Let G , D and D_r denote the generator, critic and reconstruction critic of the final latent feature model \mathbf{L} we want to train on \mathbf{X} , respectively. For the ease of exposition, we can assume that the encoder of G comprises of one input and one independent hidden (encoding) layer per modality, and a single joint encoding layer shared amongst modalities. Letting $G^{(A)}$, $D^{(A)}$ and $D_r^{(A)}$ denote the generator, critic and reconstruction critic, respectively, of a latent feature model \mathbf{L}_A trained on dataset A , and assuming that $G^{(A)}$ is a shallow 3-layer (input-encoding-output) network, the layer-wise pre-training of \mathbf{L} proceeds as follows.

iCS-GAN: Multimodal Training Procedure

1. Train one modality-specific latent feature model $\mathbf{L}_{\mathbf{X}_i}$, each with its own generator $G^{(\mathbf{X}_i)}$, critic $D^{(\mathbf{X}_i)}$ and reconstruction critic $D_r^{(\mathbf{X}_i)}$, for each of the N_M available modalities, $i = 1, \dots, N_M$. Note that this step allows us to enforce a specific prior, e.g. the normal distribution, over modality-specific encodings $G_z^{(\mathbf{X}_i)}(\mathbf{X}_i)$. As such, each modality, irrespectively of its data type or distribution, can be represented in a uniform format easing further integration. Discard $D^{(\mathbf{X}_i)}$ and $D_r^{(\mathbf{X}_i)}$.
2. Using the concatenated modality-specific encodings $\mathbf{E} = [\mathbf{E}_1 \mathbf{E}_2 \dots \mathbf{E}_{N_M}]$, obtained as $\mathbf{E}_i = G_z^{(\mathbf{X}_i)}(\mathbf{X}_i)$, for $i = 1, \dots, N_M$, train a single integra-

tion latent feature model \mathbf{L}_E consisting of a generator $G^{(E)}$, critic $D^{(E)}$ and reconstruction critic $D_r^{(E)}$. Discard $D^{(E)}$ and $D_r^{(E)}$.

3. Use the pre-trained modality-specific generators $G^{(\mathbf{X}_i)}$, for $i = 1, \dots, N_M$, and the integration generator $G^{(E)}$ to initialize the independent and shared layers, respectively, of the generator G for the final multiomic latent feature model \mathbf{L} . Keeping G frozen, train D and D_r until convergence, to avoid imbalanced training stemming from the use of a pre-trained generator and non-trained critics.

4. Unfreeze the generator G and train \mathbf{L} until convergence.

The integrated encodings obtained from G can then be used for the final multiomic subtyping analysis. Additionally, as the layer-wise pre-training involves training an independent latent feature model for each of the available modalities, we can visualise and explore modality-specific subtyping results with the use of the encodings produced by $G^{(\mathbf{X}_i)}$. Note that the clustering and survival regularizers introduced in Sections 3.1.4 and 3.1.5 can be applied in Steps 1 and 4 of the above training procedure.

3.2.3 Indian Buffet Process Prior

So far, we did not explicitly discuss the choice of a latent space prior for our proposed latent feature extraction model. In GANs [9, 65, 71] and also ALI [48] and ALICE [116], the noise vector \mathbf{z} is usually sampled from the normal distribution $\mathcal{N}(0, 1)$, potentially scaled to the $[0, 1]$ range. For iCS-GAN, we propose the use of the more domain-relevant Indian Buffet Process [68] prior, introduced earlier in Section 2.2.1. We hypothesise that such prior would allow us to interpret the extracted latent encodings as binary indicators signifying the presence or absence of the underlying biological

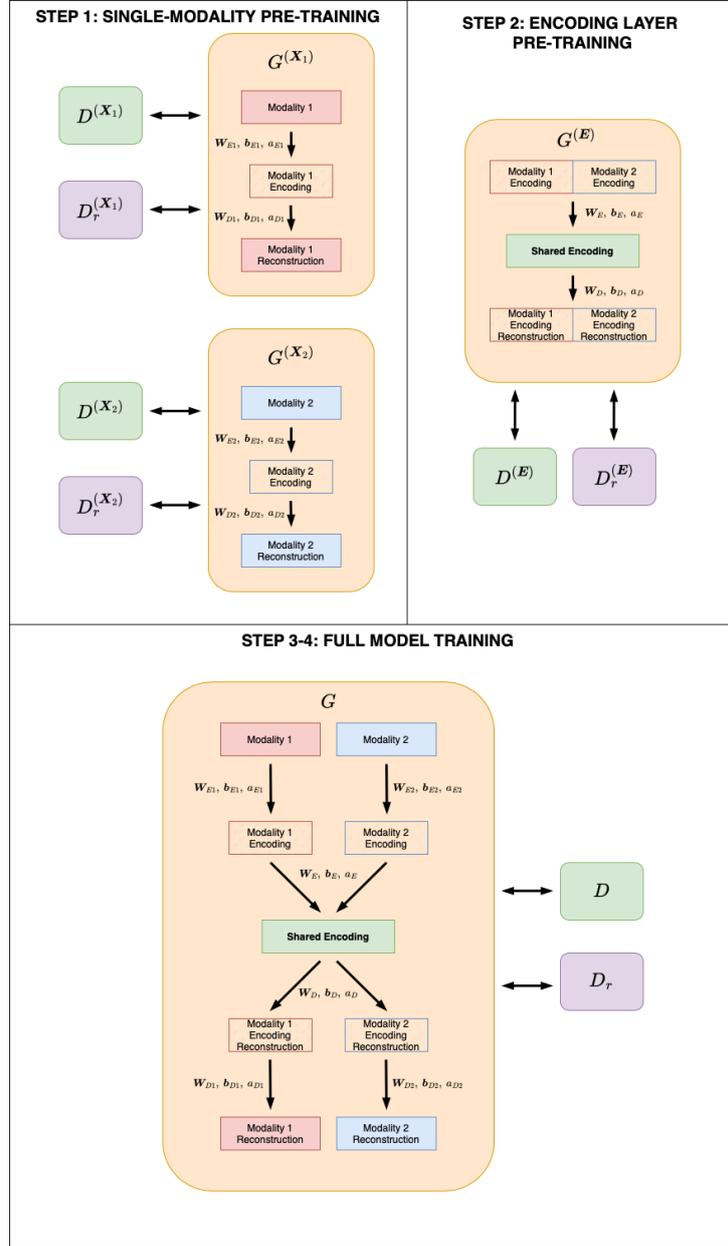


Figure 3.2: Visualisation of the layer-wise pre-training procedure, as applied in a bi-modal iCS-GAN. Step 1 involves training a stand-alone iCS-GAN models $L_{X_1} = \{G^{(X_1)}, D^{(X_1)}, D_r^{(X_1)}\}$ and $L_{X_2} = \{G^{(X_2)}, D^{(X_2)}, D_r^{(X_2)}\}$ independently for datasets X_1 and X_2 . In Step 2, we train a stand-alone integration model $L_E = \{G^{(E)}, D^{(E)}, D_r^{(E)}\}$ using as inputs the encodings obtained from $G_z^{(X_1)}$ and $G_z^{(X_2)}$. In Steps 3 and 4, the final latent feature extraction model $L = \{G, D, D_r\}$ is trained with parameters initialized based on $G^{(X_1)}$, $G^{(X_2)}$ and $G^{(E)}$.

processes, representing various disease signatures and patterns of genetic alterations.

To introduce in IBP-like prior into iCS-GAN, we propose the following procedure. Recall the stick-breaking construction process [183] for an IBP matrix Z with N rows (patient samples), K columns (latent features) and the expected number of α features active for each sample. To avoid challenges associated with optimization in the presence of fully binary variables, rather than sampling the feature activation probability μ_k for each of the K columns as $\mu_k \sim \text{Beta}(\frac{\alpha}{K}, 1)$, and then sampling each entry z_{ik} of Z as $z_{ik} | \mu_k \sim \text{Bernoulli}(\mu_k)$, we propose the following approximate smooth construction:

1. Sample the feature activation probability μ_k for each of the K columns independently as $\mu_k \sim \text{Beta}(\frac{\alpha}{K}, 1)$.
2. Set b_k as the $100 \cdot (1 - \mu_k)$ percentile of the normal distribution.
3. Sample the noise vector \mathbf{z} as $\mathbf{z} \sim \mathcal{N}(0, 1)$.
4. Given a small positive smoothing constant ϵ , activate each entry z_k in \mathbf{z} with the *mod_softsign* function (Equation 3.4), i.e.

$$z_k \leftarrow \left(1 + \frac{z_k - b_k}{|z_k - b_k| + \epsilon} \right) / 2.$$

The above gives us $z_k \approx 1$ with probability μ_k and $z_k \approx 0$ with probability $1 - \mu_k$.

With the proposed sampling procedure, the noise vector \mathbf{z} used in iCS-GAN smoothly approximates samples from an IBP prior. To match the prior with the generated latent space encodings, the *mod_softsign* function can be used as an activation function for the last encoding layer in the generator’s encoder.

We suggest the IBP prior is used in iCS-GAN in the final shared multiomic encoding layer only (Steps 2-4 of the Multimodal Training Procedure), with normal prior, scaled to the $[0, 1]$ range, utilized for the modality-specific latent spaces (Step 1 of the Multimodal Training Procedure), to prevent a decline in the quality of synthetic data or real data reconstructions caused by the excessive use of binary values.

3.3 Multiomic Data Integration with Missing Modalities

As described so far, iCS-GAN assumes that all data sources are available for all samples considered, i.e. allows for the limited, as discussed in Section 2.5, complete case analysis only. As evident from Section 1.2, the samples in the PPCG dataset are severely incomplete, and the complete case analysis would limit our study to 453 patients only. To avoid this, and to make full use of the layer-wise pre-training strategy, and the generative nature of GANs, we propose to modify the Multimodal Training Procedure from Section 3.2 as described below.

Suppose we are given a dataset \mathbf{X} that can be described as a multiomic collection of N_M single-omic datasets $\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_{N_M}$, but let us assume that \mathbf{X} is not complete, i.e. not all of the N_M modalities are available for all samples in \mathbf{X} . Let $\mathbf{X}^c = [\mathbf{X}_1^c \mathbf{X}_2^c \dots \mathbf{X}_{N_M}^c]$ denote the complete subset of \mathbf{X} , i.e. the part of \mathbf{X} where all modalities are available for all samples. It holds that $\mathbf{X}_1^c \subseteq \mathbf{X}_1, \mathbf{X}_2^c \subseteq \mathbf{X}_2, \dots, \mathbf{X}_{N_M}^c \subseteq \mathbf{X}_{N_M}$. Let G, D and D_r denote the generator, critic and reconstruction critic of the final latent feature model \mathbf{L} we want to train on \mathbf{X} , respectively. For the ease of exposition, we can assume that the encoder of G comprises of one input and one independent

hidden (encoding) layer per modality, and a single joint encoding layer shared amongst modalities. Letting $G^{(A)}$, $D^{(A)}$ and $D_r^{(A)}$ denote the generator, critic and reconstruction critic, respectively, of a latent feature model \mathbf{L}_A trained on dataset A , and assuming that $G^{(A)}$ is a shallow 3-layer (input-encoding-output) network, the layer-wise pre-training of \mathbf{L} proceeds as follows.

iCS-GAN: Multimodal Training Procedure with Missing Modalities

1. Train one modality-specific latent feature model $\mathbf{L}_{\mathbf{X}_i}$, each with its own generator $G^{(\mathbf{X}_i)}$, critic $D^{(\mathbf{X}_i)}$ and reconstruction critic $D_r^{(\mathbf{X}_i)}$, for each of the N_M available modalities, $i = 1, \dots, N_M$. For this step, utilize all the data available for a given modality, i.e. train $\mathbf{L}_{\mathbf{X}_i}$ on all samples from \mathbf{X}_i even if $\mathbf{X}_i \neq \mathbf{X}_i^c$.
2. Using the concatenated modality-specific complete encodings $\mathbf{E}^c = [\mathbf{E}_1^c \mathbf{E}_2^c \dots \mathbf{E}_{N_M}^c]$, obtained as $\mathbf{E}_i^c = G_z^{(\mathbf{X}_i)}(\mathbf{X}_i^c)$, for $i = 1, \dots, N_M$, train a single integration latent feature model $\mathbf{L}_{\mathbf{E}^c}$ consisting of a generator $G^{(\mathbf{E}^c)}$, critic $D^{(\mathbf{E}^c)}$ and reconstruction critic $D_r^{(\mathbf{E}^c)}$. Discard $D^{(\mathbf{E}^c)}$ and $D_r^{(\mathbf{E}^c)}$.
3. Use the pre-trained modality-specific generators $G^{(\mathbf{X}_i)}$, $i = 1, \dots, N_M$, and the integration generator $G^{(\mathbf{E}^c)}$ to initialize the independent and shared layers, respectively, of the generator G for the final multiomic latent feature model \mathbf{L} . Keeping G frozen, train D and D_r until convergence, to avoid imbalanced training stemming from pre-trained generator and non-trained critics, using samples from \mathbf{X}^c only.
4. Unfreeze the generator G and train \mathbf{L} until convergence using samples from \mathbf{X}^c only.

Integrated encodings for samples from \mathbf{X}^c can be easily obtained from G and used for downstream analysis. To extract the embedded representations of samples from $\mathbf{X} \setminus \mathbf{X}^c$, we suggest the imputation procedure described below.

Without the loss of generality, let us assume that $N_M = 2$ and let \mathbf{X}_1^{nc} and \mathbf{X}_2^{nc} denote the subsets of \mathbf{X} where only the first and second modalities are available, respectively, that is $\mathbf{X}_1 = \mathbf{X}_1^c \cup \mathbf{X}_1^{nc}$ and $\mathbf{X}_2 = \mathbf{X}_2^c \cup \mathbf{X}_2^{nc}$. Modality 1 and 2 specific encodings $\mathbf{z}_c^{(1)}$ and $\mathbf{z}_c^{(2)}$ of samples from \mathbf{X}^c can be obtained by projecting \mathbf{X}_1^c and \mathbf{X}_2^c onto the modality-specific encoding layers of G . Similarly, the final integrative encoding \mathbf{z}_c of \mathbf{X}_c can be obtained by projecting $\mathbf{z}_c^{(1)} \mathbf{z}_c^{(2)}$ onto the shared encoding layer of G . Encodings $\mathbf{z}_{nc1}^{(1)}$ for modality 1 for samples from \mathbf{X}_1^{nc} and $\mathbf{z}_{nc2}^{(2)}$ for modality 2 for samples from \mathbf{X}_2^{nc} can likewise be obtained by projecting samples from \mathbf{X}_1^{nc} and \mathbf{X}_2^{nc} onto the encoding layers of G for modalities 1 and 2, respectively. An imputation procedure is needed to retrieve the encodings $\mathbf{z}_{nc1}^{(2)}$ for modality 2 for samples from \mathbf{X}_1^{nc} and encodings $\mathbf{z}_{nc2}^{(1)}$ for modality 1 for samples from \mathbf{X}_2^{nc} .

Given the relatively small dimensionalities of $\mathbf{z}_c^{(1)}$ and $\mathbf{z}_c^{(2)}$, we can fit two KNN classifiers, KNN_1 on $\mathbf{z}_c^{(1)}$ and KNN_2 on $\mathbf{z}_c^{(2)}$. We propose to impute $\mathbf{z}_{nc2}^{(1)}$ as the average encoding in $\mathbf{z}_c^{(1)}$ of the nearest-neighbours of $\mathbf{z}_{nc2}^{(2)}$ identified by KNN_2 and $\mathbf{z}_{nc1}^{(2)}$ as the average encoding in $\mathbf{z}_c^{(2)}$ of the nearest-neighbours of $\mathbf{z}_{nc1}^{(1)}$ identified by KNN_1 . The final integrative encodings of samples from \mathbf{X}_1^{nc} and \mathbf{X}_2^{nc} , from which the missing data can be straightforwardly reconstructed using the decoding component of G , can then be obtained by projecting $\mathbf{z}_{nc1}^{(1)} \mathbf{z}_{nc1}^{(2)}$ and $\mathbf{z}_{nc2}^{(1)} \mathbf{z}_{nc2}^{(2)}$ onto the shared encoding layer of G .

Chapter 4

Testing and Validation

This chapter focuses on the thorough testing and validation of our proposed latent feature model, iCS-GAN, introduced earlier in Chapter 3. We designed several experiments to assess the method’s ability to integratively analyse multiomic data, even when some modalities are missing. Additionally, we evaluated the method’s interpretability, as well as its performance in recovering true clusters, reconstructing data, generating synthetic data, and aligning the latent space encodings with the desired latent space prior.

The structure of this chapter is as follows. In Section 4.1, we describe the metrics selected to evaluate the performance of iCS-GAN. Section 4.2 provides an overview of the experiments conducted, explaining their objectives and design. In Section 4.3, we present the results of testing iCS-GAN on synthetic datasets S1-S5. In Section 4.4, we validate the single-modality version of iCS-GAN on TCGA datasets, perform an ablation analysis on the model’s components, and provide a comparison with Autoencoders [165] and VAEs [103]. Section 4.5 extends the validation and ablation analysis to multiple modalities. In Section 4.6, we evaluate the multiomic subtypes identified by iCS-GAN on TCGA datasets by comparing them to the currently accepted

gold standard. Finally, in Sections 4.7 and 4.8, we explore the performance of iCS-GAN in the presence of the survival regularization or missing modalities.

4.1 Evaluation Metrics

To formally evaluate the testing and validation experiments conducted in this chapter, we have selected several evaluation metrics, summarized as follows:

- **RMSE** (Root Mean Squared Error) - this measures the reconstruction error between the original data and the model's reconstructions. A lower RMSE indicates better data reconstruction.
- **WD-Rec** (Wasserstein Distance for Reconstruction) - this metric calculates the average Wasserstein distance between the original data features and the reconstructed data features. It assesses how well the reconstructed data distributions match the original ones, with a lower value indicating a better match.
- **WD-Synth** (Wasserstein Distance for Synthetic Data) - this measures the average Wasserstein distance between the original data features and the generated synthetic data. A lower value indicates that the generated data distributions more closely match the original data distributions.
- **WD-Latent** (Wasserstein Distance for Latent Features) - this evaluates how well the distribution of the model's encoded features aligns with the desired prior distribution. A lower Wasserstein distance indicates a better match between the encoded features and the prior.
- **1-ARI** (One Minus Adjusted Rand Index) - the Adjusted Rand Index

(ARI) measures the similarity between two clustering results. For synthetic data experiments, where true labels are available, we calculated ARI to evaluate how well the model recovers the true clusters present in the data, by comparing the true labels to those obtained by applying KMeans clustering on the model’s encodings. In unsupervised experiments involving real-world datasets, we calculated ARI to compare the clustering results from two independent model runs (using different random seeds) to assess the method’s stability. A higher ARI indicates better results, but for consistency with other metrics, we report 1-ARI, where a lower value indicates better performance.

In addition to these metrics, the method’s interpretability, distribution matching properties, and the quality and separability of uncovered clusters was assessed visually using heatmaps, histograms, and UMAP visualizations. For UMAP visualizations specifically, we used the default UMAP parameters without exploring the parameter space, as our aim was to quickly illustrate the results. While parameter tuning could potentially improve the visualizations, it was not performed in this context.

4.2 Experimental Design

We have conducted a number of experiments designed to investigate the performance of iCS-GAN.

1. Experiment 1 (Section 4.3) - the first experiment involved assessing the ability of iCS-GAN in recovering true, possibly small, clusters present in the data and handling different data types and distributions. To this end, we applied the single-modality version of iCS-GAN on synthetic datasets S1-S4, and multi-modal version on synthetic dataset S5.

2. Experiment 2 (Section 4.4) - the second experiment considered validating the single-modality version of iCS-GAN on 15 TCGA datasets (5 datasets, 3 modalities each), assessing its interpretability and ability to uncover clusters, or subtypes, with significant survival differences, and comparing its performance with that of Autoencoders and VAEs which are commonly used for latent feature extraction. Additionally, the experiment involved performing an ablation analysis in which the following components were independently removed from iCS-GAN: Wasserstein training procedure (W/O WD), interpretability constraints (W/O INT), modifications for binary variables (W/O BIN) and clustering regularization (W/O CL), to justify the changes introduced to the ALICE [116] model during the development of iCS-GAN.
3. Experiment 3 (Section 4.5) - the third experiment involved exploring the performance of iCS-GAN for multiomic data integration on the 5 multi-modal TCGA datasets. We also performed an ablation analysis, by independently removing the following, multiomic data integration-specific, components of iCS-GAN: modality-specific independent layers (W/O IND), layer-wise pre-training (W/O PT), clustering regularization over both shared and independent layers (W/O CL), clustering regularization over independent layers only (W/O CLI) and the Indian Buffet Process prior (W/O IBP). In Section 4.6, building on the results from Experiment 3, we validated the multiomic subtypes identified by iCS-GAN against the gold standard currently accepted in the literature.
4. Experiment 4 (Section 4.7) - the fourth experiment explored the effects the survival regularization component has on iCS-GAN and its ability

to guide the discovery of subtypes with significant survival differences, using the TCGA KIRC dataset.

5. Experiment 5 (Section 4.8) - the final experiment validated the performance of iCS-GAN in the presence of missing modalities, which we explored by masking parts of TCGA datasets and comparing the new subtyping results with those obtained with modality-complete data in Experiment 3.

Each of the experiments outlined above was repeated 5 times, each time with a different random seed. Any metrics reported are test set averages of 5 runs.

GAN training is naturally computationally expensive. For reference, training iCS-GAN with default parameters and a pre-set clustering regularization strength on TCGA datasets using CPU only took between 1 h 2 min (BLCA) and 3 h 19 min (COAD) on a 2020 MacBook Pro, equipped with 32 GB of RAM and a 2.3 GHz Quad-Core Intel Core i7 processor. As such, no parameter search for iCS-GAN was performed for any of our experiments, except for the tuning of the clustering regularization strength and, where utilized, parameters for the survival regularization network. Instead, we trained iCS-GAN with our suggested default parameters, outlined in Appendix A. For each experiment, iCS-GAN was first trained until the training reconstruction error has converged. The model was then fine-tuned with the clustering regularization, until the clustering loss has stabilized. In each case, we considered regularization strengths of $\{5, 10, 15, 20, 25\}$ and selected number of clusters (2-10) and the regularization strength that resulted in the highest training Adjusted Rand Index (ARI) between the true labels and the labels predicted by applying the KMeans algorithm on the obtained

latent space encodings (synthetic data experiments), or the number of clusters and the regularization strength that resulted in the highest clustering stability, as measured by ARI on the training set only, among the 5 runs of the model (TCGA experiments). Where used, the survival regularization network was fully trained before its introduction into iCS-GAN, with 5-fold cross-validation parameter search performed over the survival network’s parameters.

For single-modality experiments, the Generator of iCS-GAN was a shallow tied non-negative network with a single hidden (encoding) layer, and both Critics had deep non-constrained architectures. We selected the normal distribution ($\mathcal{N}(0, 1)$), re-scaled to the $[0, 1]$ range, as the latent space prior. The number of units in the latent space (the number of latent variables) was empirically set to 20 for all experiments involving synthetic datasets, and to the square root of the number of observed features in each modality for TCGA experiments. For multi-modal experiments, the Generator of iCS-GAN was again a tied non-negative network with a single hidden (encoding) layer for each modality, pre-trained as for single-modality experiments, and an additional single multiomic encoding layer with an Indian Buffet Process prior with 20 latent variables and feature activation parameter $\alpha = 10$. Both Critics had deep, non-constrained architectures.

Where comparison with AEs and VAEs was conducted, we trained two versions of an Autoencoder: a standard version, i.e. non-constrained (AE), and non-negative version (nnAE); and a standard VAE (VAE). To ensure a fair comparison, the architectures of both Autoencoders mimicked that of the single-modality iCS-GAN’s Generator. The same approach was applied for VAE, with additional mean and standard deviation layers. Full 5-fold cross-validation parameter search was performed over batch size, learning

rate, L1 and L2 penalty strengths, and the number of training epochs. In each case, we selected the parameters that minimized the cross-validation reconstruction error, as measured by RMSE.

Details of model architectures, implementations, and parameter searches can be found in Appendix A.

4.3 Synthetic Datasets: Testing

Table 4.1 summarizes the performance of iCS-GAN on synthetic datasets S1-S5. The method accurately recovered the true clustering labels, even under challenging conditions such as highly-skewed distributions (S2), small cluster sizes (S3), binary variables (S4 and S5), and multimodal data (S5). Additionally, iCS-GAN maintained low reconstruction errors and effectively matched the desired data and prior distributions. Distributions of selected data features and their reconstructions obtained with iCS-GAN are visualised in Figure 4.1. UMAP visualisations of the results can be found in Figure 4.2.

Dataset	RMSE	WD-Rec	WD-Synth	WD-Latent	1-ARI
S1	0.08 (0.07-0.09)	0.03 (0.03-0.03)	0.03 (0.03-0.03)	0.04 (0.03-0.04)	0.20 (0.16-0.27)
S2	0.09 (0.08-0.09)	0.05 (0.04-0.05)	0.04 (0.03-0.06)	0.05 (0.05-0.06)	0.27 (0.25-0.30)
S3	0.11 (0.09-0.15)	0.04 (0.03-0.04)	0.04 (0.03-0.07)	0.05 (0.03-0.06)	0.20 (0.07-0.41)
S4	0.12 (0.09-0.15)	0.05 (0.05-0.06)	0.07 (0.07-0.08)	0.08 (0.06-0.09)	0.24 (0.21-0.28)
S5	0.13 (0.12-0.14)	0.05 (0.05-0.05)	0.06 (0.05-0.06)	0.06 (0.04-0.07)	0.23 (0.15-0.33)

Table 4.1: Performance of iCS-GAN on synthetic datasets S1-S5. All reported values represent the averages from 5 runs of the model. The values in parentheses indicate the minimum and maximum values observed for each metric across these 5 runs.

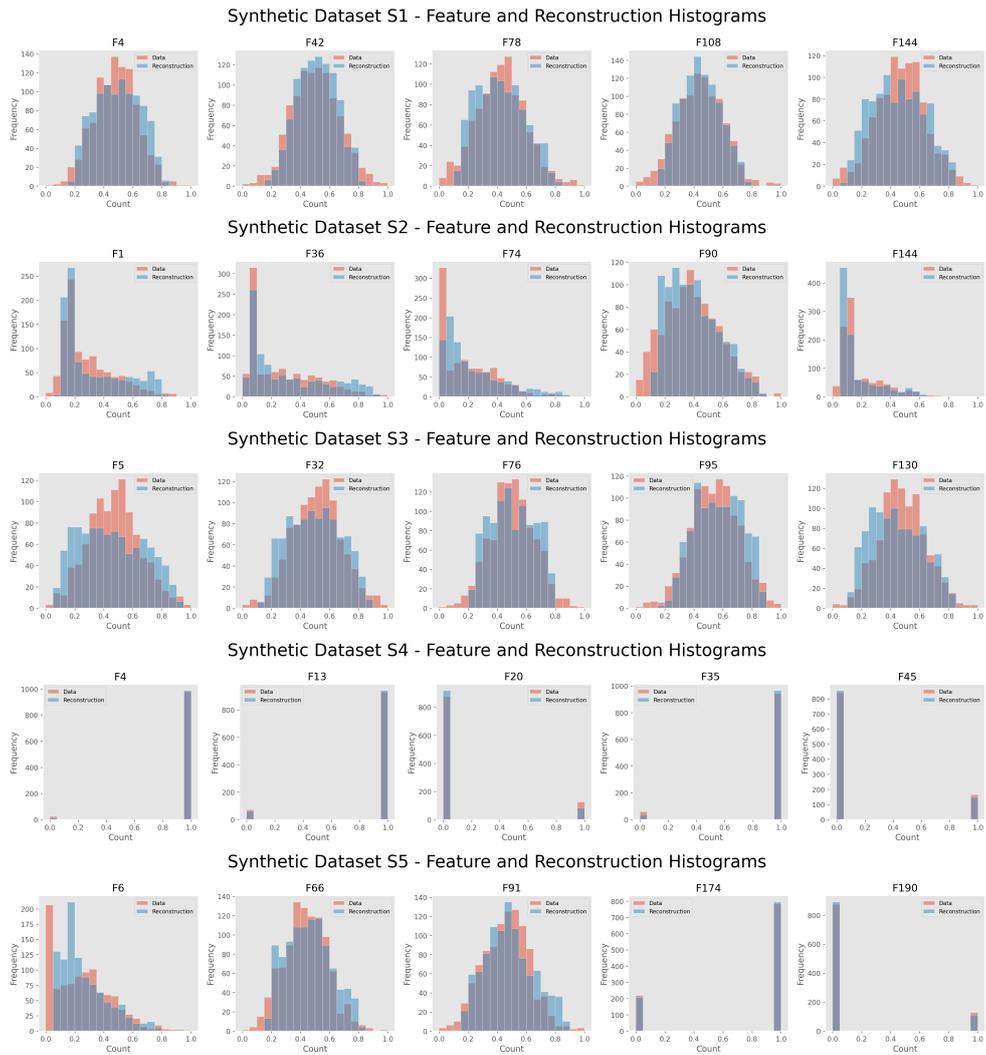


Figure 4.1: Frequency histograms for selected features from datasets S1-S5 (top to bottom), including reconstructions obtained with iCS-GAN.

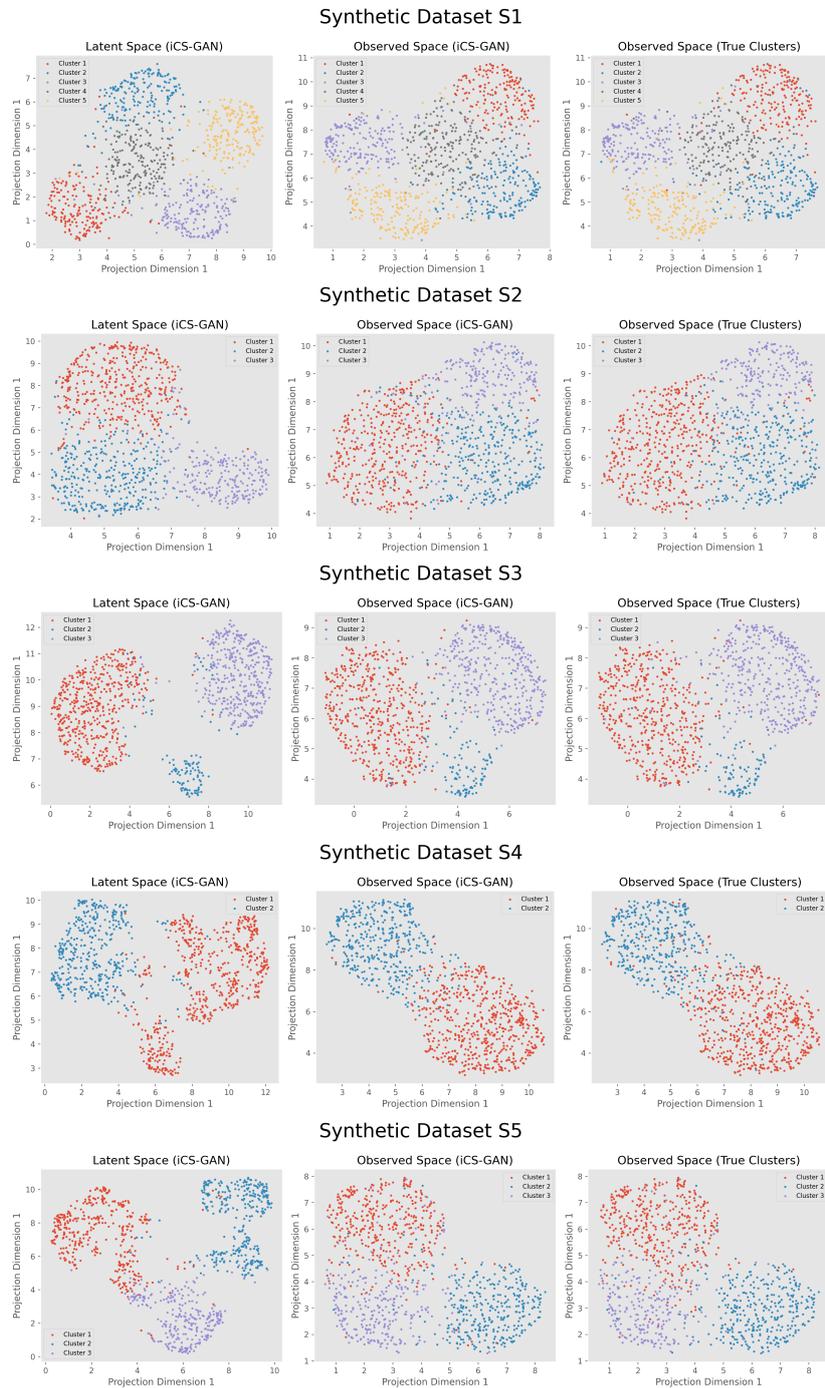


Figure 4.2: UMAP plots showing the results of applying iCS-GAN to synthetic datasets S1-S5 (top to bottom). The discovered clusters are displayed in both the latent space of iCS-GAN (left) and the observed space (middle), with the true clusters in the observed space (right) included for comparison.

4.4 TCGA: Single-Modality Validation

Table 4.2 summarizes the performance of iCS-GAN as applied on 15 single-modality TCGA datasets. The subtyping results are illustrated in Figures 4.3 - 4.7. iCS-GAN consistently performed well across all datasets, providing stable subtyping results and identifying subtypes with significant survival differences on BRCA methylation, KIRC methylation and mRNA, BLCA mRNA and HNSC mRNA datasets. This was achieved while effectively managing the varying data types and distributions (Figure 4.8) and maintaining full interpretability (Figure 4.9).

Compared to standard Autoencoders and VAEs (Table 4.3), while attaining full interpretability, iCS-GAN better matched the desired prior distributions and resulted in more stable subtyping results. VAE achieved superior reconstruction error. Both iCS-GAN and VAE well matched the data distributions with their reconstructions, with respect to which Autoencoders performed rather poorly. However, the synthetic data generation quality was significantly better for iCS-GAN than for VAE. iCS-GAN strongly outperformed the interpretable non-negative Autoencoder across all metrics. For comparison with Figure 4.9, example input to latent feature correspondence heatmap for AE, nnAE and VAE are given in Figure 4.10. Except for nnAE, these are largely non-interpretable.

Finally, we performed an ablation analysis on the components of single-modality iCS-GAN (Table 4.4). Each modification introduced to the ALICE model [116] in Section 3.1 led to an improvement in the overall performance of iCS-GAN. Specifically, adapting the training procedure to match that of a Wasserstein GAN [9, 71] reduced the reconstruction error, enhanced distribution alignment for both reconstructed and synthetically generated data, improved prior matching, and increased clustering stability. The addition of

interpretability constraints notably enhanced clustering stability with minimal impact on other performance metrics. Incorporating the modification for binary variables further optimized distribution alignment for both reconstructed and synthetic data. Lastly, the clustering regularization contributed to improved clustering stability with only a minor increase in reconstruction error.

Dataset	Modality	RMSE	WD-Rec	WD-Synth	WD-Latent	1-ARI
BRCA	Methylation	0.24 (0.23-0.24)	0.09 (0.07-0.09)	0.11 (0.10-0.12)	0.07 (0.05-0.08)	0.17 (0.10-0.25)
	mRNA	0.17 (0.16-0.17)	0.08 (0.06-0.09)	0.06 (0.06-0.07)	0.08 (0.05-0.11)	0.04 (0.02-0.08)
	CNA	0.39 (0.30-0.45)	0.14 (0.08-0.19)	0.10 (0.09-0.11)	0.07 (0.05-0.09)	0.20 (0.11-0.36)
KIRC	Methylation	0.22 (0.21-0.22)	0.06 (0.06-0.07)	0.09 (0.09-0.10)	0.05 (0.04-0.06)	0.27 (0.11-0.37)
	mRNA	0.18 (0.17-0.19)	0.09 (0.07-0.10)	0.09 (0.07-0.10)	0.05 (0.04-0.06)	0.16 (0.08-0.27)
	CNA	0.22 (0.22-0.23)	0.06 (0.06-0.06)	0.08 (0.07-0.08)	0.11 (0.10-0.11)	0.18 (0.05-0.28)
BLCA	Methylation	0.25 (0.24-0.26)	0.12 (0.10-0.13)	0.12 (0.11-0.12)	0.07 (0.06-0.08)	0.30 (0.22-0.36)
	mRNA	0.23 (0.21-0.26)	0.14 (0.12-0.16)	0.12 (0.11-0.14)	0.06 (0.06-0.08)	0.18 (0.10-0.27)
	CNA	0.42 (0.35-0.47)	0.16 (0.11-0.19)	0.12 (0.11-0.12)	0.07 (0.05-0.11)	0.21 (0.14-0.28)
COAD	Methylation	0.24 (0.23-0.25)	0.07 (0.06-0.08)	0.13 (0.12-0.13)	0.04 (0.04-0.05)	0.25 (0.14-0.33)
	mRNA	0.20 (0.19-0.22)	0.10 (0.09-0.11)	0.10 (0.09-0.11)	0.04 (0.04-0.05)	0.10 (0.04-0.14)
	CNA	0.33 (0.32-0.34)	0.09 (0.08-0.10)	0.11 (0.10-0.11)	0.05 (0.04-0.06)	0.21 (0.11-0.33)
HNSC	Methylation	0.23 (0.22-0.24)	0.07 (0.06-0.08)	0.11 (0.11-0.12)	0.05 (0.04-0.06)	0.25 (0.11-0.39)
	mRNA	0.20 (0.19-0.21)	0.09 (0.08-0.10)	0.10 (0.09-0.11)	0.05 (0.05-0.06)	0.16 (0.05-0.25)
	CNA	0.36 (0.32-0.41)	0.10 (0.09-0.13)	0.10 (0.09-0.11)	0.05 (0.04-0.07)	0.16 (0.08-0.19)

Table 4.2: Performance of iCS-GAN on single-modality TCGA datasets. All reported values represent the averages from 5 runs of the model. The values in parentheses indicate the minimum and maximum values observed for each metric across these 5 runs.

Method	RMSE	WD-Rec	WD-Synth	WD-Latent	1-ARI
iCS-GAN	0.26	0.10	0.10	0.06	0.19
AE	0.27	0.25	N/A	0.08	0.21
nnAE	0.28	0.25	N/A	0.08	0.45
VAE	0.21	0.09	0.16	0.14	0.21

Table 4.3: Performance of iCS-GAN on TCGA single modality datasets as compared to Autoencoders, non-negative Autoencoders and VAEs. The rows represent the results obtained with each method averaged across the 15 TCGA datasets (5 runs each). Full results for each dataset, with error bounds, can be found in Appendix B.

Method	RMSE	WD-Rec	WD-Synth	WD-Latent	1-ARI
iCS-GAN	0.26	0.10	0.10	0.06	0.19
W/O WD	+0.02	+0.01	+0.04	+0.02	+0.16
W/O INT	-0.02	-0.01	-0.02	-0.01	+0.40
W/O BIN	0.00	+0.02	+0.03	0.00	+0.01
W/O CL	-0.02	0.00	0.00	0.00	+0.08

Table 4.4: TCGA single modality ablation study results. The top row represents the results obtained with iCS-GAN averaged across the 15 TCGA datasets (5 runs each). The remaining rows describe the average change in each metric, for iCS-GAN with removed: Wasserstein training procedure (W/O WD), interpretability constraints (W/O INT), binary modification (W/O BIN) and clustering regularization (W/O CL). Positive change value corresponds to an increase in a given metric (i.e. worse performance). Full results for each dataset, with error bounds, can be found in Appendix B.

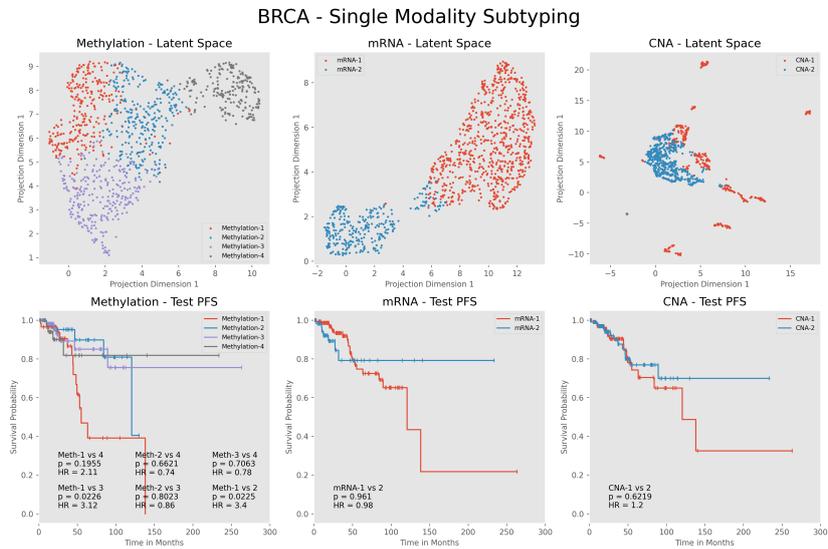


Figure 4.3: Single modality subtyping results as obtained with iCS-GAN on TCGA BRCA methylation (left), mRNA (middle) and CNA (right) datasets. UMAP visualisation of the subtypes discovered in the latent space (top) and the associated test-set PFS survival curves (bottom). For each subtype comparison, we provide the hazard ratio (HR) and p value calculated with the Cox proportional hazard test.

KIRC - Single Modality Subtyping

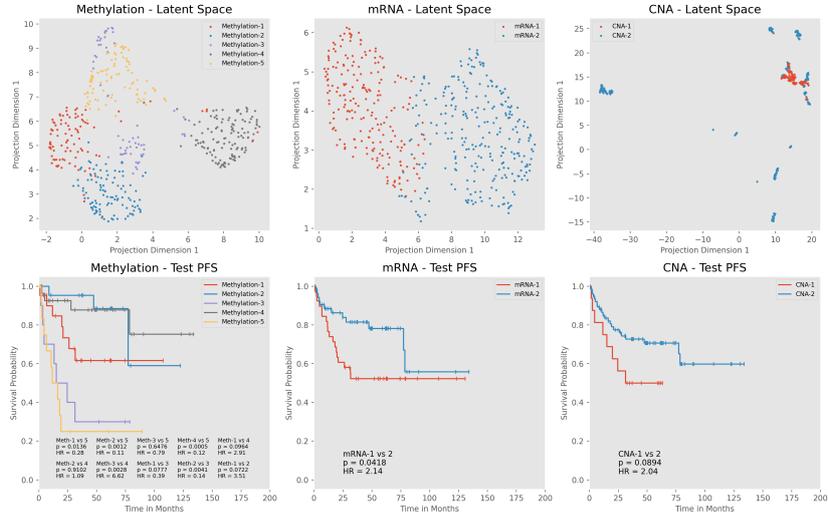


Figure 4.4: Single modality subtyping results as obtained with iCS-GAN on TCGA KIRC methylation (left), mRNA (middle) and CNA (right) datasets. UMAP visualisation of the subtypes discovered in the latent space (top) and the associated test-set PFS survival curves (bottom). For each subtype comparison, we provide the hazard ratio (HR) and p value calculated with the Cox proportional hazard test.

BLCA - Single Modality Subtyping

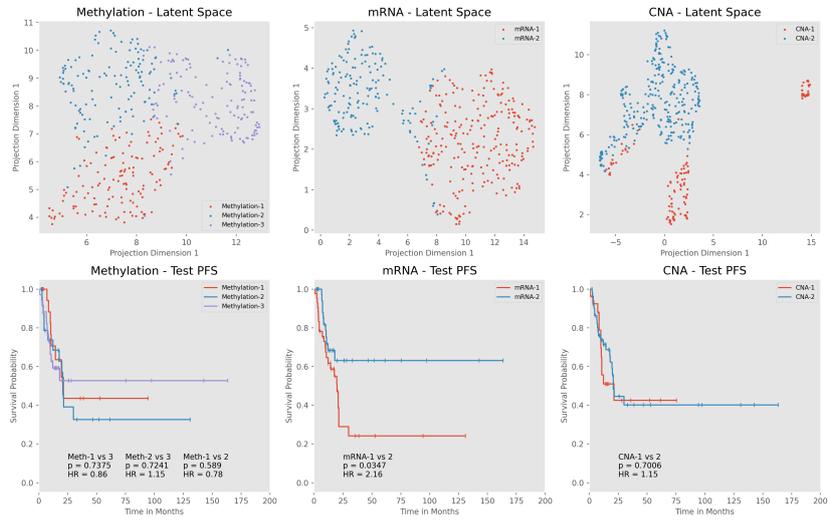


Figure 4.5: Single modality subtyping results as obtained with iCS-GAN on TCGA BLCA methylation (left), mRNA (middle) and CNA (right) datasets. UMAP visualisation of the subtypes discovered in the latent space (top) and the associated test-set PFS survival curves (bottom). For each subtype comparison, we provide the hazard ratio (HR) and p value calculated with the Cox proportional hazard test.

