Statistical Learning Methods for Bias-aware HIV Therapy Screening

Author
Jasmina Bogojeska

Dissertation

for obtaining the degree
of a Doctor of the Natural Sciences (Dr. rer. nat.)
of the Natural-technical Faculties
of the Saarland University

Saarbrücken
2011

Statistische Lernverfahren für Bias-bewusste HIV-Therapie Screening

Autor
Jasmina Bogojeska

Dissertation

zur Erlangung des Grades
des Doktors der Naturwissenschaften (Dr. rer. nat.)
der Naturwissenschaftlich–Technischen Fakultäten
der Universität des Saarlandes

Saarbrücken
2011

Tag des Kolloquiums:	08.12.2011
Dekan:	Prof. Dr. Holger Hermanns
Vorsitzender des Prüfungsausschusses:	Prof. Dr. Bernt Schiele
Berichterstatter:	Prof. Dr. Dr. Thomas Lengauer
	Prof. Dr. Jörg Rahnenführer
Beisitzer:	Dr. Ingolf Sommer

Eidesstattliche Versicherung

Hiermit versichere ich an Eides statt, dass ich die vorliegende Arbeit selbstständig und ohne Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe. Die aus anderen Quellen oder indirekt ubernommenen Daten und Konzepte sind unter Angabe der Quelle gekennzeichnet. Die Arbeit wurde bisher weder im In- noch im Ausland in gleicher oder ähnlicher Form in einem Verfahren zur Erlangung eines akademischen Grades vorgelegt.

Saarbrücken, den

Jasmina Bogojeska

Abstract

The human immunodeficiency virus (HIV) is the causative agent of the acquired immunodeficiency syndrome (AIDS) which claimed nearly 30 million lives and is arguably among the worst plagues in human history. With no cure or vaccine in sight, HIV patients are treated by administration of combinations of antiretroviral drugs. The very large number of such combinations makes the manual search for an effective therapy practically impossible, especially in advanced stages of the disease. Therapy selection can be supported by statistical methods that predict the outcomes of candidate therapies. However, these methods are based on clinical data sets that are biased in many ways. The main sources of bias are the evolving trends of treating HIV patients, the sparse, uneven therapy representation, the different treatment backgrounds of the clinical samples and the differing abundances of the various therapy-experience levels.

In this thesis we focus on the problem of devising bias-aware statistical learning methods for HIV therapy screening – predicting the effectiveness of HIV combination therapies. For this purpose we develop five novel approaches that when predicting outcomes of HIV therapies address the aforementioned biases in the clinical data sets. Three of the approaches aim for good prediction performance for every drug combination independent of its abundance in the HIV clinical data set. To achieve this, they balance the sparse and uneven therapy representation by using different routes of sharing common knowledge among related therapies. The remaining two approaches additionally account for the bias originating from the differing treatment histories of the samples making up the HIV clinical data sets. For this purpose, both methods predict the response of an HIV combination therapy by taking not only the most recent (target) therapy but also available information from preceding therapies into account. In this way they provide good predictions for advanced patients in mid to late stages of HIV treatment, and for rare drug combinations.

All our methods use the time-oriented evaluation scenario, where models are trained on data from the less recent past while their performance is evaluated on data from the more recent past. This is the approach we adopt to account for the evolving treatment trends in the HIV clinical practice and thus offer a realistic model assessment.

Kurzfassung

Das Humane Immundefizienz-Virus (HIV) ist der Erreger des erworbenes Immundefektsyndroms (AIDS), das fast 30 Millionen Menschen das Leben gekostet hat und wohl als eine der schlimmsten Seuchen in der Geschichte der Menschheit gelten kann. Da in absehbarer Zeit keine Heilung oder Impfung gegen diese Krankheit zu erwarten ist, werden HIV-Patienten durch die Verabreichung von Kombinationen von anti-retroviralen Medikamenten behandelt. Die sehr große Zahl solcher Kombinationen macht die manuelle Suche nach einer effektiven Therapie vor allem in fortgeschrittenen Stadien der Erkrankung praktisch unmöglich. Dieser Prozess der Therapieauswahl kann mit Hilfe statistischer Verfahren unterstützt werden, welche die Ergebnisse der Therapie vorherzusagen versuchen. Allerdings beruhen diese Methoden auf klinischen Datensätzen die verschiedene Biases enthalten. Die wichtigsten Quellen für Bias sind die sich entwickelnden Trends in der Behandlung von HIV-Patienten, die sparse, ungleichmäßige Repräsentation der Therapien, die verschiedenen Behandlungshintergründe der klinischen Proben sowie die variablen Häufigkeiten der Therapieerfahrungen.

In dieser Arbeit konzentrieren wir uns auf die Aufgabe, Bias-bewusste statistische Lernverfahren für das HIV-Therapie Screening zu konzipieren und die Effektivität von HIV-Kombinationstherapien vorherzusagen. Zu diesem Zweck entwickeln wir fünf neue Ansätze, welche die erwähnten Biases in klinischen Datensätzen bei der Vorhersage von HIV-Therapien berücksichtigen. Drei dieser Ansätze zielen auf eine gute Vorhersageleistung für jede Medikamentenkombination unabhängig von deren Frequenz in den klinischen Daten. Um dies zu erreichen versuchen die Ansätze die sparsen und ungleichmäßig verteilten Therapie-Repräsentationen auszugleichen, indem sie Informationen über verwandte Therapien auf verschiedenen Weise ausnutzen. Die verbleibenden zwei Ansätze berücksichtigen zudem den Bias, der von den verschiedenen Behandlungshintergründen der Proben in den klinischen Datensätzen herrührt. Zu diesem Zweck sagen die Methoden das Therapie-Ansprechen für HIV-Kombinationstherapien auf eine Weise vorher, die nicht nur die direkt vorhergehende Therapie berücksichtigt sondern auch auch Informationen über andere, zeitlich früher gelegene Therapien mit einbezieht. Auf diese Weise bieten die vorgestellten Ansätze gute Vorhersagen für fortgeschrittene Patienten im mittleren bis späten Stadium der HIV-Behandlung sowie für seltene Medikamentenkombinationen.

Alle unsere Methoden verwenden ein zeitorientiertes Evaluierungsszenario, in dem Modelle auf Daten aus der entfernteren Vergangenheit trainiert werden, während ihre Vorhersageleistung auf Daten aus der jüngeren Vergangenheit ausgewertet werden. Dieser Ansatz wurde gewählt, um die entwickelnden Trends in der klinischen HIV-Behandlung zu berücksichtigen und damit eine realistische Bewertung der vorgestellten Modelle zu ermöglichen.

Acknowledgements

First of all, I would like to thank my advisor Prof. Thomas Lengauer for offering me the great opportunity to work on a challenging problem, for the continuous support, advice and encouragement throughout the thesis and also for giving me the freedom to pursue my own ideas and goals. I am also grateful to Prof. Jörg Rahnenführer for kindly agreeing to referee this thesis.

Special thanks goes to all people behind the EuResist project (IST-2004-027173) coordinated by Maurizio Zazzi and Francesca Incardona – without the HIV clinical data provided by them this work would not have been possible. I would also like to Rolf Kaiser and all members of his group at the Institute of Virology (University of Cologne) for the motivating meetings (Arevir and Rettenstein) where I was updated on the clinical developments in HIV treatment and I had the opportunity to participate in many inspiring discussions. I am much obliged to Steffen Bickel and Tobias Sheffer for the fruitful collaboration. I learned a lot about dealing with differing training and test distributions in the numerous discussions with Steffen.

Thanks to all the former and present members of the *Department for Computational Biology and Applied Algorithmics* at the *Max Planck Institute for Informatics* in Saarbrücken for creating an inspiring and fun working environment. Especially Laura and Adrian for the interesting discussions and for being great office mates; Laura, Alex, Levi, Yassen and Glenn for proof-reading parts of this thesis; the HIV subgroup - André, Alex, Kasia, Alejandro for the support on the HIV-related topics; Daniel for the excellent work in his Master's Thesis; Yassen, Kosio, Jörg, Ingolf, Oliver, Bastian, Fidel, Lars, Gabi, Francisco, Jochen, Sven, Hagen for the numerous work-, life- and fun-related discussions. Furthermore, I am very grateful to Ruth and Achim for their continuous support.

During my time here in Saarbrücken I met many great people – Adrian, Laura, Konstantin, Yassen, Cris, Rali, Josi, Esteban, Dana, Hans, Fidel, Evangelia, André, Steffen, Irina, Monika, Bojan, Marjan, Brice, Gabi, Jochen, Alejandro, Oliver, Sven, Hagen, Kasia. Thanks for all the fun time and for being great friends. I would also like to thank all my friends in Macedonia, especially Sonja, Irena and Danja for staying in touch beside the distance and for the great time spent together whenever I go there.

I owe a big thank you to my parents, Zlatko and Milana, and my sister Gabriela for their love, invaluable support throughout my whole life and for always believing in me. A special thanks also goes to my koleska Gordana for being a great friend, for our great plans, for her constant encouragement and for our perfect understanding. Above all, I am grateful to Levi for always being there for me – sharing the happy moments and helping me through the sad ones.

Contents

1 Introduction

The HIV challenge

The human immunodeficiency virus (HIV) infects and destroys the cells of the human immune system causing the acquired immunodeficiency syndrome (AIDS). The deterioration of the immune system is eventually accompanied by opportunistic infections which typically lead to the death of the patient. With no cure or vaccine in sight, HIV is among the deadliest pathogens in the history of mankind. Since its discovery in 1981, AIDS claimed nearly 30 million lives and the current number of infected people worldwide is larger than 33 million.

In developed regions of the world a whole arsenal of antiretroviral drugs that inhibit different stages of the HIV replication cycle is available for the treatment of HIV patients. Despite the large number of available antiretroviral drugs, the virus cannot be eradicated completely from the patient's body and AIDS continues to cause high rates of mortality. The feature that makes HIV so vigorous is its high genetic diversity due to its fast replication cycle with an error-prone reverse transcription step. This brings forth a very dynamic virus population within each infected patient that is able to rapidly evolve and adapt to the selective pressure of administered drugs by developing resistant variants. While each of these drugs is insufficient by itself for substantially delaying the progression of the HIV infection towards AIDS, the administration of combinations of several drugs routinely leads to prolonged virus suppression and restoration of immunologic function. Therefore, modern HIV treatment follows an approach called highly active antiretroviral therapy (HAART) that comprises combinations of several antiretroviral drugs. More specifically, each drug cocktail consists of at least three compounds that provide at least two different mechanisms of inhibiting viral replication. The introduction of HAART was a major breakthrough in the clinical management of HIV infections, resulting in an impressive decrease of HIV-related mortality in the industrialized countries. Nonetheless, drug combination therapies are eventually defeated by the evolution of HIV to resistance as well. In such a case the physician needs to administer a new effective drug combination.

The search for an optimal therapy combination for a given patient is hard because it requires the analysis of a large pool of information. First of all, the increasing number of antiretroviral drugs leads to a very large number of putative drug combinations (hundreds to thousands), each characterized by complex drug interactions. Secondly, a large number of resistance-relevant mutations emerge in the course of the virus' response to the administered combination therapies. While such mutations disappear in the viral population found in the patient's blood as the respective drug combination is taken out, they remain present in the latent virus population in several tissues and organs. These hidden mutations are quickly accessed if this is beneficial for the virus which renders previously administered therapy combinations useless. Hence, an important goal in HIV treatment management is

keeping therapy options open, as running out of options means disease progression to AIDS followed by death of the patient. Last but not least, the amount of knowledge acquired in 30 years of HIV research and treatment is constantly expanding. Briefly, for the purpose of selecting an optimal combination therapy, all information mentioned above needs to be appropriately taken into account in a short period of time on a per-patient basis. Thus, finding optimal therapies for HIV patients becomes increasingly impractical to do manually. This illustrates the need for an automated, objective procedure able to exhaustively search through the space of available drug combinations for an optimal therapy that is selected on the basis of the relevant information available on the patient's individuality and history.

Goals

The large amount of clinical data combined with the use of advanced statistical learning methodologies offer a framework for an automated approach to utilizing the available knowledge for predicting the effectiveness of a potential antiretroviral combination therapy. Such a technology can therefore assist the screening for an optimal, effective regimen for an HIV patient and thereby enhance the clinical management of HIV infections. However, having been collected from many patients over many years, the HIV clinical data sets are biased in many ways. The main sources of these biases are the following:

- The trends of treating HIV patients evolve over time as a result of the experience gained in clinical practice and the introduction of new antiviral compounds.
- The clinical data sets comprise many different combination therapies with highly unbalanced sample representation: while for some therapies many samples exist, for others there are very few.
- The data samples originate from patients with different treatment backgrounds. Also the specific treatment histories for the majority of the therapy-experienced samples are unique.
- The various levels of therapy experience ranging from therapy-naïve to heavily pretreated are represented with widely differing sample frequencies.

Such biases influence the distribution of the data which in turn impacts the predictive power of the statistical models derived from these data.
Inspired by the aforementioned problems, the main purpose of this thesis is to develop statistical learning methods for HIV therapy screening by addressing the different kinds of bias affecting the HIV clinical data sets.

Outline

Figure 1.1 presents a schematic overview of the outline of this thesis. In the following we will describe this outline in more detail.

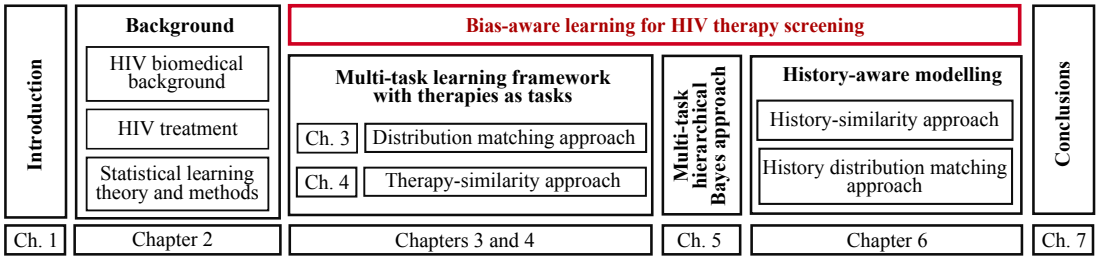


Figure 1.1: Thesis outline.

Chapter 2 provides a brief introduction to the epidemiology of HIV, including its discovery, spread and current geographical distribution in the different regions of the world. Afterwards, we give a description of the HIV virion, its replication cycle and the HIV disease progression. Then, we provide a brief summary of the available anti-HIV drugs in terms of their targets and mechanism of action followed by a description of the modern anti-HIV treatment. In this context, we also review the existing statistical learning methods developed for assisting the administration of HIV treatments. The chapter closes with a short overview of statistical learning theory and methods that play a major role throughout this thesis as they make up the foundations of the models presented in the subsequent chapters.

Chapter 3 addresses the sparse and uneven sample representation of the different drug combinations comprising the HIV clinical data sets when predicting therapy effectiveness. In order to achieve this we first develop a novel multi-task learning approach that considers each combination therapy as a separate task. Our approach trains a separate model for each therapy by using data from all available therapies with properly derived sample weights. These weights are derived such that they match the distribution of all available data to the target distribution of the therapy of interest. In this way our method compensates for the sparse representation of many therapies in the clinical data. Then, we introduce the *time-oriented evaluation scenario* by which our models are trained on the data stemming from the more distant past, while their performance is assessed on data stemming from the more recent past. In this fashion we address the existence of evolving trends in treating HIV patients over time. We close the chapter by describing the clinical data sets and report on experimental results that compare the performance of the newly introduced approach to the corresponding performance of relevant reference approaches.

In Chapter 4 we approach the task of predicting outcomes of HIV combination therapies from viral genotypes. We develop a prediction method that concentrates on producing high quality models for rare therapies, *i.e.* therapies with very few training samples, by taking the sparse therapy representation in the clinical data sets into account. For this purpose, we first introduce two different similarity measures that quantify pairwise similarities of drug combinations. Then, a separate model is trained for each distinct therapy combination by using not only the samples comprising the target therapy, but also

genotypic information from the available samples pertaining to similar therapies weighted with their appropriate similarities. Afterwards, we demonstrate how the existing method is able to utilize additional phenotypic knowledge on the therapy effectiveness stemming from resistance testing. Finally, we present the experimental results realized in the time-oriented evaluation scenario that assess the quality of the presented approach in terms of both prediction performance and interpretability.

The methods we introduce in Chapter 5 use the abundance of samples involving each individual drug to deal with the sparse and highly unbalanced therapy representation when predicting effectiveness of HIV therapies. After we review the general approach of multi-task hierarchical Bayes modeling, we devise two scenarios in the hierarchical Bayesian framework that tackle the problem of predicting the outcome of HIV combination therapies. According to the first approach, each antiretroviral drug is considered as a separate task with the assumption that the effects of the drugs comprising a combination therapy are additive. The second approach builds upon the previous one by adding information on the previous administration of each of the drugs making up the target therapy. This is achieved by creating two separate tasks for each drug in the target therapy that distinguish whether the drug was administered in earlier treatments of the patient or not. Then we describe the experimental setting and report on experimental results.

Chapter 6 presents two novel methods that account not only for the sparse, uneven therapy representation but also for the bias originating from the different treatment backgrounds of the samples making up the clinical data sets. To achieve this, both methods predict the response of an HIV combination therapy by considering not only the most recent therapies but also information from previous therapies administered to the considered patient. We again present the experimental results in the time-oriented evaluation scenario that compare the two novel approaches to the relevant reference approaches and to each other.

Chapter 7 concludes the work presented in this thesis and describes its potential future extensions.

2 Background

This chapter provides the background information necessary for the understanding of this thesis. It consists of two parts. The first part gives a brief overview of the human immunodeficiency virus (HIV) and the current standard approach to treating HIV-infected patients. The second part focuses on statistical learning in general terms and in the framework of assisting clinical management of HIV infections. There is an extensive amount of material and information on both considered topics and we will only provide aspects of direct relevance for this thesis.

2.1 HIV

The human immunodeficiency virus (HIV) is the causative agent of the acquired immunodeficiency syndrome (AIDS) observed for the first time in 1981 in the USA. Since its discovery it claimed nearly 30 million lives and is arguably among the worst plagues in human history. The AIDS pandemic has presented a great medical challenge to humanity rendering HIV the most studied virus in the medical history. HIV is also a great success story of modern medicine in that applying modern therapy has led to a dramatic decline of HIV-related mortality.

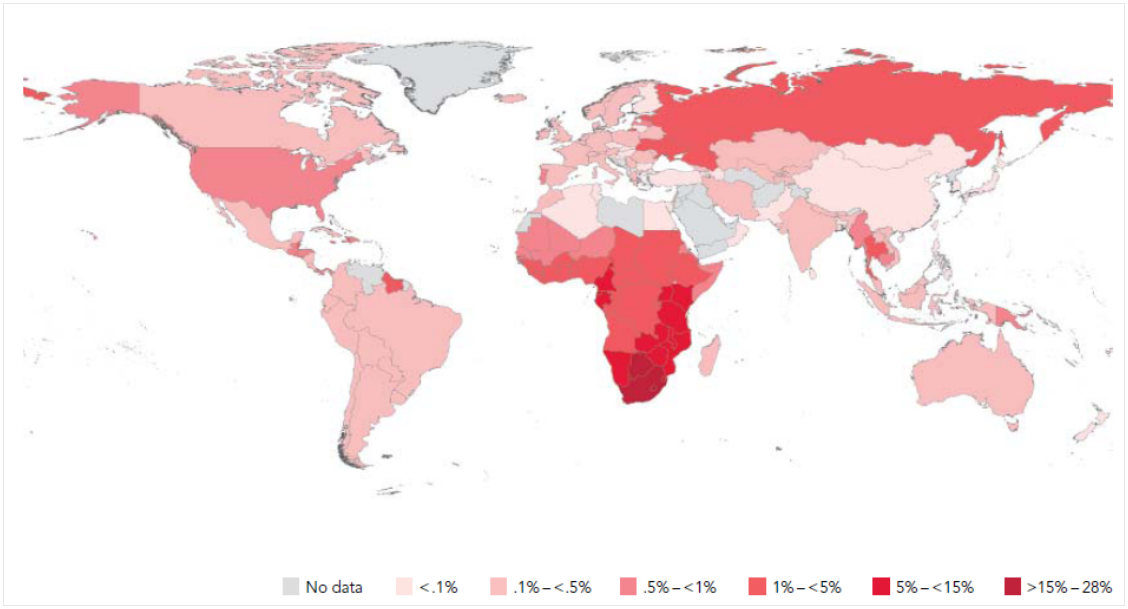
2.1.1 History and Prevalence

The first clinical cases of AIDS were observed in the USA in 1981 (Centers for Disease Control, CDC). The HIV virus was first discovered and isolated in 1983 by two research groups, the group of Robert Gallo (Gallo et al., 1983) and the group of Luc Montagnier (Barre-Sinoussi et al., 1983), independently. Popovic et al. (1984) demonstrated that HIV is the cause of AIDS, described the isolation and characterization of HIV from AIDS patients and, most importantly, developed an immunoassay for screening for HIV which enabled diagnosis that prevented new infections and thus saved many lives.
HIV is a retrovirus that is thought to have entered the human population in the early 20th century (1931 [95% CI $1915 - 1941$]) as a result of several cross-species transfers from non-human primates infected with the simian immunodeficiency virus (SIV) (Korber et al., 2000). There are two types of HIV: type-1 (HIV-1) and type-2 (HIV-2). HIV-1 is far more infective and virulent than HIV-2. Therefore, HIV-1 is responsible for the majority of the infections worldwide, whilst the spread of HIV-2 is mainly restricted to West Africa (Reeves and Doms, 2002; Azevedo-Pereira et al., 2005; de Silva et al., 2008).
HIV expanded at a fast pace from its cradle in Africa to the rest of the world and significantly affected the global health of the human population. According to UNAIDS/WHO since the beginning of the epidemic in 1981 more than 60 million people have been infected with the HIV virus and nearly 30 million people have died of AIDS (UNAIDS/WHO,

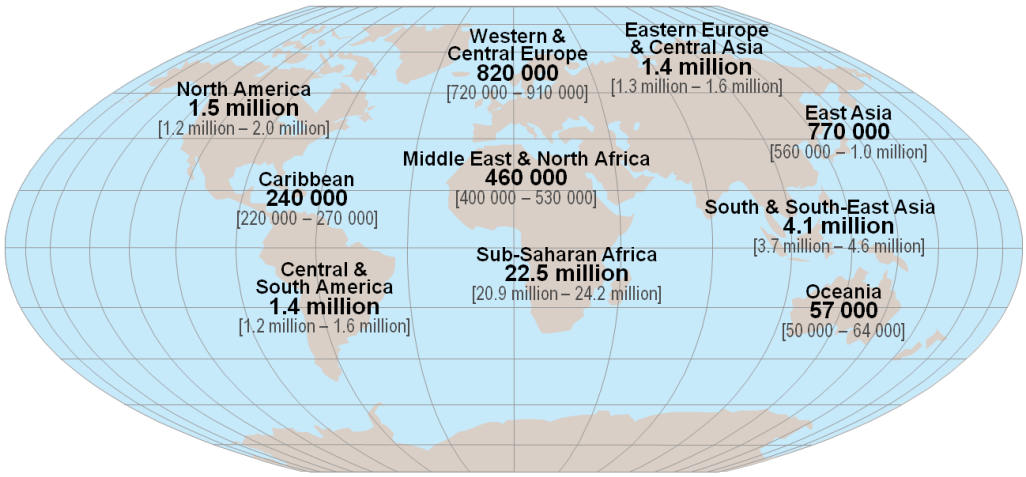
2010). Furthermore, at the end of 2009 there were an estimated 33.3 million people living with HIV, 2.64 million new infections and 1.8 million AIDS-related deaths. As depicted in Figure 2.1 (a), the HIV prevalence varies between different regions of the world with the highest rate observed in sub-Saharan Africa where two thirds of all HIV/AIDS infected people live. In four Southern African countries (Botswana, Lesotho, South Africa and Swaziland) the prevalence exceeds 15%. The numbers of infected individuals in the different parts of the world are presented in Figure 2.1 (b).
There are three major transmission routes for HIV: sexual intercourse, exchange of contaminated blood and blood products (blood transfusions, needle sharing by intravenous drug users, persons receiving medical care in third world countries, health-care workers exposed to contaminated material) and mother-to-child transmission during pregnancy, at childbirth and via breastfeeding. In the industrial countries where antiretroviral drugs are available the risk of mother-to-child transmission is only one percent (Coovadia, 2008). Also the routine prescreening of blood products renders the risk of transmission by blood transfusion negligible. The majority of HIV infections are caused by unprotected sexual intercourse, with much higher rates in low-income countries (Boily et al., 2009).

2.1.2 Virion Structure and Genome

HIV is a member of the genus of *lentiviruses* in the family of *retroviruses*. This means that it has a long incubation period (Lévy, 1993) which is believed to be the main reason for its silent spread in the human population from the time of the first cross-species transmissions from SIV (around 1931) until the first AIDS cases appeared (in 1981). Furthermore, as other retroviruses, HIV encodes its genomic information in the form of ribonucleic acid (RNA) and can only replicate in a host cell by reversely transcribing its viral RNA to deoxiribonucleic acid (DNA) and incorporating its genes into the host genome.
An HIV virion is schematically depicted in Figure 2.2. Its shape is roughly spherical with a diameter of about $100 - 120\ nm$. Protected in a cone-like capsid, the core of the virus consists of two copies of positive-sense single-stranded RNA tightly bound to nucleocapsid proteins (p6 and p7), and two viral enzymes, namely, the reverse transcriptase and the integrase. The viral capsid together with the viral enzyme protease is surrounded by the spherical matrix composed of about 2000 copies of the viral protein p17. This whole complex is engulfed by a phospholipid bilayer referred to as viral envelope. The 72 spikes embedded in the envelope consist of three copies of the viral transmembrane protein gp41 and three copies of the viral envelope protein gp120.
The viral RNA strand is approximately $9.7kb$ long and comprises nine genes: Gag, Pol, Env, Vif, Vpr, Vpu, Rev, Tat, and Nef that code for 15 viral proteins in overlapping reading frames. The genome of HIV is schematically illustrated in Figure 2.3. The first three genes code for precursor proteins that have to be cleaved into functional subunits and the rest are mainly regulatory genes. The Gag gene encodes the four viral structural proteins: p17 (matrix), p24 (capsid), p7 (nucleocapsid) and p6 (nucleocapsid). The Pol gene is the precursor of the three viral proteins: protease, reverse transcriptase and integrase. The last precursor gene Env is responsible for encoding the proteins gp41 and gp120 that build up the viral spikes. The other genes code for various regulatory proteins that are important for the replication efficiency of the virus, its ability to infect cells or to defeat the defense



(a)



Total: 33.3 million [31.4 million – 35.3 million]

(b)

Figure 2.1: (a) Global prevalence of HIV; and (b) Estimated counts of HIV infected individuals in different regions of the world in December 2009. The figures are taken from the WHO 2010 Report on the global AIDS epidemic (UNAIDS/WHO, 2010) available at `http://www.who.int/hiv/pub/global_report2010/en/`.

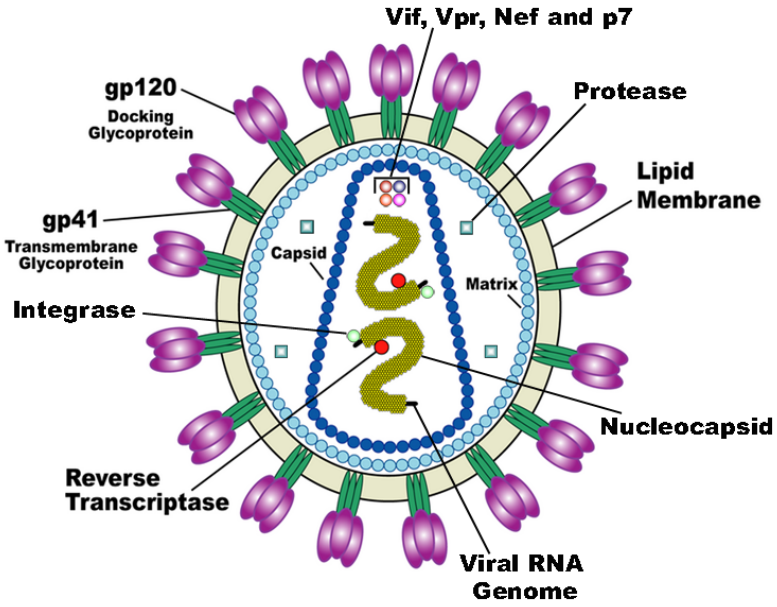


Figure 2.2: Schematic representation of an HIV virion borrowed from Wikipedia (`http://commons.wikimedia.org/wiki/File:HIV_Virion-en.png`).

mechanisms of the host.
Further details on the structure and the genome of the HIV virion can be found in Fields et al. (2007).

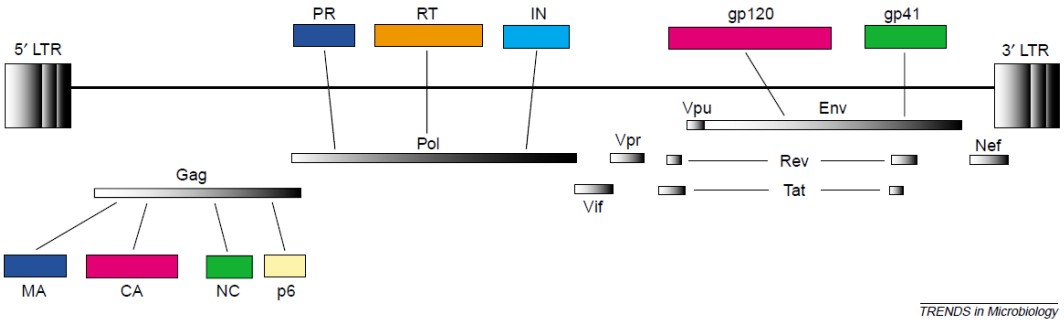


Figure 2.3: Organization of the HIV-1 genome. From Freed (2004).

2.1.3 HIV Replication Cycle

For many years now the HIV replication cycle has been the topic of extensive research that resulted in deeper understanding of its complex stages. Nevertheless, many details of this complex process still remain a mystery. In what follows we briefly describe the basic steps of the HIV life cycle important for the development of the available anti-HIV drugs. Further details and more in-depth explanations are available in Freed (2001); Fields et al. (2007). Figure 2.4 presents a graphical overview of the essential steps in the HIV replication cycle.
The process of virus replication is completed in about 1.5 days and consists of two main

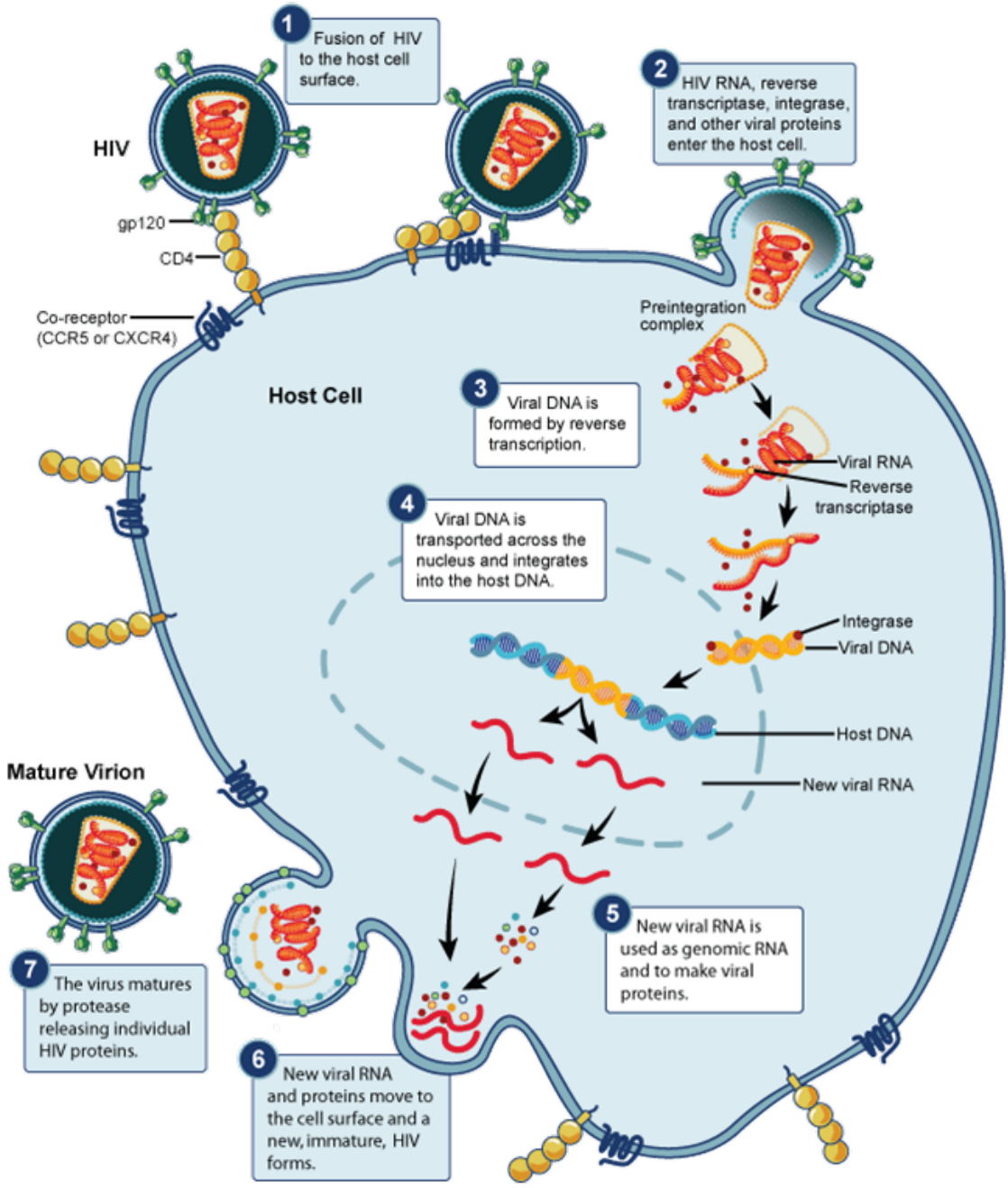


Figure 2.4: The essential steps in the replication cycle of HIV: viral entry, reverse transcription, integration, assembly, budding and maturation of new viral particles. Courtesy from the National Institute of Allergy and Infectious Diseases (`http://www.niaid.nih.gov/SiteCollectionImages/topics/hivaids/hivReplicationCycle.gif`).

stages: *the early stage*, and *the late stage*. The early stage starts with the viral entry of HIV in the host cell and ends with the integration of the viral genome in the genome of the host cell. The late stage comprises the process of production of new HIV virions. The main targets of HIV are the $CD4^+$ T cells, the dendritic cells and the macrophages. They all express the $CD4$ receptor which is very important for HIV cell entry (Dalgleish

et al., 1984). In order to enter the cell one of the viral gp120 surface proteins binds to the $CD4$ cell protein forming a complex that undergoes structural changes allowing specific domains of gp120 to interact with the target chemokine receptor (referred to as coreceptor) – a process called *anchoring*. The two important coreceptors for HIV binding in vivo are $CCR5$ and $CXCR4$ (Berger et al., 1999). Upon binding, the membranes of the virus and the host cell are fused (Esté and Telenti, 2007) and the viral core is released into the cytoplasm of the host cell. Shortly after the viral capsid is uncoated, the enzyme reverse transcriptase (RT) transcribes the single-stranded viral RNA into a double-stranded DNA. The process of reverse transcription is highly error-prone because of the lack of a proof-reading mechanism. According to Gao et al. (2004) the estimated mutation rate of HIV in one replication round is 5.4×10^{-5} mutations per nucleotide. This causes a highly diverse viral population even in a single patient enabling the virus to evade the patient's immune system or to escape the selective pressure presented by drug therapy.

Right after the transcription phase is completed the so called preintegration complex (PIC) comprising the viral DNA together with viral and host proteins is formed and transported to the nucleus. Then the viral enzyme integrase (IN) catalyzes the integration of the viral DNA into the genome of the host cell. At this point the cell is irreversibly infected and becomes a potential virus producer (Simon et al., 2006). The integrated viral DNA is referred to as the *provirus* which can enter the lysogenic cycle enabling it to remain in the cell for a long period of time (in the range of several decades) without being actively transcribed. In this way latent copies of many different viral strains remain dormant in the host cells and are thus not recognized as targets by the host immune system. Therefore, it is virtually impossible to eradicate HIV from infected patients.

Once HIV enters the late stage it exploits the transcription mechanism of the host cell to produce viral messenger RNA (mRNA). Firstly, spliced mRNA is exported from the nucleus to the cytoplasm where it is used for the production of the regulatory proteins Tat and Rev. Secondly, the Rev protein binds to the viral RNA in the nucleus allowing unspliced mRNAs to be transported to the cytoplasm (Cullen, 1991; Pollard and Malim, 1998) where it is translated to the viral structural polyproteins Gag and Env. The process of the assembly of the new viral particles begins with the glycosylation of the protein Env (gp160) in the endoplasmatic reticulum and its transport to the Golgi complex where it is cleaved into the viral glycoproteins gp120 and gp41. These are then transported to the plasma membrane of the host cell where gp41 anchors the gp120 to the cell membrane. As the Gag and GagPol polyproteins along with the viral RNA also associate with the inner surface of the plasma membrane, the new viral virion buds from the infected cell. The very last stage of the HIV replication cycle is the virus maturation which occurs either during or after the process of budding from the cell. The viral enzyme protease (PR) plays the key role during viral maturation since it cleaves the HIV precursor polyproteins to form all functional components of an HIV virion. Only mature virus particles are able to infect new host cells.

2.1.4 Course of Infection and Pathogenesis

The course of an HIV infection depicted in Figure 2.5 is commonly characterized by three consecutive stages: *acute phase*, *latent phase*, and *AIDS (last) phase*. These stages are

determined by measuring two markers, namely the level of $CD4^+$ T cells (cell count per microliter (μl) blood), and the viral load (copies of viral RNA per milliliter blood plasma cp/ml).

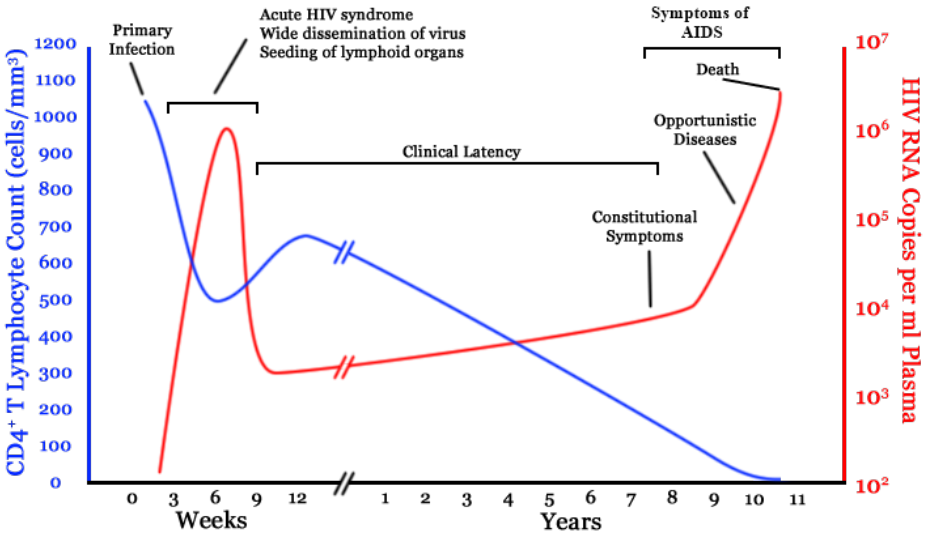


Figure 2.5: Typical course of the time progression of an untreated HIV infection depicted in terms of viral and $CD4^+$ T cell dynamics. From Wikipedia (`http://en.wikipedia.org/wiki/File:Hiv-timecourse_copy.svg`).

The acute phase starts right after the patient is infected when the virus replicates at a very high rate and the viral load is in the range of millions of copies per milliliter blood plasma (Piatak et al., 1993). Since HIV mainly targets $CD4^+$ T cells, their count during the acute phase is halved resulting in influenza-like symptoms (fever, sore throat, rash, headache, nausea) appearing at about two weeks after the initial virus exposure. These nonspecific symptoms render the HIV infection difficult to diagnose and thus increase the risk of virus transmission to new uninfected individuals. The end of the acute phase is characterized by a strong immune response that leads to a reduction of the viral load to a level of several thousand copies along with an increase of the count of $CD4^+$ T cells.

The subsequent phase is the latent phase when the viral load slowly increases and at the same time the $CD4^+$ count continuously decreases. The span of this phase ranges from several months to more than two decades.

AIDS is the final phase of the HIV infection when the immune system of the patient collapses with only up to 200 $CD4^+$ T cells per microliter blood. This causes a sharp increase in the viral load level and occurrence of various AIDS-related opportunistic diseases. These infections are typical for AIDS. In fact they are used to define the AIDS syndrome. They range from repeating respiratory tract infections, chronic diarrhea, tuberculosis and skin rashes to cancer (Moore and Chaisson, 1996; Chaisson et al., 1998). While many of these diseases are not that dangerous for healthy people they are life-threatening for the weak immune system of a progressed HIV patient. Usually the AIDS phase rather quickly ends with death caused by some of the AIDS-related infections.

The time period between the primary HIV infection and AIDS is different for different patients ranging from six months to more than 20 years. Furthermore, the disease devel-

opment of a small percentage of HIV patients deviates from the typical progression we described in the text above. These patients are able to retain a high level of $CD4^+$ T cells by keeping the replication rate of the virus low without antiretroviral treatment. They are grouped in two groups, namely, the *long-term nonprogressors* with virus load bellow 5000 cp/ml, and the *elite controllers* with virus load bellow 50 cp/ml (Grabar et al., 2009; Blankson, 2010). In recent years patients belonging to these two groups have been recruited for more detailed studies which aim at discovering the reasons for their natural virus suppression (Fellay et al., 2007).

2.1.5 Genetic Variability

One of the main characteristics of HIV is its high genetic diversity due to its fast replication cycle with an error-prone reverse transcription step. This brings forth a very dynamic virus population in each infected patient able to rapidly evolve and adapt to the selective pressure of administered drugs by developing resistant variants.

Molecular dating studies estimate that HIV entered the human population in the early 20th century (around 1931) as a result of several cross-species transmissions from nonhuman primates infected with SIV (Korber et al., 2000). The phylogenetic tree of SIV and HIV is depicted in Figure 2.6. As we already mentioned, two types of HIV are known: HIV-1 and HIV-2. It is very likely that HIV-1 originates from a zoonotic transmission from chimpanzee populations infected with SIVcmp to humans (Keele et al., 2009). HIV-2 is most probably derived from multiple cross-species transfers of SIVmm from sooty mangabeys (Heeney et al., 2006). Note that, unlike HIV, SIV is typically non-pathogenic in its natural host.

HIV-1 is divided into three groups, M (main), N (non-M, non-O) and O (outlier), each corresponding to a separate zoonosis event. There is a hypothesis for the existence of a fourth group P (Plantier et al., 2009) probably derived from gorilla SIV. Group M is responsible for the majority of observed infections and is further divided into eight genetically distinct subtypes (A to D, F to H, and J), with subtypes C, A, B and D being the most prevalent ones (Robertson et al., 2000). Subtype B is mostly found in Europe and North America, subtypes A and D are dominant in Africa and subtype C is mostly occurring in Africa and Southeast Asia (Hemelaar et al., 2004).

In the case of HIV-2 the groups A to H are known to exist with groups A and B responsible for the majority of infections in West Africa (Santiago et al., 2005).

Note that different subtypes can form recombinants. These mostly appear as unique recombinant forms (URFs) generated in single patients coinfected with two or more distinct virus subtypes. Further, there are also established circulating recombinant forms (CRFs) when at least three representative full-length genomes have been sequenced from individuals whose infections cannot be epidemiologically linked.

The most dominant subtype in the developed world is HIV-1 subtype B. This is the reason why it is the most studied subtype and why the majority of antiretroviral drugs target HIV-1 subtype B infections (Parkin and Schapiro, 2004). Furthermore, most of the available clinical data stem from Europe and North America and thus also from patients infected with subtype B. In this thesis we will focus on HIV-1 subtype B simply because the majority of the clinical data at our disposal originate from HIV-1 subtype B. Extending the conclusions of such analysis to non-B HIV-1 subtypes or to HIV-2 requires further

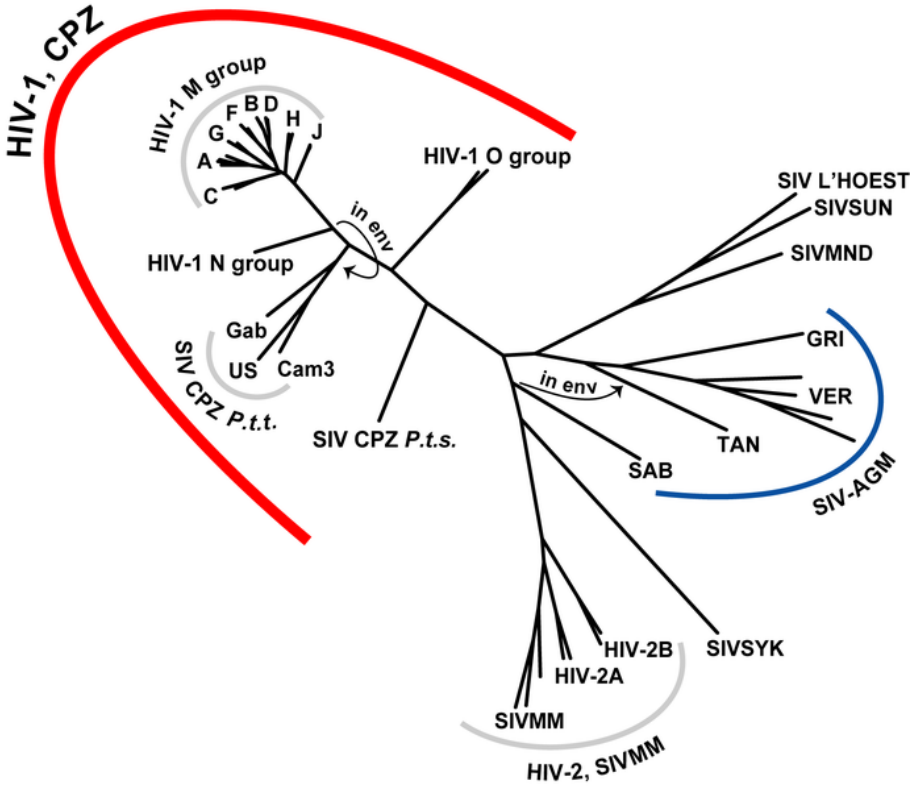


Figure 2.6: Phylogenetic tree depicting the relations among SIV and HIV types and subtypes. From Wikipedia (`http://en.wikipedia.org/wiki/File:HIV-SIV-phylogenetic-tree.png`).

careful research involving more data from the analyzed variant.
The high genetic variability is the central problem for anti-HIV therapy and for vaccine development (Walker and Burton, 2008) where conserved regions are necessary for the vaccine to be effective against all HIV variants.

