Unsupervised multiple kernel learning approaches for integrating molecular cancer patient data

Dissertation

zur Erlangung des Grades
des Doktors der Naturwissenschaften
der Fakultät für Mathematik und Informatik
der Universität des Saarlandes

von

Nora K. Speicher

Saarbrücken, 2019

Tag des Kolloquiums:	November 11, 2019
Dekan der Fakultät:	Prof. Dr. Sebastian Hack
Vorsitzender des Prüfungsausschusses:	Prof. Dr. Kurt Mehlhorn
Berichterstatter:	Prof. Dr. Nico Pfeifer
	Prof. Dr. Hans-Peter Lenhof
Akademischer Mitarbeiter:	Dr. Peter Ebert

Abstract

Cancer is the second leading cause of death worldwide. A characteristic of this disease is its complexity leading to a wide variety of genetic and molecular aberrations in the tumors. This heterogeneity necessitates personalized therapies for the patients. However, currently defined cancer subtypes used in clinical practice for treatment decision-making are based on relatively few selected markers and thus provide only a coarse classification of tumors. The increased availability in multi-omics data measured for cancer patients now offers the possibility of defining more informed cancer subtypes. Such a more fine-grained characterization of cancer subtypes harbors the potential of substantially expanding treatment options in personalized cancer therapy.

In this thesis, we identify comprehensive cancer subtypes using multidimensional data. For this purpose, we apply and extend unsupervised multiple kernel learning methods. Three challenges of unsupervised multiple kernel learning are addressed: robustness, applicability, and interpretability. First, we show that regularization of the multiple kernel graph embedding framework, which enables the implementation of dimensionality reduction techniques, can increase the stability of the resulting patient subgroups. This improvement is especially beneficial for data sets with a small number of samples. Second, we adapt the objective function of kernel principal component analysis to enable the application of multiple kernel learning in combination with this widely used dimensionality reduction technique. Third, we improve the interpretability of kernel learning procedures by performing feature clustering prior to integrating the data via multiple kernel learning. On the basis of these clusters, we derive a score indicating the impact of a feature cluster on a patient cluster, thereby facilitating further analysis of the cluster-specific biological properties. All three procedures are successfully tested on real-world cancer data. Comparing our newly derived methodologies to established methods provides evidence that our work offers novel and beneficial ways of identifying patient subgroups and gaining insights into medically relevant characteristics of cancer subtypes.

Zusammenfassung

Krebs ist eine der häufigsten Todesursachen weltweit. Krebs ist gekennzeichnet durch seine Komplexität, die zu vielen verschiedenen genetischen und molekularen Aberrationen im Tumor führt. Die Unterschiede zwischen Tumoren erfordern personalisierte Therapien für die einzelnen Patienten. Die Krebssubtypen, die derzeit zur Behandlungsplanung in der klinischen Praxis verwendet werden, basieren auf relativ wenigen, genetischen oder molekularen Markern und können daher nur eine grobe Unterteilung der Tumoren liefern. Die zunehmende Verfügbarkeit von Multi-Omics-Daten für Krebspatienten ermöglicht die Neudefinition von fundierteren Krebssubtypen, die wiederum zu spezifischeren Behandlungen für Krebspatienten führen könnten.
In dieser Dissertation identifizieren wir neue, potentielle Krebssubtypen basierend auf Multi-Omics-Daten. Hierfür verwenden wir unüberwachtes *Multiple Kernel Learning*, welches in der Lage ist mehrere Datentypen miteinander zu kombinieren. Drei Herausforderungen des unüberwachten Multiple Kernel Learnings werden adressiert: Robustheit, Anwendbarkeit und Interpretierbarkeit. Zunächst zeigen wir, dass die zusätzliche Regularisierung des Multiple Kernel Learning Frameworks zur Implementierung verschiedener Dimensionsreduktionstechniken die Stabilität der identifizierten Patientengruppen erhöht. Diese Robustheit ist besonders vorteilhaft für Datensätze mit einer geringen Anzahl von Proben. Zweitens passen wir die Zielfunktion der kernbasierten Hauptkomponentenanalyse an, um eine integrative Version dieser weit verbreiteten Dimensionsreduktionstechnik zu ermöglichen. Drittens verbessern wir die Interpretierbarkeit von kernbasierten Lernprozeduren, indem wir verwendete Merkmale in homogene Gruppen unterteilen bevor wir die Daten integrieren. Mit Hilfe dieser Gruppen definieren wir eine Bewertungsfunktion, die die weitere Auswertung der biologischen Eigenschaften von Patientengruppen erleichtert. Alle drei Verfahren werden an realen Krebsdaten getestet. Den Vergleich unserer Methodik mit etablierten Methoden weist nach, dass unsere Arbeit neue und nützliche Möglichkeiten bietet, um integrative Patientengruppen zu identifizieren und Einblicke in medizinisch relevante Eigenschaften von Krebssubtypen zu erhalten.

Acknowledgments

First and foremost, I would like to thank Nico Pfeifer for giving me the opportunity to complete my thesis under his supervision. Nico provided guidance, advice, and shared his technical expertise, but at the same time offered trust and freedom, which allowed me to pursue my own ideas. I also wish to thank Thomas Lengauer for creating such an inspiring place to work, introducing me to the basics of statistical learning and providing valuable feedback on my research. I also wish to thank all members of my thesis committee for investing time and effort in the review process.

There are many more people that influenced this thesis and made everyday life and work more fun. Thanks to my former office mates Anna Hake and Peter Ebert for the great discussions. Thanks to the other members of this department for the friendly atmosphere at work. I am particularly thankful to my friends, coffee companions, and (former) colleagues Adrin, Alejandro, Fabian, Lisa, Matthias, Matthias, Nadezhda, Olga, Prabhav, and Sarvesh.

I am very grateful to Achim Büch and Georg Friedrich for their technical support whenever needed, and to Ruth Schneppen-Christmann for taking care of the bureaucratic challenges during the last years.

Last but certainly not least, I am deeply grateful to Benedikt and my family for their support, love, and constant trust.

Contents

List of Figures

Chapter 1

Introduction

Many common diseases are not caused by mutations perturbing the function of a single gene, but instead, exhibit various molecular aberrations and are therefore called multifactorial or complex diseases. One group of them is cancer, the second leading cause of death worldwide [163]. The most common types of cancer are lung and prostate cancer in males, and breast and colorectum cancer in females [73]. Even though an estimated 30-50% of all cases could be prevented by avoiding key risk factors, such as tobacco, alcohol and obesity [163], the worldwide overall incidence is increasing. The number of new cases is predicted to increase from 18.1 millions in 2018 to 24.1 millions in 2030 [72]. At the same time, the number of cancer-related deaths is expected to increase from 9.6 millions in 2018 to 13 millions in 2030. Moreover, it has been shown that cancer is overtaking heart disease as the leading cause of death in several high-income countries [141]. Overall, this shows that despite the improvements made in the medical and pharmaceutical sector, the treatment of cancer still poses a highly relevant, major challenge.

In clinical practice, a tumor in a specific organ is usually classified based on its stage and grade [8]. The most common staging scheme is the so-called TNM system, where T describes the size and the location of the original **t**umor, N reflects to which degree lymph **n**odes are affected, and M describes whether the tumor has already formed **m**etastases in distant body parts. Other staging systems exist, e.g., for brain tumors and blood cancers, where not all of the TNM factors apply. On the other hand, the grade of a tumor is a histological measure describing the cellular appearance of the tumor. In many cases, its grade correlates with the speed at which the tumor grows or spreads, i.e., with its aggressiveness. Generally, the grade varies from 0, if the tumor looks very similar to the healthy tissue, to 4, which indicates that the tumor consists of poorly differentiated tissue. In addition to stage and

grade, known markers, which correlate with clinically relevant characteristics, are considered to make treatment decisions for certain cancer types. Such markers can be of different types, including proteins, DNA mutations, or epigenetic characteristics. One example of a protein being used as a marker is PSA, the prostate-specific antigen, which is measured for the detection and classification of prostate cancer [5]. DNA mutations are considered for instance in lung cancer, where specific mutations in the epidermal growth factor receptor are used to determine the therapy of lung cancer patients [4]. An example of an epigenetic marker is the methylation status of the promoter of the MGMT gene, which is used in glioblastoma to guide treatment decisions [2]. For some cancer types, more complex markers are considered, e.g., the PAM50 test, which is based on the expression of 50 selected genes, distinguishes four clinically relevant subtypes of breast cancer [3, 106]. The combination of stage, grade, and specific markers gives rise to a classification of the tumors, which is used to determine prognosis and treatment.

In contrast to this rough classification based on a few pieces of information, each tumor exhibits a complex landscape of genetic and molecular aberrations. These changes are caused by a sequence of oncogenic events including, e.g., genetic, epigenetic, and regulatory alterations [48]. As a result, tumors that are similar in their phenotype and status of specific markers still might vary strongly in their overall molecular composition. As the currently employed classification is mainly based on histological features and a small number of molecular features, large parts of this diversity are not captured, and, consequently, not considered in treatment decisions. In contrast, modern personalized medicine aims at taking into account the molecular foundation of a tumor [155]. One aspect of personalized medicine is the search for cancer subtypes that are based on a combination of different molecular data, instead of one individual data type. Diagnosing patients with specific subtypes could translate into targeted, and therefore more effective, therapies. The identification of such subtypes is facilitated by large consortia, such as The Cancer Genome Atlas (TCGA), which accumulate molecular data for large patient cohorts covering various cancer types. Publicly available measurements[1] include, amongst others, gene expression, somatic mutations, DNA methylation, protein expression, and miRNA expression. Due to the interconnectedness of the different cellular mechanisms, the different data types are correlated to a certain degree. However, different data types can contribute complementary information, and hence, no individual data type provides the complete information. Moreover, analyzing the different data types jointly can reveal mechanisms in which multiple molecular cellu-

[1] `https://dcc.icgc.org/projects/details`

lar processes (measured by different data types) work in concert to generate the observed outcome [15]. Overall, integrating multiple data types allows accounting for combinatorial effects in cancer or – in other words – to account for a principle that was already formulated by Aristotle: "The whole is greater than the sum of its parts."

Objective of this work

The objective of this work is the development and extension of data integration methods to enable the identification of comprehensive cancer subtypes, i.e, of subtypes identified using multi-omic data. This type of data is also called multidimensional data and enables exploiting synergies and complementary information between the different data types. Therefore, the identified subtypes should reflect both weak signals that are consistent over multiple data types and strong signals in individual data types. The integration of different data types for subtype identification is motivated by the complexity of cancer at the molecular level. For some cancer types, molecular signatures that correlate with relevant clinical parameters have been identified based on individual data types [104, 154]. However, analyzing a combination of multiple data types can lead to comprehensive subtypes, which would pave the way for more patient-specific therapy, implying higher effectiveness and fewer side effects.

Integrating biological data types harbors a number of challenges. One of them is that different data types might have different characteristics, for instance, gene expression data tend to be Gaussian distributed while DNA methylation data generally follow a bimodal distribution. Moreover, molecular data can be high-dimensional, that is, a large number of features is available. In biological settings, these features are often measured for a small number of patients. For data fusion, in which multiple data types are considered, the number of features increases even further while the number of patients remains the same or is even reduced if some patients lack measurements for one or more data types. This combination of few samples with many features leads to the so-called *curse of dimensionality*, a phenomenon of increasing data sparsity in high dimensions [62, Section 2.5]. The sparsity represents a methodical challenge as it implies that Euclidean distances between the samples all converge toward the same value, rendering some approaches ineffective, such as nearest neighbor analysis [7]. Despite these challenges, including additional data types into a biological analysis also provides advantages. Overall, the use of several data sources with certain degrees of pairwise correlations can reduce the influence of the noise that is present in the experimental data. Furthermore, such an integrative analysis

can provide a more detailed picture of the disease and help to understand the involved processes.

Considering the described challenges, we chose multiple kernel learning [54] for the data integration process. This class of methods provides substantial flexibility due to the processing of each input data using specific kernel functions, which can preserve characteristic properties of the respective data type. As mentioned above, this work aims at further increasing the usability of multiple kernel learning in biological settings by addressing different challenges, namely robustness, applicability, and interpretability. Each of the approaches presented in this thesis was implemented and evaluated on real-world data with respect to the biological relevance of the identified subtypes.

Thesis outline

Chapter 2 introduces the relevant background, starting with the biological aspects. Section 2.1 discusses cancer and its relationship to the molecular landscape of the cell. Subsequently, the methodical basis of this thesis is presented, which includes the general idea of machine learning and kernel methods, as well as the extension to multiple kernel learning (Section 2.2). Established dimensionality reduction and clustering methods are introduced in Section 2.3 and 2.4, followed by methods to evaluate clustering results according to both mathematical and biological criteria (Section 2.5). The chapter ends with Section 2.6, which provides an overview of related work in the field of cancer subtyping, as well as general methods that have been proposed for unsupervised data integration.

Chapter 3 is concerned with increasing the robustness of an existing multiple kernel learning framework, which supports the implementation of various dimensionality reduction schemes. After a short overview of the procedure and related approaches (Section 3.1), Section 3.2 describes the methods and the added regularization. Section 3.3 discusses the application of a specific multiple kernel dimensionality reduction scheme to real-world cancer data sets, showing the usefulness of the extended approach.

Chapter 4 focuses on a multiple kernel implementation of principal component analysis. Section 4.1 motivates the interest in this particular dimensionality reduction technique. Section 4.2 and Section 4.3 demonstrate the mathematical limitations of multiple kernel principal component analysis in the graph embedding framework and in general. In Section 4.4, we propose a gain function as an alternative to the traditional objective function for principal component analysis and show results for this gain function in comparison to standard approaches.

Chapter 5 addresses the problem of interpretability for multiple kernel

learning. Section 5.1 provides an introduction on the problem and current approaches aiming at better interpretability of clustering results. Section 5.2 introduces the idea of feature clustering in combination with multiple kernel learning. Section 5.3 presents the methodology and introduces a score that can be used to identify groups of features having high importance for a specific patient cluster. The results for cancer data sets, given in Section 5.4, show that this approach enables the extraction of meaningful hypotheses for the identified patient subgroups.

Chapter 6 completes this thesis with a summary of the methodological advances in the field of unsupervised data integration. Results and biological findings of this work are summarized before highlighting potential future directions for cancer subtyping.

Chapter 2

Background

This Background chapter summarizes the underlying biological and methodological concepts that are necessary to understand the approaches presented in Chapter 3, 4, and 5. Section 2.1 starts with an overview of tumor formation followed by a summary of the molecular data types measured for cancer patients. Section 2.2 provides a general introduction to machine learning and kernel methods, which form the methodological foundation of this thesis. Section 2.3 introduces the concept of dimensionality reduction and two specific algorithms before presenting graph embedding, a general framework that can implement different dimensionality reduction schemes. Clustering approaches, which will later be used to identify groups of cancer patients, are presented in Section 2.4, followed by an overview of different approaches for the evaluation of clustering results in Section 2.5. The Background chapter ends with a review summarizing related work in the field of cancer subtyping and data integration.

2.1 Biological background

2.1.1 Development of cancer

Cancer can arise from different cells in the human body (for instance epithelial cells) that divide uncontrollably. In solid tissue, cancer manifests as a tumor, or a neoplasm, which describes a *center of mass* formed by the body's own cells. Tumors are classified as benign (i.e., noncancerous) or malignant (i.e., cancerous). The latter invade neighboring tissue and can form distant metastases, which are secondary tumors caused by the spread of cells from the primary tumor through the body [35]. There are also different types of blood cancers, which generally do not form a solid tumor but are also charac-

terized by an uncontrolled growth of abnormal cells, in this case blood cells or their progenitors. Besides being named according to their primary site (e.g., lung or breast), tumors are classified based on their cell type of origin. The most common cancer categories are carcinoma (growth from epithelial cells), sarcoma (growth from connective tissue including muscles, bones, etc.), myeloma (growth from the plasma cells of bone marrow), leukemia (growth from the bone marrow), and lymphoma (growth from glands or cells of the lymphatic system) [161].

Regardless of the cell type and organ of origin, cancer cells often share a number of properties. Hanahan and Weinberg [58, 59] summarized them in the so-called *hallmarks of cancer*, which are illustrated in Figure 2.1. First, Hanahan and Weinberg [58] identified six different characteristics: (i) sustaining proliferative signaling, (ii) evading growth suppressors, (iii) resisting cell death, (iv) enabling replicative immortality, (v) inducing angiogenesis, and (vi) activating invasion and metastasis. In a follow-up work, Hanahan and Weinberg [59] extended this original set by two additional cancer hallmarks, which are deregulating cellular energetics, and avoiding immune

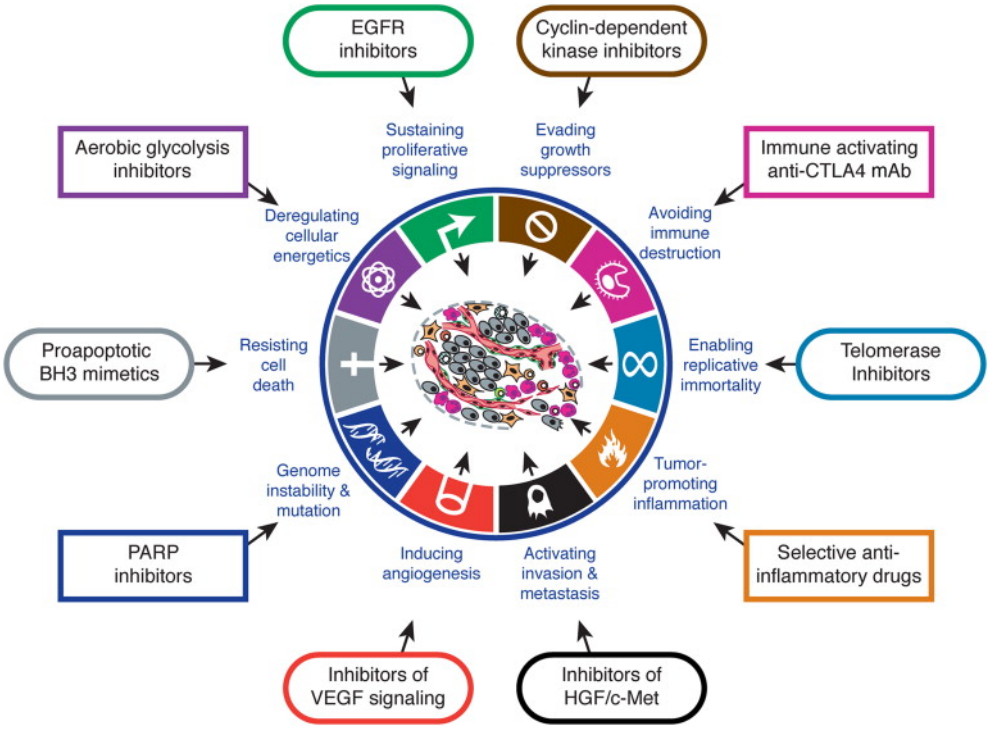


Figure 2.1: The second generation of the hallmarks of cancer identified by Hanahan and Weinberg [59] together with selected drugs targeting the respective characteristics (license #4477730950179, see Table B.1).

destruction. Together, these hallmarks enable the cancer cell population to grow without being controlled by the standard cellular mechanisms, e.g. apoptosis or cell cycle arrest. The same authors proposed two additional characteristics that help the cell acquiring the aforementioned hallmarks. These characteristics are (i) genome instability and mutation, and (ii) tumor-promoting inflammation. Genome instability is prevented in a healthy cell by a maintenance machinery, which repairs or inactivates damaged DNA. If this machinery is defective, the mutation rate increases, which leads to instability of the genome. In this case, a large number of stochastic DNA mutations can accumulate, some of which eventually lead to the formation of one or more of the aforementioned cancer hallmarks [103]. For instance, mutations in TP53 can degrade the ability of the encoded protein (p53) to trigger apoptosis as a response to cellular stress.

This explains why tumors are so diverse, even when originating from the same tissue and cell type: the transformation of a healthy cell into a cancer cell is a multi-step process in which hallmarks can be acquired at different time points via different mechanisms. Even for specific subgroups of cancer types, various driver genes have been identified. For example, the majority of cases of hereditary breast cancer can be attributed to mutations either in the BRCA1 gene, or in the BRCA2 gene, however, a minority of cases cannot be linked to any of the two genes [49]. In addition, there is a multitude of aberrations that can affect a single driver gene for a specific type of cancer. For example, the ClinVar archive for clinically relevant variants [83] lists several thousands of different mutations for BRCA1 and BRCA2 with varying clinical significances for breast cancer. The complexity of cancer is further increased by the large number of possible combinations of oncogenic events happening on different levels of the cell, which include genetic, epigenetic, and regulatory events. Characterizing tumors on the genetic and on the molecular level is hence a challenging yet relevant task. Research consortia such as The Cancer Genome Atlas and the International Cancer Genome Consortium [75] have been founded with the goal of systematically charting cancer heterogeneity by collecting cancer data from large patient cohorts. These data will be described in the next section.

2.1.2 Molecular data

Real-world cancer data sets comprise for each patient a number of different molecular data types. These data cover genomic, epigenomic, and transcriptomic measurements. Previous studies could show correlations between these data and the outcome of a patient, e.g., recurrence, or response to treatment. In the following, we will shortly introduce the different data types that were

available in the data sets used in our analysis and their relevance to oncogenesis and tumor progression.

Gene expression Gene expression belongs to the field of transcriptomics. While in general every cell contains the same genetic material, gene expression varies widely between cells and is measured by the number of copies of messenger RNA (mRNA) that are synthesized via transcription. For a given DNA segment representing, e.g., a gene or a regulatory element, expression differences in healthy cells occur, for instance, due to tissue specificity or due to time [165]. Nevertheless, some general trends have been observed, which distinguish gene expression profiles of cancer cells from those of healthy cells. First, tumor cells generally overexpress oncogenes, which are genes that are related to functions such as cell-cycle progression or apoptosis inhibition [95]. At the same time, tumor suppressor genes, which are mainly genes inhibiting cell proliferation, e.g., via DNA repair or cell-cycle control, are often silenced in tumor cells [95]. Despite these common trends, gene expression profiles vary between cancer patients. Exploiting the differences between patients, van de Vijver et al. [152] were able to train a supervised model affording a more accurate prognosis for breast cancer patients on the basis of the expression of 70 genes than using clinical and histological data. Perou et al. [110] identified subtypes of breast cancer (Luminal A, Luminal B, HER2-enriched, and Basal-like) using unsupervised analysis of gene expression data. Subsequently, a set of 50 genes (PAM50) was identified whose expression can be used to predict these subtypes accurately [106]. For various other cancer types, gene expression subtypes were also shown to add valuable insights, which could not be retrieved from the available clinical data. For gliomas, unsupervised analysis of cancer patients based on gene expression data resulted in patient groups having a higher correlation with survival than histology-based subtypes [56]; Verhaak et al. [154] defined four subtypes of the aggressive brain tumor glioblastoma (Proneural, Neural, Classical, and Mesenchymal subtype) that differ significantly concerning their response to treatment; Chung et al. [34] defined four gene expression subtypes of head and neck squamous cell carcinoma that differ significantly in their recurrence-free survival (i.e., the time to disease relapse or death).

The following sections introduce miRNA expression, copy number alterations, and DNA methylation, all of which are influenced by the expression of specific genes, such as DNA polymerase, and DNA methyltransferase. Vice versa, the expression of a gene is influenced by its copy number and regulatory processes. These regulatory processes can be induced by DNA methylation or miRNA expression, as will be described in the respective sections.

miRNA expression Like gene expression, miRNA expression levels are transcriptomic characteristics of the cell. microRNAs (miRNAs) are non-coding RNA fragments of 19-25 nucleotides length. By sequence complementarity, a miRNA can bind to an mRNA and, thereby, induce post-transcriptional silencing of the respective gene. In cancer cells, most miRNAs have low expression levels. This low abundance of miRNAs can be a selective advantage for the respective cell given that miRNAs can help inactivating oncogenes and inducing apoptosis. For instance, the miR-34 group that comprises three miRNAs involved in apoptosis or cell-cycle arrest is often inactivated in cancer cells by DNA methylation [65]. There are some exceptions, such as miR-21, which inactivates the tumor suppressor gene PTEN, and is highly expressed in different cancer types including breast, lung, prostate, and colorectal cancer [14]. Cancer type-specific miRNA expression patterns can be used to predict the tissue of origin more accurately than gene expression data, which is particularly important for cancers of unknown primary tissue [30]. Besides, differences in miRNA expression profiles exist between patients with the same cancer type (e.g., shown in lung cancer [97]). Therefore, miRNA expression data were included in integrative subtype identification approaches, e.g., a correlation between miRNA subtypes and gene expression subtypes was reported for breast tumors [146].

Copy number alterations Copy number alterations belong to the field of genomics, as this term directly refers to the DNA that is available in the cells. Human cells are usually diploid, i.e., they contain two copies of the DNA with the exception of the sex-determining chromosomes in men. Copy number alterations are structural variations in the genome, comprising duplications and deletions, that lead to either more or fewer than two copies of a DNA segment. A copy number alteration is usually related to the expression level of the gene in which it occurs, copy number aberrations in regulatory regions can even influence the expression of sets of genes. While copy number alterations do not necessarily lead to an altered phenotype [171], copy number alterations are so commonly observed in cancer genomes that this structural instability is considered one of the characteristic features of cancer (see Figure 2.1). This feature has been found to contribute to the hallmarks of cancer in various ways, e.g., some cancer cells have an amplification of the MCL1 and BCL2L1 anti-apoptotic genes, which helps the cancer cell to evade apoptosis [20]. Although some cancer type-specific patterns occur [20], even copy number profiles of patients with the same cancer type exhibit high variance. In a study of small-cell lung cancer patients, copy number profiles of circulating tumor cells could be used to predict chemosensitivity [29].

DNA methylation DNA methylation is a heritable, epigenetic modification of the DNA. The term describes the addition of methyl groups to the DNA molecule. In contrast to the sequence of the DNA, methylation patterns are dynamic and influenced by environmental factors, e.g., cigarette smoking [87]. DNA methylation occurs predominantly in cytosines that are part of a CpG dinucleotide with approximately 70-80% of CpGs being methylated [22]. However, CpG islands, which are regions of DNA of at least 200 bases with a high frequency of CpG sites, generally exhibit a low abundance of DNA methylation. CpG islands often overlap with promoter regions and are subject to tissue-specific methylation [19]. The regulatory effect of the modification on gene activity depends on the position of the methylation: DNA methylation in promoter regions generally leads to the silencing of the respective genomic region, while methylation of the gene body was found to be positively correlated with expression [167].

Cancer cells often exhibit a global hypomethylation in comparison to healthy cells, which can lead to chromosomal instability, thereby influencing the previously described copy number aberrations [45]. In contrast, high levels of methylation were found in promoters of tumor suppressor genes leading to their inactivation [46]. Because of its reversible character, DNA methylation is being explored as a potential target for cancer therapy, for example, aiming at inhibiting the methylation of promoter regions of tumor suppressor genes [74]. Subtypes with distinct clinical outcomes were identified for various cancer types using DNA methylation data, e.g., CpG island methylator phenotypes were identified in colorectal cancer [151] and in glioblastoma [104].

Overall, the considered cell properties interact in a complex machinery leading to a specific cellular phenotype. For example, a recent study reported that resistance to the drug ABT-199 in lymphoma depends on mechanisms that involve both genetic and non-mutational characteristics [175]. Due to their mutual regulatory activities, one expects correlations between the respective data types. Additionally, the different data types can harbor complementary information, for example, gene expression can change due to epigenetic regulation mechanisms (e.g., DNA methylation) or copy number variations. Data integration methods should ideally be able to uncover concordant signals in different data types and also include strong individual signals. Patients are then not only differentiated by their gene expression profile, but also by the mechanisms that influence this profile.

2.2 Machine learning and kernel methods

Due to recent developments such as autonomous driving, personalized movie recommendations, and robot assistants, there is a growing public interest in machine learning. Applications in the medical field, such as evaluation of electronic health records, diagnosis based on medical images, epidemic outbreak prediction, or robotic surgery contribute to this trend [105, 84]. Being a diverse area, the general aim of machine learning methods can be described as detecting patterns or drawing conclusions from given data.

The data consist of samples x_i (e.g., patients), which are described by features (e.g., genes). In certain scenarios, sample-specific labels or outcomes y_i (e.g., age or severity of the disease) are known. The field of machine learning can roughly be divided into three different settings: (i) supervised learning, where known outcomes y_i for the samples x_i are used to train the model; (ii) unsupervised learning, where the outcome of interest y_i is not used when training the model; and (iii) semi-supervised learning, where the outcome y_i is known only for a subset of the samples x_i.

For supervised learning, a training data set is given, which provides for each sample x_i (described by a set of features) a known outcome y_i. Using this data set, a model for predicting the label y_i given the input x_i is trained via minimizing a *loss function* or *objective function*. A classical loss function is the *squared error loss*

$$L = (\widehat{y} - y)^2 \tag{2.1}$$

with $\widehat{y}$ being the model predictions that are compared to the real outcome y [62, Chapter 2]. When training the model, there are parametric and non-parametric approaches. In the parametric setting, $\widehat{y}$ is obtained as $f(x)$, where f is a function with a predefined shape (e.g., linear) whose parameters are learned such that the given loss function L is minimized. Non-parametric methods, on the other hand, do not assume a predetermined shape of f and thus provide more flexibility. Moreover, supervised methods can be distinguished according to the type of outcome that they predict: Either classification is performed, i.e., predicting a categorical outcome, or regression, i.e., predicting a quantitative outcome. An important question when evaluating supervised learning methods is how well they generalize, i.e., how accurate the learned relationship predicts the outcome for new, unseen samples. This generalization ability should be tested on a data set that is disjoint from the previously used training data set.

In contrast to supervised methods, unsupervised learning methods do not use the outcome of interest in their optimization process. The main goal of unsupervised learning is the analysis or discovery of structures in the

data set. This can be achieved via dimensionality reduction, i.e., projecting samples into a low-dimensional subspace with minimal loss of information (Section 2.3), or clustering, i.e., identifying homogeneous groups of samples (Section 2.4).

Semisupervised learning approaches are used for data sets in which only a subset of samples has a known label y_i. Information that is available for all samples, e.g., their pairwise distances are combined with the available labels. In this way, the partial information of the labels can guide the learning process, which is not possible in unsupervised learning.

2.2.1 Kernel methods

Learning methods are often based on assumptions concerning the distribution of the data, e.g., some methods expect linearity in the mapping function f or convex compact clusters. However, in many cases these linearity assumptions do not hold for the data as described by their available features. Kernel functions provide one way of handling nonlinearity, e.g., in the boundaries between the clusters. The use of kernel functions is best known in the context of support vector machines, a supervised classification method [37], but they were also integrated into other supervised and unsupervised methods [126, 127]. Intuitively, a kernel function provides the similarity of samples mapped into a different (typically high-dimensional) feature space. Due to the change of basis vectors, linear relationships in the feature space may correspond to nonlinear relationships in the original data space. Consequently, a linear model learned in the feature space can provide a nonlinear model in the data space.

We consider a function $\phi : x_i \to \phi(x_i)$ that maps the data points into a possibly infinite-dimensional feature space. If this function is chosen "wisely", the problem of interest will be easier (e.g., linear) in the feature space. In many cases, the feature space is unknown or its construction is time-consuming. Therefore, the *kernel trick* enables making use of this mapping procedure in algorithms that can be formulated on the basis of inner products between the samples instead of using the coordinates of the data points directly. In these cases, the kernel trick means that we can replace the original inner product with the kernel function, which renders the explicit construction of the feature space obsolete. Applying a kernel function k to the data points x_i, x_j only implicitly maps these data points into the feature space. Nevertheless, the calculated kernel value is the inner product between two projected points $\phi(x_i)$ and $\phi(x_j)$:

$$k(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle. \tag{2.2}$$

Kernel functions give rise to so-called kernel matrices, or Gram matrices, defined by $K_{ij} := k(x_i, x_j)$. Kernel matrices are symmetric by construction, rotationally invariant, and capture all information of the input data set that is relevant for the application of kernel-based algorithms. As kernel matrices represent inner products, it can be shown that each kernel matrix is positive semidefinite, i.e., for any vector v holds $v^T K v \geq 0$ [128]. Vice versa, any kernel function that generates a symmetric positive semidefinite kernel matrix can be decomposed into mapping the data into a Hilbert space before calculating the inner product in that space, i.e., it can be used to construct a Hilbert space with the reproducing property, a *reproducing kernel Hilbert space* (RKHS). Further explanations and mathematical aspects of this concept can be found in Schölkopf and Smola [125] and Shawe-Taylor and Cristianini [128].

Radial basis kernel function (RBF) Given the large variety of application scenarios for kernel methods, a multitude of kernel functions exist of which many have been designed to exploit specific properties of certain data types. The discussion here is limited to the popular RBF (or Gaussian) kernel [25], which was used for the work in Chapter 3, 4, and 5. This kernel function is defined by

$$k(x_i, x_j) = \exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma^2}\right) = \exp(-\gamma\|x_i - x_j\|^2). \tag{2.3}$$

The hyperparameter σ (or γ) influences the variance of the Gaussian, i.e., how fast kernel values decrease with increasing distance of the samples. In general, for each sample x_i holds that $k(x_i, x_i) = 1$. Consequently, in the RKHS, each projected sample $\phi(x_i)$ has a distance of unit length 1 to the origin. Additionally, the mapped data points all lie in the same orthant, as all kernel values (i.e., inner products) are positive, which corresponds to an enclosed angle smaller than $\pi/2$ [125].

