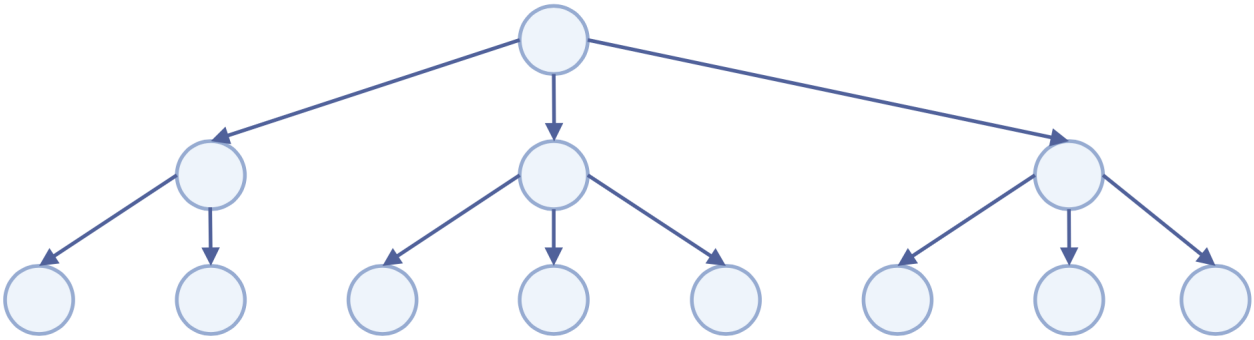
TÉCNICO LISBOA



Optimised Combination of Distance Metrics for Labelled Trees Using Survival Data

Laura Bispo Quintas

Thesis to obtain the Master of Science Degree in

Biomedical Engineering

Supervisors: Prof. Susana de Almeida Mendes Vinga Martins
Prof. Niko Beerenwinkel

Examination Committee

Chairperson: Prof. Rita Homem de Gouveia Costanzo Nunes
Supervisor: Prof. Susana de Almeida Mendes Vinga Martins
Member of the Committee: Prof. Alexandre Paulo Lourenço Francisco

November 2023

Declaration

I declare that this document is an original work of my own authorship and that it fulfills all the requirements of the Code of Conduct and Good Practices of the Universidade de Lisboa.

Acknowledgments

I would like to express my deepest gratitude to my supervisors, Prof. Susana Vinga and Prof. Niko Beerenwinkel. Prof. Susana, your unwavering support, even from a distance, has been invaluable. Prof. Niko, I am incredibly grateful for the incredible opportunity you provided, allowing me to undertake this project abroad.

My heartfelt thanks go out to the entire computational biology group in Basel, who embraced me with open arms and made my time there enriching and enjoyable. A special thank you to Monica and Xiang, your continuous support and encouragement significantly enhanced my Erasmus experience, providing me with a wealth of knowledge and good memories.

I am also thankful for the numerous opportunities that have come my way during these months, particularly for allowing me to participate in the OLISSIPO events. To all the members of the OLISSIPO project, I appreciate your warm welcome, availability, and the valuable insights you shared.

My sincere gratitude is extended to my family, who were instrumental in motivating me to seize new opportunities and settle down in a new city. Over the past five years, you have been a great comfort to me because of your unending patience and willingness to listen to my longings for the sunshine of the Algarve. When I needed comforting words and support the most, you were my rock. It has been immensely valuable for me on my journey that you have given me a haven full of inspiration and confidence in my abilities. This work is the result of your influence, made possible by the core principles you helped me develop and which have shaped my character and path. My parents in particular deserve special thanks for their unwavering encouragement and support during this journey.

To all my friends and colleagues at IST, the advice we received at the beginning—that you can't go through this journey alone—has proven to be true and even if it was possible it wouldn't have been this enjoyable and memorable. A big thank you to ACERIUS for continued friendship even after my 'betrayal' when I changed courses. And to my Doras, who embraced a confused second-phase student and turned her into a friend, I am eternally grateful.

To each and every one of you who has been part of my journey, filled with highs and

lows—Thank you.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 951970 (project OLISSIPO).

Preface

The work presented in this thesis was performed at ETH - Department of Biosystems Science and Engineering in Basel for 5 months during the period March-October 2023, under the supervision of Prof. Niko Beerenwinkel with the support of Xiang Ge Luo and Monica Dragan within the frame of the *Erasmus+* programme and project OLISSIPO - *Fostering Computational Biology Research and Innovation in Lisbon*. The thesis was supervised at INESC-ID and Instituto Superior Técnico by Prof. Susana Vinga.

Abstract

Cancer, characterized by uncontrolled cell growth and DNA mutations, presents significant treatment challenges due to intra-tumor heterogeneity, which can be represented through mutation trees. Distance metrics, capturing diverse aspects of these trees, have been used for patient clustering and associating with clinical data to identify cancer subgroups. However, a consensus on a standard metric is yet to be reached, prompting investigations into metric combinations to leverage their combined strengths. This study critically assesses optimised combinations of distance metrics, interpreting mutation trees' characteristics across Breast Cancer and Acute Myeloid Leukaemia data sets, aiming to link patient clustering with clinical outcomes, especially survival. We employ a comprehensive approach, encompassing all available distance metrics for clustering tumor trees and integrating clinical and genotypic information using the Cox model. Our findings indicate that while these metrics offer detailed insights into tumor evolution, their predictive power for survival, individually or collectively, is not significantly superior to using clinical and genotypic covariates alone. Our results effectively challenge the notion of a single distance metric for patient clustering over a variety of cancer types by showing the inconsistent way these metrics perform across various data sets. This study underscores the complexity of cancer genomics and the challenges in deriving useful information for personalised patient prognosis. It advocates for future research with standardised data sets and diverse cancer types to further explore the potential of distance metrics and to devise reliable, tailored cancer treatment strategies.

Keywords

Cancer; Mutation Trees; Distance Metrics; Hierarchical Clustering; Survival Analysis.

Resumo

O cancro é caracterizado pelo crescimento descontrolado de células e mutações no ADN, criando desafios no tratamento devido à sua heterogeneidade intra-tumoral, que pode ser representada por árvores de mutação. Utilizaram-se métricas de distância para capturar diversos aspetos dessas árvores, a fim de agrupar pacientes e estabelecer correlações com dados clínicos, permitindo a identificação de subgrupos de cancro. Contudo, a ausência de uma métrica padrão levou estudos recentes a explorar combinações de métricas para maximizar os seus pontos fortes. Este estudo propõe combinações otimizadas dessas métricas em casos de cancro da mama e leucemia mieloide aguda, procurando correlacionar agrupamentos de pacientes a resultados clínicos, focando na sobrevivência. Empregamos uma abordagem abrangente, agrupando árvores tumorais com todas as métricas disponíveis e integrando dados clínicos e genotípicos no modelo de Cox. Os resultados indicam que, apesar de fornecerem informações detalhadas sobre a evolução tumoral, as métricas não superam significativamente o uso isolado de covariáveis clínicas e genotípicas na predição da sobrevivência. Os nossos resultados desafiam efetivamente a noção de uma única métrica de distância para o agrupamento de doentes mostrando o seu desempenho inconsistente em conjunto de dados de diferentes tipos de cancro. Este estudo ressalta a complexidade genômica do cancro e os desafios para obter prognósticos personalizados, sugerindo pesquisas futuras com dados padronizados e diferentes tipos de cancro para explorar o verdadeiro potencial do uso das métricas de distância e desenvolver estratégias de tratamento mais eficazes e personalizadas.

Palavras Chave

Cancro; Árvores de Mutação; Métricas de Distância; Agrupamento Hierárquico; Análise de Sobrevivência.

Contents

List of Figures

List of Tables

Acronyms

AML	Acute Myeloid Leukaemia
MLTD	Multi-Labeled Tree Dissimilarity
CASet	Common Ancestor Set distance
DNA	Deoxyribonucleic Acid
DISC	Distinctly Inherited Set Comparison distance
MP3	Triplet-based Similarity Score
ISA	Infinite Sites Assumption
BD	Bourque Distance
kBD	k-Bourque Distance
1BD	1-Bourque Distance
2BD	2-Bourque Distance
MSA	Metaheuristic search algorithm
DE	Differential Evolution
LLR	Likelihood Ratio Test
C-Index	Concordance Index
maxCCF	Maximum Cancer Cell Fraction
PCD	Parent-Child Distance
AD	Ancestor-Descendant Distance
PD	Path Distance
CD	Clonal Distance
LOWESS	Locally Weighted Scatterplot Smoothing
WSS	Weighted Silhouette Score

AIC Akaike Information Criterion

pAIC partial Akaike Information Criterion

PI Prognostic Index

pLog-Likelihood partial Log-Likelihood

RF Robinson–Foulds

1

Introduction

Contents

1.1 Motivation

Cancer, a disease marked by unregulated cell growth and consequential genetic mutations, presents profound complexity characterised by intra-tumor heterogeneity. This diversity within tumors significantly complicates treatment responses, as dominant cancer subclones are often the primary targets of conventional therapies. However, the post-remission phase can witness the emergence of suppressed or newly evolved, treatment-resistant subclones, challenging the durability of therapeutic success [4–6].

The chronological sequence of mutations, an important factor in disease progression and treatment efficacy, brings the need for a detailed exploration of mutation trees. These trees serve as representations of the mutation sequences and evolutionary interrelationships among various subclones, offering critical insights into the tumor's evolutionary path and the emergence of drug resistance [7–9].

The use of distance metrics to analyse mutation tree has drawn interest in this complex landscape because it has the ability to reveal subtle relationships and patterns that are essential to understand the evolution of tumors. These metrics, which each highlight different aspects of the trees (e.g., label variations, structural partitions, and topological differences), naturally lead to different scores, highlighting the absence of a standard metric that can be applied to all situations and adequately capture the heterogeneity of different types of cancer [8–12].

In the past, these metrics of distance have been used in clustering patients into groups, creating associations with concrete clinical information. This utility led to an examination of the validity of these correlations using real-world data, with a particular emphasis on survival outcomes. Interestingly, Ciccolella et al. [10] created a combination of distinct metric versions, maximising their combined strengths — a method that caused the hypothesis that different metrics could produce a comparable effect. We hypothesise that this combination may be especially effective in capturing cancer-specific nuances that individual metrics would miss.

Our research aims to optimise this combination of metrics, with survival data being a crucial component. The objective is not merely to cluster data accurately; rather, it is to establish a connection between these groups and important clinical outcomes, specifically survival, both at the individual and combined metrics levels.

1.2 Objectives and Contributions

This study aimed to assess the predictive power of distance metrics for patient survival in relation to different cancer types, specifically Acute Myeloid Leukaemia (AML) and Breast Cancer.

The primary objective was to determine whether these metrics—alone or in combination—could

potentially enhance prognostic predictions beyond the standard bounds of genotype and clinical data, while also validate an association between the clusters derived from the metrics and the pertinent clinical data.

To achieve these objectives, we embarked on a detailed analysis using two real data sets, employing survival data to optimise the combination of metrics through two distinct strategies. This process involved patient stratification based on tree information derived from their Cox models, followed by a comparative analysis with models exclusively based on genotype and clinical data. It was also included suggestions for future directions in research to enhance and optimise the use of genomic data in individualised cancer treatment strategies.

The work produced during the realisation of this thesis was presented multiple times. First, it was presented in front of the entire lab during a group meeting organised by the computational biology group at ETH-DBSSE. Furthermore, during the OLISSIPO annual meeting, the work was presented in a seminar at EMBL in Heidelberg. On both occasions, a lot of ideas were shared and insightful criticism was provided.

1.3 Thesis Outline

The four important chapters in which this work's structure is divided are 'Background' (Chapter 2) which covers the theoretical foundations, key concepts, and methods associated with tumour trees, distance metrics, and related subjects. A thorough examination of the techniques employed is given in the 'Methodology' (Chapter 3), which covers the particular procedures and formulas used in the investigation. The results of the analysis are presented in the section titled 'Results and Discussion' (Chapter 4). These insights and interpretations highlight the implications of these distance metrics in two real datasets. The study concludes with the 'Conclusion' (Chapter 5), which summarises the major conclusions and offers suggestions for additional research.

2

Background

Contents

2.1 Cancer and Tumor Trees

2.1.1 Genetics of Cancer and its Evolution

Cancer is a complex disease characterised by the uncontrolled growth and dissemination of cells throughout the body, which is composed of trillions of cells. Typically, human cells grow, multiply through cell division, and replace old or damaged cells in a well-regulated manner. However, when this orderly process is disrupted, abnormal cells may proliferate when they shouldn't, potentially forming tumors [4].

Tumors can be benign or malignant, with benign tumors being non-cancerous and non-threatening. Malignant tumors, on the other hand, are cancerous and can invade tissues and spread to distant organs through metastasis. These tumors can disrupt normal functions and require aggressive treatment to effectively manage the disease [5].

Over the past two decades, cancer researchers have amassed a vast and detailed dataset, illustrating that cancer is a disease marked by dynamic changes in the genome, which eventually display various acknowledged cancer hallmarks [13]. The clonal theory, proposing that a tumor arises from an evolutionary process fueled by the accumulation of genetic mutations, stands out as a pivotal concept in comprehending cancer biology [8]. These genomic modifications are linked to genetic processes that disrupt important biochemical processes that influence cell proliferation, survival, and other cancer-related characteristics [14].

Tumour development is linked to a variety of evolutionary patterns, such as branched evolution, which is associated to widespread intra-tumor heterogeneity, reflecting diverse subclones which are cells with different mutations within a single tumour (Fig. 2.1 (a)), significantly influencing treatment strategies and tumor progression understanding [15]. A tumor typically originates from a single founder cell, which gains a growth advantage and evades the immune response. The clone from this cell expands, and descendant cells develop into subclones by acquiring additional somatic mutations, competing for resources within the tumor environment until they are out-competed by new subclones (Fig. 2.1 (b)) [14, 16].

The overall survival of patients with malignant cancer is significantly reduced as a result of tumour heterogeneity, which has been linked to unfavourable prognostic implications and outcomes. It is determined that this heterogeneity, a mosaic of cancer cells with various characteristics and sensitivity to treatments, is a significant factor in post-treatment relapse. At the time of diagnosis, therapeutic approaches frequently target the dominant subclone, and after its remission, a proliferation of previously restrained subclones or the creation of novel resistant ones may occur. In order to provide more effective, individualised cancer medicines, a thorough understanding of the genetic variety and evolutionary paths of tumours is essential [7, 17–19].

2.1.2 Labelled Tumor Trees

This intra-tumor heterogeneity can be represented using a labelled tumor tree, which is a rooted tree structure that visualises the evolutionary relationships between different subclones within a tumor [8].

A tree, defined as a connected, undirected graph devoid of cycles, when specified with a root, becomes a rooted tree, wherein one node is designated as the root, and all subsequent nodes adhere to a unique directed path towards it, with edges oriented away from the root. [20]. Different tumour tree representations, including phylogenetic and mutation trees, have been investigated in oncology.

Using a phylogenetic tree, which is based on a finite set X and has leaves that are specifically labelled with components of X - the taxon set - to represent the actual tumor samples under investigation, it is possible to see how different tumor subclones have evolved over time. Internal nodes, on the other hand, identify particular mutations within a subclone, and their labels are essential to comprehending the tumor's evolutionary history (Fig. 2.1 (c)). It is essential to understand that the taxon set for leaves and the label set for internal nodes reflect distinct concepts and are therefore incompatible, requiring a separate analysis [9].

A mutation tree, which is a specific instance of a clonal tree, is a kind of rooted, directed tree in which each vertex is labelled with one or more mutations (Fig. 2.1 (d)). In contrast, in clonal trees, each vertex represents a unique clone, and the label is made up of the mutations of the ascendants of the vertex and the new mutation added. Consequently, a mutation tree is a clonal tree with the finest granularity. [8]. These mutations can be any type of genomic variant, such as single nucleotide variants or copy number alterations. Each vertex in the tree represents a distinct tumor clone (or population) that existed at some point during the tumor's evolution. The mutations that are labeled on a vertex indicate the clone in which the mutation first appeared. As a result, the complete set of mutations in any given clone, represented by a vertex, is the set of mutations that label all of the vertices on the path from the root to the same vertex for a mutation tree [12].

These trees provide an intricate perspective of the tumor's entire temporal history, enabling researchers to acquire a thorough understanding of the genetic mutations transpiring throughout the tumor's evolution [11–13]. Similar to the significance of the mutations' presence, the order in which they occur is important in determining tumour risk and influencing the clinical direction. A notable case applies to patients diagnosed with myeloproliferative neoplasms, wherein the clinical features, responses to targeted therapies, and the behaviour of stem and progenitor cells are significantly influenced by the acquisition sequence of mutations in JAK2 and TET2. In particular, it has been demonstrated that the onset of a mutation in TET2 before JAK2

impairs JAK2's capacity to up-regulate the proliferative program, leading to modified clinical presentations, variable thrombosis risks, and unique responses to in vitro treatments. This highlights the crucial role of mutation order in determining tumour characteristics and treatment response [7].

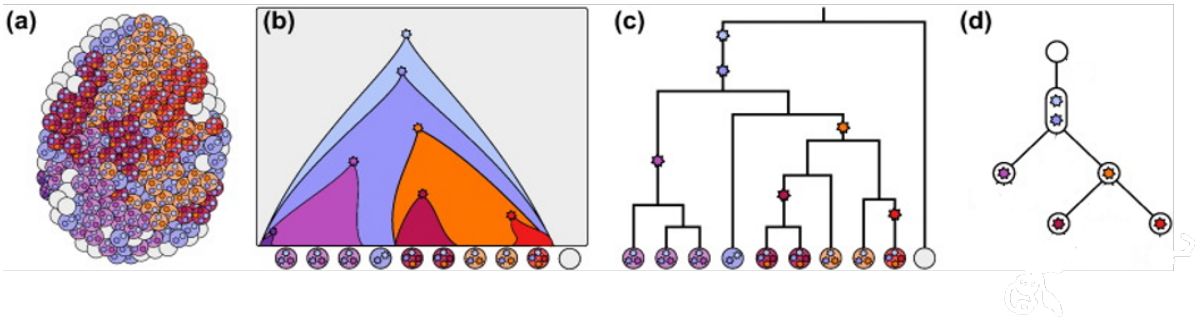


Figure 2.1: In the depicted heterogeneous tumor (a), which has evolved as shown in (b), 10 single cells are chosen for sequencing. One cell is normal tissue, while the other nine from the tumor have additional mutations, indicated by stars in the cells. These cells are part of a binary genealogical tree (c), connected at their common ancestors. The exact branching points might not always be determined by each cell's mutations; for instance, the three cells on the left can be arranged in any order as long as they all fall below the purple mutation, which sets them apart from other cells. The representation in (c) is a sample genealogical tree that focuses on the cells' relationships, while an equivalent representation in (d) encapsulates mutations in tree nodes to form a mutation tree, highlighting the mutations' ordering and evolutionary history. Adapted from Kuipers et al. [21].

The construction of evolutionary trees of tumors has been significantly enabled by advancements in Deoxyribonucleic Acid (DNA) sequencing technology and methodologies [22]. Over the past few years, many computational methods have been developed for reconstructing mutation trees using bulk sequencing data, single-cell sequencing data, or a mix of both, including SCITE [6], B-SCITE [23], PhISCS [24], MIPUP [25], and LICHeE [26]. It has been observed that tumor trees, derived from different types of a patient's sequencing data using various tree inference methods, often exhibit variability in both topology and node labels, corresponding to the mutated genes [9].

The evolution of DNA sequencing from bulk to single-cell methodologies has significantly enhanced the precision and depth of labelled trees [27]. In the realm of bulk data, while the identification of mutation prevalence within the sample is relatively accurate, challenges arise in detecting low-frequency events, which could potentially encapsulate a substantial portion of a tumor's diversity and bear relevance for treatment strategies. Enhanced sequencing depth improves the accuracy in discerning their frequency and, consequently, elucidating their evolutionary trajectory. However, a predominant challenge persists in deciphering the clonal structure amidst the DNA mixture derived from numerous cells.

On the other hand, single-cell data removes the need for deconvolution but introduces more noise to the sequencing data because it requires extensive amplification of the initial DNA material and feedback within the amplification process, complicating mutation detection. In the prospect of further developments in single-cell technologies, a comprehensive approach encompassing phylogeny and sequencing is anticipated, taking into account a wider range of mutation types and replicating the progress noted in bulk data [21]. Leveraging the detailed mutation histories obtained through single-cell sequencing presents a promising frontier for personalised cancer therapy. A particularly compelling application lies in identifying recurrent mutation patterns by comparing high-resolution mutation trees across various tumor types. This comparative analysis of mutation trees seeks to illuminate common and divergent mutational trajectories, thereby enhancing our understanding of tumorigenesis and potentially refining therapeutic strategies [6].

2.1.3 Distance Metrics for Comparing Labelled Trees

The need of quantifying the level of similarity between possible tumour histories has been brought out by the ongoing development of tumour evolution inference techniques. This is a complex and significant task for various reasons.

First, the lack of thoroughly researched ad hoc measures frequently makes it difficult to benchmark new tumour tree inference techniques against current techniques or a ground truth tree [26, 28, 29]. Second, some techniques, like the GraPhyC method, which is a tree inference algorithm, are inherently dependent on distance measures to deduce the evolutionary history of a tumour; in this case, a distance measure is employed to create a consensus tumour history from a variety of input histories [11]. Third, Bayesian methods, which attempt to account for uncertainty in inferred tree structures through sampling procedures or inference of unobserved clones, could significantly benefit from an analysis of tree similarity [6, 30, 31]. Lastly, the use of tumour evolutionary histories across patients to identify patterns of tumour evolution and emerging questions about the structure of the space of possible evolutionary histories, consistent with underlying sequence data, call for additional analysis, made possible by distance measures which can identify key features of tumour evolution histories [32–34].

Before understanding the foundation of the mutation trees distance measure, a few concepts must be understood. A tree is an undirected graph that is connected and does not contain any cycles. A rooted tree is a form of tree in which one node is identified as the root and all other nodes follow a distinct directed path to it. The edges of a rooted tree are oriented away from the root [20].

The set of nodes for a given tree T is denoted as $V(T)$, the set of leaves as $Leaf(T)$, and the

set of edges as $E(T)$. For a node i in $V(T)$, the degree of n is the number of edges connected to it. In a rooted tree, nodes with a degree of one are called leaves, while nodes with a degree greater than one are called internal nodes. The parent-child relationship defines the relationship between nodes in a tree: for nodes i and j in $V(T)$, if (i, j) is an edge in $E(T)$, then j is considered a child of i and i is considered the parent of j. Additionally, if the unique path from the root of the tree to a node j includes node i, then j is considered a descendant of i and i is considered an ancestor of j. The set of all children, ancestors, and descendants of a node i is denoted as $C_T(i)$, $A_T(i)$, and $D_T(i)$, respectively. It should be noted that i is always considered an ancestor and descendant of itself.

The set of all tumor trees is denoted as $\mathcal{T}$ and $\mathcal{T}_m$ as the set of all $m-$tumor trees that have an equal set of mutations [12]. The set of all mutations in a given tumor tree T is denoted as $M(T)$. Each vertex in the tree represents a distinct tumor clone (or population) that existed at some point during the tumor's evolution. The mutations that are labeled on a vertex indicate the clone in which the mutation first appeared. As a result, the complete set of mutations in any given clone, represented by vertex j, is the set of mutations that label all of the vertices on the path from the root to vertex j, i.e. $A(j)$ including j [12].

A multi-labeled tree T is a rooted tree where each vertex j save the root has a subset L_j of labels from a universe $M(T)$ and each label is unique to a vertex, i.e. $L_i \cap L_j = \emptyset$ for each pair of different vertices i and j. $L(T)$ denotes the set of all labels assigned to T's vertices [9].

A tumor evolution distance measure must provide a quantitative assessment of how different two tumor histories are from one another, but how to define dissimilarity is not immediately evident. There are two major elements of tumor evolutionary trees that should be considered when calculating distance: the tree's topology, which is a type of structure in which each node in a hierarchy is linked to the others, and the labels contained in the tree's vertices which correspond to the mutations [12, 35].

Mutation trees, which can be defined over varying sets of mutations, pose a significant challenge for comparison, especially when diverse data types (like single-cell versus bulk data) or differing mutation calling criteria are used. In contrast to traditional phylogenetic trees, mutation trees are fully labeled, showcasing mutations on both leaves and internal nodes.

Standard distance measurement techniques for phylogenetic trees, such as the Robinson–Foulds (RF) method and triplet distance, are not suited for mutation trees since they presuppose a shared set of leaf labels and overlook the multiple labels featured on each node [36, 37]. Therefore, classical tree distance metrics prove to be ineffective for mutation trees, requiring the creation of new measurement methods.

Govek et al. [11] delineated several simplistic distance measures for clonal trees, which ex-

panded upon previous ad hoc methodologies contingent upon the presence of a ground truth tree, Parent-Child Distance (PCD), Ancestor-Descendant Distance (AD), Clonal Distance (CD) and Path Distance (PD) [26,28,29]. However, these were not the primary focus of their research, and consequently, a profound analysis of their efficacy was not conducted. Furthermore, these distance measurements focused only on one aspect of similarity across trees, omitting to consider how these aspects affected the overall structure of the trees under consideration.

2.1.3.A Parent-Child Distance

Parent-Child Distance (PCD) quantifies the distance or difference between two trees taking into account the parent child relationship from both. $\phi_{PC}(T)$ is the set of all ordered pairs of mutations (i, j) such that i is a parent to j in T, formally $\phi_{PC}(T) = \{(i, j) \mid i \text{ is a parent of } j \text{ in } T\}$. PCD ($PCD(T_1, T_2)$) is defined

$$PCD\left(T_1, T_2\right) = \left|\phi_{PC}\left(T_1\right) \cap \phi_{PC}\left(T_2\right)\right|. \tag{2.1}$$

2.1.3.B Ancestor-Descendant Distance

Ancestor-Descendant Distance (AD) is a distance between two trees defined as the number of ancestor-descendant relationships that exist in one tree but not the other. Given $\phi_{AD}(T) = \{(i, j) \mid i \text{ is ancestor to } j \text{ in } T\}$, that is $\phi_{AD}(T)$ is the set of all ordered pairs of mutations (i, j) such that i is ancestral to j. Formally,

$$AD\left(T_1, T_2\right) = \left|\phi_{AD}\left(T_1\right) \cap \phi_{AD}\left(T_2\right)\right|. \tag{2.2}$$

2.1.3.C Path Distance

The sum of the absolute values of the differences in path lengths in T_1 and T_2 for each pair of mutations, or $path(T, i, j)$, determines the path distance (PD) between two tumour trees, T_1 and T_2, representing the length of the unique path in T that connects the vertex with label i to the vertex with label j, where $i, j \in [M(T)]$. It should be noted that this path only considers the intersection of the mutation labels in both trees, ignoring any directionality related to the graph's edges. Formally,

$$PD\left(T_1, T_2\right) = \sum_{i<j} \left|path\left(T_1, i, j\right) - path\left(T_2, i, j\right)\right|. \tag{2.3}$$

2.1.3.D Clonal Distance

A measure known as clonal distance (CD) is defined as follows: it counts the number of clones, which is $clone(i)$ representation of all the mutations labelling all vertices from the root to i that are unique to either T_1 or T_2. This allows us to compare the sets of clones underlying two tumour trees. Formally,

$$CD\,(T_1, T_2) = |\ \text{clone}\ (T_1) \cap\ \text{clone}\ (T_2)\ | \,. \tag{2.4}$$

2.1.3.E Multi-Labeled Tree Dissimilarity

The Multi-Labeled Tree Dissimilarity (Multi-Labeled Tree Dissimilarity (MLTD)) approach was proposed by Karpov et al. [8] to handle the difficulties of mutation inheritance and multi-labeled nodes; it uses an edit distance-based metric to address multi-labeled nodes in clonal trees and calculates the smallest number of moves required to convert both trees, T_1 and T_2, into a given common tree, in other words by finding on their maximum common tree.

A *Common tree* is a multi-labeled tree which can be obtained from each of T_1 and T_2 by the use of edit operations defined below and a *maximum common tree* is common tree between T_1 and T_2 that has the highest number of labels among all. To obtain this maximum common tree, the following types of edit operations are considered: deleting a label from a set, deleting an unlabeled leaf from the tree, and expanding a vertex by replacing it with two new vertices and reassigning the labels accordingly.

The MLTD is determined by subtracting from the total number of labels in $T_1$and T_2 twice the number of labels in their maximum common tree. It denotes the total number of labels that must be deleted from the two trees in order to produce the most common tree.

While MLTD is helpful for comparing trees at different resolution levels, it could fail to be the most effective way for evaluating new algorithms because it fails to contrast trees inferred by various techniques with the actual underlying tree.

2.1.3.F Common Ancestor Set distance and Distinctly Inherited Set Comparison distance

In the study by DiNardo et al. [12], implemented two novel distance metrics, the Common Ancestor Set distance (Common Ancestor Set distance (CASet)) and Distinctly Inherited Set Comparison distance (Distinctly Inherited Set Comparison distance (DISC)), offering nuanced approaches to handle mutation inheritance. CASet examines differences in mutation labelling and tree topology by comparing the common ancestors of all mutation pairs and using the

number of clones that inherit common mutations to weight the effect of mutation labelling differences. DISC, on the other hand, focuses on the set of mutations that distinguish clones from each other, placing more emphasis on recently acquired mutations. While both methods can identify complex clustering structures in a space of trees, they may not adequately penalise changes that have significant evolutionary impact or may not provide sufficient granularity when clustering sets of relatively similar trees [12].

Given two $m-$tumor trees $T_1, T_2 \in T_m$, CASet distance is a measure that compares T_1 and T_2, by finding the average similarity between the sets of common ancestors for each node in both trees. It does this by calculating the Jaccard distance between each set of common ancestors.

Jaccard similarity is a measure of how similar two sets are. It is defined as the size of the sets' intersection divided by the size of the sets' union (Eq. 2.5). The outcome is a number between 0 and 1, with 1 indicating that the sets are identical and 0 indicating that they share no items.

Formally, given two sets, A and B, the Jaccard similarity is defined as :

$$\text{Jacc}(A, B) = \frac{|A \cup B| - |A \cap B|}{|A \cup B|}, \tag{2.5}$$

where $|A|$ denotes the number of elements in set A, and $\cap$ and $\cup$ denotes the intersection and union of the sets, respectively.

It illustrates how similar the evolutionary histories of the two trees are by examining the common ancestor sets of each vertex, emphasizing differences around the root. Formally,

$$\text{CASet}\left(T_1, T_2\right) = \frac{1}{\binom{m}{2}} \sum_{\{i,j\} \subseteq [m]} \text{Jacc}\left(C_1(i,j), C_2(i,j)\right), \tag{2.6}$$

Where $C_1(i,j)$ and $C_2(i,j)$ are the *common ancestor sets* of mutations of $i, j \in V(T)$ in T_k and T_ℓ respectively, i.e. $A(i) \cap A(j)$, where $A(i)$ as define before is the set of ancestor of all labels from i until the root, and Jaccard is the Jaccard distance function between two sets.

Further extension can be applied to this metric in order to be able to compare clonal trees with different sets of mutation labels. A focus on the common mutations between the input trees is implemented in the first extension of the clonal tree approach. This enables determination of the distance between two trees by simply considering the mutations present in both. Let $I_{1,2} = M(T_1) \cap M(T_2)$, the first extension can be defined as:

$$\text{CASet}_\cap\left(T_1, T_2\right) = \frac{1}{\binom{|I_{k,\ell}|}{2}} \sum_{[i,j] \subseteq I_{12}} \text{Jacc}\left(C_1(i,j), C_2(i,j)\right), \tag{2.7}$$

Secondly, the union of the mutation sets of the input trees is considered. When computing the distance between two trees, this variant of the tumor tree approach considers variations in the sets of mutation labels. This makes it valuable for comparing evolutionary trees of tumors built using various data types, samples collected at different periods, or even patients. Let $U_{1,2} = M(T_1) \cup M(T_2)$ and if $i \notin M(T_1)$, then $A_1(i) = \emptyset$, the second variation is defined as:

$$\text{CASet}_{\cup}(T_1, T_2) = \frac{1}{\binom{|U_{1,2}|}{2}} \sum_{|[i,j] \subseteq U_{1,2}} \text{Jacc}\left(C_1(i,j), C_2(i,j)\right), \tag{2.8}$$

DISC, on the other hand, is a metric that compares the evolutionary paths of two mutation trees, T_1 and $T_2 \in T_m$, by accounting for the mutation differences in the recent tumor clones. It calculates the average Jaccard distance between the *distinctly inherited ancestor* sets of each vertex in the two trees. In simpler terms, it measures the similarity of the evolutionary paths of the two trees by comparing the sets of ancestors that are unique to each vertex of the trees. Formally,

$$\text{DISC}(T_1, T_2) = \frac{1}{m(m-1)} \sum_{\substack{(i,j) \in [m]^2 \\ i \neq j}} \text{Jacc}\left(D_1(i,j), D_2(i,j)\right), \tag{2.9}$$

Where $D_1(i,j)$ and $D_2(i,j)$ are the *distinctly inherited ancestor sets* of $i, j \in M(T)$ in T_1 and T_2 respectively.

Similarly, as previously stated in CASet, there are two possible extensions for DISC metric, one that only incorporates the intersection between the two clonal trees that are being compared and can be defined as follows:

$$\text{DISC}_{\cap}(T_1, T_2) = \frac{1}{I_{1,2}\left(I_{1,2} - 1\right)} \sum_{\substack{(i,j) \in I_{1,2} \\ i \neq j}} \text{Jacc}\left(D_1(i,j), D_2(i,j)\right), \tag{2.10}$$

Alternatively, the other extension that considers the union of both trees' mutation sets, that modifies DISC as:

$$\text{DISC}_{\cup}(T_1, T_2) = \frac{1}{U_{1,2}\left(U_{1,2} - 1\right)} \sum_{\substack{(i,j) \in U_{1,2} \\ i \neq j}} \text{Jacc}\left(D_1(i,j), D_2(i,j)\right). \tag{2.11}$$

2.1.3.G Triplet-based Similarity Score

In 2020, Ciccolella et al. [10] introduced a new distance metric for mutation trees, known as Triplet-based Similarity Score (MP3), which is based on rooted triples - a widely used distance

measure in phylogenies.