2.1.6 Antiretroviral Drugs

To date no cure has been discovered that eradicates the virus from HIV patients. However, the extensive research of the HIV virus revealed many details of its replication cycle which, in turn, facilitates the development of a whole array of anti-HIV drugs. Although these drugs are not able to completely remove the virus from the patients' body they successfully delay HIV disease progression, reduce the risk of transmission and prolong the life span and life quality of infected patients. Therefore, anti-HIV drugs are the main building blocks of modern antiretroviral therapy.
Zidovudine (ZDV) was the first anti-HIV drug approved for clinical use from the US Food and Drug Administration (FDA) in 1987. Now there are about 25 drugs approved for treating HIV patients by the FDA and the European Medicines Agency (EMEA) that target the different stages of the HIV replication cycle (see Figure 2.4). They are listed in Table 2.1. Moreover, several other drugs are under investigation and the development of novel drugs is a topic of continuous research. Based on their target and mechanism of action the anti-HIV drugs are divided into several classes: *protease inhibitors*, *reverse*

transcriptase inhibitors, *integrase inhibitors* and *entry inhibitors*. In the following we will briefly describe each of the different drug classes.

Table 2.1: Antiretroviral drugs approved by the US Food and Drug Administration (FDA). The presented information is adapted from the FDA Website.

Abbreviation	Generic name	Interval of FDA approval
Nucleoside Reverse Transcriptase Inhibitors (NRTIs)		
zidovudine, azidothymidine	AZT,ZDV	1987-
didanosine, dideoxyinosine	ddI	1991-
zalcitabine	ddC	1992-2006
stavudine	d4T	1994-
lamivudine	3TC	1995-
abacavir	ABC	1998-
didanosine	ddI	2000-
tenofovir	TDF	2001-
emtricitabine	FTC	2003-
Nonnucleoside Reverse Transcriptase Inhibitors (NNRTIs)		
nevirapine	NVP	1996-
delavirdine	DLV	1997-
efavirenz	EFV	1998-
etravirine	ETV	2008-
Protease Inhibitors (PIs)		
saquinavire	SQV	1995-
indinavir	IDV	1996-
ritonavir	RTV	1996-
nelfinavir	NFV	1997-
amprenavir	APV	1999-
lopinavir/ritonavir	LPV/RTV	2000-
fos-/amprenavir	FOS/APV	2003-
atazanavir	ATV	2003-
tipranavir	TPV	2005-
darunavir	DRV	2006-
Fusion Inhibitors (FIs)		
enfuviritide	ENF,T-20	2003-
Entry Inhibitors (EIs)		
maraviroc	MVC	2007-
Integrase Inhibitors (InIs)		
raltegravir	RAL	2007-

Protease inhibitors. The protease is essential for the virus maturation as it cleaves the precursor polyproteins Gag and GagPol into the functional HIV proteins and enzymes. The cleavage takes place in the active site of the protease which is the target of the protease inhibitors. They are small molecules that competitively bind to the active site of the protease and disrupt its cleavage function. More details on this drug class can be found in Wensing et al. (2010).

Reverse transcriptase inhibitors. The reverse transcriptase catalyzes the production of the DNA copy from the viral RNA and blocking this process is the goal of the reverse transcriptase inhibitors. Based on their mechanism of action the drugs in this drug class are further divided into two subclasses, namely the *nucleoside/nucleotide reverse transcriptase inhibitors (NRTIs)* and the *non-nucleoside reverse transcriptase inhibitors (NNRTIs)*.
The NRTIs are deoxynucleotide analogs incorporated in the newly transcribed viral DNA by the reverse transcriptase. Due to a lack of a 3′-hydroxyl group they act as chain terminators halting the further synthesis of the viral DNA after added to the DNA chain (Cihlar and Ray, 2010).
The NNRTIs are small molecules that reduce the flexibility of the reverse transcriptase by binding to a hydrophobic pocket in close proximity of its active site. In this manner they render the reverse transcriptase unable to synthesize viral DNA (Cihlar and Ray, 2010).

Integrase inhibitors. The goal of this class of drugs is to obstruct the process of integration of the viral DNA in the chromosome of the host cell. The single approved integrase inhibitor is raltegravir (RAL). It disrupts the transfer of the viral DNA by binding to the catalytic site of the integrase and thus inhibits the process of integration of the virus in the host genome (McColl and Chen, 2010).

Entry inhibitors. These drugs aim at preventing the virus from entering the host cell and can be grouped into two different groups: *fusion inhibitors* and *coreceptor antagonists*.
The fusion inhibitors interrupt the fusion of the membranes of the virus and the host cell. Currently, enfuviritide (ENF, T-20) is the only fusion inhibitor approved for treating HIV patients. It binds to a subunit of the viral transmembrane protein gp41 in order to inhibit the conformational change necessary for the fusion of the viral and host membranes (Esté and Telenti, 2007).
The only drug class that targets host rather than viral proteins is the one of coreceptor antagonists that prevent the binding of the viral protein gp120 to the coreceptor of the host cell by binding to it themselves. Maraviroc (MVC) is the first and only approved drug for clinical use from this class. It binds to the CCR5-receptor and thus prior its administration it is necessary to determine whether this coreceptor is used by the virus in the target patient (Lengauer et al., 2007; Thielen et al., 2010; Dybowski et al., 2010).

2.1.7 Highly Active Anitertoviral Therapy (HAART)

Not long after the approval of the first antiretroviral drug (ZDV) it was observed that after a prolonged treatment due to the selective drug pressure the virus develops resistance variants rendering the drug ineffective (Larder et al., 1989; Larder and Kemp, 1989). The main

reasons for the emergence of resistance are the error-prone reverse transcription process that results with high mutation rates (Gao et al., 2004), the short replication time of about 1.5 days and the high replication rate of the virus. Resistance mutations have been reported for all anti-HIV drugs introduced so far (Johnson et al., 2010). Under monotherapy consisting of a single drug compound the resistance variants appear within weeks (Simon and Ho, 2003). Moreover, due to the phenomenon of *cross resistance*, resistance mutations that emerge as a result of the selective pressure of one drug can also confer resistance to other drugs from the same drug class (Clavel and Hance, 2004). This phenomenon is especially pronounced for all but the newest compound (termed etravirine) from the class of NNRTIs where a single resistance mutation selected under monotherapy with one of the NNRTIs also confers resistance to all the other NNRTIs.

All these observations led to the idea of using drug cocktails that combine several drugs into a so called *combination therapy* or *Highly Active Antiretroviral Therapy (HAART)*. The general rule for HAART is to administer at least three drugs from at least two different classes (Clavel and Hance, 2004). In this manner, the viral replication cycle is targeted at several stages simultaneously which makes it more difficult for the virus to develop resistance as it needs to develop a specific set of resistance mutations in several drug targets. Furthermore, the use of drugs from different classes in a specific combination therapy ensures that the resistance-relevant mutations of one drug do not provide cross resistance to the other drugs in the combination. HAART was a major breakthrough in treating HIV-infected patients since it prolonged their life span by slowing down the progression of the disease substantially and thus it dramatically decreased HIV-related mortality (Clavel and Hance, 2004; Crum et al., 2006). Due to its success combination therapy has become the standard way of treating HIV patients. Although HAARTs remain effective much longer than monotherapies based on single drugs, each drug therapy is eventually defeated by the evolution of the virus to resistance. The reason for this is that even in the presence of HAART the virus continues to replicate at a very low rate and after a certain amount of time resistance mutations emerge that inhibit the effectiveness of the therapy. Eventually high viral load results constituting therapy failure. In such a case the physician needs to administer a new effective drug combination. Note that a large number of resistance-relevant mutations can emerge in the course of the response of the virus to a combination therapy. As the therapy is changed, such mutations disappear in the viral population found in the patient's blood serum, but they remain present in the latent virus population in several tissues and organs. Such hidden mutations are quickly accessed if this is beneficial for the virus. This is the reason why previously administered therapy combinations are not considered as potential therapies and an important issue in HIV treatment is keeping therapy options open.

To summarize, modern treatment of HIV patients consists of a life-long administration of different therapy combinations and is schematically illustrated in Figure 2.7. Its two main goals are to choose a combination therapy effective against the patient's current viral population, on the one hand, and to keep therapy options open as long as possible, on the other hand. Running out of therapy options eventually results in progression to AIDS followed by death of the patient.

The development of drug resistance is not the only issue of HAART. Ubiquitous adherence problems compound the issue. Since the anti-HIV drugs have many side effects ranging

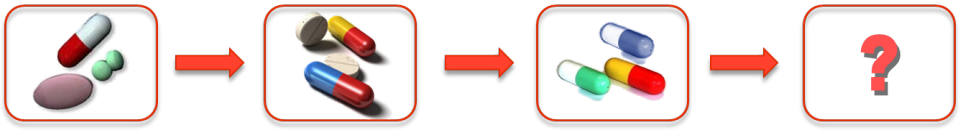


Figure 2.7: Diagram of a common life-long anti-HIV treatment comprising repeated administration of different drug combinations. Whenever the currently administered combination therapy fails a treatment change occurs and a new one needs to be prescribed.

from headache and diarrhea to peripheral nerve damage (Moore and Chaisson, 1996; Spudich and Ances, 2011), combination therapy has substantial impact on the patient's life quality which in turn gives rise to the problem of incomplete adherence. However, in order to be able to suppress virus replication over a long period of time HAART requires complete adherence. Taking the drugs in irregular time intervals, skipping some of them or taking lower doses allows the virus to escape to resistance faster and renders the therapy ineffective (Glass et al., 2008). Because of the large impact of the patients' adherence on the effectiveness of the combination therapies, the severity of side effects and the level of toxicity are gaining importance in the process of development of novel drug compounds.

2.2 Statistical Methods for Assisting the Administration of HIV Combination Therapies

Every combination therapy is eventually defeated by the evolution of the virus to resistance and in such a case the physician needs to decide on a new therapy. It is very important for the patient that the new therapy is effective. The knowledge gained over the years spend on HIV research and the practical experience of treating HIV patients demonstrate the importance of the analysis of the information acquired from *resistance tests* prior to the administration of a new HAART (Durant et al., 1999; Alcorn and Faruki, 2000). There are two types of resistance assays: *genotyping* – sequencing the genomic regions of the viral drug targets from the viral strain(s) in the patient's blood serum, and *phenotyping* – assessment of the level of resistance of each individual drug compound against a given viral strain.

The genotyping assay sequences the regions of the viral genome relevant for the effectiveness of the anti-HIV drugs. It is cheap and the results are available within days. However, the result of genotyping amounts to a set of mutations defined in reference to a wild type virus which is difficult to interpret via manual inspection because of the large number of known resistance-relevant mutations (Johnson et al., 2010). This motivated the development of many rules-based systems that rely on expert knowledge from clinicians and virologists providing the links between a mutation and the change of the virus susceptibility towards a certain drug. HIVdb (Rhee et al., 2003), Rega (Van Laethem et al., 2002), HIV-GRADE, ANRS (Meynard et al., 2002) and AntiRetroScan (Zazzi et al., 2009) are several popular examples for such rule-based systems.

Phenotyping uses a specific experimental assay (Walter et al., 1999; Petropoulos et al., 2000) to quantify the *in vitro* resistance of a given virus (isolated from a given patient) to a

specific anti-HIV drug. In this assay the replication rate of the patient's virus is compared to the one of a wild type virus strain when exposed to a varying concentration of the inspected drug. More specifically, first, the drug concentrations that cut the replication rate by half (termed IC_{50}) in the target and in the reference virus are measured. Then, the *fold-change in resistance* or the *resistance factor (RF)* for the patient's virus is determined as the ratio between the measured drug concentrations of the target and the reference virus:

$$RF = \frac{IC_{50}(target)}{IC_{50}(reference)}.$$

When deciding on a therapy combination phenotyping needs to be performed for every available drug which makes it a time-consuming (in the range of weeks) and expensive process limited to specialized labs with high security levels. Therefore, the availability of data sets comprising genotype-phenotype pairs (GPP) (Rhee et al., 2006) generated from the phenotypic tests for a set of virus sequences for the different drugs promoted the development of statistical models that aim at predicting the resistance factors for each single drug for unknown genotypes. We will refer to these models as *phenotypic models*. In clinical practice, the most widely used phenotypic models are VircoTYPE (Vermeiren et al., 2007) based on linear regression with pairwise interaction terms for the mutations, and geno2pheno[resistance] (Beerenwinkel, 2004) based on linear support vector machines. The genotyping resistance tests and their associated rules-based and statistical models predict the *in vitro* drug resistance of single drugs. Although this information is very useful and highly accepted in the clinical management of HIV infections, composing a suitable combination drug therapy from it remains a challenging task. The main reasons for this are the large number of pharmacokinetic drug interactions that occur when combining multiple drugs (Boffito et al., 2005), various host-specific characteristics, the importance of the previously administered therapies and the latent virus population they created. Furthermore, the very large number of potential drug combinations resulting from the increasing number of antiretroviral drugs makes the manual search for an optimal therapy increasingly impractical, especially in advanced stages of the disease. This illustrates the need for a systematic and quantitative procedure able to predict effectiveness of a potential HAART on the basis of the information available about the patient. An estimate of the therapy outcome can assist physicians in choosing a successful regimen for an HIV patient. The availability of large clinical data sets has paved the way for statistical methods that offer an automated procedure for predicting the outcome of a potential antiretroviral therapy. These data sets contain samples from applications of many different drug combinations in the clinical practice over many years.

In recent years a wide range of approaches for predicting the outcome of antiretroviral combination therapies has been developed. They differ in the algorithmic approach taken, the used inputs, and the scope of their prediction. Machine learning methods, including artificial neural networks (Wang et al., 2003) and fuzzy rules combined with a genetic algorithm (Prosperi et al., 2004), were used to tackle the problem of predicting virological response to a given therapy combination. Moreover, Lathrop and Pazzani (1999) applied combinatorial optimization to the same problem using features extracted from the viral genotype and the drugs in the combination, and Prosperi et al. (2005) used case-based

reasoning. The methods mentioned above predict virological response to therapy by using the viral genotype and the drugs in the applied treatment as features. Altmann et al. (2007) approaches the problem of predicting virological response by including various phenotypic and evolutionary information evaluated with several standard statistical learning techniques and demonstrated that phenotypic information improves the predictive performance of the response to antiretroviral combination therapies. Larder et al. (2007) tackle the problem of predicting virological response to a given HIV drug combination with neural networks. Several studies also include information on the patient's treatment history in their methods and demonstrate its value for predicting effectiveness of potential combination therapies (Bratt et al., 1998; Revell et al., 2010; Prosperi et al., 2010).
However, the HIV clinical data sets suffer from different kinds of bias, which can negatively affect the usefulness of the derived statistical models. First of all, the trends of treating HIV patients evolve over time due to the introduction of new anti-HIV drugs and the practical experience of treating HIV patients. There are also differences in the treatment patterns of HIV patients among the different countries. Then, the data samples originate from patients with different levels of therapy experience, from therapy naïve to heavily pretreated. Furthermore, they contain data on different combination therapies with widely differing frequencies. In particular, many therapies are only represented with very few data points. To our knowledge, none of the available statistical methods for predicting the effectiveness of HIV combination therapies take these important issues into account. Developing methods that address the different biases pertaining to the HIV clinical data sets is the main focus of this thesis.

2.3 Statistical Learning

Large amounts of data are being produced in many areas of science and industry. Statistical learning methods play a major role in the process of extracting meaningful information and drawing conclusions from such often unstructured and cryptic data. Based on their objective, statistical learning problems are classified in three groups:

- *Supervised learning problems* – The *inputs* together with their corresponding *outputs* for a set of entities are observed or measured. The aim is to use the available labeled data to derive a method that models the relationship between inputs and outputs and thus enables accurate predictions of the outputs for unseen entities based on their corresponding inputs.

- *Unsupervised learning problems* – Only the inputs of a set of entities are observed and the goal is to develop methods that discover structure in the unlabeled data by clustering them into groups according to some similarity criterion.

- *Semi-supervised learning problems* – The goal is to devise models that utilize relevant information from both labeled and unlabeled data to accurately predict outputs from inputs.

This thesis concentrates on the supervised learning problem of predicting outcomes of HIV combination therapies. Therefore, in what follows we provide more formal description of

supervised learning theory focussing mainly on aspects and methods relevant for the work we present in the thesis. For further details we refer the reader to Hastie et al. (2009).
Let $\mathbf{x} \in \mathbb{R}^p$ denote a real valued random input vector and y denote a random output variable with joint input-output distribution $p(\mathbf{x}, y)$. In supervised learning the goal is to find a function $f(\mathbf{x})$ that correctly predicts the output y given the input $\mathbf{x}$. To achieve this a loss function $\ell(f(\mathbf{x}), y)$ that quantifies the prediction quality and penalizes prediction error is required. Based on the type of the output y the supervised learning problems are divided into two groups: *regression* – when the output is *quantitative (continuous)*, and *classification* – when the output is *qualitative (categorical)*. For example, $y \in \mathbb{R}$ for regression, and $y \in \{-1, 1\}$ for binary classification. The most common loss function for regression problems is the squared error loss given by:

$$\ell(f(\mathbf{x}), y) = (y - f(\mathbf{x}))^2. \tag{2.1}$$

For classification problems the most popular loss functions are the logistic loss:

$$\ell(f(\mathbf{x}), y) = log(1 + exp(-yf(\mathbf{x}))) \tag{2.2}$$

used by logistic regression (see Subsection 2.3.1), the hinge loss:

$$\ell(f(\mathbf{x}), y) = \max(0, (1 - yf(\mathbf{x}))) \tag{2.3}$$

used in support vector machines (see Subsection 2.3.2), and the zero-one loss that assigns ones for the misclassified and zeros for the correctly classified samples. The zero-one loss is most intuitive, but it is not convex and not continuous in the parameters of f which makes it difficult to use in optimization problems. Therefore, the logistic and the hinge loss, which are convex and continuous, are most widely used in practice. For a given loss function $\ell(f(\mathbf{x}), y)$ the task of learning a prediction function $f(\mathbf{x})$ amounts to minimizing the expected loss with respect to the joint distribution $p(\mathbf{x}, y)$:

$$\mathbf{E}_{(\mathbf{x},y)\sim p(\mathbf{x},y)}[\ell(f(\mathbf{x}), y)] = \int\int \ell(f(\mathbf{x}), y)p(\mathbf{x}, y)d\mathbf{x}dy. \tag{2.4}$$

In practice the distribution $p(\mathbf{x}, y)$ is unknown and only a finite set of training samples $D = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m)\}$ drawn from $p(\mathbf{x}, y)$ is available. Therefore, the joint distribution $p(\mathbf{x}, y)$ in Equation 2.4 is replaced by the empirical distribution over the data set D. In this case, the task of finding a prediction function $f(\mathbf{x})$ amounts to minimizing the empirical expected loss given by:

$$\mathbf{E}_{(\mathbf{x},y)\sim D}[\ell(f(\mathbf{x}), y)] = \frac{1}{|D|} \sum_{(\mathbf{x}_i,y_i)\in D} \ell(f(\mathbf{x}), y). \tag{2.5}$$

Furthermore, in practice a regularization penalty $J(f)$ is added to the empirical expected loss which leads to the penalized empirical expected loss:

$$\mathbf{E}_{(\mathbf{x},y)\sim D}[\ell(f(\mathbf{x}), y)] + \lambda J(f) = \frac{1}{|D|} \sum_{(\mathbf{x}_i,y_i)\in D} \ell(f(\mathbf{x}), y) + \lambda J(f), \tag{2.6}$$

where $\lambda \geq 0$ controls the amount of penalty. The penalty functional $J(f)$ is large for ragged functions and guards against overfitting. A suitable choice for $J(f)$ facilitates maximum

generalization power (minimum test error) of the prediction function f. In the Bayesian framework J reflects the prior beliefs about f, *i.e.* J is the log-prior of the parameters of f and Equation 2.6 is the log-posterior distribution. Common regularization functions are the L_2/L_1-norm of the parameters of f or the square of the second derivative of f.
The prediction function f obtained by minimizing Equation 2.6 is then used for predicting the outcome for a target sample $\mathbf{x}_t$ as follows. In the case of regression the target outcome is given by $y_t = f(\mathbf{x}_t)$. For a binary classification task the target outcome y_t is obtained by using an application-specific threshold tr such that:

$$y_t = \begin{cases} 1, & f(\mathbf{x}_t) > tr; \\ \text{-1}, & f(\mathbf{x}_t) \leq tr. \end{cases} \tag{2.7}$$

For example, for logistic regression the threshold is typically set to $tr = 0.5$.
In what follows we describe two widely used supervised learning methods: logistic regression and support vector machines (SVMs).

2.3.1 Logistic Regression

All methods we present in this thesis are based on some form of logistic regression, so in the following we will provide a more detailed description of this method based on Hastie et al. (2009).
Assuming the existence of K classes and multinomial distribution of the outcome y, logistic regression models the log-odds of the posterior probabilities of the classes $\{P(y = k|\mathbf{x}), k = 1, \ldots, K\}$ via linear functions (hyperplanes) in the input $\mathbf{x}$. $P(y = k|\mathbf{x})$ is the conditional probability of a sample $\mathbf{x}$ to belong to class k. In order to ensure that the posterior probabilities of the K classes sum to 1 the model is specified with $K - 1$ log-odds (also termed logit transformations):

$$\begin{aligned} log\frac{P(y=1|\mathbf{x})}{P(y=K|\mathbf{x})} &= \beta_{10} + \beta_1^T\mathbf{x} \\ log\frac{P(y=2|\mathbf{x})}{P(y=K|\mathbf{x})} &= \beta_{20} + \beta_2^T\mathbf{x} \\ &\vdots \\ log\frac{P(y=K-1|\mathbf{x})}{P(y=K|\mathbf{x})} &= \beta_{(K-1)0} + \beta_{K-1}^T\mathbf{x}. \end{aligned} \tag{2.8}$$

The posterior probabilities derived from Equations 2.8 are given by:

$$\begin{aligned} P(y=k|\mathbf{x}) &= \frac{\exp(\beta_{k0} + \beta_k^T\mathbf{x})}{1 + \sum_{i=1}^{K-1}\exp(\beta_{i0} + \beta_i^T\mathbf{x})}, k = 1, \ldots, K-1 \\ P(y=K|\mathbf{x}) &= \frac{1}{1 + \sum_{i=1}^{K-1}\exp(\beta_{i0} + \beta_i^T\mathbf{x})}. \end{aligned} \tag{2.9}$$

In practice for a given training data set $D = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m)\}$ the logistic regression model is usually estimated by a maximum likelihood approach. More formally the solution of:

Optimization Problem 1. *Over parameters* $\beta = \{\beta_k, k = 1, \ldots, K-1\}$ *and* λ *maximize*

$$\sum_{(\mathbf{x}_i, y_i) \in D} \log(p(y_i|\mathbf{x}_i)) + \lambda J(\beta)$$

is a *maximum a posteriori* estimation of the logistic regression model with parameters $\{\beta_k, k = 1, \ldots, K-1\}$ and regularizer $J(\beta)$ controlled by the regularization parameter λ. In practice only a finite sample of the data is available. The regularizer is then used to counter overfitting and thus improve the generalization performance of the fitted model on unseen data (Schölkopf and Smola, 2002; Hastie et al., 2009). Throughout this thesis we will use the square of the L_2-norm of the model parameters β denoted by $\| \beta \|^2$ as a regularizer function $J(\beta)$. This is also a very common choice in practice because of its nice properties – it is convex and can be interpreted as a Gaussian log-prior of the model parameters (Hastie et al., 2009). Optimization problem 1 can be solved by gradient-based optimization procedures like the *Newton-Raphson algorithm*. Note that efficient model estimation for large high-dimensional data sets is facilitated by special optimization procedures. For more details on the optimization procedures for fitting large-scale logistic regression we refer the reader to Subsection 4.1.1 of Chapter 4.
In what follows we briefly describe two frequently used forms of logistic regression, namely *logistic regression for binary classification*, and *kernel logistic regression*.

Logistic regression for binary classification. In the binary classification scenario where $K = 2$ the logistic regression model is very simple as it only estimates a single linear function. By encoding the two classes as 1 and -1 respectively, the parameters β of this linear function are obtained by minimizing the negative log-likelihood of the data:

Optimization Problem 2. *Over parameters* β*, minimize*

$$\frac{1}{|D|} \sum_{(\mathbf{x}_i, y_i) \in D} \log(1 + exp(-y\beta^T \mathbf{x})) + \frac{\| \beta \|^2}{2\sigma_\beta^2}. \tag{2.10}$$

In the optimization problem above the square of the L_2-norm of the parameters β is used as a regularizer with $\lambda = \frac{1}{2\sigma_\beta^2}$. In this manner the regularizer can be interpreted as a Gaussian log-prior with mean 0 and isotropic covariance matrix $\sigma^2 I$ on the model parameters β (Evgeniou et al., 2000).

Kernel logistic regression. Logistic regression fits linear functions for the purpose of discriminating between the different classes. However, since the boundaries between the classes are not necessarily linear, in many applications the goal is to fit nonlinear discriminant functions. This can be achieved by casting logistic regression in the more general framework of regularization methods and reproducing kernel Hilbert spaces (Berlinet and Thomas-Agnan, 2004).

Definition 1 (Reproducing Kernel Hilbert Space (RKHS))**.** *Let* $\mathbf{x}, \mathbf{z} \in \mathbb{R}^p$*. The space* $\mathcal{H}_K$ *of functions* f *such that* f *is a linear combination of the form* $f(\mathbf{x}) = \sum_m \alpha_m K(\mathbf{x}, \mathbf{z}_m)$ *is called a reproducing kernel Hilbert space. Here each kernel term* K *is considered as a*

function of the first argument $\mathbf{x}$ *and indexed by the second argument. Given the eigen-expansion of* K*:*

$$K(\mathbf{x}, \mathbf{z}) = \sum_{i=1}^{\infty} \gamma_i \phi_i(\mathbf{x}) \phi_i(\mathbf{z}) \tag{2.11}$$

with $\gamma_i \geq 0, \sum_{i=1}^{\infty} \gamma_i^2 < \infty$ *the elements of* $\mathcal{H}_K$ *can be represented in terms of the eigen-functions as:*

$$f(\mathbf{x}) = \sum_{i=1}^{\infty} c_i \phi_i(\mathbf{x}) \tag{2.12}$$

with $\|f\|_{\mathcal{H}_K}^2 < \infty$.

Using the theory of RKHS a general regularization problem with loss function $\ell(.,.)$ has the form:

$$\min_{f \in \mathcal{H}_K} \left(\sum_{i=1}^{m} \ell(f(\mathbf{x}_i),\ y_i) + \lambda \|f\|_{\mathcal{H}_K}^2 \right) \tag{2.13}$$

According to the *representer theorem* (Kimeldorf and Wahba, 1970; Wahba, 1990) the solution to the regularization problem specified by Equation 2.13 in the RKHS $\mathcal{H}_K$ with reproducing kernel K is finite-dimensional of the form:

$$f(\mathbf{x}) = \sum_{i=1}^{m} \beta_i K(\mathbf{x}, \mathbf{x}_i) \tag{2.14}$$

where the regularization function $\|f\|_{\mathcal{H}_K}^2$ is given by:

$$\|f\|_{\mathcal{H}_K}^2 = \sum_{i=1}^{m} \sum_{j=1}^{m} K(\mathbf{x}_i, \mathbf{x}_j) \beta_i \beta_j. \tag{2.15}$$

In the theoretical framework of RKHS the optimization problem of logistic regression is specified by:

$$\min_{\beta} \ell\left(\mathbf{K}_\lambda \beta,\ \mathbf{y}\right) + \lambda \beta^T \mathbf{K}_\lambda \beta \tag{2.16}$$

where $\mathbf{K}_\lambda \in \mathbb{R}^{m \times m}$ is a matrix with entries $(\mathbf{K}_\lambda)_{ij} = K_\lambda(\mathbf{x}_i, \mathbf{x}_j)$ and $\mathbf{y}$ is an m-vector of class labels (Zhu and Hastie, 2002). In this way a feature mapping ϕ maps the samples from the original input feature space to the reproducing kernel Hilbert space (RKHS) in which the scalar product $\phi(\mathbf{x}_i)^\mathsf{T} \phi(\mathbf{x}_j)$ is determined by the kernel function K. The linear decision boundaries computed in the RKHS are nonlinear when mapped back in the original feature space. This enables fitting nonlinear logistic regression models and is referred to as *kernel logistic regression (KLR)* (Zhu and Hastie, 2002). The KLR model has comparable performance to the one of a support vector machine with the same kernel (Zhu and Hastie, 2002).

Throughout this thesis we will provide nonlinear decision boundaries for the fitted logistic regression classifiers by using a Gaussian kernel function often referred to as *radial basis function (RBF) kernel* specified by:

$$K(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\frac{||\mathbf{x}_i - \mathbf{x}_j||^2}{2\sigma^2} \right) \tag{2.17}$$

where $\|\cdot\|$ is the L_2-norm.

2.3.2 Support Vector Machines

Support vector machines are a popular statistical learning method for classification and regression. In the following we describe the standard soft margin SVM used for binary classification.

The *optimal separating hyperplane* separates two perfectly separable classes such that the distance to the closest point from either class, *i.e.* the margin between the two classes is maximized (Vapnik, 1996). The soft margin SVM (Boser et al., 1992; Cortes and Vapnik, 1995) generalizes this idea to the case of nonseparable classes by allowing for some points to be on the wrong side of the margin. Given a training data set $D = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m)\}$ for a binary classification problem ($y_i \in \{1, -1\}$) the support vector classifier solves the following optimization problem:

Optimization Problem 3. *Over parameters* $\{\beta, \beta_0\}$ *minimize*

$$\frac{1}{2}\|\beta\|^2 + C\sum_{i=1}^{m}\xi_i$$

subject to $y_i(\phi(\mathbf{x}_i)^\mathsf{T}\beta + \beta_0) \geq 1 - \xi_i, \xi_i \geq 0, i = 1, \ldots, m$, *where* $\phi(\mathbf{x}_i)$ *maps* $\mathbf{x}_i$ *into a potentially higher dimensional space and* $C > 0$ *is the regularization parameter.*

The solution for this problem is obtained by solving the following dual objective function:

Optimization Problem 4. *Over parameters* α *maximize*

$$\sum_{i=1}^{m}\alpha_i - \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y_i y_j \phi(\mathbf{x}_i)^\mathsf{T}\phi(\mathbf{x}_j)$$

subject to $\sum_{i=1}^{m}\alpha_i y_i = 0, 0 \leq \alpha_i \leq C, i = 1, \ldots, m$.

The optimal solution β^* satisfies:

$$\beta^* = \sum_{i=1}^{m} y_i\alpha_i\phi(\mathbf{x}_i) \tag{2.18}$$

The prediction function then has the form:

$$f(\mathbf{x}) = \sum_{i=1}^{m} y_i\alpha_i\phi(\mathbf{x}_i)^\mathsf{T}\phi(\mathbf{x}) + \beta_0 \tag{2.19}$$

and the label y for $\mathbf{x}$ is specified by using $tr = 0$ in Equation 2.7. It can be observed that both the dual Optimization Problem 4 and the prediction function in Equation 2.19 contain only inner products of the feature mapping ϕ. Hence, only the kernel function:

$$K(\mathbf{x}_i, \mathbf{x}_j) = \phi(\mathbf{x}_i)^\mathsf{T}\phi(\mathbf{x}_j) \tag{2.20}$$

needs to be specified and no knowledge of the mapping ϕ is required. This is often referred to as the *kernel trick*, which renders support vector machines applicable for both linear and non-linear classification. For example, by using the scalar product as kernel function $K(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i^\mathsf{T}\mathbf{x}_j$ one obtains linear classification and by using the RBF kernel one obtains non-linear classification.

Note that SVM can also be formulated as a regularization problem by using the hinge loss (Equation 2.3):

Optimization Problem 5. *Over parameters* $\{\beta, \beta_0\}$ *minimize*

$$\sum_{i=1}^{m} \max(0, (1 - yf(\mathbf{x}))) + \frac{1}{2C} \parallel \beta \parallel^2 .$$

2.4 Learning Under Differing Training and Test Distributions

A major assumption in many statistical learning methods is that the training and the test data are drawn from the same distribution $p_\gamma(\mathbf{x}, y)$ with parameters γ (Hastie et al., 2009). However, in practice the training and test data are often governed by different distributions. As an example, consider the case of experimental data (*e.g.* DNA microarrays) related to the same medical problem (*e.g.* cancer diagnosis) obtained from different labs and thus under different conditions.
Let $p_\gamma(\mathbf{x}, y)$ denote the joint distribution of the training set and $p_\theta(\mathbf{x}, y)$ the joint distribution of the test set. Then the prediction function obtained by minimizing the expected loss with respect to the training distribution does not coincide with the prediction function that minimizes the expected loss with respect to the test distribution:

$$\arg\min_f \mathbf{E}_{(\mathbf{x},y)\sim p_\gamma(\mathbf{x},y)}[\ell(f(\mathbf{x}), y)] \neq \arg\min_f \mathbf{E}_{(\mathbf{x},y)\sim p_\theta(\mathbf{x},y)}[\ell(f(\mathbf{x}), y)]. \tag{2.21}$$

In recent years many statistical learning approaches that address the problem of differing training and test distribution have emerged in the machine learning community. Their goal is to find the prediction function that minimizes the expected loss with respect to the test distribution and thus provide good prediction quality for the test data. In what follows we briefly overview two popular settings that account for the problem of differing training and test distribution: *covariate shift* and *multi-task learning*.

2.4.1 Covariate Shift

Let $D\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m)\}$ denote a labeled training set drawn from the unknown joint distribution $p(\mathbf{x}, y|\gamma) = p(y|\mathbf{x}, \gamma)p(\mathbf{x}|\gamma)$, where the inputs $\mathbf{x}$ are drawn from the training distribution $p(\mathbf{x}|\gamma)$ and the labels y are assigned based on the conditional distribution $p(y|\mathbf{x}, \gamma)$. In the *covariate shift setting*, an unlabeled test set T governed by the unknown test distribution $p(\mathbf{x}|\theta)$ is available. The two main assumptions in this setting are:

- the training and test distributions are different $p(\mathbf{x}|\gamma) \neq p(\mathbf{x}|\theta)$, and
- the training and test labels are assigned based on an identical conditional distribution $p(y|\mathbf{x}) = p(y|\mathbf{x}, \gamma) = p(y|\mathbf{x}, \theta)$.

The goal is then to find a prediction function $f : \mathbf{x} \mapsto y$ that minimizes the expected loss with respect to the unknown joint test distribution $p(\mathbf{x}, y|\theta)$.
Shimodaira (2000) shows that when the support of the test distribution $p(\mathbf{x}|\theta)$ is contained in the support of the training distribution $p(\mathbf{x}|\gamma)$, the expected loss with respect to the joint test distribution $p(\mathbf{x}, y|\theta)$ equals the expected weighted loss with respect to the joint training distribution $p(\mathbf{x}, y|\gamma)$ with sample-specific weights given by $\frac{p(\mathbf{x}|\theta)}{p(\mathbf{x}|\gamma)}$:

$$\mathbf{E}_{(\mathbf{x},y)\sim p(\mathbf{x},y|\theta)}[\ell(f(\mathbf{x}), y)] = \mathbf{E}_{(\mathbf{x},y)\sim p(\mathbf{x},y|\gamma)}\left[\frac{p(\mathbf{x}|\theta)}{p(\mathbf{x}|\gamma)}\ell(f(\mathbf{x}), y)\right]. \tag{2.22}$$

In practice, $p(\mathbf{x}|\gamma)$ and $p(\mathbf{x}|\theta)$ are unknown, but can be estimated from the data sets D and T, respectively. To this end, Shimodaira (2000) and Sugiyama and Müller (2005) use kernel density estimation to obtain empirical estimates of the mentioned distributions. Then they use these estimates to resample or reweight the training set and compute the target prediction function f. However, estimating high-dimensional input distributions requires complex modeling which suffers from the curse of dimensionality problem (Hastie et al., 2009). Therefore, there has been a line of statistical methods that estimate the density ratio $\frac{p(\mathbf{x}|\theta)}{p(\mathbf{x}|\gamma)}$, *i.e.* the sample-specific weights, directly without any explicit modeling of the training and test distributions.
Huang et al. (2007) develop the kernel mean matching procedure that estimates the sample-specific weights for the training set such that the means of the training and test sets are matched in a reproducing kernel Hilbert space. The method KLIEP introduced in Sugiyama et al. (2008) estimates these weights such that the Kullback-Leibler divergence between the test and the weighted training distribution is minimized.
Some approaches for direct estimation of the density ratio $\frac{p(\mathbf{x}|\theta)}{p(\mathbf{x}|\gamma)}$ are inspired by the so-called *sample selection bias setting* (Heckman, 1979; Zadrozny, 2004). In this setting the data generation process is modeled with a binary selector variable s that indicates whether a sample $\mathbf{x}$ drawn from the test distribution $p(\mathbf{x}|\theta)$ belongs to the training set ($s = 1$), or not ($s = -1$). Only the samples in the training set are labeled according to $p(y|\mathbf{x})$. Given that s is independent of the label y:

$$p(s = 1|\mathbf{x}, y, \gamma, \theta) = p(s = 1|\mathbf{x}, \gamma, \theta), \tag{2.23}$$

the training distribution $p(\mathbf{x}|\gamma)$ is given by:

$$p(\mathbf{x}|\gamma) \propto p(s = 1|\mathbf{x}, \gamma, \theta)p(\mathbf{x}|\theta). \tag{2.24}$$

Then, according to Zadrozny (2004) the expected loss with respect to the joint test distribution $p(\mathbf{x}, y|\theta)$ is proportional to the weighted expected loss with respect to the joint training distribution $p(\mathbf{x}, y|\theta)$ with weights $p(s = 1|\mathbf{x}, \gamma, \theta)^{-1}$:

$$\mathbf{E}_{(\mathbf{x},y)\sim p(\mathbf{x},y|\theta)}[\ell(f(\mathbf{x}), y)] \propto \mathbf{E}_{(\mathbf{x},y)\sim p(\mathbf{x},y|\gamma)}\left[\frac{1}{p(s = 1|\mathbf{x}, \gamma, \theta)}\ell(f(\mathbf{x}), y)\right]. \tag{2.25}$$

In practice, $p(s = 1|\mathbf{x}, \gamma, \theta)$ is estimated by training a model that discriminates labeled training ($s = 1$) against unlabeled test samples ($s = -1$).
Dudik et al. (2005) investigate the maximum entropy density estimation under sample selection bias. Bickel and Scheffer (2007) apply a hierarchical Bayesian model with a Dirichlet process prior on several problems with related sample selection bias. Inspired by learning under sample selection bias Bickel et al. (2007) estimate the density ratio $\frac{p(\mathbf{x}|\theta)}{p(\mathbf{x}|\gamma)}$ in the covariate shift setting directly by discriminating between training and test data with a probabilistic classifier and provide an integrated optimization problem for training all model parameters jointly.

2.4.2 Multi-task Learning

In the *multi-task learning setting* the data from several distinct but related prediction tasks are available with differing task sample abundances. Each task is governed by a distinct

task-specific joint distribution of input and output variables and for many of the tasks there are usually not many samples available. The aim is to use the available data from all tasks to develop a model that provides correct predictions for a given target task. In the context of our discussion so far, the test distribution in this setting is the input-output distribution of the task of interest and it differs from the training distribution which is the input-output distribution of all tasks. Compared to the covariate-shift setting the multi-task setting is more general because not only the distributions of the inputs but also the conditional output densities for a given task are different from the corresponding densities of all the other tasks.

A common assumption in most of the existing multi-task learning methods is that all tasks have a common model structure and all task parameters can be estimated jointly. Their objective is to simultaneously provide a good generalization across tasks and correct predictions for each task. The hierarchical Bayes model (Gelman et al., 2004) can easily be applied to multi-task modeling and, therefore, is widely used in the machine learning community. It assumes that all task parameters have the same prior probability. Bakker and Heskes (2003) use a Gaussian prior for the parameters of task-specific neural network models and Evgeniou and Pontil (2004) impose a Gaussian prior for support vector machine models. Yu et al. (2005) impose a normal-inverse Wishart hyperprior on the mean and covariance of a Gaussian process prior shared by the task-specific prediction functions. Moreover, Xue et al. (2007) use a Dirichlet process prior in the hierarchical Bayes framework for task clustering, and Teh et al. (2006) uses a hierarchical Dirichlet process prior.

Standard statistical approaches that rely on the assumption that the training and the test data are drawn from the same distribution are not directly applicable for our HIV therapy screening application due to the different biases pertaining to the HIV clinical data sets. For example, assume the goal is to develop a method that provides accurate predictions for a given therapy combination of interest. However, since the number of samples available for the target therapy is very limited the training set comprises all available samples from all therapies and its joint distribution differs from the joint distribution of the therapy of interest. Hence, throughout this thesis we will develop several methods that address the different biases in the clinical data by casting the problem of predicting effectiveness of HIV combination therapies in the multi-task learning framework. Only the methods presented in Chapter 5 utilize the assumption that all tasks share a common model structure. All other approaches presented in Chapters 3, 4 and 6 of this thesis after specifying the tasks train a separate model for each of them.

3 Multi-task Learning for HIV Therapy Screening

HIV patients are treated by administration of combinations of several antiretroviral drugs. The very large number of such combinations makes the manual search for an effective therapy increasingly impractical, especially in advanced stages of the disease. Therapy selection can be supported by statistical methods derived from HIV clinical data that predict the outcomes of candidate therapies. These data contain samples from applications of many different drug combinations over many years. The evolving trends in treating HIV patients result in a highly unbalanced representation of different therapies in the available clinical data sets: while for some therapies many samples exist, for others there are very few. This might negatively affect the usefulness of the statistical models derived from such data sets. Furthermore, due to the large number of possible combination therapies and the introduction of new antiretroviral agents, for many combinations no samples are available at all.
In this chapter we present a multi-task learning approach that considers each combination therapy as a separate task. It compensates for the lack of samples for some therapies by basing its predictions also on samples from related therapies. This approach was initially presented in Bickel et al. (2008).

3.1 Problem Setting

We tackle the problem of predicting outcomes of HIV combination therapies by considering each therapy a separate task in a *multi-task learning setting*. In this setting, each of several tasks z is characterized by an unknown joint distribution $p(\mathbf{x}, y|z)$ of input features $\mathbf{x}$ and label y given the task z. The joint distributions of different tasks may differ arbitrarily, but usually the tasks are related so they have similar joint distributions. Let $D = \{(\mathbf{x}_1, y_1, z_1), \ldots, (\mathbf{x}_m, y_m, z_m)\}$ denote the training set comprising samples from all tasks. There may be tasks with no data. For each sample, the input features $\mathbf{x}_i$, class label y_i, and its associated task z_i are known. The training set D is governed by the mixed joint density $p(\mathbf{x}, y, z) = p(z)p(\mathbf{x}, y|z)$. The prior $p(z)$ specifies the task proportions.
In the HIV therapy screening application each combination therapy is considered a task z, and each task has an associated binary vector $\mathbf{z}$ that indicates the individual drugs comprising the therapy. Since drug combinations often share a subset of identical drugs, or a subset of different drugs with similar mechanisms of action, they can be considered as related tasks. The input features $\mathbf{x}$ comprise the viral genotype and the drug history for the specific therapy sample. The input is represented with a binary vector, where the part corresponding to the viral genotype indicates the occurrence of a set of resistance-relevant mutations (Johnson et al., 2007), and the part corresponding to the drug history comprises

the drugs known to be part of previous therapies. The binary class label y indicates the success (1) or failure (-1) of each sample therapy.
The goal is to train a classifier $f_z : \mathbf{x} \mapsto y$ that correctly predicts the outcome for an HIV combination therapy z. This classifier should minimize the expected loss

$$\mathbf{E}_{(\mathbf{x},y)\sim p(\mathbf{x},y|z)}[\ell(f_z(\mathbf{x}), y)]$$

with respect to the unknown joint distribution $p(\mathbf{x}, y|z)$ for each therapy z.

