IN SILICO VACCINE DESIGN FOR HEPATITIS C: A COMPUTATIONAL PLATFORM FOR FIGHTING DISEASE

BY

GREGORY R. HART

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Physics
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2017

Urbana, Illinois

Doctoral Committee:

Assistant Professor Andrew L. Ferguson, Director of Research
Professor Nigel D Goldenfeld, Chair
Professor Emeritus John D Stack
Assistant Professor Seppe Kuehn

Abstract

Hepatitis C virus (HCV) infects 2-3% of the world's population. It is a major cause of liver morbidity and mortality and a scourge to global public health. Despite more than 20 years of research, only recently have effective treatments been developed and a vaccine remains elusive. The high cost of efficacious drug treatments severely limits their impact in the developing world, leaving prophylactic vaccination the best hope for global control. The high mutation rate of HCV coupled with its rapid replication rate enables it to rapidly escape host immune responses. In this thesis, I have employed tools from statistical physics, Bayesian inference, population dynamics, and high-performance computing to define empirical fitness landscapes and conduct viral dynamics simulations to determine vulnerable viral targets and rationally design vaccine immunogens. This computational design protocol will guide and accelerate vaccine development efforts by massively reducing the need for expensive and laborious trial-and-error experimentation. While this work has focused on HCV, the grander picture is that the tools I have developed will allow quick application of this methodology to other RNA viruses (some work has been done on HIV and influenza), magnifying the impact and implications of this work for global public health.

To my advisor, Andy, for his abundant patience and help. To my wife, Danielle, for her patience and continued efforts to help me with this endeavor

Acknowledgments

I have spent more time in Champaign-Urbana than anywhere else, other than the town I was born in. The five years we have spent here have been defining. When I was considering coming to the University of Illinois I had the opportunity to speak with a research scientist who had worked at several leading universities. He told me that while UIUC was a leading research institution it was different from the other places he had experienced. He told me UIUC was very friendly and welcoming, that people were more concerned with helping each other and working together than getting credit and prestige. I have found this to be very true: from John, the custodian who cleans my office, to Andreas the dean of the college of engineering everyone I have met has been very helpful and kind. I was surprised and honored to be accepted to UIUC. My time here has pushed me to my academic limits, making me grow and mature as a scientist. The number and variety of opportunities I have had here were more numerous than I could take advantage of. The strong academic training and wonderful atmosphere here have been a great blessing to both me and my family.

There are many people to whom I am indebted. First and foremost is Andrew Ferguson, my adviser. Andy is a wonderful scientist and an even better adviser. He is very generous with his time and knowledge. He has always allowed me to explore my ideas while still guiding me, helping overcome obstacles, and encouraging me to make progress towards my goals. As an undergraduate I was told that graduate school prepares you for research, but does little to prepare the other skills and knowledge needed after graduation. Andy has been great at looking beyond my Ph.D. In addition to helping me develop soft skills, exposing me to the whole publication process, grant writing, etc, very early on he was prompting me

to think about what career I wanted and helping me explore all my options. Any success I have obtained and obtain in the future is due to the great foundation he has laid for me. If I have any regrets, it is that I wasn't has good of a student has he deserved.

I also own a thanks to Nigel Goldenfeld, Volodymyr (Vlad) Kindratenko, and Lance Cooper. Nigel served has my co-advisor for the computation science and engineering fellowship. He is a very busy man and it was kind of him to take time to assist me in this endeavor. Vlad offered me his vast knowledge of GPU accelerators and other non-traditional computational hardware. The time he spent with me teaching me how to optimize our GPU code not only helped with the work herein, but will continue to assist me as I move forward with my research. In addition to his own research group and teaching load, Lance serves as the associate head for graduate programs. In this position he looks out for the physics graduate students as a group and each of us individually. His efforts are often not seen, but they are felt in the department's atmosphere and student life. I believe he is often underappreciated and would like thank him for all he does.

I thank the members of the Ferguson research group. Discussing research with them has been a great help and often sparked new ideas or helped find detours around road blocks. We have had many engaging and entertaining discussions. I have enjoyed getting to know them and will miss our long car rides to conferences. I hope to maintain our friendships and look forward to collaborating as we go our separate ways. I am particularly grateful for the help of Samuel Kaufman in the early development of the population dynamics algorithm and Chin-Yu "Chester" Cheng for his help in accelerating the population dynamics algorithm once it was developed.

