UNIVERSITY OF CALIFORNIA, SAN DIEGO

Modeling Complex Data in HIV-1: Kernel-based Machine Learning Approaches to Understanding Antiretroviral Drug Regimens and Neutralizing Antibody Activity

A dissertation submitted in partial satisfaction of the requirements for the degree of Doctor of Philosophy

in

Biomedical Sciences

by

Lorne William Walker

Committee in Charge:

Professor Douglas D. Richman, Chair
Professor Richard H. Haubrich
Professor Sergei L. Kosakovsky Pond
Professor Lucila Ohno-Machado
Professor Leor S. Weinberger
Professor Gene Yeo

2012

The Dissertation of Lorne William Walker is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

Chair

University of California, San Diego

2012

DEDICATION

For Laurel

There is a single light of science, and to brighten it anywhere is to brighten it everywhere.
— Isaac Asimov

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would like to thank Professor Douglas Richman for his support as my dissertation committee chair. Professor Richard Haubrich, chair of the CNICS Viral Resistance Group, is responsible for my involvement in the CNICS collaboration, which comprises much of Chapters 2-4 of this work. The research presented in Chapter 5 was done in collaboration with Professor Sergei Kosakovsky Pond. Professor Simon Frost originally proposed and implemented the application of the subset-tree kernel to HIV-1 drug regimen data. All of the above individuals have been exceedingly generous with their time, resources, attention and encouragment for which I am indeed grateful.

Professors Lucila Ohno-Machado, Leor Weinberger and Gene Yeo generously agreed to serve on my dissertation committee. All provided valuable insight and input which was always positive in tone and edifying in content.

I would also like to thank my co-workers in the UCSD AVRC Viral Evolution and CNICS research groups, past and present. This group provided insightful conversation, novel suggestions and comic relief, all of which added greatly to my graduate experience.

Chapter 2, in full is currently being prepared for submission for publication of the material. Lorne W. Walker, Sergei L. Kosakovsky Pond, Jeannette L. Aldous, Art F. Y. Poon, Shelly Sun, Sonia Jain, Simon D. W. Frost, Richard H. Haubrich. The dissertation author was the primary author of this work.

Chapter 5, in part is currently being prepared for submission for publication of the material. Lorne W. Walker, N. Lance Hepler, Dennis Burton, Douglas Richman, Sergei L. Kosakovsky Pond. The dissertation author was the primary author of this work.

VITA

2002	Bachelor of Science, Biochemistry and Applied and Computational Mathematical Sciences, University of Washington, Seattle
2002-2005	Research Scientist, Lingappa Lab, University of Washington, Seattle
2012	Doctor of Philosophy, Biomedical Sciences, University of California, San Diego
2013	Doctor of Medicine (expected)

PUBLICATIONS

Thielen BK, Klein KC, Walker LW, Rieck M, Buckner JH, Tomblingson GW, and Lingappa JR. T cells contain an RNase-insensitive inhibitor of APOBEC3G deaminase activity. PLoS Pathogens, 2007 Sep 21;3(9):1320-34.

Aldous JL, Jain S, Sun S, Walker LW, Mathews WC, Kitahata M, Kahn J, Saag M, Rodriguez B, Boswell SL, Frost SDW, Haubrich RH. Increasing prevalence of multiclass antiretroviral exposure and decreasing prevalence of virologic failure from 2000-2008: Data from the CFAR Network of Integrated Clinical Systems (CNICS) Cohort. In Review.

Walker LW, Kosakovsky Pond SL, Aldous JL, Poon AFY, Sun S, Jain S, Frost SDW, Haubrich RH. Representation and comparison of structured treatment data for statistical analyses: an application to HIV-1 antiretroviral regimens. In Preparation.

Walker LW, Hepler NL, Burton DR, Richman DD, Kosakovsky Pond SL. A 3-dimensional spatial neighborhood kernel method for learning predictive models for antibody neutralization in HIV-1. (working title) In Preparation.

ABSTRACT OF THE DISSERTATION

Modeling Complex Data in HIV-1: Kernel-based Machine Learning Approaches to Understanding Antiretroviral Drug Regimens and Neutralizing Antibody Activity

A dissertation submitted in partial satisfaction of the requirements for the degree of Doctor of Philosophy

in

Biomedical Sciences

by

Lorne William Walker

Professor Douglas D. Richman, Chair

The advent of modern antiretroviral (ARV) therapy has changed the face of the HIV/AIDS epidemic, yet many challenges still remain. Previously suppressed viral replication may rebound, while the search for an effective HIV-1 vaccine has produced meager results. Recent efforts and modern technologies have provided an unprecedented volume of clinical and molecular data, along with innovative mathematical tools for analysis. Here we apply a class of machine learning techniques called *kernel methods* to several domains in HIV-1 research. Kernel methods require the designation of a *kernel function*, which quantifies the similarity between data points in some informative fashion.

The first kernel function explored here is the subset-tree (SST) kernel. Originally developed in the natural language field for comparison of sentences formatted as trees, the SST kernel quantifies tree-similarity based on the number of shared sub-features. We apply this tool to ARV drug regimens. Modern ARV therapy is administered as combinations of three or more drugs, which are grouped

in certain patterns. SST-kernel based analysis combines knowledge of drug class, protein target and regimen composition, in addition to particular drugs employed. This SST kernel analysis was then used to inform survival models for predicting ARV regimen failure and overall disease progression. Incorporation of the SST kernel results improved prediction of ARV failure. Although more mixed, initial results indicate this may prove a fruitful approach for disease progression models as well.

Identification and characterization of HIV-1 neutralizing antibodies has long been of interest in the search for an HIV-1 vaccine. Here we develop a novel kernel method for comparison of HIV-1 envelope (Env) sequences, and apply this tool to prediction of antibody neutralization. This "structural neighborhood" kernel takes into account the primary Env sequence, as well as available knowledge of the 3-dimensional structure of Env. This method defines small 3-dimensional neighborhoods, and searches for those best able to predict antibody neutralization. Aside from incorporation of potentially important structural information in predictions, this method is robust uncertainties in protein alignment. This method was applied to six neutralizing antibodies, and produced classification accuracies at or near 80% in most cases.

Chapter 1: Introduction

Modern technologies provide scientists, physicians and the public a heretofore impossible volume and richness of data. In 2007, the worldwide digital data storage capacity was estimated at 276 exabytes ($2.76\text{x}10^{14}$ MB). In 1986, the same figure amounted to 0.0208 exabytes, a change of 10,000-fold in hardly more than two decades [1]. 1986 also marked the conclusion of the first clinical trial of zidovudine (AZT) for treatment of HIV/AIDS [2]. In the intervening 25 years, pharmacological antiretroviral (ARV) therapy has also advanced impressively, rendering HIV a chronic infection with life expectancy approaching that of non-infected individuals [3]. Despite this progress, neither a cure nor a vaccine has been developed, and numerous challenges persist in both efforts [4, 5]. The advances of the last few decades provide HIV-1 related data of unprecedented size and scope. Modern mathematical approaches may provide new insight on the clinical and molecular mechanisms of this disease.

Complex and large-scale clinical and molecular HIV-1 datasets represent both a boon and a challenge to researchers. Analysis of complex and large-scale data is increasingly important as data grows both in number of data points and number of measurements per data point. Large HIV-1 clinical cohorts, such as the CFAR Network of Integrated Clinical Sites (CNICS) provide the opportunity to study the factors driving the HIV pandemic in a real-world setting. CNICS is an electronic medical record-based clinical cohort drawn from eight academic Center for AIDS Research (CFAR) sites around the United States [6, 7]. This co-

hort includes more than 20,000 HIV-1 infected individuals, and collects many types of data, including demographics, clinical diagnoses, ARV medication use and laboratory test results. Cohorts of this size provide unique and immensely valuable data to researchers, but also present serious analytic challenges. Complex data, such as longitudinal samples of laboratory values and changes in ARV therapy over time are challenging to adequately model, and are likely intimately related to outcomes of interest. Advanced molecular assays also provide important, yet challenging opportunities to better understand HIV-1 disease. Genome-wide association studies, DNA sequencing (including massively-parallel "deep sequencing"), microarray technology, and *in vitro* methods for testing viral virulence, drug resistance, chemokine co-receptor tropism and antibody neutralization provide unprecedented large-scale and high-complexity data [8, 9, 10, 11].

Modern machine learning techniques provide a valuable toolset for approaching complex biological data. Machine learning technologies are utilized in fields including bioinformatics, econometrics, risk management, and many others [12]. These techniques aim to leverage large data sets, and learn relationships from the data with minimal prior assumptions. A particular class of machine learning techniques, called kernel methods, are particularly flexible to complex relationships in data. In the work presented here we focus on the application of kernel methods to two important problems in HIV-1 research: the clinical implications of patterns in ARV drug use, and the understanding and prediction of antibody neutralization.

1.1 Antiretroviral Regimens and HIV

The disease now known as HIV/AIDS was first described in 1981, when clinicians noted unexplained clusters of Pneumocystis carinii pneumonia and Kaposi's Sarcoma in homosexual communities in the United States [13]. This

syndrome, marked by profound immune dysfunction and subsequent opportunistic infections, was termed acquired immune deficiency syndrome (AIDS) in 1982 [14]. The following year, two research groups in the United States and France reported identification of a novel human retrovirus, now known as HIV-1 [15, 16]. Thirty years after the initial clinical descriptions in 1981, HIV-1 is a worldwide pandemic, responsible for nearly 30 million deaths in that span and an estimated 33.3 million individuals living with HIV infection in 2010 [17].

Fortunately, effective antiretroviral medications have transformed a once uniformly fatal infection into an often manageable chronic condition. With proper therapy, it is estimated that an early-stage HIV infection newly-diagnosed at the age of 55 may only subtract 1.4 years from the average life expectancy. This figure drops to 0.4 years if diagnosed at the age of 25 [3]. Zidovudine (AZT) was the first anti-HIV medication, granted FDA approval in 1987. AZT was the first in a line of nucleoside or nucleotide reverse trancriptase inhibitors (NRTIs). This class of drugs is comprised of analogs of nucleotides or nucleosides which disrupt the activity of viral reverse transcriptase (RT) when incorporated into the growing DNA strand [18]. While NRTIs still form the backbone of modern ARV therapy, treatment with one or two NRTI drugs ultimately proved insufficient. In the presence of these early single- or dual-drug treatments, the HIV-1 virus inevitably evolved drug-resistance mutations, ultimately rendering therapy ineffective [19, 18].

In late 1995 and 1996, two new classes of ARV drugs became available — the non-nucleoside/nucleotide reverse transcriptase inhibitors (NNRTIs) and protease inhibitors (PIs) [20, 21, 22]. Unlike the NRTI drugs, the NNRTIs bind outside the enzymatic pocket of RT, and inhibit the protein through allosteric effects [23]. When given alone, the HIV-1 virus rapidly evolves resistance to these agents as well [24]. However, when drugs from multiple classes are combined, durable viral suppression can be achieved in many patients [18]. In modern combination antiretroviral therapy (cART), typically two or more NRTIs are used in

Table 1.1: Summary of antiretroviral drug classes. NRTI: nucleotide/nucleoside reverse transcriptase inhibitor, NNRTI: non-nucleotide/nucleoside reverse transcriptase inhibitor, PI: protease inhibitor.

Drug Class	Drug Target	First FDA Approval	First-line Therapy?
NRTI	Reverse transcriptase	Zidovudine (AZT), 1987	Yes
PI	Protease	Saquinavir, 1995	Yes
NNRTI	Reverse transcriptase	Nevirapine, 1996	Yes
Fusion Inhibitor	Envelope (gp41)	Enfuvirtide, 2003	No
Entry Inhibitor	Envelope (CCR5)	Maraviroc, 2007	No
Integrase Inhibitor	Integrase	Raltegravir, 2007	No

combination with PIs or an NNRTI. While these drugs are individually susceptible to evolution of viral resistance, the probability of the virus generating multiple simultaneous resistance adaptations is very low [19].

cART remains the standard of care for HIV-1 infection in the developed world. Recent innovations have continued to improve the convenience and effectiveness of therapy, and provide additional therapeutic options when first-line therapies fail. Ritonavir, a PI with low anti-viral potency, was found to increase the bioavaliability of other PIs, and is now frequently used as a "boosting" agent in PI-containing regimens [25]. Multi-drug formulations have reduced pill burden and made adherence to ARV therapy easier. In 2003, the FDA approved enfuvirtide, the first HIV-1 fusion inhibitor, which prevents fusion of the viral membrane with target cells by binding the gp41 subunit of the viral envelope protein. Two additional ARV drugs with novel mechanisms became available in 2007: maraviroc, which blocks cell entry through inhibition of CCR5 binding, and raltegravir, which inhibits the viral integrase protein. While NRTI/NNRTI and NRTI/PI combination therapies remain the primary choices for initial therapy, these new drug classes provide important options for salvage therapy when first-line regimens fail (Table 1.1).

While ARV therapy has improved profoundly since the approval of AZT in 1987, challenges remain. Many factors can contribute to sub-optimal viral suppression, including non-adherence to regimen, drug-drug interactions and variations in drug metabolism [26]. Incomplete suppression of virus increases

the risk of emergent drug resistance, and extensive patterns of resistance can develop. Virologic failure of a drug regimen, defined as on-therapy rebound viral replication, marks a poor prognosis, and increases the risk of subsequent failed treatments [27]. Optimization of ARV therapy is a difficult task. With 25 ARV drugs on the market, there are nearly 15,000 possible 3- and 4-drug combinations to choose from. While not all of these combinations are clinically relevant, individual assessment of all potentially useful drug regimens remains impractical. Furthermore, individuals may exhibit complex patterns in ARV use, including gaps in treatment and multiple changes in ARV regimen. The large number of choices and high complexity of these data make analysis of the implications of patterns in ARV use difficult.

1.2 Neutralizing Antibodies and HIV

Development of a vaccine that provides protection against infection with the HIV-1 virus has long been a goal of researchers. To date, success in vaccine trials has been modest. Early efforts focused on eliciting neutralizing antibody against soluble derivatives of envelope (Env), the HIV-1 surface protein. Two soluble-Env vaccine candidates, both variants of the AIDSVAX vaccine design, began phase III clinical trials in 1998-1999 in the United States, the Netherlands and Thailand. In these trials, a total of seven injections of locale-appropriate recombinant gp120 Env-subunit protein were administered with an aluminum adjuvant. Approximately 7,900 volunteers were enrolled in these two trials, and no evidence of protection against infection, or reduction in viral levels was found [28, 29]. In 2004, Merck and the NIH sponsored the 3,000-subject STEP trial of an Adenovirus-based vaccine designed to induce cell-mediated immunity. This trial was also unsuccessful, and was halted early when HIV-infection risk was found to be higher in the vaccine group than in the placebo group for some patient

subsets [30]. The only vaccine trial to date to report any positive effect tested a combination Canarypox vector/soluable Env gp120 vaccine on 16,402 volunteers in Thailand [31]. In this study, neither the intention-to-treat, nor the per-protocol analyses yielded statistically significant protection. However, if 7 subjects (5 vaccinated, 2 placebo) found to be HIV-infected at baseline were removed from the intention-to-treat analysis, the vaccine showed a modest 31.2% effectiveness (P = 0.04).

Molecular and *in vitro* studies have provided several clues as to why an effective HIV-1 vaccine has been so difficult to develop. In its mature form, Env is cleaved into the gp41 and gp120 subunits, which bind non-covalently as a trimer of heterodimers, and are present on the virus surface. gp120 is a highly-variable, heavily-glycosylated glycoprotein which is responsible for binding CD4 and chemokine co-receptors on target T-cells. The gp41 subunits act as the transmembrane anchor for the protein, and mediate cell-virus fusion at the time of cellular infection [32]. The extensive glycosylation of gp120 is thought to act as a "shield" against antibody neutralization [33, 34]. The high-diversity of Env — up to 35% amino acid sequence divergence between hosts — and the presence of five structurally flexible variable loops (named V1-V5) in gp120 also complicate the development of a single antibody that will neutralize a broad diversity of HIV-1 (Figure 1.1) [35, 36, 37]. Testing antibody neutralization *in vitro* has also provided insight into unsuccessful vaccination efforts. Within 8-12 weeks of HIV-1 infection, neutralizing antibody active against autologous virus can often be detected [38]. However, the virus rapidly evolves to escape neutralization by these antibodies [39, 40]. Broadly neutralizing antibodies (NAb), capable of neutralizing virus from other individuals, generally do not arise for years, if at all [41].

Despite the barriers to development of a broad antibody response to HIV-1, a small number of broadly-neutralizing NAb have been discovered. The first of these to be identified, IgGb12 (b12), targets the CD4 binding site in Env [42]. A

Figure 1.1: Schematic of HIV-1 Envelope structure — SP: signal peptide, V1-V5, variable loops of gp120, TM: transmembrane region. Potential N-linked glycosylation sites in gp120 highlighted by blue circles.

number of additional broadly neutralizing antibodies have been discovered, and target various features of the Env protein. These include antibodies targeting the gp41 stem, the V3 loop of gp120 and clusters of gp120-bound sugars [43, 44, 45]. The failure of the cell-mediated STEP vaccine trial has recently reignited interest in antibody neutralization, and high-throughput screens have begun to yield a new generation of broadly-effective NAb [46, 47, 48].

The broadly-effective NAbs identified to date vary greatly in the breadth and potency of their neutralization abilities. For instance, b12 potently neutralizes the majority of subtype B and C viruses, but was only active against 33% of a subtype-A panel. The gp41 binding antibody 2F5 potently neutralizes subtype A and B, but not C virus, while another gp41-binding antibody, 4E10 neutralizes nearly all viruses regardless of subtype, but with only a modest potency [49]. The elements of viral diversity that drive these differences in NAb sensitivity remain incompletely understood. The binding epitopes for many of these antibodies have been determined, yet site-directed mutagenesis experiments have identified artificial mutations distant from the canonical epitopes that modulate antibody sensitivity [50, 51]. The effect of point mutations on neutralization phenotype in Env appears to be very sensitive to genetic context. Mutations that confer neutralization resistance or sensitivity in one genetic backbone may result in the opposite phenotype or non-viable virus in another [52]. This suggests that molecular techniques that rely on viral strains with non-natural mutants — for example, alanine scanning, where each amino acid site is sequentially mutated to alanine — may provide results that do not generalize well to natural variants.

Elucidation of the naturally-occurring Env variations that drive antibody

neutralization presents a challenging problem. Some efforts have been made to computationally map naturally-occurring mutations in Env that are associated with in vitro neutralization phenotype. Two studies have looked for amino acid variations that influence neutralization by polyclonal sera from subtype C infected individuals in India [53, 54]. These studies identified a number of potentially influential amino acids sites, including several in a specific amphipathic α-helix. Interestingly, although amino acid variation in this helical structure was tightly associated with neutralization phenotype, swapping this domain between gp120 backgrounds failed to transfer phenotype between variants. This reinforces the notion that mutations influencing neutralization may be context-specific. To our knowledge, one study has attempted to computationally map mutations in Env sequence to in vitro neutralization phenotypes for a monoclonal antibody. Gnanakaran et. al., identified a number of variations associated with neutralization with b12 [55]. In this study as well, many variations distant from the canonical binding epitope appear to play a role in b12 neutralization phenotype.

Computational prediction of antibody neutralization of HIV-1 presents a challenging analytic task. The high degree of sequence diversity of Env makes multiple sequence alignment impractical in many cases. For this reason, particularly diverse regions may be masked for analytic convenience. Furthermore, any mistakes or ambiguities in alignment of the remaining sequence may mask important information, or produce spurious signals. Analysis of primary amino acid sequence also neglects an important piece of information — the 3-dimensional structure of the target protein. Antibody-antigen binding is an interaction between two 3D objects, and the features determining NAb sensitivity may be 3D in nature [56]. For example, two recently identified NAbs, PG9 and PG16, appear to only bind the fully-assembled Env trimer, highlighting the importance of structure in understanding neutralization. Analytic tools that allow researchers

to contend with this complex situation may be necessary to fully understand antibody neutralization of HIV-1, and advance the quest to develop an effective HIV-1 vaccine.

1.3 Kernel Methods

Here we have outlined two problems in HIV-1 research where data exists, but inherent complexities complicate analysis. ARV medications are powerful antiviral tools, but the huge number of possible drug combinations and complexity of patterns in use complicate the picture. Antibodies that target the viral Env protein have been identified, but natural determinants that drive neutralization are difficult to characterize, particularly in light of the high genetic variability of Env and 3-dimensional complexity of the protein. Kernel methods present a flexible, practical, and potentially powerful toolset for analysis of such complex, high-dimensional data.

Kernel methods refer to a set of machine learning methods that have at their root the definition of a kernel matrix (or Gram matrix), a symmetric, positive semi-definite matrix that defines the inner product between data vectors in some reproducing kernel Hilbert space (reviewed in [57], first defined in [58]). Intuitively, the kernel matrix can be thought of as a matrix of similarity values that relate the data points in a sample. For any two data vectors x_i, x_j, this kernel score can be defined as by a function, $\kappa(x_i, x_j)$ Hence, for n data points, the kernel matrix K is simply the nxn matrix with the (i, j)-th entry $K_{i,j} = \kappa(x_i, x_j)$. In the simplest case (the linear kernel), this function is simply the inner (or dot) product between data vectors, or $\kappa(x_i, x_j) = \langle x_i, x_j \rangle$. To provide more flexibility, we can introduce a function $\phi(x)$, imposing some (possibly non-linear) transformation on x, and let $\kappa(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle$. Given an appropriate choice for ϕ, learning problems, such as the classification of X's and O's in Figure 1.2A may be

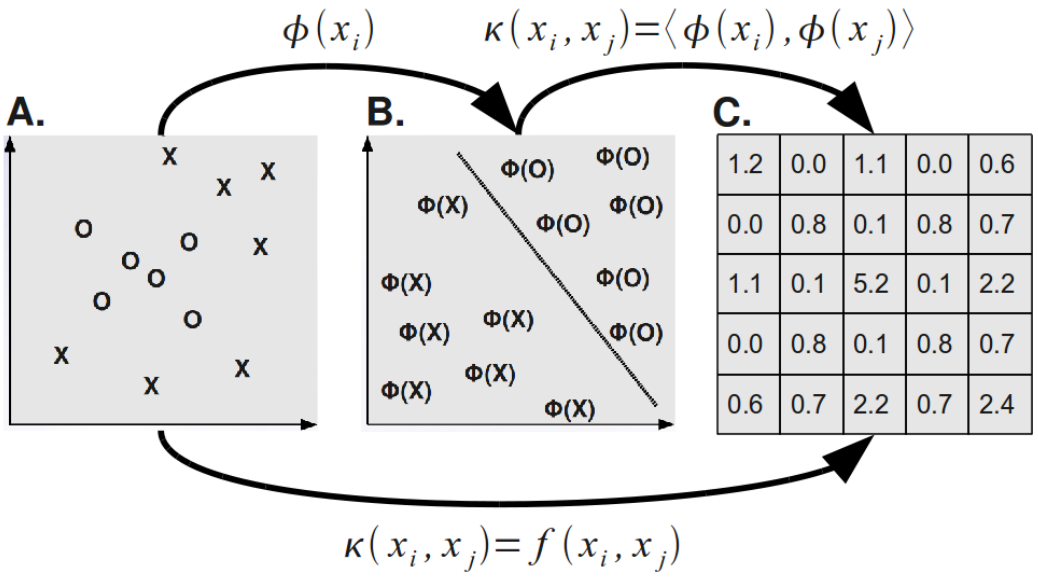


Figure 1.2: Outline of kernel data transformations. Schematic of the relationship between input space (A) feature space (B) and the kernel matrix (C).

transformed from difficult non-linear problems in the original input space into more tractable linear tasks in the so-called "feature space" (Figure 1.2B)

The great power and utility of kernel methods stems from two additional relationships. First, Mercer's Theorem, originally proven in the context of integral equations, establishes the symmetric relationship between a positive semi-definite kernel function $\kappa(x_i, x_j) = f(x_i, x_j)$ and the existence of $\langle \phi(x_i), \phi(x_j) \rangle$ as a valid inner-product space [57, 59]. The practical implication of this relationship is that one can design a kernel function without specifying or characterizing the corresponding transformation ϕ. This allows us to circumvent the top path from A to C in Figure 1.2, opting instead for the bottom path, and leaving the definition of ϕ implicit. This is sometimes referred to as the "kernel trick." The only theoretical requirement is that the kernel function satisfy Mercer's condition, or equivalently that the resulting kernel matrix is symmetric and positive semi-definite. Second, the representer theorem states that many convex optimization problems, which are generally stated in a so-called "primal" form, can be transformed into a "dual" representation, where the solution can be expressed

as a weighted sum of the input data. These dual-form optimization problems can often be solved directly from the kernel matrix. This opens up a wide array of optimization algorithms to use with kernels, such as the support-vector machine (SVM), a powerful algorithm for classification and regression problems.

This combination of flexibility and power has driven the recent rise in popularity of kernel methods. A number of general-purpose kernel functions for continuous numerical input have been well described, and include the linear, polynomial and Gaussian kernels [57]. There has been a rapid proliferation of specialized kernels that handle specific types of data, including kernels for text and image classification, graph and tree kernels, and the Fisher kernel for generative probabilistic models among many others [60, 61, 62, 63]. In addition to SVM methods for regression and classification, kernel-based methods have been developed for many applications, including clustering, ranking, principal components analysis, multi-dimensional scaling and outlier detection [57]. An additional advantage of kernel methods is that the complexity of these analytic methods (SVMs, clutering, etc.) only increase in complexity as the size of the kernel matrix (ie. the size of the dataset). The data transformation functions κ or ϕ (depending on the approach) can be made as complex as desired, without effecting the computational complexity of the final analysis step.

In spite of their increasing popularity, there are several drawbacks to kernel methods. The calculation of a kernel matrix requires a transformation of the input data that generates an abstraction of the original input that may be difficult to interpret. For instance, maximum-variance data projections can be generated by kernel principal components analysis. These components will be guaranteed to capture as much of the transformed data variance as possible, and may be very effective representations of the information of interest. At the same time, for most kernels,6 the data transformation can not be practically reversed — ie, what those principal components *mean* with respect to the input data may not

be clear. Similarly, support-vector methods for regression and classification will produce solutions in the form of linear combinations of the rows (or equivalently columns) of the kernel matrix. If the objective is to produce classification rules that are easily interpretable, or can be applied manually, other tools, such as logistic regression or decision trees, may be more appropriate. Finally, as with all statistical modeling approaches, high flexibility comes at the risk of overfitting data. Care must be taken to avoid this pitfall through techniques such as regularization and cross-validation.

1.4 Dissertation Outline

The goal of the ensuing chapters of this work is to determine if non-linear kernel approaches can improve parameterization of complex, high-dimensional data relevant to HIV-1. For ARV drug regimens, this may allow better prediction of the clinical consequences of patterns in ARV use. When applied to the 3-dimensional protein targets of antibody neutralization, these tools may provide an efficient computational approach to characterizing novel antibodies and predicting the neutralization of untested viral strains. In Chapter 2 we define the subset-tree (SST) kernel, describe how it can be applied to ARV drug regimens, and determine if this approach can improve simple models for average regimen performance metrics. In Chapter 3 we extend this SST-kernel approach to models estimating the hazard of ARV regimen failure, as applied to the CNICS cohort data set. In Chapter 4 we develop models for overall disease progression, and determine if the SST-kernel can improve predictions of the risk of HIV-1 infection progressing to AIDS or death. In Chapter 5, we describe the a novel kernel function for 3-dimensional protein structures, and apply it to the prediction of antibody neutralization sensitivity. Finally, in Chapter 6 we summarize the results presented here, and discuss possible extensions of this work.

Chapter 2: Representation and comparison of structured treatment data for statistical analyses: an application to HIV-1 antiretroviral regimens

2.1 Introduction

Beginning with the approval of AZT as the first anti-HIV medication in 1987, the arsenal of available antiretroviral (ARV) agents has expanded to 25 by 2010, with several more in the approval pipeline [18]. Modern ARV therapy typically includes 3 or 4 antiviral drugs applied in combination. While relatively few of the 14,950 such combinations are commonly prescribed, the number of utilized combinations is sufficiently large to complicate analysis. Paradigms of regimen construction also evolve over time as newer, more effective and tolerable agents are developed and toxic and difficult regimens fall out of favor. Early single- and dual-drug therapies proved ultimately insufficient due to the rapid generation of drug resistance mutations in the target protein and were replaced by combinations of three or more drugs in the mid-1990s. Combination therapy proved effective as the simultaneous evolution of multiple drug resistance mutations is much less likely than evolution of resistance to one or two agents [22, 20, 64, 19]. Regimens are carefully selected in the hope of achiev-

ing complete viral suppression and staving off development of drug resistance. Despite this remarkable progress, regimen failure and on-therapy viral rebound continue to occur, and are markers of poor overall prognosis [27].