3.2 Related Work

A straightforward approach for multi-task learning is to train an independent model for each target task t by using only the data associated with the task $D_t = \{(\mathbf{x}_i, y_i, z_i) \in D : z_i = t\}$. The other extreme is a *one-size-fits-all model* $f_*(\mathbf{x})$ trained on the data from all tasks.
In many applications, task-level descriptions or prior knowledge on task similarity encoded in a kernel are available. Bonilla et al. (2007) study an extension of the one-size-fits-all model and find that training with a kernel defined as the multiplication of an input feature kernel and a task-level kernel outperforms a gating network. Task-level features have also been utilized for task clustering and for a task-dependent prior on the model parameters (Bakker and Heskes, 2003).
Another simple extension to the one-size-fits-all model would be to train a model for a target task from all data with weighted examples from other tasks, using one fixed uniform weight for each task. Such a model is described in Wu and Dietterich (2004).
Our work is inspired by learning under covariate shift (see Section 2.4 in Chapter 2). In the covariate shift setting the marginals $p_{train}(\mathbf{x})$ and $p_{test}(\mathbf{x})$ of training and test distributions differ, but the conditionals are identical $p_{train}(y|\mathbf{x}) = p_{test}(y|\mathbf{x})$. If training and test distributions were known, then the loss on the test distribution could be minimized by weighting the loss on the training distribution with an instance-specific factor. Shimodaira (2000) illustrates that the scaling factor has to be:

$$\frac{p_{test}(\mathbf{x})}{p_{train}(\mathbf{x})}. \tag{3.1}$$

Bickel et al. (2007) derive a discriminative expression for this marginal density ratio that can be estimated – without estimating the potentially high-dimensional densities of training and test distributions – by discriminating training against test data.
Hierarchical Bayesian models for multi-task learning are based on the assumption that task-specific model parameters are drawn from a common prior. The task dependencies are captured by estimating the common prior. Yu et al. (2005) impose a normal-inverse Wishart hyperprior on the mean and covariance of a Gaussian process prior that is shared by all task-specific regression functions. Mean and covariance of the Gaussian process are estimated using the Expectation Maximization (EM) algorithm (Hastie et al., 2009). A Dirichlet process can serve as prior in a hierarchical Bayesian model and cluster the tasks (Xue et al., 2007); all tasks in one cluster share the same model parameters. Evgeniou and Pontil (2004) derive a kernel that is based on a hierarchical Bayesian model with Gaussian prior (covariance matrix is scalar) on the parameters of a regularized regression.

Various statistical learning methods, including artificial neural networks, decision trees, random forests, support vector machines (SVMs) and logistic regression (Lathrop and Pazzani, 1999; Wang et al., 2003; Prosperi et al., 2005; Altmann et al., 2007; Larder et al., 2007; Rosen-Zvi et al., 2008; Altmann et al., 2009a; Prosperi et al., 2009), have been used to predict the virological response to HIV combination therapies. All these methods supply the drugs comprising the corresponding therapy as part of the input feature space.

3.3 Methods

The most accurate model for the target task t, the so-called target model, is the one that minimizes the loss with respect to the conditional probability $p(\mathbf{x}, y|t)$. The straightforward way to do this minimization is to fit a model by only using the portion of samples from the training set pertaining to the therapy t. However, for most therapies the number of available samples is not sufficient for generating accurate individual therapy models. Therefore, we exploit the available data from all therapies distributed according to the sample density $\sum_z p(z)p(\mathbf{x}, y|z)$ to train the target model for therapy t as follows. Each sample in the training set D is assigned a therapy-specific weight $r_t(\mathbf{x}, y)$ such that the training distribution $\sum_z p(z)p(\mathbf{x}, y|z)$ is matched to the target distribution $p(\mathbf{x}, y|t)$. In this way, the weighted sample is governed by the correct target distribution, but is still larger as it draws from the sample pool for all tasks. Formally, the expected loss with respect to the training distribution $\sum_z p(z)p(\mathbf{x}, y|z)$ weighted by $r_t(\mathbf{x}, y)$ equals the expected loss with respect to the target distribution $p(\mathbf{x}, y|t)$:

$$\mathbf{E}_{(\mathbf{x},y)\sim p(\mathbf{x},y|t)}[\ell(f(\mathbf{x}, t), y)] = \mathbf{E}_{(\mathbf{x},y)\sim\sum_z p(z)p(\mathbf{x},y|z)}\left[r_t(\mathbf{x}, y)\ell(f(\mathbf{x}, t), y)\right]. \tag{3.2}$$

In the following we will show that the equation above holds if:

$$r_t(\mathbf{x}, y) = \frac{p(\mathbf{x}, y|t)}{\sum_z p(z)p(\mathbf{x}, y|z)}$$

.

$$\begin{aligned}
&\mathbf{E}_{(\mathbf{x},y)\sim p(\mathbf{x},y|t)}[\ell(f(\mathbf{x}, t), y)] && (3.3)\\
&= \int \frac{\sum_z p(z)p(\mathbf{x}, y|z)}{\sum_{z'} p(z')p(\mathbf{x}, y|z')} p(\mathbf{x}, y|t)\ell(f(\mathbf{x}, t), y)d\mathbf{x}dy \\
&= \int \sum_z \left(p(z)p(\mathbf{x}, y|z)\frac{p(\mathbf{x}, y|t)}{\sum_{z'} p(z')p(\mathbf{x}, y|z')}\ell(f(\mathbf{x}, t), y)\right) d\mathbf{x}dy && (3.4)\\
&= \mathbf{E}_{(\mathbf{x},y)\sim\sum_z p(z)p(\mathbf{x},y|z)}\left[\frac{p(\mathbf{x}, y|t)}{\sum_{z'} p(z')p(\mathbf{x}, y|z')}\ell(f(\mathbf{x}, t), y)\right] && (3.5)
\end{aligned}$$

In the derivation above Equation 3.3 expands the expectation and introduces a fraction that equals one. Then, Equation 3.4 expands the sum over z in the numerator to run over the entire expression because the integral over $(\mathbf{x}, y)$ is independent of z. Finally, Equation 3.5 is the expected loss over the distribution of all tasks weighted by $\frac{p(\mathbf{x},y|t)}{\sum_z p(z)p(\mathbf{x},y|z)}$.
Equation 3.5 signifies that the model for task t that minimizes the expected loss with respect to the target distribution $p(\mathbf{x}, y|t)$ can be trained by minimizing the expected loss over the distribution of all tasks weighted by $r_t(\mathbf{x}, y)$.

What remains is the problem of estimating the density ratio:

$$r_t(\mathbf{x}, y) = \frac{p(\mathbf{x}, y|t)}{\sum_z p(z)p(\mathbf{x}, y|z)}. \tag{3.6}$$

Obtaining estimators for the high-dimensional densities $p(\mathbf{x}, y|t)$ and $\sum_z p(z)p(\mathbf{x}, y|z)$ directly is a difficult modeling task. In order to avoid such estimation we derive a discriminative model that directly evaluates the resampling weights $r_t(\mathbf{x}, y)$ as follows. We reformulate the density ratio $\frac{p(\mathbf{x},y|t)}{\sum_z p(z)p(\mathbf{x},y|z)}$ in terms of a conditional model $p(t|\mathbf{x}, y)$:

$$\begin{aligned} r_t(\mathbf{x}, y) &= \frac{p(\mathbf{x}, y|t)}{\sum_z p(z)p(\mathbf{x}, y|z)} && (3.7)\\ &= \frac{p(t|\mathbf{x}, y)p(\mathbf{x}, y)}{p(t)} \frac{1}{\sum_z p(z)\frac{p(z|\mathbf{x},y)p(\mathbf{x},y)}{p(z)}} && (3.8)\\ &= \frac{p(t|\mathbf{x}, y)}{p(t)\sum_z p(z|\mathbf{x}, y)} && (3.9)\\ &= \frac{p(t|\mathbf{x}, y)}{p(t)} && (3.10) \end{aligned}$$

The derivation above underlies the assumption that the prior on the size of the target sample is greater than zero, $p(t) > 0$. In Equation 3.8 Bayes' rule is applied twice and in Equation 3.9 $p(\mathbf{x}, y)$ and $p(z)$ are canceled out. Equation 3.10 follows by $\sum_z p(z|\mathbf{x}, y) = 1$ and shows how the resampling weights $r_t(\mathbf{x}, y) = \frac{p(\mathbf{x},y|t)}{\sum_z p(z)p(\mathbf{x},y|z)}$ can be determined without any knowledge of the task densities $p(\mathbf{x}, y|z)$.

Intuitively, the conditional $p(t|\mathbf{x}, y)$ quantifies the probability that a sample $(\mathbf{x}, y)$ randomly drawn from the training set D of samples of all therapies belongs to the target therapy t, *i.e.* how much more likely $(\mathbf{x}, y)$ is to occur in the target distribution than it is to occur in the mixture distribution of all tasks. This conditional probability can be estimated with a model that discriminates the labeled samples of the target therapy from the labeled samples of all therapies.

We realize this with a multi-class version of logistic regression (Hastie et al., 2009), the so called soft-max model, with model parameters $\mathbf{v}$ that estimates the discriminative models for all therapies in the training set simultaneously. The model parameter $\mathbf{v}$ is a concatenation of the therapy-specific vectors $\mathbf{v}_z$, one for every therapy z. With this model an estimate for $p(t|\mathbf{x}, y)$ is given by the evaluation of the soft-max model with respect to task t, *i.e.* $p(z = t|\mathbf{x}, y, \mathbf{v})$. Formally, the soft-max model is given by:

$$p(z|\mathbf{x}, y, \mathbf{v}) = \frac{\exp(\mathbf{v}_z^\mathsf{T}\Phi(\mathbf{x}, y))}{\sum_{z'} \exp(\mathbf{v}_{z'}^\mathsf{T}\Phi(\mathbf{x}, y))} \tag{3.11}$$

where Equation 3.11 requires a problem-specific feature mapping $\Phi(\mathbf{x}, y)$. Without loss of generality we define this mapping for binary labels $y \in \{+1, -1\}$ as:

$$\Phi(\mathbf{x}, y) = \begin{bmatrix} \delta(y, +1)\Phi(\mathbf{x}) \\ \delta(y, -1)\Phi(\mathbf{x}) \end{bmatrix} \tag{3.12}$$

where δ is the Kronecker delta ($\delta(a, b) = 1$, if $a = b$, and $\delta(a, b) = 0$, if $a \neq b$). In the absence of prior knowledge about the similarity between the successful and failing

combination therapies, input features $\mathbf{x}$ of samples with different class labels y are mapped to disjoint subsets of the feature vector. With this feature mapping the models for positive and negative examples do not interact and can be trained independently.
The soft-max model is trained by maximizing the regularized log-likelihood of the training data:

Optimization Problem 6. *Over parameters* $\mathbf{v}$*, maximize*

$$\sum_{(\mathbf{x}_i,y_i,z_i)\in D} \log(p(z_i|\mathbf{x}_i,y_i,\mathbf{v})) - \mathbf{v}^\mathsf{T}\Sigma^{-1}\mathbf{v}.$$

The solution of Optimization Problem 6 is a *maximum a posteriori* (MAP) estimation of the soft-max model (Equation 3.11) over the model parameters $\mathbf{v}$ using a Gaussian prior $N(\mathbf{0},\mathbf{\Sigma})$ on the parameter vector.
Available prior knowledge on the similarity of tasks, represented as a positive semi-definite kernel function $k(z,z')$, can be encoded in the covariance matrix Σ of the Gaussian prior $N(0,\Sigma)$. We set all main diagonal entries of Σ to the scalar parameter $\sigma^2_\mathbf{v}$ and set the secondary diagonal entries corresponding to the covariances between $\mathbf{v}_z$ and $\mathbf{v}'_z$ to $k(z,z')\rho\sigma^2_\mathbf{v}$ (assuming kernel values $0 \leq k(z,z') \leq 1$). Parameter $\sigma^2_\mathbf{v}$ specifies the variance of each element in $\mathbf{v}$. $k(z,z')\rho$ is the correlation coefficient between elements of subvectors $\mathbf{v}_z$ and $\mathbf{v}'_z$; parameter ρ specifies the strength of this correlation. The covariance matrix Σ is required to be invertible and therefore $0 \leq \rho < 1$. All other entries of Σ are set to zero.
When prior knowledge on the therapy similarities is encoded in the prior on the model parameters, then this prior knowledge dominates the optimization criterion for small samples (*e.g.* therapies with very few available samples) while the data-driven portion of the criterion becomes dominant and overrides prior beliefs as more data arrives. Furthermore, for therapies with no training examples the Gaussian prior with the task kernel $k(z,z')$ encoded in the covariance matrix determines the model.
Usually there are several hundred different therapies. In order to obtain good predictions we need a non-linear version of the soft-max model. Therefore, we use a kernelized variant of Optimization Problem 6 by applying the representer theorem. Details on the kernelization of multi-class logistic regression are found in Zhu and Hastie (2002) and section 2.4 in Chapter 2.
From the results of Optimization Problem 6 we can obtain the sample weights $r_t(x,y)$ for the target therapy t. Using these weights we can evaluate the expected loss over the weighted training data as displayed in Equation 3.13:

$$\mathbf{E}_{(\mathbf{x},y)\sim D}\left[\frac{p(t|\mathbf{x},y,\mathbf{v})}{p(t)}\ell(f(\mathbf{x},t),y)\right] + \frac{\mathbf{w}_t^\mathsf{T}\mathbf{w}_t}{2\sigma^2_\mathbf{w}}. \tag{3.13}$$

We can train the final model for the target therapy t by minimizing the weighted regularized loss (Equation 3.13) over the training samples. This is realized with a standard logistic regression model with a Gaussian log-prior with variance $\sigma^2_\mathbf{w}$ on the parameters $\mathbf{w}_t$:

Optimization Problem 7. *For task* t*: over parameters* $\mathbf{w}_t$*, minimize*

$$\frac{1}{|D|}\sum_{(\mathbf{x}_i,y_i)\in D}\frac{p(t|\mathbf{x}_i,y_i,\mathbf{v})}{p(t)}\ell(f(\mathbf{x}_i,\mathbf{w}_t),y_i) + \frac{\mathbf{w}_t^\mathsf{T}\mathbf{w}_t}{2\sigma^2_\mathbf{w}}.$$

In the Optimization Problem 7 each example is weighted by the discriminatively estimated density fraction from Equation 3.10 using the solution of Optimization Problem 6. An instance of this problem is solved for each task independently to produce a separate model for this task.
To summarize, our approach trains an individual model for each therapy by using the available data from all therapies with proper sample weights. This enables the therapy-specific model to exploit data from related therapies and base its predictions on samples relevant for the target therapy. The method is summarized in Algorithm 1.

Algorithm 1: Multi-task learning method

Input: Training data D and sample $\mathbf{x}$ associated with a target therapy t.

1. Estimate sample weights $r_t(\mathbf{x}, y)$ – train a discriminative model for $p(t|\mathbf{x}, y)$ that discriminates labeled instances of the target therapy against labeled instances of the pool of samples for all therapies.
2. Use the weights $r_t(\mathbf{x}, y)$ to estimate the final model for the target therapy t – regularized logistic regression model that minimizes the weighted loss on the training data D.

3.4 HIV Therapy Screening

In this chapter we describe an approach that models the problem of HIV therapy screening in a multi-task learning framework, where each antiretroviral therapy is considered a separate task. In the next subsections we describe the clinical data sets, the validation setting, the reference methods, and the empirical results of the computational experiments.

3.4.1 Clinical Data Sets

The training data are extracted from the EuResist database that contains information on 52846 antiretroviral therapy samples administered to 16999 HIV-1 (subtype B) patients from several countries in the period from 1988 to 2007. This information includes the individual drugs that comprise a therapy, virus load measurements (copies of viral RNA per ml blood plasma, cp/ml) during the course of a therapy, all available therapies administered to each patient, as well as consensus sequences of the predominant viral strains in the patients' blood. We include a therapy as a sample in the training data if there is a viral sequence obtained shortly before the therapy was started (up to 90 days before) and if it can be assigned a label (success or failure) based on the virus load values measured during its course.
We use two different definitions of therapeutic success and failure to label the data: *virus load labeling* and *multi-conditional labeling*. According to the virus load labeling represented in Figure 3.1 (a), a therapy is successful if the viral load drops below 400 cp/ml during the time of the treatment. Otherwise the treatment is a failure. In the multi-conditional

labeling illustrated in Figure 3.1 (b), a therapy is successful if at least one of the following conditions is fulfilled:

- the viral load measured in the time range between 28 and 84 days after the start of the therapy decreases by at least two orders of magnitude compared to the most recent viral load measured one to three months before the start of the therapy,
- the viral load drops below 400 cp/ml 56 days after the start of the therapy.

A drawback of this definition is that due to the strict time intervals it imposes on the measurements, class labels that adhere to this labeling are only available for a small number of records. The virus load labeling does not require these strict time intervals by making use of any viral load measurement during the course of therapy to label it.

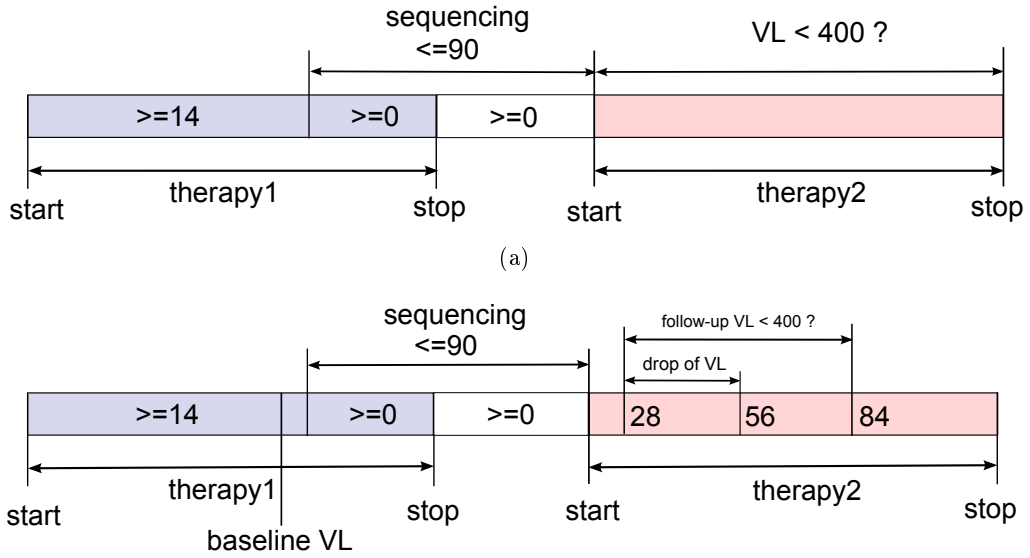


Figure 3.1: Assigning a label and a viral sequence to *therapy2*, where *therapy1* and *therapy2* are two consecutive therapy administered to a patient: (a) the virus load labeling, and (b) the multi-conditional labeling.

We extract two types of features for each therapy sample: a genotypic description of the virus and information about the treatment history of the patient. We use the viral genotype taken from the patient shortly before the therapy start (up to 90 days before) and represent it by a binary vector indicating the presence of any from a set of predefined resistance-relevant mutations. These mutations are derived from the list in Johnson et al. (2007). Drug-resistant viral quasi-species evolve during the course of the therapy due to selective pressure imposed by the drugs. As they remain in the patient's body, the treatment history plays an important role for predicting the outcome of a potential treatment. Hence, we extract all drugs given to the patient in all known previous treatments and use a binary vector representation indicating the occurrence of the drugs given to the patient in the treatment history. The 82-dimensional feature vector $\mathbf{x}$ for each data point results from the concatenation of 65 genotypic and 17 treatment history features.

Finally, out of all available treatment records we extract two different data sets using the two labelings. With the virus load labeling we extract 3260 and with the multi-conditional labeling 2011 treatment records with corresponding ratios of 65.7% and 64.1% successful treatments. The size of these data sets is much smaller than the size of the original data due to missing viral load measurements, or missing virus sequence information.
Figure 3.2 depicts a histogram of the frequencies of the different combination therapies in the two training data sets. A number of 545 distinct drug combinations (tasks z) occur at least once in the virus load data set; 433 occur in the multi-conditional data set. For many combinations, only a few examples occur in the data. For instance, in the virus load data set we observe 253 out of 545 drug combinations with only one data point and 411 with less than five instances. Similarly, the multi-conditional data set has 213 out of 433 drug combinations with a single data point and 331 with less than five observations.

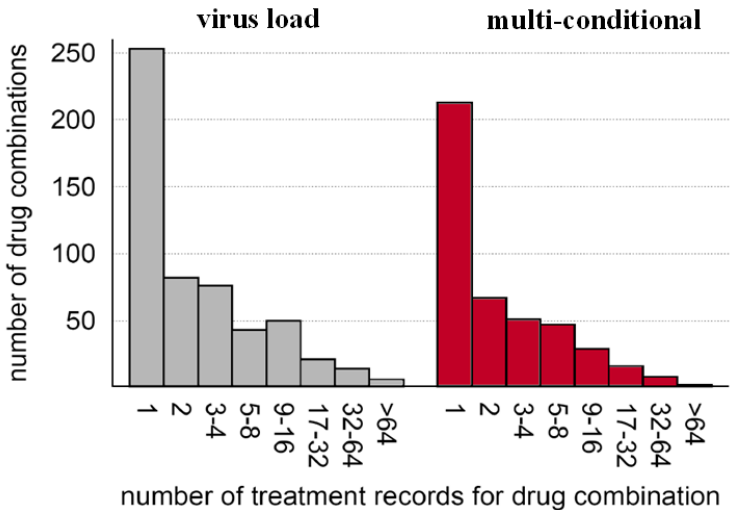


Figure 3.2: Histogram over number of treatment records for drug combinations (tasks) in the virus load data set (gray) and multi-conditional data set (red).

3.4.2 Prior Knowledge on Therapy Similarity

We encode the prior knowledge about the similarity of different drug combinations with two kernels: the *drug indicator kernel*, and the *mutation table kernel*.
The drug indicator kernel is based on the number of common drugs that two combination therapies share. Let $\mathbf{u}_z$ and $\mathbf{u}_{z'}$ be binary vectors indicating the distinct drugs comprising the therapies z and z', respectively. The similarity $k_d(z, z')$ between the combination therapies z and z' is given by:

$$k_d(z, z') = \frac{\mathbf{u}_z^T \mathbf{u}_{z'}}{max(\|\mathbf{u}_z\|^2, \|\mathbf{u}_{z'}\|^2)}. \tag{3.14}$$

where $\mathbf{x}^T\mathbf{y}$ is the scalar product of the vectors $\mathbf{x}$ and $\mathbf{y}$ and $\|\cdot\|$ is the L_2-norm. According to this kernel the more drugs therapies z and z' have in common the higher their similarity. Its values are in the $[0, 1]$-interval.
The *mutation table kernel* uses the table of resistance-associated mutations of each drug afforded by the International AIDS society (Johnson et al., 2007). First, we construct

binary vectors indicating resistance-relevant mutations for the set of drugs occurring in a combination. Then, in the same way as the drug indicator kernel, the mutation table kernel computes the normalized inner product between such binary vectors for two drug combinations.

3.4.3 Validation Setting and Reference Methods

Reference methods. The first reference method is training of a separate logistic regression model for each task without any interaction ("separate"). Tasks without any training examples get a constant classifier that assigns each test example with 50% to each of both classes.

The next baseline is a one-size-fits-all model; all examples are pooled and only one common logistic regression is trained for all tasks ("pooled"). For the experiments with prior knowledge on task similarity we multiply the feature kernel with the task kernel values $k(\mathbf{x}, \mathbf{x}')(k(z, z') + 1)$ and train one model using this kernel (Bonilla et al., 2007). For task kernels that can have a value of zero we include a "+1" term to ensure that the feature kernel does not vanish.

The third reference method ("hier. Bayes kernel") is a logistic regression with the hierarchical Bayesian kernel of Evgeniou and Pontil (2004):

$$k_{hBayes}(\mathbf{x}, \mathbf{x}') = (\lambda + \delta(z, z'))k(\mathbf{x}, \mathbf{x}'), \tag{3.15}$$

where $\delta(z, z')$ is the Kronecker delta and λ is a tuning parameter. For the experiments with task similarity kernel the hierarchical Bayes and the task kernel are multiplied. As second hierarchical Bayesian method ("hier. Bayes Gauss. proc.") we use the Gaussian process regression of Yu et al. (2005).

Time-oriented validation scenario. The trends of treating HIV patients change over time as a result of the gathered practical experience with the drugs and the introduction of new antiretroviral drugs. In order to account for this phenomenon we use a *time-oriented validation scenario* which makes a time-oriented split when selecting the training and the test set. First, we order all available training samples by their corresponding therapy starting dates. We then make a time-oriented split by selecting the most recent 20% of the samples as the test set and the rest as the training set. This procedure is depicted in Figure 3.3 and yields 653 and 403 test examples for the virus load and multi-conditional data set, respectively. For the model selection we split the training set further in a similar manner. We take the most recent 25% of the training set for selecting the best model parameters and refer to this set as tuning set. We use it to tune the prior and regularization parameters of all methods, the Dirichlet parameter γ, and the variance of the RBF kernels. In this way, our models are trained on the data from the more distant past, while their performance is measured on the data from the more recent past.

The performance of all considered approaches is quantified by the accuracy of predicting the correct label (success or failure of a treatment) on the test set. In order to compare the accuracies of two methods on a separate test set, the significance of the difference of two accuracies are calculated based on a paired t-test. In the following we explain the details of these calculations.

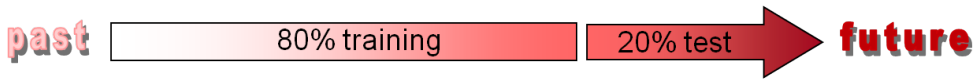


Figure 3.3: Time-oriented validation scenario. The arrow depicts the therapy starting times of the therapy samples.

Let S and T be the training and test data of a binary classification problem, respectively. Let A and B denote two classification methods trained on S and let $\mathbf{v}^A$ denote a binary vector with length $|T|$ such that:

$$v_i^A = \begin{cases} 1, & \text{if method A correctly predicts the outcome for the i-th test sample;} \\ 0, & \text{otherwise.} \end{cases} \tag{3.16}$$

$\mathbf{v}^B$ has the same definition as $\mathbf{v}^A$, but for method B. Then, $\mathbf{v}^A$ and $\mathbf{v}^B$ can be considered as realizations of two Bernoulli random variables. The significance of their difference can be evaluated by using a paired t-test and the standard error of this difference is calculated by:

$$SE(\mathbf{v}^A - \mathbf{v}^B) = \frac{sd(\mathbf{v}^A - \mathbf{v}^B)}{\sqrt{|T|}}, \tag{3.17}$$

where sd stands for standard deviation. The standard error of a single method (for example the method A) is given by:

$$SE(\mathbf{v}^A) = \frac{sd(\mathbf{v}^A)}{\sqrt{|T|}}. \tag{3.18}$$

3.4.4 Experimental Results and Discussion

In our experiments we study the benefit of distribution matching for HIV therapy screening compared to the reference methods described in the previous subsection. Optimization Problem 6 is solved with limited-memory BFGS (see Chapter 4) and Optimization Problem 7 with Newton gradient descent using a logistic loss. We use RBF kernels for all methods. For the prior term $p(t)$ required in Optimization Problem 7 we use a maximum a posteriori (MAP) estimate with a symmetric Dirichlet prior given by:

$$\frac{|D_t| + \gamma}{\sum_z (|D_z| + \gamma)}. \tag{3.19}$$

Table 3.1 summarizes the prediction accuracies for all methods over both data sets without and with the two different types of prior knowledge on combination similarity. The columns "ste.Δ" placed next to the accuracy columns display the standard errors of the differences to the distribution matching method.

Multi-task learning by distribution matching outperforms, or is as good as, the best alternative method in all cases. The improvement over the separate model baseline is about $10-14\%$. We can reject the null hypothesis that the pooled and the hierarchical Bayesian kernel baseline is at least as accurate as distribution matching in four and five cases respectively out of six according to a paired t-test at significance level $\alpha = 0.05$.

As can be observed in Figure 3.4, prior knowledge does not improve the accuracy for the distribution matching method. The pooled baseline benefits from prior knowledge for

Table 3.1: Classification accuracies with standard errors of differences to the distribution matching method (ste.Δ). Symbols (•,○,∗,◇) indicate statistical significance according to a paired t-test with significance level $\alpha = 0.05$, (•) compared to separate baseline, (○) compared to pooled baseline, (∗) compared to hierarchical Bayesian kernel baseline, (◇) compared to hierarchical Bayesian Gaussian process baseline.

virus load data set

method	prior					
	none	ste.Δ	drugs	ste.Δ	mutations	ste.Δ
separate	67.87%	1.80	67.87%	1.76	67.87%	1.78
pooled	75.00%	1.47	75.46%	1.39	75.61%	1.37
hierarchical Bayes	76.69%	1.39	75.31%	1.34	76.84%	1.16
hierarchical Bayes GP	76.53%	1.36				
distribution matching	• ○ ∗ ◇ **79.14%**		• ○ ∗ **77.91%**		• ○ ∗ **79.29%**	

multi-condition data set

method	prior					
	none	ste.Δ	drugs	ste.Δ	mutations	ste.Δ
separate	64.64%	2.41	64.64%	2.29	64.64%	2.38
pooled	76.67%	1.13	**78.41%**	1.63	**78.66%**	1.11
hierarchical Bayes	77.17%	1.29	75.19%	1.44	**77.42%**	1.24
hierarchical Bayes GP	76.43%	1.44				
distribution matching	• ○ ∗ ◇ **79.40%**		• ∗ **78.16%**		• **79.16%**	

the multi-condition data set. For the case without prior knowledge we do not observe a statistically significant difference of the two hierarchical Bayesian methods, but they are both significantly worse than distribution matching according to the paired t-test. Note that the Gaussian process baseline is a regression model; all other methods are classification models. Furthermore, we do not use any prior knowledge on therapy similarity in the Gaussian process model, since it is unclear how to formally introduce such knowledge in this model.

Table 3.2: Sample counts in the bins grouping the test samples based on their corresponding number of available training instances.

data set	**multi-condition**				**virus load**			
bin	$0-1$	$2-5$	$6-20$	>20	$0-2$	$3-9$	$10-38$	>38
count	87	112	96	108	164	162	177	149

Figures 3.5 and 3.6 display the accuracy over the combinations in the test set grouped by the number of available examples for the settings without and with the mutation table kernel. For instance, an accuracy of 74% for the first group $0-2$ means, that only test

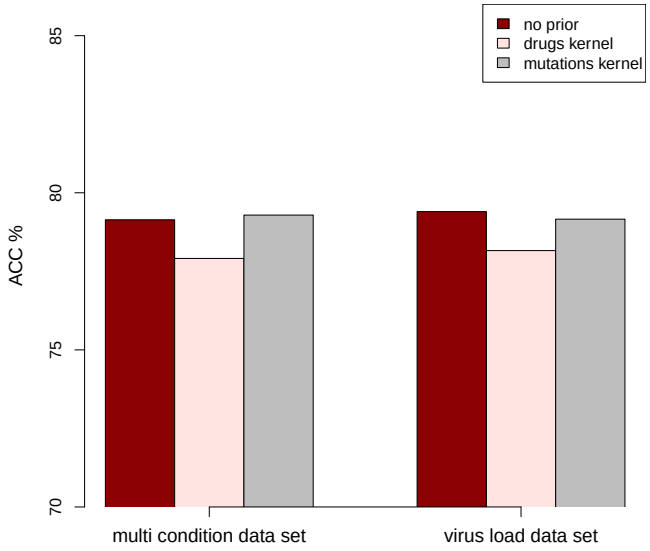
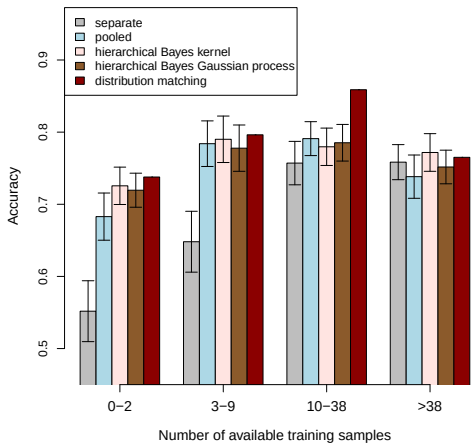


Figure 3.4: Classification accuracies for the distribution matching method.

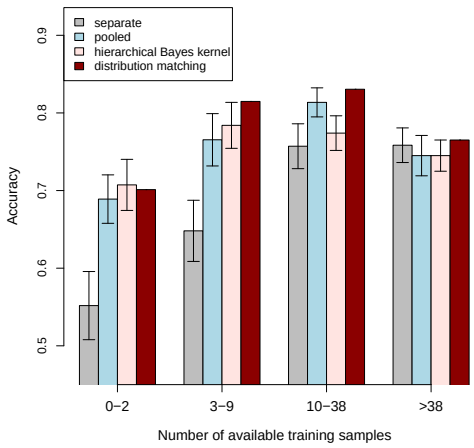
examples from combinations are selected that have zero, one, or two training examples each, and the accuracy on this subset of the test examples is 74%. The error bars indicate the standard error of the differences to the distribution matching method. Note, that the statistical tests described above are based on all test data and are not directly related to the group-specific error bars in the diagrams. The sample counts on each of the four groups for the two considered data sets are given in Table 3.2.
All methods benefit from larger numbers of training examples per drug combination. The slightly decreasing accuracy for the virus load data set for test samples with more than 38 training examples is surprising. Further analysis reveals that in this case there is an accumulation of test examples with history profiles very different from the training examples of the same combination.
For all methods that generalize over the tasks the benefit compared to the separate model baseline is the largest for the smallest group ($0-2$ and $0-1$ training examples respectively). It is worth noting that the approach of training individual models for each drug combination also remedies potential deviations that can occur in the HIV treatment patterns originating from different countries. For example, Figure 3.7 depicts slight differences between the HIV treatment patterns from Italy and Germany for our clinical data set.

3.5 Conclusions

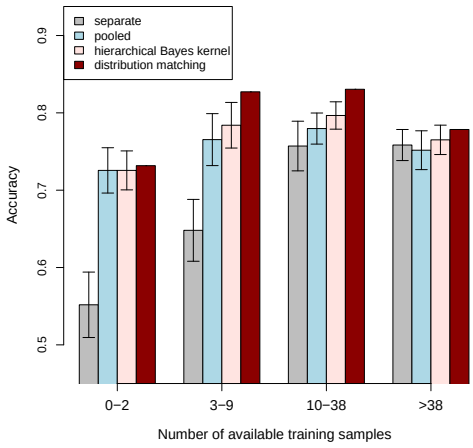
In this chapter, we devised a multi-task learning method that centers around resampling weights which match the distribution of the pool of examples of multiple tasks to the target distribution for a given task at hand. The method creates a weighted sample that reflects the desired target distribution and exploits the entire corpus of training data for all tasks. We showed how appropriate weights can be obtained by discriminating the labeled sample



(a) no prior



(b) drugs kernel



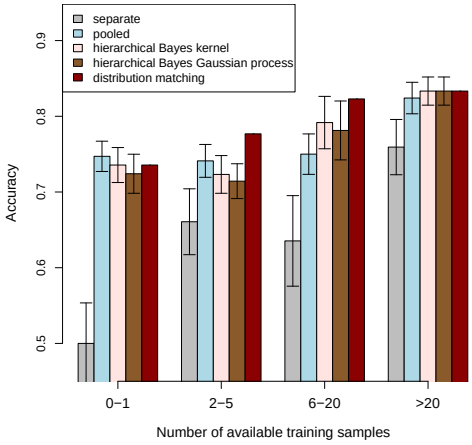
(c) mutations kernel

Figure 3.5: Accuracies for all considered methods over different number of training examples for test therapy sample for the virus load data set. Error bars indicate the standard error of the differences to the distribution matching method.

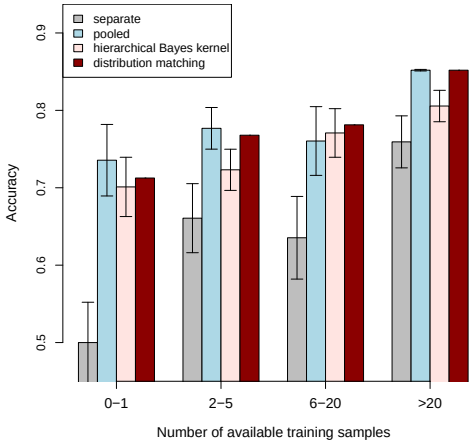
for a given task against the pooled sample. After weighting the pooled sample, a classifier for the given task can be trained.

In our experiments on HIV therapy screening we found that the distribution matching method improves on the prediction accuracy over independently trained models by $10-14\%$. According to a paired t-test, distribution matching is significantly better than the reference methods for 17 out of 20 experiments.

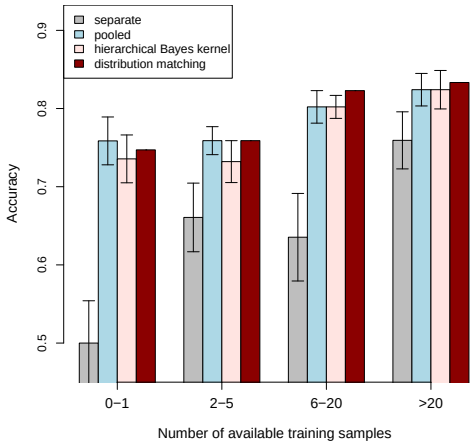
A combination of drugs is the standard way of treating HIV patients. The accuracy to which the likely outcome of a combination therapy can be anticipated can therefore directly impact the quality of HIV treatments.



(a) no prior



(b) drugs kernel



(c) mutations kernel

Figure 3.6: Accuracies for all considered methods over different number of training examples for test therapy sample for the multi-condition data set. Error bars indicate the standard error of the differences to the distribution matching method.

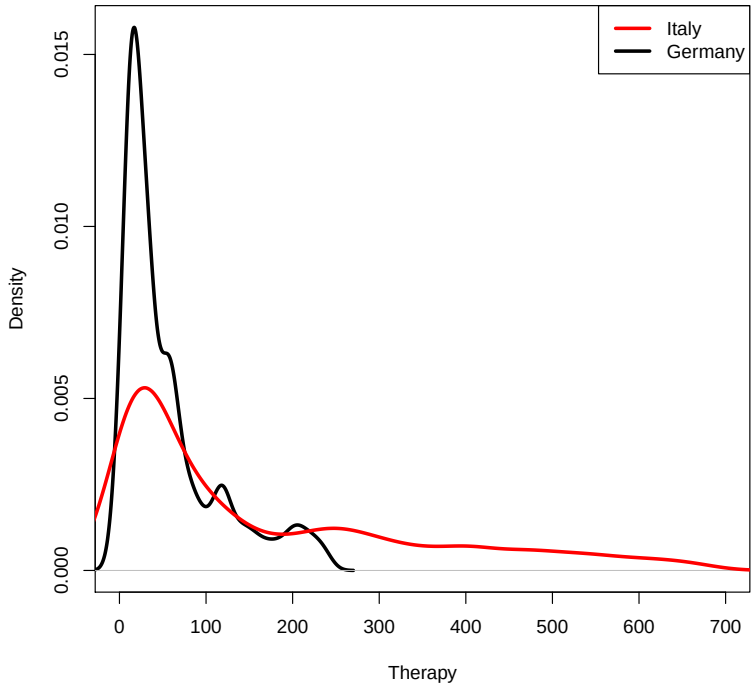


Figure 3.7: Distribution of the different combination therapies in the Italian and German subsets of our clinical data set. The numbers on the x-axis represent the different therapy combinations ordered by their first appearance in our clinical data: from older to newer. The y-axis depicts the density.

4 Therapy-similarity Method for Predicting Effectiveness of HIV Therapies

The multi-task learning approach for HIV therapy screening presented in the previous chapter trains a separate model for each combination therapy from all available samples with properly derived sample weights. The weights are computed by matching the distribution of all samples to the distribution of the samples of each individual therapy. In this way it tackles the highly unbalanced representation of different therapies in the available clinical data sets and compensates for the lack of samples for the rare therapies by basing their predictions also on samples from related therapies. However, this method computes the similarities between therapies and the prediction model in a single integrated procedure. While being statistically sound, the method is also quite compute-intensive, as it involves a multi-class logistic regression with as many classes as there are therapies (usually several hundred).

In this chapter we follow this line of research and present a somewhat more heuristic but much more efficient and configurable alternative model, referred to as *therapy-similarity model*, which also utilizes information from similar therapies to train an individual model for each target therapy. However it separates the computation of the therapy similarities in a preprocessing step from the estimation of the model. First of all, this reduces the computation time of the therapy similarities to a few seconds. Secondly, it has the advantage that the resulting similarities can be reconfigured at the users discretion and then used to train a new model. Finally, by training a separate model for each therapy by using data from similar drug combinations, the therapy-similarity approach balances the uneven therapy representation in the data sets and produces higher quality models for drug combinations with very few training samples. This model was initially presented in Bogojeska et al. (2010).

4.1 Methods

The therapy-similarity approach to the problem of predicting outcomes of HIV combination therapies trains a separate model for each of them by using the viral genotypic information from similar therapies. This is done in the model-fitting procedure by using precomputed weights that up-weight samples originating from therapies similar to the therapy of interest. In this way the separate models are tuned to focussing on information coming from drug combinations akin to the target therapy. The mathematical concept of therapy similarity is governed by a specific predefined understanding of what similar therapies are. This paper describes and evaluates two different approaches for quantifying pairwise therapy similarity. The therapy-specific models that we describe in this chapter also offer the possibility of incorporating phenotypic information on the effectiveness of the individual

drugs comprising the therapy of interest. All this sums up to easy-to-interpret linear logistic regression models fitted for each individual drug combination with a learning procedure that takes advantage of additional information (similar therapies, phenotypic information) and therefore can deal with therapies that have only few training samples available.
Let $\mathbf{x}$ denote the input features that comprise the viral genotype encoded as a binary vector indicating the occurrence of a set of resistance-relevant mutations (Johnson et al., 2008). The drug combinations are denoted by z – each of them is represented by a binary vector that indicates the individual drugs administered in the combination. The binary class label y marks each therapy sample with *success* (1) or *failure* (-1). Let $D = \{(\mathbf{x}_1, z_1, y_1), \ldots, (\mathbf{x}_m, z_m, y_m)\}$ denote the training set and t denote the therapy of interest.
We model the problem of predicting the outcome of the therapy t with weighted linear logistic regression using a logarithm of a Gaussian prior with mean μ_t and isotropic covariance matrix $\sigma^2\mathbf{I}$ shared by all therapies (see Chapter 2 and Evgeniou et al. (2000)). The model parameters $\mathbf{w}_t$ for therapy t are obtained by solving the optimization problem given as follows.

Optimization Problem 8. *Over parameters* $\mathbf{w}_t$*, minimize*

$$\frac{1}{|D|} \sum_{(\mathbf{x}_i, z_i, y_i) \in D} k(z_i, t)^\gamma \cdot \ell(f(\mathbf{x}_i, \mathbf{w}_t), y) + \frac{(\mu_t - \mathbf{w}_t)^T(\mu_t - \mathbf{w}_t)}{2\sigma^2}. \tag{4.1}$$

$k(z_i, t)$ is a function that provides sample-specific weights that quantify the similarity of the therapy of interest t with the therapy z_i from the i-th sample (see Subsection 4.1.2), and γ is its smoothing parameter. The expression:

$$\ell(f(\mathbf{x}, \mathbf{w}_t), y) = \ln(1 + \exp(-y\mathbf{w}_t^T\mathbf{x})) \tag{4.2}$$

is the loss of linear logistic regression. μ_t is phenotypic prior knowledge on the outcome of therapy t as explained in Subsection 4.1.3. We will refer to this model as *therapy similarity model*. The large number of distinct therapies and our approach of training a separate model for each of them demand an efficient method for solving Optimization Problem 8. To achieve this, we apply a *trust region Newton method* for training logistic regression (Lin et al., 2008) that takes advantage of the sparseness of our feature space. This enables fast model fitting which results in efficient model selection. In what follows we first give a short overview of the optimization methods used for training large-scale logistic regression followed by a detailed description of the therapy similarity kernels and the prior phenotypic knowledge on the therapy outcomes.

4.1.1 Optimization methods for large-scale logistic regression

Many unconstrained optimization methods, such as iterative scaling (Darroch and Ratcliff, 1972; Pietra et al., 1997; Goodman, 2002; Jin et al., 2003), conjugate gradient (Nocedal and Wright, 2006), quasi-Newton (Dong and Nocedal, 1989; Benson and Moré, 2001) and truncated Newton (Komarek and A.W., 2005; Lin et al., 2008), have been used for training large-scale logistic regression. According to Malouf (2002); Sutton and McCallum (2006) the limited memory BFGS – a quasi-Newton approach that uses a limited memory variation of the Broyden - Fletcher - Goldfarb - Shanno (BFGS) update to approximate the

inverse Hessian matrix – is the most efficient method. However, for large and sparse high-dimensional training data sets Lin et al. (2008) show that for fitting logistic regression a trust region Newton method (truncated Newton method adapted from Lin and Moré (1999)) is more efficient than the quasi-Newton approach.

Newton's method. The Newton's method is an iterative procedure for finding roots of equations and critical points of twice-differentiable functions. Thus, for a given initial guess $\mathbf{x}^0$ and a twice-differentiable function $f(\mathbf{x})$ with gradient $\nabla f(\mathbf{x})$ and Hessian $\nabla^2 f(\mathbf{x})$ the Newton's update rule is given by:

$$\mathbf{x}^{k+1} = \mathbf{x}^k + \mathbf{p}^k \tag{4.3}$$

where k is the iteration index and the Newton direction (step) $\mathbf{p}^k$ is the solution of the system of linear equations:

$$\nabla^2 f(\mathbf{x}^k)\mathbf{p}^k = -\nabla f(\mathbf{x}^k). \tag{4.4}$$

In this form, the sequence $\mathbf{x}^k$ produced by the update rule in Equation 4.3 is not guaranteed to converge to an optimal solution. Convergence is assured by adjusting the length of the Newton direction in either a line search or a trust region framework. Moreover, when the number of features is large the Hessian cannot be explicitly stored in main memory. Therefore, for large-scale logistic regression, efficient approaches that do not require second derivatives, like the limited memory BFGS method (Dong and Nocedal, 1989), or efficient approaches that do not store the entire Hessian matrix, like the trust region Newton method (Lin et al., 2008), are more suitable. In the following we will briefly describe these two methods.