I owe a debt of gratitude to the computation science and engineering program, XSEDE, and Bluewaters. The CSE program helped fund my research for 2 years, and has given me access to countless hours of computer time. XSEDE and Bluewaters have also granted tens of thousands of CPU hours. Without these high performance computing resources and support from the departments of Physics and Material Science and Engineering this work

would not have been possible.

I need to thank all those who help run the various departments, handle travel, and generally make it much easier to navigate graduate school. In particularly Sandra Helregel and Jay Menacher from the material science department, Elizabeth Stull from CSE, and Sandy Johnson, Kate Shunk, and Dashawnique Long from the physics department have put up with me a lot. They have done much to make things more comfortable and easier on me and my family.

Lastly I would like to thank my family. Due to the size of my immediate family I will not list all their names here, but I would like to mention my brother Gus. On several occasions Gus has talked with me about his own experience as a graduate student and doing research, allaying my self doubt, and encouraging me. Victoria, my daughter, has formed her earliest memories here. Her smile and joy has often given me energy that I needed. Danielle, my housekeeper, secretary, editor, and wife, has been a constant strength and support. Though her degree and passion is music, she understands this work better than most.

Contents

List of Tables

List of Figures

List of Abbreviations

HIV	Human Immunodeficiency Virus
HCV	Hepatitis C Virus
HSV	Herpes Simplex Virus
HLA	Human Leukocyte Antigen
NS5B	Nonstructural Protein 5B
CTL	Cytotoxic T Lymphocyte
MSA	Multiple Sequence Alignment

Chapter 1

Introduction and Background

Viruses are one of the simplest biological agents on Earth. Viruses exist as small particles tens to hundreds of nanometers in diameter, carry very little genetic material, and are incapable of replicating themselves. They are commonly considered to live on the edge of life. Yet they are more prevalent than any living organism on Earth, outnumbering cellular organisms by at least an order of magnitude, occupy every ecological niche, and infect all known forms of life [17–20]. Emerging with the oldest known life forms some viruses appear to be very old, while others are less than a century old. For example, it is thought that human immunodeficiency virus (HIV) evolved from a primate virus and moved to humans in the 1930s. Viruses infect the cells of other organisms and co-opt their replication machinery to produce daughter virions that get released to infect other cells. Infection damages host tissues by inducing pathological changes to the host cell, and ultimately cell death by lysis or apoptosis (programmed cell death) [21].

Viral infections are responsible for many diseases and ailments from cold sores, the common cold, and influenza to HIV, several forms of hepatitis, and zika. There are medicines that can be used to relieve the symptoms of viral diseases and a few antiviral drugs have been developed in recent years for specific diseases, but the best way to deal with viruses is prevention through vaccination. The protection against disease offered by vaccination is a hallmark of modern medicine. Vaccination has lead to the eradication of smallpox [22], which is estimated to have killed hundreds of millions of people in the 20th century alone [23]. Vaccination has also lead to the eradication of rinderpest, a disease afflicting livestock [24]. Other diseases are well on their way to eradication or are largely controlled. For example, today polio only circulates in a handful of countries infecting tens of people each year

while 40 years ago it afflicted hundreds of thousands each year [25].

The development of a vaccine usually takes 10 to 15 years of dedicated effort and costs €1.7 billion ($2.0 billion) [26, 27], however many viruses such as HIV, hepatitis C virus (HCV), and Ebola virus have no vaccine after several decades of research. There are many obstacles that can arise in the development of a vaccine, but there is a commonality with almost every virus currently evading vaccination efforts: rapid evolution [28]. DNA viruses (e.g. Herpes simplex virus (HSV), Smallpox virus) have mutation rates ranging between $10^{-8} - 10^{-6}$ substitutions per nucleotide per replication while RNA viruses (e.g. HCV, HIV) range between $10^{-6} - 10^{-4}$ [29]. Viruses also have rapid replication cycles. This high virion production rate combined with the rapid mutation rate can result in large genetic diversity of viral strains within an infected individual, enabling rapid escape from host immune responses. In hepatitis C virus, for example, a high mutation rate of $\sim 1.2 \times 10^{-4}$ substitutions per nucleotide per replication conspires with a high virion production rate of $\sim 10^{12}$ viral progeny per day to produce all possible point mutations at all positions in the 9600-nucleotide genome every single day [29–31].