MP3 is based on the distance proposed by Jansson and Rajaby [38], where the rooted triplet distance measures the dissimilarity between two leaf-labeled trees with identical labels and this dissimilarity is given by the number of rooted triplets that induce different minimal topologies in the two trees over the total number of triplets. It was extended to include multi-labeled trees and poly-occurring labels, considering the recurrence and loss of mutations. MP3 is unique in that it shows a monotonic decline in similarity as the number of poly-occurring labels grows and distinguishes between intra- and inter-similarity better compared to previous metrics. However, it has a significantly more dispersed distribution for intrasimilarity scores, indicating potential limitations in its applicability across varied scenarios [10].

Two common assumptions are usually made in tumor evolution, the first being the Infinite Sites Assumption (ISA) which says that no mutation happens more than once in the history of a tumor and that once attained, a mutation is never lost [39]. The second assumption is that all tumor cells descend from a single founder tumor cell, thus the development of the tumor may be defined as monoclonal [12]. Most metrics make use of these assumptions, but research has shown that the first assumption is not always true due to the possibility of poly-occurring mutations and a loss of existing mutations. As mentioned above, not only MP3 can include this ISA violation and account for it in its calculation, but is the only metric capable of handling this, being in this way a very powerful metric.

Given three leaves (a, c, e) in V(T), the minimal tree topology they induce on T, denoted as $MTT_T(a, c, e)$, is the smallest subtree of T that includes the nodes $V_T(a, c, e) = a, c, e \cup LCA(a, c) \cup LCA(a, e) \cup LCA(c, e)$, where $LCA(a, c)$ represents the lowest common ancestor of nodes a and c and all the nodes with degree 2 that are different from (a, c, e) are contracted (Fig. 2.2 (a)).

However, because tumor trees are both completely labelled and multi-labelled, such a metric cannot be used directly. As a result, there was a need to develop an extension, MP3, that includes those as well as the case of poly-occurring labels. Ciccolella et al. [10] began by determining that there are only nine possible configurations for $MTT_T(a, b, c)$ given T multi-labelled and $a, b, c \in M(T)$ (Fig. 2.2 (b)), being able in this case to extend for the fully and multi-labelled trees. In order to make a meaningful comparison and not overlook the poly-occurring labels the minimal tree topology definition needed to be extended to $MIN_T(a, b, c)$ for poly-occurring labels (a,b,c) (Eq. 2.12), note that if a label occurs multiple times in the tree, then N maps each label to one or more nodes in V_T and it includes a multiset union ($\sqcup$) to account for the

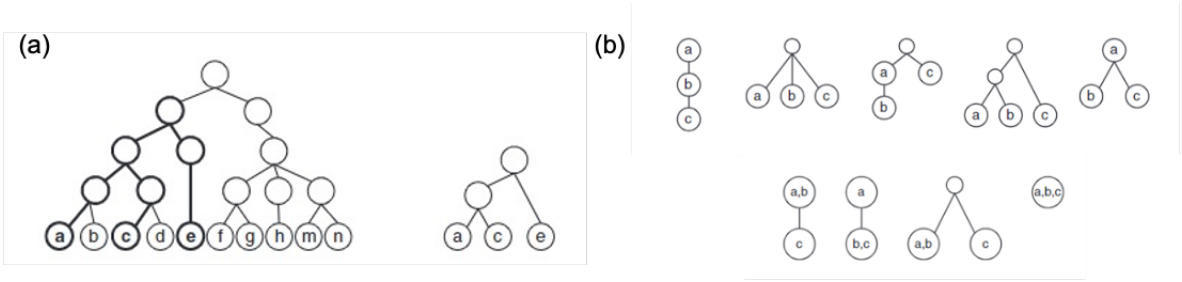


Figure 2.2: Illustration of (a) rooted triplet on labels (a,c and e). (Left) Tree T, highlighting the smallest subtree that has all three labels. (Right) The resulting minimum topology by (a,c and e); (b) The nine possible configurations for the minimal tree topology of multi-labelled trees induced by an unordered set of three labels. Adapted from [10].

multiplicity of each element in each set.

$$\mathrm{MIN}_T(a,b,c) = \bigsqcup_{u\in \mathrm{N}(a), v\in \mathrm{N}(b), z\in \mathrm{N}(c)} \mathrm{MTT}_T(u,v,z), \tag{2.12}$$

The similarity measure-based can then be split into two variants where the cardinality of the multiset intersection is $N(a,b,c) = |\mathrm{MIN}_{T_1}(a,b,c) \sqcap \mathrm{MIN}_{T_2}(a,b,c)|$ and the maximum number of configurations of the triplet in the trees is $D(a,b,c) = \max\{|\mathrm{MIN}_{T_1}(a,b,c)|, |\mathrm{MIN}_{T_2}(a,b,c)|\}$. The first variation (Eq. 2.13) includes only the mutations that appear on both trees, resulting in I being the set of triples in $M(T_1) \cap M(T_2)$, and the second (Eq. 2.14) includes all mutations, resulting in I being the set of triples in $M(T_1) \cup M(T_2)$.

$$\mathrm{MP3}_{\cap} = \frac{\sum_{(a,b,c)\in I} N(a,b,c)}{\sum_{(a,b,c)\in I} D(a,b,c)}, \tag{2.13}$$

$$\mathrm{MP3}_{\cup} = \frac{\sum_{(a,b,c)\in J} N(a,b,c)}{\sum_{(a,b,c)\in J} D(a,b,c)}, \tag{2.14}$$

Both these variations have their advantages and disadvantages, $\mathrm{MP3}_{\cap}$ shows how effectively the two progressions may be condensed into an identical subsequence of common mutations and $\mathrm{MP3}_{\cup}$ evaluates the impact of acquired or lost mutations in only one progression. To benefit from both versions, a weighted mean combination with an intended bias toward $\mathrm{MP3}_{\cap}$ is used, in order to find inner parallels in various trees (Eq. 2.15).

$$\mathrm{MP3}_{\sigma} = \mathrm{MP3}_{\mathrm{U}} + \sigma\left(\mathrm{MP3}_{\cap}\right) \cdot \min\left\{\mathrm{MP3}_{\cap} - \mathrm{MP3}_{U}, \mathrm{MP3}_{U}\right\}, \tag{2.15}$$

where $\sigma(x) = \frac{1}{1+\exp\left(-\mu\left(x-\frac{1}{2}\right)\right)}$, which is a sigmoid function i.e. a type of mathematical function that has an 'S' shaped curve.

Another combination of the two versions was available, being a geometrical combination of the union and intersection outputs, defined as in Eq. 2.16.

$$\mathrm{MP3}_{geo} = \sqrt{\mathrm{MP3}_{\mathrm{U}} \cdot (\mathrm{MP3}_{\cap})}. \tag{2.16}$$

2.1.3.H Bourque Distance and k-Bourque Distance

Recently, Jahn et al. [9] proposed a collection of distance measures called Bourque distances, which can be used to measure the topological dissimilarity between rooted and unrooted labeled trees with different label sets. These distances are based on the RF distance, and are closely related to the edge contraction and decontraction operations introduced by Bourque for leaf-labeled unrooted trees, as well as the nearest-neighbor interchange distance. The Bourque distance for leaf-labeled unrooted tree, also known as the 'edge contraction distance', was first introduced in 1978. Bourque et al. [40] proposed a method for comparing and measuring the similarity between two phylogenetic trees by considering the number of edge contractions needed to transform one tree into the other. This distance metric is based on the idea that two trees are considered more similar if they require fewer edge contractions to be transformed into one another. Let T_1 and T_2 be two rooted labeled trees, the RF distance denoted as $\mathrm{RF}(T_1, T_2)$ can be defined as follows:

$$RF(T_1, T_2) = |\mathcal{O}P(T_1) \triangle \mathcal{O}P(T_2)|, \tag{2.17}$$

where $\mathcal{O}P(T_1)$ is defined as the set of partitions caused for each unique edge entering the node $u \in V(T_1)$ (Eq. 2.18) and Δ is the symmetric difference operator which is the set of elements that are in the sets, but not in the intersection.

$$\mathcal{O}P(T_1) = \{P(u)|\ u \in V(T_1)\}, \tag{2.18}$$

where $P_u = \{L_{T_1}(u), M(T_1) \backslash L_{T_1}(u)\}$, which denotes $L_{T_1}(u)$ as the set of labels of u and its descendants $L_{T_1}(u) = \cup_{x \in \{u\} \cup D_{T_1}(u)} \ell(x)$ and $M(T_1)$ the set of labels appearing in T_1.

When both trees have different sizes or labels, the RF distance is simply equal to the total number of edges in both trees; to avoid this issue, Bourque Distance (BD) generalises to be able to identify those instances. Let $\mathcal{P} = \{\{C', C \backslash C'\} : \emptyset \neq C' \subset C, C' \neq C\}$, with $C = M(T_1) \cap M(T_2)$, the Bourque metric between T_1 and T_2 is defined as:

$$BD(T_1, T_2) = |\mathcal{O}P(T_2) \cup \mathcal{O}P(T_1)| - \sum_{P \in \mathcal{P}} \min\left(\left|\mathcal{O}'_{T_1}(P)\right|, \left|\mathcal{O}''_{T_2}(P)\right|\right), \tag{2.19}$$

where

$$\begin{aligned} \mathcal{O}'_{T_1}(P) &= \left\{P' \in \mathcal{O}P(T_1) : P' \sim P\right\} \\ \mathcal{O}''_{T_2}(P) &= \left\{P'' \in \mathcal{O}P(T_2) : P'' \sim P\right\}, \end{aligned} \tag{2.20}$$

So BD rectifies RF by using the similarity relationship ($\sim$) between P and both $\mathcal{P}(T_2)$ and $\mathcal{P}(T_1)$, in other words, by using partitions that would be shared by both trees if labels unique to each tree were discarded. The ordered partitioned induced by the edges entering u and v are only denoted as similar if $L_{T_1}(u) \cap C, [M(T_1)\ L_{T_1}(u)] \cap C) = (L_{T_2}(v) \cap C, [M(T_2)\ L_{T_2}(v)] \cap C)$ and $\emptyset \neq L_{T_1}(u) \cap C \neq C$.

The BD, like the RF distance, has a propensity to overpenalize some labeling variations, so Jahn et al. [9] created other distance metrics, k-Bourque Distance (kBD), derivated from BD that take more values than it, it uses local subtrees and a matching algorithm. For an integer $k > 0$, a subtree composed by the node set $\{v \in V(T_1) : d(u, v) \leq k\}$ centered at u is the k-star subtree $C_k(u)$. For any pair of labeled trees T_1 and T_2 of n and $n\prime$ nodes, respectively, such that $n \leq n\prime$, $BG_k(T_1, T_2)$ is complete weighted bipartite graph with two node parts $\emptyset_1, \ldots \emptyset_{n\prime-n} \cup V(T_1)$ and $V(T_2)$, where each $\emptyset_i$ is just a copy of the empty graph for the unexisting nodes in S. For each edge of $BG_k(T_1, T_2)$, the Bourque distance $B(C_k(x), C_k(y))$ is assigned as the weight for every $x \in V(T_1)$ and $y \in V(T_2)$ and a weight of $|E(C_k(y))|$ is assigned to the edge $(\emptyset_i, y)$ for any $\emptyset_i$ and $y \in V(T_2)$. The kBD $B_k(T_1, T_2)$ is then defined as the minimum weight of a perfect matching in $BG_k(T_1, T_2)$.

The metrics mentioned above are summarised in Table 2.1. In this work, the terms 'distance metrics' and 'distance measures' are used interchangeably without specifying which ones meet the criteria to be called metrics.

However, a standardised method for comparing mutation trees has not yet been developed, and when the same trees are compared using some of the current measures described, they have shown shortcomings and produced distinct results (Fig. 2.3) [9]. Their calculation is described in detail in Appendix A. In addition, although distance measures for tumour trees seem necessary, very few have been meticulously developed and thoroughly evaluated on real data.

Table 2.1: Summary of the features of different complex distance measures.

Distances	Time Complexity	Metric	Strengths	Downfalls	References
PCD	Linear	Pseudo	Able to capture and distinguish trees with specific mutation order differences	Ineffective for trees with a depth of 1 unless the Root node is included in the analysis. Can't take into account the poly-occurence of mutations in a tree.	[11]
AD	Quadratic	Pseudo	Capable of identifying and differentiating trees based on specific mutation order variations throughout the whole tree, extending beyond PCD.		
CD	Quadratic	No	Strongly penalises discrepancies that are found nearer the tree's root than its leaves	Can't take into account the poly-occurence of mutations in a tree.	
PD	Cubic	No	Can detect structural similarity between non-descendants	It overlooks information regarding the ancestral relationships between mutations and only considers the intersection of the mutations.	
MLTD	Polynomial	No	Useful for tree comparison at different resolution levels	It is not useful to compare trees that are very similar. Fails on trees that break the ISA.	[8]
CASet	Cubic	No	Penalises more the differences near the root, useful when very different trees are expected	Can't discern complex poly-occuring labels.	[12]
DISC	Cubic	No	More emphasis n recently acquired mutations, achieves more granularity when clustering sets of relatively similar trees		
MP3	Cubic	No	It can distinguish well between the combination of topology and mutations	Not applicable to datasets with shallow trees with fewer than three intersected mutation labels trees.	[10]
BD	Polynomial	Yes	It captures well differences in both ancestors and descendants in trees	Overpenalises some labeling variations and can saturate quickly.	[9]
kBD	Cubic	Yes	Refines BD and is able to discern better between trees	Unable to detect differences in trees with poly-occuring labels.	

	Scores
CASet	0.11
DISC	0.33
BD	0.33
MP3	0.5
PCD	0.8
CD	0.4
AD	0.4

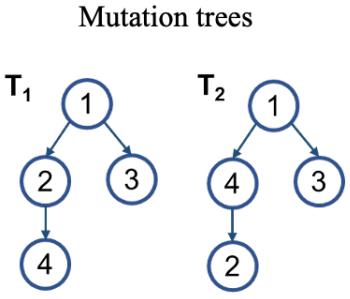


Figure 2.3: Comparison of the different metric scores for the same trees, shown on the left, and the mutation trees, shown on the right.

2.1.4 Distance Combination and Optimisation Methods

In developing the MP3 distance measure, Ciccolella et al. [10] pioneered an innovative path by combining together two distinct versions of the metric and maximising the strengths of each. This effective combination not only increased the MP3 distance metrics robustness but also broadened the limits of its applicability, creating a precedent for improved metric development. The success of this union highlights the opportunity to use similar approaches to combine various distance measurements. It is then possible to create a more sophisticated and robust metric that takes use of the combined strengths by carefully combining the key elements of several metrics. This opens up a wide range of options in the area of distance metric creation and optimisation.

Optimisation algorithms are central to numerous computational and engineering fields, coordinate a systematic search for viable solutions to choose the optimal one based on predetermined criteria. These algorithms aim to either minimise or maximise a specific function that encapsulates the problem being addressed. Over time, a wide array of algorithmic strategies has been developed, each embodying unique attributes designed for various optimisation challenges. The essence of optimisation algorithms lies in the iterative enhancement of solutions, aiming towards an optimal result that complies with a set of predetermined guidelines or constraints. The broad domain of optimisation algorithms includes techniques like gradient-based algorithms, apt for problems with continuous derivatives, and derivative-free algorithms or metaheuristics, suitable for situations devoid of gradient information or involving non-convex functions [41, 42].

Traditional mathematical programming methods, such as linear programming [43], dynamic programming [44], and Newton's methods [45], have conventionally been used by practitioners to navigate these optimisation challenges. However, these conventional optimisers exhibit limitations such as limited global strength, poor initial solution guessing, and a strong reliance on gradient information, which can further reduce their usefulness in solving a variety of current engineering optimisation problems with rising complexity.

Metaheuristic search algorithms (MSAs) emerge as promising candidates to solve challenging modern optimisation problems by leveraging their search mechanisms, which are inspired by various natural phenomena. MSAs can perform searches with different levels of exploration and exploitation strengths during the optimisation process to locate the global or near-global optimum solution. Unlike mathematical programming methods, MSAs are generally more flexible and can locate the near-global optima of given optimisation problems more efficiently without requiring substantial modifications of algorithmic frameworks and the derivative information of given problems. The stochastic nature of MSAs also enables them to exhibit better robustness in handling local entrapment issues commonly encountered in real-world global optimisation problems [42].

An interesting example of an MSA is the Differential Evolution (DE) algorithm proposed by Storn et al. [46], defined as a heuristic strategy for globally optimising non-differentiable, nonlinear, and multi-dimensional functions (Fig. 2.4). It excels at identifying the global minimum of complex multidimensional functions, especially when the function landscape consists of numerous local minima. The general procedure consists of:

1. **Initialisation**: A population of candidate solutions is initialised randomly within the provided bounds.

2. **Mutation**: For each candidate in the population, a mutant vector is generated by combining vectors from the current population, typically using vector addition and difference:

$$v_i = x_{r1} + F \cdot (x_{r2} - x_{r3}), \tag{2.21}$$

where v_i is the mutant vector, x_{r1}, x_{r2}, and x_{r3} are vectors randomly selected from the population and F is a scale factor that controls the rate at which the population evolves.

3. **Crossover**: Elements of the mutant vector are probabilistically mixed with elements from the original candidate to produce a trial vector:

$$u_{i,j} = \begin{cases} v_{i,j} & \text{if rand}(j) \leq CR \\ x_{i,j} & \text{otherwise} \end{cases}, \tag{2.22}$$

in which $u_{i,j}$ is the j-th parameter of the trial vector, $v_{i,j}$ is the j-th parameter of the mutant vector, $x_{i,j}$ is the j-th parameter of the target vector, CR is the crossover rate,rand(j) is a random number between 0 and 1.

4. **Selection**: The trial vector replaces the original candidate in the population if it yields a lower objective function value:

$$x_i = \begin{cases} u_i & \text{if } f(u_i) \leq f(x_i) \\ x_i & \text{otherwise} \end{cases}, \tag{2.23}$$

where x_i is the target vector, u_i is the trial vector and f is the objective function.

The above steps are iteratively repeated until a termination criterion is met, such as a maximum number of iterations or an acceptable convergence level [47]. Uniquely, it requires minimal control parameters, making it user-friendly and practical. Mutation strategies follow a convention, with variations affecting population diversity and convergence. The process continues until a termination criterion, ensuring only the fittest solutions survive.

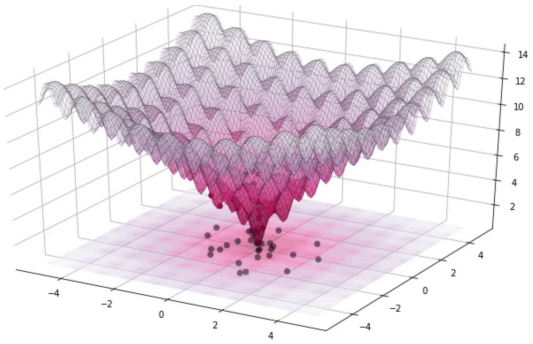


Figure 2.4: Example of DE iteratively optimizing the 2D Ackley function (generated using Yabox) [1].

2.1.5 Clustering in Biological Data Analysis

The metrics, displayed on Table 2.1, with the exception of BD and kBD, have been used for clustering mutation trees, most in artificial data sets [10,12] and in only one real data set [10], and have demonstrated significant potential, especially in the domain of cancer research. Given the complexity and diversity inherent in cancer, clustering trees that correlate with similar clinical outcomes can enhance treatment efficacy and deepen our understanding of the underlying genetic mutations. Consequently, the evolution and refinement of distance measures for mutation trees not only stand to enhance our comparative and analytical capabilities but also to substantively impact patient care and treatment outcomes in oncology.

Clustering algorithms are a fundamental tool in data analysis that groups data objects into subsets (clusters) based on similarity or dissimilarity measures. The key idea behind clustering is that patterns within a valid cluster are more similar to each other than they are to patterns belonging to a different cluster. This process can be performed using different techniques, such as unsupervised, semi-supervised, or supervised methods [48].

The measurement of similarity, dissimilarity, or distance between features or objects, followed by the use of an algorithm to discover underlying clusters, are the two main steps that must be completed in a clustering analysis. These steps may be carried out individually or concurrently [49].

The algorithms may be generally categorised into groups like Hierarchical, Partitional, Density-based, and Grid-based depending on the approach used to generate clusters [48]. Hierarchical clustering is generally used when it comes to techniques that employ the pairwise distance metric for tree similarity as seen in several studies [10, 12]. Through this technique, a distance matrix of pairwise similarity measurements between all items is transformed into a hierarchy of nested groups. The layered grouping of patterns and the similarity thresholds at which groupings change are displayed in a dendrogram that resembles a binary tree to indicate the

hierarchy. It can either cluster through agglomeration or division, the first being the case where it starts with each tree as a cluster and then proceeds to successively cluster them, and the latter starts with one big cluster and then iteratively splits it. Specifically, agglomerative hierarchical clustering works by first assigning each tree to a single cluster, and then joining the closest clusters together until every tree is in one big cluster. Different linkage or amalgamation techniques, including single-linkage, complete-linkage, average-linkage, and Ward's approach, can be used to recalculate the distances between two clusters. For example, complete-linkage distance is the largest distance between any two members of distinct clusters, whereas, single-linkage defines the distance between two clusters to be the lowest distance between any member of each cluster. Any data type can be used in hierarchical clustering, which makes no assumptions about its distribution [48].

The silhouette score is a measure of the similarity between an object and the other objects in its cluster (cohesion) normalized by the distance to the closest cluster (separation), used to evaluate the performance of a clustering algorithm and to determine the optimal number of clusters for a given data set. The silhouette score [50] for each sample, $s(i)$, is computed using:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}, \tag{2.24}$$

where $a(i)$ is the mean distance from the i-th sample to the other samples in the same cluster, and $b(i)$ is the smallest mean distance from the i-th sample to samples in a different cluster, minimised over clusters. Resulting in a value between -1 and 1, where a score of 1 indicates good clustering, a score of -1 indicates poor clustering, and a score of 0 indicates that the object is on the boundary between two clusters.

For many years, researchers have been interested in distance metric learning, with the goal of finding the best way to compute the similarity of data points. A good distance measure can enhance the performance of a machine learning model, whether it is used for classification or clustering. Advances in this field have led to the development of new distance methods to infer (dis)similarity between different types of data, such as mutation trees. This allows for the clustering of these, which can help us better understand the patterns of genetic mutations similarities [48, 51].

In real dataset applications, a paper performed a clustering analysis on 36 medulloblastoma patients from another study [10, 52]. The patients were clustered according to four different distance metrics and its alternate versions and were compared against the clinical data regarding the tumor subgroups. Using scRNA-seq data, the researchers inferred the cancer phylogeny of each patient using the method SCITE [6]. They then computed the similarities between all the inferred trees and used them to perform a hierarchical clustering. The results, shown in

Fig. 2.5, indicate that using the measure MP3 was able to distinctly group the patients into their relative subtypes with only a few mismatched trees. A similar result was achieved by the measure CASet∩, while the other measures tended to cluster together subtypes SHH and WNT, without a clear distinction between them. This study, which used distance measurements on

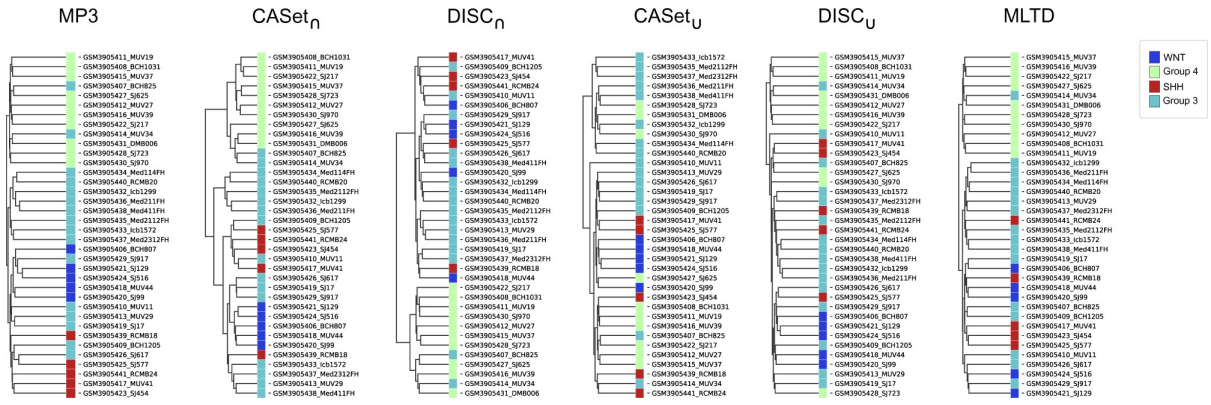


Figure 2.5: Results for the clustering experiment: Hierarchical clustering obtained from 36 medulloblastoma patients using different distance metrics and where the different colors represent the different subgroups from the clinical data (WNT, Group 4, SHH and Group 3). From [10]

.

actual data, was only able to distinguish between 36 patients' cancer subtypes. In order to fully understand the value and importance of these distance metrics for patient clustering, further actual dataset studies using various forms of clinical data, such as survival data, are required.

2.2 Survival Analysis in Cancer Research

Physicians are increasingly interested in using accurate prognostic tools to inform medical care, which has led to an increase in attention from the machine learning community on survival data. Clustering patients with associated survival data in cancer cases is a particularly interesting area for comparing the future success of treatment, for example.

Survival analysis is a way of investigating the distribution of lifetimes, or the period between an initial event like birth, the start of treatment and a terminal event like death or disease. The effect of an intervention is measured in clinical or community studies by counting the number of individuals who survive or are saved following the intervention over time [53].

In some cases, such as cancer therapy, this length of time can be quite long, in which case the number of occurrences, such as death, can be examined per unit amount of time. In other cases, the period before a cancer reappears or an infection arises can be calculated. The period from a defined point to the occurrence of a specific event is known as the survival time, and survival analysis is the examination of that group data [54]. A portion of the survival times of

interest will frequently be unknown due to censoring, which is the non-observation of the event of interest during the period of follow-up [55].

In cancer research, survival analysis has several important applications. First off, it helps to clarify how many elements, including medical interventions, societal trends, and coexisting diseases, affect the length of time that cancer patients live. It is helpful in predicting survival probabilities and comparing survival distributions among various patient groups, which is essential for assessing the effectiveness of therapies and treatments. Secondly, by analysing survival statistics, professionals are better equipped to decide on treatment strategies and patient counselling. Additionally, it aids in the identification of patient populations at high risk and is essential to the planning of clinical studies. The ability to predict the time until an event, such as death or a cancer recurrence, is provided by survival analysis, which is crucial for patient care and treatment planning [55–57].

2.2.1 Models for Survival Analysis

Among the key techniques in Survival Analysis are the Kaplan-Meier Estimation and the Cox Proportional Hazards Model. The Kaplan-Meier Estimation is a non-parametric method employed to estimate the survival function from lifetime data. It is particularly useful in quantifying the fraction of subjects living for a certain duration post-treatment. Its survival curve is defined as the likelihood of surviving in a given amount of time when time is divided into numerous small intervals [53].

$$\widehat{S}(t) = \prod_{t(i)\leq t} \frac{n_i - d_i}{n_i}, \tag{2.25}$$

where $\widehat{S}(t)$ is the Kaplan-Meier estimate, $t(i)$ is the time passed to the next observation from the beginning of the study, n_i the number of patients that are still alive at the time i and d_i refers to the number of deaths at time i.

This method finds its application predominantly in clinical or community trials to assess the impact of an intervention by measuring the number of subjects survived over a defined period (Fig. 2.6).

On the other hand, the Cox Proportional-Hazards Model is a semi-parametric model utilised to study the effect of several variables upon the time a specified event takes to occur. Unlike the Kaplan-Meier Estimation, this model allows for the inclusion of covariates, enabling adjustments for other variables. It is often employed to identify risk factors and obtain adjusted hazard ratios, which are crucial for understanding the underlying mechanisms affecting survival [58].

The fundamental component of the Cox model is the hazard function, denoted as $h(t)$, which

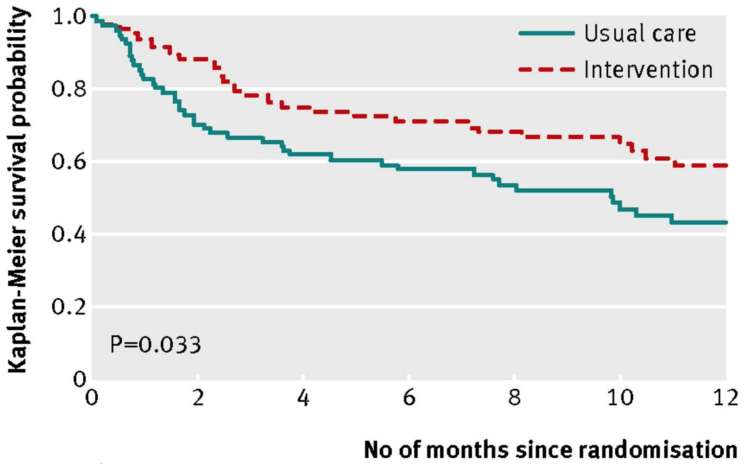


Figure 2.6: Kaplan-Meier survival curves for the intervention and control treatment groups until the occurrence of death from any cause or hospital readmission for heart failure following randomisation. From [2].

describes the risk of experiencing the event of interest at time t, given survival up to that time. The hazard function for a patient i given his profile $\boldsymbol{X}_i$ is represented as follows:

$$h\left(t \mid \boldsymbol{X}_i\right) = h_0(t) \exp\left(\boldsymbol{X}_i' \boldsymbol{\beta}\right), \tag{2.26}$$

where $h_0(t)$ is the baseline hazard, representing the hazard function when all covariates are zero, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)$ are the coefficients representing the influence of the covariates and where $\boldsymbol{X}_i$, with $i = 1, \ldots, n$, is the profile of the patient i over P covariates (clinical data, mutations and clusters), $\boldsymbol{X}_i' = (X_{i1}, \ldots, X_{iP})$.

The coefficients β are usually estimated by maximising the method's total log-likelihood, which is a function that induces partial log-likelihood of the data that involves the hazard functions for only those individuals who experienced the event, $\delta_i = 1$, and a baseline hazard which is given by the Breslow estimator:

$$\hat{h}_0\left(t_i\right) = \frac{1}{\sum_{j:y_j \geq t_i}^{n} \exp\left(\boldsymbol{X}' \boldsymbol{\beta}\right)}, \tag{2.27}$$

The overall partial log-likelihood function is given by:

$$l\left(\beta, h_0\right) = \sum_{i=1}^{n} - \exp\left(\boldsymbol{X}_i' \beta\right) H_0\left(t_i\right) + \delta_i \left[\log\left(h_0\left(t_i\right)\right) + \boldsymbol{X}_i' \beta\right], \tag{2.28}$$

with $H_0\left(t_i\right) = \sum_{t_k \leq t_i} h_0\left(t_k\right)$, the inference of the optimal coefficients alternates between maximising with β and updating the $h_o(t)$ estimation.

The Cox model is called a proportional hazards model because the primary assumption of

the model is that the hazard ratios are constant over time. This can be assessed using various diagnostic plots and tests, such as performing a test based on the scaled Schoenfeld residuals. Although a breach of the proportional hazards assumption indicates a potential unsuitability of the Cox proportional hazards model due to a change in a predictor's effect on the hazard rate over time, it can be addressed by using alternative models such as stratified Cox models or extended Cox models. [59].

Particularly when employing the Cox Proportional-Hazards Model, regularisation plays an important role in enhancing model precision and interpretability. Regularisation techniques, such as Lasso (L1 regularisation) (Fig. 2.7 (a)) and Ridge (L2 regularisation) (Fig. 2.7 (b)), are deployed to mitigate overfitting, especially when navigating scenarios with various of features or multicollinearity. Lasso regularisation performs feature selection in addition to regularisation by adding the absolute value of the coefficient's magnitude as a penalty term to the loss function. This can result in some feature coefficients being exactly zero. This technique has found its application in addressing issues in accelerated failure time models, Cox's model, and semi-parametric relative risk models. Conversely, Ridge regularisation reduces the coefficients but does not set them to zero by adding the squared magnitude of the coefficient as a penalty term to the loss function [60].

Group Lasso (Fig. 2.7 (c)), a particular case of L1 regularisation, has attracted an interest in the literature due to its suitability in situations requiring the consideration of grouped structures of covariates, such as clusters derived from a particular distance metric [61–63]. The integration of Group Lasso alongside the Cox model has shown promise in elevating survival prediction accuracy by incorporating gene-level group prior knowledge into the model training process [64]. Moreover, algorithmic considerations for Cox regression with Group Lasso penalty have been discussed, emphasising the identification of influential genes and clinical covariates, which are indispensable in analysing survival data [65]. In the context of the Cox proportional hazards model, the inclusion of the Group LASSO penalty leads to the following optimisation problem:

$$\min_{\beta} -l\left(\beta, h_0\right) + \lambda \sum_{l=1}^{G} \sqrt{p_l} \|\beta^{(l)}\|_2, \tag{2.29}$$

In the equation above, the first term is the negative log partial likelihood that we aim to maximise, as described in Equation 2.28. The parameter λ serves as the regularisation constant, influencing the degree of shrinkage applied to the coefficients. The total number of distinct groups, represented by G, can correspond to the aggregated clusters derived from distance metrics. Non-category covariates are treated as singular groups, thereby simplifying their regularisation into a standard LASSO framework. The variable p_l specifies the count of

clusters within the l-th group. The vector $\beta^{(l)}$ encompasses the coefficients associated with the l-th group. The term $\|\beta^{(l)}\|_2$ denotes the Euclidean norm of the coefficients within the l-th group, computed as the square root of the sum of squared coefficients. Selection of the optimal λ is typically achieved through a cross-validation process and chosen as the minimal mean squared error based on the log-likelihood deviance.