Limited memory LBFGS method. This method belongs to the class of quasi-Newton methods (Nocedal and Wright, 2006) where the gradient evaluations are utilized for approximating the inverse of the Hessian matrix in an iterative fashion. Unlike the popular BFGS update rule which stores the approximation of the entire Hessian matrix, the limited memory BFGS (Dong and Nocedal, 1989) uses only a small number of updates from the previous iterations (the most recent ones) to represent the approximation of the Hessian implicitly. Furthermore, using the most recent updates this method provides an efficient calculation of $\nabla^2 f(\mathbf{x}^k)\mathbf{p}^k$. These properties make the limited memory BFGS approach very well suited for problems with large number of features like the large-scale logistic regression problem. In order to apply it one needs to specify the objective function and its gradient.

Trust region Newton method. Truncated Newton methods are doubly iterative procedures where the outer iteration focuses on the optimization problem and the inner iteration computes the Newton direction. Conjugate gradients are the most used methods for finding the Newton direction. However, since they often produce lengthy iterations, in the early stages of the outer iteration the inner iteration is stopped (truncated) before the solution to the Newton equations is obtained. Global convergence is achieved by specifying certain stopping conditions for the inner iterations. Although they use an approximate Newton direction they still use the exact Hessian matrix and compared to the limited

memory quasi-Newton methods, which use approximate Hessian matrices, tend to have faster convergence. Further details on truncated Newton methods can be found in Nash (2000); Nocedal and Wright (2006).

In Lin et al. (2008) it is shown that a simplified version of the trust region method from Lin and Moré (1999), which is a truncated Newton method for bound-constrained optimization problems, is more efficient than a quasi-Newton method for training sparse large-scale logistic regression. For a given iterate $\mathbf{x}^k$ and a size of the trust region ∇_k this method first approximates the value $f(\mathbf{x}^k + \mathbf{p}) - f(\mathbf{x}^k)$ with a quadratic function $q_k(\mathbf{p})$. Then, it uses the conjugate gradient method to find the step $\mathbf{p}^k$ that minimizes $q_k(\mathbf{p})$ subject to the constraint $\|\mathbf{p} < \nabla_k\|$. Finally, the direction $\mathbf{p}^k$ is accepted by checking if the ratio between the actual reduction of the function to the predicted reduction in the quadratic model q_k is larger than a specified value. More details on the algorithm and the implementation can be found in Lin et al. (2008). Because of its efficiency for sparse large-scale linear logistic regression we use the trust region Newton method for solving the Optimization Problem 8. In order achieve this, beside the objective function given in Optimization Problem 8, we also need to specify its gradient and Hessian. Let L denote the objective function specified in Optimization Problem 8. Then the gradient has the form:

$$\frac{\partial L}{\partial \mathbf{w}_t} = \sum_{i=1}^{m} \{\mathbf{x}_i (p(\mathbf{x}_i; \mathbf{w}_t) - y_i)\} + \frac{\mathbf{w}_t - \mu_t}{\sigma^2}, \tag{4.5}$$

and the Hessian is given by:

$$\frac{\partial^2 L}{\partial \mathbf{w}_t \partial \mathbf{w}_t^T} = \sum_{i=1}^{m} \{\mathbf{x}_i \mathbf{x}_i^T p(\mathbf{x}_i; \mathbf{w}_t)(1 - p(\mathbf{x}_i; \mathbf{w}_t))\} + diag(\mathbf{1}/\sigma^2), \tag{4.6}$$

where $p(\mathbf{x}; \mathbf{w}_t) = \frac{1}{1+exp(-\mathbf{w}_t^T \mathbf{x})}$ and $diag(\mathbf{1}/\sigma^2)$ is a diagonal matrix with all diagonal elements equal to $1/\sigma^2$.

Once the gradient and Hessian are available we can use the trust region Newton method from Lin and Moré (1999) to efficiently solve the Optimization Problem 8.

4.1.2 Therapy Similarity Kernels

We quantify the pairwise similarity between the different drug combinations with two kernels: the *drugs kernel* and the *groups additivity kernel*. The method also allows for alternative definitions of pairwise therapy similarity that can include different types of additional information, *e.g.* expert knowledge or information obtained from phenotypic drug resistance tests.

The drugs kernel similarity is identical to the *drug indicator kernel* described in Section 3.4 of Chapter 3. It is based on the number of common drugs that two combination therapies share – the higher this number the higher the similarity of the considered therapies. Its values are in the $[0, 1]$-interval.

The groups additivity kernel assumes that the similarity between different drug groups is additive. This is a reasonable assumption since drugs belonging to different groups have different targets and/or modes of inhibiting virus replication and thus can be assumed to act independently (Beerenwinkel et al., 2003b). Let G denote the set of different drug groups. In our data set we have three drug groups: NRTIs (Nucleoside Reverse Transcriptase

Inhibitors), NNRTIs (Non-Nucleoside Reverse Transcriptase Inhibitors) and PIs (Protease Inhibitors). Let $\mathbf{u}_{zg}$ and $\mathbf{u}_{z'g}$ be binary vectors indicating the set of drugs occurring in drug group $g \in G$ of the therapies z and z', respectively. The similarity between the group-g drugs of the two therapies z and z' is then calculated by:

$$sim_g(z, z') = \frac{\mathbf{u}_{zg}^T \mathbf{u}_{z'g}}{\max(\|\mathbf{u}_{zg}\|^2, \|\mathbf{u}_{z'g}\|^2)}, \tag{4.7}$$

where $\mathbf{x}^\mathsf{T}\mathbf{y}$ denotes the scalar product of the vectors $\mathbf{x}$ and $\mathbf{y}$, and $\|\cdot\|$ is the L_2-norm. Intuitively, the larger the number of common drugs making up a specific drug group of the two therapies of interest, the higher their group similarity.
We derive the similarity $k_a(z, z')$ between the therapies z and z' by averaging the similarities of their corresponding drug groups:

$$k_a(z, z') = \sum_{g \in G} (sim_g(z, z'))/|G|. \tag{4.8}$$

Since the group similarities $sim_g(z, z')$ lie in the interval $[0, 1]$, $k_a(z, z')$ also has values within $[0, 1]$.
Note that while fitting a single multi-class logistic regression model with several hundred classes required for the distribution matching approach described in Chapter 3 can take up to five days for some values of its respective tuning parameter, the computation of the therapy similarity kernels for all therapies is performed in several seconds.

4.1.3 Phenotypic Prior Knowledge on Therapy Outcome

Phenotypic resistance tests are laboratory experiments that produce continuous values, referred to as resistance factors, that measure the effectiveness of individual drugs against a given viral strain. Genotype-phenotype pairs (GPP) are sequences with the associated resistance factor measured in a phenotypic test using the virus defined by the sequence. The models trained on GPP data aim at predicting the resistance factors for each single drug for unknown genotypes. We will refer to these models as *phenotypic models*. One such model is described in Beerenwinkel et al. (2003a). Furthermore, this paper reveals the bimodal nature of the distribution of the resistance factors common to all drugs. Such a distribution can be approximated with a two-component Gaussian mixture model. We derive drug-specific resistance cut-offs from the intersection of the two mixture components. The cut-offs can then be used to infer the effectiveness of each drug against a given genotype: a drug is *effective* when its resistance value is smaller than its resistance cut-off, otherwise it is *ineffective*.
Unlike other approaches that add predictions facilitated by phenotypic models as additional features in their input feature space (Altmann et al., 2007, 2009b), we incorporate the models themself via a logarithm of a Gaussian prior on the model parameters for each therapy combination. We choose a Gaussian prior for two reasons. First, it is easy to integrate into regularized logistic regression as can be seen from Optimization Problem 8. Second, the *trust region Newton method* (Lin et al., 2008) which affords an efficient solution of the problem requires this prior. For a given therapy t we do this as follows.

- Consider the subset of the GPP data comprising the virus genotypes that have an associated resistance factor for all individual drugs that appear in the clinical data;

- Label each virus sequence based on the effectiveness of the best performing drug from the drugs comprising therapy t: *success* if effective, *failure* if ineffective;
- Fit a logistic regression model to the labeled data with model parameters (weights) p_t;
- Use the model parameters p_t from the fitted logistic regression as means $\mu_t = p_t$ for the Gaussian prior ($N(\mu_t, \sigma^2\mathbf{I})$) on model parameters $\mathbf{w}_t$ in Optimization Problem 8.

We repeat the procedure described above for every individual combination therapy. We will show that with this procedure one can better utilize the phenotypic knowledge compared to using the prediction of the phenotypic model as additional input feature in a single logistic regression model estimated for all therapies. Instead of using the effectiveness of the best performing drug to label a drug combination, one can also use other quantities such as for example the average of the effectiveness of the drugs comprising the therapy of interest. For further details on this see Section 4.2.
GPP data provide knowledge on the efficiency of individual drugs against HIV that is especially valuable for assessing therapies for which not many clinical samples are available. For example, while there can be a considerable amount of available GPP data for newly introduced drugs, clinical data for therapies that include these newly introduced drugs may be very sparse. This is the case simply because after the approval of a new antiretroviral agent it is spared as an option for highly treatment-experienced patients.

4.2 Results and Discussion

We described a model that targets the problem of predicting the outcome of HIV combination therapies from the genotype of the most abundant virus strain in the patient's blood serum. In the next sections we describe the data sets, the details of the computational study that assesses the quality of our model, as well as its results.

4.2.1 Data Sets

Our clinical data stem from an updated version the EuResist database (Rosen-Zvi et al., 2008) described in Chapter 3. It incorporates information of 88469 antiretroviral therapies administered to 18255 HIV-1 (subtype B) patients from several European countries in the period from 1988 through 2008. This information includes the combination of drugs given to the patients, the sequence of the predominant viral variant, and virus load measurements at different time points during a therapy.
The data set we use to train our models is derived from the EuResist database as described in Chapter 3. Each sample of the data corresponds to a therapy given to a patient and contains information describing the viral sequence obtained shortly before the respective therapy was administered. The virus genotype is represented by a binary vector indicating the occurrence or absence of a set of predefined resistance-relevant mutations based on the mutation list reported in Johnson et al. (2008). Each therapy is denoted by a binary vector indicating the presence or absence of the individual drugs comprising it. Note that we only consider the combination therapies composed of drugs for which GPP data are available.

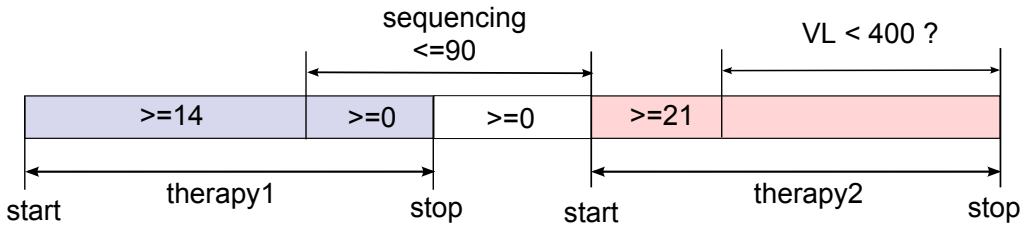


Figure 4.1: Assigning a label and a viral sequence to *therapy2*, where *therapy1* and *therapy2* are two consecutive therapies administered to a patient.

While in Chapter 3 we use two different definitions to assign a label to each therapy sample, namely the *virus load labeling* and the *multi-conditional labeling*, in this chapter and in the rest of the thesis we will focus on one definition. It is a modified version of the virus load labeling and is defined as follows. We label each therapy sample as success or failure based on the virus load measured in the course of the therapy. If the virus load drops bellow 400 cp/ml in the period from 21 days after the start of the therapy to the end of the therapy we label it with *success* (1); otherwise with a *failure* (-1). Figure 4.1 depicts the labeling procedure. In this way we create a labeled data set that includes 6336 therapy samples with 638 distinct therapy combinations. Note that having a single labeling definition enables efficient process of conducting computational experiments and comprehensive presentation of results.

As we mentioned in the previous chapter the different drug combinations are not evenly represented in the clinical data sets: while some therapies are represented with many samples, others have only few. The histogram in Figure 4.2, shows that this observation is valid for the updated version of the data set as well where for most therapies we have fewer than 50 samples available. Additionally, almost 500 antiretroviral therapies are represented by fewer than five examples.

We take the GPP data from the Arevir database (Roomp et al., 2006). As described in Subsection 4.1.3, the GPP data comprise drug resistance factors that characterize the effectiveness of individual drugs on specific viral variants. We consider only the virus sequences that have an associated resistance factor value in the database for all individual drugs that are relevant for our clinical data set. This is necessary because the construction of the prior knowledge on the therapy outcome for a specific therapy requires resistance factors for all drugs comprising the respective therapy. We label the GPP viral sequences as *success* or *failure* with respect to a given therapy of interest based on the resistance factor of the most effective drug. After the filtering we end up with 200 samples associated to 17 drugs. For representing the genotypic information describing the virus we use the same binary encoding as for the clinical data.

4.2.2 Validation Setting and Reference Methods

Validation setting. The practical experience with the drugs acquired over time and the introduction of new antiretroviral drugs affect the treatment trends for HIV patients. Our data collected over a period of two decades cannot be representative for a given time point. In order to account for the changing treatment trends over time we use the *time-oriented*

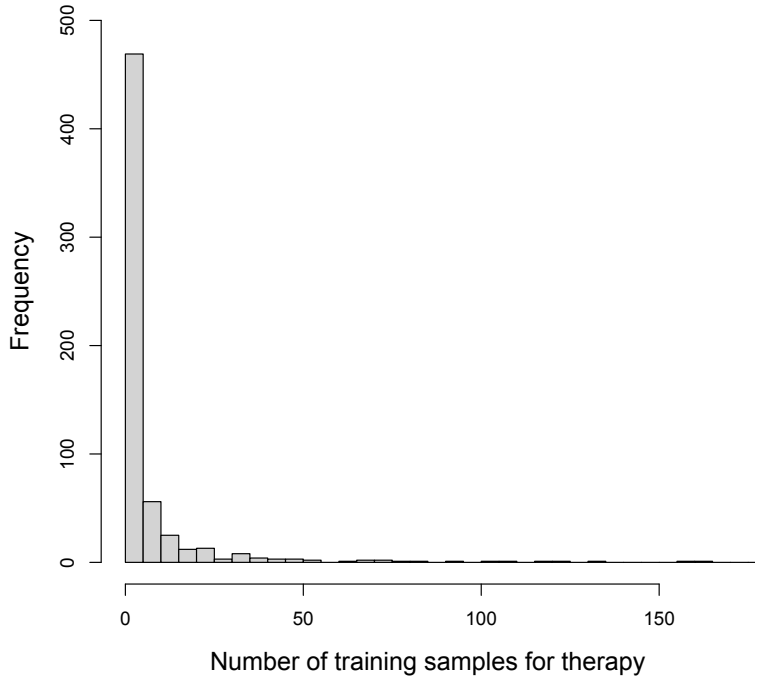


Figure 4.2: Histogram that groups the HIV combination therapies based on the number of samples present in the clinical data set.

validation scenario from Chapter 3 where our models are trained on the data from the more distant past, while their performance is measured on more recent data. Such a setting is realistic since it captures how a given model would perform on the recent trends of selecting combinations of drugs from established drug classes. We apply it as follows. First, we order all available training samples by their corresponding therapy starting dates. We then make a time-oriented split by selecting the most recent 20% of the samples (from June 2006 to January 2008) as the test set and the rest as the training set. For the model selection we split the training set further in a similar manner. We take the most recent 25% of the training set for selecting the best model parameters and refer to this set as tuning set. Figure 4.3 depicts the different treatment trends in the training, tuning and test sets, defined as explained in the text above. One can observe that, unlike the treatment trends in the training set, the treatment trends in the tuning set closely resemble those in the test set. This justifies the choice of the tuning set. Figure 4.3 also shows the changing treatment trends over time in the clinical data.

Reference methods. In the computational experiments we compare the performance of our therapy similarity methods to the performances achieved by three other reference methods.

The first reference method consists of training a separate logistic regression model for each combination therapy using only the samples from the target therapy. If we had enough data for each therapy combination this would be the best choice as the separate model captures the characteristics specific to the corresponding therapy and therefore can also

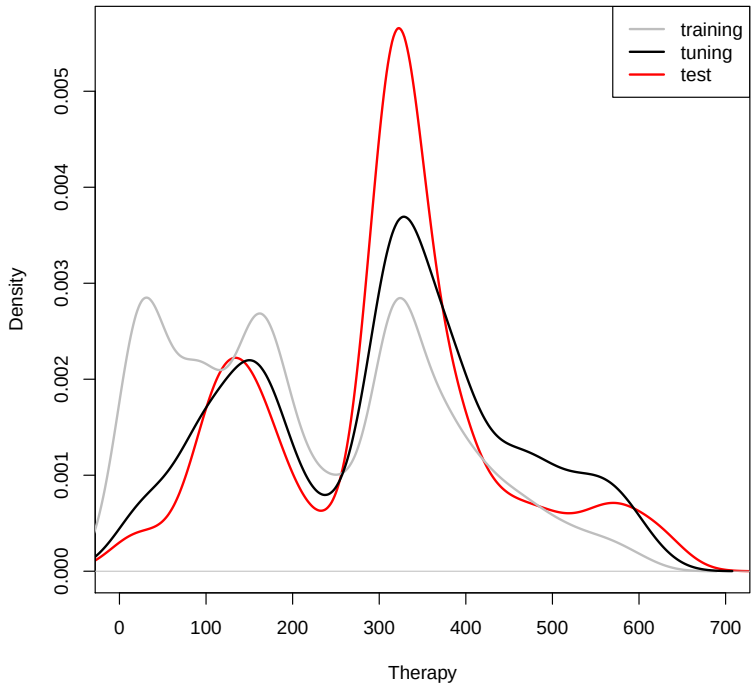


Figure 4.3: Distribution of the different combination therapies in the training, tuning and test set chosen in the time-oriented scenario. The numbers on the x-axis represent the different therapy combinations ordered by their first appearance in our clinical data: from older to newer. The y-axis depicts the density.

make the best predictions for it. In our case we fit the separate models by using the data available for each individual therapy in the clinical database. The therapies with no samples available are either randomly classified as *success* or *failure* both with equal probability of 50%, or the phenotypic model (Subsection 4.1.3) is used to assign their labels. For therapies represented only with successful or only with failing examples we assign success probabilities of 1 or 0, respectively. We will refer to this approach as *partitioned evaluation scenario*.

The second reference method, referred to as *transfer evaluation scenario*, implements the distribution matching approach by Bickel et al. (2008) described in the previous chapter. In order to provide a fair comparison we use a linear instead of a nonlinear logistic regression for the separate models estimated for each therapy. The sample-specific weights for each therapy are obtained by using a nonlinear multi-class logistic regression as described in Bickel et al. (2008).

The last reference method, referred to as *one-for-all evaluation scenario*, fits a single logistic regression model to the viral genotypes of all therapies. The information about the drugs comprising the corresponding therapies is added to the input feature space. This is the most common approach in the field (Larder et al., 2007; Altmann et al., 2009b).

Table 4.1: Classification accuracies (ACCs) and AUCs with their corresponding standard errors (SE) for our therapy similarity model (drugs kernel) and the three reference models (one-for-all, partitioned, transfer) that predict the outcomes of drug combination therapies.

method	ACC ± SE		AUC ± SE	
	no prior	with prior	no prior	with prior
drugs kernel	0.850 ± 0.010	0.848 ± 0.010	0.703 ± 0.022	0.695 ± 0.023
one-for-all	0.850 ± 0.010	0.856 ± 0.010	0.700 ± 0.023	0.703 ± 0.022
partitioned	0.830 ± 0.011	0.822 ± 0.011	0.624 ± 0.025	0.608 ± 0.026
transfer	0.848 ± 0.010	–	0.625 ± 0.024	–

Performance measures. The goal of each model is to predict the outcomes of the combination therapies for the most recent samples in the data set. We are primarily interested in the accuracy as a measure of the quality of the considered models. However, sometimes one is not only interested in the absolute results, but also in the quality of the ranking of the therapies based on their success probability. This is especially important when choosing a future therapy for a patient. Therefore we also carry out model selection based on AUC (Area Under the ROC Curve) performance and report AUC results. Resampling techniques (*e.g.* bootstrap) to estimate the standard errors of these measurements are not readily applicable in the time-oriented validation scenario in which the data samples are ordered by the starting times of their corresponding therapies. Therefore, we resort to calculating standard errors (SE) of the accuracies as detailed in Section 3.4 of Chapter 3. Standard errors of the AUCs are computed as described in Hanley and McNeil (1983).

4.2.3 Experimental Results and Discussion

In this subsection we present and discuss the results of the validation experiments. We first show the results pertaining to the therapy similarity models that use the drugs kernel followed by those pertaining to the therapy similarity models that use the group additivity kernel. Both therapy similarity approaches and each of the different reference models, as explained in the previous subsection, are trained on the EuResist clinical data set using the time-oriented validation scenario. As we mentioned before, efficient model fitting is important for the approaches that train a separate model for each combination therapy. By using the trust region Newton method for training logistic regression (Lin et al., 2008) we fit a single model in a fraction of a second. For example, fitting 154 separate models for the different therapies in the test set takes about 13 seconds on a normal desktop computer. Finally, we discuss different aspects of model interpretability for the therapy similarity approach.

Therapy similarity model with drug additivity kernel. Table 4.1 summarizes the classification accuracies (ACCs) and the AUCs for the considered methods: *drugs kernel* denotes our therapy similarity models with the drugs kernel therapy similarities as sample-specific weights; *partitioned* denotes the reference models fitted for each distinct therapy using only

Table 4.2: AUCs for the therapy similarity model (drugs kernel) and the two reference models (one-for-all, transfer) with their corresponding standard errors (SE) for two groups of test therapies: with $0-20$ and >20 available training samples.

method	drugs kernel		one-for-all		transfer
	no prior	with prior	no prior	with prior	
$0-20$ (SE)	0.659(0.041)	0.690(0.039)	0.641(0.041)	0.642(0.041)	0.608(0.043)
>20 (SE)	0.694(0.028)	0.681(0.028)	0.697(0.029)	0.700(0.029)	0.637(0.029)

the samples belonging to the target therapy; the label *with prior (no prior)* can refer to any of the previously described models with (without) additional phenotypic knowledge encoded as a Gaussian prior; *one-for-all* refers to the reference method that fits a single logistic regression model to the samples from all combination therapies where the therapy information is encoded as a part of the input feature space; the label *with prior* for the *one-for-all* approach refers to encoding the prediction of the phenotypic model described in Subsection 4.1.3 as an additional feature; *transfer* refers to the linear version of the transfer model by Bickel et al. (2008). In what follows we will first discuss the performance of the different models with respect to the accuracy and then continue with a similar discussion for the AUC.

As shown in Table 4.1, our approach (drugs kernel) of utilizing information of similar therapies, the transfer model and the model from the one-for-all scenario perform significantly better than training separate models by solely using the samples from the target therapies (partitioned scenario). We assess the significance of the accuracy with a paired t-test where we observe *p-values* ≤ 0.01 for all pairwise comparisons between the models from the partitioned scenario and the other models. The performance of the partitioned models is worse because for many therapies there are only few samples in the data set (see Figure 4.2). The therapy similarity kernels in our approach and the sample-specific weights in the transfer scenario compensate for this lack of data by utilizing the samples from similar therapies for making the predictions.

In order to further investigate the performance of the models we take the uneven representation of the different therapies into account. We do this by grouping the therapies in the test set based on the number of samples they have in the training set and then computing the accuracies of all the groups. The results are depicted in Figure 4.4. All models, including the ones derived in the partitioned scenario, deliver very good predictions for therapies for which there is a reasonable number of samples (≥ 15) available in the training data set. In such cases the models have enough samples to capture the characteristics of each different combination therapy.

As can be anticipated the models derived in the partitioned scenario achieve much worse performance compared to the other models for the therapies that have fewer samples in the training set ($0-14$). Our model (drugs kernel) and the transfer model that utilize therapy similarity in the learning process significantly outperform the one-for-all model for the therapies that have only very few ($0-3$) samples in the training set. Although this group comprises only about 8% of the test set, it contains 61 of the 154 different drug combinations in the test set which is around 40% of all distinct drug combinations

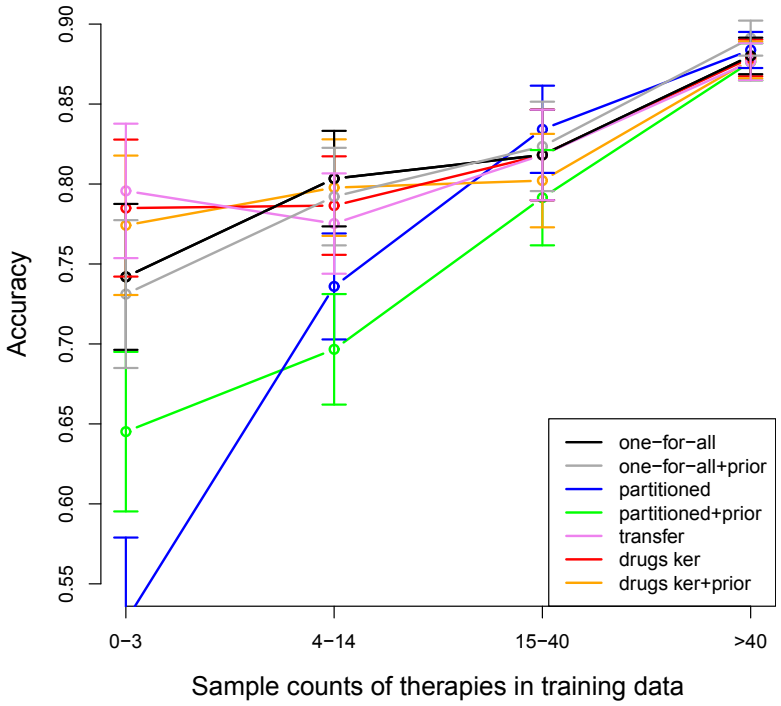


Figure 4.4: Classification accuracy of the different models over groups of test samples grouped by the number of training examples for their corresponding therapy combinations. Error bars indicate the standard errors of the accuracies.

occurring among the test therapies. We verified the significance of the improvements with paired t-test: $p\text{-}value = 0.04$ for the drugs kernel and $p\text{-}value = 0.09$ for the transfer model. For the group of therapies with 4 to 14 samples the therapy similarity model and the transfer model perform slightly worse than the one-for-all model. However, according to paired t-test this difference is only significant ($p\text{-}value = 0.03$) for the transfer model. The $p\text{-}value$ for the comparison with the drugs kernel model is 0.31.

The results of the partitioned models for therapies with only very few $(0-3)$ training samples significantly improve by incorporating the phenotypic prior knowledge: the paired t-test shows significant improvement of the accuracy ($p\text{-}value = 0.006$). However, including this prior knowledge makes the results worse for the therapies with $14-40$ training samples. It also does not improve the accuracy results for the therapy similarity or the one-for-all models. The reason for this may be that with respect to accuracy the samples from the similar therapies in the clinical data probably carry at least as much relevant information as the prior itself. It is encouraging to see that the added phenotypic information does not deteriorate the similarity models either.

Inspecting the overall AUC performance in Table 4.1 we can observe that all models except for the partitioned models and the transfer model have comparable performance. The reason for the poor performance of the partitioned models is that they often assign probabilities of 1 or 0 for therapies with very few training samples because they are often either all successful or all failing. That is why we do not look at the AUC performance of the partitioned models into more detail. As to the transfer model, the decision functions of

the individual therapy models are not guaranteed to be identically calibrated and thus the AUC performance calculated over all test samples is low.

We take the uneven therapy representation into account by splitting the data into two groups: one with $0-20$ training samples and another with more than 20 samples. The AUC results for the different methods are shown in Table 4.2. In this case our therapy similarity model with phenotypic prior knowledge significantly improves the AUC results compared to those of the transfer model ($p\text{-}value = 0.002$) for the therapies with $0-20$ available training samples. This group of therapies comprises about 25% of the test data. Integrating the prediction of the phenotypic model as additional input feature in the one-for-all model does not boost its AUC performance for the therapies with fewer training samples. Our therapy similarity model that incorporates the model parameters of the phenotypic model via a Gaussian prior outperforms the one-for-all model with prior knowledge for the therapies with $0-20$ available training samples ($p\text{-}value = 0.04$). This demonstrates the ability of our approach to better integrate the additional information provided with the phenotypic model. The differences between the AUCs of the one-for-all model (with and without prior) and the therapy similarity model (with and without prior) for the therapies with > 20 available training samples were not significant (all $p\text{-}values > 0.149$). We compute the significance of the difference of the AUCs and its standard error as described in DeLong et al. (1988).

Note that there are different alternatives to using the effectiveness of the best performing drug to label a combination therapy for the phenotypic prior models. For example, another approach is to compute the labeling based on the average effectiveness of all drugs comprising the combination therapy. As can be seen in Figure 4.5 the accuracy results obtained using the average are similar to those obtained when using the effectiveness of the best performing drug. However, the AUC performance for the test therapies with $0-20$ available training samples is significantly lower with a p-value$= 0.04$ (see Table 4.3).

Table 4.3: AUCs pertaining to the therapy similarity models (drugs kernel) with their corresponding standard errors (SE) for two groups of test therapies: with $0-20$, and > 20 available training samples. The labeling based on the best performing drug is denoted as *max prior* and the prior knowledge with labeling based on the average of the effectiveness of all drugs comprising the target therapy is referred to as *avg prior*.

method	drugs kernel + max prior	drugs kernel + avg prior
$0-20$ (SE)	0.690 (0.039)	0.667 (0.041)
> 20 (SE)	0.681 (0.028)	0.680 (0.028)

We should also mention that considering the effectiveness of the most ineffective drug from a target therapy might not be a well justified labeling approach. This is the case because the therapies administered to therapy-experienced patients (patients that have had at least one antiretroviral therapy) mostly include drug(s) already contained in the previous therapy. These previously administered drugs render low in-vitro effectiveness of

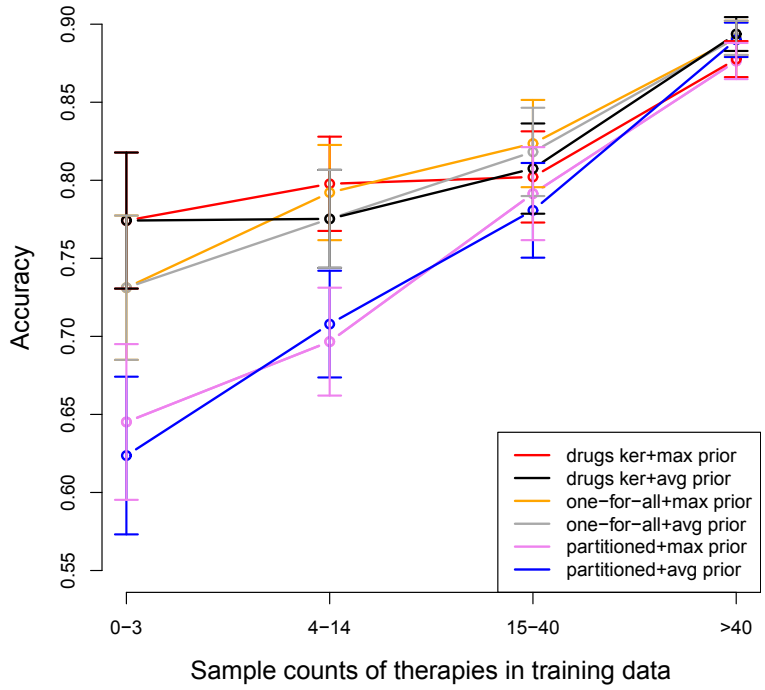


Figure 4.5: Classification accuracy of the different models using different priors over groups of test samples grouped by the number of training examples for their corresponding therapy combinations. Error bars indicate the standard errors of the accuracies.

the whole drug combination.
To summarize, with respect to measured accuracies our therapy-similarity approach that trains separate models for each therapy has its prime advantage for therapies with few (less than four) training samples. The transfer model also achieves good prediction results for this group of therapies. However, it has lower accuracy for the group of therapies with $4 - 14$ available training samples and has a significantly worse AUC performance due to the potentially different calibrations pertaining to the decision functions of the individual therapy models. Another disadvantage of this method is the compute-intensive calculation of the sample-specific weights: fitting a single multi-class logistic regression model for a large number of classes (in our case several hundreds) and training samples is a very time-consuming task – in order to update such model with new data one has to fit many multi-class logistic regression models since they also have a tuning parameter.
The phenotypic prior knowledge added to the therapy similarity model significantly improves the AUC performance for therapies with $0 - 20$ available training samples. Here the added phenotypic information is essential in bringing the performance of the model to a level, which is also achieved for the abundant therapies.

Therapy similarity model with drug additivity kernel. The results obtained when using the group additivity kernel in our therapy similarity models are comparable to those obtained when using the drugs additivity kernel. The only difference is the slightly higher *p-value* (0.08) for the accuracy performance of the group of therapies with few $(0 - 3)$ training

samples compared to the one-for-all model.

We should also point out that the calculated correlation between the values of the two similarity kernels is higher than 0.6 for 90% of the therapies in our data set.

Table 4.4 and Figure 4.6 summarize the detailed results for the group additivity kernel. They contain the same information as Table 4.2, and Figure 4.4 of the drugs kernel, respectively. A visual comparison of the detailed accuracy results of both kernels is depicted in Figure 4.7.

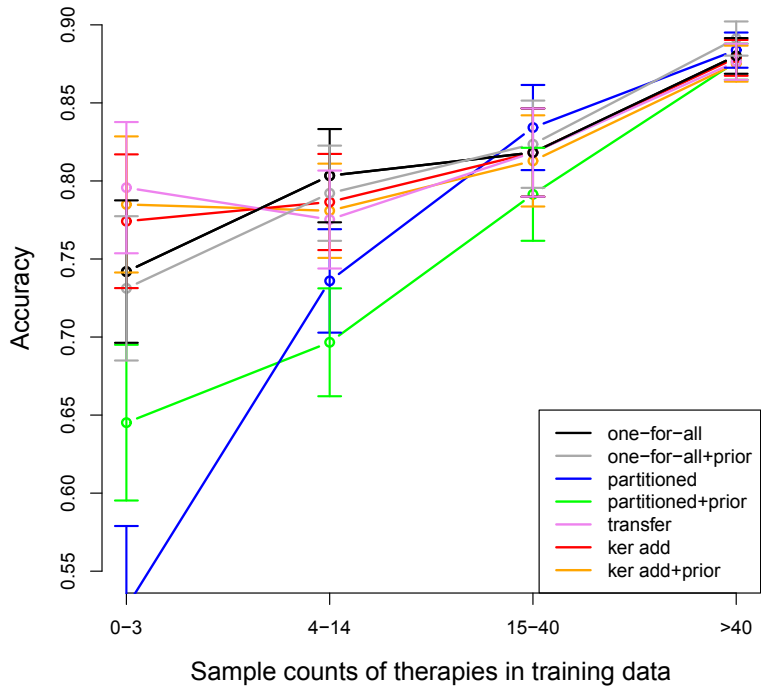


Figure 4.6: Classification accuracy of the different models over groups of test samples grouped by the number of training examples for their corresponding therapy combinations. Error bars indicate the standard errors of the accuracies.

Table 4.4: AUCs for the therapy similarity model (using the groups additivity kernel) and the two reference models (one-for-all, transfer) with their corresponding standard errors (SE) for two groups of test therapies: with $0-20$, and > 20 available training samples.

method	additivity kernel		one-for-all		transfer
	no prior	with prior	no prior	with prior	
$0-20$ (SE)	0.654(0.041)	0.689(0.038)	0.641(0.041)	0.642(0.041)	0.608(0.043)
> 20 (SE)	0.691(0.028)	0.685(0.028)	0.697(0.029)	0.700(0.029)	0.637(0.029)

Our method also allows for alternative definitions of therapy similarity. For example, one can derive a similarity measure that includes phenotypic knowledge.

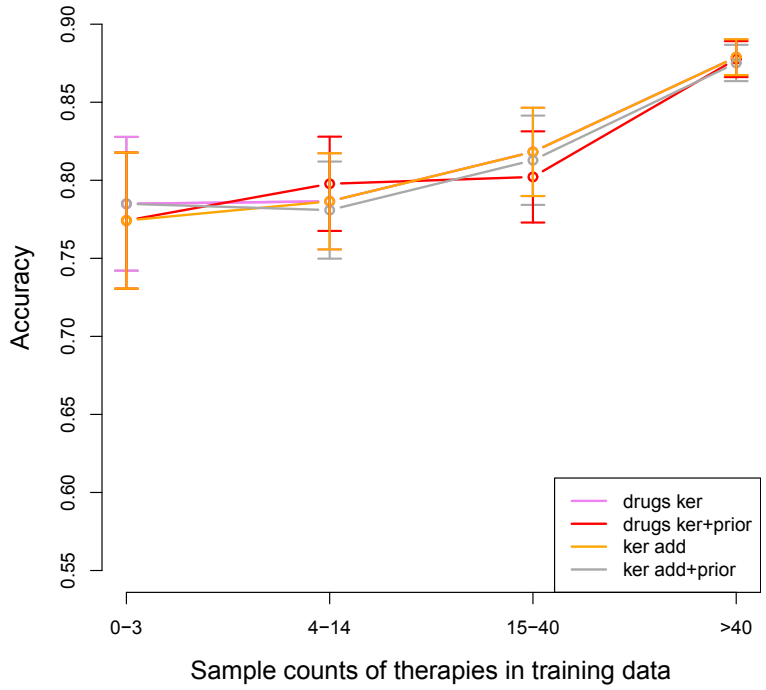


Figure 4.7: Classification accuracy over groups of test samples grouped by the number of training examples for their respective therapy combinations for the two different therapy similarity measures: drugs kernel (drugs ker) and additivity kernel (ker add).

Model interpretability. In a medical setting the model interpretability is important since it enables access to the argumentative basis of the predictions. The therapy similarity approach makes two contributions to model interpretability.

On the one hand, one can analyze the profile of similar therapies associated to the model of a given target therapy that harbors relevant information. An interesting example is illustrated in Figure 4.8. This figure shows a heatmap that depicts the magnitude of similarity of each of six randomly chosen test drug combinations from the group of therapies with no available training samples when compared with all therapies from the training set with similarity values greater than 0.5 according to the drugs kernel. In this way we can easily see how much each of the drug combinations contributed to predicting the outcome for the test treatments with no training samples available. As an example, let us consider the model for the combination therapy *ZDV 3TC ABC TDF RTV ATV*. From the heatmap we can easily see that the model assigns the highest weights to the examples from the therapies: *ZDV 3TC TDF RTV LPV*, *3TC ABC TDF RTV ATV*, *ZDV 3TC ABC TDF SQV RTV ATV*, and *ZDV 3TC ABC RTV ATV*.

On the other hand, we can assess how different mutations in the viral genome contribute to predicting the outcome of a given target therapy. Since we have a separate model for the target therapy we can simply do this by quantifying the importance of the model features. One way to do this is to calculate z-scores for each of the model coefficients, which corresponds to a test of the null hypothesis that the coefficient of interest is zero, while all the others are not. The coefficients with the highest z-scores are the most significant ones.

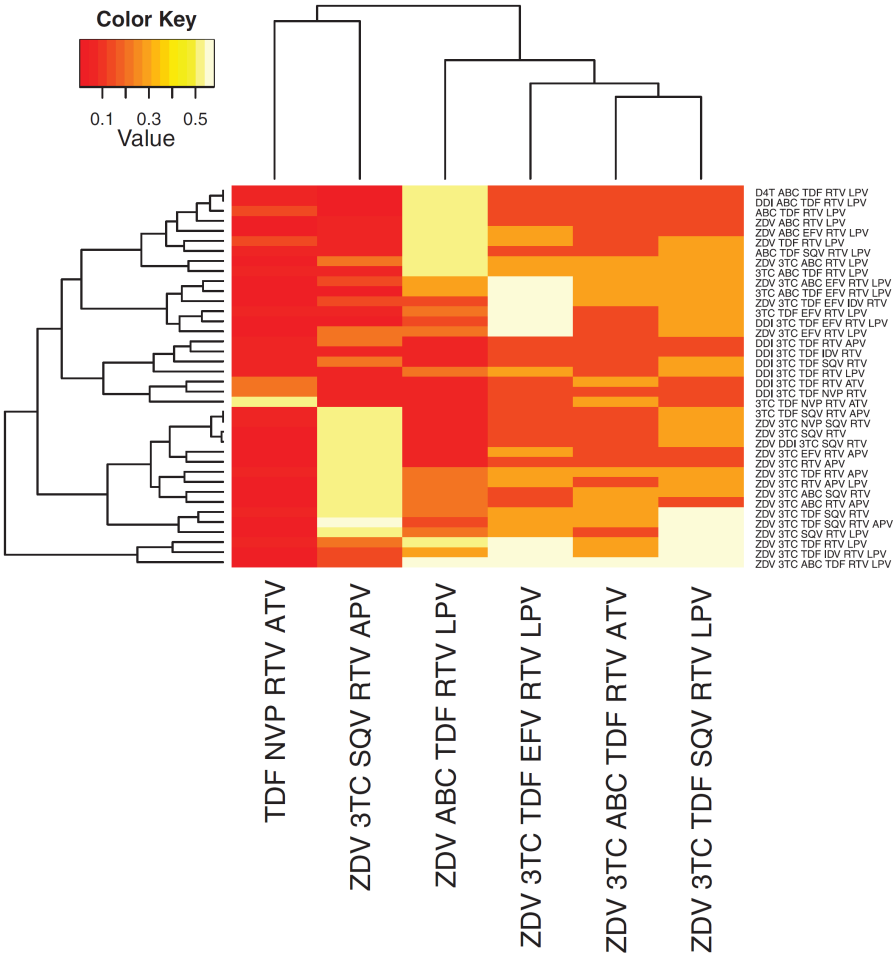


Figure 4.8: Heatmap of the similarity profile of six test therapies with no training samples available. The profile considers only the training therapies with similarity values greater than 0.5. The test therapies are depicted on the horizontal axis and the training therapies are depicted on the vertical axis. The similarity values are derived according to the drugs kernel.

Table 4.5 contains the z-scores for the coefficients of the therapy similarity model (using the drugs kernel) for the combination therapy *ZDV 3TC ABC TDF RTV ATV*. According to them, for this therapy the three most important positions in the protease sequence are 54, 90 and 10; and in the reverse transcriptase sequence the most relevant positions are 215, 210 and 41.

4.3 Conclusion

This paper presents an approach that tackles the problem of predicting virological response to combination therapies by training a separate logistic regression model for each different combination therapy. Each model is fitted using not only the data from the target therapy but also the information from therapies similar to it. For this purpose we introduce and evaluate two different measures of pairwise therapy similarity which are used as weights in

the logistic regression models. The model is also able to incorporate phenotypic knowledge on the therapy outcomes through a Gaussian prior. With such an approach we balance the uneven therapy representation in the clinical data sets and produce higher quality models for the therapies with very limited number of training samples. Our approach is not only advantageous for therapies with few training samples, but also for all other therapies. Having a separate model for each drug combination increases the interpretability of the fitted models in that users have access to the argumentative basis of the prediction. On the one hand, the scores of the mutations contributing to therapy effectiveness which result from the linear predictions are derived in a therapy-specific manner and can therefore be considered more informative than for a general model. On the other hand, the therapy similarity kernel affords information on which similar therapies were most informative for the prediction. It has to be stressed that interpretability of the prediction is a prime requirement for use of a prediction method in a medical setting. Finally, the use of an efficient optimization method that takes advantage of the sparseness of our input data ensures very fast model fitting and selection, although we train a separate model for each combination therapy.

In terms of accuracy, our therapy similarity model performs significantly better (at the 1% significance level) than training separate models for each therapy by using solely its samples. Furthermore, the therapy similarity model significantly outperforms (at the 4% significance level for the drugs kernel and the 8% significance level for the groups additivity kernel) the one-for-all scenario for the group of therapies with fewer than four training samples. Although this group comprises only about 8% of the test data, it contains around 40% of the different drug combinations in the test set. For therapies with a sizeable number of samples (above four) both the similarity and the one-for-all models have comparable performance. Our model achieves similar accuracy and significantly better AUC performance than the transfer model, which uses a compute-intensive distribution matching approach to quantify the therapy similarities. This demonstrates the quality of our therapy similarity measures.

The phenotypic prior knowledge included via a Gaussian prior does not improve the prediction accuracies of the similarity models, but it significantly improves (at the 4% level) the AUC of the test therapies that have $0-20$ available training samples. This group comprises about 25% of the test data and contains around 77% of the different drug combinations in the test set.