The goal of the research presented herein is to build a computational platform to accelerate the vaccine creation process. By taking advantage of increasingly abundant sequence data we can construct empirical fitness landscapes. A fitness landscapes allows us to quantitatively evaluate the replicative fitness of different viral strains. This allows us to quickly and cheaply compare vaccine candidates, systematically search for viral vulenerabilities, and identify potential escape pathways from a given immune response.

In developing these tools I mainly used HCV as a test system. The reminder of this chapter will cover some information on HCV. Then in chapter 2 we will dicuss more about fitness landscapes and our model for them. In chapter 3 we will discuss using a fitness landscape to design vaccines based on an unevolving population. In chapter 4 we will discuss ways to use the fitness landscapes to explore alternative treatments. Then in chapter 5 we refine our vaccine predictions with an evolving viral population.

HCV infects 2% of the worlds population and 700,000 people die each year from HCV-related liver diseases [32]. Prevalences in the US and European populations are around 1%, but can reach 20% in parts of Northern Africa [33]. Chronic infection is responsible for over a quarter of worldwide cases of cirrhosis and hepatocellular carcinoma [33] and is the leading cause of liver transplantation in the developed world [34]. Despite many challenges remarkable progress has been made in treating HCV [35, 36]. In the last 7 years success rates of treatment has gone from $\sim 50\%$ [37] to 90% [35]. However the impact of these new treatments are severally limited by their cost [38–40]. Furthermore, recent concerns have been raised about the safety of the new drugs [41]. Accordingly vaccination remains "the most cost-effective and realistic method of controlling HCV globally" [42].

While more than two decades of research has not lead to a vaccine, there is a body of evidence supporting the viability of a vaccine [43–46] and hope that a vaccine is within reach. In 20-30% of cases, the host aborts HCV infection in the absence of treatment [47, 48]. Clinical studies have identified particular human leukocyte antigen (HLA) alleles, such as HLA-B57, that are correlated with HCV control, and identified epitopes (i.e. parts of the viral proteins attacked by the immune system) presented by these HLA molecules [4, 8]. It appears that these epitopes are particularly vulnerable to immune attack and cannot easily escape by mutation [8, 12, 34, 48]. The first prophylactic HCV T-cell vaccine trial is underway [48–50], which seeks to prime host T-cell responses – the cellular arm of the adaptive immune system that seeks to identify and destroy infected cells – by delivering complete NS3, NS4, and NS5 proteins. In analogy with recent findings in HIV [12, 51, 52], there are concerns that by presenting whole proteins, potent responses against susceptible viral targets may be drowned out by ineffective responses against poor targets, thereby failing to confer protective immunity. Other HCV vaccines currently under development try to address this issue [53–58]. Knowledge of which regions of the viral proteome are most vulnerable to immune pressure can inform which regions of the HCV proteome should be included in a vaccine immunogen to elicit only efficacious immune responses.

Chapter 2

Fitness Landscapes

2.1 Introduction

Viruses are the most prevalent biological agents on Earth, occupying every ecological niche, infecting all known forms of life, and outnumbering cellular organisms by at least an order of magnitude [17–20]. Derived from the Latin word meaning "venom" [21, 60], viruses are often considered organisms "on the edge of life" carrying hereditary genetic material and subject to natural selection, but which are themselves unable to reproduce, instead reliant on hijacking the replication machinery of an infected host [21, 61, 62]. Viruses exist as small pathogenic particles tens to hundreds of nanometers in size known as virions [17]. A virion comprises a single or double stranded RNA or DNA genome shrouded in a proteinaceous coat known as a capsid, which itself may be encapsulated by a lipid bilayer envelope containing additional proteins and/or carbohydrates [17, 62, 63]. Viruses are communicated by many means, including air, water, sexual contact, and vectors such as mosquitos. Common viruses infecting humans include influenza, hepatitis B virus (HBV), hepatitis C virus (HCV), human immunodeficiency virus (HIV), and Zika virus. Upon entering the body, innate and adaptive immune responses seek to destroy the virus and prevent an infection, which – depending on the viral strain, health of the host, intensity of infection, and efficacy of treatment – can induce symptoms ranging from mild discomfort, to organ failure, to death.