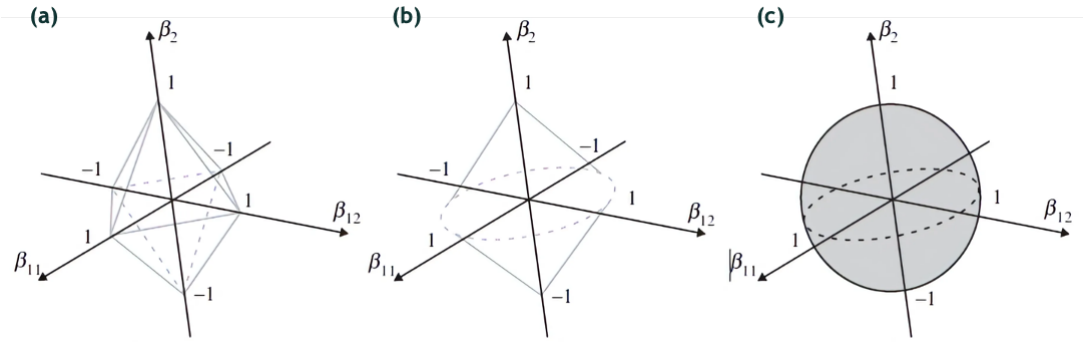


Figure 2.7: Graphical visualisation of the regularisation techniques L1, Group Lasso and L2, in the respective order. It displays an easy problem with the three coefficients β_2, β_{11} and β_{12}. Group lasso (middle image) takes into account the fact that the last two coefficients constitute a group while lasso (left image) does not. Group lasso can thus be understood as a ridge within groups and a lasso between groups. If a group is significant, the entire group is chosen. It sends it to zero if it isn't. Still, Ridge reduces the coefficients but never to zero (right image). [66]

Through these regularisation techniques, the Cox model's efficacy in analysing survival data can be significantly enhanced and simplified, providing a robust framework to navigate the complex landscapes of cancer research data.

2.2.2 Models Comparison and Assessment

Model evaluation is an essential aspect of using survival models because it ensure the models can generalise to new data and enables the comparison of different models to identify the most suitable one. Testing the model on unseen data is important to ascertain the efficacy of the hazard ratios deduced by the model, requiring the random division of data into training and test sets.

For a more nuanced evaluation of the regularised versions of the Cox model discussed earlier, on the training data, k-fold cross-validation is commonly employed. This technique partitions the data into k subsets, training the model on *k-1* of these subsets while validating it on the remaining subset. This procedure is conducted k times, with each subset serving as the validation set precisely once, thus ensuring a thorough evaluation. Cross-validation is instrumental in lowering the risk of overfitting and enables the fine-tuning of parameters of the models [67].

An established statistical hypothesis test can be used for the overall model assessment. When comparing the survival distributions of two or more groups, the log-rank test is employed. When analysing Kaplan-Meier survival curves to determine whether there are significant differences in survival curves between distinct groups, this test is especially useful. It makes it possible to analyse how different cohorts' survival probabilities change over time; however, it is unable to test the impact of other independent factors [53, 68]. The test statistic for the log-rank test is:

$$\chi^2 = \frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2}, \tag{2.30}$$

where O_1 and O_2 are the observed number of events in the high and low risk groups, respectively. E_1 and E_2 are the expected number of events in the high and low risk groups, respectively, under the null hypothesis of no difference between the groups. The test statistic follows a chi-squared distribution with 1 degree of freedom under the null hypothesis.

A p-value less than 0.05 from the log-rank test traditionally indicates a statistically significant difference between the survival curves of high and low risk groups, suggesting that the prognostic index effectively stratifies individuals based on their risk. This implies that the coefficients of the fitted model are adept at identifying crucial covariates for risk assessment.

However, when p-values are obtained from multiple comparisons, as is common in extensive analyses, the risk of false positives increases—a phenomenon known as the multiple testing problem. To address this, various multiple testing correction methods are employed to adjust the p-values, ensuring more reliable and accurate results [69].

The Bonferroni Correction is a straightforward yet conservative approach, adjusting the p-value by multiplying it by the total number of tests. Other methods, such as controlling the False Positive Rate, the Family-Wise Error Rate, or the False Discovery Rate, manage different aspects of error rates to mitigate the risk of false positives [69, 70].

While not all researchers adopt multiple testing corrections, and some might not report adjusted p-values in their publications, implementing these adjustments is crucial for maintaining the integrity of statistical findings, especially in the context of multiple comparisons. This ensures that the identified significant differences are indeed valid and not products of statistical chance, thereby enhancing the credibility of the research [71].

In addition to the log-rank test, Likelihood Ratio Test (LLR), introduced by Neyman and Pearson in 1928 [72], is a hypothesis test used to compare two nested models: a simpler null model and a more complex alternative model. It determines if the additional parameters in the complex model significantly improve the fit to the data. It is a more powerful test than the Wald test, which is used to test whether any linear combination of estimated parameter values is significantly different from a specified null hypothesis, when the proportional hazards assumption

is satisfied. In the field of survival analysis, the LLR is commonly employed to assess the importance of covariates or to compare various prognostic models, providing a straightforward method to determine which model is best for a given dataset.

The test statistic for the LLR is given by:

$$LLR = 2\left(\ell_0 - \ell_1\right), \tag{2.31}$$

where ℓ_0 is the log likelihood of the data under the base model we want to compare to and ℓ_1 is the log likelihood of the data under the alternative model. Under the null hypothesis (assuming the simpler model is true), LLR approximately follows a chi-squared distribution with degrees of freedom equal to the difference in the number of parameters between the two models.

The Akaike Information Criterion (AIC) is a metric used for model selection, developed by Hirotugu Akaike [73], in the case of Cox Regression model since its the case of a partial likelihood, the AIC is also partial, partial Akaike Information Criterion (pAIC). The pAIC is a valuable tool for comparing models, especially when there's no clear hypothesis about which model is best. It provides a balance between model complexity and the goodness-of-fit, helping to select the most appropriate model for the data at hand. Given a statistical model with some estimated parameters, the pAIC is defined as:

$$\text{pAIC} = 2k - 2\ln(\ell), \tag{2.32}$$

where k is the number of estimated parameters in the model and ℓ is the maximised value of the partial log likelihood function for the estimated model.

Furthermore, the Prognostic Index (PI) is recognised as an useful assessment method, serving as a significant statistical measure used to predict clinical outcomes based on certain pre-treatment factors. The PI is a scalar value derived for each individual in a study, often computed as a linear combination of their covariates weighted by the coefficients from a survival model.

The PI_i for the i-th individual is typically defined as:

$$PI_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}, \tag{2.33}$$

where x_{ij} is the j-th covariate for the i-th individual and β_j is the coefficient of the j-th covariate from a fitted survival model.

Individuals can be stratified into high and low risk groups based on a threshold value of the prognostic index. Commonly, the median prognostic index value is used as the threshold. Those with a prognostic index above this threshold are categorised as high risk, and those below as low risk. The log-rank test is then employed to assess the differences in survival between these

groups.

A particular study illustrated the construction of a PI model aimed at predicting prognosis, employing Kaplan-Meier estimation in conjunction with the log-rank test to contrast the overall survival difference between low-risk and high-risk patient groups [74]. Additionally, another study explored the statistical methodologies for the external validation of a published Cox model against other models using the PI alongside Kaplan-Meier curves for risk groups, thereby demonstrating the intertwined evaluation of these survival analysis models [75].

Simultaneously, the Concordance Index (C-Index) is a widely adopted metric for appraising the predictive accuracy of survival models. It's a measure used to evaluate the discriminatory capability of survival models [76]. Essentially, the C-Index assesses the model's ability to correctly rank order individuals based on their survival times or event risks (Fig 2.8). For a set of survival times, the C-Index is computed as:

$$C = \frac{\text{number of concordant pairs}}{\text{number of permissible pairs}}. \tag{2.34}$$

A pair is considered in the equation if one of the two subjects has an event. It's termed concordant if, among the two subjects in the permissible pair, the one with the higher risk score (from the model) actually had the event first. The C-Index ranges between 0.5 and 1.0: $C = 0.5$ means that model's predictions are no better than random guessing and $C = 1.0$ means that the model's predictions are perfect. In survival analysis, the C-Index is especially valuable as it accounts for censoring. When comparing two subjects, if one is censored before the other has an event, the pair is not included in the denominator of the formula since it's not clear how their true survival times compare.

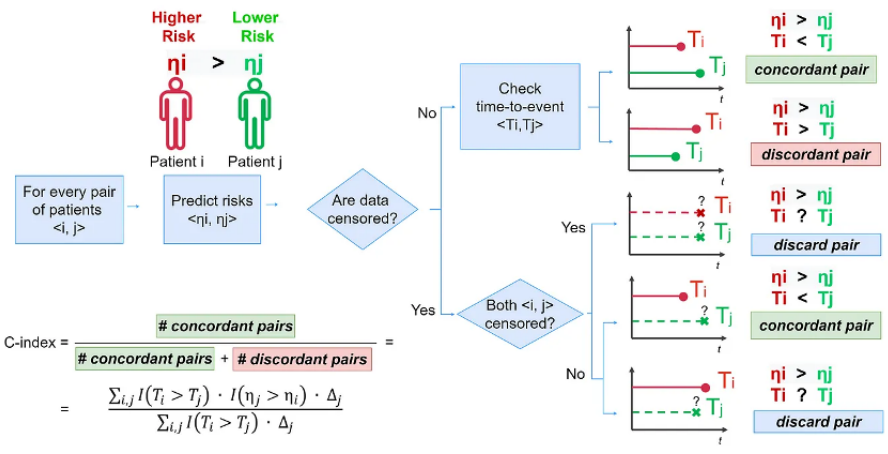


Figure 2.8: Illustration showing how the C-Index is calculated, indicating which pairs are used as concordant and discordant in the computation and which are discarded. From [3].

Through the combination of these evaluation methods, a robust scheme can be made to evaluate the performance of survival models such as Kaplan-Meier estimations and Cox models, thereby anchoring the reliability and validity of the findings derived from survival analysis in cancer research.

3

Methodology

Contents

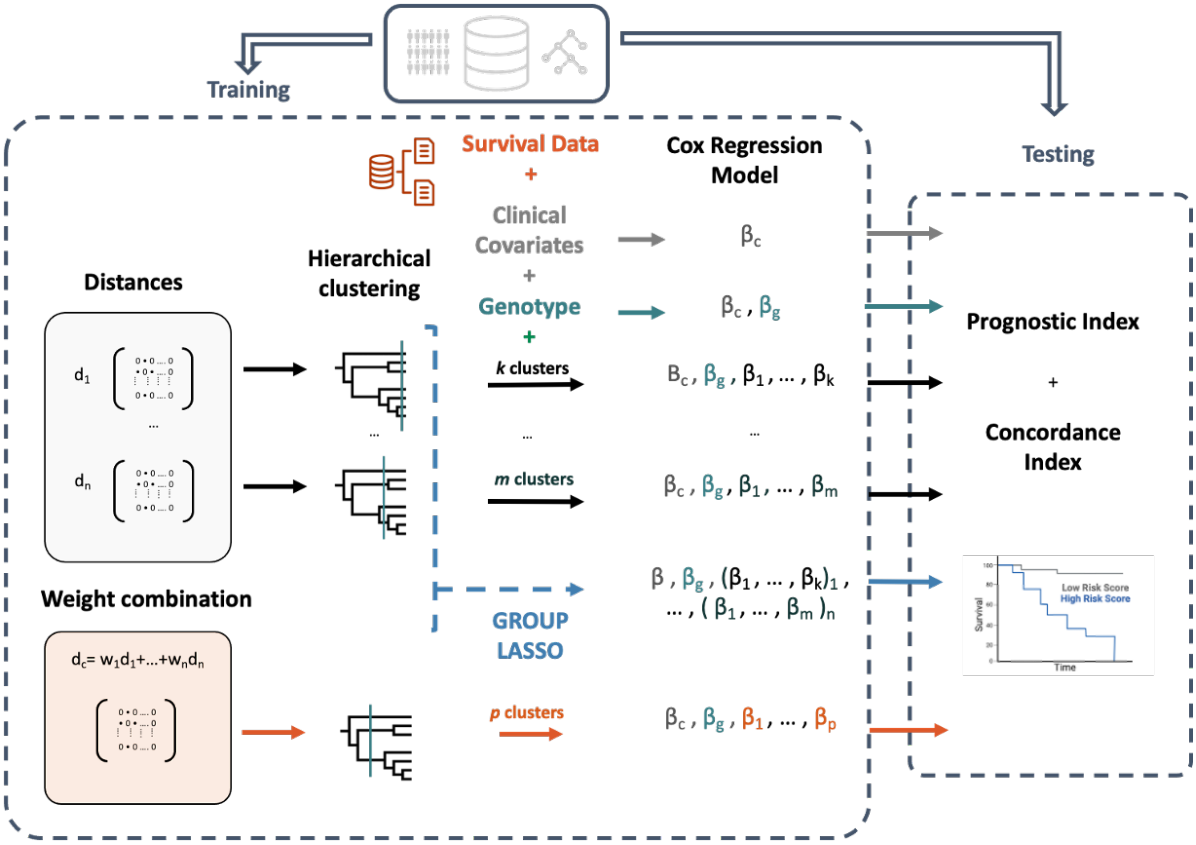


Figure 3.1: Schematic Illustration of the Thesis' Methodological Framework.

3.1 Data availability and processing

In this thesis, two data sets were considered: Breast Cancer and Acute Myeloid Leukemia (AML). The `anytree` library in Python was used to represent the tumour trees from data sets from both cancers.

3.1.1 Breast Cancer

We utilised tumor trees processed and deduced by Luo et al. [77], drawn from the breast cancer data in [78], bulk sequencing data, and employing phylogenetic trees from SPRUCE [32]. The preprocessing completed by Luo et al. [77]confined the analysis to mutations observed in a minimum of 10% of patients. This resulted in 19 mutations across 1152 patients with 1232 phylogenetic trees, where trees from identical patient samples were given equal weight.

The dataset's clinical data encompassed patient details on age, tumor grade and stage, hormone receptor status, vital status, and overall survival in months.

3.1.2 Acute Myeloid Leukemia

Based on the data from Morita et al. [79], which included 154 samples from 123 AML patients with discernible somatic mutations via scDNA-seq, containing a total of 31 mutated genes, we used tumor trees processed by Luo et al. [77]. Our study's tumor trees were inferred through SCITE [6], a tool that leverages Bayesian inference to generate mutation trees aligned with observed single-cell genotypes. These trees were special due to the fact that they allowed parallel mutations, thereby violating the ISA. Nodes that were repeated in the tree became clonal nodes, which meant they included ancestral mutations. On the other hand, nodes that shared the last mutation and had a Jaccard similarity greater than 50% experienced mutation shifts to already existing mutations. The number of distinct mutations increased from the initial set of 31 to 53 as a result of this clonal node modification.

Clinical data encompassed details like age, gender, Maximum Cancer Cell Fraction (maxCCF), treatment classifications (Tx_group), specific AML treatment regimens, AML subtypes (Diagnosis), vital status, and overall survival in days calculated from the difference between 'Service Date' and 'Last Contact Date'.

3.2 Distance Metrics

In our implementation, several distance metrics were employed; some of which were implemented by us, while others were employed from existing packages.

1. **Implemented Metrics**:
    - We implemented the metrics introduced by Govek et al. [11], namely PCD, AD, PD, and CD in Python, following the distance equations specified in Eqs. 2.1 , 2.2, 2.3 and 2.4, respectively.
    - Additionally, we included and implemented the metrics outlined by Jahn et al. [9], specifically BD, 1-Bourque Distance (1BD), and 2-Bourque Distance (2BD) in Python described in section 2.1.3.H.
2. **Employed Packages**:
    - The MLTD approach by Karpov et al. [8] was utilized, with its implementation available at `https://github.com/khaled-rahman/MLTED`.
    - From the study by DiNardo et al. [12], we employed two novel distance metrics, CASet) and DISC, accessible at `https://bitbucket.org/oesperlab/stereodist`.

- We also used the MP3 metric for mutation trees from Ciccolella et al. [10], which is available at `https://github.com/AlgoLab/mp3treesim`.

Through the combination of our implemented metrics and the employed packages, we were able to carry out a comprehensive analysis as part of our methodology.

3.3 Clustering

3.3.1 Hierarchical Clustering

This research utilised Hierarchical Clustering via the `scipy` library, specifically employing the `linkage` function from `scipy.cluster.hierarchy`. The Ward method, introduced by Ward et al. [80], was chosen for its efficacy in minimising the total within-cluster variance. The Ward method, an agglomerative clustering technique, starts with each data point as a separate cluster and iteratively merges them. The objective is to minimise the total within-cluster variance, or the squared Euclidean distance between points in a cluster. The process is mathematically defined as:

$$\Delta = \sum_{i=1}^{n} \sum_{j=1}^{n} ||x_i - x_j||^2, \tag{3.1}$$

where Δ is the within-cluster variance, n is the number of data points, and x_i and x_j are individual data points within the cluster.

3.3.2 Optimal Number of Clusters

This study utilised a modified silhouette score from `sklearn.metrics` to determine the optimal number of clusters. The Weighted Silhouette Score (WSS) is calculated as follows:

$$WSS = \sum_{j=1}^{K} \left(\frac{n_j}{N} \cdot \text{median}\{s(i) : i \in C_j\} \right), \tag{3.2}$$

where K is the number of clusters, n_j is the number of patients in the j-th cluster, N is the total number of patients, C_j is the set of samples in the j-th cluster, and $s(i)$ is the silhouette score of the i-th sample, in Eq. 2.24. This metric adapts the traditional silhouette score by incorporating information about cluster size, thereby providing a more nuanced evaluation of cluster quality.

The optimal number of clusters, k^*, was determined using a two-step approach. First, the differences between consecutive WSS values, $\Delta WSS(k) = WSS(k) - WSS(k-1)$, were smoothed using Locally Weighted Scatterplot Smoothing (LOWESS) for each possible number

of clusters k. The LOWESS smoothing [81] was applied using the `lowess` function from the `statsmodels` library in Python, it performs a locally weighted polynomial regression, providing smoothed values $\hat{y}_k$:

$$\hat{y}_k = \text{LOWESS}(\Delta WSS(k);\ \text{fraction} = 0.66,\ \text{iters} = 3), \tag{3.3}$$

where *fraction* refers to the fraction of data used when estimating each y-value, and *iters* is the number of residual-based reweightings to perform. The first cluster number for which the smoothed value fell below a specific threshold, k_{LOWESS}, was identified:

$$k_{\text{LOWESS}} = \min\{k : \hat{y}_k < \text{threshold}\}, \tag{3.4}$$

where $\hat{y}_k$ is the LOWESS-smoothed Δ WSS(k). Then, k_{LOWESS} was compared with k_{WSS}, the cluster number with the highest WSS, and the minimum was selected as the optimal number of clusters:

$$k^* = \min(k_{\text{LOWESS}}, k_{\text{WSS}}). \tag{3.5}$$

3.3.3 Clusters Analysis

The Jaccard index is a metric employed to quantify the similarity between two sets. In the context of clustering, it measures the extent of overlap between two partitions of a dataset [82].

A common way to visualise the agreement between two partitions is through a mismatch matrix. The matrix, as shown in Table 3.1, tabulates the number of unique pairs of data points that are grouped similarly or differently between two clusterings.

Table 3.1: Mismatch matrix for two different partitions where a, b, c and d represent the amount of unique pairs in the partitions.

Number of pairs	In the same cluster	In different clusters	Sums
In the same cluster	a	b	$a+b$
In different clusters	c	d	$c+d$
	$a+c$	$b+d$	M

The Jaccard index, defined for two sets A and B, is given by the equation:

$$\text{Jaccard}_{AB} = \frac{a}{a+b+c}. \tag{3.6}$$

The index reflects the ratio of the number of pairs that remain in the same cluster across both partitions (a) to the total number of pairs that are either together in one partition but separated in the other or vice versa ($a+b+c$). A higher Jaccard index indicates greater

similarity between the two partitions. An online tool was used to calculate this index (http://www.comparingpartitions.info/index.php?link=Tool).

3.4 Survival Models

3.4.1 Cox Regression Model and Group Lasso Regularisation

We applied the Cox Proportional Hazards Model for our survival analysis tasks using the open-source Python `lifelines` library.

Additionally, we chose to use the R package `grpreg`, specifically the `grpsurv` function, to perform Cox regression modelling with Group LASSO regularisation.

3.4.2 Model Comparison and Assessment

The `lifelines` library in Python was heavily utilised by us for a variety of survival analysis calculations.

The partial Akaike Information Criterion (pAIC), which is essential for model comparison and selection in survival analysis scenarios, was computed using the library's `CoxPHFitter` function.

Additionally, the partial Log-Likelihood (pLog-Likelihood) required for the Log Likelihood Ratio (LLR) computation was also obtained using the `CoxPHFitter` function. For the LLR, we established two reference models: the first incorporated only clinical data, and the second combined both clinical and genotype data. The subsequent p-value calculation was facilitated through the `scipy.stats` package, utilising the chi-squared distribution. For adjusting the results in the context of multiple comparisons, the Bonferroni correction was manually applied by multiplying the p-value by the number of comparisons being executed.

For the assessment of the Concordance Index (C-Index), which was used to evaluate the models in the testing data, we once again utilised the `CoxPHFitter` function from the `lifelines` library.

The Prognostic Index (PI) was computed using the β output from the `lifelines`, adhering to the previously defined equation 2.33.

Lastly, to conduct the log-rank test, an essential tool for comparing survival curves, we employed the relevant function from the `lifelines` library.

3.5 Optimised Weight Combination

The DE algorithm is conveniently implemented in the `scipy.optimize` library in Python. The function 'differential_evolution' provides a straightforward interface to optimize a given objective function.

The main parameters of DE are F (=(0.5, 1)) which is a scaling factor that regulates the amplification of the differential variation when dithering is used on a tuple, as in our case. Generation by generation, dithering modifies the mutation constant in an unpredictable way. U[min, max] provides the mutation constant for that generation. CR (=0.7) which corresponds to the crossover probability, that determines the chance of elements from the mutant vector being selected over those from the original candidate.

The DE strategy used was 'randtobest1bin' and initialisation used was 'Sobol'. Both the 'randtobest1bin' strategy and the Sobol sequence are advanced techniques that can enhance the performance of Differential Evolution. The 'randtobest1bin' strategy offers a directed mutation approach combining the best solution's influence with random perturbations. Simultaneously, the Sobol sequence provides a more structured initialisation, ensuring a broad and uniform exploration of the solution space right from the start.

The `differential_evolution` function, when applied, mandates inputs of the objective function and variable bounds, set to [0,1] for all metrics. It outputs the best-found solution alongside its objective function value. Our target was minimising the pAIC derived from the Cox model and outputs were the minimal pAIC found and a list of the optimal weights, as depicted in Fig. 3.2.

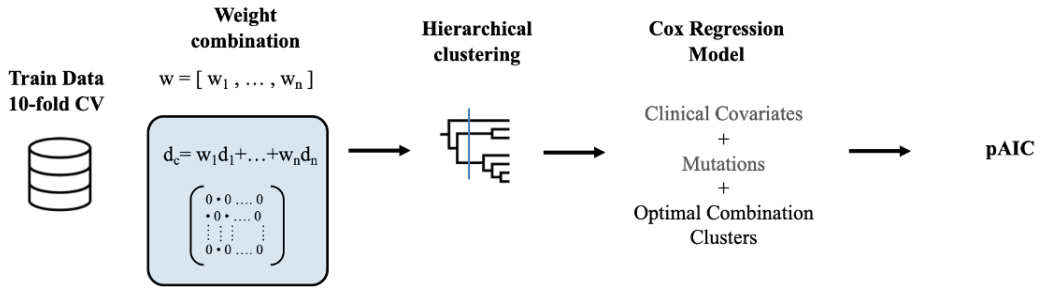


Figure 3.2: Objective function used on the DE to find the minimal output.

The code of the methods and functions are available at https://github.com/laurabquintas/CombinedDistances.

4

Results and Discussion

Contents

4.1 Results

4.1.1 Breast Cancer

4.1.1.A Mutation Trees and Distance Metrics

The mutation trees used in our research were derived from bulk sequencing data. These trees were curated to retain only those mutations observed in at least 10% of the patient cohort. The data set comprised 1232 mutation trees corresponding to 1152 patients, with some patients contributing multiple trees due to repeated sample extractions.

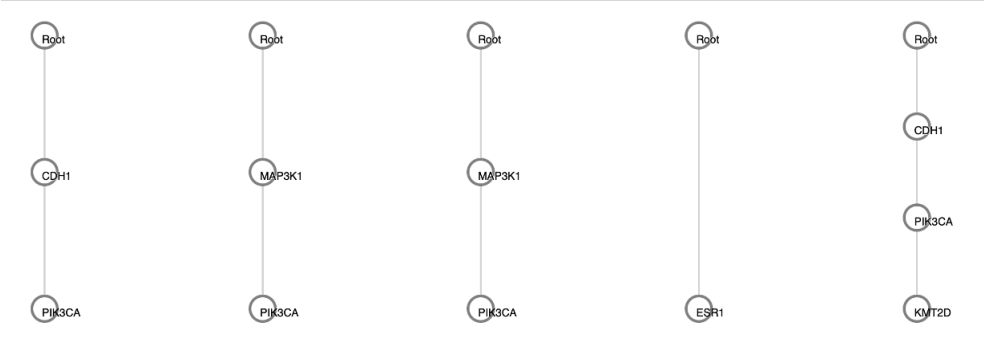


Figure 4.1: Mutation trees examples from the Breast Cancer data set.

This selective inclusion criterion resulted in relatively compact mutation trees, exemplified in Fig. 4.1, with an average of merely 2.9 nodes per tree. The data set encompassed a total of 19 distinct mutations, featuring genes such as *PIK3CA*, *NF1*, *ESR1*, *GATA3*, *MAP3K1*, *KMT2D*, *KMT2C*, *FOXA1*, *RB1*, *EPHA7*, *TSC2*, *RHOA*, *PIK3R1*, *PRDM1*, *PBRM1*, and *CD79A*. Notably, several of these genes, including *TP53*, *PTEN*, and *CDH1*, have been previously associated with elevated risk, and *NF1* with moderate risk, in Breast Cancer, as documented in [83].

The distance metrics employed in this data set included DISC ∪, DISC ∩, CASet ∪, CASet ∩, MLTD, BD, 1BD, 2BD, PD, PCD, AD, and CD, all of which are detailed in Chapter 3.

Upon application of these metrics to our trees, we generated a 1052x1052 distance matrix for each metric. In these matrices, values ranged from 0 (indicating similar trees) to 1 (indicating dissimilar trees), with the diagonal representing the distance between identical patients (self-distance).

Fig. 4.2 displays selected examples of the distributions of distance metrics between patients, including CASet ∩, DISC ∩, and PD (other metrics histograms shown in Fig. B.1). The histograms reveal a visible distribution of similar trees, yet they are predominantly characterised by dissimilarity. This significant presence of dissimilarity within the data suggests promising potential for the formation of well-defined, distinct clusters.

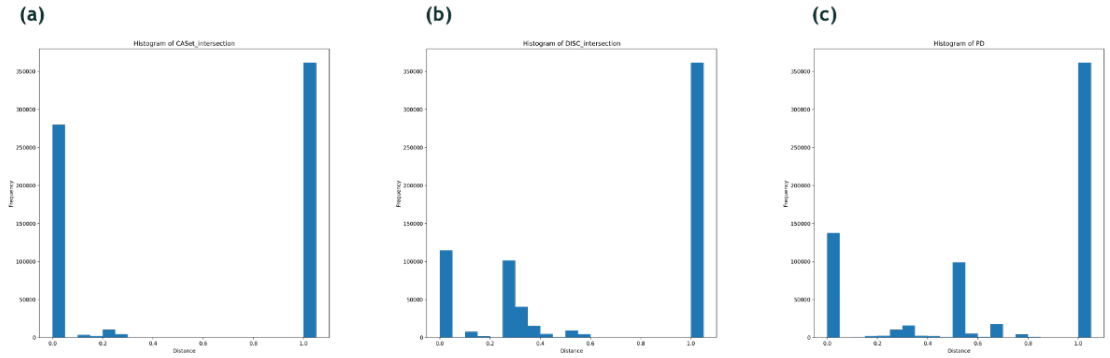


Figure 4.2: Distance distributions applied on the Breast Cancer mutation trees (a) CASet ∩, (b) DISC ∩, and (c) PD.

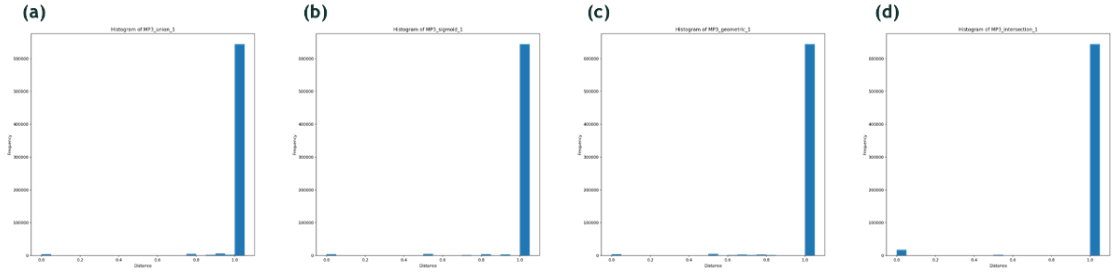


Figure 4.3: Distance distributions applied on the Breast Cancer mutation trees (a) MP3 ∪, (b) MP3 σ, (c) MP3 *geo* and (c) MP3 ∩.

The distance metric MP3, in all its variants (union, intersection, sigmoid, and geometric), relies on triplet comparisons. However, with the average node size in our data being less than three, these metrics resulted predominantly in uniform outputs, mostly ones, as illustrated in Fig. 4.3. Consequently, these distances were deemed unsuitable for our analysis on this data set and were thus excluded.

4.1.1.B Clustering Analysis

In their investigation of the correlation between distance metrics and clinical data, Ciccolella et al. [10] used hierarchical clustering, a method also reproduced by other studies employing simulated data to compare various methodologies [8, 12].

Drawing inspiration from this approach, we similarly applied hierarchical clustering in our analysis. To determine the most suitable linkage method, we limited the number of potential clusters to a maximum of 25 and selected the method that, on average, yielded the highest Silhouette Score across all distances, shown in Table 4.1. The Ward linkage method, which minimises cluster variation, emerged as the best choice, demonstrating a higher Silhouette score

Linkage Method	Silhouette Score
Average	0.40
Complete	0.36
Single	0.18
Ward	0.53

Table 4.1: Analysis of linkage methods in hierarchical clustering, showcasing average silhouette scores for optimal clusters.

compared to the other methods, and was therefore selected for our analysis.

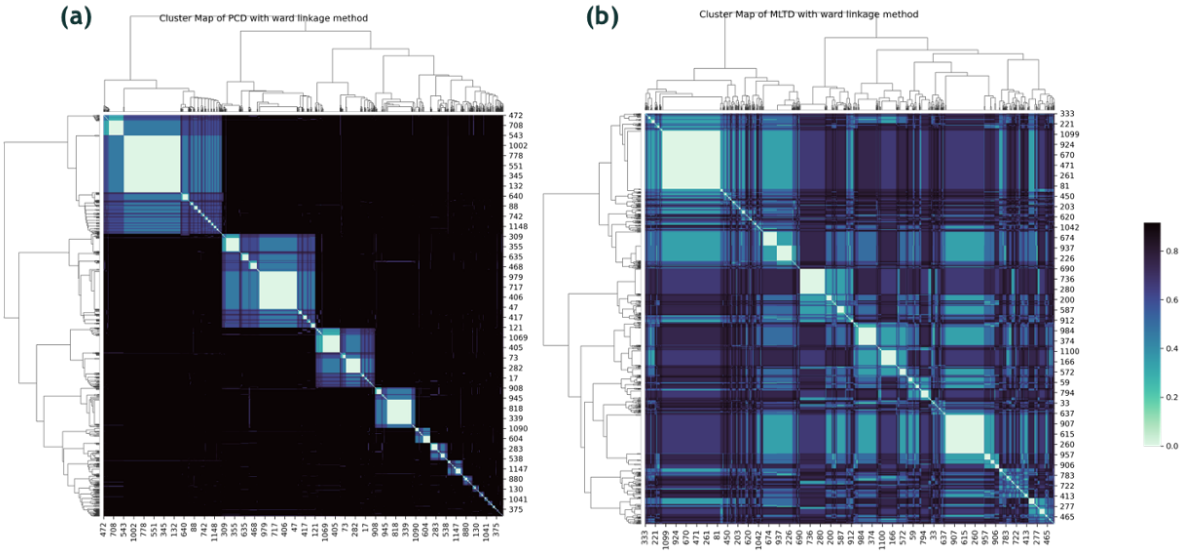


Figure 4.4: Cluster maps of the hierarchical clustering with Ward linkage method for (a) PCD and (b) MLTD.

We employed a cluster map to visually assess the clusters and their cohesion (Fig. B.2), featuring a dendrogram and a distance matrix with a colour gradient from light blue (denoting similar trees) to dark blue (indicating dissimilar trees). Fig. 4.4 presents two examples using the PCD and MLTD metrics. One example clearly depicts well-differentiated clusters, as evidenced by stark contrasts between groups (outlined in black) and high similarity within the clusters (lighter colours). Conversely, the other example reveals less cohesion among clusters, with discernible separation only in smaller segments near the diagonal.