Modern ARV therapy typically combines drugs of different mechanisms or against different targets. These regimens typically include a "backbone" of two nucleotide or nucleoside analogues (NRTIs), which inhibit the viral reverse transcriptase (RT). Additionally, regimens often include either a non-NRTI RT inhibitor (NNRTI) or an inhibitor of viral protease (PI) whose bioavailability is boosted with ritonavir. Recently-approved agents target viral envelope glycoprotein and integrase, but are generally reserved for individuals ineligible for standard regimens due to viral resistance or drug intolerability [18]. We can view the composition of these regimens as a hierarchy of information, spanning general features such as the proteins targeted by a regimen and detailed features such as the individual drugs included.

The representation of these regimens in analytic models remains relatively simplistic, despite our rich knowledge about ARV drugs. While this may not be a challenge in randomized controlled trials comparing a small number of alternative therapies, representation of ARV data in observational cohort studies is problematic. Analysis of cohort data requires modeling response to particular regimens, but also controlling for confounding by regimen selection when studying other outcomes, such as co-infections or development of cardiovascular disease. Our goal is to better parameterize ARV use for HIV outcomes research. In existing analyses, drug regimens are typically lumped together into general regimen types (e.g. NRTI + boosted PI), considered on the basis of the individual drugs included, or sorted into other general categories based on expert opinion [27, 65, 66, 67]. In each of these approaches, much of our knowledge concerning the regimen in question is discarded for the sake of analytic tractability. Unfortunately, depending on the question being asked, it may be unclear *a priori* what regimen

features are most influential. Perhaps particular drugs are important in predicting an outcome of interest, or alternatively the individual drug identities are not as important as the regimen type. Of course, combinations or subsets of these features may be influential and analytic tools that are flexible to all possibilities would be ideal.

We propose an approach that captures the information-rich nature of ARV regimens by representing them as trees, and apply this technique to observational data from the Center for AIDS Research (CFAR) Network of Integrated Clinical Systems (CNICS) cohort [6]. This cohort is comprised of data from point-of-care medical records from clinical sites in the United States, and reflects the complexity of ARV use in observational clinical data. The main tool of our approach — the subset-tree (SST) kernel — was originally developed in the context of natural language processing, where it can be used to represent sentences in a way that considers both the grammatical sentence structure (as laid out in a tree), and the individual words that comprise the sentence [61, 68]. In a similar sense, we wish to represent ARV regimens in a way that synthesizes what is known about regimen structure and the drugs included into a single mathematical representation. The SST-kernel achieves this by calculating similarity scores between regimens that capture features from all levels of this hierarchy of regimen information. Once this kernel calculation has been made, many kernel-based algorithms exist to perform common tasks, such as categorization, regression and ranking (several are described in [57]). This provides an approach to parameterizing ARV regimens for analysis that is both information-rich and practical.

2.2 Results

Data Summary

25,508 records of ARV treatment initiation or change were retrieved from the CNICS data repository and passed our data quality filters (see Methods). Since our analyses depended on measures of clinical outcome averaged over regimen instances, we limited the dataset to those regimens observed five or more times with associated HIV-1 plasma RNA viral load (VL) measurements. The remaining 19,129 records encompassed 9,184 unique individuals and represented 393 distinct drug combinations (see Table 2.1). The median regimen initiation date was 27 October 2003, and regimens had a median duration of 396 days. The cohort was largely male (78%) and white (51%), and members had a median age of 39 years at enrollment. Subjects were spread across six member sites of CNICS, with the most common site being the University of California at San Diego (26%) and the least common the Fenway Community Health Center (10%).

For each unique ARV regimen selected for analysis, summary statistics of regimen performance were calculated. Median duration, percent of subjects who achieved viral suppression (defined as VL $<$200 copies/mL at any point more than seven days post-initiation), or experienced virologic failure (on-therapy VL $\geq$ 200 copies/mL at any point more than 24 weeks post-initiation) were computed. Overall, 73% of regimen instances achieved viral suppression, 29% experienced virologic failure, 18% experienced both and 16% neither. Regimens which achieved neither suppression nor failure were generally unsuccessful regimens terminated or changed before the 24-week minimum to qualify as "failed." When separated into distinct drug regimens, the median duration of treatment on a particular regimen (measured as a median of median durations for each distinct regimen) was 331 days (IQR [224-442]), the median percent of subjects achieving viral suppression for a particular regimen was 67% (IQR [50%-80%]), and the

median of those experiencing virologic failure for a particular regimen was 30% (IQR [20%-42%]) (Table 2.1).

Kernel Data Transformations

We applied four different kernel functions, which represent inner products in different feature spaces, to the 393 unique ARV regimens observed in the filtered dataset. The SST-kernel function applies a non-linear transformation to the data, while each of the other three kernels calculates the linear inner product of some summary variable. The SST-kernel accounts for the structured regimen features included in the rooted tree graph (Figure 2.1A-C). The SST kernel function includes an adjustable kernel-parameter (λ) which modifies the relative influence of subset trees near the root or the tips of the regimen trees depicted in Figure 2.1. An optimum value for $\lambda = 0.55$ was estimated from the data to maximize the statistical performance of a linear model incorporating kernel-transformed data, and is used throughout the manuscript. The linear kernel functions score regimen similarity by calculating the inner product between vectors of summary variables. The "linear drug", "linear class" and "linear regimen type" kernels consider only a list of indicators of individual drugs, a count of included drugs in each class, or a categorization by broad regimen class respectively. Because no non-linear data transformation is being made, the linear kernel-transformed data can be seen as equivalent to the input data, albeit in a different format. For instance, the linear regimen type kernel is equivalent to parameterizing regimen as a categorical variable with each of the regimen types as a possible value. Differences between the kernels functions are summarized in Table 2.2.

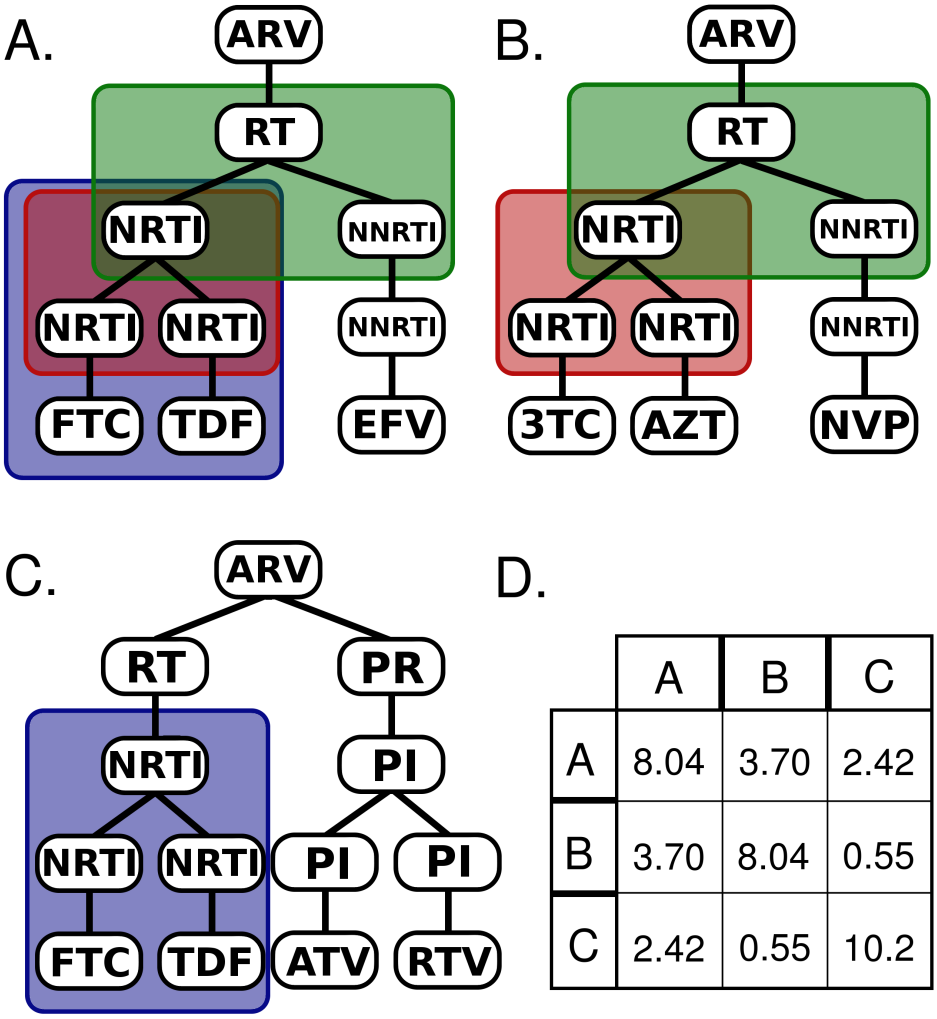
Unlike the linear kernels, the SST-kernel is capable of incorporating regimen features from all levels of this hierarchy into a single framework. For instance, regimens A and B in Figure 2.1 share no drugs — and would therefore bear no similarity according to the linear drug kernel — yet the subset trees out-

Table 2.1: Summary of cohort data. NRTI: nucleot(s)ide reverse transcriptase inhibitor, NNRTI: non-nucleotide reverse transcriptase inhibitor, PI: protease inhibitor, IQR: interquartile range.

Category	**N**	
All Records	38,953	
Filtered Records	25,508	
Records with qualifying regimens	19,129	
Unique Individuals	9,184	
Unique Regimens	393	
Regimen Records (N=19,129)	**N (%)**	**Median (IQR)**
Start Date		9/27/2003 (9/2/2000-5/7/2006)
Regimen Duration (days)		396 (168-846)
Viral Suppression	13,936 (73)	
Virologic Failure	5,193 (29)	
Unique Individuals (N = 9,184)	**N** (%)	**Median (IQR)**
Age (data missing for 856)		39 (33-45)
Gender:		
Male	7,145 (78)	
Female	1,183 (13)	
No Record	856 (9)	
Site:		
UCSD	2,429 (26)	
UW	1,894 (21)	
UAB	1863 (20)	
CWRU	1193 (13)	
UCSF	928 (10)	
Fenway	877 (10)	
Ethnicity/Race:		
White Non-Hispanic	4,638 (51)	
Black	2,168 (24)	
Hispanic	1,085 (12)	
Other	437 (5)	
No Record	856 (9)	
Unique Regimens (N = 393)	**N (%)**	**Median (IQR)**
Median Duration		331 (224-442)
Median Percent Suppressed		67 (50-80)
Median Percent Failed		30 (20-42)
Regimen Type:		
NRTI + NNRTI	53 (13)	
NRTI + boosted-PI	125 (32)	
NRTI + PI (unboosted)	56 (14)	
NRTI only	18 (5)	
Other	141 (36)	

Table 2.2: Comparison of drug regimen representations. Subset-tree (SST) kernel utilizes tree representations outlined in Figure 2.1. The three linear kernels are equivalent to existing regimen representations.

Kernel Name	**Linear?**	**Information**	**Representation of FTC/TDF/ATV/RTV**	**Max Data Dimension**
Subset-tree	No	Drugs, classes, combinations	See Figure 2.1C	393
Drug	Yes	Drug Identities	FTC, TDF, ATV, RTV	25
Drug class	Yes	Drug Classes	2 NRTI, 2 PI	7
Regimen Type	Yes	Typical Drug Combinations	Boosted PI-containing regimen	5



	A	B	C
A	8.04	3.70	2.42
B	3.70	8.04	0.55
C	2.42	0.55	10.2

Figure 2.1: Outline of SST kernel calculation. A-C Regimen-trees for three common ARV regimens: A:FTC/TDF/EFV, B:3TC/AZT/NVP C:FTC/TDF/ATV/RTV Examples of subset trees shared between trees in A and B are outlined in green and red. An example of subset trees shared between trees in A and C are outlined in blue. D: SST-kernel scores calculated for the regimen trees in A-C. Higher scores indicate a greater degree of similarity between regimens and/or larger regimen size. FTC: emtricitabine , TDF: tenofovir, 3TC: lamivudine, AZT: zidovudine, EFV: efavirenz, NVP: nevirapine, ATV: atazanavir , RTV: ritonavir

lined in red and green are shared between these regimens, and would contribute to their SST kernel similarity score. The red-highlighted and green-highlighted subset trees correspond to the regimen features "contains two NRTI drugs" and "contains both NRTI and NNRTI" respectively. Additional subset trees shared between trees A and B would identify further shared features. In a similar sense, the SST kernel approach accounts for similarities between regimens of different types that nonetheless share some drugs. For instance, regimens A and C in Figure 2.1 are different types of regimen: regimen A is an NNRTI-containing regimen, while regimen C is a boosted-PI regimen. These regimens share nothing in common according to the linear regimen type kernel, yet they share the same NRTI backbone, and consequently share all subset trees contained in blue-highlighted sub-tree as scored by the SST-kernel.

The manner in which the SST-kernel function scores the similarity between the three regimens represented in Figure 2.1A-C can be seen in Figure 2.1D. The regimen FTC/TDF/EFV (A) is more similar to regimen 3TC/AZT/NVP (B) (i.e. has a higher kernel function score) than when either NNRTI-containing regimen is compared to FTC/DTF/ATV/RTV (C). This matches our expectation that regimens A and B belong to the same type of regimen (NRTI+NNRTI), while regimen C is a boosted protease-containing regimen. However, all of these regimens share at least one feature — in this case, the inclusion of two NRTI drugs — hence, none of the values in Figure 2.1D are zero, as would sometimes be the case for the linear kernels. In this manner, the SST kernel description of ARV regimen is more informative than summary variables that focus solely on drug composition or regimen type.

Kernel Principal Components Analysis (KPCA)

KPCA was applied to the four different kernel-transformed datasets described above. The SST-kernel KPCA yielded 337 significant principal compo-

nents (of a maximum of 393 — the number of distinct regimens) the first 90 of which captured 95% of the data variability. The drug linear kernel produced 25 principal components (of a maximum of 25 — the number of drugs), the first 16 of which captured 95% of the data variability. The linear class kernel produced 7 principal components (of 7 possible), the first 4 of which captured 95% of the data variability. The linear regimen type kernel produced 4 principal components (of 4 possible), of which all 4 were required to capture more than 95% of the data variability.

The 393 unique ARV regimens distributed according to the first two principal components of the SST kernel matrix, which accounted for 22% of the data variability, are plotted in Figure 2.2A. Visual segregation of these points into clusters is apparent and confirmed by kernel K-means clustering. Kernel K-means clustering considers distance in all 393 dimensions (instead of just the two represented in the plot), taking advantage of the full rank of the data, but producing clusters that may appear counter-intuitive in two dimensions. In contrast, an analogous plot for the linear drug kernel reveals no obvious clustering patterns (Figure 2.2B). The clusters identified by the SST-kernel KPCA analysis correspond to well-known types of ARV drug regimens. They include regimens containing an NRTI backbone and an NNRTI (cluster 1), non-boosted PI (cluster 3) and ritonavir-boosted PI (clusters 2 and 4). 95% of patient-years fall into these 4 clusters, which reflect known real-life ARV regimen classes (Figure 2.2C). Kernel K-means clustering was applied to the linear drug kernel seen in Figure 2.2B, examining the full rank of the data, and failed to separate regimens into these general regimen types (data not shown). This pattern of regimen clustering apparent in the SST-kernel transformed data not is reproduced by the linear drug kernel, and cannot be discerned from the drug identities alone.

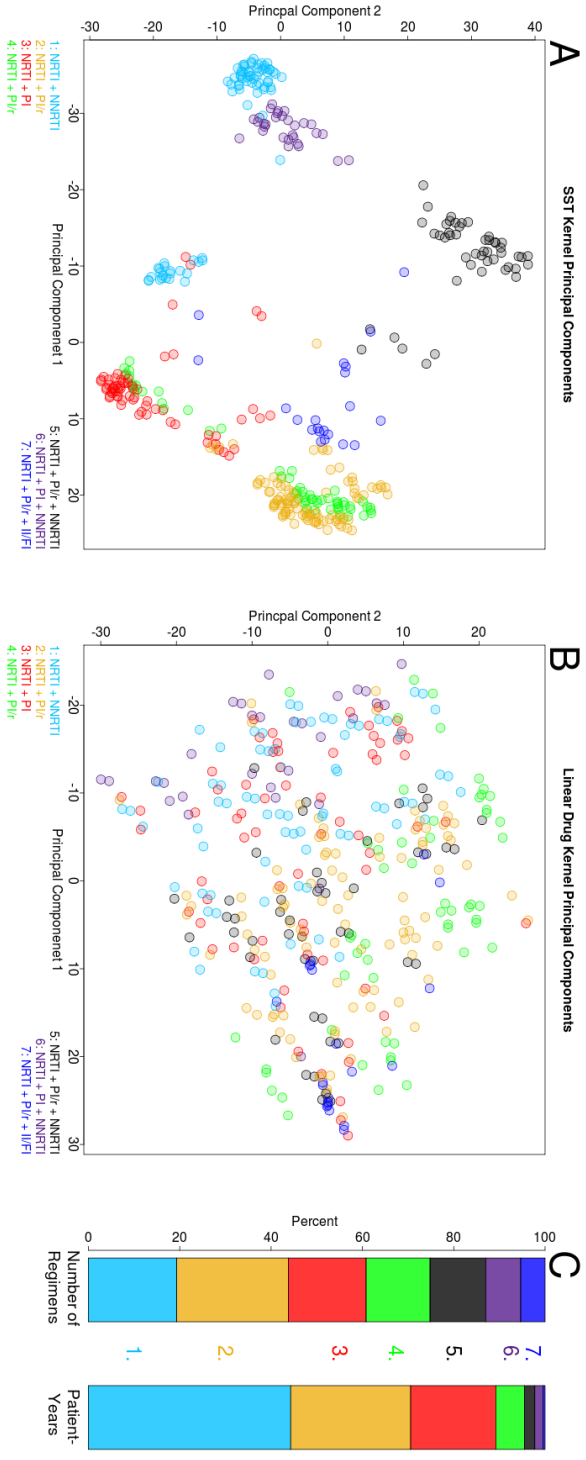


Figure 2.2: Kernel principal components analysis and clustering. A: plot of first two SST-kernel principal components. Kernel k-means clusters denoted by color-coding. B: A plot of first two linear-drug kernel principal components. Color-coding from 2A is preserved. C: Distribution of unique regimens and total patient-years among the clusters.

Linear Models for Regimen Performance

We assessed the ability of the four different kernel-transformed datasets described above to account for variation in regimen performance in multivariate linear and support vector regression (SVR) models. Clinical performance of regimens was summarized by three basic measures: median regimen duration, proportion of regimen instances achieving viral suppression, and proportion of regimen instances experiencing virologic failure (see *Data Summary* for definitions). The ability of the SST kernel and the three linear kernels to model differences in clinical outcome was assessed in two ways. First, KPCA was used to extract principal components, which then became covariates in a standard linear model. Akaike's Information Criterion (AIC) was used to assess the model fit for the KPCA-based models. Principal components were sequentially added to a base model containing no covariates. At each step, the next principal component that would provide the best (lowest) AIC value was added. The AIC values and optimum number of principal components are plotted in Figure 2.3A for all 12 models. In the models for regimen duration and regimen failure, SST-kernel based variables consistently provided a better fit to the data (lower AIC scores). For instance, in the model of virologic suppression, the linear drug kernel provided better AIC values for the first 15 variables added, although the SST kernel model was able to reach lower AIC values with more variables. As more than one record per individual is included in this full dataset, non-independence of samples represents a potential source of bias. When analysis was restricted to only the last regimen per individual, similar patterns were observed (Figure 2.4A).

Second, an SVR model was fitted directly using the kernel matrices. The predictive power of SVR models was assessed by mean squared error (MSE) in a 10-fold cross validation (Figure 2.3B). For the regimen duration and viral suppression models the SST kernel model provided lower MSE, while the linear drug

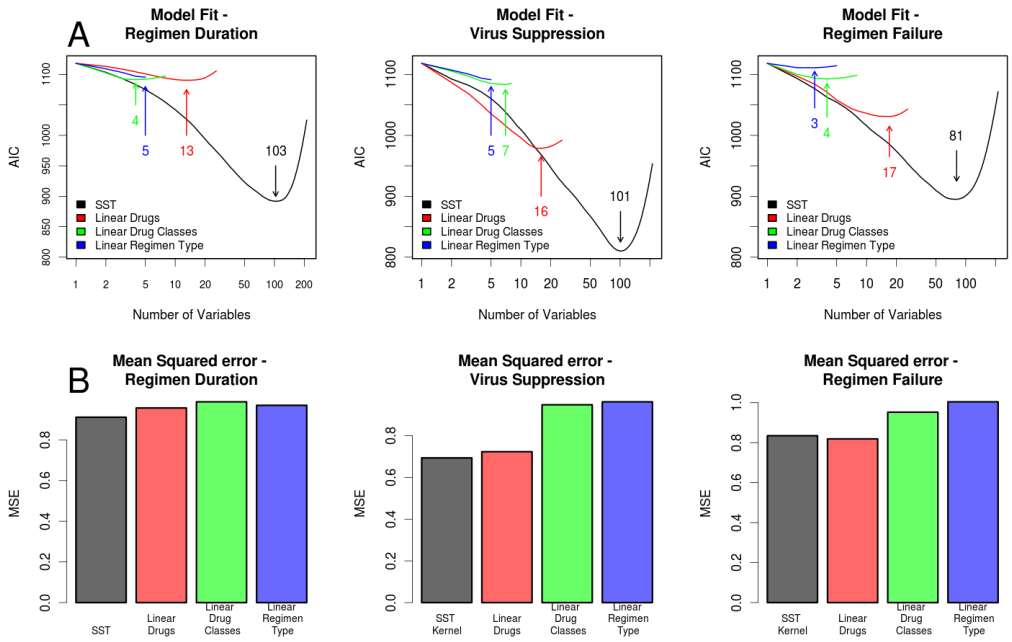


Figure 2.3: Linear model fit, all data. A: Akaike's information criterion of principal-components driven models for regimen duration, virus suppression and regimen failure for all records in data set. Arrows indicate the number of covariates in the best fit (minimum AIC) model. B: Residual mean-squared error for support-vector regression driven models for regimen duration, viral suppression and regimen failure.

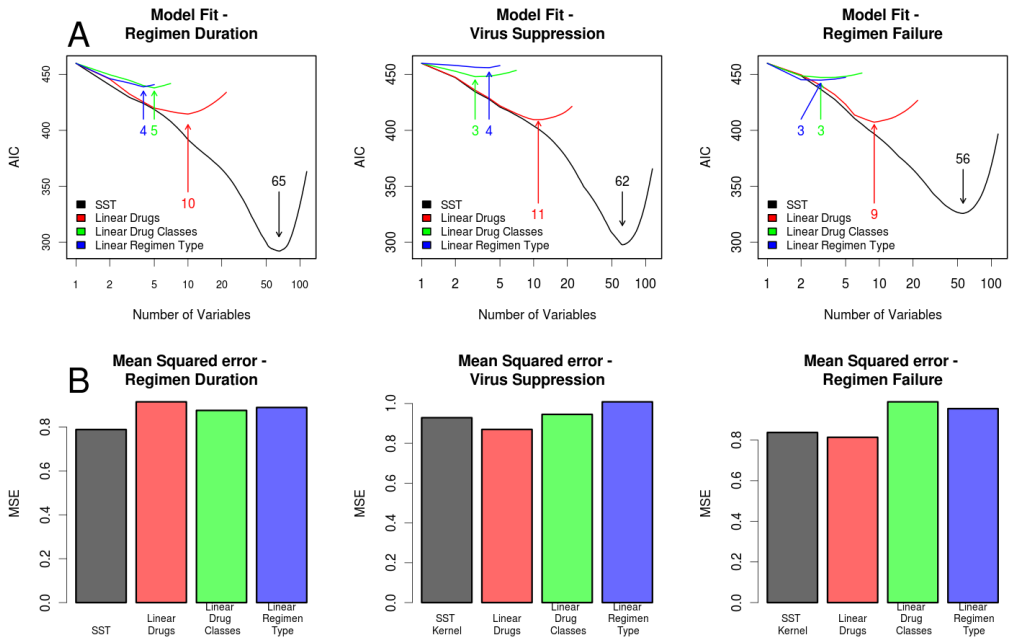


Figure 2.4: Linear model fit, last regimen only. A: Akaike's information criterion of principal-components driven models for regimen duration, virus suppression and regimen failure for individual last records only. Arrows indicate the number of covariates in the best fit (minimum AIC) model. B: Residual mean-squared error for support-vector regression driven models for regimen duration, viral suppression and regimen failure.

kernel provided lower MSE for the regimen failure model. Again, when analysis was restricted to the last regimen per individual, the results were largely consistent (Figure 2.4B), with the exception of the virus suppression model, for which the Linear Drugs SVR had a lower MSE than the SST-kernel model. All other MSE values were consistent between the full and last regimen only datasets. Bias due to multiple measurements per individual does not appear to be the driving force behind the trends observed here.

2.3 Discussion

The SST-kernel is a powerful tool for comparing regimens, and kernel-based analytic tools provide a framework for incorporating this SST-kernel representation into models of observational cohort data. The SST-kernel based KPCA

(Figure 2.2A) highlights the flexibility and information-richness of this approach. The first two SST-kernel principal components, but not the matching linear drug kernel components (Figure 2.2B), exhibit grouping of the regimens into clusters that recapitulate known types of drug combinations. The clusters labeled 1-4 represent drug regimen types in contemporary or recent use. These four regimen clusters account for 95% of the patient-years in the dataset, while only containing 75% of the unique regimens (Figure 2.2C). Notably, the SST kernel and KPCA approach, without being informed of the principles of combination ARV therapy, nor the unique use of ritonavir as a boosting agent amongst protease inhibitors, correctly detected the difference between boosted and unboosted PI-containing regimens, and separated regimens into the main types used in clinical practice. Here the SST kernel provides the "best of both worlds" — correctly suggesting that regimens can be sorted into clusters which share common features, but recognizing that not all regimens within a type are equivalent.

If the SST kernel captures relevant data features missed by simpler encodings, we expect the SST kernel to better account for variation in clinical outcome. To evaluate this hypothesis, we trained two different types of linear models to predict three different measures of clinical outcome — duration of regimen use, virus suppression and virologic failure. These models were informed by either by the SST kernel, or one of three linear kernels. In the first type of model, kernel principal components were extracted, and fitted to clinical outcome data with linear models, using AIC to assess model fit. Second, kernel support-vector regression (KSVR) models were inferred directly from the kernel-transformed data. KSVR model fit was and assessed by cross-validation. The KSVR models estimate a linear predictor in the transformed feature space, and are make no assumption on which projection of the data might be most informative. These two models highlight the flexibility of the kernel approach. The KPCA approach generates continuous covariates that can be incorporated into existing modeling ap-

proaches, while the KSVR approach minimizes modeling assumptions.

In four of the six cases (Figure 2.3), the SST kernel was the best-performing model. Clearly, factors other than regimen composition, such as age, CD4+ lymphocyte count, etc., will influence clinical outcome. However, these factors are not relevant to the comparison of ARV regimen parameterization, and their omission explains the relatively large amount of residual MSE left unexplained in Figure 2.3B. The "linear drug" kernel provided the best performance amongst the linear kernels, and is akin to considering an independent binary variable for each individual ARV drug available. Of particular interest is the relatively poor performance of the "linear regimen type" approach. This approach amounts to sorting regimens into broad bins based on generally-known types of regimens (used here: NRTI+NNRTI, NRTI+boosted PI, NRTI+unboosted PI, NRTI only and Other), and was the worst performer amongst the compared approaches. Stratifying regimens by type is common in the HIV modeling literature [27]. Our results suggest that lumping regimens together into broad *a priori* categories is a poor representation of the heterogeneity in ARV options. This implies that not all regimens within a class are "created equal;" if there were little residual intra-category variation, we would expect the linear regimen type kernel to perform well. Therefore, we expect models predicting response to ARV regimens would be improved by employing the SST-kernel. Similarly, the ability to control for ARV regimen in studies of other independent variables would also be improved by the SST-kernel. The SST kernel supplies a method for representing regimens in models that is substantially more informative than existing approaches.