COAD - Single Modality Subtyping

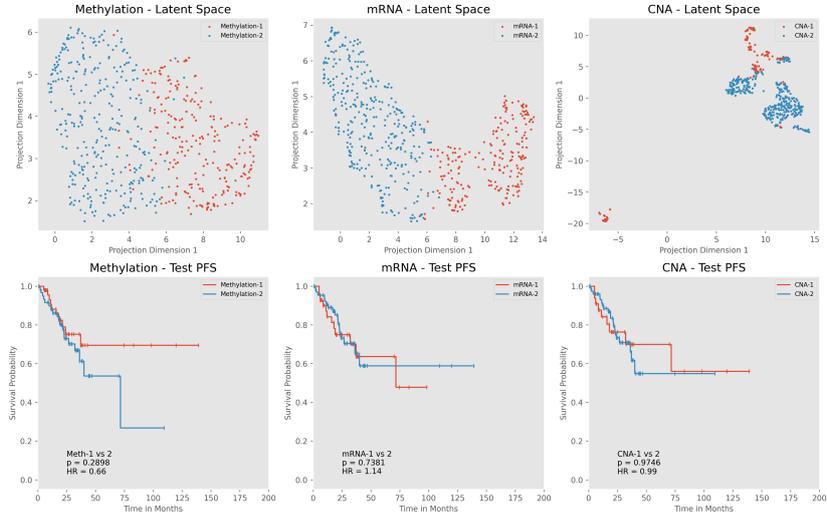


Figure 4.6: Single modality subtyping results as obtained with iCS-GAN on TCGA COAD methylation (left), mRNA (middle) and CNA (right) datasets. UMAP visualisation of the subtypes discovered in the latent space (top) and the associated test-set PFS survival curves (bottom). For each subtype comparison, we provide the hazard ratio (HR) and p value calculated with the Cox proportional hazard test.

HNSC - Single Modality Subtyping

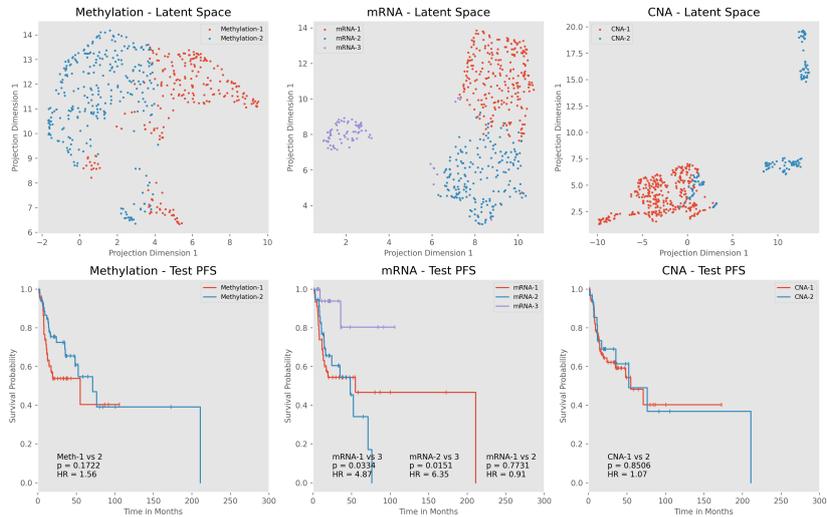


Figure 4.7: Single modality subtyping results as obtained with iCS-GAN on TCGA HNSC methylation (left), mRNA (middle) and CNA (right) datasets. UMAP visualisation of the subtypes discovered in the latent space (top) and the associated test-set PFS survival curves (bottom). For each subtype comparison, we provide the hazard ratio (HR) and p value calculated with the Cox proportional hazard test.

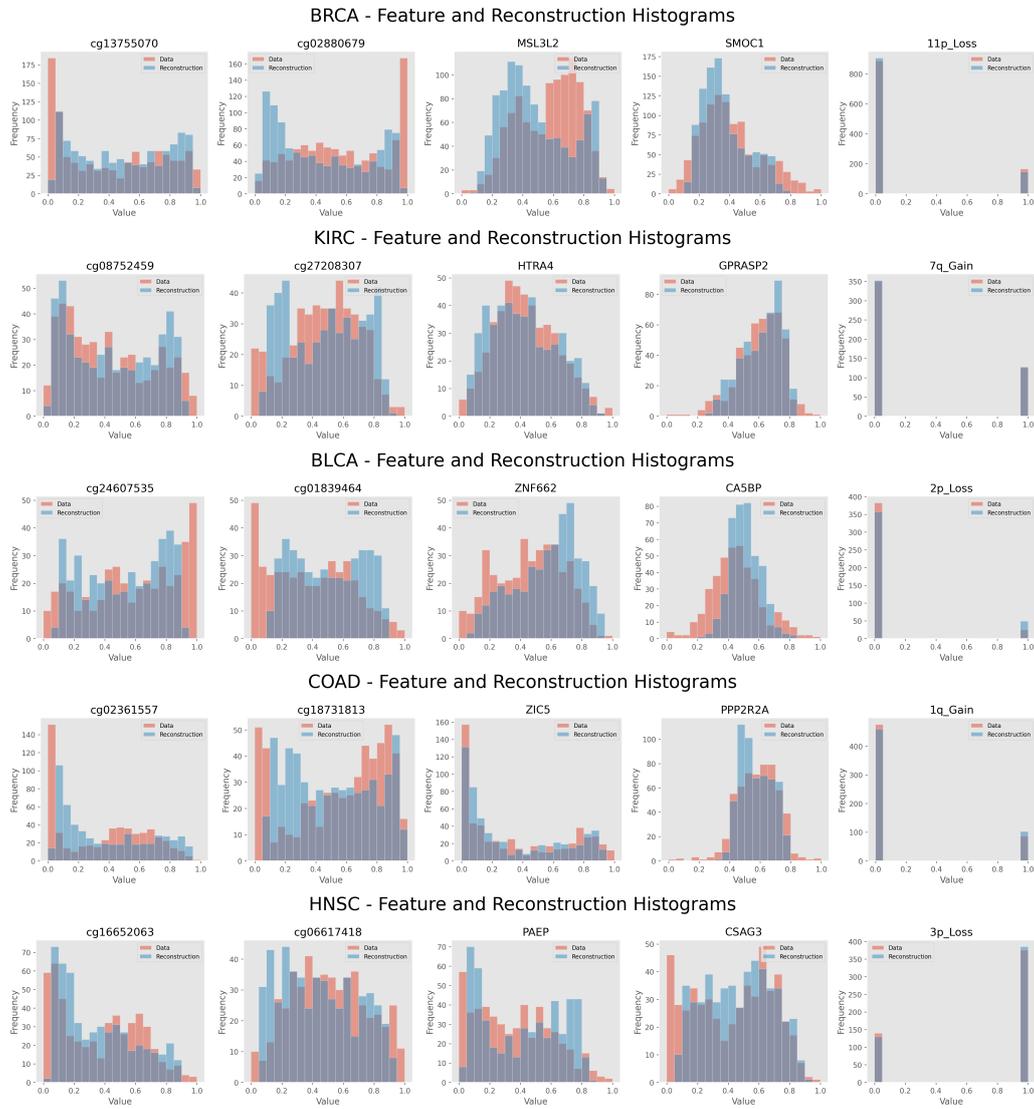


Figure 4.8: Frequency histograms for selected methylation (first two columns), mRNA (next two columns) and CNA (last column) features from BRCA, KIRC, BLCA, COAD and HNSC datasets (top to bottom), including reconstructions obtained with single-modality iCS-GAN.

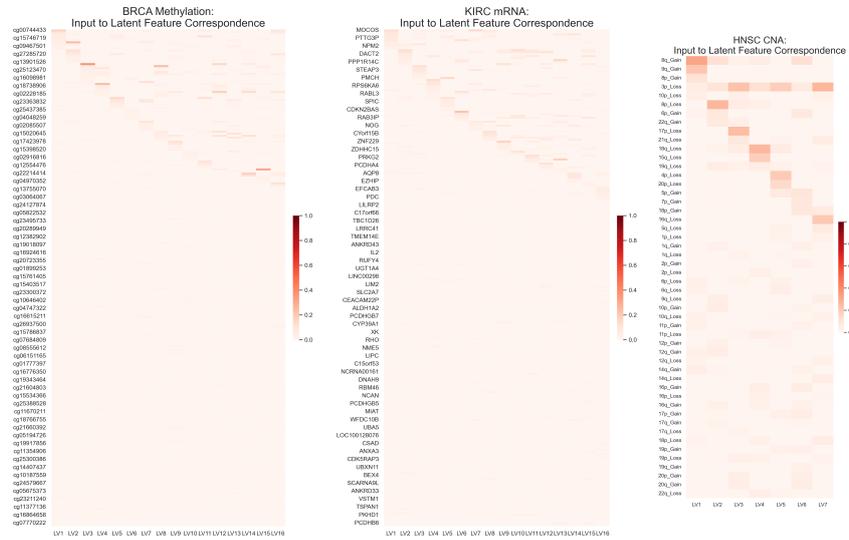


Figure 4.9: Example input to latent feature correspondence heatmaps, i.e. the encoding weights of iCS-GAN, as applied on the BRCA methylation (left), KIRC mRNA (middle) and HNSC CNA (right) datasets. Heatmap rows represent input features, columns represent latent variables. Each non-zero entry indicates that a given input feature contributed to the creation of a given latent variable. With most heatmap elements being 0, we can identify clear links between the observed and latent variables.

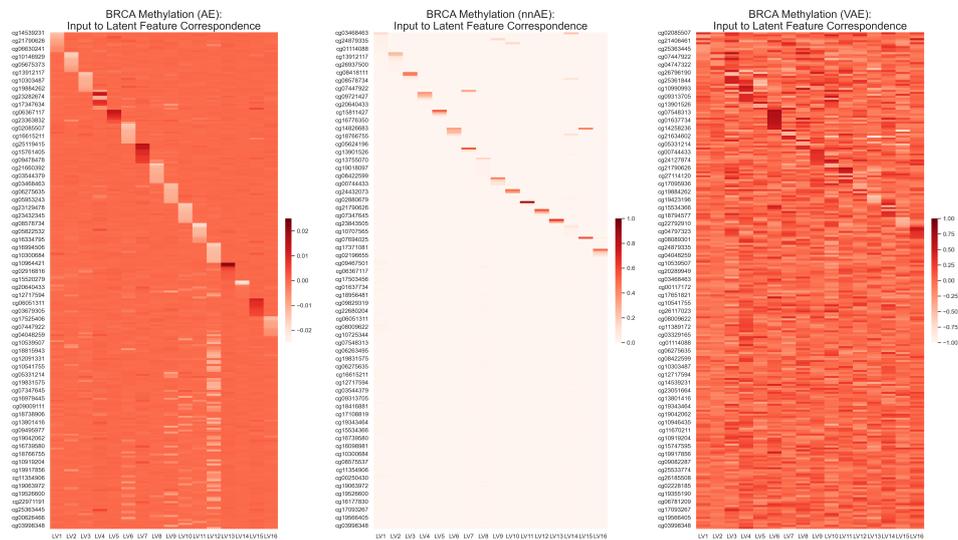


Figure 4.10: Example input to latent feature correspondence heatmaps, i.e. the encoding weights of AE (left), nnAE (middle) and VAE (right), as applied on the BRCA methylation dataset. Heatmap rows represent input features, columns represent latent variables. Each entry indicates how much a given input feature contributed to the creation of a given latent variable. Except for nnAE, the weight maps are largely non-interpretable.

4.5 TCGA: Multiomic Integration Validation

Table 4.5 summarizes the performance of iCS-GAN as applied on multiomic TCGA datasets. Subtyping results are illustrated in Figures 4.11 - 4.15. The clusters found by iCS-GAN were stable on all datasets except BLCA, with significant survival differences identified on KIRC and BLCA datasets. Despite reconstructing the data from near-binary encodings, the distributions of data reconstructions obtained with iCS-GAN well matched the distributions of the observed features (Figure 4.16). iCS-GAN allowed us to represent patient data in a form of easily interpretable binary encodings which can be viewed as binary indicators signifying the presence or absence of underlying biological events (Figure 4.17), with clear links between input features and these events being identifiable via SHAP visualisations (Figure 4.18).

Table 4.6 summarizes the results of ablation analysis conducted on the components of the multiomic version of iCS-GAN. Except for the IBP prior, each of the modifications introduced to the ALICE model [116] in Section 3.2 largely enhanced the stability of iCS-GAN’s clustering results. Furthermore, the inclusion of shared and independent layers and layer-wise pre-training improved the method’s reconstruction error and distribution matching capabilities. Performance differences between iCS-GANs with normal and IBP priors were marginal.

Dataset	RMSE	WD-Rec	WD-Synth	WD-Latent	1-ARI
BRCA	0.22 (0.21-0.23)	0.11 (0.09-0.12)	0.09 (0.09-0.10)	0.07 (0.05-0.09)	0.08 (0.06-0.14)
KIRC	0.21 (0.21-0.21)	0.08 (0.08-0.09)	0.10 (0.09-0.10)	0.05 (0.04-0.05)	0.28 (0.12-0.41)
BLCA	0.24 (0.23-0.25)	0.12 (0.11-0.12)	0.12 (0.12-0.13)	0.06 (0.04-0.07)	0.43 (0.23-0.58)
COAD	0.24 (0.24-0.25)	0.10 (0.10-0.11)	0.11 (0.11-0.12)	0.05 (0.04-0.06)	0.20 (0.10-0.30)
HNSC	0.23 (0.22-0.23)	0.10 (0.10-0.11)	0.10 (0.10-0.11)	0.06 (0.06-0.07)	0.12 (0.08-0.19)

Table 4.5: Performance of iCS-GAN on multiomic TCGA datasets. All reported values represent the averages from 5 runs of the model. The values in parentheses indicate the minimum and maximum values observed for each metric across these 5 runs.

Method	RMSE	WD-Rec	WD-Synth	WD-Latent	1-ARI
iCS-GAN	0.23	0.10	0.10	0.06	0.22
W/O IND	+0.03	0.00	+0.01	+0.01	+0.24
W/O PT	+0.02	+0.01	+0.01	+0.02	+0.18
W/O CL	0.00	0.00	0.00	-0.01	+0.24
W/O CLI	0.00	0.00	0.00	0.00	+0.13
W/O IBP	-0.01	0.00	-0.01	0.00	-0.03

Table 4.6: TCGA multimodal ablation study results. The top row represents the results obtained with iCS-GAN averaged across the 5 TCGA datasets (5 runs each). The remaining rows describe the average change in each metric, for iCS-GAN with removed: independent layers (W/O IND), layer-wise pre-training (W/O PT), clustering regularization on shared and independent layers (W/O CL), clustering regularization on independent layers only (W/O CLI) and the Indian Buffet Process prior (W/O IBP). Positive change value corresponds to an increase in a given metric (i.e. worse performance). Full results for each dataset, with error bounds, can be found in Appendix B.

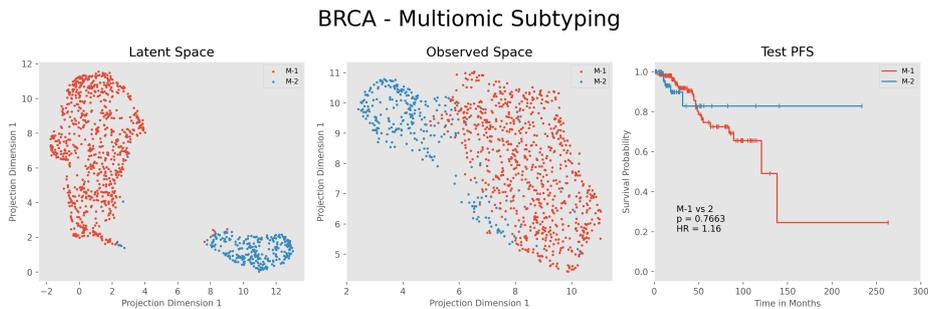


Figure 4.11: Multiomic subtyping results as obtained with iCS-GAN on TCGA BRCA dataset. UMAP visualisation of the subtypes discovered in the latent (left) and observed (middle) spaces, and the associated test-set PFS survival curves (right). For each subtype comparison, we provide the hazard ratio (HR) and p value calculated with the Cox proportional hazard test.

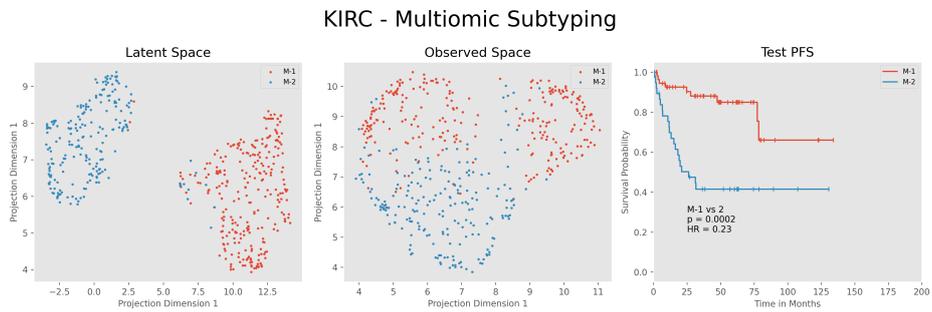


Figure 4.12: Multiomic subtyping results as obtained with iCS-GAN on TCGA KIRC dataset. UMAP visualisation of the subtypes discovered in the latent (left) and observed (middle) spaces, and the associated test-set PFS survival curves (right). For each subtype comparison, we provide the hazard ratio (HR) and p value calculated with the Cox proportional hazard test.

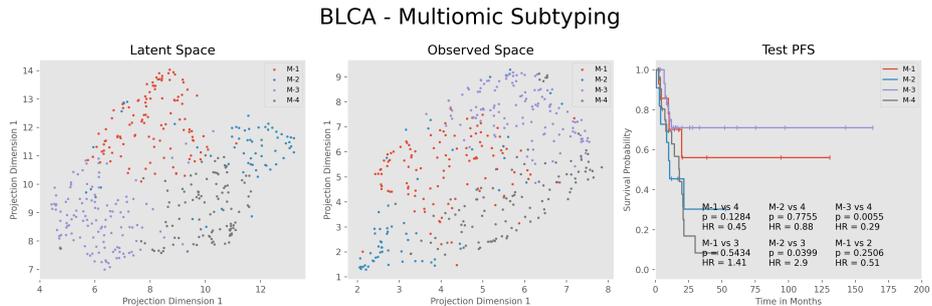


Figure 4.13: Multiomic subtyping results as obtained with iCS-GAN on TCGA BLCA dataset. UMAP visualisation of the subtypes discovered in the latent (left) and observed (middle) spaces, and the associated test-set PFS survival curves (right). For each subtype comparison, we provide the hazard ratio (HR) and p value calculated with the Cox proportional hazard test.

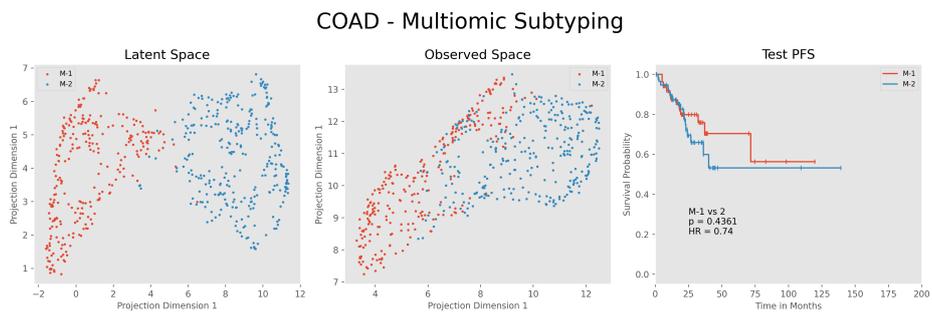


Figure 4.14: Multiomic subtyping results as obtained with iCS-GAN on TCGA COAD dataset. UMAP visualisation of the subtypes discovered in the latent (left) and observed (middle) spaces, and the associated test-set PFS survival curves (right). For each subtype comparison, we provide the hazard ratio (HR) and p value calculated with the Cox proportional hazard test.

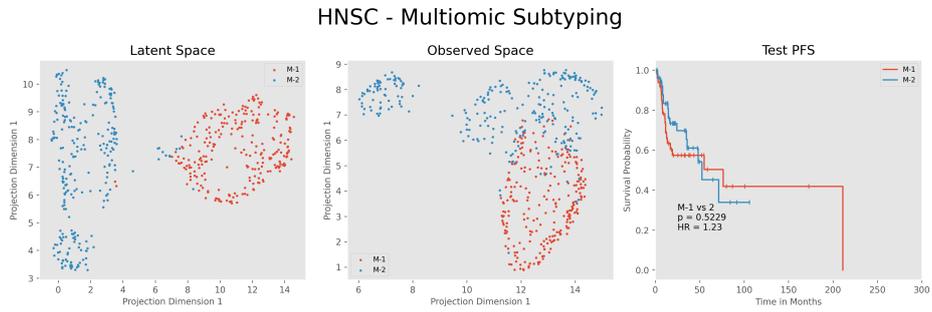


Figure 4.15: Multiomic subtyping results as obtained with iCS-GAN on TCGA HNSC dataset. UMAP visualisation of the subtypes discovered in the latent (left) and observed (middle) spaces, and the associated test-set PFS survival curves (right). For each subtype comparison, we provide the hazard ratio (HR) and p value calculated with the Cox proportional hazard test.

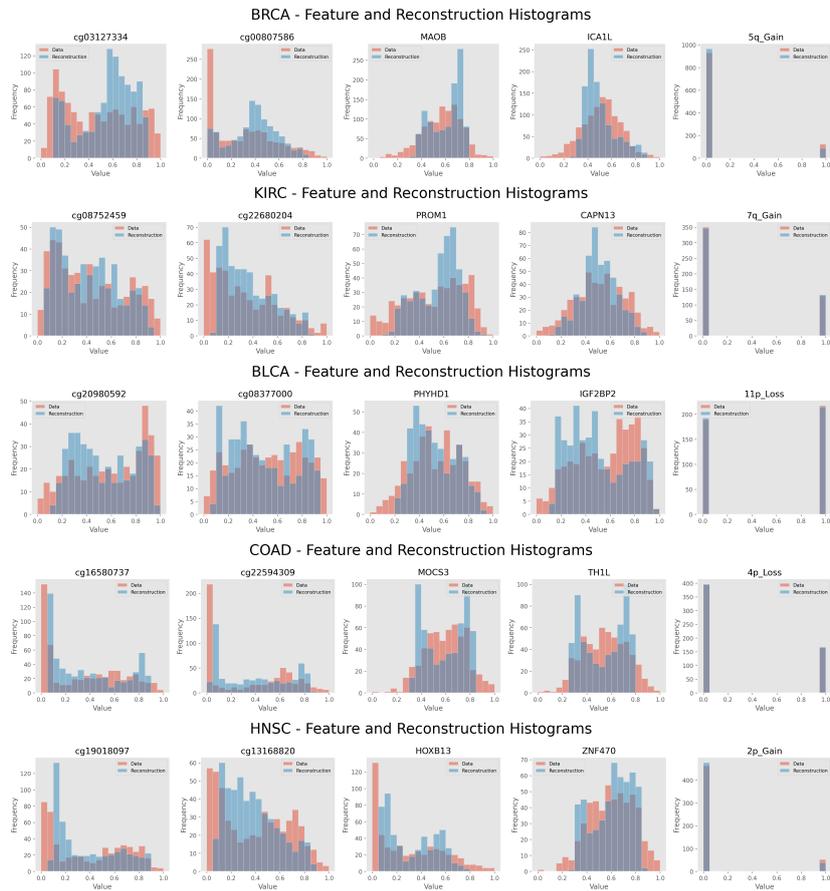


Figure 4.16: Frequency histograms for selected methylation (first two columns), mRNA (next two columns) and CNA (last column) features from BRCA, KIRC, BLCA, COAD and HNSC datasets (top to bottom), including reconstructions obtained with multiomic iCS-GAN.

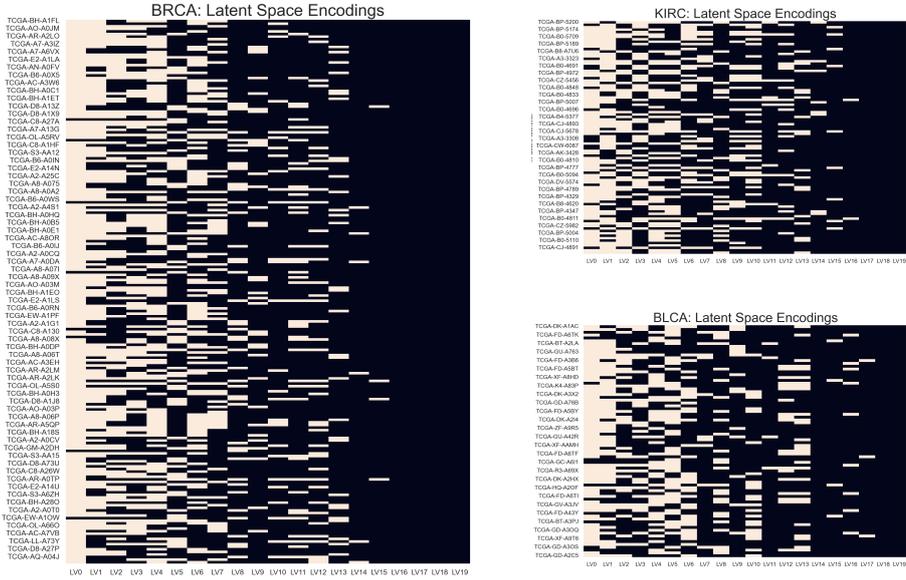


Figure 4.17: Heatmaps visualising the binarized test set latent space encodings as obtained on BRCA (left), KIRC (top right) and BLCA (bottom right) datasets. Heatmap rows represent latent variables, columns represent patient samples. White mark indicates that a given latent variable is active for a given patient sample.

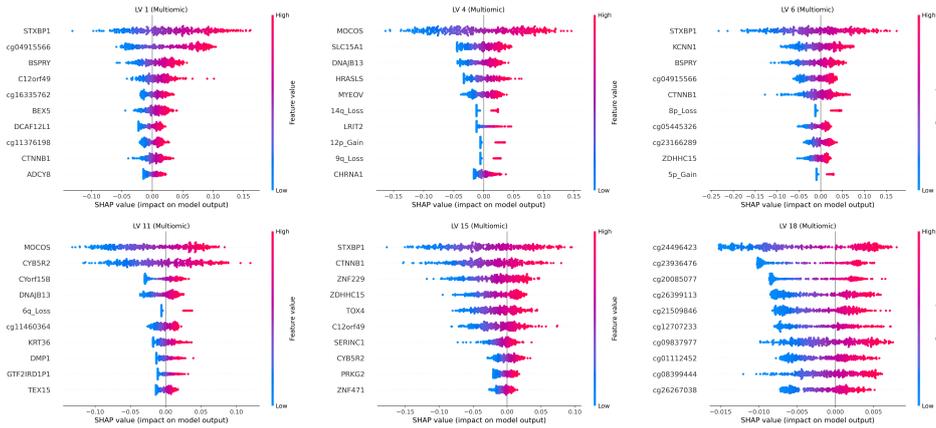


Figure 4.18: SHAP plots describing the importance of input features in the creation of selected of multiomic latent variables (KIRC dataset). For each latent variable, we plot the top 10 most important inputs. The LVs encapsulate relationships both within (LV15, LV18) and between modalities (LV1, LV4, LV6, LV11).

4.6 TCGA: Gold Standard Comparison

In this section, to further validate our proposed method, we compare the multiomic subtypes identified by iCS-GAN on the TCGA datasets with the gold standard currently accepted in the literature. As clustering is an unsupervised exploratory analysis, we do not necessarily expect our method to replicate previously identified subtypes, although some alignment with established classification schemes would increase our confidence in the model and its outputs. For the remainder of this section, we perform this comparison for the BRCA, COAD, and HNSC datasets, as well-established “gold standard” assignments for the other datasets were not readily available.

4.6.1 BRCA - ER, PR and HER2 Subtypes

Breast cancers are typically classified based on the presence or absence of oestrogen receptors (ER), progesterone receptors (PR), and human epidermal growth factor receptor 2 (HER2), or the absence of all of these (triple-negative). Tumours that express oestrogen receptors are classified as ER positive (ER+), while those that do not are ER negative (ER-). Similarly, cancers with progesterone receptors are PR positive (PR+), and those without are PR negative (PR-). Likewise, tumours expressing HER2 are HER2 positive (HER2+), whereas those without are HER2 negative (HER2-). Triple negative (i.e. ER-/PR-/HER2-) tumours are denoted TN. These subtypes are not mutually exclusive; for example, a tumour can be ER+, PR+, and HER2+ simultaneously. Triple-negative breast cancer is associated with worst overall and disease-free survival outcomes [146].

We obtained subtype assignments for ER/PR/HER2 breast cancer status from the cBioPortal for Cancer Genomics [25] using the Nature 2012 version

of the dataset [188]. These assignments were not available for all samples in our analysis, and some were labelled as ‘Indeterminate’ for ER/PR/HER2 status. Therefore, we restricted our comparison to samples with a clearly defined ‘Positive’ or ‘Negative’ status. Figure 4.19 presents a comparison between the multiomic BRCA subtypes M-1 and M-2 and the subtypes classified by ER/PR/HER2 status. Table 4.7 displays the confusion matrices for the comparisons.

As shown in Figure 4.19, the multiomic subtype M-1 identified by iCS-GAN predominantly corresponds to the ER+ and/or PR+ breast cancer subtype, while M-2 aligns mostly with the ER-/PR- subtype. However, no clear association was observed between iCS-GAN’s assignments and HER2 status, as HER2+ and HER2- samples were distributed across both M-1 and M-2 subtypes. The TN (triple-negative) subtype is almost entirely contained within M-2.

These findings align with our understanding of breast cancer biology and the aetiology of the established subtypes. Oestrogen and progesterone are both hormones that drive similar transcriptional programs, so it is unsurprising that the most pronounced distinction in the clusters arises between cases where at least one of these hormones is present and those where neither are. Conversely, HER2 status is classified as a subtype because it signifies the overexpression of a receptor that can be targeted by specific drugs, such as Herceptin [61], rather than reflecting a distinct cellular state. Since HER2 overexpression is not mutually exclusive with ER or PR expression, it is expected that the broad transcriptional changes driven by these hormones dominate the clustering, rather than the overexpression of a single gene. Despite this, we note that the ER-/PR-/HER2- subgroup is largely co-localized within M-2, suggesting that it would be straightforward to extract the TN

subtype from the rest of this cluster with minimal manual curation. This shows that iCS-GAN has performed as well as can be reasonably expected in identifying the established breast cancer subtypes.

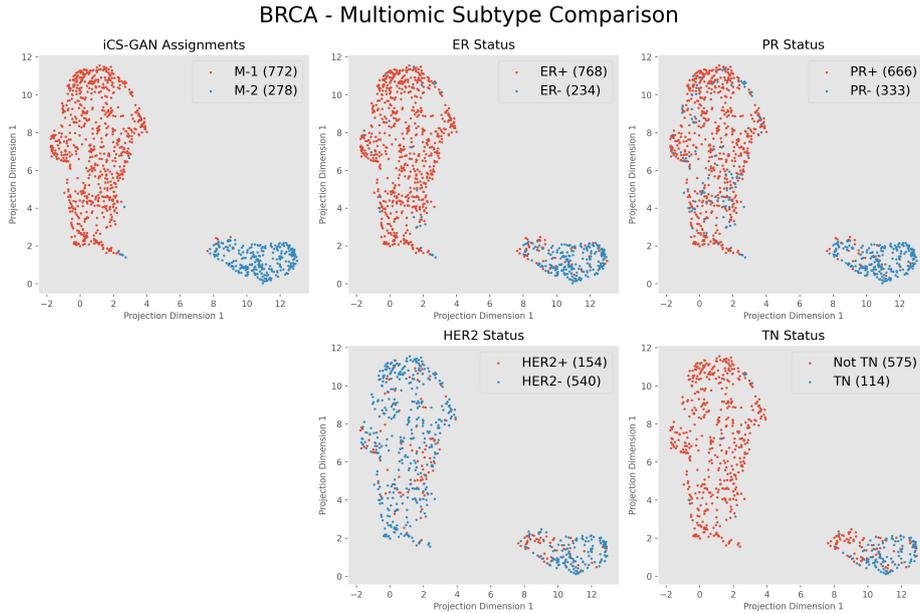


Figure 4.19: Latent space UMAP visualisation for the multiomic subtypes M-1 and M-2 identified with iCS-GAN on the TCGA BRCA dataset, ER, PR, HER2, and TN status sample assignments in the latent space of iCS-GAN (left to right, top to bottom). For clarity, samples with unavailable ER/PR/HER2/TN assignments were removed from their respective visualisations.

	ER+	ER-		PR+	PR-
M-1	713	21	M-1	634	99
M-2	55	213	M-2	32	234

	HER2+	HER2-		Not TN	TN
M-1	101	399	M-1	491	6
M-2	53	141	M-2	84	108

Table 4.7: Confusion matrices comparing the multiomic subtypes M-1 and M-2, identified by iCS-GAN on the TCGA BRCA dataset, with the ER, PR, HER2, and TN status sample assignments (left to right, top to bottom). Each matrix includes only samples with available assignments from both iCS-GAN and the respective ER, PR, HER2, or TN status.

4.6.2 COAD - CMS Subtypes

The gold standard framework for colorectal cancer subtyping is the Consensus Molecular Subtype (CMS) classification system [70], which divides colorectal cancer into four molecular subtypes:

- CMS1 (MSI Immune) subtype characterized by microsatellite instability, hypermutations and low prevalence of somatic copy number alterations,
- CMS2 (Canonical) subtype characterized by epithelial differentiation (the process by which unspecialized cells acquire features of epithelial cells), chromosomal instability and WNT (development and stemness regulation) and MYC (adjustment of cellular functions such as DNA damage response) signalling activations,
- CMS3 (Metabolic) subtype characterized by evident metabolic dysregulation,
- CMS4 (Mesenchymal) subtype characterized by significant TGF β activations, stromal invasion and angiogenesis (the formation of new blood vessels via the migration, growth, and differentiation of endothelial cells).

The CMS classification can be a useful prognostic marker. Of the four subtypes, CMS4 tumours display worse overall and relapse-free survival [70]. Conversely, CMS2 patients have superior survival after relapse, which in turn is very poor for CMS1 patients [70].

We retrieved subtype assignments for the CMS colorectal cancer classification from the Synapse portal¹. These annotations were originally generated

¹Found at doi:10.7303/syn2623706

and published alongside the original CMS study [70]. However, CMS labels were not available for all COAD samples used in our experiments. Figure 4.20 presents a comparison between the multiomic COAD subtypes, M-1 and M-2, and the CMS subtypes. Table 4.8 displays the confusion matrix for the comparison.

We found that our multiomic subtypes M-1 and M-2 did not precisely recover the CMS subtypes. However, we noted that CMS1 and CMS3 almost entirely fall within the multiomic subtype M-1, while subtype M-2 almost entirely consists of CMS2 and CMS4 tumours. There was a small pocket of CMS2 and CMS4 tumours classified as M1 that were co-located in the centre of the UMAP. Interestingly, CMS1 is co-localised within the M-1 cluster, which appears reasonable as these MSI tumours display a hyper-mutated, low CNA profile that is unique to this CMS type. The co-assignment with CMS3 is not surprising because, as noted in the paper that introduced the CMS subtypes [70], many of these tumours also have a hyper-mutated phenotype, similar to MSI tumours. We hypothesise the main distinction between M-1 and M-2 is the hypermutation and associated molecular changes.

The lack of direct alignment between the CMS framework and the subtypes identified with iCS-GAN is not unexpected, as the CMS classification was derived through a hand-curated approach, integrating results from six independent colorectal cancer subtyping algorithms with domain knowledge [70]. Given that our method relies on a single clustering approach, it would be unlikely to replicate the CMS subtypes exactly. However, the multiomic features extracted by iCS-GAN still strongly correlate with the CMS classification. Notably, a random forest classifier trained on iCS-GAN encodings achieved a test set accuracy of 0.72 in classifying patient samples into CMS subtypes.

COAD - Multiomic Subtype Comparison

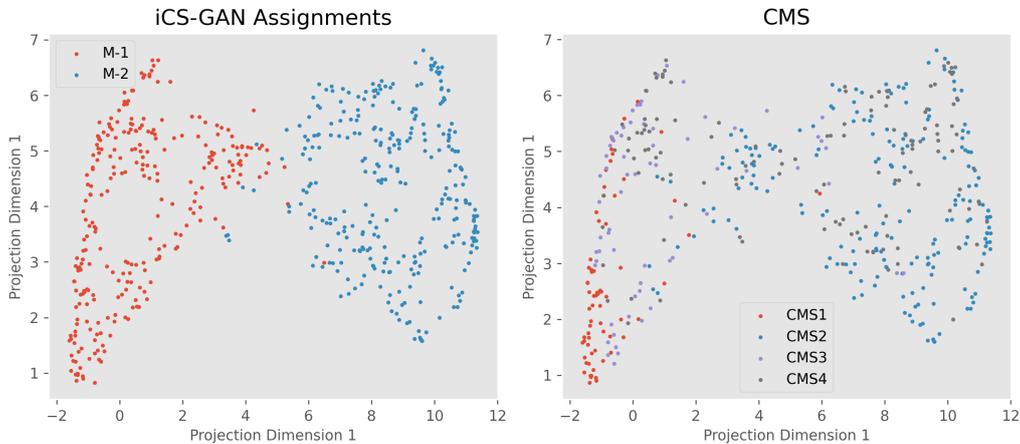


Figure 4.20: Latent space UMAP visualisation for the multiomic subtypes M-1 and M-2 identified with iCS-GAN on the TCGA COAD dataset (left), CMS classification sample assignments mapped in the latent space of iCS-GAN (right). For clarity, samples with unavailable CMS labellings were removed from their respective visualisation.

	CMS1	CMS2	CMS3	CMS4
M-1	67	36	62	48
M-2	1	162	7	78

Table 4.8: Confusion matrix comparing the multiomic subtypes M-1 and M-2, identified by iCS-GAN on the TCGA COAD dataset, with the CMS framework. The matrix includes only samples with both CMS and iCS-GAN’s labellings available.

4.6.3 HNSC - HPV Subtypes

Head and neck cancers are typically categorized based on their human papillomavirus (HPV) status. Tumours associated with HPV infection are classified as HPV-positive (HPV+), while those without HPV involvement are considered HPV-negative (HPV-). HPV status defines two distinct types of head and neck squamous cell carcinoma (HNSC) and is the only clinically validated biomarker for predicting survival in these cancers [166].

Subtype assignments for the HPV head and neck cancer status were avail-

able in the HNSC dataset directly, though HPV status was not provided for all tumours. Figure 4.21 compares the multiomic HNSC subtypes uncovered with iCS-GAN with subtypes defined by HPV status. Figure 4.22 provides a similar comparison for the mRNA-defined subtypes mRNA-1, mRNA-2 and mRNA-3. Table 4.9 displays confusion matrices for the comparisons.

As shown in Figure 4.21, the HPV+ subtype of head and neck cancer is largely a localized subset of the multiomic subtype M-2 identified by iCS-GAN, albeit it is not a separate distinct cluster. However, increasing the number of multiomic subtypes to five (potentially at the expense of clustering stability) would result in HPV+ tumours forming their own cluster, designated as the new multiomic subtype M-4. Additionally, as illustrated in Figure 4.22, the HPV+ subtype aligns directly with the mRNA-3 subtype identified by iCS-GAN through single-modality mRNA-based analysis. This demonstrates that our method successfully identified two major HNSC subtypes, reinforcing confidence in the model’s outputs.

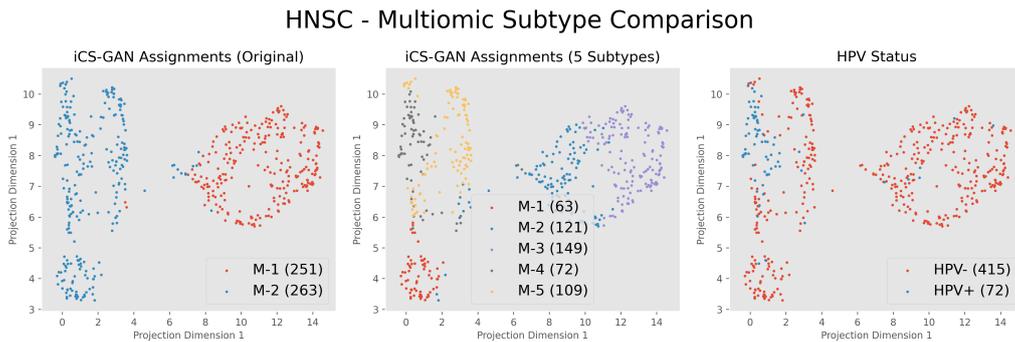


Figure 4.21: Latent space UMAP visualisation for the multiomic subtypes M-1 and M-2 identified with iCS-GAN on the TCGA HNSC dataset (left), latent space UMAP visualisation for the multiomic subtypes M-1 to M-5 identified with iCS-GAN with increased number of clusters (middle), HPV status sample assignments mapped in the latent space of iCS-GAN (right). For clarity, samples with unavailable HPV assignments were removed from their respective visualisation.

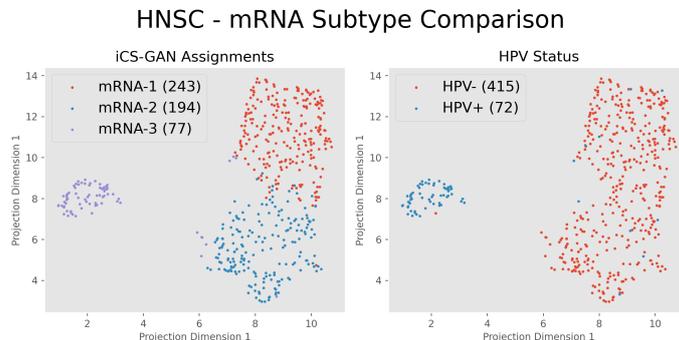


Figure 4.22: Latent space UMAP visualisation for the mRNA-based subtypes mRNA-1, mRNA-2 and mRNA-3 identified with iCS-GAN on the TCGA HNSC mRNA dataset (left), HPV status sample assignments mapped in the latent space of iCS-GAN (right). For clarity, samples with unavailable HPV assignments were removed from their respective visualisation.

	HPV+	HPV-
M-1	6	237
M-2	66	178

	HPV+	HPV-
mRNA-1	4	227
mRNA-2	5	180
mRNA-3	63	8

Table 4.9: Confusion matrices comparing the multiomic subtypes M-1 and M-2 (left), and the mRNA subtypes mRNA-1, mRNA-2 and mRNA-3 (right) identified by iCS-GAN on the TCGA HNSC dataset, with the HPV status sample assignments. Both matrices includes only samples with available assignments from both iCS-GAN and the HPV status.

4.6.4 Remarks

Comparisons of breast, colorectal, and head and neck cancer subtypes identified using iCS-GAN with the established gold standards from the literature validated our proposed method. iCS-GAN demonstrated its ability to extract key details relevant to these classification frameworks in an unsupervised manner. For the BRCA dataset, the multiomic subtype M-2 aligned with the ER and/or PR-negative subtype of breast cancer, with triple-negative tumours representing a localized subset of M-2. In colorectal cancer, while the method did not directly identify the CMS subtypes, the latent variables extracted by iCS-GAN were highly indicative of these assignments. For head

and neck cancer, the mRNA-based subtype mRNA-3 identified by iCS-GAN corresponded directly to HPV-positive tumours. These findings highlight iCS-GAN’s ability to uncover critical details relevant to established subtyping frameworks. However, as clustering is inherently an unsupervised analysis, additional manual curation may be necessary to fully recover the known subtypes.