The RBF kernel is defined on real-valued data, however, there are kernel functions for other data types, such as graph kernels [156, 132] or sequence kernels [88]. Since the cancer patients in the data sets we used were described by numerical data, these other types of kernels are not considered in this thesis.

2.2.2 Multiple kernel learning

Kernel matrices have a number of closure properties that enable the construction of new kernels based on one or more known kernel matrices. Mathemat-

ical operations that preserve the properties of positive semidefiniteness and symmetry include addition and multiplication with a positive scalar [128].

Multiple kernel learning (MKL) uses this fact by optimizing a weight vector β that linearly combines a set of input kernel matrices $\{K_1, ..., K_M\}$ to generate a unified *ensemble kernel matrix* $\mathbb{K}$, such that

$$\mathbb{K} \;=\; \sum_{m=1}^{M} \beta_m K_m, \qquad \beta_m \geq 0 \quad \forall m \in \{1, ..., M\}. \tag{2.4}$$

Consequently, a kernel approach can be extended to handle several kernels by additionally optimizing the kernel weight vector β. The optimization is then performed according to the objective function of the respective algorithm subject to the additional constraints concerning the kernel weights.

There are different possibilities to generate the individual kernel matrices that should be integrated via MKL. Chapter 3 and 4 present results where each kernel matrix is generated on all features of one specific data type. Chapter 5 extends this idea by using different sets of features to generate multiple kernel matrices per data type.

Centering data in the RKHS Equivalently to the mean in the original space, the mean vector, or center of mass, in the RKHS is defined as

$$\overline{\phi(X)} = \frac{1}{N} \sum_{i=1}^{N} \phi(x_i). \tag{2.5}$$

Consequently, the data can be centered in the RKHS by moving the mean to the origin. As shown in the following, this can be done implicitly by exploiting the fact that the kernel function $k(a, b)$ calculates the inner product of $\phi(a)$ and $\phi(b)$, with a and b being two arbitrary data points.

$$\begin{aligned} k_c(a,b) \;&=\; \langle \phi_c(a), \phi_c(b) \rangle = \left\langle \phi(a) - \overline{\phi(X)},\; \phi(b) - \overline{\phi(X)} \right\rangle \\ &= \left\langle \phi(a) - \frac{1}{N}\sum_{i=1}^{N} \phi(x_i),\; \phi(b) - \frac{1}{N}\sum_{i=1}^{N} \phi(x_i) \right\rangle \\ &= k(a,b) - \frac{1}{N}\sum_{i=1}^{N} k(a,x_i) - \frac{1}{N}\sum_{i=1}^{N} k(b,x_i) + \frac{1}{N^2}\sum_{i,j=1}^{N} k(x_i,x_j) \end{aligned} \tag{2.6}$$

Unlike the uncentered kernel matrix K, the centered kernel matrix K_c contains negative entries and has a row, column, and matrix mean of zero.

Spectral normalization of the data in the RKHS The norm of a data point that is mapped into the RKHS $\phi(a)$ can be calculated by

$$\|\phi(a)\|_2 = \sqrt{\langle \phi(a), \phi(a) \rangle} = \sqrt{K(a,a)}. \tag{2.7}$$

This equation facilitates different normalizations of the kernel matrix. We use the spectral normalization, which ensures that the distances of the samples to the origin in the RKHS are equal to one.

$$\begin{aligned} k_{\text{norm}}(a,b) &= \left\langle \frac{\phi(a)}{\|\phi(a)\|_2}, \frac{\phi(b)}{\|\phi(b)\|_2} \right\rangle \\ &= \frac{\langle \phi(a), \phi(b) \rangle}{\|\phi(a)\|_2 \, \|\phi(b)\|_2} = \frac{k(a,b)}{\sqrt{k(a,a)k(b,b)}} \end{aligned} \tag{2.8}$$

The normalized kernel matrix K_{norm} has a variance of N, the number of samples, as can be seen by the diagonal entries, which are set to one when applying Equation (2.8).

When integrating multiple kernel matrices, each kernel matrix is normalized to avoid arbitrary differences in variance, which would influence the optimization of the kernel weights. For the application of all approaches that are presented in the following chapters, the kernel matrices were first centered using Equation (2.6) and then normalized using Equation (2.8).

2.3 Dimensionality reduction

Data available for machine learning in medical settings often have more features than samples, i.e., the samples are distributed in a high-dimensional space. In particular for cases where the number of features is much larger than the number of samples (short: $p \gg N$), the *curse of dimensionality* results in sample sparsity, which makes local neighborhoods hard to identify [62, Section 2.5]. The challenging situation of the curse of dimensionality is commonly remedied by performing feature selection. Feature selection has the purpose of reducing the total number of features as much as possible while retaining most of the information in the data, i.e., uninformative features are discarded and informative ones are kept in the data set. Alternatively, dimensionality reduction procedures identify a low-dimensional subspace into which the data are projected while keeping the loss of information as small as possible. Dimensionality reduction can be used as a preprocessing step before applying the algorithm of interest or for visualization purposes. The approach can be performed in a supervised manner (i.e., including the knowledge of

all sample labels), in an unsupervised manner (i.e., not including any information on the sample labels), or in a semi-supervised manner (i.e., including the knowledge of some of the sample labels). Depending on the availability of labeled data, on the aim of the analysis, and on the assumptions about the data, different parts of the structure in the data need to be preserved. This gives rise to a number of dimensionality reduction schemes based on different objective functions. In the following sections, the two unsupervised approaches *locality preserving projections* [63] and *principal component analysis* [109] will be discussed. Both operate by optimizing a set of new basis vectors according to the respective objective function. The data are projected onto these basis vectors using a mapping function. Dimensionality reduction is achieved by reducing the number of basis vectors. In both approaches, new data points can still be easily included after the optimization, in contrast to other methods, where the mapping function is not known (e.g., *t-distributed stochastic neighbor embedding* [153]). The inability to project unseen samples into the same space is often referred to as the *out-of-sample problem*. Dimensionality reduction methods depend on a parameter p, which is the number of dimensions that is used for the projection. Usually, the user has to choose a reasonable value either according to practical reasons (for instance if the dimensionality reduction is performed for visualization), prior knowledge, or a heuristic such as the elbow method [6]. In the following sections, we will mainly formulate the dimensionality reduction methods using a projection vector, i.e., $p = 1$. However, projection into a multidimensional space is possible by optimizing a projection matrix.

2.3.1 Locality preserving projections

As the name suggests, locality preserving projections (LPP) [63] is a method that projects the data points into a subspace with the aim of preserving local structures. Locality is defined on the basis of neighborhood graphs in which two neighboring data points are connected by an edge. To construct the graph He and Niyogi [63] propose to use either ϵ-neighborhoods (i.e., two nodes x_i and x_j are neighbors if $\|x_i - x_j\|^2 < \epsilon$) or k-nearest neighbors (i.e., two samples x_i and x_j are neighbors if x_i is among the k data points that are closest to x_j or vice versa). However, other neighborhood graphs are possible depending on the data of interest. Irrespective of its construction, the aim of the approach is the preservation of the defined neighborhood graph. The weight w_{ij} of the edge connecting x_i and x_j can also be set in different ways: either in a uniform manner where the weight for each existing edge w_{ij} is set

to 1, or by using a heat kernel[1] $w_{ij} = \exp\left(-\frac{\|x_i - x_j\|^2}{t}\right)$, which considers the squared euclidean distance between x_i and x_j and depends on the parameter t. The weights between non-connected (non-neighboring) nodes are always zero, resulting in a sparse symmetric weight matrix W of size $n \times n$. The first basis vector v for the projection optimizes the following objective function

$$\begin{aligned} &\arg\min_{\boldsymbol{v}} \sum_{i,j \in N} (v^T x_i - v^T x_j)^2 \, W_{ij} \\ &\text{subject to } \sum_i (v^T x_i)^2 D_{ii} = 1, \end{aligned} \tag{2.9}$$

with $D_{ii} = \sum_j W_{ij}$, which is the degree of sample x_i in the neighborhood graph. Intuitively, the objective function penalizes two neighbors with a high edge weight W_{ij} being projected far away from each other while imposing no restrictions on the projection of non-neighboring points. The constraint is necessary to avoid the trivial solution of obtaining the minimal sum of distances by projecting each point to the origin.

The solution of Problem (2.9) are the eigenvectors corresponding to the smallest eigenvalues for the generalized eigenvalue problem:

$$XLX^T v = \lambda XDX^T v, \tag{2.10}$$

with $L = D - W$ being the Laplacian matrix and λ the corresponding generalized eigenvalue.

This approach results in a linear projection of the data points, that is, $\text{proj}(x_i) = v^T x_i$. Using the kernel trick, LPP can be conducted in the reproducing kernel Hilbert space, thereby facilitating nonlinear projections. The results of kernel LPP are equivalent to those obtained by Laplacian Eigenmaps [16]. However, as opposed to Laplacian Eigenmaps, LPP affords also projecting new samples into the learned space because LPP generates the actual projection vectors, and not just the new coordinates of the projected samples.

2.3.2 Principal component analysis

Principal component analysis (PCA) [109] can be used for dimensionality reduction. The data points are transformed from the original coordinate system into a new orthogonal basis spanned by the principal components.

[1] He and Niyogi [63] use the term *heat kernel*, which is motivated from the physical process of heat diffusion over time t. As we can see in their definition, it is equivalent to the RBF kernel with $t = 2\sigma^2$ (cf. Formula 2.3).

The first principal component is the axis along which the data points have the highest variance, with the variance of a random variable X being defined as $\mathrm{Var}(X) = E[(X - E(X))^2]$. The first principal component v optimizes

$$\arg\max_{v} \ \mathrm{Var}(Xv), \quad \|v\| = 1, \tag{2.11}$$

with $X \in \mathbb{R}^{N \times d}$ being the data matrix describing N samples and d features. Subsequent principal components are orthogonal to each other and sorted by decreasing variance in the samples. The first p principal components are the eigenvectors associated to the p largest eigenvalues of the sample covariance matrix. The respective eigenvalues give the variance in each of these components. The objective function of PCA is motivated by the assumption that the directions with the smallest variance mainly represent noise. Therefore, a projection that only uses the first p principal components, with p being smaller than the original dimensionality of the data, will still contain large parts of the (biologically) meaningful variation. As a consequence of this choice of projection basis vectors, the reconstruction error, which is the sum of the euclidean distances between each data point x_i and its projection $v^T x_i$, is minimized. In contrast to LPP, which preserves local structures, PCA is a global dimensionality reduction method since the variance is a global characteristic of all data points combined.

As for LPP, a kernel version of PCA exists, which identifies the directions of maximum variance in the reproducing kernel Hilbert space [127]. PCA can thus be used for linear, kernel PCA for nonlinear projections.

2.3.3 Graph embedding

The dimensionality reduction methods described above optimize distinct objective functions with different constraints. Yan et al. [166] showed that many dimensionality reduction approaches can be formulated under the common framework of so-called graph embeddings. The graph embedding framework affords a straightforward implementation of dimensionality reduction methods that only requires adjusting two parameters. These parameters are an adjacency matrix W defined by a similarity graph G, and a constraint matrix D (or W'), which will be described in the following.

Assume G being an undirected similarity graph with edge weights W, which reflect similarities that are to be preserved. Then, the projection vector v (for the projection into a one-dimensional subspace) or the projection matrix $V \in \mathbb{R}^{d \times p}$ (for the projection into p dimensions) is optimized based

on the following graph-preserving criterion:

$$\arg\min_{v} \sum_{i,j=1}^{N} \|v^T x_i - v^T x_j\|^2 w_{ij} \tag{2.12}$$

$$\text{subject to} \sum_{i=1}^{N} \|v^T x_i\|^2 d_{ii} = C, \quad \text{or} \tag{2.13}$$

$$\sum_{i,j=1}^{N} \|v^T x_i - v^T x_j\|^2 w'_{ij} = C \tag{2.14}$$

with C being a positive constant. Equations (2.13) and (2.14) are two alternative constraints used for different dimensionality reduction techniques. These constraints are either based on penalty graph G' with edge weights W' representing similarities that are to be suppressed in the learned projection or on a diagonal matrix D for scale normalization. The minimization problem preserves the given graph structure because pairs of data points with high similarity w_{ij} are forced to be projected close together. At the same time, smaller and especially negative similarities cause larger distances in the resulting projections.

As mentioned above, the choice of adjacency matrix W and constraint matrix D (or W') determines the dimensionality reduction scheme that is implemented. For the methods used in this work, the respective matrices are listed in Table 2.1. The formulation of LPP is based on either one of the two neighborhood graphs introduced in Section 2.3.1, while PCA is based on a complete graph with uniform weights reflecting the global optimization criterion. Furthermore, both methods are formulated using a diagonal matrix D in the constraint for scale normalization.

2.3.3.1 Nonlinear extension of graph embedding

In Formula (2.12), dimensionality reduction is achieved by linear projection of the data point, i.e., $\text{proj}(x_i) = v^T x_i$. Extending the formulation using the kernel trick enables the optimization of nonlinear projections. The kernel trick corresponds to implicitly mapping the samples into a RKHS using a function $\phi : x_i \to \phi(x_i)$ (see Section 2.2). It can be shown that the optimal projection vector v lies in the span of the data points, the projection vector v can thus be represented as a weighted linear combination of the data points in the RKHS, i.e.,

$$v = \sum_{n=1}^{N} \alpha_n \phi(x_n), \tag{2.15}$$

Table 2.1: Similarity and constraint matrices for different dimensionality reduction schemes in the graph embedding framework. $\mathcal{N}_k(x_i)$ represents the set of the k nearest neighbors of sample x_i.

Algorithm	Similarity matrix	Constraint matrix
LPP	$w_{ij} = \begin{cases} s, & \text{if } \|x_i - x_j\|^2 < \epsilon \\ 0, & \text{else,} \end{cases}$ or $w_{ij} = \begin{cases} s, & \text{if } x_i \in \mathcal{N}_k(x_j) \vee x_j \in \mathcal{N}_k(x_i) \\ 0, & \text{else} \end{cases}$ with $s = 1$ or $s = exp\left(\frac{-\|x_i - x_j\|^2}{t}\right)$	$d_{ii} = \sum_{i \neq j} w_{ij}$
PCA	$w_{ij} = \begin{cases} \frac{1}{N}, & \text{if } i \neq j \\ 0, & \text{else} \end{cases}$	$d_{ii} = 1$

where α is the sample coefficient vector for the specific projection vector v. As the kernel matrix consists of the pairwise inner products of points in the RKHS, i.e., $K_{ij} = \phi(x_i)^T \phi(x_j)$, Equation (2.15) can be used to formulate the projection of a point x_i into a one-dimensional space using only the kernel matrix as follows:

$$\text{proj}(x_i) = \left(\sum_{n=1}^{N} \alpha_n \phi(x_n) \right)^T \phi(x_i) = \alpha^T K_i \tag{2.16}$$

with K_i being the ith column of the kernel matrix K. Consequently, nonlinear sample projections can be obtained by kernelization of the graph embedding framework as follows:

$$\begin{aligned} \arg\min_{\alpha} \quad & \sum_{i,j=1}^{N} \|\alpha^T K_i - \alpha^T K_j\|^2 w_{ij} \\ \text{subject to} \quad & \sum_{i=1}^{N} \|\alpha^T K_i\|^2 d_{ii} = C, \quad \text{or} \\ & \sum_{i,j=1}^{N} \|\alpha^T K_i - \alpha^T K_j\|^2 w'_{ij} = C \end{aligned} \tag{2.17}$$

with

$$\alpha \;=\; [\alpha_1 \cdots \alpha_N]^T \in \mathbb{R}^N. \tag{2.18}$$

Using the definition of the Laplacian matrix L of a graph, $L = D - W$ with $d_{ii} = \sum_j w_{ij}$ and W being the adjacency matrix as defined before, the objective function can be reformulated as

$$\begin{aligned}\sum_{i,j=1}^{N} \|\alpha^T K_i - \alpha^T K_j\|^2 w_{ij} &= 2\left(\sum_{i=1}^{N} \alpha^T K_i d_{ii} K_i^T \alpha - \sum_{i,j=1}^{N} \alpha^T K_i w_{ij} K_j^T \alpha\right) \\ &= 2\left(\alpha^T K(D-W)K^T\alpha\right) \\ &= 2\alpha^T KLK^T\alpha. \end{aligned} \tag{2.19}$$

Including the first of the two alternative constraints, this leads to the optimization problem

$$\begin{aligned} \arg\min_{\alpha} \quad & \alpha^T KLK^T\alpha \\ \text{subject to} \quad & \alpha^T KDK^T\alpha = C. \end{aligned} \tag{2.20}$$

A widely used strategy for solving constrained optimization problems is the Lagrangian function and Karush-Kuhn-Tucker conditions (further details can be found in [23, Appendix E]). In the considered setting, the corresponding Lagrangian function and the derived Karush-Kuhn-Tucker conditions show that the problem can be solved via the generalized eigenvalue problem

$$KLK^T\alpha = \lambda KDK^T\alpha. \tag{2.21}$$

The optimal sample coefficient vector α is the generalized eigenvector associated with the minimum generalized eigenvalue. In general, projections into p-dimensional spaces are achieved using the eigenvectors that correspond to the p smallest generalized eigenvalues.

2.3.3.2 Multiple kernel extension of graph embedding

The kernelized version of the constrained optimization problem (2.17) can be extended to integrate several kernel matrices via multiple kernel learning [92]. Replacing the kernel matrix in the optimization problem with the ensemble kernel matrix (cf. Formula 2.4) yields the following optimization problem:

$$\begin{aligned} \arg\min_{\alpha,\beta} \quad & \sum_{i,j=1}^{N} \|\alpha^T \mathbb{K}^{(i)}\beta - \alpha^T \mathbb{K}^{(j)}\beta\|^2 w_{ij} \\ \text{subject to} \quad & \sum_{i=1}^{N} \|\alpha^T \mathbb{K}^{(i)}\beta\|^2 d_{ii} = C \qquad \text{or} \end{aligned}$$

$$\sum_{i,j=1}^{N} \|\alpha^T \mathbb{K}^{(i)}\beta - \alpha^T \mathbb{K}^{(j)}\beta\|^2 w'_{ij} = C$$
$$\beta_m \geq 0, \quad m = 1, 2, ..., M. \tag{2.22}$$

where

$$\beta = [\beta_1 \cdots \beta_M]^T \in \mathbb{R}^M, \tag{2.23}$$
$$\mathbb{K}^{(i)} = \begin{pmatrix} K_1(1,i) & \cdots & K_M(1,i) \\ \vdots & \ddots & \vdots \\ K_1(N,i) & \cdots & K_M(N,i) \end{pmatrix} \in \mathbb{R}^{N \times M}. \tag{2.24}$$

This problem requires the optimization of two entities, the projection vector α (or more generally the projection matrix $A \in \mathbb{R}^{N \times p}$ for reduction into p dimensions) and the kernel weight vector β. This is often achieved via coordinate descent, which iteratively optimizes the two variables in an alternating manner [92].

2.4 Clustering

Unlike dimensionality reduction, which aims at projecting samples into a low-dimensional space, clustering is the task of identifying groups of samples in the data. Generally, each cluster should have high intra-cluster similarity and high inter-cluster dissimilarity. If both properties are fulfilled, the clusters are described as dense and well-separated.

2.4.1 K-means clustering

Given a data matrix $X \in \mathbb{R}^{N \times d}$ describing N samples x_i with d features. K-means [61] is a widely used algorithm that identifies K clusters, which are sets of samples $\boldsymbol{C} = \{C_1, ..., C_K\}$, by minimizing the objective function

$$\arg\min_{\boldsymbol{C}} \sum_{k=1}^{K} \sum_{x_i \in C_k} \|x_i - \mu_k\|^2 \tag{2.25}$$

with μ_k representing the mean of cluster C_k. The number of clusters K is a hyperparameter that needs to be set a priori by the user. Intuitively, k-means learns cluster memberships such that the within-cluster scatter, which is defined using the euclidean distance, is minimized. Since the optimization problem is computationally difficult, an iterative, local heuristic (i.e., a greedy approach) is used. The heuristic repeats the following two steps:

1. identify the cluster means (in the first step at random, later according to the current cluster assignment of the samples),

2. assign each sample to the cluster whose mean is closest to the sample.

Due to the greedy approach, the random initialization of the cluster means in Step 1 might result in different locally optimal cluster assignments. Therefore, the procedure is usually repeated several times in order to find the global optimum of the objective function.

Due to the euclidean distance in the objective function, the approach favors spherical clusters of approximately the same size, while it is not suited for long, snake-like clusters or clusters that strongly differ in size. To deal with this limitation when using k-means, one can apply a nonlinear dimensionality reduction method beforehand. Alternatively, kernel k-means clusters the samples after projection in the RKHS, analogous to previously discussed applications of the kernel trick [126].

2.4.2 Fuzzy c-means clustering

Similar to k-means, fuzzy c-means [44, 21] identifies K cluster centers μ_k. However, instead of learning a binary cluster assignment (also called hard clustering), where each sample is assigned to exactly one cluster, a fuzzy clustering is generated. In a fuzzy clustering, each sample x_i has a membership probability for each cluster C_k, the so-called degree of membership $u_{i,k}$. These probabilities are obtained by minimizing the following objective function:

$$\begin{aligned} \underset{U,\mu}{\arg\min} \quad & \sum_{i=1}^{N}\sum_{k=1}^{K} u_{i,k}^{f}\|x_i-\mu_k\|^2 \\ \text{subject to} \quad & \sum_{k=1}^{K} u_{i,k} = 1 \quad \forall i \\ & u_{i,k} \geq 0 \quad \forall i,k \\ & \sum_{i=1}^{N} u_{i,k} > 0 \quad \forall k. \end{aligned} \tag{2.26}$$

The resulting U is a matrix of size $N \times K$ containing the degrees of cluster memberships $u_{i,k}$. The result depends on $f \geq 1$, a parameter of the method that controls the degree of fuzzification. Choosing $f = 1$ results in a hard clustering of the samples, i.e., $u_{i,k} \in \{0, 1\}$, whereas choosing $f \to \infty$

results in uniform cluster probabilities $u_{i,k} = 1/K$ for all samples x_i and clusters C_k. Similar to k-means, the optimization problem is solved by a local optimization heuristic, such that the algorithm might return a local optimum. Repeating the method multiple times generates more stable results. In contrast to k-means, fuzzy clustering provides additional information on the reliability of the cluster assignment for each sample. A measure that quantifies "confidence" in sample-to-cluster assignments is particularly relevant for studies in which the expected clusters are unlikely to show a clear separation, e.g. subgroups of patients with the same cancer type.

2.5 Cluster evaluation

Since clustering is an unsupervised method, external labels are not available to validate the identified clusters. For this reason, a number of methods have been developed to assess the quality of a clustering. One approach is to re-use the same information that was used for clustering, i.e. the distances or similarities between the samples, resulting in so-called internal evaluation methods (Section 2.5.1). External methods leverage additional information that presumably correlates with the outcome of interest, e.g., survival data (Section 2.5.2). Finally, enrichment methods can be applied to sets of features that are relevant for a cluster to obtain a biologically meaningful interpretation of the clustering (Section 2.5.3).

2.5.1 Internal cluster evaluation measures

Internal measures evaluate a given clustering on the basis of the data that were used for the clustering process itself. Pairwise sample similarities can be separated into intra-cluster and inter-cluster similarity, depending on the cluster assignments of the respective samples. A classical measure, which uses this information to quantify how dense and how separated the different clusters are, is the silhouette score (Section 2.5.1.1). Another important property of a result is its robustness, which describes how similar the clusterings identified on slightly modified data sets are. This property can be evaluated using a cross-validation technique (Section 2.5.1.2), as applied in Chapter 3. Even though it is not a traditional internal cluster evaluation approach, we discuss cross-validation in this section since no external information, such as class labels, is used for its application.

2.5.1.1 Silhouette value

Given a cluster assignment, the silhouette score S_i [121] indicates how similar a sample x_i is to all samples that are assigned to the same cluster compared to all samples in the neighboring cluster. The score is calculated by

$$S_i = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \qquad \forall i : S_i \in [-1, 1], \tag{2.27}$$

where $a(i)$ denotes the average dissimilarity of x_i to all other samples in the same cluster and $b(i)$ denotes the average dissimilarity of x_i to all samples in the neighboring cluster. Here, the neighboring cluster is defined as the one with the lowest dissimilarity to the considered sample x_i. In general, the higher the silhouette value, the higher is the quality of this cluster assignment:

$$S_i > 0 \implies a(i) < b(i) \quad \text{(good cluster assignment)} \tag{2.28}$$
$$S_i = 0 \implies a(i) = b(i) \tag{2.29}$$
$$S_i < 0 \implies a(i) > b(i) \quad \text{(bad cluster assignment)} \tag{2.30}$$

Calculating the average silhouette score over all samples

$$\overline{S_i} = \frac{1}{N} \sum_{i=1}^{N} S_i \tag{2.31}$$

quantifies how dense and how well separated the identified clusters are.

In Chapter 3 and 4, the silhouette score is used for identifying the number of clusters in the data set, i.e., we vary the number of clusters K for the clustering process and choose K such that the average silhouette score $\overline{S_i}$ is maximized.

2.5.1.2 Cross-validation

In supervised learning, overtraining or overfitting refers to a phenomenon where an algorithm fits very closely the data set on which it was trained, but lacks the ability to generalize on unseen data. Since no labels are used, clustering is not prone to overtraining. However, it is possible that a clustering result was obtained by chance and that slight variations in the training data or in the parameter setting lead to completely different results. Therefore, it can be reasonable to assess the stability of the obtained cluster assignment with respect to small changes in the clustering scenario. For this purpose, we apply the concept of K-fold cross-validation (CV), a method widely used in supervised learning to estimate the test error [62, Section 7.10], and combine

it with the Rand index, which measures similarity between two clusterings (see Section 2.5.2.1).

To mimic the existence of independent training and test sets, CV splits the complete data set into K disjoint parts of roughly the same size. Then, $K-1$ parts are used as training data while one part is left out and serves as independent test data to assess the performance of the trained model. This procedure is repeated K times such that each part was used once as test set. The cross-validation error can then be calculated based on all samples:

$$\mathrm{CV}(f) = \frac{1}{N} \sum_{i=1}^{N} L(y_i, f^{-k_i}(x_i)), \tag{2.32}$$

with L being the loss function, y_i the known label for sample x_i, and $f^{-k_i}(x_i)$ the prediction for sample x_i of the model trained without the partition k_i that contains sample x_i. This approach provides an estimation of the generalization error even when no separate test set is available. In the special case where K, the parameter determining into how many parts the data set is split, equals N (the total number of samples), the approach is also called leave-one-out cross-validation (LOOCV).

In order to evaluate the robustness of a clustering method, we apply LOOCV by leaving out one sample for the clustering procedure and assigning it afterwards to the cluster with the mean closest to the respective sample. Analogous to supervised learning, the test point was not involved in learning the model, i.e., learning the cluster centers, still, we obtain a cluster assignment for this sample. In this scenario, there are no labels y available to calculate the cross-validation error. Therefore, we also generate a clustering of the complete data set without leaving out any sample. This cluster assignment of the full data set is then compared to each LOOCV clustering result using the Rand index, a measure quantifying the similarity between two clustering assignments (see Section 2.5.2.1). These comparisons of the LOOCV result and the full clustering indicate how strong the results vary when they are based on slightly different training sets.

2.5.2 External cluster evaluation measures

Instead of evaluating the clustering based on the structure in the data, external measures use additional information to assess the quality of a cluster assignment. This approach is reasonable even if the external information used is not exactly the outcome of interest, as long as a correlation to the outcome of interest is expected. For instance, one would assume a correlation between reasonable cancer subtypes and the survival times of the patients.

Furthermore, external measures can be used to compare different clustering results, e.g., to show an overlap with previous results or the robustness of the clustering method.

2.5.2.1 Rand index

The Rand index [114] measures pairwise cluster similarity. The Rand index is an adaptation of the accuracy measure [99], which is a commonly employed model performance measure in supervised learning:

$$\text{acc} = \frac{TP + TN}{TP + TN + FP + FN} \tag{2.33}$$

with TP being the number of correctly predicted positive samples, TN the number of correctly predicted negative samples, FP the number of falsely predicted positive samples, and FN the number of falsely predicted negative samples. Since the Rand index has been developed for cluster evaluation, that is for unlabeled data, it formulates the notion of accuracy of the cluster assignment using a pairwise definition of the considered properties (TP, TN, FP, and FN) in the following way:

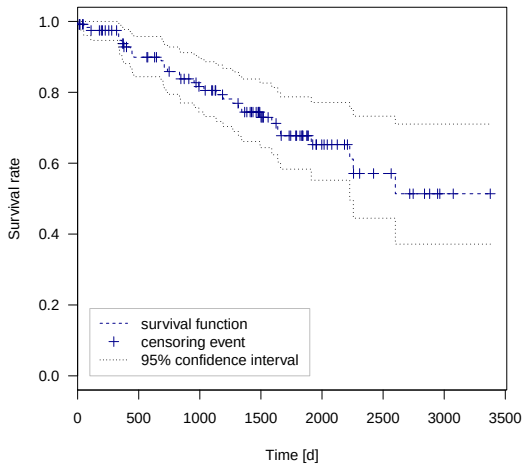
$$R = \frac{a + b}{a + b + c + d} \qquad R \in [0, 1] \tag{2.34}$$

where a, b, c, and d replace the previous entities to highlight the change of definition. Comparing two different cluster assignments C_1 and C_2, a is the number of sample pairs where both partners belong to the same cluster in both assignments C_1 and C_2 (representing TP), b is the number of sample pairs belonging to different clusters in both C_1 and C_2 (representing TN), c is the number of sample pairs belonging the same cluster in C_1 but to different clusters in C_2 (representing FP), and d is the number of sample pairs belonging to different clusters in C_1 and the same cluster in C_2 (representing FN).

The Rand index ranges from zero to one, where values close to one indicate strong similarity between the two clusterings. A Rand index of zero indicates that no pair of samples has the same relation in both cluster assignments, a Rand index of one signifies that the two clusterings are exactly identical.

2.5.2.2 Survival analysis

Medical studies analyzing patient cohorts are often focused on the occurrence of certain events in the course of patient treatment, e.g., the recurrence of a disease or a disease-related death. Such an analysis of patient cohort data



patient	time	status
1	16	0
2	1190	1
3	735	1
4	1493	0
5	1491	0
6	683	1
7	599	0
⋮	⋮	⋮

Figure 2.2 & Table 2.2: Example of a Kaplan-Meier graph depicting the survival of kidney cancer patients and, exemplarily, survival data for the first seven patients. Each patient without occurrence of the event (status=0) is censored at the given time point.

is called survival analysis. One important characteristic of data for survival analysis is the so-called (right) censoring. Censoring happens if a patient leaves the study at a time point before he had an event, or if a patient did not experience the event until the end of the study. An example would be a breast cancer study on recurrence-free survival of the patients: the data for each patient leaving the study for medically irrelevant reasons would be censored after that point in time, but would still provide useful information for any prior time point. An illustration of these data is given in Table 2.2, where the status reports for each patient if the event occurred and at which time the event or the censoring happened.

Kaplan-Meier graph The Kaplan-Meier estimator $\hat{S}(t)$ [77], or *product limit estimator*, estimates the conditional probability of survival at time t via

$$\hat{S}(t) = \prod_{i:t_i \leq t} \left(1 - \frac{d_i}{n_i}\right). \tag{2.35}$$

For each time point t_i, the number of deaths observed ($= d_i$) are compared to the number of patients at risk of an event ($= n_i$). The calculation accounts for censoring because a censored patient will not be counted in n_i after the time of the censoring. The estimated survival function can be displayed in a Kaplan-Meier graph as exemplified in Figure 2.2. The confidence intervals are estimated at each time point t based on the Greenwood formula for the

variance [47][2]. As the 0.95% confidence intervals illustrate, the estimate of the survival function is associated with uncertainties that increase with a decrease in the number of patients being considered at point t_i.

Log-rank test The Kaplan-Meier graph depicting estimated survival functions for two or more groups of patients can give a visual impression of possible differences between the sample groups in terms of survival time. Statistical tests evaluate how likely observed differences between two or more curves appear due to chance under the null hypothesis that the curves are equal. These tests differ from standard approaches, such as the one-way analysis of variance, in their ability to handle censored observations. For two groups of size n_1 and n_2, the Mantel-Cox test or *log-rank test* [68] compares the expected with the observed events (e.g., deaths) by computing

$$\mathrm{LR} = \frac{(O_2 - E_2)^2}{\mathrm{Var}(O_2 - E_2)}. \tag{2.36}$$

Here, $O_2 = \sum_{t=1}^{T} o_{2t}$, which is the number of observed events in Group 2 over all times t, and E_2 is the expected number of events for Group 2, calculated by

$$E_2 = \sum_{t=1}^{T} \mathrm{risk} \times n_{2t} = \sum_{t=1}^{T} \frac{o_t}{n_t} \times n_{2t}, \tag{2.37}$$

with n_{2t} being the number of patients belonging to Group 2 that are at risk at time t, i.e., the size of Group 2 at time t. o_t is the number of observed events at time t and n_t the sum of the two group sizes n_{1t} and n_{2t}. The risk of an event, used for the expected number of events, is calculated on the whole data set and not for each group separately. The variance in the denominator of the test statistic LR can be calculated by

$$\begin{aligned} \mathrm{Var}(O_2 - E_2) &= \sum_{t=1}^{T} \frac{n_{1t} n_{2t} (o_{1t} + o_{2t})(n_{1t} + n_{2t} - o_{1t} - o_{2t})}{(n_{1t} + n_{2t})^2 (n_{1t} + n_{2t} - 1)} \\ &= \sum_{t=1}^{T} \frac{n_{1t} n_{2t} o_t (n_t - o_t)}{n_t^2 (n_t - 1)}. \end{aligned} \tag{2.38}$$

To increase the number of groups tested from 2 to K, one needs to include covariances in addition to the variances resulting in

$$\mathrm{LR} = d^T V^{-1} d, \tag{2.39}$$

[2]The variance formula was originally published in Major Greenwood's report on "The natural duration of cancer"(1926). This publication is not available online, however, the cited publication provides the relevant paragraphs in Appendix B.

with d being a vector of length K with $d_i = O_i - E_i$ and V being the covariance matrix where $V_{ij} = \mathrm{Cov}(O_i - E_i, O_j - E_j)$ for $i, j \in \{1, ..., K\}$.