The SST kernel is not limited to ARV regimen data, and has broad applicability in incorporating structured data, as illustrated by existing uses in the natural language processing field [69]. In fact, the SST kernel has been employed in a strategy to automatically extract protein-protein interaction data from published literature [70]. The SST-kernel could also be used for analysis of data from

high-throughput gene expression technologies, such as microarrays and "deep" sequencing of mRNA which provide large volumes of hierarchical data amenable to application of kernel methods. For instance, the SST-kernel could be used to combine gene expression data with Gene Ontology (GO) annotations for gene function [71]. This would present an information-rich upgrade over existing gene "fingerprint" approaches used, for instance, for characterization of pathogenic microbes, diagnosis of heart failure, and cancer diagnosis [72, 73, 74]. In one approach, gene products in an SST-kernel fingerprint could be organized into a tree first by the cellular compartment, then by biological process, molecular function (all provided by GO annotations), and finally by individual gene identity. This would provide a measure of similarity between fingerprints that not only captures shared genes, but features of similar construction as described by the GO annotations. Just as we have adapted a tool from the language processing literature to analysis of ARV regimen, the approach described here can be applied to a diverse range of structured data problems to construct a useful, information-rich description of complex data.

The HIV dataset in this study spans six clinical sites around the United States and provides a realistic depiction of more than a decade of HIV clinical care. We expect the lessons learned here to be applicable to any observational study where regimen initiation and selection is non-random and changes with evolving standards of care and available ARV agents. Our linear model results suggest that existing approaches of condensing these complex, multi-level data into broad treatment types or a list of individual drugs for analysis can be improved upon. While previous cohort studies have contributed greatly to our understanding of HIV from an observational perspective, the *ad hoc* representations of ARV regimen may miss or obscure relevant features in the data. In this study we propose a powerful approach to encode the richness of this information into a computationally convenient and statistically powerful set of features

defined using a subset tree (SST) kernel.

While the SST kernel-based methods provide an information-rich approach for characterizing ARV regimen, it increases computational cost, and may be more difficult to interpret than traditional approaches. Non-linear kernel transformations obscure the relationship between input variables and principal components or support vectors. In the models presented here, the features of the SST kernel have the strongest associations with clinical features of interest in 4 of 6 models, but the relative contributions of individual regimen features cannot be easily determined. In contrast, the linear kernels (which correspond to generally-used ARV regimen descriptions) did not associate as well with clinical outcome, although the relative contributions of specific regimen features to the models could be extracted. For the SST-kernel to be of utility in other data sets, this additional computational challenge and difficulty in interpretation must be justified by its superior predictive power. The AIC-based and cross-validation tests used here to test our linear models provide two approaches to making this assessment.

The SST-kernel approach described here provides a mathematically powerful, yet practical approach to describing complex data. Researchers possess a great deal of knowledge about ARV regimens, but have been forced resort to *ad hoc* summaries for analytic convenience. Using kernel methods has allowed us to capture much of the available information about ARV regimen in a streamlined framework. Techniques like KPCA and KSVR allow us to predict clinical outcome without making *a priori* assumptions about what features of drug regimen may be important. A similar approach may prove fruitful for any data amenable to this sort of hierarchical structure. Other complex data types specific to the HIV field, such as HIV drug resistance genotype and phenotype tests, could be similarly categorized and described in a manner more informative than a list alone could offer. Here we describe an SST-kernel approach to describing ARV drug

combinations that is more informative, and a better correlate of outcome than existing approaches.

2.4 Methods Summary

Data Preparation

Antiretroviral (ARV) drug use and HIV-1 plasma RNA viral load (VL) records were retrieved from the CFAR Network of Integrated Clinical Systems (CNICS) data repository, described previously [6]. We retrieved 38,953 records of ARV regimen use from the CNICS data repository collected between 1996 and 2009. Records were carefully filtered for quality assurance. Some drug regimens temporarily omitted a single drug, only to have it subsequently reinstated, and were deemed more likely bookkeeping errors than actual drug omissions. Such regimens were treated as if therapy was uninterrupted, removing 491 records. Other short interruptions between identical regimens (≤ 7 days) were disregarded, and treated as if therapy were uninterrupted (671 records removed). 3,003 treatments of 7 or fewer days in duration were disregarded. 5,062 remaining regimens were marked as suspicious and also discarded. "Suspicious" regimens met one or more of the following criteria: single- or dual-drug therapy after 1997, more than one NNRTI, more than 4 NRTI, more than 2 non-ritonavir PI, reported exposure more than 1 year prior to licensure, simultaneous use of stavudine and zidovudine and simultaneous use of emtricitabine and lamivudine. These cases included regimens inappropriate to their labeled date (single- or dual- drug regimens prescribed after the broad availability of combination ART, drugs prescribed before availability), formulations not typically used, but easily constructed by inadvertent combination of legitimate regimens (excessive use of drugs from a single class in one regimen), and regimens containing drugs too similar to be productive together in the same regimen (stavudine/zidovudine and emtricita-

bine/lamivudine). Finally, 4,218 records collected retrospectively by subject recollection were removed. After these filters, 25,508 records remained.

For each unique drug combination associated with five or more viral load (VL) records, three correlates of regimen performance were calculated: median duration of regimen use, percent of regimen instances that achieved virus suppression, and percent of regimen instances that exhibited virologic failure. Virus suppression was defined as at least one VL measurement $<$ 200 copies/mL assessed more than 7 days after starting therapy. Virologic failure was defined as at least one on-therapy VL measurement $\geq$ 200 copies/mL assessed more than 24 weeks after starting therapy. The final dataset included 19,129 records, representing 393 distinct regimens prescribed to 9,184 unique individuals, allowing more than one record per individual to be included in the dataset.

Kernel Design

Various kernels were used to calculate sets of pairwise scalar products (or similarity scores) between qualifying regimens. Features of drug regimen composition were encoded using four different kernels for comparison: the subset-tree (SST) kernel and three linear kernels. To calculate the SST kernel, regimens were first represented as trees, incorporating information from each level of the target protein/drug class/drug composition hierarchy. (Figure 2.1, Table 2.2). Each individual regimen tree begins at the root node labeled "ARV." Subsequent branching points in the tree represent particular aspects of the regimen, specifically: protein target, drug class and individual drug name, yielding a tree of depth four (i.e. four levels of edges between nodes) for each treatment, with each leaf corresponding uniquely to an individual drug. The SST kernel function evaluates a similarity score for two trees by counting shared subset trees weighted by a decay factor (λ). The minimum score is zero (for example for an NRTI monotherapy vs PI monotherapy), and the maximum score is the weighted total of all subset

trees, and will occur only when applying the kernel function to two identical regimens. The kernel score for regimen pair (X,Y) will always be less than the kernel score for regimen X with itself (and also regimen Y with itself), i.e. no regimen will be scored as more similar to another than to itself. Note that regimens with more drugs/branches correspond to a greater maximum score than those with few drugs/branches. An efficient algorithm for computing the SST kernel function has been described elsewhere [68], and is summarized briefly here. The kernel score, $\kappa(X_a, X_b)$, between two regimen trees X_a and X_b is defined as:

$$\kappa(X_a, X_b) = \sum_{n_i \in X_a} \sum_{n_j \in X_b} \Delta(n_i, n_j). \tag{2.1}$$

Here, the kernel function (2.1) is the sum of a function Δ over all pairs of nodes in the two trees, where $n_i \in X_a$ denotes the set of all nodes in tree X_a. The contribution of each particular node pair is given by:

$$\Delta(n_i, n_j) = \begin{cases} 0 & \text{if } c(n_i) \neq c(n_j) \text{ or } c(n_i) = c(n_j) = \varnothing \\ \lambda \prod_{k=1}^{nc(n_1)} (1 + \Delta(c_i^k, c_i^k)) & \text{otherwise} \end{cases} \tag{2.2}$$

where $c(n_i)$ are the children of node n_i, $nc(n_i)$ is the number of children node n_i possesses, and c_i^k is the k^{th} child of node n_i. This function has the value zero if the nodes have different sets of children, or if one is a "leaf" nodes without children (ie. $c(n_i) = \varnothing$). The value λ is a decay factor that prevents contributions from nodes high in the tree from overwhelming the kernel score. This algorithm was implemented here using a custom Python script with $\lambda = 0.55$. An optimum value for λ was determined by grid search from the data. Principal-components based linear models were fit as described below for various values of λ, and the value giving the lowest total mean-squared error was used subsequently.

Linear kernels were also used as comparisons for representing drug regimen. The linear kernel is simply calculated by taking the inner product between the input data vectors. Here we used three different sets of input data in linear kernels for comparison with the SST approach. The "linear drug" kernel repre-

sents regimens as a vector of binary values that each indicate the presence or absence of an individual drug and is equivalent to characterizing regimens as a drug list. The "linear class" kernel represents regimens as a vector of drug counts for each drug class (NRTI, NNRTI, PI, etc.) and describes regimen composition, but does not distinguish between different drugs within a class. The "linear regimen" type kernel classifies regimens into types specified *a priori*. Here the regimen types are 1) NRTI + NNRTI, 2) NRTI + ritonavir-boosted PI, 3) NRTI + unboosted PI, 4) NRTI only, and 5) Other. These kernel types are summarized in Table 2.2.

Kernel Principal Components

Kernel principal components analysis (KPCA) was applied to each kernel matrix in R using *kernlab* [75, 76]. Briefly, KPCA calculates the eigenvalues and normalized eigenvector of the variance-covariance matrix in feature space. Data points in feature space can be projected on any single principal component to yield a KPCA coordinate. The magnitude of each eigenvalue is proportional to the amount of data variance explained by the matching projection. Eigenvalues below a certain threshold are generally disregarded as negligible (here 0.0001). KPCA can be carried out without the knowledge of actual feature values, but only by pairwise scalar products in the feature space, i.e. the kernel matrix.

Linear Models for Regimen Performance

Two modeling approaches were employed — linear models fit with kernel principal components (KPC), and support-vector regression models (SVR) inferred directly from the kernel matrices. The KPC models were fit in R using the *lm* function. Variables were selected using a stepwise forward procedure. Beginning with an empty model, variables that give the lowest model AIC were added one at a time. AIC was calculated as $2k - 2\ln(L)$ where k is the number of variables, and L is the model partial likelihood.

SVR models were fit using epsilon-support vector regression using R and the *kernlab* package. The regularization parameters C and ϵ were fit empirically from the data for each model to minimize residual error in 10-fold cross-validation sets. Model fit was assessed by mean-squared error in 10-fold cross validation.

As individual subjects were allowed to contribute multiple records to the dataset, we constructed a parallel dataset consisting of only the last regimen for each individual. The linear models were re-fit on this data to determine if non-independence of records contributed a substantial bias to the full dataset analysis.

Chapter 2, in full is currently being prepared for submission for publication of the material. Lorne W. Walker, Sergei L. Kosakovsky Pond, Jeannette L. Aldous, Art F. Y. Poon, Shelly Sun, Sonia Jain, Simon D. W. Frost, Richard H. Haubrich. The dissertation author was the primary author of this work.

Chapter 3: Subset-tree Kernel-based Models for Regimen Failure

3.1 Introduction

Modern ARV therapy provides a broad set of therapeutic choices, encompassing multiple drug classes which target distinct steps in the viral life cycle (reviewed in chapters 1 and 2). In the face of regimen failure, in which viral replication rebounds despite ongoing treatment, the clinician's challenge is to select a new drug regimen with a maximum probability of providing lasting suppression of viral replication. This presents a complex puzzle, as viral adaptation to previously failed therapies limits the range of therapeutic choices, and regimen failure is a maker of poorer prognosis [77, 27]. Furthermore, viral adaptations which provide cross-resistance within a drug class may further narrow the field of available drugs [78]. The very large number of potential drug combination options make it impossible to test every potential drug regimen independently. The subset-tree (SST) kernel developed in the previous chapter provides a ready approach for describing the relationships between these options, and may provide insight helpful in optimizing further therapy.

Much effort has been made in the development of tools for aiding in selection of drug regimen. *In vitro* systems allow the direct measurement of drug susceptibility phenotype, and several approaches exist for predicting this phenotype

from the viral genotype [79, 80, 81]. Both genotypic and phenotypic tests have been shown to have predictive value for the initial response to subsequent regimens [82, 83]. More complex mathematical models, employing machine learning tools have been developed to improve these predictions, and include the use of artificial neural networks, random forests and kernel methods, among others [84, 85, 86, 87]. Some of these approaches have been implemented as clinician-oriented tools, and may prove to have real clinical utility [88, 89].

Most of the predictive models developed to date focus on viral genotype predictions of response — defined as a treatment-induced drop in plasma viral RNA concentration (viral load or VL) to some set threshold — based on viral genotype. This is of obvious clinical utility, as it measures the initial success or failure of a new therapy. However, the ultimate goal of treatment is to induce a durable suppression of viral replication and prevent or delay disease progression. In this sense, our interest lies not only in predicting which drug combinations will halt the virus, but also which are most likely to provide lasting suppression.

Furthermore, while the viral genotype is clearly a powerful clue to elucidating viral susceptibility, it may only be available in certain economic or clinical situations. While less resource-intensive than phenotypic testing, viral genotyping is beyond the reach of clinicians in resource-limited settings [90]. The genotype test may also miss rare drug-resistant variants, particularly if not timed correctly. Treatment experienced patients may harbor low-abundance drug-resistance variants that correlate with previous ARV use, but are not typically detected by a genotype test [91]. Furthermore, even in settings where a drug-resistant virus is the dominant variant, “wild-type” (generally non-resistant) virus re-emerges when treatment is stopped [92]. Therefore, if a genotype test is not performed prior to cessation of a failed therapy, results may not be representative of the underlying resistance phenotype. While genotype remains a powerful clinical tool, the information provided may be incomplete, or unavailable due to economic or

Table 3.1: Dataset description, failure models TCE: Treatment change event, (N)NRTI: (Non) Nucleot(s)ide reverse transcriptase inhibitor, PI: Protease inhibitor. UCSD: Univ. of California San Diego, UW: Univ. of Washington, UAB: Univ of Alabama Birmingham, CWRU: Case Western Reserve University, UCSF: Univ. of California: San Francisco. MSM: Men who have sex with men, IDU: Injection drug users.

Sample Size	**N**	**Demographic/Risk Factor**	**Percent**
Total TCE	23,324	Male Gender	79
Failure Events	3,036 (13%)	White, Non-Hispanic	47
Unique Individuals	11,537	Black	40
Unique Drug Regimens	1,645	Hispanic	10
		Other Race/Ethnicity	4
Regimen Types (N = 10,367)	N (Percent)	MSM	57
NRTI + NNRTI	7,919 (34)	IDU	20
NRTI + PI (Boosted)	8,470 (36)		
NRTI + PI (Unboosted)	2,817 (12)	Demographic/Risk Factor	Median [IQR]
NRTI Only	691 (3)	Age	42 [36-48]
Other	3427 (15)	Year of TCE	2004 [2002-2007]
		Follow-Up Duration (days)	388 [182-826]
Study Site	Percent		
JHU	31	Previous Exposure	Percent
UW	16	Previous Virus Suppression	57
UCSD	16	Previous Regimen Failure	21
UAB	9		
UNC	8		
CWRU	7		
UCSF	7		
Fenway	6		

clinical constraints.

In light of this, we aim to develop a mathematical approach to estimation of virologic failure risk informed by current drug regimen, past drug regimen, and other clinical or demographic covariates. In conjunction with existing work predicting initial viral response to treatment, understanding risk of virologic failure may help to better optimize ARV therapy, particularly in highly treatment experienced patients. A recent publication by Robbins, et. al. proposes a model for predicting 1-year virologic failure based on seven clinical variables: suboptimal adherence, CD4+ lymphocyte count, drug and/or alcohol abuse, high ART-experience, missing one or more appointment, prior virologic failure and viral suppression less than 12 months [93]. This provides a good starting example for this work, however several of the variables (adherence, clinic attendance, drug and alcohol abuse) may not be available in large cohort datasets, so we aim to tailor our approach to the data available in the CNICS cohort study. In Chapter 2

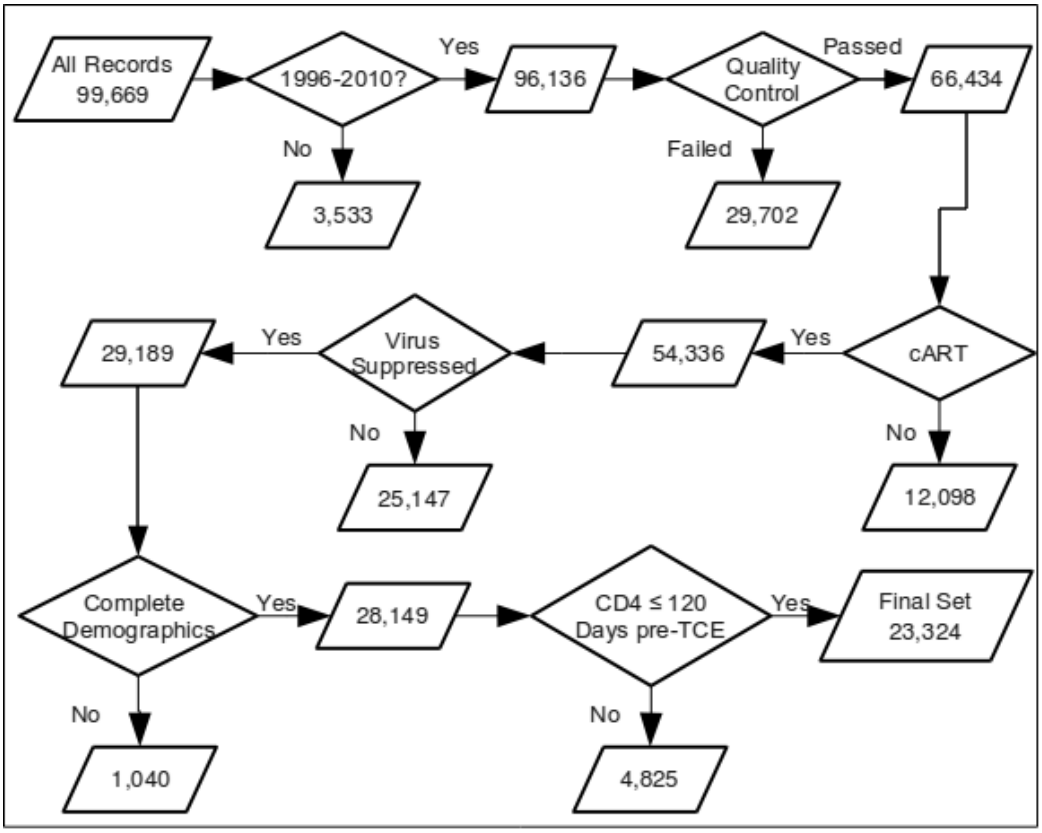


Figure 3.1: Data preparation flowchart — 23,324 treatment change event records were filtered from an original set of 99,669 drug regimen records. cART: combination antiretroviral therapy

we developed the notion that the SST-kernel is a more information-rich method for capturing the complexities of ARV regimen design. We also wish to determine if this tool improves our ability to mathematically describe current and previous ARV use, and therefore better predict failure. Ideally, a thorough characterization of current and previous ARV exposure may complement the information gained from genotypic testing, or suffice when it is unavailable.

3.2 Results

Data Description

A total of 23,324 treatment change events (TCEs) — defined as initialization of or change in ARV therapy — were selected for failure analysis from the

Table 3.2: Regimen failure model descriptions — Alternative formulations for predictive models of regimen failure. FTC/TDF/EFV: examples of how the durg combination emtricitabine/tenofovir/efavirenz is parameterized in each case.

Current Regimen Strategy	**Regimen Description**	**FTC/TDF/EFV**
SST-KPCA	Top 10 SST-kernel Principal components	See Chapter 2, Figure 3.1A
Drug Counts	Count drugs per class	1 NNRTI, 2 NRTI
Combined	SST-KPCA and Drug Counts combined	
Other Covariates (All Models)	Age, gender, race/ethnicity, MSM risk group, IDU risk group, previous regimen failure, study site	

CNICS cohort data repository. Regimens were selected from the years 1996-2010, passed quality control checks, met a general description of modern combination ARV therapy (cART), and succeeded in viral suppression (VL $\leq$ 200 copies/mL more than 7 days post TCE) (See Methods, Figure 3.1). These records were drawn from 11,537 unique individuals, and included 1,645 distinct antiretroviral drug combinations. Of these, 3,036 (13%) experienced virologic failure, defined as a VL >200 more than 24 weeks post-TCE, confirmed by a second failing VL or subsequent TCE within 120 days. The non-failing records were censored when non-failing therapy was discontinued or changed, the patient was lost to follow-up, or at the end of follow-up for this analysis (December 2010). This population was distributed amongst eight study sites in the CNICS cohort. Forty-seven precent of subjects were white, most (79%) were male, and the most common risk factor was men having sex with men (MSM, 57%) (Table 3.1).

Of the 1,645 distinct ARV drug combinations, the most commonly observed regimen types were NNRTI+NRTI (defined as $\geq$ 2 NRTI and $\geq$ 1 NNRTI) and bioavailability-boosted NRTI+PI (defined as $\geq$ 2 NRTI and $\geq$ 1 non-ritonavir PI plus ritonavir), while unboosted NRTI+PI ($\geq$ 2 NRTI and $\geq$ 1 non-ritonavir PI or ritonavir alone) and NRTI only ($\geq$ 3 NRTI) were less common. Other regimens comprised of non-standard formulations, or containing recently-approved agents (raltegravir, enfuvirtide, maraviroc) accounted for 15% of TCE records (Table 3.1). There were a small number of very common regimens in this dataset, and a large number of very rare regimens (Figure 3.2). The most common regi-

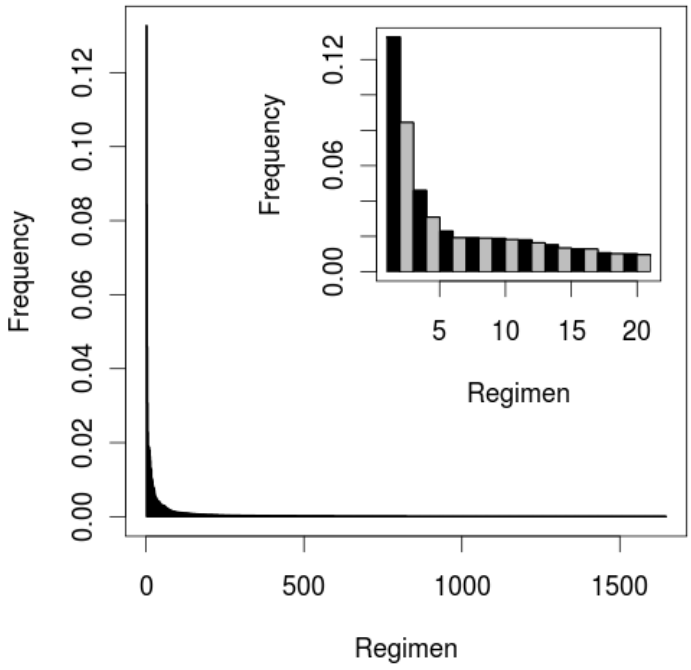


Figure 3.2: Regimen frequency distribution. A: Frequency of the 904 regimens in the filtered dataset. Inset: magnified view of the 20 most common regimens, which account for 61% of records.

men, emtricitabine/tenofovir/efavirenz, often prescribed as the single-pill combination Atripla, accounted for 16% of all TCE records, while the top 5 regimens account for almost 40% of all TCEs, and the top 20, 61% of all TCEs (Figure 3.2 inset).

Regimen Failure Model — Kernel Representation of Current Regimen

Relative risk of experiencing post-suppression virologic failure was estimated for the TCEs in the dataset using the consensus of 10-fold cross-validation fit Cox proportional hazards models (see Methods). All models included demographic, risk factor and clinical covariates, including age, gender, race/ethnicity, risk group (MSM or IDU), previous regimen failure and study site. The subset-tree kernel principal components anlaysis (SSTK-PCA) model parameterizes the current drug regimen in use as the first 10 principal components (PC) of the SST kernel (Chapter 2). The “Drug Counts” model simply describes current regimen

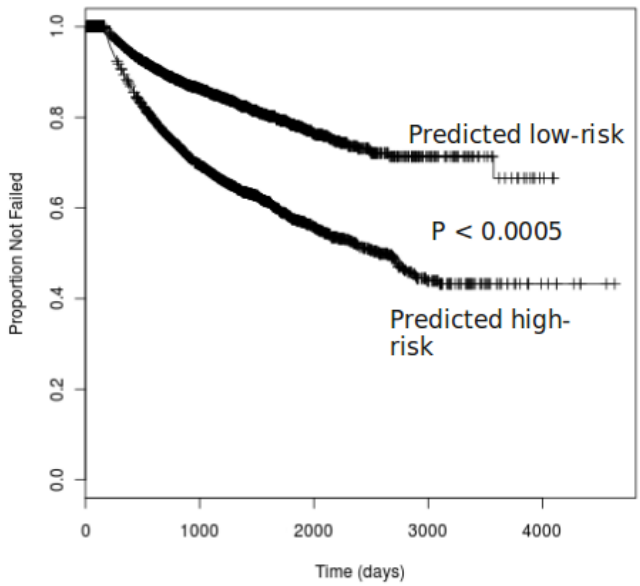


Figure 3.3: Representative Kaplan-Meier curve for TCEs predicted at above or below average risk for virologic failure. Predictions from Combined model.

as a count of the drugs in each included drug class (ie. the number of NNRTI, PI, etc.), and the "Combined" model combines both approaches (Table 3.2). As not all covariates may be good predictors of failure, parsimonious models were selected by searching for sub-models minimizing the Akaike information criterion (AIC), a model selection tool maximizing model likelihood while penalizing for excessive parameterization. The final models are summarized in Table 3.3. Dividing the data into high- and low-risk groups based on model predictions provides statistically significant differences in failure hazard rate (Figure 3.3).

The AIC model selection algorithm selected 6 of 10 SST PCs for the final SSTK-PCA model, while counts of NNRTI, NRTI, PI and fusion inhibitor (FI) drugs were retained in the Drug Counts model. When the model selection algorithm had both kernel and drug-count options to choose among, it retained 6 of 10 SST principal-components (overlapping some but not all of the components selected in the SSTK-PCA model), and counts of NNRTI, NRTI, PI, FI, and integrase inhibitors (II). As of 2010, the II and FI classes each only contained one drug (ral-

tegravir and maraviroc, respectively).

In the Drug Count and Combined models, NNRTI count significantly decreased hazard rate, while NRTI count increased the risk of failure (Table 3.3). This may reflect the sub-optimality of NRTI-only therapy which can be rescued by combination with an NNRTI. On the other hand, PIs are frequently used in combination with NRTIs, but also appear to increase risk of failure. The mixture of boosted and un-boosted PI-containing regimens in this dataset may partially account for this.