Upon infecting a susceptible cell, a virus releases its genetic material and proteins and

Most of this chapter is an excerpt from ref. [59]: G.R. Hart and A.L. Ferguson "Viral fitness landscapes: A physical sciences perspective" in "Systems Immunology: An introduction to modeling methods for scientists" J. Das, C. Jayaprakash (eds.) Taylor and Francis (in press, 2017) [ISBN-10: 1498717403]

co-opts the host replication machinery to produce daughter virions that are released to go on and infect other cells. Infection damages host tissues by inducing pathological changes to the host cell, and ultimately cell death by lysis or apoptosis (programmed cell death) [21]. Some viruses, however, can remain dormant within infected cells for years, providing a latent reservoir of infection [21]. Most viruses, RNA viruses in particular, are highly error prone in copying their genetic material, introducing random mutations into the genome of the daughter virions [62, 64]. Most mutations are deleterious to the virus by impairing the activity of viral proteins, others are neutral having negligible impact on viral function, while a small number may be beneficial by enabling the virus to escape from host immune pressure or develop resistance to an antiviral drug. Mutation rates are a strong function of genome size, and can vary between $10^{-8} - 10^{-6}$ substitutions per nucleotide per cell infection for DNA viruses to $10^{-6} - 10^{-4}$ for RNA viruses [29, 65]. Together with high virion production rates, the genetic diversity of viral strains within an infected host can be exceedingly high, enabling rapid escape from host immune responses. In hepatitis C virus, for example, a high mutation rate of $\sim 1.2 \times 10^{-4}$ substitutions per nucleotide per cell infection conspires with a high virion production rate of $\sim 10^{12}$ viral progeny per day to produce all possible point mutations at all positions in the 9600-nucleotide genome every single day [29–31].

2.2 Sequence space and viral fitness landscapes

The genotype of each viral strain – the genetic sequence of its RNA or DNA genome – together with its interaction with its environment, determines its phenotype – the characteristics and performance of the virus [64]. The phenotype, in turn, dictates the viral fitness [64, 66–68]. The definition of fitness can be somewhat slippery [66, 69, 70]. For viral replication within an infected host the replicative fitness is the appropriate measure, and has been succinctly defined as "the capacity of a virus to produce infectious progeny in a given environment" [66, 68]. (This should be distinguished from transmission fitness describing the

capacity of a viral strain to jump between hosts, and the epidemiological fitness defining the capacity of a particular viral strain to come to dominate within a particular host population [66].) The replicative fitness is a positive real number that can be empirically measured *in vitro* or *in vivo* [66, 70–73].

For a viral genome of length M nucleotides, each position can be occupied by one of four nucleobases, leading to 4^M possible distinct viral strains. Alternatively, for a N-residue proteome in which each position can be occupied by one of 20 naturally occurring amino acids, there are 20^N distinct sequences. For hepatitis C virus, $M \approx 9600$ and $N \approx 3000$, such that there are more than 10^{5700} distinct viral genomes and 10^{3900} distinct proteomes [31]. For comparison, there are "only" around 10^{80} protons in the universe [1, 62, 74]. The size of the known universe may be astronomically big, but the sequence space accessible to the virus is "genomically" large [75]!

Mathematically, the ensemble of viral genomes or proteomes define a set S. Distances between sequences are naturally measured by the Hamming distance, d_H, quantifying the number of point mutations by which a pair of sequences differ. The set S and distance d_H together define a M-dimensional metric space (S,d_H) known as sequence space [62, 64, 65, 76]. The sequence space can be conceived as a M-dimensional hypercube with genome sequences residing on the vertices and edges connecting nearest neighbors differing by a single point mutation [64, 65]. This space is high-dimensional, highly-connected, and dense [62] (figure 2.1)

Superposing the (replicative) fitness of each strain onto the sequence space defines a (M+1)-dimensional fitness landscape [62, 64, 76, 77]. First posited by Sewall Wright in the 1930s [1] and refined by J. Maynard Smith [78], Manfred Eigen, and Peter Schuster [79, 80] in the 1970s, the fitness landscape is a cornerstone of population genetics coupling the key concepts of sequence space and differential fitness. In essence, the fitness landscape can be considered the "playing field" over which the virus is constrained to evolve, defining a quantitative map describing how its fitness changes as point mutations arise by the error