The log-rank test is based on the proportional hazards assumption, where the hazard is the slope of the survival curve. Proportional hazards require the ratio of the hazards to be constant over time with deviations only due to random sampling. In some cases, a violation of this assumption can be detected visually, for example, when two survival curves cross. For sufficiently many events and under the assumptions of proportional hazards and group-independent censoring, the resulting test statistic is approximately χ^2-distributed with $K-1$ degrees of freedom, where K is the number of groups. The degrees of freedom implicitly correct for the number of groups compared, such that multiple testing correction is not necessary when applying this test to data sets with more than two groups. The resulting p-value indicates how likely it is to observe at least such extreme differences simply due to chance if, in fact, all curves follow the same survival function.

2.5.3 Enrichment analysis

Enrichment analysis tests whether certain biological functions (or entities of a different category of interest) are over-represented in a specified set of genes compared to a background set of genes. For Chapter 3 and 5, we used the categories of Gene Ontology (GO). GO provides a standardized set of terms for describing gene products with respect to their molecular function, their participation in biological processes, and their localization in certain cellular components [10]. Using these GO terms, we applied over-representation analysis (ORA), which operates on unsorted gene lists and is implemented using the hypergeometric test [42].

Assuming, we are given N genes of which M belong to a specific GO term C. In the following, this set will be called background or reference set. When choosing K out of N genes at random, the expected number of genes belonging to C can be calculated as

$$k' = \frac{M * K}{N}. \tag{2.40}$$

With K genes being randomly chosen, the probability of having x genes of category C can be calculated using the hypergeometric distribution

$$P(X = x|N, M, K) \;=\; \frac{\binom{M}{x}\binom{N-M}{K-x}}{\binom{N}{K}}. \tag{2.41}$$

The hypergeometric test with the null hypothesis that genes labeled with C and not labeled with C are chosen from the background with the same

probability can be formulated as

$$p = \begin{cases} \sum_{i=x}^{\min(K,M)} P(X = i|N, M, K), & \text{if } k' < x \\ \sum_{i=\max(K+M-N,0)}^{x} P(X = i|N, M, K), & \text{else,} \end{cases} \tag{2.42}$$

where x is the number of genes drawn from category C and k' the expected number of chosen genes belonging to C as defined in Equation (2.40). In summary, ORA checks if the chosen subset contains more or less genes from category C than would be expected based on the complete set of N genes. The p-value is then the probability of observing an outcome at least as extreme as in the given sample if the null hypothesis were true.

In this thesis, ORA was used to gain insights into the biological characteristics of the identified patient clusters (Chapter 3 and 5). Meaningful subsets of genes for the cluster were identified for instance according to differential methylation or differential expression in the respective cluster. To identify these genes, statistical hypothesis testing determines whether the null hypothesis (i.e., the distribution of the expression of a particular gene a in the cluster is the same as in the remaining data set) can be rejected on the basis of the available data. Gene sets that have been identified in this way can be tested for connections to specific functions using enrichment analysis.

2.5.3.1 Multiple testing correction

A p-value is considered significant, if it is smaller than a chosen significance threshold α, i.e., if $p < \alpha$. When performing a large number of hypothesis tests, the probability of obtaining a significant result purely by chance increases even if the null hypothesis is actually true, i.e., if there is no detectable effect in the data, and all variation is simply due to noise. This scenario occurs, for instance, when testing a gene set for enrichment with a large number of different terms (e.g., as provided by GO). In these cases, one needs to correct for the number of tests executed to avoid random, significant results. In this thesis, we used two different methods for this purpose: Bonferroni correction and Benjamini-Hochberg correction.

Bonferroni correction The Bonferroni correction [159] controls the family-wise error rate, which is the probability of rejecting the null hypothesis if it were actually true. The Bonferroni correction method tries to avoid such false-positive discoveries by multiplying each p-value p with the number of tests, i.e., $p_{\text{adj}} = mp$ with m being the number of tests performed. Equivalently, one can also adjust the significance threshold α such that the null hypothesis is rejected if $p < \frac{1}{m}\alpha$. The Bonferroni method is one of the most conservative correction methods.

Benjamini-Hochberg correction When using the Benjamini-Hochberg method [17], the significance threshold is defined as the largest p-value p_i for which $p_i < \frac{i}{m}\alpha$ holds (where i is the rank of the p-value when all obtained p-values are sorted increasingly). With this adjustment, the false discovery rate, i.e., the expected number of false rejections of the null hypothesis, remains thus at most α. When we report p-values that are corrected via Benjamini-Hochberg p_i^{adj}, they are calculated by

$$p_i^{\text{adj}} \;=\; \min\{p_i \frac{m}{i},\; p_{i+1}^{\text{adj}}\} \tag{2.43}$$

with i being the rank of the p-value and m the number of tests performed.

2.6 Related work

This section provides an overview of previously proposed cancer subtypes and methodical approaches that integrate multidimensional cancer patient data in an unsupervised manner. We include integrative clustering methods but also integrative dimensionality reduction methods, because the latter can provide a good basis for applying simple clustering algorithms, e.g., k-means, resulting in sample groups that are influenced by all input data. Additionally, the representation of data in an integrative, reduced-dimensional space can facilitate other valuable applications, such as visualizations, which are not discussed in this work.

Single-omics cancer subtypes

As mentioned in Section 2.1.2, molecular subtypes have been identified for a few cancer types on the basis of different, individual data types. Four breast cancer subtypes, which were determined based on hierarchical clustering of gene expression data [110], are currently considered in the treatment guidelines. Subtypes for glioblastoma have been identified via consensus hierarchical clustering of gene expression data [154] and via consensus k-means clustering of DNA methylation data [104]. One of the methylation subtypes appears as a subgroup of one of the previously identified gene expression subtypes [104]. Subtyping efforts based on single-omics have been made for other cancer types, including head and neck squamous cell carcinoma [34], and lung adenocarcinoma [50].

Multi-view approaches

Since the combination of several data sources could provide more comprehensive views on the patients, a number of unsupervised multi-view methods

have been developed over the last years. Some approaches aim specifically at integrating biological data and make use of known relationships between the data types. Besides, general approaches for the application in diverse scenarios, e.g., the combination of different views in computer vision scenarios [174], have been proposed.

In general, multi-omics or multi-view methods can be classified based on the time point when the data integration is performed, i.e., one distinguishes between *early*, *intermediate* and *late integration* methods [108].

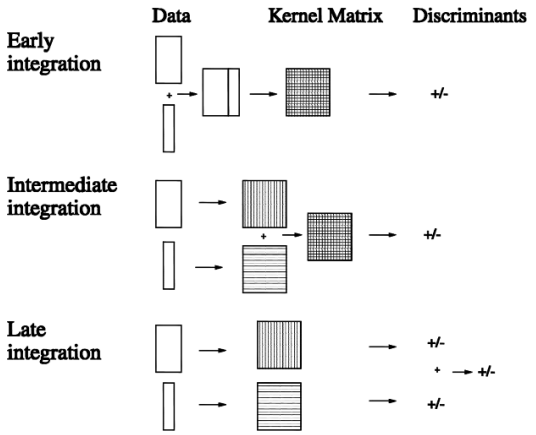


Figure 2.3: **Illustration of early, intermediate, and late integration** on the example of a kernel-based classifier. Reproduction from Pavlidis et al. [108] (license #4462490879804, see Table B.1).

As illustrated in Figure 2.3, *early integration* corresponds to a simple concatenation of the available data and subsequent execution of a single-omic approach. LRAcluster [164], a representative of early integration, models each feature as a random variable with a hidden parameter before decomposing the concatenated parameter matrix to identify candidate sample groups. For a pan-cancer data set, the authors integrate somatic mutations, copy number variations, DNA methylation, and gene expression data. While this method is able to handle conceptually different data types (e.g., real-valued and binary data), it ignores differences in dimensionality between the data types. Another example for early integration is MEREDITH [142]: by reducing the number of features for each data type to 50 via principal component analysis before the concatenation, the authors circumvent the problem of differences in dimensionality. Based on the concatenation of the remaining features, the samples are subsequently projected into a two-dimensional space using t-distributed stochastic neighborhood embedding and clustered using DBSCAN. The authors applied MEREDITH to a pan-cancer data set and could identify known, tissue-specific clusters, as well as new clusters comprising tumors from different tissues. However, this method does not account for differences regarding the statistical properties of the data types, such as their distribution.

Late integration is a frequently applied method for combining biological data, which consists of separate sample clustering for each data type and subsequent integration of the different cluster assignments. The latter step

can be performed either manually or automatically, e.g., using consensus clustering [102]. Consensus clustering combines several clusterings into a consensus matrix indicating for each pair of samples, in how many of the clusterings they were assigned to the same cluster. Based on this consensus matrix, a consensus clustering is derived. This approach was used, for instance, for the identification of comprehensive breast cancer subtypes on TCGA data [146]. Although both manual and automatic integration might maintain strong signals in the data, late integration approaches cannot capture weak but concordant structures in different data types because these signals already vanish during the initial clustering. Additionally, manual integration tends to be biased, leading to inconsistent results. For these reasons, other approaches bring forward the step of data integration leading to *intermediate integration*.

The following paragraphs present different intermediate integration techniques categorized according to their algorithmic approach.

Co-regularization To respect the structure of the different data types at the time of data integration, co-regularization extends the objective function of a specific algorithm such that the clustering of each single source is regularized by all other sources. Cai et al. [26] apply co-regularization to k-means, i.e., the samples are clustered based on each data source separately, however, the objective function of k-means is constrained by the clusterings derived from the other data sources. In this way, the approach, which is called multi-view k-means clustering, converges to a common clustering result. Cai et al. [26] additionally use the $l_{2,1}$-norm in the objective function instead of the l_2-norm, which is usual for k-means clustering, to render the approach robust to outliers.

Kumar et al. [81] proposed two versions of multi-view spectral clustering. The co-regularization term in the objective function either penalizes pairwise differences between the kernel matrices or the difference of each kernel matrix to a common centroid. Here, kernel matrices measure the similarity between two eigenvector embeddings based on the different data that are used for the clustering. Both approaches are solved iteratively with each kernel matrix being updated separately. Moreover, both formulations contain a hyperparameter weighting the influence of each data type. To make a reasonable choice for the parameter, the user would need prior knowledge on the importance of the different data types.

Matrix factorization A classic approach for analyzing two sets of variables, or two data types, is Canonical Correlation Analysis (CCA) [69]. As the name suggests, this method identifies the directions in the data in which

the correlation between the two data matrices is maximized. In the initial formulation, the number of samples is required to be larger than the number of features, which makes an application to high-dimensional datasets prohibitive. Several approaches have been proposed that make CCA applicable for these high-dimensional data and enable analyzing correlations between more than two different data types, e.g., using sparse multiple CCA [160], regularized generalized CCA [143], or sparse generalized CCA [144].

Similar to CCA, intermediate integration methods are often based on the idea of identifying shared or correlated information in different data sources. Consequently, numerous methods were developed that decompose the input matrices into common structure and data type-specific structure. Shen et al. [129] introduced iCluster for the integration of continuous data, which is based on a Gaussian latent variable model. The latent variables are shared between the data types and interpreted as the driving factors for the tumor in the example of cancer patient data. Therefore, these latent variables can be used to project the data into a low-dimensional, integrative subspace, as well as for sample clustering. Whereas the original version of iCluster uses l_1-penalties to perform feature selection, extensions include variance weighted penalty terms [130] and specific penalty terms for different data types [131]. Furthermore, the extension iClusterPlus enables the integration of count, binary and categorical data in addition to continuous data [100]. Finally, iClusterBayes employs a Bayesian latent variable model, which additionally returns a posterior probability for each feature [101]. Despite improvements in the running time, this series of methods remains computationally expensive and requires a feature preselection. Similar methods use (group) factor analysis to decompose the data matrices, where factors represent the shared information between the data sources. Multi-omics factor analysis [9] implements a factor- and source-specific regularization with the level of regularization learned automatically during model training. This procedure improves the interpretability of the results by inducing sparsity of the used features. In contrast to iCluster and its extensions, multi-omics factor analysis can include samples in the analysis for which some measurements are missing without requiring explicit data imputation. Other Bayesian approaches include the use of mixture models and extensions thereof [13, 78].

Data integration has also been realized based on non-negative matrix factorization (NMF), which learns a decomposition of a matrix $X \approx VH$ subject to the constraint that the constituent matrices need to be non-negative [86]. NMF is often applied for clustering as the matrix H can be used for cluster identification. Extensions to multiple input data types include collective NMF, where a common matrix H is enforced for all data types [133]. Besides, Liu et al. [93] proposed multi-view NMF, which learns the two components

V and H specifically for each data type but iteratively regularizes H towards a common consensus. The methods described above decompose the matrices such that reconstruction is possible via multiplication of the identified components. Lock et al. [94] introduced a matrix factorization method called Joint and Individual Variation Explained (JIVE), which affords a reconstruction of the original matrices via a summation of the three learned entities. These components represent the joint variation in the data, individual variation per data type, and residual noise. Joint and individual variations are optimized in an alternating manner by minimizing the squared residual noise. In general, matrix factorization approaches might require some additional processing steps to handle non-numeric data types.

Deep learning In recent years, deep learning has proven to be a promising field of study in machine learning [123]. These methods, which are based on neural networks, have initially been utilized mainly in the context of image data. However, in more recent work, deep learning methods have successfully been applied to biological problems, e.g., to discover specific patterns in genomic sequences [89]. Chaudhary et al. [32] proposed a deep learning-based early integration scheme for cancer subtype discovery. Their approach uses variational autoencoders for joint dimensionality reduction resulting in 100 transformed features that contain all input data types. The subsequent patient clustering is, however, not entirely unsupervised as the features are preselected based on their correlation to the survival time. In a similar approach, Ronen et al. [120] use stacked variational autoencoders and subsequent clustering to identify integrative subgroups of colorectal cancer. The identified subgroups differ significantly in their survival rates and represent a refinement of the state-of-the-art subtypes, which are defined on gene expression. However, model selection was performed based on a score that considers both the survival times and the state-of-the-art subtypes.

Multi-modal deep neural networks have been proposed, which correspond to the intermediate integration scheme. The main idea of multi-modal deep neural networks is to stack different learning machines, such that the network performing data integration is fed by an input layer of one learning machine per data source (e.g., Deep Boltzmann Machines [139]). Liang et al. [91] use a multi-modal Deep Believe Network, a hierarchical model of several restricted Boltzmann machines (see [66] for details) for clustering cancer patients based on genomic data. Multi-modal deep neural networks benefit from their architecture, which enables pretraining for each data type separately, while still being able to identify global effects due to the integrative layers. Moreover, they can also handle samples with missing values without the need for imput-

ing missing data before model training. While a noteworthy characteristic of deep learning methods is their ability to learn complex structures, this flexibility comes at the cost of potential overtraining and thus poor generalization. This issue appears particularly in biological applications, where the number of available samples is often very limited. Therefore, reasonable regularization schemes have to be employed.

Similarity-based integration The methods discussed so far are concerned with direct data integration. The following section introduces methods that perform the integration on the basis of similarities or kernel matrices derived from the individual data types. These methods have the advantage that they are not restricted to specific types of data, e.g., numeric or binary data. Furthermore, similarity matrices or networks are an efficient way to represent sample information, especially in high-dimensional settings when the number of features is higher than the number of samples.

For these reasons, similarity network fusion (SNF) uses pairwise sample similarities from each data type [158]. The similarities are generated by a scaled exponential similarity kernel, which is similar to a radial basis kernel with an additional normalization for local density structures. Sample similarity networks derived from these kernels are then fused iteratively by using a message-passing algorithm until convergence to a common, integrated network. In this way, SNF implements a nonlinear integration of the similarities. Finally, the patient or sample clusters are identified using spectral clustering [157] on the integrated network.

Another possibility for similarity-based data integration is multiple kernel learning, where one weight is optimized for each input kernel matrix leading to a combined ensemble kernel matrix (see Section 2.2.2). In the supervised setting, various methods of optimizing the ensemble kernel have been proposed, including kernel-target alignment [39] and idealized kernels [82]. All of these aim at finding the best kernel for a given supervised learning task. However, without labels for the samples, different strategies and objective functions need to be adopted with the simplest approach being a fixed kernel combination. Here, fixed refers to the fact that no data-dependent parameter needs to be learned because the ensemble kernel is generated as the unweighted sum or product of the input kernels. NEMO, a recently proposed data integration method, uses the average kernel, a commonly chosen fixed kernel combination [116]. Similar to SNF, the method uses radial basis kernel functions with a normalization for the density in the respective neighborhoods. The strength of NEMO lies in its simplicity, which results in robust solutions and low computational complexity. Additionally, NEMO

can be used to integrate partial data sets as long as each pair of samples has measurements for at least one common data type. However, noisy kernel matrices without a clear structure cannot be automatically excluded or downweighted when using a fixed and data-independent kernel combination.

Therefore, other approaches for integrating several kernel matrices optimize specific kernel weights by extending existing dimension reduction or clustering approaches. These methods often employ an iterative procedure that alternates between the optimization of the kernel weights and the optimization of the cluster assignment or projection matrix (for clustering or dimensionality reduction approaches, respectively). Yu et al. [170] developed a multiple kernel version of k-means clustering. The optimization of the kernel weights and cluster memberships is a non-convex problem, potentially having local optima. This is tackled by an alternating minimization procedure. The multiple kernel k-means uses an additional parameter δ that controls the sparsity of the kernel weights by imposing the weight vector to have an l_δ-norm of one. Further extensions aim at increasing the robustness of multiple kernel k-means [43] or implement a fuzzy cluster assignment [70]. In addition to the previously discussed co-regularization extension of spectral clustering [81], Huang et al. [71] propose affinity aggregation for spectral clustering. Both methods extend spectral clustering, however, they differ in the data integration approach: whereas Kumar et al. [81] apply co-regularization to generate a common result, Huang et al. [71] directly fuse the kernel matrices using learned kernel weights.

The multiple kernel learning methods discussed above optimize a linear combination of kernel matrices using one weight per matrix. Gönen and Margolin [55] moved one step further by introducing localized multiple kernel k-means, which uses sample-specific weights instead of optimizing one weight per kernel. This leads to a nonlinear integration of kernel matrices, which provides more flexibility to account for sample-specific characteristics or noise in some measurements. However, the additional flexibility requires optimizing a notably larger set of parameters and can therefore lead to instabilities in the obtained clustering results and optimized weights.

Contributions

This section summarizes the contributions in this thesis in the light of the given related work.

1. Multiple kernel learning aims at optimizing kernel- or sample-specific weights, which provides increased flexibility in comparison to fixed kernel combinations. However, this technique bears the risk that the results depend strongly on the used data set and might thus not be robust

to outliers or small changes in the data set. We demonstrate that regularizing the kernel weights in unsupervised multiple kernel learning increases the robustness of the final results (Chapter 3).

2. Many dimensionality reduction schemes have been extended such that they can handle multiple input data types. We prove that for kernel PCA, one of the most widely used dimensionality reduction methods, a sensible extension into a multiple kernel setting is not possible. Moreover, we provide an alternative formulation, which combines the basic concept of PCA with the aim of data integration (Chapter 4).

3. Despite the large number of different approaches that optimize kernel- or sample-specific weights, interpretation remains a difficult issue for multiple kernel learning as well as for most of the similarity based integration methods. The biological interpretation of the results is commonly done retrospectively, e.g., by filtering for genes that are differentially expressed between the patient groups and associating these genes with common biological functions. However, it remains unclear how the learning machine has come to the final result. Therefore, we present a general extension for kernel learning methods that yields better interpretability of the obtained results (Chapter 5).

Chapter 3

Regularization of unsupervised multiple kernel learning

This chapter presents the extension and application of current multiple kernel learning approaches in the context of dimensionality reduction. We provide evidence that the graph embedding framework that incorporates several input kernel matrices gains robustness by using an additional regularization constraint. Furthermore, we show that this regularized approach can also be used with a larger number of input kernels and thereby enables implicit kernel parameter selection.

The content of this chapter was published in Speicher and Pfeifer [136] in the proceedings of the conference *Intelligent Systems for Molecular Biology* (ISMB 2015).

3.1 Overview

For the identification of cancer subtypes, we propose to apply nonlinear, kernel-based dimensionality reduction with subsequent patient clustering. To this end, we adopt the multiple kernel learning for dimensionality reduction framework (MKL-DR; see Section 2.3.3.2) that enables dimensionality reduction and data integration at the same time. In order to avoid overfitting, especially in scenarios with many distinct input matrices, we extend the MKL-DR approach by adding a regularizing constraint resulting in rMKL-DR. The samples, in our case cancer patients, are projected into a low-dimensional, integrated space where they can be further analyzed. We show that this representation captures meaningful information, which we use for clustering the samples.

The outlined procedure offers several advantages: multiple kernel learning

provides high flexibility concerning the input data type, which enables the combination of qualitatively different data, such as sequences or numerical matrices. Moreover, in case one does not have enough information to choose the best kernel function for a data type or the best parameter (combination) for a given kernel beforehand, it is possible to input several kernel matrices per data type, based on different kernel functions or parameter settings. The multiple kernel learning approach automatically upweights the matrices providing more information with respect to the objective function while downweighting those with less information. Moreover, the framework provides high flexibility concerning the choice of the dimensionality reduction method, which does not need to be unsupervised, but also various supervised and semi-supervised methods can be adopted. Finally, by capturing the nonlinearity in the dimensionality reduction step, we can apply afterwards a simple clustering algorithm, such as k-means, to identify the patient subgroups.

The evaluation of our method on five different cancer sets shows that the results gain robustness due to the regularization of the kernel weights (Section 3.3.1). The identified patient clusters reflect characteristics from distinct input data types and show differences concerning their response to treatment with a standard chemotherapy drug. Furthermore, we observe that kernel matrices with less information have less influence on the final result. A comparison of the survival differences between our clusters and a state-of-the-art method shows that our method yields comparable results while using a simpler, and therefore potentially more robust, data integration process.

Related work A general overview on multi-omics and multi-view data integration approaches is provided in Section 2.6. Here, we focus on approaches that are similar to the proposed rMKL-DR in the sense that they build on the graph embedding framework (see Section 2.3.3) and intend to increase its stability. Jiang and Chung [76] developed MKL-TR, an approach that optimizes multiple kernel graph embedding using the trace ratio optimization problem. This problem maximizes the ratio of inter-class (or inter-cluster) variation to intra-class (or intra-cluster) variation. Here, the definition of *class* or *cluster* is given by the user, which makes MKL-TR a framework that supports the implementation of different dimension reduction techniques. In this approach, the authors tackle the problem of overfitting, and thereby reduced robustness, by adding a regularization term to the denominator of the objective function, i.e., the trace ratio maximization. Later, Li et al. [90] further extended the formulation by adding a regularization term to the numerator

of the trace ratio maximization to improve the performance with sparse input data in high dimensions. Going into a different direction, an adaptive extension of MKL-DR was proposed by Thiagarajan et al. [148]. After the complete optimization of the low-dimensional representation via MKL-DR, their algorithm re-calculates the affinity matrices on the basis of the learned projection, thereby updating the objective function. The iterative execution of these two steps aims at reducing the sensitivity of the affinities to perturbations in the data. To our knowledge, none of these methods have been applied to cancer or molecular data to prove their usefulness in this specific setting. Moreover, for none of the three approaches an implementation is publicly available.

Therefore, and because of its popularity in the bioinformatics community, we used similarity network fusion (see Section 2.6: Similarity-based integration) for comparison to our method. Similarity network fusion first determines similarities between the samples using a kernel function that is related to the RBF kernel we are using. In contrast to our method, the subsequent combination of the different sources is performed in a nonlinear manner whereas our approach uses a linear combination that learns one weight per data type.

3.2 Methods

In order to integrate several data types, we extend the MKL-DR approach (multiple kernel learning for dimensionality reduction, see Section 2.3.3.2). This method is based, on the one hand, on multiple kernel learning, and, on the other hand, on the graph embedding framework for dimensionality reduction. We add a constraint that leads to the regularization of the vector controlling the kernel combinations. We call this method rMKL-DR (regularized multiple kernel learning for dimensionality reduction) in the following discussion.

3.2.1 Regularization in the graph embedding framework

As described in Section 2.3.3.2, the multiple kernel graph embedding framework combines the optimization of an ensemble kernel matrix $\mathbb{K}$ with the objective of dimensionality reduction schemes. The ensemble kernel matrix $\mathbb{K}$ is defined as a weighted linear combination of the available input kernel

matrices $\{K_1, ..., K_M\}$

$$\mathbb{K} = \sum_{m=1}^{M} \beta_m K_m, \tag{3.1}$$

with β_m being the weight of the kernel matrix K_m. The graph embedding framework optimizes the projection matrix $A \in \mathbb{R}^{N \times p}$ that leads to the reduction of the dimensionality of the data, and the vector β that weights the input kernels at the same time. Having learned these two parameters A and β, we obtain the low-dimensional (here p-dimensional) representation of the samples in the data set $X \in \mathbb{R}^{N \times d}$ via

$$\text{proj}(X) = A^T \sum_{m=1}^{M} \beta_m K_m \quad \in \mathbb{R}^{N \times p}, \tag{3.2}$$

where M is the number of integrated kernel matrices considered. However, when M is large, the risk of overfitting during parameter training, a phenomenon leading to high variance in the results for slightly varying input data, increases. To avoid this behavior, we add the constraint $\|\beta\|_1 = 1$ to the original optimization problem (Formula 2.22), which restricts the search space for the kernel weights. Had we added the constraint $\|\beta\|_1 \leq 1$, this would amount to an l_1-regularization, which is regularly used in supervised learning approaches (e.g. the lasso [150]). However, in our case, the equality in the constraint ensures variance preservation in the ensemble kernel matrix and thus avoids solutions with very small weights for all kernel matrices. For N samples described by M different kernel matrices $\{K_1, ..., K_M\}$, the full optimization problem for rMKL-DR is given by:

$$\arg\min_{A,\beta} \quad \sum_{i,j=1}^{N} \|A^T \mathbb{K}^{(i)} \beta - A^T \mathbb{K}^{(j)} \beta\|^2 w_{ij} \tag{3.3}$$

$$\text{subject to} \quad \sum_{i=1}^{N} \|A^T \mathbb{K}^{(i)} \beta\|^2 d_{ii} = C \tag{3.4}$$

$$\|\beta\|_1 = 1 \tag{3.5}$$

$$\beta_m \geq 0, \quad m = 1, 2, ..., M \tag{3.6}$$

with C being a positive constant, β and $\mathbb{K}^{(i)}$ defined as in Section 2.3.3.2:

$$\beta = [\beta_1 \cdots \beta_M]^T \in \mathbb{R}^M, \tag{3.7}$$

$$\mathbb{K}^{(i)} = \begin{pmatrix} K_1(1,i) & \cdots & K_M(1,i) \\ \vdots & \ddots & \vdots \\ K_1(N,i) & \cdots & K_M(N,i) \end{pmatrix} \in \mathbb{R}^{N \times M}. \tag{3.8}$$

The projection matrix $A \in \mathbb{R}^{N \times p}$ consists of p vectors $\{\alpha_1, ..., \alpha_p\}$ to project the samples into p dimensions. The weight matrix W and the diagonal matrix D determine which dimensionality reduction scheme is implemented.

3.2.1.1 Iterative optimization

Since the simultaneous optimization of the two variables of interest is difficult, coordinate descent is employed as suggested for the MKL-DR framework [92]. In this technique, A and β are iteratively optimized in an alternating manner until convergence or until a maximum number of iterations is reached. One can start either with the optimization of A, then β is initialized to equal weights for all kernel matrices summing up to one (such that $\beta_i = 1/M,\ \forall i \in \{1, ..., M\}$), or with the optimization of β, then AA^T is initialized to the identity matrix I.

Optimizing A For the optimization of the projection matrix A, β is fixed. As for a vector u holds $\|u\|^2 = \text{trace}(uu^T)$, the problem can be reformulated as follows:

$$\arg\min_A \ \text{trace}(A^T S_W^\beta A) \tag{3.9}$$

$$\text{subject to} \ \ \text{trace}(A^T S_D^\beta A) = C \tag{3.10}$$

with

$$S_W^\beta = \sum_{i,j=1}^{N} w_{ij} \, (\mathbb{K}^{(i)} - \mathbb{K}^{(j)}) \beta\beta^T (\mathbb{K}^{(i)} - \mathbb{K}^{(j)})^T, \tag{3.11}$$

$$\text{and} \quad S_D^\beta = \sum_{i=1}^{N} d_{ii} \, \mathbb{K}^{(i)} \beta\beta^T (\mathbb{K}^{(i)})^T. \tag{3.12}$$

The additional Constraints (3.5) and (3.6) only concern β and are thus irrelevant for the optimization of A. The formulated problem corresponds to a *trace ratio problem*, i.e.

$$\arg\min_A \ \frac{\text{trace}(A^T S_W^\beta A)}{\text{trace}(A^T S_D^\beta A)}, \tag{3.13}$$

for which no closed-form solution exists. However, the trace ratio problem can be relaxed into a ratio trace problem, that is

$$\arg\min_A \ \ \text{trace}[(A^T S_D^\beta A)^{-1} (A^T S_W^\beta A)], \tag{3.14}$$

which can be efficiently solved using the generalized eigenvalue decomposition

$$S_W^\beta \alpha = \lambda S_D^\beta \alpha. \tag{3.15}$$

The projection vectors of A are the eigenvectors $\{\alpha_1, ..., \alpha_p\}$ corresponding to the p smallest generalized eigenvalues λ.

In case of reducing the dimensionality to one, the solution of the ratio trace problem is equal to the solution of the trace ratio problem. For higher dimensions, solving the ratio trace problem can be seen as a greedy approach to the trace ratio problem.

Optimizing $\boldsymbol{\beta}$ For the optimization of the kernel weight vector β, the projection matrix A is fixed. By using the fact that $\|u\|^2 = u^T u$ (for a given vector u), the problem becomes

$$\arg\min_{\beta} \quad \beta^T S_W^A \beta \tag{3.16}$$

$$\text{subject to} \quad \beta^T S_D^A \beta = C \tag{3.17}$$

$$\|\beta\|_1 = 1 \tag{3.18}$$

$$\beta_m \geq 0, m = 1, 2, ..., M. \tag{3.19}$$

with

$$S_W^A = \sum_{i,j=1}^{N} w_{ij} \, (\mathbb{K}^{(i)} - \mathbb{K}^{(j)}) A A^T (\mathbb{K}^{(i)} - \mathbb{K}^{(j)})^T, \tag{3.20}$$

$$\text{and} \quad S_D^A = \sum_{i=1}^{N} d_{ii} \, \mathbb{K}^{(i)} A A^T (\mathbb{K}^{(i)})^T. \tag{3.21}$$

Due to the additional constraints on β, this non-convex problem cannot be solved using a generalized eigenvalue decomposition as was possible for the optimization of A. However, this quadratically constrained quadratic programming problem can still be solved efficiently using a semidefinite programming relaxation [92].

Complexity of the approach The runtime of the algorithm can be decomposed into the dimensionality reduction step and the k-means clustering. The dimensionality reduction is performed by iteratively updating the projection matrix A and the kernel weight vector β (in this implementation, at most 20 iterations were performed with a convergence threshold of 1E-05). The optimization of β uses semidefinite programming where the number of

constraints is linear in the number of input kernel matrices and the number of variables is quadratic in the number of input kernel matrices. However, if $M \ll N$, the dominating term is the optimization of A. This optimization involves solving a generalized eigenvalue problem having a complexity of $\mathcal{O}(N^3)$, where N is the number of samples in the data set.

3.2.2 Leave-one-out cross-validation for rMKL-DR

In order to assess the stability of the resulting clusterings, we applied a leave-one-out cross-validation approach (see Section 2.5.1.2). After learning the projection matrix A, the kernel weights β, and the cluster assignment on the reduced data set (without the ith patient), the projection of the left-out sample x_i can be calculated as $\text{proj}(x_i) = \text{A}^T\mathbb{K}^i\beta \in \mathbb{R}^p$. The leave-one-out clustering is obtained by assigning patient x_i to the cluster, which has the mean that is closest to $\text{proj}(x_i)$ in the dimensionality-reduced space. This cluster assignment is performed in concordance with the idea of k-means. Finally, we compare this leave-one-out clustering to the clustering of the full data set using the Rand index, which measures the similarity of two clusterings (see Section 2.5.2.1). It should be noted that the leave-one-out cross-validation can only be applied because the used dimensionality reduction procedure does not suffer from the out-of-sample problem, i.e., new samples can be projected into the new space using A (see Section 2.3).