The SST principal components included in the SSTK-PCA and Combined model do not have a convenient interpretation, but represent the diversity of regimens in the dataset. It is interesting to note that when combined with the drug count data, a different set of components was selected, and covariate effects differed between the two models. This suggests that the two regimen representations, SSTK-PCA and Drug Counts, have contain overlapping, but non-identical information. If the kernel-based and drug count variables were entirely independent, they would not interact in this manner. These differences indicate that the PCs selected for the Combined model may be only significant when controlling for NRTI, NNRTI, II and FI counts but perhaps not otherwise. This can be most easily illustrated in the case of PC1. In the SSTK-PC model, increased PC1 indicates lower risk of failure, but in the Combined model PC1 increases failure risk. PC1 appears to act as a trade-off variable between NNRTI and PI-containing regimens. Positive values of PC1 indicate NNRTI use, while negative values of PC1 indicate PI use (P $<2.2\text{x}10^{-6}$ for correlation between PC1 and NNRTI count or PI count by Pearson's product moment correlation)(Figure 3.4A). Therefore, while controlling for NNRTI count, low PC1 likely acts as an indicator of PI use, and in this case is protective rather than hazardous. We expect other, possibly more subtle relationships to exist between the principal components and other variables. Ideally, this nuanced — albeit difficult to interpret — parameterization

Table 3.3: Hazard rate ratios for failure model parameters — 95% Confidence intervals presented in brackets, covariates with P <0.05 in bold. NS: Not Selected, Hyphen denotes covariate not available in this model, Ref: reference value for categorical variables. II: Integrase inhibitors, FI: Fusion inhibitors, EI: Entry inhibitors

Covariate	**SST-KPCA**	**Drug Counts**	**Combined**
Gender	NS	NS	NS
Age	**0.99 [0.98-0.99]**	**0.98 [0.98-0.99]**	**0.99 [0.98-0.99]**
Race/Ethnicity = Black	Ref	Ref	Ref
= Hispanic	**0.79 [0.69-0.91]**	**0.78 [0.68-0.89]**	**0.79 [0.69-0.91]**
= White	**0.74 [0.68-0.80]**	**0.75 [0.69-0.81]**	**0.73 [0.67-0.80]**
= Other	**0.63 [0.51-0.78]**	**0.62 [0.50-0.76]**	**0.64 [0.52-0.79]**
Study Site = CWRU	Ref	Ref	Ref
= Fenway	**0.72 [0.60-0.87]**	**0.70 [0.58-0.85]**	**0.73 [0.60-0.89]**
= Hopkins	0.89 [0.77-1.02]	0.91 [0.79-1.05]	0.87 [0.76-1.00]
= UAB	0.99 [0.86-1.14]	1.04 [0.91-1.20]	1.00 [0.87-1.16]
= UCSD	0.91 [0.79-1.05]	**0.83 [0.72-0.96]**	0.91 [0.79-1.05]
= UCSF	1.00 [0.84-1.19]	0.91 [0.76-1.08]	1.01 [0.85-1.20]
= UNC	**0.67 [0.57-0.79]**	**0.64 [0.55-0.75]**	**0.68 [0.58-0.79]**
= UW	**0.838 [0.73-0.96]**	**0.841 [0.73-0.97]**	**0.821 [0.72-0.94]**
MSM	NS	NS	NS
IDU	**1.18 [1.07-1.30]**	**1.21 [1.10-1.33]**	**1.20 [1.09-1.32]**
Previous Failure	**1.64 [1.47-1.82]**	**1.60 [1.43-1.77]**	**1.59 [1.43-1.78]**
Previous Suppression	**0.64 [0.59-0.70]**	**0.62 [0.56-0.68]**	**0.64 [0.59-0.71]**
SST PC1	**0.84 [0.80-0.88]**	—	**1.47 [1.14-1.88]**
SST PC2	**0.92 [0.86-0.97]**	—	**0.70 [0.63-0.79]**
SST PC3	NS	—	NS
SST PC4	**0.92 [0.86-0.97]**	—	NS
SST PC5	**1.08 [1.04-1.12]**	—	NS
SST PC6	NS	—	NS
SST PC7	**0.87 [0.84-0.90]**	—	**0.87 [0.84-0.91]**
SST PC8	**1.27 [1.21-1.33]**	—	**1.08 [1.00-1.16]**
SST PC9	NS	—	**0.82 [0.77-0.88]**
SST PC10	NS	—	**0.95 [0.92-0.99]**
# NNRTI	—	**0.66 [0.59-0.74]**	**0.49 [0.29-0.83]**
# NRTI	—	**1.13 [1.06-1.22]**	**1.25 [1.16-1.34]**
# PI	—	**1.08 [1.02-1.15]**	**1.17 [1.10-1.25]**
# II	—	NS	**3.12 [2.00-4.89]**
# FI	—	**1.93 [1.47-2.53]**	**5.30 [3.34-8.39]**
# EI	—	NS	NS

of drug regimen will improve modeling of the relationship between ARV therapy and clinical response.

Several methods were used to assess the regimen-failure models developed here. AIC, which was also used to select covariates, can be used to compare different models of the same data, as is done here. The Combined model gave lowest AIC value (indicating the best fit of the options), with the SSTK-PCA 62 points higher, and the Drug Count model 162 points higher yet. AIC differences above 10 points are considered to confer "almost none" support for the less desirable model [94]. During model selection, the data was divided into 10 equal cross-validation sets, providing "test set" hazard ratio estimates for each datapoint. These model estimates were used to inform a receiver operating characteristic (ROC) curve for each model predicting regimen failure by 1 year (Figure 3.4B). Model performance can be assessed by estimating the area under the ROC curve (AUC), where higher area indicates better model predictions. Consistent with the AIC data, the Combined model gave the best results, followed by the SST-KPC model (Figure 3.4C). Finally, we calculated partial correlations between the final model and the recorded virologic failures, which estimate the proportion of the total residual sum of squares (RSS) accounted for by the models [95]. The RSS explained by the full models, as well as by the current-regimen covariates alone was calculated. As with the other model diagnostics, the Combination model explained the largest amount of data variation, with the SSTK-PCA model out-performing the Drug Counts model.

Regimen Failure Model — Kernel Representation of Previous Regimen

We also expected that prior exposure to antiretroviral drugs would effect the probability of experiencing subsequent failure. Several methods for representing past ARV exposure in models for virologic failure were evaluated. Three kernel-based methods were assessed, called "Recent Kernel," "Mean Kernel" and

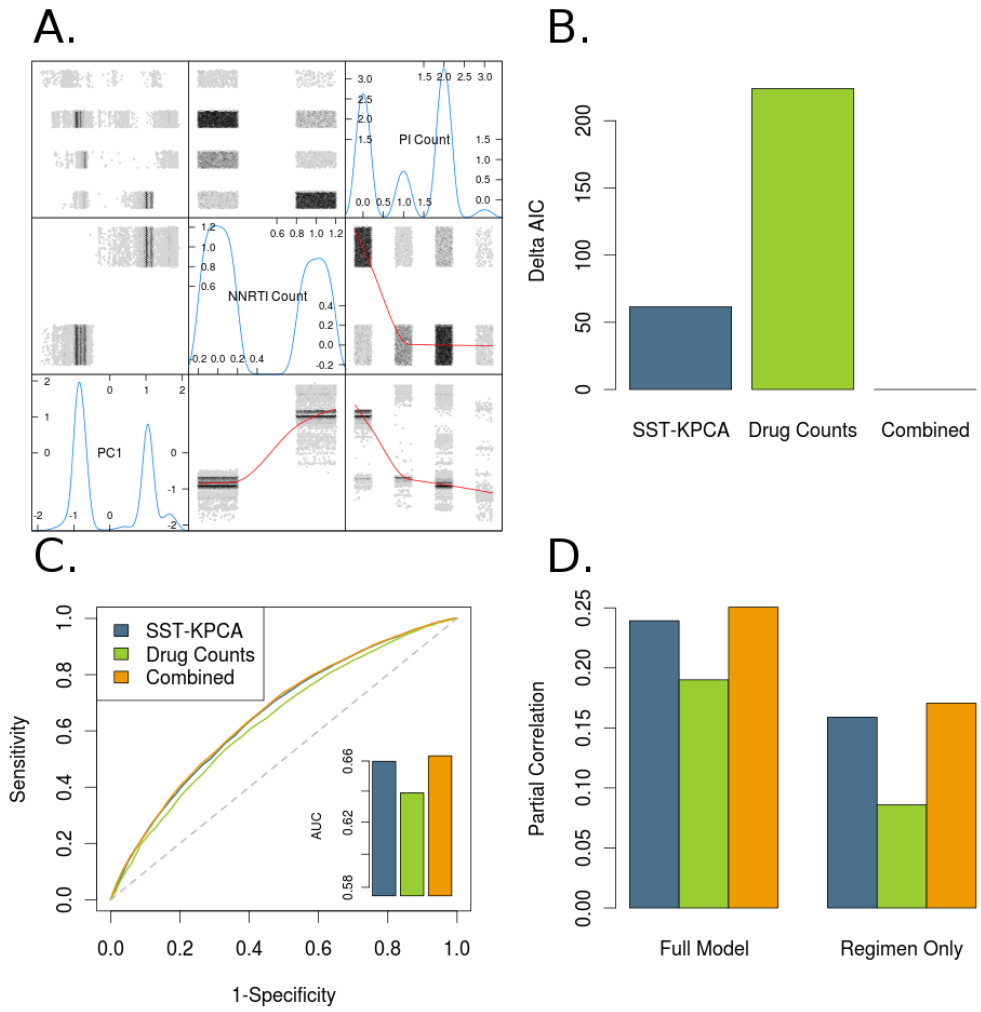


Figure 3.4: Regimen failure model diagnostics — A. Correlation plots between Principal component 1 (PC1), and the number of PIs and NNRTIs in a regimen. Individual variable density estimates are shown by blue lines on the diagonal, jitter has been added to the adjoining density plot to separate identical data points. Red lines in the lower half of the chart are loess trendlines for the relationships between variables. B. Delta AIC: difference between model AIC and minimum AIC. C. ROC curve, inset figure area under AUC curve. D. Partial correlation between linear predictor for full model, or the current regimen data only.

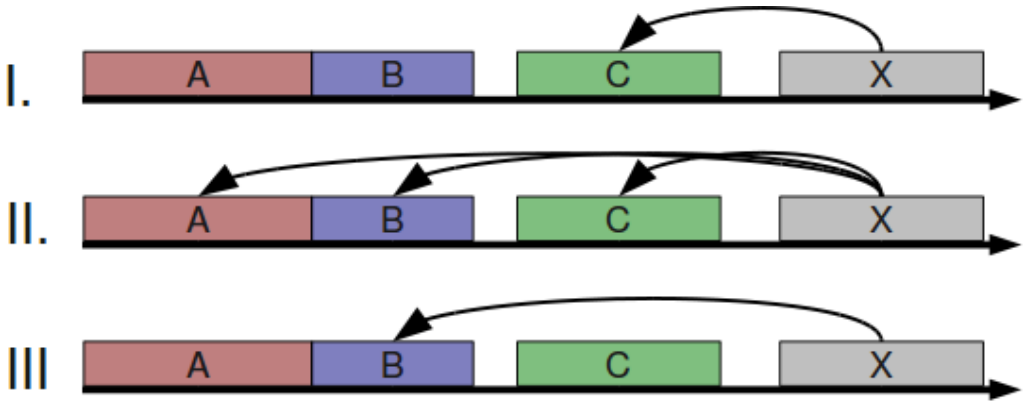


Figure 3.5: Approaches to description of previous ARV regimens. I. Most Recent Kernel Score: kernel (similarity) score between current regimen (X) and most recent regimen(C). II. Mean Kernel Score: average of similarity between current and all previous regimens. III. Maximum Kernel Score: maximum of all pairwise similarity scores between current and previous regimens.

"Maximum Kernel." These measures of similarity between current and previous regimens are defined as the kernel score between the most recent regimen and the current one, or the mean or maximum of kernel scores between all previous regimens and the current regimen, respectively (Figure 3.5, Table 3.4). For comparison, counts of previously-used drugs per drug class (NRTI, NNRTI, PI, etc.) were used, as well as combinations between the drug counts and kernel scores. cross-validation AIC-based model selection was used to fit models and construct consensus final models (see Methods).

The final models for the Recent, Mean and Maximum kernel approaches all gave hazard rate ratios between 1.2-1.4, demonstrating an elevated risk of virologic failure when current regimen is similar to past regimen. Non kernel-models also showed that extensive previous ARV-exposure increased the risk of subsequent virologic failure. In each model where drug count variables were also available to the model-selection algorithm, the kernel scores longer had statistically significant associations with virologic failure at the α <0.05 level. Final models were assessed by AIC and ROC-AUC. All three kernel-only models had higher AIC

Table 3.4: Regimen failure models formulations, previous ARV regimen — alternative model formulations for accounting for past antiretroviral regimen use. Examples based on hypothetical sequence of regimens {A,B,C,X}. κ(x,y) denotes the kernel (similarity) score between dataponts x and y. See Figure 3.5.

	Previous Regimens Strategy	**Description of Previous Regimen(s)**	**Example: Current X, past A, B, C**
1.	Most Recent Kernel Score	Kernel score, most recent regimen	κ(X,C)
2.	Mean Kernel Score	Mean of kernel scores, all regimens	1/3 * (κ(X,A), κ(X,B), κ(X,C))
3.	Maximum Kernel Score	Maximum kernel score, all regimens	max((κ(X,A), κ(X,B), κ(X,C))
4.	Drug Class Counts	Count of previous NRTI, NNRTI, PI, etc.	5NRTI, 2PI , 1NNRTI
5.	Combine Counts, Recent Kernel	Combine drug counts, most recent kernel (rows 1, 4)	
6.	Combine Counts Mean Kernel	Combine drug counts, mean kernel (rows 2, 4)	
7.	Combine Counts, Max Kernel	Combine drug counts, max kernel (rows 3, 4)	
8.	Combined All	Combine all previous regimen variables (rows 1-4)	

than the non-kernel models. The tree kernel-only models had similar AIC values to each other. However, when compared by ROC-AUC all models evaluated had very similar predictive value (Figure 3.6).

3.3 Discussion

The models presented here support the notion that the SST-kernel provides a practical, information-rich approach to improving modeling of regimen failure. All of the model diagnostics performed concur that the SST-kernel encoding provided improved model fit and predictions over a conventional representation of current ARV regimen. This model was further improved upon by allowing the model selection algorithm to choose among both conventional and kernel-based variables. Similarly, we evaluated the use of kernel-based similarity scores between current and previous regimen as predictors of failure. The results for these past-regimen kernel models were more equivocal, with non-kernel models scoring better by AIC, but having very similar predictive performance on cross-validation data (Figure 3.6).

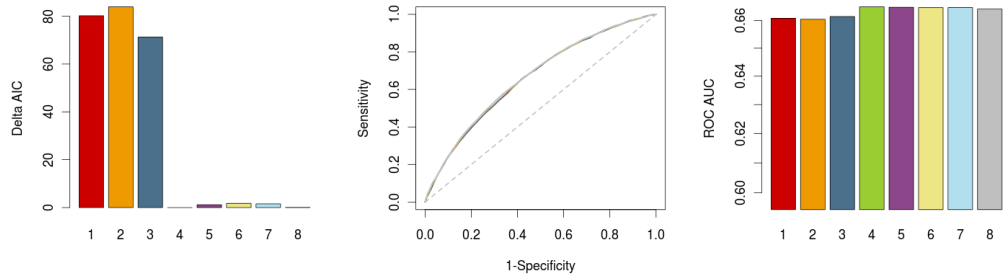


Figure 3.6: Regimen failure model diagnostics, different accountings for past regimen. A. Difference between model AIC and minimum among considered models. B. ROC curve for considered models. C. Area under ROC curves for considered models. 1: Most Recent Kernel Score, 2: Mean Kernel Score, 3: Maximum Kernel Score, 4: Drug Class Counts, 5-7: Combine Drug Class Counts and Recent, Mean and Maximum Kernel Score respectively, 8: All Combined

These models also provide a general portrait of the features associated with regimen failure. There were variations among the various study sites, with cohort participants from different cities having different failure rates. Potential differences in cohort characteristics across the sites, including duration of infection and socioeconomic status may account for some of these differences. Other patient sub-populations had elevated risk for failure, including African Americans and injected drug users. Inclusion of newer ARV drugs (integrase, fusion and entry inhibitors) in current regimen was associated with higher failure rates. For this study period, these agents were largely used as salvage therapies for highly-ARV experienced patients. The high failure rate observed with these new agents may be due to their used as fall-back options in particularly high-risk individuals. Previous outcomes on therapy were particularly powerful predictors of future outcome. Suppression of the virus by a previous cART regimen afforded a 36% lower failure rate, while previous virologic failure was associated with a 59% higher failure rate. These effects may stem from unmeasured individual-level differences, such as general health quality, variations in viral virulence or immune parameters (such as MHC type) and degree of reliability in taking medication as prescribed (Table 3.4).

Table 3.5: Hazard Rate Ratios for covariates of previous regimen history. Model parameters with P <0.05 for significant effect on hazard rate printed in bold. Hyphen denotes covariates not available in this model. NS: not selected. 95% confidence intervals in brackets.

Covariate	Recent Kernel	Mean Kernel	Maximum Kernel	Drug Class Counts	Count + Recent	Count + Mean	Count + Max	Combined All
Recent Score	**1.26 [1.11-1.42]**	—	—	—	0.95 [0.82-1.09]	—	—	NS
Mean Score	—	**1.26 [1.09-1.46]**	—	—	—	1.03 [0.88-1.20]	—	1.03 [0.88-1.20]
Max Score	—	—	**1.36 [1.20-1.54]**	—	—	—	0.96 [0.83-1.12]	NS
# NRTI	—	—	—	**1.10 [1.07-1.14]**	**1.11 [1.07-1.14]**	**1.10 [1.07-1.14]**	**1.11 [1.07-1.15]**	**1.10 [1.07-1.14]**
# NNRTI	—	—	—	NS	NS	NS	NS	NS
# PI	—	—	—	**1.06 [1.02-1.10]**	**1.06 [1.02-1.10]**	**1.06 [1.02-1.10]**	**1.06 [1.02-1.10]**	**1.06 [1.02-1.10]**
# II	—	—	—	NS	NS	NS	NS	NS
# EI	—	—	—	**2.61 [1.07-6.40]**	**2.61 [1.07-6.40]**	**2.61 1.07-6.41]**	**2.60 [1.06-6.39]**	**2.61 [1.07-6.41]**
# FI	—	—	—	**1.44 [1.06-1.96]**	**1.44 [1.06-1.95]**	**1.45 [1.07-1.96]**	**1.44 [1.06-1.95]**	**1.45 [1.07-1.96]**

There are also disadvantages to the SST-kernel methods outlined here. Computation of the kernel scores, and extraction of principal components from the kernel matrix can be be computationally expensive. The models outlined in this chapter could be run on a contemporary desktop computer (4GB RAM, Quad-core 2.0GHz processor), although data preparation, kernel computations and model estimations required more than an hour to complete. Interpretation of kernel models can also be challenging. The SST-kernel function transforms the input data in a manner that is designed to capture information about drug regimen on several levels (individual medications, types of regimens, drug classes, etc.), but also obscures the details of these data. Continuous covariates suitable for use in these proportional hazards models can be extracted by principal components analysis, but these components can not be expected *a priori* to have a known interpretation. We can see that, for instance, the first principal compo-

nent describing the current regimen correlates with the choice of a PI or NNRTI to accompany the NRTI backbone in cART (Figure 3.4A). This interpretation of the "meaning" of PC1 was not know ahead of time, and we do not expect it to hold for all possible inputs.

We also evaluated the use of kernel-based similarity for estimating the influence of past regimen use on the risk of subsequent regimen failure. The three measures investigated are all based on the hypothesis that regimens highly similar to past drug combinations will be more likely to fail than those markedly different from previous exposure. The three kernel approaches differ in their assumptions of which past regimens are most important: the "Recent kernel" approach considers the immediately preceding regimen as most important, the "Mean kernel" weights each previous regimen equally, and the "Maximum kernel" only considers the previous regimen most closely matching current ARV therapy (Figure 3.5). The models containing only these kernel measures gave similar results, with hazard rate ratios 1.2-1.4 (all $P < 0.05$)(Table 3.5). It should be noted that in the case of an individual's second regimen all three of these measures will be identical — ie. the last value, mean and maximum of an array of length one are equivalent. This may provide partial explanation for the similar results seen here. Interestingly, when previous ARV drug use is simultaneously controlled for via counts of previous drugs, these kernel scores lose their association with failure. This suggests that the Recent/Mean/Maximum kernel scores recapitulate much of the information represented by the previous drug counts. While the AIC scores were inferior for these kernel-based metrics, the cross-validation ROC curves were essentially identical (Figure 3.6). The kernel strategies may be appealing, as they provide a parsimonious description of previous ARV exposure, summarizing the increased risk due to prior ARV exposure in a single number, rather than the four variables selected or the non-kernel models.

ARV therapy options continue to expand. The II, EI and FI classes of drugs

extend the list of viral functions targeted by therapy, and new drugs continue to be added to the commonly-used drug classes [18]. In this dataset, we observed 1,645 unique drug combinations, of which 821 were only prescribed once, and 1,433 were prescribed 10 times or fewer. As more drugs and drug classes are added to the ARV toolkit, these options will continue to expand. The large number of and sparse data regarding these rare drug regimens excludes the possibility of separately estimating failure risk for each potentially useful combination. While this problem has been recognized elsewhere [96], the SST-kernel provides a ready approach by quantifying the similarity between ARV regimens, allowing those with similar compositions (and therefore similar kernel principal components) to influence their neighbors. The SST-kernel function recognizes similarity between regimens with similar compositions without completely disregarding intra-regimen type differences (see Figure 2.2A). The incorporation of novel drugs or drug types into this approach is straightforward, as they can simply be added to regimen trees according to the targeted protein and drug type.

Regimen failure continues to be a serious clinical problem. On-therapy rebound of viral replication allows the evolution of drug-resistant virus, and ultimately narrows the range of therapy available to ARV-experienced individuals. In the setting of an expanding range of ARV options the number of potential drug combinations grows super-exponentially — individual evaluation of every possible cART drug regimen becomes impossible. The use of the SST-kernel to represent these complex objects provides a practical approach that captures much of the relevant information, and improves our ability to predict virologic failure.

3.4 Methods Summary

Data preparation

ARV regimen, CD4+ cell count, HIV-1 VL and demographic data were retrieved from the CNICS data repository. A total of 23,324 TCEs were ultimately selected from an original pool of 99,669 ARV regimen records. These records were matched by study site and patient ID to VL, CD4 and demographic data. Records were rejected or condensed for several reasons. Selected records had start dates between 1996 and 2010, and passed several quality control filters. Specifically, regimens which temporarily dropped, then re-gained a single drug, or were interrupted by short intervals ($\leq$ 7 days) were treated as uninterrupted therapy, as they were deemed more likely inaccuracies in drug start/stop dates than legitimate gaps. Regimens 7 days or fewer in duration, regimens containing more than one NNRTI, more than 4 NRTIs or more than 2 non-ritonavir PIs, regimens with simultaneous use of zidovudine and stavudine, or emtricitabine and lamivudine or regimens containing a drug more than 1 year prior to licensure were also excluded from analysis. These criteria identified regimens inappropriate to their stated date (pre-licensure dates), formulations typically avoided, but easily constructed by inadvertent concatenation of legitimate regimens (too many drugs in the same class), and drugs too similar to be constructively grouped in the same regimen (zidovudine/stavudine, emtricitabine/lamivudine). As we are investigating the failure of effective modern ARV therapy, regimens were selected that met the description of cART (also referred to as highly active antiretroviral therapy, or HAART), defined as "three or more antiretroviral medications, one of which has to be a PI, an NNRTI, one of the NRTIs abacavir or tenofovir, an integrase inhibitor (e.g., raltegravir), or an entry inhibitor (e.g., Maraviroc or enfuvirtide) ." [97] In an effort to distinguish virologic failure from ineffective regimens, we discarded regimens that did not achieve viral suppression, defined as

VL 200 copies/mL, 7 or more days post-treatment initiation. Finally, records that contained complete covariate data, including age, gender, study site and one or more CD4+ cell counts in the 120 days prior to treatment initiation were selected (outlined in Figure 3.1).

Kernel Calculations

Individual regimens were represented as tree structures beginning at a root node labeled "ARV," and branching down to final nodes representing individual drugs. Similarities between these trees was calculated as a weighted sum of shared subset trees as described previously (see Chapter 2). This produced a 1,645 x 1,645 kernel matrix of pairwise similarities between regimens. The *kernlab* package in R was used to calculate the kernel principal components of this matrix. The first ten principal components were selected as potential covariates in regimen failure models. The SST-kernel was also used to represent the similarity between a particular regimen and that individual's past ARV exposure. The three kernel-based representations of past ARV exposure were function of the array of kernel scores between the current regimen and each past regimen. Investigated options included the most recent kernel score, maximum kernel score and the mean of all kernel scores (Figure 3.5). In each of these three cases, regimens given to individuals with no previous ARV exposure were assigned scores of zero.

Kernel scores were also used to quantify the similarity between current and previous regimens. A similar kernel matrix was calculated, in this case including the filtered regimens described above plus regimens not meeting the potent cART definition or necessarily achieving viral suppression. The latter two classes of regimens were not modeled for regimen failure, but prior exposure to these regimens may effect the future likelihood of failure. This created a 3,117 x 3,117 matrix, which contained pairwise kernel (similarity) scores between ARV regimens. For each TCE, we defined a list of previous ARV regimens, and deter-

mined the kernel score between the current and most recent regimen, the mean or maximum of all kernel scores between current and previous regimen, as appropriate (Figure 3.5)

Proportional Hazards Model

Filtered regimen records were analyzed using Cox proportional hazards models, with virologic failure as the event of interest, and failure-free treatment change, loss to follow-up and end of study period (December 2010) as right-censoring events. Virologic failure was defined as a VL >200 copies/mL confirmed by a second VL >200 copies/mL or treatment-change event within 120 days. Variables included in the models were selected by first fitting a model with all available variables, then eliminating variables until a minimum in AIC (see "Model Diagnostics" below) was reached. cross-validation data was generated by random partition of data into 10 groups, performing an independent variable selection and model fit with each group, and making model predictions on the holdout set. Final modelss were fit on the full dataset using variables included in the majority of the cross-validation models. Models were fit using the *survival* package in the R statistical computing language [98]. In models specifically assessing the performance of the "Recent kernel", "Maximum kernel" and "Mean kernel" measures of previous regimen, we specified that these variables should be retained by the model selection algorithm. The final model selection algorithm was permitted to eliminate any variables for the final "Combined All" model.

Model Diagnostics

Several methods were used to test model fit and predictive accuracy. Kaplan-Meier curves were generated from the cross-validation predictions divided into high- and low-risk groups by the mean of the model linear predictors, signifi-

cance was tested using a class-rank test. Receiver-operating characteristic (ROC) curves were generated from the cross-validation predictions by defining a time cutoff of 1 year, and calculating the sensitivity (or true positive rate) and 1-specificity (or false positive rate) for classifying failure or non-failure among the samples not yet censored at 1 year using varying cutoffs for the proportional hazards model linear predictor. Area under this curve was calculated using Simpson's rule. The Akaike information criterion (AIC) was used both for model evaluation and variable selection, and is defined as $AIC = 2k - 2ln(L)$ where k is the number of parameters in the model, and L is the model partial likelihood. Partial correlation was calculated to assess the proportion of the residual sum of squares explained by a model using the method of O'Quigley et. al[95].

Chapter 4: Subset-tree Kernel-based Models for HIV Disease Progression

4.1 Introduction

Barring the discovery of a cure for HIV-1 infection, the ultimate goal of ARV therapy is to reduce HIV-associated morbidity and mortality by suppression of viral replication. In the era of modern cART, progression of HIV-associated disease to the late-stage clinical syndrome AIDS and ultimately to death may occur after a complex pattern of ARV therapy has occurred. In chapter 2 we highlighted the complexity of these ARV regimens, and suggested that the SST-kernel can improve the mathematical characterization of that complexity. In chapter 3 we examined the use of the SST-kernel to improve prediction of virologic failures of ARV therapy. As the sequence of ARV therapy prior to progression to AIDS or death is also complex, and the accurate prediction of disease progression is of obvious clinical utility, we now examine whether the SST-kernel can also improve our ability to model this process.

Previously developed prognostic models have improved our understanding of HIV disease progression. These studies have provided many invaluable clues to the optimization of ARV therapy. However, as we find in our models for regimen failure, the influence of ARV expousure has remained relatively simple. Prognostic models based on cohort-data in the early 2000s demonstrated the

powerful effect that the introduction of high-potency cART had on survival rates among HIV-infected individuals [99, 100, 101]. These studies generally modeled ARV use as being either highly-active combination therapy or not (HAART yes or no) [101, 99] or split cART into broad categories, for instance discriminating between PI-containing cART and other formulations [100]. Marginal structural models provide one tool for handling for time-varying treatment, and have been used to support the hypothesis that earlier administration of cART is preferable to deferred treatment [102, 103].

The puzzle of managing ARV therapy becomes more complicated in the highly antiretroviral-exposed patient. Life expectancy among HIV-1 infected individuals has improved in the age of less toxic and more convenient fixed-dose drug combinations [104, 105]. Nevertheless, evolution of drug resistance, likely secondary to viral replication in the context of virologic failure, remains a challenge [65, 27]. Cross-resistance between drugs in a class may narrow therapeutic options for the ARV-experienced patient, forcing the use of more expensive, more toxic or more difficult regimens [78, 77]. Even in the absence of a identified drug resistance mutation, undetected, low-abundance resistant variants may exist, and may reduce the probability of success on future therapy [91]. While steadily improving tools exist for guiding treatment in these individuals [88], better characterization of the contribution of accumulated drug exposure to risk of disease progression can shed additional light on this complex problem

To better characterize the influence of regimen sequences on disease progression, we have modified the SST-kernel approach outlined in previous chapters. As with our previous methods, the object is to define a kernel function that scores the similarity between data objects in a manner that captures the variation of interest. Here we continue to use a tree-structure to represent sequences of ARV regimens, and have taken the simple step of linking several regimen trees, as described in Chapter 2, by a set of numbered nodes to a common root (Fig-

ure 4.1). The SST-kernel function can then be applied to this tree as described previously. As with the single-regimen trees, sharing subset-trees contributes to the overall kernel score, and correspond to certain shared features. Exposure to the same drug combinations, types of regimens, drug classes and individual ARV drugs is scored in the same manner as in the single-tree situation. Additionally, regimen sequences containing the same number of distinct regimens, or the same regimen in the same point in the sequence will also contribute to kernel score. For instance, Figure 4.1 considers a hypothetical individual who has used three distinct ARV regimens, with a gap following the first regimen. The drug sequence outlined in Figure 4.1A would be represented by the tree drawn in Figure 4.1B. The subset-tree outlined in red recognizes the number of distinct regimen changes in the sequence, and the sub-tree outlined in blue corresponds to the feature "emtricitabine/tenofovir/efavirenz was the third regimen administered." We will test whether this approach to describing a sequence of ARV regimens improves prediction of disease progression.