4.7 TCGA: Survival Experiments

In Section 3.1.5, we introduced the optional CQRNN [152] based survival regularization component that we hypothesized could be used in iCS-GAN to encourage latent space encodings, and thus subtypes, more indicative of patient survival. To verify this hypothesis, we trained iCS-GAN with survival regularization on the KIRC dataset. Note that we explored the survival regularization for KIRC dataset only due to the subpar performance of the CQRNN network on the remaining TCGA datasets (Table 4.10). We decided to avoid regularizing iCS-GAN with survival network with close to random predictive accuracy.

Inclusion of the survival regularization modified our results for the KIRC dataset from Experiments 2 and 3 as follows.

- The optimal number of subtypes changed from 5 to 4 for the KIRC methylation dataset, resulting in a better organization of the latent space in terms of cluster separability. Significant survival differences were retained (Figure 4.23 vs Figure 4.4).
- Survival differences between the two mRNA-defined subtypes increased (HR = 2.51 vs HR = 2.14, $p = 0.0154$ vs $p = 0.0418$; Figure 4.23 vs Figure 4.4).

- Subtypes with significant survival differences were additionally uncovered on the CNA dataset (Figure 4.23 vs Figure 4.4).
- The optimal number of subtypes increased from 2 to 3 for the multiomic analysis, providing greater granularity of the results while retaining the significant survival differences (Figure 4.24 vs Figure 4.12).
- No major differences in performance metrics were observed (Table 4.11).

Dataset	Methylation	mRNA	CNA	Multiomic
BRCA	0.59	0.64	0.57	0.58
KIRC	0.72	0.73	0.65	0.74
BLCA	0.60	0.61	0.59	0.63
COAD	0.55	0.59	0.55	0.58
HNSC	0.58	0.61	0.54	0.60

Table 4.10: 5-fold cross-validation performance of the CQRNN survival network on TCGA datasets, as measured with Harrell’s Concordance Index independently on the methylation, mRNA and CNA modalities, as well as on the concatenated multiomic data.

Method	Modality	RMSE	WD-Rec	WD-Synth	WD-Latent	1-ARI
iCS-GAN	Methylation	0.22 (0.21-0.22)	0.06 (0.06-0.07)	0.09 (0.09-0.10)	0.05 (0.04-0.06)	0.27 (0.11-0.37)
iCS-GAN + Surv	Methylation	0.22 (0.21-0.22)	0.06 (0.06-0.07)	0.09 (0.09-0.10)	0.05 (0.04-0.06)	0.36 (0.27-0.43)
iCS-GAN	mRNA	0.18 (0.17-0.19)	0.09 (0.07-0.10)	0.09 (0.07-0.10)	0.05 (0.04-0.06)	0.16 (0.08-0.27)
iCS-GAN + Surv	mRNA	0.18 (0.17-0.19)	0.08 (0.07-0.10)	0.08 (0.07-0.10)	0.05 (0.05-0.06)	0.10 (0.04-0.16)
iCS-GAN	CNA	0.22 (0.22-0.23)	0.06 (0.06-0.06)	0.08 (0.07-0.08)	0.11 (0.10-0.11)	0.18 (0.05-0.28)
iCS-GAN + Surv	CNA	0.25 (0.22-0.29)	0.06 (0.06-0.07)	0.09 (0.08-0.09)	0.10 (0.09-0.11)	0.26 (0.13-0.40)
iCS-GAN	Multimodal	0.21 (0.21-0.21)	0.08 (0.08-0.09)	0.10 (0.09-0.10)	0.05 (0.04-0.05)	0.28 (0.12-0.41)
iCS-GAN + Surv	Multimodal	0.21 (0.20-0.21)	0.09 (0.08-0.09)	0.09 (0.09-0.10)	0.06 (0.05-0.07)	0.26 (0.15-0.38)

Table 4.11: Performance of iCS-GAN with (iCS-GAN + Surv) and without (iCS-GAN) survival regularization on KIRC dataset. All reported values represent the averages from 5 runs of the model. The values in parentheses indicate the minimum and maximum values observed for each metric across these 5 runs.

KIRC - Single Modality Subtyping (Survival Reg)

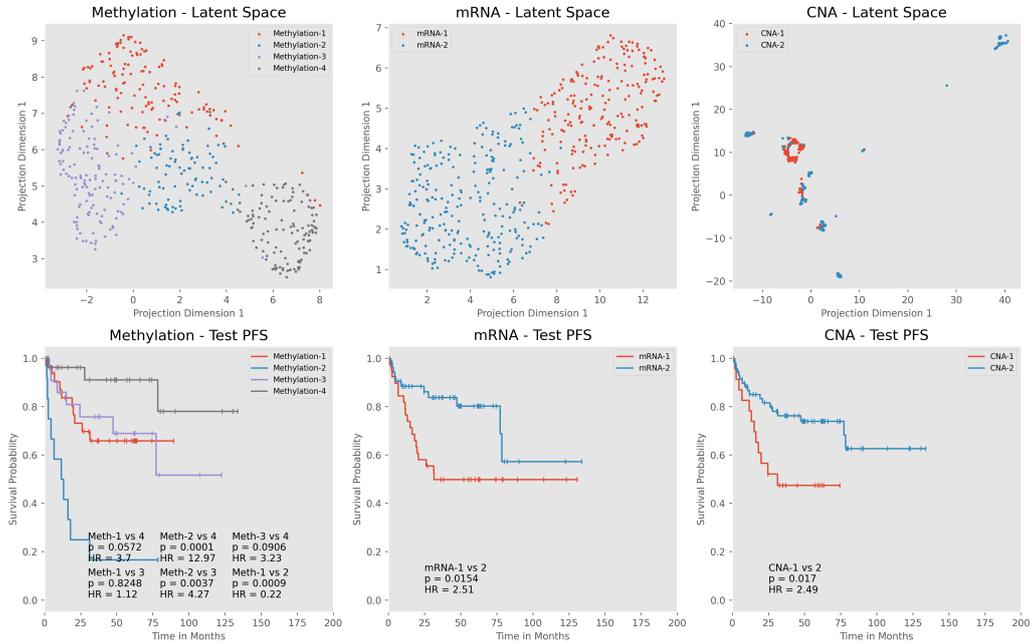


Figure 4.23: Single modality subtyping results as obtained with iCS-GAN with survival regularization on TCGA KIRC methylation (left), mRNA (middle) and CNA (right) datasets. UMAP visualisation of the subtypes discovered in the latent space (top) and the associated test-set PFS survival curves (bottom). For each subtype comparison, we provide the hazard ratio (HR) and p value calculated with the Cox proportional hazard test.

KIRC - Multiomic Subtyping (Survival Reg)

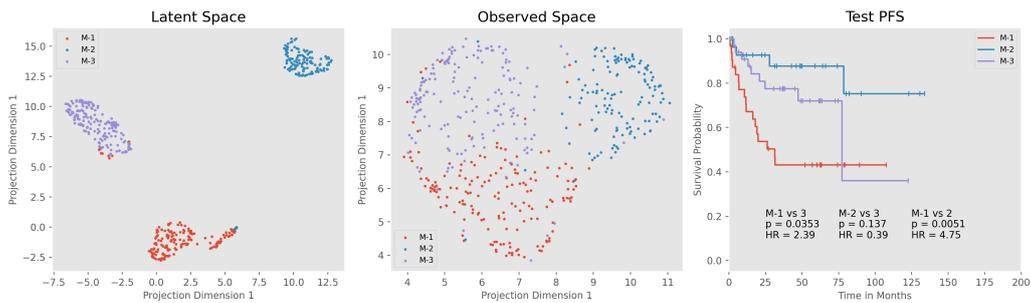


Figure 4.24: Multiomic subtyping results as obtained with iCS-GAN with survival regularization on TCGA KIRC dataset. UMAP visualisation of the subtypes discovered in the latent (left) and observed (middle) spaces, and the associated test-set PFS survival curves (right). For each subtype comparison, we provide the hazard ratio (HR) and p value calculated with the Cox proportional hazard test.

4.8 TCGA: Experiments with Missing Modalities

To assess the effectiveness of our model for integrating multiomic data with missing modalities, for our final validation experiment, we trained iCS-GAN on all TCGA datasets, modified so that a different proportion of data was missing for each dataset (as detailed in Table 4.12). We then compared the newly obtained results with those from Experiment 3, evaluating how well the model recovered the subtypes identified previously with complete data. The true subtype assignments for each dataset were defined as the majority vote from five runs of iCS-GAN in Experiment 3. For the new subtype assignments, we used the majority vote from the new predictions made with potentially imputed encodings. The accuracy between these two labellings was then computed for each datasets.

The imputation process followed the steps described in Section 3.3, with 10 nearest neighbors used for the KNN imputation. KMeans clustering was fitted on the complete (non-imputed) training set encodings only, with subtype predictions generated for the entire dataset.

As presented in Table 4.13, iCS-GAN recovered the initial subtypes almost perfectly on BRCA, KIRC and COAD datasets with 40%, 60% and 50% of modality-complete data available, respectively. Performance was significantly lower for HNSC, which was expected given that we removed 80% of modality-complete data for this dataset, and for BLCA, where iCS-GAN’s subtyping results were unstable in the first place.

Our method, which utilized all the data available for each given modality to pre-train its modality-specific layer, outperformed the complete-case analysis where only modality-complete data was used to train all components

of iCS-GAN (Table 4.14). Subtype recovery accuracy of the complete case analysis was lower on the BLCA and HNSC datasets, significantly lower for BRCA and KIRC, and comparable in the case of COAD.

Dataset	Methylation Missing	mRNA Missing	CNA Missing	Complete Subset
BRCA	20%	20%	20%	40%
KIRC	5%	15%	20%	60%
BLCA	10%	10%	10%	70%
COAD	15%	10%	25%	50%
HNSC	40%	20%	20%	20%

Table 4.12: Percentage of missing data for each modality in the TCGA datasets, along with the percentage of complete data available. Missing data was introduced by randomly masking selected samples from each dataset.

Dataset	Test Accuracy	Test Accuracy (IMP)
BRCA	0.94	0.93
KIRC	0.93	0.95
BLCA	0.54	0.54
COAD	0.93	0.95
HNSC	0.62	0.61

Table 4.13: iCS-GAN test set accuracy for recovering subtype assignments from Experiment 3 with missing modalities. The columns show the accuracy across the entire test set, as well as specifically for the modality-incomplete samples within the test set.

Dataset	Test Accuracy	Test Accuracy (IMP)
BRCA	0.73	0.72
KIRC	0.59	0.63
BLCA	0.48	0.43
COAD	0.92	0.95
HNSC	0.56	0.58

Table 4.14: iCS-GAN test set accuracy for recovering subtype assignments from Experiment 3 with missing modalities, considering only the complete case analysis. The columns show the accuracy across the entire test set, as well as specifically for the modality-incomplete samples within the test set.

Chapter 5

Results: Prostate Cancer

Subtyping

In this chapter, we apply iCS-GAN to the PanProstate Cancer Group prostate cancer dataset introduced in Section 1.2 and present the results. The main objective of this chapter is to conduct an integrative analysis of prostate cancer data and provide insights into the composition of the disease by deriving multiomic subtypes, using the now fully tested and validated latent feature model developed in this thesis. Multimodal data integration with iCS-GAN requires us to first train the single-modality variant of the model independently on each of the available modalities before moving onto multiomic analysis. Therefore, we first apply the single-modality version of iCS-GAN to the summary measurements, driver genes, copy number alterations and RNA modalities in Section 5.1. In Section 5.2, we extend our analysis to the multimodal case by applying iCS-GAN to the entire PPCG dataset, deriving comprehensive multiomic prostate cancer subtypes. Finally, in Section 5.3, we compare our results with a number of previously established prostate cancer subtyping schemas.

For each experiment in Sections 5.1 and 5.2, we applied iCS-GAN as described in Chapter 3, utilizing both the clustering and survival regularization components. To ensure optimal performance and stability of the subtyping results, we conducted a hyper-parameter search, details of which are provided in Appendix A. The number of latent features in each single-modality model was set to the square root of the number of input features, and for multimodal integration, we selected an Indian Buffet Process prior with 20 latent variables and $\alpha = 10$ feature activation probability. We ran iCS-GAN five times for each experiment using different random seeds to assess the stability of the results. All runs produced highly similar subtyping outcomes, so we focused our subtype characterization analysis on the ‘consensus’ assignments, where the label (subtype assignment) for each sample was defined as the majority vote across the five runs. For exploring the latent features extracted by the model or UMAP visualisation, to conserve space, we selected the results from a single run of iCS-GAN for inclusion in this chapter.

5.1 Single Modality Subtyping

In this section, we present the single-modality subtyping results obtained with iCS-GAN for the summary measurements (Section 5.1.1), driver genes (Section 5.1.2), CNAs (Section 5.1.3) and RNA (Section 5.1.4) modalities available in the PPCG dataset. In each case, we briefly summarize the uncovered single-modality subtypes, providing insights into survival differences and subtype characteristics. We additionally investigate the relevance of the latent features extracted by iCS-GAN. Furthermore, to illustrate how the uncovered subtypes might be used in a real-world clinical test, for each modality, we use the obtained subtype assignments as true labels for the development of

RFC-based (random forest classifier) predictive tests, to can classify patient samples into disease subtypes directly from the data. Successful development of such predictive tests would eliminate the need to repeatedly use latent feature extraction models on new data, making the subtyping process more clinically efficient and practical. The random forest classifier was chosen as a baseline model due to its insensitivity to input data types and distributions, as well as the known superior performance of tree-based models for tabular data classification [69].

5.1.1 Subtyping with Summary Measurements

Applied to the summary measurements dataset, 3 distinct prostate cancer subtypes returned by iCS-GAN, here named SM-1, SM-2 and SM-3, fulfilled the stability criteria. The average test set reconstruction error was 0.07 for the 5 runs of the model, and the test set subtyping stability, as assessed with ARI, was 0.67.

Figure 5.1 shows the survival differences between the subtypes. Tumours classified as SM-1 ($n = 241$) exhibit worse patient prognosis when compared to those in SM-2 ($n = 340$) and SM-3 ($n = 276$), both in terms of both relapse-free (SM-1 vs SM-2: HR = 2.09, $p = 0.0151$; SM-1 vs SM-3: HR = 2.42, $p = 0.0116$) and metastasis-free survival (SM-1 vs SM-2: HR = 3.71, $p = 0.0062$; SM-1 vs SM-3: HR = 8.7, $p = 0.0052$).

In terms of subtype characteristics (Figure 5.2), SM-3 is a “quiet” subtype with a generally low prevalence of each genetic aberration. Tumours classified as SM-1 more commonly exhibit kataegis, chromothripsis and Whole Genome Duplication (as measured by ploidy), and are characterized by a higher percentage genome altered, as well as the number of clonal and subclonal CNAs (PGA (Clonal) and PGA (Subclonal)), higher numbers of structural variants,

including inversions, deletions, duplications and rearrangements, more prevalent complex indels and higher total and average numbers of breakpoints, higher numbers of breakpoints in chains and numbers of chains. Tumours classified as SM-2 are characterized by positive *ETS* status, more prevalent *ERG* gene fusions, higher median number of chromosomes involved in breakpoint chains and higher inter to intra chromosomal breakpoint ratio - metrics linked to the occurrence of chromoplexy.

A predictive RFC-based test classified data samples as SM-1, SM-2 or SM-3 with 0.93 test set accuracy.

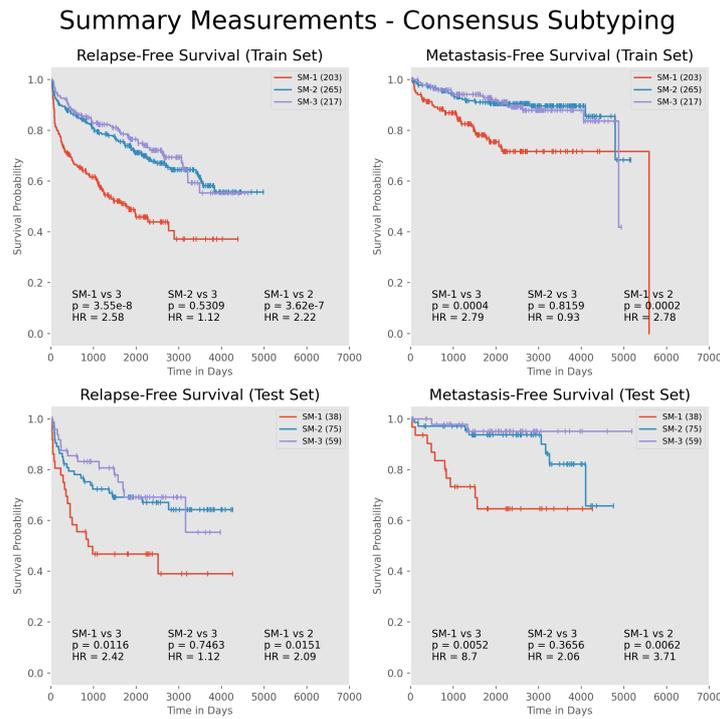


Figure 5.1: Kaplan-Meier plots visualising the relapse-free (left) and metastasis-free (right) survival for the summary measurements subtypes SM-1, SM-2 and SM-3, separately for the training (top) and test set (bottom) samples. For each pairwise comparison, we provide the hazard ratio (HR) and the p value calculated with the Cox proportional hazard test. The number of samples in each subtype is given in brackets, next to the subtype label.

Summary Measurements Subtypes: Subtype Characteristics

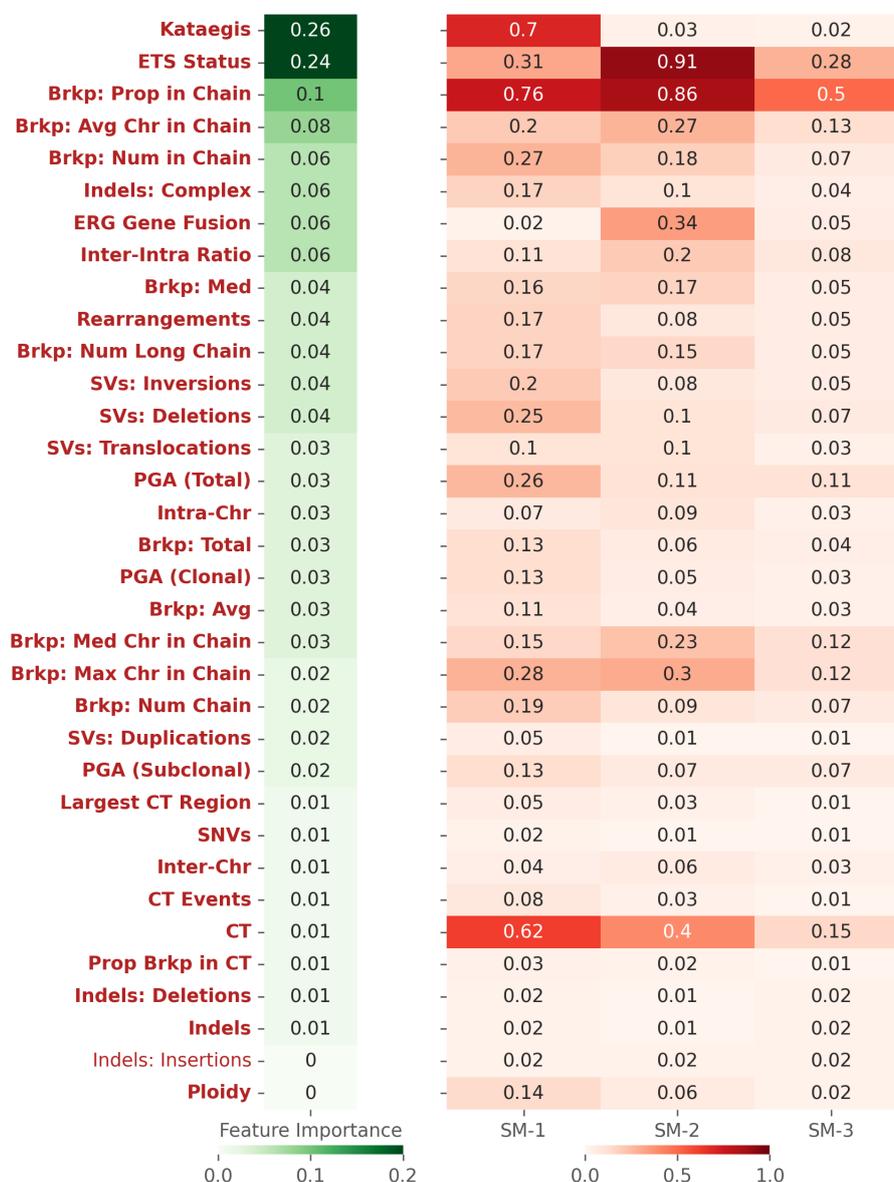


Figure 5.2: Characteristics of the prostate cancer subtypes uncovered from summary measurements. Main heatmap visualises mean feature value in each subtype, as measured on the training set. Features are ordered by SHAP-based importance when predicting cluster assignments using RFC applied on the original data features. Aberrations with significant differences between subtypes are marked in bold red for features significant on both training and test sets or in light red for features significant on the training set only (FDR-adjusted $p < 0.05$, Chi-squared (categorical variables) or Kruskal-Wallis (continuous variables) test).

Notably, the latent features created by iCS-GAN from the summary measurements were both fully interpretable and biologically-relevant, with each latent variable describing a specific set of genetic aberrations. As evident from Figures 5.3 and 5.4, the 6 latent variables broadly corresponded to: the percentage of genome altered by CNAs, including clonal and subclonal CNAs (LV1), chromothripsis and inversions (LV2), DNA breakpoints (LV3 and LV6), *ETS* / *ERG* aberrations (LV4) and kataegis, chromthripsis and ploidy (LV5).

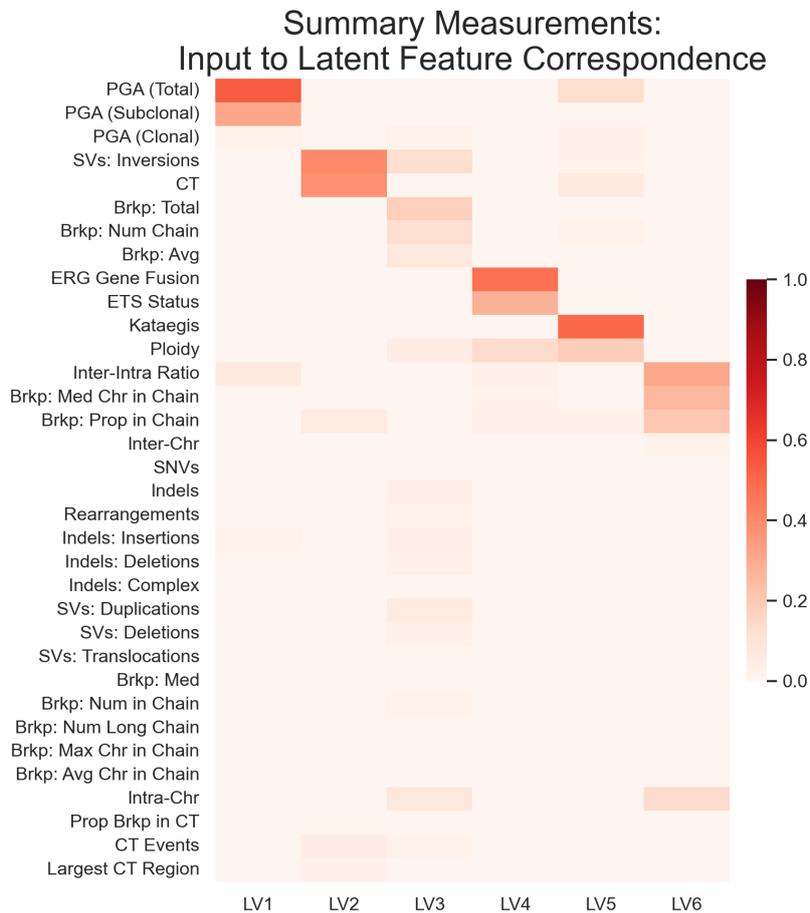


Figure 5.3: Input to latent feature correspondence heatmap, i.e. the encoding weights of iCS-GAN, as applied on the summary measurements dataset.

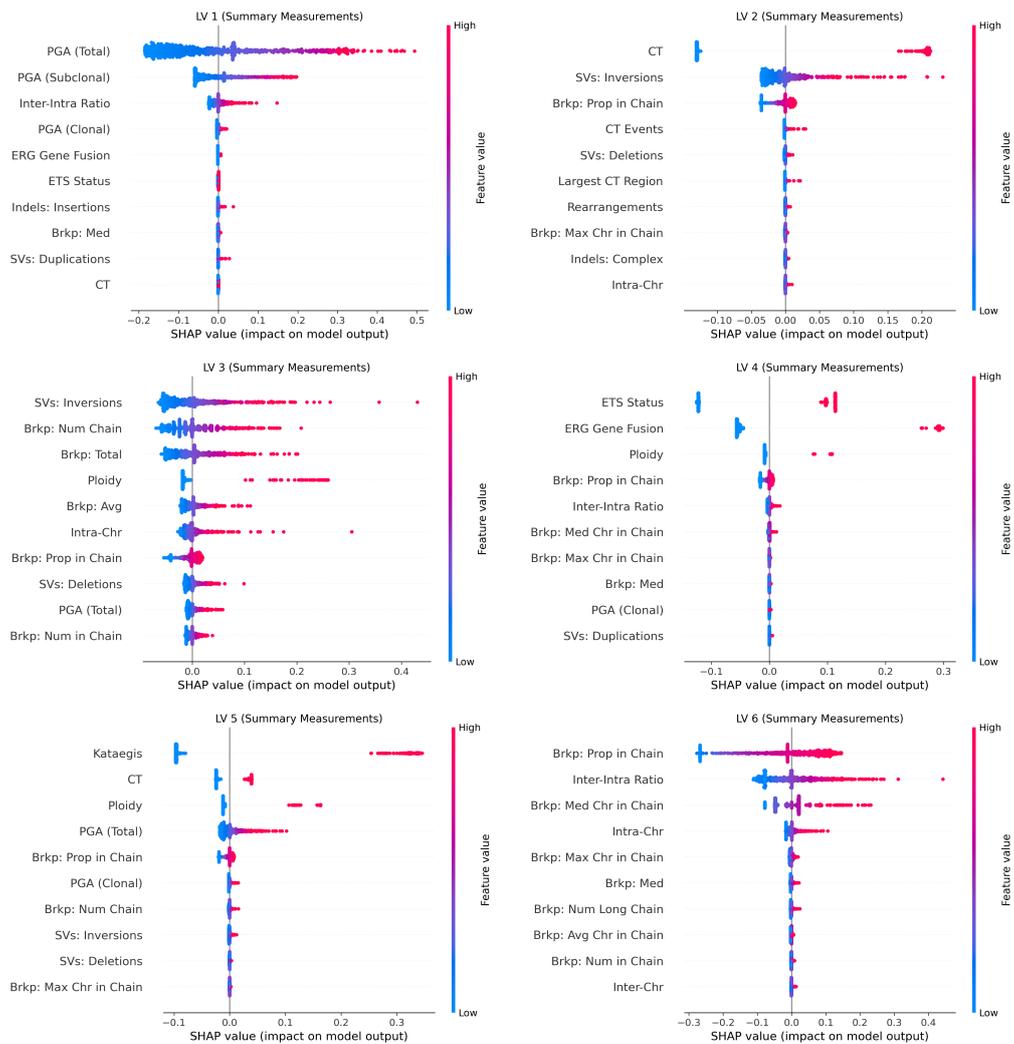


Figure 5.4: SHAP plots describing the importance of input features in the creation of each latent variable as applied on the summary measurements dataset. For each latent variable, we plot the top 10 most important inputs.

5.1.2 Subtyping with Drivers

Applied to the driver genes dataset, iCS-GAN returned 5 distinct prostate cancer subtypes, here named Dr-1, Dr-2, Dr-3, Dr-4 and Dr-5. The average test set reconstruction error was 0.08 for the 5 runs of the model, and the test set subtyping stability, as assessed with ARI, was 0.97.

Figure 5.5 visualises Kaplan-Meier survival plots for subtypes Dr-1 ($n = 55$), Dr-2 ($n = 48$), Dr-3 ($n = 74$), Dr-4 ($n = 47$) and Dr-5 ($n = 633$). No subtypes with significant survival differences were found using the driver genes dataset. This is rather unsurprising, given the low prevalence of driver gene mutations in prostate cancer - the most common driver gene mutation in the PPCG dataset (*SPOP*) affects less than 6.5% of all patients.

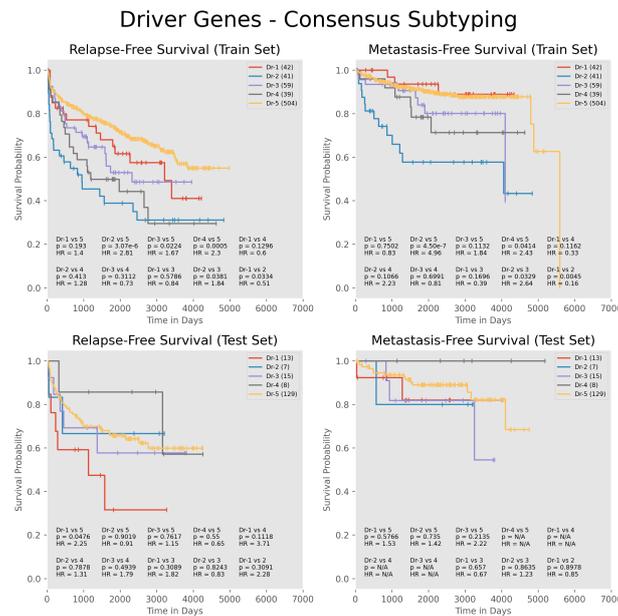


Figure 5.5: Kaplan-Meier plots visualising the relapse-free (left) and metastasis-free (right) survival for the driver genes subtypes Dr-1, Dr-2, Dr-3, Dr-4, Dr-5, separately for the training (top) and test set (bottom) samples. For each pairwise comparison, we provide the hazard ratio (HR) and the p value calculated with the Cox proportional hazard test. MFS test set comparisons with subtype Dr-4 were not meaningful as there were no events in the group. The number of samples in each subtype is given in brackets, next to the subtype label.

In terms of subtype characteristics (Figure 5.6), Dr-5 is a “quiet” subtype with no driver gene mutations. All tumors classified as Dr-2 had a *TP53* mutation, and all tumors classified as Dr-4 had an *SPOP* mutation. Subtype Dr-1 is characterized by a higher prevalence of mutations affecting *FOXA1* and *KMT2D*, and subtype Dr-3 by a higher prevalence of mutations affecting *PTEN*, *PIK3CA*, *CDK12* and *KDM6A* genes.

An RFC was able to classify patient samples as subtypes Dr-1 to Dr-5 with 0.98 test set accuracy.

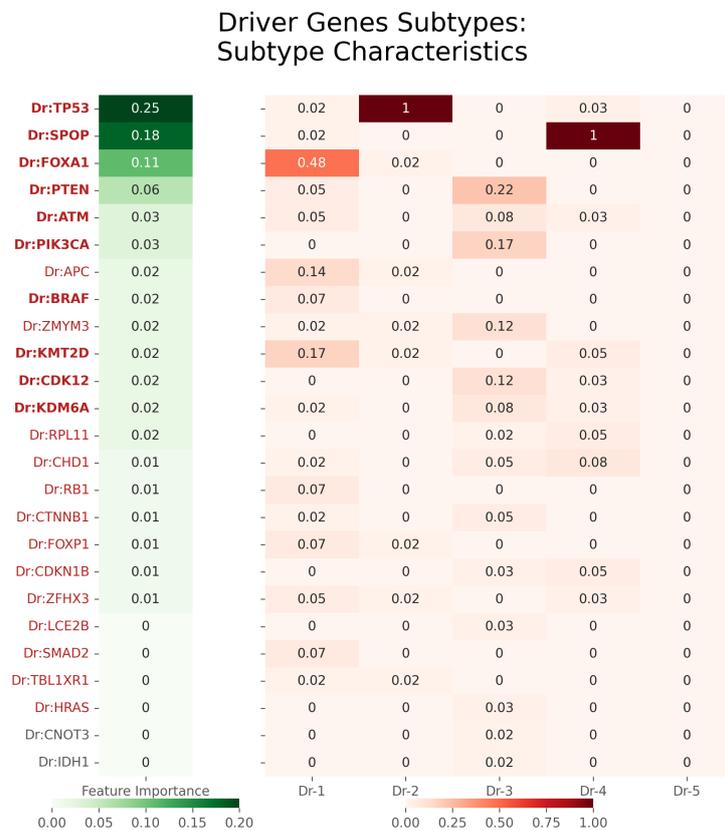


Figure 5.6: Characteristics of the prostate cancer subtypes uncovered from driver genes. Main heatmap visualises mean feature value in each subtype, as measured on the training set. Features are ordered by SHAP-based importance when predicting cluster assignments using RFC applied on the original data features. Aberrations with significant differences between subtypes are marked in bold red for features significant on both training and test sets or in light red for features significant on the training set only (FDR-adjusted $p < 0.05$, Chi-squared test).

The latent features created by iCS-GAN from the driver genes dataset (Figures 5.7 and 5.8) were fully interpretable, but given the sparsity of the drivers data, each extracted LV, rather than describing a specific set of mutations, corresponded more to the overall mutational burden.

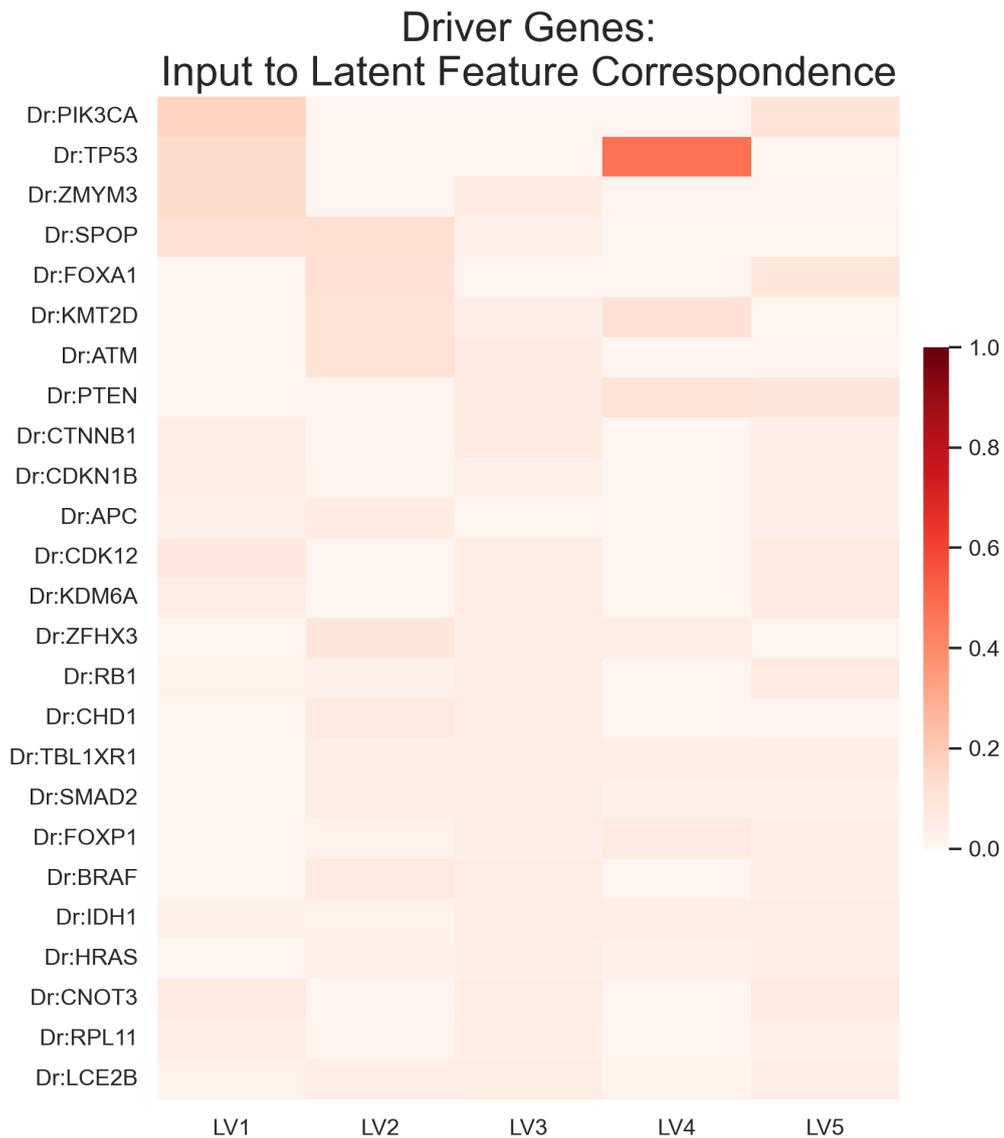


Figure 5.7: Input to latent feature correspondence heatmap, i.e. the encoding weights of iCS-GAN, as applied on the driver genes dataset.

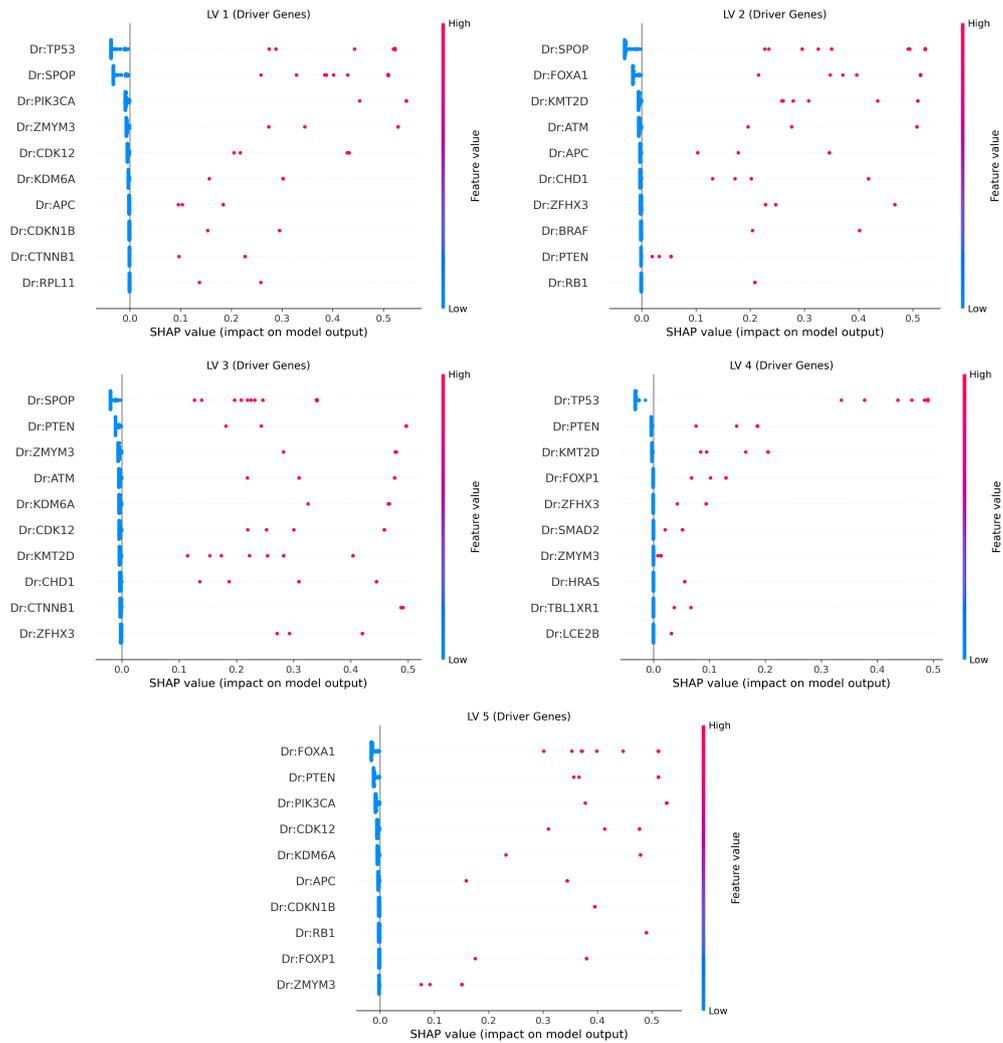


Figure 5.8: SHAP plots describing the importance of input features in the creation of each latent variable as applied on the driver genes dataset. For each latent variable, we plot the top 10 most important inputs.

5.1.3 Subtyping with CNAs

Applied to the CNA dataset, iCS-GAN returned 3 distinct prostate cancer subtypes, here named CNA-1, CNA-2 and CNA-3. The average test set reconstruction error was 0.32 for the 5 runs of the model, and the test set subtyping stability, as assessed with ARI, was 0.77.

Figure 5.9 visualises the survival differences between the subtypes. Compared to samples from CNA-3 ($n = 292$), tumours classified as CNA-1 ($n = 215$) and CNA-2 ($n = 314$) exhibit worse patient prognosis in terms of relapse-free survival (CNA-1 vs CNA-3: HR = 2.44, $p = 0.0197$; CNA-2 vs CNA-3: HR = 2.35, $p = 0.0193$), and metastasis-free survival for CNA-1 (CNA-1 vs CNA-3: HR = 5.73, $p = 0.0273$).

As for subtype characteristics (Figure 5.10), CNA-3 is a “quiet” subtype with a generally low CNA mutational burden. Compared to samples from other subtypes, tumours classified as CNA-1 are more commonly affected by losses of heterozygosity on chromosomes 1, 2, 5, 6 and 13, gains on chromosomes 3, 7 and 8, and homozygous deletions on chromosomes 5 and 13. Losses on chromosomes 3, 8, 10, 16, 17, 18 and 21, and deletions on chromosome 10 are more prevalent in tumours classified as CNA-2.

An RFC classified patient samples as subtypes CNA-1, CNA-2 and CNA-3 with 0.94 test set accuracy.

The latent features created by iCS-GAN from the CNA data were both fully interpretable and biologically-relevant, with each latent variable describing a specific set of genetic aberrations on specific locations. As evident from Figures 5.11 and 5.12, the 8 latent variables extracted broadly corresponded to: losses on chromosomes 3, 17 and 21 (LV1), gains on chromosomes 3, 7, 8, 19 and 22 (LV2), losses on chromosomes 8 and 12 (LV3), losses on chromosomes 1, 5, 17 and 18 (LV4), aberrations on chromosomes 3 and 10 (LV5),

losses on chromosomes 5 and 13 and deletions on chromosome 5 (LV6), losses on chromosomes 2, 5 and 6 (LV7), and losses on chromosomes 1, 8, 16 and 17 and gains on chromosomes 8 and 19 (LV8).