3.2.3 Materials

We used data from five different cancer types from The Cancer Genome Atlas (TCGA) [1] that were preprocessed by Wang et al. [158][1]. The cancer types comprise breast invasive carcinoma (BIC), colon adenocarcinoma (COAD), glioblastoma multiforme (GBM), kidney renal clear cell carcinoma (KRCCC), and lung squamous cell carcinoma (LSCC). For each cancer type, we clustered the patients based on gene expression, DNA methylation, and miRNA expression data (see Section 2.1.2 for an introduction on the biological relevance of these data types in the context of cancer). Additionally, we leveraged patient survival data for the subsequent evaluation. In the preprocessing, all patients with more than 20% missing values over all features were removed. Similarly, all features with more than 20% missing values over all patients were removed. Remaining missing values were imputed using k-nearest neighbor imputation. Finally, the data were normalized by subtracting the mean and dividing by the standard deviation. Table 3.1

[1] downloaded from `http://compbio.cs.toronto.edu/SNF/SNF/Software.html`

Table 3.1: Overview over the number of features per data type in each cancer type.

Cancer type	Gene expression	DNA methylation	miRNA expression
BIC	17 814	23 094	354
COAD	17 814	23 088	312
GBM	12 042	1 305	534
KRCCC	17 899	24 960	329
LSCC	12 042	23 074	352

summarizes the numbers of features for each data type. The numbers display a large difference between gene expression and DNA methylation with tens of thousands of features on the one hand, and miRNA expression with only a few hundreds of features on the other hand. The provided clinical data contained the overall survival of the patients. For most cancer types, this was measured by the number of days to the last follow-up. For COAD, a combination of the number of days to last known alive and the number of days to the last follow-up was provided because of many missing values in the latter attribute. Furthermore, the vital status of the patients (i.e., alive or deceased) was used in the survival analysis. Table 3.2 summarizes the number of samples analyzed, as well as the survival data in terms of the number of events, which refers here to a cancer-related deaths, for each cancer type. Each patient without event represents a censored data point.

Table 3.2: Overview over the number of samples N per cancer type and the respective number of events (i.e., cancer-related deaths) that are used for the survival analysis.

	Cancer type	N	W/ event	W/o event
BIC	Breast invasive carcinoma	105	18	87
COAD	Colon adenocarcinoma	92	9	83
GBM	Glioblastoma multiforme	213	197	16
KRCCC	Kidney renal clear cell carcinoma	122	33	89
LSCC	Lung squamous cell carcinoma	106	66	40

Implementation

The data integration approach rMKL-DR was implemented in Matlab version R2016b [147]. For the optimization of the kernel weights β the modeling and optimization toolbox *YALMIP* [96] was used. The evaluation including leave-one-out cross-validation, survival analysis, and comparison to established subtypes was performed using custom scripts in R version 3.1.3 [112]. The enrichment analysis was performed using GeneTrail2 [140] via a custom Python script.

3.3 Regularized multiple kernel locality preserving projections

Using the presented framework, we applied the unsupervised local dimensionality reduction algorithm Locality Preserving Projections (LPP, see Section 2.3.1 and Table 2.1), which aims to preserve the distances of each sample to its local neighborhood. For LPP, there are different possibilities of defining the matrices W and D controlling the dimensionality reduction in the graph embedding framework. We defined the neighborhood graph of each data point x_i by its nearest neighbors. This neighborhood is denoted as $\mathcal{N}_k(x_i)$, with $k_{\mathcal{N}}$ being a parameter controlling the size of the neighborhood[2]. Moreover, we chose uniform weights between neighboring samples, resulting in the following definitions for W and D:

$$w_{ij} = \begin{cases} 1, & \text{if } x_i \in \mathcal{N}_k(x_j) \vee x_j \in \mathcal{N}_k(x_i) \\ 0, & \text{else} \end{cases} \tag{3.22}$$

$$d_{ij} = \begin{cases} \sum_{n=1}^{N} w_{in}, & \text{if } i = j \\ 0, & \text{else.} \end{cases} \tag{3.23}$$

The rMKL-DR approach implementing LPP will be called rMKL-LPP from now on.

Workflow We applied rMKL-LPP to each of the five cancer data sets separately. For each available data type, we used the Gaussian radial basis kernel function to calculate the kernel matrices, then centered and normalized them in the RKHS (see Section 2.2.2). In order to investigate how well the method is able to handle multiple input kernels for single data types, we generated two different scenarios:

[2]The number of nearest neighbors is denoted by $k_{\mathcal{N}}$ to avoid confusion with the number of clusters for k-means.

- **Scenario 1 (3K)**: We generated one kernel matrix per data type, resulting in a total number of three kernels; the kernel parameter γ was chosen according to the heuristic $\gamma = 1/2d^2$, with d being the number of features of the respective data matrix [51].

- **Scenario 2 (15K)**: We generated five kernel matrices per data type, resulting in a total number of 15 kernels; the kernel parameters γ_n were derived by scaling the previously used heuristic with a constant factor such that $\gamma_n = f_\gamma \, 1/2d^2$, with $f_\gamma \in \{10^{-6}, 10^{-3}, 1, 10^3, 10^6\}$ and d being the number of features.

For each cancer type, the data types were combined using rMKL-LPP with both possible initializations, i.e., starting with the optimization of A and with the optimization of β. The number of retained dimensions was fixed to 5 for several reasons. First, due to the curse of dimensionality, samples with many dimensions tend to lie far apart from each other, leading to sparse and dispersed clusterings. Second, we wanted only a medium number of subtypes, such that very high dimensionality was not necessary. However, reducing the dimensions to two or three, such that visualization would have been possible, could be too simplistic given the heterogeneity of cancer data. After dimensionality reduction, the integrated data points were clustered using k-means (see Section 2.4.1). To decide on the number of clusters, we used the average silhouette width of all samples (see Section 2.5.1.1), a measure that indicates how compact the clusters are and how well they are separated. This enabled to identify the most coherent result among the clusterings with $K = \{2, ..., 15\}$ clusters. The average silhouette width was then also utilized to select the best clustering among the two different initializations.

3.3.1 Results and discussion

In this section, we first evaluate the dependence of the results on the parameter settings and the convergence behavior of the iterative optimization, before discussing the results obtained using the cancer data and their biological implications.

3.3.1.1 Model training

Robustness with respect to parameter settings The iterative optimization of rMKL-LPP either starts with the optimization of the projection matrix A or with the kernel weights β. Concerning these two possibilities,

Table 3.3: Optimal number of clusters (determined by the average silhouette score) when initializing A or β in the 3K and 15K scenario.

	3K		**15K**	
	Initialized		Initialized	
Cancer type	A	β	A	β
BIC	7	6	6	7
COAD	3	2	6	6
GBM	5	5	6	6
KRCCC	6	5	14	7
LSCC	3	2	6	6

the application to the biological data sets showed that initializing β to uniform weights led to slightly better (i.e. higher) silhouette values in most cases. However, the final results for both initializations were highly similar concerning the number of identified clusters (see Table 3.3) and the cluster assignment. Figure 3.1 depicts pairwise similarities between cluster assignments with the same number of clusters but generated using different initializations, i.e., we varied the number of cluster K from 2 to 15 and compared the results of the two possible initializations using the Rand index. The figure illustrates that in most cases, at least 90% of all sample pairs have

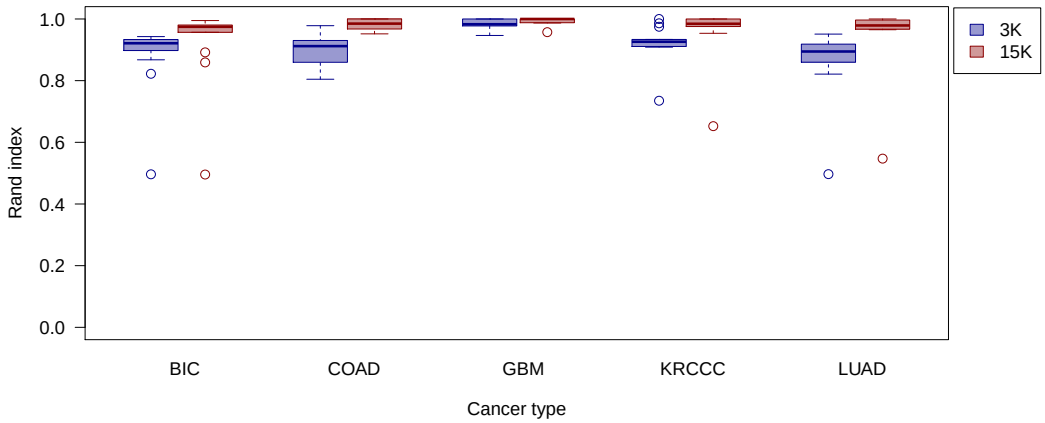


Figure 3.1: **Comparison of rMKL-LPP results with different initializations.** For each number of clusters $K \in \{2, ..., 15\}$ the Rand index was calculated comparing the result obtained when initializing A vs. initializing β.

the same relationship in both results (either they belong to the same cluster in both assignments or to different clusters in both assignments). Furthermore, we can already observe that, despite the increased number of learned parameters in the 15K scenario, the results seem more stable. This finding is supported by higher similarities between the two modes measured according to the number of clusters (Table 3.3) as well as the specific cluster assignments (Figure 3.1).

To evaluate the influence of the number of neighbors $k_{\mathcal{N}}$ on the clustering results, we varied this parameter between 5 and 15. The number of dimensions was fixed at 5, the number of clusters and the initialization were chosen according to the silhouette score. The consistency of the results is visualized in Figure 3.2. For most cancer types, similar cluster assignments are identified when using different numbers of neighbors, resulting in merely slight variations in the Rand index. Only for KRCCC, there seem to be two different possible cluster assignments with relatively low similarity (only approximately 30% percent of sample pairs have the same relationship in the two compared clusterings), which appears as two separate accumulations of Rand indices with rather low variations within the groups.

Motivated by the general stability, subsequent results for all cancer types

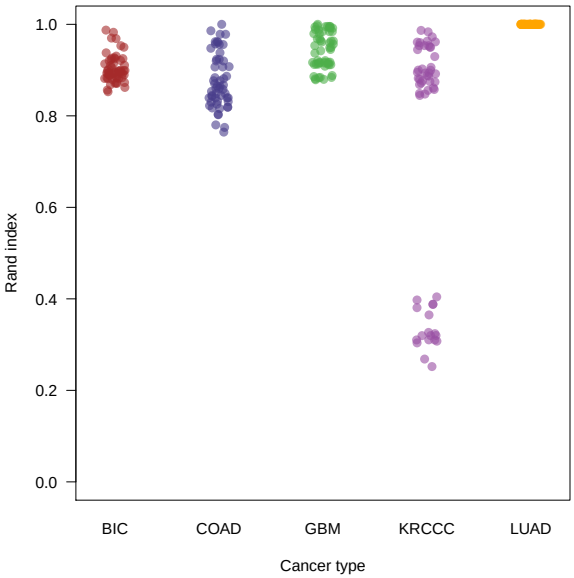


Figure 3.2: **Comparison of rMKL-LPP results with different numbers of neighbors considered.** Each point represents the similarity between two clusterings generated with varied number of neighbors $k_{\mathcal{N}}$ (from 5 to 15).

are based on nine nearest neighbors, although specific optimization would be feasible in terms of running time and memory requirements.

Iterative optimization For the GBM data set with three kernels in total, Figure 3.3 shows the value of the objective function as well as the optimized kernel weights in each iteration of the optimization procedure. In this examples, the kernel weights were initialized to $1/3$ such that the projection matrix A was optimized first. We can see that the objective value improves quickly in the beginning, and after 6 iterations reaches almost the final value. In numbers, the total improvement after the 6th iteration is lower than 5, which is less than 0.14% of the final objective value. However, the convergence threshold of 1E-05 is only reached after 19 iterations. The same findings also hold for the learned kernel weights of the three data types, which remain relatively stable after the 6th iteration. We observed a similar trend when analyzing other cancer types. This suggests that, if necessary, a higher convergence threshold could be used to reduce the running time of the method without strong perturbations of the final results.

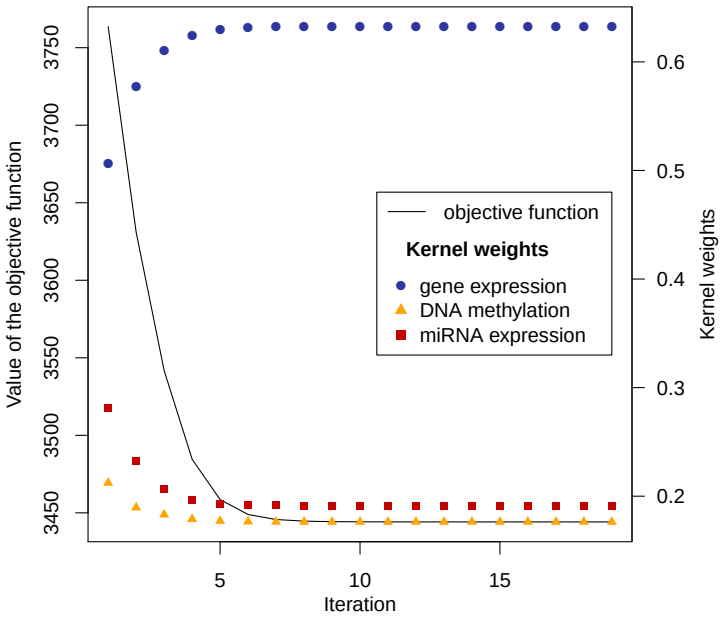


Figure 3.3: **Iterative optimization of the objective value and the kernel weights for GBM** with nine nearest neighbors and projection into five dimensions in the 3K scenario.

3.3.1.2 Comparison to state-of-the-art

We compared the identified clusterings with the results of Similarity Network Fusion (SNF) in terms of survival differences between the clusters (see Table 3.4). Regarding the p-value for the log-rank test (see Section 2.5.2.2), rMKL-LPP with one kernel per data type (3K) has a comparable performance to SNF. Only for KRCCC, the result was not considered significant with a significance level α of 0.05. As can be seen in the last column (15K), the significance for four out of the five data sets increased when using a set of different values for the kernel parameter γ. This indicates that our method is able to capture more information if it is provided. A further observation when increasing the number of kernel matrices from 3 to 15 is the higher optimal number of clusters determined. A possible explanation for this is that more detailed information is contributed by the different kernel matrices. Depending on the kernel parameter setting, similarities between particular groups of patients can appear stronger in the respective kernel matrix while others diminish, leading to a more fine-grained clustering. Moreover, in the 15K scenario, we obtain the least significant p-values for BIC and COAD. This could be a consequence of the composition of the data sets: these two are the cancer types with the smallest number of samples and the lowest number of events (see Table 3.2). One can generally assume an improved performance with higher sample numbers, since they likely cover the under-

Table 3.4: Survival analysis of clustering results of similarity based network fusion (SNF) and the proposed rMKL-LPP with one and five kernels per data type. The numbers in brackets denote the number of clusters. For SNF, these are determined using the eigenrotation method [158], and for rMKL-LPP using the silhouette value.

		rMKL-LPP	
Cancer type	SNF	3K	15K
BIC	1.1E-3 (5)	3.0E-4 (6)	3.4E-3 (7)
COAD	8.8E-4 (3)	2.8E-2 (2)	2.8E-3 (6)
GBM	2.0E-4 (3)	4.5E-2 (5)	6.5E-6 (6)
KRCCC	2.9E-2 (3)	0.23 (6)	4.0E-5 (14)
LSCC	2.0E-2 (4)	2.2E-3 (2)	2.4E-4 (6)
median	1.1E-3	2.8E-2	2.4E-4
product	1.1E-13	1.9E-10	5.9E-19

lying distribution better. Moreover, small numbers of events implicate lower power of the survival analysis. Overall, the performance of rMKL-LPP with five kernel matrices per data type was best, shown by the smallest median and product p-value. Note that the higher number of clusters of rMKL-LPP is controlled in the calculation of the log-rank test p-value by the higher number of degrees of freedom of the χ^2-distribution.

An advantage of the rMKL-LPP method with five kernels per data type is that one does not have to decide on the best similarity measure for each data type beforehand, which makes this method more applicable out of the box. Additionally, the results suggest that it might even be beneficial in some scenarios to have more than one kernel matrix per data type to capture different degrees of similarity between data points (patients in this application scenario).

Runtime As shown by Wang et al. [158], the runtime of the probabilistic approach iCluster (see Section 2.6: Matrix factorization) scales exponentially in the number of genes, which makes the analysis of the cancer data sets infeasible if no gene preselection is performed. For SNF, this preprocessing step is not necessary and it is significantly faster than iCluster. We compared the runtime for the data integration in SNF and rMKL-LPP (15K), which precedes the clustering step in both methods. The SNF approach with the standard parameter settings completes the network fusion procedure for each cancer type within a few seconds, while the data integration with rMKL-LPP (15K) was slightly slower with running times up to one minute. However, just like SNF, rMKL-LPP does not require a gene preselection, which suggests that using data sets with a higher number of samples as well as including more kernel matrices would be feasible in terms of runtimes.

3.3.1.3 Contribution of individual kernel matrices

For rMKL-LPP (15K), Figure 3.4 shows the influence of every kernel matrix on the final integrated matrix. The top bar shows what the graphic would look like for an equal contribution of all kernel matrices. In comparison to this, we can see that kernel matrices using high values for the parameter $\gamma_n = \gamma * 10^6$ have a very low impact for all cancer types. These results agree with the heuristic that the parameter should be chosen in the order of magnitude of $\frac{1}{2d^2}$ or lower, which was used for the choice of γ. Furthermore, most data types contribute significantly to the combined kernel supporting the hypothesis that different kernel parameters lead to different but useful information in the generated kernel matrix. Finally, there are differences between the cancer types, e.g., for BIC, DNA methylation data have a higher

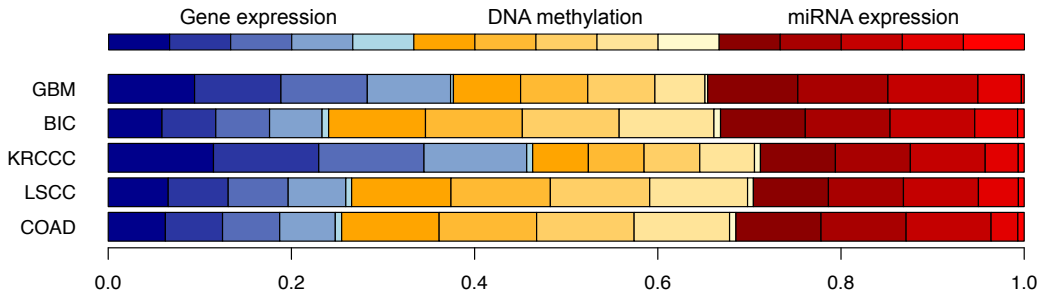


Figure 3.4: **Contribution of the different kernel matrices to each entry in the unified ensemble kernel matrix.** The three colors represent the considered data types: gene expression (blue), DNA methylation (yellow), and miRNA expression (red). The color intensities represent the used kernel parameters γ_n, starting from $\gamma_n = \frac{1}{2d^2} * 10^{-6}$ (high intensity) to $\frac{1}{2d^2} * 10^{6}$ (low intensity), with d being the number of features of the data type.

impact while for KRCCC, there is more information taken from the gene expression data.

3.3.1.4 Robustness analysis

In order to assess the robustness of the approach to small changes in the data set, we performed a leave-one-out cross-validation approach (see Section 3.2.2). Figure 3.5 shows the stability of the clustering when using one kernel matrix per data source (3K) and five kernel matrices per data source (15K). While almost no perturbation in cluster structure appears for GBM and LSCC in 3K, for the other three cancer types there is some deviation concerning the left-out sample but also the clustering of the training samples observable. Especially for the COAD data set, we obtained a number of leave-one-out clusterings in which, compared to the full clustering, one of the clusters was split up into two distinct groups. This separation increases the overall number of clusters from two to three and leads to a strong decrease in the Rand index. The opposite happens for BIC, where we have a full clustering consisting of six groups, while in some of the leave-one-out runs two of the groups collapsed, resulting in five different clusters and, therefore, a lowered Rand index. However, when using five kernel matrices per data source, the results seem to be more stable: they show an increased agreement with the full clustering and a generally reduced variance among the leave-one-out results.

To further investigate the impact of the regularization constraint, we compared the robustness of the results obtained using rMKL-LPP to the

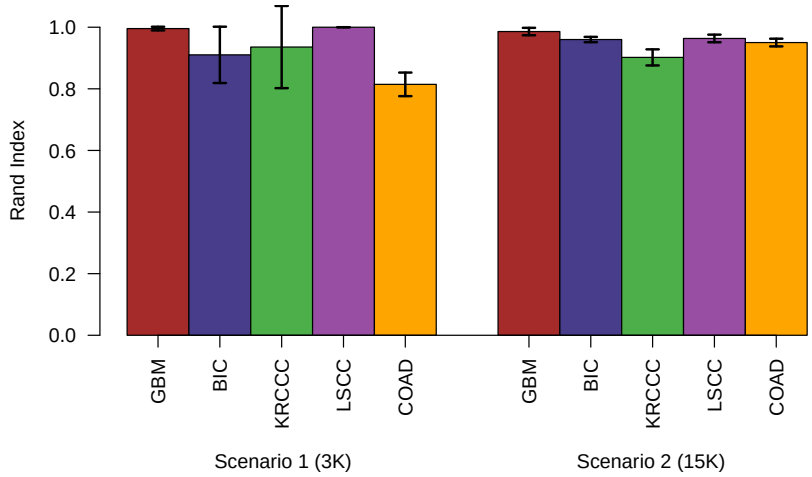


Figure 3.5: **Robustness of clustering results assessed by leave-one-out cross-validation.** Each patient is left out once in the dimensionality reduction and clustering procedure and afterwards added to the cluster with the closest mean based on the learned projection for this data point, which is given by $\text{proj}(x_i) = A^T \mathbb{K}^i \beta$. The depicted Rand index indicates the similarity between the leave-one-out result and the clustering of the whole data set, the error bars represent one standard deviation.

robustness of the results from MKL-LPP, which were generated by simply omitting the regularization constraint. For both approaches, we used the 15K scenario as it seems to result in more stable clusterings. In general, overfitting is expected especially for data sets with a small number of samples or a high number of predictors. To account for this, we created from each cancer data set smaller data sets by randomly sampling without replacement 50% of the patients and repeated the sampling 20 times. On these reduced data, the unregularized MKL-LPP showed the highest instabilities for GBM and KRCCC, with an overall increase in variance among the clustering results compared to the regularized version for most cancer types (see Figure 3.6). This trend continued when the number of samples was further reduced, as Figure 3.7 summarizes for all data types. Although the results without regularization seem to be robust when using the complete data set for each cancer type, the robustness decreased with the number of available samples. The regularized approach, however, showed only a slight decrease in robustness when half the samples of each cancer type were considered, and remained at this level when only one third or one quarter of the data were used. This suggests that rMKL-LPP has advantages in scenarios where MKL-LPP would overfit, while being comparable when no regularization is required.

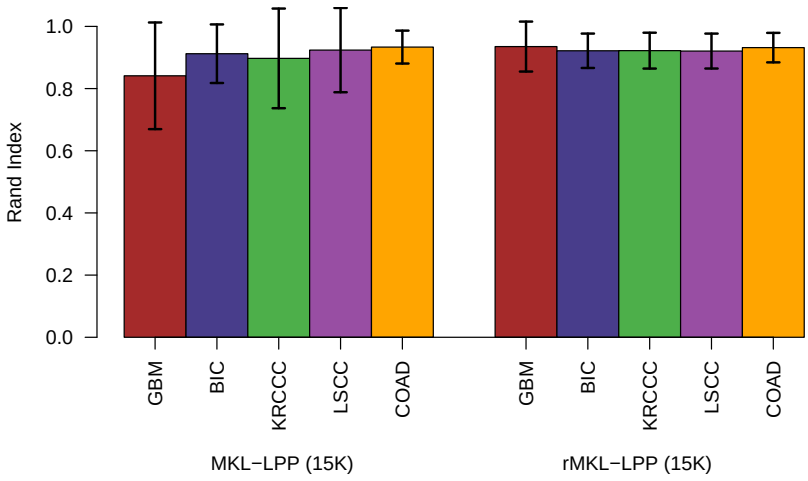


Figure 3.6: **Impact of the regularization of MKL-LPP on the robustness of the clusterings.** For each cancer type, we sampled 20 times half of the patients and applied leave-one-out cross-validation as described in Section 3.2.2. The error bars represent one standard deviation.

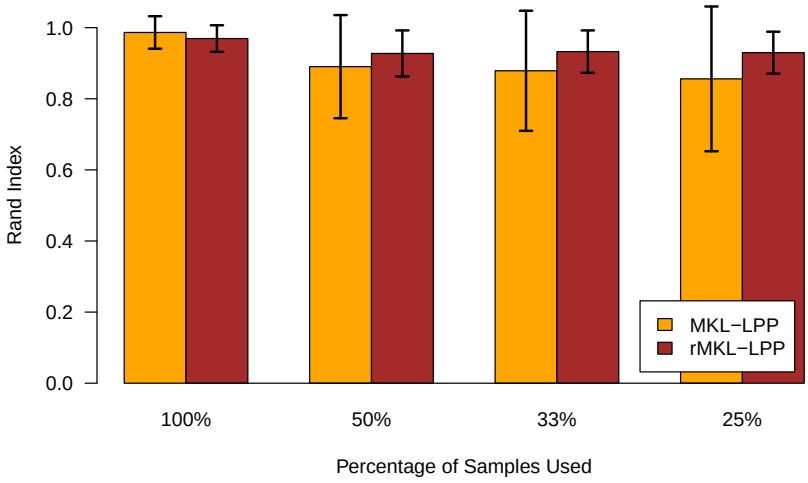


Figure 3.7: **Robustness of the method with and without regularization on data sets of varying size** averaged over all cancer types. The percentage on the x-axis denotes how many patients were used for generating a smaller data set on which leave-one-out cross-validation was performed. For each cancer type and each fraction of patients, we repeated the process 20 times. The error bars represent one standard deviation.

Table 3.5: Comparison of clusters identified by rMKL-LPP to gene expression and DNA methylation subtypes of GBM (Rand indices of 0.75 and 0.64, respectively).

		Subtypes based on:						
		Gene Expression Verhaak et al. [154]				DNA methylation Noushmehr et al. [104]		
		Classical	Mesenchymal	Neural	Proneural	G-CIMP+	#2	#3
rMKL-LPP subtypes	1	0	36	5	1	0	7	37
	2	31	7	13	2	0	46	6
	3	1	0	1	15	16	1	1
	4	1	1	5	22	0	13	27
	5	9	8	2	3	0	19	18
	6	6	1	2	9	3	7	9

3.3.1.5 Comparison of clusterings to established subtypes

In the following, we look further into the results generated by the 15K scenario for the glioblastoma multiforme (GBM) data set. For this cancer type, four subtypes determined by their gene expression profiles [154] have been published. Additionally, Noushmehr et al. [104] identified three patient subgroups via clustering of DNA methylation data, one of them was characterized as the Glioma-CpG island methylator phenotype (G-CIMP)[3]. The comparison of our GBM clustering to these existing subtypes shows that our result does not only reflect evidence from one individual data type, but finds a clustering that combines information from both gene expression and DNA methylation data. This is reflected by a similarity (measured by Rand index) of 0.75 between our result and the gene expression subtypes, and a similarity of 0.64 to the DNA methylation subtypes.

Table 3.5 illustrates that Cluster 1 is strongly enriched for the mesenchymal subtype, whereas Cluster 2 contains mainly samples that belong to the

[3]In this study, only G-CIMP was further characterized, we will thus refer to the other two clusters as Cluster #2 and #3.

classical and the neural subtype. Moreover, Cluster 1 and 2 differ in their predominant methylation cluster assignment. Samples of the proneural subtype are mainly distributed over Cluster 3 and Cluster 4, which are two clusters that can be distinguished at the DNA methylation level by their G-CIMP status. While Cluster 3 consists almost only of G-CIMP positive samples, Cluster 4 contains samples that belong to the proneural subtype but are G-CIMP negative. This shows that in this scenario, evaluating gene expression and DNA methylation data together can provide useful additional information since the analysis based on gene expression data alone could not distinguish Cluster 3 and Cluster 4. For Cluster 5 and 6, we cannot see an enrichment of one of the already established subtypes.

3.3.1.6 Clinical implications of the clusterings

To gain further insights into the biological characteristics of the identified clusters, we investigated how patients of the different clusters respond to the same treatment exemplified by GBM. Analysis and refinement of GBM subtypes is particularly relevant because it is at the same time the most common (3.2/100 000 persons a year) and the most aggressive (survival rate after two years approximately 14%) of primary brain tumors [2]. According to the WHO, GBM tumors are currently mainly classified according to their IDH status as wildtype or mutant [162]. Of the 213 GBM patients, 94 were treated with temozolomide, a drug that forms part of the standard therapy for gliomas. Temozolomide constitutes an alkylating agent, which leads to thymine mispairing during DNA replication, thereby eliminating rapidly duplicating cancer cells [107].

Figure 3.8 shows for each cluster the survival time of patients treated versus those not treated with this drug. Patients belonging to Cluster 5 had a significantly increased survival time when treated with temozolomide with a p-value < 0.01 after Bonferroni correction. The multiple testing correction was applied due to the six tests performed (see Section 2.5.3.1). For Cluster 1 and Cluster 2, we can see a weaker tendency of treated patients living longer than untreated ones (p-value after Bonferroni correction < 0.05), while for the other clusters, we did not detect significant differences in survival time between treated and untreated patients after correcting for multiple testing. To summarize, we detect a difference between the two groups (treated vs. untreated) only in a subset of the identified clusters indicating that the treatment success might depend on the molecular foundation of the tumor. Survival analysis for other medications could show their effectiveness in different groups.

Cluster 3 consists mainly of patients belonging to the proneural expression

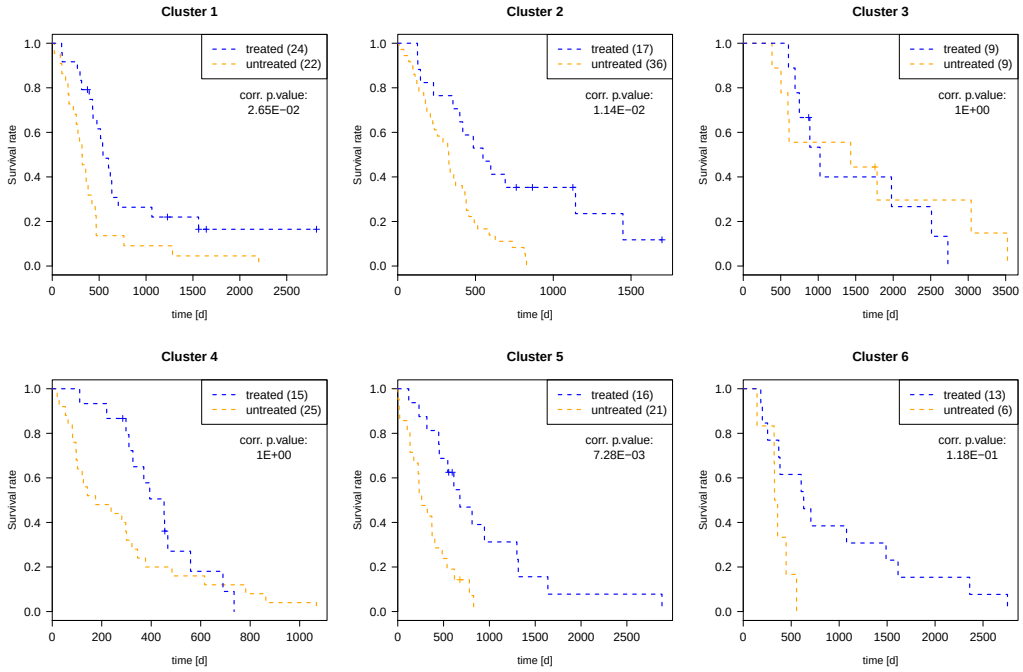


Figure 3.8: **Survival analysis of GBM patients separated according to treatment with temozolomide in the different clusters.** The numbers in brackets denote the number of patients in the respective group. The specified p-values are corrected for multiple testing using the Bonferroni method.

subtype and the G-CIMP methylation subtype. Patients from this cluster show in general an increased survival time; however, they do not benefit significantly from the treatment with temozolomide. We have determined differentially expressed genes between these patients and all other patients using the Kruskal-Wallis rank sum test [80]. As the name suggests, this test is based on the ranks of the data and, therefore, does not assume a specific distribution in the data. It is based on the null hypothesis that the means of the ranks of each group are equal. Under this hypothesis, the test statistic follows a χ^2-distribution with $K-1$ degrees of freedom, where K refers to the number of groups that are compared. The genes that were identified as differentially expressed, were divided into over- and underexpressed and were tested for enrichment of Gene Ontology (GO) terms using over-representation analysis (ORA, see Section 2.5.3).

Table 3.6 shows GO terms of the category Biological Process enriched in the overexpressed set. When comparing the identified terms to those found to

Table 3.6: Enriched GO terms identified by ORA using the category Biological Process for overexpressed genes of GBM Cluster 3.