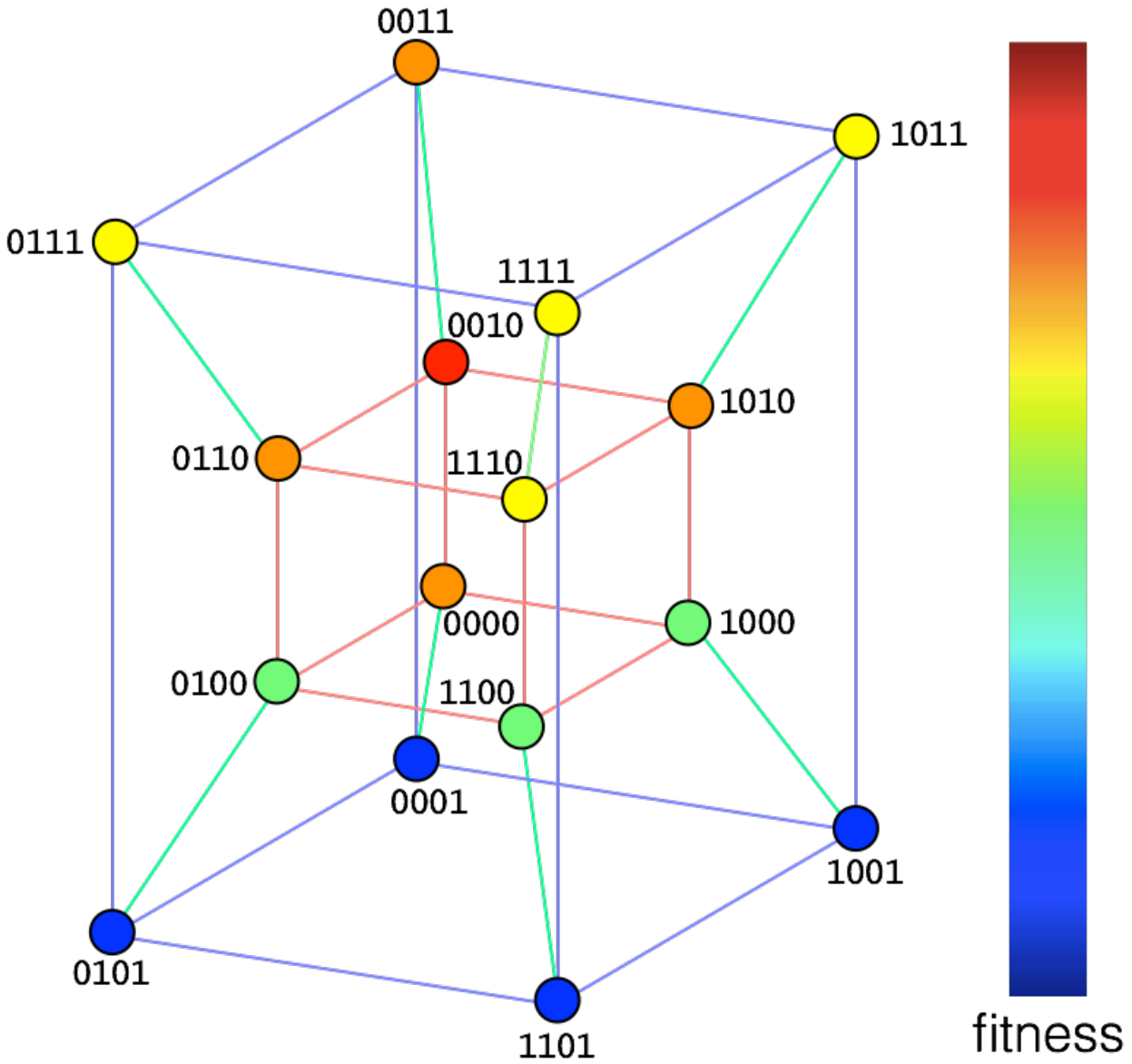


Figure 2.1: Illustrative fitness landscape for a hypothetical $M = 4$ position virus, each of which can take on $z = 2$ values {0,1}. The $K = z^M = 2^4 = 16$ distinct viral genomes comprising the sequence space S define the nodes of a 4-dimensional hypercube (tesseract) with edges linking genomes differing by a single point mutation (Hamming distance, $d_H = 1$). Superposing the fitness of each mutant over the sequence space defines the $(M + 1)$-dimensional fitness landscape represented here as a heat map. Although low-dimensional pictures provide "a very inadequate representation of such a field" [1], the mathematical concept indicated by this schematic straightforwardly generalizes to arbitrary M and z.