4.2 Results

Data Description

ARV drug records, along with clinical and demographic covariates were retrieved from the CNICS data repository and filtered for quality control, resulting in a study group of 11,908 HIV-1 positive individuals (see Methods). The CNICS patient population includes individuals who initiated care at one of eight academic clinical centers from the United States after 1 January, 2005 [6]. Disease progression events were identified when development of an AIDS-defining illness (ADI), according to the 1993 CDC definition [106], or death from any cause occurred. Individuals who had no disease progression, or their first disease progression event after initiation of ARV therapy were selected, and constituted the

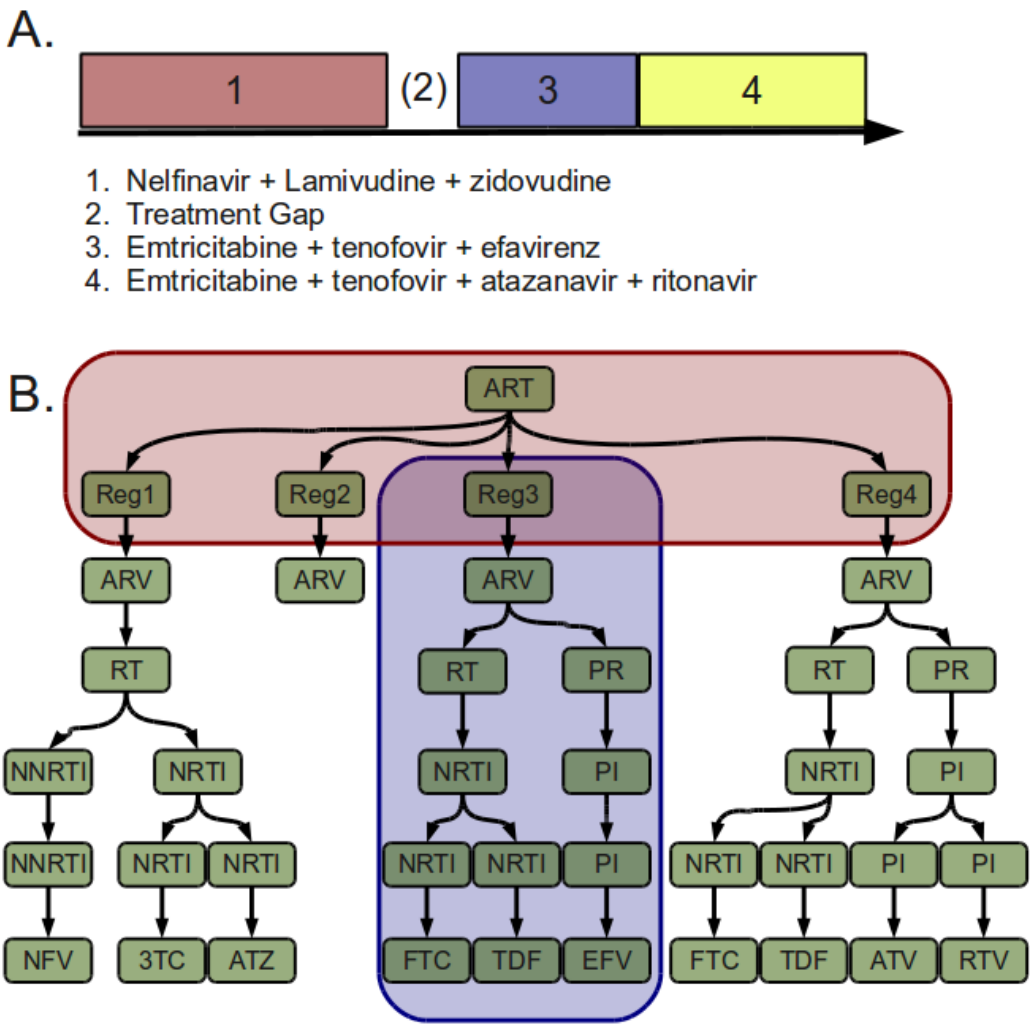


Figure 4.1: Outline of subset-tree kernel description of regimen history. A: Timeline of regimen use for a hypothetical patient. Drug combinations labeled by number. B: Regimen tree for same hypothetical patient. Two subset-trees are outlined in red and blue. ART: Antiretroviral therapy, ARV: Antiretroviral regimen, NFV: nelfinavir, 3TC: lamivudine, AZT: zidovudine, FTC: emtricitabine, TDF: tenofovir, EFV: efavirenz, ATV: atazanavir, RTV: ritonavir.

Table 4.1: Disease progression dataset description — ADI: AIDS Defining Illness, ARV: Antiretroviral JHU: Johns Hopkins Univ., UCSD: Univ. of California, San Diego, UW: Univ. of Washington, UAB: Univ. of Alabama, Birmingham, UCSF: Univ. of California, San Diego, UNC: Univ. of North Carolina, CWRU: Case Western Reserve Univ., Fenway: Fenway Health. MSM: Men having sex with men, IDU: Injection drug use. IQR: Intra-quartile range.

Sample Size	**N**	**Demographics/Risk Factor**	**Percent**
Individuals	11,908	Male Gender	80
Progress to ADI or Death	3,179 (27%)	White, Non-Hispanic	47
Distinct ARV histories	16,719	Black	39
		Hispanic	10
Study Site	**Percent**	Other Race/Ethnicity	4
JHU	23	MSM	58
UCSD	18	IDU	21
UW	15		**Mean [IQR]**
UAB	12	Age at Cohort Entry	37 [31-43]
UCSF	9	Year Entered Cohort	2000 [1997-2004]
UNC	8	Duration of Follow-Up (years)	6 [3-10]
CWRU	8		
Fenway	7		

11,908-person population for this study. Among this group, 3,179 experienced a disease progression event, over 63,896 person-years of follow-up, giving a raw disease progression risk of 5.0% per person-year (Table 4.1).

Disease Progression Models

Time-dependent Cox Proportional Hazards models were used to determine the factors associated with disease progression in ARV-therapy exposed individuals in the CNICS cohort. Subjects were included in the model from the time of first ARV regimen, VL or CD4+ cell count record, until the time of first disease progression event, or were censored when losfit to follow-up, or at the end of the study period for this analysis (December 2010). An updated set of covariates was calculated every subsequent year. Three alternative approaches to describing past ARV-exposure were examined, SST-kernel principal components calculated from regimen-sequence trees (Figure 4.1B), binary variables indicating prior exposure to each ARV class (NNRTI, NRTI, PI, II, EI, FI), or a combination of both. This dataset included 16,719 distinct ARV exposure histories amongst the 11,908 (note that since regimen history is re-calculated each year,

an individual may contribute multiple unique regimen histories (Table 4.1). Indicator variables providing a “yes” or “no” to prior drug-class exposure provided better model performance than raw counts of drugs in each class previously prescribed (data not shown).

Final model covariates are shown in Table 4.2. As with the regimen failure models presented previously, there was site-to-site variation in disease progression risk, possibly due to differences in patient population. Disease progression risk increased with age, and was consistently greater in white patients than those of other racial or ethnic backgrounds. As expected, higher recent and lagged CD4+ values were associated with lower risk. Surprisingly, while 9-month lagged VL was associated with higher progression rates, a higher VL value at the most recent tests was mildly protective. In all models, current regimen was represented by SST-kernel principal components (current PC), as described in Chapter 3. In the three models presented here, 4-6 current PCs were included in the final models, and these varied slightly from model to model. The data projections selected in kernel principal components analysis are dependent on the input data, so the PCs extracted here do not have a one -to-one correspondence with those identified in Chapter 3, as the data has been prepared differently.

The principal difference between the three models presented here is the representation of regimen history. In the regimen-sequence kernel model, 7 of the 10 past-regimen principal components (past PCs) were selected by the model-selection algorithm, of which 5 had statistically-significant associations with disease progression. In the drug indicators model, all variables were retained by model selection, and previous exposure to NNRTI, NRTI, PI, II and FI drugs were significantly associated with progression. Of these, prior exposure to NRTI drugs was associated with higher progression risk, while exposure to NNRTI, PI, FI and II drug classes appeared protective. This may be an expression of the sub-optimality of NRTI-only therapy, particularly historic mono- and dual-therapy. Addi-

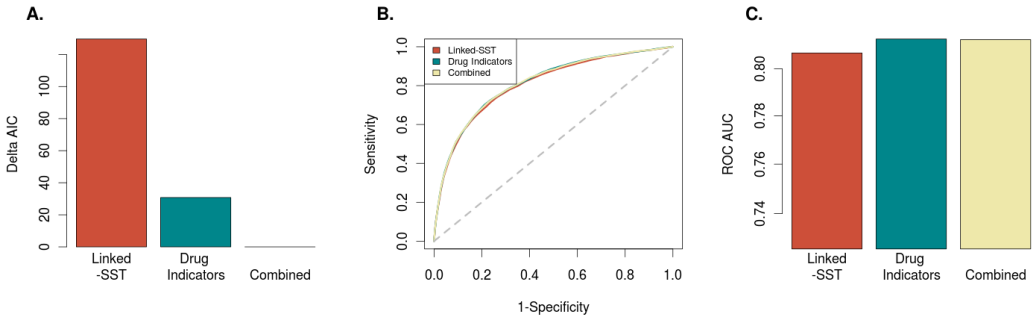


Figure 4.2: Disease Progression Model Diagnostics: A. Delta AIC: difference between model AIC and minimum (best) AIC, B: ROC curves, C: Area under each ROC curve.

tional exposure to other drug classes (as occurs in combination therapy) lessens this risk. In the combination model, all of the drug exposure variables were selected for inclusion in the Combined model, as were 5 of the 10 past PCs. Models were evaluated using the partial likelihood-based AIC measure, and cross-validation based receiver-operating characteristic (ROC) curves. The Combined model provided the best (lowest) AIC, with the model based on drug-indicator variables alone out-performing the regimen-sequences SSTK-model (Figure 4.2A). Although the non-kernel model out-performs the regimen-history SST-kernel method by this metric, the lower AIC for the Combined model suggests that the kernel representation may provide additional information not represented by the drug indicators method. There appears to be little difference between the three methods as measured by ROC curves, either by curve shape, or overall area under the curve (AUC) (Figure 4.2B-C).

4.3 Discussion

Here we have presented a kernel-based strategy for describing longitudinal ARV regimen use in a model for HIV disease progression. Use of ARV over time is a very complex phenomenon, with an extremely large feature space of possible regimen sequences. Kernel approaches are well-suited for such high dimensional problems, as the complexity of the model scales with the number

Table 4.2: Hazard Rate Ratios for disease progression model parameters — 95% Confidence intervals presented in brackets. Covariates with p <0.05 in bold. NS: not selected, hyphen dennotes variable not avalible in selected model. Ref: Reference value for categorical variable. II: Integrase inhibitor, EI: Entry inhibitor, FI: Fusion inhibitor, VL: Viral load (RNA), Current PC: kernel-based description of currently-used ARV regimen (See Chapter 2-3), Past PC: kernel-based description of ARV use history (See Figure 4.1).

Covariate	**Sequence-SST**	**Non-Kernel**	**Combined**	**Covariate**	**Sequence-SST**	**Non-Kernel**	**Combined**
Gender	NS	NS	NS	Current PC1	NS	NS	NS
Age	1.03 [1.02-1.03]	1.03 [1.02-1.0]	1.03 [1.02-1.03]	Current PC2	NS	0.995 [0.993-0.998]	0.996 [0.993-0.998]
Cohort Entry Year	NS	0.98 [0.97-0.99]	NS	Current PC3	NS	NS	NS
Race/Ethnicity = Black	Ref	Ref	Ref	Current PC4	0.997 [0.995-0.999]	0.997 [0.995-0.999]	0.997 [0.995-0.999]
= Hispanic	0.97 [0.82-1.14]	0.98 [0.83-1.14]	0.98 [0.83-1.15]	Current PC5	0.996 [0.994-0.999]	NS	NS
= White	1.11 [1.02-1.21]	1.11 [1.01-1.20]	1.11 [1.02-1.21]	Current PC6	NS	1.007 [1.004-1.011]	1.007 [1.003-1.01]
= Other	1.09 [0.88-1.35]	1.07 [0.87-1.32]	1.08 [0.88-1.34]	Current PC7	1.002 [1.000-1.004]	1.003 [1.000-1.005]	1.003 [1.000-1.005]
Study Site = CWRU	Ref	Ref	Ref	Current PC8	NS	1.002 [1.000-1.004]	1.002 [1.000-1.004]
= Fenway	0.52 [0.41-0.66]	0.53 [0.41-0.68]	0.53 [0.41-0.68]	Current PC9	1.003 [1.001-1.005]	1.008 [1.005-1.010]	1.008 [1.005-1.011]
= JHU	1.39 [1.20-1.60]	1.30 [1.13-1.50]	1.31 [1.13-1.51]	Current PC10	NS	NS	NS
= UAB	1.08 [0.92-1.28]	1.06 [0.90-1.25]	1.08 [0.91-1.27]	Past PC1	0.998 [0.997-0.999]	—	NS
= UCSD	1.29 [1.10-1.52]	1.27 [1.08-1.49]	1.26 [1.07-1.48]	Past PC2	0.997 [0.995-0.998]	—	NS
= UCSF	0.99 [0.81-1.20]	0.96 [0.79-1.18]	0.96 [0.79-1.17]	Past PC3	0.998 [0.996-0.999]	—	0.997 [0.995-0.998]
= UNC	1.04 [0.87-1.25]	1.05 [0.87-1.26]	1.03 [0.86-1.24]	Past PC4	0.998 [0.996-1.000]	—	NS
= UW	1.01 [0.86-1.19]	1.01 [0.86-1.19]	1.01 [0.86-1.19]	Past PC5	1.004 [1.001-1.006]	—	1.004 [1.001-1.007]
MSM	NS	NS	NS	Past PC6	NS	—	0.997 [0.994-1.000]
IDU	NS	NS	NS	Past PC7	0.998 [0.995-1.000]	—	0.997 [0.995-0.999]
Most Recent CD4	0.82 [0.77-0.87]	0.82 [0.78-0.87]	0.82 [0.78-0.87]	Past PC8	NS	—	NS
CD4 — 3 Months	0.82 [0.77-0.88]	0.82 [0.77-0.88]	0.82 [0.77-0.88]	Past PC9	0.995 [0.992-0.997]	—	0.995 [0.992-0.997]
CD4 — 6 Months	0.94 [0.86-0.97]	0.91 [0.86-0.97]	0.91 [0.86-0.97]	Past PC10	NS	—	NS
CD4 — 9 Months	0.96 [0.91-1.00]	0.96 [0.91-1.01]	0.96 [0.91-1.01]	Past NNRTI	—	0.84 [0.77-0.91]	0.80 [0.73-0.87]
Most Recent VL	0.88 [0.85-0.92]	0.89 [0.86-0.93]	0.89 [0.85-0.92]	Past NRTI	—	2.49 [2.08-2.99]	2.60 [2.17-3.12]
VL — 3 Months	NS	NS	NS	Past PI	—	0.91 [0.82-1.01]	0.88 [0.79-0.98]
VL — 6 Months	NS	NS	NS	Past II	—	0.26 [0.17-0.42]	0.26 [0.17-0.41]
VL — 9 Months	1.15 [1.10-1.20]	1.16 [1.11-1.21]	1.16 [1.11-1.21]	Past FI	—	0.26 [0.17-0.42]	0.71 [0.52-0.97]
Regimen Count	0.98 [0.97-1.00]	0.98 [0.96-1.00]	0.98 [0.96-0.999]	Past EI	—	0.36 [0.09-1.46]	0.36 [0.09-1.44]

of unique examples (and therefore the dimension of the kernel matrix) rather than the dimension of the feature space. In this case, the kernel-based model for disease progression performed on a par with simpler non-kernel methods, as assessed by ROC-AUC, albeit not as well when measured by AIC.

The kernel function described here accounts for exposed drugs, drug classes, regimen types, number and sequence of prescribed regimens. At present, this method does not account for regimen duration, or differentiate between regimens that suffered virologic failure or were changed for other reasons. Nor does it allow any flexibility in the regimen sequence. For instance, if two individuals had received the same regimen, but offset by one position in their sequences — e.g., subject A was given Regimen X as a first regimen, and subject B received the same regimen as their second ARV regimen, there is no additional contribution to the kernel score for being close in order. Accounting for these details may be possible as future improvements to this approach.

There were some differences in both the current and past PCs selected for the SST-kernel model and the combination model. This is likely due to interdependence between these variables — the drug-exposure and past PC variables are overlapping but non-identical representations of the underlying regimen history data, and will both likely be correlated with current ARV use. For instance, dropping past PC1 and PC2 from the Combined model could be explained by overlapping information between these variables and the drug-class exposure variables.

The Linked SST kernel representation of a sequence of ARV regimens presented here does not greatly improve our ability to estimate risk of disease progression in HIV-1. The Combination model selected 5 of 10 past-regimen principal components, of which four had statistically significant effects on disease progression risk. The Combination model also had an AIC score modestly better than the Drug Indicators model (Figure 4.2A). Taken together, this indicates that

the Linked SST approach encodes some useful information not available in the simpler model. On the other hand, the Linked-SST model AIC was substantially worse than the Drug Indicators model, and all three models preformed very similarly when assessed by cross-validation ROC AUC (Figure 4.2). Computation of the kernel scores for large trees, as depicted in Figure 4.1B, as well as the kernel principal components analysis a kernel matrix of this size are very computationally expensive. Therefore, it is unlikely that the Linked-SST model is likely to be useful as described here. Despite this result, all three models make good predictions of disease progression risk, with ROC AUC $>$0.8. This estimate implies that two randomly selected samples, one destined to experience disease progression in the subsequent year, and the other experiencing no event, this model will assign higher risk to the progressing individual 80% of the time. The Linked-SST kernel presented here focuses on particular aspects of regimen — particularly regimen sequence and composition. It is possible that these regimen history features are simply not good correlates of disease progression, or that other variables in the model (such as age, year of cohort entry and duration of follow-up) are sufficiently confounded with regimen history as to render the linked-SST kernel variables redundant. Covariates not examined here that focus on other aspects of ARV history, such as duration of individual therapies and virologic failure of previous therapy could prove more useful.

4.4 Methods Summary

Data Preparation

A total of 99,699 ARV regimen records from 18,532 individuals were retrieved from the CNICS data repository and filtered for quality control as described previously (see Chapter 3). Briefly, regimen data was first filtered for temporary gaps in a single drug, and short interruptions in therapy that were

more likely bookkeeping errors than legitimate changes in therapy. ARV use was regarded as uninterrupted in these cases. Regimens were also filtered for appropriate calendar year (1996-2010), sufficient regimen duration ($>$7 days), therapy appropriate to recorded date (no individual drug use $>$1 year prior to licensure) and reasonable regimen construction ($\leq$ 4 NRTI, $\leq$ 1 NNRTI, $\leq$ 2 non-ritonavir PI, no simultaneous use of zidovudine/stavudine or emtricitabine/lamivudine). Additionally, single- or dual-drug regimens after 1997 were excluded as inappropriate therapy for stated date (this filter was redundant with the potent cART requirement imposed in Chapter 3). Records were matched with CD4+ cell count, viral RNA (VL), demographic, diagnosis and mortality data by patient id number and study site. Disease progression events were defined as death or development of an AIDS defining illness (ADI), as defined by the 1993 CDC definition [106]. Subjects who experienced a disease progression event prior to first ARV use were excluded from this analysis. For the remaining 11,908 individuals, time-zero was defined as the date of first ARV therapy, VL or CD4+ cell count on record.

Kernel Calculations

Current ARV regimen was parameterized using the first 10 principal components from the SST-kernel matrix, as described extensively in Chapters 2 and 3. Past regimen was entered in to the model in a similar fashion. At each follow-up time, past regimens were built into a regimen-sequence tree formed from linking each individual regimen tree to a common node via a layer of numbered nodes (Figure 4.1B). If a gap of greater than seven days occurred between regimens, this was included as an empty regimen tree. A kernel matrix of all observed regimen sequences was calculated for this set of trees using the SST-kernel described in Chapter 2, with λ = 0.55. The first 10 principal components of this kernel matrix were included as descriptors of regimen history in models for disease progression.

Proportional Hazards Model

A time-dependent Cox Proportional Hazards model was used to estimate the effects of covariates on risk of disease progression. The regression dataset contained a record at time-zero for each individual, one follow-up record for each year after time-zero until a disease progression event, or the individual is censored due to loss to follow-up or end of study period (December 2010). This gave 63,896 person-years of follow-up. Baseline covariates included age at time-zero, study site, gender, risk group (MSM or IDU), race/ethnicity, and year of entry to cohort. Time-dependent covariates included years since time-zero, the most recent CD4+ cell count and VL measurement, lagged CD4+ and VL (defined as the most recent measurements as of 3, 6, and 9 months prior to follow-up date) ARV therapy in use at time of follow-up, and past ARV regimen use. As described in Chapter 3, a cross-validation approach was used, covariates were selected by iterative elimination of covariates until a minimum AIC value was reached, and final models were constructed from covariates included in a majority of cross-validation models. Model fit was assessed by cross-validation ROC curves and AIC, as described in Chapter 3. Briefly, the AIC is a measure of model fit that prefers high-partial likelihood models while penalizing for the number of parameters in an attempt to avoid over-fitting. The area under the ROC is based here on 10-fold cross validation predictions, and estimates the probability of correctly ordering a random positive and random negative sample.

Chapter 5: Three-Dimensional Spatial Kernel Models for Characterization of Antibody Neutralization

5.1 Introduction

Neutralizing antibodies (NAb) capable of inhibiting a wide array of HIV-1 strains are central to the search for an effective HIV-1 vaccine. Anti-HIV-1 antibodies generally target the viral surface protein (Env), a highly-variable glycoprotein that mediates viral entry into the host cell. Env is composed of a non-covalently bound trimer of heterodimers. Each dimer consists of two cleavage products of the gp160 precursor: gp120, which is responsible for CD4 and co-receptor binding, and the gp41 transmembrane protein, responsible for cell fusion [32]. The high level of population diversity, dense glycosylation and structural flexibility of Env are barriers to development of a broadly-effective vaccine [36, 37]. Env amino acid sequences can differ by 20% between individuals within a single subtype of HIV-1, and as much as 35% between subtypes [35]. Infected individuals generally mount an antibody response that neutralizes autologous virus from early in their own infection, but viral evolution allows rapid escape from this pressure [39, 40]. Broadly-neutralizing antibodies, capable of neutralization of heterologous virus are generally produced later in infection, if at all [41]. The broad diversity of Env and rapid escape have prompted a search

for antibodies that neutralize a broad diversity of HIV-1 strains, with a focus on those targeting conserved sites in the Env protein.

The first such antibody identified, IgG1b12 (b12) was discovered in 1991 and targets the CD4 binding site in Env [42]. Several other NAbs have since been identified and characterized, variously binding the third variable (V3) loop of gp120 (447-52D), the gp41 stem (2F5), and gp120-bound glycans (2G12) [43, 44, 45]. More recently, focus on individuals infected with non-subtype B virus and high-throughput screens for desirable antibodies have re-ignited the interest in discovery of NAbs [46, 47, 48]. Despite this array of anti-HIV-1 antibodies capable of virus neutralization, progress in vaccine development has been modest. Early efforts to develop an antibody-based vaccine elicited antibody response to laboratory-adapted strains but not to primary isolates of circulating virus [107]. Subsequent attempts to develop a cytotoxic T-lymphocyte based vaccine were ultimately unsuccessful in phase III clinical trials [30]. The recent RV144 vaccine trial conducted in Thailand reported a 31.2% efficacy (P=0.04) in a modified intention-to-treat analysis [31]. In this trial, induction of HIV-1 binding antibody was reported, but data on neutralization of autologous or heterologous virus was not available. This remains the only study to date to show protective results in an HIV-1 vaccine.

Recombinant DNA techniques have contributed a great deal to our understanding of NAb neutralization and its complex relationship with the Env protein. Site-directed mutagenesis and domain swapping or deletion have helped identify antibody binding sites and other determinants of neutralization. Binding sites have been identified for some NAb, however there is also strong evidence that Env variations outside these epitopes strongly influence neutralization. Deletion of the V1/V2 domains of Env was shown early on to increase the susceptibility to some antibodies that bind outside the deleted region [108]. Further research has suggested that the V1/V2 loops in particular may contribute

to a "glycan shield," capable of protecting other regions of Env from neutralization [33, 34]. Alanine scans, in which each site in the Env protein is separately mutated to alanine, have also identified neutralization determinants outside canonical epitopes [50, 51]. One limitation of these approaches is that they ultimately assess artificially mutated virus variants that may never be observed in the wild. Furthermore, the surrounding genetic context is likely to influence the implications of particular variations. Mutations may confer a particular neutralization phenotype in one HIV-1 backbone, but yield the opposite result or produce non-viable virus in another[52]. Clearly, there is a complex, context-dependent relationship between variation in Env and unacceptability to NAb. While informative, the results of SDM or alanine scan approaches may only be relevant to the genetic context investigated, and may not faithfully recapitulate variations that drive neutralization phenotype in the wild.

Recent advances in high-throughput searches for NAbs [47] motivate a similarly efficient process for their evaluation and characterization. Computational methods possess advantages in speed and cost over laboratory techniques, and development of effective computational approaches would facilitate the search useful NAbs. Some progress has been made in the development of such computational methods. Genetic signatures that influence neutralization of viral strains by the NAb b12 or polyclonal sera have been identified [53, 54, 55]. However, several features of the Env protein complicate this effort. The high degree of sequence diversity in the Env gene makes multiple sequence alignment difficult. For this reason, regions of particularly high diversity may be excluded from analysis, potentially missing determinants of neutralization. Furthermore, ambiguities in sequence alignment may mis-align residues, producing erroneous signals, or masking relevant ones. Approaches based on examination of primary amino acid sequence also neglect the potentially important 3-dimensional structure of the Env protein. Antibody neutralization occurs due to the interaction

between two 3-dimensional objects, and determinants of binding such as epitope masking, steric hindrance and conformational change, as well as the binding epitope itself, are 3-dimensional in nature [56]. An analytic approach that examines naturally-occuring virus and accounts for 3-dimensional structure and uncertainty in high-diversity regions may better characterize real-world antibody behavior.

Kernel-based approaches may provide a an approach for analysis of an object as complicated as the highly-variable, 3-dimensional Env protein. Kernel methods provide tools such as regression and classification via the computation of the "kernel matrix:" a symmetric, positive semi-definite matrix of pairwise similarity scores between data points. This provides a practical approach to parameterization of complex data — if we define a function for scoring the similarity between 3-dimensional Env proteins, we can side-step the difficult task of adapting these data for traditional parametric models. Here we describe a kernel approach for prediction of NAb neutralization from structural neighborhoods in the 3-dimensional Env protein. This approach takes advantage of existing knowledge of the Env structure, includes highly-variable regions in analysis, and is robust to alignment uncertainty.

5.2 Methods

Antibody Neutralization Assay

Primary HIV-1 Env amino acid sequences were obtained, along with matched antibody neutralization data for six different NAbs: IgG1b12 (b12), 447-52D, 2G12, 2F5, PG9 and PG16. The antibody concentration capable of neutralizing 50% of HIV-1 virus infection activity (IC50) was determined for each sequence using an *in vitro* assay. The acquisition of this data has been described elsewhere [49, 46]. Briefly, Env-pseudotyped virus was produced by co-transfection of the

Env of interest with an Env-deficient HIV-1 backbone. A single round of replication was then performed in target cells, and the antibody concentration reducing virus production by 50% (IC50) was determined. Depending on the experiment, the maximum antibody concentration tested was 25 or 50 g/mL.