Table 4.5: Table of z-scores with their corresponding *p-values* showing the importance of the different positions in the reverse transcriptase sequence (left column) and the protease sequence (right column) for the combination therapy *ZDV 3TC ABC TDF RTV ATV*.

reverse transcriptase(RT)			protease(PR)		
sequence position	z-score	p-value	sequence position	z-score	p-value
10	-6.982	1.458043e-12	41	-4.527	2.990632e-06
11	-0.702	2.411966e-01	62	-0.495	3.102540e-01
13	-2.436	7.423252e-03	65	0.855	1.963060e-01
16	-0.363	3.584558e-01	67	-3.338	4.220042e-04
20	-4.067	2.379967e-05	70	-0.137	4.455815e-01
24	-2.188	1.435101e-02	74	-1.369	8.556330e-02
30	4.481	3.715210e-06	75	-2.122	1.693959e-02
32	-2.823	2.376032e-03	77	-2.130	1.658176e-02
33	-5.703	5.876144e-09	90	-1.042	1.486348e-01
34	-2.617	4.441481e-03	98	0.117	4.533519e-01
35	1.469	7.094346e-02	00	-0.889	1.871202e-01
36	-3.438	2.929115e-04	01	-1.273	1.015777e-01
43	-2.178	1.470533e-02	03	-0.800	2.119379e-01
46	-6.823	4.454905e-12	06	0.571	2.840376e-01
47	-2.436	7.425982e-03	08	-0.088	4.648635e-01
48	-1.206	1.139030e-01	15	-0.645	2.594725e-01
50	-2.209	1.357249e-02	16	-2.934	1.673395e-03
53	-3.910	4.622426e-05	38	-0.298	3.827428e-01
54	-10.91	5.022132e-28	51	-3.579	1.722739e-04
58	-2.699	3.472476e-03	79	-2.577	4.977403e-03
60	-0.715	2.374222e-01	81	-1.181	1.188574e-01
62	0.015	4.939747e-01	84	-2.231	1.283414e-02
63	-2.975	1.462568e-03	88	-3.663	1.244663e-04
64	0.201	4.204933e-01	90	-1.360	8.692458e-02
69	-2.480	6.568336e-03	10	-5.379	3.740906e-08
71	-6.374	9.222044e-11	15	-9.336	5.023530e-21
73	-4.013	3.000900e-05	19	-2.636	4.197186e-03
74	-2.383	8.589538e-03	25	0.686	2.463370e-01
76	-2.500	6.206607e-03	30	2.831	2.319453e-03
77	-0.672	2.506833e-01	36	-0.489	3.123181e-01
82	-6.640	1.563237e-11			
83	0.575	2.826479e-01			
84	-5.231	8.423786e-08			
85	-1.849	3.223458e-02			
88	0.827	2.041891e-01			
89	-2.446	7.216124e-03			
90	-7.322	1.219247e-13			
93	-0.621	2.673961e-01			

5 Hierarchical Bayes Model for Predicting Outcomes of HIV Combination Therapies

The method we present here affords a simple, direct, effective and efficient approach to modeling the response to HIV combination therapies based on the hierarchical Bayes paradigm. The individual drugs comprising each therapy combination are considered as separate tasks in a multi-task model that estimates their additive effects on the therapy outcome from the available clinical data. In this way, the model makes use of the abundance of samples involving each individual drug. Doing so improves the predictive power on target therapies that are scarcely represented in the clinical database. Note that this chapter elaborates on the work initially presented in Bogojeska and Lengauer (2011).

5.1 Related Work

The approaches in Bickel et al. (2008) and Bogojeska et al. (2010) (explained in detail in Chapters 3 and 4, respectively) deal with the problem of uneven and sparse therapy representation in the HIV data by training a separate model for each combination therapy which uses the available samples from all therapies with properly derived sample weights. The weights reflect the similarities between the target therapy and the corresponding therapies of all training samples. While these therapy-specific models achieve very good accuracy (Bogojeska et al., 2010), their AUC (Area Under the ROC Curve) performance can be improved.

The hierarchical Bayes paradigm (Gelman et al., 2004) can easily be applied to multi-task modeling and, therefore, is widely used in the machine learning community (Evgeniou and Pontil, 2004; Yu et al., 2005; Dudik et al., 2005; Teh et al., 2006). Our work is inspired by the work of Evgeniou and Pontil (2004) who present a feature mapping method for multi-task learning with support vector machines based on a hierarchical Bayes approach. Bickel (2009) shows that a modified version of this method with a logistic loss function is equivalent to a hierarchical Bayes model. In this chapter we adapt this method to the problem at hand which yields a novel method that models the individual effects of the drugs on therapy outcome.

5.2 Methods

Here we derive the multi-task hierarchical Bayes learning method for the problem of predicting the outcomes of HIV drug combination therapies. Our goal is to model the effects of the drugs making up a target combination therapy on its outcome by using the viral genotype information and the available information on previously administered drugs as input features. Since the individual drugs comprising the combination therapies appear in

many samples we can consider each drug as a separate task in a multi-task setting. We use an additivity assumption to model the combined effects of the individual drugs comprising a target therapy on its response. Clearly this assumption is a gross simplification of the complex and little understood process of drug interaction. Still, the drug additivity approach is a widely used simple assumption in a situation where little information is available on actual interactions, and it exhibits good prediction performance. The sum of the drug-specific contributions provides a score quantifying the propensity of the therapy to be effective. For each drug model we have a comparatively data-rich scenario, thus avoiding the necessity to make predictions on the basis of only very few informative samples.

5.2.1 Hierarchical Bayes Model

In the hierarchical Bayes setting the posterior probability $p(\mathbf{w}, \varphi|D)$ is computed by using the likelihood $p(D|\mathbf{w})$ of the training data D under model parameters $\mathbf{w}$, the prior probability $p(\mathbf{w}|\varphi)$ of model parameters $\mathbf{w}$ under hyperparameters φ, and the prior $p(\varphi)$ of the hyperparameters φ:

$$p(\mathbf{w}, \varphi|D) \propto p(D|\mathbf{w})p(\mathbf{w}|\varphi)p(\varphi). \tag{5.1}$$

Then, the maximum a posteriori (MAP) estimate of the model parameters:

$$(\hat{\mathbf{w}}, \hat{\varphi}) = \arg\max_{\mathbf{w},\varphi} p(\mathbf{w}, \varphi|D) \tag{5.2}$$

is used for the final prediction $\hat{y} = \arg\max_y p(y|\mathbf{x}, \hat{\mathbf{w}})$ for a target sample $\mathbf{x}$.
A multi-task problem with several related tasks that share a common prior can easily be realized in the hierarchical Bayes framework. Let $\mathbf{w}_1, \ldots, \mathbf{w}_T$ denote the task parameters of each of the T different tasks appearing in the training data $D = \{(\mathbf{x}_1, y_1, t_1), \ldots, (\mathbf{x}_m, y_m, t_m)\}$, and $D_t = \{(\mathbf{x}_i, y_i, t_i)|t_i = t\}$ is the training data for task t. All task parameters have the same prior probability $p(\mathbf{w}_t|\varphi)$ and are conditionally independent given the prior. The posterior is then given by:

$$p(\mathbf{w}_1, \ldots, \mathbf{w}_T, \varphi|D_1, \ldots, D_T) = p(\varphi) \prod_t p(D_t|\mathbf{w}_t)p(\mathbf{w}_t|\varphi) \tag{5.3}$$

where the parameters are approximated with a MAP estimate. Intuitively, the prior models what all tasks have in common, while the task parameters capture task-specific information. The structure of such a hierarchical Bayes model is schematically depicted in Figure 5.1.

5.2.2 Outcome Prediction for HIV Combination Therapies

Let $\mathbf{x}$ denote the input features that comprise the viral genotype and the drug history for the specific therapy example. The input is represented with a binary vector, where the part corresponding to the viral genotype indicates the occurrence of a set of resistance-relevant mutations (Johnson et al., 2008), and the part corresponding to the drug history comprises the drugs known to be part of previous therapies. Let $\mathbf{z}$ denote the therapy combination encoded as a binary vector that indicates the individual drugs comprising the therapy. The label y indicates the success (1) or failure (-1) of each sample therapy. Let $D = \{(\mathbf{x}_1, y_1, \mathbf{z}_1), \ldots, (\mathbf{x}_m, y_m, \mathbf{z}_m)\}$ denote the training data set.

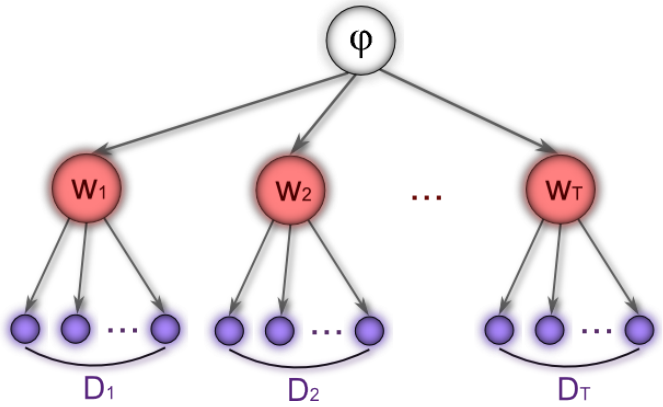


Figure 5.1: Image of the hierarchical structure of a multi-task Bayes model with T tasks.

The most common approach in the field trains a single statistical model (*e.g.* a linear logistic regression model) for all available therapy samples in the data set. Here the information on the individual drugs comprising the target therapy is encoded in a binary vector and supplied together with the other input features. In what follows we will present the details of the derivation of a hierarchical Bayes model that predicts the outcomes of HIV combination therapies.

The goal is to train a classifier $f_{\mathbf{z}} : \mathbf{x} \mapsto y$ that correctly predicts the outcome for an HIV combination therapy $\mathbf{z}$. We model the class likelihood $p(y|\mathbf{x}, \mathbf{z})$ with a logistic regression model that calculates predictions of the effectiveness of drug combinations by using the assumption that the drugs making up the therapy have a cumulative effect on the therapy outcome. This is reflected in the formula:

$$p(y|\mathbf{x}, \mathbf{z}, \mathbf{w}) = \frac{1}{(1 + \exp(-y \sum_{d \in \mathbf{z}} \mathbf{w}_d^{\mathsf{T}} \mathbf{x}))} \tag{5.4}$$

where $\mathbf{z}$ denotes the set of drugs comprising the combination therapy and $\mathbf{w}_d$ are the model parameters of the individual drugs, *i.e.* the drug-specific weights pertaining to the resistance-relevant mutations and the previously administered drugs. These drug parameters are trained via the multi-task hierarchical Bayes framework where each drug is considered a separate task. The model parameters $\mathbf{w}_d$ for each drug are drawn from a common Gaussian prior $\mathbf{w}_d \sim N(\mathbf{w}_0, \sigma_{\mathbf{w}}^2 \mathbf{I})$ with a mean drawn from a Gaussian hyperprior $\mathbf{w}_0 \sim N(\mathbf{0}, \sigma_{\mathbf{w_0}}^{\mathbf{2}} \mathbf{I})$. In this way all tasks (drugs) are related and their similarity is modeled with the common Gaussian prior. In fact, some drugs are more similar than others in that they belong to the same drug class or evoke a similar genomic fingerprint in terms of viral resistance mutations. More formally, all task parameters $\mathbf{w}_d$ deviate to some extent from a mean function $\mathbf{w}_0$ (in our case the mean of the Gaussian prior). The smaller the distance between two distinct drug parameters $\mathbf{w}_{d_1}$ and $\mathbf{w}_{d_2}$ the more similar the effect of drugs d_1 and d_2.

Let n denote the number of different drugs in our data set and $D_d = \{(\mathbf{x}_i, y_i, \mathbf{z}_i) \in D | d \in \mathbf{z}_i\}$ denote all training samples whose corresponding therapies contain the drug d. In the following we derive the log-posterior of all parameters given the data in accordance with Equation 5.3 and the assumptions made in the previous paragraph.

$$
\begin{aligned}
&\log p(\mathbf{w}_1, \ldots, \mathbf{w}_n, \mathbf{w}_0 | D_1, \ldots, D_n, \sigma^2_{\mathbf{w}_0}, \sigma^2_{\mathbf{w}}) \\
&\quad \propto \quad \log N(\mathbf{w}_0 | \mathbf{0}, \sigma^2_{\mathbf{w_0}} \mathbf{I}) + \sum_{\mathbf{d=1}}^{\mathbf{n}} \log \mathbf{N}(\mathbf{w_d} | \mathbf{w_0}, \sigma^2_{\mathbf{w}} \mathbf{I}) \\
&\qquad + \sum_{d=1}^{n} \sum_{(\mathbf{x},y,\mathbf{z}) \in D_d} \log p(y | \mathbf{x}, \mathbf{z}, \mathbf{w_z}) \qquad (5.5) \\
&\quad \propto - \frac{\|\mathbf{w}_0\|^2}{2\sigma^2_{\mathbf{w}_0}} - \sum_{d=1}^{n} \frac{\|\mathbf{w}_d - \mathbf{w}_0\|^2}{2\sigma^2_{\mathbf{w}}} \\
&\qquad - \sum_{d=1}^{n} \sum_{(\mathbf{x},y,\mathbf{z}) \in D_d} \log(1 + \exp(-y \sum_{d \in \mathbf{z}} \mathbf{w}_d^{\mathsf{T}} \mathbf{x})) \qquad (5.6) \\
&\quad = - \frac{\|\mathbf{w}_0\|^2}{2\sigma^2_{\mathbf{w}_0}} - \sum_{d=1}^{n} \frac{\|\mathbf{v}_d\|^2}{2\sigma^2_{\mathbf{w}}} \\
&\qquad - \sum_{d=1}^{n} \sum_{(\mathbf{x},y,\mathbf{z}) \in D_d} \log(1 + \exp(-y(|\mathbf{z}| \mathbf{w}_0 + \sum_{d \in \mathbf{z}} \mathbf{v}_d)^{\mathsf{T}} \mathbf{x})) \qquad (5.7) \\
&\quad = - \frac{\|\mathbf{v}_0\|^2}{2\sigma^2_{\mathbf{w}}} - \sum_{d=1}^{n} \frac{\|\mathbf{v}_d\|^2}{2\sigma^2_{\mathbf{w}}} \\
&\qquad - \sum_{d=1}^{n} \sum_{(\mathbf{x},y,\mathbf{z}) \in D_d} \log(1 + \exp(-y(\frac{|\mathbf{z}| \sigma_{\mathbf{w}_0}}{\sigma_{\mathbf{w}}} \mathbf{v}_0 + \sum_{d \in \mathbf{z}} \mathbf{v}_d)^{\mathsf{T}} \mathbf{x})) \qquad (5.8) \\
&\quad = - \frac{\|\mathbf{v}\|^2}{2\sigma^2_{\mathbf{w}}} - \sum_{(\mathbf{x},y,\mathbf{z}) \in D} \log(1 + \exp(-y \mathbf{v}^{\mathsf{T}} \Phi(\mathbf{x}, \mathbf{z}))) \qquad (5.9)
\end{aligned}
$$

Equation 5.5 uses Equation 5.3 to derive the logarithm of the posterior probability. In Equation 5.6 the Gaussian density functions are expanded up to constant terms. Since $\mathbf{w}_d \sim N(\mathbf{w}_0, \sigma^2_{\mathbf{w}} \mathbf{I})$ each individual drug parameter $\mathbf{w}_d$ can be replaced by $\mathbf{w}_d = \mathbf{w}_0 + \mathbf{v}_d$ yielding Equation 5.7 where $|\mathbf{z}|$ is the number of drugs comprising therapy $\mathbf{z}$. In Equation 5.8 $\mathbf{w}_0$ is replaced with $\frac{\sigma_{\mathbf{w}_0}}{\sigma_{\mathbf{w}}} \mathbf{v}_0$. Finally, in the last Equation 5.9 the vector $\mathbf{v}$ denotes the concatenation of all parameter vectors $\mathbf{v} = [\mathbf{v}_0, \ldots, \mathbf{v}_n]$ and $\Phi(\mathbf{x}, \mathbf{z})$ is a new feature mapping defined as follows. Let $\mathbf{c_z} = [\frac{|\mathbf{z}| \sigma_{\mathbf{w_0}}}{\sigma_{\mathbf{w}}}, \mathbf{z}]$ denote an extension of the therapy vector $\mathbf{z}$, where each vector component is a vector itself with dimension equal to the dimension of the input feature vector $\mathbf{x}$. The new feature mapping is then given by $\Phi(\mathbf{x}, \mathbf{z}) = \mathbf{c_z} \cdot [\mathbf{x}, \ldots, \mathbf{x}]$ (where $\cdot$ denotes componentwise vector multiplication). In other words, it maps the input features of the training samples to a new feature space that provides a separate set of dimensions for each drug comprising the target therapy: the feature vector $\mathbf{x}$ for a given training sample is copied to the sections corresponding to the drugs comprising the target therapy $\mathbf{z}$; all other sections except the first one are filled with zeros; the first section is shared by all drugs and models their similarity. For example, let us assume that a target drug combination $\mathbf{z}$ comprises only two drugs $\mathbf{z} = \{d_1, d_2 | d_1 < d_2, d_1, d_2 \in 1, \ldots, n\}$. Then for given input features $\mathbf{x}$ the feature mapping $\Phi(\mathbf{x}, \mathbf{z})$ is given by:

$$
\Phi(\mathbf{x}, \mathbf{z}) = \left[\frac{2\sigma_{\mathbf{w}_0}}{\sigma_{\mathbf{w}}} \mathbf{x}, \underbrace{\mathbf{0}, \ldots, \mathbf{0}}_{1,\ldots,d_1-1}, \underbrace{\mathbf{x}}_{d_1}, \underbrace{\mathbf{0}, \ldots, \mathbf{0}}_{d_1+1,\ldots,d_2-1}, \underbrace{\mathbf{x}}_{d_2}, \underbrace{\mathbf{0}, \ldots, \mathbf{0}}_{d_2+1,\ldots,n} \right]. \qquad (5.10)
$$

As can be observed from Equation 5.9, by using the feature mapping $\Phi(\mathbf{x}, \mathbf{z})$ we obtain the objective function of a logistic regression model with model parameters $\mathbf{v}$. The dimensionality of the new input feature space is the dimension of $\mathbf{x}$ multiplied by $(n+1)$.
To summarize, the MAP estimate of the parameters of a hierarchical Bayes model with Gaussian prior and hyperprior applied to the problem of predicting outcomes of HIV therapies with drug additivity assumption is given by:

$$\hat{\mathbf{v}} = \arg\max_{\mathbf{v}} \{ - \sum_{(\mathbf{x},y,\mathbf{z}) \in D} \log(1 + \exp(-y\mathbf{v}^{\mathsf{T}}\Phi(\mathbf{x}, \mathbf{z}))) - \frac{\|\mathbf{v}\|^2}{2\sigma_{\mathbf{w}}^2} \}. \tag{5.11}$$

We obtain the maximum $\hat{\mathbf{v}}$ with logistic regression. This also yields the prediction of the label (success probability) of a target therapy $\mathbf{z}$ administered to a sample $\mathbf{x}$:

$$\hat{y} = \arg\max_{y} \frac{1}{(1 + \exp(-y\hat{\mathbf{v}}^{\mathsf{T}}\Phi(\mathbf{x}, \mathbf{z})))}. \tag{5.12}$$

We will refer to this method as *drug additivity Bayes*. A slightly modified version of the drug additivity Bayes is the *drug additivity + hist Bayes* described as follows. For each of the drugs comprising a target combination therapy two tasks are created: one for the case when the drug is administered for the first time to the considered patient, and another one for the case when the drug was administered previously in the patient's drug history. Once the tasks are defined, a task additivity assumption is applied, and the model is derived in the same way as the *drug additivity Bayes*. The dimensionality of the input feature space of the new model is the dimension of $\mathbf{x}$ multiplied by $(2n+1)$. The two hierarchical Bayes scenarios for predicting effectiveness of HIV combination therapies are sketched in Figure 5.2.

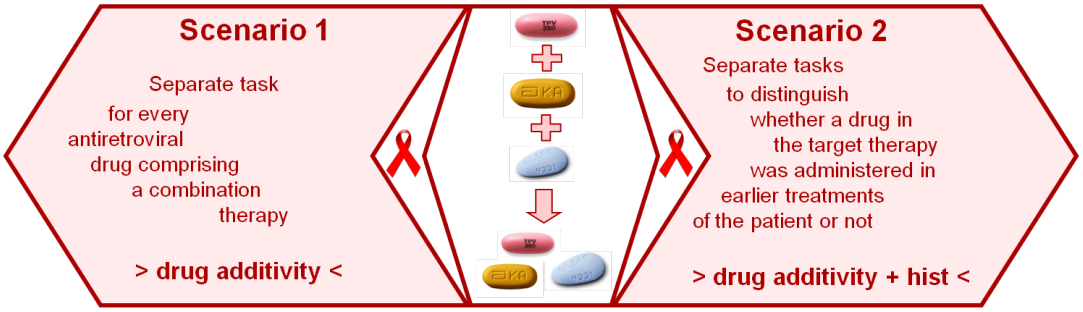


Figure 5.2: Multi-task hierarchical Bayes models for the problem of predicting effectiveness of HIV combination therapies.

Since we use Gaussian priors in our Bayes models, we can employ the trust-region Newton method for training logistic regression (Lin et al., 2008). This is an efficient implementation for sparse data sets with large number of features and samples (see Chapter 4 for more details). With this approach our models are trained in about one second, albeit the increased dimensionality of the input feature space and the large number of training samples. The Bayes methods have two tuning parameters: one replacing the fraction $\frac{|\mathbf{z}|\sigma_{\mathbf{w}_0}}{\sigma_{\mathbf{w}}}$ in the feature mapping $\Phi(\mathbf{x}, \mathbf{z})$ and one for the regularizer in Equation 5.11.

Note that by using kernel logistic regression (Zhu and Hastie (2002) and briefly described in Chapter 2) one can train non-linear hierarchical Bayes models as follows. The corresponding kernel function $k_\Phi((\mathbf{x},\mathbf{z}),(\mathbf{x}',\mathbf{z}'))$ for the feature mapping Φ derived from Equation 5.10 is given by:

$$k_\Phi((\mathbf{x},\mathbf{z}),(\mathbf{x}',\mathbf{z}')) = \Phi((\mathbf{x},\mathbf{z}))^\mathsf{T}\Phi((\mathbf{x}',\mathbf{z}')) = (\frac{\sigma^2_{\mathbf{w}_0}}{\sigma^2_{\mathbf{w}}}|\mathbf{z}||\mathbf{z}'| + \mathbf{z}^\mathsf{T}\mathbf{z}')\mathbf{x}^\mathsf{T}\mathbf{x}'. \tag{5.13}$$

Now replacing the linear kernel $\mathbf{x}^T\mathbf{x}'$ in Equation 5.13 with any non-linear kernel function $k(\mathbf{x},\mathbf{x}')$ renders a nonlinear version of the hierarchical Bayes models. However, we observed that training non-linear Bayes models does not improve the methods' prediction performance for our HIV therapy screening application. Therefore, in the following we only consider linear multi-task Bayes methods.

5.3 Experiments and Results

5.3.1 Data Sets

The training data are again extracted from the EuResist database – the same version as in Chapter 4. We include a therapy as a sample in the training data if there is a viral sequence obtained shortly before the therapy was started (up to 90 days before) and if it can be assigned a label (success or failure) based on the virus load values measured during its course. The information on the viral genotype is given in terms of the presence of any from a set of predefined resistance-relevant mutations (based on the list in Johnson et al. (2008)) encoded with a binary vector. The therapy label is determined as in Chapter 4: if the virus load drops below 400 cp/ml in the period from 21 days after the start of the therapy to its end we label it successful (1); otherwise we label it failing (-1). We represent the individual drugs comprising each therapy by a binary vector indicating the presence or absence of all drugs appearing in the data set. Finally, we end up with a training set that includes 6750 labeled therapy samples with 805 distinct therapy combinations. Note that the labeled data set is larger than the one in Chapter 4. The reason for this is that we no longer need to remove the combination therapies comprising drugs for which GPP (genotype-phenotype pairs) data are not available.

In the two previous chapters we observed that the HIV clinical data sets have uneven and sparse therapy representation. The histogram of therapy frequencies in Figure 5.3 confirms this observation for our current labeled data set: almost 500 therapies occur less than five times; for almost all therapies there are no more than 50 samples. While there are many rare therapies, there is a reasonable number of samples in which each of the different drugs appear. This can be observed in Figure 5.4, where the majority of the drugs are involved in hundreds of samples.

5.3.2 Validation Settings

The quality of our approach is assessed in two validation scenarios: the *therapy-stratified cross-validation scenario*, and the *time-oriented scenario*.

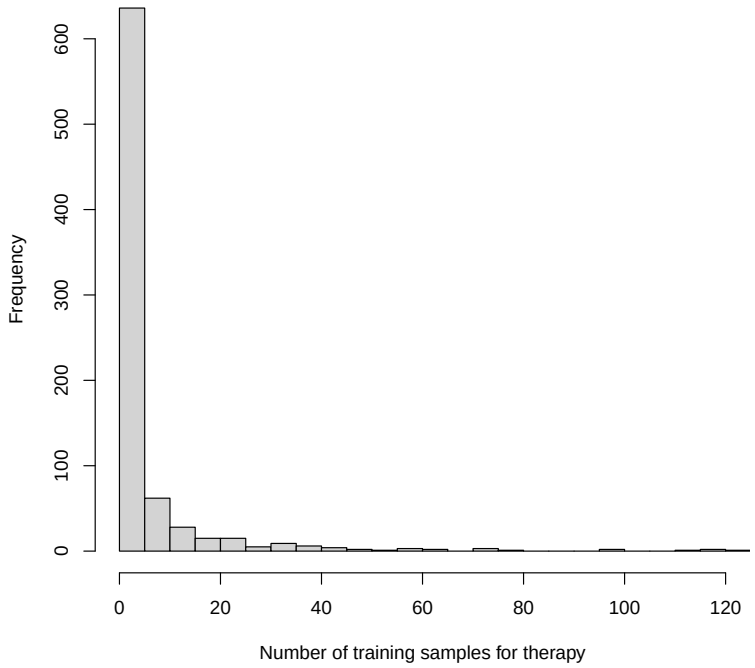


Figure 5.3: Histogram that groups the 805 distinct combination therapies in our labeled training data set based on their corresponding number of available training examples. The image displays the uneven therapy representation in the data where almost 500 therapies are represented with less than five samples.

Therapy-stratified cross-validation scenario. In order to provide an assessment of the performance of a target method that stratifies for therapy abundance in the training data set, we introduce the therapy-stratified cross-validation scenario. We start by describing the procedure of creating therapy-stratified cross-validation folds. First, all available samples are grouped in therapy bins based on their corresponding therapies. Then, we populate the cross-validation folds with samples: the folds are repeatedly visited one after the other and one sample is assigned to each fold at a time; the assigned samples are chosen at random from the therapy bins, which are traversed in a round-robin fashion. In this way we make sure that both infrequent and abundant therapy samples are distributed evenly among the cross-validation folds. In the following we detail the therapy-stratified cross-validation scenario that we applied for our computational experiments. We first construct a separate test set, that comprises 20% of the available data, by selecting one fold from a five-fold therapy-stratified cross validation. Then, we conduct a 10-fold therapy-stratified cross validation on the remaining data and use it for the model selection. At the end, we first report the cross-validation results and then evaluate the selected model on the separate test set.

Time-oriented scenario. The practical experience with the drugs acquired over time and the introduction of new antiretroviral drugs affect the trends of treating HIV patients. In

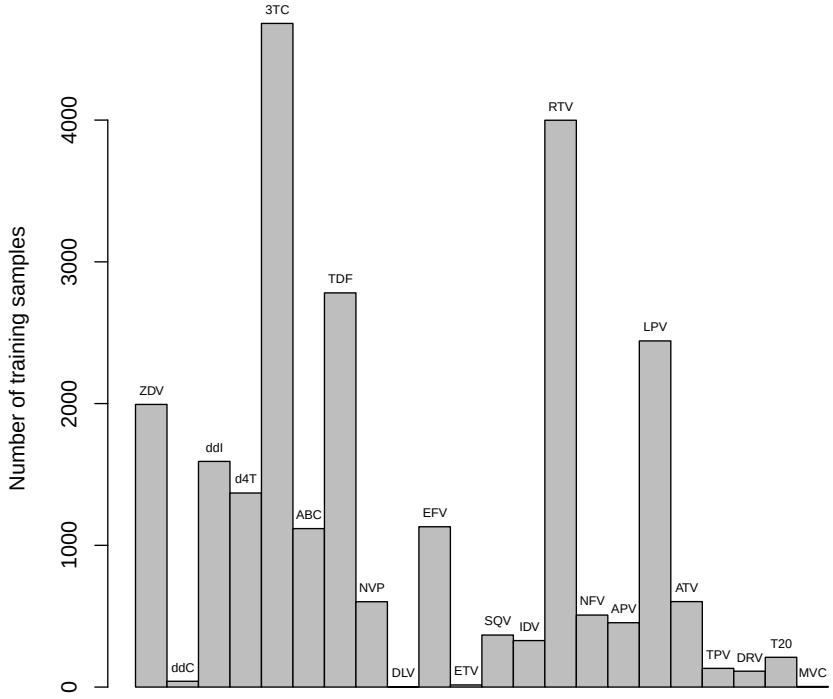


Figure 5.4: Sample abundances for each of the distinct drugs that appear in our labeled clinical data.

order to account for the changing treatment trends over time we use the time-oriented scenario introduced in Chapter 3. In this setting we select the training and test data with a time-oriented approach. So in order to make sure that we have a reasonable amount of training samples for each individual drug that appears in the test set, we remove the therapy samples which contain very recent drugs from our data set. Our multi-task Bayes approach utilizes the high frequencies of samples involving the individual drugs to address the problem of the low frequencies of samples corresponding to specific combination therapies. As a side remark, in practice one cannot expect quality predictions for therapy samples comprising a drug for which there are only few training samples. The resulting data set contains 6336 samples.

Model performance. We assess the performance of the target models by taking the uneven representation of the different therapies into account. We do this by grouping the therapies in the test set based on the number of samples they have in the training set, and then measuring the model performance on each of the groups. We thereby assess the performance of the models for the rare and the abundant therapies, separately. We carry out the model selection based on AUC (Area Under the ROC Curve) results and use the AUCs to assess the model performance. In this way we evaluate the quality of

the ranking of the therapies based on their success probabilities. For the comparison of the cross-validation AUC performances of two methods we use a paired t-test. In order to compare the performance of two methods on a separate test set, the standard errors of the AUC values and the significance of the difference of two AUCs are estimated as described in Hanley and McNeil (1983). We use the ROCR package to plot the ROC curves (Sing et al., 2005).
In this chapter, we are primarily interested in improving the quality of the ranking of the therapies based on their success probability. However, for the sake of completeness, on the one hand, and in order to demonstrate the quality of the absolute results, on the other hand, we also carry out the model selection using the accuracy (ACC) as a performance measure and report the accuracy results. For the comparison of the cross-validation accuracy performances of two methods we again use a paired t-test. Furthermore, in order to compare the performance of two methods on a separate test set, the significance of the difference of two accuracies as well as their standard errors are calculated as described in Section 3.4 of Chapter 3.

Reference methods. In our computational experiments we compare the performance of the two multi-task Bayes methods described in this chapter to those of two reference approaches, namely the *one-for-all model* and the *therapy-specific model*. The *one-for-all* method mimics the most common approach in the field where a single linear logistic regression model is trained on all available therapy samples in the data set. The information on the individual drugs comprising each of the therapies is encoded in a binary vector and supplied together with the other input features. The therapy-specific model represents the approaches that deal with the uneven and sparse therapy representation by training a separate model for each combination therapy using not only the samples from the target therapy but also the available samples from similar therapies with appropriate sample weights. It implements the drugs kernel therapy similarity model (Bogojeska et al., 2010) on the input feature space defined in this chapter. Since training separate models for every different therapy in a cross-validation setting is very time-consuming, we only consider this approach as a reference model in the time-oriented validation scenario. Note also that in the papers where they are introduced (Bickel et al., 2008; Bogojeska et al., 2010) the performance of the therapy-specific approaches is evaluated in the time-oriented validation scenario.

5.3.3 Experimental Results

In this subsection we first present the results of the computational experiments for the therapy-stratified cross-validation scenario, followed by the results of the time-oriented scenario.

Therapy-stratified cross-validation scenario. Table 5.1 summarizes the cross-validation AUC performance of the considered methods: *drug additivity Bayes*; *drug additivity + history Bayes*; and *one-for-all* as the reference method. The two Bayes approaches significantly outperform the *one-for-all method* for the therapies that have few $(0-7)$ available samples in the training set. We verified the significance of the improvements with the

Table 5.1: AUCs with their corresponding standard errors for our two multi-task Bayes models (drug additivity, drug additivity + hist) and the reference (one-for-all) method. Generated by a 10-fold therapy-stratified cross validation for three groups of test therapies: with $0-7$, $8-30$, and more than 30 training samples, they summarize the AUC performance for both the rare and the abundant test therapy samples.

method	multi-task Bayes		one-for-all
	drug additivity	drug additivity + hist	
$0-7$ (SE)	0.771 (0.016)	0.774 (0.016)	0.749 (0.015)
$8-30$ (SE)	0.745 (0.011)	0.738 (0.012)	0.732 (0.010)
>30 (SE)	0.772 (0.017)	0.765 (0.018)	0.759 (0.012)

paired t-test: *p-value* $= 0.05$ for the *drug additivity* and *p-value* $= 0.06$ for the *drug additivity + hist model*. Moreover, for the group of therapies with $8-30$ training samples the *drug additivity* approach also shows significantly better cross-validation performance than the reference model (*p-value* $= 0.05$). All models deliver comparable predictions for the group of therapy samples for which there is a reasonable number (more than 30) of available samples in the training set. According to the AUC results for the separate test set, depicted in Figure 5.5, the *drug additivity* model has better AUC performance than the reference method for all three therapy groups ($0-7$, $8-30$, and more than 30). However, the improvements are only significant for the test therapies with $0-7$ and more than 30 training samples, with *p-values* of 0.045 and 0.002, respectively. The *p-value* for the therapies with $8-30$ training samples is 0.157. The *drug additivity + hist model* shows significantly better AUC performance than the *one-for-all method* for the rare therapies with a *p-value* $= 0.034$. Figure 5.6 depicts the ROC curves for all considered methods for the rare test therapies in the separate test set.
The accuracy results for the therapy-stratified cross validation and the separate test set for all considered methods are shown in Table 5.2 and Figure 5.7, respectively. It can be observed that all approaches have comparable accuracy performance for all considered groups of test therapies, *i.e.* for both the rare and the abundant ones. This observation is confirmed by the relevant statistical tests with *p-values* larger than 0.1 for all pairwise method comparisons.

Time-oriented scenario. The AUC results for the time-oriented scenario are summarized in Figure 5.8. Note that in this case both the *one-for-all* and the *therapy-specific* models are considered as reference methods. As can be observed, the *drug additivity* method outperforms the *one-for-all* method for the test therapies with $0-7$ and $8-30$ training samples. According to the paired difference test described in (Hanley and McNeil, 1983), the improvement is significant only for the test therapies with $0-7$ samples (*p-value*$=$ 0.078). The *p-value* for the test therapies with $8-30$ training samples is 0.132. Compared to the *therapy-specific* model the *drug additivity* model has better AUC performance for the

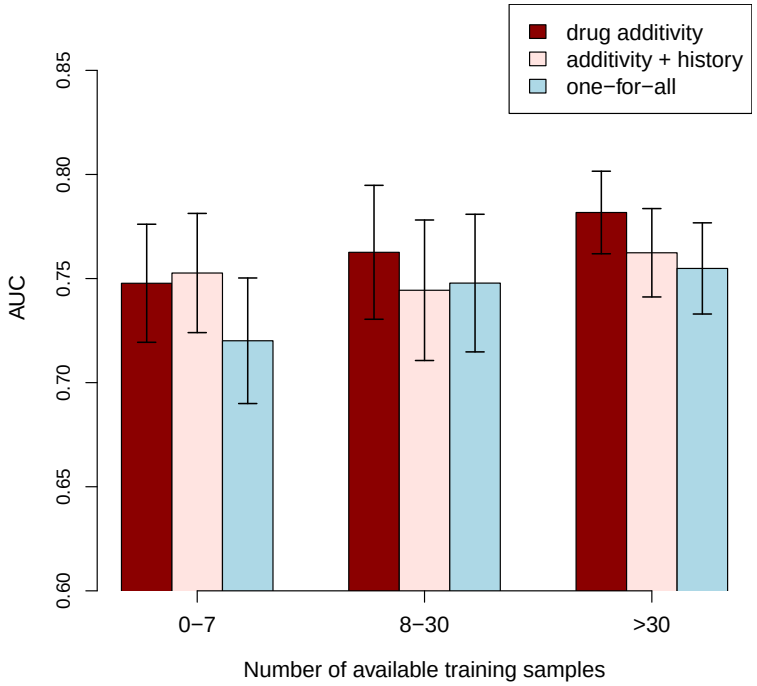


Figure 5.5: AUC results of the different models obtained on the separate test set in the cross-validation therapy-stratified scenario. The test samples are grouped based on the number of training examples for their corresponding therapy combinations. Error bars indicate the standard errors of the AUCs.

Table 5.2: Accuracies (ACC) with their corresponding standard errors for our two multi-task Bayes models (drug additivity, drug additivity + hist) and the reference (one-for-all) method. Generated by a 10-fold therapy-stratified cross validation for three groups of test therapies: with $0-7$, $8-30$, and more than 30 training samples, they summarize the accuracy performance for both the rare and the abundant test therapy samples.

method	multi-task Bayes		one-for-all
	drug additivity	drug additivity + hist	
$0-7$ (SE)	0.723 (0.041)	0.719 (0.039)	0.714 (0.034)
$8-30$ (SE)	0.748 (0.039)	0.753 (0.032)	0.746 (0.037)
>30 (SE)	0.849 (0.019)	0.847 (0.023)	0.841 (0.022)

test therapies with $0-7$ training samples, yet this improvement is not significant (*p-value* $= 0.253$). For the test therapies with $8-30$ training samples both the *therapy-specific* and the *drug additivity* models have comparable performance. The *drug additivity + hist model* outperforms all considered approaches for the rare test therapies (with $0-7$ training

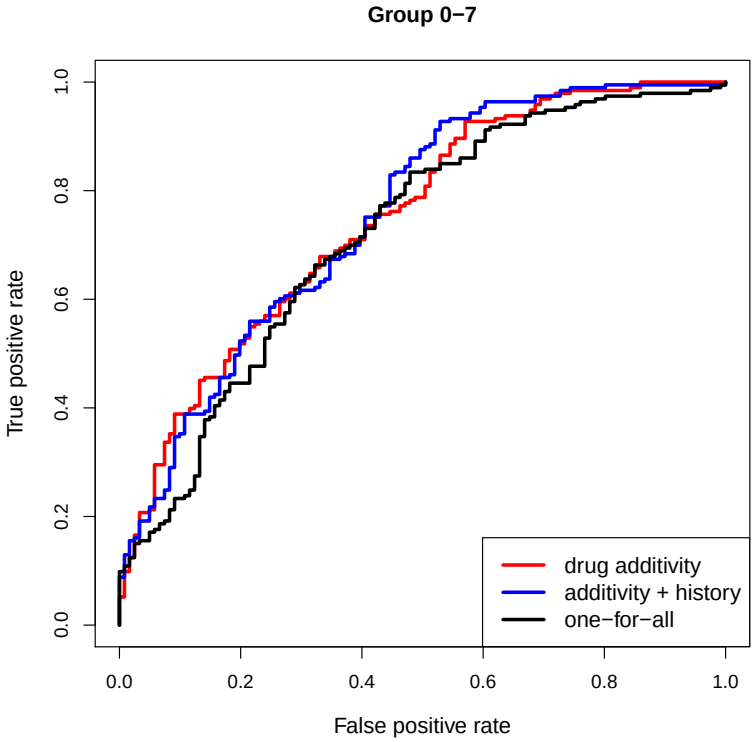


Figure 5.6: ROC curves displaying the performance of the different methods on the rare therapies (with $0-7$ training samples) of the separate test set obtained in the cross-validation therapy-stratified scenario.

samples) with estimated *p-values* of 0.007 for the *one-for-all*, 0.042 for the *therapy-specific* and 0.033 for the *drug additivity* model; for the test therapies with $8-30$ training samples it delivers similar performance as the *one-for-all* method. The AUC results of the *drug additivity + hist model* for the test therapies with $8-30$ training samples are slightly worse compared to the *therapy-specific* and the *drug additivity* models. However the respective differences in performance are not significant (*p-values* > 0.1). Considering the abundant test therapies (with more than 30 training samples) all approaches deliver comparable results. The relevant ROC curves for the rare test therapies are shown in Figure 5.9.
Figure 5.10 depicts the accuracy results for the two Bayes approaches and the two considered reference methods in the time-oriented scenario. All methods deliver comparable accuracies for the test therapies with $0-7$ and $8-30$ available training samples with *p-values* from all pairwise comparisons larger than 0.1. Considering the abundant test therapies with more than 30 available training samples the *drug additivity* Bayes model and the *therapy-specific* model achieve significantly better accuracies than the *one-for-all* method with estimated *p-values* of 0.052 and 0.002, respectively. All other pairwise method comparisons for this group yield *p-values* larger than 0.2.

To summarize, according to the presented AUC results, the two multi-task Bayes approaches have their prime advantage for therapies with few (less than eight) available training samples in both validation scenarios. The *drug additivity Bayes* performs better

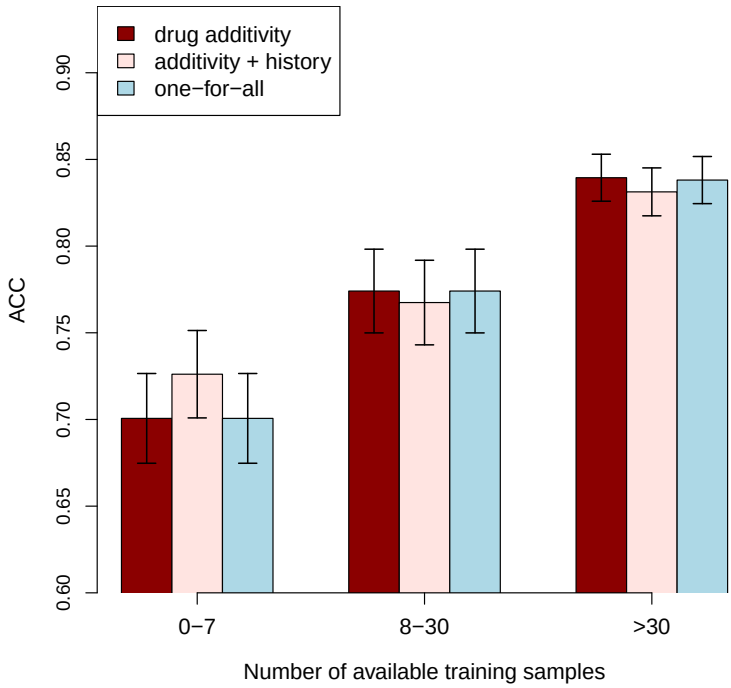


Figure 5.7: Accuracy (ACC) results with their corresponding standard errors for the different models obtained on the separate test set in the cross-validation therapy-stratified scenario. The test samples are grouped based on the number of training examples for their corresponding therapy combinations.

than the *one-for-all* and the *therapy-specific* methods for the therapies with $8-30$ available samples, however the improvement is statistically significant (corresponding *p-value* < 0.1) only for the cross-validation results. For the abundant test therapies (with more than 30 training samples) all considered methods have comparable performance in almost all validation scenarios – one exception is the significantly better performance of the *drug additivity Bayes* method for the separate test set in the cross-validation scenario. Finally, the accuracy performance of the multi-task Bayes approaches is at least as good as the accuracy performance of all considered reference approaches.

5.4 Discussion

This chapter presents an approach to predicting virological response to HIV combination therapies by considering each individual antiretroviral drug as a separate task in a multi-task hierarchical Bayes framework. With our method the additive effects of the individual drugs comprising each combination therapy on its response are modeled from the data. It is worth noting that the most common approaches in the field that use linear models and encode the therapy information in the input feature space, also implicitly use a drug additivity assumption. However, in this case, the effects of the drugs comprising each therapy on its response are not explicitly modeled. Instead, a generic statistical learning method

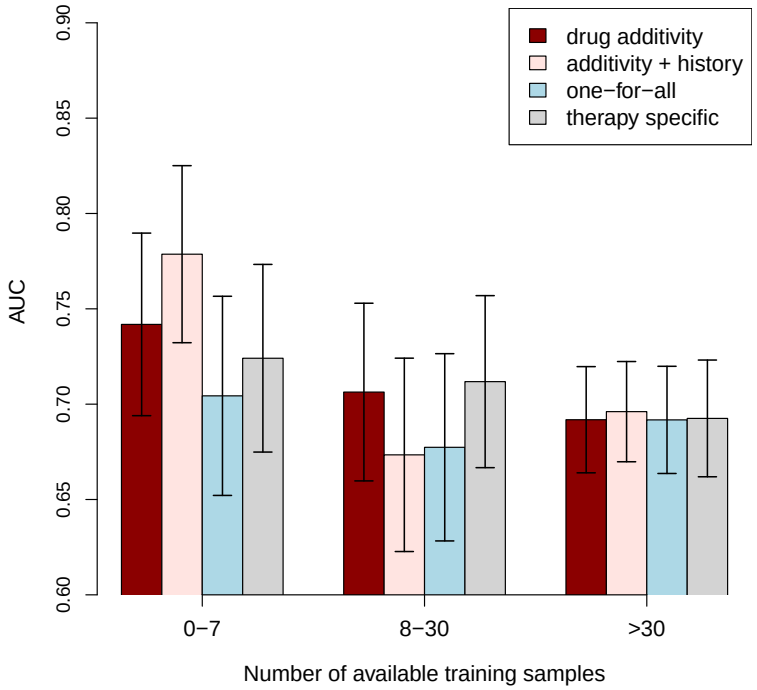


Figure 5.8: AUC results of the different models obtained on the test set in the time-oriented validation scenario. The test samples are grouped based on the number of training examples for their corresponding therapy combinations. Error bars indicate the standard errors of the AUCs.

that simultaneously models the contributions of all available information (e.g. therapy, viral genotype) on the therapy outcome is used. Above all, such methods do not take the uneven therapy representation in the clinical data sets into account. By considering each drug as a separate task, our Bayes approach uses the abundance of samples that pertain to each drug to circumvent the lack of samples for the specific combination therapies. In this way we provide more accurate predictions (rankings) for rare therapies by maintaining the prediction quality for the more frequent therapies. The samples corresponding to rare therapies (represented with $0-7$ samples in our clinical data) make up only around 18% of the available data, but they contain 83% of the different therapies i.e. they make up the therapy variety in our data set. Moreover, our approach uses an extended input feature space where each drug has a separate range and it thereby models the interactions among the input features of the different drugs.

The use of an efficient optimization method (Lin et al., 2008) that takes advantage of the sparseness of our input data ensures very fast model fitting (one second) and model selection. For example, the model selection procedure performed with a 10-fold cross validation for the drugs additivity model screens 289 different value combinations for the two model selection parameters specified in the Methods section and is completed in about ten hours.