K=3	GO Term (Benjamini-Hochberg corrected p-value < 0.05)
	tumor necrosis factor production
	positive regulation of lymphocyte differentiation
	positive regulation of transcription factor import into nucleus
	positive regulation of NF kappaB transcription factor activity
	transcription factor import into nucleus
	regulation of transcription factor import into nucleus
	RNA export from nucleus
	regulation of mRNA metabolic process
	mRNA catabolic process
	nucleobase containing compound transport
	regulation of mRNA processing
	tRNA metabolic process

be significant for the G-CIMP positive subtype [104], both results cover similar processes related to the regulation of gene expression in general. Besides, the results for the underexpressed genes, given in Table 3.7, confirmed the G-CIMP-specific downregulation of the extracellular matrix. Additionally, the analysis of the underexpressed genes revealed associations to the immune system and inflammation processes. Although the immune system plays an important role in tumor prevention, chronic inflammation has been associated with tumor formation and progression [59]. Hanahan and Weinberg [59] suggested tumor-promoting inflammation as a characteristic enabling tumors to acquire further hallmarks of cancer. Therefore, the downregulation of immune response-related genes in this patient group might contribute to their favorable outcome.

Table 3.7: Enriched GO terms identified by ORA using the category Biological Process for underexpressed genes of GBM Cluster 3.

K=3	GO Term (Benjamini-Hochberg corrected p-value < 0.05)
	antigen receptor mediated signaling pathway
	regulation of T cell proliferation
	synapse assembly
	leukocyte apoptotic process
	negative regulation of neurogenesis
	cell activation involved in immune response
	negative regulation of mRNA metabolic process
	pallium development
	DNA methylation or demethylation
	chronic inflammatory response
	regulation of extracellular matrix assembly
	negative regulation of inflammatory response
	regulation of DNA damage response signal transduction by p53 class mediator

3.3.2 External validation

Our approach was reviewed in a benchmark paper by Rappoport and Shamir [115]. The authors tested nine different data integration methods covering different integration strategies, such as early integration based on concatenation (LRAcluster, k-means, and spectral clustering), intermediate (MCCA, MultiNMF, iCluster, SNF, rMKL-LPP) and late integration (PINS). Ten cancer types from TCGA were used. The data for each sample consist of DNA methylation, gene expression, and miRNA expression measurements. Our approach rMKL-LPP was used with the same parameters as described previously. Each data type was represented by five different kernel matrices, similar to the 15K scenario, with kernel widths of $\{10^{-6}, 10^{-3}, 1, 10^{3}, 10^{6}\}$. In the absence of a known label, the performance was evaluated on the basis of survival analysis using the log-rank test. Furthermore, the authors tested the identified clusterings for an enrichment of six clinical labels (gender, age at diagnosis, pathologic T, pathologic M, pathologic N, and pathologic stage). The p-values for survival differences and enrichment were estimated based on permutation tests. Figure 3.9 summarizes the performances of the compared

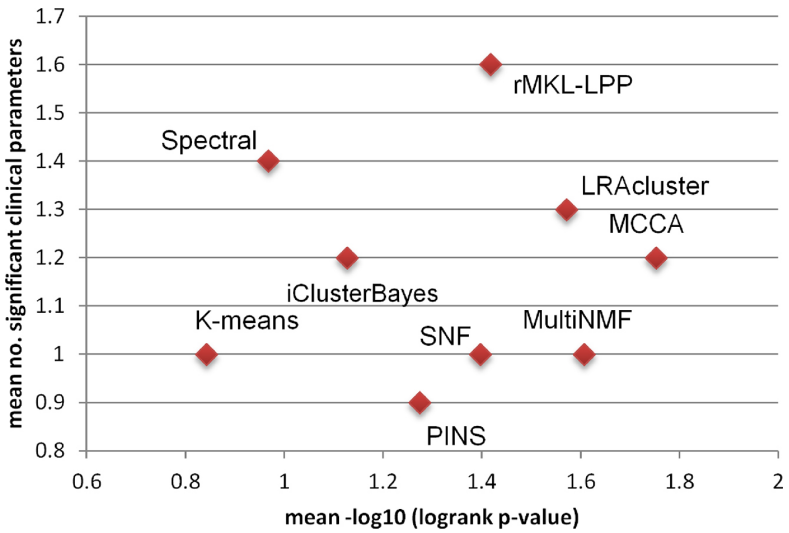


Figure 3.9: **Performance of nine integrative clustering approaches** evaluated via survival p-values and the number of enriched clinical parameters averaged over ten cancer data sets (Figure from Rappoport and Shamir [115], license #4455320736016, see Table B.1).

approaches with respect to these two measures. The optimal method would identify clusterings that are both enriched for clinical parameters and have significant differences in survival time between the clusters, i.e., be located in the top right corner of the plot. One can see that rMKL-LPP provides reasonable results in terms of survival analysis. Additionally, the rMKL-LPP clusters have higher concordance with the chosen clinical labels on average than the clusters identified by other approaches.

3.4 Conclusion

The availability of multidimensional data for cancer patients enables studying this complex disease in a more comprehensive manner. For the unsupervised analysis of patients that aims at identifying interesting subgroups, it is in general not clear how to weight the importance of the different types of information. In this chapter, we have proposed to use a regularized version of the multiple kernel graph embedding framework, which does not only automatically learn an appropriate weight for each source of information but can also be used to implement of various dimensionality reduction techniques.

For patient data from five different cancers, we have shown that our approach implementing locality preserving projections can find subgroups that

provide a better separation with respect to patient survival according to the log-rank test than the ones found by state-of-the-art methods. Furthermore, we have demonstrated that we can utilize several kernel matrices per data type, not only to improve performance but also to remove the burden of manually selecting the optimal kernel matrix. The visualizations of the contributions of the individual kernels and the survival analyses of the final clusterings suggest that kernel matrices based on different parameter settings add valuable information. Moreover, the stability analysis shows that the method does not overfit when more kernels are added. In contrast to the unregularized MKL-DR, rMKL-DR remains stable also for small data sets. For a wide applicability of the method, this is especially important, since in many potential application scenarios the number of available samples is smaller than in this study. The good performance of our method was endorsed in a later benchmark study with an evaluation via survival analysis and enrichment for clinical parameters. The application of the methods to ten different cancer data sets confirmed the stability of our approach.

Our clustering of GBM patients displayed concordance to previous clusterings based on gene expression as well as on DNA methylation data, which shows that rMKL-LPP is able to capture diverse information within one clustering. For the same clustering, we also analyzed the response of the patients to the drug temozolomide, revealing that patients belonging to specific clusters significantly benefit from this therapy while others do not. These results suggest that integrative subtypes can support personalized treatment decisions. The exemplary GO enrichments for one cluster of GBM patients showed, on the one hand, similar results to what was known from the biological literature. On the other hand, the enrichment also indicated a down-regulation of the immune system in the subgroup of cancer patients who survived longer. This suggests that down-regulation of parts of the immune system could be beneficial in some scenarios and points at the controversial role of the immune system in cancer. Follow-up studies for the different clusterings are necessary to assess their biological significance and medical implications.

Furthermore, given that the learned data representation unites information from all sources in a low-dimensional space, other follow-up analyses, e.g., the visualization of new patients, could also benefit from the data integration. Additionally, semi-supervised dimensionality reduction is straightforward in this framework, which makes different analyses possible, e.g., the treatment data can be used as labels, where available, to evaluate how unlabeled data points are distributed in the projected space or over different clusters.

Chapter 4

Multiple kernel principal component analysis

This chapter focuses on extending one particular dimensionality reduction technique, namely principal component analysis (PCA), to enable a joint analysis of multi-omics input. We point out intrinsic problems when the objective function of PCA is used to optimize kernel weights and present an alternative heuristic, which generates promising results on cancer data.

The main parts of Section 4.3 - 4.5 were published in Speicher and Pfeifer [137] as a part of the *13th International Symposium on Integrative Bioinformatics* (IB 2017).

4.1 Overview

Similar to the last chapter, the aim of this chapter is the identification of cancer subtypes after integrative projection of the patients into a low-dimensional space. However, this chapter specifically concentrates on extending the global dimensionality reduction approach PCA, a widely used algorithm, which benefits from valuable advantages: the application to an arbitrary data set is easy, as one does not need to determine parameters that define the size of a neighborhood as is necessary for local dimensionality reduction techniques. Still, due to the possibility of using a kernel function (i.e., using kernel PCA), the method provides enough flexibility to model different types of data with different characteristics also in a nonlinear fashion. Moreover, both linear PCA and kernel PCA learn a projection matrix and hence, do not suffer from the out-of-sample problem. Consequently, new test samples can easily be projected into the same coordinate system.

In the following, we will first show that although kernel PCA (kPCA) can

be implemented in the graph embedding framework (see Section 2.3.3), multiple kernel PCA cannot be solved using the extended framework presented in the previous chapter due to an ill-posed eigenvalue problem. Section 4.3 demonstrates why a direct transfer of the optimization problem for PCA in the multiple kernel scenario only yields a trivial solution that chooses exactly one kernel matrix. Previous observations in molecular cancer subtyping suggest the usefulness of combining several data types instead of considering only one data type or kernel matrix [158], thereby motivating the development of an alternative objective function, which is introduced in Section 4.4. Results on cancer data are presented in Section 4.5.

Related work Several approaches have been proposed to make PCA applicable with multiple inputs. Guo [57] applied a late integration kPCA approach to medical image data. The data types that are combined correspond here to clusters of voxels, each cluster giving rise to a kernel matrix reflecting similarities between subjects. Then, kPCA is applied on each of these matrices separately, and the identified principal components are used in a prediction model. However, when learning this model, the weight of each kernel matrix (or rather the influence of the respective principal components) is determined in a supervised manner based on the outcome of interest. Another version of multiple kernel PCA combines a local scoring term similar to the nearest neighbor-based objective function of LPP (see Section 2.3.1) to the global variance maximizing objective function of PCA to learn kernel weights [118]. Finally, some approaches first optimize kernel weights such that the local topology or the consensual information is preserved and independently apply kPCA to the combined ensemble kernel matrix [98].

In many scenarios, methods using the uniformly weighted average kernel as input give good results or even outperform standard approaches that learn kernel-specific weights [79]. Moreover, using uniform weights does not require making assumptions on the individual data types. For these reasons, we use kPCA on the average kernel matrix as a baseline in this chapter.

4.2 PCA in the graph embedding framework

As described in Section 2.3.3, graph embedding provides a flexible framework for the implementation of different dimensionality reduction schemes including PCA and kPCA. For this algorithm, the similarity matrix W is set to uniform weights, i.e., $w_{ij} = 1/N$, for all $i, j \in \{1, ..., N\}$, and the diagonal constraint matrix D is set to the identity matrix, i.e., $d_{ii} = 1$ for all $i \in \{1, ..., N\}$ (see Table 2.1). Using uniform values in W seems intuitively reasonable as it

reflects that PCA is a global method seeking to preserve as much variance as possible. Moreover, one needs to consider that the introduced minimization problem of graph embedding (Problem (2.12) for linear and Problem (2.22) for multiple kernel formulation) identifies the directions of minimum variance. Instead of solving the minimization problem and discarding these directions, one can formulate a maximization problem to directly identify the directions of maximum variance. The final optimization problem is given by

$$\arg\max_{A,\beta} \quad \frac{1}{N}\sum_{i,j=1}^{N} \|A^T\mathbb{K}^{(i)}\beta - A^T\mathbb{K}^{(j)}\beta\|^2 \tag{4.1}$$

$$\text{subject to} \quad \sum_{i=1}^{N} \|A^T\mathbb{K}^{(i)}\beta\|^2 = const. \tag{4.2}$$

$$\|\beta\|_1 = 1 \tag{4.3}$$

$$\beta_m \geq 0, \quad m = 1, 2, ..., M, \tag{4.4}$$

with $A \in \mathbb{R}^{N\times p}$, $\beta \in \mathbb{R}^M$ and

$$\mathbb{K}^{(i)} = \begin{pmatrix} K_1(1,i) & \cdots & K_M(1,i) \\ \vdots & \ddots & \vdots \\ K_1(N,i) & \cdots & K_M(N,i) \end{pmatrix} \in \mathbb{R}^{N\times M} \tag{4.5}$$

as introduced in Section 2.3.3.2.

PCA is used on centered data, consequently, the projections $A^T\mathbb{K}^{(i)}\beta$ for all data points x_i are centered around zero as well. In this case, the objective function (4.1) is the same as the first constraint (4.2), as both are, up to a multiplicative normalization constant, formulations for the variance.

Theorem 1. Consider m centered kernel matrices $K_m \in \mathbb{R}^{N\times N}$, which are combined into N matrices $\mathbb{K}^{(i)}$ with $i \in \{1, ..., N\}$ as shown in Equation (4.5), a projection matrix $A \in \mathbb{R}^{N\times p}$, and a weight vector $\beta \in \mathbb{R}^M$ with $\|\beta\| = 1$. Then,

$$\frac{1}{N}\sum_{i,j=1}^{N} \|A^T\mathbb{K}^{(i)}\beta - A^T\mathbb{K}^{(j)}\beta\|^2 = 2\sum_{i=1}^{N} \|A^T\mathbb{K}^{(i)}\beta\|^2 \tag{4.6}$$

holds.

Proof. For the sake of brevity, we replace the term $A^T\mathbb{K}^{(i)}\beta$, which is the projection of sample x_i, in the following with pr_i; m_{pr} is used to represent the mean of the projected data.

$$\frac{1}{N}\sum_{i,j=1}^{N} \|pr_i - pr_j\|^2$$

$$
\begin{aligned}
&= \frac{1}{N}\sum_{i,j=1}^{N} \|(pr_i - m_{pr}) - (pr_j - m_{pr})\|^2 \\
&= \frac{1}{N}\sum_{i,j=1}^{N} \left(\|pr_i - m_{pr}\|^2 - 2(pr_i - m_{pr})^T(pr_j - m_{pr}) + \|pr_j - m_{pr}\|^2\right) \\
&= \sum_{i=1}^{N}\|pr_i - m_{pr}\|^2 + \sum_{j=1}^{N}\|pr_j - m_{pr}\|^2 - \frac{2}{N}\underbrace{\sum_{i,j=1}^{N}(pr_i - m_{pr})^T(pr_j - m_{pr})}_{=0,\text{ by the definition of the mean}} \\
&= 2\sum_{i=1}^{N}\|pr_i - m_{pr}\|^2 + 0 \\
&\overset{(m_{pr}=0)}{=} 2\sum_{i=1}^{N}\|pr_i\|^2 \qquad (4.7)
\end{aligned}
$$

As the data were centered beforehand, the mean m_{pr} is equal to zero, resulting in Term (4.7), which was derived the left-hand side of Equation (4.6) and is equal to the right-hand side of Equation (4.6). □

This problem manifests in an ill-posed eigenvalue problem, when optimizing the projection matrix A, as will be shown in the following.

In general, A and β are optimized iteratively, as described in Section 3.2.1.1. The optimization of A is formulated as a trace ratio problem involving the matrices S_W^β and S_D^β. For kPCA, the setting of W and D leads to

$$S_W^\beta = \frac{1}{N}\sum_{i,j=1}^{N}(\mathbb{K}^{(i)} - \mathbb{K}^{(j)})\beta\beta^T(\mathbb{K}^{(i)} - \mathbb{K}^{(j)})^T, \qquad (4.8)$$

$$\text{and} \quad S_D^\beta = \sum_{i=1}^{N}\mathbb{K}^{(i)}\beta\beta^T(\mathbb{K}^{(i)})^T. \qquad (4.9)$$

As $\mathbb{K}^{(i)}\beta$ is equal to $\mathbb{K}_i$, the ith column of the ensemble kernel matrix $\mathbb{K}$, we can rewrite the equations into

$$S_W^\beta = \frac{1}{N}\sum_{i,j=1}^{N}(\mathbb{K}_i - \mathbb{K}_j)(\mathbb{K}_i - \mathbb{K}_j)^T, \qquad (4.10)$$

$$\text{and} \quad S_D^\beta = \sum_{i=1}^{N}\mathbb{K}_i\mathbb{K}_i^T. \qquad (4.11)$$

For two arbitrary but fixed indices $a, b \in \{1, ..., N\}$, we consider $S_W^\beta[a, b]$ and $S_D^\beta[a, b]$, which are entries of the matrices S_W^β and S_D^β, respectively. In the following, we will use m_i for the mean of row i of $\mathbb{K}$, i.e., $m_i = \frac{1}{N}\sum_{j=1}^{N}\mathbb{K}[i, j]$.

$$\begin{aligned}
S_W^\beta[a,b] &= \frac{1}{N}\sum_{i,j=1}^{N}\left(\mathbb{K}[a,i]-\mathbb{K}[a,j]\right)\left(\mathbb{K}[b,i]-\mathbb{K}[b,j]\right)\\
&= \frac{1}{N}\sum_{i,j=1}^{N}\left(\mathbb{K}[a,i]\mathbb{K}[b,i]-\mathbb{K}[a,i]\mathbb{K}[b,j]-\mathbb{K}[a,j]\mathbb{K}[b,i]+\mathbb{K}[a,j]\mathbb{K}[b,j]\right)\\
&= \sum_{i=1}^{N}\left(\mathbb{K}[a,i]\mathbb{K}[b,i]-\mathbb{K}[a,i]m_b-m_a\mathbb{K}[b,i]+\frac{1}{N}\sum_{j=1}^{N}\mathbb{K}[a,j]\mathbb{K}[b,j]\right)\\
&= \sum_{i=1}^{N}\mathbb{K}[a,i]\mathbb{K}[b,i]-2Nm_am_b+\sum_{j=1}^{N}\mathbb{K}[a,j]\mathbb{K}[b,j]\\
&= 2\sum_{i=1}^{N}\mathbb{K}[a,i]\mathbb{K}[b,i]-2Nm_am_b\\
\text{and}\quad S_D^\beta[a,b] &= \sum_{i=1}^{N}\mathbb{K}[a,i]\mathbb{K}[b,i].
\end{aligned}$$

Since we are interested in variation from the mean, and not from the origin, each kernel matrix is centered using Formula (2.6). Consequently, the ensemble kernel matrix $\mathbb{K}$, being a weighted linear combination of the input kernels, will be centered as well. This causes m_a and m_b to be zero, which leads to

$$S_W^\beta = 2S_D^\beta, \tag{4.12}$$

when choosing the parameter setting for kPCA in the graph embedding framework. Therefore, the generalized eigenvalue problem $S_W\alpha = \lambda S_D\alpha$ (or $S_WA = \lambda S_DA$, for matrices) cannot provide a unique, meaningful solution for the projection vector α. For this reason, graph embedding is not suitable for multiple kernel PCA.

4.3 Direct extension of kernel principal component analysis

Instead of making use of the graph embedding framework, the following formulation extends the original optimization problem for the first principal

component in kPCA (Formula 2.11) such that multiple input kernels are incorporated:

$$\begin{aligned} \underset{\alpha,\beta}{\arg\max} \quad & \mathrm{Var}(\alpha^T \sum_{m=1}^{M} \beta_m K_m) \\ \text{subject to} \quad & \beta_m \geq 0, \quad m = 1, ..., M \\ & \sum_{m=1}^{M} \beta_m = 1. \end{aligned} \tag{4.13}$$

λ is the eigenvalue corresponding to the eigenvector α and α is normalized in length such that $\|\alpha\| = 1/\sqrt{N\lambda}$. The latter follows from the orthonormality constraint $A^T \mathbb{K} A = I$, with I being the identity matrix. Yet, this direct implementation does not allow for data integration. The variance of the data in each principal component is given by the respective eigenvalue. In this context, the fact that the direct implementation does not lead to an integration of the different data becomes clear when looking at Thompson's inequality concerning the eigenvalues of sums of matrices [172].

Theorem 2 (Thompson)**.** Consider A and B being $n \times n$ Hermitian matrices and $C = A + B$, with their respective eigenvalues $\lambda_i(A)$, $\lambda_i(B)$, and $\lambda_i(C)$ sorted decreasingly. Then, for any $p \geq 1$ holds

$$\sum_{i=1}^{p} \lambda_i(C) \;\leq\; \sum_{i=1}^{p} \lambda_i(A) \;+\; \sum_{i=1}^{p} \lambda_i(B). \tag{4.14}$$

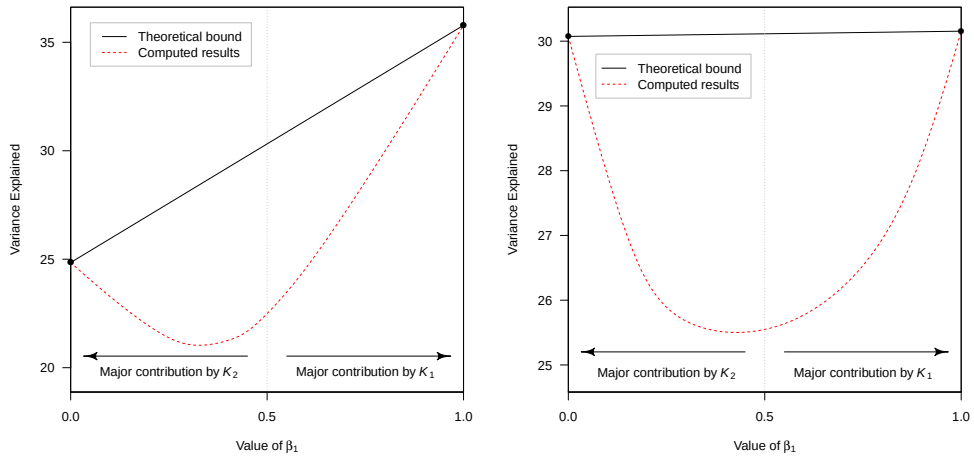
For the special case of the multiple kernel scenario, we extend this formula by including the kernel weight β_1, which simply scales the eigenvalues of the respective matrices. Having $C = \beta_1 A + (1 - \beta_1)B$ and $0 \leq \beta_1 \leq 1$ leads to the following inequality

$$\sum_{i=1}^{p} \lambda_i(C) \;\leq\; \beta_1 \sum_{i=1}^{p} \lambda_i(A) \;+\; (1 - \beta_1) \sum_{i=1}^{p} \lambda_i(B). \tag{4.15}$$

The right hand side is maximized if the kernel matrix with the highest sum of the p largest eigenvalues has a weight of one. In that setting, the right hand side is equal to the left hand side and, thus, this would also be the maximum of the left hand side. Consequently, weights that maximize the sum of the first p eigenvalues are binary, and do not lead to data integration.

Figure 4.1 illustrates the theoretical upper bound for the sum of eigenvalues of weighted combinations of two example data sets[1] calculated using

[1]This chapter uses the same data sets as Chapter 3. Therefore, these are described in Section 3.2.3



(a) Results for kidney renal clear cell carcinoma with K_1 based on gene expression and K_2 based on DNA methylation data.

(b) Results for breast invasive carcinoma with K_1 based on DNA methylation and K_2 based on miRNA expression data.

Figure 4.1: **Theoretical upper bound and practical results for the variance of the ensemble kernel matrix** $\mathbb{K} = \beta_1 K_1 + (1 - \beta_1) K_2$ in the first three principal components.

Equation (4.15). In comparison to this upper bound, the practically achieved sum of eigenvalues for weighted combinations of the example data are shown. The sums of eigenvalues correspond to the variance explained in the first three dimensions and are to be maximized in kPCA. In most cases, the sum of the first p eigenvalues will differ between the individual kernel matrices as depicted in Figure 4.1a. Occasionally, the sums of eigenvalues might be very similar or even the same (see Figure 4.1b). Even in the latter case, we can observe a decrease in the curve depicting variance between the two special cases of β_1 (being either zero or one). This shows that, in any case, a PCA-like algorithm maximizing the variance would choose one of the two trivial solutions instead of a combination of the kernel matrices.

The extension of the theorem to more than two kernel matrices can be made by recursively partitioning the involved matrices, e.g., A can already be the sum of two matrices, which would C make the sum of three matrices. It therefore follows that optimizing Problem (4.13) leads to weight vectors β with $\beta_i = 1$ and $\beta_j = 0$ for all $j \neq i$, where i is the index of the matrix for which the sum of the p largest eigenvalues is maximal.

The constraint $\|\beta\| = 1$ is a major factor for the observed behavior, however, this constraint is necessary to ensure that the ensemble kernel matrix

will be normalized in the RKHS, just like the input kernel matrices are normalized in the RKHS. Dropping the constraint could lead to an arbitrary increase in variance in the ensemble kernel matrix and consequently in the projections, without being based on the input data but due to an increase in the kernel weights.

4.4 Scoring function

A kernel weight vector that chooses one kernel matrix (i.e., an indicator vector) instead of combining the available matrices maximizes the variance in the first p principal components. However, it might not be the best choice for biological data, where we assume that different data types can give complementary information and should therefore be considered jointly. Hence, in the following, we will introduce a scoring function that combines the idea of maximizing the variance with the assumption of different data supplementing each other. Proceeding in the spirit of kPCA, the aim of this approach is to find the ensemble kernel matrix that preserves the global variance best, but also integrates data from different sources. Consequently, new information can be added even if it does not lead to an increase in the total variance of the projection.

For integrating M different kernel matrices $\{K_1, ..., K_M\}$ to an ensemble kernel matrix

$$\mathbb{K} = \sum_{m=1}^{M} \beta_m K_m \quad \text{with} \quad \sum_{m=1}^{M} \beta_m = 1,$$

we propose the following gain function:

$$g_i = \exp\left(\frac{\lambda_i(\mathbb{K})}{\max\{\max_m\{\lambda_i(K_m)\}, 1\}} - 1\right) \tag{4.16}$$

for each dimension i, with $\lambda_i(K_m)$ being the ith eigenvalue of K_m. In the following, we will assume that the eigenvalues are sorted decreasingly, i.e., $\lambda_i \geq \lambda_{i+1}$. The overall score G_p for a projection into a p-dimensional space is then calculated as

$$G_p = \frac{1}{p}\sum_{i=1}^{p} g_i, \tag{4.17}$$

i.e., the average gain over the retained p dimensions. The main idea is the definition of a baseline for each dimension i, i.e.

$$\max\{\max_m\{\lambda_i(K_m)\}, 1\} \tag{4.18}$$

that represents the variance we can have by using only one matrix, more specifically, the matrix with the highest ith eigenvalue. It should be noted that the matrix considered in this baseline might be different with varying i. Due to the use of the exponential function, gains of variance in comparison to this baseline have a strong positive impact on the score while losses of variance are penalized only slightly. Thereby, we can account for the fact that small losses of variance in one direction often do not change the global structure of the data, but allow for more variation in a subsequent direction. Additionally, we ensure that the baseline is not smaller than one, which is the variance each direction would have in case of an equal distribution of the variance. Finally, the best kernel weights β optimize the following objective function:

$$\arg\max_{\beta} \; G_p. \tag{4.19}$$

The projection matrix A is optimized such that it maximizes the variance in the projections given the ensemble kernel matrix $\mathbb{K}$. Although this procedure constitutes a method in which the optimizations of A and β are performed independently and based on different objective functions, the proposed gain function is strongly geared to the variance maximization rationale of PCA.

Figure 4.2 illustrates a scenario where maximizing the variance does not yield the best result in terms of dimensionality reduction. Despite the summative variance in the maximum variance kPCA projection (Figure 4.2 B) being larger than in the gain function kPCA result (Figure 4.2 C) with 10.79% vs. 9.45% of the variance explained by the first two dimensions, the cluster structure is preserved better using the latter approach. This can be explained by the fact that increasing the distance between two already separated clusters does increase the variance, but does not provide additional information in terms of cluster structure. Moreover, using the gain function leads to a more balanced distribution of variance between the two shown principal components. This is not the case in the maximum variance approach, where the variance is mainly concentrated in the first principal component with the second only adding 0.71%.

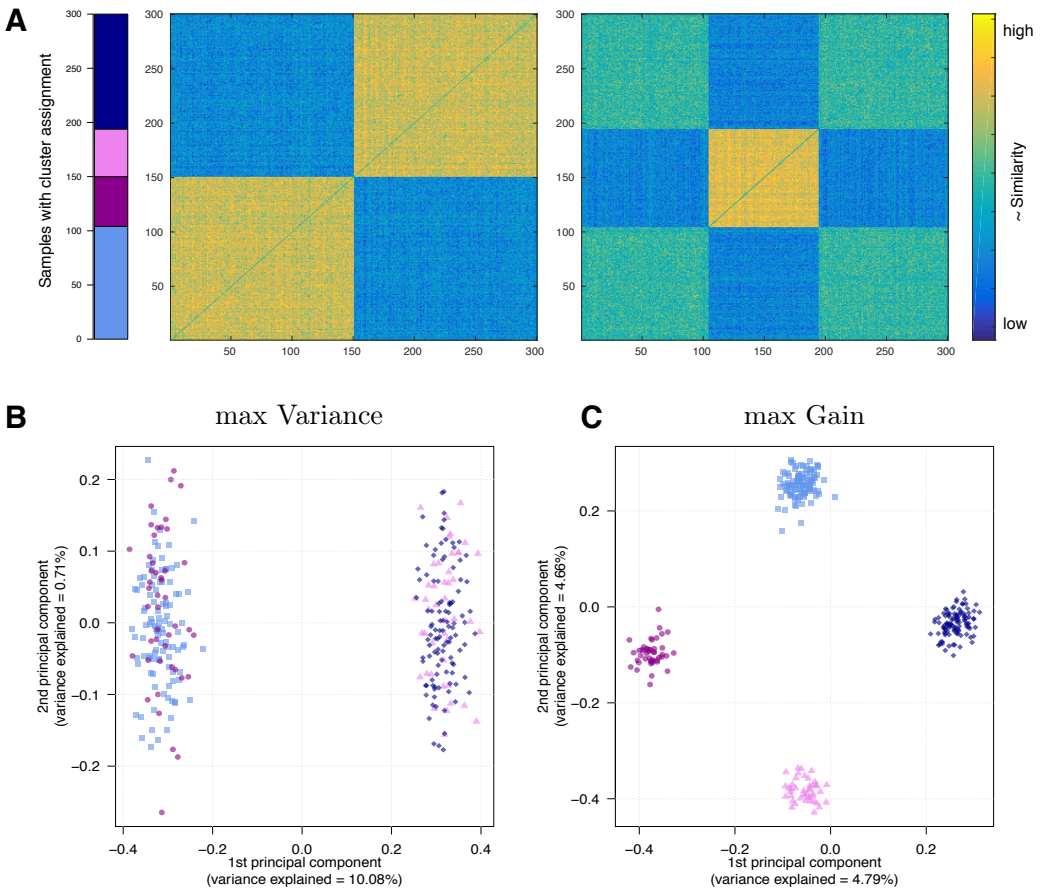


Figure 4.2: **Illustrating example:** different versions of PCA on synthetic data. (A) shows the given cluster assignment of the patients and the two kernel matrices to be combined, where high kernel values (depicted in yellow) roughly correspond to high similarity. (B) and (C) show the results obtained by maximum variance kPCA and by gain function kPCA. The third principal component only harbors less than 0.6% of the variance for both approaches.

4.5 Application to cancer patient data

4.5.1 Materials

The analyses in this chapter were performed on data for five cancer types: breast invasive carcinoma (BIC), colon adenocarcinoma (COAD), glioblastoma multiforme (GBM), kidney renal clear cell carcinoma (KRCCC), and lung squamous cell carcinoma (LSCC). For each cancer type, gene expression, DNA methylation, and miRNA expression data were available. Survival

data were additionally used for the subsequent validation. As the same data were used to test rMKL-LPP (Chapter 3), they are introduced in more detail in Section 3.2.3, with descriptive summaries on the number of samples, features, and events per cancer type given in Tables 3.1 and 3.2.

Implementation

The optimization of gain function kPCA for data integration was implemented in Matlab (version R2016b) [147] via random sampling of the kernel weights. The evaluation including survival analysis was done using custom scripts in R version 3.4.0 [112].

4.5.2 Workflow

For each cancer type, we generated results using a two-step procedure, i.e., we performed integrative dimensionality reduction and subsequent clustering of the patients.

In the first step, we optimized the kernel weights according to the proposed scoring function and ran the dimensionality reduction approach in order to integrate the three data types and reduce the noise in the final projection. In similar approaches, the number of projection dimensions is usually determined either by using the elbow method [6] or based on a chosen threshold for the remaining variance [27]. Here, we could benefit from the proposed scoring function (Section 4.4), which indicates for each dimension if we gain variance by combining matrices in comparison to using only one matrix. The scoring function can thus be used to measure whether adding the subsequent principal component increases or decreases the gain. Since this is a non-convex function, we started with a projection into one dimension and increased the number of considered dimensions. Then, we used the first local maximum of the average gain G_p to determine the number of projection directions. Thereby, we avoided adding directions with no gain in combined variance.

In the second step, we clustered the projected samples using k-means (see Section 2.4.1) to identify potential cancer subtypes. The number of clusters was chosen according to the average silhouette width (see Section 2.5.1.1) of all results from 2 to 15 clusters. For each cancer type, we evaluated the resulting clusterings by comparing the survival of the patients among the different groups using the log-rank test (see Section 2.5.2.2). This test is based on a χ^2-distribution whose degrees of freedom are equal to the number of clusters, such that correction for multiple testing is not necessary. We

compared our approach (gain function kPCA) to two different versions of kPCA:

- **average kPCA** based on the average kernel (i.e. fixed kernel weights of $1/M$), and
- **max variance kPCA** based on the kernel with the highest variance in the first p dimensions. As shown in Section 4.3, this corresponds to the trivial solution to multiple kernel PCA.

4.5.3 Results and discussion

Variance preservation Given that the main idea behind PCA is retaining as much variance as possible, we inspected the variance preserved in the first p principal components derived by the different methods. Figure 4.3 depicts the percentage of variance preserved in the first p principal components, i.e., the sum of the p largest eigenvalues of the respective kernel matrix divided by the total sum of the eigenvalues, for each cancer type. Each kernel matrix was normalized beforehand such that the overall variance equals the number of patients available in the respective data set for the individual kernel matrices and for the ensemble kernel matrix. This means that the reference value of 100% variance remains the same for each cancer type. The number of dimensions p was chosen according to the average gain G_p as described previously.