CNAs - Consensus Subtyping

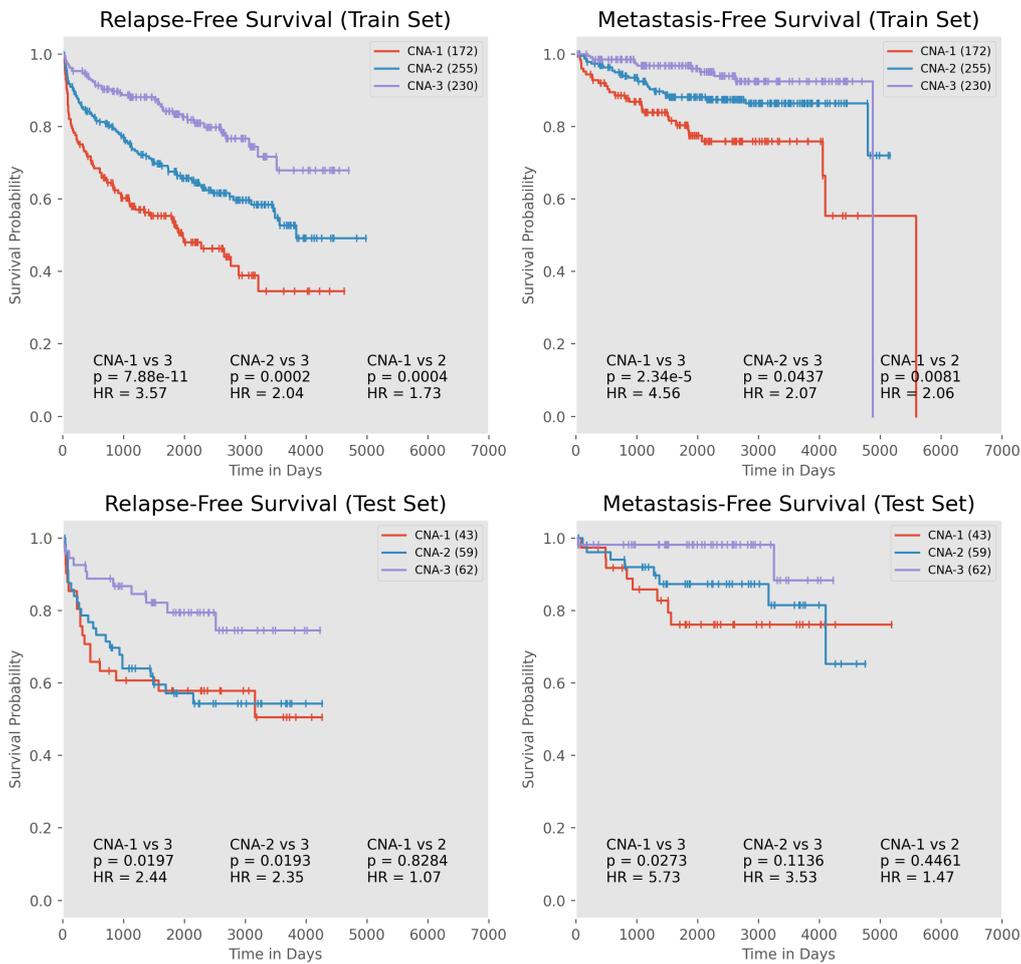


Figure 5.9: Kaplan-Meier plots visualising the relapse-free (left) and metastasis-free (right) survival for CNA subtypes CNA-1, CNA-2 and CNA-3, separately for the training (top) and test set (bottom) samples. For each pairwise comparison, we provide the hazard ratio (HR) and the p value calculated with the Cox proportional hazard test. The number of samples in each subtype is given in brackets, next to the subtype label.

CNA Subtypes: Subtype Characteristics

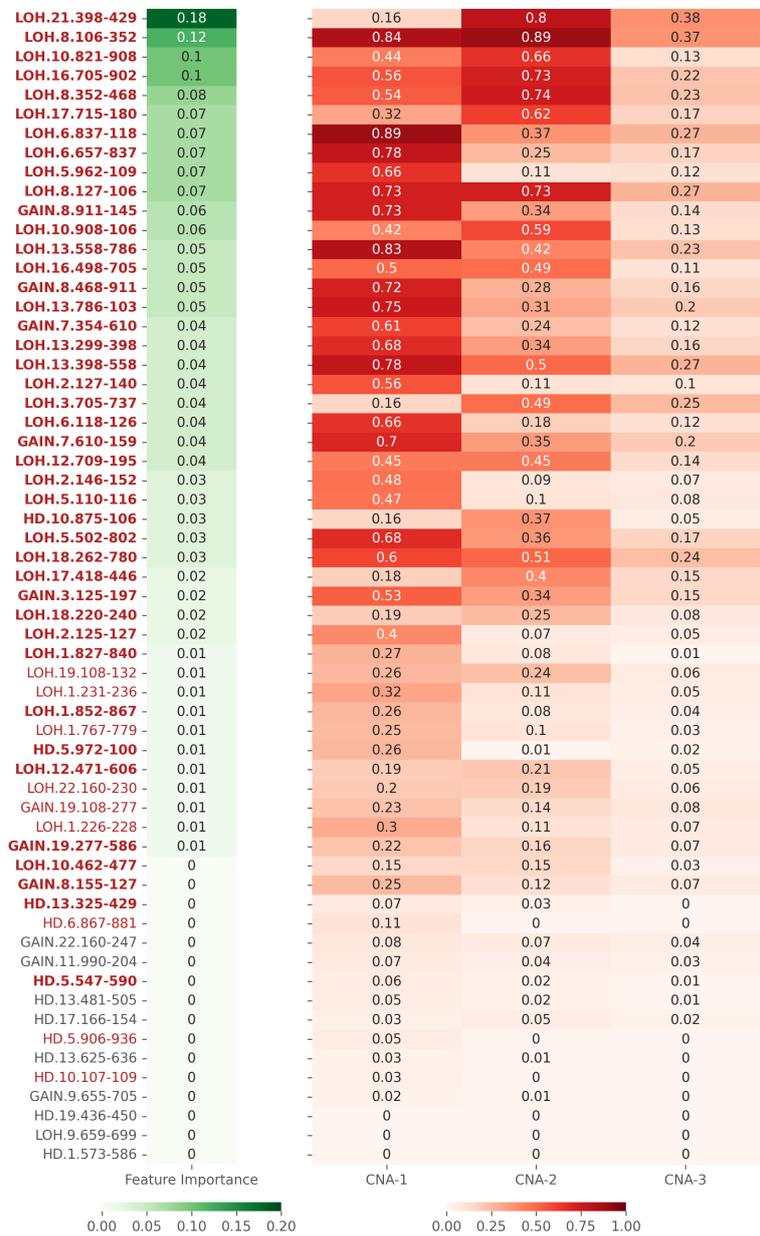


Figure 5.10: Characteristics of the prostate cancer subtypes uncovered from CNAs. Main heatmap visualises mean feature value in each subtype, as measured on the training set. Features are ordered by SHAP-based importance when predicting cluster assignments using RFC applied on the original data features. Aberrations with significant differences between subtypes are marked in bold red for features significant on both training and test sets or in light red for features significant on the training set only (FDR-adjusted $p < 0.05$, Chi-squared test).



Figure 5.11: Input to latent feature correspondence heatmap, i.e. the encoding weights of iCS-GAN, as applied on the CNA dataset.

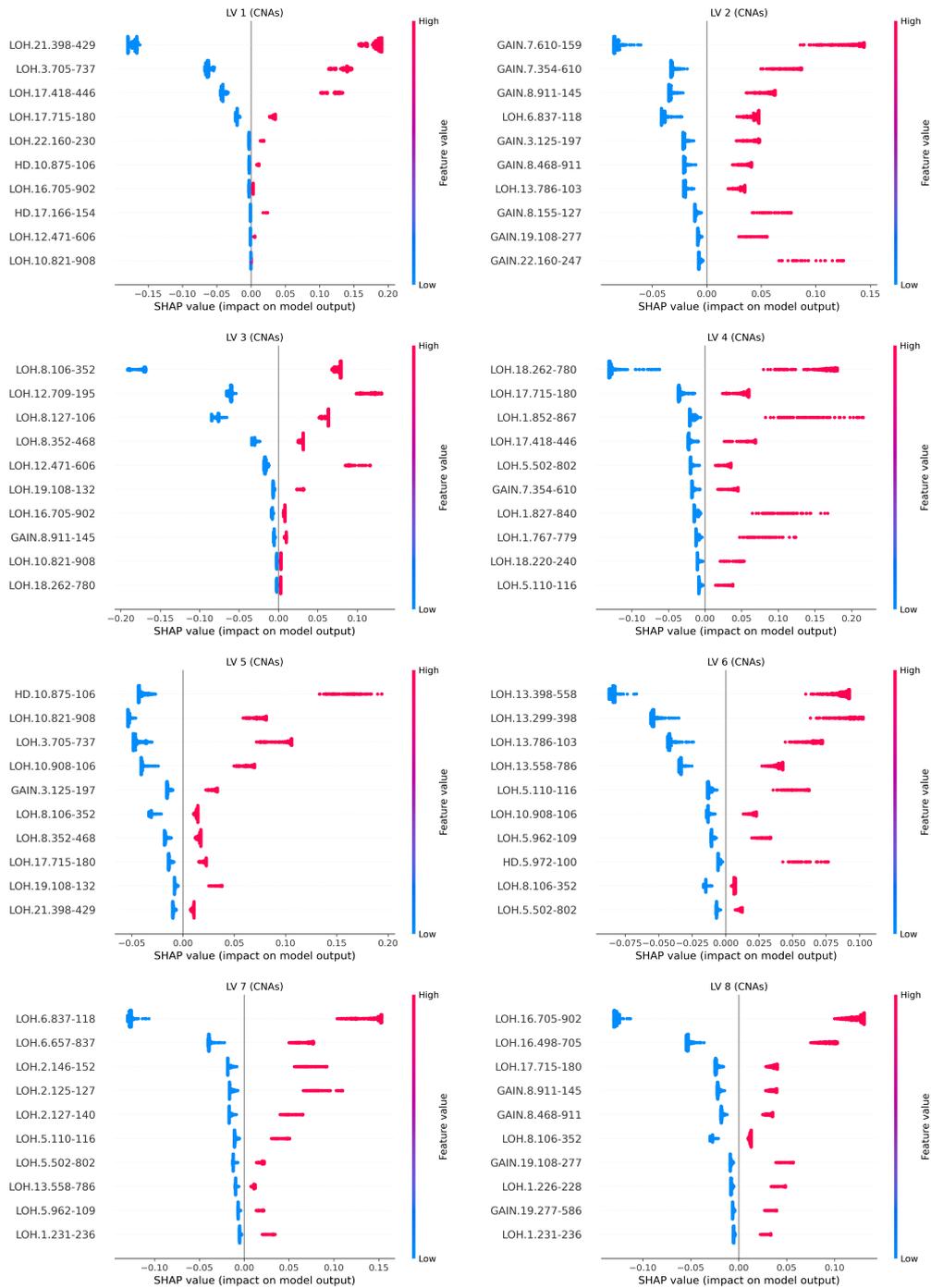


Figure 5.12: SHAP plots describing the importance of input features in the creation of each latent variable as applied on the CNA datasets. For each latent variable, we plot the top 10 most important inputs.

5.1.4 Subtyping with RNA

Applied to the RNA dataset, iCS-GAN returned 3 distinct prostate cancer subtypes, here named RNA-1, RNA-2 and RNA-3. The average test set reconstruction error was 0.15 for the 5 runs of the model, and the test set subtyping stability, as assessed with ARI, was 0.82.

Figure 5.13 visualises Kaplan-Meier survival plots for subtypes RNA-1 ($n = 371$), RNA-2 ($n = 373$) and RNA-3 ($n = 491$). No subtypes with significant survival differences were found using the RNA dataset. This is unsurprising given the intrinsic noisiness and the large dimensionality of the RNA dataset.

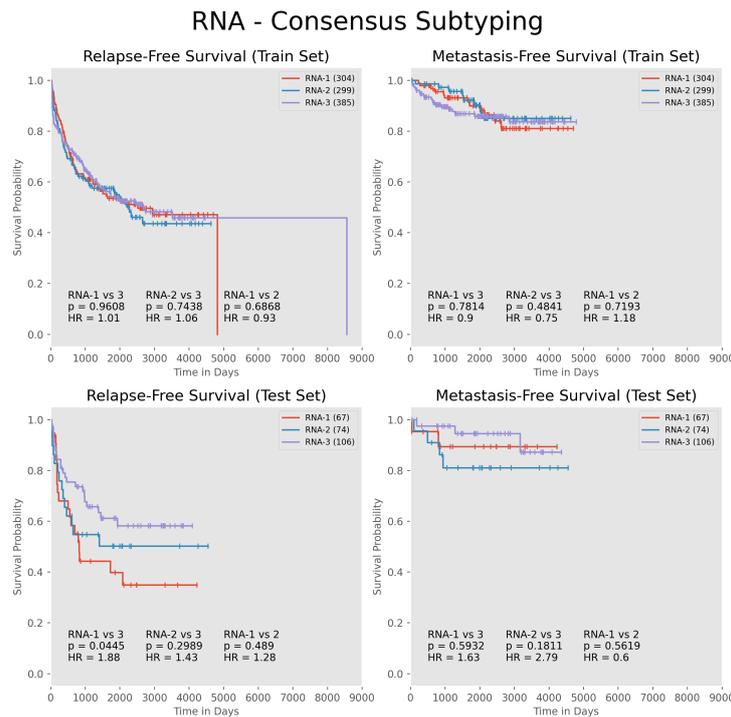


Figure 5.13: Kaplan-Meier plots visualising the relapse-free (left) and metastasis-free (right) survival for RNA subtypes RNA-1, RNA-2 and RNA-3, separately for the training (top) and test set (bottom) samples. For each pairwise comparison, we provide the hazard ratio (HR) and the p value calculated with the Cox proportional hazard test. The number of samples in each subtype is given in brackets, next to the subtype label.

In terms of subtype characteristics (Figure 5.14), RNA-1 is characterized by the over-expression of *COL17A1*, *S100A14*, *LGR6*, *EVX2*, *GPX2*, *PGM5-AS1*, *FAT2*, *GJB5*, *KRT5*, *IGSF1*, *WFDC2*, *CPA6*, *IP6K3* and *ADRA1D* genes. In RNA-2, *TFF3*, *POTEH-AS1*, *SLC26A3*, *ANPEP*, *ALOX15B*, *RLN1*, *SLC38A4*, *P3H2*, *LOC100508046*, *GMNC* are over-expressed and *ERG*, *EPIC1*, *OGDHL*, *CDH7*, *NKAIN1*, *PAX1*, *PCAT5*, *KIAA0087*, *ALOX15*, *ANKRD34B*, *KCNH8*, *EML6*, *GREM1-AS1*, *LINC03095*, *INSM1*, *LINC02418*, *CD8B2*, *LINC01019*, *ARSH*, *DACT2*, *KCNN4*, *KCNG3*, *MAGED4B*, *SLC18A3*, *LOC101930421*, *CPA6* are under-expressed. Finally, *ERG*, *EPIC1*, *OGDHL*, *CDH7*, *NKAIN1*, *PAX1*, *PCAT5*, *KIAA0087*, *ALOX15*, *ANKRD34B*, *KCNH8*, *EML6*, *GREM1-AS1*, *LINC03095*, *INSM1*, *LINC02418*, *LINC01019*, *ARSH*, *KCNG3*, *SLC18A3*, *LOC101930421* are over-expressed and *TFF3*, *ANPEP*, *ALOX15B*, *RLN1*, *SLC38A4*, *P3H2* are under-expressed in RNA-3.

An RFC classified patient samples as subtypes RNA-1, RNA-2 and RNA-3 with 0.89 test set accuracy.

The latent features extracted by iCS-GAN from the RNA data corresponded to cancer-relevant biological signatures. For example (Figure 5.15), LV2 described mostly the protocadherins (*PCDHA*) gene family, primarily expressed in the nervous system and associated with cell proliferation and death [42, 49, 131]. LV5 corresponded mostly to the immunoglobulin heavy chain variable region gene family (*IGHV*), which form part of the receptors in B-cells - possibly hinting at immunogenic behaviour, something not normally associated with prostate cancer. LV11 described the *ERG* prostate cancer oncogene - part of the *ETS* gene family whose relationship to prostate cancer prognosis is still not fully understood with studies showing contradictory results [51, 66, 142, 154, 202] - and *PCAT5*, a potential prostate cancer

therapeutic target [225]. LV16 corresponded mostly to the *FAM223A* gene known to facilitate colorectal cancer progression [233]. LV25 described genes from the *HERC* family, a suspected influencer of breast carcinogenesis [216]. Finally, LV26 focused on the olfactory receptor (*OR*) family, a potential breast-cancer biomarker [11].

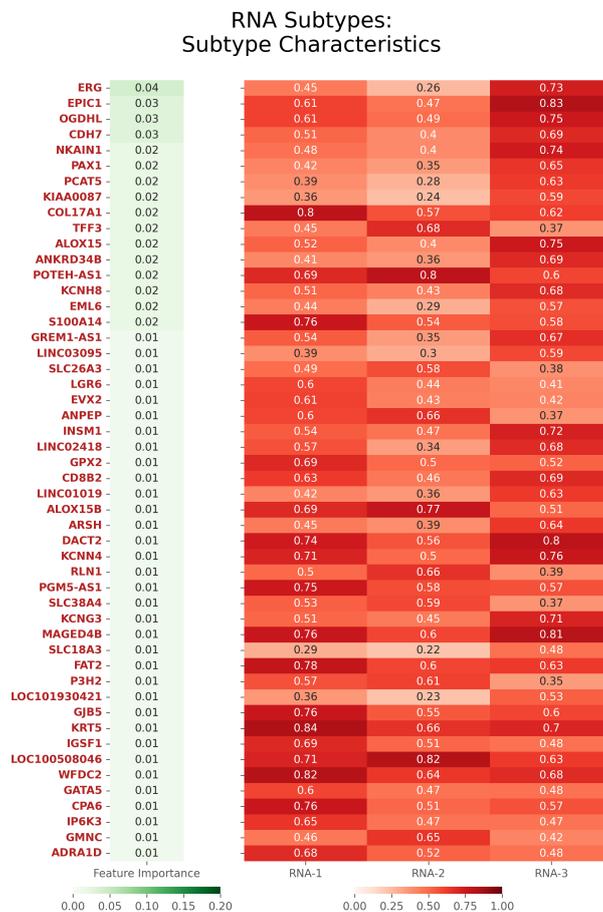


Figure 5.14: Characteristics of the prostate cancer subtypes uncovered from RNA. Main heatmap visualises mean feature value in each subtype, as measured on the training set. Features are ordered by SHAP-based importance when predicting cluster assignments using RFC applied on the original data features. Aberrations with significant differences between subtypes are marked in bold red for features significant on both training and test sets or in light red for features significant on the training set only (FDR-adjusted $p < 0.05$, one-way anova test). For clarity, we limit the visualisation to 50 most important features.

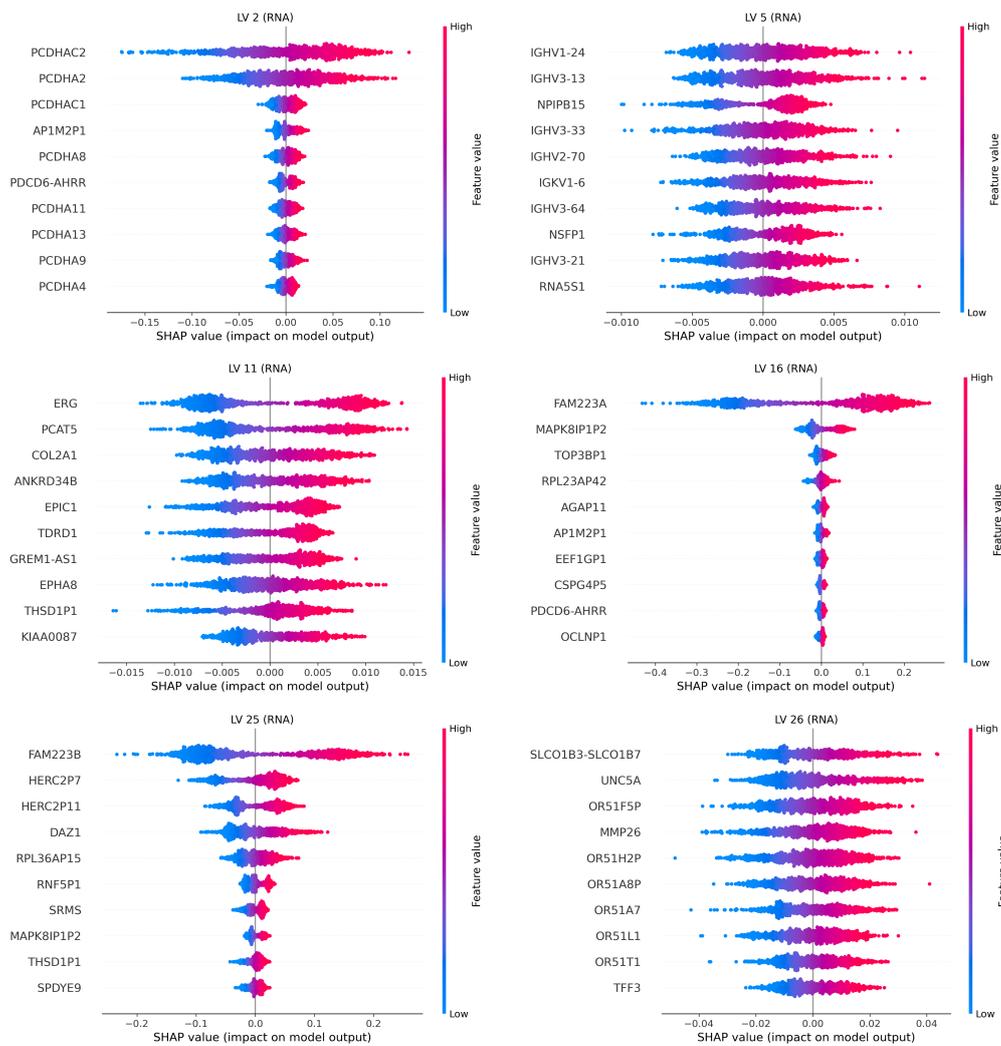


Figure 5.15: SHAP plots describing the importance of input features in the creation of selected latent variables as applied on the RNA dataset. For each latent variable, we plot the top 10 most important inputs. Note that we do not show the encoding weight heatmap for the RNA dataset as it is largely unreadable with 1000 inputs and 32 latent variables.

5.2 Multiomic Subtyping

Multiomic analysis of the PPCG dataset with iCS-GAN returned 3 distinct prostate cancer subtypes, here named M-1, M-2 and M-3. The average test set reconstruction error (measured on the complete samples only) was 0.17 for the 5 runs of the model, and the test set subtyping stability, as assessed with ARI, was 0.73. Figure 5.16 illustrates the subtypes identified in a single run of the model, considering only the complete data samples. The clusters are clearly distinguished in UMAP derived from the latent space, unlike the observed space. Figure 5.17 extends the latent space visualization to include the entire dataset, encompassing both complete and incomplete samples, with the latter having their encodings imputed. Notably, both complete and imputed samples are distributed roughly evenly across the three subtypes, regardless of the available modalities for each sample. This demonstrates the effectiveness of the imputation process.

Each of the remaining 4 runs of the model returned subtyping results similar to those visualised in Figures 5.16 and 5.17. As such, for our multiomic subtyping analysis, we use consensus subtype assignments. These assignments were determined by grouping each sample into a cluster based on the majority vote from the 5 runs of the model. Figure 5.18 visualises the survival differences between the subtypes. Compared to samples from M-2 ($n = 450$) and M-3 ($n = 734$), tumours classified as M-1 ($n = 443$) exhibit worse patient prognosis in terms of relapse-free survival (SM-1 vs SM-2: HR = 1.88, p = 0.0452; SM-1 vs SM-3: HR = 1.9, p = 0.0097), making subtype M-1 most aggressive.

For the remainder of this section, we describe the molecular characteristics of the multiomic subtypes M-1, M-2, and M-3 (Section 5.2.1), analyse the latent variables identified by iCS-GAN (Section 5.2.2), examine the clinical

profiles associated with each subtype (Section 5.2.3), compare the multiomic subtypes with those uncovered via single-modality analysis (Section 5.2.4), and illustrate the potential for the development of predictive clinical tests (Section 5.2.5).

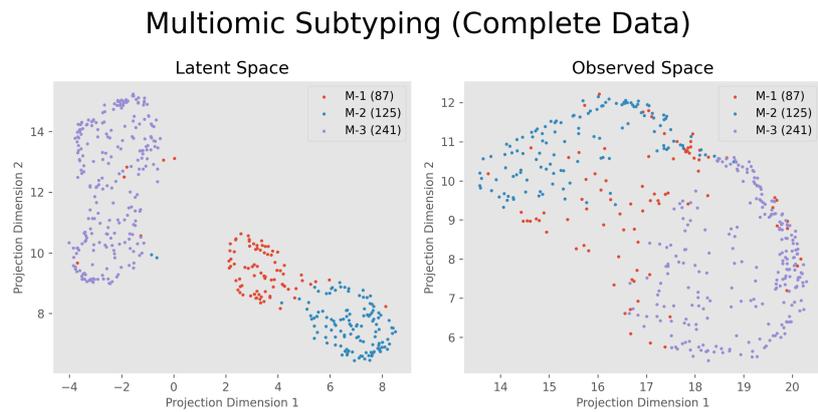


Figure 5.16: UMAP visualisation of the multiomic prostate cancer subtypes uncovered with iCS-GAN. Latent (left) and observed (right) space visualisations. The number of samples in each subtype is given in brackets, next to the subtype label. The plots correspond to the modality-complete subset of the PPCG dataset only, and a single run of iCS-GAN.

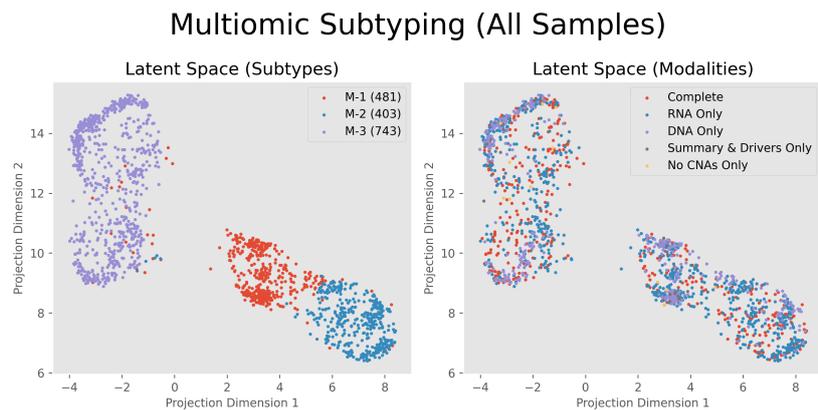


Figure 5.17: UMAP visualisation of the multiomic prostate cancer subtypes uncovered with iCS-GAN. Latent space visualisation for the subtypes (left) and imputation classes (right). The number of samples in each subtype is given in brackets, next to the subtype label. The plots correspond to the entire PPCG dataset, and a single run of iCS-GAN.

Multimomic Analysis - Consensus Subtyping

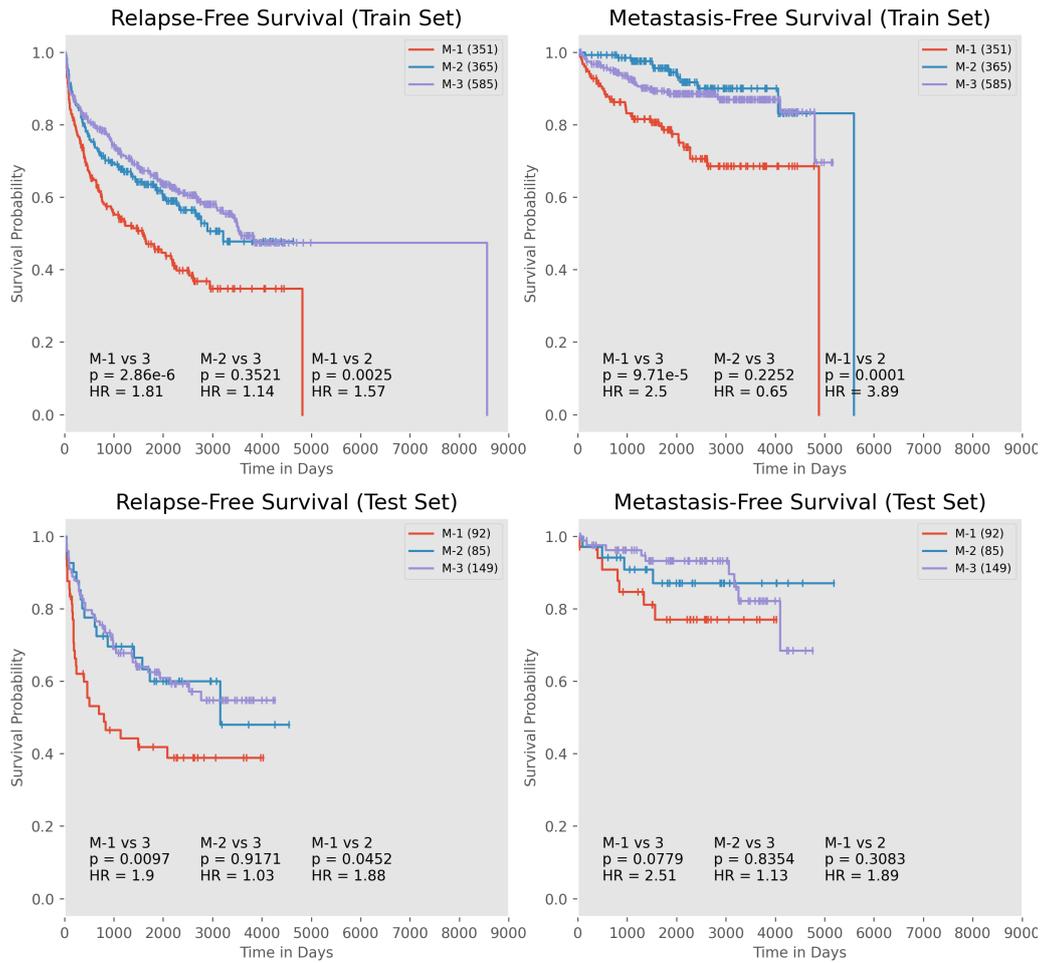


Figure 5.18: Kaplan-Meier plots visualising the relapse-free (left) and metastasis-free (right) survival for multimomic subtypes M-1, M-2 and M-3, separately for the training (top) and test set (bottom) samples. For each pairwise comparison, we provide the hazard ratio (HR) and the p value calculated with the Cox proportional hazard test. The number of samples in each subtype is given in brackets, next to the subtype label.

5.2.1 Subtype Characteristics

In terms of subtype characteristics (Figures 5.19 - 5.22), clear definitions of subtypes M-2 and M-3 emerge. Subtype M-3 is characterized by positive *ETS* status, prevalent *ERG* gene fusions, highest inter to intra chromosomal breakpoint ratio, highest average, median and maximum number of chromosomes involved in breakpoint chains, as well as highest median number of breakpoints, proportion of breakpoints in chains, intra-chromosomal events and the number of translocations. Except for the *ERG* gene fusions, which are not present in M-1 and M-2, and the intra-chromosomal events, where the mean value is equal for subtypes M-1 and M-2, the mean value of the aforementioned features is generally lowest for samples from M-2, and intermediate for samples from M-1.

SPOP mutations are present in a fractions of samples from M-2, and absent in M-1 and M-3.

As for copy number alterations, tumours classified as M-3 are more commonly affected by losses on chromosomes 3, 10, 17 and 21, and deletions on chromosome 10. Losses on chromosomes 2, 5 and 6, gains on chromosome 8, and deletions on chromosome 5 are more prevalent in tumours classified as M-2. M-1 is once again the ‘intermediate’ subtype with higher ratio of losses on chromosomes 3, 10, 17 and 21 and deletions on chromosome 10 than in samples from M-2, and more prevalent losses on chromosome 5 and gains on chromosome 8 than for samples from M-3.

In terms of the RNA data, *ERG*, *PCAT5*, *LOC101930421*, *KIAA0087*, *GREM1-AS1*, *TDRD1*, *GDA*, *CHRM3* genes are under-expressed in M-1 and M-2. Additionally, *SERPINA11*, *C1orf87*, *OR51F5P*, *OR51H2P*, *PHGR1*, *MMP26*, *TRPC7-AS1*, *OR51A8P*, *SMIM32* are under-expressed in M-1 and *OGDHL*, *EPIC1*, *ALOX15*, *KCNH8*, *EML6*, *NKAIN1*, *ELFN1*-

AS1, *LINC02418*, *FAM87A*, *CD8B2*, *ADARB2*, *MDFI*, *DACT2*, *SLC18A3*, *CDH7*, *LINC03095*, *KCNG3*, *DLX6*, *ANKRD34B* are under-expressed in M-2. All of the aforementioned genes, as well as *MAGED4B* and *PLPPR1* are over-expressed in M-3, where *TFF3*, *PCOTH* and *ANPEP* are under-expressed. None of the 50 most important (as measured with SHAP scores for subtype predictions) genes are over-expressed in M-1, and *TFF3*, *PCOTH*, *POTEH-AS1*, *ALOX15B*, *ANPEP*, *RLN1*, *POTEB3*, *HCAR1*, *HAUS1P2*, *LOC100508046* and *UNC5A* are overexpressed in M-2.

Multiomic Subtypes:
Subtype Characteristics (Summary Measurements)

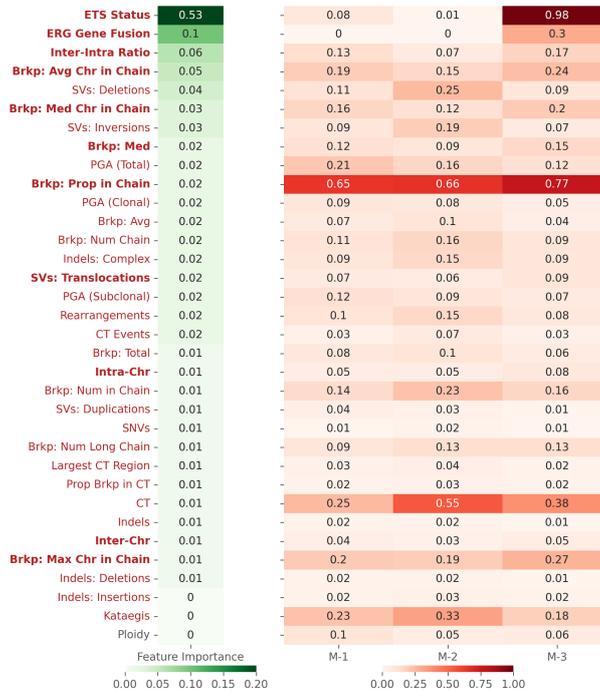


Figure 5.19: Summary measurements based characteristics of the multiomic prostate cancer subtypes. Main heatmap visualises mean feature value in each subtype, as measured on the training set. Features are ordered by SHAP-based importance when predicting cluster assignments using RFC applied on the original data features. Aberrations with significant differences between subtypes are marked in bold red for features significant on both training and test sets or in light red for features significant on the training set only (FDR-adjusted $p < 0.05$, Chi-squared (categorical variables) or Kruskal-Wallis (continuous variables) test).

Multiomic Subtypes: Subtype Characteristics (Driver Genes)

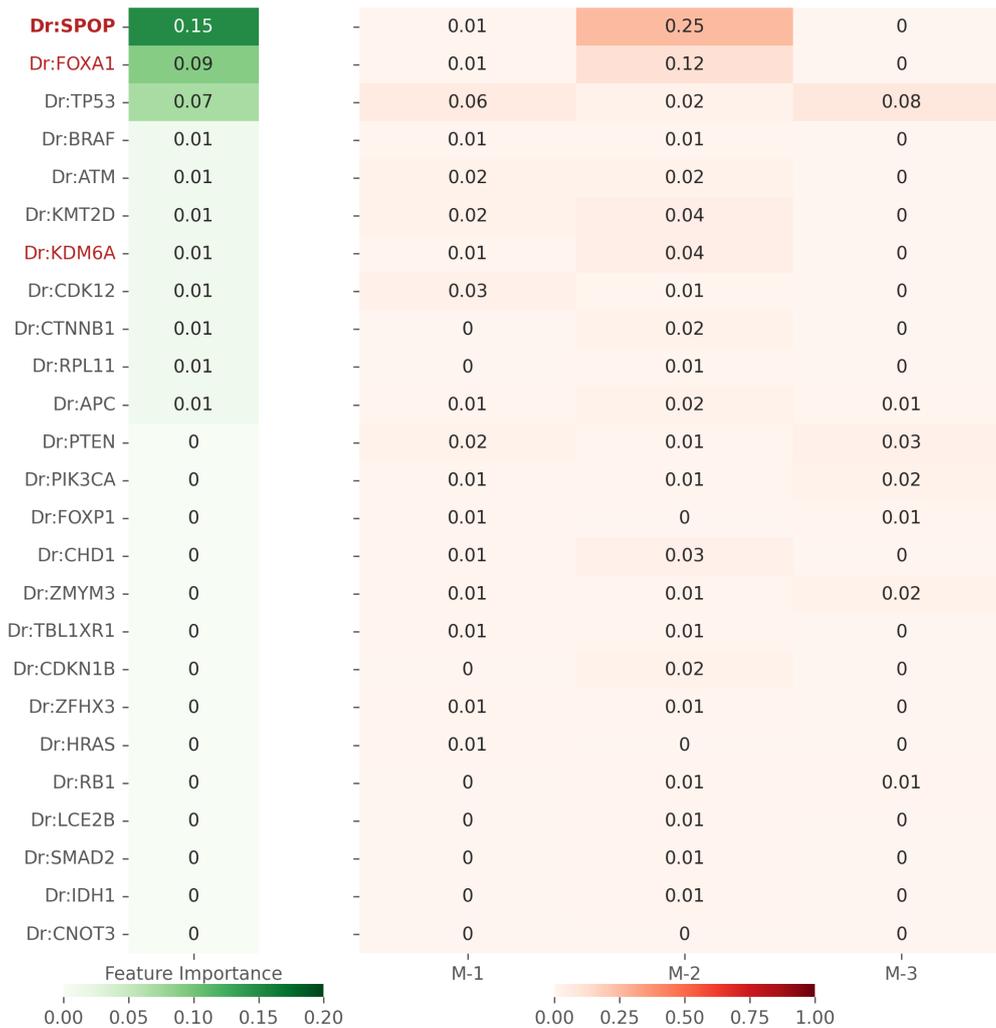


Figure 5.20: Driver genes based characteristics of the multiomic prostate cancer subtypes. Main heatmap visualises mean feature value in each subtype, as measured on the training set. Features are ordered by SHAP-based importance when predicting cluster assignments using RFC applied on the original data features. Aberrations with significant differences between subtypes are marked in bold red for features significant on both training and test sets or in light red for features significant on the training set only (FDR-adjusted $p < 0.05$, Chi-squared test).

Multiomic Subtypes: Subtype Characteristics (CNAs)

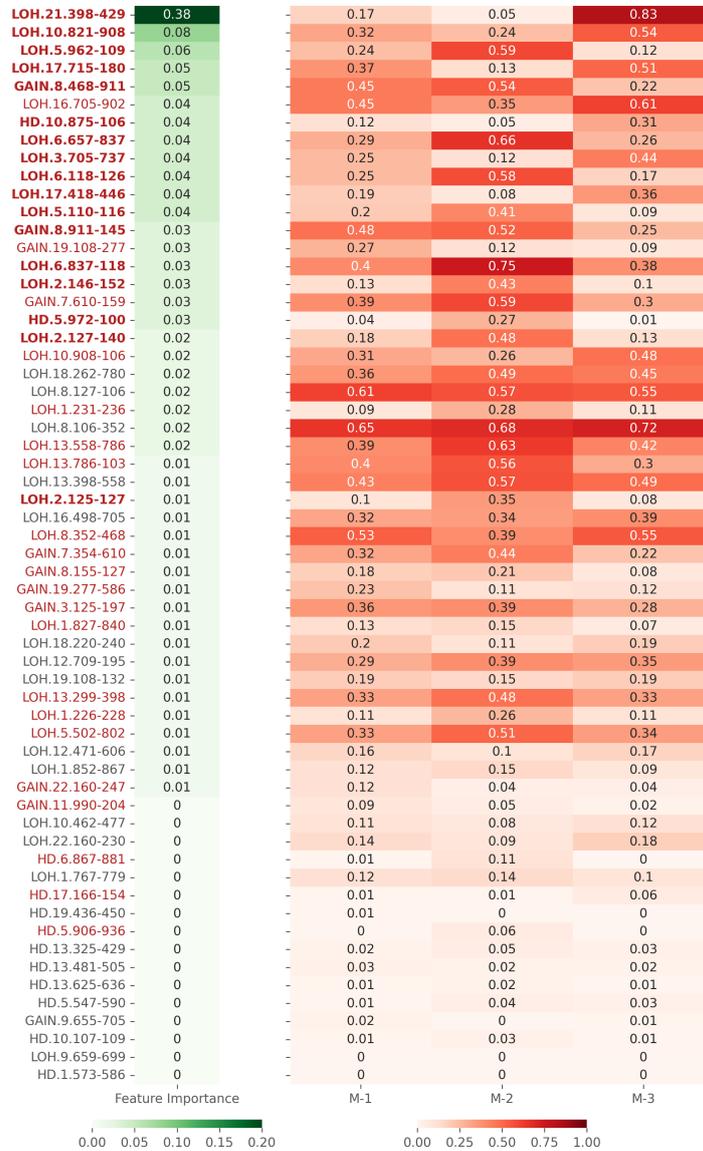


Figure 5.21: CNA based characteristics of the multiomic prostate cancer subtypes. Main heatmap visualises mean feature value in each subtype, as measured on the training set. Features are ordered by SHAP-based importance when predicting cluster assignments using RFC applied on the original data features. Aberrations with significant differences between subtypes are marked in bold red for features significant on both training and test sets or in light red for features significant on the training set only (FDR-adjusted $p < 0.05$, Chi-squared test).

Multiomic Subtypes: Subtype Characteristics (RNA)

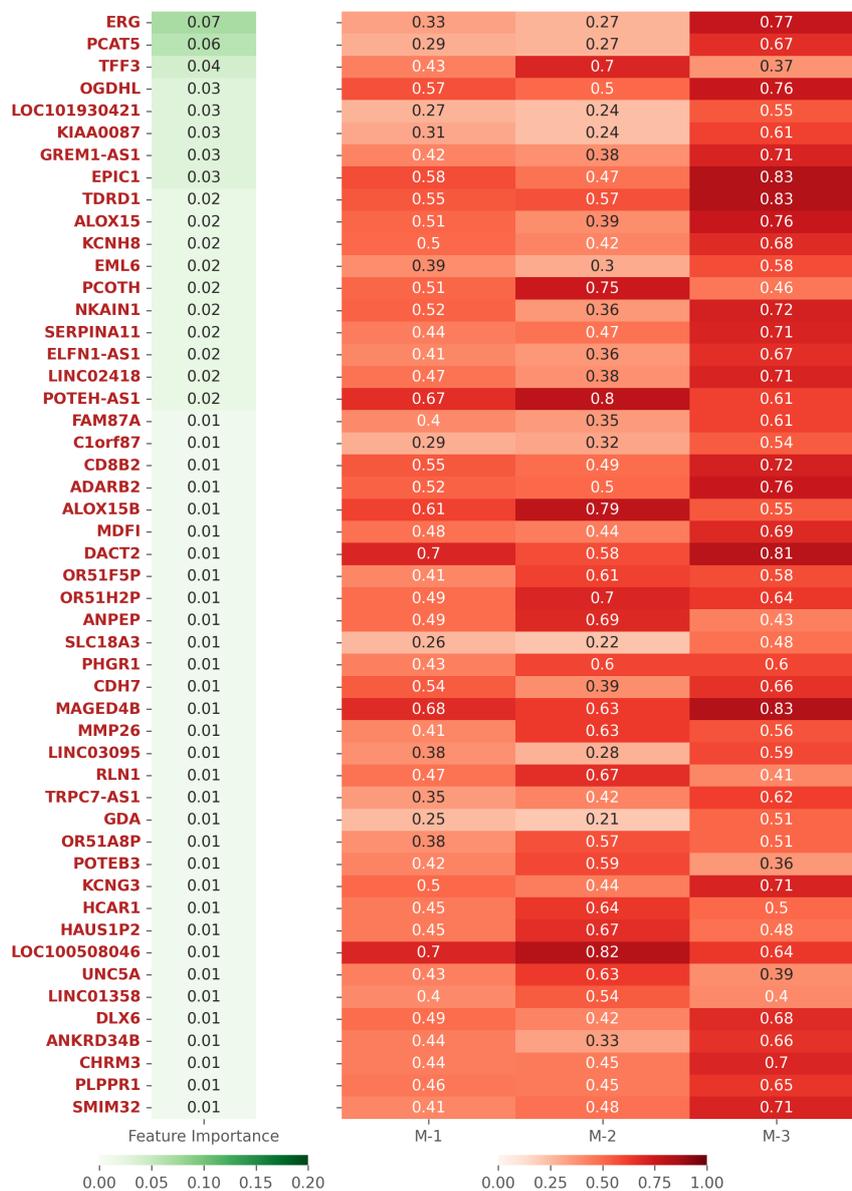


Figure 5.22: RNA based characteristics of the multiomic prostate cancer subtypes. Main heatmap visualises mean feature value in each subtype, as measured on the training set. Features are ordered by SHAP-based importance when predicting cluster assignments using RFC applied on the original data features. Aberrations with significant differences between subtypes are marked in bold red for features significant on both training and test sets or in light red for features significant on the training set only (FDR-adjusted $p < 0.05$, one-way anova test). For clarity, we limit the visualisation to 50 most important features.