According to the cross-validation AUC results both our multi-task Bayes models perform

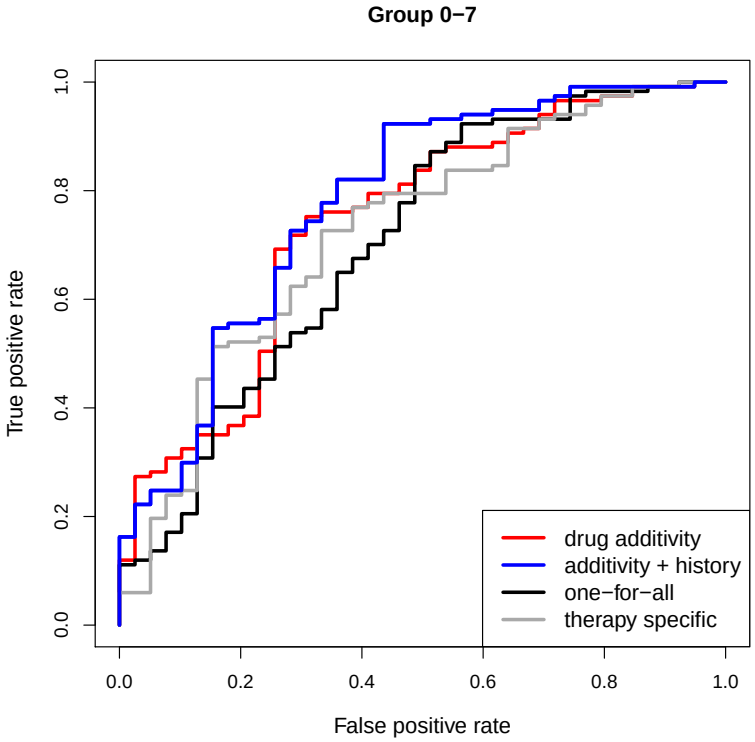


Figure 5.9: ROC curves displaying the performance of the different methods on the rare therapies (with $0-7$ training samples) of the test set obtained in the time-oriented validation scenario.

significantly better (at the 5% significance level for the *drugs additivity* and the 6% level for the *drugs additivity + hist* model) than the *one-for-all* model for rare test therapies (with $0-7$ available training samples). The *drugs additivity* model also significantly outperforms the *one-for-all scenario* for the group of therapies with $8-30$ training samples. For therapies with a sizeable number of samples (above 30) all approaches show comparable cross-validation performance. The AUC results on the left-out set in the cross-validation scenario confirms the advantage of both multi-task Bayes models for the less frequent therapies (the significance level is 5% for the *drugs additivity* and the 3% level for the *drugs additivity + hist* model). Furthermore, the *drugs additivity* model achieves better AUC performance for the other two groups of test therapies (with $8-30$ and more than 30 training samples). However, the improvement is only significant for the test therapies with more than 30 available training samples.

According to the time-oriented scenario both Bayes models significantly outperform (at the 8% significance level for the *drugs additivity* and the 1% significance level for the *drugs additivity + hist* model) the *one-for-all* model for the test therapies with less than eight available training samples. Moreover, the *drugs additivity + hist* model also outperforms the *drugs additivity* model (at the 3% significance level) and the *therapy-specific* model (at the 4% significance level) for the group of rare test therapies. All models show comparable AUC performance for the abundant test therapies.

It is also worth noting that the accuracy performance of the multi-task Bayes approaches is

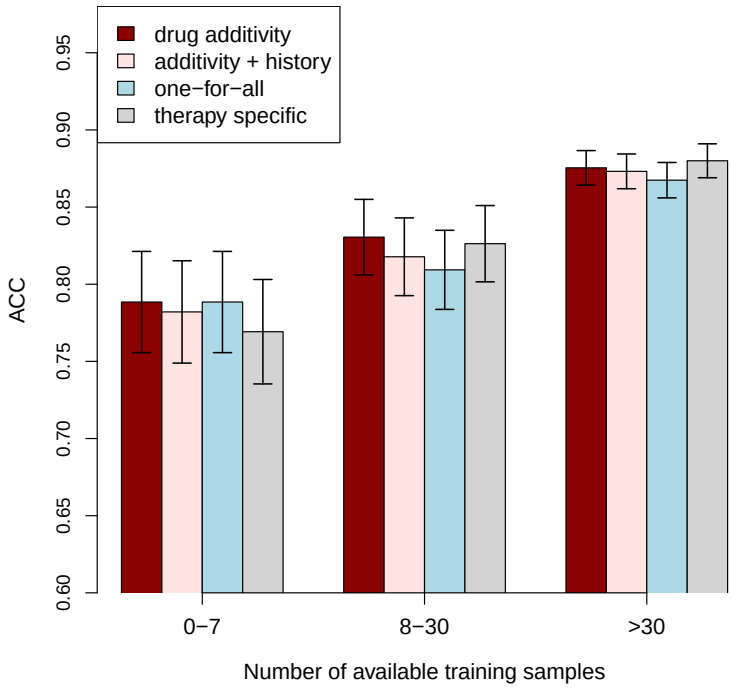


Figure 5.10: Accuracy (ACC) results with their corresponding standard errors for the different models obtained on the test set in the time-oriented validation scenario. The test samples are grouped based on the number of training examples for their corresponding therapy combinations.

at least as good as the accuracy performance of all considered reference approaches for all test therapies in both validation scenarios. This demonstrates the quality of the absolute results achieved by the Bayes approaches.

In summary, the approach presented in this chapter models the effects of the individual drugs comprising an HIV combination therapy on its effectiveness by using a multi-task hierarchical Bayes approach. The AUC performance of this approach is at least as good as an approach that encodes therapy information in the input feature space for the abundant therapies and significantly better for therapies with few training samples. The same observation holds when comparing the AUC performance of the hierarchical Bayes approaches to the therapy-specific approach which trains a separate model for every different drug combination. In this case the Bayes models have the additional advantage of being more time-efficient compared to the therapy-specific approach. Note that the group of rare therapies is very important as it makes up the therapy variety in the available clinical data.

6 History-aware Methods for Predicting Virological Response to HIV Combination Therapy

In previous chapters (Chapters 3 to 5) we observed that the clinical data sets contain data on different combination therapies with widely differing frequencies. In particular, many therapies are only represented with very few data points. Nonetheless, beside the sparse therapy representation there are several other important issues affecting the HIV clinical data sets. First of all, they comprise only viral strain(s) that can be detected in the patients' blood serum. However, past therapies leave a genomic fingerprint of associated resistance mutations among the latent viral population stored in different organs of the patient. This information is important for making accurate predictions for therapy-experienced HIV patients. Second, the clinical data comprise therapy samples that originate from patients with different treatment backgrounds. Also the specific treatment histories for the majority of the therapy-experienced samples are unique. Furthermore, the various levels of therapy experience ranging from therapy-naïve to heavily pretreated are represented with different sample abundances – especially samples stemming from patients with higher therapy-experience levels are underrepresented. All these issues create what we will refer to as *treatment bias* in the data sets which influences the predictions of the derived statistical models, especially the predictions pertaining to samples originating from therapy-experienced patients and to samples associated with rare therapies.

This chapter presents two novel methods that account not only for the sparse, uneven therapy representation but also for the bias originating from the different treatment backgrounds of the samples making up the clinical data sets. More specifically, in the first half of the chapter we present a statistical learning method that incorporates information on the latent virus population, the specific therapies previously given to a patient and the order in which they were administered to deal with the different kinds of bias present in the clinical data. In the second half of the chapter we consider two additional problems pertaining to the available HIV clinical data sets. The first one is that the clinical data do not necessarily have the complete information on all administered HIV therapies for all patients and the information on whether all administered therapies are available or not is also missing for the majority of the patients. The second issue is the increasing imbalance between the effective and ineffective therapies over time. The quality of treating HIV patients has largely increased in the recent years due to the knowledge acquired from HIV research and clinical practice. This renders the amount of effective therapies in recently collected data samples much larger than the amount of ineffective ones. To this end, we present an approach that addresses all previously mentioned problems simultaneously. To tackle the issues of the uneven therapy representation and the different treatment back-

grounds of the samples, we use information on both the current therapy and the patient's treatment history. Additionally, our method uses a distribution matching approach to account for the problems of missing information in the treatment history and the growing gap between the abundances of effective and ineffective HIV therapies over time in the clinical data sets.
We should point out that this chapter provides an extended version of the work presented in Bogojeska et al. and Bogojeska (2011).

6.1 Related Work

In the recent years there have been several statistical learning methods (Bickel et al., 2008; Rosen-Zvi et al., 2008) that utilize information from previous therapies when predicting the outcome of a potential antiretroviral therapy. Moreover, two recent studies (Revell et al., 2010; Prosperi et al., 2010) show that a substantial amount of information about the effectiveness of a therapy can be deduced from the treatment history even if viral genotypic information is absent. In the aforementioned publications the information on treatment history has been flattened to the set of different drugs that have been administered in any of the therapies that comprise the relevant treatment history record. While this simple approach can easily be incorporated in every statistical learning method, it neglects the information on the specific makeup of the drug combinations comprising the patient's treatment history, their resulting viral genetic fingerprints in the latent virus population and the order in which they were administered. There is medical evidence that the order in which therapies are administered affects therapy response (Robbins et al., 2003). Moreover, Saigo et al. (2010) present an approach denoted as *sequence boosting* for predicting therapy effectiveness targeted at therapy-experienced patients with completely recorded treatment history. It uses novel feature encoding, referred to as sequence representation, for capturing all available history information: previous therapies and their corresponding responses, previous viral genotypes. Then, it searches for subsequences discriminative for therapy response by first enumerating all sequence features using a tree structure and then pruning the tree based on a gain function. The sequence boosting method incorporates information on the order in which the therapies were administered and shows the importance of such information for treatment-experienced patients. The two main shortcomings of this method are given as follows. Firstly, in the available HIV clinical data the information on whether the available treatment history is complete or not as well as information on previous genotypes and outcomes of past therapies is missing for the majority of the samples. Secondly, sequence boosting is non-linear and thus computationally very demanding – in terms of complexity it is NP-hard.
None of the approaches mentioned above tackles the bias introduced by the different treatment backgrounds of the samples and their sparse representation in the clinical data sets.

6.2 History-similarity Model for HIV Therapy Screening

We present an approach, referred to as *history-similarity model*, that tackles the treatment bias in the HIV clinical data by introducing a notion of treatment similarity which includes not only information on the current therapy but also detailed information on the

treatment history. More specifically, it considers two treatments as similar if they have similar treatment patterns and their genomic fingerprint in the latent viral population is similar. Our approach trains a separate model for each sample of interest by using all available training samples, each with a specific weight, that reflects the similarity of the corresponding treatment pattern to the treatment pattern of the target sample. In this way we address the different treatment backgrounds of the clinical samples, their differing sample abundances, the hidden latent virus population and the uneven therapy representation in the clinical data sets. In what follows we first describe the problem setting, then provide detailed description of the similarity measure of therapy sequences and finally present the history-similarity model.

6.2.1 Problem Setting

Let $\mathbf{x}$ denote the viral genotype represented as a binary vector indicating the occurrence of a set of resistance-relevant mutations, let $\mathbf{z}$ denote the therapy combination encoded as a binary vector that indicates the individual drugs comprising the current therapy and let $\mathbf{h}$ denote a binary vector representing the drugs administered in all known previous therapies for the specific therapy example. The label y indicates the success (1) or failure (-1) of each therapy sample. Let $D = \{(\mathbf{x}_1, \mathbf{z}_1, \mathbf{h}_1, y_1), \ldots, (\mathbf{x}_m, \mathbf{z}_m, \mathbf{h}_m, y_m)\}$ denote the training set and let $\mathbf{t}$ denote the therapy sample of interest. Let $start(\mathbf{t})$ denote the point of time when the therapy $\mathbf{t}$ was started and $patient(\mathbf{t})$ denote the patient identifier corresponding to the therapy sample $\mathbf{t}$. Then:

$$r(\mathbf{t}) = \{\mathbf{z} \mid (start(\mathbf{z}) \leq start(\mathbf{t})) \; and \; (patient(\mathbf{z}) = patient(\mathbf{t}))\}$$

denotes the complete treatment record associated with the therapy sample $\mathbf{t}$ and is referred to as *therapy sequence*. It contains all known therapies administered to $patient(\mathbf{t})$ not later than $start(\mathbf{t})$ ordered by their corresponding starting times, from older to newer. Note that each therapy sequence also contains the current therapy, *i.e.* the most recent therapy in the therapy sequence $r(\mathbf{t})$ is $\mathbf{t}$. Our goal is to train a model $f(\mathbf{x}, \mathbf{t}, \mathbf{h})$ that correctly predicts the outcome of the target therapy $\mathbf{t}$ for given viral genotypes by utilizing the information from its associated therapy sequence.

6.2.2 Similarity of Therapy Sequences

Our main objective when quantifying the similarity of therapy sequences is to consider two therapy sequences similar if they consist of similar drug combinations administered in a similar order and producing similar genomic fingerprints in the latent viral population. We first quantify the pairwise similarity between different drug combinations and then use it together with the order in which the therapies were administered to compute the overall similarity between two therapy sequences. Since we lack primary data on the latent virus population, the pairwise therapy similarity measure considers the genomic fingerprint the therapies leave in the viral genome as a surrogate. This fingerprint comprises resistance-relevant mutations of the drugs making up the therapy.

We quantify the pairwise similarities between different therapy combinations with the *resistance mutations kernel*, which uses the table of resistance-associated mutations of each drug afforded by the International AIDS society (Johnson et al., 2008). The kernel

assumes that the similarity between different drug groups is additive. This is a reasonable assumption since drugs belonging to different groups have different targets and/or modes of action and thus can be assumed to act independently (Beerenwinkel et al., 2003b). Formally, the kernel is defined as follows. Let G denote the set of different drug groups. In our clinical data set we have three drug groups: NRTIs (Nucleoside Reverse Transcriptase Inhibitors), NNRTIs (Non-Nucleoside Reverse Transcriptase Inhibitors) and PIs (Protease Inhibitors). Let $\mathbf{u}_{zg}$ and $\mathbf{u}_{z'g}$ be binary vectors indicating the resistance-relevant mutations for the set of drugs occurring in drug group $g \in G$ of the therapies $\mathbf{z}$ and $\mathbf{z}'$, respectively. The similarity between the drug-g mutations of the two therapies $\mathbf{z}$ and $\mathbf{z}'$ is then calculated by:

$$sim_g(\mathbf{z}, \mathbf{z}') = \frac{\mathbf{u}_{zg}^\mathsf{T} \mathbf{u}_{z'g}}{\max(\|\mathbf{u}_{zg}\|^2, \|\mathbf{u}_{z'g}\|^2)},$$

where $\mathbf{x}^\mathsf{T}\mathbf{y}$ denotes the scalar product of the vectors $\mathbf{x}$ and $\mathbf{y}$, and $\|\cdot\|$ is the L_2-norm. We derive the similarity $k_m(\mathbf{z}, \mathbf{z}')$ between the therapies $\mathbf{z}$ and $\mathbf{z}'$ by averaging the similarities of their corresponding drug groups:

$$k_m(\mathbf{z}, \mathbf{z}') = \sum_{g \in G} \frac{sim_g(\mathbf{z}, \mathbf{z}')}{|G|}.$$

Since the group similarities $sim_g(\mathbf{z}, \mathbf{z}')$ lie in the interval $[0, 1]$, the values of the resistance mutations kernel are also within $[0, 1]$. Intuitively, the higher the number of common resistance relevant mutations associated with the corresponding sets of drugs making up the two therapies of interest, the higher their similarities. In this way the therapy similarity also accounts for the similarity of the genetic fingerprint of the potential latent virus populations of the compared therapies. Furthermore, our kernel represents drugs in terms of their mutation profile and, by doing so, allows for high group similarity for non-identical drugs that have very similar resistance mutation profiles. In this way we take the high level of cross resistance within the same drug class into account.

Once we have determined the pairwise similarities of different drug combinations, we will use them to quantify the pairwise similarities between complete therapy sequences. We need a similarity score that accounts for both the similarity of the different therapies comprising the therapy sequences and the order in which they were administered. Thus we can adapt the score commonly used for assessing the quality of an alignment of protein or nucleic acid sequences. In what follows we give the details of how to align therapy sequences.

Let $X = [x_1, \ldots, x_{|X|}]$ and $Y = [y_1, \ldots, y_{|Y|}]$ be two therapy sequences defined over a finite alphabet Σ with lengths $|X|$ and $|Y|$, respectively. The pair of sequences (X', Y') defined over the alphabet $\{\Sigma \cup "-"\}$ that includes the gap character "$-$" denotes their sequence alignment when the following conditions are fulfilled:

- $|X'| = |Y'|$.
- X' and Y' become X and Y, respectively, after deleting all gap characters "$-$".
- There is no position i such that $x'_i = y'_i = -$.

Each alignment can be associated with a score that determines its quality:

$$S(X', Y') = \sum_{i=1}^{|X'|} s(x'_i, y'_i),$$

where s is a similarity function that quantifies all pairwise similarities of all letters in the alphabet $\{\Sigma \cup "-"\}$. Of course only good alignments with as few gaps as possible are of interest. In this sense an optimal alignment (X^*, Y^*) is the one that maximizes the alignment score S:

$$(X^*, Y^*) = \arg \max_{(X', Y')} S(X', Y').$$

The solution of this maximization problem is obtained by applying the Needleman-Wunsch algorithm (Needleman and Wunsch, 1970) specified with the following recursion:

$$\begin{aligned} S_{00} &= 0 \\ S_{i,0} &= S_{i-1,0} + s(x_i, -) \\ S_{0,j} &= S_{0,j-1} + s(-, x_j) \\ S_{i,j} &= \max \begin{cases} S_{i-1,j} + s(x_i, -) \\ S_{i-1,j-1} + s(x_i, y_j), \\ S_{i,j-1} + s(-, x_j) \end{cases} \end{aligned}$$

where $S_{i,j}$ is the score of the optimal alignment of the subsequences $x_1, \ldots, x_i$ and $y_1, \ldots, y_j$, and $s(x, -)$ and $s(-, x)$ are the *gap costs*.

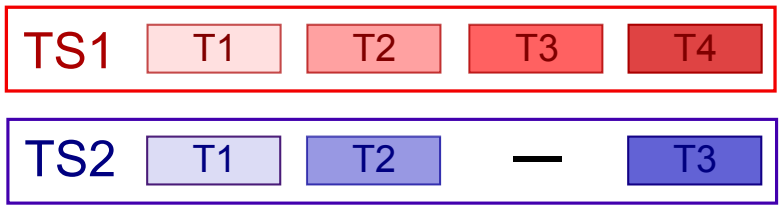


Figure 6.1: Alignment of the blue therapy sequence TS2 to the red therapy sequence TS1 comprising three and four therapies, respectively. Their most recent therapies – T4 in TS1 and T3 in TS2 – are matched. Since the therapy sequence TS2 is shorter than TS1 the alignment contains one gap.

The alphabet used for the therapy sequence alignment comprises all distinct drug combinations making up the clinical data set. The mutations kernel determines the pairwise similarities s between its letters. Each therapy sequence ends with the current (most recent) therapy – the one that determines the label of the sample. Therefore, we adapt the sequence alignment algorithm such that the rightmost (most recent) therapies (characters) are always matched, *i.e.* we do not allow for gaps at the right end of an alignment. In this way we also address the problem of the sparse, uneven representation of the different therapies. We apply linear gap cost penalty. The parameter specifying the gap cost is

selected in the model selection procedure. The score of such an optimal alignment quantifies the pairwise similarity of therapy sequences and is referred to as *alignment similarity kernel*. An example alignment of two therapy sequences is depicted in Figure 6.1. It should also be pointed out that since it is a sum that uses the mutations kernel values, the alignment similarity kernel also reflects the similarity of the accumulated mutations (genetic fingerprints) of the latent virus population of the compared therapy sequences.

6.2.3 History-similarity Method

The history-similarity model utilizes the alignment similarity kernel to train a separate model for every sample of interest. The details of this method for a given target sample are summarized in Algorithm 2.

Algorithm 2: History-similarity method

Input: Target sample with corresponding current therapy $\mathbf{t}$ and therapy sequence $r(\mathbf{t})$.

1. Calculate the weights for all training samples $\{S(r(\mathbf{z}_i), r(\mathbf{t})), i = 1, \ldots, m\}$.
2. Apply linear rescaling to normalize the alignment similarity weights to the range of $[0, 1]$:
$$S(r(\mathbf{z}_i), r(\mathbf{t})) = \frac{S(r(\mathbf{z}_i), r(\mathbf{t})) - \min_i S(r(\mathbf{z}_i), r(\mathbf{t}))}{\max_i S(r(\mathbf{z}_i), r(\mathbf{t})) - \min_i S(r(\mathbf{z}_i), r(\mathbf{t}))}.$$
3. Use the weights $\{S(r(\mathbf{z}_i), r(\mathbf{t})), i = 1, \ldots, m\}$ to estimate the final model for the target sample – minimize weighted loss on training data.

The first step utilizes the alignment similarity kernel: the therapy sequence of the target sample $r(\mathbf{t})$ is aligned to the corresponding therapy sequences of all training samples $\{r(\mathbf{z}_i), i = 1, \ldots, m\}$ and the resulting alignment scores $\{S(r(\mathbf{z}_i), r(\mathbf{t})), i = 1, \ldots, m\}$ are the weights for the training samples. Then, the second step applies linear rescaling to normalize the sample weights to the range of $[0, 1]$. Once the sample weights are available we can proceed to step three and train the final model that predicts the therapy response for the sample of interest. For this purpose we use regularized logistic regression model (described in Chapter 2 and Evgeniou et al. (2000)) that minimizes the loss over the weighted training samples:

$$\arg\min_{\mathbf{w}_t} \frac{1}{|D|} \sum_{(\mathbf{x}_i, \mathbf{z}_i, \mathbf{h}_i, y_i) \in D} S(r(\mathbf{z}_i), r(\mathbf{t}))^{\gamma} \cdot \ell(f(\mathbf{x}_i, \mathbf{z}_i, \mathbf{h}_i, \mathbf{w}_t), y_i) + \sigma \mathbf{w}_t^T \mathbf{w}_t, \tag{6.1}$$

where σ is the regularization parameter, γ is the smoothing parameter and $\mathbf{w}_t$ is the model parameter. In the minimization above we use all available training samples, from therapy-naïve to heavily pretreated, to produce a separate model for each sample of interest or, if we have a specific test set, for each test sample. Intuitively, the history-alignment approach estimates a model tailored towards the sample of interest such that it up-weights those

samples that are relevant for the target sample and down-weights the remaining samples. In this manner the method accounts for the various treatment backgrounds associated with the samples making up the clinical data sets, the different abundances of the levels of therapy experience, the latent virus population and the sparse therapy representation. Note also that by using the alignment similarity kernel which allows for gaps enables our method to utilize information from samples with incomplete treatment histories.
As an important aspect in every biomedical application, interpretability should be one of the properties of our prediction models. We thus use linear logistic regression and the loss function in the formula above is given by:

$$\ell(f(\mathbf{x}, \mathbf{z}, \mathbf{h}, \mathbf{w}_t), y) = \ln(1 + \exp(-y\mathbf{w}_t^T[\mathbf{x}, \mathbf{z}, \mathbf{h}])).$$

Our approach of training a separate model for each target sample demands an efficient method for minimizing the loss function. The choice for linear models and the sparse input feature space, provided by the binary input features, offer the possibility to use the trust region Newton method for training linear logistic regression (for more details see Chapter 3 and Lin et al. (2008)). In this way we ensure real-time model fitting (in the range of few milliseconds) and time-efficient model selection.

6.2.4 Validation Setting

Data set. Same as in the previous chapters, the data source for our models is the EuResist database (Rosen-Zvi et al., 2008) that contains information on 93014 antiretroviral therapies administered to 18325 HIV (subtype B) patients from several countries in the period from 1988 to 2008. We point out that the clinical data do not necessarily have the complete information on all administered HIV therapies for all patients. Furthermore, the information on whether all administered therapies are available or not is also missing for the majority of the patients. Therefore, the statistical methods utilize only the available information. The viral sequence assigned to each therapy sample is obtained shortly before the respective therapy was started (up to 90 days before). The response to a given therapy is quantified with a label (success or failure) based on the virus load values measured during its course. The label assignment is identical to the one described in Chapter 4. The information on the viral genotype is given in terms of a binary vector indicating the presence (1) or absence (0) of a set of predefined resistance-relevant mutations derived from the list given in Johnson et al. (2008). The currently administered therapy is also encoded by a binary vector that indicates the presence or absence of all drugs appearing in the data set. The set of drugs administered in all available therapies preceding the current therapy is represented in the same manner. Finally, our training set comprises all samples providing a viral sequence and a label; it includes 6537 labeled therapy samples from 690 distinct therapy combinations.

Time-oriented validation scenario. The trends of treating HIV patients change over time as a result of the gathered practical experience with the drugs and the introduction of new antiretroviral drugs. As in the previous chapters, in order to account for this phenomenon we use the time-oriented validation scenario introduced in Chapter 3 which makes a time-oriented split when selecting the training and the test set. In this way, our models are

trained on the data from the more distant past, while their performance is measured on the data from the more recent past. This scenario is more realistic than other scenarios since it captures how a given model would perform on the recent trends of combining the drugs.

The therapy samples gathered in the HIV clinical data sets are associated with patients whose treatment histories differ in length: while some patients receive their first antiretroviral treatment, others are heavily pretreated. Moreover, these different sample groups, from treatment-naïve to heavily pretreated, are represented with different abundances in the HIV clinical data. Figure 6.2 depicts a histogram of the frequencies of the previously mentioned sample groups in the training data set, where it can be observed that the number of samples stemming from patients in early stages of HIV treatment is much higher than the number of samples from therapy-experienced patients (with more than five or more than ten previously administered therapies). The numbers are based on the therapy-history information in our clinical data set. We should also point out that most of the therapy sequences associated with patients in the mid or late stages of HIV treatment are unique, *i.e.* the representation of specific longer therapy sequences in the clinical data sets is very sparse.

The search for an effective HIV therapy is particulary challenging for patients in the mid to late stages of antiretroviral therapy when the number of therapy options is reduced and effective therapies are increasingly hard to find because of the accumulated drug resistance mutations from all previous therapies. Therefore, in our computational experiments we want to elucidate the predictive power of the models in dependence on the level of therapy experience. In order to do this, we group the therapy samples in the test set into different bins based on the number of therapies administered prior to the therapy of interest – the current therapy. Note that for some patients some therapy information might be missing. Thus, with the sample binning we make sure that the samples in the treatment-experienced bin (denoted by > 5) originate from patients that had at least five previous therapies. Some details on each of the bins are given in Table 6.1. We assess the quality of a given target model by reporting its performance for each of the bins.

Table 6.1: Details on the bins grouping the test samples based on their corresponding number of previous therapies.

Bin	$0-2$	$3-5$	>5
Sample count	807	225	275
Success rate	89%	82%	68%

Another important property of our approach is its ability to address the uneven and sparse representation of the different therapies. This property arises from the definition of similarities of therapy sequences where the current therapies are always matched. In order to consider the sparse representation of the different therapies when assessing the performance of our models we adopt the validation scenario from Chapters 4 and 5: the therapies in the test set are grouped based on the number of samples they have in the training set, and then the model performance on each of the groups is measured. The details on the sample

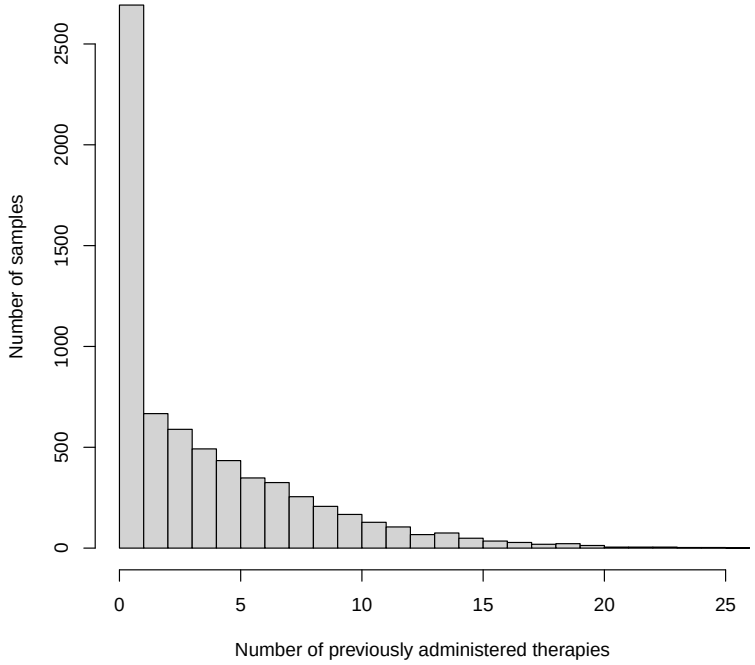


Figure 6.2: Histogram that groups all labeled samples in our clinical data set based on their corresponding number of known previous therapies. The histogram illustrates the uneven representation with respect to the length of the treatment history in the data, where the largest group with size 2600 consists of samples with none or only one known previous therapy.

counts and the success rates in each of the bins are given in Table 6.2.

Table 6.2: Details on the bins grouping the test samples based on the number of training examples for their corresponding therapy combinations.

Bin	$0-7$	$8-30$	>30
Sample count	217	242	848
Success rate	77%	82%	85%

In order to be able to assist the selection of a potential combination therapy for HIV patients our method should provide a good ranking based on the probability of therapy success. For this reason, we carry out the model selection based on AUC (Area Under the ROC Curve) results and use AUC to assess the model performance. The standard errors of the AUC values and the significance of the difference of two AUCs used for the pairwise method comparison are estimated as described in Hanley and McNeil (1983).

Reference methods. In our computational experiments we compare the results of our history-similarity approach, denoted as *history-similarity validation scenario*, to those of the *one-for-all validation scenario* and the *one-for-all + hist mutations validation scenario*, which are used as reference methods. The one-for-all reference method mimics the most common approach in the field where a single linear logistic regression model is trained on all available therapy samples in the data set. The information on the individual drugs comprising the target (most recent) therapy and the drugs administered in all its available preceding therapies are encoded in a binary vector and supplied as input features. On the one hand, this is a very simple way of incorporating knowledge on treatment history into a statistical model. On the other hand, however, it neglects the information on the specific drug combinations comprising the patients' treatment history and the order in which they were administered. We should also point out that removing the similarity score weights from the history-similarity approach yields the one-for-all method. The *one-for-all + hist mutations* approach is a modified version of one-for-all approach where the drugs from the drug indicator representation of the treatment history are replaced with their respective cumulative resistance-mutation profiles. In this way the accumulated mutations of the latent virus population are encoded in the input feature space.

When assessing the ability of our history-similarity model to address the uneven representation of the different therapies in the clinical data sets we also consider the *therapy-specific model* as a second reference method. It represents the approaches that deal with the uneven, sparse therapy representation by training a separate model for each combination therapy by using not only the samples from the target therapy but also the available samples from similar therapies with appropriate sample weights. It implements the drugs kernel therapy similarity model as described in Bogojeska et al. (2010) on the input feature space defined in the beginning of this section.

6.2.5 Experimental Results

In what follows, we first present the results of the validation experiments of the time-oriented validation scenario stratified for the length of treatment history, followed by the results stratified for the abundance of the different therapies.

The experimental results for the history-similarity method and the two one-for-all reference methods stratified for the length of treatment history are summarized in Figure 6.3. For samples with a small number of previously administered therapies (≤ 5), *i.e.* with short treatment histories, the three considered models have comparable performance. The low AUC values of all methods for the group of samples with very short history lengths (≤ 2) are to be expected. Based on the information available in our clinical data this group comprises samples from therapy-naïve patients (about 75%) and samples from patients who had only one or two previous HIV therapies. Therefore, most of them are successful – the success rate is 89%. The main reason for ineffectiveness of initial therapies is lack of adherence. An additional reason for observing failing therapies in the bin of samples with short treatment histories is the wrong assignment of treatment history lengths due to the incomplete records on patient histories in the database. All these issues may be causes for the low AUCs for the samples with short treatment history. One should also point out that there are specific guidelines for both treating therapy-naïve patients with

first-line therapy and administering the first couple of follow-up therapies, which normally are successfully applied. This is also reflected in the high success rate in our clinical data for this group of therapy samples (see Table 6.1). Thus assistance is mainly necessary for therapy-experienced patients. According to the paired difference test described in Hanley and McNeil (1983) the history-similarity model that incorporates knowledge on the specific therapies comprising the treatment history, their latent virus population and the order in which they were applied significantly improves the performance for the test samples stemming from patients with longer treatment histories (> 5) over the two reference models with $p\text{-}value = 0.001$ for the one-for-all and $p\text{-}value = 0.005$ for the one-for-all + hist mutations model. Figure 6.4 depicts the ROC curves for this group created by using ROCR (Sing et al., 2005).

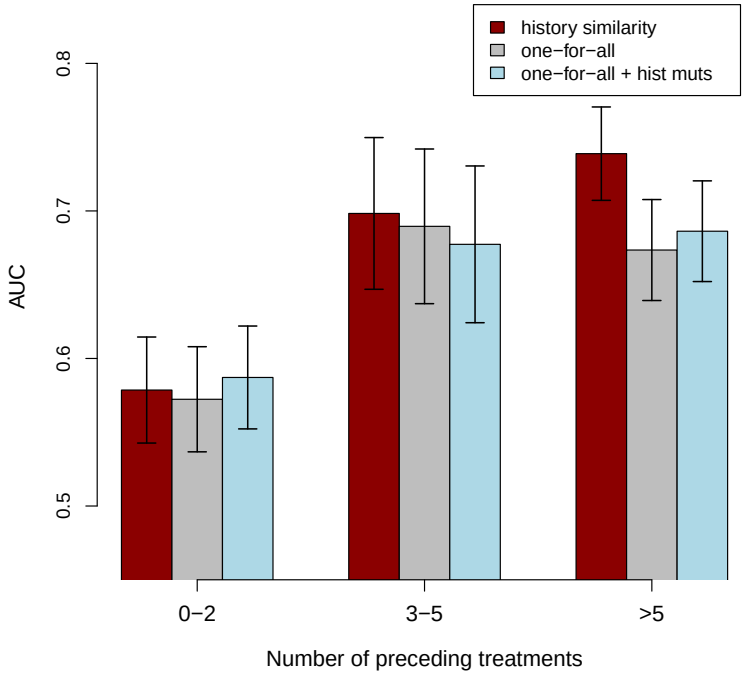


Figure 6.3: Experimental results stratified for the length of treatment history obtained on the test set in the time-oriented validation scenario. The depicted barplot represents the AUC values with their corresponding standard errors for the history-similarity approach and the reference (one-for-all, one-for-all + hist mutations) models. The test samples are grouped by their corresponding number of available previously administered therapies – length of treatment history.

The experimental results stratified for the abundance of the therapies are summarized in Figure 6.5. As can be observed, the history-similarity method achieves better results than the three reference methods for the test therapies with $0 - 7$ available training samples. According to the paired difference test described in Hanley and McNeil (1983), the improvement is significant with estimated $p\text{-}value = 0.018$ for the one-for-all, $p\text{-}value = 0.050$ for the one-for-all + hist mutations, and $p\text{-}value = 0.008$ for the therapy-specific model.

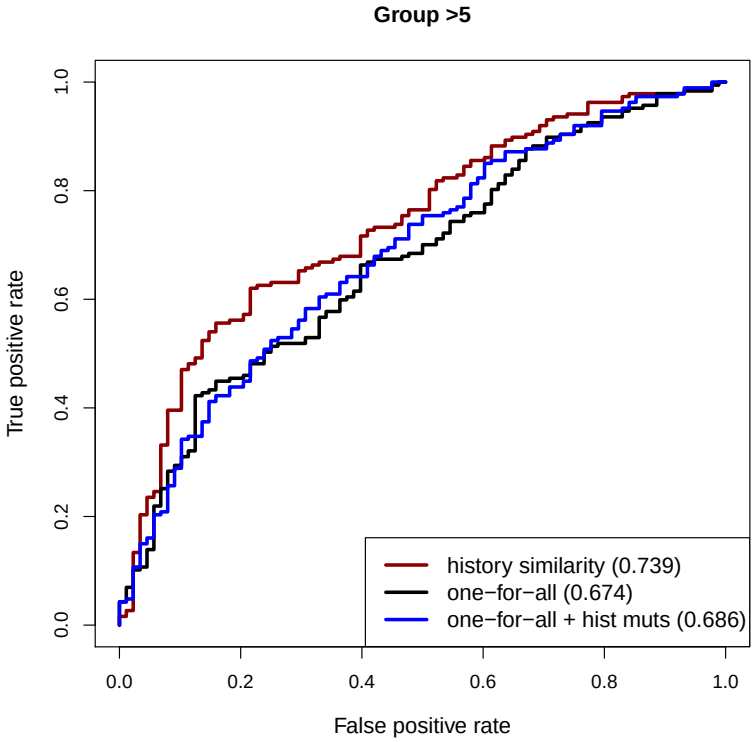


Figure 6.4: ROC curves depicting the performance of the history-similarity approach and the reference models for the group of test samples with more than five previously administered therapies (> 5). The AUC values for each method are given in the legend.

Considering the test therapies with $8-30$ and more than 30 training samples all considered approaches deliver comparable results with no significant differences. The relevant ROC curves for the rare test therapies are shown in Figure 6.6.

6.2.6 Discussion

We presented the history-similarity learning approach for predicting the outcome of combination therapies that trains individual model for each target sample. Each of these models weights different training samples differently: the more similar the respective therapy sequences to the target therapy sequence the higher their importance for the respective model. The similarity of the therapy sequences is quantified by means of sequence alignment which incorporates information on the resistance-relevant mutations as described in Subsection 6.2.2. In this way we account for the bias imposed by the different treatment histories of the samples in the clinical data and we extract information on the genetic fingerprint of the latent virus population. According to the experimental results this approach significantly outperforms the reference method for test therapies associated with treatment-experienced patients (with at least five previous treatments) and exhibits comparable performance for the rest of the test therapies. Moreover, the comparison of the history-similarity approach to the one-for-all + hist mutations method demonstrates once

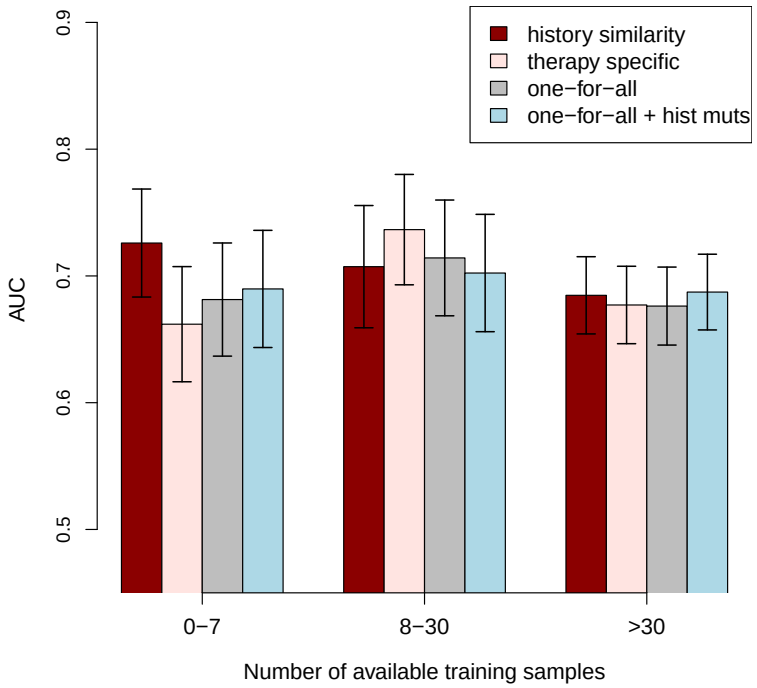


Figure 6.5: Experimental results stratified for the abundance of therapies obtained on the test set in the time-oriented validation scenario. The depicted barplot represents the AUC values with their corresponding standard errors for the history-similarity approach and the three reference models: the one-for-all, the one-for-all + hist mutations, and the therapy-specific model. The test samples are grouped based on the number of training examples for their corresponding therapy combinations.

again that the detailed information on all therapies previously given to a patient and the order in which they were administered reflected in the alignment similarity kernel is important for the performance gain of our approach. Considering the available guidelines for choosing the several initial HIV treatments and their high success rates, on the one hand, and the difficulty of choosing successful therapies for heavily pretreated patients, on the other hand, availability of statistical methods that focus on providing high-quality models for treatment-experienced patients is becoming increasingly important. We should also point out that the history-similarity approach is very patient specific since it trains sample-specific models that use very detailed treatment history information. In this manner it makes one step further in the direction of personalized HIV treatment.

Our model also addresses the uneven therapy representation in the clinical data sets and outperforms the reference methods for rare test therapies. This is an important feature because the rare therapies comprise 61% of the different therapies in the test set.

An example of how the history-similarity approach can tackle the bias in the clinical data sets introduced from the different treatment backgrounds of the samples and their sparse representation is illustrated in Figure 6.7. From the image of the therapy sequence

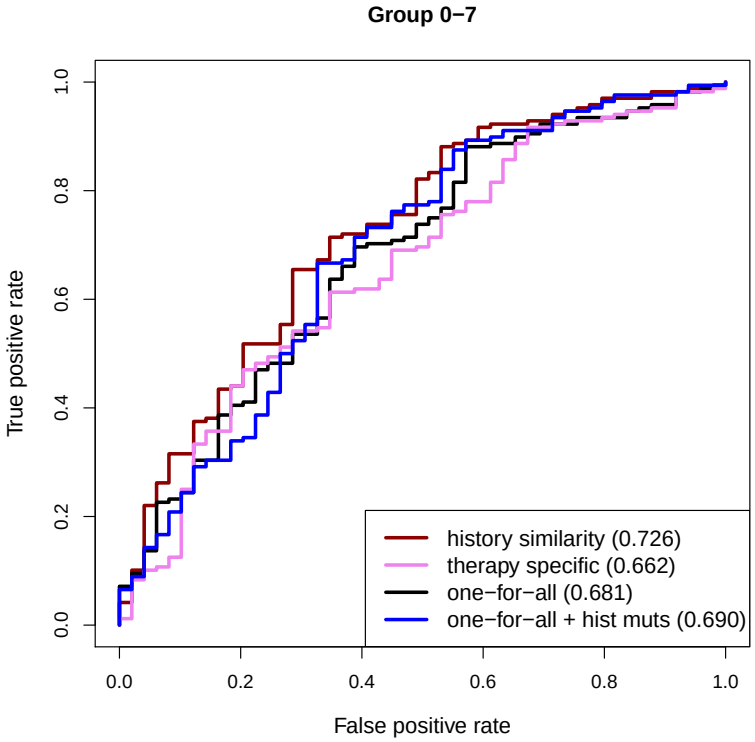


Figure 6.6: ROC curves displaying the performance of the history-similarity approach and the reference models for the rare therapies (with $0-7$ training samples) of the test set. The AUC values for each method are given in the legend.

corresponding to the sample of interest (Figure 6.7 (a)) we can observe that the target model predicts the outcome of the therapy *ZDV 3TC SQV TDF RTV LPV* – this is the most recent therapy in the therapy sequence, and the therapy sequence has a length of nine. Furthermore, Figure 6.7 (b) depicts the three most relevant therapy sequences for this specific model. Here the relevance is reflected in the similarity of the training therapy sequences to the target therapy sequence. One can observe that the most recent therapies in these sequences are similar to the most recent target therapy. Moreover, the corresponding training samples originate from pretreated patients. Also the average length of the therapy sequences for the 100 most relevant training samples for the considered model is 11. In this way the target model assigns the highest relevance to the training samples originating from therapy-experienced patients with therapy sequences similar to the target therapy sequence and thereby compensates for the bias caused by the different treatment backgrounds of the training samples and the sparse representation of therapy sequences. Furthermore, the available information on the contribution of each training combination therapy to predicting the outcome of the sample of interest is an important aspect of model interpretability. Such information details the most relevant training therapy sequences for a given target therapy sequence and thereby enables access to the argumentative basis of the predictions. More detailed insight of the impact of the complete training sample on the predictions for the target sample (Figure 6.7 (a)) is depicted in Figure 6.8. It images the distribution of the training sample weights for the therapy sequence of the target sample.

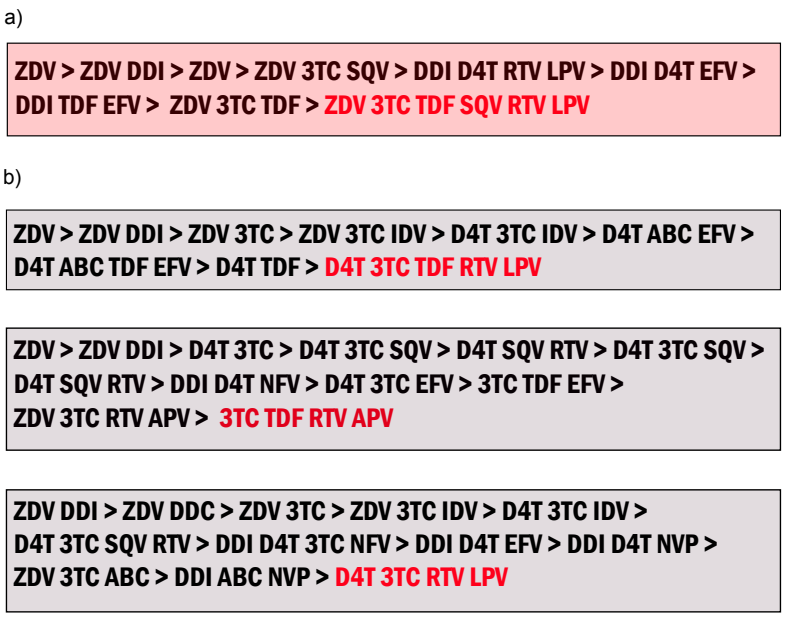


Figure 6.7: Example demonstrating the ability of the history similarity approach to tackle the bias introduced by the different treatment backgrounds of the samples. (a) target therapy sequence; (b) therapy sequences of the three most relevant training therapies for the given target therapy sequence. The therapies comprising each therapy sequence are given from older to newer where the current (latest) therapy is depicted in red and > denotes a treatment change.