prone replication machinery. Replicatively fit strains that rapidly produce progeny reside at the peaks of the fitness landscape, whereas unfit strains containing mutations that im-

pair protein function reside in low-fitness valleys [12, 81]. Viral evolution is the process of mutational motion over the fitness landscape under applied selective pressure [78].

2.3 Quasispecies theory

For rapidly mutating viruses, the picture of viral evolution as a hill-climbing process in which the virus seeks to maximize fitness is not quite correct [69, 77]. Due to the high rate at which mutations are introduced during viral replication, a viral strain residing in sequence space at a particular vertex of the M-dimensional hypercube does not generate identical copies of itself but produces daughter progeny that "leak" along the edges of the hypercube into neighboring mutational states. Recognizing the importance of this effect upon natural selection, Manfred Eigen and Peter Schuster proposed that the unit of natural selection was not a particular viral strain, but a swarm of closely related mutant strains in sequence space known as a quasispecies arising from a balance of mutation and selection [79, 80]. The deterministic evolution of the quasispecies over the fitness landscape in the infinite population size limit is specified by a first order nonlinear differential equation known as Eigen's equation, or the quasispecies equation [2, 62, 64, 79, 80, 82, 83],

$$\dot{n}_i = \sum_{j=1}^{K} n_j f_j q_{ij} - \phi(\{n_k\}) n_i, \tag{2.1}$$

where there are K distinct genomes $i = 1 \dots K$, n_i is the fraction of the population with genome i and the population fractions are normalized such that $\sum_{i=1}^{K} n_i = 1$, $\dot{n}_i$ is the rate of change of strain i, f_i is the relative replicative fitness of strain i, q_{ij} is the mutational probability that replication of strain j produces strain i under replication and are normalized such that $\sum_{i=1}^{K} q_{ij} = 1$, and $\phi(\{n_k\}) = \sum_{i=1}^{K} n_i f_i$ is the average fitness of the population defining the fitness-independent removal rate of each strain to keep the total population size fixed. This equation may be succinctly expressed in matrix-vector form as [2],

$$\dot{\vec{n}} = \mathbf{QF}\vec{n} - (\vec{n} \cdot \vec{f})\vec{n}, \tag{2.2}$$

where $\vec{n} = [n_1, n_2 \ldots, n_k]^T$ is a column vector specifying the instantaneous structure of the viral population, $\dot{\vec{n}}$ is the rate of change of the population structure, $\mathbf{Q} = [q_{ij}]$ is the left stochastic mutation matrix, $f = [f_1, f_2 \ldots, f_k]^T$, and $\mathbf{F} = diag(f)$ is the diagonal fitness matrix. The nonlinear quasispecies equation can be transformed into a linear differential equation and solved exactly using a path integral formulation [84, 85]. In the limit of perfect replication fidelity (i.e. a diagonal mutation matrix), the quasispecies equation reduces to the standard replicator equation [64].

The quasispecies equation is central to the study of viral dynamics, and makes a number of important predictions (figure 2.2). First, it asserts that for rapidly mutating viruses, natural selection acts on the level of the quasispecies rather than the individual strain [65, 79, 80]. Second, it predicts that while at low mutation rates the equilibrium quasispecies will be centered on the global fitness maximum, at sufficiently high mutation rates it may sacrifice a narrow high peak in favor of a broad lower peak [62, 64, 79, 80]. This result can be understood as natural selection on the level of the quasispecies, which seeks to maximize the mean fitness of the swarm of closely related strains. A virus with a low-error rate can exist as a tight cloud of member strains within a narrow high fitness peak, resulting in a high mean fitness of the population. A high-error rate virus exists as a diffuse cloud of strains due to mutational "leakage" along the edges of the sequence space hypercube. In the region of a high narrow peak, many members of the high-error rate ensemble will exist in low fitness states outside the peak, pulling down the mean fitness of the population. In the region of a lower but broader peak, more members of the quasispecies occupy high fitness states, leading to a net elevation of the mean fitness. This phenomenon has been termed "survival of the flattest" [86, 87]. Third, it predicts the existence of a maximum error rate beyond which the quasispecies loses cohesion and cannot adapt to the fitness landscape