The inspection of the variances shows that different data types harbor most variance in the first p principal components in different cancer types.

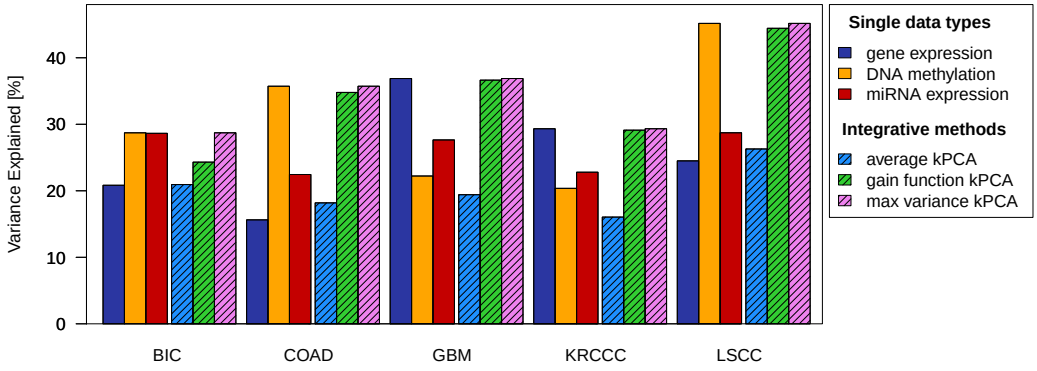


Figure 4.3: **Variances preserved in the first p dimensions of individual data types and combined data.** For each cancer type, the parameter p was chosen according to the average gain G_p (for BIC, $p = 3$; for COAD, $p = 2$; for GBM, $p = 3$; for KRCCC, $p = 3$; and for LSCC, $p = 4$).

As expected, max variance kPCA has the highest variance of the integrative methods because it simply chooses the one data type with maximum variance. Average kPCA preserves much less variance, in some cases even less than any individual data type. This observation is not surprising given the results on the integration of two data types: Figure 4.1a and 4.1b show a decrease of variance below the theoretical upper bound and also below the variances of the individual kernel matrices. For gain function kPCA, the preserved variances are often only slightly below the maximum variance, showing that the overall variance still has a strong impact on the scoring function. This is especially true if there is one data type that clearly dominates the others in terms of variance. For BIC, where DNA methylation and miRNA expression account for approximately the same amount of variance, we can see a stronger decrease in variance for gain function kPCA compared to the max variance kPCA.

Survival analysis The number of retained dimensions p with the highest gain G_p and the p-values of the survival analysis for all three approaches are provided in Table 4.1. For all cancer types, p was rather small (at most four), which could be due to the fact that we used three input data types. If these data types are not strongly correlated, it would be possible that each of them contributed its strongest information thereby adding a useful direction. Nevertheless, this low-dimensional projection harbors meaningful information and facilitates the identification of biologically significant clusters within the cancer types, as is shown in the survival analysis.

Using the conservative significance threshold of $\alpha \leq 0.01$, our method was able to find significant clusters in three cancer types (BIC, COAD, and LSCC). Both other methods identified significant clusters only for two out of the five cancer types (BIC and LSCC for average kPCA; COAD and LSCC for max variance kPCA). This fact indicates that it can be beneficial to combine the two goals, namely data integration and variance maximization, in the objective function. In the GBM data, the gene expression kernel is dominant in terms of variance, therefore, it obtains a high weight in the gain function kPCA. However, there is no clear group structure in this matrix, such that neither max variance kPCA nor gain function kPCA is able to find a clustering that correlates with the survival of the patients. For KRCCC, there is a very small group of patients whose survival behavior differs from the other patients. The survival analysis of the KRCCC clustering shows a trend for max variance kPCA and gain function kPCA (p-values ≤ 0.05) but due to the small number of samples in each cluster, the result is not significant according to $\alpha \leq 0.01$. However, in this example, one can see that

Table 4.1: P-values of survival analysis for the clustering results of kPCA used with an average kernel (average kPCA), a weighted integrated kernel (gain function kPCA), and the kernel with the largest variance in the first p dimensions (max variance kPCA). p denotes the number of considered dimensions. The numbers of clusters was determined by the silhouette value and are given in brackets.

Cancer	p	average kPCA		gain function kPCA		max variance kPCA	
BIC	3	5.7E-4	(4)	6.65E-3	(4)	0.59	(2)
COAD	2	3.28E-2	(3)	6.47E-3	(2)	6.47E-3	(2)
GBM	3	1.59E-2	(4)	0.11	(3)	0.11	(5)
KRCCC	3	0.17	(8)	1.37E-2	(14)	2.27E-2	(14)
LSCC	4	9.22E-3	(3)	7.52E-3	(3)	7.52E-3	(3)
median		1.59E-2		7.52E-3		2.27E-2	
product		4.66E-10		4.86E-10		7.17E-8	

with the unweighted average of the kernel functions, the signal of interest is not captured (p-value = 0.17).

The results for the LSCC data are very stable for all three methods. For all other cancer types, at least one of the naive approaches results in a clustering with no significant difference in survival times between the patient groups, while gain function kPCA results only for GBM in a non-significant clustering. These results show that using the gain function can be beneficial in cases, where only one of the naive approaches identifies a significant clustering. In these cases, gain function kPCA provides the flexibility to determine appropriate weights for the different kernel matrices. Overall, gain function kPCA has the minimal median p-value. One can also see that the average kPCA constitutes a reasonable baseline with a similar product of the p-values. Both integrative approaches outperform max variance kPCA, which only considers one kernel matrix.

Visualization In addition to clustering the cancer patients, the use of dimensionality reduction also enables inspecting the results visually. Figure 4.4 shows the first two principal components learned via gain function kPCA on the LSCC data set. In this graphic, we can see that the black cluster clearly separates from the other two. However, the cluster assignment generated for the other patients seems to be too strict, given that the exact location of the

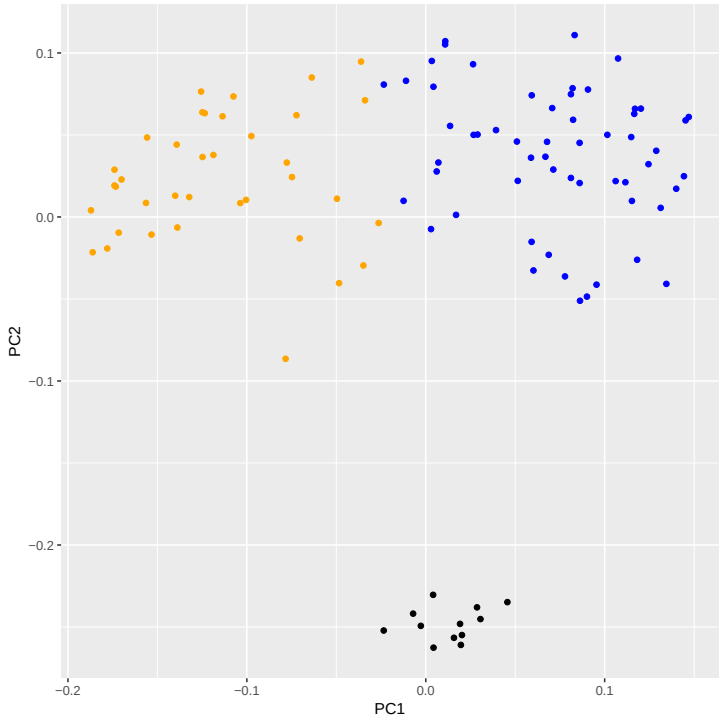


Figure 4.4: **Scatter plot of LSCC patients on the first two principal components identified by gain function kPCA.** Colors represent the different clusters identified by k-means clustering.

border between the clusters might vary with the training data set. Patients close to this border probably share molecular traits from both subgroups and would thus be better described by a mixture of the characteristics of the subgroups.

4.6 Conclusion

This chapter discussed options to extend kernel PCA in order to enable the integration of different data sources. We showed that the objective function of PCA (i.e., maximizing the variance of the projection) in combination with the side constraint for β in multiple kernel learning (i.e., $\sum_m \beta_m = 1$) does not allow for data integration. This objective function is therefore not suitable to optimize kernel weights β for the integration of biological data, where different data types are assumed to add supplementary information. To this end, we introduced a gain function that scores possible ensemble kernel ma-

trices. The function does not use the sum of variances in the projection directions, but instead is based on the increase of variance in individual directions. Our results indicate that kernel weights that maximize this scoring function generally enable the integration of different kernel matrices. Application to five cancer data sets with three different input data types showed that, despite preserving less variance than possible, this approach can perform meaningful data integration in cases where the standard approaches, namely generating an average kernel matrix or choosing the one with the highest variance, cannot capture the relevant patterns.

Future work could focus on improving the heuristic scoring function. Different choices for the denominator in the baseline would be possible, for instance, consistently using the same kernel matrix K_{best} for all dimensions, where K_{best} is the matrix with the maximum sum of the p largest eigenvalues. For increasing number of dimensions, this choice would still consider previous eigenvalues, therefore, the baseline might in some cases be smaller than our currently used one. This would promote a more balanced weight distribution and move the approach farther away from the strict variance maximization. In addition, the exponential function could be replaced by various other functions with similar characteristics. A systematic evaluation of these possibilities on a large number of test sets could either identify the best overall scoring function or determine properties of the test set that can guide the choice of a data-specific scoring function. Moreover, the presented approach disconnects the optimization of the kernel weights from the calculation of the projection directions. It would be promising to develop a multiple kernel PCA where the kernel weights and the projection matrix optimize the same objective function, without preventing data integration as it is the case for the straightforward approach.

Chapter 5

Increased interpretability of unsupervised multiple kernel learning

This chapter addresses the difficulties in the interpretability of multiple kernel learning methods. The extension proposed in this chapter involves the identification of feature sets, and thus represents a version of kernel-based biclustering. Moreover, motivated by previous observations, we assign cluster probabilities to each patient using fuzzy clustering instead of generating a hard clustering.

The content of this chapter is publicly available on arXiv (Speicher and Pfeifer [138]).

5.1 Introduction

Due to the flexibility of the kernel functions, multiple kernel learning methods enable the integration of distinct data types, such as numerical, sequence, or graph data. Additionally, the individual up- or downweighting of kernel matrices can account for differences in the quality or relevance for the learning task at hand. However, these advantages come along with deficiencies of interpretability of the learning process and final results. The general process of unsupervised multiple kernel clustering comprises the calculation of kernel matrices for the input data, subsequent data integration, and clustering of the samples. After obtaining the final result, it is not clear how the learning machine has come to this conclusion, as for nonlinear kernels it is in general not straightforward to identify features that were important for solving the particular learning task. Therefore, evaluation and interpretation of the clus-

tering are difficult, especially when there is no known outcome of interest that can be used as ground truth. In this chapter, we present a general extension for existing unsupervised multiple kernel learning approaches that aims at increasing their interpretability. Therefore, our approach combines feature clustering with subsequent data integration using multiple kernel learning before clustering the samples. Due to the learned kernel weights, we are able to identify which feature clusters were especially influential for each identified sample cluster, making a further biological characterization of the different sample clusters possible.

When analyzing cancer data, identified patient clusters are considered interesting or meaningful if they correlate with clinically relevant characteristics, e.g., if the groups show group-specific responses to treatments. Different clinically relevant patient characteristics are related to different subsets of features, e.g., the success of a particular therapy might primarily depend on the activity of the targeted gene or pathway. Additionally, a therapy-specific set of enzymes processing and transporting the drug might be relevant for the efficacy of the treatment (see for example [124]). In the example of breast cancer treatment, some drugs (e.g., tamoxifen) bind to the estrogen receptor and, therefore, are also influenced by the activity of the estrogen receptor pathway in the tumor cells [111], whereas other drugs rather depend on the ERBB2 signaling pathway (e.g., trastuzumab [169]). Consequently, the available features might be of varying importance for different, clinically relevant patient groups. Therefore, we propose a procedure that is based on the assumption that patient clusters are not defined by all features, but rather by a subset of features. This assumption is also the basis of biclustering or co-clustering, i.e., the simultaneous clustering of samples and features [60]. The approach starts with clustering the features of each data type in order to identify groups of features with similar patterns in the respective data type, e.g. similar expression or methylation patterns over all patients. Based on each feature cluster, a kernel matrix is calculated reflecting sample similarities. Such a kernel matrix can reveal a clear cluster structure for those patients with consistent patterns in the underlying feature set. Since each kernel matrix uses a different set of features, different biological aspects are covered, which can potentially help to identify different groups of patients. In this way, we can reduce the noise in the kernel matrices compared to a kernel calculation based on all available features.

Another new aspect in this chapter is the choice of the clustering method used for the samples. Most clustering algorithms, including k-means, which was used in the two previous chapters, generate a hard clustering, i.e., each sample is assigned to exactly one cluster. This approach is reasonable if the clusters are well separated, or if the assignment of samples into distinct

groups is necessary, e.g., for the sizing of clothes. Samples representing cancer patients might lie between two or more clusters because they exhibit characteristics of each of them, as illustrated in Figure 4.4 for lung cancer patients. In such cases, fuzzy or soft clustering methods give additional information, because they express the patient-to-cluster assignment in a probabilistic – rather than binary – manner. Therefore, we apply fuzzy c-means (see Section 2.4.2), which provides a soft clustering version of the prominent k-means clustering method. A multiple kernel extension for fuzzy c-means has been proposed by Huang et al. [70]. However, we will apply fuzzy c-means to the projection of the samples obtained from regularized multiple kernel based LPP (as introduced in Section 3.3), as this integrative dimensionality reduction approach has been shown to perform very well in comparison to other data integration approaches (see Section 3.3.2).

Related work A general overview of data integration methods is given in Section 2.6. In the following paragraph, we will focus specifically on data integration methods that perform biclustering or aim at improving the interpretability of the clustering results.

A common approach for biclustering is matrix factorization. SRF (*Subtyping with Ranked Factors*) [85] integrates gene expression with Boolean mutation data by transforming both into ranked matrices either directly (expression data), or by using a known network of molecular gene-gene interactions (mutation data). Via rank matrix factorization, Le Van et al. [85] are able to identify patient subgroups associated with one set of genes per data type. In line with the idea of biclustering, the chosen genes show a consistent behavior in the respective data source for the group of patients. On breast cancer data, the method could refine the established subtypes, which are currently used to guide treatment decisions. However, the current approach is limited to the integration of the two data types used. Other approaches are based on non-negative matrix tri-factorization (NMTF), which extends NMF such that a matrix X is decomposed into three non-negative low-dimensional factors, i.e., $X = FSG^T$ [41]. Gligorijević et al. [52] used this approach to co-cluster three different entities, namely patients, genes, and drugs, which are described by mutation data and drug-target interactions. The common dimension in these two binary data matrices are the genes, which makes the joint clustering of the three entities possible. The decompositions of the two input matrices share a common matrix associated with the genes, while the other two derived matrices are unique to the respective data type, and can be used to identify clusters of patients and drugs via a simple matrix binarization approach. The authors constrained the decomposition using known

molecular or chemical similarities within the genes and drugs. Just as in this example, NMTF is often used for relational data. Unlike multi-view data, where each data source describes the same set of samples using different features, relational data additionally contains different types of samples (e.g., genes or patients). Relational data and multi-view data can be understood as a graph, where a node corresponds to either one dimension of a data matrix, that is a set of features or a set of samples, and an edge between two nodes v_1 and v_2 indicates the existence of a data type relating the entities represented by v_1 and v_2. In this representation, a multi-view data graph is star-shaped (i.e., every feature set is connected to the set of samples), while a graph representing relational data is connected but not restricted in its shape.

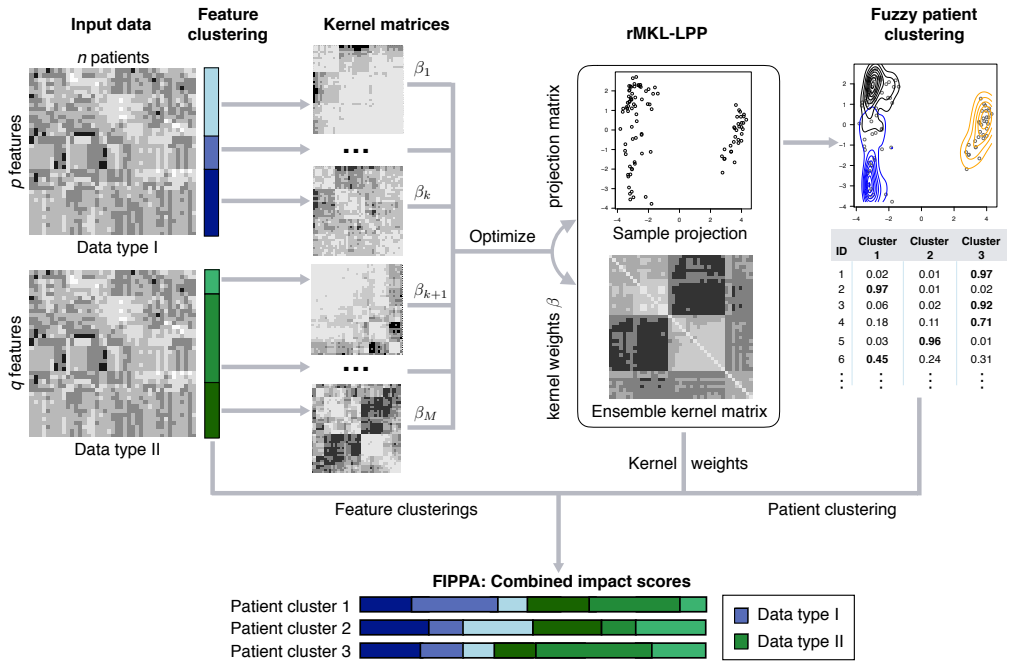
Kernel approaches similar to ours – in the sense that they subdivide the features before applying multiple kernel learning – were presented by Sinnott and Cai [134] for supervised survival prediction of cancer patients and by Rahimi and Gönen [113] for discriminating early- from late-stage cancers. The two methods offered better interpretability compared to standard approaches, because the learned kernel weights provide information concerning the importance of the respective feature sets for the learning task. However, in both cases, the known outcome of interest is used to train the models, which is not the case for our unsupervised clustering. Moreover, the feature groups are identified using prior knowledge, i.e., gene groups reflecting the memberships to pre-defined biological pathways. Thereby, the approaches necessarily exclude genes, methylation patterns, or whole data types that are not already well studied.

5.2 Conceptual outline

We propose a procedure that combines feature clustering with sample clustering based on multiple kernels. Thereby, we increase the potential of interpreting the result without losing the power of the multiple kernel learning approach.

First, we cluster the features of each data type using k-means such that we can generate one kernel matrix based on each feature cluster (see Figure 5.1). The kernel matrices are then integrated using a multiple kernel learning approach. For this purpose, we use regularized multiple kernel learning for locality preserving projections (rMKL-LPP), however, other multiple kernel learning approaches could be used. The increased homogeneity of the features within a feature cluster can reduce the noise in each kernel matrix. In this way, a signal that is generated by only a few features can still signifi-

cantly influence the final result if these features behave consistently over a subset of samples. Based on the low-dimensional projection, we cluster the samples using fuzzy c-means. The availability of the feature clusters and respective kernel weights helps interpreting the obtained patient clusters, as each patient cluster can be traced back to those feature clusters that had the highest influence on the sample similarities. We introduce the score FIPPA (Feature cluster ImPact on PAtient cluster) that can be used to quantify this influence. To our knowledge, this is the first extension of a multiple kernel clustering algorithm to integrative biclustering.



ID	Cluster 1	Cluster 2	Cluster 3
1	0.02	0.01	**0.97**
2	**0.97**	0.01	0.02
3	0.06	0.02	**0.92**
4	0.18	0.11	**0.71**
5	0.03	**0.96**	0.01
6	**0.45**	0.24	0.31

Figure 5.1: **Overview of the proposed approach exemplified for two input data matrices.** First, feature clustering is performed, here with $K = 3$. Each feature cluster gives rise to one kernel matrix, which are integrated using rMKL-LPP. This method optimizes one weight for each kernel matrix and a projection matrix, leading to a low-dimensional representation of the samples. Using these learned coordinates, the samples are clustered using fuzzy c-means. Finally, the feature clusters, kernel weights per feature cluster, and the patient clusters are used to calculate FIPPA scores, which indicate the feature cluster impact on a patient cluster.

5.3 Methods

As described in the Section 5.2, the presented approach combines k-means clustering (see Section 2.4.1) with rMKL-LPP (see Section 3.3). The latter is a multiple kernel extension of locality preserving projections, a dimensionality reduction scheme aiming at preserving local neighborhoods. Using this objective, rMKL-LPP optimizes one weight β_m per kernel matrix and a projection matrix $A \in \mathbb{R}^{N \times p}$, such that the information in the different kernels can be combined into an ensemble kernel matrix $\mathbb{K} = \sum_{m=1}^{M} \beta_m K_m$, and projected via $\text{proj}(x_i) = \mathbf{A}^T \mathbb{K}^{(i)} \boldsymbol{\beta}$. The result is a p-dimensional, integrative projection of the samples, which are subsequently clustered using fuzzy c-means (see Section 2.4.2).

The following section describes the calculation of FIPPA, a score that measures the impact of the feature groups on the density and separation of the individual patient clusters.

5.3.1 Feature cluster impact on patient cluster

After the data integration, each feature cluster FC_m with $m \in \{1, .., M\}$ is associated with a kernel matrix K_m and a kernel weight β_m. For each patient cluster C_k with $k \in \{1, ..., K\}$, we denominate the set of all indices of samples assigned to this cluster by I_{C_k}, i.e.,

$$I_{C_k} = \{l | x_l \in C_k\}. \tag{5.1}$$

The impact of each feature cluster FC_m on each identified patient cluster C_k (in the following: $\text{FIPPA}_{k,m}$) can be calculated based on K_m and β_m using the following equation

$$\text{FIPPA}_{k,m} = \frac{2}{|C_k|^2 - |C_k|} \sum_{\substack{i,j \in I_{C_k} \\ j > i}} \frac{\beta_m K_m[i,j]}{\mathbb{K}[i,j]}, \tag{5.2}$$

with $\mathbb{K}$ being the ensemble kernel matrix. Intuitively, we calculate for each sample pair with both partners being assigned to C_k how much of the kernel value in the ensemble kernel can be attributed to the specific feature cluster. The factor in the beginning accounts for the $\frac{1}{2}(|C_k|^2 - |C_k|)$ pairs of samples considered. Due to the symmetry of each kernel matrix and the ensemble kernel matrix, the order of the samples in the pair is not relevant and we use either (x_i, x_j) or (x_j, x_i) for the calculation of FIPPA.

In Equation (5.2), we suppose a hard clustering of the patients, which can be generated using a hard clustering algorithm, such as k-means, or using

the modal class of a fuzzy clustering. However, fuzzy clustering provides additional information concerning the reliability of the cluster assignment for each sample. Using these probabilities can make the calculated scores more robust, because it takes into account that some samples might have an ambiguous signature and therefore lie between two or more clusters. Given the probability $p_k(x_i \wedge x_j) = p(x_i \in C_k \wedge x_j \in C_k)$ that patient x_i and patient x_j both belong to the same cluster C_k, the fuzzy FIPPA can be calculated as follows:

$$\text{fFIPPA}_{k,m} = \frac{2}{N^2 - N} \sum_{i=1}^{N} \sum_{j=i+1}^{N} p_k(x_i \wedge x_j) \frac{\beta_m K_m[i,j]}{\mathbb{K}[i,j]}, \tag{5.3}$$

where N is the total number of samples. The incorporation of the joint probability $p_k(x_i \wedge x_j)$ replaces the selection of sample pairs performed in Equation (5.2), such that the effects of sample pairs where at least one partner is unlikely to belong to C_k are reduced. For all sample pairs (x_i, x_j) with $i \neq j$, we assume independence between $p_k(x_i)$ and $p_k(x_j)$, such that $p_k(x_i \wedge x_j) = p_k(x_i)p_k(x_j)$.

We generally calculate the positive and the negative part of the importances separately to avoid that terms cancel each other out, which is possible when the kernel matrices are centered in the feature space. For this purpose, we define

$$\mathbb{K}^+ = \sum_{m=1}^{M} \beta_m K_m^+ \quad \text{and} \quad \mathbb{K}^- = \sum_{m=1}^{M} \beta_m K_m^- \tag{5.4}$$

with K_m^+ being the positive part of the matrix K_m (all negative values set to zero) and vice versa for K_m^-. Moreover, we can calculate the fFIPPA based on the positive values ($\text{fFIPPA}_{k,m}^+$) such that it is related to high intra-cluster similarity, by using the joint probability $p_k(x_i \wedge x_j)$ as probability factor for each summand. In this way, sample pairs where both partners are likely to belong to cluster C_k have higher influence on the feature cluster-specific score.

In contrast to $\text{fFIPPA}_{k,m}^+$, the fFIPPA score based on the negative values ($\text{fFIPPA}_{k,m}^-$) is calculated to disentangle high dissimilarity between two clusters, i.e., this score emphasizes the differences between two clusters. This is achieved by choosing the *exclusive or*, defined by

$$p_k(x_i \oplus x_j) = p_k(x_i) + p_k(x_j) - 2p_k(x_i \wedge x_j), \tag{5.5}$$

as probability factor. This choice results in an increased factor for pairs of samples of which exactly one partner has a high probability of belonging to

C_k. When combining Formula (5.4) with Formula (5.3) and the adjusted factors, the calculations of $\text{fFIPPA}^{+}_{k,m}$ and $\text{fFIPPA}^{-}_{k,m}$ are given by

$$\text{fFIPPA}^{+}_{k,m} = \frac{2}{N^2 - N} \sum_{i=1}^{N} \sum_{j=i+1}^{N} p_k(x_i \wedge x_j) \frac{\beta_m K^{+}_m[i,j]}{\mathbb{K}^{+}[i,j]}, \quad \text{and}$$

$$\text{fFIPPA}^{-}_{k,m} = \frac{2}{N^2 - N} \sum_{i=1}^{N} \sum_{j=i+1}^{N} p_k(x_i \oplus x_j) \frac{\beta_m K^{-}_m[i,j]}{\mathbb{K}^{-}[i,j]}. \tag{5.6}$$

Due to the construction of the fFIPPA scores, the calculated values facilitate the identification of feature clusters that contribute more than average to the similarity of the samples within a sample cluster, and also the identification of features clusters that contribute more than average to the dissimilarity of a cluster to all other clusters. This information can help in revealing the underlying basis of the generated clustering result.

5.3.2 Materials

We applied our approach to six different cancer data sets from The Cancer Genome Atlas (TCGA), which were downloaded from the UCSC Xena browser [53][1]. The cancer types included in the analysis are breast invasive carcinoma (BIC), lung adenocarcinoma (LUAD), head and neck squamous cell carcinoma (HNSC), lower grade glioma (LGG), thyroid carcinoma (THCA), and prostate adenocarcinoma (PRAD). For each cancer patient, we used DNA methylation, gene expression data, copy number variations, and miRNA expression data for clustering. For the evaluation, we further leveraged the survival times of the patients. Table 5.1 provides an overview of the number of events (i.e., deaths) that could be used for a survival analysis. Due to the small number of events, we excluded the PRAD and THCA data set from the evaluation via survival analysis.

Data preprocessing

To avoid gender bias, we removed all male patients from the BIC data. Since breast cancer is very rare in males, this step reduced the sample size only marginally. The PRAD data set did not contain any female patients to be removed. Features were excluded if they had more than 20 % missing values, and, similarly, patients were excluded if they had more than 20% missing features. This threshold was chosen in analogy to the preprocessing performed

[1] date accessed: 2017/10/24

Table 5.1: Number of samples N and number of events for each cancer type. Discrepancies between N and the sum over the last two columns occur due to the fact that for some patients the vital status (dead/alive) is missing.

	Cancer type	N	W/ event	W/o event
BIC	Breast invasive carcinoma	600	76	522
HNSC	Head and neck squamous cell carcinoma	464	188	274
LGG	Lower grade glioma	503	124	375
LUAD	Lung adenocarcinoma	439	150	278
PRAD	Prostate adenocarcinoma	482	8	473
THCA	Thyroid carcinoma	490	14	475

Table 5.2: Number of features per data type for each cancer type after the removal of samples and patients with more than 20% missing values.

Cancer type	Gene expression	DNA methylation	Copy number variation	miRNA expression
BIC	20 209	59 097	24 776	516
HNSC	19 433	57 159	23 817	581
LGG	19 414	57 159	23 817	592
LUAD	19 351	57 156	23 817	567
PRAD	20 219	59 214	24 776	484
THCA	19 366	57 158	23 817	556

by Wang et al. [158] for the data used in Chapters 3 and 4. Table 5.2 displays the number of remaining features for each data type. For this data set, missing values were imputed as the average of the three nearest neighbors. Gene expression, copy number variation, and miRNA data were standardized to a mean of zero and a variance of one. The DNA methylation data was quantile normalized via BMIQ [145]. The measurements, which were available at the level of methylation sites, were summarized for gene promoter and gene body regions using RnBeads [11]. Using GeneTrail2 [140], we mapped the miRNAs to their target genes (i.e., the gene that is regulated by the miRNA). Finally, due to the high number of features in the data set, we applied the method to a reduced data set for each cancer type containing the 10% most variable

features of each data type, in any case keeping at least 500 features.

Implementation

Preprocessing of the data was done using custom scripts in R version 3.4.0 [112] and Matlab version R2016b [147]. The feature clustering with subsequent application of rMKL-LPP and patient clustering was implemented in Matlab. The calculation of the fFIPPA scores, as well as the subsequent gene filtering, was done using custom scripts in R. GO enrichment of the resulting gene lists was performed using GeneTrail2 [140].

5.4 Results and discussion

5.4.1 Parameter selection

When applying our method to a data set, the user needs to choose the number of feature clusters per data type, as well as the number of patient clusters. For the validation of our method, we set both parameters to the same value ($K \in \{2, ..., 6\}$). Choosing the number of feature clusters according to the silhouette score would result in the same feature clusters independent of the number of sample clusters. To increase the differences in the used sets of features (and consequently kernel matrices), we interlinked these two parameters, i.e., the number of clusters for features and samples. Feature clustering was performed using k-means before generating the kernel matrices using the Gaussian radial basis function (RBF) kernel. The kernel parameter γ was chosen according to the heuristic of setting $\gamma = \frac{1}{2d^2}$ [51]. In other words, the parameter γ varies depending on the number of features in the respective feature set. We generated three kernels per feature set by multiplying γ with a factor $f_\gamma \in \{0.5, 1, 2\}$, and only used the one kernel matrix with the highest variance in the first p principal components. This parameter p also refers to the number of retained dimensions for rMKL-LPP and was set to five. The number of neighbors for rMKL-LPP was set to nine[2]. The fuzzification degree f of the soft-clustering algorithm fuzzy c-means was set to the default value of 2 in concordance with Dunn [44]. If necessary for the subsequent analysis (for instance survival analysis), we assigned each patient to its modal cluster (i.e., the cluster with the highest probability). Otherwise, we used the cluster membership probabilities returned by fuzzy c-means.

[2]The number of retained dimensions and the number of neighbors are both parameters of rMKL-LPP and discussed in more detail in Section 3.3.

5.4.2 Robustness of the final clusterings

Since each kernel matrix in our method is constructed on the basis of a previously identified feature cluster, slight changes in the initial feature clusters will propagate through the algorithm and might have an effect on the final result. We repeated the complete analysis 50 times with different random initializations for the clustering steps to analyze the robustness of the final patient clustering. The pairwise similarities between the patient clusterings measured by the Rand index are depicted in Figure 5.2.

When including all patients according to their modal class, we observed high reproducibility of the clusterings for all cancer types except LUAD, for which the average Rand index is approximately 0.85. We then used the cluster probabilities for each sample to exclude patients where the assignment has low confidence, here defined as the probability for the modal cluster being more than one standard deviation lower than the mean over all patients. Using this smaller set of patients, the cluster assignments for the remaining patients were more stable for all cancer types including LUAD. Overall, this suggests that, despite the random initializations, the results are repro-

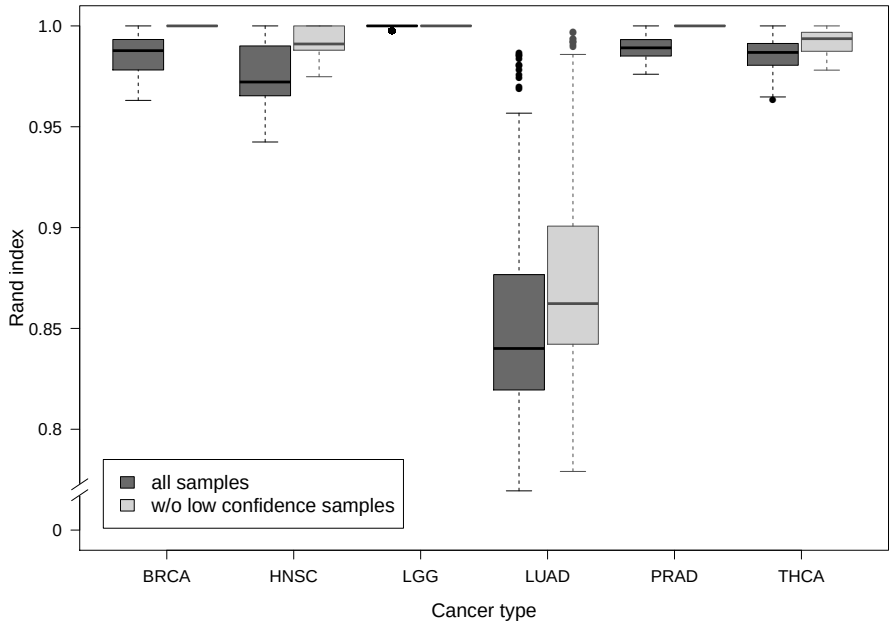


Figure 5.2: **Robustness of FC+rMKL-LPP** in 50 repetitions. Dark grey boxes indicate Rand indices on the basis of all patients. Light grey boxes indicate Rand indices calculated without low confidence samples, i.e., patients with a modal class probability p_k more than one standard deviation smaller than the mean of the modal class probabilities.

ducible for most samples. Additionally, the cluster probabilities can be used to remove outliers or patients that are located between several clusters and, thereby, increase the robustness of the approach.