After that, as shown in Fig. 4.5, we conducted a comparison analysis between the dendrograms—which were created using a variety of distance metrics—and the relevant clinical and genotype data. A detailed examination revealed a correlation within the cluster of trees positioned on the right side of the dendrogram. This specific cluster exhibited a pronounced association with category III High Grade in the Overall Tumor Grade and with Receptor Statuses that are not marked as HR+/HER2- and a less obvious relationship was also found with

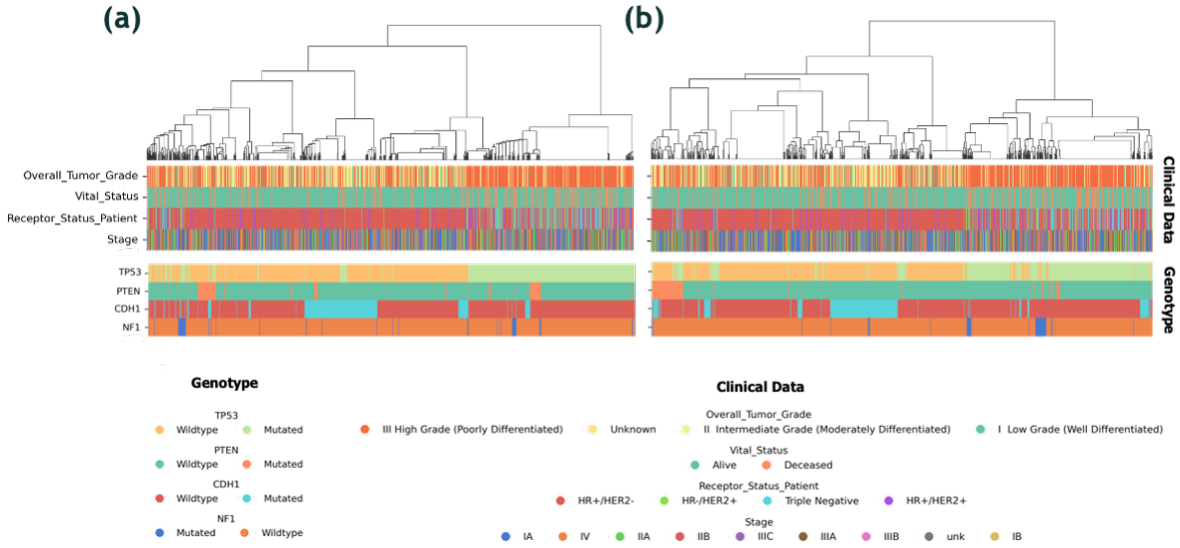


Figure 4.5: Dendograms from hierarchical clustering using the Ward linkage method, juxtaposed with relevant clinical data (Overall Tumor Grade, Vital Status, Receptor Status, and Stage) and genotype data (TP53, PTEN, CDH1 and NF1). Panels (a) and (b) represent the results for PCD and MLTD metrics, respectively.

the Deceased Vital Status. A similar observation applies to the *TP53* and *CDH1* mutations, with a clear clustering pattern for patients carrying these genetic alterations. Furthermore, a correlation was discerned between TP53 and the aforementioned clinical markers — the High Grade and non-HR+/HER2- statuses. This suggests the possibility that these associations could be attributed solely to the presence of this mutation, rather than the topological configurations of the trees as well which requires more analysis.

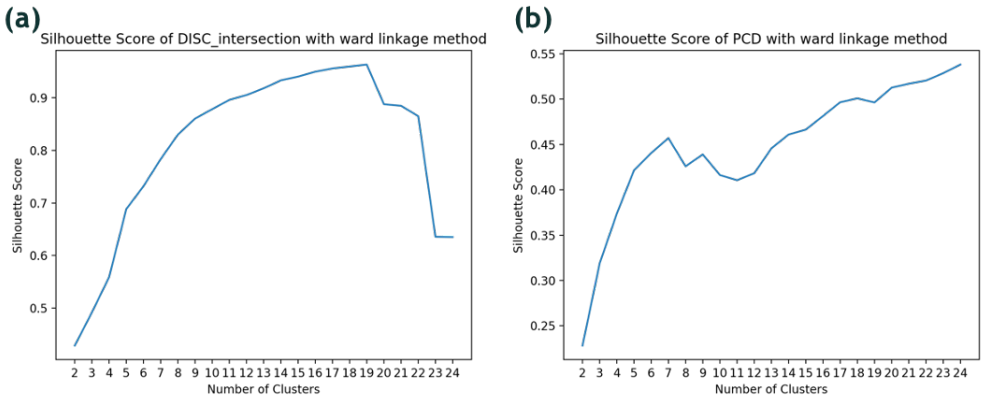


Figure 4.6: Silhouette Score for a max of 25 clusters for (a) DISC ∩ and (b) PCD

To determine the optimal number of clusters, prior studies have relied on the silhouette score, a metric assessing clustering quality through measures of cohesion and separation [10,

12]. However, our analysis presented a unique challenge; for most metrics, the silhouette score consistently increased as more clusters were permitted, as depicted in Fig.4.6 (b). In instances where this trend did not hold, there was a rise to a substantial number of clusters, accompanied by only marginal improvements in the score with each additional cluster, as shown in Fig.4.6 (a). It's important to note that we also considered other evaluation metrics, such as the Calinski Harabasz score and the Diversity index. However, these metrics exhibited a bias toward the minimum number of clusters permitted, as they are grounded in assessing variation differences. Due to this inherent bias, we did not use these metrics in our analysis.

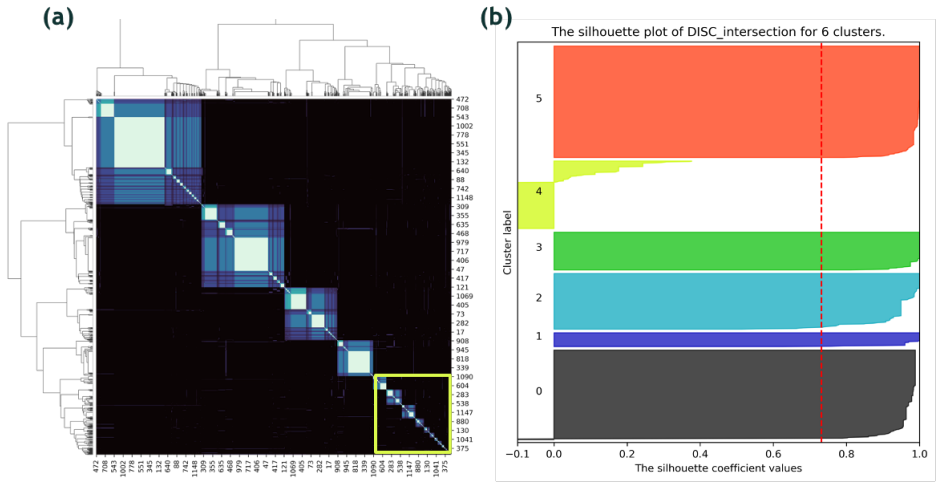


Figure 4.7: Close Analysis of the Silhouette Score for 6 clusters for DISC ∩ (a) cluster map and (b) silhouette analysis.

We examined in detail the silhouette scores associated with each clusters form by our metrics, as illustrated in Fig. 4.7. According to our analysis, a single, distinct cluster made up of small, dissimilar clusters are the cause of the low silhouette score for fewer clusters as well as the slight increases seen with the addition of each additional cluster. To reduce the influence of this smaller cluster on the overall score, we introduced a modification to the conventional silhouette score calculation. Instead of the standard method, which averages the silhouette scores of all samples, our adjusted approach incorporated a weighting factor based on the number of patients in each cluster (Eq. 3.2). In order to avoid situations where the ideal cluster count would be one with clusters with an unsuitable small number of patients, such as four, as was happening when choosing the cluster number with the highest traditional silhouette score, this adjustment was made with the intention of lessening the overall impact of smaller clusters on the overall score and reducing the score increases contributed by these clusters.

This modification produced notable improvements in the overall score, specially for lower numbers of clusters, particularly evident in Fig 4.8. This change effectively amplified the influ-

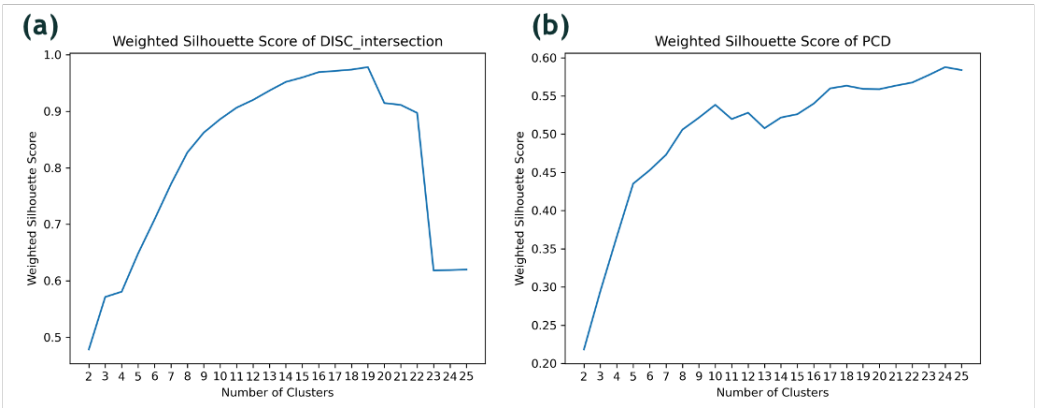


Figure 4.8: Weighted Silhouette Score for a max of 25 clusters for (a) DISC ∩ and (b) PCD

ence of higher scores from larger cohesive clusters. For instance, in the case of PCD, the score for nine clusters ascended above 0.5, a significant increase from the previous sub-0.45 value (Fig.4.8 (b) and Fig. 4.6 (b), respectively). Despite this improvement, selecting a large number of clusters continued to present the challenge of a steady increase in scores with the addition of clusters.

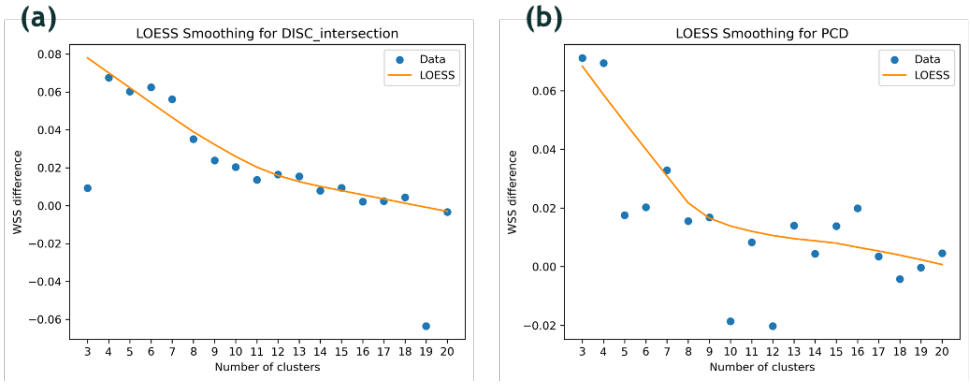


Figure 4.9: LOWESS smoothing applied to WSS difference for (a) DISC ∩ and (b) PCD

To navigate this issue, we applied a LOWESS smoothing technique to the differences between the weighted silhouette scores (WSS) for x and $x-1$ clusters. We established a threshold (t), selecting the number of clusters at the point where the smoothed values fell below this predefined value. This strategy effectively culminated in a configuration where one cluster ended up essentially aggregating the numerous smaller clusters of patients that made the small increases in the score. Although we explored alternative strategies, such as applying LOWESS specifically to the newly added clusters WSS, the outcomes closely mirrored those of the more straightforward method previously described. Consequently, we opted to proceed with the simpler approach.

To choose the threshold t, we examined the number of clusters for different values of t and

the respective smoothed regressions, Fig. 4.9 and Fig. B.4. We observed a consistent trend: the difference remained stable below the 0.02 threshold. Table B.1 specifically enumerates the number of clusters determined for each metric at the different thresholds. It illustrates that the additional number of clusters remains at a 25 increase for thresholds from $t = 0.04$ down to $t = 0.02$, beyond which the difference jumps to 59.

Table 4.2: Number of clusters in relation to the respective WSS for selected metrics, using a LOWESS threshold of 0.02.

	No. Clusters	WSS
DISC ∪	10	0.38
DISC ∩	12	0.92
CASet ∪	10	0.48
CASet ∩	4	0.72
BD	10	0.41
1BD	8	0.4
2BD	9	0.43
MLTD	3	0.27
AD	8	0.32
PD	7	0.53
CD	7	0.47
PCD	9	0.52
Optimal Combination	8	0.38

This stability, coupled with a close comparison of the clusters against the WSS for the different thresholds, guided us to select a threshold of $t = 0.02$, presenting the results for the number of clusters chosen and the respective WSS in Table 4.2. Importantly, this decision was also guided by the need to preserve valuable information inherent in the distance metrics, a risk that increases with a low number of clusters.

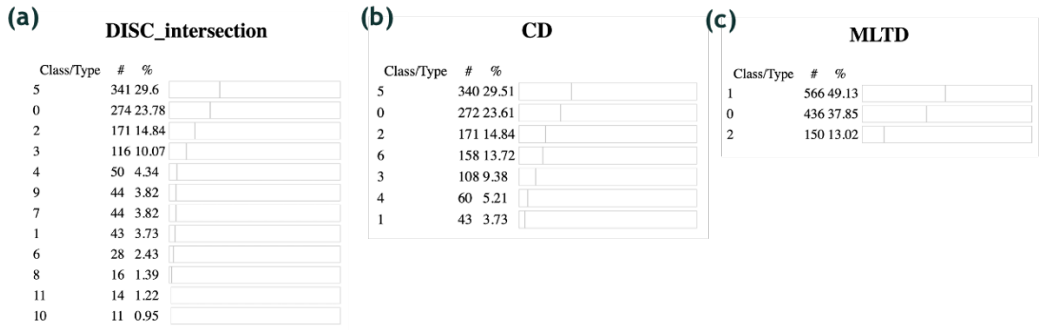


Figure 4.10: Cluster Patient Count of the distance with the highest number of cluster (a) DISC ∩, an example of average number of clusters (b) CD, and the lowest number of clusters (c) MLTD.

Fig. 4.10 illustrates the distribution of patients across clusters for three distinct distance metrics (other metrics in Fig. B.5). Notably, CD showcases an average number of clusters,

indicating our success in preventing excessively small clusters for most distance metrics. This is further exemplified by MLTD, the metric with the lowest number, which features clusters encompassing a substantial number of patients. However, DISC ∩ stands as an exception, presenting the highest number of clusters, including three that comprise low number of patients, around 1% of the total. Despite this, the overall distribution reflects a satisfactory cluster count per distance metric.

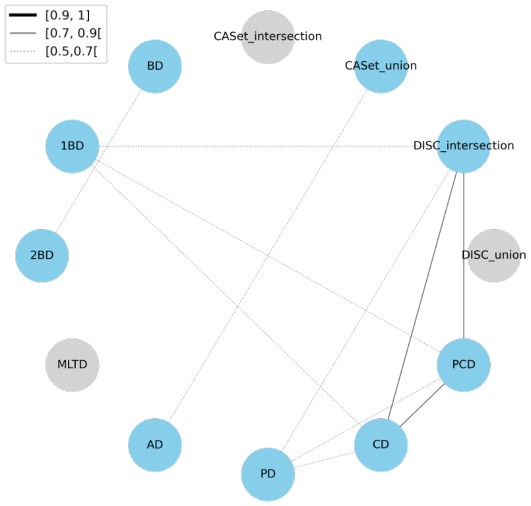


Figure 4.11: Relevant Jaccard Index correlations between the clusters of the distance metrics.

Having established the optimal number of clusters, we proceeded to analyse the inter-correlation of clusters across different distance metrics and their association with clinical and genotype data, using the Jaccard Index for this purpose. Fig. 4.11 shows the Jaccard correlations among clusters derived from the distance metrics in Table B.2, revealing a pronounced correlation specifically among DISC ∩, CD, and PCD.

The underlying reason for this substantial overlap can be attributed to the metrics' simplicity and the trees' limited size. These factors contribute to a high correlation score, particularly near the leaves in the case of DISC, making the parent-child relationships and subclones across trees more reliant on the mutations themselves rather than their sequence or structural variances.

However, it's important to note that the Jaccard index didn't uncover any significant correlations between the clusters of the metrics and the clinical or genotype data, Table B.3 and B.4, respectively. We looked into the relation between genotypes and clusters further as a result of the lack of anticipated correlations.

To achieve a more granular understanding, we constructed bar plots for each cluster an example is shown in Fig. 4.12 for PCD, focusing on mutations present in at least 50% of the patients within a given cluster (other metrics in Fig. B.6). These plots not only highlighted

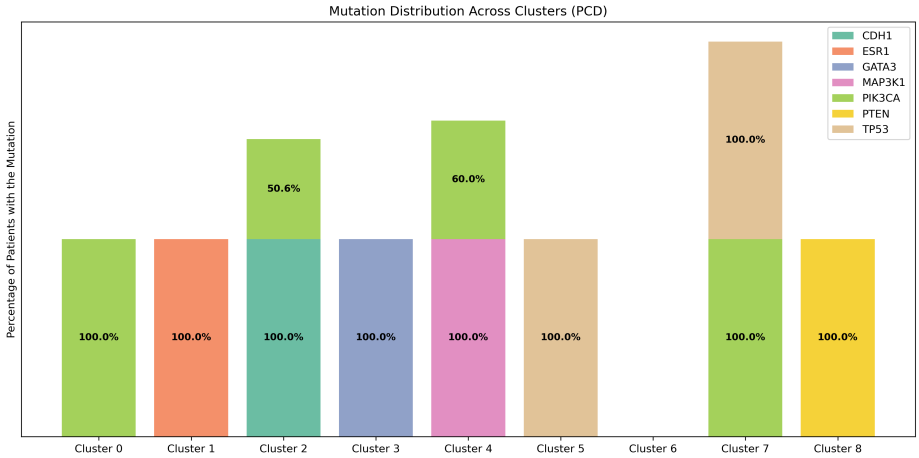


Figure 4.12: Mutations present in over 50% of patients, distributed across clusters, for PCD.

the prevalent mutations but also indicated their respective percentages within each cluster. The detailed visualisation facilitated a more thorough assessment of the genotype data within the framework of our established clusters, shedding light on patterns that were not immediately discernible from the preliminary correlation metrics.

Expected correlations between the organisation of clusters, as defined by distance metrics, and the incidence of specific mutations were found when the mutation distribution across clusters was examined. For some distance metrics it could only be seen this correlation related to the existence of the mutations but for others we were also able to uncover a new layer of association. Such as for the case of the PCD the joint occurrence of certain mutations — *TP53* with *PIK3CA* — within a large cluster of patients. This pattern provides more information than just identifying the mutations; it highlights their potential collaborative significance in patient groups, bringing another angle to our knowledge that may be pertinent to patient survival.

Additionally, we identified a specific cluster characterised by a diverse mix of mutations, none of which individually appeared in more than 50% of the patients within those clusters. This lack of a dominant mutation is indicative of the heterogeneity within these clusters, contributing to the lower silhouette score discussed before.

4.1.1.C Survival Analysis Using Cox Models

In our survival analysis, we sought to determine whether the clusters formed through these distance metrics could provide significant insights correlated with patient survival. We compared these clusters against Clinical Data and Genotype — encompassing clinical data and mutation information — to discern if they offer additional prognostic value.

Within the clinical data, we selected 'Overall_Tumor_Grade' and 'Invasive_Carcinoma_Dx_Age'

based on their compliance to the proportional hazards assumption.This selection was important because the Group Lasso algorithm employed does not accommodate stratification for variables violating this assumption. Furthermore, these specific data were also chosen for their combined contribution to the optimal partial Log-Likelihood value.

Regarding mutation information, we incorporated all mutations featured in the data set's mutation trees, encompassing a total of 19 distinct mutations.

For this analysis, we divided our data set 70% for Cox Model fitting and 30% for model evaluation. We proceeded with the application of the Cox model to our set of covariates, which included 'Only Clinical', 'Genotype', and each individual distance metrics clusters combined with clinical and mutation data, as well as, Group Lasso and Group Lasso CV regularised models.

Table 4.3: Cox regression model outcomes displaying pAIC and partial Log Likelihood scores for training data. In the pAIC column, green indicates scores superior to the clinical pAIC, while grey denotes inferior results. For the Log Likelihood column, blue signifies scores better than the clinical value, with grey marking those that are worse.

	pAIC	pLog-Likelihood
Only Clinical	1795.1	-893.6
Genotype	1781.2	-867.6
DISC ∪	1783	-859.5
DISC ∩	1788.2	-860.1
CASet ∪	1775.5	-855.7
CASet ∩	1780.5	-864.2
BD	1789.6	-862.8
1BD	1787.2	-863.6
2BD	1787.2	-862.6
MLTD	1784.3	-867.1
AD	1774.8	-857.4
PD	1785.8	-863.9
CD	1786.3	-864.1
PCD	1784.4	-861.2
Optimal Combination	1780.6	-860.3
Group Lasso	1756	-857
Group Lasso CV	1769.1	-854.5

In Table 4.3, we outline the pAIC and pLog-Likelihood for the models fitted, alongside the corresponding number of clusters. The pAIC, essentially a penalised version of the pLog-Likelihood, increases with the number of clusters, with the goal of achieving a low pAIC. Our focus leaned more toward the it because of concerns that the smaller clusters, potentially having a high percentage of deaths (or vice versa), could bias the pLog-Likelihood. This concern became evident when comparing the pLog-Likelihood and pAIC of MLTD and BD — with BD appearing more favourable per the Log-Likelihood yet less so per the pAIC. Therefore, we deemed the pAIC a

more reliable metric for the combination algorithm, which we will elaborate on in the following subsection 4.1.1.D.

Table 4.4: Results of the Likelihood Ratio Test (LLR) for our metrics in comparison to the clinical (LLR Clinical) and genotype models (LLR Genotype), accompanied by their p-value and Bonferroni adjusted p-values (adj p-value).

	LLR Clinical	p-value	adj p-value	LLR Genotype	p-value	adj p-value
Only Clinical						
Genotype	52	6.62E-05	1.06E-03			
DISC ∪	68.2	3.31E-05	5.30E-04	16.2	0.06	0.78
DISC ∩	67	1.21E-04	1.94E-03	15	0.18	1
CASet ∪	75.8	2.77E-06	4.43E-05	23.8	0.005	0.065
CASet ∩	58.8	3.36E-05	5.38E-04	6.8	0.08	1
BD	61.6	2.53E-04	4.05E-03	9.6	0.38	1
1BD	60	1.68E-04	2.69E-03	8	0.33	1
2BD	62	1.43E-04	2.29E-03	10	0.27	1
MLTD	53	1.38E-04	2.21E-03	1	0.61	1
AD	72.4	2.95E-06	4.72E-05	20.4	0.005	0.065
PD	59.4	1.26E-04	2.02E-03	7.4	0.29	1
CD	59	1.43E-04	2.29E-03	7	0.32	1
PCD	64.8	5.99E-05	9.58E-04	12.8	0.12	1
Optimal Combination	66.6	2.05E-05	3.28E-04	14.6	0.04	0.52
Group Lasso	73.2	6.00E-09	9.60E-08			
Group Lasso CV	78.2	3.96E-07	6.34E-06			

To further analyse the models, we employed the Likelihood Ratio test (LLR), a statistical tool designed to compare the fit of two models and assess whether additional covariates enhance the model's data fit. This test was particularly applied to the analysis of models using training data. Our findings, detailed in Table 4.4, indicate that compared to the 'Only Clinical' model, all other models introduce significant features, affirming that distance metrics do indeed contribute additional insights. This improvement was not consistent, though; when these models were compared to the 'Genotype' model, only three approaches—CASet ∪, BD, and Optimal Combination—showed that statistically significant information was added to their p-values, which, after being adjusted using the Bonferroni technique, were higher than 0.05, leaving no significant metrics. These results not only suggest that certain distance metrics provide more valuable information but also highlight specific ones that could be more effective for this type of cancer.

4.1.1.D Combined Metrics and Their Optimisation

In the search for an optimal combination distance metrics, we adhered to the methodology delineated in Chapter 3. Our objective was minimising the pAIC output of the Cox model for reasons previously elaborated. We opted for the Differential Evolution algorithm (DE) over other approaches, chiefly due to its superior time and memory efficiencies. Unlike the brute force algorithm, which was limited by memory only provided a granularity of weight options [0

, 0.25 , 0.5 , 0.75 , 1], another challenge was encountered with gradient-based functions. These functions were notably sensitive to their initial vector, causing the solution space exploration to be constrained, as the pAIC exhibited negligible shifts in response to minor weight modifications. This often resulted in solutions that remained in close proximity to their initial parameters. Conversely, DE, particularly when enhanced with Sobol initialisation, overcame these limitations by producing initial vectors with a more uniform distribution across the space. This approach significantly broadened the search radius, enabling a more thorough investigation into the array of potential solutions.

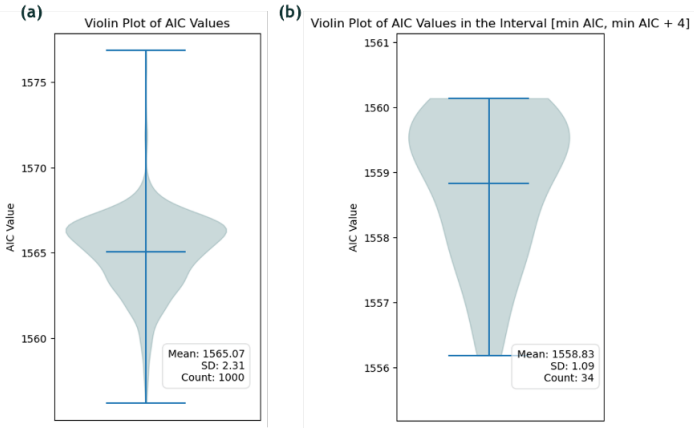


Figure 4.13: Violin plots illustrating the distribution of pAIC values: (a) across all 1000 runs and (b) centered around the minimum pAIC value.

Executing a total of 1,000 runs yielded the pAIC outputs depicted in Fig. 4.13 (a). The necessity of this volume of runs is evidenced by the substantial number of results persisting at a level 10 points above the minimum value identified, with merely 34 results approximating this minimal pAIC (Fig. 4.13 (b)).

Concerned that this combination might merely average the distances of the metrics, thereby failing to extract any substantive information, we proceeded to analyse the weights corresponding to the values near the minimal pAIC, examining their variance (Fig. 4.14). The boxplot of the weights reveals a consistent trend: the algorithm effectively nullifies certain distances while adjusting the weights for a smaller subset, thereby confirming that the combination is not a mere arbitrary average of weights.

Displayed in a radar chart format in Fig. 4.15, the weights given by the DE algorithm for the minimal pAIC conspicuously align with the previously observed trend in the weights' variance for lower pAIC outputs (Fig. 4.14). Most prominent among these weights is CASet ∪, followed by MLTD. This distribution of weights is particularly noteworthy, especially considering that CASet ∪ also secured the lowest p-value of the LLR when contrasted with 'Genotype' data, as

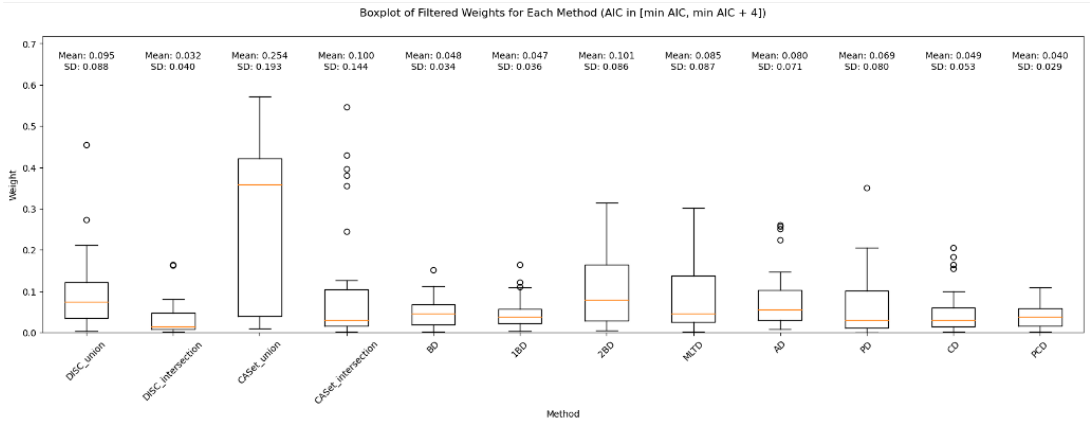


Figure 4.14: Boxplot depicting the variability in metric weights across runs near the minimum pAIC value.

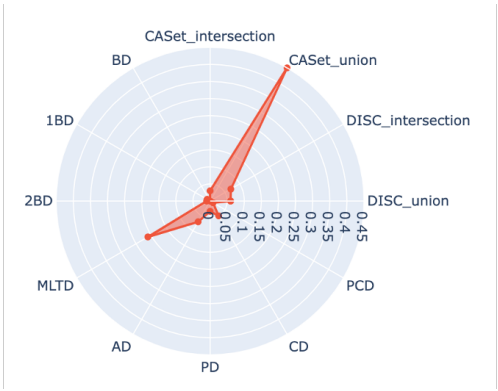


Figure 4.15: Radar chart of the DE distance metrics output weights for the minimal pAIC.

evidenced in Table 4.4. This optimal combination of distance metrics was also one of the few metrics that showed a significant contribution compared to that model.

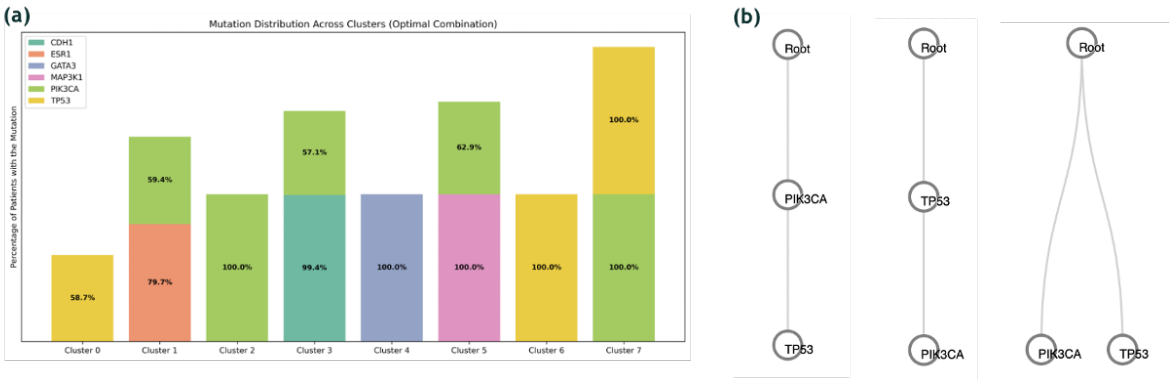


Figure 4.16: Analysis of the clusters of the optimal combination: (a) distribution of mutations across clusters and (b) illustrative trees from the significant cluster of the applied Cox model.

An in-depth analysis of the Cox model parameters revealed that Cluster 7, comprising 87 patients, was distinct in our optimal combination due to having the only significant p-value. This cluster had a hazard ratio of 0.44, suggesting a decreased risk. A closer investigation into the trees of this cluster uncovered a consistent co-occurrence of the *TP53* and *PIK3CA* mutations (Fig. 4.16 (a)). The defining feature of the trees was the exclusive presence of these two mutations, irrespective of their order, as illustrated in Fig. 4.16 (b). This co-occurrence potentially unveils an intriguing relationship between the *TP53* and *PIK3CA* mutations that may be indicative of a favorable prognosis.

4.1.1.E Group Lasso for Metric Importance

An alternative strategy for integrating the unique aspects of each distance metric involved combining all metrics within a single Cox model. However, this direct approach was impracticable due to the high collinearity observed among certain clusters, a complication anticipated during the correlation analysis with the metrics. To navigate this challenge, we opted for a penalised Cox model that adopted an all-or-nothing approach with the clusters of the distances: it either considered all clusters or turned them to zero. This method not only avoided the risk of the model being influenced by smaller clusters but also preserved the model's interpretability. The optimal value of λ was chosen with cross validation by selecting the minimum mean cross-validated error, which in the case of the survival model refers to the Log Likelihood.

The penalisation process distinguished several covariates as nonzero coefficients, among them 'Invasive_Carcinoma_Dx_Age', 'Overall_Tumor_Grade', *PIK3CA*, *PIK3R1*, *RHOA*, *TP53*, *TSC2*, CASet $\cap$, and CASet $\cup$, as illustrated in Fig. 4.17. A closer look to the p-values associated with the model's covariate coefficients revealed an absence of significant impact from CASet $\cap$, possibly suggesting a confounding effect. On the other hand, CASet $\cup$ — previously acknowledged for its informational contribution to the 'Genotype' model — demonstrated significant associations within several clusters (Clusters 2, 5, 7, and 8). Moreover, the mutations *TP53* and *PIK3R1*, along with 'Overall_Tumor_Grade' indicative of III High Grade, were deemed as significant as well.

To mitigate the risk of overfitting, we enhanced our approach by integrating a variant of group Lasso, 'Group Lasso CV', that not only used the optimal λ but also incorporated a 5-fold cross-validation technique. This method aimed to identify the covariates that consistently appeared in more than one fold, thereby ensuring their relevance. This process led to the inclusion of PCD clusters and the *MAP3K1* mutation as potential covariates. However, upon subsequent analysis, these newly included variables did not exhibit statistical significance, with all p-values exceeding the 0.05 threshold.

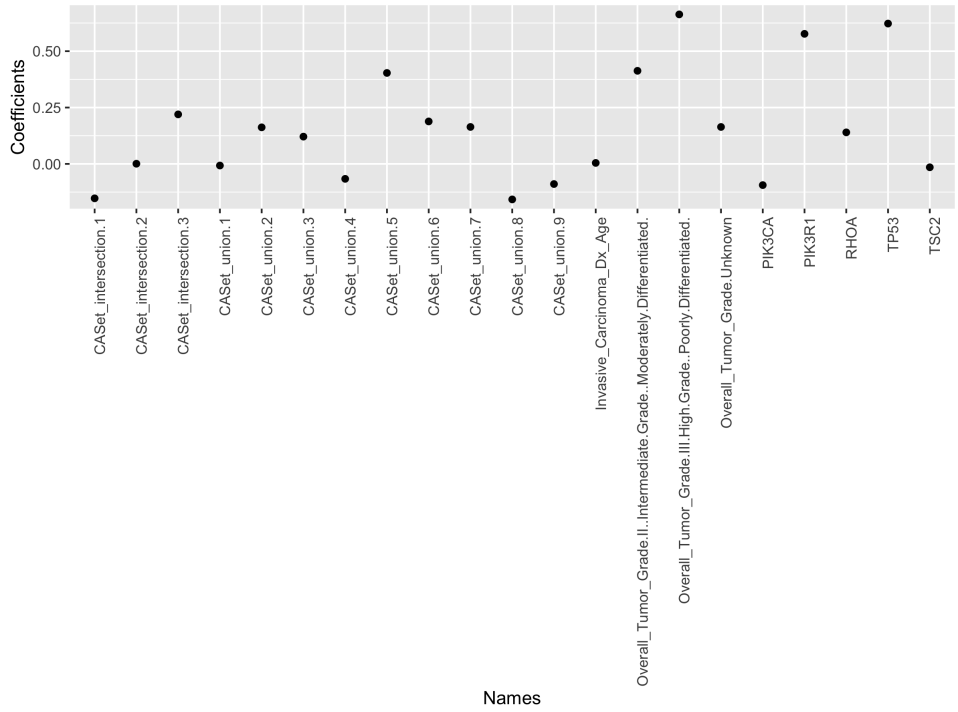


Figure 4.17: Coefficients corresponding to the covariates chosen through Group Lasso in the Cox regression model.