An additional contribution to model interpretability is achieved by assessing the relevance of the different input features, namely, the mutations in the viral sequence, the drugs comprising the current therapy and the drugs appearing in all previous therapies. This can easily be accomplished if we observe that we have a separate model for each sample and each of these models is trained on features describing the viral genome, the current therapy and the drugs from all previous therapies. In such a setting the importance of an input feature of the target model quantifies its relevance. One way to quantify feature importance is by computing the z-scores for each model coefficient: the higher the magnitude of the z-score the more significant the feature. In this manner we perform a statistical test for each model coefficient that checks the null hypothesis that the considered coefficient is zero, while all others are not. Figure 6.9 and 6.10 illustrate an interesting example for the relevance of the different input features for the sample with therapy sequence depicted in Figure 6.7 (a). In Figure 6.9 one can observe the importance of the different mutations for the considered therapy sequence. Thus, the three most important mutations are given as follows: 16, 30 and 54 for the protease sequence; and 151, 70 and 230 for the reverse transcriptase sequence. According to Figure 6.10 the drugs comprising the current therapy ordered by their relevance are given by: *LPV TDF 3TC RTV ZDV SQV*, and the three most important drugs from the drugs administered in the history of the considered therapy sequence are: *LPV*, *RTV* and *DDI*.

One disadvantage of the history-similarity method is that it is quite compute-intensive, since it trains an individual model for each target sample. To overcome this problem we

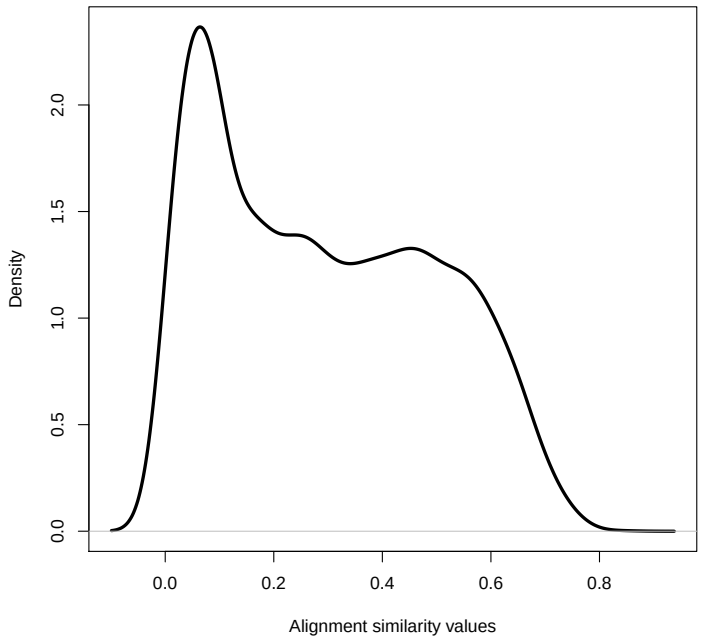
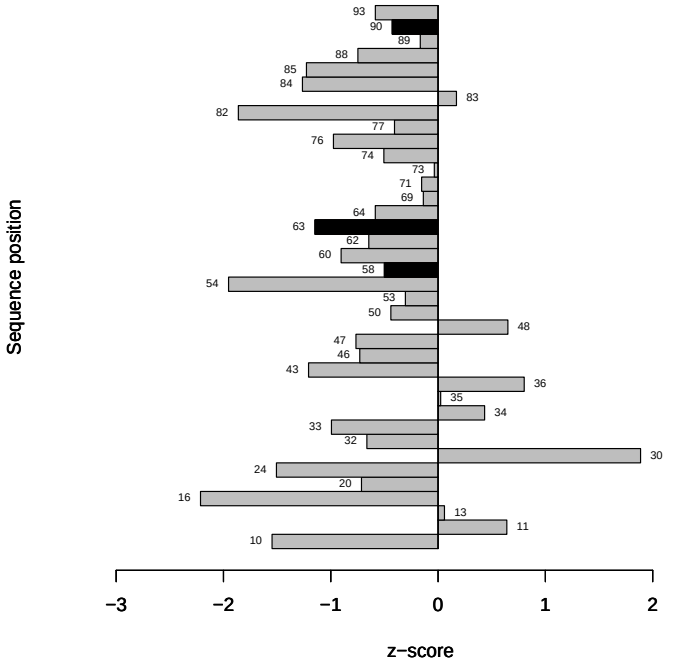
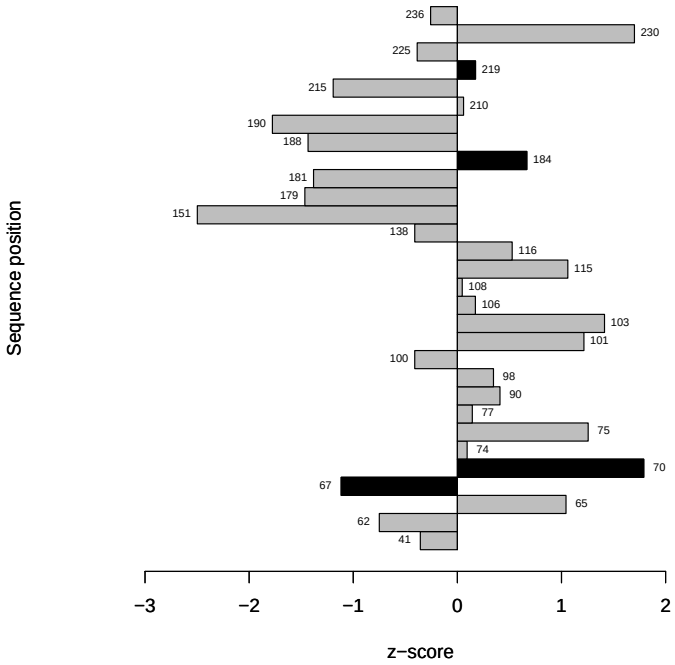


Figure 6.8: Distribution density of the sample-specific similarity weights for the training set corresponding to the therapy sequence depicted in Figure 6.7 (a).

use the trust region Newton optimization for logistic regression (Lin et al., 2008) and thus provide an efficient way for training the individual models – one model is trained in a fraction of a second. Moreover, the alignment similarities of the target therapy sequence to all training therapy sequences is computed in about four minutes for our training data with 5230 samples. The only bottleneck is computing the pairwise similarity alignments for all training samples in the model selection procedure. However, they can be precomputed and stored for all different therapy sequences in the available clinical data set. Thus, new alignment scores need to be computed only if the training set is extended with new samples whose corresponding therapy sequences are not among the ones appearing in the previous version of the training data. Whenever we encounter such sequences we can compute their alignment scores for all training therapy sequences and store them together with the others. This enables fast model selection procedure whenever there is an update of the training data. More specifically, our tuning set comprises 261 different combination therapies, and thus for a given precomputed similarity alignment kernel the model selection procedure screens 456 different value combinations for the two model selection parameters (the regularization and the smoothing parameter) specified in Optimization Problem 6.1 in Subsection 6.2.3. In total, we train 119016 logistic regression models. Our implementation completes the model selection procedure in less than 12 hours and this procedure only needs to be repeated whenever there is an update of the training data set.

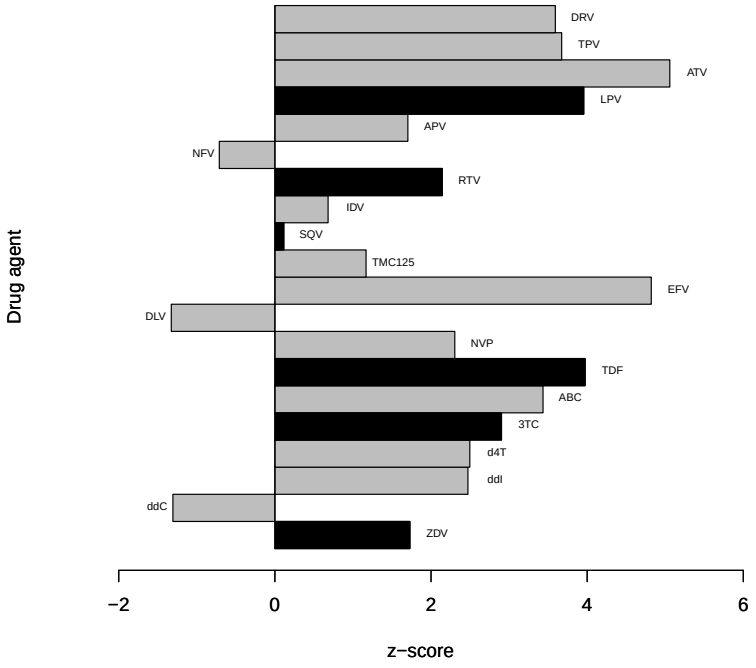


(a)

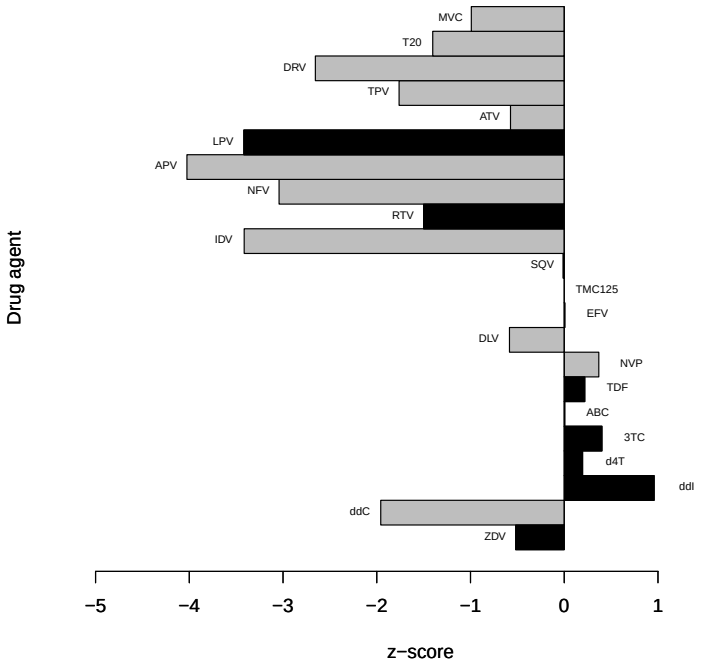


(b)

Figure 6.9: Barplot of z-scores showing the importance of the different viral sequence related input features: (a) protease mutations; and (b) reverse transcriptase mutations for the therapy sequence depicted in Figure 1 (a). The features appearing in the specific target sample are depicted in black.



(a)



(b)

Figure 6.10: Barplot of z-scores showing the importance of the different drug input features: (a) drugs comprising the current therapy; and (b) drugs appearing in treatment history for the therapy sequence depicted in Figure 1 (a). The features appearing in the specific target sample are depicted in black.

6.3 History Distribution Matching Method

The history distribution matching approach predicts the effectiveness of HIV combination therapies by simultaneously addressing several problems regarding the available HIV clinical data sets: the different treatment backgrounds of the samples, the uneven representation of the different levels of therapy experience, the missing treatment history information, the uneven therapy representation and the unbalanced therapy outcome representation especially pronounced in recently collected samples. The main idea behind this method is that the predictions for a given patient should originate from a model trained using samples from patients with treatment backgrounds similar as the one of the target patient. The details of this method are summarized in Algorithm 3. In what follows, we explain each step of this algorithm.

Algorithm 3: History distribution matching method

1. Cluster the training samples by using the pairwise dissimilarities of their corresponding therapy sequences.
2. For each (target) cluster:
   - Match the distribution of all available training samples to the distribution of samples in the target cluster.
   - Train a sample-weighted linear logistic regression model with the sample weights computed in the distribution matching step.

6.3.1 Clustering Based on Similarities of Therapy Histories

Clustering partitions a set of objects into clusters, such that the objects within each cluster are more similar to one another than to the objects assigned to different clusters (Hastie et al., 2009). In the first step of Algorithm 3, all available training samples are clustered based on the pairwise dissimilarity of their corresponding therapy sequences. In the following, we first describe a similarity measure for therapy sequences and then present the details of the clustering.

Similarity of therapy sequences. In order to quantify the pairwise similarity of therapy sequences we use a slightly modified version of the *alignment similarity kernel* introduced in the first part of this chapter. It adapts sequence alignment techniques (Needleman and Wunsch, 1970) to the problem of aligning therapy sequences by considering the specific therapies given to a patient, their respective resistance-relevant mutations, the order in which they were applied and the length of the therapy sequence. It has one parameter that specifies the gap cost.
For the history distribution matching method, we modified the alignment similarity kernel described in the paragraph above such that it also takes the importance of the different resistance-relevant mutations into account. This is achieved by updating the resistance

mutations kernel, where instead of using binary vectors that indicate the occurrence of a set of resistance-relevant mutations, we use vectors that indicate their importance. If two or more drugs from a certain drug group that comprise a target therapy share a resistance mutation, then we consider its maximum importance score. This is the most intuitive way of assigning importance for the resistance mutations relevant for a set of drugs that target the same virus protein. Importance scores for the resistance-relevant mutations are derived from in-vivo experiments and can be obtained from the Stanford University HIV Drug Resistance Database (Liu and Shafer, 2006). Furthermore, we want to keep the cluster similarity measure parameter-free, such that in the process of model selection the clustering Step 1 in Algorithm 3 is decoupled from the Step 2 and is computed only once. This results in time-efficient model selection procedure and is achieved by computing the alignments with zero gap costs. However, in this case only the similarities of the matched therapies comprising the two compared therapy sequences contribute to the similarity score and thus the differing lengths of the therapy sequences are not accounted for. Having a clustering similarity measure that takes the differing therapy lengths into account is important for tackling the uneven sample representation with respect to the level of therapy experience. In order to achieve this we normalize each pairwise similarity score with the length of the longer therapy sequence. This yields pairwise similarity values in the interval $[0, 1]$ which can easily be converted to dissimilarity values in the same range by subtracting them from 1.

Clustering. Given the measure of dissimilarity of therapy sequences, we cluster our data using the most popular version of K-medoids clustering (Hastie et al., 2009), referred to as *partitioning around medoids* (PAM) (Kaufman and Rousseeuw, 1990). The details of PAM are presented in Algorithm 4.
After an initial cluster assignment based on randomly chosen K data samples as cluster medoids (centers), PAM exchanges each medioid with non-medoid data samples and selects the exchange that results in maximum decrease of the objective function. This process is repeated until no exchanges beneficial for the objective function can be found. The main reason why we choose this approach instead of the simpler K-means clustering (Hastie et al., 2009) is that it can use any precomputed dissimilarity matrix. We select the number of clusters with the *silhouette validation technique* (Rousseeuw, 1987), which uses the so called *silhouette value* defined as follows.

Definition 2 (Silhouette Value)**.** *Let $C_1, \ldots, C_k$ denote a cluster assignment for a given data set $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ based on a dissimilarity matrix $\mathbf{d}$ with ij-th element denoted as $\mathbf{d}(i, j)$. Let*

$$a(i) := \frac{1}{|C_k - 1|} \sum_{j \in C_k, j \neq i} \mathbf{d}(i, j)$$

be the average dissimilarity of $\mathbf{x}_i$ to the data points in its cluster C_k. Let

$$b(i) := \min_{C_j \neq C_k} dc(i, C_j),$$

where

$$dc(i, C) := \frac{1}{|C|} \sum_{j \in C} \mathbf{d}(i, j)$$

Algorithm 4: Partitioning Around Medoids (PAM)

Input: An arbitrary dissimilarity matrix $\mathbf{d}$ and the number of clusters K.

1. Initialize medoids $\{\mu_j, j = 1, \ldots, K\}$
   by randomly choosing K numbers (data samples) from $\{1, \ldots, n\}$.

2. Assign each data point i to the closest medoid by computing:

$$C(i) = \arg \min_{j=1,\ldots,K} \mathbf{d}(i, \mu_j).$$

3. Repeat until convergence:
   For each $\{\mu_j, j = 1, ..., K\}$:
   For each $\{i = 1, \ldots, n \text{ and } i \neq \mu_j\}$:
   Swap i with μ_j.
   Compute the objective function:

$$\sum_{i=1}^{n} \min_{j=1,\ldots,K} \mathbf{d}(i, \mu_j).$$

   Select the swap that yields the maximum decrease of the objective function.

4. Compute the resulting cluster assignment as in step 2.

denotes the average dissimilarity of $\mathbf{x}_i$ *to all data samples in the cluster* C*. The silhouette value is then defined as:*

$$s(i) := \frac{b(i) - a(i)}{\min(a(i), b(i))}.$$

The silhouette values lie in the interval $[-1, 1]$, where higher values indicate better clustering of a given target sample. Intuitively, the more similar the target data sample $\mathbf{x}_i$ is to the samples in its respective cluster compared to the samples in any other cluster the higher its corresponding silhouette value $s(i)$. By averaging over the silhouette values of all data samples one obtains the cluster's *silhouette width*. Finally, the silhouette validation technique uses the magnitude of the average silhouette width to assess the quality of the clustering and select the optimal number of clusters.

6.3.2 Cluster Distribution Matching

The clustering step of the history distribution matching method groups the training data into different bins based on their corresponding therapy sequences. However, the complete treatment history (all previously administered combination therapies) is not necessarily available for all patients in our clinical data set. Moreover, the information regarding the completeness of a patient's treatment history is also missing for many patients. Therefore, by restricting the prediction model for a target sample only to the data from its corresponding cluster, the model might ignore relevant information from the other clusters. The approach we use to deal with this issue is inspired by the multi-task learning

with distribution matching method introduced in Bickel et al. (2008). More details on this method are given in Chapter 3.

For the sake of simplicity, in the rest of the chapter we will denote the set of all input features by $\mathbf{x}$, *i.e.* $\mathbf{x} = (\mathbf{x}, \mathbf{t}, \mathbf{h})$ (see Subsection Problem Setting). In our current problem setting, the goal is to train a prediction model $f_c : \mathbf{x} \rightarrow y$ for each cluster c of similar therapy sequences, where $\mathbf{x}$ denotes the input features and y denotes the label. The straightforward approach to achieve this is to train a prediction model by using solely the samples from cluster c. However, since the available treatment history for some samples might be incomplete, totally excluding the samples from all other clusters ($\neq c$) ignores relevant information about the model f_c. Furthermore, the cluster-specific tasks are related and the samples from the other clusters especially those close to the cluster boundaries of cluster c also carry valuable information for the model f_c. Therefore, we use a multi-task learning approach where a separate model is trained for each cluster by not only using the training samples from the target cluster, but also the available training samples from the remaining clusters with appropriate sample-specific weights. These weights are computed by matching the distribution of all samples to the distribution of the samples in the target cluster and they thereby reflect the relevance of each sample for the target cluster. In this way, the model for the target cluster uses information from the input features to extract relevant knowledge from the other clusters. Such knowledge is available because of the missing information from the treatment histories and the similarity of the tasks f_c associated with each cluster.

More formally, let $D = \{(\mathbf{x}_1, y_1, c_1), \ldots, (\mathbf{x}_m, y_m, c_m)\}$ denote the training data, where c_i denotes the cluster associated with the training sample $(\mathbf{x}_i, y_i)$ in the history-based clustering. The training data are governed by the joint training distribution $\sum_c p(c)p(\mathbf{x}, y|c)$. The most accurate model for a given target cluster t minimizes the loss with respect to the conditional probability $p(\mathbf{x}, y|t)$ referred to as the target distribution. In Bickel et al. (2008) and Chapter 3 we have shown that:

$$E_{(\mathbf{x},y)\sim p(\mathbf{x},y|t)}[\ell(f_t(\mathbf{x}))] = E_{(\mathbf{x},y)\sim \sum_c p(c)p(\mathbf{x},y|c)}[r_t(\mathbf{x}, y)\ell(f_t(\mathbf{x}))], \tag{6.2}$$

where:

$$r_t(\mathbf{x}, y) = \frac{p(\mathbf{x}, y|t)}{\sum_c p(c)p(\mathbf{x}, y|c)}. \tag{6.3}$$

In other words, by using sample-specific weights $r_t(\mathbf{x}, y)$ that match the training distribution $\sum_c p(c)p(\mathbf{x}, y|c)$ to the target distribution $p(\mathbf{x}, y|t)$ we can minimize the expected loss with respect to the target distribution by minimizing the expected loss with respect to the training distribution. In this way we train the model for the target cluster t by using all available training samples and thereby tackle the problem of missing data that arises from the incomplete treatment history information. The weighted training data is governed by the correct target distribution $p(\mathbf{x}, y|t)$ and the sample weights reflect the relevance of each training sample for the target model. The weights are derived based on information from the input features. If a sample was assigned to the wrong cluster due to the incompleteness of the treatment history, by matching the training to the target distribution it can still receive high sample weight for the model of its correct cluster.

In order to avoid the estimation of the high-dimensional densities $p(\mathbf{x}, y|t)$ and $p(\mathbf{x}, y|c)$ in Equation 6.3, we follow the example of Bickel et al. (2007, 2008) and compute the sample

weights $r_t(\mathbf{x}, y)$ using a discriminative model for a conditional distribution with a single variable:

$$r_t(\mathbf{x}, y) = \frac{p(t|\mathbf{x}, y)}{p(t)}, \tag{6.4}$$

where $p(t|\mathbf{x}, y)$ quantifies the probability that a sample $(\mathbf{x}, y)$ randomly drawn from the training set D belongs to the target cluster t. $p(t)$ is the prior probability which can easily be estimated from the training data.
As in Bickel et al. (2008), $p(t|\mathbf{x}, y)$ is modeled for all clusters jointly using a kernelized version of multi-class logistic regression with feature mapping that separates the effective from the ineffective therapies:

$$\Phi(\mathbf{x}, y) = \begin{bmatrix} \delta(y, +1)\mathbf{x} \\ \delta(y, -1)\mathbf{x} \end{bmatrix}, \tag{6.5}$$

where δ is the Kronecker delta ($\delta(a, b) = 1$, if $a = b$, and $\delta(a, b) = 0$, if $a \neq b$). In this way, we can train the cluster-discriminative models for the effective and the ineffective therapies independently, and thus, by proper time-oriented model selection address the increasing imbalance in their representation over time. Formally, the multi-class model is trained by maximizing the log-likelihood over the training data using a Gaussian log-prior on the model parameters:

$$\arg\max_{\mathbf{v}} \sum_{(\mathbf{x}_i, y_i, c_i) \in D} \log(p(c_i|\mathbf{x}_i, y_i, \mathbf{v})) - \mathbf{v}^{\mathsf{T}} \Sigma^{-1} \mathbf{v}.$$

In the equation above $\mathbf{v}$ are the model parameters (a concatenation of the cluster specific parameters $\mathbf{v}_c$), and Σ is the covariance matrix of the Gaussian prior.

6.3.3 Sample-weighted Linear Logistic Regression Method

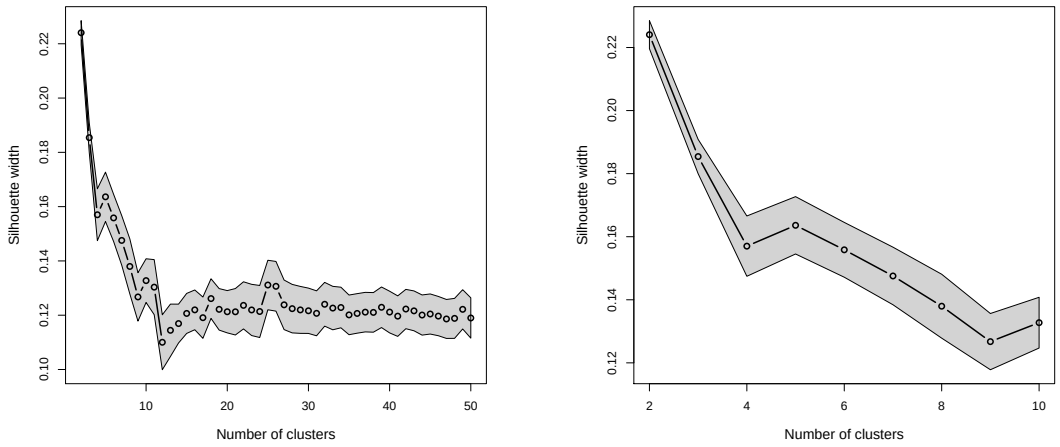
As described in the previous subsection, we use a multi-task distribution matching procedure to obtain sample-specific weights for each cluster, which reflect the relevance of each sample for the corresponding cluster. Then, a separate linear logistic regression model that uses all available training data with the proper sample weights is trained for each cluster. More formally, let t denote the target cluster and let $r_t(\mathbf{x}, y)$ denote the weight of the sample $(\mathbf{x}, y)$ for the cluster t. Then, the prediction model for the cluster t that minimizes the loss over the weighted training samples is given by:

$$\arg\min_{\mathbf{w}_t} \frac{1}{|D|} \sum_{(\mathbf{x}_i, y_i) \in D} r_t(\mathbf{x}_i, y)^{\gamma} \cdot \ell(f_t(\mathbf{x}_i), y_i) + \sigma \mathbf{w}_t^T \mathbf{w}_t, \tag{6.6}$$

where $\mathbf{w}_t$ are the model parameters, σ is the regularization parameter, γ is a smoothing parameter for the sample-specific weights and $\ell(f_t(\mathbf{x}), y) = \ln(1 + \exp(-y\mathbf{w}_t^T \mathbf{x}))$ is the loss of linear logistic regression.
All in all, the history distribution matching method first clusters the training data based on their corresponding treatment sequences and then trains a separate model for each cluster by using relevant data from the remaining clusters. By doing so it tackles the problems of the different treatment backgrounds of the samples and the uneven sample representation in the clinical data sets with respect to the level of therapy experience.

Since the alignment kernel considers the most recent therapy and the drugs comprising this therapy are encoded as a part of the input feature space our method also deals with the differing therapy abundances in the clinical data sets.

Once we have the models for each cluster, we use them to predict the label of a given test sample $\mathbf{x}$ as follows. First of all, we use the treatment sequence of the target sample to calculate its dissimilarity to each of the cluster centers. Then, we assign the sample $\mathbf{x}$ to the cluster c with the closest cluster center. Finally, we use the logistic regression model trained for cluster c to predict the label y for the target sample $\mathbf{x}$. Note that the target therapy sequence is only aligned to the therapy sequences of the cluster centers which enables very fast prediction – in the range of a couple of seconds.



(a) (b)

Figure 6.11: Silhouette widths with their corresponding standard deviations for different number of clusters: (a) two to fifty; and (b) two to ten. The standard deviations are estimated with the bootstrap method ($B = 100$) (Hastie et al., 2009).

6.3.4 Validation Setting

Time-oriented validation scenario. The validation setting for the history distribution matching method is very similar to the one described in the first part of this chapter. The computational experiments are performed on the same clinical data set by using the time-oriented validation scenario, where the test data are first stratified with respect to the length of their corresponding treatment histories and then with respect to the therapy abundance of their corresponding therapies. Since the history distribution matching method aims at addressing the problem of growing imbalance in the therapy outcome representation in the clinical data sets over time, we display the success rates of the training, tuning and test data sets generated in the time-oriented scenario in Table 6.3. It can be observed that there is a large gap between the abundances of the effective and ineffective therapies, especially for the most recent data. The training of the cluster-discriminative models for the effective and the ineffective therapies independently in concert with the se-

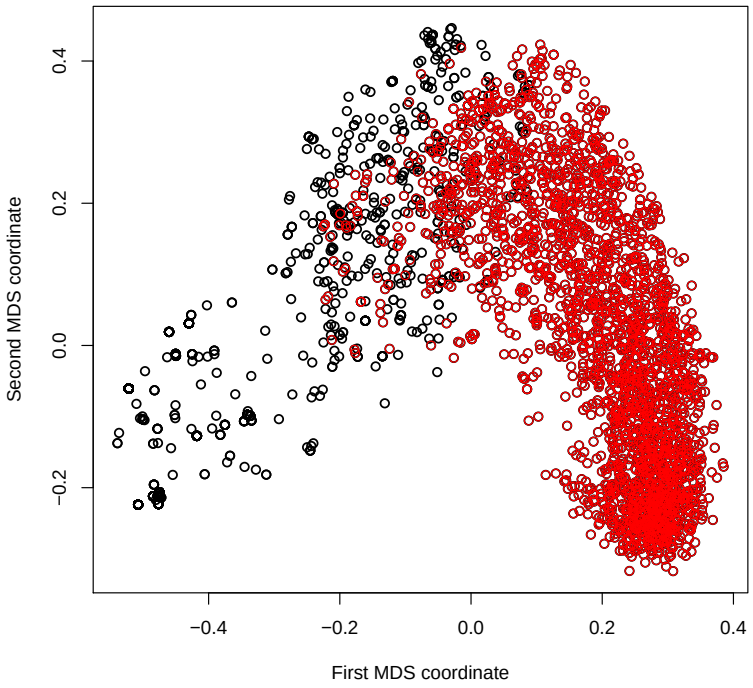


Figure 6.12: Clustering results for the training data set.

lection of the model parameters on the most recent training data (the tuning set chosen in the time-oriented validation scenario), enables our history distribution matching approach to tackle the growing gap between the abundances of the successful and failing therapies over time.

Table 6.3: Details on the data sets generated in the time-oriented validation scenario.

Data set	training	tuning	test
Sample count	3596	1634	1307
Success rate	69%	79%	83%

Reference methods. In our computational experiments we compare the results of our history distribution matching approach, denoted as *transfer history validation scenario*, to those of four reference approaches: the *one-for-all validation scenario*, the *therapy-specific validation scenario* and the *history-similarity validation scenario*, all explained in the first part of this chapter, and the *history-clustering validation scenario*. The history-clustering method implements a modified version of the algorithm of the history aware distribution matching method which skips the distribution matching step. In other words, a separate model is trained for each cluster by using only the data from the respective cluster.

Performance measures. The performances of all considered methods are assessed by reporting their corresponding accuracies (ACC) and AUCs (Area Under the ROC Curve) obtained on the test set. The accuracy reflects the ability of the methods to make correct predictions, *i.e.* to discriminate between successful and failing HIV combination therapies. With the AUC we are able to assess the quality of the ranking based on the probability of therapy success. For this reason, we carry out the model selection based on both accuracy and AUC and then use accuracy or AUC, respectively, to assess the model performance. In order to compare the performance of two methods on a separate test set, the significance of the difference of two accuracies as well as their standard errors are calculated as described in Section 3.4 of Chapter 3. The standard errors of the AUC values and the significance of the difference of two AUCs used for the pairwise method comparison are estimated as described in Hanley and McNeil (1983).

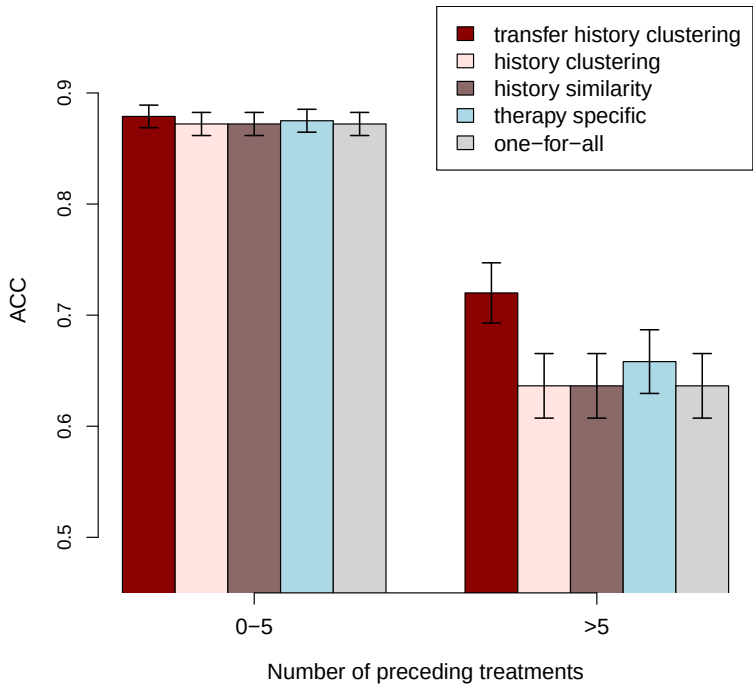


Figure 6.13: Accuracy results with their corresponding standard errors for the different models obtained on the test set in the time-oriented validation scenario. The test samples are grouped based on their corresponding number of known previous therapies.

6.3.5 Experimental Results

According to the results from the silhouette validation technique (Rousseeuw, 1987) displayed in Figure 6.11, the first clustering step of Algorithm 3 divides our training data into two clusters – one comprises the samples with longer therapy sequences (with average treatment history length of 5.507 therapies), and the other one with shorter therapy

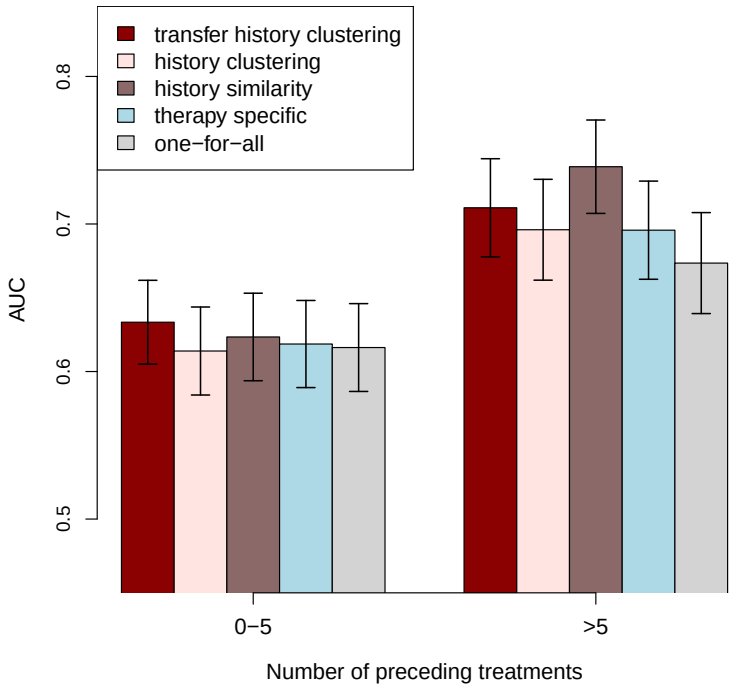


Figure 6.14: AUC results with their corresponding standard errors for the different models obtained on the test set in the time-oriented validation scenario. The test samples are grouped based on their corresponding number of known previous therapies.

sequences (with average treatment history length of 0.308 therapies). Thus, the transfer history distribution matching method trains two models, one for each cluster. The clustering results are depicted in Figure 6.12. In what follows, we first present the results of the validation experiments of the time-oriented validation scenario stratified for the length of treatment history, followed by the results stratified for the abundance of the different therapies. In both cases we report the test accuracies and AUCs for all considered methods.

The computational results for the transfer history method and the four reference methods stratified for the length of therapy history are summarized in Figures 6.13 and 6.14, where Figure 6.13 depicts the accuracies and Figures 6.14 depicts the AUCs. For samples with a small number of previously administered therapies ($\leq$ 5), *i.e.* with short treatment histories, all considered models have comparable accuracies. For test samples from patients with longer treatment histories ($>$ 5) the transfer history approach achieves significantly better (*p-values* $\leq$ 0.004) accuracy compared to the accuracies of all considered reference methods. According to the paired difference test described in Hanley and McNeil (1983), the history-similarity method achieves significantly better AUC than the transfer history approach for test samples with long treatment histories ($>$ 5) with estimated *p-value*= 0.021. The transfer history approach has significantly better AUC performance for test samples with longer ($>$ 5) treatment histories compared to the one-for-all (*p-value*= 0.043) and the history-clustering (*p-value*= 0.044) reference methods. It also has better AUC

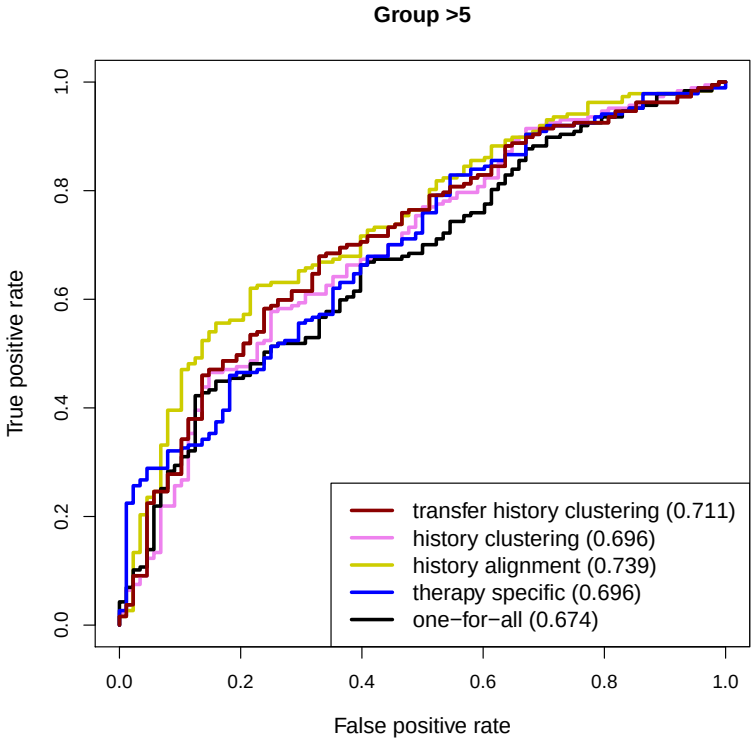


Figure 6.15: ROC curves depicting the performance of the history-similarity approach and the reference approaches for the group of test samples with more than five previously administered therapies (> 5). The AUC values for each method are given in the legend.

performance compared to the one of the therapy-specific model, yet this improvement is not significant ($p\text{-}value = 0.253$). Furthermore, the transfer history approach achieves better AUCs for test samples with less than five previously administered therapies compared to all reference methods. However, the improvements are only significant for the one-for-all method ($p\text{-}value = 0.007$) and the history-clustering method ($p\text{-}value = 0.080$). The corresponding *p-values* for the history-similarity and the therapy-specific methods are 0.189 and 0.178, respectively. Figure 6.15 depicts the ROC curves for the group of test samples with more than five previously administered therapies corresponding to the AUC results presented in Figure 6.14.

The experimental results stratified for the abundance of the therapies summarizing the test accuracies and AUCs for all considered methods are depicted in Figures 6.16 and 6.17, respectively. As can be observed from Figure 6.16, all considered methods have comparable accuracies for the test therapies with more than seven samples. The transfer history method achieves significantly better ($p\text{-}values \leq 0.0001$) accuracy compared to all reference methods for the test therapies with few ($0 - 7$) available training samples. Considering the AUC results displayed in Figure 6.17, the transfer history approach outperforms the one-for-all, the therapy-specific and the history-clustering models for the rare test therapies (with $0 - 7$ training samples) with estimated *p-values* of 0.05, 0.042 and 0.1, respectively. According to the paired difference test described in Hanley and McNeil (1983), the slightly

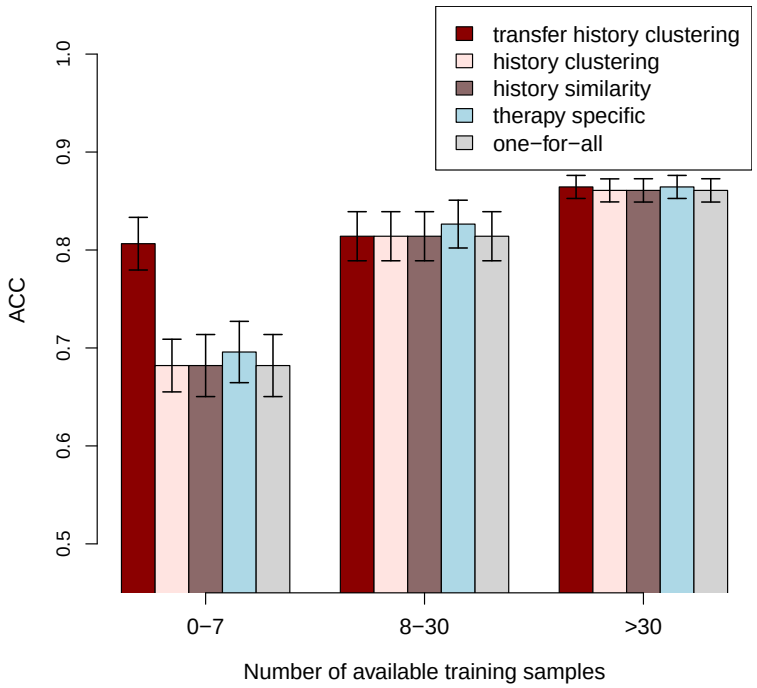


Figure 6.16: Accuracy results with their corresponding standard errors for the different models obtained on the test set in the time-oriented validation scenario. The test samples are grouped based on the number of available training examples for their corresponding therapy combinations.

better AUC value of the history-similarity method for the rare test therapies compared to the one of the transfer history model is not significant ($p\text{-}value = 0.170$). The one-for-all, the therapy-specific and the history similarity models have slightly better AUC performance than the transfer history and the history-clustering approaches for test therapies with $8-30$ available training samples. However, the improvements are not significant with *p-values* larger than 0.141 for all pairwise comparisons. Considering the test therapies with more than 30 training samples the transfer history approach significantly outperforms the one-for-all and the history-clustering reference approaches with estimated *p-values* of 0.037 and 0.064, respectively. It also has slightly better AUC performance compared to those of the history-similarity and the therapy-specific models, however these improvements are not significant (with $p\text{-}values \geq 0.136$). Figure 6.18 depicts the relevant ROC curves for all considered methods for the rare test therapies (with $0-7$ available training samples) corresponding to the AUC results from Figure 6.17.

To summarize, for test samples stemming from patients with long treatment history and for test samples associated with rare therapies the transfer history approach achieves significantly better accuracy than all considered reference approaches. Furthermore, its AUC performance for this group of test samples is significantly better than those of the one-for-all and the history-clustering methods. The history-similarity approach has better AUC performance than the one of the transfer history method for the test therapies with long

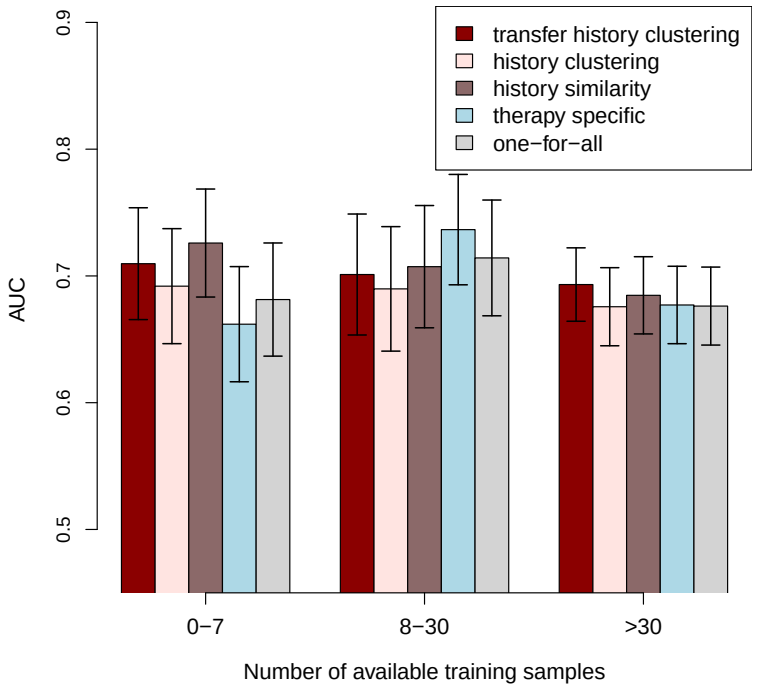


Figure 6.17: AUC results with their corresponding standard errors for the different models obtained on the test set in the time-oriented validation scenario. The test samples are grouped based on the number of available training examples for their corresponding therapy combinations.

treatment histories and the rare test therapies, however the improvement is only significant for the test therapies with long treatment histories. For the remaining test samples both the accuracy and the AUC performance of the transfer history method is comparable to the corresponding performance of the reference methods.

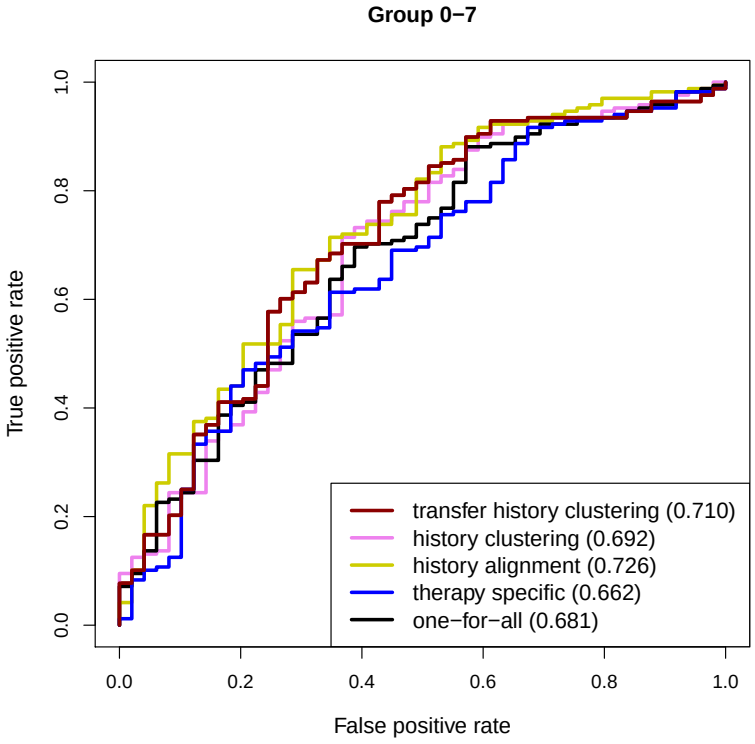


Figure 6.18: ROC curves displaying the performance of the different methods on the rare therapies (with $0-7$ training samples) of the test set in the time-oriented validation scenario. The AUC values for each method are given in the legend.