5.2.2 Latent Variables

The latent features extracted from the entire PPCG dataset by iCS-GAN were fully interpretable and biologically relevant. As evident from Figures 5.23 - 5.25, the features encapsulated relationships both within and between data from different modalities. For example, LV2 described kataegis, chromothripsis and losses on chromosomes 6, 8 and 12. LV4 characterized the *ETS* status, *ERG* gene fusions, *ERG* expression levels, and losses on chromosome 21, where the *ERG* gene is located. LV16 focused on gains on chromosomes 7, 8, 19 and 22, LV17 described various gene expression levels, including the SAA (Serum Amyloid) family and the Serpin family. Finally, LV19 characterized the events on chromosomes 2, 16, 17 and 18, including the *TP53* and *SPOP* mutations.

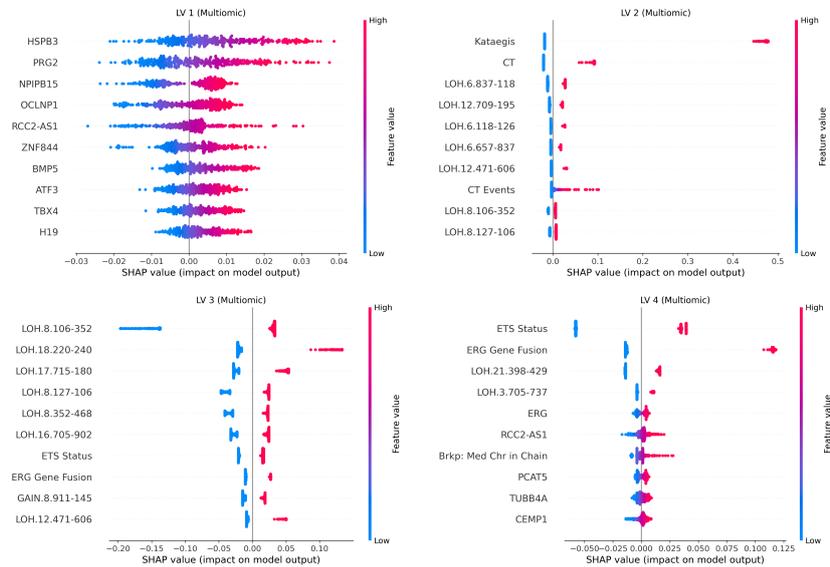


Figure 5.23: SHAP plots describing the importance of input features in the creation of each of multitomic latent variables LV1-LV4. For each latent variable, we plot the top 10 most important inputs.

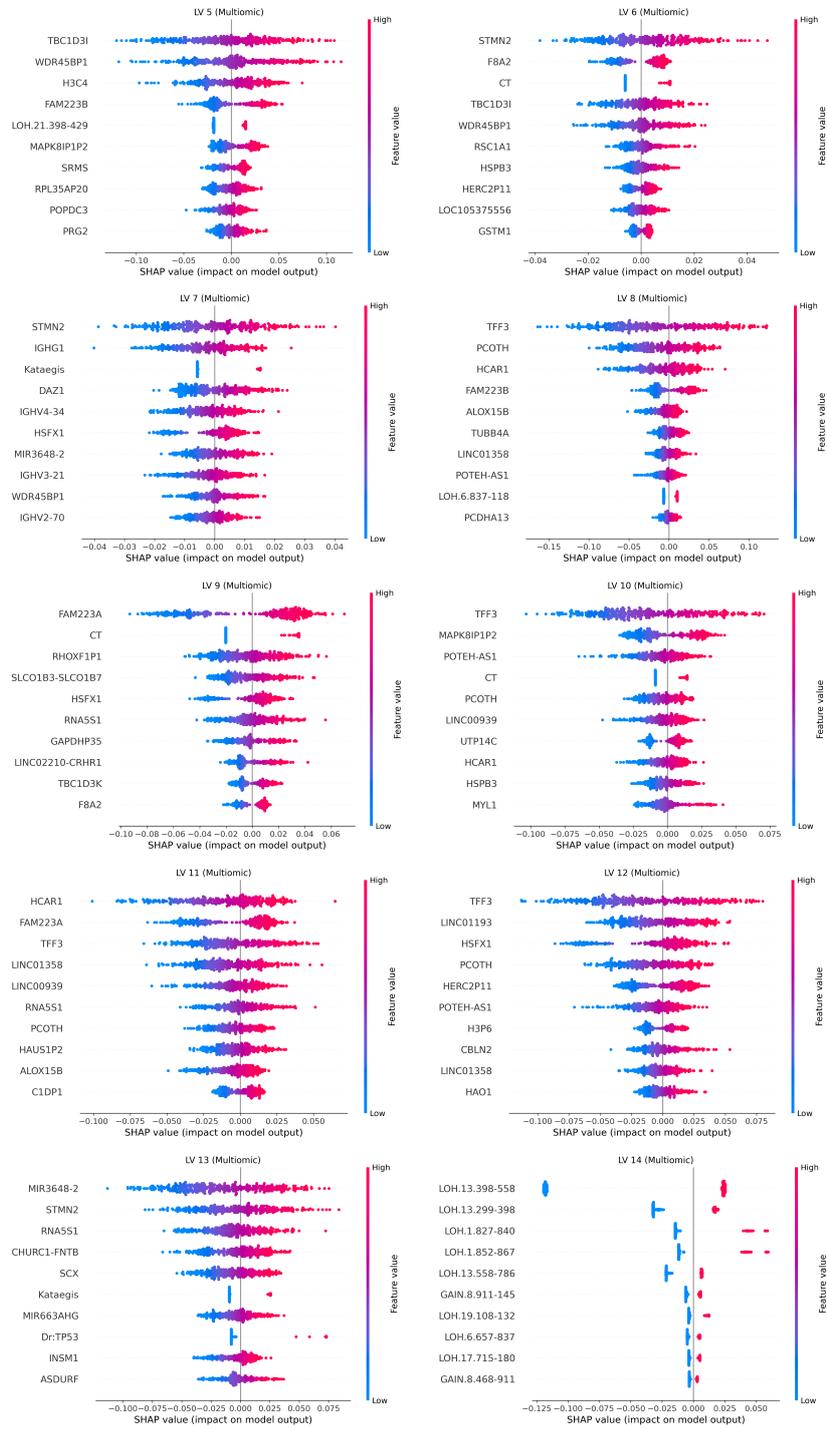


Figure 5.24: SHAP plots describing the importance of input features in the creation of each of multiomic latent variables LV5-LV14. For each latent variable, we plot the top 10 most important inputs.

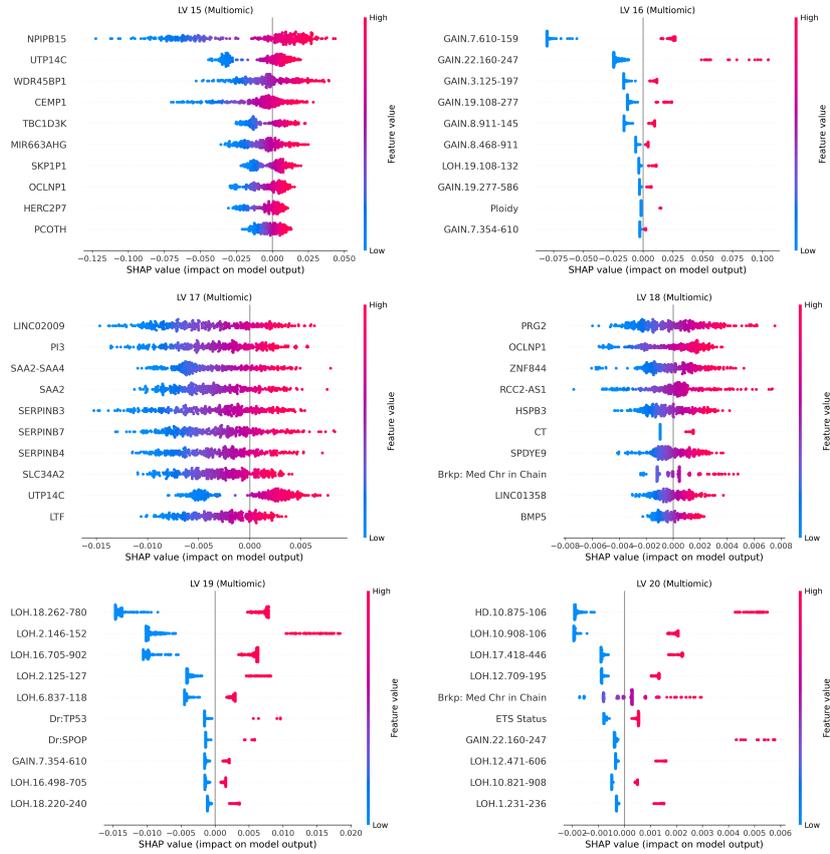


Figure 5.25: SHAP plots describing the importance of input features in the creation of multiomic latent variables LV15-LV20. For each latent variable, we plot the top 10 most important inputs.

As is evident from Figure 5.26, iCS-GAN allowed us to represent patient samples in a form of easily interpretable binary encodings, approximating the Indian Buffet Process representation. Furthermore, some of the most important characteristics of each subtype could have been identified directly from the encoding heatmaps and the corresponding LV SHAP plots. The heatmap in Figure 5.26 clearly points towards the higher activation of LV4 (positive ETS Status, *ERG* Gene fusions, losses on chromosome 21, overexpression of *ERG*) in M-3 and LVs 8, 10 and 12 (over-expression of *TFF3*, *PCOTH* and *HAR1*) in M-2. None of these LVs are highly activated in M-1, suggesting the under-expression of *ERG* and *TFF3*.

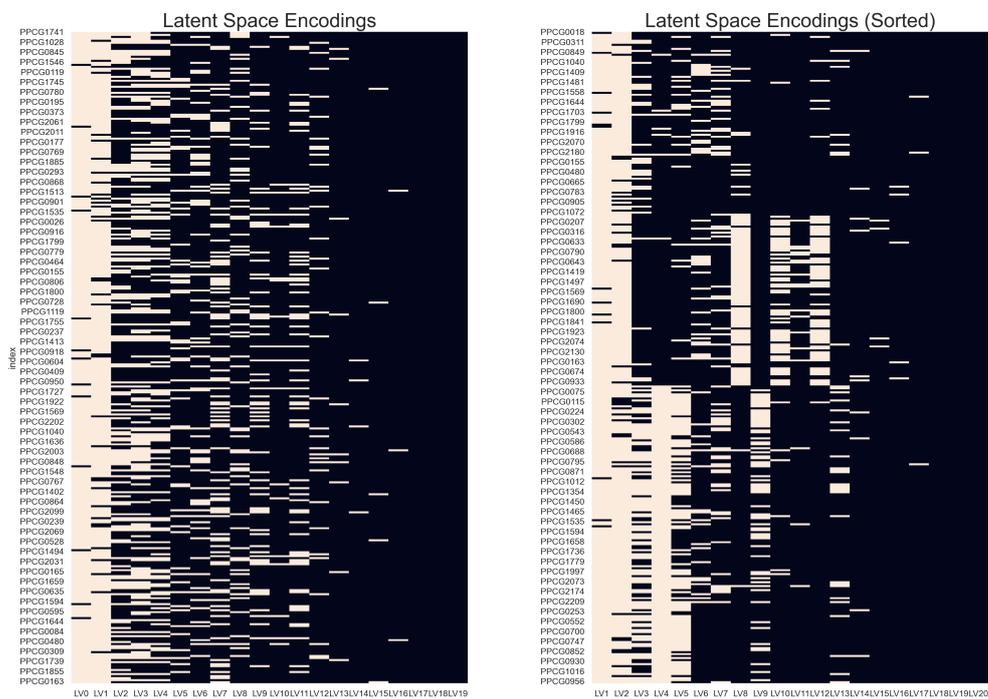


Figure 5.26: Heatmaps visualising the binarized test set latent space encodings as obtained on the entire PPCG dataset, in a random order (left) and sorted by subtype membership (right). Heatmap rows represent latent variables, columns represent patient samples. White mark indicates that a given latent variable is active for a given patient sample.

5.2.3 Clinical Characteristics

Figure 5.27 summarizes the clinical characteristics of multiomic prostate cancer subtypes M-1, M-2 and M-3. As evident from the figure, tumours from the M-1 subtype tend to be more advanced (stage 3 or 4) and correspond to a higher Gleason grade (9 or 10). In turn, tumours from the M-3 subtype tend to be less advanced (stage 2, Gleason grade 5-7). Subtype M-2 appears to be more prevalent in older men (ages 60-89), which is further supported by the lower proportion of the early onset German cohort being classified as M-2. No major differences in PSA levels among the subtypes are apparent.

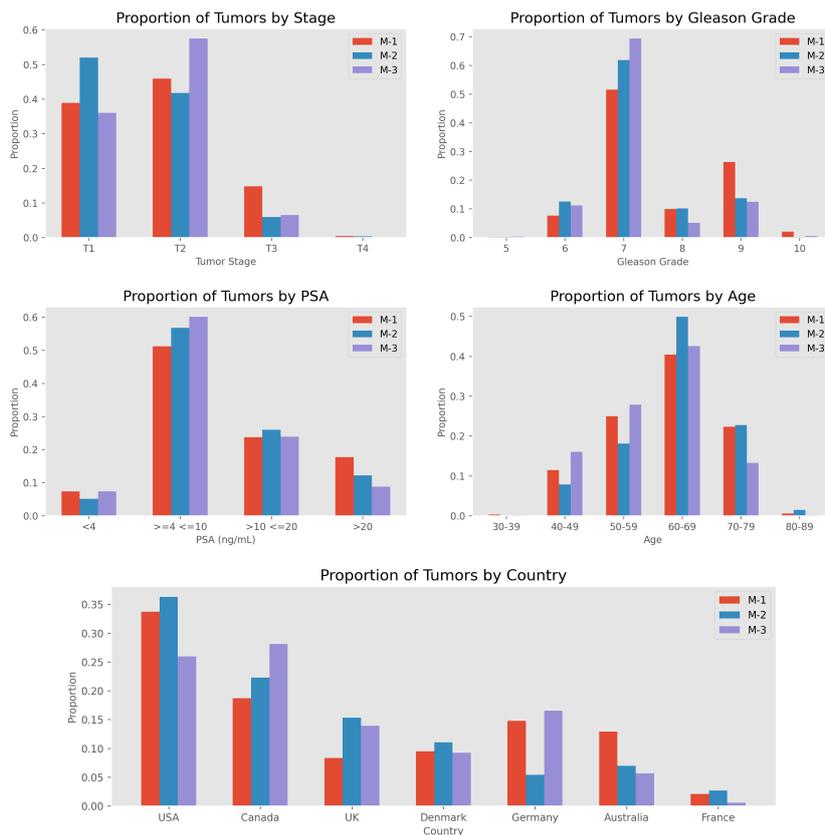


Figure 5.27: Clinical characteristics of the multiomic subtypes M-1, M-2 and M-3. Proportion of tumours at each tumour stage, Gleason grade, PSA, age and country (left to right, top to bottom).

5.2.4 Comparison with Single-Modality Subtypes

In this section, to illustrate both the similarities and the differences between single- and multi- omics subtyping, we compare the multiomic subtypes (M-1, M-2 and M-3) with the single-modality subtypes uncovered in Section 5.1. We limit this comparison to samples with modality-complete data only. Figures 5.28 - 5.31 visualise the comparisons via Sankey diagrams, for the DNA summary measurements, driver genes, CNAs and RNA data, respectively.

Multiple pairwise relationships become apparent when analysing Figures 5.28 - 5.31. For the summary measurement DNA subtypes, almost all samples from SM-2 (positive *ETS* status, frequent *ERG* gene fusions) were classified as subtype M-3 (positive *ETS* status, frequent *ERG* gene fusions, over-expression of *ERG*). In terms of the subtypes defined by driver gene mutations, subtype Dr-4 defined exclusively by the *SPOP* mutation was predominantly a subset of M-2 (frequent *SPOP* mutation). For the CNA-based subtypes, the majority of samples from CNA-2 (losses on chromosome 21) were classified as M-3 (losses on chromosome 21), while more than half of the samples from CNA-1 (losses on chromosome 6, high CNA burden) were assigned to M-2 (losses on chromosome 6). The strongest alignment between single- and multi- omics subtype assignments is visible for the RNA-based subtypes. Here, RNA-2 (under-expression of *ERG*) is almost entirely a subset of M-2 (*TFF3*-positivity, *ERG*-negativity) and RNA-3 (over-expression of *ERG*) of M-3 (*ERG*-positivity, *TFF3*-negativity), while RNA-1 is distributed across all three multiomic subtypes.

The similarities observed between all four subtyping schemas based on single-modality analysis and the multiomic schema indicate that the multiomic integration incorporated data from all modalities in the final classification. Additionally, the lack of a one-to-one correspondence between any two

schemas suggests that the model leveraged complementary information from each modality to determine the final assignments, which is further evidenced by the presence of fully multiomic latent variables (Figures 5.23 - 5.25).

Summary DNA - Based vs Multiomic Subtypes

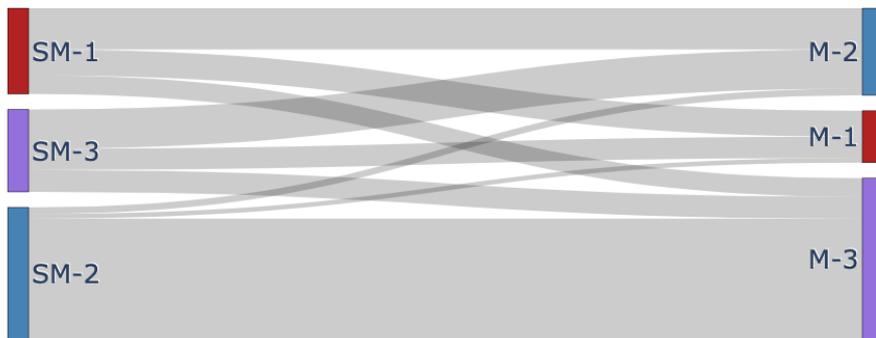


Figure 5.28: Sankey diagram visualising the comparison between the multiomic prostate cancer subtypes M-1, M-2 and M-3 and subtypes SM-1, SM-2 and SM-3 derived from the summary DNA measurements only.

Drivers - Based vs Multiomic Subtypes

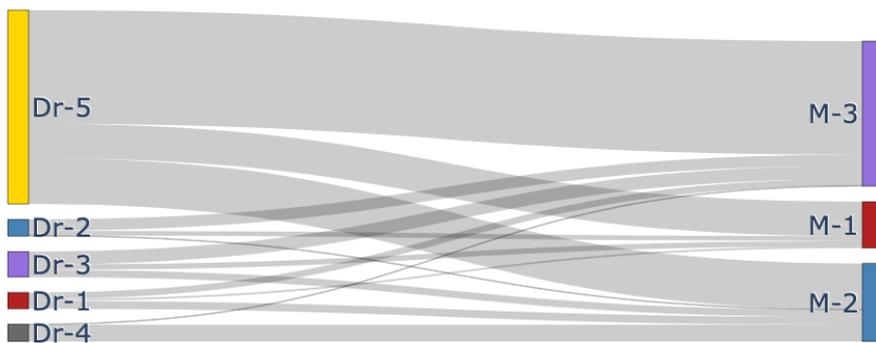


Figure 5.29: Sankey diagram visualising the comparison between the multiomic prostate cancer subtypes M-1, M-2 and M-3 and subtypes Dr-1, Dr-2, Dr-3, Dr-4 and Dr-5 derived from the driver gene data only.

CNA - Based vs Multiomic Subtypes

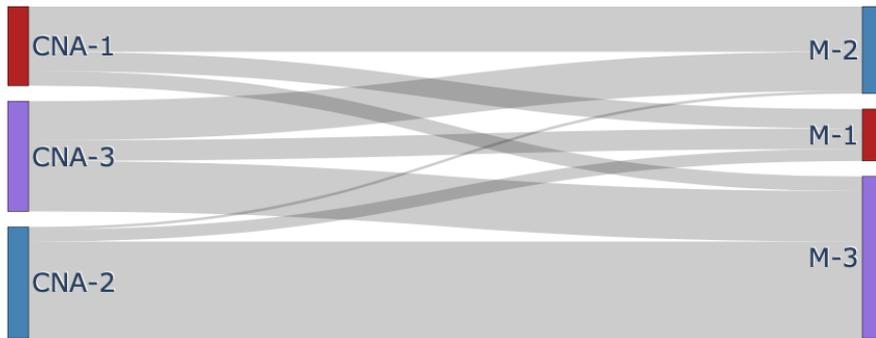


Figure 5.30: Sankey diagram visualising the comparison between the multiomic prostate cancer subtypes M-1, M-2 and M-3 and subtypes CNA-1, CNA-2 and CNA-3 derived from the CNA data only.

RNA - Based vs Multiomic Subtypes

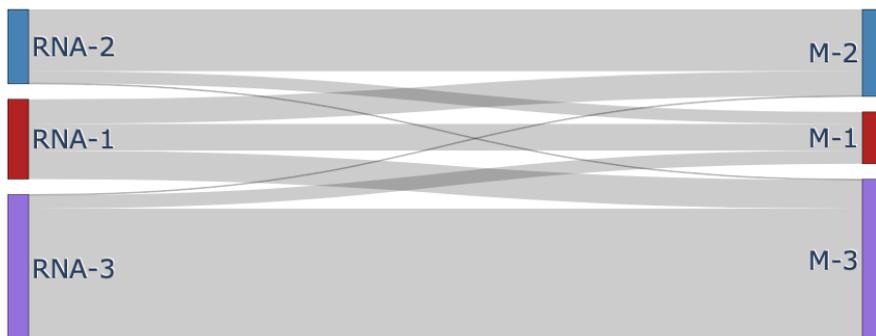


Figure 5.31: Sankey diagram visualising the comparison between the multiomic prostate cancer subtypes M-1, M-2 and M-3 and subtypes RNA-1, RNA-2 and RNA-3 derived from the RNA data only.

5.2.5 Predictive Tests

After identifying the three multiomic prostate cancer subtypes, we attempted to illustrate the potential of creating predictive subtyping tests, that would allow us to assign new patient samples to these subtypes directly from the data. To do this, we developed a predictive RFC test separately for each modality. For each test, we used a portion of the PPCG dataset where the relevant modality was available, along with the corresponding subtype assignments for the same samples. Predictive tests are most clinically useful when based on a single modality, as sequencing multiple modalities is expensive and often not feasible in practice.

An RFC predicted subtypes M-1, M-2, and M-3 with an overall test set accuracy of 0.83 and an F1 score of 0.72 for subtype M-1 when using all summary measurement features. However, using only driver genes data, the model's performance was significantly lower, achieving an overall test set accuracy of 0.59 and an F1 score of 0.11 for subtype M-1. For CNA data, the model yielded an overall accuracy of 0.73 and an F1 score of 0.49 for subtype M-1. Finally, an RFC model based solely on RNA data classified samples into subtypes M-1, M-2, and M-3 with an overall test accuracy of 0.91 and an F1 score of 0.81 for subtype M-1, using all 1000 RNA features.

Following from the above results, we applied a SHAP explainer to determine the 15 most important RNA features for predicting each subtype. This analysis resulted in a final panel of 24 genes. Using these 24 genes, we developed a new RFC predictor, which achieved the same 0.91 overall test accuracy and improved the F1 score for subtype M-1 to 0.83. This finding is clinically valuable as it allows us to accurately identify patients in the high-risk M-1 subtype based on just 24 RNA measurements.

Figure 5.32 visualises the importance of each gene from the 24-gene panel

for the prediction of each subtype. Two genes emerge as most important for the predictions: *ERG* and *TFF3*. Higher expression levels of *ERG* and lower expression levels of *TFF3* lead to an assignment to subtype M-3. The opposite is true for subtype M-2. Low expression levels of both *ERG* and *TFF3* characterize subtype M-1.

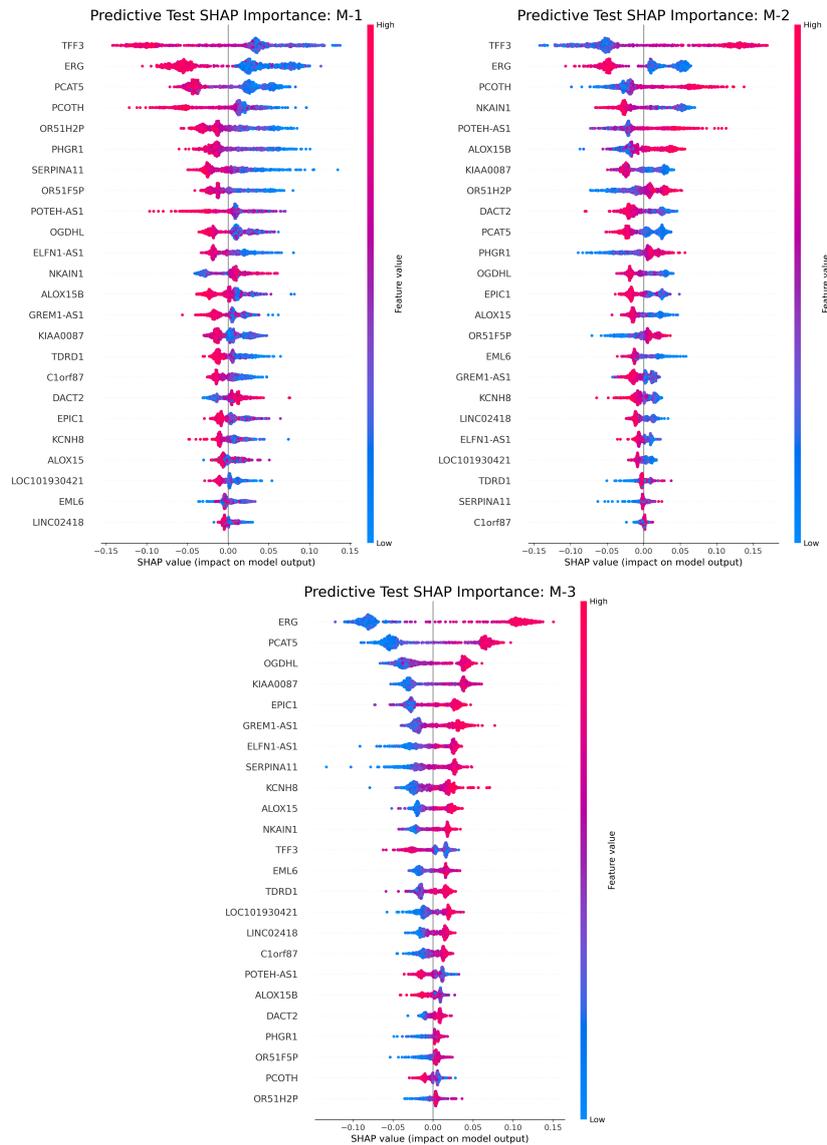


Figure 5.32: SHAP scores visualising the importance of each gene from the 24 RNA gene panel, for the prediction of subtypes M-1, M-2 and M-3.

Notably, the previously identified survival differences amongst subtypes M-1, M-2 and M-3 remain significant if labels obtained from the 24 gene RFC test are used as subtype assignments (Figure 5.33). Additionally, metastasis-free survival differences between subtypes M-1 and M-3 become significant.

Predictive Test Subtypes

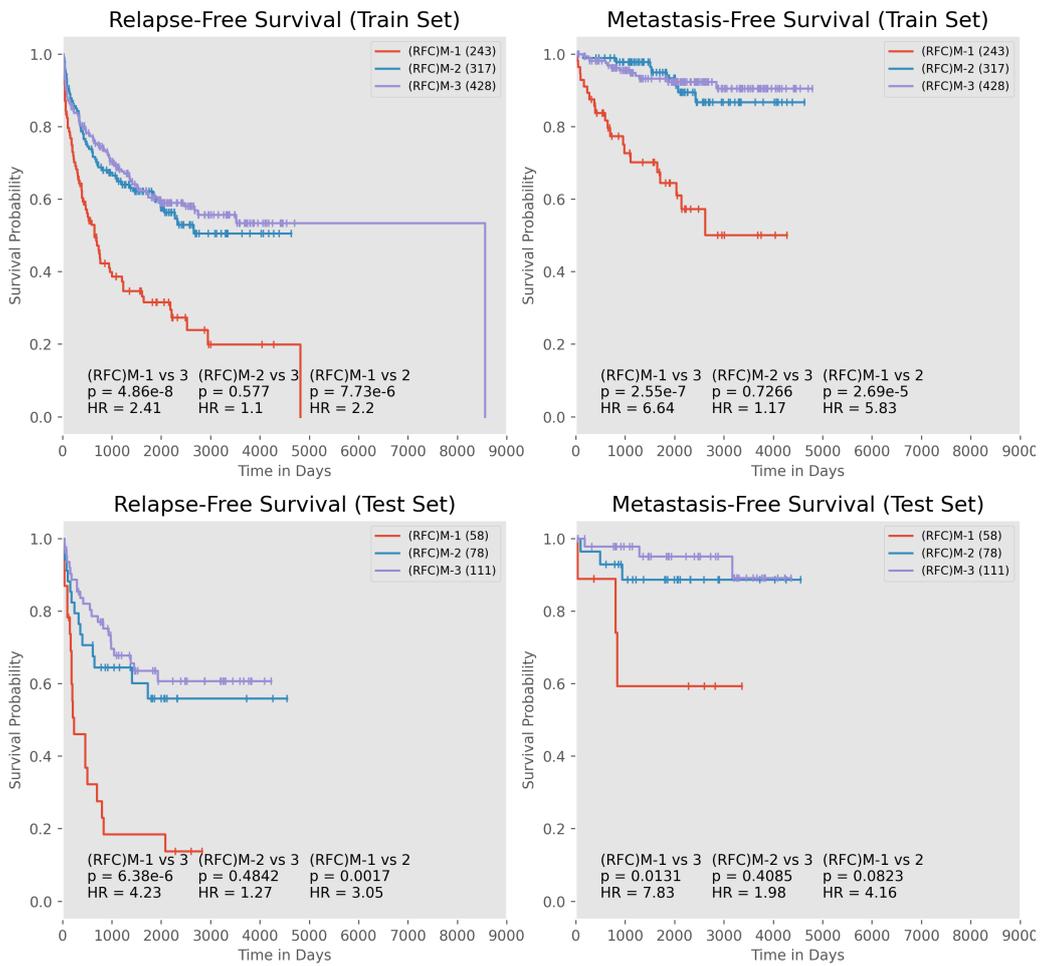


Figure 5.33: Kaplan-Meier plots visualising the relapse-free (left) and metastasis-free (right) survival for multiomic subtypes M-1, M-2 and M-3 as predicted with the 24 gene panel RFC test, separately for the training (top) and test set (bottom) samples. For each pairwise comparison, we provide the hazard ratio (HR) and the p value calculated with the Cox proportional hazard test. The number of samples in each subtype is given in brackets, next to the subtype label.

5.3 Comparison with Previously Established Subtypes

In this section, we compare our multiomic subtypes M-1, M-2 and M-3 with previously established prostate cancer subtyping schemas, namely the DESNT [125], You et al. [227] and Evotypes [213] classifications, reviewed earlier in Section 1.2.6. We have obtained PPCG sample assignments for these schemas from other PPCG members¹.

Figures 5.34 - 5.36 visualise the comparisons via Sankey diagrams. As evident, although more (33 out of 54) of the DESNT positive samples were classified as M-1, no majorly clear similarities between our subtypes and the DESNT [125], or You et al. [227] classifications are apparent. However, the Alternative evotype [213] appears to be almost entirely a subset of M-2 (104 out of 131 samples). We therefore explored this relationship further.

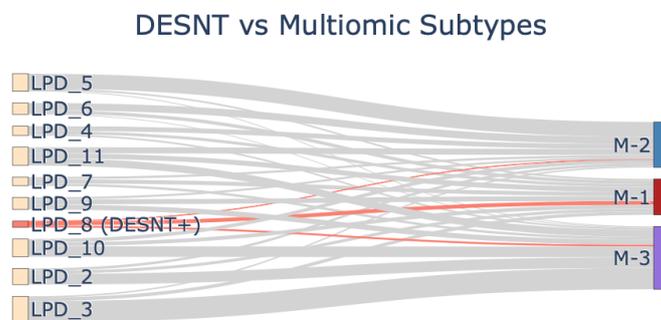


Figure 5.34: Sankey diagram visualising the comparison between the multiomic prostate cancer subtypes M-1, M-2 and M-3 and the DESNT [125] classification. The DESNT assignments were available for 1237 of the 1627 samples classified as multiomic subtypes. The LPDs correspond to the components of the Latent Process Decomposition procedure applied to identify the DESNT samples [125]. LPD_1 was omitted from this visualisation as no samples used in our analysis were assigned to this component.

¹Subtype assignments were obtained from: Sergio Llana Lago (DESNT), Valeriia Haberland (You et al.) and Emre Esenturk (Evotypes).

You et al. vs Multiomic Subtypes

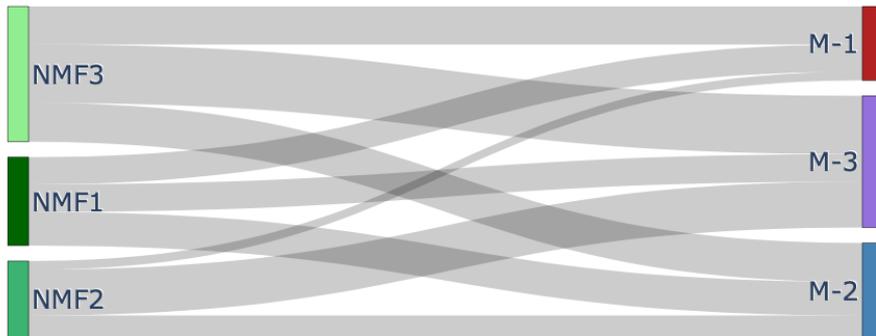


Figure 5.35: Sankey diagram visualising the comparison between the multiomic prostate cancer subtypes M-1, M-2 and M-3 and the You et al. [227] classification. The You et al. assignments were available for 1081 of the 1627 samples classified as multiomic subtypes.

Evotypes vs Multiomic Subtypes

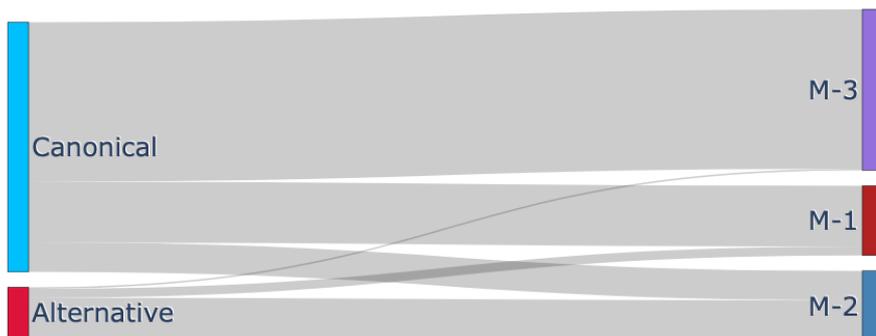


Figure 5.36: Sankey diagram visualising the comparison between the multiomic prostate cancer subtypes M-1, M-2 and M-3 and the Evotypes [213] classification. The Evotypes assignments were available for 776 of the 1627 samples classified as multiomic subtypes.

5.3.1 Evotypes and Multiomic Subtypes

The Sankey diagram in Figure 5.36 suggested a potential relationship between the Evotype [213] classification and the multiomic prostate cancer subtypes identified earlier. In Figure 5.37, we plotted the Evotype assignments in the latent space of iCS-GAN. In this UMAP, the Alternative evotype members are largely co-localized within multiomic subtype M-2.

Based on this result, we expanded our multiomic prostate cancer subtypes into four categories: M-1, M-3, M-2A (alternative subset of M-2), and M-2C (canonical subset of M-2). Since Evotypes assignments were not available for all samples in the PPCG dataset, we developed this new classification in the following way. For each of the five independent runs of our model, we selected the subset of the PPCG dataset that had complete data for all modalities and included Evotypes assignments. We then manually reassigned these samples to the new subtypes (M-1, M-2A, M-2C, and M-3) by comparing the original subtype labels from each run with the Evotypes assignments. Next, we trained a KNN classifier using the latent space encodings and the newly assigned labels. This classifier was then used to predict the new Multiomic-Evotypes classification labels for the remaining data. Finally, the majority vote from the five runs determined the final label for each sample.

In this joint classification schema, classes M-1, M-2C and M-3 can be viewed as subtypes of the Canonical evotype and class M-2A corresponds to the Alternative evotype. Figure 5.38 visualises the survival differences between the subtypes. Subtype M-1 ($n = 453$) remains the most aggressive, exhibiting worse patient prognosis than subtypes M-2C ($n = 140$) and M-3 ($n = 754$) in terms of relapse-free survival (M-1 vs M-2C: HR = 3.7, $p = 0.0318$; M-1 vs M-3: HR = 1.76, $p = 0.0231$). We found no significant survival differences between M-1 and M-2A ($n = 280$). Survival differences

between subtypes M-3, M-2A and M-2C were significant on the training set only, however, these relationships could not have been verified on the test set. Nevertheless, the Canonical (M-2C) and Alternative (M-2A) subsets of M-2 appear to have rather different survival outcomes (test set relapse-free HR = 2.75, p = 0.1179).

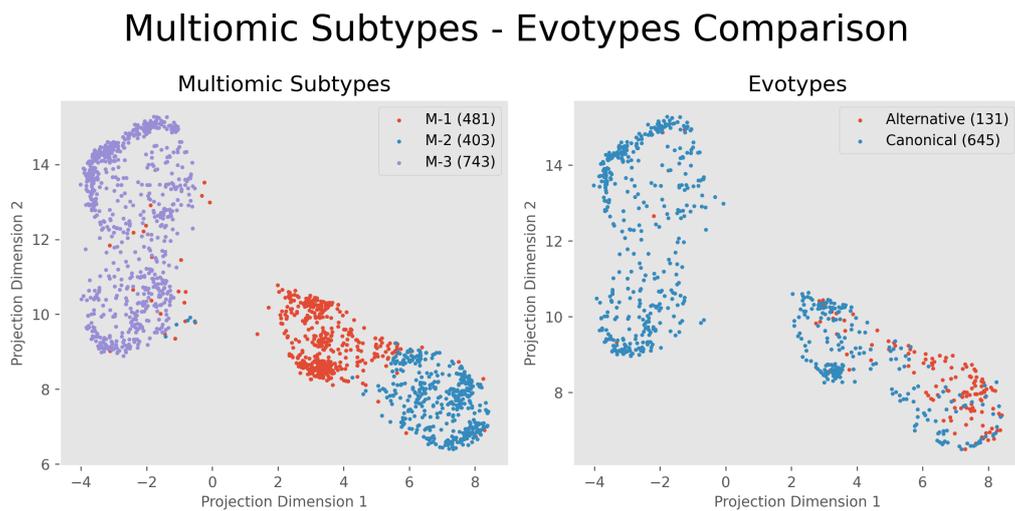


Figure 5.37: UMAP visualisation comparing the multiomic prostate cancer subtypes uncovered with iCS-GAN with the Evotypes [213] classification. Latent space visualisation for the multiomic subtypes (left). Samples with available Evotypes assignments visualised in the same latent space (right).

Multiomic Analysis (Evotypes) - Consensus Subtyping

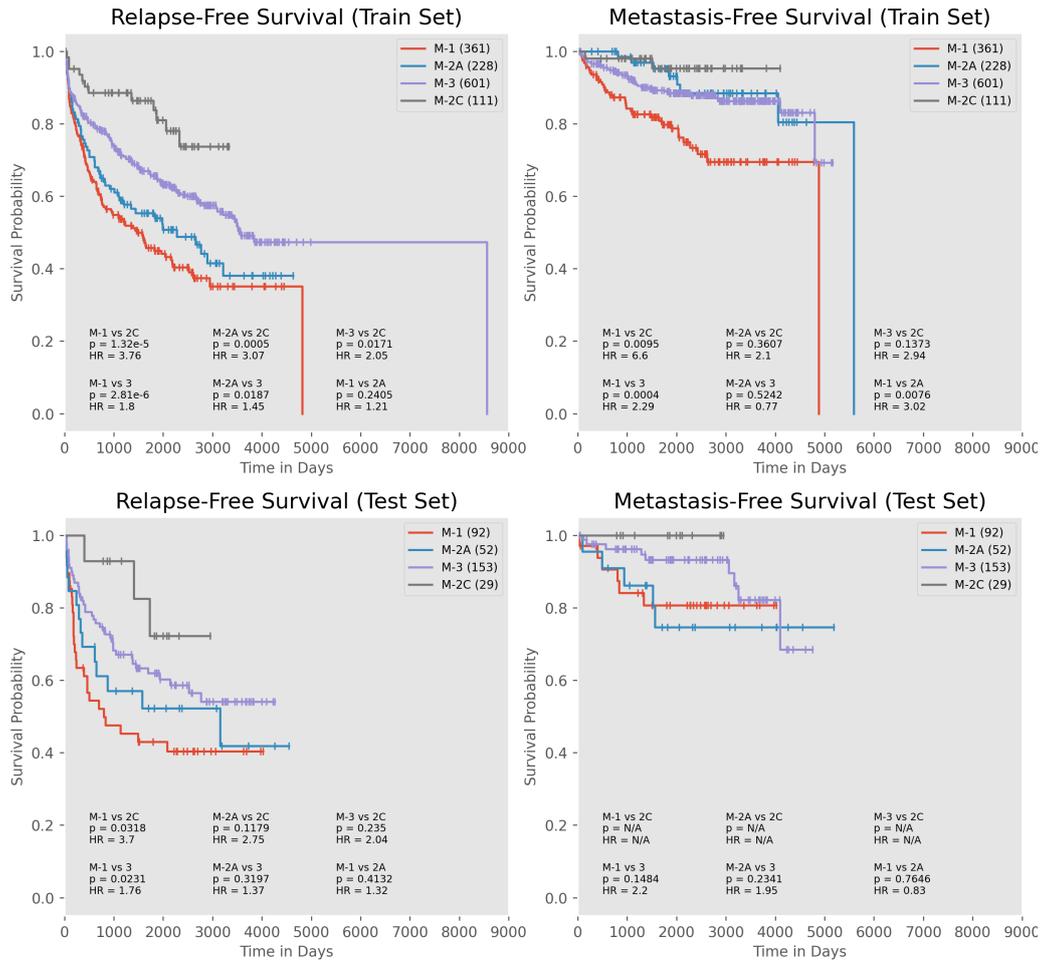


Figure 5.38: Kaplan-Meier plots visualising the relapse-free (left) and metastasis-free (right) survival for multiomic-evotypes subtypes M-1, M-2A, M-2C and M-3, separately for the training (top) and test set (bottom) samples. For each pairwise comparison, we provide the hazard ratio (HR) and the p value calculated with the Cox proportional hazard test. MFS test set comparisons with subtype 2C were not meaningful as there were no events in the group. The number of samples in each subtype is given in brackets, next to the subtype label.

Figures 5.39 - 5.42 characterize the combined subtypes, providing further insights into our initial classification schema. As before, subtype M-3 is characterized by positive *ETS* status, prevalent *ERG* gene fusions, highest inter to intra chromosomal breakpoint ratio, average, median and maximum number of chromosomes involved in breakpoint chains, as well as highest intra-chromosomal and inter-chromosomal events. Samples from M-2C exhibit higher numbers of SNVs and lowest median number of breakpoints, proportion of breakpoints in chains, maximum number of chromosomes involved in a chain and number of translocations. Compared to the other two groups, tumours from M-1 and M-2A have a higher percentage of genome altered by CNAs, as well as clonal CNAs, and higher numbers of duplications. Samples from M-2A are additionally characterized by higher numbers of deletions, inversions, rearrangements and complex indels, average and total numbers of breakpoints, as well as higher numbers of breakpoint chains, breakpoints in chains and breakpoints in longest chain, and more commonly exhibit chromothripsis.

SPOP mutations are present in a fraction of samples from M-2A, and absent in M-1, M-2C and M-3.