Despite these results, both these approaches demonstrated to have lower pAIC values than the previously discussed models, even the optimal combination, shown in Table 4.3.

4.1.1.F Model Evaluation

To assess these models, we applied them to our test data set that was not used during the model fitting process. Our evaluation metrics included the C-Index, which measures the predictive accuracy of the models, and p-values derived from log-rank tests, which determine the statistical significance of the differences in survival between high and low-risk patient groups. These groups, which comprised half of the patient population each, were equally divided and determined by means of the prognostic index.

Analysis of the pAIC revealed an unexpected contrast: in the validation set, none of the distance metrics alone or in combination with each other outperformed the 'Only Clinical' model, as opposed to the training data where all metrics had previously outperformed it. Regarding the C-Index scores, only the PCD metric exhibited a marginally higher value, with a minimal difference of 0.02, indicating negligible improvement. The C-Index, often utilised for Cox model evaluation, quantifies the model's discriminatory ability — essentially, its capacity to correctly predict which patients will experience the event of interest. However, a minimal increase in the

Table 4.5: Test data results for pAIC and C-Index using pre-fitted Cox models. In the C-Index column, green signifies values superior to the clinical model, while grey indicates those that are inferior.

	pAIC	C-Index
Only Clinical	668.589	0.669
Genotype	738.678	0.648
DISC ∪	766.333	0.631
DISC ∩	760.369	0.655
CASet ∪	762.64	0.641
CASet ∩	748.108	0.643
BD	756.544	0.657
1BD	749.031	0.657
2BD	751.346	0.662
MLTD	741.86	0.644
AD	755.022	0.63
PD	749.363	0.662
CD	749.864	0.654
PCD	752.37	0.671
Optimal Combination	755.605	0.647
Group Lasso	714.777	0.655
Group Lasso CV	736.242	0.653

C-Index suggests that the additional covariates introduced by the distances and combinations do not significantly enhance the model's predictive accuracy compared to the clinical data alone.

Furthermore, the C-Index has a known limitation: it frequently disregards a substantial portion of data, particularly patients who are censored early. To compensate for this shortcoming, we expanded our analysis to include a comparison of survival curves for high-risk and low-risk groups, determined by the prognostic index. We employed the log-rank test to assess the statistical significance of the disparities between these curves, providing a more comprehensive evaluation of patient outcomes over time.

An examination of the survival curves for 'Only Clinical', 'Genotype', 'Optimal Combination' (which displayed the lowest p-value and a discernible divergence between curves), and PCD (notable for the highest C-Index), reveals findings that are not immediately apparent from the C-Index alone (other survival curves can be seen in Fig B.7). As depicted in Fig. 4.18, there's a pronounced distinction when comparing the survival curves of the 'Only Clinical' and other models. However, when these curves are compared with those derived from the 'Genotype' data, the disparities are not sufficiently substantial to conclusively demonstrate that the additional information provided by the distance metrics significantly influences the outcomes within this data set, although a bigger difference can be seen in the survival curves of the 'Optimal Combination', not reflected in the C-Index.

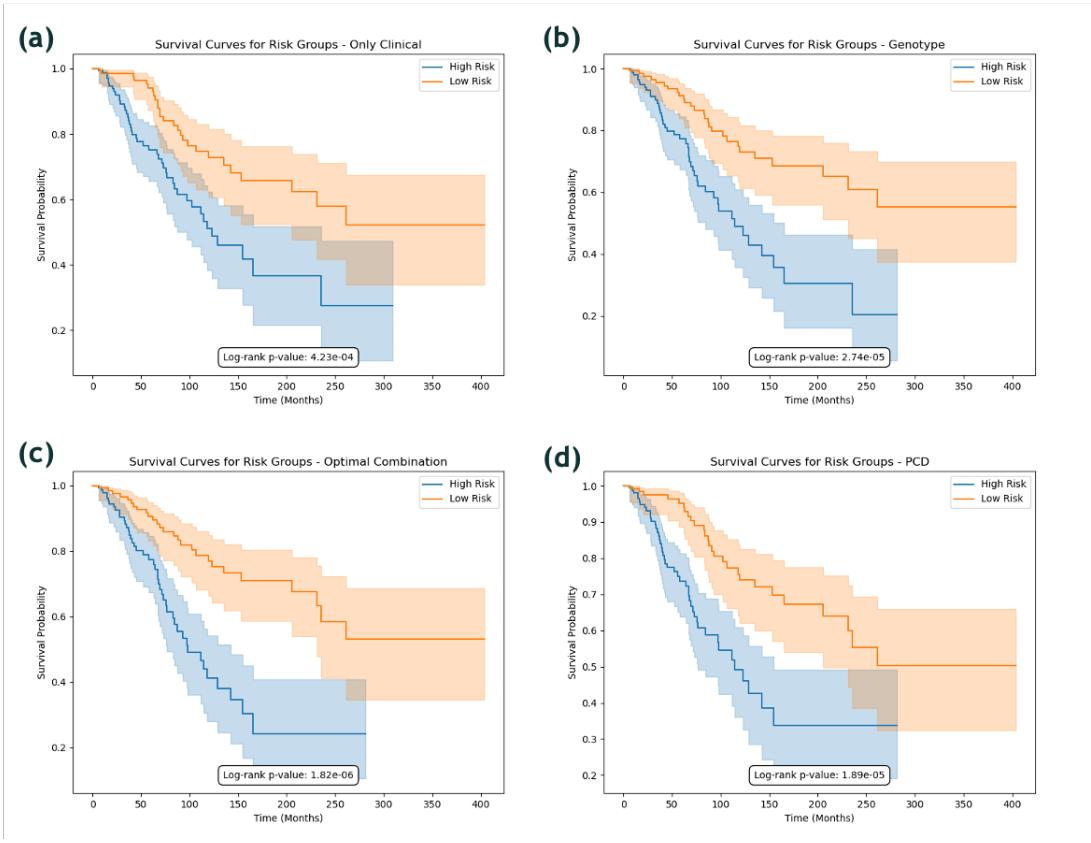


Figure 4.18: Survival curves for the high and low risk for (a) 'Only Clinical', (b) 'Genotype', (c) 'Optimal Combination' and (d) PCD.

4.1.2 Acute Myeloid Leukaemia

We aimed to explore the same research question within the context of a different cancer type, specifically Acute Myeloid Leukaemia (AML), to ascertain whether the same metrics emerge within the combinations and to evaluate the clarity of the contribution provided by these distance metrics. This new data set will undergo an identical sequence of methods and results presentation as previously applied.

4.1.2.A Mutation Trees and Distance Metrics

The mutation trees utilised in our study originated from single-cell sequencing data. The data set comprised of 154 trees from 123 patients, the mutation trees, shown in Fig. 4.19, of this data set had an average 4.8 nodes per tree, higher than the previous cancer data.

Uniquely, these trees defied the Infinite Sites Assumption (ISA) by permitting parallel mutations. As previously outlined, nodes encountering repetition were transformed into clonal nodes, meaning they preserved ancestral mutations. However, certain nodes underwent mutation alter-

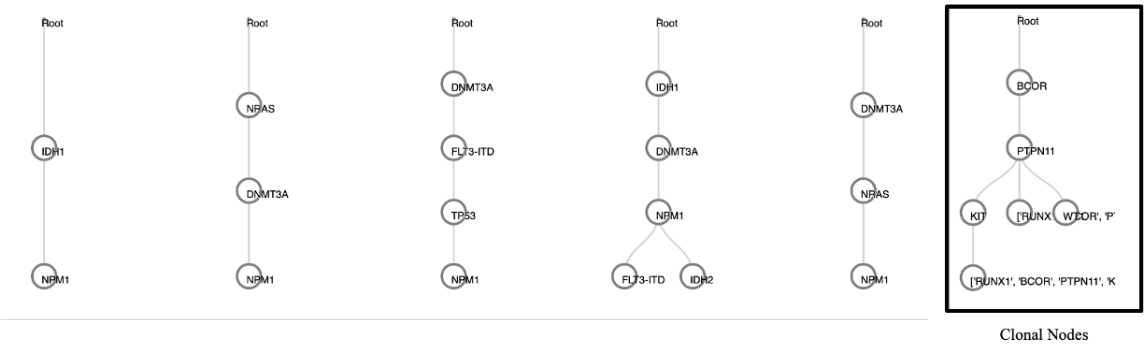


Figure 4.19: Tree Examples of AML data set.

ations if they shared the final mutation and exhibited a Jaccard similarity exceeding 50%. This approach was adopted to recapture potentially lost information consequent to the introduction of clonal nodes. This clonal node modification resulted in an expansion of the original set of 31 unique mutations to a total of 53. Among these 31, several mutations frequently associated with AML were identified, including *FLT3*, *SF3B1*, *TP53*, *SRSF2*, *NPM1*, *ASXL1*, *PTPN11*, *NRAS*, *IDH2*, *IDH1*, *KRAS*, *DNMT3A*, *TET2*, *GATA2*, *KIT*, *STAG2*, *RUNX1*, *EZH2*, and *SMC3* [84]. In contrast to the prior data set, the present analysis incorporated all distance metrics, encompassing every variant of MP3. Again we employed this metrics to our trees and obtained distance matrices 123x123.

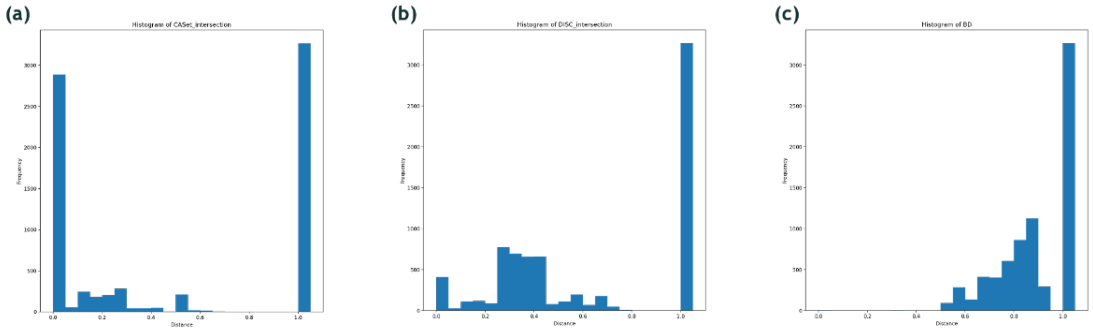


Figure 4.20: Histograms depicting the distance distributions of (a) CASet ∩, (b) DISC ∩ and (c) BD.

An analysis of our metrics' distance distributions, as depicted in Fig. 4.20, reveals a notable leftward skew in most of the histograms, examples being (b) and (c). This skewness towards the higher end of the scale, closer to 1, signifies a prevalence of dissimilarity among patients. In contrast, CASet ∩ presents a more even distribution, suggesting a potential for more effective clustering characterised by distinct separation and internal cohesion.

4.1.2.B Clustering Analysis

After implementing hierarchical clustering using the Ward linkage method, we generated clustermaps that illustrate the dendograms alongside color-coded distance matrices (Fig. 4.21).

Figure 4.21: Cluster maps with Ward linkage method of (a) CASet ∩, (b) DISC ∩ and (c) PCD.

For CASet ∩, the metric that initially seemed promising based on its histogram, the cluster map reveals substantial similarities among samples but lacks clear separation and cohesion among clusters. Conversely, DISC ∩ presents the most visually coherent cluster map, with pronounced differentiation between clusters. PCD exemplifies the cluster maps for the majority of distances, where there is a scarcity of similarities among trees, evidenced by predominantly darker clusters with marginally lighter shades proximate to the diagonal. These outcomes were somewhat anticipated, considering the larger tree sizes, reduced patient count in the data set, and the introduction of clonal nodes (other metrics cluster maps in Fig. C.2).

Upon examining the dendograms in relation to clinical data such as 'PriorMalig,' indicative of patients with a history of malignancy, 'Diagnosis,' denoting the specific type of AML, 'Gender,' and 'VitalStatus' (Fig. 4.22), no striking correlations emerge from this visual scrutiny. A subtle gender-based segregation is slightly perceptible in the BD metric (Fig. 4.22 (b)), but it's minimally discernible.

Additionally, we juxtaposed the dendograms with specific mutations including TP53, FLT3, NPM1, NRAS, IDH2, and KRAS, as shown in the same figure (Fig. 4.22). Notable disparities become evident upon contrasting different distance metrics, particularly with IDH2 appearing predominantly in the first big distinct cluster for DISC ∩ and NPM1 exhibiting a similar pattern but for BD.

In this analysis, we adhered to the previously described methodology, using LOWESS regression on the WSS differences between clusters to determine the optimal number of clusters, setting the threshold at $t = 0.02$. Table 4.6 delineates the optimal number of clusters for each metric, alongside the corresponding WSS values. It's observable that the cluster quantities are

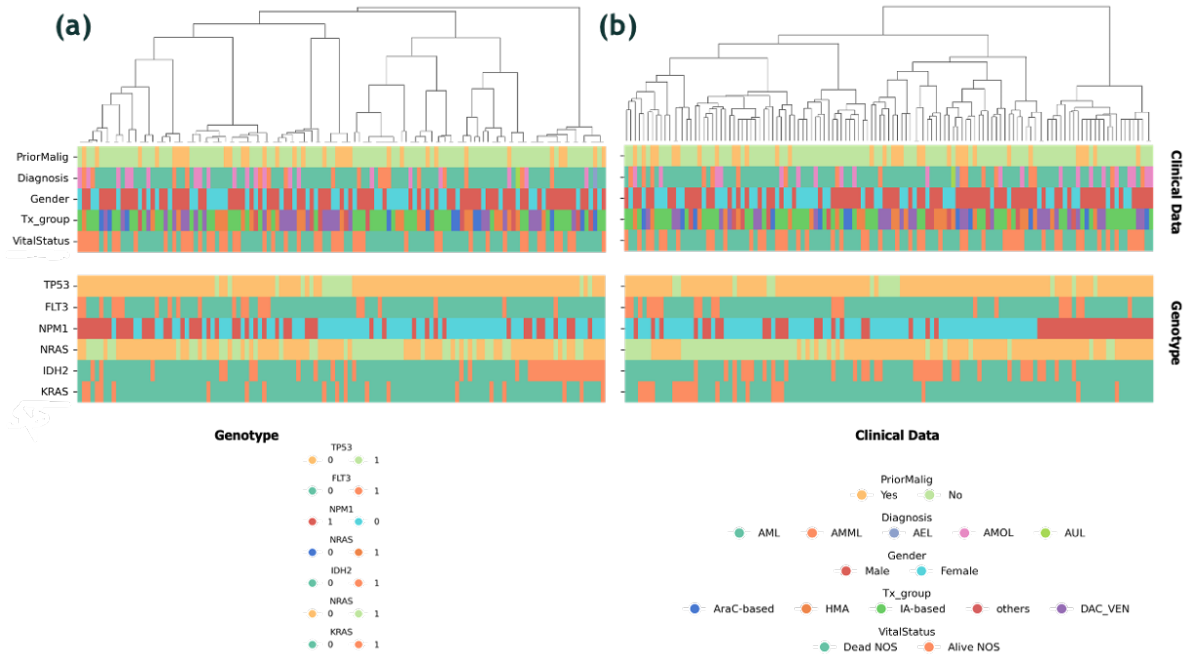


Figure 4.22: Dendograms from hierarchical clustering using the Ward linkage method, juxtaposed with relevant clinical data (PriorMalig, Diagnosis, Gender, Tx_group and VitalStatus) and genotype data (TP53, FLT3, NPM1, IDH2, NRAS and KRAS). Panels (a) and (b) represent the results for DISC ∩ and BD.

Table 4.6: Number of clusters in relation to the respective WSS for selected metrics, using a LOWESS threshold of 0.02.

	No. Clusters	**WSS**
DISC ∪	3	0.05
DISC ∩	14	0.68
CASet ∩	4	0.13
CASet ∪	3	0.03
BD	3	0.1
1BD	3	0.09
2BD	3	0.07
AD	3	0.05
CD	3	0.06
PD	8	0.4
PCD	3	0.06
MLTD	3	0.08
MP3 ∪	3	0.02
MP3 σ	3	0.03
MP3 ∩	3	0.12
MP3 *geo*	3	0.03
Optimal Combination	3	0.14

relatively low, with the WSS predictably low as well, a deduction drawn from the initial inspection of the cluster maps. Exceptionally, DISC ∩ registers a high WSS, a consequence of the

notable cohesion and separation discernible in its cluster map, and it's also the metric that our approach permitted to have the highest number of clusters.

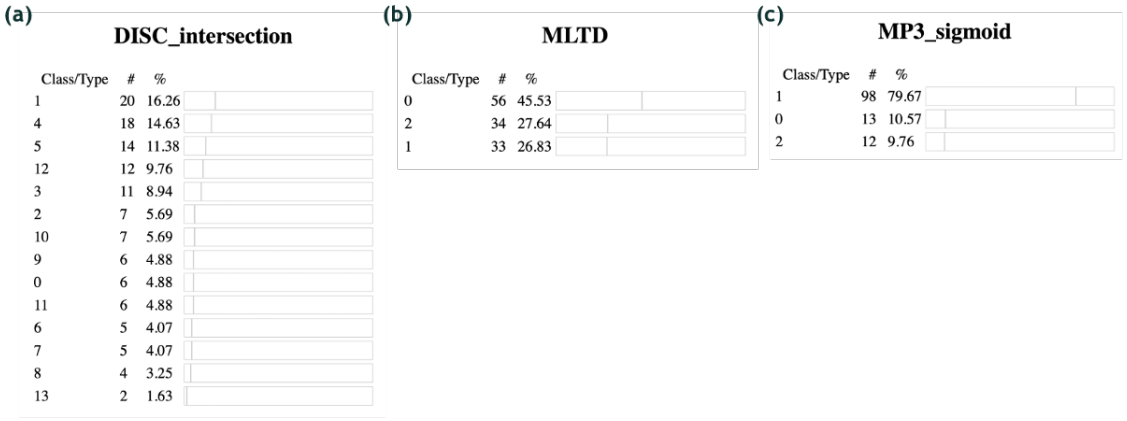


Figure 4.23: Cluster Patient Count of the distance with the highest number of cluster (a) DISC ∩, an example of average number of cluster with a balanced distribution (b) MLTD and with an imbalanced one (c) MP3 σ.

Diving deeper into the analysis of cluster counts, as illustrated in Fig. 4.23, we observe that DISC ∩ (a) culminates in clusters with a mere two patients each, representing 1% — a situation mirroring the ten-patient clusters in our previous data set (other metrics cluster counts in Fig. C.3). In the majority of metrics, which predominantly exhibit just three clusters, there are two primary scenarios. One is manifested by metrics like MLTD (b), which demonstrates an equitable distribution of patients across clusters. Conversely, metrics such as MP3 σ depict a pronounced presence of one extensive cluster with the emergence of two markedly smaller ones.

We proceeded to explore potential correlations among the distances using the Jaccard index, as depicted in Fig. 4.24, with detailed outcomes provided in Table C.1. Notably, a strong correlation reemerged between CD and PCD, both of which are relatively simplistic metrics. Additionally, an anticipated correlation was observed between two variants of MP3 — union and sigmoid. This specific linkage did not extend to other versions, thereby showing a distinct preference for the sigmoid version within the union construct in this particular data set.

We extended the use of the Jaccard index to probe for associations with clinical data, but this yielded no significant correlations (Table C.2). Conversely, when examining genotype data, a notable correlation emerged with numerous mutations specifically for MP3 σ and MP3 ∪ in Table C.3, potentially attributable to the smaller clusters depicted in Fig. 4.23 (c). However, this method did not reveal any considerable correlations for the clusters derived from other metrics.

To uncover the expected correlations between the clusters formed by the metrics and the mutations, we delved into the mutations present in over 50% of patients within each metric's

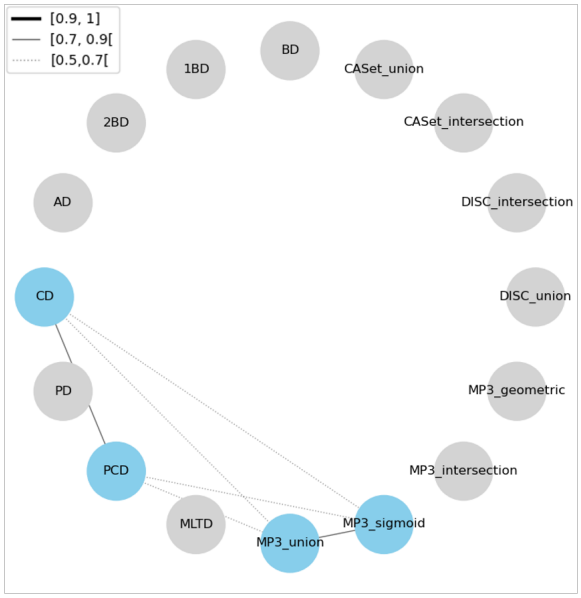


Figure 4.24: Relevant Jaccard Index correlations between the clusters of the distance metrics, highlighted in blue the metrics that had a score higher than 0.5.

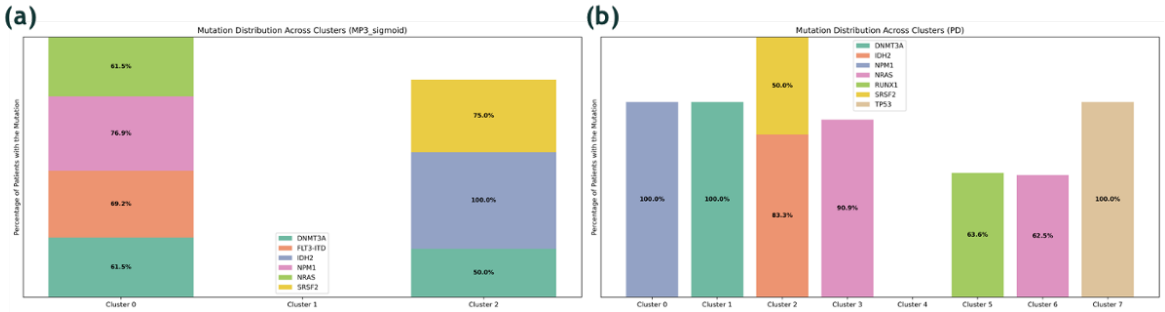


Figure 4.25: Mutations present in over 50% of patients, distributed across clusters, for (a) MP3 σ and (b) PD.

clusters, as illustrated in Fig. 4.25. Fig. 4.25 (a) presents clusters from MP3 σ and confirms our assumption: the smaller clusters exhibit a high percentage of various mutations, likely because a presence in merely six patients is sufficient to achieve a significant percentage. Metrics yielding more uniformly distributed patients across clusters typically revealed one prevailing mutation, alongside a cluster—as also observed in the MP3 example—with a larger patient count that doesn't display any mutation exceeding 50%, indicating a blend of numerous types (other metrics mutation distribution in Fig. C.4). Notably, PD stood out in potentially discovering additional insights, just by looking at its plot, depicted in Fig. 4.25 (b). This metric reveals two clusters with high occurrences of NRAS, suggesting a division in the trees that might be attributed to their topology.

4.1.2.C Survival Analysis Using Cox Models

In the Cox model analysis for this data set, the included clinical data were 'Age', 'Gender', and 'maxCCF' (maximum cancer cell fraction). These variables were chosen because they complied with the Proportional Hazards Assumption and facilitated the highest Log Likelihood for the 'Only Clinical' model. Regarding the 'Genotype' model, it encompassed the aforementioned clinical covariates along with mutations present in no less than 10% of patients. This criterion was essential since the Cox model requires variance in the covariates, which was not achievable for mutations falling below this threshold. Using then a total of 15 mutations in model corresponding to *TP53*, *IDH1*, *SRSF2*, *DNMT3A*, *FLT3*, *PTPN11*, *NRAS*, *IDH2*, *KRAS*, *ASXL1*, *TET2*, *NPM1*, *WT1*, *RUNX1* and *FLT3-ITD*.

Table 4.7: Cox regression model outcomes displaying pAIC and Log Likelihood scores for training data. In the pAIC column, green indicates scores superior to the clinical pAIC, while grey denotes inferior results. For the pLog-Likelihood column, blue signifies scores better than the clinical value, with grey marking those that are worse.

	pAIC	pLog-Likelihood
Only Clinical	388	-191.0
Genotype	394.2	-179.1
DISC ∪	398.1	-179.0
DISC ∩	399.7	-169.8
CASet ∩	399.7	-178.8
CASet ∪	396.7	-178.3
BD	395.2	-177.6
1BD	388.5	-174.2
2BD	397.6	-178.8
AD	395.2	-177.6
CD	397.4	-178.7
PD	400.1	-175.1
PCD	397.1	-178.5
MLTD	394.9	-177.4
MP3 ∪	397.5	-178.8
MP3 σ	396.9	-178.5
MP3 ∩	398.1	-179.1
MP3 *geo*	397.5	-178.8
Optimal Combination	395.8	-178.9
Group Lasso	371.5	-178.7

In contrast to the prior data set, merely two metrics managed to secure an pAIC that was lower or akin to that of the 'Only Clinical' model — these were 'Group Lasso' and 1BD, respectively. This outcome didn't imply an inferior Log Likelihood, as all metrics demonstrated a Log Likelihood that was better. However, the penalty associated with the number of covariates precluded them from surpassing the clinical model. Unexpectedly, our Optimal Combination was unable to identify a combination that showed a lower pAIC, being even worst than some

individual metrics scores.

Table 4.8: Results of Likelihood Ratio test (LLR) for our metrics in comparison to the clinical (LLR Clinical) and genotype models (LLR Genotype), accompanied by their p-values. Metrics highlighted in green and blue demonstrate a superior fit to the data compared to the clinical and genotype model, respectively.

	LLR Clinical	**p-value**	**adj p-value**	**LLR Genotype**	**p-value**	**adj p-value**
Only Clinical						
Genotype	23.8	0.07	1			
DISC $\cup$	24	0.12	1	0.2	0.9	1
DISC $\cap$	42.4	0.03	0.54	18.6	0.1	1
CASet $\cap$	24.4	0.14	1	0.6	0.9	1
CASet $\cup$	25.4	0.09	1	1.6	0.45	1
BD	26.8	0.06	1	3	0.22	1
1BD	33.6	0.01	0.18	9.8	0.01	0.17
2BD	24.4	0.11	1	0.6	0.74	1
AD	26.8	0.06	1	3	0.22	1
CD	24.6	0.1	1	0.8	0.67	1
PD	31.8	0.08	1	8	0.33	1
PCD	25	0.09	1	1.2	0.55	1
MLTD	27.2	0.06	1	3.4	0.18	1
MP3 $\cup$	24.4	0.11	1	0.6	0.74	1
MP3 σ	25	0.09	1	1.2	0.55	1
MP3 $\cap$	23.8	0.12	1	0	1	1
MP3 *geo*	24.4	0.11	1	0.6	0.74	1
Optimal Combination	24.2	0.09	1	0.4	0.53	1

Turning our attention to the Likelihood Ratio test (LLR), with the clinical and genotype models serving as baseline comparisons, the findings are detailed in Table 4.8. Unexpectedly, not even the genotype model showed a discernible increase in information or better fit when compared to the clinical model. The same goes for comparing the models against the genotype only. Important to note, that Group Lasso isn't present because it didn't select all the clinical covariates as it did in the previous data set, therefore was not suitable for LLR calculation.

4.1.2.D Combined Metrics and Their Optimisation

In the optimisation process for this data set, we managed to obtain 992 outcomes, falling short of the intended thousand, potentially due to the Differential Evolution (DE) algorithm reaching its maximum iteration cap of 20,000 without convergence, as illustrated in Fig. 4.26 (a).

The lowest pAIC identified was distinct, with a 10-point gap separating it from the next lowest score. There appeared to be a consistent pattern where the algorithm frequently settled on weights corresponding to pAICs approximately 25 points higher, positioning the minimal pAIC as an outlier rather than a standard (Fig. 4.26 (b)).

When examining the variance among results within close proximity to the minimum—expanded to encompass scores within 'min pAIC + 20' to ensure a substantial sample size, in contrast

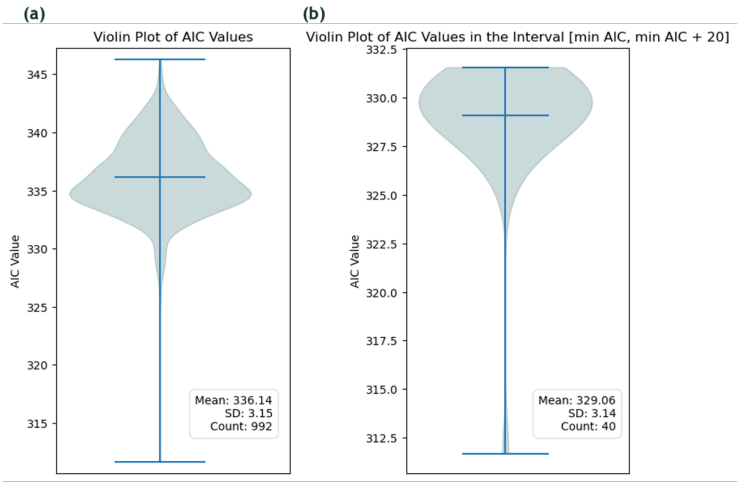


Figure 4.26: Violin plots illustrating the distribution of pAIC values: (a) across all 1000 runs and (b) the closer values to the minimum pAIC value

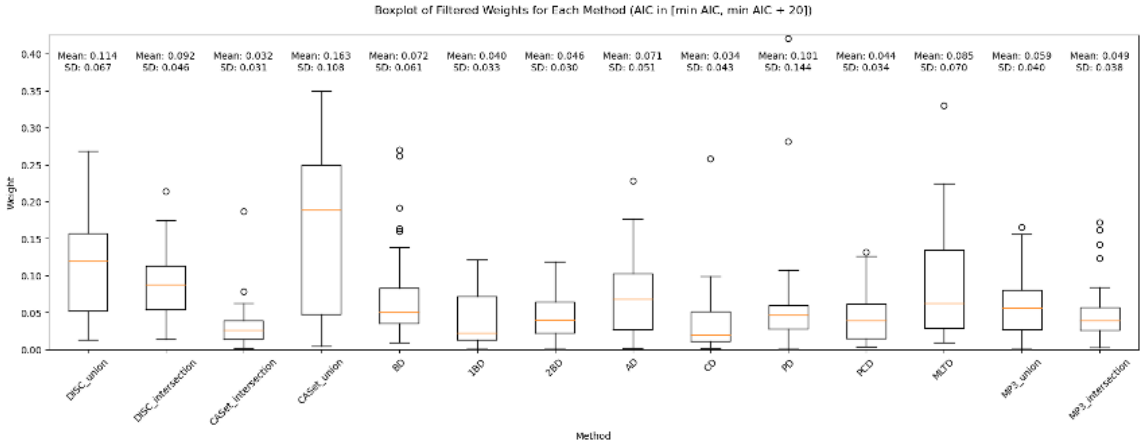


Figure 4.27: Boxplot depicting the variability in metric weights across runs which are the closest to the minimum pAIC value.

to 'min pAIC + 4' in the previous data set—the dispersion in the weights isn't as indicative of certain distances consistently trending toward zero. However, there's an evident inclination for specific weights to register at higher values.

The weight distribution, shown in Fig. 4.28, in the optimal combination with the minimal pAIC deviates from the general trend observed in the variance analysis. Given the substantial difference exceeding 10 points from the nearest AICs, it's plausible that with more results congregating around that optimal pAIC, a more pronounced preference for the PD distance might have emerged. Unfortunately, a detailed examination of the Cox Regression Model revealed that none of the clusters held significant relevance, contributing minimal additional insight. This suggests that for this particular data set, our methodology didn't yield the anticipated

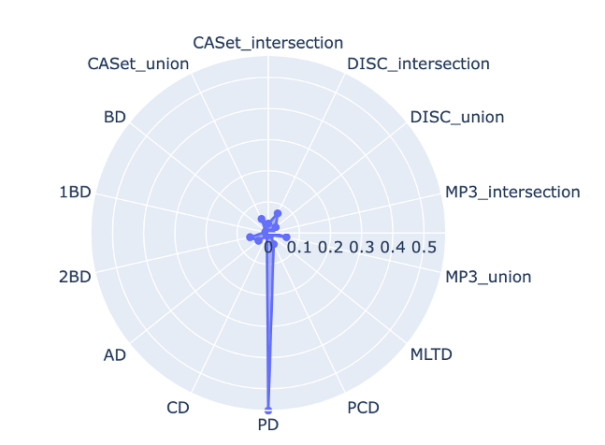


Figure 4.28: Radar chart of the DE distance metrics output weights for the minimal pAIC.

level of success.

4.1.2.E Group Lasso for Metric Importance

Employing the Group Lasso technique on our data set led to the selection of 'maxCCF' from the clinical data, mutations such as *FLT3*, *FLT3-ID*, *NPM1*, and *TP53*, and the 1BD distance metric, as depicted in Fig. 4.29, this time there was no overlap with the DE weight distribution. A meticulous inspection of the Cox regression model pinpointed *FLT3*, *NPM1*, *TP53*, and cluster 2 of BD as bearing significant p-values.

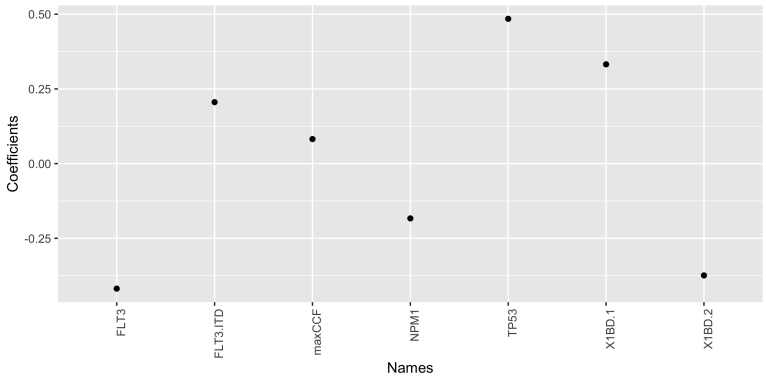


Figure 4.29: Coefficients corresponding to the covariates chosen through Group Lasso in the Cox regression model.