[62, 64, 82, 88–91]. This *error catastrophe* can be shown to be analogous to a first order phase transition in a finite system [92], corresponding to a loss of locality of the quasispecies within sequence space [93–97]. We further discuss this in chapter 4 where we show that some proteins exist on the edge of this phase transition and inducing the error catastrophe offers alternative treatment options. Evidence suggests that many RNA viruses possess mutation rates close to the error threshold, which is thought to provide a survival advantages to the virus by offering a reservoir of viral phenotypes, enhancing adaptability, and aiding in the development of immunological escape mutations [62, 90, 98, 99]. Drug therapies that elevate the viral mutation rate above the error threshold are under investigation as novel treatments for HIV [100–103].

Despite the central importance of quasispecies theory in the theoretical understanding of viral evolution, it has not enjoyed strong experimental support largely due to the technical difficulties associated with sequencing a substantial fraction of the strains constituting the quasispecies [83]. Nevertheless, support exists for the existence of the error catastrophe [104], and recent advances in deep sequencing and ultra deep sequencing are expected to enable more rigorous testing of its predictions [83]. One significant deficiency of the theory is that it is inherently deterministic, therefore pertaining to formally infinite viral populations. Viruses can have relatively small effective population sizes, such that stochastic effects can play an important role in their evolutionary dynamics [82, 97]. Numerical population genetics simulations provide a means to explicitly account for stochasticity, and also straightforwardly incorporate other effects such as co-infection, recombination, spatial heterogeneity, and drug or immune pressure [82, 97, 105–108]. Finally, we note the ingenious observation by Guy Sella and Aaron Hirsh of an isomorphism – under relatively restrictive conditions – between population genetics and equilibrium statistical thermodynamics [109]. Similar analogies to statistical mechanics and information theory have been made by Michael Deem, Arup Chakraborty, Bill Bialek, Hendrik Richter, and ourselves [12, 76, 81, 96, 106, 110–115].

A prerequisite to simulating viral dynamics using either quasispecies theory or numerical