Structural Neighborhood Kernel Calculation

Kernel matrices of pairwise similarity scores were calculated for the amino acid sequences in a four-step algorithm (Figure 5.1). Each kernel matrix is specific to a particular point in the 3-dimensional protein structure, and depends on the structural neighborhood around that point. First, each Env amino acid sequence in the sample was mapped to a partial Env 3-dimensional structure. To perform this mapping, each amino acid sequence was aligned to the reference sequence HXB2 [109]. Partial structures for the HXB2 gp41 (PDB accession: 1AIK [110]) and gp120 (PDB accession 3DNN) [111] subunits were then used to estimate the spatial location of each residue in the sample sequence. Insertions and deletions were handled by linear interpolation of the effected residues to space them evenly between flanking sequences. For regions not included in the solved 3-dimensional structure, each residue was assumed to be 3.8 Å from it's immediate neighbors — the average $\alpha - \alpha$ carbon distance in the reported structure. Second, for each residue in the reference sequence, a structural neighborhood was determined, composed of the amino acid residues within a set radius of the site in question.

In the third step of the kernel algorithm, we calculated pairwise similarity scores between the structural neighborhoods defined in step 2. The kernel function(5.1) depends on the overlap between amino acids in the two neighborhoods and the 3-dimensional (euclidean) distance between these matching residues. Essentially, we sum over all amino acid pairs that are identical or from the same amino acid class (with a penalty of 0.5), then penalize those pairs by

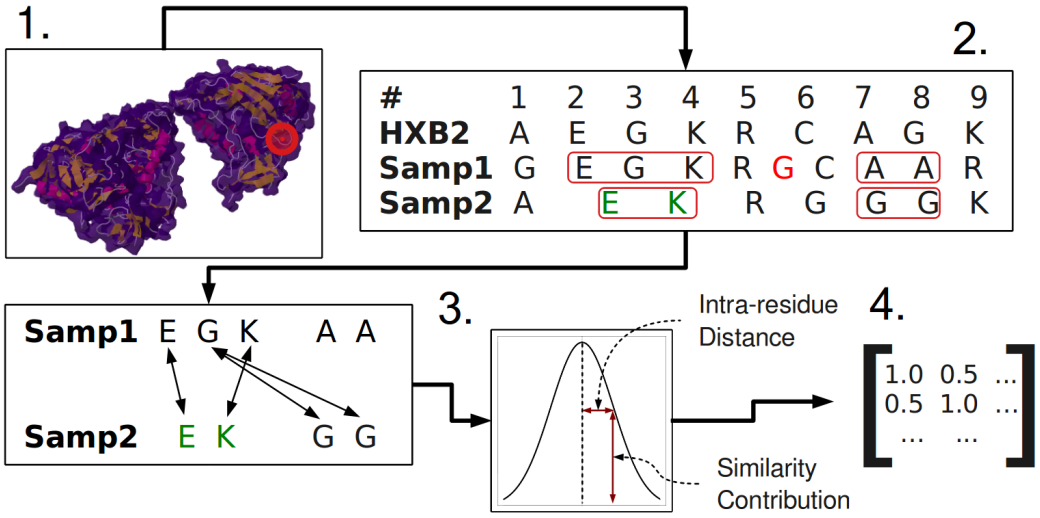


Figure 5.1: Outline of structural neighborhood kernel calculation. 1. Sample sequences are mapped to 3-dimensional structure of HXB2 Env. 2. A neighborhood of amino acids is highlighted in red. Note residues in neighborhood need not be contiguous in primary sequence. 3. Similarity score is calculated. For illustration only exact residue matches are highlighted by arrows. Relatively distant matching "G" residues will contribute less than nearby matching "E" and "K" residues. 4. Matrix is normalized, similarities between each sequence and itself is 1.

a negative exponential (Gaussian) function of the euclidean distance between them. Hence, identical residues in the same location will make a maximum contribution to the overall similarity function, residues merely from the same class, or at a larger distance contribute less. Amino acid classes were defined as: positively charged (Arg, His, Lys), negatively charged (Asp, Glu), polar uncharged (Ser, Thr, Asn, Gln), and hydrophobic (Ala, Ile, Leu, Met, Phe, Trp, Tyr, Val). Three amino acids were not assigned to a class and only contributed to the similarity score when matched exactly (Cys, Gly, Pro). Finally, in the fourth step, the kernel matrix was normalized according to equation (5.2).

$$\kappa(n_i, n_j) = \sum_{a_k \in n_i} \sum_{a_l \in n_i} \delta(a_k, a_l) e^{\frac{-\|a_k - a_l\|^2}{2\sigma^2}}$$

Where: n_i is a structural neighborhood, and $a_k \in n_i$ are the amino acids in n_i, and $\|a_k - a_l\|$is the euclidean distance between residues k and l.

$$\delta(a_k, a_l) = \begin{cases} 1 & \text{if } a_k \text{ and } a_l \text{identical} \\ 0.5 & \text{if } a_k \text{ and } a_l \text{from same amino acid class} \\ 0 & \text{otherwise} \end{cases} \tag{5.1}$$

$$\bar{\kappa}(n_i, n_j) = \frac{\kappa(n_i, n_j)}{\sqrt{\kappa(n_i, n_i)\kappa(n_j, n_j)}} \tag{5.2}$$

This leads to a matrix where the similarity of any sequence with itself is 1, and the similarity between two non-identical sequences is ≤ 1. This algorithm has two adjustable tuning parameters: the neighborhood radius, r, and the variance of the Gaussian distance penalty function, σ^2. These parameters were determined independently for each NAb model as described below. Calculations were performed using a custom program in the Python programming language. This algorithm was applied to each amino acid site in the reference strain, HXB2. This yielded 856 kernel matrices, each corresponding to a structural neighborhood centered on a residue in the HIV-1 Env protein reference sequence.

Model Selection

A model for prediction of neutralization sensitivity was constructed for each of the six NAbs evaluated. In each case, a nu-regularized kernel support vector machine (kSVM) was used to predict sensitivity (IC50 $\leq$ 25 μg/mL) or resistance (IC50 $>$ 25 μg/mL) to neutralization using the libsvm software package.[Chang, 2001] The optimum regularization parameter, , and kernel tuning parameters (r and σ^2) were estimated from the data using a grid search. For each of the six NAbs we selected the combination of tuning and regularization parameters that gave the best 10-fold cross-validation accuracy for a kSVM based on a single structural neighborhood (for each NAb, all 856 available neighborhoods were evaluated as candidates in this step). After tuning, a final model was assembled for each of the NAbs consisting of a combination of several structural neighborhoods from the Env protein. Neighborhoods were added to the model by a greedy search, in which each iteration added the structural neighborhood that maximized the 10-fold cross-validation accuracy in combination with the

previously-selected elements. Each final model consisted of a kSVM trained to predict neutralization sensitivity from the sum of the selected kernel matrices. For comparison, the proportion of Env sequences with an exact match to the published linear epitope was calculated, when available (b12, 447-52D, 2F5).

Model Evaluation

Final models were evaluated by a independent run of 10-fold cross-validated kSVMs. This provided a new random partitioning of the data into cross-validation sets independent of the sets used to build the models. Accuracy, sensitivity and specificity of classification were calculated for these models. Model performance was further evaluated by calculating the area under the receiver operator characteristic (ROC) curve, comparing change in the true positive and false positive rates as the cutoff for the kSVM decision boundary varied, again in the context of an independent 10-fold cross-validation.

5.3 Results

HIV-1 Env amino acid sequences with matching neutralization IC50 values were obtained for six NAbs targeting various regions of the Env protein: b12, 2F5, 447-52D, 2G12, PG9 and PG16. The number of sequences for each NAb varied from 88 for 447-52D to 217 for 2G12. The balance between sensitive and resistant viral strains varied from 18.2% sensitive for 447-52D to 71.1% sensitive for PG9. The distribution of sequences by HIV-1 subtype also varied across the NAbs, with subtype B as the most common for most antibodies (Table 5.1).

Structural information was available 65% of the gp120 sequence, and 42% of gp41. Regions not included in the known 3-dimensional structure included the highly-variable V1/V2 loops, a portion of the V4 domain, and the intracellular and transmembrane domains of gp41 (Figure 5.2). In regions lacking structural data, each amino acid was assumed to be 3.8Åfrom its immediate neighbors (the

Table 5.1: Summary of neutralizing antibody data. Sample size, proportion of sequences sensitive to neutralization and breakdown by subtype for each of the six NAbs studied. All subtypes representing more than 10% of any sample shown. Other: all other subtypes and sequences of unknown subtype. AE: CRF_AE

			% by Subtype						
Antibody	**N**	**% Sensitive**	**A**	**B**	**C**	**D**	**G**	**AE**	**Other**
b12	127	40.2	8.7	38.6	13.4	7.9	0.0	12.6	18.8
2F5	126	60.0	8.7	38.9	13.5	7.9	1.0	12.0	18.0
2G12	217	25.3	9.2	23.0	20.9	6.6	1.0	13.8	25.5
447-52D	88	18.2	9.1	30.7	12.5	11.4	1.1	11.4	23.8
PG9	142	71.1	9.9	11.3	10.6	12.0	12.7	6.3	37.2
PG16	142	66.2	9.9	11.3	10.6	12.0	12.7	6.3	37.2

average $\alpha - \alpha$ carbon distance in the known structure). For a neighborhood radius of 7Å, structural neighborhoods contained a mean of 5 amino acids. With a distance penalty variance of $\sigma^2 = 5\mathring{A}^2$ the mean normalized pairwise similarity (ie. the value of the kernel function) was 0.065 (Figure 5.2). Average neighborhood size was expectedly higher in regions of structural information than in those without, as protein folding may bring amino acids into closer proximity. Mean pairwise similarity varied across regions with and without structural data, but exhibited high similarity just proximal to the gp120/gp41 junction, and particularly low similarity in the V1/V2 and V4 variable loops.

The variable parameters in the algorithm — the neighborhood radius r, the width of the Gaussian penalty function σ^2, and the SVM regularization paramater ν, were determined by an empirical grid search for each antibody. The selected values maximized the cross-validation accuracy of a single-neighborhood SVM in each case. Selected values for r ranged between 6-12 Angstroms (Å), 2-8 $\mathring{A}^2$ for σ^2, and 0.05 to 0.3 for ν (Table 5.2). The final selected model contained between three (for PG16) and seven (for 2F5) structural neighborhoods. The neighborhoods for 2F5 and 2G12 overlapped known antibody-binding regions — the 2F5 epitope (neighborhoods 666 and 667) [44] and a potential N-linked glycosylation site necessary for 2G12 binding (neighborhood 364) [45] (Table 5.2).

In independent cross-validation runs, the classification accuracy ranged from 78% for b12 to 94% for 2F5. Sensitivity and specificity were generally greater

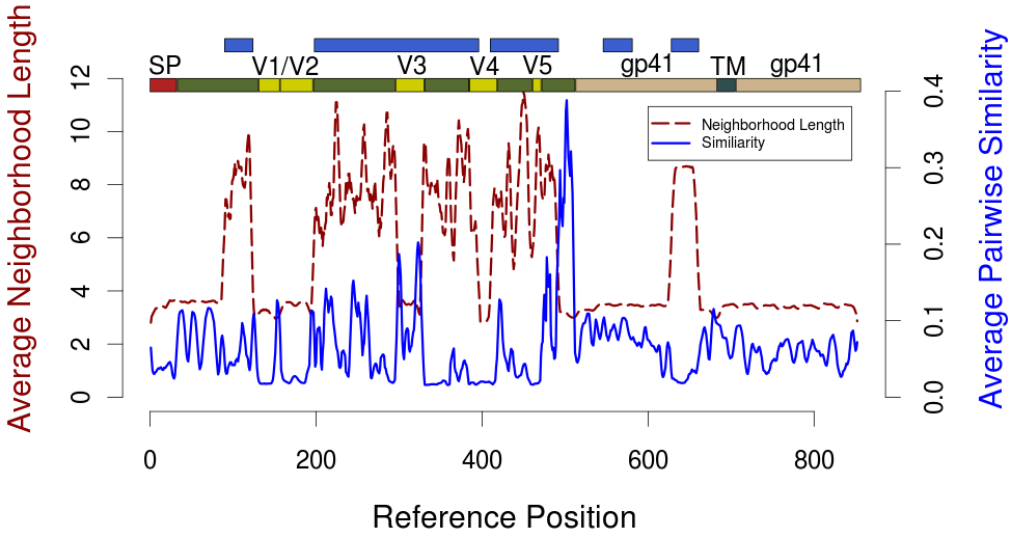


Figure 5.2: Average number of amino acids and average pairwise similarity across sequences in gp120. Regions of gp120 labeled above plot, regions with associated 3-dimensional structure denoted by blue bars above regional annotation. SP: signal peptide, V1-5 Variable regions 1-5, TM: trans-membrane region. Curves smoothed by a Daniell smoothing kernel with bandwidth 5.

Table 5.2: Antibody neutralization model summary. Acc: accuracy, Spec: specificity, Sens: sensitivity, Model: predictions from kernel model, Exact: predictions from exact matching of published epitope r: structural neighborhood radius, σ^2: Gaussian penalty function width, ν: SVM regularization parameter. Bold entries denote neighborhoods that overlap known determinants of neutralization by NAb or polyclonal sera. Selected neighborhoods are listed in order of their addition to the model. [a] No known linear epitope. [b] Overlaps "canonical" epitope or antibody binding determinant.

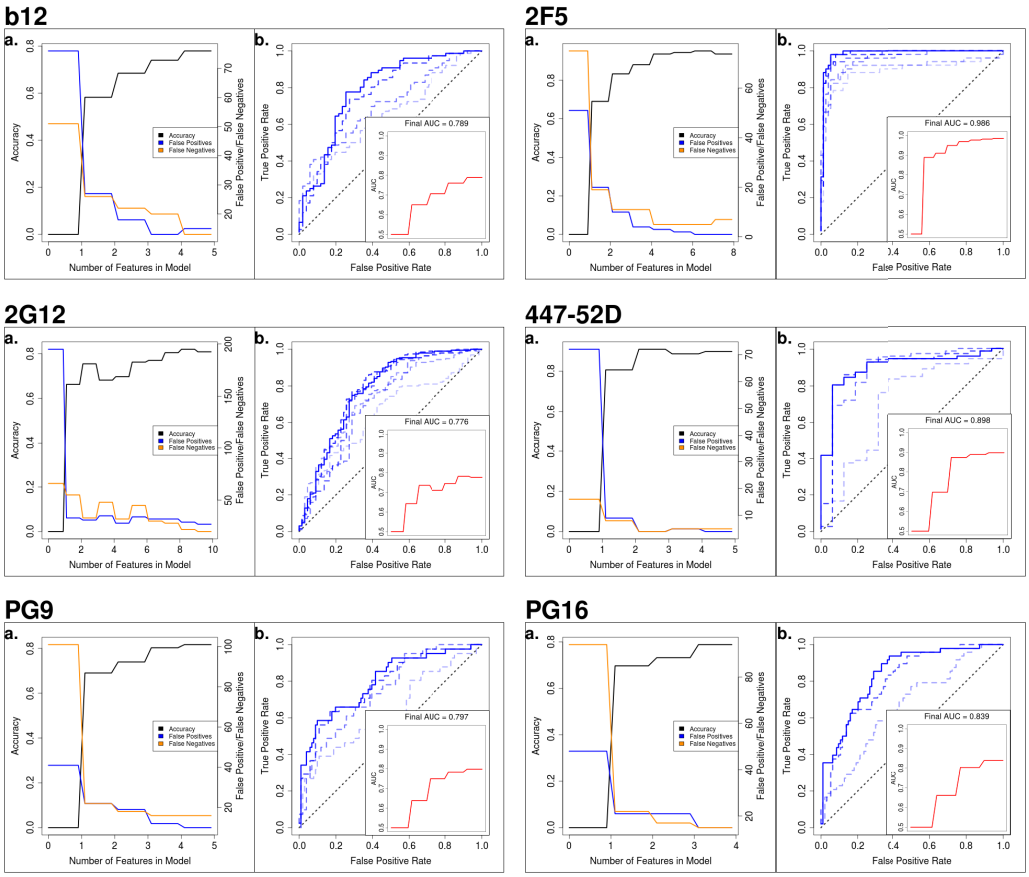
Antibody	**Model Acc**	**Exact Acc**	**Model Spec**	**Exact Spec**	**Model Sens**	**Exact Sens**	**r (Å)**	σ^2 **(Å)**	ν	**Selected Neighborhoods**
b12	0.78	0.65	0.71	0.22	0.83	0.95	12	4	0.10	133, 533, 610, 9
2F5	0.94	0.72	0.99	0.53	0.86	1.00	8	7	0.05	**666**[b], **393**, **642**, **802**, **667**[b], 109, 427
2G12	0.81	NA[a]	0.58	NA	0.89	NA	7	5	0.15	382, **376**[b], **336**[b], **666**, **364**[b], **652**
447-52D	0.90	0.81	0.75	0.88	0.93	0.79	6	8	0.30	**622**, 171, 617, 430
PG9	0.82	NA[a]	0.90	NA	0.61	NA	7	5	0.15	268, 18, 381, **269**
PG16	0.79	NA[a]	0.84	NA	0.69	NA	10	2	0.20	**655**, 627, 38

than 70% for each model, with the exceptions of specificity for 2G12 and sensitivity for PG9 and PG16 (Table 5.2). Addition of new structural neighborhoods to models generally increased accuracy and decreased the number of false positives and false negatives in a 10-fold cross-validation analysis independent of the model selection cross-validation sets (Figure 3). By comparison, for the three antibodies listed with a published linear epitope, — B12, 2F5 and 447-52D — the accuracy of simply searching the primary amino sequences for the exact epitope sequence was 65%, 72% and 80% respectively. In the cases of b12 and 2F5, looking for exact epitope match had a high sensitivity, but rather low specificity for detecting neutralization resistant virus.

For each of the models, ROC curves were generated by allowing the kernel predictor cutoff to vary, and examining the change in the 10-fold cross-validation true positive rate (equivalent to sensitivity) and false positive rate (equivalent to 1-specificity). The estimated area under the ROC curves increased as more sites were included in the models, further indicating that the additions to the model improved predictive capability (Figure 5.3).

5.4 Discussion

Predictive models for NAb neutralization provide multiple insights into the dynamic between the HIV-1 virus and the human adaptive immune system. Here we develop an approach for predicting antibody neutralization that incorporates 3-dimensional structure, is robust to sequence alignment uncertainty, and ultimately classifies sequences as susceptible or resistant with high accuracy. The high accuracy, low false positive and false negative rates, and high areas under ROC curves in a cross-validation setting indicate a strong ability to predict neutralization phenotype. The antibodies examined here include those with simple linear epitopes (2F5, 447-52D), conformational epitopes (b12) as well as those



[Neutralization prediction evaluation]Neutralization prediction evaluation. For each of the 6 antibodies examined (b12, 2F5, 2G12, 447-52D, PG9, PG16), model performance was evaluated in independent cross-validation analyses. a: classification accuracy and number of false positives or false negatives as more structural neighborhoods are added to the model. A sample resistant to neutralization is considered "positive." b. receiver operating characteristic (ROC) curves as more structural neighborhoods are added to the model. Solid blue line is the final model, dotted blue lines are sub-models, the darker the line, the more neighborhoods included. Inset graphs show the area under the ROC curve (AUC) as neighborhoods are added.

that bind Env-associated glycans (2G12) or complex "quaternary epitopes" (PG9, PG16) [44, 43, 42, 45, 46]. The 3-dimensional structural neighborhoods selected to build predictive models of antibody neutralization reflect this complexity and diversity.

Several of the structural neighborhoods selected here overlap known binding epitopes, or have been described elsewhere as neutralization determinants, while others are previously undescribed. The neighborhoods around amino acid 666 and 667 overlap the known 2F5 epitope in the membrane-proximal external region (MPER) of gp41, [44] and are selected here amongst regions predicting 2F5 neutralization. Several of the neighborhoods selected for the 2G12 model also overlap known determinants of 2G12 neutralization. The neighborhood at site 364 overlaps the potential N-linked glycosylation site at position 392, as well as a point mutation at position 470, both described as influential in 2G12 neutralization [45]. Additionally, the neighborhoods centered at sites 336 and 376 both overlap point mutations known to be important for neutralization by 2G12 at residues 334 and 256 respectively [45]. In contrast, the models for b12 and 447-52D neutralization performed well despite containing no neighborhoods that overlap the described epitiopes. This highlights the complex nature of neutralization patterns in Env. Unlike recombinant techniques, such as alanine scanning, our approach assess only patient-derived viable virus variants. It is possible that artificial mutations which abrogate antibody function, but are not found in the wild, differ from the variations that drive real-world diversity in neutralization phenotype.

This distinction between the previously described "cannonical" binding epitopes and the predictors of neutralization determined here is highlighted by the differences between our kernel-based predictions and a simple epitope-based approach. As a simple comparison, we searched for an exact match to the previously described epitopes for b12, 2F5 and 447-52D in the appropriate place in

the Env amino acid sequence. For b12 and 2F5, the simple "exact match" approach achieved very high sensitivity for detecting neutralization resistant virus (95% and 100% respectively), albeit with a low specificity (22% and 53%)(Table 5.2). This indicates that for both of these antibodies, most of the sequences in the sample did not exactly match the canonical epitope regardless of neutralization phenotype. Hence, nearly all resistant viruses had non-canonical sequences in the binding epitope (resulting in high sensitivity), but many sensitive viruses departed from the epitope sequence as well (providing low specificity). This is unsurprising for b12, which has a long epitope (RPVVSTQLLLNGSLAEEEVV at sites 252-271), of which many of the amino acids may not be necessary for binding. In fact, only 15 of the 127 sequences in the b12 sample possessed an exact match, of which 4 were neutralization resistant according to the in-vitro assay. A similar pattern was observed for 2F5, despite its shorter and simpler epitope (ELDKWA at sites 662-667). The canonical epitope for 447-52D is shorter yet (GPXR at sites 312-314), and 29 of 88 samples provided an exact match. Only 16 samples were sensitive to 447-52D, of which 14 matched the defined epitope exactly. This leaves 15 neutralization resistant samples also matching the known epitope, indicating that neutralization may be blocked by variation elsewhere in Env. Despite the moderate specificity (88%), and specificity (79%) achieved for the exact epitope matches in 447-52D (Table 5.2), the high number of "false negatives" for neutralization resistance result in a negative predictive value of only 48%. These examples clearly demonstrate that exact possession of previously described epitope sequences is neither necessary nor sufficient to provide neutralization sensitivity in all cases.

The kernel-based predictors developed here compare favorably to the "exact match" approach. In each of the three cases, the overall accuracy of the kernel predictors is 9-21% higher. For b12 and 2F5, the exact match approach provided a poor balance between specificity and sensitivity, this was much improved for

the kernel-based predictors (Table 2). Of these three antibodies, only the kernel model for 2F5 selected neighborhoods that overlap the known epitope. Predictive models for antibodies without a linear peptide epitope — 2G12, PG9 and PG16 — provided similarly high accuracy, specificity and sensitivity, with accuracy near or above 80%, and only the specificity of predicting 2G12 resistance falling below 60% (Table 5.2). In all cases, the neighborhoods selected included regions not thought to directly bind the relevant NAb. There are several possible reasons that non-binding features predict neutralization sensitivity well for these antibodies. It is possible that these neighborhoods effect NAb action through steric or conformational effects, they may be compensatory mutations that restore viral fitness impacted by adaptations more directly associated with resistance elsewhere in the protein, or they may be associated with neutralization phenotypes in a non-mechanistic manner, for instance due to a founder effect in the viral ancestry.

These models were built iteratively by adding individual neighborhoods one at a time, therefore model performance across this process can be assessed. For each of the models, area under the cross-validation receiver operating characteristic (ROC) curve increased as more sites were added to the model (Figure reffig:nabroc). These curves, along with accuracy and false positive and false negative rates were calculated based on an independent round of 10-fold cross-validation. As cross-validation was also used to select neighborhoods and tune the models, re-randomizing the cross-validation sets was important to ensure that the performance observed was not dependent on any idiosyncrasies of the original cross-validation sets. Two of the antibodies, 2G12 and 447-52D, were particularly unbalanced between sensitive and resistant groups. In these cases, a high prediction accuracy could be achieved by simply predicting all viruses will be resistant. It is important to note that despite this fact, low false negative rates were achieved for both of these models (Figure 5.3), indicating that these models did

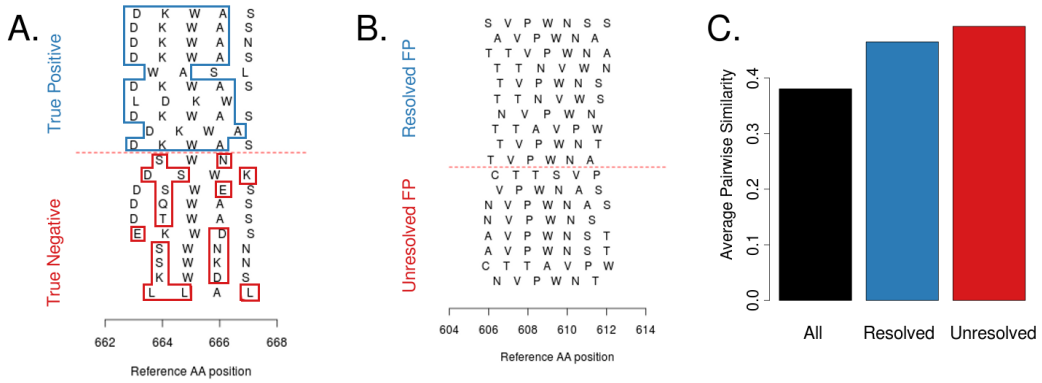


Figure 5.3: Two examples of neighborhoods that improve neutralization prediction. A. 20 randomly-chosen correctly-classified samples from the first neighborhood (666) in the 2F5 model, 10 sensitive and 10 resistant. Residues outlined in blue are consistent with an exact match to the published epitope (ELDKWA), those in red are inconsistent with an exact match. B. Previously false positive samples either resolved or not resolved by the addition of the third neighborhood (610) in the b12 model. C. Average pairwise similarities for the resolved, pooled and unresolved samples from the third site in the b12 model.

not fall prey to this particular pitfall.

A closer examination of two of the neighborhoods selected provides additional insight into the types of patterns recognized by this approach. The first site added to the 2F5 model (centered at site 666) alone provided an accuracy of 69%. This neighborhood overlaps the known epitope for 2F5, and largely categorizes virus on the basis of match to the 2F5 epitope (Figure 5.3A). In this sense, when a selected neighborhood overlaps the known epitope, it may partially recapitulate the behavior of the "exact match" approach. In contrast, the third neighborhood incorporated into the b12 model matches no known determinant of antibody neutralization, yet its inclusion increases classification accuracy from 69% to 74%, increases the area under the ROC curve from 0.71 to 0.76, and decreases the number of false positives (ie. samples scored as resistant while actually sensitive) from 18 to 13. Note, 10 false positives were resolved at this step, while 5 new false positives were created, for a net decrease of 5. Stratifying the samples previously scored incorrectly as resistant into those resolved by the new neighborhood, and those not resolved reveals no obvious pattern (Figure 5.3B). However, when the average kernel (similarity) score for the resolved and unre-

solved groups is calculated, both are greater than the average kernel score for both groups pooled together. (Figure 5.3C) This indicates that the samples resolved by the additional information share some set of characteristics in common, as do those unresolved, although the nature of these characteristics is not apparent upon inspection. Furthermore, in both of these examples the mapping of sample sequences onto 3-dimensional structure causes the amino acids not to align on a rigid grid, as they would in multiple sequence alignment, due to handling of indels. It should be noted that any amino acid matches, even those not precisely aligned vertically will still contribute to the overall similarity score. This flexibility allows this approach to detect determinants of neutralization, even in the face of considerable alignment uncertainty.

Locations of selected structural neighborhoods on the 3-dimensional structure of Env (where structural data is available) or on the primary Env amino acid sequence (where structure is not solved) reveal regions of the protein which influence neutralization (Figure 5.4). Neighborhoods that predict neutralization by 2F5 (393), 2G12 (336, 364, 376, 382) and PG9 (268, 269, 381) cluster on the relatively exposed outer domain of gp120 (Figure 5.4). These neighborhoods are largely centered in the C2, C3 and V4 regions of gp120, and also overlap the α2 helix (336) and CD4 binding site (364). Several residues included in these neighborhoods have been described as influential in two studies of neutralization of subtype C virus with polyclonal sera from Rong et al, and Kulkarni et al [53, 54]. These include residue 270, overlapped by PG9 neighborhood 269, residue 393, overlapped by 2F5 neighborhood 393 and 2G12 neighborhood 364, and residues 335-337 overlapped by 2G12 neighborhood 336. Two other neighborhoods included in the 2F5 model, centered at residues 109 and 427 are far apart in the primary sequence, but actually sit very close to each other in the 1 helix and bridging sheet, respectively. To our knowledge, these residues have not been previously described as relevant to NAb neutralization, however, the 1 helix lies just distal to

the cluster of gp41-interacting residues in the gp120 inner domain, providing a possible link between this amino acid cluster and the 2F5 epitope in gp41.