6.4 Conclusions

This chapter presents two approaches that deal with the bias introduced from the treatment history when predicting the effectiveness of HIV combination therapies: the history-similarity approach, which contributes information on the latent viral population and deals with the various treatment backgrounds associated with the samples making up the clinical data sets, the uneven sample representation with respect to the level of therapy experience, the sparse therapy representation; and the history distribution matching method, which addresses the different treatment backgrounds of the samples, the uneven representation of the different levels of therapy experience, the incomplete treatment history information, the uneven therapy representation and the increasing imbalance between the abundances of the effective and the ineffective therapies over time. While the history-similarity approach uses detailed information on the treatment history, available through the alignment similarity kernel, to provide patient(sample)-specific predictions, the history distribution matching method uses the alignment kernel only in the first clustering step. However, by doing so this method trades some detail of the treatment history information for the ability to also explicitly account for the problems of missing treatment history information and the increasing imbalance between the abundances of effective and ineffective therapies in the clinical data sets over time. It should also be pointed out that the model selection procedure for the history distribution matching method is more efficient compared to the history similarity approach because it does not train a separate model for each therapy sequence and it uses a parameter-free version of the alignment similarity kernel. Furthermore, while computing the prediction for a given target sample using the history-similarity requires several minutes, the same computation takes only a couple of seconds for the history distribution matching approach. This is the case because when using the latter approach the target therapy sequence needs to be aligned only to the cluster centers and not to the therapy sequences of all training samples.

The computational experiments show that both history methods significantly outperform the most commonly used approach – the one-for-all reference method, that does not account for the issues mentioned above. Both history models have their prime advantage for samples stemming from patients with long treatment histories and for samples associated with rare therapies. When compared to each other, the history similarity method has significantly better AUC performance (ranking) for the test samples associated with long treatment histories, while the transfer history method achieves significantly better accuracy for both the test samples associated with long treatment histories and the rare test samples. The reasons that makes these two groups particulary interesting are given as follows. Since there are specific guidelines for both treating therapy-naïve patients with first-line therapy and administering the first couple of follow-up therapies, which normally are successfully applied, assistance is mainly necessary for therapy-experienced patients. Moreover, because of the lack of data and practical experience for the rare HIV combination therapies, predicting their effectiveness is more challenging compared to estimating the effectiveness of the frequent therapies. It is worth noting that the performance of both history approaches for samples stemming from patients with shorter treatment history, or samples associated with abundant therapies is at least as good as the one of the considered reference methods.

7 Conclusions

This chapter presents the complete puzzle that assembles the different pieces presented in all previous chapters. In other words, it provides an overview of the statistical methods for predicting effictiveness of HIV therapies presented throughout this thesis together with the insights they provide for the considered application. The chapter closes by sketching potential extensions of the presented work first in the area of HIV research and then also to other challenging applications.

Summarizing Remarks

The main objective of the work presented in this thesis is to develop statistical learning approaches which are able to enhance the clinical management of HIV infections. For this purpose we take up the challenge of devising bias-aware statistical learning methods for HIV therapy screening — predicting the effectiveness of HIV combination therapies. The methods we develop are able to deal with various kinds of bias relevant for the available HIV clinical data sets, such as:

- the evolving trends of treating HIV patients,
- the sparse, unbalanced therapy representation,
- the different treatment backgrounds of the clinical samples,
- the uneven sample representation of the various levels of treatment experience.

If these biases are not accounted for they propagate to the derived statistical models and influence their predictions. Throughout the thesis we introduce five novel approaches for HIV therapy screening which depending on their objective tackle the aforementioned issues. The first three approaches – the distribution matching approach presented in Chapter 3, the therapy-similarity approach described in Chapter 4 and the multi-task hierarchical Bayes approach detailed in Chapter 5 – aim for balancing the sparse and uneven therapy representation in the HIV clinical data sets by using different routes of sharing common knowledge among related therapies. This results in a good prediction performance for every drug combination independent of its abundance in the clinical data set. The remaining two approaches – the history-similarity approach and the history distribution matching approach presented in Chapter 6 – address the bias originating from the different treatment backgrounds of the samples making up the clinical data sets. For this purpose, both methods predict the response of an HIV combination therapy by taking not only the most recent (target) therapy but also available information from preceding therapies into account. In this way they achieve good predictions for rare therapies and for advanced patients in mid to late stages of HIV treatment.

All these methods use the time-oriented evaluation scenario, where models are trained on data from the less recent past while there performance is evaluated on data from the more recent past. This is the approach we have chosen to address the evolving treatment trends in HIV clinical practice and thus offer a realistic model assessment.

Distribution matching approach. This first approach considers each therapy as a separate task in a multi-task learning setting. It can handle arbitrarily different data distributions for the different therapies without making assumptions about the data generation process or the relation between therapies. Briefly, it first computes therapy-specific resampling weights which match the joint input-output distribution of the available data from all therapies to the joint input-output distribution of a given target therapy. Then, it uses the weighted training data governed by the desired target distribution to train a separate predictive model for each target therapy. In this way, the distribution matching approach exploits the entire corpus of training data for all therapies to compensate for the sparse therapy representation in the clinical data sets. In the computational experiments we demonstrate that this method significantly improves the overall prediction accuracy compared to the relevant reference methods. Note however that while being statistically sound, the method is also quite compute-intensive, as it involves a multi-class logistic regression with as many classes as there are therapies (usually several hundred) – for some values of its respective tuning parameter fitting one such multi-class logistic regression model can take up to five days.

Therapy-similarity approach. This method focuses on predicting effectiveness of HIV therapies from genotypic information. Like the distribution matching approach, in order to address the sparse therapy representation it also trains separate models for each distinct therapy combination by using not only the samples from the target therapy but also the available samples from related therapies with appropriate sample weights. However, instead of the time-consuming distribution matching step we use two different application-specific similarity kernels – the drugs kernel and the groups additivity kernel – that quantify the pairwise relatedness of the therapies. These kernels are computed in a few seconds in a preprocessing step. This, together with the use of an efficient optimization method that takes advantage of the sparseness of our input data, ensures very fast model fitting and model selection, although a separate model is trained for each combination therapy. The therapy-similarity model is also able to directly integrate parameters of phenotypic models that give information on the in vitro effectiveness of each drug as prior knowledge through a Gaussian prior. According to the accuracy performance from the computational experiments, this approach has its prime advantage for rare therapies. While integrating phenotypic prior knowledge did not improve the accuracies of the therapy-similarity models, it significantly improved the AUC performance for the rare test therapies. Note that for the abundant therapies, the model has a performance comparable to the considered reference methods. Last but not least, the therapy-similarity approach provides increased interpretability of the fitted models. On the one hand, the scores of the mutations contributing to therapy effectiveness are derived in a therapy-specific manner and can, therefore, be considered more informative than for a general model. On the other hand, the therapy similarity kernel affords information on which similar therapies were

most informative for the prediction.

Multi-task hierarchical Bayes approach. The third approach casts the problem of predicting virological response to HIV combination therapies in a hierarchical Bayes framework. Here we devise a novel approach that considers the individual drugs comprising each therapy combination as separate tasks in a multi-task model that learns their additive effects on the therapy outcome from the available clinical data. In this manner, our Bayes approach uses the abundance of samples involving each individual drug to circumvent the problem of data scarcity pertaining to some target therapies. Moreover, it allows for interactions among the input features of the different drugs by using an extended input feature space where each drug has a separate set of dimensions. The advantage of the Bayes approach compared to the two previous therapy-specific methods is that it achieves significantly better AUC performance (better ranking) for therapies with very few training samples and is at least as powerful for abundant therapies. Furthermore, since it fits a single model for all therapies the Bayes method has the additional advantage of being much more time-efficient compared to the therapy-specific approaches.

History-aware modeling. The remaining two approaches account not only for the sparse, uneven therapy representation but also for the bias originating from the differing treatment backgrounds of the samples making up the clinical data. By doing so they provide good quality predictions for treatment-experienced patients in mid to late stages of HIV treatment and for rare therapies. Furthermore, they also maintain good quality predictions for the remaining samples.

- **History-similarity approach** – This approach trains a separate model for each sample of interest by using all available training samples, each with a specific weight, that reflects the similarity of their corresponding therapy sequences to the therapy sequence of the target sample. For this purpose, it introduces a novel quantitative notion of pairwise similarity of therapy sequences (termed alignment similarity kernel) that adapts techniques from sequence alignment to the problem of aligning sequences of therapies. This similarity measure also incorporates information on the similarity of the corresponding genetic fingerprints in the latent virus population of the compared therapy sequences. In this way the alignment kernel captures information on the latent virus population, all available therapies given to a patient and the order in which they were administered. More importantly, this kernel enables the history-similarity approach to deal with several biases relevant for the clinical data sets: the sparsity of the various therapy sequences, the uneven sample representation with respect to the level of therapy experience and the uneven therapy representation. The history-similarity approach is also patient specific since it trains sample-specific models that use very detailed treatment history information. In this manner it makes one step further in the direction of personalized HIV treatment. Additionally, such models are also more interpretable compared to one-for-all approaches that train a single model for all samples which is very important property in medical applications.

- **History distribution matching approach** – It first clusters the training data based on their corresponding treatment histories and then trains a separate model

for each cluster by using relevant information from the remaining clusters. The relevance of each sample for a corresponding target cluster is reflected by sample-specific weights obtained with a multi-task distribution matching procedure. These weights match the distribution of the entire training set to the distribution of the target cluster. While the history-similarity approach uses detailed information on the treatment history, available through the alignment similarity kernel, to provide patient-specific predictions, the history distribution matching method uses the alignment kernel only in the first clustering step. By doing so, this method trades some detail of the treatment history information for the ability to also account for the problems coming with the incomplete treatment history information, the increasing imbalance between the effective and ineffective therapies over time and to provide time-efficient prediction (in the range of a couple of seconds).

Computational experiments show that both history-aware methods significantly outperform the most commonly used approach that fits a single model for all therapies by encoding therapy information in the input feature space and does not account for the issues mentioned above. Both history models have their prime advantage for samples stemming from patients with long treatment histories and for samples associated with rare therapies. When compared to each other, the history-similarity method has significantly better AUC performance for the test samples associated with long treatment histories, while the transfer history method achieves significantly better accuracy for both the test samples associated with long treatment histories and the rare test samples.

To summarize, this thesis is devoted to the challenge of developing predictive models for HIV therapy screening while addressing the different kinds of data bias relevant for the problem at hand. More specifically, we begin in Chapters 3 through 5 by developing statistical methods that consider the bias introduced by the sparse, uneven representation of the different therapies in the HIV clinical data sets. Then, in Chapter 6 we extended this initial idea further by introducing the history-aware methods that take not only the bias introduced by the most recent (target) therapy but also the bias which originates from the different treatment backgrounds of the clinical samples. By doing so the history-aware models achieve better predictions for the samples associated with rare drug combinations and long treatment histories than the considered reference methods. There are a number of reasons that make these two groups especially interesting. For example, the rare HIV combination therapies account for most of the therapy variety in the clinical data sets. Moreover, because of lack of data and practical experience with administering such therapies, predicting their effectiveness is more challenging than estimating the effectiveness of frequent therapies. The search for an effective treatment is particulary challenging for patients in the mid to late stages of antiretroviral therapy when the number of therapy options is reduced and effective therapies are increasingly hard to find because of the accumulated drug resistance mutations from all previous therapies. Further, since there are specific guidelines for both treating therapy-naïve patients with first-line therapy and administering the first couple of follow-up therapies, which normally are successfully applied, assistance is mainly necessary for therapy-experienced patients. Note that the history-aware models do not require completely recorded treatment history of the patients and can be utilized to enhance the clinical management of HIV patients. While the history

distribution matching method provides real-time prediction for a given sample of interest, the history-similarity method computes this prediction in several minutes. However, the history-similarity method trains sample-specific models which makes the predictions more patient-specific and more interpretable. Therefore, we expect that for the treating physicians the history-similarity model will be the preferred alternative for the use in clinical practice.

Outlook

HIV-1 subtype B is the best-studied variant of HIV and the main target of the developed antiretroviral compounds due to its prevalence in the industrialized countries. Since the majority of the available HIV clinical data stem from HIV-1 subtype B, the prediction methods devised in this thesis focused on this subtype. In recent years, however, non-B HIV-1 subtypes have been gaining attention owing to their dominance in the countries in Africa and Asia with high HIV prevalence. Once data collection efforts in these regions produce a reasonable amount of data some of the methodology developed in this thesis can be adapted to the problem of non-B HIV therapy screening. For example, each HIV-1 subtype can be considered as a separate task in a multi-task learning setting. Like this, the relation among the different HIV-1 variants can be exploited to transfer the available knowledge from the subtype B, for which a large amount of clinical data is available, to non-B subtypes.

The need for devising bias-aware prediction models is very general and extends far beyond the boundaries of the HIV application presented in this thesis. A similar line of research can be conducted for many other statistical learning applications after recognizing the main sources of bias in their respective data sets.

One direction would be to apply the developed approaches to other biomedical applications. One example that resembles the HIV application considered in this thesis is treatment optimization for Hepatitis B and Hepatitis C. Note however that in this case the employment of our HIV tailored methods requires further research that will take the biomedical background and the clinical expertise of the new application into account to modify and extend these methods accordingly.

Another example is the challenge of cancer diagnosis where often the available data sets for a specific cancer type are very limited and originate from different labs. One can use the available data from all labs by considering each of the labs as a separate task in a multi-task learning framework. In this way one can obtain good predictions for data produced in a specific target lab by utilizing the joint information available from all labs in a proper way.

List of Figures

List of Tables

Bibliography

T. Alcorn and H. Faruki. HIV resistance testing: methods, utility, and limitations. *Molecular Diagnostics*, 5(3):159–168, 2000.

A. Altmann, N. Beerenwinkel, T. Sing, I. Savenkov, M. Däumer, R. Kaiser, S. Rhee, W.J. Fessel, W.R. Shafer, and T. Lengauer. Improved prediction of response to antiretroviral combination therapy using the genetic barrier to drug resistance. *Antiviral Therapy*, 12: 169–178, 2007.

A. Altmann, M. Däumer, N. Beerenwinkel, E. Peres, Y. Schülter, A. Büch, S. Rhee, A. Sönnerborg, W.J. Fessel, M. Shafer, W.R. Zazzi, R. Kaiser, and T. Lengauer. Predicting response to combination antiretroviral therapy: retrospective validation of geno2phenotheo on a large clinical database. *Journal of Infectious Diseases*, 199:999–1006, 2009a.

A. Altmann, T. Sing, H. Vermeiren, B. Winters, E. Van Craenenbroeck, K. Van der Borght, S.Y. Rhee, R.W. Shafer, E. Schülter, R. Kaiser, Y. Peres, A. Sönnerborg, W.J. Fessel, F. Incardona, M. Zazzi, L. Bacheler, H. Van Vlijmen, and T. Lengauer. Advantages of predicted phenotypes and statistical learning models in inferring virological response to antiretroviral therapy from hiv genotype. *Antiviral Therapy*, 14:273–283, 2009b.

J.M. Azevedo-Pereira, Q. Santos-Costa, and J. Moniz-Pereira. HIV-2 infection and chemokine receptors usage - clues to reduced virulence of HIV-2. *Current HIV Research*, 3(1):3–16, 2005.

B. Bakker and T. Heskes. Task clustering and gating for Bayesian multitask learning. *The Journal of Machine Learning Research*, 4:83–99, 2003.

F. Barre-Sinoussi, J.C. Chermann, F. Rey, M.T. Nugeyre, S. Chamaret, J. Gruest, C. Dauguet, C. Axler-Blin, F. Vezinet-Brun, C. Rouzioux, W. Rozenbaum, and L. Montagnier. Isolation of a T-Lymphotropic Retrovirus from a Patient at Risk for Acquired Immune Deficiency Syndrome (AIDS). *Science*, 220(4599):868–871, 1983.

N. Beerenwinkel. *Computational Analysis of HIV Drug Resistance Data*. PhD thesis, Universität des Saarlandes, 2004.

N. Beerenwinkel, M. Däumer, M. Oette, K. Korn, D. Hoffmann, R. Kaiser, T. Lengauer, J. Selbig, and H. Walter. Geno2pheno: Estimating phenotypic drug resistance from HIV-1 genotypes. *Nucleic Acids Res*, 31:3850–3855, 2003a.

N. Beerenwinkel, T. Lenaguer, M. Däumer, R. Kaiser, H. Walter, K. Korn, D. Hoffmann, and J. Selbig. Methods for optimizing antiviral combination therapies. *Bioinformatics*, 19:i16–i25, 2003b.

S. Benson and J.J. Moré. A limited memory variable metric method for bound constrained minimization. Technical report, Preprint MCS-P909-0901, Mathematics and Computer Science Division, Argonne National Laboratory, Argonne, Illinois, 2001.

E.A. Berger, P.M. Murphy, and J.M. Farber. Chemokine receptors as HIV-1 coreceptors: roles in viral entry, tropism, and disease. *Annual Review of Immunology*, 17:657–700, 1999.

A. Berlinet and C. Thomas-Agnan. *Reproducing kernel Hilbert spaces in probability and statistics*. Kluwer Academic Publishers, 2004.

S. Bickel. *Learning under Differing Training and Test Distributions*. PhD thesis, Universität Potsdam, 2009.

S. Bickel and T. Scheffer. Dirichlet-enhanced spam filtering based on biased samples. In *Advances in Neural Information Processing Systems*, 2007.

S. Bickel, M. Brückner, and T. Scheffer. Discriminative learning for differing training and test distributions. In *Proceedings of the International Conference on Machine Learning*, 2007.

S. Bickel, J. Bogojeska, T. Lengauer, and T. Scheffer. Multi-task learning for hiv therapy screening. In *Proceedings of the International Conference on Machine Learning*, 2008.

J.N. Blankson. Control of HIV-1 replication in elite suppressors. *Discovery Medicine*, 9 (46):261–266, 2010.

M. Boffito, E. Acosta, D. Burger, C.V. Fletcher, C. Flexner, R. Garaffo, G. Gatti, M. Kurowski, C.F. Perno, G. Peytavin, M. Regazzi, and D. Back. Therapeutic drug monitoring and drug-drug interactions involving antiretroviral drugs. *Antiviral Therapy*, 10(4):469–477, 2005.

J. Bogojeska. History distribution matching method for predicting effectiveness of HIV combination therapies. In *Advances in Neural Information Processing Systems (NIPS)*, 2011.

J. Bogojeska and T. Lengauer. Hierarchical Bayes model for predicting effectiveness of HIV combination therapies. *submitted*, 2011.

J. Bogojeska, D. Stöckel, M. Zazzi, R. Kaiser, F. Incardona, M. Rosen-Zvi, and T. Lengauer. History-alignment models for bias-aware prediction of virological response to HIV combination therapy.

J. Bogojeska, S. Bickel, A. Altmann, and T. Lengauer. Dealing with sparse data in predicting outcomes of hiv combination therapies. *Bioinformatics*, 26:2085–2092, 2010.

M.C. Boily, R.F. Baggaley, L. Wang, B. Masse, R.G. White, R.J. Hayes, and M. Alary. Heterosexual risk of HIV-1 infection per sexual act: systematic review and meta-analysis of observational studies. *The Lancet Infectious Diseases*, 9(2):118Ű–129, 2009.

E.V. Bonilla, F. Agakov, and C. Williams. Kernel multi-task learning using task-specific features. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, 2007.

B.E. Boser, I. Guyon, and V. Vapnik. A training algorithm for optimal margin classfiers. In *In Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, 1992.

G. Bratt, A. Karlsson, A.C. Leandersson, J. Albert, B. Wahren, and E. Sandström. Treatment history and baseline viral load, but not viral tropism or CCR-5 genotype, influence prolonged antiviral efficacy of highly active antiretroviral treatment. *AIDS*, 12(16):2193–2202, 1998.

CDC. Centers for Disease Control(CDC). A cluster of Kaposi's sarcoma and Pneumocystis carinii pneumonia among homosexual male residents of Los Angeles and Orange Counties, California. *MMWR Morb. Mortal. Wkly. Rep.*, 31(23):305–307, 1982.

R.E. Chaisson, J. Gallant, J. Keruly, and R.D. Moore. Impact of opportunistic disease on survival in patients with HIV infection. *AIDS*, 12(1):29–33, 1998.

T. Cihlar and A.S. Ray. Nucleoside and nucleotide HIV reverse transcriptase inhibitors: 25 years after zidovudine. *Antiviral Research*, 85(1):39–58, 2010.

F. Clavel and A.J. Hance. HIV drug resistance. *The New England Journal of Medicine*, 350(10):1023–1035, 2004.

H. Coovadia. Antiretroviral Agents Ů How Best to Protect Infants from HIV and Save Their Mothers from AIDS. *The New England Journal of Medicine*, 351:289–292, 2008.

C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20:273–297, 1995.

N.F. Crum, R.H. Riffenburgh, S. Wegner, B.K. Agan, S.A. Tasker, K.M. Spooner, A.W. Armstrong, S. Fraser, M.R. Wallace, and T.A.C. Consortium. Comparisons of causes of death and mortality rates among HIV-infected persons: analysis of the pre-, early, and late HAART (highly active antiretroviral therapy) eras. *Journal of Acquired Immune Deficiency Syndromes*, 41(2):194–200, 2006.

B.R. Cullen. Regulation of HIV-1 gene expression. *The FASEB Journal*, 5(10):2361–2368, 1991.

A.G. Dalgleish, P.C. Beverley, P.R. Clapham, D.H. Crawford, M.F. Greaves, and R.A. Weiss. The CD4 (T4) antigen is an essential component of the receptor for the AIDS retrovirus. *Nature*, 312(5996):763–767, 1984.

J.N. Darroch and D. Ratcliff. Generalized iterative scaling for log-linear models. *The Annals of Mathematical Statistics*, 43(5):1470–1480, 1972.

T.I. de Silva, M. Cotten, and S.L. Rowland-Jones. HIV-2: the forgotten AIDS virus. *Trends in Microbiology*, 16(12):588–595, 2008.

E.R. DeLong, D.M. DeLong, and D.L. Clarke-Pearson. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*, 44:837–845, 1988.

C.L. Dong and J. Nocedal. On the limited memory BFGS method for large scale optimization. *Mathematical Programming*, 45(1):503–528, 1989.

M. Dudik, R. Schapire, and S. Phillips. Correcting sample selection bias in maximum entropy density estimation. In *Advances in Neural Information Processing Systems*, 2005.

J. Durant, P. Clevenbergh, P. Halfon, P. Delgiudice, S. Porsin, P. Simonet, N. Montagne, C.A. Boucher, J.M. Schapiro, and P. Dellamonica. HIV resistance testing: methods, utility, and limitations. *Lancet*, 353(9171):2195–2199, 1999.

J.N. Dybowski, D. Heider, and D. Hoffmann. Prediction of co-receptor usage of HIV-1 from genotype. *PLoS Computational Biology*, 6(4):e1000743, 2010.

J.A. Esté and A. Telenti. HIV entry inhibitors. *Lancet*, 370(9581):81–88, 2007.

T. Evgeniou and M. Pontil. Regularized multi–task learning. *Proceedings of the International Conference on Knowledge Discovery and Data Mining*, pages 109–117, 2004.

T. Evgeniou, M. Pontil, and T. Poggio. Regularization networks and support vector machines. *Advances in Computational Mathematics*, 13:1–50, 2000.

J. Fellay, K.V. Shianna, D. Ge, S. Colombo, B. Ledergerber, M. Weale, K. Zhang, C. Gumbs, A. Castagna, A. Cossarizza, A. Cozzi-Lepri, A.D. Luca, P. Easterbrook, P. Francioli, S. Mallal, J. Martinez-Picado, J.M. Miro, N. Obel, J.P. Smith, J. Wyniger, P. Descombes, S.E. Antonarakis, N.L. Letvin, A.J. McMichael, B.F. Haynes, A. Telenti, and D.B. Goldstein. High levels of HIV-1 in plasma during all stages of infection determined by competitive PCR. *Science*, 317(5840):944–947, 2007.

B.N. Fields, D.M. Knipe, and P.M. Howley. *Fields Virology*. Lippincott Williams and Wilkins, 2007.

E.O. Freed. HIV-1 replication. *Somatic Cell and Molecular Genetics*, 26:13–33, 2001.

E.O. Freed. HIV-1 and the host cell: an intimate association. *Trends in Microbiology*, 12 (4):170–177, 2004.

R.C. Gallo, P.S. Sarin, E.P. Gelmann, M. Robert-Guroff, E. Richardson, V.S. Kalyanaraman, D. Mann, G.D. Sidhu, R.E. Stahl, S. Zolla-Pazner, J. Leibowitch, and M. Popovic. Isolation of human T-cell leukemia virus in acquired immune deficiency syndrome (AIDS). *Science*, 220(4599):865–867, 1983.

F. Gao, Y. Chen, D.N. Levy, J.A. Conway, T.B. Kepler, and H. Hui. Resistance to HIV-1 infection in caucasian individuals bearing mutant alleles of the CCR-5 chemokine receptor gene. *Journal of Virology*, 78(5):2426–2433, 2004.

A. Gelman, J.B. Carlin, H.S. Stern, and D.B. Rubin. *Bayesian Data Analysis*. Chapman & Hall/CRC, 2004.

T.R. Glass, S.D. Geest, B. Hirschel, M. Battegay, H. Furrer, M. Covassini, P.L. Vernazza, E. Bernasconi, M. Rickenboch, R. Weber, H.C. Bucher, and S.H.C. Study. Self-reported non-adherence to antiretroviral therapy repeatedly assessed by two questions predicts treatment failure in virologically suppressed patients. *Antiviral Therapy*, 13(1):77–85, 2008.

J. Goodman. Sequental conditional generalized iterative scaling. *ACL*, pages 9–16, 2002.

S. Grabar, H. Selinger-Leneman, S. Abgrall, G. Pialoux, L. Weiss, and D. Costagliola. Prevalence and comparative characteristics of long-term nonprogressors and HIV controller patients in the French Hospital Database on HIV. *AIDS*, 23(9):1163–1169, 2009.

J. Hanley and B. McNeil. A Method of comparing the Areas under Receiver Operating Characteristic Curves Derived from the Same Cases. *Radiology*, 148:839–843, 1983.

T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer, 2009.

J. Heckman. Sample selection bias as a specification error. *Econometrica*, 47:153–161, 1979.

J.L. Heeney, A.G. Dalgleish, and R.A. Weiss. Origins of HIV and the Evolution of Resistance to AIDS. *Science*, 313(5786):462–466, 2006.

J. Hemelaar, E. Gouws, P.D. Ghys, and S. Osmanov. Global and regional distribution of HIV-1 genetic subtypes and recombinants in 2004. *AIDS*, 20(16):13–23, 2004.

J. Huang, A. Smola, A. Gretton, K. Borgwardt, and B. Schölkopf. Correcting sample selection bias by unlabeled data. In *Advances in Neural Information Processing Systems*, 2007.

R. Jin, R. Yan, J. Zhang, and A.G. Hauptmann. A faster iterative scaling algorithm for conditional exponential model. In *Proceedings of the Twentieth International Conference on Machine Learning (ICML)*, 2003.

V.A. Johnson, F. Brun-Vezinet, B. Clotet, H.F. Günthrad, D.R. Kuritzkes, D. Pillay, J.M. Schapiro, A. Telenti, and D.D. Richman. Update of the drug resistance mutations in HIV-1: 2007. *Topics in HIV Medicine*, 15:119–125, 2007.

V.A. Johnson, F. Brun-Vezinet, B. Clotet, H.F. Günthrad, D.R. Kuritzkes, D. Pillay, J.M. Schapiro, and D.D. Richman. Update of the drug resistance mutations in HIV-1: December 2008. *Topics in HIV Medicine*, 16:138–145, 2008.

V.A. Johnson, F. Brun-Vezinet, B. Clotet, H.F. Günthrad, D.R. Kuritzkes, D. Pillay, J.M. Schapiro, and D.D. Richman. Update of the Drug Resistance Mutations in HIV-1:December 2010. *Topics in HIV Medicine*, 18(5):156–163, 2010.

L. Kaufman and P.J. Rousseeuw. *Finding Groups in Data. An introduction to cluster analysis*. John Wiley and Sons, Inc., 1990.

B.F. Keele, J.H. Jones, K.A. Terio, J.D. Estes, R.S. Rudicell, M.L. Wilson, Y. Li, G.H. Learn, T.M. Beasley, J. Schumacher-Stankey, E. Wroblewski, A. Mosser, J. Raphael, S. Kamenya, E.V. Lonsdorf, D.A. Travis, T. Mlengeya, M.J. Kinsel, J.G. Else, G. Silvestri, J. Goodall, P.M. Sharp, G.M. Shaw, A.E. Pusey, and B.H. Hahn. Increased mortality and AIDS-like immunopathology in wild chimpanzees infected with SIVcpz. *Nature*, 460(7254):515–519, 2009.

G.S. Kimeldorf and G. Wahba. A correspondence between Bayesian estimation on stochastic processes and smoothing by splines. *Annals of Mathematical Statistics*, 41:495–502, 1970.

P. Komarek and Moore A.W. A limited memory variable metric method for bound constrained minimization. Technical report, Technical Report TR-05-27, Robotics Institute, Carnegie Mellon University, 2005.

B. Korber, M. Muldoon, J. Theiler, F. Gao, R. Gupta, A. Lapedes, B.H. Hahn, S. Wolinsky, and T. Bhattacharya. Timing the ancestor of the HIV-1 pandemic strains. *Science*, 288 (5472):1789–1796, 2000.

B. Larder, D. Wang, A. Revell, J. Montaner, R. Harrigan, F. De Wolf, J. Lange, S. Wegner, L. Ruiz, M.J. Pérez-Elías, S. Emery, J. Gatell, A. DŚArminio Monforte, C. Torti, M. Zazzi, and C. Lane. The development of artificial neural networks to predict virological response to combination HIV therapy. *Antiviral Therapy*, 12:15–24, 2007.

B.A. Larder and S.D. Kemp. Multiple mutations in HIV-1 reverse transcriptase confer high-level resistance to zidovudine (AZT). *Science*, 246(4934):1155–1158, 1989.

B.A. Larder, G. Darby, and D.D. Richman. HIV with reduced sensitivity to zidovudine (AZT) isolated during prolonged therapy. *Science*, 243(4899):1731–1734, 1989.

R.H. Lathrop and M.J. Pazzani. Combinatorial optimization in rapidly mutating drug-resistant viruses. *Journal of Combinatorial Optimization*, 3:301–320, 1999.

T. Lengauer, O. Sander, S. Sierra, A. Thielen, and R. Kaiser. Bioinformatics prediction of HIV coreceptor usage. *Nature Biotechnology*, 25(12):1407–1410, 2007.

J.A. Lévy. HIV pathogenesis and long-term survival. *AIDS*, 7(11):1401–1410, 1993.

C.J. Lin and J. Moré. Newton's method for large-scale bound constrained problems. *SIAM Journal on Optimization*, 9:1100–1127, 1999.

C.J. Lin, R.C. Weng, and S.S. Keerthi. Trust region newton method for large-scale logistic regression. *Journal of Machine Learning Research*, 9:627–650, 2008.

T.F. Liu and R.W. Shafer. Web resources for HIV type 1 genotypic-resistance test interpretation. *Clinical Infectious Diseases*, 42, 2006.

R. Malouf. A comparisson of algorithms for maximum entropy parameter estimation. In *Proceedings of the 6th Conference on Natural Language Learning*, 2002.

D.J. McColl and X. Chen. Strand transfer inhibitors of HIV-1 integrase: bringing IN a new era of antiretroviral therapy. *Antiviral Research*, 85(1):101–118, 2010.

J.-L. Meynard, M. Vray, L. Morand-Joubert, E. Race, D. Descamps, G. Peytavin, S. Matheron, C. Lamotte, S. Guiramand, D. Costagliola, F. Brun-Vézinet, F. Clavel, P.-M. Girard, and N.T. Group. Phenotypic or genotypic resistance testing for choosing antiretroviral therapy after treatment failure: a randomized trial. *AIDS*, 16(5):727–736, 2002.

R.D. Moore and R.E. Chaisson. Natural History of Opportunistic Disease in an HIV-Infected Urban Clinical Cohort. *Annals of Internal Medicine*, 124:633–642, 1996.

S.G. Nash. On the limited memory BFGS method for large scale optimization. *Journal of Computational and Applied Mathematics*, 124(1–2):45–59, 2000.

S. Needleman and C. Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, 48(3):443–453, 1970.

J. Nocedal and S.J. Wright. *Numerical Optimization*. Springer, 2006.

N.T. Parkin and J.M. Schapiro. Antiretroviral drug resistance in non-subtype B HIV-1, HIV-2 and SIV. *Antiviral Therapy*, 9(1):3–12, 2004.

C.J. Petropoulos, N.T. Parkin, K.L. Limoli, Y.S. Lie, T. Wrin, W. Huang, H. Tian, D. Smith, G.A. Winslow, D.J. Capon, and J.M. Whitcomb. A novel phenotypic drug susceptibility assay for human immunodeficiency virus type 1. *Antimicrobial Agents Chemotherapy*, 44(4):920–928, 2000.

M. Piatak, M.S. Saag, L.C. Yang, S.J. Clark, J.C. Kappes, K.C. Luk, B.H. Hahn, G.M. Shaw, and J.D. Lifson. High levels of HIV-1 in plasma during all stages of infection determined by competitive PCR. *Science*, 159(5102):1749–1754, 1993.

S.D. Pietra, V.D. Pietra, and J. Lafferty. Inducing features of random fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(4):380–393, 1997.

J.C. Plantier, M. Leoz, J.E. Dickerson, F.D. Oliveira, F. Cordonnier, V. Lemée, F. Damond, D.L. Robertson, and F. Simon. A new human immunodeficiency virus derived from gorillas. *Nature Medicine*, 15(8):871–871, 2009.

V.W. Pollard and M.H. Malim. The HIV-1 Rev protein. *Annual Review of Microbiology*, 52:491–532, 1998.

M. Popovic, M.G. Sarngadharan, E. Read, and R.C. Gallo. Detection, isolation, and continuous production of cytopathic retroviruses (HTLV-III) from patients with AIDS and pre-AIDS. *Science*, 224(4648):497–500, 1984.

M. Prosperi, S. Di Giambenedetto, M.P. Trotta, A. Cingolani, L. Ruiz, J.D. Baxter, P. Clevenbergh, C F. Perno, R. Cauda, G. Ulivi, A. Antinori, and A. De Luca. A fuzzy relational system trained by genetic algorithms and HIV-1 resistance genotypes/virological

response data from prospective studies usefully predicts treatment outcomes. *Antiviral Therapy*, 9:U89–U89, 2004.

M. Prosperi, M. Zazzi, C.F. Perno, S. Di Giambenedetto, J. Baxter, L. Ruiz, P. Clevenbergh, G. Ulivi, A. Antinori, and A. De Luca. 'Common law' applied to treatment decisions for drug resistant HIV. *Antiviral Therapy*, 10:S62–S62, 2005.

M. Prosperi, A. Altmann, M. Rosen-Zvi, E. Aharoni, G. Borgulya, F. Bazso, A. Sönnerborg, E. Schülter, D. Struck, G. Ulivi, A. Vandamme, J. Vercauteren, and M. Zazzi. Investigation of expert rule bases, logistic regression, and non-linear machine learning techniques for predicting response to antiretroviral treatment. *Antiviral Therapy*, 14: 433–442, 2009.

M. Prosperi, M. Rosen-Zvi, A. Altmann, M. Zazzi, S. Giambenedetto, R. Kaiser, E. Schülter, D. Struck, P. Slot, A.D. Van De Vijver, A.M. Vandamme, and A. Sönnerborg. Antiretroviral therapy optimisation without genotype resistance testing: A perspective on treatment history based models. *PLoS ONE*, 65:5–10, 2010.

J.D. Reeves and R.W. Doms. Human Immunodeficiency Virus Type 2. *Journal of General Virology*, 83(6):1253–1265, 2002.

A.D. Revell, D. Wang, R. Harrigan, R.L. Hamers, A.M.J. Wensing, F. DeWolf, M. Nelson, A.M. Geretti, and B.A. Larder. Modelling response to hiv therapy without a genotype: an argument for viral load monitoring in resource-limited settings. *Journal of Antimicrobial Chemotherapy*, 65:605–607, 2010.

S.-Y. Rhee, M.J. Gonzales, R. Kantor, B.J. Betts, J. Ravela, and R.W. Shafer. Human immunodeficiency virus reverse transcriptase and protease sequence database. *Nucleic Acids Research*, 31(1):298–303, 2003.

S.-Y. Rhee, J. Taylor, G. Wadhera, A. Ben-Hur, D.L. Brutlag, and R.W. Shafer. Genotypic predictors of human immunodeficiency virus type 1 drug resistance. *Proceedings of the National Academy of Sciences USA*, 103(46):17355–17360, 2006.

G. Robbins, V. De Gruttola, R. Shafer, L. Smeaton, S. Snyder, C. Pettinelli, M. Dubé, M. Fischl, R. Pollard, R. Delapenha, L. Gedeon, C. Van Der Horst, R. Murphy, M. Becker, R. D'Aquila, S. Vella, T. Merigan, and M. Hirsch. Comparison of sequential three-drug regimens as initial therapy for hiv-1 infection. *New England Journal of Medicine*, 349(24):2293–2303, 2003.

D.L. Robertson, J.P. Anderson, J.A. Bradac, J.K. Carr, B. Foley, R.K. Funkhouser, F. Gao, Kalish M.L. Hahn, B.H., C. Kuiken, G.H. Learn, T. Leitner, F. McCutchan, S. Osmanov, M. Peeters, D. Pieniazek, M. Salminen, P.M. Sharp, S. Wolinsky, and B. Korber. HIV-1 nomenclature proposal. *Science*, 288(5463):55–56, 2000.

K. Roomp, N. Beerenwinkel, T. Sing, E. Schülter, J. Büch, S. Sierra-Aragon, M. Däumer, D. Hoffmann, R. Kaiser, T. Lengauer, and J. Selbig. Arevir: A secure platform for designing personalized antiretroviral therapies against hiv. In *Lecture Notes in Computer Science: Data Integration in the Life Sciences*, pages 185–194, 2006.

M. Rosen-Zvi, A. Altmann, M. Prosperi, E. Aharoni, H. Neuvirth, A. Sönnerborg, E. Schülter, D. Struck, Y. Peres, F. Incardona, R. Kaiser, M. Zazzi, and T. Lengauer. Selecting anti-hiv therapies based on a variety of genomic and clinical factors. *Proceedings of the ISMB*, 2008.

P. J. Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65, 1987.

H. Saigo, A. Altmann, J. Bogojeska, F. Müller, S. Nowozin, and T. Lengauer. Learning from past treatments and their outcome improves prediction of in vivo response to anti-hiv therapy. *Statistical Applications in Genetics and Computational Biology*, 10, 2010.

M.L. Santiago, F. Range, B.F. Keele, Y. Li, E. Bailes, F. Bibollet-Ruche, C. Fruteau, R. Noë, M. Peeters, J.F.Y. Brookfield, G.M. Shaw, P.M. Sharp, and B.H. Hahn. Simian immunodeficiency virus infection in free-ranging sooty mangabeys (Cercocebus atys atys) from the Taï Forest, Côte dŠIvoire: implications for the origin of epidemic human immunodeficiency virus type 2. *Journal of Virology*, 79(19):12515Ű–12527, 2005.

B. Schölkopf and A. Smola. *Learning with kernels*. MIT Press, 2002.

H. Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, 90:227–244, 2000.

V. Simon and D.D. Ho. HIV-1 dynamics in vivo: implications for therapy. *Nature Reviews in Microbiology*, 1(3):181Ű–190, 2003.

V. Simon, D.D. Ho, and Q.A. Karim. HIV/AIDS epidemiology, pathogenesis, prevention, and treatment. *Lancet*, 368(9534):489–504, 2006.

T. Sing, O. Sander, N. Beerenwinkel, and T. Lengauer. Rocr: visualizing classifier performance in r. *Bioinformatics*, 21:3940, 2005.

S.S. Spudich and B.M. Ances. Central nervous system complications of HIV infection. *Topics in Antiviral Medicine*, 19(2):48–57, 2011.

M. Sugiyama and K.-R. Müller. Model selection under covariate shift. In *Proceedings of the International Conference on Artificial Neural Networks*, 2005.

M. Sugiyama, S. Nakajima, H. Kashima, P. von Bunau, and M. Kawanabe. Direct importance estimation with model selection and its application to covariate shift adaptation. In *Advances in Neural Information Processing Systems*, 2008.

C. Sutton and A. McCallum. *An introduction to conditional random fields for relational learning*. MIT Press, 2006.

Y. Teh, M. Jordan, M. Beal, and D. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006.

A. Thielen, N. Sichtig, R. Kaiser, J. Lam, P.R. Harrigan, and T. Lengauer. Improved Prediction of HIV-1 Coreceptor Usage with Sequence Information from the Second Hypervariable Loop of gp120. *Journal of Infectious Diseases*, 202(9):1435–1443, 2010.

UNAIDS/WHO. Report on the global aids epidemic: 2010. 2010.

K. Van Laethem, A. De Luca, A. Antinori, A. Cingolani, C.F. Perno, and A.-M. Vandamme. A genotypic drug resistance interpretation algorithm that significantly predicts therapy response in HIV-1-infected patients. *Lancet*, 7(2):123–129, 2002.

V. Vapnik. *The Nature of Statistical Learning Theory*. Springer, 1996.

H. Vermeiren, E. Van Craenenbroeck, P. Alen, L. Bacheler, G. Picchio, P. Lecocq, and V.C.R.C. Team. Prediction of HIV-1 drug susceptibility phenotype from the viral genotype using linear regression modeling. *Journal of Virological Methods*, 145(1):47–55, 2007.

G. Wahba. *Spline models for observational data*. Society for Industrial Mathematics, 1990.

B.D. Walker and D.R. Burton. Toward an AIDS vaccine. *Science*, 320(5877):760–764, 2008.

H. Walter, B. Schmidt, K. Korn, A.M. Vandamme, T. Harrer, and K. Überla. Rapid, phenotypic HIV-1 drug sensitivity assay for protease and reverse transcriptase inhibitors. *Journal of Clinical Virology*, 13(1-2):71–80, 1999.

D. Wang, B.A. Larder, A. Revell, R. Harrigan, and J. Montaner. A neural network model using clinical cohort data accurately predicts virological response and identifies regimens with increased probability of success in treatment failures. *Antiviral Therapy*, 8:U99–U99, 2003.

A.M.J. Wensing, N.M. van Maarseveen, and M. Nijhuis. Fifteen years of HIV Protease Inhibitors: raising the barrier to resistance. *Antiviral Research*, 85(1):59–74, 2010.

P. Wu and T.G. Dietterich. Improving SVM accuracy by training on auxiliary data sources. *Proceedings of the International Conference on Machine Learning*, 2004.

Y. Xue, X. Liao, L. Carin, and B. Krishnapuram. Multi-task learning for classification with Dirichlet process priors. *Journal of Machine Learning Research*, 8:35–63, 2007.

K. Yu, V. Tresp, and A. Schwaighofer. Learning Gaussian processes from multiple tasks. *Proceedings of the International Conference on Machine Learning*, 2005.

B. Zadrozny. Learning and evaluating classifiers under sample selection bias. In *Proceedings of the International Conference on Machine Learning*, 2004.

M. Zazzi, M. Prosperi, I. Vicenti, S.D. Giambenedetto, A. Callegaro, B. Bruzzone, F. Baldanti, A. Gonnelli, E. Boeri, E. Paolini, S. Rusconi, A. Giacometti, F. Maggiolo, S. Menzo, A. De Luca, and A.C. Group. Rules-based HIV-1 genotypic resistance interpretation systems predict 8 week and 24 week virological antiretroviral treatment outcome and benefit from drug potency weighting. *Journal of Antimicrobial Chemotherapy*, 64 (3):616–624, 2009.

J. Zhu and T. Hastie. Kernel Logistic Regression and the Import Vector Machine. In *Advances in Neural Information Processing Systems 14*. MIT Press, 2002.

List of Own Publications

Jasmina Bogojeska, Daniel Stöckel, Maurizio Zazzi, Kaiser Rolf, Incardona Francesca, Rosen-Zvi Michal, Thomas Lengauer (2012). History-alignment models for bias-aware prediction of virological response to HIV combination therapy. *Proceedings of the 15th International Conference on Artificial Intelligence and Statistics (AISTATS).*

Jasmina Bogojeska (2011). History distribution matching method for predicting effectiveness of HIV combination therapies. *Advances in Neural Information Processing Systems (NIPS).*

Jasmina Bogojeska, Thomas Lengauer (2011). Hierarchical Bayes model for predicting outcomes of HIV combination therapies. *(submitted).*

Hiroto Saigo*, André Altmann*, **Jasmina Bogojeska**, Fabian Müller, Sebastian Nowozin, Thomas Lengauer (2011). Learning from past treatments and their outcome improves prediction of in vivo response to anti-HIV therapy *Statistical Applications in Genetics and Molecular Biology (SAGMB) 10(1): art6.*

Jasmina Bogojeska, Steffen Bickel, André Altmann, Thomas Lengauer (2010). Dealing with sparse data in predicting outcomes of HIV combination therapies. *Bioinformatics, 26(17):2085–2092.*

Jasmina Bogojeska, Adrian Alexa, André Altmann, Thomas Lengauer, Jörg Rahnenführer (2008). Rtreemix: an R package for estimating evolutionary pathways and genetic progression scores. *Bioinformatics 24(20):2391.*

Steffen Bickel, **Jasmina Bogojeska**, Thomas Lengauer, Tobias Scheffer (2008). Multi-Task Learning for HIV Therapy Screening. *Proceedings of the International Conference on Machine Learning (ICML).*

Jasmina Bogojeska, Thomas Lengauer, Jörg Rahnenführer (2008). Stability analysis of mixtures of mutagenetic trees. *BMC Bioinformatics 9:165.*