Given that it was the sole method with an pAIC lower than that of the clinical data, we delved deeper into the constituent trees. Fig. 4.30 (a) revealed a notable co-occurrence of *IDH2* and *SRSF2* within cluster 2, correlating with a hazard ratio of 0.29. This implies that patients exhibiting these mutations concurrently tended to have a more favorable prognosis compared to

cluster 0, which was influenced by the *NPM1* mutations. Interestingly, the specific topology and sequential order of these mutations didn't exert an influence, as evidenced by the tree examples in Fig. 4.30 (b). Instances where *SRSF2* appeared first were observed, as well as cases where it followed, negating the hypothesis that the linear arrangement of trees or the co-occurrence alone was impactful, given the diversity in tree types, as illustrated in the final tree example. For this data set, 'Group Lasso CV' was deemed unnecessary due to the complete overlap of selected covariates with the regular 'Group Lasso'.

Figure 4.30: Analysis of the clusters of the 1BD: (a) distribution of mutations across clusters and (b) illustrative trees from the significant cluster of the applied Cox model.

4.1.2.F Model Evaluation

In assessing these models on the testing data in Table 4.9, we found that none surpassed the clinical one in terms of pAIC, not even 'Group Lasso,' which had achieved this in the training data. The C-Index did identify two metrics with superior values, 1BD and DISC ∩, though it's unexpected that Group Lasso didn't score higher despite incorporating 1BD. This discrepancy might be attributed to the exclusion of certain mutations that proved significant in the testing data, as evidenced when revisiting the 1BD Cox model where *NRAS*, omitted by Group Lasso, emerged as significant. The difference in C-Index was more pronounced in this data set, with an increase of 0.058 for 1BD and 0.025 for DISC ∩.

To determine whether these distances facilitated a more effective stratification of patients into high and low risk categories, we examined the Kaplan-Meier curves for these groups, calculated using the prognostic index derived from the fitted Cox model coefficients. However, the log-rank test for these curves yielded no significant p-values for any method, and the survival curves in Fig. 4.31 elucidate this outcome. The 'Only Clinical' model (a) displayed a modest separation, while the 'Genotype' model (b) fared the worst, with numerous instances of curve overlap over time. The 1BD model (d) didn't demonstrate any improvement over the clinical one in terms of separation. Only DISC ∩ showed a noteworthy enhancement in separation, prompting a deeper

Table 4.9: Test data results for pAIC and C-Index using pre-fitted Cox models. In the C-Index column, green signifies values superior to the clinical model, while grey indicates those that are inferior.

	pAIC	C-Index
Only Clinical	143.956	0.561
Genotype	185.607	0.498
DISC ∪	189.548	0.5
DISC ∩	323.555	0.586
CASet ∩	193.234	0.509
CASet ∪	184.331	0.541
BD	197.63	0.496
1BD	189.725	0.619
2BD	190.351	0.502
AD	192.94	0.522
CD	192.527	0.504
PD	218.256	0.474
PCD	191.387	0.53
MLTD	192.263	0.489
MP3 ∪	191.016	0.517
MP3 σ	191.155	0.506
MP3 ∩	189.719	0.502
MP3 *geo*	190.878	0.522
Optimal Combination	189.874	0.498
Group Lasso	158.933	0.541

investigation into its Cox regression model to identify the influential clusters, which were clusters 4 (comprising 18 patients) and 6 (encompassing 5 patients).

Fig. 4.32 illustrates a tree sample for each cluster, including cluster 0, which serves as the reference for calculating the Hazard Ratio (HR). Both clusters 4 and 6 exhibited a HR below 1, indicating a lower risk compared to patients in cluster 0. A commonality among all trees in cluster 4 was the presence of *IDH2* as the initial or one of the initial mutations, whereas for cluster 6, it was *WT1*, in contrast to *FLT3-ID* in cluster 0. This suggests the potential significance of the first mutation in AML prognosis. However, given the small size of these clusters and the data set at large, definitive conclusions remain elusive.

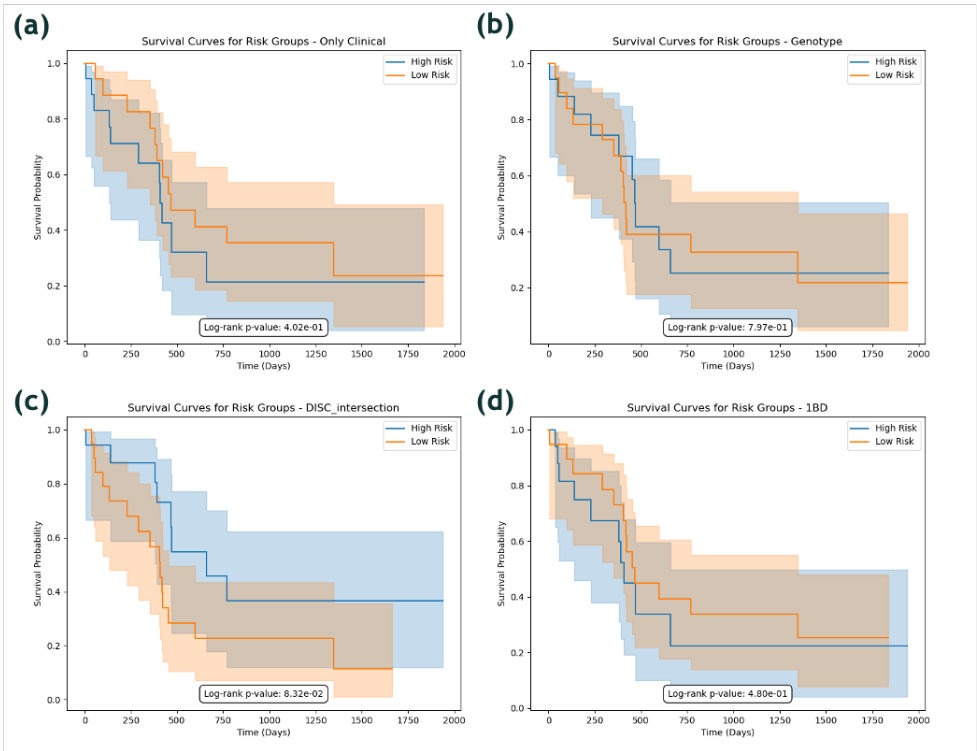


Figure 4.31: Survival curves for the high and low risk for (a) 'Only Clinical', (b) 'Genotype', (c) DISC ∩ and (d) 1BD.

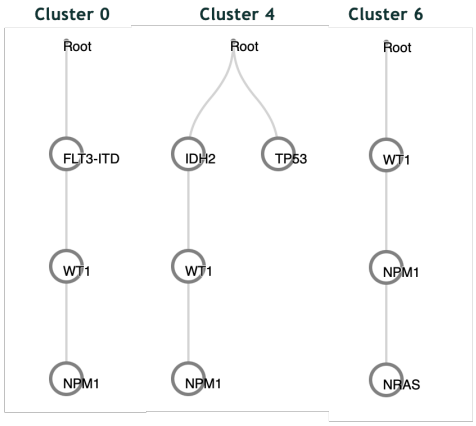


Figure 4.32: Tree examples for the clusters 0, 4 and 6 of DISC ∩.

4.2 Discussion

As mentioned before, this comprehensive analysis was undertaken to explore whether the clustering derived from various distance metrics could enhance our understanding of patient survival across different cancer types, specifically by identifying unique aspects of tumor topology and mutation patterns not fully captured by clinical and genotype data alone. Moreover, we

aimed to discern whether an optimal combination of these metrics, each emphasising different tree characteristics, could more accurately correlate with patient survival, potentially revealing cancer-specific nuances such as particular topologies or mutations indicative of increased risk.

The ongoing search for new approaches to enhance prognostic predictions and inform more individualised treatment plans based on the complex characteristics of each patient's tumour is the motivation behind this. To this end, we delved into an in-depth analysis of two distinct cancer types: Breast Cancer and AML. Our investigations shed light on the complexity of cancer genomics, reinforcing the notion that creating effective patient stratification and prognostic tools is an intricate task, resistant to generic, one-size fits all approaches.

When it comes to patient clustering and the selection of an optimal number of clusters, there are significant insights to be taken. The linkage cluster outputs for Breast Cancer demonstrated enhanced cohesion and separation across various metrics compared to those in AML, a phenomenon potentially more attributable to the larger patient cohort rather than the inherent efficacy of the distances for one cancer type over the other.

Our pipeline was first developed using the Breast Cancer data set in order to avoid the problem of too small clusters that prevented the silhouette score from peaking at a reasonable number of clusters. This data set comprised a substantially larger number of patients compared to AML (1152 vs 123). Consequently, while the methodology the choice of a satisfactory number of clusters with a generally robust WSS for the initial data set, it didn't translate as effectively to the AML context. In AML, the approach drastically reduced the number of clusters for nearly all distances, often to a minimum of just three. This limitation potentially led to an information loss, as the reduced cluster allowance might have restricted the depth of insights the distances could offer. However, certain metrics, such as DISC $\cap$, still managed to yield a significant cluster count, totalling 14 in AML — even surpassing the number identified within the Breast Cancer data set (12).

This brings to light the potential need for a methodology that bypasses the clustering step altogether, thereby preserving the rich information that might otherwise be lost due to the constraints imposed by cluster selection. Such an approach could offer a more comprehensive understanding and use of the data, ensuring that no valuable insights are overlooked in the quest to enhance patient stratification and treatment outcomes.

Now, we shift our focus to both metric combinations strategies, designed to integrate the metrics most suited to each specific cancer type. In our novel application of the differential evolution (DE) algorithm for metric optimisation, we observed notable disparities between data sets. The Breast Cancer data exhibited a consistent trend, where the weights of the minimal pAIC solution aligned with the variance observed in alternative runs. In contrast, the AML

data presented an optimal solution that deviated significantly from the variance pattern of other solutions and boasted a substantially lower value. This discrepancy highlights a critical limitation within our algorithm: despite a broad search distribution, there remains a possibility that the algorithm might miss the true optimal solution, which could potentially be uncovered through an exhaustive exploration of all pAIC values for every possible weight combination. This realisation calls into question whether a more advanced algorithm or an enhanced brute force method, capable of surmounting the existing challenges to enable a more granular solution search, might be necessary.

The Group Lasso method revealed an overlap with the DE combination for the Breast Cancer data set, suggesting the efficacy of the DE approach despite the aforementioned constraints. However, this consistency was not seen in the AML data set, with Group Lasso and DE selecting entirely different metrics—1BD and PD, respectively. The selection of PD, in particular, was surprising, as this metric theoretically holds limited biological significance. For instance, two trees with sequences a-b-c and c-b-a, with a, b and c representing different mutations, would be considered identical under PD, despite their clear differences, pointing out the possibility that the topology of the mutations may not be detected, only their existence.

In this section, we delve into the comparison with clinical and genotype data. For the Breast Cancer data set, every method surpassed the 'Only Clinical' model in terms of pAIC, and the LLR indicated a more precise fit with the inclusion of additional covariates. On the other hand, the AML data set painted a different picture, with only the Group Lasso method improving upon the pAIC of the 'Only Clinical' model, and a mere two individual distances demonstrating a more significant fit compared to both clinical and genotype models. Notably, the optimal combination faltered in these data set, underperforming even individual metrics.

Our evaluative metrics underscored distinct disparities across the data sets, potentially supporting our hypothesis that no 'one-size-fits-all' metric exists; rather, different metrics may better represent different cancer types. In the Breast Cancer data set, the C-Index only noted a marginal improvement with PCD when compared with the clinical model. However, an examination of the survival curves for high- and low-risk patients revealed that, relative to the clinical data, most metrics appeared to enhance patient stratification. This contrast was less pronounced against the Genotype curves, with PCD not offering any discernible enhancement, while the Optimal Combination suggested a subtle improvement. This slight increase in stratification might show the potential importance of the co-occurrence of *TP53* and *PIK3CA* mutations analysed which defies Lin et al. [85] which states that the combination of this two mutations leads to a worst prognosis even than the *TP53* mutant alone, that is the opposite of our findings. However, it's crucial to emphasize that the results didn't unequivocally confirm whether they were

genuinely superior to the 'Only Clinical' and 'Genotype' models.

In the AML data set, the C-Index improved for two metrics, 1BD and DISC ∩, showing a more pronounced enhancement compared to the clinical model than was observed in the Breast Cancer data set. The survival curves, even for the clinical data, did not illustrate a robust separation, with the genotype data behaving even worse. For the metric with the highest C-Index, 1BD, the survival curves did not exhibit a clear enhanced separation, while DISC ∩ indicated a slightly better stratification. Examination of the significant clusters revealed a trend regarding the initial mutation: depending on whether it was *FLT3-ID*, *IDH2*, or *WT1*, the risk was correspondingly higher or lower. This insight could have been also identified by PCD, reinforcing the notion that our choice of clustering and number potentially led to a loss of critical information. Moreover, inferring risk based on this data set for AML might not be accurate, as it could merely be a byproduct of cluster representation in the testing data — a representation that might not be present for other metrics, thereby yielding poorer results upon evaluation.

5

Conclusion and Future Work

Contents

5.1 Conclusion

This study represents a significant advancement in the field of distance metrics analysis, providing a robust framework capable of processing tree data and calculating distance matrices for a variety of metrics. This development paves the way for future, simplified research in this field. Especially in data sets with a large patient population, like the Breast Cancer data set, the strategies developed to navigate the complexities of real-world data—particularly in determining out the optimal number of clusters—have proven to be effective by excluding the very small clusters.

While our exploration into the use of distance metrics clustering aimed to enhance survival predictions for cancer patients, the results, eventhough the metrics offer nuanced insights into tumor evolution, their contribution, individual or in combination, to survival prediction was not significantly superior to that of clinical and genotypical covariates alone. It is crucial to recognise that the results are closely related to the specific characteristics and preprocessing of the datasets that were used, highlighting an important area that requires careful consideration in future studies.

Despite the obstacles encountered, this study successfully challenges the notion of a universal distance metric for diverse cancer types, illustrating the variability in metric efficacy across different datasets. For Breast Cancer, significant metrics included the Optimal Combination and PCD, while for AML, 1BD and DISC $\cap$ proved most effective. This observation supports our theory regarding the viability of developing a combined distance metric, even in the cases where the combination of the metrics did not consistently outperform the individual metrics in terms of prediction.

The study introduces and validates two promising strategies for optimal metric combination based on survival data: the established Group Lasso and an innovative approach employing the Differential Evolution (DE) algorithm. Both methods prove their efficacy, with the DE algorithm demonstrating strategic utilisation of higher-weight metrics which showed good results on training data and avoiding arbitrary combinations; and the use of survival data for both method optimisation was a novel approach that, in theory, holds great promise. However, it is vital to address the limitations of our chosen optimisation algorithm, balancing its extensive search capabilities with the acknowledgement that it may not always reach the true optimal solution.

This thesis provides a logical framework for utilising distance metrics, integrating them with survival data to potentially unveil tumor patterns associated with varying prognoses in the future. Despite inconclusive results from our data sets, the application of patient stratification based on coefficients derived from the Cox Regression model using the clusters from distance

metrics holds promise. While this study establishes a solid foundation for future research, it also underscores the complexity of enhancing cancer survival predictions. These methods have the potential to significantly contribute to personalised medicine, aiding clinicians in selecting the most effective treatment strategies tailored to individual patient characteristics.

5.2 Future Work

Our investigation into mutation trees comparison in cancer prognosis has highlighted key areas for future exploration. One pivotal strategy involves rethinking the reliance on hierarchical clustering. Future studies should consider alternative methods or potentially bypass this step altogether, leveraging techniques capable of utilising the complete information embedded in distance matrices. This approach could unveil subtleties missed by our methodology.

Moreover, a promising avenue for subsequent research is the application of these metrics to cancer types for which a 'ground truth' is established — particularly cancers where specific mutations order are indicative of prognosis variance. Such an approach would offer a good test of whether these distance metrics can genuinely discern critical genomic differences.

In addition, it is paramount for future studies to standardise sequencing approaches or, at minimum, recognise and control for biases inherent in varying methodologies. This standardisation is important for generating reliable outcomes that are applicable across studies and cancer types and is essential to uncover the true value and limitations of these metrics in the realm of personalised cancer prognosis and treatment.

Bibliography

[1] P. Rodriguez-Mier, "A tutorial on Differential Evolution with Python — pablormier.github.io," https://pablormier.github.io/2017/09/05/a-tutorial-on-differential-evolution-with-python/, 2017, [Accessed 11-10-2023].

[2] P. Sedgwick and K. Joekes, "Kaplan-meier survival curves: interpretation and communication of risk," *BMJ*, vol. 347, no. nov29 1, pp. f7118–f7118, nov 2013. [Online]. Available: https://doi.org/10.1136%2Fbmj.f7118

[3] N. C. Albanese, "How to Evaluate Survival Analysis Models — towardsdatascience.com," https://towardsdatascience.com/how-to-evaluate-survival-analysis-models-dd67bc10caae, 2022, [Accessed 25-08-2023].

[4] "What is Cancer?" https://www.cancer.gov.

[5] "Cancer Research UK," https://www.cancerresearchuk.org/about-cancerl.

[6] K. Jahn, J. Kuipers, and N. Beerenwinkel, "Tree inference for single-cell data," *Genome Biology*, vol. 17, p. 86, 12 2016.

[7] C. A. Ortmann, D. G. Kent, J. Nangalia, Y. Silber, D. C. Wedge, J. Grinfeld, E. J. Baxter, C. E. Massie, E. Papaemmanuil, S. Menon, A. L. Godfrey, D. Dimitropoulou, P. Guglielmelli, B. Bellosillo, C. Besses, K. Döhner, C. N. Harrison, G. S. Vassiliou, A. Vannucchi, P. J. Campbell, and A. R. Green, "Effect of mutation order on myeloproliferative neoplasms," *New England Journal of Medicine*, vol. 372, no. 7, pp. 601–612, feb 2015. [Online]. Available: https://doi.org/10.1056%2Fnejmoa1412098

[8] N. Karpov, S. Malikic, M. K. Rahman, and S. C. Sahinalp, "A multi-labeled tree dissimilarity measure for comparing "clonal trees" of tumor progression," *Algorithms for Molecular Biology*, vol. 14, p. 17, 12 2019.

[9] K. Jahn, N. Beerenwinkel, and L. Zhang, "The bourque distances for mutation trees of cancers," *Algorithms for Molecular Biology*, vol. 16, p. 9, 12 2021.

[10] S. Ciccolella, G. Bernardini, L. Denti, P. Bonizzoni, M. Previtali, and G. D. Vedova, "Triplet-based similarity score for fully multilabeled trees with poly-occurring labels," *Bioinformatics*, vol. 37, pp. 178–184, 4 2021.

[11] K. Govek, C. Sikes, and L. Oesper, "A consensus approach to infer tumor evolutionary histories," *Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, pp. 63–72, 8 2018.

[12] Z. DiNardo, K. Tomlinson, A. Ritz, and L. Oesper, "Distance measures for tumor evolutionary trees," *Bioinformatics*, vol. 36, pp. 2090–2097, 4 2020.

[13] E. Oláh, "Basic concepts of cancer: Genomic determination." *EJIFCC*, vol. 16, pp. 10–15, 5 2005.

[14] N. B. Jamieson, D. K. Chang, and A. V. Biankin, "Cancer genetics and implications for clinical management," *Surgical Clinics of North America*, vol. 95, pp. 919–934, 10 2015.

[15] H. A. Pallikonda and S. Turajlic, "Predicting cancer evolution for patient benefit: Renal cell carcinoma paradigm," *Biochimica et Biophysica Acta (BBA) - Reviews on Cancer*, vol. 1877, no. 5, p. 188759, sep 2022. [Online]. Available: https://doi.org/10.1016%2Fj.bbcan.2022.188759

[16] L. R. Yates and P. J. Campbell, "Evolution of the cancer genome," *Nature Reviews Genetics*, vol. 13, no. 11, pp. 795–806, oct 2012. [Online]. Available: https://doi.org/10.1038%2Fnrg3317

[17] M. Greaves and C. C. Maley, "Clonal evolution in cancer," *Nature*, vol. 481, no. 7381, pp. 306–313, jan 2012. [Online]. Available: https://doi.org/10.1038%2Fnature10762

[18] M. R. Stratton, P. J. Campbell, and P. A. Futreal, "The cancer genome," *Nature*, vol. 458, no. 7239, pp. 719–724, apr 2009. [Online]. Available: https://doi.org/10.1038%2Fnature07943

[19] S. R. y Cajal, M. Sesé, C. Capdevila, T. Aasen, L. D. Mattos-Arruda, S. J. Diaz-Cano, J. Hernández-Losa, and J. Castellví, "Clinical implications of intratumor heterogeneity: challenges and opportunities," *Journal of Molecular Medicine*, vol. 98, no. 2, pp. 161–177, jan 2020. [Online]. Available: https://doi.org/10.1007%2Fs00109-020-01874-2

[20] E. Williamson, *Lists, Decisions and Graphs*. S. Gill Williamson, 2010. [Online]. Available: https://books.google.pt/books?id=vaXv_yhefG8C

[21] J. Kuipers, K. Jahn, and N. Beerenwinkel, "Advances in understanding tumour evolution through single-cell sequencing," *Biochimica et Biophysica Acta (BBA) - Reviews on Cancer*, vol. 1867, no. 2, pp. 127–138, apr 2017. [Online]. Available: https://doi.org/10.1016%2Fj.bbcan.2017.02.001

[22] S. Ciccolella, C. Ricketts, M. Soto Gomez, M. Patterson, D. Silverbush, P. Bonizzoni, I. Hajirasouliha, and G. Della Vedova, "Inferring cancer progression from Single-Cell Sequencing while allowing mutation losses," *Bioinformatics*, vol. 37, no. 3, pp. 326–333, 08 2020. [Online]. Available: https://doi.org/10.1093/bioinformatics/btaa722

[23] S. Malikic, K. Jahn, J. Kuipers, S. C. Sahinalp, and N. Beerenwinkel, "Integrative inference of subclonal tumour evolution from single-cell and bulk sequencing data," *Nature Communications*, vol. 10, p. 2750, 6 2019.

[24] S. Malikic, F. R. Mehrabadi, S. Ciccolella, M. K. Rahman, C. Ricketts, E. Haghshenas, D. Seidman, F. Hach, I. Hajirasouliha, and S. C. Sahinalp, "Phiscs: a combinatorial approach for subperfect tumor phylogeny reconstruction via integrative use of single-cell and bulk sequencing data," *Genome Research*, vol. 29, pp. 1860–1877, 11 2019.

[25] E. Husić, X. Li, A. Hujdurović, M. Mehine, R. Rizzi, V. Mäkinen, M. Milanič, and A. I. Tomescu, "MIPUP: minimum perfect unmixed phylogenies for multi-sampled tumors via branchings and ILP," *Bioinformatics*, vol. 35, no. 5, pp. 769–777, 08 2018. [Online]. Available: https://doi.org/10.1093/bioinformatics/bty683

[26] V. Popic, R. Salari, I. Hajirasouliha, D. Kashef-Haghighi, R. B. West, and S. Batzoglou, "Fast and scalable inference of multi-sample cancer lineages," *Genome Biology*, vol. 16, p. 91, 12 2015.

[27] M. Griffith, C. A. Miller, O. L. Griffith, K. Krysiak, Z. L. Skidmore, A. Ramu, J. R. Walker, H. X. Dang, L. Trani, D. E. Larson, R. T. Demeter, M. C. Wendl, J. F. McMichael, R. E. Austin, V. Magrini, S. D. McGrath, A. Ly, S. Kulkarni, M. G. Cordes, C. C. Fronick, R. S. Fulton, C. A. Maher, L. Ding, J. M. Klco, E. R. Mardis, T. J. Ley, and R. K. Wilson, "Optimizing cancer genome sequencing and analysis," *Cell Systems*, vol. 1, no. 3, pp. 210–223, sep 2015. [Online]. Available: https://doi.org/10.1016%2Fj.cels.2015.08.015

[28] M. El-Kebir, L. Oesper, H. Acheson-Field, and B. J. Raphael, "Reconstruction of clonal trees and tumor composition from multi-sample sequencing data," *Bioinformatics*, vol. 31, no. 12, pp. i62–i70, jun 2015. [Online]. Available: https://doi.org/10.1093%2Fbioinformatics%2Fbtv261

[29] S. Malikic, A. W. McPherson, N. Donmez, and C. S. Sahinalp, "Clonality inference in multiple tumor samples using phylogeny," *Bioinformatics*, vol. 31, no. 9, pp. 1349–1356, jan 2015. [Online]. Available: https://doi.org/10.1093%2Fbioinformatics%2Fbtv003

[30] Y. Jiang, Y. Qiu, A. J. Minn, and N. R. Zhang, "Assessing intratumor heterogeneity and tracking longitudinal and spatial clonal evolutionary history by next-generation sequencing," *Proceedings of the National Academy of Sciences*, vol. 113, no. 37, aug 2016. [Online]. Available: https://doi.org/10.1073%2Fpnas.1522203113

[31] E. M. Ross and F. Markowetz, "OncoNEM: inferring tumor evolution from single-cell sequencing data," *Genome Biology*, vol. 17, no. 1, apr 2016. [Online]. Available: https://doi.org/10.1186%2Fs13059-016-0929-9

[32] M. El-Kebir, G. Satas, L. Oesper, and B. J. Raphael, "Inferring the mutational history of a tumor using multi-state perfect phylogeny mixtures," *Cell Systems*, vol. 3, no. 1, pp. 43–53, jul 2016. [Online]. Available: https://doi.org/10.1016%2Fj.cels.2016.07.004

[33] D. Pradhan and M. El-Kebir, "On the non-uniqueness of solutions to the perfect phylogeny mixture problem," in *Comparative Genomics*. Springer International Publishing, 2018, pp. 277–293. [Online]. Available: https://doi.org/10.1007%2F978-3-030-00834-5_16

[34] K. Tomlinson and L. Oesper, "Examining tumor phylogeny inference in noisy sequencing data," in *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, dec 2018. [Online]. Available: https://doi.org/10.1109%2Fbibm.2018.8621437

[35] "What is tree topology? definition and explanation." [Online]. Available: https://www.javatpoint.com/what-is-tree-topology

[36] D. F. Robinson and L. R. Foulds, "Comparison of phylogenetic trees."

[37] D. Critchlow, D. Pearl, and C. Qian, "The triples distance for rooted bifurcating phylogenetic trees," *Systematic Biology - SYST BIOL*, vol. 45, pp. 323–334, 09 1996.

[38] J. Jansson and R. Rajaby, "A more practical algorithm for the rooted triplet distance," *Journal of Computational Biology*, vol. 24, no. 2, pp. 106–126, 2017.

[39] M. Kimura, "THE NUMBER OF HETEROZYGOUS NUCLEOTIDE SITES MAINTAINED IN A FINITE POPULATION DUE TO STEADY FLUX OF MUTATIONS," *Genetics*, vol. 61, no. 4, pp. 893–903, 04 1969. [Online]. Available: https://doi.org/10.1093/genetics/61.4.893

[40] M. Bourque, "Arbres de steiner et reseaux dont certains sommets sont a localisation variable," *Universite de Montreal, [Montreal]*, 1978.

[41] M. J. Kochenderfer and T. A. Wheeler, *Algorithms for Optimization*. The MIT Press, 2019.

[42] M. F. Ahmad, N. A. M. Isa, W. H. Lim, and K. M. Ang, "Differential evolution: A recent review based on state-of-the-art works," *Alexandria Engineering Journal*, vol. 61, no. 5, pp. 3831–3872, may 2022. [Online]. Available: https://doi.org/10.1016%2Fj.aej.2021.09.013

[43] N. Oladejo, A. Abolarinwa, S. Salawu, A. Lukman, and H. Bukari, "Optimization principle and its' application in optimizing landmark university bakery production using linear programming," *International Journal of Civil Engineering and Technology*, vol. 10, pp. 183–190, 02 2019.

[44] J. Kim and K.-K. K. Kim, "Dynamic programming for scalable just-in-time economic dispatch with non-convex constraints and anytime participation," *International Journal of Electrical Power*, vol. 123, p. 106217, dec 2020. [Online]. Available: https://doi.org/10.1016%2Fj.ijepes.2020.106217

[45] M. TURKYİLMAZOGLU, "A simple algorithm for high order newton iteration formulae and some new variants," *Hacettepe Journal of Mathematics and Statistics*, vol. 49, no. 1, pp. 425–438, feb 2020. [Online]. Available: https://doi.org/10.15672%2Fhujms.459810

[46] R. Storn and K. Price, "Minimizing the real functions of the icec'96 contest by differential evolution," in *Proceedings of IEEE international conference on evolutionary computation*. IEEE, 1996, pp. 842–844.

[47] B. S. C. F. Leite, "Differential Evolution: An alternative to nonlinear convex optimization — towardsdatascience.com," https://towardsdatascience.com/differential-evolution-an-alternative-to-nonlinear-convex-optimization-690a123f3413, 2022, [Accessed 20-09-2023].

[48] I. Frades and R. Matthiesen, "Overview on techniques in cluster analysis," *Bioinformatics methods in clinical research*, pp. 81–107, 2010.

[49] A. Ahmad and S. S. Khan, "Survey of state-of-the-art mixed data clustering algorithms," *IEEE Access*, vol. 7, pp. 31 883–31 902, 2019.

[50] P. J. Rousseeuw, "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis," *Journal of computational and applied mathematics*, vol. 20, pp. 53–65, 1987.

[51] L. Yang and R. Jin, "Distance metric learning: A comprehensive survey," *Michigan State Universiy*, vol. 2, no. 2, p. 4, 2006.

[52] V. Hovestadt, K. S. Smith, L. Bihannic, M. G. Filbin, M. L. Shaw, A. Baumgartner, J. C. DeWitt, A. Groves, L. Mayr, H. R. Weisman, A. R. Richman, M. E. Shore, L. Goumnerova, C. Rosencrance, R. A. Carter, T. N. Phoenix, J. L. Hadley, Y. Tong, J. Houston, R. A. Ashmun, M. DeCuypere, T. Sharma, D. Flasch, A. Silkov, K. L. Ligon, S. L. Pomeroy, M. N. Rivera, O. Rozenblatt-Rosen, J. M. Rusert, R. J. Wechsler-Reya, X.-N. Li, A. Peyrl, J. Gojo, D. Kirchhofer, D. Lötsch, T. Czech, C. Dorfer, C. Haberler, R. Geyeregger, A. Halfmann, C. Gawad, J. Easton, S. M. Pfister, A. Regev, A. Gajjar, B. A. Orr, I. Slavc, G. W. Robinson, B. E. Bernstein, M. L. Suvà, and P. A. Northcott, "Resolving medulloblastoma cellular architecture by single-cell genomics," *Nature*, vol. 572, pp. 74–79, 8 2019.

[53] M. K. Goel, P. Khanna, and J. Kishore, "Understanding survival analysis: Kaplan-meier estimate," *International journal of Ayurveda research*, vol. 1, no. 4, p. 274, 2010.

[54] V. Bewick, L. Cheek, and J. Ball, "Statistics review 12: survival analysis," *Critical care*, vol. 8, no. 5, pp. 1–6, 2004.

[55] T. G. Clark, M. J. Bradburn, S. B. Love, and D. G. Altman, "Survival analysis part i: Basic concepts and first analyses," *British Journal of Cancer*, vol. 89, no. 2, pp. 232–238, jul 2003. [Online]. Available: https://doi.org/10.1038%2Fsj.bjc.6601118

[56] S. Belciug, "A survival analysis guide in oncology," in *Intelligent Systems Reference Library*. Springer International Publishing, sep 2022, pp. 29–46. [Online]. Available: https://doi.org/10.1007%2F978-3-031-11170-9_2

[57] J. Emmerson and J. Brown, "Understanding survival analysis in clinical trials," *Clinical Oncology*, vol. 33, no. 1, pp. 12–14, jan 2021. [Online]. Available: https://doi.org/10.1016%2Fj.clon.2020.07.014

[58] L. L. Johnson and J. H. Shih, "An introduction to survival analysis," in *Principles and practice of clinical research*. Elsevier, 2007, pp. 273–282.

[59] S. V. Deo, V. Deo, and V. Sundaram, "Survival analysis—part 2: Cox proportional hazards model," *Indian journal of thoracic and cardiovascular surgery*, vol. 37, pp. 229–233, 2021.

[60] J. Bradic, J. Fan, and J. Jiang, "Regularization for cox's proportional hazards model with NP-dimensionality," *The Annals of Statistics*, vol. 39, no. 6, dec 2011. [Online]. Available: https://doi.org/10.1214%2F11-aos911

[61] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 68, no. 1, pp. 49–67, 2006.

[62] T. T. Cai, "Discussion of "regularization of wavelet approximations"(by a. antoniadis and j. fan)," *J. Am. Statist. Ass*, vol. 96, pp. 960–962, 2001.

[63] A. Antoniadis and J. Fan, "Regularization of wavelet approximations," *Journal of the American Statistical Association*, vol. 96, no. 455, pp. 939–967, 2001.

[64] G. Malenová, D. Rowson, and V. Boeva, "Exploring pathway-based group lasso for cancer survival analysis: A special case of multi-task learning," *Frontiers in Genetics*, vol. 12, nov 2021. [Online]. Available: https://doi.org/10.3389%2Ffgene.2021.771301

[65] J. C. Utazirubanda, T. M. León, and P. Ngom, "Variable selection with group LASSO approach: Application to cox regression with frailty model," *Communications in Statistics - Simulation and Computation*, vol. 50, no. 3, pp. 881–901, mar 2019. [Online]. Available: https://doi.org/10.1080%2F03610918.2019.1571605

[66] Álvaro Méndez Civieta, "Sparse Group Lasso in Python — towardsdatascience.com," https://towardsdatascience.com/sparse-group-lasso-in-python-255e379ab892, 2020, [Accessed 10-10-2023].

[67] T. Hastie, R. Tibshirani, J. H. Friedman, and J. H. Friedman, *The elements of statistical learning: data mining, inference, and prediction.* Springer, 2009, vol. 2.