Models for neutralization by b12, 2F5, 2G12, 447-52D and PG16 all included neighborhoods in the extracellular region of the gp41 peptide. The epitope for 2F5 is located in gp41, and PG16 is thought to only bind the properly assembled Env protein (but not soluble gp120) [46], providing explanations for the importance of gp41 in the function of these antibodies. The mechanism for gp41 modulation of neutralization by b12, 2G12 and 447-52D remains less clear. It is possible that changes in the conformation of gp120 could require compensatory changes in gp41 or vice-versa via non-covalent interactions. Several of the gp41 neighborhoods implicated in neutralization by NAb in our models contain sites also found to be important in neutralizing by polyclonal sera by Kulkarni et al [54]. These include residue 644, overlapping 2F5 neighborhood 642, residue 651 overlapping PG16 neighborhood 655 and 2G12 neighborhood 642 and residue 667, overlapping 2G12 neighborhood 666 (as well as 2F5 neighborhood 666, although this is already explained by overlap with the 2F5 epitope). The model for 2F5 also includes one neighborhood from the gp41 intracellular domain centered at residue 802. This neighborhood also contains a residue identified in the Kulkarni et al as influential study at site 803.

While many of the structural regions of Env identified in this study as important for NAb neutralization have known functions or overlap with previous results, some results and notable omissions were surprising, and warrant future investigation. Our model for b12 neutralization did not overlap any residues in or around the known epitope or CD4 binding site, and also had no sites in common with alanine scans for determinants of b12 neutralization or other neutralization panel-based studies of b12[50, 55]. Despite this, our model for b12 neutralization provides an independent-run cross-validation accuracy of 78% with specificity and sensitivity of 70% and 82%, respectively. This performance is on a par with

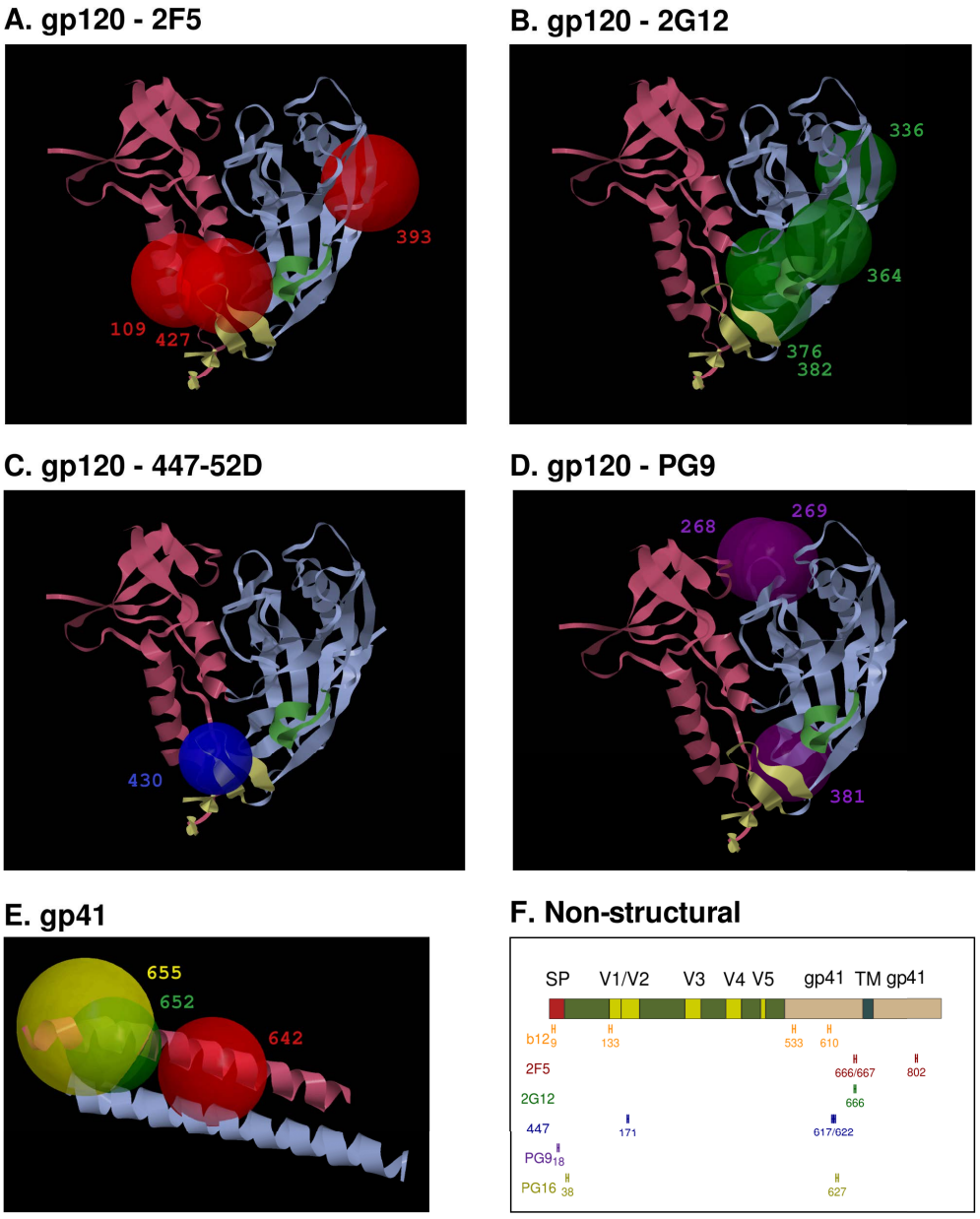


Figure 5.4: Locations of predictive structural neighborhoods in HIV-1 envelope. 3D ribbon diagrams of gp120 are annotated by color. Red ribbons represent the inner domain, blue ribbons the outer domain, yellow ribbons the bridging sheet, and green ribbons the CD4 binding site. Colored spheres represent the structural neighborhoods selected for the final models. Spheres are centered at the alpha-carbon for the residue indicated by the numeric labels, and have radii as determined by model tuning (Table 2). Selected neighborhoods outside of the solved 3D structure are indicated on the non-structural regions schematic of the envelope polypeptide.

model performance reported elsewhere.[Gnanakaran, 2010] Two models, for antibodies PG9 and b12, contain sites in the Env signal peptide, which is cleaved from the expressed protein. The possible significance of these sites is unclear. Co-evolution with regions expressed on the viral surface provide one possible explanation for their apparent significance. Finally, only one neighborhood in the V2 variable loop (centered at residue 171 in the 447-52D) was identified in this analysis. The V2 loop has been extensively described as important for development NAb resistance, but is not strongly associated with neutralization in this study [33, 34].

This technique is naturally limited by the information provided by the analyzed sequences. A site or neighborhood in Env may be vital to antibody binding, but if it does not vary in the model training data, it will not be included in the selected model. In this sense, the neighborhoods selected here do not represent an exhaustive list of regions in Env important for important in NAb neutralization, but rather an estimate of the regions driving variations in neutralization in this sample. The approach described here makes several assumptions that may be relaxed by future improvements in the algorithm. We assume here that single values for the model tuning parameters (r and σ^2) are suitable for a neutralization model. The "greedy" neighborhood selection algorithm may identify a local optimum, and is not guaranteed to locate a globally ideal solution. Site-to-site variation in tuning parameters, or a more thorough model selection approach could be implemented, but would incur higher computational costs.

The structural-neighborhood kernel approach outlined here can be applied to many problems where 3-dimensional structure or protein alignment uncertainty is likely to be important. Tertiary protein structure may influence a variety of protein features and functions, including protein-protein interaction, ligand binding and drug resistance. In any case where the outcome of interest can be described as a binary variable (in this case neutralization sensitivity or resis-

tance), this approach can be applied directly. In situations where a continuous outcome variable is desired, the same kernel function could be used in conjunction with support vector regression.

The structural neighborhood kernel approach outlined here compares proteins on the basis of their similarities in discrete 3-dimensional regions. This method is designed on the expectation that outcome depends on a combination of protein motifs, which may or may not be only apparent in their real-world 3-dimensional arrangement. This harnesses structural information, when available, to provide a more realistic portrait of protein function than can be achieved by looking only at primary amino acid sequence. An added advantage of this approach is the alignment flexibility it provides. Amino acids offset slightly in space will still contribute to the kernel score, enabling analysis of highly-variable regions with low alignment confidence — regions that are often removed prior to analysis.[Rong, 2007], [Gnanakaran, 2010] The application of this method to *in vitro* NAb neutralization data created models that predict neutralization sensitivity and resistance with high accuracy, and in most cases high sensitivity and specificity. Interestingly, for several of the antibodies examined, determinants of neutralization clustered in the gp41 stem and on the outer domain of gp120. These sites largely overlap with previously described determinants of neutralization with polyclonal sera [53, 54]. This raises the possibility that antibodies with wildly varying binding locations may be escaped from by common pathways in these domains. This possibility warrants further study. As continued progress is made in the discovery of novel NAbs, [47, 48, 112] we hope that this approach will provide an efficient and informative tool for the characterization of neutralization.

Chapter 5, in part is currently being prepared for submission for publication of the material. Lorne W. Walker, N. Lance Hepler, Dennis Burton, Douglas Richman, Sergei L. Kosakovsky Pond. The dissertation author was the primary

author of this work.

Chapter 6: Conclusions

In this work we have examined the use of kernel-based Machine Learning methods for modeling two systems in HIV biology. First, we described the Subset-Tree kernel, and used it to describe ARV regimens. The SST-kernel integrates multiple layers of drug regimen information, including the proteins targeted, drug mechanisms (i.e. ARV drug class) involved, and individual drugs included. Second, we defined a novel kernel function for analysis of 3-dimensional protein structure. This tool allowed us to develop models to predict antibody neutralization and identify likely real-world determinants of neutralization phenotype. These approaches sharpen our understanding of two of the key methods by which the HIV-1 virus may be impeded. Pharmacological control of viral replication has been the most successful anti-HIV intervention to date, while antibody neutralization is one of the immune system's main weapons against HIV-1 infection, and is one basis for hope for an effective vaccine.

In Chapter 2 we introduced the Subset-tree kernel, and described how it is calculated. In the SST-kernel application to drug regimen, a particular drug combination is first expressed as a tree. The various layers of nodes in these trees describe many of the relevant features of drug regimens, including individual drugs, drug classes, the number of included drugs per class, and drug protein targets. By integrating this information into a single framework, we can then use this tool to model the implications of the use of various drug combinations, without making *a priori* assumptions about the most relevant features to a particular problem.

We then applied this tool to a set of linear and support vector machine models describing the differences between various drug regimens in terms of average treatment duration, rate of success in suppressing viral replication, and rate of eventual virologic failure. In these cases, the SST-kernel provided better model fit, and smaller residual error than non-SST kernel alternatives.

In the third chapter, we developed a set of models estimating the risk of virologic failure of ARV regimens. By including a model that was free to include both kernel (principal components) and non-kernel covariates, we observed an interesting and important interaction between these variables. The first kernel principal component was strongly associated with the tradeoff between NNRTI and PI drugs, essentially acting as a switch between the two. This variable, therefore, is an expression of one of the most common clinical choices — whether to include an NNRTI or PI with an NRTI backbone. When included with other non-kernel covariates (in this case, the number of NNRTIs included in a regimen), the coefficient for PC1 changed drastically, as NNRTI use was already controlled for in another variable. Ultimately, we saw that kernel-covariate containing models provided modestly better performance than non-kerneo models. Also in Chapter 3 we examined the use of the kernel (i.e. similarity) score between the current and past regimens as an additional predictor of virologic failure. While these kernel-based previous-ARV covariates did not out-perform non-kernel options, they were able to provide similarly powerful predictions, while only describing past regimen use with a single variable.

In Chapter 4 we extended the SST-kernel approach to look at a series of regimens used over time, instead of a single combination. While we hoped that this technique would improve models of risk for progression to AIDS or death, results for this model were mixed. Whether this is because the “Linked SST-kernel” derrived variables are simply not good predictors of outcome, or if this is due to the sort of variable-interaction we saw in Chapter 3 remains unclear. Despite

this, our prognostic model for disease progression provided strong predictions, with a better than 0.80 ROC AUC for predicting 1-year progression risk. Many of the regimen history features that may be of interest to disease progression, including regimen duration, and prior on-therapy failure were not included in this kernel function, and may provide an opportunity for improvements.

Chapter 5 focused on the application of a novel kernel function to predicting antibody neutralization. Here, the kernel function incorporates what is known of the 3-dimensional Env structure, and defines "structural neighborhoods" that are taken to be the functional units of protein-protein interaction. Any amino acid match within this neighborhood contributes to the kernel score, allowing a degree of flexibility in light of the very high diversity of Env, and resulting sequence alignment uncertainty. This approach was applied to six monoclonal neutralizing antibodies. For many of these antibodies, models with high accuracy, specificity and sensitivity were derived. Interestingly, many of the most important determinants of neutralization in these real-world samples were outside of, and often quite distant from the described epitopes. Here we provide a novel approach to the computational determination of influential structural neighborhoods based only on amino acid sequence and observed neutralization titer. Interestingly, many of the neighborhoods described have been described elsewhere in the context of polyclonal sera neutralization, however there are also several notable exceptions. Sites in the V1/V2 loop, and known epitopes for b12 and 447-52D were not identified as important by this analysis. Whether this implies that intra-epitope variation is simply not the mechanism by which HIV-1 escapes from these antibodies, or if further refinement of this technique would detect eiptiope-associated signals remains an interesting question.

The vast majority of this work was done using command-line packages, such as libsvm and svm-light-TK or custom-programmed solutions in R and Python [113, 114, 98, 115]. While these tools can be powerful, this degree of compu-

tational and programming inconvenience may not be acceptable to many scientists and clinicians. Graphical-interface software for data mining and machine learning is available. Although they afford less flexibility than programming languages like Python and R, they provide these powerful tools to a much broader audience. Commercial packages, such as SAS, Statistica and SPSS are well-established powerful statistical platforms [116, 117, 118]. More recently, free and open source machine learning suites, such as Weka, Orange and Shogun have become avalible [119, 120, 121]. While open source software occasionally lacks the stability of commercial products, the active (and often academic) developer bases make for appealing research tools.

Here we have shown several examples of how kernel-based machine learning can be applied to HIV-1 research. These methods have allowed us to improve the prediction of several HIV-1 outcomes, including virologic failure of ARV regimens, and antibody neutralization titer. The structural neighborhoods selected for antibody neutralization prediction in Chapter 5 also shed light on what Env features may be associated with antibody escape *in vivo*. In an era of rapidly increasing computational capacity, machine learning techniques are increasingly employed in fields as far-ranging as credit scores and movie recommendations [122, 123]. As the size and complexity of biological and HIV-1 data sources continues to increase, these tools may help bring new light to problems of great importance to human health and well-being.

Bibliography

[1] Martin Hilbert and Priscila López. The worlds technological capacity to store, communicate, and compute information. *Science*, 332(6025):60–65, 2011.

[2] Margaret A. Fischl, Douglas D. Richman, Michael H. Grieco, Michael S. Gottlieb, Paul A. Volberding, Oscar L. Laskin, John M. Leedom, Jerome E. Groopman, Donna Mildvan, Robert T. Schooley, George G. Jackson, David T. Durack, and Dannie King. The efficacy of Azidothymidine (AZT) in the treatment of patients with AIDS and AIDS-related complex. *New England Journal of Medicine*, 317(4):185–191, 1987.

[3] Ard van Sighem, Luuk Gras, Peter Reiss, Kees Brinkman, Frank de Wolf, and on behalf of the ATHENA national observational cohort study. Life expectancy of recently diagnosed asymptomatic HIV-infected patients approaches that of uninfected individuals. *AIDS*, 24(10):1527–1535, 2010.

[4] Douglas D. Richman, David M. Margolis, Martin Delaney, Warner C. Greene, Daria Hazuda, and Roger J. Pomerantz. The challenge of finding a cure for HIV infection. *Science*, 323(5919):1304–1307, 2009.

[5] M. Juliana McElrath and Barton F. Haynes. Induction of immunity to Human Immunodeficiency Virus Type-1 by vaccination. *Immunity*, 33(4):542–554, 2010.

[6] Mari M Kitahata, Benigno Rodriguez, Richard Haubrich, Stephen Boswell, W Christopher Mathews, Michael M Lederman, William B Lober, Stephen E Van Rompaey, Heidi M Crane, Richard D Moore, Michael Bertram, James O Kahn, and Michael S Saag. Cohort profile: the Centers for AIDS Research Network of Integrated Clinical Systems. *International Journal of Epidemiology*, 37(5):948–955, 2008.

[7] CNICS: CFAR network of integrated clinical sites. http://www.cnics.net/, 2011.

[8] Christos J. Petropoulos, Neil T. Parkin, Kay L. Limoli, Yolanda S. Lie, Terri Wrin, Wei Huang, Huan Tian, Douglas Smith, Genine A. Winslow, Daniel J. Capon, and Jeannette M. Whitcomb. A novel phenotypic drug susceptibility assay for Human Immunodeficiency Virus Type 1. *Antimicrob. Agents Chemother.*, 44(4):920–928, 2000.

[9] Gary K. Geiss, Roger E. Bumgarner, Mahru C. An, Michael B. Agy, Angélique B. van 't Wout, Erick Hammersmark, Victoria S. Carter, David Upchurch, James I. Mullins, and Michael G. Katze. Large-scale monitoring of host cell gene expression during HIV-1 infection using cDNA microarrays. *Virology*, 266(1):8–16, 2000.

[10] Chunlin Wang, Yumi Mitsuya, Baback Gharizadeh, Mostafa Ronaghi, and Robert W. Shafer. Characterization of mutation spectra with ultra-deep pyrosequencing: application to HIV-1 drug resistance. *Genome Res.*, 17(8):1195–1201, 2007.

[11] Jacques Fellay, Kevin V. Shianna, Dongliang Ge, Sara Colombo, Bruno Ledergerber, Mike Weale, Kunlin Zhang, Curtis Gumbs, Antonella Castagna, Andrea Cossarizza, Alessandro Cozzi-Lepri, Andrea De Luca, Philippa Easterbrook, Patrick Francioli, Simon Mallal, Javier Martinez-Picado, José M. Miro, Niels Obel, Jason P. Smith, Josiane Wyniger, Patrick Descombes, Stylianos E. Antonarakis, Norman L. Letvin, Andrew J. McMichael, Barton F. Haynes, Amalio Telenti, and David B. Goldstein. A whole-genome association study of major determinants for host control of HIV-1. *Science*, 317(5840):944–947, 2007.

[12] Ethem Alpaydin. *Introduction to Machine Learning (Adaptive Computation and Machine Learning)*. The MIT Press, Cambridge, MA, USA, 2004.

[13] Pneumocystis pneumonia – Los Angeles. *Morb Mortal Wkly Rep.*, 30(21):250–252, 1981.

[14] JL Marx. New disease baffles medical community. *Science*, 217(4560):618–621, 1982.

[15] F Barre-Sinoussi, JC Chermann, F Rey, MT 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.

[16] RC Gallo, SZ Salahuddin, M Popovic, GM Shearer, M Kaplan, BF Haynes, TJ Palker, R Redfield, J Oleske, B Safai, and al. et. Frequent detection and isolation of cytopathic retroviruses (HTLV-III) from patients with AIDS and at risk for AIDS. *Science*, 224(4648):500–503, 1984.

[17] UNAIDS. UNAIDS report on the global AIDS epidemic 2010. http://www.unaids.org/globalreport/global_report.htm, 2010.

[18] Melanie A. Thompson, Judith A. Aberg, Pedro Cahn, Julio S. G. Montaner, Giuliano Rizzardini, Amalio Telenti, José M. Gatell, Huldrych F. Günthard, Scott M. Hammer, Martin S. Hirsch, Donna M. Jacobsen, Peter Reiss, Douglas D. Richman, Paul A. Volberding, Patrick Yeni, and Robert T. Schooley.

Antiretroviral treatment of adult HIV infection. *JAMA: The Journal of the American Medical Association*, 304(3):321–333, 2010.

[19] Jan Balzarini, Heidi Pelemans, Anna Karlsson, Erik DeClercq, and Jörg-Peter Kleim. Concomitant combination therapy for HIV infection preferable over sequential therapy with 3TC and non-nucleoside reverse transcriptaseinhibitors. *Proceedings of the National Academy of Sciences*, 93(23):13152–13157, 1996.

[20] Ann C. Collier, Robert W. Coombs, David A. Schoenfeld, Roland L. Bassett, Joseph Timpone, Alice Baruch, Michelle Jones, Karen Facey, Caroline Whitacre, Vincent J. McAuliffe, Harvey M. Friedman, Thomas C. Merigan, Richard C. Reichman, Carol Hooper, and Lawrence Corey. Treatment of Human Immunodeficiency Virus infection with Saquinavir, Zidovudine, and Zalcitabine. *New England Journal of Medicine*, 334(16):1011–1018, 1996.

[21] Andrew Carr and David A. Cooper. HIV protease inhibitors. *AIDS*, 10:S151–158, 1996.

[22] Scott M. Hammer, Kathleen E. Squires, Michael D. Hughes, Janet M. Grimes, Lisa M. Demeter, Judith S. Currier, Joseph J. Eron, Judith E. Feinberg, Henry H. Balfour, Lawrence R. Deyton, Jeffrey A. Chodakewitz, Margaret A. Fischl, John P. Phair, Louise Pedneault, Bach-Yen Nguyen, and Jon C. Cook. A controlled trial of two nucleoside analogues plus Indinavir in persons with Human Immunodeficiency Virus infection and CD4 cell counts of 200 per cubic millimeter or less. *New England Journal of Medicine*, 337(11):725–733, 1997.

[23] Erik De Clercq. The role of non-nucleoside reverse transcriptase inhibitors (NNRTIs) in the therapy of HIV-1 infection. *Antiviral Research*, 38(3):153–179, 1998.

[24] Johnson VA, Brun-Vezinet F, Clotet B, Gunthard HF, Kuritzkes DR, Pillay D, Schapiro JM, and Richman DD. Update of the drug resistance mutations in HIV-1: December 2009. *Top HIV Med.*, 17(5):138–145, 2009.

[25] GJ Moyle and D Back. Principles and practice of HIV-protease inhibitor pharmacoenhancement. *HIV Medicine*, 2(2):105–113, 2001.

[26] Mark A. Boyd and Andrew M. Hill. Clinical management of treatment-experienced, HIV/AIDS patients in the combination antiretroviral therapy era. *PharmacoEconomics*, 28:17–34, 2010.

[27] Steven G. Deeks, Stephen J. Gange, Mari M. Kitahata, Michael S. Saag, Amy C. Justice, Robert S. Hogg, Joseph J. Eron, John T. Brooks, Sean B. Rourke, M. John Gill, Ronald J. Bosch, Constance A. Benson, Ann C. Collier, Jeffrey N. Martin, Marina B. Klein, Lisa P. Jacobson, Benigno Rodriguez,

Timothy R. Sterling, Gregory D. Kirk, Sonia Napravnik, Anita R. Rachlis, Liviana M. Calzavara, Michael A. Horberg, Michael J. Silverberg, Kelly A. Gebo, Margot B. Kushel, James J. Goedert, Rosemary G. McKaig, Richard D. Moore, North American AIDS Cohort Collaboration on Research, and Design. Trends in multidrug treatment failure and subsequent mortality among antiretroviral therapy-experienced patients with HIV infection in North America. *Clinical Infectious Diseases*, 49(10):1582–1590, 2009.

[28] The rgp120 HIV Vaccine Study Group. Placebo-controlled phase 3 trial of a recombinant glycoprotein 120 vaccine to prevent HIV-1 infection. *Journal of Infectious Diseases*, 191(5):654–665, 2005.

[29] Punnee Pitisuttithum, Peter Gilbert, Marc Gurwith, William Heyward, Michael Martin, Fritz vanGriensven, Dale Hu, JordanW. Tappero, and Bangkok Vaccine Evaluation Group. Randomized, doubleblind, placebo-controlled efficacy trial of a bivalent recombinant glycoprotein 120 HIV1 vaccine among injection drug users in Bangkok, Thailand. *Journal of Infectious Diseases*, 194(12):1661–1671, 2006.

[30] Susan P. Buchbinder, Devan V. Mehrotra, Ann Duerr, Daniel W. Fitzgerald, Robin Mogg, David Li, Peter B. Gilbert, Javier R. Lama, Michael Marmor, Carlos del Rio, M. Juliana McElrath, Danilo R. Casimiro, Keith M. Gottesdiener, Jeffrey A. Chodakewitz, Lawrence Corey, and Michael N. Robertson. Efficacy assessment of a cell-mediated immunity HIV-1 vaccine (the Step Study): a double-blind, randomised, placebo-controlled, test-of-concept trial. *The Lancet*, 372(9653):1881–1893, Nov 2008.

[31] Supachai Rerks-Ngarm, Punnee Pitisuttithum, Sorachai Nitayaphan, Jaranit Kaewkungwal, Joseph Chiu, Robert Paris, Nakorn Premsri, Chawetsan Namwat, Mark de Souza, Elizabeth Adams, Michael Benenson, Sanjay Gurunathan, Jim Tartaglia, John G. McNeil, Donald P. Francis, Donald Stablein, Deborah L. Birx, Supamit Chunsuttiwat, Chirasak Khamboonruang, Prasert Thongcharoen, Merlin L. Robb, Nelson L. Michael, Prayura Kunasol, and Jerome H. Kim. Vaccination with ALVAC and AIDSVAX to prevent HIV-1 infection in Thailand. *New England Journal of Medicine*, 361(23):2209–2220, 2009.

[32] Richard Wyatt, Peter D. Kwong, Elizabeth Desjardins, Raymond W. Sweet, James Robinson, Wayne A. Hendrickson, and Joseph G. Sodroski. The antigenic structure of the HIV gp120 envleope glycoprotein. *Nature*, 393(6686):705–711, Jun 1998.

[33] Amy Ly and Leonidas Stamatatos. V2 loop glycosylation of the Human Immunodeficiency Virus Type 1 SF162 Envelope facilitates interaction of this protein with CD4 and CCR5 receptors and protects the virus from neutralization by anti-V3 loop and anti-CD4 binding site antibodies. *J. Virol.*, 74(15):6769–6776, 2000.

[34] Britt Losman, Anders Bolmstedt, Kristian Schønning, Åsa Björndal, Charlotta Westin, Eva Maria Fenyö, and Sigvard Olofsson. Protection of neutralization epitopes in the V3 loop of oligomeric Human Immunodeficiency Virus Type 1 glycoprotein 120 by N-linked oligosaccharides in the V1 region. *AIDS Research and Human Retroviruses*, 17(11):1067–1076, 2001.

[35] Brian Gaschen, Jesse Taylor, Karina Yusim, Brian Foley, Feng Gao, Dorothy Lang, Vladimir Novitsky, Barton Haynes, Beatrice H. Hahn, Tanmoy Bhattacharya, and Bette Korber. Diversity considerations in HIV-1 vaccine selection. *Science*, 296(5577):2354–2360, 2002.

[36] Dennis R. Burton, Ronald C. Desrosiers, Robert W. Doms, Wayne C. Koff, Peter D. Kwong, John P. Moore, Gary J. Nabel, Joseph Sodroski, Ian A. Wilson, and Richard T. Wyatt. HIV vaccine design and the neutralizing antibody problem. *Nat Immunol*, 5(3):233–236, Mar 2004.

[37] Barton F Haynes and David C Montefiori. Aiming to induce broadly reactive neutralizing antibody responses with HIV-1 vaccine candidates. *Expert Review of Vaccines*, 5(3):347–363, 2006.

[38] Xiping Wei, Julie M. Decker, Shuyi Wang, Huxiong Hui, John C. Kappes, Xiaoyun Wu, Jesus F. Salazar-Gonzalez, Maria G. Salazar, J. Michael Kilby, Michael S. Saag, Natalia L. Komarova, Martin A. Nowak, Beatrice H. Hahn, Peter D. Kwong, and George M. Shaw. Antibody neutralization and escape by HIV-1. *Nature*, 422(6929):307–312, Mar 2003.

[39] Douglas D. Richman, Terri Wrin, Susan J. Little, and Christos J. Petropoulos. Rapid evolution of the neutralizing antibody response to HIV Type 1 infection. *Proceedings of the National Academy of Sciences*, 100(7):4144–4149, 2003.