5.4.3 Survival analysis

The clinical relevance of a patient clustering can be evaluated, for example, using survival analysis (see Section 2.5.2.2). We use this analysis to investigate if the performance of the chosen multiple kernel learning approach (here rMKL-LPP) degrades when prior feature clustering is performed. Therefore, we also generated patient clusterings using rMKL-LPP without feature clustering. For comparison, we also include kLPP based on the average kernel into the analysis. As this method does not involve feature clustering or kernel weight optimization, one kernel matrix per data type is used to calculate the unweighted average kernel. In general, multiple kernel methods using uniform weights have been shown to perform well in many scenarios [79], but lack the additional information gained through the kernel weights. For all three approaches, we evaluated the clustering results for $K \in \{2, ..., 6\}$ and chose the number of clusters that resulted in the lowest p-value for the log-rank test.

The p-value for the log-rank test indicates if there is at least one group of patients among the identified ones with a significant difference in the survival rate compared to the other groups. While the number of clusters is reflected in the degrees of freedom that are used in computing the test statistic, we still corrected for the number of tests performed due to the variation of K using Benjamini-Hochberg correction.

Table 5.3 summarizes the obtained results for the three methods. For our approach, we report the median p-value from 50 runs to account for variation due to the feature clustering. The results indicate that, on the four cancer types where the number of events allows for a meaningful survival analysis, our method performs comparably to the established methods, i.e., it is able to find biologically relevant groups of patients. Moreover, despite the additional flexibility that is achieved by the feature clustering, the optimal number of clusters does not increase in comparison to less complex methods. Besides, average kLPP performs well on three cancer types, but does not lead to significant clusters with respect to survival differences for LUAD. This finding supports the assumption that in some scenarios, the unweighted average cannot capture the provided information.

However, for some data sets, the small number of events does not permit a robust survival analysis, as it is the case for the PRAD and THCA data. Additionally, the survival data can be biased as not all patients receive optimal

Table 5.3: Survival analysis of the clustering results obtained by different methods. Average kLPP stands for kernel locality preserving projections on the (uniformly weighted) average kernel, rMKL-LPP for the standard multiple kernel learning approach with one kernel per data type, and FC + rMKL-LPP represents the proposed approach for which the reported p-values are the median of 50 runs. All three approaches are combined with subsequent fuzzy c-means clustering. For each cancer type and method, we chose the number of clusters (given in brackets) according to the best survival result.

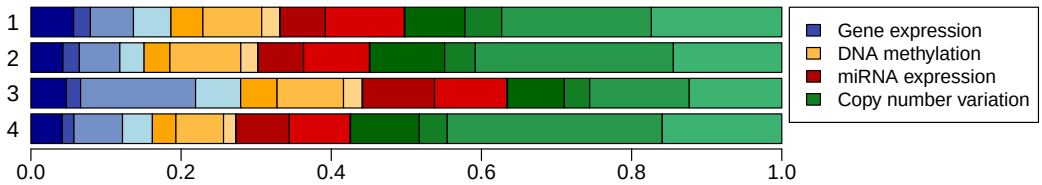
Cancer	average kLPP		rMKL-LPP		FC + rMKL-LPP	
BIC	3.7E-2	(6)	7.3E-2	(6)	5.0E-2	(4)
HNSC	1.4E-3	(6)	1.4E-3	(6)	9.96E-3	(5)
LGG	<1.0E-16	(3:6)	<1.0E-16	(3:6)	<1.0E-16	(3:6)
LUAD	0.15	(2)	2.9E-2	(2)	3.1E-2	(6)

treatment. Therefore, FIPPA provides an additional strategy to nevertheless interpret the clustering results without relying on survival data.

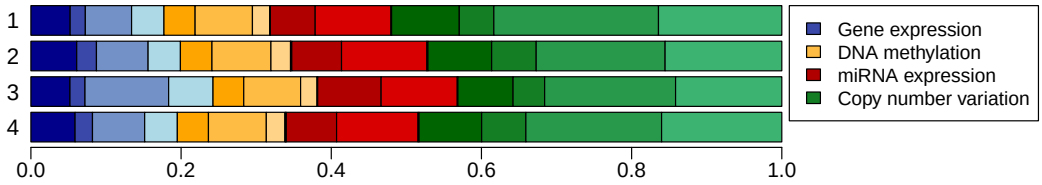
5.4.4 Interpretation

As described in Section 5.3.1, the weights for each feature cluster in combination with the kernel matrices facilitate calculating the fFIPPA scores. As a result, we obtain one fFIPPA^+ score and one fFIPPA^- score for each pair of feature cluster and patient cluster. Figure 5.3 visualizes the calculated fFIPPA scores exemplified for the BIC clustering with $K = 4$. While the underlying assignment of features in clusters does not change with the patient cluster, we can see clear differences between the fFIPPA scores for the same feature cluster in the different patient clusters. For all patient clusters, copy number variations have the strongest impact, but the exact contributions of the data types vary (e.g., the fFIPPA of gene expression, especially feature cluster 3, is strongest in patient cluster 3). When comparing fFIPPA^+ (Figure 5.3a) and fFIPPA^- (Figure 5.3b), we observe in general similar patterns. Differences still exist, e.g., the fFIPPA^+ of CNV is larger than fFIPPA^- for patient clusters 2 and 4, meaning that CNV contributes more to the intra-cluster similarity (i.e., what makes the patient cluster dense) than to the dissimilarity to other patients.

Identification of cancer subtype-specific features Deriving the impact of each data type is also possible with traditional multiple kernel learn-



(a) fFIPPA^{+} scores for intra-cluster similarity of the patient clusters.



(b) fFIPPA^{-} scores for inter-cluster dissimilarity of the patient clusters.

Figure 5.3: **fFIPPA scores of each feature cluster and patient cluster** for BIC with $K = 4$. Each row represents one patient cluster C_k, i.e., it is derived using the sample probabilities for this specific cluster. Different colors represent the different data types. Each bar segment represents one feature cluster FC_m, where the feature clusters remain the same independent of the patient clustering considered. The width of a segment is defined by $\mathrm{fFIPPA}^{+}_{k,m}$ (or $\mathrm{fFIPPA}^{-}_{k,m}$) for the respective patient cluster C_k and feature cluster FC_m.

ing. However using our method, we can additionally analyze the impact of the individual feature clusters on specific patient clusters, and thereby further characterize the potential cancer subtypes. We could describe each feature cluster by the main functions in which the included features participate, i.e., by a set of representative functions. These representatives could be directly associated to the calculated fFIPPA scores. However, in doing so, we would ignore synergies between different feature clusters and could not detect functionalities that are spread over different feature clusters. Therefore, we instead generate lists of relevant features for each patient cluster in two steps:

1. For each patient cluster, we identify high impact feature clusters.
2. We identify genes within the high impact feature clusters, that exhibit constant patterns in the respective data type over the patients of the related cluster.

In the first step, we chose all feature clusters that contributed more than average to the patient cluster. For this purpose, the average was calculated

separately for $\text{fFIPPA}^{+}_{k,m}$ and $\text{fFIPPA}^{-}_{k,m}$. Consequently, we generated two lists of feature clusters for each patient cluster C_k: the first list is associated with high dissimilarity to other patient clusters (based on fFIPPA^{-}), the second list is associated with high similarity within the patient cluster (based on fFIPPA^{+}). The fFIPPA calculation assigns the same value to all genes in one cluster, but these clusters might not be entirely homogeneous due to the limited number of feature clusters and the impossibility to exclude outliers. After identification of the high-fFIPPA feature clusters, we therefore filtered the associated features in the second step. The aim was to keep only those features with a consistent pattern over the patients of the specific cluster. The consistency was checked using the data of the respective type, i.e., gene expression, methylation, or copy number data. Besides the used data, the proceeding of this filtering step was the same for every feature. We counted in how many patients of the respective patient cluster the gene was consistent in its expression, methylation, or copy number (i.e., either above or below the average, whichever happened more often). Based on these counts, we obtained a distribution representing the homogeneity of the features with respect to a patient cluster. For further analysis, we kept every feature with a count larger than the mean of this distribution. In other words, we kept all genes from a high-fFIPPA feature cluster that showed a more homogeneous expression, methylation, or copy number than the average gene in the considered patient cluster.

Finally, we distinguished between features with a higher expression, methylation, or copy number than the average and features with a lower expression, methylation, or copy number than the average. In this way, we generated four feature lists for each patient cluster describing a) active genes leading to high intra-cluster similarity, b) inactive genes leading to high intra-cluster similarity, c) active genes leading to high inter-cluster dissimilarity and d) inactive genes leading to high inter-cluster dissimilarity.

Insights from cancer subtype-specific features In the assembled lists, we found genes with known implications to cancer formation or progression. Some of the genes were related to several cancer types, e.g. PTEN, which regulates the AKT/PKB signaling pathway and thereby represents a tumor suppressor in various tumors [28] was present in lists for BIC, THCA, and PRAD. Some genes were related to several clusters of the same cancer type, e.g. SOX2, which has been shown to control tumor initiation in squamous-cell carcinomas [24] was present in several clusters of HNSC patients. Other genes were only found for one specific cluster of a cancer type (e.g., among the four clusters of BIC, only one gene list included GRPR, a G-protein cou-

pled receptor that has been shown to influence the viability of breast cancer cells [36]). From a biological perspective, this finding is consistent with our expectation that the lists are not disjoint, since cancer does not form due to an individual mutation, but rather as a consequence of a number of different aberrations (see Section 2.1.1). These can be shared between different subtypes, e.g., if they happened early in the process of tumorigenesis, or even between different cancer types.

Besides examining the lists at the level of individual genes, we also inferred biological functions ascribed to the genes based on the GO annotation using GeneTrail2 [140]. We performed over-representation analysis (ORA) testing if significantly more genes in the list are associated with a specific category than would be expected based on a background gene list (see Section 2.5.3). All genes that entered the clustering were used as background and we tested for enrichment of KEGG pathways and gene ontology categories (GO).

Given that we detected some variation in the cluster assignment (see Section 5.4.2), a certain variation among the enriched terms can be expected. To quantify the stability of the result, we compared the identified terms of the best result to the terms of the median result, where "best" and "median" was defined according to the survival p-value. The similarity of the GO terms was determined using the relevance score [122], for the similarity between two sets of KEGG pathways we used the Jaccard index. The average similarities based on GO terms and KEGG pathways correlate with the observed stabilities of the cluster assignments: the value for LUAD is slightly lower than for BIC and HNSC (0.48 vs. 0.58 and 0.68, respectively). However, for all three cancer types, the scores indicate that similar functions or pathways are found. For LGG, the results were stable such that the best and median were equivalent, for PRAD and THCA, the best and median result could not be identified due to the lack of adequate survival data.

Similar to these two latter examples, there are many other biological scenarios in which the user is not able to identify the optimal result of an unsupervised problem. Therefore, the following analysis is based on the intersection of the best and median result, as this intersection contains terms that are found consistently over different runs. Table 5.4 and Table 5.5 show findings for HNSC for Cluster 3, which has the poorest prognosis (median survival time 1079 days), and for Cluster 4, which has the best prognosis (median survival time $>$ 4241 days), respectively. The terms are very different for the two clusters. In Cluster 3, the downregulated, enriched terms were mainly associated with muscle cells, and more specifically with their development or their activation. Skeletal muscle invasion has been reported to be correlated with the recurrence of the tumor [31]. This finding agrees with our observation that patients with a distinct signature in the genes involved

Table 5.4: Significant GO terms associated with high intra-cluster similarity identified by over-representation analysis for patient cluster 3 (median survival time 1079 days) for HNSC. The terms are based on inactive genes.

$k = 3$	GO - Cellular component (p-value $< 1.0E-3$)
	contractile fiber
	myofibril
	sarcomere
	I band
	myosin II complex
	muscle myosin complex
	A band
	myosin complex
	Z disc
	myofilament
	GO - Biological Process (p-value $< 1.0E-3$)
	myofibril assembly
	striated muscle cell development
	muscle filament sliding&actin myosin filament sliding
	striated muscle contraction
	actomyosin structure organization
	muscle cell development

in these functions have a poor survival.

For patient cluster 4, we identified terms related to the *phosphorylation of STAT protein*, which forms part of the JAK/STAT signaling pathway regulating cell growth, and has been shown to play a role in tumor formation and progression [135]. On the other hand, the identified terms mainly refer to functions of the immune system, including the *regulation of type I interferon mediated signaling pathway*, which is involved in the regulation of the innate immune response. Inflammation processes and the immune system seem to play an important yet controversial role in cancer [38], however, specific regulation of the immune system could have a positive effect on survival times in the respective patient group. For HNSC, the terms associated with high inter-cluster dissimilarity and high intra-cluster similarity differ only

Table 5.5: Significant GO terms associated with high intra-cluster similarity identified by over-representation analysis for patient cluster 4 (median survival time > 4241 days) for HNSC. ORA was performed using active genes.

$k = 4$	GO - Biological process (p-value $< 1.0E-3$)
	regulation of peptidyl serine phosphorylation of STAT protein
	positive regulation of peptidyl serine phosphorylation of STAT protein
	serine phosphorylation of STAT protein
	positive regulation of peptidyl serine phosphorylation
	natural killer cell activation involved in immune response
	response to exogenous dsRNA
	natural killer cell activation
	regulation of type I interferon mediated signaling pathway

for Cluster 1, where we identified terms, such as *B cell differentiation* and *proliferation*, that are associated with high dissimilarity to other clusters but not to high intra-cluster similarity.

We also applied our approach to patient groups of PRAD and THCA. These two cancer types share the common property that only relatively few patients die from the cancer, which renders a meaningful survival analysis difficult. For both cancer types, we set $K = 4$ and could identify several significantly enriched GO terms for the patient clusters. For PRAD, we identified, amongst others, a number of terms related to the activity of olfactory receptors, which have been shown to participate in the process of tumor cell proliferation and apoptosis [33]. Several hits for THCA, e.g., *DNA deamination*, suggest an influence of epigenetic regulatory mechanisms, which is in line with current reports in the literature [119].

5.5 Conclusion

The extension presented in this chapter provides a step toward improved interpretability of multiple kernel clustering. Extracting feature importances is only straightforward for a few kernel functions, leaving a vast number of kernels with limited usefulness for scenarios where this level of interpretability is essential. Our method closes this gap as we designed it with the objective of providing a framework that could be used with any type of kernel function.

We demonstrated the utility of our method in an exemplary study of diverse cancer types, where we could show that our method delivers state-of-the-art performance for survival analyses while providing biologically interpretable results. We developed the FIPPA score as an informative measure of feature cluster impact on patient clusterings. The FIPPA score enabled us to identify the underlying high impact feature groups that contributed to the formation of the respective patient cluster. The high impact features, filtered for homogeneity in the respective data type and patient cluster, were enriched for certain biological functions or pathways, which supports hypothesis generation concerning potential deregulations in the tumor cells of the respective patient group. The initial screening for overlap between the enriched terms and biological literature showed very promising results: for HNSC, the identified terms differed strongly between the two clusters with the best and the worst survival prognosis, respectively, hinting at the involvement of muscle cells and of the immune system in this particular type of cancer. Overall, our method provides impact information from the actual learning process instead of a retrospective analysis of the patient clusters. These results could lead to deeper insights on individual subtypes, which can be tested in suitably designed follow-up studies. Finally, in line with the literature, we found that each patient seems to represent a unique history of initial cancer formation and progression, which is reflected in the numerous patient samples that are placed in-between clusters.

In the current implementation of our method, the feature clusters influence the patient clusters but not vice versa. While these feature clusters provide valuable information, simultaneous optimization of both feature and patient clusters could lead to even more specific selections of features. In addition, the feature clusters are currently restricted to be non-overlapping and to cover the whole feature space. In many real-world scenarios, we might want a more flexible approach, that is, relaxing these constraints, e.g., to be able to exclude outlier features. This additional flexibility in the feature clustering process could be achieved, for example, by employing a fuzzy clustering approach.

Besides, fFIPPA combines, in the proposed version, the information on the feature clusters directly with the fuzzy cluster assignment of the patients, while the projected coordinates of the samples are ignored. Alternatively, an inspection of the individual patients would be possible such that a FIPPA vector is calculated for each patient indicating the impact of the used kernel matrices on the basis of the low-dimensional projection. For this purpose, the fuzziness in FIPPA would be dropped, since we would not consider cluster membership probabilities. This individual analysis would be particularly interesting in cases where highly different FIPPA vectors are derived for

patients that appear overall very similar to each other. In these cases, different molecular bases of the tumors might lead to different, clinically relevant characteristics, e.g., response to treatment. A similar approach would be an interactive application enhancing the interpretability of dimensionality reduction methods. This application could help to understand which factors are important by updating feature-group specific weights according to a user-defined movement of a sample. For instance, when moving a patient closer to a well-defined cluster, a decreased weight for one data type might indicate intrinsic differences between the two entities, possibly hinting at treatment options to overcome these differences. Such interactive scenarios could also be useful when partial additional information is available, that is, in combination with semi-supervised learning.

Chapter 6

Conclusions and outlook

6.1 Summary

Due to the large amount of biological measurements, it is has become possible to study complex diseases on many different levels, such as comparing differences in DNA methylation, gene expression, or copy number variation. In cancer, aberrations have been observed on different levels of the cell and markers of different data types are already used to guide treatment decisions. However, the currently established cancer subtypes with their respective markers do not cover the heterogeneity of the disease. Therefore, data integration approaches aim at developing a comprehensive understanding of the different facets of a tumor.

In this work, we focused on multiple kernel learning (MKL) to combine the available data types with the aim of identifying integrative cancer subtypes. MKL enables a flexible integration of qualitatively distinct data types, which is valuable in the field of bioinformatics, where, amongst others, numerical, categorical, and sequence data naturally occur. In MKL, one data type is commonly represented as one kernel matrix and the integration is performed as a weighted linear combination of these kernel matrices. We addressed three important issues of MKL, namely robustness, applicability, and interpretability.

We extended the multiple kernel graph embedding framework, which implements numerous dimensionality reduction schemes, by adding a constraint that regularizes the kernel weights. We showed that this restriction of the search space for the weights leads to increased robustness of the method with respect to perturbations in the training data (Section 3.3.1). The gained stability is beneficial since it also enables the analysis of smaller data sets or the use of a larger number of input kernels without impairment of the per-

formance. One use case, which is made possible by the increased stability, is to represent each data type by multiple kernel matrices, which enables an implicit parameter learning as the user does not need to know the optimal kernel function or parameters for each data type beforehand.

Despite the flexibility of the regularized multiple kernel graph embedding framework, one limitation we observed is the nonapplicability of principal component analysis. This widely used global dimensionality reduction scheme cannot be implemented in the framework due to an ill-posed generalized eigenvalue problem (Section 4.2). Moreover, the direct multiple kernel extension of principal component analysis results in kernel selection instead of kernel combination (Section 4.3). However, incorporating into the objective function the assumption that biological data types supplement each other, we were able to generate a method that performs data integration without abandoning the core of principal component analysis, which is the variance maximization. The good performance of this method in comparison to standard methods supports the biological assumption that was made.

Finally, a common drawback of kernel-based approaches is their lack of interpretability: after the implicit mapping into a potentially high- or even infinite-dimensional feature space, the feature contributions to the result cannot easily be sorted out. Therefore, we applied a feature clustering before the MKL approach, on the basis of which we could calculate patient cluster-specific feature importances. Our results showed that this provides additional means to interpret the result from the perspective of the method (Chapter 5) instead of performing a retrospective analysis only considering the final result.

From the biological perspective, an important factor that emerged in the analysis of several cancer types is the immune system and inflammation processes. In recent years, the functionality of the immune system has been of particular interest due to the research on cancer immunotherapy, a therapy stimulating the patient's immune system to fight tumor cells [168]. In our analysis, we found deregulations in immune system-related genes in glioblastoma (Section 3.3.1.6), and head and neck squamous cell carcinoma (Section 5.4.4). In both cases, the respective group of patients had a relatively good prognosis. However, for glioblastoma, the genes related to the immune response were underexpressed while they were overexpressed for head and neck cancer. This indicates the controversial role of the immune system in cancer: on the one hand, acute, temporary activation of the immune system protects against cancer development. In studies, immune-suppressed patients showed an increased risk for certain cancer types, such as viral-associated cancers [40]. On the other hand, chronic inflammation increases the risk of cancer development and might lead to a poor prognosis [40]. These different

aspects were used by Thorsson et al. [149], who identified six immune cancer subtypes based on published immune expression signatures. For multiple cancer types, these immune subtypes correlated to the survival times of the patients.

Summarizing, our results show that multiple kernel learning provides a useful and stable framework for the integration of various biological data types in unsupervised settings. Whereas some methods could be directly translated into a multiple kernel version, others required some adaptations to serve as a reasonable data integration method, as was the case for principal component analysis. Moreover, some common issues of kernel methods can be tackled using simple, known approaches, e.g., feature clustering for improved interpretability. Overall, our biological findings agreed with current literature. The developed methodologies can thus be used to generate new hypotheses potentially leading to new and clearly defined cancer subtypes.

6.2 Perspectives

The aim of this work was the development of methods that facilitate the identification of clinically relevant cancer subtypes. Despite the ever-increasing amount of molecular data for cancer patients, it would be misguided to believe that we have complete data sets. Therefore, extending the current methodologies such that they enable the incorporation of data types, even if measurements are missing for some patients, would greatly increase the number of utilizable samples and data types, thereby leading to more representative results. The recently proposed NEMO, a method using the uniformly weighted average kernel, is able to handle missing data types by calculating the average over each available data type for the respective samples and thus simply ignores the missing data type [116]. A similar approach could be adapted in our MKL scenario by preserving the relation of the weights for the available data types and setting the other weights to zero for the incomplete samples. To some extent, this corresponds to the use of per-sample weights, however, the analogy is limited since these introduced sample-specific kernel weights do not optimize an objective function but are only adapted in case of missing measurements. Therefore, the kernel weights will be more stable and less patient-specific compared to localized multiple kernel learning (e.g., as proposed by Gönen and Margolin [55] as an extension of k-means). Localized versions of our proposed MKL methods could combine both, the incorporation of incomplete data and the ability to learn sample-specific weights for each data type. However, as this implies an enormous increase in flexibility due to the increase in the number of learned parameters, a sensible regular-

ization of the parameters would be necessary.

In general, the approach of dividing the samples into strict subgroups might not be suitable for cancer patient data. Therefore, Chapter 5 started a transition from a binary cluster assignment toward considering the distribution of the cancer patients via fuzzy clustering. In this chapter, cluster probabilities were only used for the calculation of the feature impact scores, however, other aspects of the analysis could also benefit from this more fine-grained information. One such aspect would be the choice of the number of clusters, which was in this thesis largely performed using the average silhouette score. Using a fuzzy silhouette score would penalize samples less that have a low cluster membership probability. This approach might help to identify clusters that have a dense core, and are therefore biologically interesting, without being influenced strongly by outliers in the data set.

Thinking more globally, it could be beneficial to study multiple cancer types jointly, instead of each of them individually. Concatenating the data measured for the different cancer types (as done for instance by Taskesen et al. [142]) often leads to a clustering that is dominated by the tumor type [18]. Hoadley et al. [67] also obtained clustering results according to histology, tissue type, or anatomic origin for various individual data types as well as for their integrative analysis, which suggests high tissue specificity in the used data. However, some studies indicate partially high genetic similarity between different cancer types [64, 12]. Exploiting these partial correlations between tumor samples in different tissues, a pan-cancer analysis could be formulated as a multi-task problem, where each cancer type defines one task. Multi-task learning aims at improving the generalization performance for each individual task and has been extended in recent years from supervised to unsupervised application scenarios (see [173] for an overview of multi-task learning approaches). Transferred to the pan-cancer setting, the analysis of each cancer type could benefit from the additional, available data leading to increased power of the analysis. In analogy to the terms used for the integration of different data types, this approach would move the combination of patients with different cancer types from an early integration, namely concatenation as performed, e.g., in [142], to an intermediate integration.

In general, the approaches presented in this thesis as well as possible extensions could also provide valuable insights when applied to other complex diseases. For this purpose, interpretability plays an important role, which motivates the extension of the current methodology towards feature selection. Especially for poorly understood diseases with treatments that are limited in their efficacy, e.g., Alzheimer's disease [117], the exploratory analysis could help to generate new hypotheses concerning the molecular foundations of the disease and possibly hint to new drug targets.

Appendix A

List of publications

[1] B. Röder, N. Kersten, M. Herr, **N. K. Speicher**, and N. Pfeifer. webrMKL: A web server for dimensionality reduction and sample clustering of multi-view data based on unsupervised multiple kernel learning. *Nucleic Acids Research*, 2019. doi: 10.1093/nar/gkz422

[2] **N. K. Speicher** and N. Pfeifer. An interpretable multiple kernel learning approach for the discovery of integrative cancer subtypes. *arXiv*, 2018. URL `http://arxiv.org/abs/1811.08102`.

[3] **N. K. Speicher** and N. Pfeifer. Towards multiple kernel principal component analysis for integrative analysis of tumor samples. *Journal of Integrative Bioinformatics*, 14(2), 2017. doi: 10.1515/jib-2017-0019.

[4] **N. K. Speicher** and N. Pfeifer. Integrating different data types by regularized unsupervised multiple kernel learning with application to cancer subtype discovery. *Bioinformatics*, 31(12), 2015. doi: 10.1093/ bioinformatics/btv244.

[5] P. Sun, **N. K. Speicher**, R. Roettger, J. Guo and J. Baumbach. Bi-Force: Large-scale bicluster editing and its application to gene expression data biclustering. *Nucleic Acids Research*, 42(9), 2014. doi: 10.1093/ nar/gku201.

Appendix B

Licensing, copyright, and plagiarism prevention

B.1 Manuscripts

If applicable, license and copyright information for material reused for Chapters 3, 4, and 5 is listed below.

Regularization of unsupervised multiple kernel learning

The manuscript Speicher and Pfeifer [136] has been published in *Bioinformatics*. The article has been published under a Creative Commons license. This license grants the following rights:

> This article is available under the Creative Commons CC-BY-NC license and permits non-commercial use, distribution and reproduction in any medium, provided the original work is properly cited. (See link "Permissions" in the online version of the article.)

License: #4455330381284

Multiple kernel principal component analysis

The main parts of Section 4.3–4.5 have been published in Speicher and Pfeifer [137] in the *Journal of Integrative Bioinformatics*. This work is licensed under the Creative Commons Attribution-NonCommercial-NoDerivatives 3.0 License (cf. BY-NC-ND 3.0).

Increased interpretability of multiple kernel learning

The manuscript [138] describing the approach of Chapter 5 is publicly available as a preprint at arXiv: https://arxiv.org/abs/1811.08102.

B.2 Figure reprints

Table B.1: Licensing information for figure reprints

Fig.	License No.	Publisher	Source
2.1	4477730950179	Elsevier	Hanahan and Weinberg [59]
2.3	4462490879804	Mary Ann Liebert, Inc.	Pavlidis et al. [108]
3.9	4455320736016	Oxford University Press	Rappoport and Shamir [115]

B.3 Plagiarism prevention

The contents of each chapter of this thesis were screened for plagiarism using the software *iThenticate* on July, 18. The papers [136, 137, 138] mentioned in Section B.1 were excluded from the corpus. Table B.2 summarizes the results obtained using the standard settings of *iThenticate*.

Table B.2: Results of the plagiarism detection software *iThenticate* for each chapter of this thesis.

	Similarity index
Chapter 1	1%
Chapter 2	7%
Chapter 3	9%
Chapter 4	4%
Chapter 5	4%
Chapter 6	1%

Bibliography

[1] The Cancer Genome Atlas, Website, Available from: http://cancergenome.nih.gov/.

[2] Leitlinien für Diagnostik und Therapie in der Neurologie, Gliome, 2014. AWMF-Registernummer: 030/099, https://www.dgn.org/leitlinien/2977-ll-76-gliome [Accessed: 2018/11/27].

[3] Leitlinienprogramm Onkologie (Deutsche Krebsgesellschaft, Deutsche Krebshilfe, AWMF): S3 Leitlinie Früherkennung, Diagnose, Therapie und Nachsorge des Mammakarzinoms, Kurzversion 4.1, 2018. AWMF Registernummer: 032-0450L http://leitlinienprogramm-onkologie.de/leitlinien/mammakarzinom [Accessed: 2018/11/27].

[4] Leitlinienprogramm Onkologie (Deutsche Krebsgesellschaft, Deutsche Krebshilfe, AWMF): S3 Leitlinie Prävention, Diagnostik, Therapie und Nachsorge des Lungenkarzinoms, Langversion 1.0, 2018. AWMF-Registernummer: 020/007OL, http://leitlinienprogramm-onkologie.de/Lungenkarzinom.98.0.html [Accessed: 2018/11/27].

[5] Leitlinienprogramm Onkologie (Deutsche Krebsgesellschaft, Deutsche Krebshilfe, AWMF): Interdisziplinäre Leitlinie der Qualität S3 zur Früherkennung, Diagnose und Therapie der verschiedenen Stadien des Prostatakarzinoms, Kurzversion 5.0, 2018. AWMF Registernummer: 043/022OL http://leitlinienprogramm-onkologie.de/Prostatakarzinom.58.0.html [Accessed: 2018/11/27].

[6] H. Abdi and L. J. Williams. Principal component analysis. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(4):433–459, Jul 2010. doi: 10.1002/wics.101.

[7] C. C. Aggarwal, A. Hinneburg, and D. A. Keim. On the surprising behavior of distance metrics in high dimensional space. In J. Van den Bussche and V. Vianu, editors, *Database Theory — ICDT 2001. Lecture Notes in Computer Science, vol 1973*. Springer, 2001. doi: 10.1007/3-540-44503-X_27.

[8] M. B. Amin, D. M. Gress, L. R. Meyer Vega, S. B. Edge, et al. *AJCC Cancer Staging Manual, Eighth Edition*. American College of Surgeons, 8 edition, 2018.

[9] R. Argelaguet, B. Velten, D. Arnol, S. Dietrich, et al. Multi-Omics Factor Analysis-a framework for unsupervised integration of multi-omics data sets. *Molecular systems biology*, 14(6):e8124, Jun 2018. doi: 10.15252/MSB.20178124.

[10] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, et al. Gene Ontology: tool for the unification of biology. *Nature Genetics*, 25(1):25–29, May 2000. doi: 10.1038/75556.

[11] Y. Assenov, F. Müller, P. Lutsik, J. Walter, et al. Comprehensive analysis of DNA methylation data with RnBeads. *Nature Methods*, 11(11):1138–1140, 2014. doi: 10.1038/nmeth.3115.

[12] M. H. Bailey, C. Tokheim, E. Porta-Pardo, S. Sengupta, et al. Comprehensive characterization of cancer driver genes and mutations. *Cell*, 173(2): 371–385.e18, Apr 2018. doi: 10.1016/J.CELL.2018.02.060.

[13] Y. Barash and N. Friedman. Context-specific bayesian clustering for gene expression data. *Journal of Computational Biology*, 9:169–191, 2002. doi: 10.1089/10665270252935403.

[14] E. Barbarotto, T. D. Schmittgen, and G. A. Calin. MicroRNAs and cancer: Profile, profile, profile. *International Journal of Cancer*, 122(5):969–977, Mar 2008. doi: 10.1002/ijc.23343.

[15] S. B. Baylin, M. Esteller, M. R. Rountree, K. E. Bachman, et al. Aberrant patterns of DNA methylation, chromatin formation and gene expression in cancer. *Human Molecular Genetics*, 10(7):687–692, Apr 2001. doi: 10.1093/hmg/10.7.687.

[16] M. Belkin and P. Niyogi. Laplacian eigenmaps and spectral techniques for embedding and clustering. In *Advances in Neural Information Processing Systems 14*, pages 585–591, 2002. doi: 10.7551/mitpress/1120.003.0080.

[17] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300, 1995. doi: 10.1111/j.2517-6161.1995.tb02031.x.

[18] A. C. Berger, A. Korkut, R. S. Kanchi, A. M. Hegde, et al. A comprehensive pan-cancer molecular study of gynecologic and breast cancers. *Cancer Cell*, 33(4):690–705.e9, Apr 2018. doi: 10.1016/J.CCELL.2018.03.014.

[19] B. E. Bernstein, A. Meissner, and E. S. Lander. The mammalian epigenome. *Cell*, 128(4):669–681, Feb 2007. doi: 10.1016/J.CELL.2007.01.033.

[20] R. Beroukhim, C. H. Mermel, D. Porter, G. Wei, et al. The landscape of somatic copy-number alteration across human cancers. *Nature*, 463(7283): 899–905, Feb 2010. doi: 10.1038/nature08822.

[21] J. C. Bezdek. *Pattern recognition with fuzzy objective function algorithms*. Plenum Press, 1981. doi: 10.1007/978-1-4757-0450-1.