[68] T. M. H. T. M. Kamarul Imran, Wan Nor Arifin, "Chapter 11 Survival Analysis: Kaplan-Meier and Cox Proportional Hazard (PH) Regression — Data Analysis in Medicine and Health using R — bookdown.org," https://bookdown.org/drki_musa/dataanalysis/survival-analysis-kaplan-meier-and-cox-proportional-hazard-ph-regression.html#:~:text=11.9.1%20Log,peto%20tests, 2022, [Accessed 10-10-2023].

[69] M. Jafari and N. Ansari-Pour, "Why, when and how to adjust your p values?" *Cell Journal (Yakhteh)*, vol. 20, no. 4, p. 604, 2019.

[70] S. Lydersen, "Adjustment of p-values for multiple hypotheses," *Tidsskrift for Den norske legeforening*, 2021.

[71] R. J. Feise, "Do multiple outcome measures require p-value adjustment?" *BMC medical research methodology*, vol. 2, pp. 1–4, 2002.

[72] J. Neyman and E. S. Pearson, "On the use and interpretation of certain test criteria for purposes of statistical inference: Part i," *Biometrika*, pp. 175–240, 1928.

[73] H. Akaike, "Factor analysis and aic," *Psychometrika*, vol. 52, pp. 317–332, 1987.

[74] P. Xue, L. Zhu, Z. Wan, W. Huang, N. Li, D. Chen, J. Hu, H. Yang, and L. Wang, "A prognostic index model to predict the clinical outcomes for advanced pancreatic cancer patients following palliative chemotherapy," *Journal of Cancer Research and Clinical Oncology*, vol. 141, no. 9, pp. 1653–1660, mar 2015. [Online]. Available: https://doi.org/10.1007%2Fs00432-015-1953-y

[75] E. Longato, M. Vettoretti, and B. D. Camillo, "A practical perspective on the concordance index for the evaluation and selection of prognostic time-to-event models," *Journal of Biomedical Informatics*, vol. 108, p. 103496, aug 2020. [Online]. Available: https://doi.org/10.1016%2Fj.jbi.2020.103496

[76] F. E. Harrell, R. M. Califf, D. B. Pryor, K. L. Lee, and R. A. Rosati, "Evaluating the yield of medical tests," *Jama*, vol. 247, no. 18, pp. 2543–2546, 1982.

[77] X. G. Luo, J. Kuipers, and N. Beerenwinkel, "Joint inference of exclusivity patterns and recurrent trajectories from tumor mutation trees," *Nature Communications*, vol. 14, no. 1, p. 3676, 2023.

[78] P. Razavi, M. T. Chang, G. Xu, C. Bandlamudi, D. S. Ross, N. Vasan, Y. Cai, C. M. Bielski, M. T. Donoghue, P. Jonsson *et al.*, "The genomic landscape of endocrine-resistant advanced breast cancers," *Cancer cell*, vol. 34, no. 3, pp. 427–438, 2018.

[79] K. Morita, F. Wang, K. Jahn, T. Hu, T. Tanaka, Y. Sasaki, J. Kuipers, S. Loghavi, S. A. Wang, Y. Yan *et al.*, "Clonal evolution of acute myeloid leukemia revealed by high-throughput single-cell genomics," *Nature communications*, vol. 11, no. 1, p. 5327, 2020.

[80] J. H. Ward Jr, "Hierarchical grouping to optimize an objective function," *Journal of the American statistical association*, vol. 58, no. 301, pp. 236–244, 1963.

[81] W. S. Cleveland, "Robust locally weighted regression and smoothing scatterplots," *Journal of the American statistical association*, vol. 74, no. 368, pp. 829–836, 1979.

[82] L. Hubert and P. Arabie, "Comparing partitions," *Journal of classification*, vol. 2, pp. 193–218, 1985.

[83] "Genetics — breastcancer.org," https://www.breastcancer.org/risk/risk-factors/genetics, [Accessed 01-10-2023].

[84] C. D. DiNardo and J. E. Cortes, "Mutations in AML: prognostic and therapeutic implications," *Hematology*, vol. 2016, no. 1, pp. 348–355, Dec. 2016. [Online]. Available: https://doi.org/10.1182/asheducation-2016.1.348

[85] X.-Y. Lin, L. Guo, X. Lin, Y. Wang, and G. Zhang, "Concomitant pik3ca and tp53 mutations in breast cancer: An analysis of clinicopathologic and mutational features, neoadjuvant therapeutic response, and prognosis," *Journal of Breast Cancer*, vol. 26, no. 4, p. 363, 2023.

A

Distance Metrics

Applying CASet distancing to the trees on Fig. 2.3, the common ancestors of all nodes pair combination is as follows in Fig. A.1.

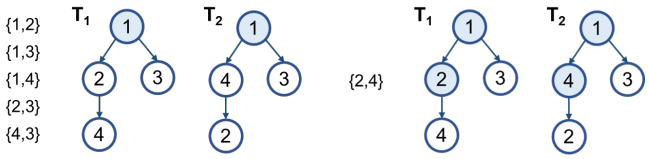


Figure A.1: Common Ancestor sets in the two trees for every combination of nodes.

Now applying Jaccard Similarity (Eq. 2.5) the left combination of nodes in Fig. A.1, $Jacc(C_1(i,j), C_2(i,j)) = 0$ for all $\{i,j\} \neq \{2,4\}$ and $\{i,j\} \in M(T)$, where i≠j. When applied to $\{2,4\}$, we get $Jacc(\{1,2\},\{1,4\}) = 0.66(6)$, so by inputting the sum of the Jaccard results for all combination of nodes in Eq. 2.6, we have $CASet(T_1,T_2) = \frac{0.66(6)}{6} = 0.11$.

Applying DISC distancing to the trees on Fig. 2.3, the distinctly inherited sets of all nodes pair combination is as follows, note that unlike CASet the order of the pairs matters, having in this case 12 different pairs:

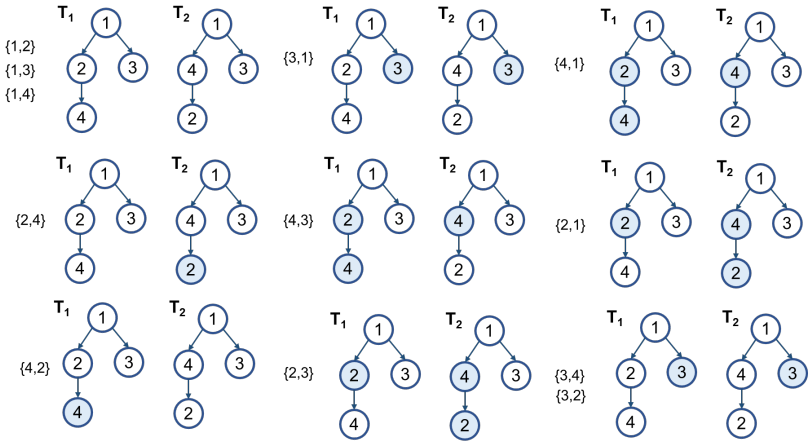


Figure A.2: Distinctly Inherited sets in the two trees for every ordered combination of nodes.

Respectively, the Jaccard results are: $0, 0, \frac{1}{2}, 1, \frac{1}{2}, \frac{1}{2}, 1, \frac{1}{2}, 0$. That makes $DISC = \frac{4}{12} = 0.33$, so the same tree is more dissimilar for DISC than for CASet, highlighting that CASet indeed gives more weight to changes in the root, whereas DISC penalises more differences near the leaves.

For BD, we can compute the different partitions for each tree as follows, given that left side correspond to T_1 partitions and the right side correspond to T_2:

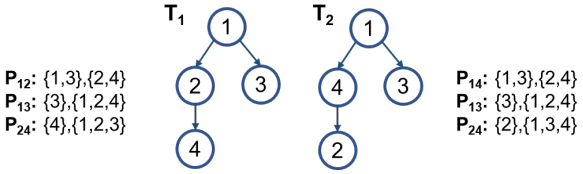


Figure A.3: Partition of each edge for the trees in Fig. 2.3.

The BD distance in these trees is simply $BD(T_1, T_2) = 2$ since the edge (2,4) induces a partition in both trees that are not found in the other then normalised by the total number of the partitions in both trees we achieved the final result $BD(T_1, T_2) = 0.33$.

For MP3 it was calculated by finding the minimal tree topologies:

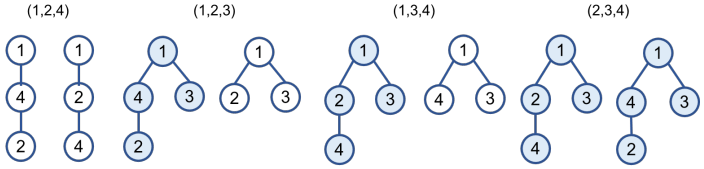


Figure A.4: Minimal tree topology for the triplet combinations for the trees in Fig. 2.3.

Given that the maximum number of minimal topology trees is 4 (no poly-ocurring labels)

and with 2 intersected trees, the MP3 equation can translate to $MP3 = \frac{2}{4} = 0.5$, higher difference than DISC and CASet, note that MP3 provides a measure of similarity, to transform into distance we need to perform 1 - MP3 output.

For PCD, the Parent-Child relationships of $T_1 = \{(1,2),(2,4),(1,3)\}$ and for $T_2 = \{(1,4),(4,2),(1,3)\}$, doing the intersection and normalising with the union gives us a measure of similarity so again we had to subtract our score from 1. Gving us a total of $PCD(T_1,T_2) = 1 - \frac{1}{5} = 0.8$

In AD case, the Ancestor-Descendant relationships for $T_1 = \{(1,2),(2,4),(1,3),(1,4)\}$ and for $T_2 = \{(1,4),(4,2),(1,3),(1,2)\}$, gives us a normalised distance of $AD(T_1,T_2) = 1 - \frac{3}{5} = 0.4$.

Finally for CD, the clones present in $T_1 = \{\{1\},\{1,2\},\{1,3\},\{1,2,4\}\}$ and for $T_2 = \{\{1\},\{1,4\},\{1,3\},\{1,4,2\}\}$, which then translates to a score of $CD(T_1,T_2) = 1 - \frac{3}{5} = 0.4$.

B

Breast Cancer

Table B.1: Cluster count per metric at different LOESS thresholds: $t = 0.04$, $t = 0.03$, $t = 0.02$ and $t = 0.01$

	t = 0.04	**t = 0.03**	**t = 0.02**	**t = 0.01**
DISC ∪	3	7	10	14
DISC ∩	7	10	12	15
CASet ∪	3	7	10	16
CASet ∩	4	4	4	4
BD	3	3	10	16
1BD	4	7	8	13
2BD	4	7	9	14
MLTD	3	3	3	20
AD	3	4	8	14
PD	4	6	7	8
CD	4	6	7	9
PCD	5	8	9	13

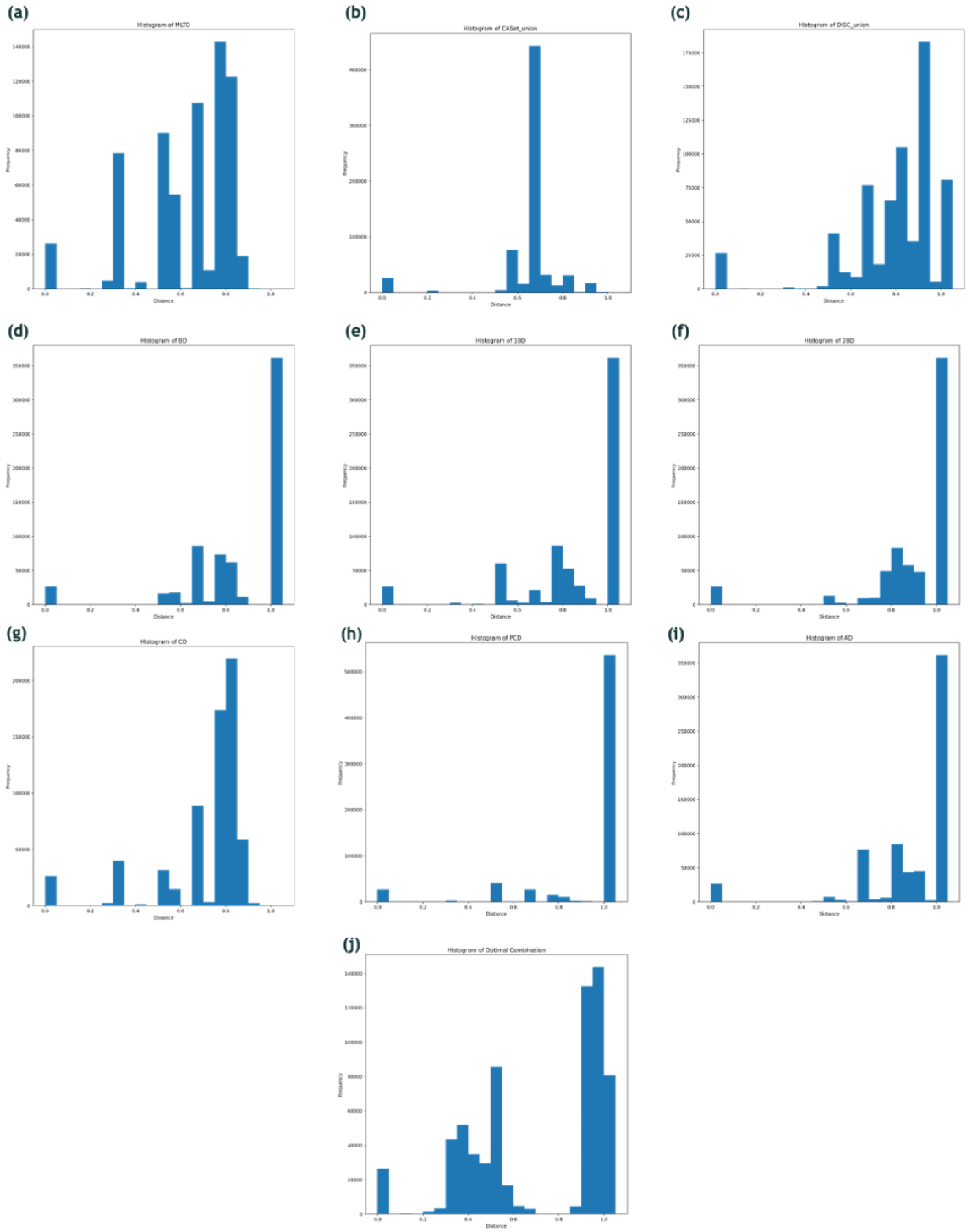


Figure B.1: Distance distributions of the distance metrics (a) BD, (b) 1BD, (c) 2BD, (d) CD, (e) CASet ∪, (f) CASet ∩, (g) DISC ∩, (h) DISC ∪, (i) AD, (j) PD and (k) Optimal Combination.

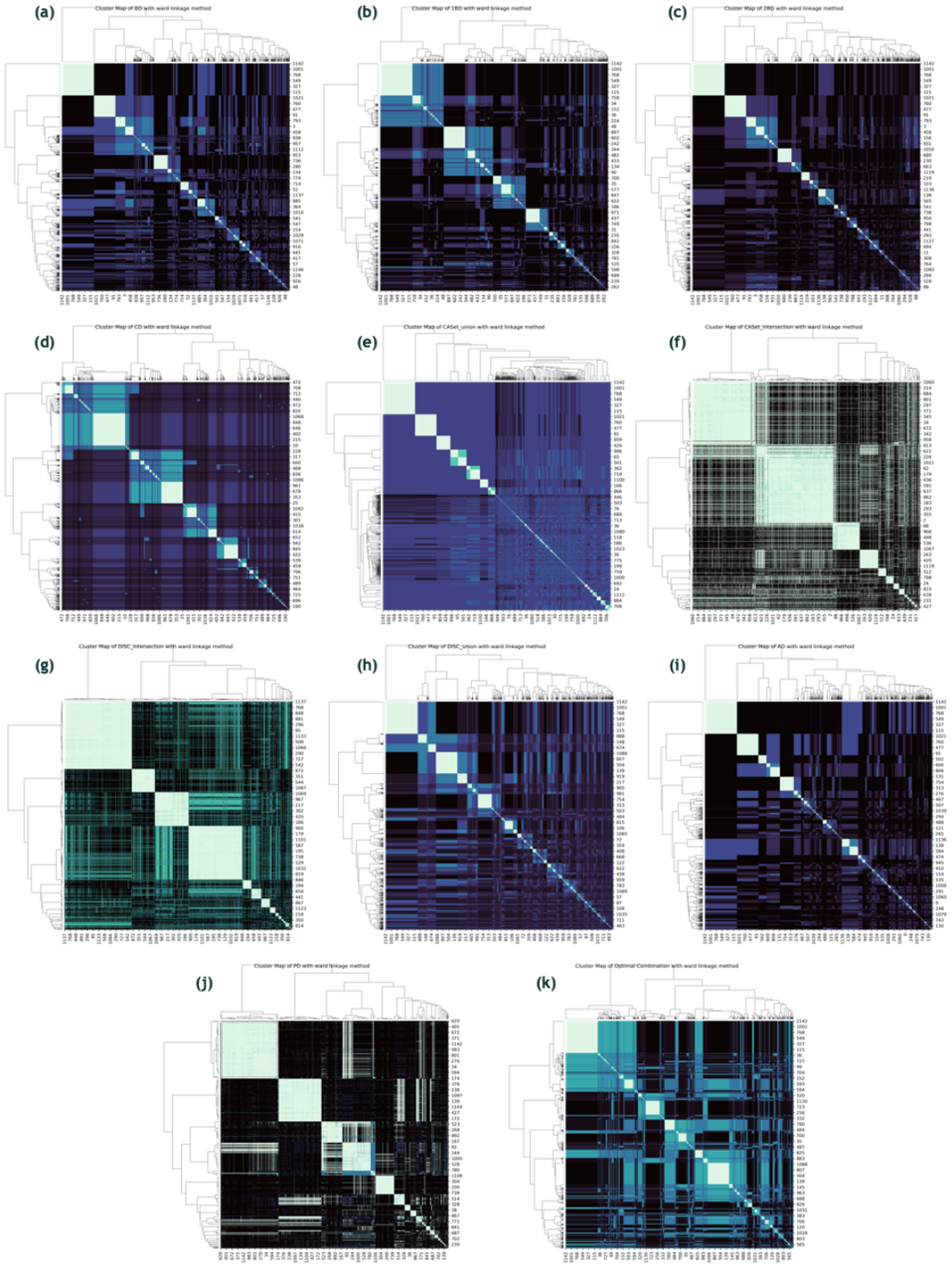


Figure B.2: Cluster maps of the distance metrics (a) MLTD, (b) CASet ∪, (c) DISC ∪, (d) BD, (e) 1BD, (f) 2BD, (g) CD, (h) PCD, (i) AD and (j) Optimal Combination.

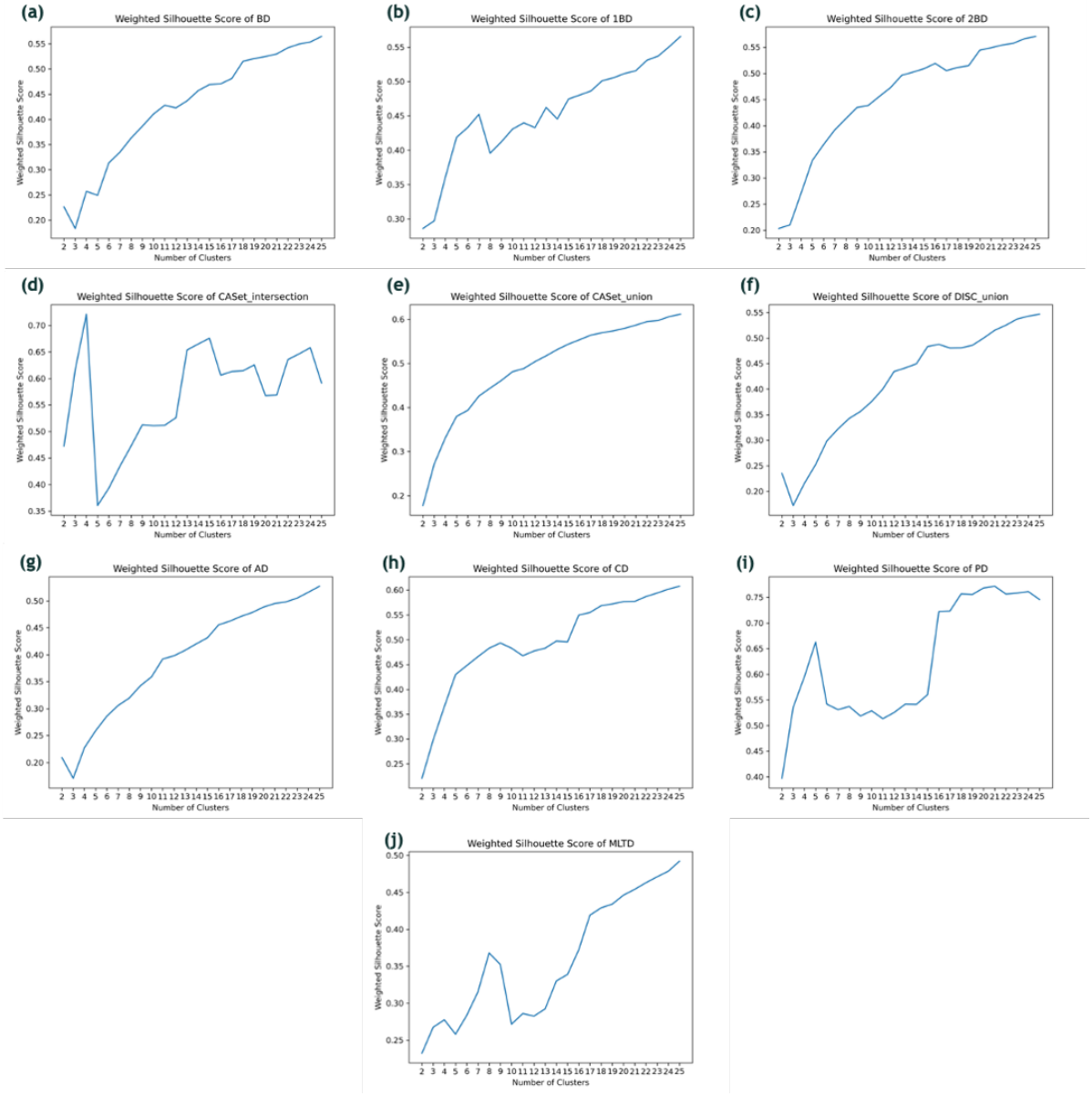


Figure B.3: WSS for a max of 25 clusters of the distance metrics (a) BD, (b) 1BD, (c) 2BD, (d) CASet ∩, (e) CASet ∪, (f) DISC ∪, (g) AD, (h) CD, (i) PD, and (j) MLTD.

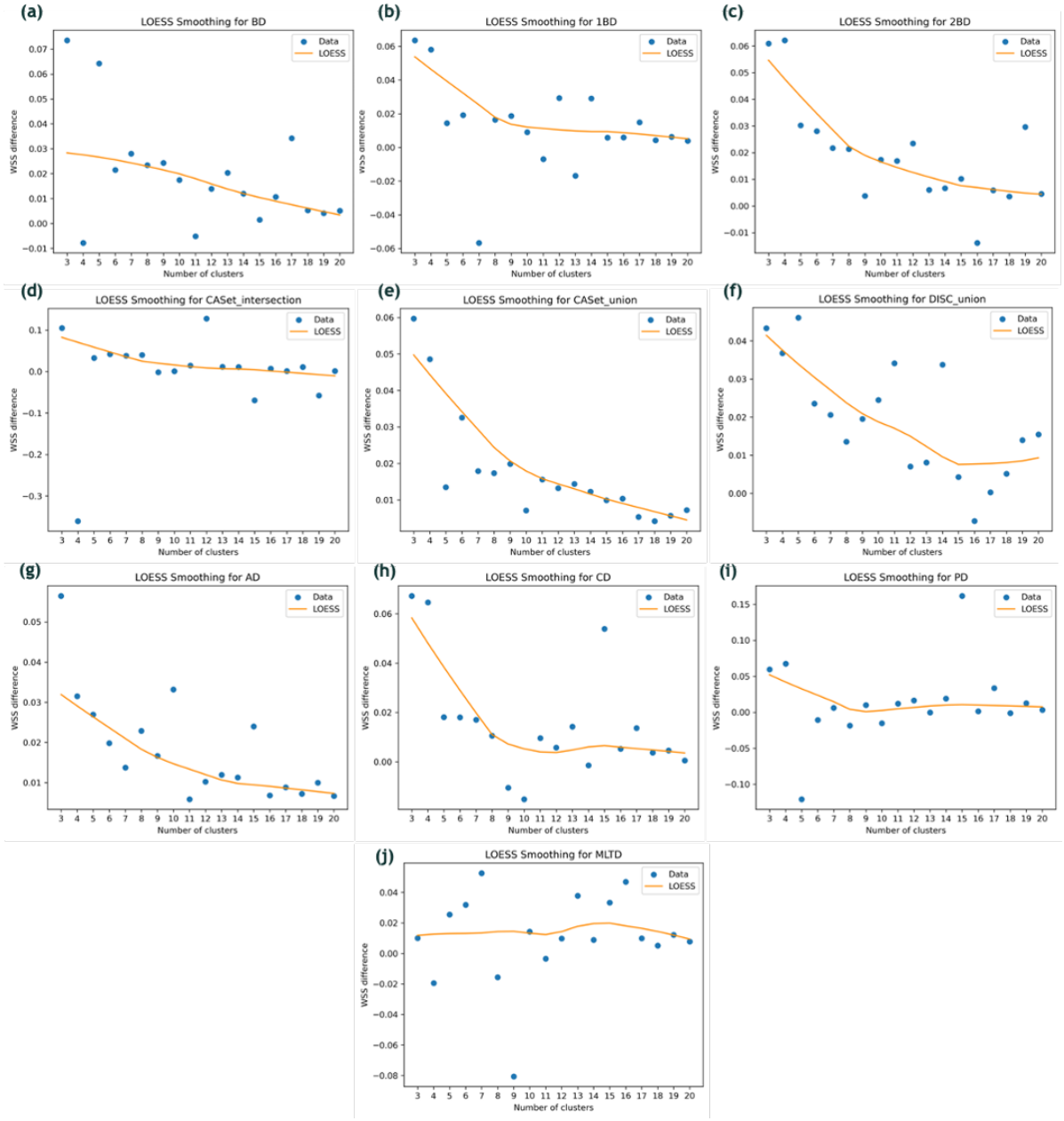


Figure B.4: LOESS applied to the WSS difference of the distance metrics (a) BD, (b) 1BD, (c) 2BD, (d) CASet ∩, (e) CASet ∪, (f) DISC ∪, (g) AD, (h) CD, (i) PD, and (j) MLTD.

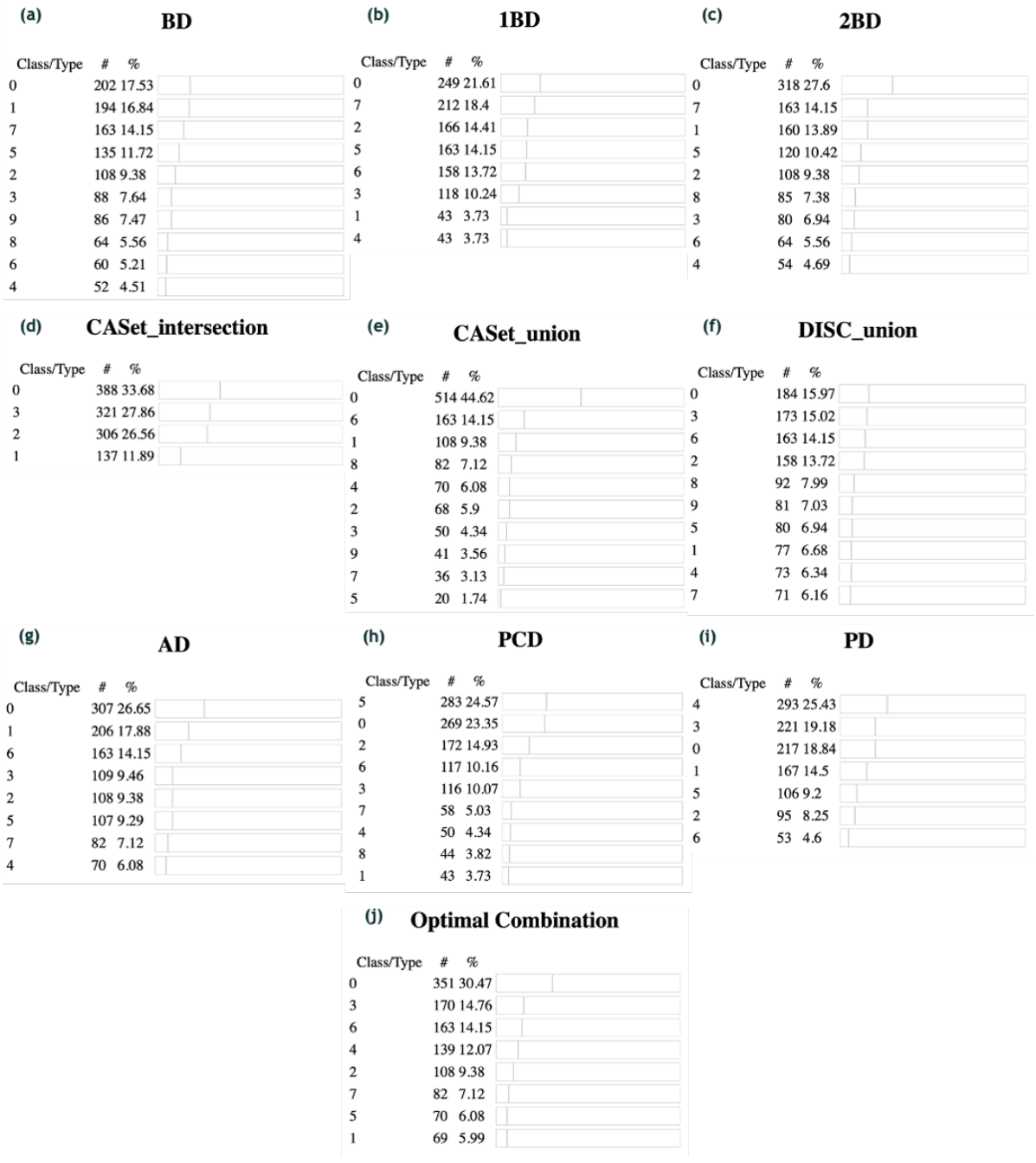


Figure B.5: Cluster patient's distribution of the distance metrics (a) BD, (b) 1BD, (c) 2BD, (d) CASet ∩, (e) CASet ∪, (f) DISC ∪, (g) AD, (h) PCD, (i) PD, and (j) Optimal Combination.

Table B.2: Jaccard Index between the metrics clusters.

	DISC∪	DISC∩	CASet∪	CASet∩	BD	1BD	2BD	MLTD	AD	PD	CD	PCD
DISC∪												
DISC∩	0.216											
CASet∪	0.296	0.178										
CASet∩	0.213	0.326	0.245									
BD	0.45	0.189	0.277	0.203								
1BD	0.277	0.562	0.277	0.271	0.229							
2BD	0.386	0.189	0.417	0.214	0.609	0.246						
MLTD	0.215	0.266	0.259	0.356	0.183	0.211	0.198					
AD	0.408	0.229	0.51	0.219	0.331	0.308	0.419	0.245				
PD	0.21	0.517	0.23	0.313	0.246	0.376	0.245	0.238	0.23			
CD	0.217	0.899	0.216	0.333	0.19	0.627	0.195	0.271	0.243	0.529		
PCD	0.243	0.817	0.205	0.321	0.206	0.603	0.207	0.235	0.257	0.515	0.816	

Table B.3: Jaccard Index between the metrics clusters and the clinical data.

	Overall_Tumor_Grade	Stage	Receptor_Status_Patient	Vital_Status
DISC ∪	0.113	0.077	0.113	0.11
DISC ∩	0.194	0.1	0.159	0.165
CASet ∪	0.194	0.118	0.216	0.212
CASet ∩	0.219	0.124	0.251	0.243
BD	0.115	0.079	0.114	0.114
1BD	0.142	0.092	0.151	0.145
2BD	0.136	0.09	0.135	0.142
MLTD	0.284	0.145	0.359	0.339
AD	0.146	0.092	0.146	0.143
PD	0.17	0.1	0.159	0.158
CD	0.202	0.104	0.168	0.175
PCD	0.164	0.094	0.153	0.151

Table B.4: (Jaccard Index between the metrics clusters and all the mutated genes.