[40] Simon D. W. Frost, Terri Wrin, Davey M. Smith, Sergei L. Kosakovsky Pond, Yang Liu, Ellen Paxinos, Colombe Chappey, Justin Galovich, Jeff Beauchaine, Christos J. Petropoulos, Susan J. Little, and Douglas D. Richman. Neutralizing antibody responses drive the evolution of Human Immunodeficiency Virus Type 1 envleope during recent HIV infection. *Proceedings of the National Academy of Sciences of the United States of America*, 102(51):18514–18519, 2005.

[41] C Cheng-Mayer, J Homsy, L A Evans, and J A Levy. Identification of Human Immunodeficiency Virus subtypes with distinct patterns of sensitivity to serum neutralization. *Proceedings of the National Academy of Sciences*, 85(8):2815–2819, 1988.

[42] D R Burton, C F Barbas, M A Persson, S Koenig, R M Chanock, and R A Lerner. A large array of human monoclonal antibodies to Type 1 Human Immunodeficiency Virus from combinatorial libraries of asymptomatic seropositive individuals. *Proceedings of the National Academy of Sciences*, 88(22):10134–10137, 1991.

[43] M K Gorny, A J Conley, S Karwowska, A Buchbinder, J Y Xu, E A Emini, S Koenig, and S Zolla-Pazner. Neutralization of diverse Human Immunodeficiency Virus Type 1 variants by an anti-V3 human monoclonal antibody. *J. Virol.*, 66(12):7538–7542, 1992.

[44] T Muster, F Steindl, M Purtscher, A Trkola, A Klima, G Himmler, F Ruker, and H Katinger. A conserved neutralizing epitope on gp41 of Human Immunodeficiency Virus Type 1. *J. Virol.*, 67(11):6642–6647, 1993.

[45] A Trkola, M Purtscher, T Muster, C Ballaun, A Buchacher, N Sullivan, K Srinivasan, J Sodroski, JP Moore, and H Katinger. Human monoclonal antibody 2G12 defines a distinctive neutralization epitope on the gp120 glycoprotein of Human Immunodeficiency Virus Type 1. *J. Virol.*, 70(2):1100–1108, 1996.

[46] Laura M. Walker, Sanjay K. Phogat, Po-Ying Chan-Hui, Denise Wagner, Pham Phung, Julie L. Goss, Terri Wrin, Melissa D. Simek, Steven Fling, Jennifer L. Mitcham, Jennifer K. Lehrman, Frances H. Priddy, Ole A. Olsen, Steven M. Frey, Phillip W. Hammond, Protocol G Principal Investigators, Stephen Kaminsky, Timothy Zamb, Matthew Moyle, Wayne C. Koff, Pascal Poignard, and Dennis R. Burton. Broad and potent neutralizing antibodies from an african donor reveal a new HIV-1 vaccine target. *Science*, 326(5950):285–289, 2009.

[47] Xueling Wu, Zhi-Yong Yang, Yuxing Li, Carl-Magnus Hogerkorp, William R. Schief, Michael S. Seaman, Tongqing Zhou, Stephen D. Schmidt, Lan Wu, Ling Xu, Nancy S. Longo, Krisha McKee, Sijy ODell, Mark K. Louder, Diane L. Wycuff, Yu Feng, Martha Nason, Nicole Doria-Rose, Mark Connors, Peter D. Kwong, Mario Roederer, Richard T. Wyatt, Gary J. Nabel, and John R. Mascola. Rational design of envleope identifies broadly neutralizing human monoclonal antibodies to HIV-1. *Science*, 329(5993):856–861, 2010.

[48] Tongqing Zhou, Ivelin Georgiev, Xueling Wu, Zhi-Yong Yang, Kaifan Dai, Andrés Finzi, Young Do Kwon, Johannes F. Scheid, Wei Shi, Ling Xu, Yongping Yang, Jiang Zhu, Michel C. Nussenzweig, Joseph Sodroski, Lawrence Shapiro, Gary J. Nabel, John R. Mascola, and Peter D. Kwong. Structural basis for broad and potent neutralization of HIV-1 by antibody VRC01. *Science*, 329(5993):811–817, 2010.

[49] James M. Binley, Elizabeth A. Lybarger, Emma T. Crooks, Michael S. Seaman, Elin Gray, Katie L. Davis, Julie M. Decker, Diane Wycuff, Linda Harris, Natalie Hawkins, Blake Wood, Cory Nathe, Douglas Richman, Georgia D. Tomaras, Frederic Bibollet-Ruche, James E. Robinson, Lynn Morris, George M. Shaw, David C. Montefiori, and John R. Mascola. Profiling the specificity of neutralizing antibodies in a large panel of plasmas from patients chronically infected with Human Immunodeficiency Virus Type 1 subtypes b and c. *J. Virol.*, 82(23):11651–11668, 2008.

[50] Ralph Pantophlet, Erica Ollmann Saphire, Pascal Poignard, Paul W. H. I. Parren, Ian A. Wilson, and Dennis R. Burton. Fine mapping of the interaction of neutralizing and nonneutralizing monoclonal antibodies with the CD4 binding site of Human Immunodeficiency Virus Type 1 gp120. *J. Virol.*, 77(1):642–658, 2003.

[51] Michael B. Zwick, Richard Jensen, Sarah Church, Meng Wang, Gabriela Stiegler, Renate Kunert, Hermann Katinger, and Dennis R. Burton. Anti-Human Immunodeficiency Virus Type 1 (HIV-1) antibodies 2F5 and 4E10 require surprisingly few crucial residues in the membrane-proximal external region of glycoprotein gp41 to neutralize HIV-1. *J. Virol.*, 79(2):1252–1261, 2005.

[52] Sara M. O'Rourke, Becky Schweighardt, Pham Phung, Dora P. A. J. Fonseca, Karianne Terry, Terri Wrin, Faruk Sinangil, and Phillip W. Berman. Mutation at a single position in the V2 domain of the HIV-1 envleope protein confers neutralization sensitivity to a highly neutralization-resistant virus. *J. Virol.*, 84(21):11200–11209, 2010.

[53] Rong Rong, S. Gnanakaran, Julie M. Decker, Frederic Bibollet-Ruche, Jesse Taylor, Jeffrey N. Sfakianos, John L. Mokili, Mark Muldoon, Joseph Mulenga, Susan Allen, Beatrice H. Hahn, George M. Shaw, Jerry L. Blackwell, Bette T. Korber, Eric Hunter, and Cynthia A. Derdeyn. Unique mutational patterns in the envleope alpha2 amphipathic helix and acquisition of length in gp120 hypervariable domains are associated with resistance to autologous neutralization of subtype C Human Immunodeficiency Virus Type 1. *J. Virol.*, 81(11):5658–5668, 2007.

[54] Smita S. Kulkarni, Alan Lapedes, Haili Tang, S. Gnanakaran, Marcus G. Daniels, Ming Zhang, Tanmoy Bhattacharya, Ming Li, Victoria R. Polonis, Francine E. McCutchan, Lynn Morris, Dennis Ellenberger, Salvatore T. Butera, Robert C. Bollinger, Bette T. Korber, Ramesh S. Paranjape, and David C. Montefiori. Highly complex neutralization determinants on a monophyletic lineage of newly transmitted subtype C HIV-1 Env clones from India. *Virology*, 385(2):505–520, 2009.

[55] S. Gnanakaran, Marcus G. Daniels, Tanmoy Bhattacharya, Alan S. Lapedes, Anurag Sethi, Ming Li, Haili Tang, Kelli Greene, Hongmei Gao, Barton F. Haynes, Myron S. Cohen, George M. Shaw, Michael S. Seaman, Amit Kumar, Feng Gao, David C. Montefiori, and Bette Korber. Genetic signatures in the envleope glycoproteins of HIV-1 that associate with broadly neutralizing antibodies. *PLoS Comput Biol*, 6(10):e1000955, 10 2010.

[56] Peter D. Kwong, Michael L. Doyle, David J. Casper, Claudia Cicala, Stephanie A. Leavitt, Shahzad Majeed, Tavis D. Steenbeke, Miro Venturi, Irwin Chaiken, Michael Fung, Hermann Katinger, Paul W. I. H. Parren, James Robinson, Donald Van Ryk, Liping Wang, Dennis R. Burton, Ernesto Freire,

Richard Wyatt, Joseph Sodroski, Wayne A. Hendrickson, and James Arthos. HIV-1 evades antibody-mediated neutralization through conformational masking of receptor-binding sites. *Nature*, 420(6916):678–682, Dec 2002.

[57] John Shawe-Taylor and Nello Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge University Press, New York, NY, USA, 2004.

[58] N. Aronszajn. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68(3):337–404, 1950.

[59] J. Mercer. Functions of positive and negative type, and their connection with the theory of integral equations. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 209(441-458):415–446, 1909.

[60] Huma Lodhi, Craig Saunders, John Shawe-Taylor, Nello Cristianini, and Chris Watkins. Text classification using string kernels. *J. Mach. Learn. Res.*, 2:419–444, March 2002.

[61] Michael Collins and Nigel Duffy. New ranking algorithms for parsing and tagging: kernels over discrete structures, and the voted perceptron. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 263–270, Stroudsburg, PA, USA, 2002. Association for Computational Linguistics.

[62] Simon Tong and Edward Chang. Support vector machine active learning for image retrieval. In *Proceedings of the ninth ACM international conference on Multimedia*, MULTIMEDIA '01, pages 107–118, New York, NY, USA, 2001. ACM.

[63] Tommi S. Jaakkola and David Haussler. Exploiting generative models in discriminative classifiers. In *Proceedings of the 1998 conference on Advances in neural information processing systems II*, pages 487–493, Cambridge, MA, USA, 1999. MIT Press.

[64] Roy M. Gulick, John W. Mellors, Diane Havlir, Joseph J. Eron, Charles Gonzalez, Deborah McMahon, Douglas D. Richman, Fred T. Valentine, Leslie Jonas, Anne Meibohm, Emilio A. Emini, Jeffrey A. Chodakewitz, Paul Deutsch, Daniel Holder, William A. Schleif, and Jon H. Condra. Treatment with Indinavir, Zidovudine, and Lamivudine in adults with Human Immunodeficiency Virus infection and prior antiretroviral therapy. *New England Journal of Medicine*, 337(11):734–739, 1997.

[65] Robert S Hogg, David R Bangsberg, Viviane D Lima, Chris Alexander, Simon Bonner, Benita Yip, Evan Wood, Winnie W. Y Dong, Julio S. G Montaner, and P. Richard Harrigan. Emergence of drug resistance is associated with an increased risk of death among patients first starting HAART. *PLoS Med*, 3(9):e356, 09 2006.

[66] Viktor von Wyl, Sabine Yerly, Jürg Böni, Philippe Bürgisser, Thomas Klimkait, Manuel Battegay, Enos Bernasconi, Matthias Cavassini, Hansjakob Furrer, Bernard Hirschel, Pietro L. Vernazza, Patrick Francioli, Sebastian Bonhoeffer, Bruno Ledergerber, Huldrych F. Günthard, and Swiss HIV Cohort Study. Long-term trends of HIV Type 1 drug resistance prevalence among antiretroviral treatmentexperienced patients in Switzerland. *Clinical Infectious Diseases*, 48(7):979–987, 2009.

[67] Sonia Napravnik, Jessica R Keys, E Byrd Quinlivan, David A Wohl, Oksana V Mikeal, and Joseph J Jr Eron. Triple-class antiretroviral drug resistance: risk and predictors among HIV-1-infected patients. *AIDS*, 21(7):825–834, 2007.

[68] Alessandro Moschitti. Making tree kernels practical for natural language learning. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 113–120, 2006.

[69] Alessandro Moschitti, Daniele Pighin, and Roberto Basili. Tree kernels for semantic role labeling. *Comput. Linguist.*, 34:193–224, June 2008.

[70] Makoto Miwa, Rune Stre, Yusuke Miyao, and Jun'ichi Tsujii. Protein-protein interaction extraction by leveraging multiple kernels and parsers. *International Journal of Medical Informatics*, 78(12):e39–e46, 2009. Mining of Clinical and Biomedical Text and Data Special Issue.

[71] Michael Ashburner, Catherine A. Ball, Judith A. Blake, David Botstein, Heather Butler, J. Michael Cherry, Allan P. Davis, Kara Dolinski, Selina S. Dwight, Janan T. Eppig, Midori A. Harris, David P. Hill, Laurie Issel-Tarver, Andrew Kasarskis, Suzanna Lewis, John C. Matese, Joel E. Richardson, Martin Ringwald, Gerald M. Rubin, and Gavin Sherlock. Gene ontology: tool for the unification of biology. *Nat Genet*, 25(1):25–29, May 2000.

[72] Douglas R. Call, Monica K. Borucki, and Thomas E. Besser. Mixed-genome microarrays reveal multiple serotype and lineage-specific differences among strains of listeria monocytogenes. *J. Clin. Microbiol.*, 41(2):632–639, 2003.

[73] Fen-Lai Tan, Christine S. Moravec, Jianbo Li, Carolyn Apperson-Hansen, Patrick M. McCarthy, James B. Young, and Meredith Bond. The gene expression fingerprint of human heart failure. *Proceedings of the National Academy of Sciences*, 99(17):11387–11392, 2002.

[74] Sridhar Ramaswamy, Pablo Tamayo, Ryan Rifkin, Sayan Mukherjee, Chen-Hsiang Yeang, Michael Angelo, Christine Ladd, Michael Reich, Eva Latulippe, Jill P. Mesirov, Tomaso Poggio, William Gerald, Massimo Loda, Eric S. Lander, and Todd R. Golub. Multiclass cancer diagnosis using tumor gene expression signatures. *Proceedings of the National Academy of Sciences*, 98(26):15149–15154, 2001.

[75] R Development Core Team. *R: A language and Environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria, 2011.

[76] Alexandros Karatzoglou, Technische Universität Wien, Alex Smola, Kurt Hornik, and Wirtschaftsuniversität Wien. kernlab an s4 package for kernel methods in r. *Journal of Statistical Software*, pages 1–20, 2004.

[77] Alessandro Cozzi-Lepri, Andrew N. Phillips, Lidia Ruiz, Bonaventura Clotet, Clive Loveday, Jesper Kjaer, Helene Mens, Nathan Clumeck, Ludmila Viksna, Francisco Antunes, Ladislav Machala, Jens D. Lundgren, and for the EuroSIDA Study Group. Evolution of drug resistance in HIV-infected patients remaining on a virologically failing combination antiretroviral therapy regimen. *AIDS*, 21(6):721–732, 2007.

[78] Somnuek Sungkanuparph, Weerawat Manosuth, Sasisopin Kiertiburanakul, Bucha Piyavong, Noppanath Chumpathat, and Wasun Chantratita. Options for a second-line antiretroviral regimen for HIV Type 1-infected patients whose initial regimen of a fixed-dose combination of Stavudine, Lamivudine, and Nevirapine fails. *Clinical Infectious Diseases*, 44(3):447–452, 2007.

[79] Niko Beerenwinkel, Martin Daumer, Mark Oette, Klaus Korn, Daniel Hoffmann, Rolf Kaiser, Thomas Lengauer, Joachim Selbig, and Hauke Walter. Geno2pheno: estimating phenotypic drug resistance from HIV-1 genotypes. *Nucleic Acids Research*, 31(13):3850–3855, 2003.

[80] Tommy F. Liu and Robert W. Shafer. Web resources for HIV Type 1 genotypic-resistance test interpretation. *Clinical Infectious Diseases*, 42(11):1608–1618, 2006.

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

[82] Victor DeGruttola, Lynn Dix, Richard DAquila, Dan Holder, Andrew Phillips, Mounir Ait-Khaled, John Baxter, Philippe Clevenbergh, Scott Hammer, Richard Harrigan, David Katzenstein, Randall Lanier, Michael Miller, Michael Para, Sabine Yerly, Andrew Zolopa, Jeffrey Murray, Amy Patick, Veronica Miller, Steven Castillo, Louise Pedneault, and John Mellors. The relation between baseline HIV drug resistance and response to antiretroviral therapy: re-analysis of retrospective and prospective studies using a standardized data analysis plan. *Antiviral Therapy*, 5:41–48, 2000.

[83] Mark Oette, Rolf Kaiser, Martin Daumer, Ruth Petch, Gerd Fatkenheuer, Horst Carls, Jurgen Kurt Rockstroh, Dirk Schmaloer, Jurgen Stechel, Torsten Feldt, Herbert Pfister, Dieter Haussinger, and the RESINA Study Team. Primary HIV drug resistance and efficacy of first-line antiretroviral therapy

guided by resistance testing. *JAIDS Journal of Acquired Immune Deficiency Syndromes*, 41(5):573–581, 2006.

[84] Dechao Wang, Brendan Larder, Andrew Revell, Julio Montaner, Richard Harrigan, Frank De Wolf, Joep Lange, Scott Wegner, Lidia Ruiz, María Jésus Pérez-Elías, Sean Emery, Jose Gatell, Antonella D'Arminio Monforte, Carlo Torti, Maurizio Zazzi, and Clifford Lane. A comparison of three computational modelling methods for the prediction of virological response to combination HIV therapy. *Artificial Intelligence in Medicine*, 47(1):63–74, 2009.

[85] Mattia CF Prosperi, Andre Altmann, Michal Rosen-Zvi, Ehud Aharoni, Gabor Borgulya, Fulop Bazso, Anders Sönnerborg, Eugen Schülter, Daniel Struck, Giovanni Ulivi, Anne-Mieke Vandamme, Jurgen Vercauteren, Maurizio Zazzi, EuResist, and Virolab study groups. 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.

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

[87] Hiroto Saigo, Takeaki Uno, and Koji Tsuda. Mining complex genotypic features for predicting HIV-1 drug resistance. *Bioinformatics*, 23(18):2455–2462, 2007.

[88] André Altmann, Martin Däumer, Niko Beerenwinkel, Yardena Peres, Eugen Schülter, Joachim Büch, Soo-Yon Rhee, Anders Sönnerborg, W. Jeffrey Fessel, Robert W. Shafer, Maurizio Zazzi, Rolf Kaiser, and Thomas Lengauer. Predicting the response to combination antiretroviral therapy: Retrospective validation of geno2pheno-THEO on a large clinical database. *Journal of Infectious Diseases*, 199(7):999–1006, 2009.

[89] Brendan A. Larder, Andrew Revell, JoAnn M. Mican, Brian K. Agan, Marianne Harris, Carlo Torti, Ilaria Izzo, Julia A. Metcalf, Migdalia Rivera-Goba, Vincent C. Marconi, Dechao Wang, Daniel Coe, Brian Gazzard, Julio Montaner, and H. Clifford Lane. Clinical evaluation of the potential utility of computational modeling as an HIV treatment selection tool by physicians with considerable HIV experience. *AIDS Patient Care and STDs*, 25(1):29–36, 2011.

[90] World Health Organization. TOWARDS UNIVERSAL ACCESS: Scaling up priority HIV/AIDS interventions in the health sector. progress report 2009. http://www.who.int/entity/HIV/pub/tuapr_2009_en.pdf, 2009.

[91] Thuy Le, Jennifer Chiarella, Birgitte B. Simen, Bozena Hanczaruk, Michael Egholm, Marie L. Landry, Kevin Dieckhaus, Marc I. Rosen, and Michael J.

Kozal. Low-abundance HIV drug-resistant viral variants in treatment-experienced persons correlate with historical antiretroviral use. *PLoS ONE*, 4(6):e6079, 06 2009.

[92] Jacques Izopet, Patrice Massip, Corinne Souyris, Karine Sandres, Bénédicte Puissant, Martine Obadia, Christophe Pasquier, Eric Bonnet, Bruno Marchou, and Jacqueline Puel. Shift in HIV resistance genotype after treatment interruption and short-term antiviral effect following a new salvage regimen. *AIDS*, 14(15):2247–2255, 2000.

[93] Gregory K. Robbins, Kristin L. Johnson, Yuchiao Chang, Katherine E. Jackson, Paul E. Sax, James B. Meigs, and Kenneth A. Freedberg. Predicting virologic failure in an HIV clinic. *Clinical Infectious Diseases*, 50(5):779–786, 2010.

[94] Kenneth P. Burnham and David Raymond Anderson. *Model selection and multimodel inference : a practical information-theoretic approach*. Springer, New York [u.a.], 2002.

[95] John O'Quigley and Ronghui Xu. Explained variation in proportional hazards regression. In John Crowley and Donna Pauler Ankerst, editors, *Handbook of Statistics in Clinical Oncology*, pages 347–364. Chapman and Hall/CRC, 2006.

[96] Steffen Bickel, Jasmina Bogojeska, Thomas Lengauer, and Tobias Scheffer. Multi-task learning for HIV therapy screening. In *Proceedings of the 25th international conference on Machine learning*, ICML '08, pages 56–63, New York, NY, USA, 2008. ACM.

[97] DHHS and Kaiser Family Foundation Panel on Clinical Practices. DHHS/henry j. kaiser family foundation panel on clinical practices for the treatment of HIV infection. guidelines for the use of antiretroviral agents in HIV-infected adults and adolescents. october 2008 revision. http://AIDSinfo.nih.gov/ContentFiles/AboutHIVTreatment Guidelines_FS_en.pdf, 2008.

[98] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2011. ISBN 3-900051-07-0.

[99] Matthias Egger, Margaret May, Geneviève Chêne, Andrew N Phillips, Bruno Ledergerber, François Dabis, Dominique Costagliola, Antonella D'Arminio Monforte, Frank de Wolf, Peter Reiss, Jens D Lundgren, Amy C Justice, Schlomo Staszewski, Catherine Leport, Robert S Hogg, Caroline A Sabin, M John Gill, Bernd Salzberger, and Jonathan AC Sterne. Prognosis of HIV-1-infected patients starting highly active antiretroviral therapy: a collaborative analysis of prospective studies. *The Lancet*, 360(9327):119–129, 2002.

[100] Jens D. Lundgren, Amanda Mocroft, Jose M. Gatell, Bruno Ledergerber, Antonella DArminio Monforte, Philippe Hermans, Frank-Detlef Goebel, Anders Blaxhult, Ole Kirk, and Andrew N. Phillips. A clinically prognostic scoring system for patients receiving highly active antiretroviral therapy: Results from the EuroSIDA study. *Journal of Infectious Diseases*, 185(2):178–187, 2002.

[101] CASCADE Collaboration. Determinants of survival following HIV-1 seroconversion after the introduction of HAART. *The Lancet*, 362(9392):1267–1274, 2003.

[102] Miguel A. Hernán, Stephen R. Cole, Joseph Margolick, Mardge Cohen, and James M. Robins. Structural accelerated failure time models for survival analysis in studies with time-varying treatments. *Pharmacoepidemiology and Drug Safety*, 14(7):477–491, 2005.

[103] Mari M. Kitahata, Stephen J. Gange, Alison G. Abraham, Barry Merriman, Michael S. Saag, Amy C. Justice, Robert S. Hogg, Steven G. Deeks, Joseph J. Eron, John T. Brooks, Sean B. Rourke, M. John Gill, Ronald J. Bosch, Jeffrey N. Martin, Marina B. Klein, Lisa P. Jacobson, Benigno Rodriguez, Timothy R. Sterling, Gregory D. Kirk, Sonia Napravnik, Anita R. Rachlis, Liviana M. Calzavara, Michael A. Horberg, Michael J. Silverberg, Kelly A. Gebo, James J. Goedert, Constance A. Benson, Ann C. Collier, Stephen E. Van Rompaey, Heidi M. Crane, Rosemary G. McKaig, Bryan Lau, Aimee M. Freeman, and Richard D. Moore. Effect of early versus deferred antiretroviral therapy for HIV on survival. *New England Journal of Medicine*, 360(18):1815–1826, 2009.

[104] Viviane D. Lima, Robert S. Hogg, P. Richard Harrigan, David Moore, Benita Yip, Evan Wood, and Julio SG Montaner. Continued improvement in survival among HIV-infected individuals with newer forms of highly active antiretroviral therapy. *AIDS*, 21(6):685–692, 2007.

[105] The Antiretroviral Therapy Cohort Collaboration. Life expectancy of individuals on combination antiretroviral therapy in high-income countries: a collaborative analysis of 14 cohort studies. *The Lancet*, 372(9635):293–299, 2008.

[106] Kenneth G. Castro, John W. Ward, Laurence Slutsker, James W. Buehler, Harold W. Jaffe, Ruth L. Berkelman, and James W. Curran. 1993 revised classification system for HIV infection and expanded surveillance case definition for AIDS among adolescents and adults. *Clinical Infectious Diseases*, 17(4):pp. 802–810, 1993.

[107] John R. Mascola, Stuart W. Snyder, Owen S. Weislow, Seifu M. Belay, Robert B. Belshe, David H. Schwartz, Mary Lou Clements, Raphael Dolin, Barney S. Graham, Geoffrey J. Gorse, Michael C. Keefer, M. Juliana McElrath, Mary Clare Walker, Kenneth F. Wagner, John G. McNeil, Francine E.

McCutchan, and Donald S. Burke. Immunization with envleope subunit vaccine products elicits neutralizing antibodies against laboratory-adapted but not primary isolates of Human Immunodeficiency Virus Type 1. *Journal of Infectious Diseases*, 173(2):340–348, 1996.

[108] J Cao, N Sullivan, E Desjardin, C Parolin, J Robinson, R Wyatt, and J Sodroski. Replication and neutralization of Human Immunodeficiency Virus Type 1 lacking the v1 and V2 variable loops of the gp120 envleope glycoprotein. *J. Virol.*, 71(12):9808–9812, 1997.

[109] C.P. van Beveren, J. Coffin, and S. Hughes. Appendix B: HTLV-3/LAV genome. In R.L. Weiss, N. Teich, H. Varmus, and J. Coffin, editors, *RNA Tumor Viruses, Second Edition*, pages 1102–1123. Cold Spring Harbor, 1985.

[110] David C. Chan, Deborah Fass, James M. Berger, and Peter S. Kim. Core structure of gp41 from the HIV envleope glycoprotein. *Cell*, 89(2):263 – 273, 1997.

[111] Jun Liu, Alberto Bartesaghi, Mario J. Borgnia, Guillermo Sapiro, and Sriram Subramaniam. Molecular architecture of native HIV-1 gp120 trimers. *Nature*, 455(7209):109–113, Sep 2008.

[112] Davide Corti, Johannes P. M. Langedijk, Andreas Hinz, Michael S. Seaman, Fabrizia Vanzetta, Blanca M. Fernandez-Rodriguez, Chiara Silacci, Debora Pinna, David Jarrossay, Sunita Balla-Jhagjhoorsingh, Betty Willems, Maria J. Zekveld, Hanna Dreja, Eithne O'Sullivan, Corinna Pade, Chloe Orkin, Simon A. Jeffs, David C. Montefiori, David Davis, Winfried Weissenhorn, áine McKnight, Jonathan L. Heeney, Federica Sallusto, Quentin J. Sattentau, Robin A. Weiss, and Antonio Lanzavecchia. Analysis of memory B cell responses and isolation of novel monoclonal antibodies with neutralizing breadth from HIV-1-infected individuals. *PLoS ONE*, 5(1):e8805, 01 2010.

[113] Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2(3):27:1–27:27, 2011.

[114] Alessandro Moschitti. A study on convolution kernels for shallow semantic parsing. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, ACL '04, Stroudsburg, PA, USA, 2004. Association for Computational Linguistics.

[115] Guido van Rossum. *Python Library Reference.*, 1995.

[116] SAS business analytics and business intelligence software, 2011.

[117] SPSS, data mining, statistical analysis software, predictive analysis, predictive analytics, decision support systems, 2011.

[118] Statistical analysis software, data mining, predictive analytics, credit scoring, 2011.

[119] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. The WEKA data mining software: an update. *SIGKDD Explor. Newsl.*, 11:10–18, November 2009.

[120] Janez Demšar, Blaž Zupan, Gregor Leban, and Tomaz Curk. Orange: from experimental machine learning to interactive data mining. In *Proceedings of the 8th European Conference on Principles and Practice of Knowledge Discovery in Databases*, PKDD '04, pages 537–539, New York, NY, USA, 2004. Springer-Verlag New York, Inc.

[121] Soeren Sonnenburg, Gunnar Raetsch, Sebastian Henschel, Christian Widmer, Jonas Behr, Fabio de Bona Alexander Zien, Alexander Binder, Christian Gehl, , and Vojtech Franc. The SHOGUN machine learning toolbox. *Journal of Machine Learning Research*, 11:1799–1802, June 2010.

[122] Lyn C. Thomas. A survey of credit and behavioural scoring: forecasting financial risk of lending to consumers. *International Journal of Forecasting*, 16:149–172, 2010.

[123] J Bennett and S Lanning. The netflix prize. In *In KDDCup Workshop*, 2007.