In terms of the copy number alterations, CNA mutational burden for tumours classified as M-2C is generally low. As before, samples from M-3 are more commonly affected by losses on chromosomes 3, 10, 17 and 21, however losses on chromosome 10 are also prevalent in M-1 and M-2A. Losses on chromosomes 2, 5, 6 and 13, gains on chromosomes 7 and 8 and deletions on chromosome 5 affect tumours classified as M-2A. Additionally, losses on chromosome 5 and gains on chromosomes 7 and 8 are prevalent in M-1.

No new major insights into the RNA profiles of the multiomic subtypes were uncovered by combining our results with the Evotypes classification,

with the Canonical and Alternative subsets of M-2 sharing the same RNA profile.

Multiomic Subtypes (Evotypes): Subtype Characteristics (Summary Measurements)

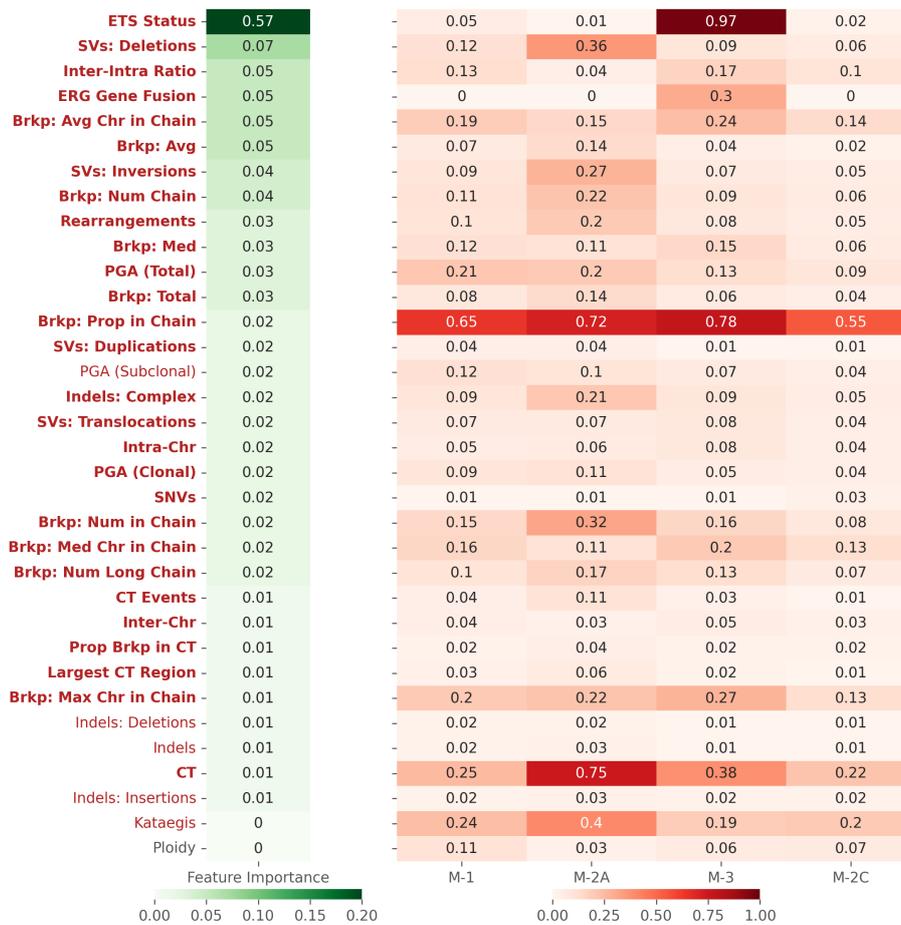


Figure 5.39: Summary measurements based characteristics of the multiomic-evotypes prostate cancer subtypes. Main heatmap visualises mean feature value in each subtype, as measured on the training set. Features are ordered by SHAP-based importance when predicting cluster assignments using RFC applied on the original data features. Aberrations with significant differences between subtypes are marked in bold red for features significant on both training and test sets or in light red for features significant on the training set only (FDR-adjusted $p < 0.05$, Chi-squared (categorical variables) or Kruskal-Wallis test (continuous variables)).

Multiomic Subtypes (Evotypes): Subtype Characteristics (Driver Genes)

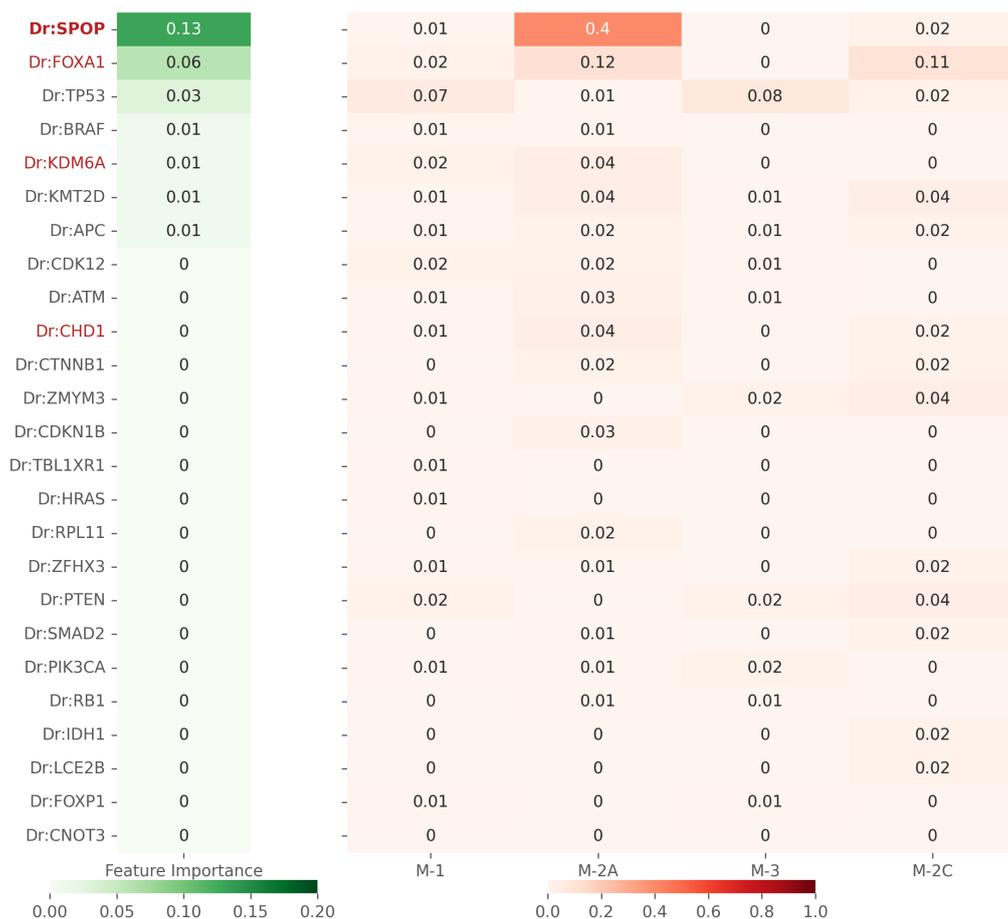


Figure 5.40: Driver genes based characteristics of the multiomic-evotypes prostate cancer subtypes. Main heatmap visualises mean feature value in each subtype, as measured on the training set. Features are ordered by SHAP-based importance when predicting cluster assignments using RFC applied on the original data features. Aberrations with significant differences between subtypes are marked in bold red for features significant on both training and test sets or in light red for features significant on the training set only (FDR-adjusted $p < 0.05$, Chi-squared test).

Multioomic Subtypes (Evotypes): Subtype Characteristics (CNAs)

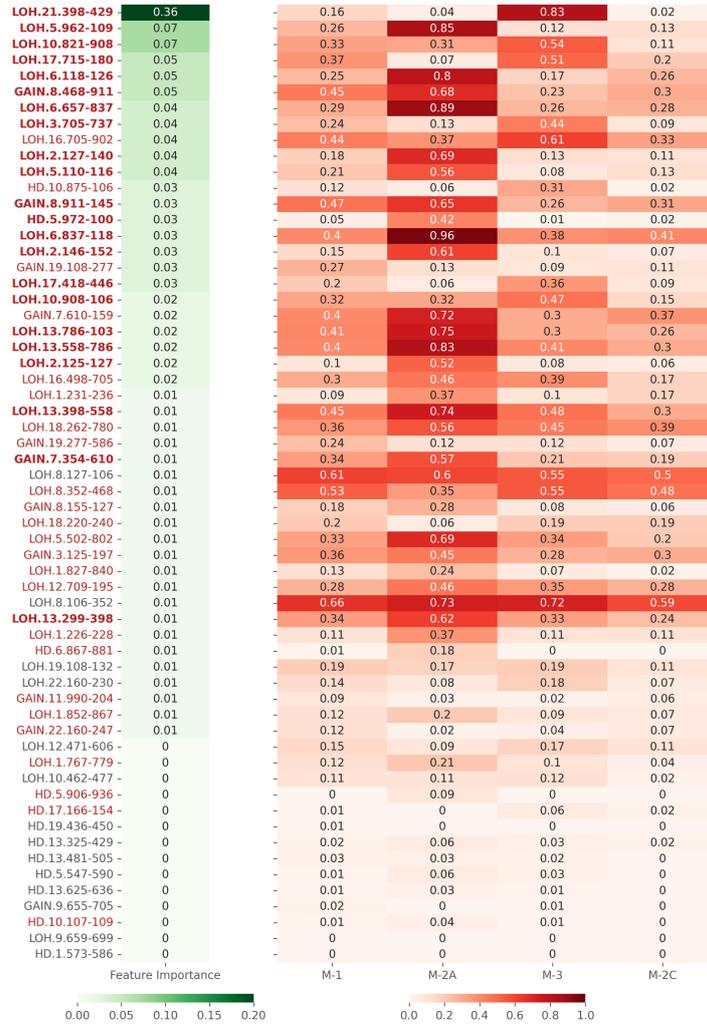


Figure 5.41: CNA based characteristics of the multiomic-evotypes prostate cancer subtypes. Main heatmap visualises mean feature value in each subtype, as measured on the training set. Features are ordered by SHAP-based importance when predicting cluster assignments using RFC applied on the original data features. Aberrations with significant differences between subtypes are marked in bold red for features significant on both training and test sets or in light red for features significant on the training set only (FDR-adjusted $p < 0.05$, Chi-squared test).

Multimic Subtypes (Evotypes): Subtype Characteristics (RNA)

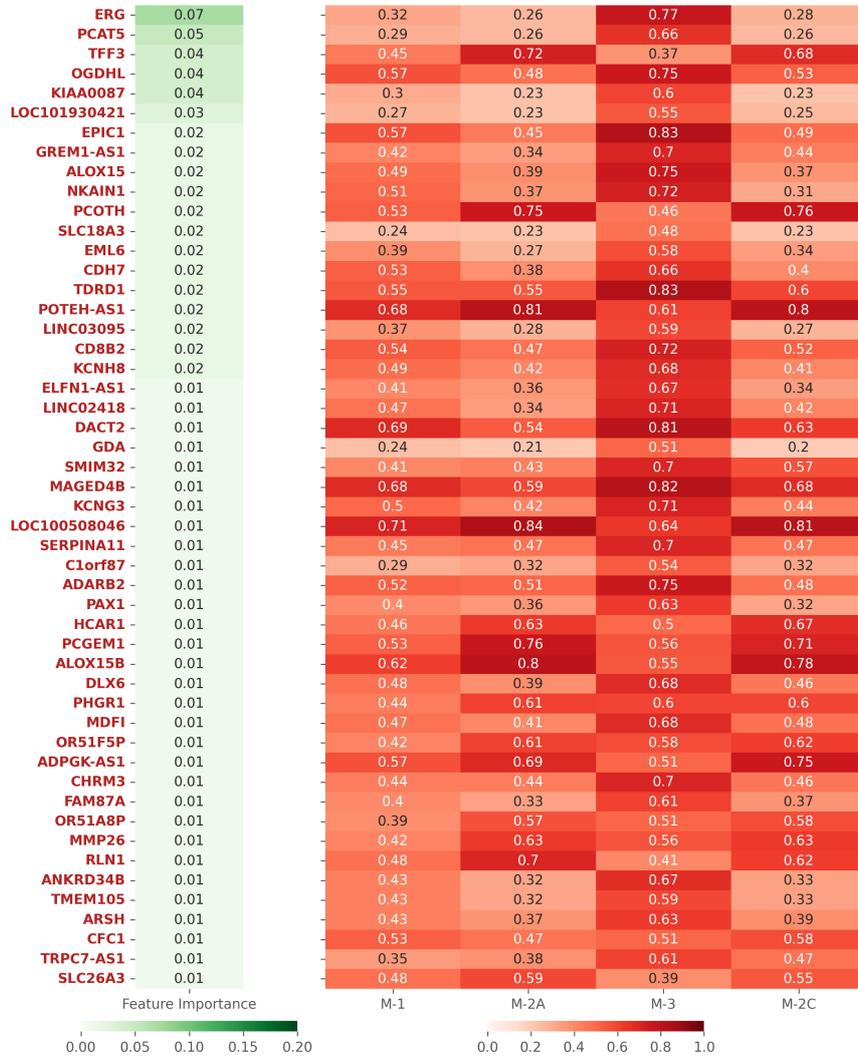


Figure 5.42: RNA based characteristics of the multimic-evotypes prostate cancer subtypes. Main heatmap visualises mean feature value in each subtype, as measured on the training set. Features are ordered by SHAP-based importance when predicting cluster assignments using RFC applied on the original data features. Aberrations with significant differences between subtypes are marked in bold red for features significant on both training and test sets or in light red for features significant on the training set only (FDR-adjusted $p < 0.05$, one-way anova test). For clarity, we limit the visualisation to 50 most important features.

5.3.2 Risk Stratification Utility Comparison

As our final experiment, we compared the patient risk stratification utility of the subtyping schemas considered so far. Figure 5.43 visualises the relapse-free Kaplan-Meier survival curves for the DESNT [125], You et al. [227] and Evotypes [213] classification frameworks, as well as for the multiomic subtypes uncovered in this DPhil, the multiomic subtypes as predicted with a 24-gene panel RFC, and the multiomic-evotypes subtypes. As evident from the figure, the 24-gene panel test resulted in best patient stratification (lowest p values).

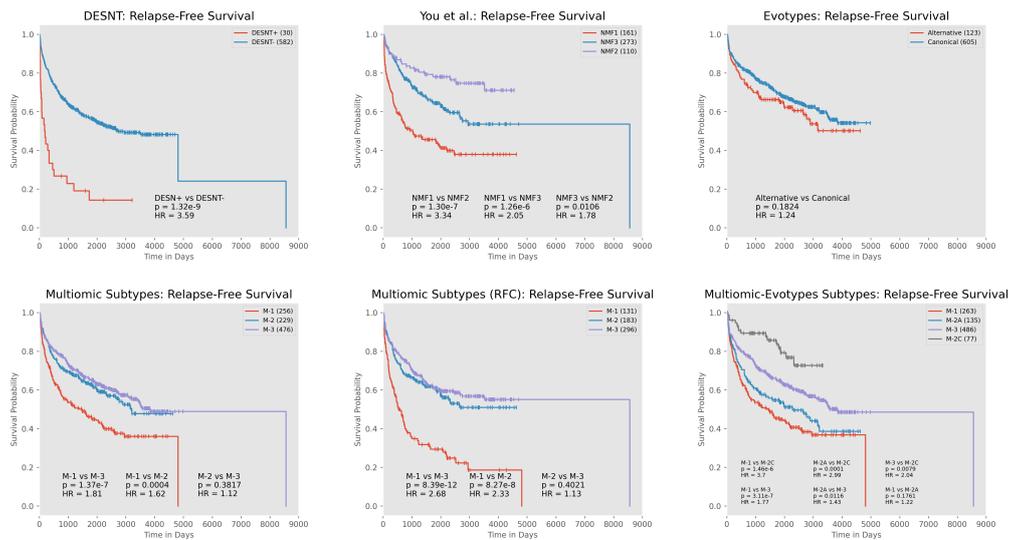


Figure 5.43: Kaplan-Meier plots visualising the relapse-free survival the DESNT [125], You et al. [227], Evotypes [213], multiomic subtypes, multiomic subtypes as obtained with the 24 gene panel predictive RFC test and multiomic-evotypes subtypes classification schemas (left to right, top to bottom). For each pairwise comparison, we provide the hazard ratio (HR) and the p value calculated with the Cox proportional hazard test. The number of samples in each subtype is given in brackets, next to the subtype label. Each plot was fitted for all the samples in the PPCG dataset for which the corresponding subtype assignments and survival variables were available.

Chapter 6

Discussion

In this thesis we developed a bespoke latent feature model for multiomic data integration and used it to perform subtyping on the largest molecular prostate cancer dataset to date. The main contributions of this DPhil are twofold:

1. We proposed, developed, and validated iCS-GAN, a fully interpretable, adversarially learned, latent feature model specifically designed for the integrative analysis of multiomic data with missing modalities. The model effectively extracts biologically relevant signatures in an unsupervised manner, without the need for prior domain knowledge, and guides the analysis toward subtypes that are clinically-relevant.
2. We applied iCS-GAN to conduct a comprehensive analysis of the Pan-Prostate Cancer Group prostate cancer dataset. This approach identified three distinct molecular subtypes of prostate cancer, one of which displays genetic features that have not been described previously and we therefore consider it as a novel subtype.

In Sections 6.1 and 6.2, we discuss these two contributions and their

significance in detail, before explaining the limitations of our work in Section 6.3 and providing suggestions for further research in Section 6.4. In Section 6.5, we discuss the ethical considerations surrounding this DPhil. Finally, in Section 6.6, we conclude this thesis with closing remarks.

6.1 Contributions to Methods for Multiomic Data Analysis

To the best of our knowledge, iCS-GAN is the first adversarially learned latent feature model designed for multiomic data analysis. We outline the key contributions that iCS-GAN brings to the field of multiomic data integration.

6.1.1 Adversarially Learned Inference Models are an Effective Tool for Multiomic Data Integration

In Chapter 4, we rigorously tested and validated iCS-GAN, demonstrating its effectiveness in multiomic data integration and subtyping analysis. Specifically, we showed that iCS-GAN can handle the diverse and skewed distributions typical of multiomic data and can accurately match its latent space encodings to various prior distributions, even beyond the standard normal or uniform priors. We showed that these two properties, in conjunction with a mixture of shared and independent layers in the generator, and the layer-wise pre-training strategy, effectively integrate multiomic data despite the heterogeneities present among different omics outputs, allowing for the extraction of fully integrative latent variables that encompass relationships both within and between modalities. Additionally, we confirmed the method’s full interpretability and assessed its performance when integrating data with miss-

ing modalities, showing that iCS-GAN can infer these missing components. Therefore, we have shown that iCS-GAN overcomes many of the challenges associated with multiomic data analysis. However, these improvements came with the typical computational expense associated with training GAN-based models. One mitigating factor that offsets this to some degree is the finding that iCS-GAN performs well using its default parameters, which significantly reduces the need for time-consuming hyperparameter searches.

These findings highlight the potential of adversarially learned latent feature models for effective multiomic data integration, positioning them as strong alternatives to commonly used statistical tools and Machine Learning methods such as Autoencoders and VAEs.

6.1.2 Survival Regularization Can Guide the Discovery of Clinically-Relevant Disease Subtypes

By incorporating the survival regularization component, we ensured iCS-GAN has the potential to guide the latent variables - and thus the resulting subtyping analysis - toward identifying subtypes with significant survival differences. This was demonstrated in our analysis in Section 4.7, where survival regularization enabled the discovery of subtypes with notable survival differences in the KIRC CNA dataset and provided greater granularity of the subtyping results for the KIRC multiomic analysis.

While latent feature models have previously been guided to produce encodings useful for downstream classification or clustering tasks [e.g. 72, 111, 217], to the best of our knowledge, applying this approach to survival outcomes is a novel contribution. Although our validation of the survival regularization component was limited, due to the quality of the available data, we believe this offers a strong foundation for future research.

6.1.3 Interpretable Models Extract Domain-Relevant Latent Variables

iCS-GAN, by incorporating interpretability constraints, extracts fully interpretable and biologically relevant latent variables in an unsupervised manner, without relying on prior domain knowledge. This is evidenced by the latent variables identified in Chapter 5, which correspond to known biological processes. For instance, the RNA LV5 corresponds to the immunoglobulin genes that are crucial to the formation of B-cell receptors, and the multiomic LV4 encompasses the *ETS* status, *TMPRSS2/ERG* gene fusions, *ERG* expression levels, and the LOH on chromosome 21 - which reflects intergenic region between the *TMPRSS2* and *ERG* genes that is deleted when the gene fusion occurs. The recovery of known interdependencies without prior biological information indicates that iCS-GAN could be valuable for exploratory analysis of biological signatures, with the potential to uncover previously unknown associations between genetic alterations by analyzing the extracted latent variables.

The addition of these interpretability constraints into iCS-GAN had only a minimal negative effect on the overall performance of the method, as demonstrated in Section 4.4. However, it significantly improved the stability of the outputs, showcasing that constraining adversarial models in such a way actually facilitates the analysis of complex, heterogeneous data.

The interpretability of iCS-GAN boosts confidence in its outputs and simplifies translation for clinical applications. This is particularly effective when the latent variables are viewed as straightforward binary indicators, signifying the presence or absence of groups of underlying genetic alterations - a feature enabled by the application of an IBP prior in iCS-GAN. This level of interpretability is invaluable in clinical settings, such as subtype tests,

where clinicians must fully understand the rationale behind any algorithmic recommendation before using it to make potentially life-changing decisions.

6.2 Contributions to Prostate Cancer Subtyping

By applying iCS-GAN to the PPCG dataset and analysing the results, we have made a number of observations related to prostate cancer subtypes and disease progression, both through single-source and multiomic analysis. These both support and unify previous observations, and enhance our understanding of how these are interlinked.

6.2.1 PGA, Kataegis, Chromothripsis and Ploidy Predict Negative Prostate Cancer Outcomes

The summary measurements subtyping analysis in Section 5.1.1 identified a distinct prostate cancer subtype characterized by higher prevalence of kataegis, chromothripsis, ploidy and DNA breakpoints, and a higher percentage of genome altered by CNAs (SM-1). We found SM-1 subtype membership to be a useful relapse predictor ($p = 0.0116$ as compared to the quiet SM-3 subtype; $p = 0.0151$ as compared to the *ETS* positive SM-2 subtype). This result unifies several previous observations, including that of Woodcock et al. [213] who associated kataegis, chromothripsis and DNA breakpoint burden with the more aggressive, in terms of time to biochemical recurrence, Alternative evotype, Lennartz et al. [115] who showed that the risk of PSA recurrence increases with ploidy status and Hieronymus et al. [81] who associated the percentage of genome affected by CNAs with prostate cancer

biochemical recurrence and metastasis.

6.2.2 ETS Status and ERG Fusions Are Not Indicators of Prostate Cancer Patient Prognosis

The clinical significance of the commonly occurring genetic events of the *ETS* and *TMPRSS2 / ERG* gene fusions in prostate cancer has been widely investigated, but with conflicting results [51, 66, 142, 154, 202]. For instance, Esgueva et al. [51] found no link between *TMPRSS2 / ERG* fusions and either pathological features or clinical outcomes. Similarly, Gopalan et al. [66] reported that the presence of these rearrangements does not predict prostate cancer aggressiveness. In contrast, Nam et al. [142] associated the fusions with higher propensity for biochemical recurrence, and Perner et al. [154] linked them to risk factors for disease progression.

Our findings, based on single-modality summary measurements and the largely *ETS* positive multiomic subtype M-3, support the findings of Esgueva et al. [51] and Gopalan et al. [66] and thus contradicts those of Nam et al. [142] and Perner et al. [154]. Specifically, the *ETS* positive subtype SM-2 (Section 5.1.1) with prevalent *ERG* fusions did not show significant differences in relapse-free or metastasis-free survival when compared to the subtype SM-3 ($p = 0.7463$ and $p = 0.3656$). This suggests that while the *ETS* and *TMPRSS2 / ERG* gene fusions define a distinct molecular subtype in prostate cancer, this subtype is not a clinically-useful prognostic marker. Further evidence of this comes from our multiomic subtyping analysis (Section 5.2). In this case, the *ERG* over-expressed, *ETS* positive with frequent *ERG* fusions subtype M-3 showed no survival differences compared to the *TFF3* over-expressed, *ETS* negative subtype M-2 ($p = 0.9171$ and $p = 0.8354$).

6.2.3 CNA Burden Predicts Prostate Cancer Relapse

The CNA subtyping analysis in Section 5.1.3 identified three distinct prostate cancer subtypes: “quiet” subtype CNA-3, subtype CNA-1 characterized by more prevalent losses on chromosomes 1, 2, 5, 6 and 13, gains on chromosomes 3, 7, 8, and deletions on chromosomes 5 and 15, and subtype CNA-2 with a higher propensity of losses on chromosomes 3, 8, 10, 16, 17, 18 and 21, and deletions on chromosome 10. Compared to the low CNA burden subtype CNA-3, membership in CNA-1 and CNA-2 was predictive of worse, in terms of relapse-free survival, patient outcomes ($p = 0.0197$ and $p = 0.0193$). These findings are in line with the works of Hieronymus et al. [81] who showed an association of prostate cancer CNA burden with biochemical recurrence, and Salachan et al. [168] who found higher CNA burden to be significantly associated with BCR, with an emphasis on strong associations between losses on chromosomes 13q (CNA-1) and 16q (CNA-2) and gains on 8q (CNA-1), and BCR-free relapse. Furthermore, Wang et al. [206] showed the association of gains on chromosomes 3 and 8 with prostate cancer metastasis. These gains characterize subtype CNA-1, which we found to be a significant predictor of metastatic prostate cancer as compared to CNA-3 ($p = 0.0273$).

6.2.4 Multiomic Analysis Reveals 3 Distinct Prostate Cancer Subtypes

Our multiomic prostate cancer analysis in Section 5.2 identified three distinct prostate cancer subtypes. For two of these subtypes, clear molecular characteristics emerged: M-3 was characterized by a positive *ETS* status and common *ERG* gene fusions, more prevalent losses on chromosomes 3, 10, 17 and 21, and deletions on chromosome 10, over-expression of *ERG*

and *PCAT5* and under-expression of *TFF3* and *PCOTH* while subtype M-2 was characterized by a higher likelihood of *SPOP* mutations, more prevalent losses on chromosomes 2, 5 and 6, gains on chromosome 8, and deletions on chromosome 5, over-expression of *TFF3* and *PCOTH* and under-expression of *ERG* and *PCAT5*. To briefly summarize, this makes M-3 an *ERG* positive, *TFF3* negative subtype and M-2 a *TFF3* positive, *ERG* negative subtype. This exclusivity of *ERG* and *TFF3* is broadly consistent with previous observations, for example those of Terry et al. [187] who suggested that *ERG* and *TFF3* characterize two distinct subsets of prostate cancer, and who also did not observe any prognostic significance of these subtypes ($p = 0.9171$ for relapse-free survival between M-2 and M-3).

While M-3 and M-2 do show some cluster-defining features, the characteristics of the aggressive subtype M-1 was less clear, with tumours classified as M-1 sharing some similarities with both subtypes, but with enough differentiation not to be categorised alongside them. What can be used to distinguish M-1 was the downregulation of genes and absence of genetic alterations that defined the other two subtypes, defining it by omitting it from either of the other two categories. Therefore, the M-1 subtype of prostate cancer could be considered as a ‘double-negative’ prostate cancer, analogous to how a ‘triple-negative’ breast cancer is defined by the lack of expression of the three receptors that can be used to define other breast cancer subtypes: estrogen receptor, progesterone receptor, or human epidermal growth factor receptor 2 (HER2). Note that ‘double-negative’ prostate cancer is sometimes used to describe the disease state after developing resistance to androgen deprivation therapy, so we do not label the M-1 subtype as such to avoid confusion.

To the best of our knowledge, the M-1 subtype has not been described

in the literature before, possibly due to the conceptual difficulties in defining something by what it is not. It could be argued that these tumours could represent ‘less advanced’ forms of the disease, perhaps ones that have not yet developed a sufficient number of genetic alterations to be defined as one of the other two subtypes, but the association with poor survival (Figure 5.18), high Gleason grade and tumour stage (Figure 5.27) at least partially counteract that argument. More work is therefore necessary to identify the biological and clinical underpinnings of this subtype.

6.2.5 Prostate Cancer Subtypes Can Be Accurately Predicted from 24 RNA Expressions

In Section 5.2.5, we developed a highly accurate predictive test capable of classifying patients into subtypes M-1, M-2, and M-3 using only 24 RNA gene expressions. Such streamlined approach could significantly reduce the complexity and cost of the gene expression analysis necessary for disease subtyping. The potential clinical utility of this test is substantial, as it could pave the way for more accessible and cost-effective stratification tools for prostate cancer. By enabling accurate risk stratification at the point of diagnosis, such tests could guide personalized treatment decisions, thus improving patient outcomes while reducing unnecessary interventions and healthcare costs.

6.2.6 Multiomic Analysis Provides Additional Insights into Prostate Cancer Evotypes

The Evotypes comparison analysis in Section 5.3.1 expanded our understanding of the multiomic subtypes M-1, M-2, and M-3 and offered new insights

into the established Evotypes [213] classification system. In particular, by refining the multiomic subtypes as M-1, M-2A, M-2C, and M-3, where M-2A corresponds to the Alternative evotype, M-1, M-2C, and M-3 therefore represent distinct subsets of the Canonical evotype, we provided more granularity to the Evotype classification which does not divide the Canonical subtype further.

Comparing subtypes 2-A and 2-C reveals different genomic profiles (Figures 5.39, 5.40 and 5.41), but almost identical transcriptomic profiles (Figure 5.42). As the Alternative evotype is defined by tumours that share the same evolutionary trajectory [213], and our method has identified M-2A and M-2C as similar through their shared transcriptomic profiles, it therefore follows that these subtypes belong to the same (Alternative) evolutionary trajectory. We hypothesise that M-2C corresponds to tumours at an earlier stage of that evolutionary trajectory than tumours in M-2A, which explains why they were undetectable from the used in the Evotypes [213] study genomic data alone. This hypothesis has some profound ramifications, not least that we could possibly *infer the evolutionary fate* of the disease far earlier than previously thought. Furthermore, the original Evotype model describes how evolutionary divergence is potentially driven by acquired genetic alterations, with some evidence that this is the case with *CHD1* loss [213], but this hypothesis in conjunction with similar RNA profiles implies that the transcriptional program is fixed before the genetic alterations that contribute to the evolutionary divergence are acquired. This could indicate that there are other unobserved factors in play, for instance methylation events, or that some tumours are always destined to become the Alternative evotype and so follow a completely separate evolutionary path rather than a divergent one.

Unfortunately, testing a hypothesis on the same data used to generate the

hypothesis is not statistically valid and so investigating this further remains outside the scope of this thesis.

6.3 Limitations

Despite the contributions and strengths of this research, certain limitations must be considered when interpreting our results. First of all, although our analysis was performed on the largest prostate cancer dataset compiled to date, the countries involved in the PPCG consortium (USA, Canada, UK, Denmark, Germany, Australia and France) are predominantly populated by men of White ethnicity, and so other ethnic groups are notably under-represented in the data. This is a significant weakness considering that, for instance, Black men are at twice the risk of getting prostate cancer as are White men, and experience worse outcomes of the disease [122]. Relating our subtypes to members of non-White ethnic groups is therefore of paramount importance.

Secondly, when characterizing our multiomic subtypes, we observed that they are predominantly associated with transcriptional programs, with some genomic information (such as that identified in the CNA-only subtyping analysis) missing from the multiomic subtypes. This suggests that some information imbalance may still persist, perhaps due to the larger number of features in the RNA dataset, which might have contributed more heavily to the adversarial learning process. However, the fact that we did not simply derive the same RNA and multiomic subtypes indicates that genomic information was incorporated to some extent. Moving forward, we should explore ways to balance or weight the inputs to the multimodal layer, ensuring that as much of the individual-modality information as possible is preserved in the

multiomic subtyping.

Finally, we observed that selecting the number of clusters based on a stability-related heuristic still leads to limited granularity in the resulting subtypes. To address this, we should consider refining the proposed method or the approach for selecting the number of clusters so that it enables the discovery of more granular and potentially more coherent subtypes.

6.4 Future Work

We have identified several avenues for extending the work presented in this thesis, beyond those previously mentioned in the discussion of its limitations.

First of all, incorporating methylation data could deepen our understanding of the molecular prostate cancer subtypes characterized in this research. We intend to pursue this as soon as the methylation data passes quality control and becomes available for analysis.

Second, we aim to extend the validation of the survival regularization technique proposed in this thesis by rigorously testing it on other cancer datasets with stronger survival predictive utility than that observed in TCGA. Additionally, we plan to explore robust approaches for applying survival regularization directly to the embedding, rather than reconstruction, layer of iCS-GAN.

Next, we would like to expand our comparison of the molecular subtypes M-1, M-2 and M-3, with other prostate cancer classification schemas beyond DESNT [125], You et al. [227] and Evotypes [213]. Members of the PPCG consortium are currently working on replicating various studies, including e.g. those of Luca et al. [126], Fraser et al. [57] or Taylor et al. [182], using the PPCG dataset. We anticipate that comparing our results with these

frameworks will deepen our understanding of the subtypes identified in this thesis.

Finally, having observed the benefits of combining classification schemas through our comparison with the Evotypes [213] framework, we would like to introduce a ‘consensus classification loss’ into iCS-GAN, to incorporate prior knowledge about known disease subtypes into the latent feature extraction process. This approach would be akin to how Autoencoders are regularized for downstream classification tasks [111], with the difference that we would use multiple classifiers, with each one trained to predict labels for an independent subtyping schema. We hypothesize that if multiple, potentially contradicting, subtyping schemas can be accurately predicted from the latent space encodings of iCS-GAN, the subsequent subtyping analysis could be directed towards uncovering consensus subtypes that represent a high-degree of agreement among different schemas.

6.5 Ethical Considerations

This section summarizes the ethical considerations related to this DPhil, the PPCG dataset, and general ML research in prostate cancer. Specifically, we discuss the key issues including patient consent, privacy and data security, training data bias, environmental impact of generative models, transparency and explainability of ML models, and concerns about equal access to any clinical tests that may be developed as a result of this research.

6.5.1 PPCG Dataset: Consent, Privacy and Security

All patients in the PPCG dataset provided consent according to International Cancer Genome Consortium (ICGC) standards. Based on this, we

can reasonably assume that valid consent was obtained. Adherence with ICGC guidelines guarantees that patients were informed that participation in sample collection was voluntary, made aware of withdrawal procedures, the nature of research using their samples, and potential privacy risks related to re-identification [87].

The remote risk of being re-identified from genomic data is particularly important in the context of whole genome sequencing (WGS) data, with research showing that patients can be uniquely identified from just 30 to 80 statistically independent single nucleotide polymorphisms (SNPs) [118]. While researchers adhering to ICGC standards are required to attest that, unless otherwise agreed in consent, they will not attempt to re-identify the participants of the study [87], significant threats still exist should data leaks occur. The PPCG consortium has therefore implemented strong data security measures, including encryption keys and access controls. Additionally, the summary measurements generated to describe the phenomena observed in the WGS, used in place of the raw WGS data, largely reduce the possibility of patient re-identification, thus ensuring a high degree of data security.

6.5.2 PPCG Dataset: Sample Bias

The PPCG dataset mainly includes samples from White men, with Black men being significantly under-represented. As previously outlined, this is a major limitation of this and many other studies, given that Black men are twice as likely to be diagnosed with prostate cancer compared to White men (1 in 4 vs. 1 in 8) and have a higher risk of dying from it (1 in 12 vs. 1 in 24) [122]. This bias in training data could worsen the existing disparities in prostate cancer diagnosis and care, and reduce the potential to effectively help the group most affected by this disease. However, significant

effort is being made within the PPCG consortium to address this limitation, with numerous studies focusing on prostate cancer presentation in Black men specifically [e.g. 78, 79, 95, 96, 190].

6.5.3 Environmental Impact of GANs

The growing carbon footprint of machine learning and artificial intelligence raises concerns about their environmental impact and resource use [e.g. 17, 145, 197]. While we believe that the potential of ML to enhance cancer care outweighs the relatively minor environmental costs associated with smaller architectures like iCS-GAN - especially when compared to, for example, large language models [151] - steps were taken during this DPhil to minimize any negative effects. Given the computational intensity of GAN-based models, we limited hyperparameter searches to the necessary minimum, reducing training time. Additionally, our use of shallow networks - although primarily for interpretability - also minimized the computational burden, further reducing the environmental impact of this study.

6.5.4 ML Models: Transparency and Explainability

Machine learning models, and deep learning models in particular, are often seen as ‘black boxes’ because their decision-making processes are difficult to explain [76]. However, interpretability is crucial for clinical use. Clinicians need to fully understand the reasoning behind any algorithmic recommendations before applying them to make potentially life-altering decisions for patients. Clear and transparent models are therefore necessary to ensure that healthcare professionals can trust and rely on these systems in clinical practice. Throughout the development of iCS-GAN, we have ensured that the model and its outputs remain fully interpretable. By prioritizing inter-

pretability, we aimed to build trust in the model’s predictions and ensure its potential for trustworthy use in clinical settings.

6.5.5 Commercialization vs Improving Patient Care

Finally, we must consider the potential impact of any clinical tests that could be developed based on our research. While such tests could significantly enhance patient care and diagnosis, there is a risk that companies may prioritize commercial profits over patient well-being. Although this is a broader concern within the healthcare system [e.g. 41, 74, 155], it is essential to emphasize that improving patient outcomes and ensuring equal access to care should always be a top priority.

6.6 Final Remarks

In conclusion, this thesis presents a robust, customized, and fully interpretable alternative to existing methods for multiomic data analysis. To the best of our knowledge, iCS-GAN is the first adversarially learned latent feature model developed for this application. By applying iCS-GAN to the PPCG dataset, we have demonstrated its potential and utility in integrative cancer subtyping, establishing strong baselines for future research. Furthermore, we have thoroughly characterized three distinct molecular subtypes of prostate cancer, including an aggressive, an to the best of our knowledge novel, ‘double negative’ subtype, thus opening avenues for further clinical research and exploration.

References

- [1] Abdi, H., Williams, L.J., Valentin, D. “Multiple factor analysis: principal component analysis for multitable and multiblock data sets”. In: *WIREs Computational Statistics* 5.2 (2013), pp. 149–179.
- [2] Adamo, P., Lodomery, M.R. “The oncogene ERG: a key factor in prostate cancer”. In: *Oncogene* 35 (2016), pp. 403–414.
- [3] Ahmed, K.T., Sun, J., Cheng, S. et al. “Multi-omics data integration by generative adversarial network”. In: *Bioinformatics* 38.1 (2022), pp. 179–186.
- [4] Akhmedov, M., Arribas, A., Montemanni, R. et al. “OmicsNet: Integration of Multi-Omics Data using Path Analysis in Multilayer Networks”. 2017. bioRxiv: 238766.
- [5] Alexandrov, L.B., Kim, J., Haradhvala, N.J. et al. “The repertoire of mutational signatures in human cancer”. In: *Nature* 578 (2020), pp. 94–101.
- [6] Allen, D.M. “Mean Square Error of Prediction as a Criterion for Selecting Variables”. In: *Technometrics* 13.3 (1971), pp. 469–475.
- [7] Argelaguet, R., Arnol, D., Bredikhin, D. et al. “MOFA+: a statistical framework for comprehensive integration of multi-modal single-cell data”. In: *Genome Biology* 21 (2020).
- [8] Argelaguet, R., Velten, B., Arnol, D. et al. “Multi-Omics Factor Analysis—a framework for unsupervised integration of multi-omics data sets”. In: *Molecular Systems Biology* 14.6 (2018).

- [9] Arjovsky, M., Chintala, S., Bottou, L. “Wasserstein Generative Adversarial Networks”. In: *Proceedings of the 34th International Conference on Machine Learning*. Vol. 70. 2017, pp. 214–223.
- [10] Arora, K., Barbieri, C.E. “Molecular Subtypes of Prostate Cancer”. In: *Current Oncology Reports* 20.8 (2018), p. 58.
- [11] Asadi, M., Ahmadi, N., Ahmadvand, S. et al. “Investigation of olfactory receptor family 51 subfamily j member 1 (OR51J1) gene susceptibility as a potential breast cancer-associated biomarker”. In: *PLoS One* 16.2 (2021).
- [12] Ayinde, B.O., Zurada, J.M. “Deep Learning of Constrained Autoencoders for Enhanced Understanding of Data”. In: *IEEE Transactions on Neural Networks and Learning Systems* 29.9 (2018), pp. 3969–3979.
- [13] Baca, S.C., Prandi, D., Lawrence, M.S. et al. “Punctuated Evolution of Prostate Cancer Genomes”. In: *Cell* 153.3 (2013), pp. 666–677.
- [14] Barbieri, C.E., Baca, S.C., Lawrence, M.S. et al. “Exome sequencing identifies recurrent SPOP, FOXA1 and MED12 mutations in prostate cancer”. In: *Nature Genetics* 44.6 (2012), pp. 685–689.
- [15] Bibikova, M., Barnes, B., Tsan, C. et al. “High density DNA methylation array with single CpG site resolution”. In: *Genomics* 98.4 (2011), pp. 288–295.
- [16] Bishop, C.M., Tipping, M.E. “A hierarchical latent variable model for data visualization”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 20.3 (1998), pp. 281–293.
- [17] Bolón-Canedo, V., Morán-Fernández, L., Cancela, B. et al. “A review of green artificial intelligence: Towards a more sustainable future”. In: *Neurocomputing* 599 (2024).
- [18] Bramer, L.M., Irvahn, J., Piehowski, P.D. et al. “A Review of Imputation Strategies for Isobaric Labeling-Based Shotgun Proteomics”. In: *Journal of Proteome Research* 20.1 (2021), pp. 1–13.

- [19] Burstein, M.D., Tsimelzon, A., Poage, G.M. et al. “Comprehensive Genomic Analysis Identifies Novel Subtypes and Targets of Triple-Negative Breast Cancer”. In: *Clinical Cancer Research* 21.7 (2015), pp. 1688–1698.
- [20] Cancer Research UK. Worldwide cancer statistics [Online]. Accessed: March 2024. URL: <https://www.cancerresearchuk.org/health-professional/cancer-statistics/worldwide-cancer>.
- [21] Cancer Research UK. Cancer Statistics for the UK [Online]. Accessed: March 2024. URL: <https://www.cancerresearchuk.org/health-professional/cancer-statistics-for-the-uk>.
- [22] Cancer Research UK. Prostate cancer statistics [Online]. Accessed: March 2024. URL: <https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/prostate-cancer>.
- [23] Cao, Y., Ding, G.W., Lui, K.Y-C. et al. “Improving GAN Training via Binarized Representation Entropy (BRE) Regularization”. 2018. arXiv: 1805.03644.
- [24] Caroline, R., Grob, J.J., Stroyakovskiy, D. et al. “Five-Year Outcomes with Dabrafenib plus Trametinib in Metastatic Melanoma”. In: *New England Journal of Medicine* 381.7 (2019), pp. 626–636.
- [25] Cerami, E., Gao, J., Dogrusoz, U. et al. “The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data”. In: *Cancer Discovery* 2.5 (2012), pp. 401–404.
- [26] Chai, H., Shi, X., Zhang, Q. et al. “Analysis of cancer gene expression data with an assisted robust marker identification approach”. In: *Genetic Epidemiology* 41.8 (2017), pp. 779–789.
- [27] Chalise, P., Fridley, B.L. “Integrative clustering of multi-level ‘omic data based on non-negative matrix factorization algorithm”. In: *PLoS One* 12.5 (2017).
- [28] Chaudhari, P., Agrawal, H., Kotecha, K. “Data augmentation using MG-GAN for improved cancer classification on gene expression data”. In: *Soft Computing* 24 (2020), pp. 11381–11391.