[22] A. Bird. DNA methylation patterns and epigenetic memory. *Genes & Development*, 16:6–21, 2002. doi: 10.1101/gad.947102.

[23] C. M. Bishop. *Pattern recognition and machine learning*. Springer Science+Business Media, LLC, 2006. doi: 10.1016/c2009-0-22409-3.

[24] S. Boumahdi, G. Driessens, G. Lapouge, S. Rorive, et al. SOX2 controls tumour initiation and cancer stem-cell functions in squamous-cell carcinoma. *Nature*, 511(7508):246–250, 2014. doi: 10.1038/nature13305.

[25] M. D. Buhmann. *Radial basis functions : theory and implementations*. Cambridge University Press, 2003. doi: 10.1017/cbo9780511543241.

[26] X. Cai, F. Nie, and H. Huang. Multi-view k-means clustering on big data. In *Proceedings of the Twenty-Third international joint conference on Artificial Intelligence*, pages 2598–2604, Beijing, China, 2013. AAAI Press.

[27] R. Cangelosi and A. Goriely. Component retention in principal component analysis with application to cDNA microarray data. *Biology direct*, 2:2, Jan 2007. doi: 10.1186/1745-6150-2-2.

[28] L. C. Cantley and B. G. Neel. New insights into tumor suppression: PTEN suppresses tumor formation by restraining the phosphoinositide 3-kinase/AKT pathway. *Proceedings of the National Academy of Sciences of the United States of America*, 96(8):4240–5, Apr 1999. doi: 10.1073/PNAS.96.8.4240.

[29] L. Carter, D. G. Rothwell, B. Mesquita, C. Smowton, et al. Molecular analysis of circulating tumor cells identifies distinct copy-number profiles in patients with chemosensitive and chemorefractory small-cell lung cancer. *Nature Medicine*, 23(1):114–119, Jan 2017. doi: 10.1038/nm.4239.

[30] E. Chan, D. E. Prado, and J. B. Weidhaas. Cancer microRNAs: from subtype profiling to predictors of response to therapy. *Trends in molecular medicine*, 17(5):235–43, May 2011. doi: 10.1016/j.molmed.2011.01.008.

[31] K. Chandler, C. Vance, S. Budnick, and S. Muller. Muscle invasion in oral tongue squamous cell carcinoma as a predictor of nodal status and local recurrence: just as effective as depth of invasion? *Head and neck pathology*, 5(4):359–63, Dec 2011. doi: 10.1007/s12105-011-0296-5.

[32] K. Chaudhary, O. B. Poirion, L. Lu, and L. X. Garmire. Deep learning-based multi-omics integration robustly predicts survival in liver cancer. *Clinical cancer research*, 24(6):1248–1259, Mar 2018. doi: 10.1158/1078-0432.CCR-17-0853.

[33] Z. Chen, H. Zhao, N. Fu, and L. Chen. The diversified function and potential therapy of ectopic olfactory receptors in non-olfactory tissues. *Journal of Cellular Physiology*, 233(3):2104–2115, Mar 2018. doi: 10.1002/jcp.25929.

[34] C. H. Chung, J. S. Parker, G. Karaca, J. Wu, et al. Molecular classification of

head and neck squamous cell carcinomas using patterns of gene expression. *Cancer Cell*, 5(5):489–500, May 2004. doi: 10.1016/S1535-6108(04)00112-6.

[35] G. M. Cooper. *The cell: A molecular approach*. Sinauer Associates, 2nd edition, 2000.

[36] D. B. Cornelio, C. B. DE Farias, D. S. Prusch, T. E. Heinen, et al. Influence of GRPR and BDNF/TrkB signaling on the viability of breast and gynecologic cancer cells. *Molecular and clinical oncology*, 1(1):148–152, Jan 2013. doi: 10.3892/mco.2012.7.

[37] C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20 (3):273–297, Sep 1995. doi: 10.1007/BF00994018.

[38] L. M. Coussens and Z. Werb. Inflammation and cancer. *Nature*, 420(6917): 860–7, 2002. doi: 10.1038/nature01322.

[39] N. Cristianini, A. Elisseeff, J. Shawe-Taylor, and J. Kandola. On kernel-target alignment. In T.G. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems 14 (NIPS 2001)*, 2001. doi: 10.7551/mitpress/1120.003.0052.

[40] K. E. de Visser, A. Eichten, and L. M. Coussens. Paradoxical roles of the immune system during cancer development. *Nature Reviews Cancer*, 6:24–37, 2006. doi: 10.1038/nrc1782.

[41] C. Ding, T. Li, W. Peng, and H. Park. Orthogonal nonnegative matrix t-factorizations for clustering. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 126–135. ACM Press, 2006. doi: 10.1145/1150402.1150420.

[42] S. Drăghici, P. Khatri, R. P. Martins, G. Ostermeier, and S. A. Krawetz. Global functional profiling of gene expression. *Genomics*, 81(2):98–104, Feb 2003. doi: 10.1016/S0888-7543(02)00021-6.

[43] L. Du, P. Zhou, L. Shi, H. Wang, et al. Robust multiple kernel K-means using $l_{2,1}$-norm. In *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence (IJCAI)*, pages 3476–3482, 2015.

[44] J. C. Dunn. A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters. *Journal of Cybernetics*, 3(3):32–57, 1974. doi: 10.1080/01969727308546046.

[45] A. Eden, F. Gaudet, A. Waghmare, and R. Jaenisch. Chromosomal instability and tumors promoted by DNA hypomethylation. *Science*, 300(5618): 455, Apr 2003. doi: 10.1126/science.1083557.

[46] M. Esteller. Epigenetics in cancer. *New England Journal of Medicine*, 358 (11):1148–1159, 2008. doi: 10.1056/nejmra072067.

[47] V. Farewell and T. Johnson. Major Greenwood (1880-1949): a biographical

and bibliographical study. *Statistics in Medicine*, 35(5):645–70, Feb 2016. doi: 10.1002/sim.6772.

[48] S. L. Floor, J. E. Dumont, C. Maenhaut, and E. Raspe. Hallmarks of cancer: of all cancer cells, all the time? *Trends in Molecular Medicine*, 18(9):509–515, Sep 2012. doi: 10.1016/J.MOLMED.2012.06.005.

[49] D. Ford, D. Easton, M. Stratton, S. Narod, et al. Genetic heterogeneity and penetrance analysis of the BRCA1 and BRCA2 genes in breast cancer families. *The American Journal of Human Genetics*, 62(3):676–689, Mar 1998. doi: 10.1086/301749.

[50] M. E. Garber, O. G. Troyanskaya, K. Schluens, S. Petersen, et al. Diversity of gene expression in adenocarcinoma of the lung. *Proceedings of the National Academy of Sciences*, 98(24):13784–13789, 2001. doi: 10.1073/pnas.241500798.

[51] T. Gärtner, P. A. Flach, A. Kowalczyk, and A. J. Smola. Multi-instance kernels. In *Proc. 19th International Conf. on Machine Learning*, pages 179–186. Morgan Kaufmann, 2002.

[52] V. Gligorijević, N. Malod-Dognin, and N. Pržulj. Patient-specific data fusion for cancer stratification and personalised treatment. In *Proccedings of the Pacific Symposium on Biocomputing 2016*, pages 321–332. World Scientific Publishing Company, Jan 2016. doi: 10.1142/9789814749411_0030.

[53] M. Goldman, B. Craft, J. Zhu, and D. Haussler. The UCSC Xena system for cancer genomics data visualization and interpretation [abstract]. *Proceedings of the American Association for Cancer Research Annual Meeting 2017; Cancer Research*, 77(13 Supplement):2584–2584, Jul 2017. doi: 10.1158/1538-7445.AM2017-2584. Available from http://xena.ucsc.edu [Accessed: 2017/10/24].

[54] M. Gönen and E. Alpaydın. Multiple kernel learning algorithms. *Journal of Machine Learning Research*, 12:2211–2268, 2011. ISSN 1532-4435.

[55] M. Gönen and A. A. Margolin. Localized data fusion for kernel k-means clustering with application to cancer biology. In *Advances in Neural Information Processing Systems 27 (NIPS 2014)*, pages 1305–1313, 2014.

[56] L. A. M. Gravendeel, M. C. M. Kouwenhoven, O. Gevaert, J. J. de Rooi, et al. Intrinsic gene expression profiles of gliomas are a better predictor of survival than histology. *Cancer Research*, 69(23):9065–72, Dec 2009. doi: 10.1158/0008-5472.CAN-09-2307.

[57] Y. Guo. A weighted cluster kernel PCA prediction model for multi-subject brain imaging data. *Statistics and Its Interface*, 3(1):103–112, 2010. doi: 10.4310/sii.2010.v3.n1.a9.

[58] D. Hanahan and R. A. Weinberg. The hallmarks of cancer. *Cell*, 100(1):57–70, Jan 2000. doi: 10.1016/S0092-8674(00)81683-9.

[59] D. Hanahan and R. A. Weinberg. Hallmarks of cancer: The next generation. *Cell*, 144(5):646 – 674, Mar. 2011. doi: 10.1016/j.cell.2011.02.013.

[60] J. A. Hartigan. Direct clustering of a data matrix. *Journal of the American Statistical Association*, 67(337):123–129, 1972. doi: 10.1080/01621459.1972.10481214.

[61] J. A. Hartigan. *Clustering algorithms*. John Wiley & Sons, 1975.

[62] T. Hastie, R. Tibshirani, and J. Friedman. *The elements of statistical learning: Data mining, inference, and prediction*. Springer Series in Statistics, 2009. doi: 10.1007/978-0-387-84858-7_2.

[63] X. He and P. Niyogi. Locality preserving projections. In S. Thrun, L. Saul, and B. Schölkopf, editors, *Advances in Neural Information Processing Systems 16*, pages 153–160. MIT Press, 2004.

[64] D. Heim, J. Budczies, A. Stenzinger, D. Treue, et al. Cancer beyond organ and tissue specificity: Next-generation-sequencing gene mutation data reveal complex genetic similarities across major cancers. *International Journal of Cancer*, 135:2365–2369, 2014. doi: 10.1002/ijc.28882.

[65] H. Hermeking. The miR-34 family in cancer and apoptosis. *Cell Death & Differentiation*, 17:193–199, May 2009. doi: 10.1038/cdd.2009.56.

[66] G. E. Hinton and R. R. Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science (New York, N.Y.)*, 313(5786):504–507, jul 2006. doi: 10.1126/science.1127647.

[67] K. A. Hoadley, C. Yau, T. Hinoue, D. M. Wolf, et al. Cell-of-origin patterns dominate the molecular classification of 10,000 tumors from 33 types of cancer. *Cell*, 173(2):291–304.e6, Apr 2018. doi: 10.1016/J.CELL.2018.03.022.

[68] D. W. Hosmer, Jr., S. Lemeshow, and S. May. *Applied survival analysis: regression modeling of time to event data*. Wiley, 2011.

[69] H. Hotelling. Relations between two sets of variates. *Biometrika*, 28(3/4): 321–377, Dec 1936. doi: 10.2307/2333955.

[70] H.-C. Huang, Y.-Y. Chuang, and C.-S. Chen. Multiple kernel fuzzy clustering. *IEEE Transactions on Fuzzy Systems*, 20(1):120–134, 2012. doi: 10.1109/tfuzz.2011.2170175.

[71] H.-C. Huang, Y.-Y. Chuang, and C.-S. Chen. Affinity aggregation for spectral clustering. In *Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 773–780. IEEE Computer Society, 2012. doi: 10.1109/cvpr.2012.6247748.

[72] International Agency for Research on Cancer (IARC). Cancer Tomorrow, 2018. URL `http://gco.iarc.fr/tomorrow/`.

[73] International Agency for Research on Cancer (IARC). Cancer Today, 2018. URL `http://gco.iarc.fr/today/online-analysis-pie`.

[74] J.-P. J. Issa. DNA methylation as a therapeutic target in cancer. *Clinical cancer research : an official journal of the American Association for Cancer Research*, 13(6):1634–7, Mar 2007. doi: 10.1158/1078-0432.CCR-06-2076.

[75] J. L. Jennings and T. J. Hudson. Abstract 2988: International cancer genome consortium (ICGC)[abstract]. *Proceedings of the 106th Annual Meeting of the American Association for Cancer Research*, 75(15 Supplement):2988–2988, Apr 2015.

[76] W. Jiang and F.-l. Chung. A trace ratio maximization approach to multiple kernel-based dimensionality reduction. *Neural Networks*, 49:96–106, Jan 2014. doi: 10.1016/j.neunet.2013.09.004.

[77] E. L. Kaplan and P. Meier. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53(282):457–481, 1958.

[78] P. Kirk, J. E. Griffin, R. S. Savage, Z. Ghahramani, and D. L. Wild. Bayesian correlated clustering to integrate multiple datasets. *Bioinformatics*, 28(24): 3290–3297, Dec 2012. doi: 10.1093/bioinformatics/bts595.

[79] M. Kloft. *l_p-norm multiple kernel learning*. PhD thesis, Technische Universitaet Berlin, 2011.

[80] W. H. Kruskal and W. A. Wallis. Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association*, 47(260):583–621, Dec 1952. doi: 10.2307/2280779.

[81] A. Kumar, P. Rai, and H. Daume. Co-regularized multi-view spectral clustering. In J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 24*, pages 1413–1421. Curran Associates, Inc., 2011.

[82] J. T. Kwok and I. W. Tsang. Learning with idealized kernels. In *Proceedings of the 20th International Conference on Machine Learning*, pages 400–407, Washington, USA, 2003.

[83] M. J. Landrum, J. M. Lee, M. Benson, G. Brown, et al. ClinVar: public archive of interpretations of clinically relevant variants. *Nucleic Acids Research*, 44(D1):D862–D868, Jan 2016. doi: 10.1093/nar/gkv1222.

[84] A. R. Lanfranco, A. E. Castellanos, J. P. Desai, and W. C. Meyers. Robotic surgery: a current perspective. *Annals of surgery*, 239(1):14–21, Jan 2004. doi: 10.1097/01.sla.0000103020.19595.7d.

[85] T. Le Van, M. van Leeuwen, A. Carolina Fierro, D. De Maeyer, et al. Simultaneous discovery of cancer subtypes and subtype features by molecular data integration. *Bioinformatics*, 32(17):i445–i454, 2016. doi: 10.1093/bioinformatics/btw434.

[86] D. D. Lee and H. S. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401:788–791, Oct 1999. doi: 10.1038/44565.

[87] K. W. K. Lee and Z. Pausova. Cigarette smoking and DNA methylation. *Frontiers in genetics*, 4:132, 2013. doi: 10.3389/fgene.2013.00132.

[88] C. Leslie, E. Eskin, and W. S. Noble. The spectrum kernel: A string kernel for SVM protein classification. In *Pacific Symposium on Biocomputing 7*, pages 566–575. World Scientific Publishing, 2002. doi: 10.1142/9789812799623_0053.

[89] M. K. K. Leung, H. Y. Xiong, L. J. Lee, and B. J. Frey. Deep learning of the tissue-regulated splicing code. *Bioinformatics*, 30(12):i121–129, Jun 2014. doi: 10.1093/bioinformatics/btu277.

[90] S. Li, B. Liu, and C. Zhang. Regularized embedded multiple kernel dimensionality reduction for mine signal processing. *Computational Intelligence and Neuroscience*, 2016:1–12, 2016. doi: 10.1155/2016/4920670.

[91] M. Liang, Z. Li, T. Chen, and J. Zeng. Integrative data analysis of multi-platform cancer data with a multimodal deep learning approach. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 12(4):928–937, Jul 2015. doi: 10.1109/TCBB.2014.2377729.

[92] Y.-Y. Lin, T.-L. Liu, and C.-S. Fuh. Multiple kernel learning for dimensionality reduction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(6):1147–1160, 2011. doi: 10.1109/tpami.2010.183.

[93] J. Liu, C. Wang, J. Gao, and J. Han. Multi-view clustering via joint nonnegative matrix factorization. In *Proceedings of the 2013 SIAM International Conference on Data Mining*, pages 252–260. Society for Industrial and Applied Mathematics, Philadelphia, PA, May 2013. doi: 10.1137/1.9781611972832.28.

[94] E. F. Lock, K. A. Hoadley, J. S. Marron, and A. B. Nobel. Joint and individual variation explained (jive) for integrated analysis of multiple data types. *The Annals of Applied Statistics*, 7(1):523–542, 2013. doi: 10.1214/12-aoas597.

[95] H. Lodish, A. Berk, S. L. Zipursky, P. Matsudaira, et al. Molecular cell biology. chapter 24. W. H. Freeman, New York, 4th edition, 2000.

[96] J. Löfberg. Yalmip : A toolbox for modeling and optimization in matlab. In *In Proceedings of the CACSD Conference*, Taipei, Taiwan, 2004. doi: 10.1109/CACSD.2004.1393890.

[97] F. D. Mairinger, S. Ting, R. Werner, R. F. H. Walter, et al. Different micro-RNA expression profiles distinguish subtypes of neuroendocrine tumors of the lung: results of a profiling study. *Modern Pathology*, 27(12):1632–1640, Dec 2014. doi: 10.1038/modpathol.2014.74.

[98] J. Mariette and N. Villa-Vialaneix. Unsupervised multiple kernel learning for heterogeneous data integration. *Bioinformatics*, 34(6):1009–1015, Mar 2018. doi: 10.1093/bioinformatics/btx682.

[99] C. E. Metz. Basic principles of ROC analysis. *Seminars in Nuclear Medicine*, 8(4):283–298, 1978. doi: 10.1016/s0001-2998(78)80014-2.

[100] Q. Mo, S. Wang, V. E. Seshan, A. B. Olshen, et al. Pattern discovery and cancer gene identification in integrated cancer genomic data. *Proceedings of the National Academy of Sciences*, 110(11):4245–4250, Mar 2013. doi: 10.1073/pnas.1208949110.

[101] Q. Mo, R. Shen, C. Guo, M. Vannucci, et al. A fully Bayesian latent variable model for integrative clustering analysis of multi-type omics data. *Biostatistics*, 19(1):71–86, Jan 2018. doi: 10.1093/biostatistics/kxx017.

[102] S. Monti, P. Tamayo, J. Mesirov, and T. Golub. Consensus clustering – a resampling-based method for class discovery and visualization of gene expression microarray data. In *Machine Learning, Functional Genomics Special Issue*, pages 91–118, 2003.

[103] S. Negrini, V. G. Gorgoulis, and T. D. Halazonetis. Genomic instability — an evolving hallmark of cancer. *Nature Reviews Molecular Cell Biology*, 11 (3):220–228, Mar 2010. doi: 10.1038/nrm2858.

[104] H. Noushmehr, D. J. Weisenberger, K. Diefes, H. S. Phillips, et al. Identification of a CpG island methylator phenotype that defines a distinct subgroup of glioma. *Cancer Cell*, 17(5):510 – 522, May 2010. doi: 10.1016/s1040-1741(10)79529-4.

[105] Z. Obermeyer and E. J. Emanuel. Predicting the future - big data, machine learning, and clinical medicine. *The New England Journal of Medicine*, 375 (13):1216–1219, Sep 2016. doi: 10.1056/NEJMp1606181.

[106] J. S. Parker, M. Mullins, M. C. U. Cheang, S. Leung, et al. Supervised risk predictor of breast cancer based on intrinsic subtypes. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology*, 27 (8):1160–7, Mar 2009. doi: 10.1200/JCO.2008.18.1370.

[107] M. A. Patel, J. E. Kim, J. Ruzevick, G. Li, and M. Lim. The future of glioblastoma therapy: Synergism of standard of care and immunotherapy. *Cancers*, 6(4):1953–1987, Jan 2014. doi: 10.3390/cancers6041953.

[108] P. Pavlidis, J. Weston, J. Cai, and W. S. Noble. Learning Gene Functional Classifications from Multiple Data Types. *Journal of Computational Biology*, 9(2):401–411, 2002. doi: 10.1089/10665270252935539.

[109] K. Pearson. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine and Journal of Science*, 2:559–572, 1901. doi: 10.1080/14786440109462720.

[110] C. M. Perou, T. Sørlie, M. B. Eisen, M. van de Rijn, et al. Molecular portraits of human breast tumours. *Nature*, 406(6797):747–752, Aug 2000. doi: 10.1038/35021093.

[111] R. J. Pietras and D. C. Márquez-Garbán. Membrane-associated estrogen receptor signaling pathways in human cancers. *Clinical Cancer Research*, 13 (16):4672–4676, Aug 2007. doi: 10.1158/1078-0432.CCR-07-1373.

[112] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2017. URL `https://www.R-project.org/`.

[113] A. Rahimi and M. Gönen. Discriminating early- and late-stage cancers using multiple kernel learning on gene sets. *Bioinformatics*, 34(13):i412–i421, Jul 2018. doi: 10.1093/bioinformatics/bty239.

[114] W. M. Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 66(336):847–850, Dec 1971. doi: 10.2307/2284239.

[115] N. Rappoport and R. Shamir. Multi-omic and multi-view clustering algorithms: review and cancer benchmark. *Nucleic Acids Research*, 46(20): 10546–10562, Nov 2018. doi: 10.1093/nar/gky889.

[116] N. Rappoport and R. Shamir. NEMO: Cancer subtyping by integration of partial multi-omic data. *Bioinformatics*, Jan 2019. doi: 10.1093/bioinformatics/btz058.

[117] C. Reitz. Toward precision medicine in Alzheimer's disease. *Annals of translational medicine*, 4(6):107, Mar 2016. doi: 10.21037/atm.2016.03.05.

[118] S. Ren, P. Ling, M. Yang, Y. Ni, and Z. Song. Multi-kernel PCA with discriminant manifold for hoist monitoring. *Journal of Applied Sciences*, 13 (20):4195–4200, Dec 2013. doi: 10.3923/jas.2013.4195.4200.

[119] S. Rodríguez-Rodero, E. Delgado-Álvarez, L. Díaz-Naya, A. Martín Nieto, and E. Menéndez Torre. Epigenetic modulators of thyroid cancer. *Endocrinología, Diabetes y Nutrición*, 64(1):44–56, Jan 2017. doi: 10.1016/J.ENDINU.2016.09.006.

[120] J. Ronen, S. Hayat, and A. Akalin. Evaluation of colorectal cancer subtypes and cell lines using deep learning. *bioRxiv*, 2018. doi: 10.1101/464743.

[121] P. J. Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65, Nov 1987. doi: 10.1016/0377-0427(87)90125-7.

[122] A. Schlicker, F. S. Domingues, J. Rahnenführer, and T. Lengauer. A new measure for functional similarity of gene products based on Gene Ontology. *BMC Bioinformatics*, 7(1):302, Jun 2006. doi: 10.1186/1471-2105-7-302.

[123] J. Schmidhuber. Deep learning in neural networks: An overview. *Neural Networks*, 61:85–117, Jan 2015. doi: 10.1016/J.NEUNET.2014.09.003.

[124] L. Schneider, T. Kehl, K. Thedinga, N. L. Grammes, et al. ClinOmicsTrail^{bc}:

a visual analytics tool for breast cancer treatment stratification. *Bioinformatics*, Apr 2019. doi: 10.1093/bioinformatics/btz302.

[125] B. Schölkopf and A. J. Smola. *Learning with Kernels : Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, 2001.

[126] B. Schölkopf, A. Smola, and K.-R. Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, Jul 1998. doi: 10.1162/089976698300017467.

[127] B. Schölkopf, A. J. Smola, and K.-R. Müller. Advances in kernel methods. chapter Kernel Principal Component Analysis, pages 327–352. MIT Press, 1999.

[128] J. Shawe-Taylor and N. Cristianini. *Kernel methods for pattern analysis*. Cambridge University Press, 2004.

[129] R. Shen, A. B. Olshen, and M. Ladanyi. Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. *Bioinformatics*, 2009. doi: 10.1093/bioinformatics/btp543.

[130] R. Shen, Q. Mo, N. Schultz, V. E. Seshan, et al. Integrative subtype discovery in glioblastoma using iCluster. *PloS ONE*, 7, Apr 2012. doi: 10.1371/journal.pone.0035236.

[131] R. Shen, S. Wang, and Q. Mo. Sparse integrative clustering of multiple omics data sets. *The annals of applied statistics*, 7(1):269–294, Apr 2013. doi: 10.1214/12-AOAS578.

[132] N. Shervashidze, P. Schweitzer, E. J. van Leeuwen, K. Mehlhorn, and K. M. Borgwardt. Weisfeiler-Lehman graph kernels. *Journal of Machine Learning Research*, 12(Sep):2539–2561, 2011.

[133] A. P. Singh and G. J. Gordon. Relational learning via collective matrix factorization. In *Proceeding of the 14th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 650–658, Las Vegas, Nevada, USA, 2008. ACM Press. doi: 10.1145/1401890.1401969.

[134] J. A. Sinnott and T. Cai. Pathway aggregation for survival prediction via multiple kernel learning. *Statistics in Medicine*, 2018. doi: 10.1002/sim.7681.

[135] J. I. Song and J. R. Grandis. STAT signaling in head and neck cancer. *Oncogene*, 19(21):2489–2495, May 2000. doi: 10.1038/sj.onc.1203483.

[136] N. K. Speicher and N. Pfeifer. Integrating different data types by regularized unsupervised multiple kernel learning with application to cancer subtype discovery. *Bioinformatics*, 31(12):i268–i275, Jun 2015. doi: 10.1093/bioinformatics/btv244.

[137] N. K. Speicher and N. Pfeifer. Towards multiple kernel principal component

analysis for integrative analysis of tumor samples. *Journal of Integrative Bioinformatics*, 14(2), 2017. doi: 10.1515/jib-2017-0019.

[138] N. K. Speicher and N. Pfeifer. An interpretable multiple kernel learning approach for the discovery of integrative cancer subtypes. Nov 2018. URL `http://arxiv.org/abs/1811.08102`.

[139] N. Srivastava and R. Salakhutdinov. Multimodal learning with deep Boltzmann machines. *Journal of Machine Learning Research*, 15:2949–2980, 2014.

[140] D. Stöckel, T. Kehl, P. Trampert, L. Schneider, et al. Multi-omics enrichment analysis using the GeneTrail2 web service. *Bioinformatics*, 32(10):1502–1508, May 2016. doi: 10.1093/bioinformatics/btv770.

[141] S. Stringhini and I. Guessous. The shift from heart disease to cancer as the leading cause of death in high-income countries: A social epidemiology perspective. *Annals of Internal Medicine*, Nov 2018. doi: 10.7326/M18-2826.

[142] E. Taskesen, S. M. H. Huisman, A. Mahfouz, J. H. Krijthe, et al. Pan-cancer subtyping in a 2D-map shows substructures that are driven by specific combinations of molecular characteristics. *Scientific Reports*, 6, Apr 2016. doi: 10.1038/srep24949.

[143] A. Tenenhaus and M. Tenenhaus. Regularized generalized canonical correlation analysis. *Psychometrika*, 76(2):257–284, Apr 2011. doi: 10.1007/s11336-011-9206-8.

[144] A. Tenenhaus, C. Philippe, V. Guillemot, K.-A. Le Cao, et al. Variable selection for generalized canonical correlation analysis. *Biostatistics*, 15(3):569–583, Jul 2014. doi: 10.1093/biostatistics/kxu001.

[145] A. E. Teschendorff, F. Marabita, M. Lechner, T. Bartlett, et al. A beta-mixture quantile normalization method for correcting probe design bias in Illumina Infinium 450 k DNA methylation data. *Bioinformatics*, 29(2):189–196, Jan 2013. doi: 10.1093/bioinformatics/bts680.

[146] The Cancer Genome Atlas. Comprehensive molecular portraits of human breast tumors. *Nature*, 490:61–70, Oct 2012. doi: 10.1038/nature11412.

[147] The MathWorks, Inc. *Matlab 2015b*. Natick, Massachusetts, United States. URL `https://www.mathworks.com/`.

[148] J. J. Thiagarajan, K. N. Ramamurthy, and A. Spanias. Multiple kernel sparse representations for supervised and unsupervised learning. *IEEE Transactions on Image Processing*, 23(7):2905–2915, Jul 2014. doi: 10.1109/TIP.2014.2322938.

[149] V. Thorsson, D. L. Gibbs, S. D. Brown, D. Wolf, et al. The Immune Landscape of Cancer. *Immunity*, 48(4):812–830.e14, Apr 2018. doi: 10.1016/J.IMMUNI.2018.03.023.

[150] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal*

of the Royal Statistical Society. Series B, 58(1):267–288, Jan 1996. doi: 10.1111/j.2517-6161.1996.tb02080.x.

[151] M. Toyota, N. Ahuja, M. Ohe-Toyota, J. G. Herman, et al. CpG island methylator phenotype in colorectal cancer. *Proceedings of the National Academy of Sciences of the United States of America*, 96(15):8681–8686, 1999.

[152] M. J. van de Vijver, Y. D. He, L. J. van 't Veer, H. Dai, et al. A gene-expression signature as a predictor of survival in breast cancer. *New England Journal of Medicine*, 347(25):1999–2009, Dec 2002. doi: 10.1056/NEJMoa021967.

[153] L. Van Der Maaten and G. Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605, 2008.

[154] R. G. W. Verhaak, K. A. Hoadley, E. Purdom, V. Wang, et al. Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. *Cancer Cell*, 17(1):98–110, Jan 2010. doi: 10.1016/j.ccr.2009.12.020.

[155] M. Verma. Personalized medicine and cancer. *Journal of personalized medicine*, 2(1):1–14, Mar 2012. doi: 10.3390/jpm2010001.

[156] S. Vishwanathan, N. N. Schraudolph, R. Kondor, and K. M. Borgwardt. Graph kernels. *Journal of Machine Learning Research*, 11(Apr):1201–1242, 2010.

[157] U. von Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17(4):395–416, Dec 2007. doi: 10.1007/s11222-007-9033-z.

[158] B. Wang, A. M. Mezlini, F. Demir, M. Fiume, et al. Similarity network fusion for aggregating data types on a genomic scale. *Nature Methods*, 11 (3):333–337, 2014. doi: 10.1038/nmeth.2810.

[159] L. Wasserman. *All of statistics : a concise course in statistical inference*. Springer, New York, 1 edition, 2004.

[160] D. M. Witten and R. J. Tibshirani. Extensions of sparse canonical correlation analysis with applications to genomic data. *Statistical applications in genetics and molecular biology*, 8(1):28, Jan 2009. doi: 10.2202/1544-6115.1470.

[161] World Health Organization. The international classification of diseases for oncology (ICD-O-3.1). Website: http://codes.iarc.fr/codegroup/2, 2011.

[162] World Health Organization. WHO Classification of Tumours of the Central Nervous System, Revised. Fourth Edition. Technical report, 2016.

[163] World Health Organization. Fact sheet: Cancer. Technical report, World Health Organization, Sep 2018. URL `http://www.who.int/news-room/fact-sheets/detail/cancer`.

[164] D. Wu, D. Wang, M. Q. Zhang, and J. Gu. Fast dimension reduction and integrative clustering of multi-omics data using low-rank approximation: application to cancer molecular classification. *BMC Genomics*, 16(1):1022, Dec 2015. doi: 10.1186/s12864-015-2223-8.

[165] J. Yan, H. Wang, Y. Liu, and C. Shao. Analysis of gene regulatory networks in the mammalian circadian rhythm. *PLoS Computational Biology*, 4(10): e1000193, Oct 2008. doi: 10.1371/journal.pcbi.1000193.

[166] S. Yan, D. Xu, B. Zhang, H.-J. Zhang, et al. Graph embedding and extensions: A general framework for dimensionality reduction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(1):40–51, 2007. doi: 10.1109/TPAMI.2007.12.

[167] X. Yang, H. Han, D. D. De Carvalho, F. D. Lay, et al. Gene body methylation can alter gene expression and is a therapeutic target in cancer. *Cancer cell*, 26(4):577–90, Oct 2014. doi: 10.1016/j.ccr.2014.07.028.

[168] Y. Yang. Cancer immunotherapy: harnessing the immune system to battle cancer. *The Journal of Clinical Investigation*, 125(9):3335–3337, Sep 2015. doi: 10.1172/JCI83871.

[169] Y. Yarden and M. X. Sliwkowski. Untangling the ErbB signalling network. *Nature Reviews Molecular Cell Biology*, 2:127–137, Feb 2001. doi: 10.1038/35052073.

[170] S. Yu, L.-C. Tranchevent, X. Liu, W. Glanzel, et al. Optimized data fusion for kernel k-means clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(5):1031–1039, 2012. doi: 10.1007/978-3-642-19406-1_4.

[171] M. Zarrei, J. R. MacDonald, D. Merico, and S. W. Scherer. A copy number variation map of the human genome. *Nature Reviews Genetics*, 16(3):172–183, Mar 2015. doi: 10.1038/nrg3871.

[172] F. Zhang. *Matrix Theory: Basic Results and Techniques*. Springer, 1999.

[173] Y. Zhang and Q. Yang. A survey on multi-task learning. *arXiv preprint arXiv:1707.08114v2*, Jul 2017.

[174] J. Zhao, X. Xie, X. Xu, and S. Sun. Multi-view learning overview: Recent progress and new challenges. *Information Fusion*, 38:43–54, Nov 2017. doi: 10.1016/J.INFFUS.2017.02.007.

[175] X. Zhao, Y. Ren, M. Lawlor, B. D. Shah, et al. BCL2 amplicon loss and transcriptional remodeling drives ABT-199 resistance in B cell lymphoma models. *Cancer cell*, 35(5):752–766.e9, may 2019. ISSN 1878-3686. doi: 10.1016/j.ccell.2019.04.005.