	DISC ∪	DISC ∩	CASet ∪	CASet ∩	BD	1BD	2BD	MLTD	AD	PD	CD	PCD
EPHA7	0.11	0.18	0.24	0.27	0.12	0.15	0.15	0.40	0.15	0.17	0.20	0.16
FOXA1	0.11	0.19	0.22	0.27	0.12	0.15	0.14	0.38	0.15	0.17	0.20	0.16
RB1	0.11	0.18	0.22	0.27	0.12	0.15	0.14	0.39	0.15	0.17	0.19	0.16
GATA3	0.15	0.22	0.24	0.33	0.15	0.18	0.17	0.48	0.18	0.20	0.23	0.20
PTEN	0.13	0.19	0.22	0.26	0.13	0.15	0.16	0.36	0.17	0.17	0.19	0.17
TP53	0.17	0.29	0.24	0.36	0.16	0.21	0.18	0.54	0.22	0.26	0.30	0.25
CDH1	0.13	0.24	0.23	0.25	0.13	0.19	0.16	0.33	0.17	0.21	0.25	0.21
KMT2D	0.11	0.18	0.22	0.27	0.12	0.15	0.14	0.39	0.15	0.17	0.19	0.16
CD79A	0.11	0.18	0.24	0.28	0.12	0.15	0.15	0.40	0.16	0.17	0.20	0.16
PRDM1	0.11	0.18	0.24	0.27	0.12	0.15	0.15	0.40	0.16	0.17	0.20	0.16
TSC2	0.11	0.18	0.23	0.27	0.12	0.15	0.15	0.40	0.15	0.17	0.19	0.16
MAP3K1	0.13	0.20	0.21	0.26	0.11	0.16	0.13	0.36	0.15	0.17	0.21	0.18
PBRM1	0.11	0.18	0.24	0.27	0.12	0.15	0.15	0.40	0.16	0.17	0.20	0.16
PIK3CA	0.16	0.25	0.24	0.39	0.17	0.21	0.19	0.32	0.18	0.26	0.26	0.26
KMT2C	0.13	0.19	0.20	0.26	0.13	0.15	0.13	0.36	0.14	0.17	0.19	0.16
ESR1	0.13	0.20	0.22	0.26	0.13	0.16	0.16	0.37	0.16	0.17	0.21	0.17
NF1	0.11	0.18	0.22	0.27	0.12	0.15	0.14	0.39	0.15	0.17	0.19	0.16
RHOA	0.11	0.18	0.24	0.27	0.12	0.15	0.15	0.40	0.16	0.17	0.20	0.16
PIK3R1	0.11	0.18	0.23	0.27	0.12	0.15	0.14	0.39	0.15	0.17	0.19	0.16

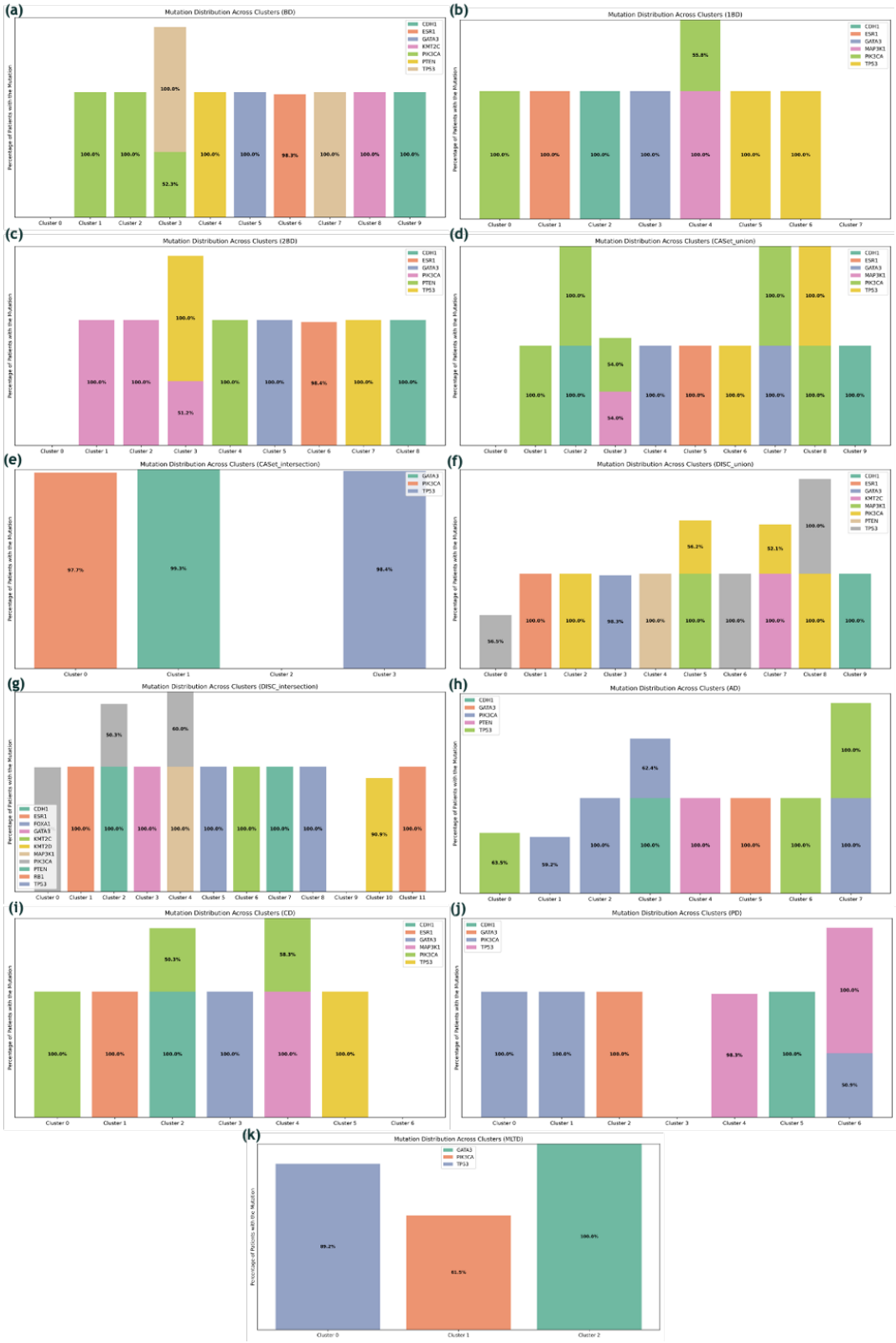


Figure B.6: Mutation distribution across the clusters of the distance metrics (a) BD, (b) 1BD, (c) 2BD, (d) CASet ∪, (e) CASet ∩, (f) DISC ∪, (g) DISC ∩, (h) AD, (i) CD, (j) PD and (k) MLTD.

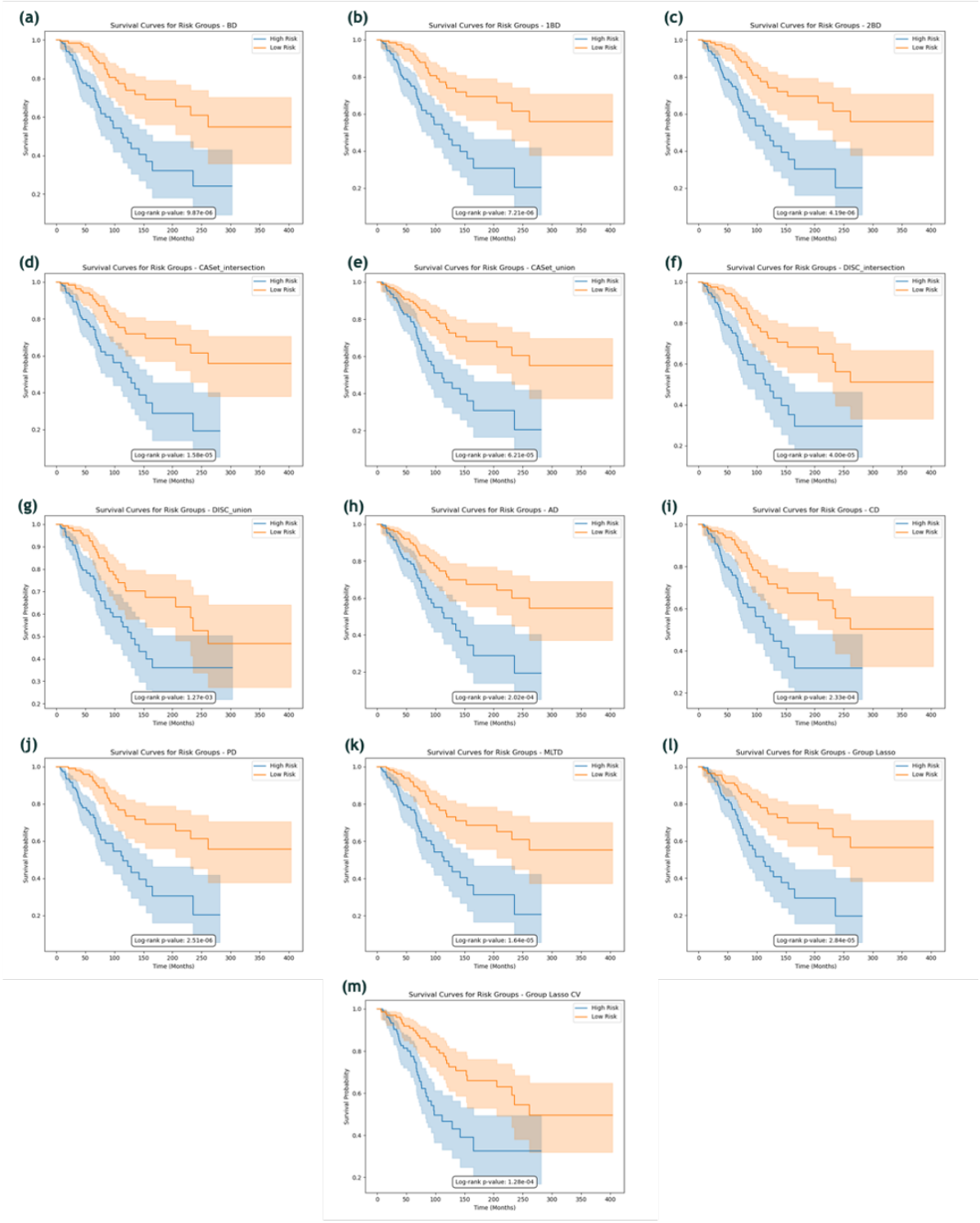


Figure B.7: Survival curves of the high and low risk groups of the distance metrics (a) BD, (b) 1BD, (c) 2BD, (d) CASet ∩, (e) CASet ∪, (f) DISC ∩, (g) DISC ∪, (h) AD, (i) CD, (j) PD, (k) MLTD, (l) Group Lasso and (m) Group Lasso CV.

Table B.5: Adjusted Log rank test p-values.

	log rank p-values
Only Clinical	0.00719
Genotype	0.00047
DISC_union	0.02166
DISC_intersection	0.00068
CASet_union	0.00106
CASet_intersection	0.00027
BD	0.00017
1BD	0.00012
2BD	0.00007
MLTD	0.00028
AD	0.00344
PD	0.00004
CD	0.00396
PCD	0.00032
Optimal Combination	0.00003
Group Lasso	0.00048
Group Lasso CV	0.00217

C

Acute Myeloid Leukaemia

Table C.1: Jaccard index between the distance metric's clusters.

	DISC ∪	DISC ∩	CASet ∩	CASet ∪	BD	1BD	2BD	AD	CD	PD	PCD	MLTD	MP3 ∪	MP3 σ	MP3 ∩	MP3 *geo*
DISC ∪																
DISC ∩	0.092															
CASet ∩	0.245	0.094														
CASet ∪	0.229	0.09	0.197													
BD	0.388	0.077	0.273	0.211												
1BD	0.258	0.148	0.245	0.208	0.26											
2BD	0.349	0.087	0.252	0.212	0.464	0.317										
AD	0.324	0.102	0.24	0.232	0.3	0.355	0.347									
CD	0.293	0.135	0.247	0.275	0.286	0.339	0.336	0.459								
PD	0.13	0.219	0.126	0.12	0.152	0.167	0.162	0.134	0.158							
PCD	0.288	0.135	0.249	0.277	0.287	0.342	0.327	0.449	0.885	0.165						
MLTD	0.323	0.138	0.232	0.219	0.253	0.387	0.286	0.287	0.338	0.154	0.345					
MP3 ∪	0.354	0.081	0.287	0.309	0.336	0.35	0.366	0.498	0.657	0.133	0.648	0.358				
MP3 σ	0.345	0.083	0.264	0.28	0.315	0.33	0.345	0.432	0.632	0.134	0.624	0.341	0.79			
MP3 ∩	0.312	0.086	0.225	0.235	0.306	0.268	0.268	0.29	0.345	0.113	0.345	0.305	0.402	0.383		
MP3 *geo*	0.321	0.111	0.242	0.238	0.254	0.314	0.272	0.276	0.344	0.126	0.343	0.45	0.375	0.403	0.402	

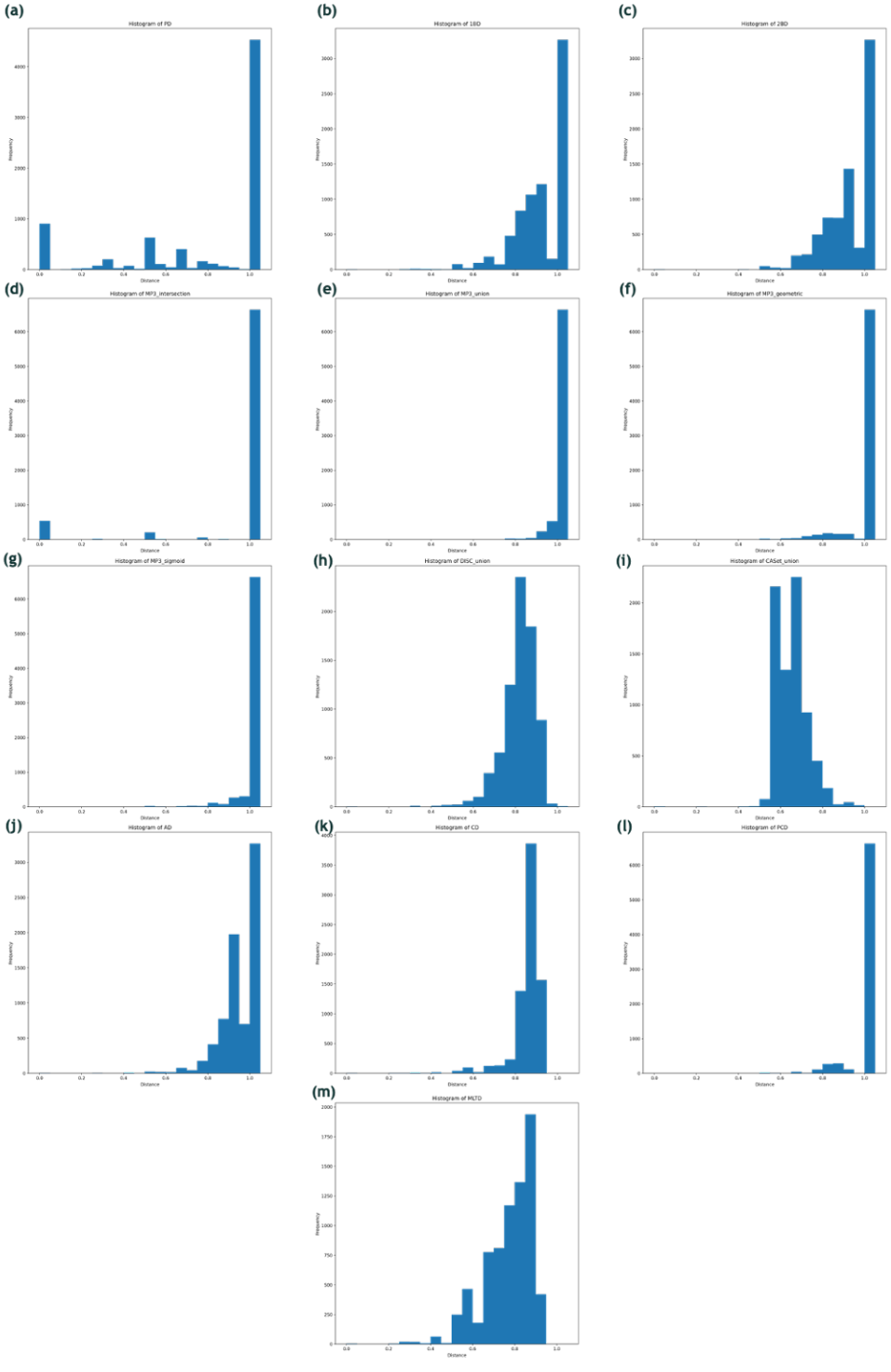


Figure C.1: Distance distributions of the distance metrics (a) PD, (b) 1BD, (c) 2BD,(d) MP3 $\cap$, (e) MP3 $\cup$, (f) MP3 *geo*, (g) MP3 σ, (h) DISC $\cup$, (i) CASet $\cup$, (j) AD, (k) CD, (l) PCD and (m) MLTD.

¡¡

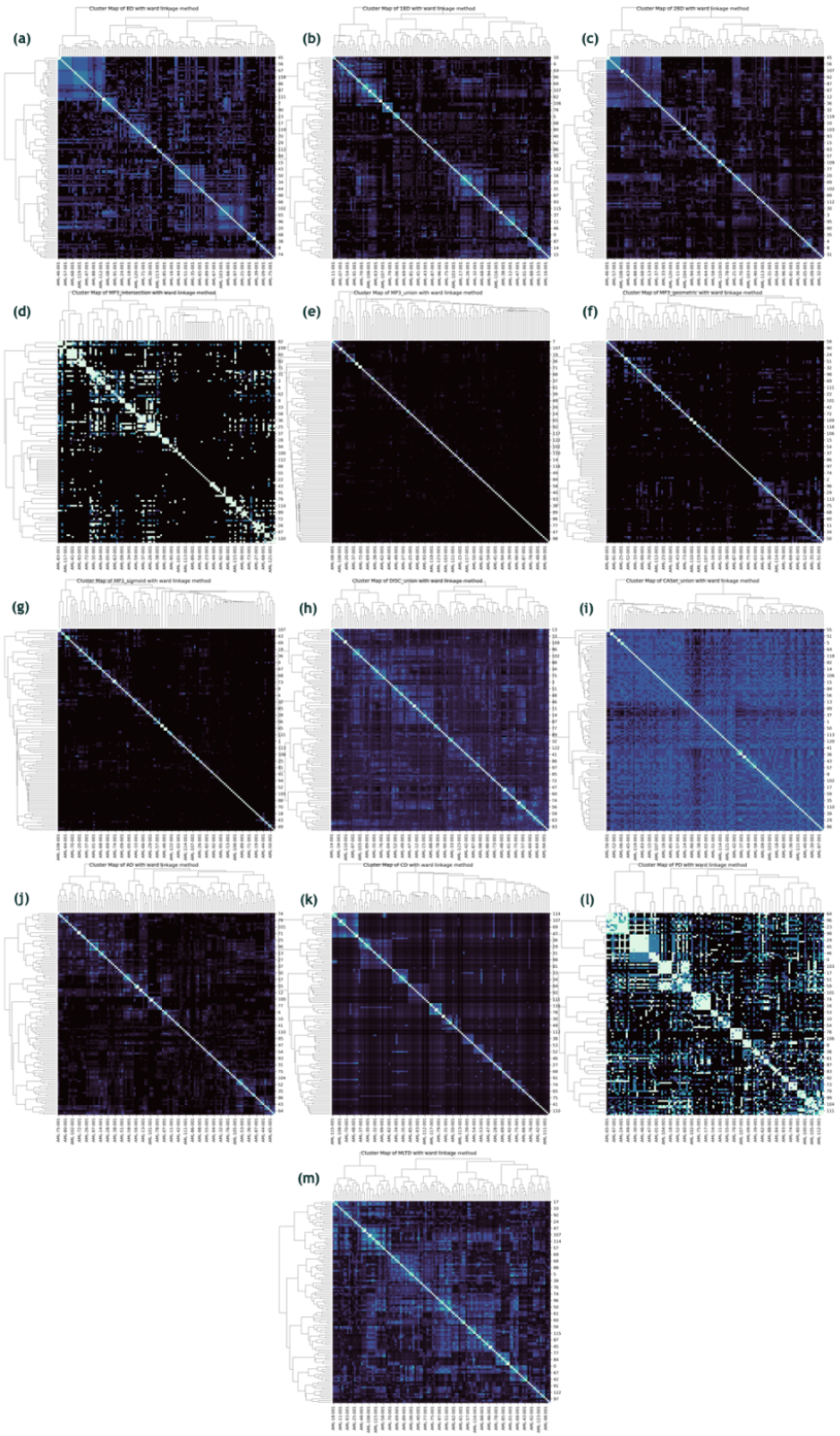


Figure C.2: Cluster maps of the distance metrics (a) BD, (b) 1BD, (c) 2BD, (d) MP3 ∩, (e) MP3 ∪, (f) MP3 *geo*, (g) MP3 σ, (h) DISC ∪, (i) CASet ∪, (j) AD, (k) CD, (l) PD and (m) MLTD.

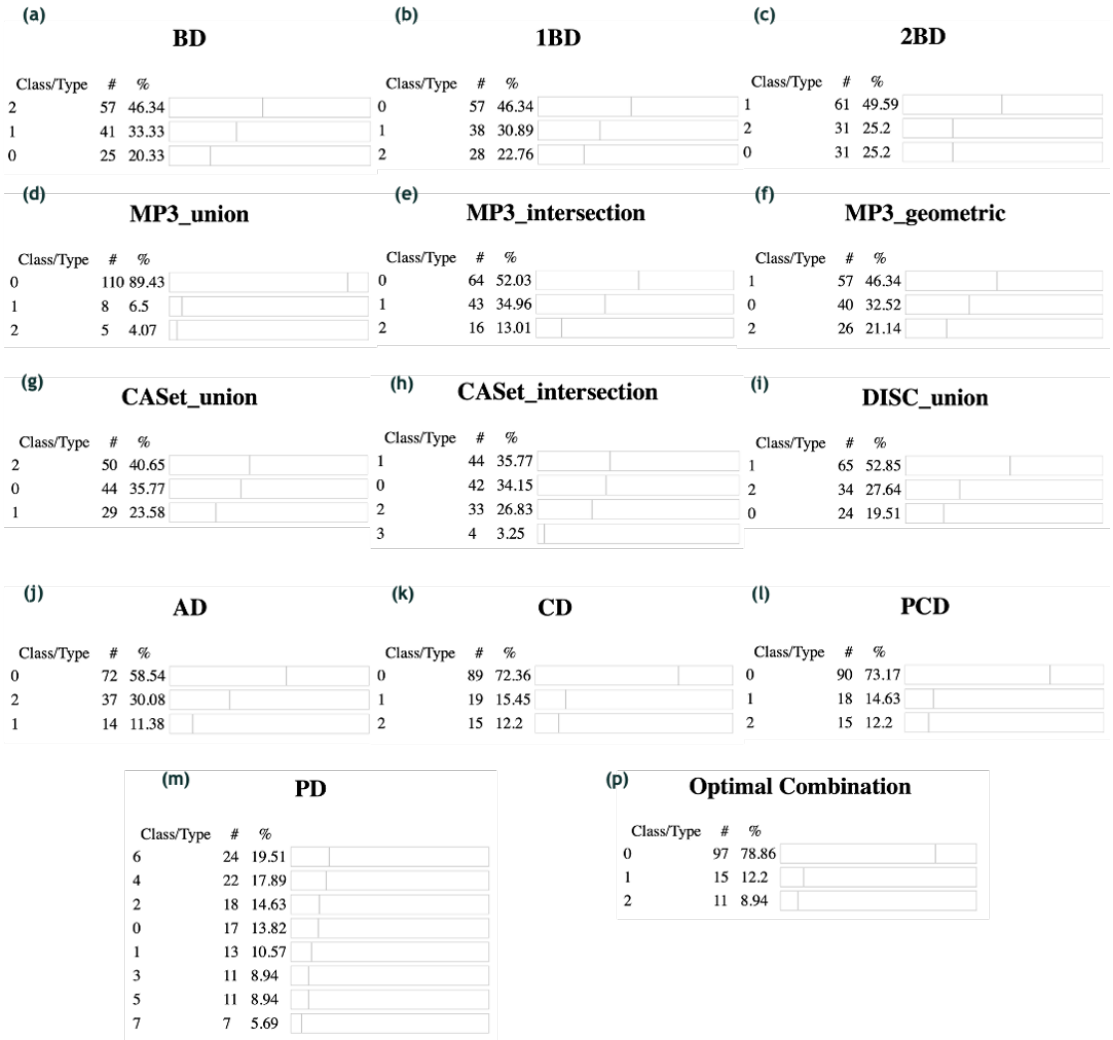


Figure C.3: Cluster patients distribution of the distance metrics (a) BD, (b) 1BD, (c) 2BD, (d) MP3 ∩, (e) MP3 ∪, (f) MP3 *geo*,(g) CASet ∪,(h) CASet ∩, (i) DISC ∪, (j) AD, (k) CD, (l) PCD, (m) PD and (o) Optimal Combination.

Table C.2: Jaccard index between distance metrics and the clinical data.

	Gender	VitalStatus
DISC ∪	0.292	0.295
DISC ∩	0.08	0.08
CASet ∩	0.245	0.239
CASet ∪	0.258	0.267
BD	0.272	0.269
1BD	0.265	0.275
2BD	0.27	0.275
AD	0.313	0.345
CD	0.374	0.384
PD	0.119	0.117
PCD	0.381	0.385
MLTD	0.264	0.259
MP3 ∪	0.465	0.465
MP3 σ	0.438	0.407
MP3 ∩	0.298	0.292
MP3 *geo*	0.271	0.274
Optimal Combination	0.416	0.402

Table C.3: Jaccard index between the genes and distance metrics clusters.

	DISC ∪	DISC ∩	CASet ∩	CASet ∪	BD	1BD	2BD	AD	CD	PD	PCD	MLTD	MP3 ∪	MP3 σ	MP3 ∩	MP3 *geo*
EZH2	0.372	0.09	0.303	0.342	0.348	0.355	0.359	0.422	0.536	0.134	0.546	0.35	0.753	0.611	0.388	0.357
GATA2	0.367	0.093	0.304	0.33	0.347	0.345	0.353	0.416	0.514	0.131	0.523	0.334	0.741	0.601	0.381	0.34
TP53	0.34	0.095	0.32	0.319	0.335	0.332	0.326	0.377	0.467	0.143	0.475	0.31	0.64	0.546	0.366	0.318
STAG2	0.385	0.089	0.309	0.341	0.359	0.348	0.363	0.435	0.54	0.132	0.55	0.342	0.778	0.631	0.395	0.352
IDH1	0.353	0.092	0.29	0.322	0.338	0.312	0.33	0.403	0.452	0.124	0.45	0.307	0.621	0.517	0.357	0.313
ETV6	0.389	0.088	0.31	0.344	0.358	0.356	0.363	0.435	0.564	0.134	0.574	0.35	0.791	0.642	0.404	0.359
SRSF2	0.313	0.094	0.289	0.3	0.319	0.413	0.363	0.381	0.465	0.133	0.473	0.383	0.632	0.557	0.403	0.385
U2AF1	0.367	0.086	0.296	0.333	0.338	0.332	0.34	0.4	0.51	0.13	0.532	0.348	0.717	0.581	0.393	0.352
DNMT3A	0.302	0.117	0.262	0.288	0.298	0.297	0.307	0.345	0.534	0.146	0.527	0.28	0.544	0.489	0.322	0.295
SF3B1	0.361	0.097	0.298	0.334	0.337	0.321	0.331	0.387	0.489	0.128	0.498	0.317	0.672	0.545	0.364	0.323
FLT3	0.37	0.085	0.274	0.323	0.316	0.298	0.31	0.378	0.461	0.13	0.47	0.308	0.568	0.464	0.364	0.316
PTPN11	0.358	0.094	0.276	0.344	0.313	0.307	0.311	0.407	0.432	0.126	0.431	0.303	0.593	0.517	0.395	0.327
NRAS	0.421	0.093	0.258	0.28	0.369	0.283	0.305	0.315	0.374	0.139	0.375	0.442	0.477	0.433	0.414	0.365
CSF3R	0.383	0.088	0.31	0.344	0.361	0.358	0.368	0.446	0.564	0.134	0.574	0.35	0.791	0.661	0.409	0.361
IDH2	0.306	0.112	0.277	0.281	0.303	0.397	0.369	0.43	0.546	0.148	0.556	0.394	0.601	0.545	0.391	0.389
PPM1D	0.373	0.088	0.313	0.336	0.359	0.351	0.353	0.423	0.532	0.134	0.541	0.338	0.765	0.621	0.39	0.348
CBL	0.383	0.089	0.311	0.344	0.358	0.352	0.363	0.435	0.549	0.133	0.559	0.346	0.791	0.642	0.4	0.356
MYC	0.378	0.091	0.308	0.338	0.354	0.355	0.358	0.429	0.54	0.134	0.55	0.342	0.778	0.631	0.398	0.352
KRAS	0.379	0.085	0.286	0.312	0.315	0.321	0.308	0.371	0.461	0.129	0.47	0.317	0.593	0.517	0.369	0.333
KIT	0.382	0.087	0.304	0.339	0.356	0.353	0.359	0.433	0.537	0.13	0.546	0.342	0.776	0.628	0.388	0.354
PHF6	0.378	0.09	0.309	0.345	0.356	0.353	0.363	0.429	0.54	0.132	0.55	0.346	0.778	0.631	0.395	0.357
SMC3	0.391	0.089	0.31	0.344	0.358	0.352	0.368	0.442	0.549	0.133	0.559	0.35	0.791	0.642	0.404	0.359
ASXL1	0.327	0.099	0.286	0.321	0.333	0.369	0.365	0.365	0.461	0.13	0.458	0.377	0.631	0.512	0.377	0.371
BCOR	0.378	0.089	0.307	0.345	0.359	0.351	0.363	0.429	0.54	0.132	0.55	0.342	0.778	0.631	0.395	0.352
SETBP1	0.378	0.088	0.305	0.338	0.355	0.351	0.358	0.423	0.545	0.134	0.555	0.345	0.765	0.621	0.393	0.35
TET2	0.346	0.109	0.285	0.322	0.347	0.314	0.34	0.375	0.443	0.125	0.45	0.307	0.621	0.505	0.348	0.318
NPM1	0.411	0.091	0.285	0.256	0.38	0.372	0.426	0.628	0.387	0.145	0.381	0.275	0.488	0.434	0.292	0.288
WT1	0.329	0.097	0.282	0.308	0.313	0.311	0.311	0.377	0.427	0.128	0.442	0.3	0.599	0.509	0.337	0.304
RUNX1	0.309	0.097	0.277	0.309	0.341	0.329	0.342	0.347	0.427	0.133	0.444	0.312	0.552	0.461	0.329	0.326
FLT3-ITD	0.332	0.088	0.255	0.275	0.289	0.274	0.283	0.317	0.389	0.128	0.397	0.277	0.485	0.47	0.316	0.307
JAK2	0.378	0.09	0.305	0.338	0.355	0.351	0.353	0.423	0.545	0.132	0.555	0.342	0.765	0.621	0.393	0.348

Figure C.4: Mutations present in over 50% of patients, distributed across clusters, for (a) DISC ∩, (b) DISC ∪, (c) MP3 ∩, (d) MP3 ∪, (e) MP3 *geo*, (f) CASet ∪, (g) CASet ∩, (h) CD, (i) AD, (j) BD, (k) 2BD, (l) PCD and (m) MLTD.

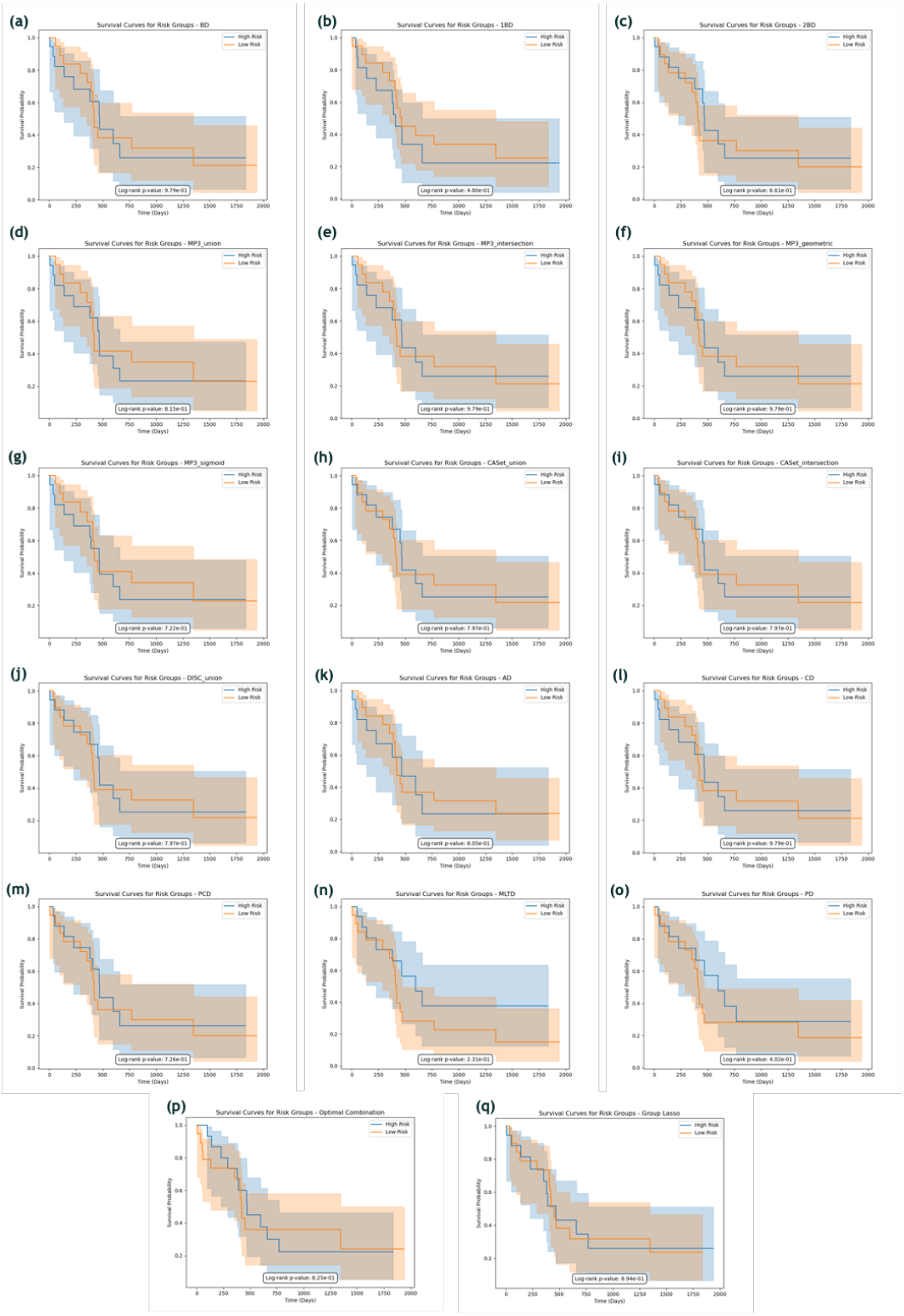


Figure C.5: Survival risk curves for high and low risk groups, for (a) BD, (b) 1BD, (c) 2BD, (d) MP3 ∪, (e) MP3 ∩, (f) MP3 *geo*, (g) MP3 σ,(h) CASet ∪, (i) CASet ∩, (j) DISC ∪, (k) AD, (l) CD, (m) PCD, (n) MLTD, (o) PD, (p) Optimal Combination and (q) Group Lasso.