- [29] Chaudhary K., Poirion O.B., Lu L. et al. “Deep Learning-Based Multi-Omics Integration Robustly Predicts Survival in Liver Cancer”. In: *Clinical Cancer Research* 24.6 (2018).
- [30] Chen, X., Duan, Y., Houthoofd, R. et al. “InfoGAN: Interpretable Representation Learning by Information Maximizing Generative Adversarial Nets”. In: *Advances in Neural Information Processing Systems*. Vol. 29. 2016.
- [31] Chen, Y., Gao, Q., Wang, X. “Inferential Wasserstein Generative Adversarial Networks”. In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 84.1 (2021), pp. 83–113.
- [32] Chierici, M., Bussola, N., Marcolini, A. et al. “Integrative Network Fusion: A Multi-Omics Approach in Molecular Profiling”. In: *Frontiers in Oncology* 10 (2020).
- [33] Choi, E., Biswal, S., Malin, B.A. et al. “Generating Multi-label Discrete Patient Records using Generative Adversarial Networks”. In: *Machine Learning in Health Care*. 2017.
- [34] Chollet, F. et al. “Keras”. 2015. URL: <https://keras.io>.
- [35] Choudhury, A.D., Eeles, R., Freedland, S.J. et al. “The Role of Genetic Markers in the Management of Prostate Cancer”. In: *European Urology* 62.4 (2012), pp. 577–587.
- [36] Collins, M., Dasgupta, S., Schapire, R.E. “A Generalization of Principal Components Analysis to the Exponential Family”. In: *Advances in Neural Information Processing Systems*. Vol. 14. 2001.
- [37] Comon, P. “Independent component analysis, A new concept?” In: *Signal Processing* 36 (1994), pp. 287–314.
- [38] Cooper, C., Eeles, R., Wedge, D. et al. “Analysis of the genetic phylogeny of multifocal prostate cancer identifies multiple independent clonal expansions in neoplastic and morphologically normal prostate tissue”. In: *Nature Genetics* 47 (2015), pp. 367–372.
- [39] Crick, F. “Central Dogma of Molecular Biology”. In: *Nature* 227 (1970), pp. 561–563.

- [40] Cunningham, H., Ewart, A., Riggs, L. et al. “Sparse Autoencoders Find Highly Interpretable Features in Language Models”. 2023. arXiv: 2309.08600.
- [41] Cylus, J., Smith, P.C. “The economy of wellbeing: what is it and what are the implications for health?” In: *BMJ* 369 (2020).
- [42] Dallosso, A.R., Hancock, A.L., Szemes, M. et al. “Frequent Long-Range Epigenetic Silencing of Protocadherin Gene Clusters on Chromosome 5q31 in Wilms’ Tumor”. In: *PLOS Genetics* (2009).
- [43] Dempster, A.P., Laird, N.M., Rubin, D.B. “Maximum Likelihood from Incomplete Data Via the EM Algorithm”. In: *Journal of the Royal Statistical Society: Series B (Methodological)* 39.1 (1977), pp. 1–22.
- [44] Diao, P., Dai, Y., Wang, A. et al. “Integrative Multiomics Analyses Identify Molecular Subtypes of Head and Neck Squamous Cell Carcinoma with Distinct Therapeutic Vulnerabilities”. In: *Cancer Research* 84.18 (2024), pp. 3101–3117.
- [45] Dienstmann, R., Salazar, R., Tabernero, J. “Molecular Subtypes and the Evolution of Treatment Decisions in Metastatic Colorectal Cancer”. In: *American Society of Clinical Oncology Educational Book* (2018), pp. 231–238.
- [46] Ding, R.B., Chen, P., Rajendran, B.K. et al. “Molecular landscape and subtype-specific therapeutic response of nasopharyngeal carcinoma revealed by integrative pharmacogenomics”. In: *Nature Communications* 12 (2021).
- [47] Donahue, J. Krähenbühl, P., Darrell, T. “Adversarial Feature Learning”. In: *International Conference on Learning Representations*. 2017.
- [48] Dumoulin, V., Belghazi, I., Poole, B. et al. “Adversarially Learned Inference”. In: *International Conference on Learning Representations*. 2017. URL: <https://openreview.net/forum?id=B1E1R4cgg>.
- [49] El Hajj, N., Dittrich, M., Haaf, T. “Epigenetic dysregulation of protocadherins in human disease”. In: *Seminars in Cell & Developmental Biology* 69 (2017), pp. 172–182.

- [50] Eltager, M., Abdelaal, T., Mahfouz, A. et al. “scMoC: single-cell multi-omics clustering”. In: *Bioinformatics Advances* 2.1 (2022).
- [51] Esgueva, R., Perner, S., LaFargue, C.J., et al. “Prevalence of TMPRSS2-ERG and SLC45A3-ERG gene fusions in a large prostatectomy cohort”. In: *Modern Pathology* 23.4 (2010), pp. 539–546.
- [52] Fang, Z., Ma, T., Tang, G. et al. “Bayesian integrative model for multi-omics data with missingness”. In: *Bioinformatics* 34.22 (2018), pp. 3801–3808.
- [53] Feng, C., Wang, H., Lu, N. et al. “Log-transformation and its implications for data analysis”. In: *Shanghai Archives of Psychiatry* 26.2 (2014), pp. 105–109.
- [54] Ferguson, T.S. “A Bayesian Analysis of Some Nonparametric Problems”. In: *The Annals of Statistics* 1.2 (1973), pp. 209–230.
- [55] Fix, E., Hodges, J.L. “Discriminatory Analysis. Nonparametric Discrimination: Consistency Properties”. In: *International Statistical Review / Revue Internationale de Statistique* 57.3 (1989), pp. 238–247.
- [56] Flores, J.E., Claborne, D.M., Weller, Z.D. et al. “Missing data in multi-omics integration: Recent advances through artificial intelligence”. In: *Frontiers in Artificial Intelligence* 6 (2023).
- [57] Fraser M., Sabelnykova, V.Y., Yamaguchi, T.N. et al. “Genomic hallmarks of localized, non-indolent prostate cancer”. In: *Nature* 541 (2017), pp. 359–364.
- [58] Gadd, C., Nirantharakumar, K., Yau, C. “mmVAE: multimorbidity clustering using Relaxed Bernoulli β -Variational Autoencoders”. In: *Proceedings of the 2nd Machine Learning for Health symposium*. Vol. 193. 2022, pp. 88–102.
- [59] Gaudet, M.M., Press, M.F., Haile, R.W. et al. “Risk factors by molecular subtypes of breast cancer across a population-based study of women 56 years or younger”. In: *Breast Cancer Research and Treatment* 130 (2011), pp. 587–597.
- [60] Gelman, A., Hill, J. *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press, 2006, pp. 529–542.

- [61] Gemmete, J.J., Mukherjia, S.K. “Trastuzumab (Herceptin)”. In: *American Journal of Neuroradiology* 32.8 (2011), pp. 1373–1374.
- [62] Giansanti, V., Giannese, F., Botrugno, O.A. et al. “Scalable Integration of Multiomic Single Cell Data Using Generative Adversarial Networks”. In: *bioRxiv* (2023). DOI: 10.1101/2023.06.26.546547.
- [63] Goldhirsch, A., Wood, W.C., Coates, A.S. et al. “Strategies for subtypes—dealing with the diversity of breast cancer: highlights of the St Gallen International Expert Consensus on the Primary Therapy of Early Breast Cancer 2011”. In: *Annals of Oncology* 22.8 (2011), pp. 1736–1747.
- [64] Gonen, M., Alpaydin, E. “Multiple Kernel Learning Algorithms”. In: *Journal of Machine Learning Research* 12 (2011), pp. 2211–2268.
- [65] Goodfellow, I., Pouget-Abadie, J., Mirza, M. et al. “Generative Adversarial Nets”. In: *Advances in Neural Information Processing Systems*. Vol. 27. 2014.
- [66] Gopalan, A., Leversha, M.A., Satagopan, J.M. et al. “TMPRSS2-ERG gene fusion is not associated with outcome in patients treated by prostatectomy”. In: *Cancer Research* 69.4 (2009), pp. 1400–1406.
- [67] Grasso, C.S., Wu, Y.M., Robinson, D.R. et al. “The Mutational Landscape of Lethal Castrate Resistant Prostate Cancer”. In: *Nature* 487 (2012), pp. 239–243.
- [68] Griffiths, T.L., Ghahramani, Z. “The Indian Buffet Process: An Introduction and Review”. In: *Journal of Machine Learning Research* 12 (2011), pp. 1185–1224.
- [69] Grinsztajn, L., Oyallon, E., Varoquaux, G. “Why do tree-based models still outperform deep learning on typical tabular data?” In: *Proceedings of the 36th International Conference on Neural Information Processing Systems*. 2024.
- [70] Guinney, J., Dienstmann, R., Wang, X. et al. “The consensus molecular subtypes of colorectal cancer”. In: *Nature Methods* 21 (2015), pp. 1350–1356.

- [71] Gulrajani, I., Ahmed, F., Arjovsky, M. et al. “Improved Training of Wasserstein GANs”. In: *Advances in Neural Information Processing Systems*. Vol. 30. 2017.
- [72] Guo, X., Liu, X., Zhu, E. et al. “Deep Clustering with Convolutional Autoencoders”. In: *Neural Information Processing*. 2017, pp. 373–382.
- [73] Gyawali, P., Li, Z., Knight, C. et al. “Improving Disentangled Representation Learning with the Beta Bernoulli Process”. In: *2019 IEEE International Conference on Data Mining (ICDM)*. 2019, pp. 1078–1083.
- [74] Haier, J., Schaefer, J. “Economic Perspective of Cancer Care and Its Consequences for Vulnerable Groups”. In: *Cancers (Basel)* 14.13 (2022).
- [75] Hanahan, D., Weinberg, R.A. “The Hallmarks of Cancer”. In: *Cell* 100.1 (2000), pp. 57–70.
- [76] Hassija, V., Chamola, V., Mahapatra, A. et al. “Interpreting Black-Box Models: A Review on Explainable Artificial Intelligence”. In: *Cognitive Computation* 4 (2024), pp. 45–74.
- [77] Hawinkel S., Bijmens L., Cao K.L. et al. “Model-based joint visualization of multiple compositional omics datasets”. In: *NAR Genomics and Bioinformatics* 2.3 (2020).
- [78] Hayes, V., Bornman, M.S.R. “Prostate cancer in Southern Africa: does Africa hold untapped potential to add value to the current understanding of a common disease?” In: *Journal of Global Oncology* 4 (2018).
- [79] Hayes, V., Jiang, J., Tapinos, A. et al. “Kataegis associated mutational processes linked to adverse prostate cancer presentation in African men”. [Preprint]. 2024.
- [80] Heo, Y.J., Hwa, C., Lee G-H. et al. “Integrative Multi-Omics Approaches in Cancer Research: From Biological Networks to Clinical Subtypes”. In: *Molecules and Cells* 44.7 (2021), pp. 433–443.

- [81] Hieronymus, H., Schultz, Gopalan, A. et al. “Copy number alteration burden predicts prostate cancer relapse”. In: *Proceedings of the National Academy of Sciences of the United States of America* 111.30 (2014), pp. 11139–11144.
- [82] Hinton, G.E. “Products of Experts”. In: *Proceedings of the Ninth International Conference on Artificial Neural Networks*. Vol. 1. 1999, pp. 1–6.
- [83] Hinton, G.E. “Training Products of Experts by Minimizing Contrastive Divergence”. In: *Neural Computation* 14.8 (2002), pp. 1771–1800.
- [84] Hira, M.T., Razzaque, M.A., Angione, C. et al. “Integrated multi-omics analysis of ovarian cancer using variational autoencoders”. In: *Scientific Reports* 11.1 (2021), p. 6265.
- [85] Hotelling, H. “Analysis of a complex of statistical variables into principal components”. In: *Journal of Educational Psychology* 24.6 (1933), pp. 417–441.
- [86] Howey, R., Clark, A.D., Naamane, N. et al. “A Bayesian network approach incorporating imputation of missing data enables exploratory analysis of complex causal biological relationships”. In: *PLOS Genetics* 17.9 (2021), pp. 1–28.
- [87] International Cancer Genome Consortium. ICGC ARGO Policies and Guidelines [Online]. Accessed: October 2024. URL: <https://www.icgc-argo.org/page/75/e1-ethics-and-informed-consent>.
- [88] Iqbal, N., Iqbal, N. “Human Epidermal Growth Factor Receptor 2 (HER2) in Cancers: Overexpression and Therapeutic Implications”. In: *Molecular Biology International* (2014).
- [89] Islam, J., Zhang, Y. “GAN-based synthetic brain PET image generation”. In: *Brain Informatics* 7.3 (2020).
- [90] Islam, M.M., Huang, S., Ajwad, R. et al. “An integrative deep learning framework for classifying molecular subtypes of breast cancer”. In: *Computational and Structural Biotechnology Journal* 18 (2020), pp. 2185–2199.

- [91] Isola, P., Zhu, J.-Y., Zhou, T. et al. “Image-to-Image Translation with Conditional Adversarial Networks”. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition CVPR*. 2017, pp. 5967–5976.
- [92] Jadon, A., Patil, A., Jadon, S. “A Comprehensive Survey of Regression-Based Loss Functions for Time Series Forecasting”. In: *Data Management, Analytics and Innovation*. 2024, pp. 117–147.
- [93] Jamaspishvili, T., Berman, D.M., Ross, A.E. et al. “Clinical implications of PTEN loss in prostate cancer”. In: *Nature Reviews Urology* 15 (2018), pp. 222–234.
- [94] Jarada, T.N., Rokne, J.G., Alhajj, R. “SNF-NN: computational method to predict drug-disease interactions using similarity network fusion and neural networks”. In: *BMC Bioinformatics* 22 (2021).
- [95] Jaratlerdsiri, W., Chan, E.K.F., Gong, T. et al. “Whole-genome sequencing reveals elevated tumor mutational burden and initiating driver mutations in African men with treatment-naïve, high-risk prostate cancer”. In: *Cancer Research* 78.24 (2018), pp. 6736–6746.
- [96] Jaratlerdsiri, W., Jiang, J., Gong, T. et al. “African-specific molecular taxonomy of prostate cancer”. In: *Nature* 609 (2022), pp. 552–559.
- [97] Jiang, S., Pang, G., Wu, M. et al. “An improved K-nearest-neighbor algorithm for text categorization”. In: *Expert Systems with Applications* 39.1 (2012), pp. 1503–1509.
- [98] Jolliffe I.T. *Principal component analysis*. NY: Springer-Verlag, 2002.
- [99] Kaffenberger, S.D., Barbieri, C.E. “Molecular Subtyping of Prostate Cancer”. In: *Current Opinion in Urology* 26.3 (2016), pp. 213–218.
- [100] Kamoun, A., Cancel-Tassin, G., Fromont, G. et al. “Comprehensive molecular classification of localized prostate adenocarcinoma reveals a tumour subtype predictive of non-aggressive disease”. In: *Annals of Oncology* 29.8 (2018), pp. 1814–1821.
- [101] Kaplan, E.L., Meier, P. “Nonparametric Estimation from Incomplete Observations”. In: *Journal of the American Statistical Association* 53 (1958), pp. 457–481.

- [102] Karras, T., Laine, S., Aila, T. “A Style-Based Generator Architecture for Generative Adversarial Networks”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 43.12 (2021), pp. 4217–4228.
- [103] Kingma, D.P., Welling, M. “Auto-Encoding Variational Bayes”. 2013. arXiv: 1312.6114.
- [104] Konno, H., Yamauchi, S., Berglund, A. et al. “Suppression of STING signaling through epigenetic silencing and missense mutation impedes DNA damage mediated cytokine production”. In: *Oncogene* 37 (2018), pp. 2037–2051.
- [105] Kovtun, I.V., Murphy, S.J., Johnson, S.H. et al. “Chromosomal catastrophe is a frequent event in clinically insignificant prostate cancer”. In: *Oncotarget* 6.30 (2015).
- [106] Lapointe, J., Li, C., Higgins, J.P. et al. “Gene expression profiling identifies clinically relevant subtypes of prostate cancer”. In: *Proceedings of the National Academy of Sciences of the United States of America*. Vol. 101. 3. 2004, pp. 811–816.
- [107] Larsen, A.B.L., Sønderby, S.K., Larochelle, H. et al. “Autoencoding beyond pixels using a learned similarity metric”. In: *Proceedings of The 33rd International Conference on Machine Learning*. Vol. 48. 2016, pp. 1558–1566.
- [108] Latchman, D.S. “Inhibitory transcription factors”. In: *The International Journal of Biochemistry & Cell Biology* 28.9 (1996), pp. 965–974.
- [109] Lazar, C., Gatto, L., Ferro, M. et al. “Accounting for the Multiple Natures of Missing Values in Label-Free Quantitative Proteomics Data Sets to Compare Imputation Strategies”. In: *Journal of Proteome Research* 15.4 (2016), pp. 1116–1125.
- [110] Le, D.T., Durham, J.N., Smith, K.N. et al. “Mismatch-repair deficiency predicts response of solid tumors to PD-1 blockade”. In: *Science* 357 (2017), pp. 409–413.

- [111] Le, L., Patterson, A., White, M. “Supervised autoencoders: Improving generalization performance with unsupervised regularizers”. In: *Advances in Neural Information Processing Systems*. Vol. 31. 2018.
- [112] Ledig, C., Theis, L., Huszar, F. et al. “Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network”. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition CVPR*. 2017, pp. 105–114.
- [113] Lee, C., van der Schaar, M. “A Variational Information Bottleneck Approach to Multi-Omics Data Integration”. In: *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*. Vol. 130. 2021, pp. 1513–1521.
- [114] Lee, D.D., Seung, H.S. “Learning the parts of objects by non-negative matrix factorization”. In: *Nature* 401 (1999), pp. 788–791.
- [115] Lennartz, M., Minner, S., Brasch, S. et al. “The Combination of DNA Ploidy Status and PTEN/6q15 Deletions Provides Strong and Independent Prognostic Information in Prostate Cancer”. In: *Clinical Cancer Research* 22.11 (2016), pp. 2802–2811.
- [116] Li, C., Liu, H., Chen, C. et al. “ALICE: Towards Understanding Adversarial Learning for Joint Distribution Matching”. In: *Advances in Neural Information Processing Systems*. Vol. 30. 2017.
- [117] Lièvre, A. Bachet, J-B., Le Corre, D. et al. “KRAS mutation status is predictive of response to cetuximab therapy in colorectal cancer”. In: *Cancer Research* 66.8 (2006), pp. 3992–3995.
- [118] Lin, Z., Owen, A.B., Altman, R.B. “Genomic Research and Human Subject Privacy”. In: *Science* 305 (2004), pp. 183–183.
- [119] Little, R.J.A., Rubin, D.B. *Statistical Analysis with Missing Data*. John Wiley & Sons, 2002, pp. 11–19.
- [120] Liu, D., Augello, M.A., Grbesa, I. et al. “Tumor subtype defines distinct pathways of molecular and clinical progression in primary prostate cancer”. In: *Journal of Clinical Investigation* 131.10 (2021).

- [121] Liu, J., Kim, S.Y., Shin, S. et al. “Overexpression of TFF3 is involved in prostate carcinogenesis via blocking mitochondria-mediated apoptosis”. In: *Experimental & Molecular Medicine* 50 (2018), pp. 1–11.
- [122] Lloyd, T., Hounsome, L., Mehay, A. et al. “Lifetime risk of being diagnosed with, or dying from, prostate cancer by major ethnic group in England 2008-2010”. In: *BMC Medicine* 13 (2015).
- [123] Lu, M., Zhan, X. “The crucial role of multiomic approach in cancer research and clinically relevant outcomes”. In: *EPMA Journal* 9 (2018), pp. 77–102.
- [124] Lubin, J.H., Colt, J.S., Camann D. et al. “Epidemiologic evaluation of measurement data in the presence of detection limits”. In: *Environmental Health Perspectives* 112.17 (2004), pp. 1691–1696.
- [125] Luca, B.A., Brewer, D.S., Edwards, D.R. et al. “DESNT: A Poor Prognosis Category of Human Prostate Cancer”. In: *European Urology Focus* 4.6 (2018), pp. 842–850.
- [126] Luca, B.A., Moulton, V., Ellis, C. et al. “A novel stratification framework for predicting outcome in patients with prostate cancer”. In: *British Journal of Cancer* 22 (2020), pp. 1467–1476.
- [127] Lundberg, S.M., Lee, S-I. “A Unified Approach to Interpreting Model Predictions”. In: *Advances in Neural Information Processing Systems*. Vol. 30. 2017.
- [128] Lundby, A., Franciosa, G., Emdal, K.B. et al. “Oncogenic Mutations Rewire Signaling Pathways by Switching Protein Recruitment to Phosphotyrosine Sites”. In: *Cell* 179.2 (2019), pp. 543–560.
- [129] Luo, Y., Tao, D., Ramamohanarao, K. et al. “Tensor Canonical Correlation Analysis for Multi-View Dimension Reduction”. In: *IEEE Transactions on Knowledge and Data Engineering* 27.11 (2015), pp. 3111–3124.
- [130] Ma, T., Zhang, A. “Affinity network fusion and semi-supervised learning for cancer patient clustering”. In: *Methods* 145 (2018), pp. 16–24.

- [131] Mah, K.M., Houston, D.W., Weiner, J.A. “The γ -Protocadherin-C3 isoform inhibits canonical Wnt signalling by binding to and stabilizing Axin1 at the membrane”. In: *Scientific Reports* 6 (2016).
- [132] Makhzani, A., Shlens, J., Jaitly, N. et al. “Adversarial Autoencoders”. In: *International Conference on Learning Representations*. 2016. URL: <http://arxiv.org/abs/1511.05644>.
- [133] Markert, E.K., Mizuno, H., Vazquez, A. et al. “Molecular classification of prostate cancer using curated expression signatures”. In: *Proceedings of the National Academy of Sciences of the United States of America*. Vol. 108. 52. 2011, pp. 21276–21281.
- [134] Marouf, M., Machart, P., Bansal, V. et al. “Realistic in silico generation and augmentation of single-cell RNA-seq data using generative adversarial networks”. In: *Nature Communications* 11 (2020).
- [135] McCullagh, P. *Generalized Linear Models (2nd ed.)* Routledge, 1989.
- [136] Meng, C., Kuster, B., Culhane, A.C. et al. “A multivariate approach to the integration of multi-omics datasets”. In: *BMC Bioinformatics* 15 (2014).
- [137] Mo, Q., Shen, R., Guo, C. et al. “A fully Bayesian latent variable model for integrative clustering analysis of multi-type omics data”. In: *Biostatistics* 19.1 (2018), pp. 71–86.
- [138] Mo, Q., Wang, S., Seshan, V. E. et al. “Pattern discovery and cancer gene identification in integrated cancer genomic data”. In: *Proceedings of the National Academy of Sciences of the United States of America*. Vol. 110. 11. 2013, pp. 4245–4250.
- [139] Molania, R., Gagnon-Bartsch, J.A., Dobrovic, A. et al. “A new normalization for Nanostring nCounter gene expression data”. In: *Nucleic Acids Research* 47.12 (2019), pp. 6073–6083.
- [140] Nabney, I.T., Sun, Y., Tino, P. et al. “Semisupervised learning of hierarchical latent trait models for data visualization”. In: *IEEE Transactions on Knowledge and Data Engineering* 17.3 (2005), pp. 384–400.

- [141] Nakazawa, M., Fang, M., Marshall, C.H. et al. “Clinical and genomic features of SPOP-mutant prostate cancer”. In: *Prostate* 82.2 (2022), pp. 260–268.
- [142] Nam, R.K., Sugar, L., Wang, Z. et al. “Expression of TMPRSS2:ERG gene fusion in prostate cancer cells is an important prognostic factor for cancer progression”. In: *Cancer Biology & Therapy* 6.1 (2007).
- [143] Nguyen, T.D., Tran, T., Phung, D. et al. “Learning Parts-based Representations with Nonnegative Restricted Boltzmann Machine”. In: *Proceedings of the 5th Asian Conference on Machine Learning*. Vol. 29. 2013, pp. 133–148.
- [144] Nocedal, J., Wright, S.J. *Numerical Optimization*. Springer, 2020, pp. 497–506.
- [145] Nordgren, A. “Artificial intelligence and climate change: ethical issues”. In: *Journal of Information, Communication and Ethics in Society* 21.1 (2023).
- [146] Onitilo, A.A., Engel, J.M., Greenlee, R.T. et al. “Breast Cancer Subtypes Based on ER/PR and Her2 Expression: Comparison of Clinicopathologic Features and Survival”. In: *Clinical Medicine & Research* 7.1-2 (2009), pp. 4–13.
- [147] Pantanowitz, A., Marwala, T. “Missing Data Imputation Through the Use of the Random Forest Algorithm”. In: *Advances in Computational Intelligence*. 2009, pp. 53–62.
- [148] Park, J-E., Mu, W., Jiao, Y. et al. “MultImp: Multiomics Generative Models for Data Imputation”. In: *The 2021 ICML Workshop on Computational Biology*. 2021.
- [149] Park, N., Mohammadi, M., Gorde, K. et al. “Data synthesis based on generative adversarial networks”. In: *Proceedings of the VLDB Endowment* 11 (2018), pp. 1071–1083.
- [150] Pascal, L.E., True, L.D., Campbell, D.S. et al. “Correlation of mRNA and protein levels: Cell type-specific gene expression of cluster designation antigens in the prostate”. In: *BMC Genomics* 9 (2008).

- [151] Patterson, D., Gonzalez, J., Hölzle, U. et al. “The Carbon Footprint of Machine Learning Training Will Plateau, Then Shrink”. In: *Computer* 55.7 (2022), pp. 18–28.
- [152] Pearce, T., Jeong, J-H., jia, y. et al. “Censored Quantile Regression Neural Networks for Distribution-Free Survival Analysis”. In: *Advances in Neural Information Processing Systems*. Vol. 35. 2022.
- [153] Pearson, K. “LIII. On lines and planes of closest fit to systems of points in space”. In: *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 2.11 (1901), pp. 559–572.
- [154] Perner, S., Demichelis, F., Beroukhim, R. et al. “TMPRSS2:ERG fusion-associated deletions provide insight into the heterogeneity of prostate cancer”. In: *Cancer Research* 66.17 (2006), pp. 8337–8341.
- [155] Perry, T., Bernasek, A. “Profits over care? An analysis of the relationship between corporate capitalism in the healthcare industry and cancer mortality in the United States”. In: *Social Science & Medicine* 349 (2024).
- [156] Picard, M., Scott-Boyer, MP., Bodein, A. et al. “Integration strategies of multi-omics data for machine learning analysis”. In: *Computational and Structural Biotechnology Journal* 19 (2021), pp. 3735–3746.
- [157] Portnoy, S. “Censored Regression Quantiles”. In: *Journal of the American Statistical Association* 98 (2003), pp. 1001–1012.
- [158] Racle, J., Gfeller, D. “EPIC: A Tool to Estimate the Proportions of Different Cell Types from Bulk Gene Expression Data”. In: *Bioinformatics for Cancer Immunotherapy: Methods and Protocols*. Springer US, 2020, pp. 233–248.
- [159] Ramalingam, S.S., Vansteenkiste, J., Planchard, D. et al. “Overall Survival with Osimertinib in Untreated, EGFR-Mutated Advanced NSCLC”. In: *New England Journal of Medicine* 382.1 (2020), pp. 41–50.
- [160] Rappoport, N., Safra, R., Shamir, R. “MONET: Multi-omic module discovery by omic selection”. In: *PLOS Computational Biology* 16.9 (2020).

- [161] Raufaste-Cazavieille, V., Santiago, R., Droit, A. “Multi-omics analysis: Paving the path toward achieving precision medicine in cancer treatment and immuno-oncology”. In: *Frontiers in Molecular Biosciences* 9 (2022).
- [162] Reis, R.J.S., Bharill, P., Tazearslan, C. et al. “Extreme-longevity mutations orchestrate silencing of multiple signaling pathways”. In: *Biochimica et Biophysica Acta (BBA) - General Subjects* 1790.10 (2009), pp. 1075–1083.
- [163] Richiardi, L., Barone-Adesi, F., Pearce, N. “Cancer subtypes in aetiological research”. In: *European Journal of Epidemiology* 32 (2017), pp. 353–361.
- [164] Rubin, D.B. “Inference and missing data”. In: *Biometrika* 63.3 (1976), pp. 581–592.
- [165] Rumelhart, D.E., Hinton, G.E., Williams, R.J. “Learning Internal Representations by Error Propagation”. In: *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Vol. 1: Foundations*. 1986, pp. 318–362.
- [166] Sabatini, M.E., Chiocca, S. “Human papillomavirus as a driver of head and neck cancers”. In: *British Journal of Cancer* 122 (2020), pp. 306–314.
- [167] Sahm, F., Schrimpf, D., Stichel, D., et al. “DNA methylation-based classification and grading system for meningioma: a multicentre, retrospective analysis”. In: *The Lancet Oncology* 18.5 (2017), pp. 682–694.
- [168] Salachan, P.V., Ulhøi, B.P., Borre, M. et al. “Association between copy number alterations estimated using low-pass whole genome sequencing of formalin-fixed paraffin-embedded prostate tumor tissue and cancer-specific clinical parameters”. In: *Scientific Reports* 13 (2023).
- [169] Sanger, F., Nicklen, S., Coulson, A.R. “DNA sequencing with chain-terminating inhibitors”. In: *Proceedings of the National Academy of Sciences of the United States of America*. Vol. 74. 12. 1977, pp. 5463–5467.

- [170] Sharifi-Noghabi, H., Zolotareva, O., Collins, C.C. et al. “MOLI: multi-omics late integration with deep neural networks for drug response prediction”. In: *Bioinformatics* 35.14 (2019), pp. 501–509.
- [171] Shen, M.M. “Chromoplexy: A New Category of Complex Rearrangements in the Cancer Genome”. In: *Cancer Cell* 23.5 (2013), pp. 567–569.
- [172] Shen, R., Olshen, A.B., Ladanyi, M. “Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis”. In: *Bioinformatics* 25.22 (2009), pp. 2906–2912.
- [173] Sikka, A., Skand, Virk, J.S. et al. “MRI to PET Cross-Modality Translation using Globally and Locally Aware GAN (GLA-GAN) for Multi-Modal Diagnosis of Alzheimer’s Disease”. 2021. arXiv: 2108.02160.
- [174] Singh D., Febbo P.G., Ross K. et al. “Gene expression correlates of clinical prostate cancer behavior”. In: *Cancer Cell* 1.2 (2002), pp. 203–209.
- [175] Slamon, D.J., Clark, G.M., Wong, S.G. et al. “Human breast cancer: correlation of relapse and survival with amplification of the HER-2/neu oncogene”. In: *Science* 235 (1987), pp. 177–182.
- [176] Spicker, J.S., Brunak, S., Frederiksen, K.S., et al. “Integration of clinical chemistry, expression, and metabolite data leads to better toxicological class separation”. In: *Toxicological Sciences* 102.2 (2008), pp. 444–454.
- [177] Strehl, A., Ghosh, J. “Cluster Ensembles – A Knowledge Reuse Framework for Combining Multiple Partitions”. In: *Journal of Machine Learning Research* 3 (2002), pp. 583–617.
- [178] Succop, P.A., Clark, S., Chen, M. et al. “Imputation of data values that are less than a detection limit”. In: *Journal of Occupational and Environmental Hygiene* 1.7 (2004), pp. 436–441.

- [179] Sun, D., Wang, M., Li, A. “A Multimodal Deep Neural Network for Human Breast Cancer Prognosis Prediction by Integrating Multi-Dimensional Data”. In: *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 16.3 (2019), pp. 841–850.
- [180] Sunkel, B., Wu, D., Chen, Z. et al. “Integrative analysis identifies targetable CREB1/FoxA1 transcriptional co-regulation as a predictor of prostate cancer recurrence”. In: *Nucleic Acids Research* 44.9 (2016), pp. 4105–4122.
- [181] Taeksoo, K., Moonsu C., Hyunsoo K. et al. “Learning to Discover Cross-Domain Relations with Generative Adversarial Networks”. In: *Proceedings of the 34th International Conference on Machine Learning*. Vol. 70. 2017, pp. 1857–1865.
- [182] Taylor, B.S., Schultz, N., Hieronymus, H. et al. “Integrative genomic profiling of human prostate cancer”. In: *Cancer Cell* 18.1 (2010), pp. 11–22.
- [183] Teh, Y.W., Grür, D., Ghahramani, Z. “Stick-breaking Construction for the Indian Buffet Process”. In: *Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics*. Vol. 2. 2007, pp. 556–563.
- [184] ten Hoorn, S., de Back, T.R., Sommeijer, D.W. et al. “Clinical Value of Consensus Molecular Subtypes in Colorectal Cancer: A Systematic Review and Meta-Analysis”. In: *JNCI: Journal of the National Cancer Institute* 114.4 (2021), pp. 503–516.
- [185] Tenenhaus, M., Tenenhaus, A., Groenen, P.J.F. “Regularized Generalized Canonical Correlation Analysis: A Framework for Sequential Multiblock Component Methods”. In: *Psychometrika* 82.3 (2017), pp. 737–777.
- [186] Teroerde, M., Nientiedt, C., Duensing, A. et al. *Prostate Cancer*. Exon Publications, 2021.
- [187] Terry, S., Nicolaiew, N., Basset, V. et al. “Clinical value of ERG, TFF3, and SPINK1 for molecular subtyping of prostate cancer”. In: *Cancer* 121.9 (2015), pp. 1422–1430.

- [188] The Cancer Genome Atlas Network. “Comprehensive molecular portraits of human breast tumours”. In: *Nature* 490 (2012), pp. 61–70.
- [189] The ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium. “Pan-cancer analysis of whole genomes”. In: *Nature* 578 (2020), pp. 82–93.
- [190] Tindall, E.A., Monare, L.R., Petersen, D.C. et al. “Clinical presentation of prostate cancer in black South Africans”. In: *The Prostate* 74.8 (2014), pp. 880–891.
- [191] Tini, G., Marchetti, L., Priami, C. et al. “Multi-omics integration—a comparison of unsupervised clustering methodologies”. In: *Briefings in Bioinformatics* 20.4 (2019), pp. 1269–1279.
- [192] Tomlins, S.A., Laxman, B., Dhanasekaran, S.M. et al. “Distinct classes of chromosomal rearrangements create oncogenic ETS gene fusions in prostate cancer”. In: *Nature* 448 (2007), pp. 595–599.
- [193] Tomlins, S.A., Laxman, B., Varambally, S. et al. “Role of the TMPRSS2-ERG gene fusion in prostate cancer”. In: *Neoplasia* 10.2 (2008), pp. 177–188.
- [194] Tomlins, S.A., Mehra, R., Rhodes, D.R. et al. “Integrative molecular concept modeling of prostate cancer progression”. In: *Nature Genetics* 39.1 (2007), pp. 41–51.
- [195] Tran, T., Nguyen, T.D., Phung, D. et al. “Learning vector representation of medical objects via EMR-driven nonnegative restricted Boltzmann machines (eNRBM)”. In: *Journal of Biomedical Informatics* 54 (2015), pp. 96–105.
- [196] Ullmann, T., Hennig, C., Boulesteix, A.-L. “Validation of cluster analysis results on validation data: A systematic framework”. In: *WIREs Data Mining and Knowledge Discovery* 12.3 (2022).
- [197] van Wynsberghe, A. “Sustainable AI: AI for sustainability and the sustainability of AI”. In: *AI and Ethics* 1 (2021), pp. 213–218.
- [198] Villani, C. *Optimal Transport: Old and New*. Grundlehren der mathematischen Wissenschaften. Springer Berlin Heidelberg, 2008.

- [199] Viñas, R., Andrés-Terré, H., Liò, P. et al. “Adversarial generation of gene expression data”. In: *Bioinformatics* 38.3 (2022), pp. 730–737.
- [200] Wallace, C.S., Dowe, D.L. “MML clustering of multi-state, Poisson, von Mises circular and Gaussian distributions”. In: *Statistics and Computing* 10 (2000), pp. 73–83.
- [201] Wang, B., Mezlini, A., Demir, F. et al. “Similarity network fusion for aggregating data types on a genomic scale”. In: *Nature Methods* 11.3 (2014), pp. 333–337.
- [202] Wang, J., Cai, Y., Ren, C. et al. “Expression of variant TMPRSS2/ERG fusion messenger RNAs is associated with aggressive prostate cancer”. In: *Cancer Research* 66.17 (2006), pp. 8347–8351.
- [203] Wang, Q., Powell, M.A., Geisa, A. et al. “Why do networks have inhibitory/negative connections?” 2023. arXiv: 2208.03211.
- [204] Wang, T., Shao, W., Huang, Z. et al. “MOGONET integrates multi-omics data using graph convolutional networks allowing patient classification and biomarker identification”. In: *Nature Communications* 12 (2021).
- [205] Wang, W., Baladandayuthapani, V., Morris, J.S. et al. “iBAG: integrative Bayesian analysis of high-dimensional multiplatform genomics data”. In: *Bioinformatics* 29.2 (2013), pp. 149–159.
- [206] Wang, X., Grasso, C.S., Jordahl, K.M., et al. “Copy number alterations are associated with metastatic-lethal progression in prostate cancer”. In: *Prostate Cancer and Prostatic Diseases* 23 (2020), pp. 494–506.
- [207] Wang, Z., Gerstein, M., Snyder, M. “RNA-Seq: a revolutionary tool for transcriptomics”. In: *Nature Reviews Genetics* 10 (2009), pp. 57–63.
- [208] Way, G.P., Greene, C.S. “Extracting a biologically relevant latent space from cancer transcriptomes with variational autoencoders”. In: *Pacific Symposium on Biocomputing* 23 (2018), pp. 80–91.

- [209] Webb-Robertson, B.J., Wiberg H.K., Matzke, M.M. et al. “Review, evaluation, and discussion of the challenges of missing value imputation for mass spectrometry-based label-free global proteomics”. In: *Journal of Proteome Research* 14.5 (2015), pp. 1993–2001.
- [210] Wedge, D.C., Gundem, G., Mitchell, T. et al. “Sequencing of prostate cancers identifies new cancer genes, routes of progression and drug targets”. In: *Nature Genetics* 50.5 (2018), pp. 682–692.
- [211] Wei, R., Wang, J., Su, M. et al. “Missing Value Imputation Approach for Mass Spectrometry-based Metabolomics Data”. In: *Scientific Reports* 8 (2018).
- [212] Wen, Y., Song, X., Yan, B. et al. “Multi-dimensional data integration algorithm based on random walk with restart”. In: *BMC Bioinformatics* 22 (2021).
- [213] Woodcock, D.J., Sahli, A., Teslo, R. et al. “Genomic evolution shapes prostate cancer disease type”. In: *Cell Genomics* 4.3 (2024).
- [214] Wu, C., Zhang, Q., Jiang, Y. et al. “Robust network-based analysis of the associations between (epi)genetic measurements”. In: *Journal of Multivariate Analysis* 168 (2018), pp. 119–130.
- [215] Wu, M., Goodman, N. “Multimodal Generative Models for Scalable Weakly-Supervised Learning”. In: *Advances in Neural Information Processing Systems*. Vol. 31. 2018.
- [216] Wu, W., Sato, K., Koike, A. et al. “HERC2 Is an E3 Ligase That Targets BRCA1 for Degradation”. In: *Cancer Research* 70.15 (2010), pp. 6384–6392.
- [217] Xie, J., Girshick, R., Farhadi, A. “Unsupervised Deep Embedding for Clustering Analysis”. In: *Proceedings of The 33rd International Conference on Machine Learning*. Vol. 48. 2016, pp. 478–487.
- [218] Xu, J., Wu, P., Chen, Y. et al. “A hierarchical integration deep flexible neural forest framework for cancer subtype classification by integrating multi-omics data”. In: *BMC Bioinformatics* 20 (2019).

- [219] Xu, L., Skoularidou, M., Cuesta-Infante, A. et al. “Modeling Tabular data using Conditional GAN”. In: *Advances in Neural Information Processing Systems*. Vol. 32. 2019.
- [220] Yan, J., Risacher, S.L., Shen, L. et al. “Network approaches to systems biology analysis of complex disease: integrative methods for multi-omics data”. In: *Briefings in Bioinformatics* 19.6 (2018), pp. 1370–1381.
- [221] Yang, H., Chen, R., Li, D. et al. “Subtype-GAN: a deep learning approach for integrative cancer subtyping of multi-omics data”. In: *Bioinformatics* 37.16 (2021), pp. 2231–2237.
- [222] Yang, L., Wang, S., Zhou, M. et al. “Molecular classification of prostate adenocarcinoma by the integrated somatic mutation profiles and molecular network”. In: *Scientific Reports* 7 (2017).
- [223] Yang, Z., Michailidis, G. “A non-negative matrix factorization method for detecting modules in heterogeneous omics multi-modal data”. In: *Bioinformatics* 32.1 (2016), pp. 1–8.
- [224] Yi, Z., Zhang, H., Tan, P. et al. “DualGAN: Unsupervised Dual Learning for Image-to-Image Translation”. In: *2017 IEEE International Conference on Computer Vision (ICCV)*. 2017, pp. 2868–2876.
- [225] Ylipää, A., Kivinummi, K., Kohvakka, A. et al. “Transcriptome Sequencing Reveals PCAT5 as a Novel ERG-Regulated Long Noncoding RNA in Prostate Cancer”. In: *Cancer Research* 74.19 (2015), pp. 4026–4031.
- [226] Yoon, J., Jordon, J., van der Schaar, M. “GAIN: Missing Data Imputation using Generative Adversarial Nets”. In: *Proceedings of the 35th International Conference on Machine Learning*. Vol. 80. 2018, pp. 5689–698.
- [227] You, S., Knudsen, B.S., Erho, N. et al. “Integrated classification of prostate cancer reveals a novel luminal subtype with poor outcome”. In: *Cancer Research* 76.17 (2016), pp. 4948–4958.

- [228] Zarayeneh, N., Ko, E., Oh, J.H. et al. “Integration of multi-omics data for integrative gene regulatory network inference”. In: *International Journal of Data Mining and Bioinformatics* 13.3 (2017), pp. 223–239.
- [229] Zass, R., Shashua, A. “Nonnegative Sparse PCA”. In: *Advances in Neural Information Processing Systems*. Vol. 19. 2006.
- [230] Zhang, L., Lv, C., Jin, Y. et al. “Deep Learning-Based Multi-Omics Data Integration Reveals Two Prognostic types in High-Risk Neuroblastoma”. In: *Frontiers in Genetics* 9 (2018).
- [231] Zhang, Q., Wang, H., Lu, H., et al. “Medical Image Synthesis with Generative Adversarial Networks for Tissue Recognition”. In: *2018 IEEE International Conference on Healthcare Informatics (ICHI)*. 2018, pp. 199–207.
- [232] Zhang, S., Liu, CC., Li, E. et al. “Discovery of multi-dimensional modules by integrative analysis of cancer genomic data”. In: *Nucleic Acids Research* 40.19 (2012), pp. 9379–9391.
- [233] Zhang, X., Shi, H., Yao, J. et al. “FAM225A facilitates colorectal cancer progression by sponging miR-613 to regulate NOTCH3”. In: *Cancer Medicine* 9.12 (2020), pp. 4339–4349.
- [234] Zhang, X., Zhang, J., Sun, K. et al. “Integrated Multi-omics Analysis Using Variational Autoencoders: Application to Pan-cancer Classification”. In: *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. 2019.
- [235] Zhang, Y., Zhao, Y., Wang, G. et al. “Mean square cross error: performance analysis and applications in non-Gaussian signal processing”. In: *EURASIP Journal on Advances in Signal Processing* (2021).
- [236] Zhao, Z., Kunar, A., Birke, R. et al. “CTAB-GAN: Effective Table Data Synthesizing”. In: *Proceedings of The 13th Asian Conference on Machine Learning*. Vol. 157. 2021, pp. 97–112.
- [237] Zheng, K., Hai, Y., Xi, Y. et al. “Integrative multi-omics analysis unveils stemness-associated molecular subtypes in prostate cancer and pan-cancer: prognostic and therapeutic significance”. In: *Journal of Translational Medicine* 21 (2023).

- [238] Zhou, X., Chai, H., Zhao, H. et al. “Imputing missing RNA-sequencing data from DNA methylation by using a transfer learning–based neural network”. In: *GigaScience* 9.7 (2020).
- [239] Zhu, J-Y., Park, T., Isola, P. et al. “Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks”. In: *2017 IEEE International Conference on Computer Vision (ICCV)*. 2017, pp. 2242–2251.
- [240] Zhu, R., Zhao, Q., Zhao, H. et al. “Integrating multidimensional omics data for cancer outcome”. In: *Bioinformatics* 17.4 (2016), pp. 605–618.
- [241] Zou, H., Hastie, T., Tibshirani, R. “Sparse Principal Component Analysis”. In: *Journal of Computational and Graphical Statistics* 15.2 (2006), pp. 265–286.

Appendix A

Implementation Details and Experimental Setup

A.1 iCS-GAN - Implementation and Default Hyperparameters

In this section, we describe the PyTorch implementation of iCS-GAN and outline the default hyperparameters used in both the synthetic data experiments and the TCGA experiments.

Optimization

iCS-GAN uses RMSprop as the optimizer, with the following default settings: a learning rate of $5e-5$, a batch size of 64, an L2 penalty strength of $1e-4$, and an L1 regularization strength of $1e-3$.

For single-modality training (Step 1 of the Multimodal Training Procedure) or the integration generator training (Step 2 of the Multimodal Training Procedure), the iCS-GAN training (without clustering fine-tuning, which is described later) followed this process: the model was first trained for 500 epochs as a warmup. Afterward, the training RMSE was recorded at each

epoch. If the RMSE was lower than the previous minimum, the model was saved. If the lowest RMSE remained unchanged for 100 consecutive epochs, training was stopped.

For multimodal training (Step 3 of the Multimodal Training Procedure), the generator, initialized with parameters from single-modality and integration training, was kept frozen while both critics were trained until convergence. The critic-only training proceeded as follows: after a 500-epoch warmup, we began recording the sum of both critics' losses at each epoch. When the lowest loss remained unchanged for 100 consecutive epochs, critic training was completed, and the generator was unfrozen (moving us to Step 4 of the Multimodal Training Procedure). From that point, the model underwent another 50-epoch warmup, after which the training RMSE was recorded at each epoch. As before, if the RMSE at the current epoch was lower than the previous minimum, the model was saved. If the RMSE remained unchanged for 100 consecutive epochs, the multimodal training process was halted.

Generator

Each single-modality generator (Step 1 of the Multimodal Training Procedure) was a shallow, fully-connected input-encoding-output network with batch normalization, tied weights, and non-negativity constraints. The encoding layer used a sigmoid activation function. For the output layer, sigmoid activation was applied to continuous data types, while a modified softsign activation (with a smoothing constant of $\epsilon = 0.1$) was used for binary data types. By default, the number of units in the encoding layer was set to the square root of the number of input features for each modality in the PPCG and TCGA experiments, and to 20 for synthetic data experiments.

Input noise was sampled using the following function:

```
1 def get_noise(n_samples, n_features):
2     return torch.sigmoid(torch.tensor(np.random.normal(0, 1, size=[n_samples
, n_features])))
```

The integration generator (Step 2 of the Multimodal Training Procedure: Step 2) was also a shallow fully-connected input-encoding-output network with batch normalization, tied weights, and non-negativity constraints. The output layer used a sigmoid activation function. For the encoding layer, to allow for an approximation of the IBP prior, we used the modified softsign activation function. By default, the number of units in the integration encoding later was set to 20. Input noise was sampled using the following function:

```
1 def get_noise_ibp(n_samples, n_features):
2     noise = torch.tensor(np.random.normal(0, 1, size=[n_samples, n_features
]))
3     noise_active = (1+((noise-beta)/(abs(noise-beta)+epsilon)))/2
4     return noise_active
```

where the smoothing constant epsilon was set to 0.5 and beta was obtained as:

```
1 probs = list(np.random.beta(0.5, 1, size=20))
2 probs.sort(reverse = True)
3 normal_sample = np.random.normal(0, 1, 100000)
4 beta = [np.percentile(normal_sample, 100 * (1 - p)) for p in probs]
```

to approximate an IBP prior with 20 LVs and expected feature activation 10. The multiomic generator (Steps 3 and 4 of the Multimodal Training Procedure) was then constructed as a combination of the corresponding single-modality generators and the integration generator. The tied weights were enforced using:

```
1 self.Decoder[0].weight = nn.Parameter(weight)
2 self.Encoder[0].weight = nn.Parameter(weight.transpose(0, 1))
```

for the respective encoder and decoder networks, and the non-negativity was achieved by including:

```

1 with torch.no_grad():
2     weights = self.Encoder[0].weight
3     self.Encoder[0].weight.copy_(weights.where(weights>=0, torch.zeros_like(
4         weights)))
5     weights = self.Encoder[0].weight
6     self.Encoder[0].weight.copy_(weights/torch.repeat_interleave(weights.sum(
7         axis=1), weights.shape[1]).reshape(weights.shape[0], weights.shape[1]))

```

in each forward pass of the respective encoders.

Critic

The architecture of the critic remained the same regardless of whether it was used for single-modality or multimodal experiments. The critic received a pair of inputs, (\mathbf{z}, \mathbf{x}) , where \mathbf{z} represents encoding or noise, and \mathbf{x} represents real or synthetic data. Initially, fully-connected layers with LeakyReLU activations were applied separately to \mathbf{z} and \mathbf{x} . The resulting outputs were then concatenated, followed by additional fully-connected layers with LeakyReLU activations for hidden layers and linear activations for the output, as detailed below:

```

1 self.Layers_z = nn.Sequential(
2     nn.Linear(z_size, x_hidden_2),
3     nn.LeakyReLU(leaky_relu_slope)
4 )
5 self.Layers_x = nn.Sequential(
6     nn.Linear(x_size, x_hidden_1),
7     nn.LeakyReLU(leaky_relu_slope),
8     nn.Linear(x_hidden_1, x_hidden_2),
9     nn.LeakyReLU(leaky_relu_slope)
10 )
11 self.Layers = nn.Sequential(
12     nn.Linear(2 * x_hidden_2, x_hidden_2),
13     nn.LeakyReLU(leaky_relu_slope),
14     nn.Linear(x_hidden_2, 1)
15 )

```

We did not use batch normalization or weight constraints in the critic. By default, we set $x_hidden_1 = 256$, $x_hidden_2 = 16$ and $leaky_relu = 0.25$. Parameters of the critic were updated 5 times using the same batch at each optimization iteration. The gradient penalty strength was set to 10.

Reconstruction Critic

The architecture of the reconstruction critic remained the same regardless of whether it was used for single-modality or multimodal experiments. The reconstruction critic received a concatenated pair of inputs, $(\mathbf{x}, \mathbf{x}')$, where \mathbf{x} represents the real data and \mathbf{x}' represents its reconstruction. Fully-connected layers with LeakyReLU activations for hidden layers and linear activations for the output were applied to the concatenated input, as described below:

```
1 self.Layers = nn.Sequential(  
2     nn.Linear(input_size, x_hidden_1),  
3     nn.LeakyReLU(leaky_relu_slope),  
4     nn.Linear(x_hidden_1, x_hidden_2),  
5     nn.LeakyReLU(leaky_relu_slope),  
6     nn.Linear(x_hidden_2, 1)  
7 )
```

We did not use batch normalization or weight constraints in the reconstruction critic. By default, we set `x_hidden_1 = 256`, `x_hidden_2 = 16` and `leaky_relu = 0.25`. Parameters of the reconstruction critic were updated 5 times using the same batch at each optimization iteration. The gradient penalty strength was set to 10.

Weight Initialization

Weights of all iCS-GAN components were initialized as below.

```
1 def weights_init(m):  
2     if isinstance(m, nn.Linear):  
3         torch.nn.init.normal_(m.weight, mean = 0.0, std = 0.02)  
4         torch.nn.init.constant_(m.bias, val = 0)
```

Clustering Regularization

The clustering regularization implementation was adapted from https://github.com/michael94/torch_DCEC. We choose the target distribution to be updated every 25 epochs by default and experimented with a number of

possible strengths for the weighting of the clustering loss ($\{5, 10, 15, 20, 25\}$ for all TCGA and synthetic data experiments). Where applied, after the initial iCS-GAN training, fine-tuning with clustering regularization proceeded as follows: the model was first trained for 10 epochs as a warmup. Afterward, the clustering loss was recurred at each epoch. If the clustering loss exceeded the previous maximum, the model was saved. However, if the highest clustering loss remained unchanged for 100 consecutive epochs, the training process was halted.

Survival Regularization

To implement the survival regularization component, we adapted the CQRNN network implementation from https://github.com/TeaPearce/Censored_Quantile_Regression_NN. Where applied, the survival regularization strength was set to 0.1. For experiments involving survival regularization, the survival network was pre-trained prior to iCS-GAN training. Hyperparameters were selected to maximize Harrell’s concordance index, using 5-fold cross-validation on the training set data only. The following hyperparameter search space was considered:

- hidden layers size: $\{5, 10, 25, 50, 100\}$,
- learning rate: $\{0.1, 0.01, 0.001\}$,
- batch size: $\{32, 64, 128\}$,
- number of epochs: $\{10, 25, 50, 100\}$.

Additionally, we set the number of quantiles to 10 and weight decay to $1e-4$. The Adam optimizer was used to train the survival network.

A.2 AE - Implementation and Hyperparameter Search

This section outlines the implementation and hyperparameter search of the Autoencoder used for iCS-GAN comparison experiments. The Autoencoder’s architecture was chosen to resemble that of the single-modality iCS-GAN generator, without modifications for binary variables or non-negativity constraints, as detailed below.

```
1 class Autoencoder(nn.Module):
2
3     def __init__(self, num_latent, num_observed, weight):
4
5         super(Autoencoder, self).__init__()
6
7         self.Encoder = nn.Sequential(
8             nn.Linear(num_observed, num_latent),
9             nn.BatchNorm1d(num_latent),
10            nn.Sigmoid()
11        )
12
13        self.Decoder = nn.Sequential(
14            nn.Linear(num_latent, num_observed),
15            nn.BatchNorm1d(num_observed),
16            nn.Sigmoid()
17        )
18
19        self.Decoder[0].weight = nn.Parameter(weight)
20        self.Encoder[0].weight = nn.Parameter(weight.transpose(0, 1))
21
22
23    def forward(self, x_real):
24
25        encoding = self.Encoder(x_real.float())
26        reconstruction = self.Decoder(encoding.float())
27
28        return encoding, reconstruction
```

Hyperparameters were selected to minimize the RMSE, using 5-fold cross-validation on the training set data only. The following hyperparameter search space was considered:

- batch size: {32, 64, 128}

- learning rate: {0.01, 0.001}
- L1 penalty strength: {0.01, 0.001}
- L2 penalty strength: {0.01, 0.001}

Each cross-validation fold was trained until the minimal validation RMSE was reached, with early stopping after 10 epochs of no improvement. The final number of training epochs was determined by averaging the numbers of training epochs across the 5 cross-validation folds corresponding to the selected optimal hyperparameters. The Adam optimizer was used to train the Autoencoder. We used the same weight initialization strategy as for iCS-GAN. We set the number of units in the encoding layer to the square root of the number of inputs.

A.3 nnAE - Implementation and Hyperparameter Search

This section outlines the implementation and hyperparameter search of the non-negative Autoencoder used for iCS-GAN comparison experiments. The nnAE’s architecture was chosen to resemble that of the single-modality iCS-GAN generator, without modifications for binary variables, as detailed below.

```

1 class nnAutoencoder(nn.Module):
2
3     def __init__(self, num_latent, num_observed, weight):
4
5         super(nnAutoencoder, self).__init__()
6
7         self.Encoder = nn.Sequential(
8             nn.Linear(num_observed, num_latent),
9             nn.BatchNorm1d(num_latent),
10            nn.Sigmoid()
11        )
12
13        self.Decoder = nn.Sequential(
14            nn.Linear(num_latent, num_observed),

```

```

15         nn.BatchNorm1d(num_observed),
16         nn.Sigmoid()
17     )
18
19     self.Decoder[0].weight = nn.Parameter(weight)
20     self.Encoder[0].weight = nn.Parameter(weight.transpose(0, 1))
21
22
23     def forward(self, x_real):
24
25         with torch.no_grad():
26
27             # Force non-negativity
28             weights = self.Encoder[0].weight
29             self.Encoder[0].weight.copy_(weights.where(weights>=0, torch.
zeros_like(weights)))
30             weights = self.Encoder[0].weight
31             self.Encoder[0].weight.copy_(weights/torch.repeat_interleave(
weights.sum(axis = 1), weights.shape[1]).reshape(weights.shape[0],
weights.shape[1]))
32
33             encoding = self.Encoder(x_real.float())
34             reconstruction = self.Decoder(encoding.float())
35
36         return encoding, reconstruction

```

Hyperparameters were selected to minimize the RMSE, using 5-fold cross-validation on the training set data only. The following hyperparameter search space was considered:

- batch size: {32, 64, 128}
- learning rate: {0.01, 0.001}
- L1 penalty strength: {0.01, 0.001}
- L2 penalty strength: {0.01, 0.001}

Each cross-validation fold was trained until the minimal validation RMSE was reached, with early stopping after 10 epochs of no improvement. The final number of training epochs was determined by averaging the numbers of training epochs across the 5 cross-validation folds corresponding to the

selected optimal hyperparameters. The Adam optimizer was used to train the non-negative Autoencoder. We used the same weight initialization strategy as for iCS-GAN. We set the number of units in the encoding layer to the square root of the number of inputs.

A.4 VAE - Implementation and Hyperparameter Search

This section outlines the implementation and hyperparameter search of the VAE used for iCS-GAN comparison experiments. The VAE’s architecture was chosen to resemble that of the single-modality iCS-GAN generator, without modifications for binary variables, tied weights or non-negativity constraints, with additional mean and standard deviation layers, as detailed below.

```
1 class VAE(nn.Module):
2
3     def __init__(self, num_latent, num_observed):
4
5         super(VAE, self).__init__()
6
7         self.Encoder = nn.Sequential(
8             nn.Linear(num_observed, num_latent),
9             nn.BatchNorm1d(num_latent),
10            nn.Sigmoid()
11        )
12
13        self.Encoder_Mu = nn.Sequential(
14            nn.Linear(num_latent, num_latent),
15            nn.BatchNorm1d(num_latent),
16        )
17
18        self.Encoder_Logvar = nn.Sequential(
19            nn.Linear(num_latent, num_latent),
20            nn.BatchNorm1d(num_latent),
21        )
22
23        self.Decoder = nn.Sequential(
24            nn.Linear(num_latent, num_observed),
25            nn.BatchNorm1d(num_observed),
26            nn.Sigmoid()
```

```

27     )
28
29     self.num_latent = num_latent
30
31     def reparameterize(self, mu, logvar):
32         std = torch.exp(0.5 * logvar)
33         eps = torch.randn_like(std)
34         return eps * std + mu
35
36     def sample(self, num_samples):
37
38         with torch.no_grad():
39             z = torch.randn(num_samples, self.num_latent)
40             samples = self.Decoder(z)
41
42         return samples
43
44     def forward(self, x_real):
45
46         hidden = self.Encoder(x_real.float())
47         mu = self.Encoder_Mu(hidden)
48         logvar = self.Encoder_Logvar(hidden)
49         encoding = self.reparameterize(mu, logvar)
50         reconstruction = self.Decoder(encoding.float())
51
52         return encoding, mu, logvar, reconstruction

```

Hyperparameters were selected to minimize the RMSE, using 5-fold cross-validation on the training set data only. The following hyperparameter search space was considered:

- batch size: {32, 64, 128}
- learning rate: {0.01, 0.001}
- L1 penalty strength: {0.01, 0.001}
- L2 penalty strength: {0.01, 0.001}

Each cross-validation fold was trained until the minimal validation RMSE was reached, with early stopping after 10 epochs of no improvement. The final number of training epochs was determined by averaging the numbers of training epochs across the 5 cross-validation folds corresponding to the

selected optimal hyperparameters. The Adam optimizer was used to train the VAE. We used the same weight initialization strategy as for iCS-GAN. We set the number of units in the encoding layer to the square root of the number of inputs. We used 0.001 for the weighting of the KL Divergence as compared to the reconstruction loss.

A.5 PPCG - Experimental Setup

This section outlines the experimental setup for applying iCS-GAN to the PPCG dataset in Chapter 5.

iCS-GAN was primarily used with its default parameters, as previously detailed in this appendix. To ensure optimal performance, we conducted a hyperparameter search focusing on the number of hidden units in the layers of both critics. The following values were considered: 4 and 16, 4 and 32, 8 and 64, 8 and 128, 16 and 256, 16 and 512. For each single modality experiment, we trained six model variants using the aforementioned critic layer sizes and assessed the stability of the results, as measured across 5 runs with different random seeds. Instead of choosing the model and the number of clusters that yielded the most stable results, we selected the configuration that provided the highest number of clusters (between 2 and 5) while maintaining stability, as measured by the ARI, with a threshold of no less than 0.65. This selected model was then fine-tuned for the chosen number of clusters with clustering regularization. We evaluated the following strengths for the clustering regularization: [0, 5, 10, 15, 20, 25], selecting the value that yielded the most stable clustering results, ensuring that it did not increase the initial correlation of the extracted LVs by more than 30% to avoid significantly disrupting the latent space.

A similar approach was taken for training the IBP integration layer, al-

though clustering regularization was not applied in this case. For the final integration model, we again considered the critic layer sizes of: 4 and 16, 4 and 32, 8 and 64, 8 and 128, 16 and 256, 16 and 512. Two clusters met the stability criteria, prompting us to initially post-process the model for two clusters while considering clustering regularization strengths of [0, 5, 10, 15, 20, 25]. We selected the strength that provided the most stable clustering results, without exceeding a 30% increase in the initial correlation of the extracted LVs. At this stage, we observed that the subtyping results were comparably stable for three clusters. Consequently, to further increase the granularity, we fine-tuned the model for three subtypes, following the same procedure to determine the clustering regularization strength, considering clustering regularization strengths of [0, 5, 6, 10, 15, 20, 25]. This final model accounted for our ultimate results.

With the exception of training the IBP layer, we applied survival regularization with a strength of 0.1 to each component. The 5-fold training set cross-validation performance of the survival network, as measured by Harrell’s concordance was:

- 0.71 for summary measurements,
- 0.57 for driver genes,
- 0.67 for CNAs,
- 0.78 for RNA,
- 0.77 for multimodal data.

KMeans clustering was applied only to the encodings of the complete samples from the training set, and predictions were then made for the remaining samples in the dataset. For samples with missing modalities, the

encodings were imputed using the KNN procedure outlined in Section 3.3. We used 15 nearest neighbours, which was determined to provide the highest stability in subtyping within the training set.

Finally, for the predictive RFC tests, we conducted a 5-fold cross-validation hyperparameter search. The following parameters were considered:

- number of estimators: 50, 100, 250, 500, 1000,
- max depth: 2, 4, 8, None,
- minimum samples split: 2, 4,
- class weight: None, ‘balanced’.

We selected the configuration that yielded the highest accuracy during the 5-fold cross-validation.

Appendix B

Supplementary Results

The tables in this appendix present the complete results for all experiments conducted in Sections 4.4 and 4.5, specifically the validation experiments using TCGA datasets. All reported values represent the averages from 5 runs of the model. The values in parentheses indicate the minimum and maximum values observed for each metric across these 5 runs.

Dataset	Modality	RMSE	WD-Rec	WD-Synth	WD-Latent	1-ARI
BRCA	Methylation	0.25 (0.24-0.25)	0.18 (0.17-0.18)	N/A	0.04 (0.03-0.05)	0.17 (0.02-0.25)
	mRNA	0.20 (0.20-0.20)	0.15 (0.15-0.16)	N/A	0.06 (0.05-0.09)	0.13 (0.10-0.20)
	CNA	0.31 (0.30-0.33)	0.34 (0.34-0.34)	N/A	0.12 (0.11-0.12)	0.16 (0.03-0.29)
KIRC	Methylation	0.24 (0.24-0.24)	0.18 (0.18-0.18)	N/A	0.06 (0.05-0.07)	0.18 (0.10-0.28)
	mRNA	0.27 (0.26-0.27)	0.22 (0.22-0.23)	N/A	0.08 (0.06-0.12)	0.42 (0.22-0.61)
	CNA	0.22 (0.21-0.23)	0.30 (0.29-0.30)	N/A	0.14 (0.13-0.14)	0.02 (0.00-0.06)
BLCA	Methylation	0.26 (0.25-0.26)	0.19 (0.18-0.20)	N/A	0.08 (0.06-0.09)	0.30 (0.21-0.43)
	mRNA	0.27 (0.27-0.28)	0.23 (0.22-0.23)	N/A	0.07 (0.07-0.08)	0.25 (0.19-0.36)
	CNA	0.41 (0.40-0.43)	0.35 (0.34-0.37)	N/A	0.10 (0.09-0.11)	0.32 (0.14-0.51)
COAD	Methylation	0.26 (0.26-0.27)	0.19 (0.18-0.20)	N/A	0.05 (0.04-0.05)	0.15 (0.04-0.27)
	mRNA	0.27 (0.26-0.27)	0.22 (0.22-0.22)	N/A	0.06 (0.05-0.06)	0.20 (0.14-0.30)
	CNA	0.31 (0.30-0.33)	0.36 (0.36-0.37)	N/A	0.11 (0.11-0.12)	0.09 (0.05-0.16)
HNSC	Methylation	0.25 (0.25-0.26)	0.19 (0.19-0.19)	N/A	0.05 (0.05-0.06)	0.19 (0.09-0.30)
	mRNA	0.27 (0.27-0.28)	0.23 (0.22-0.23)	N/A	0.06 (0.05-0.08)	0.15 (0.08-0.20)
	CNA	0.31 (0.30-0.32)	0.36 (0.35-0.36)	N/A	0.12 (0.11-0.12)	0.49 (0.15-0.64)

Table B.1: TCGA single modality results as obtained with a standard Autoencoder.

Dataset	Modality	RMSE	WD-Rec	WD-Synth	WD-Latent	1-ARI
BRCA	Methylation	0.25 (0.25-0.25)	0.17 (0.17-0.17)	N/A	0.03 (0.03-0.04)	0.22 (0.16-0.28)
	mRNA	0.21 (0.21-0.21)	0.17 (0.16-0.17)	N/A	0.05 (0.04-0.06)	0.40 (0.21-0.63)
	CNA	0.31 (0.30-0.33)	0.35 (0.35-0.36)	N/A	0.16 (0.15-0.16)	0.47 (0.15-0.72)
KIRC	Methylation	0.25 (0.25-0.25)	0.17 (0.17-0.18)	N/A	0.04 (0.04-0.05)	0.62 (0.52-0.72)
	mRNA	0.27 (0.27-0.27)	0.22 (0.22-0.22)	N/A	0.05 (0.04-0.05)	0.55 (0.34-0.80)
	CNA	0.22 (0.20-0.24)	0.29 (0.29-0.29)	N/A	0.15 (0.14-0.16)	0.38 (0.21-0.58)
BLCA	Methylation	0.26 (0.26-0.26)	0.18 (0.18-0.18)	N/A	0.04 (0.04-0.05)	0.40 (0.29-0.56)
	mRNA	0.28 (0.28-0.28)	0.23 (0.22-0.23)	N/A	0.05 (0.04-0.05)	0.50 (0.32-0.72)
	CNA	0.38 (0.35-0.40)	0.35 (0.34-0.35)	N/A	0.14 (0.13-0.14)	0.50 (0.27-0.72)
COAD	Methylation	0.27 (0.26-0.27)	0.19 (0.18-0.19)	N/A	0.04 (0.04-0.04)	0.23 (0.10-0.35)
	mRNA	0.27 (0.27-0.27)	0.22 (0.22-0.22)	N/A	0.04 (0.04-0.05)	0.56 (0.46-0.66)
	CNA	0.33 (0.32-0.34)	0.37 (0.37-0.37)	N/A	0.15 (0.14-0.15)	0.61 (0.46-0.73)
HNSC	Methylation	0.26 (0.26-0.26)	0.18 (0.18-0.18)	N/A	0.04 (0.04-0.05)	0.37 (0.26-0.50)
	mRNA	0.28 (0.28-0.28)	0.23 (0.22-0.23)	N/A	0.04 (0.04-0.05)	0.50 (0.29-0.61)
	CNA	0.31 (0.30-0.31)	0.36 (0.36-0.36)	N/A	0.16 (0.15-0.17)	0.47 (0.00-0.78)

Table B.2: TCGA single modality results as obtained with a non-negative Autoencoder.

Dataset	Modality	RMSE	WD-Rec	WD-Synth	WD-Latent	1-ARI
BRCA	Methylation	0.19 (0.19-0.19)	0.09 (0.09-0.09)	0.18 (0.18-0.18)	0.13 (0.12-0.14)	0.09 (0.06-0.11)
	mRNA	0.13 (0.13-0.13)	0.05 (0.05-0.06)	0.10 (0.10-0.11)	0.13 (0.12-0.13)	0.08 (0.00-0.14)
	CNA	0.29 (0.28-0.29)	0.11 (0.11-0.11)	0.17 (0.17-0.17)	0.13 (0.12-0.13)	0.32 (0.25-0.38)
KIRC	Methylation	0.19 (0.18-0.19)	0.08 (0.08-0.09)	0.16 (0.16-0.16)	0.13 (0.12-0.14)	0.28 (0.20-0.38)
	mRNA	0.15 (0.14-0.15)	0.06 (0.06-0.06)	0.12 (0.11-0.12)	0.13 (0.12-0.14)	0.29 (0.16-0.41)
	CNA	0.19 (0.19-0.19)	0.09 (0.08-0.11)	0.15 (0.14-0.15)	0.14 (0.11-0.17)	0.13 (0.06-0.25)
BLCA	Methylation	0.20 (0.20-0.20)	0.09 (0.09-0.10)	0.19 (0.19-0.19)	0.15 (0.14-0.17)	0.25 (0.14-0.33)
	mRNA	0.16 (0.16-0.16)	0.08 (0.07-0.08)	0.13 (0.12-0.13)	0.13 (0.13-0.13)	0.31 (0.17-0.42)
	CNA	0.35 (0.34-0.35)	0.16 (0.16-0.17)	0.24 (0.24-0.24)	0.15 (0.14-0.16)	0.29 (0.19-0.47)
COAD	Methylation	0.22 (0.22-0.22)	0.10 (0.09-0.11)	0.21 (0.21-0.21)	0.14 (0.13-0.15)	0.09 (0.04-0.17)
	mRNA	0.16 (0.16-0.16)	0.07 (0.07-0.08)	0.13 (0.13-0.13)	0.14 (0.13-0.15)	0.13 (0.11-0.17)
	CNA	0.28 (0.28-0.29)	0.11 (0.11-0.12)	0.18 (0.18-0.18)	0.15 (0.14-0.16)	0.26 (0.20-0.35)
HNSC	Methylation	0.19 (0.19-0.19)	0.09 (0.09-0.09)	0.18 (0.18-0.18)	0.14 (0.13-0.14)	0.21 (0.14-0.27)
	mRNA	0.16 (0.16-0.16)	0.07 (0.07-0.07)	0.13 (0.13-0.13)	0.13 (0.12-0.14)	0.14 (0.10-0.20)
	CNA	0.31 (0.30-0.31)	0.14 (0.13-0.14)	0.17 (0.17-0.18)	0.13 (0.12-0.16)	0.32 (0.26-0.45)

Table B.3: TCGA single modality results as obtained with VAE.

Dataset	Modality	RMSE	WD-Rec	WD-Synth	WD-Latent	1-ARI
BRCA	Methylation	0.23 (0.22-0.25)	0.09 (0.08-0.11)	0.13 (0.09-0.15)	0.08 (0.06-0.10)	0.41 (0.16-0.61)
	mRNA	0.20 (0.18-0.21)	0.08 (0.07-0.09)	0.11 (0.10-0.12)	0.11 (0.08-0.14)	0.54 (0.31-0.71)
	CNA	0.31 (0.30-0.35)	0.10 (0.10-0.10)	0.11 (0.10-0.11)	0.06 (0.05-0.07)	0.15 (0.11-0.22)
KIRC	Methylation	0.24 (0.22-0.26)	0.09 (0.08-0.11)	0.14 (0.11-0.16)	0.09 (0.08-0.10)	0.51 (0.18-0.73)
	mRNA	0.28 (0.26-0.29)	0.14 (0.13-0.15)	0.18 (0.17-0.20)	0.08 (0.07-0.09)	0.42 (0.06-0.77)
	CNA	0.21 (0.21-0.23)	0.06 (0.06-0.07)	0.11 (0.10-0.11)	0.12 (0.11-0.13)	0.19 (0.06-0.33)
BLCA	Methylation	0.24 (0.24-0.25)	0.11 (0.11-0.12)	0.17 (0.15-0.18)	0.10 (0.08-0.13)	0.20 (0.08-0.33)
	mRNA	0.28 (0.26-0.31)	0.13 (0.11-0.14)	0.20 (0.19-0.20)	0.09 (0.07-0.12)	0.37 (0.20-0.54)
	CNA	0.38 (0.34-0.41)	0.13 (0.11-0.15)	0.13 (0.13-0.13)	0.07 (0.05-0.09)	0.29 (0.05-0.47)
COAD	Methylation	0.26 (0.25-0.27)	0.11 (0.10-0.11)	0.16 (0.15-0.17)	0.09 (0.07-0.10)	0.60 (0.23-0.80)
	mRNA	0.30 (0.29-0.31)	0.15 (0.14-0.16)	0.19 (0.18-0.20)	0.05 (0.04-0.06)	0.28 (0.00-0.44)
	CNA	0.31 (0.30-0.32)	0.09 (0.08-0.09)	0.11 (0.10-0.11)	0.05 (0.04-0.06)	0.13 (0.07-0.20)
HNSC	Methylation	0.24 (0.23-0.25)	0.12 (0.11-0.12)	0.15 (0.14-0.16)	0.10 (0.09-0.11)	0.68 (0.32-0.89)
	mRNA	0.31 (0.29-0.31)	0.16 (0.16-0.17)	0.19 (0.19-0.20)	0.04 (0.04-0.05)	0.14 (0.04-0.19)
	CNA	0.32 (0.32-0.33)	0.09 (0.09-0.10)	0.11 (0.10-0.11)	0.06 (0.05-0.06)	0.29 (0.19-0.45)

Table B.4: TCGA single modality results as obtained with iCS-GAN without the Wasserstein training procedure.

Dataset	Modality	RMSE	WD-Rec	WD-Synth	WD-Latent	1-ARI
BRCA	Methylation	0.21 (0.21-0.22)	0.08 (0.07-0.08)	0.07 (0.06-0.08)	0.05 (0.04-0.05)	0.68 (0.56-0.77)
	mRNA	0.17 (0.15-0.17)	0.06 (0.04-0.07)	0.05 (0.03-0.06)	0.05 (0.04-0.06)	0.68 (0.64-0.73)
	CNA	0.31 (0.31-0.31)	0.08 (0.08-0.08)	0.08 (0.08-0.09)	0.04 (0.04-0.05)	0.42 (0.37-0.49)
KIRC	Methylation	0.21 (0.21-0.22)	0.08 (0.08-0.08)	0.08 (0.08-0.08)	0.05 (0.04-0.06)	0.31 (0.20-0.41)
	mRNA	0.18 (0.17-0.18)	0.06 (0.05-0.07)	0.06 (0.06-0.07)	0.04 (0.04-0.05)	0.71 (0.61-0.80)
	CNA	0.20 (0.18-0.22)	0.06 (0.05-0.06)	0.09 (0.07-0.11)	0.06 (0.06-0.07)	0.27 (0.17-0.39)
BLCA	Methylation	0.22 (0.22-0.23)	0.12 (0.11-0.13)	0.09 (0.08-0.10)	0.10 (0.08-0.11)	0.37 (0.19-0.67)
	mRNA	0.21 (0.21-0.22)	0.12 (0.10-0.14)	0.11 (0.10-0.12)	0.06 (0.05-0.09)	0.82 (0.73-0.92)
	CNA	0.36 (0.35-0.37)	0.11 (0.10-0.12)	0.11 (0.10-0.12)	0.06 (0.04-0.07)	0.56 (0.43-0.64)
COAD	Methylation	0.24 (0.23-0.24)	0.09 (0.07-0.11)	0.09 (0.07-0.11)	0.04 (0.03-0.05)	0.83 (0.74-0.91)
	mRNA	0.19 (0.18-0.21)	0.08 (0.06-0.11)	0.07 (0.05-0.11)	0.05 (0.04-0.05)	0.78 (0.71-0.82)
	CNA	0.29 (0.28-0.30)	0.08 (0.08-0.08)	0.09 (0.09-0.10)	0.04 (0.03-0.05)	0.57 (0.47-0.68)
HNSC	Methylation	0.23 (0.22-0.23)	0.09 (0.08-0.11)	0.09 (0.08-0.09)	0.05 (0.04-0.06)	0.71 (0.45-0.89)
	mRNA	0.20 (0.18-0.22)	0.08 (0.06-0.12)	0.08 (0.06-0.12)	0.05 (0.04-0.06)	0.46 (0.34-0.57)
	CNA	0.32 (0.31-0.32)	0.09 (0.09-0.09)	0.09 (0.09-0.09)	0.04 (0.04-0.04)	0.60 (0.53-0.66)

Table B.5: TCGA single modality results as obtained with iCS-GAN without interpretability constraints.

Dataset	Modality	RMSE	WD-Rec	WD-Synth	WD-Latent	1-ARI
BRCA	Methylation	0.24 (0.23-0.24)	0.09 (0.07-0.09)	0.11 (0.10-0.12)	0.07 (0.05-0.08)	0.17 (0.10-0.25)
	mRNA	0.17 (0.16-0.17)	0.08 (0.06-0.09)	0.06 (0.06-0.07)	0.08 (0.05-0.11)	0.04 (0.02-0.08)
	CNA	0.35 (0.32-0.42)	0.14 (0.11-0.19)	0.14 (0.13-0.14)	0.08 (0.06-0.09)	0.16 (0.04-0.27)
KIRC	Methylation	0.22 (0.21-0.22)	0.06 (0.06-0.07)	0.09 (0.09-0.10)	0.05 (0.04-0.06)	0.27 (0.11-0.37)
	mRNA	0.18 (0.17-0.19)	0.09 (0.07-0.10)	0.09 (0.07-0.10)	0.05 (0.04-0.06)	0.16 (0.08-0.27)
	CNA	0.22 (0.20-0.25)	0.11 (0.08-0.16)	0.18 (0.14-0.26)	0.12 (0.11-0.13)	0.19 (0.07-0.28)
BLCA	Methylation	0.25 (0.24-0.26)	0.12 (0.10-0.13)	0.12 (0.11-0.12)	0.07 (0.06-0.08)	0.30 (0.22-0.36)
	mRNA	0.23 (0.21-0.26)	0.14 (0.12-0.16)	0.12 (0.11-0.14)	0.06 (0.06-0.08)	0.18 (0.10-0.27)
	CNA	0.45 (0.41-0.49)	0.27 (0.26-0.29)	0.35 (0.32-0.37)	0.11 (0.10-0.14)	0.56 (0.48-0.65)
COAD	Methylation	0.24 (0.23-0.25)	0.07 (0.06-0.08)	0.13 (0.12-0.13)	0.04 (0.04-0.05)	0.25 (0.14-0.33)
	mRNA	0.20 (0.19-0.22)	0.10 (0.09-0.11)	0.10 (0.09-0.11)	0.04 (0.04-0.05)	0.10 (0.04-0.14)
	CNA	0.32 (0.31-0.33)	0.15 (0.13-0.17)	0.16 (0.15-0.17)	0.05 (0.04-0.05)	0.06 (0.00-0.10)
HNSC	Methylation	0.23 (0.22-0.24)	0.07 (0.06-0.08)	0.11 (0.11-0.12)	0.05 (0.04-0.06)	0.25 (0.11-0.39)
	mRNA	0.20 (0.19-0.21)	0.09 (0.08-0.10)	0.10 (0.09-0.11)	0.05 (0.05-0.06)	0.16 (0.05-0.25)
	CNA	0.33 (0.31-0.35)	0.15 (0.13-0.17)	0.16 (0.15-0.17)	0.06 (0.04-0.09)	0.17 (0.04-0.26)

Table B.6: TCGA single modality results as obtained with iCS-GAN without modifications for binary variables.

Dataset	Modality	RMSE	WD-Rec	WD-Synth	WD-Latent	1-ARI
BRCA	Methylation	0.21 (0.21-0.22)	0.08 (0.07-0.09)	0.11 (0.11-0.12)	0.04 (0.04-0.05)	0.23 (0.17-0.28)
	mRNA	0.16 (0.15-0.17)	0.07 (0.07-0.07)	0.06 (0.05-0.07)	0.07 (0.06-0.08)	0.13 (0.06-0.21)
	CNA	0.30 (0.29-0.31)	0.08 (0.08-0.09)	0.09 (0.08-0.09)	0.06 (0.05-0.07)	0.23 (0.13-0.39)
KIRC	Methylation	0.22 (0.21-0.22)	0.09 (0.08-0.10)	0.11 (0.10-0.12)	0.05 (0.05-0.05)	0.49 (0.39-0.58)
	mRNA	0.19 (0.17-0.20)	0.10 (0.07-0.11)	0.10 (0.07-0.12)	0.06 (0.05-0.07)	0.28 (0.24-0.38)
	CNA	0.22 (0.20-0.23)	0.06 (0.06-0.06)	0.08 (0.08-0.10)	0.10 (0.09-0.12)	0.19 (0.05-0.35)
BLCA	Methylation	0.22 (0.22-0.22)	0.10 (0.09-0.10)	0.12 (0.12-0.13)	0.05 (0.05-0.06)	0.26 (0.13-0.37)
	mRNA	0.22 (0.20-0.23)	0.13 (0.11-0.14)	0.13 (0.10-0.15)	0.08 (0.06-0.09)	0.32 (0.23-0.47)
	CNA	0.35 (0.34-0.37)	0.11 (0.10-0.11)	0.13 (0.12-0.13)	0.05 (0.05-0.05)	0.34 (0.19-0.51)
COAD	Methylation	0.23 (0.23-0.23)	0.10 (0.09-0.12)	0.13 (0.13-0.14)	0.04 (0.04-0.05)	0.29 (0.04-0.59)
	mRNA	0.21 (0.19-0.23)	0.10 (0.09-0.12)	0.10 (0.09-0.13)	0.06 (0.05-0.06)	0.18 (0.14-0.27)
	CNA	0.31 (0.30-0.31)	0.09 (0.08-0.09)	0.10 (0.09-0.11)	0.04 (0.04-0.05)	0.22 (0.04-0.33)
HNSC	Methylation	0.22 (0.21-0.22)	0.10 (0.09-0.10)	0.12 (0.11-0.13)	0.05 (0.05-0.06)	0.30 (0.19-0.42)
	mRNA	0.21 (0.19-0.23)	0.11 (0.10-0.13)	0.11 (0.10-0.14)	0.06 (0.06-0.07)	0.28 (0.18-0.38)
	CNA	0.33 (0.32-0.33)	0.09 (0.09-0.09)	0.10 (0.09-0.10)	0.05 (0.04-0.05)	0.30 (0.11-0.53)

Table B.7: TCGA single modality results as obtained with iCS-GAN without clustering regularization.

Dataset	RMSE	WD-Rec	WD-Synth	WD-Latent	1-ARI
BRCA	0.23 (0.22-0.24)	0.09 (0.09-0.10)	0.09 (0.08-0.10)	0.07 (0.05-0.09)	0.50 (0.37-0.59)
KIRC	0.23 (0.22-0.23)	0.10 (0.09-0.11)	0.11 (0.08-0.14)	0.08 (0.07-0.10)	0.54 (0.45-0.63)
BLCA	0.29 (0.28-0.30)	0.13 (0.12-0.13)	0.13 (0.13-0.15)	0.06 (0.05-0.07)	0.37 (0.14-0.54)
COAD	0.27 (0.27-0.28)	0.11 (0.10-0.13)	0.11 (0.10-0.12)	0.06 (0.05-0.08)	0.40 (0.27-0.56)
HNSC	0.25 (0.24-0.26)	0.10 (0.10-0.11)	0.11 (0.10-0.12)	0.06 (0.05-0.07)	0.52 (0.45-0.59)

Table B.8: TCGA multimodal results as obtained with iCS-GAN without independent layers.

Dataset	RMSE	WD-Rec	WD-Synth	WD-Latent	1-ARI
BRCA	0.22 (0.21-0.24)	0.10 (0.09-0.12)	0.09 (0.08-0.09)	0.09 (0.05-0.12)	0.18 (0.14-0.25)
KIRC	0.22 (0.22-0.23)	0.11 (0.10-0.11)	0.11 (0.10-0.12)	0.08 (0.06-0.08)	0.37 (0.20-0.50)
BLCA	0.26 (0.25-0.28)	0.13 (0.12-0.16)	0.15 (0.14-0.17)	0.08 (0.06-0.09)	0.64 (0.49-0.83)
COAD	0.26 (0.25-0.26)	0.11 (0.10-0.12)	0.11 (0.11-0.12)	0.06 (0.05-0.08)	0.40 (0.23-0.54)
HNSC	0.26 (0.25-0.26)	0.11 (0.10-0.11)	0.12 (0.10-0.13)	0.07 (0.05-0.08)	0.42 (0.24-0.51)

Table B.9: TCGA multimodal results as obtained with iCS-GAN without layer-wise pre-training.

Dataset	RMSE	WD-Rec	WD-Synth	WD-Latent	1-ARI
BRCA	0.20 (0.20-0.20)	0.09 (0.09-0.10)	0.09 (0.09-0.09)	0.05 (0.04-0.06)	0.53 (0.04-0.98)
KIRC	0.21 (0.20-0.21)	0.09 (0.08-0.10)	0.10 (0.09-0.10)	0.05 (0.04-0.05)	0.39 (0.20-0.49)
BLCA	0.25 (0.23-0.26)	0.12 (0.11-0.12)	0.13 (0.12-0.14)	0.05 (0.04-0.07)	0.68 (0.33-0.88)
COAD	0.24 (0.24-0.24)	0.10 (0.09-0.11)	0.11 (0.10-0.12)	0.04 (0.04-0.05)	0.44 (0.24-0.68)
HNSC	0.22 (0.22-0.23)	0.10 (0.09-0.11)	0.10 (0.09-0.11)	0.05 (0.04-0.06)	0.28 (0.11-0.47)

Table B.10: TCGA multimodal results as obtained with iCS-GAN without clustering regularization.

Dataset	RMSE	WD-Rec	WD-Synth	WD-Latent	1-ARI
BRCA	0.21 (0.20-0.22)	0.10 (0.08-0.12)	0.09 (0.09-0.09)	0.06 (0.05-0.09)	0.41 (0.04-0.80)
KIRC	0.21 (0.20-0.21)	0.09 (0.08-0.09)	0.09 (0.09-0.10)	0.05 (0.05-0.06)	0.22 (0.13-0.33)
BLCA	0.26 (0.23-0.28)	0.12 (0.11-0.13)	0.13 (0.12-0.14)	0.06 (0.04-0.07)	0.67 (0.41-0.79)
COAD	0.24 (0.24-0.25)	0.10 (0.09-0.11)	0.11 (0.11-0.12)	0.05 (0.05-0.06)	0.33 (0.23-0.46)
HNSC	0.23 (0.22-0.23)	0.10 (0.09-0.11)	0.10 (0.10-0.11)	0.06 (0.05-0.07)	0.11 (0.04-0.19)

Table B.11: TCGA multimodal results as obtained with iCS-GAN without clustering regularization on independent layers only.

Dataset	RMSE	WD-Rec	WD-Synth	WD-Latent	1-ARI
BRCA	0.21 (0.20-0.22)	0.10 (0.07-0.11)	0.08 (0.08-0.09)	0.08 (0.05-0.10)	0.07 (0.02-0.12)
KIRC	0.21 (0.20-0.21)	0.09 (0.08-0.09)	0.09 (0.08-0.10)	0.05 (0.05-0.06)	0.05 (0.00-0.08)
BLCA	0.23 (0.22-0.25)	0.12 (0.12-0.13)	0.12 (0.11-0.13)	0.07 (0.06-0.08)	0.36 (0.22-0.55)
COAD	0.23 (0.23-0.23)	0.10 (0.09-0.10)	0.10 (0.09-0.10)	0.05 (0.05-0.06)	0.20 (0.10-0.27)
HNSC	0.22 (0.22-0.22)	0.09 (0.09-0.10)	0.09 (0.08-0.09)	0.06 (0.05-0.07)	0.28 (0.15-0.45)

Table B.12: TCGA multimodal results as obtained with iCS-GAN without the Indian Buffet Process Prior.