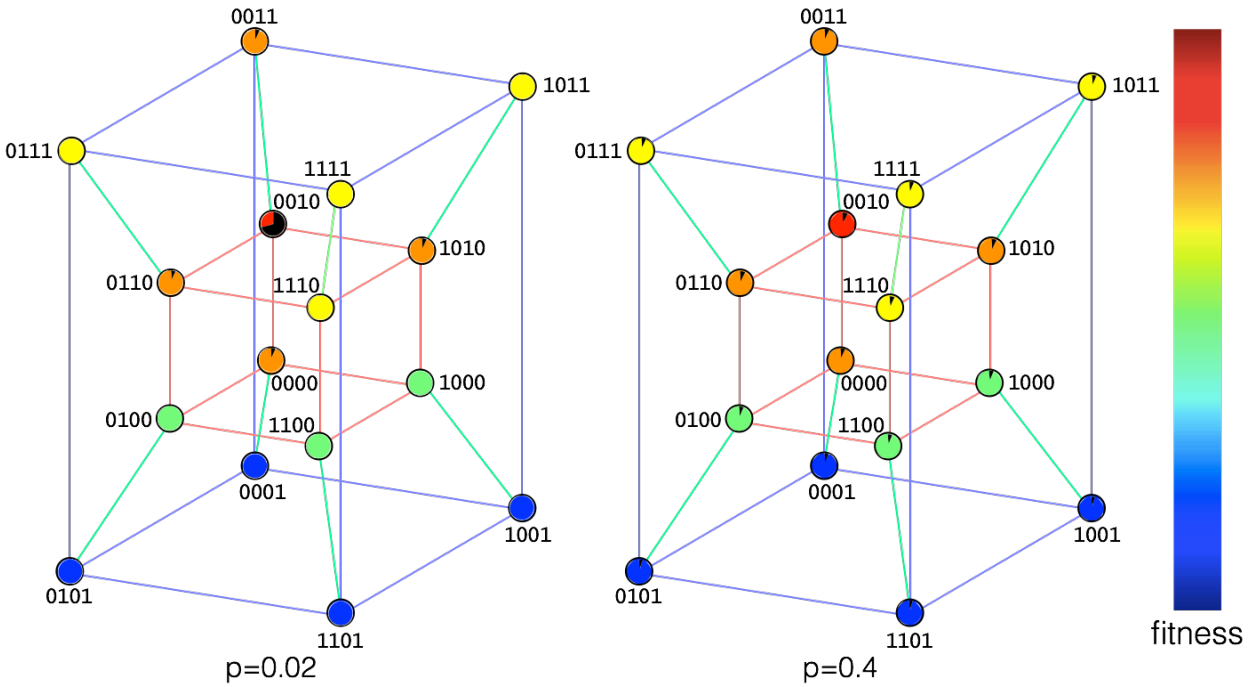


Figure 2.2: Equilibrium quasispecies distribution over the sequence space n^4 as a function of mutation rate. Continuing the example of a $M = 4$ position virus with $z = 2$ values {0,1} per position, we can simulate its quasispecies dynamics over its $K = 16$ state fitness landscape in figure 2.1. The fitness landscape is specified by the (relative) fitness vector f =[0.7, 0.1, 0.9, 0.7, 0.5, 0.1, 0.7, 0.6, 0.5, 0.1, 0.7, 0.6, 0.5, 0.1, 0.6, 0.6] with elements arranged in standard order (i.e., $1 = 0000$, $2 = 0001$, ..., $15 = 1110$, $16 = 1111$). The mutation matrix is specified as $\mathbf{Q} = q_{ij}$, where $q_{ij} = p^{H_{ij}}(1-p)^{(M-H_{ij})}$ (eqn. 2.3), H_{ij} is the Hamming distance between strains i and j, and p is the mutation rate per position per replication cycle. We solve for the equilibrium quasispecies distribution over the sequence space by solving for the steady state solution of the quasispecies equation (eqn. 2.2) as a function of mutation rate [2]. In the left panel we reproduce the fitness landscape illustrated in figure 2.1 superposed with pie segments indicating the equilibrium fraction of the quasispecies population partitioning into each mutational state at a low mutation rate of $p = 0.02$. Under these conditions, the virus possesses relatively high copying fidelity and the quasispecies localizes around the fitness peak at state [0010], with a small fraction of the population leaking into the neighboring strains. In the right panel we illustrate the equilibrium distribution at an elevated mutation rate of $p = 0.4$, where the quasispecies has experienced an error catastrophe such that it cannot adapt to the topography of the fitness landscape and has delocalized across sequence space.

simulations is specification of the fitness landscape f and the mutation matrix Q. The mutation matrix is typically assumed to be non-negative, symmetric, and stochastic [2]. Assuming independent identically distributed (iid) mutations, the mutation matrix $Q = [q_{ij}]$

can be straightforwardly specified as [2, 62],

$$q_{ij} = \left(\frac{p}{z-1}\right)^{H_{ij}} (1-p)^{(M-H_{ij})}, \tag{2.3}$$

where q_{ij} is the mutational probability that replication of strain j produces strain i, p is the mutation rate per position per replication cycle, z is the size of the alphabet at each position ($z = 4$ for genomes, $z = 20$ for proteomes), M is the number of positions in the strain, and H_{ij} is the Hamming distance (i.e., number of substitutions) between strains i and j. For $p = 0$, the Q matrix becomes the identity matrix corresponding to perfect replication fidelity. For sufficiently small values of p, daughter strains containing two or more mutations within the same replication cycle ($H_{ij} \geq 2$) may be considered vanishingly rare. Accordingly, Q may be approximated as $q_{ij} = (1-p)^M$, $q_{ij} = \left(\frac{p}{1-p}\right)(1-p)^{(M-1)}$ for $H_{ij} = 1$, and $q_{ij} = 0$ for $H_{ij} \geq 2$, corresponding to a situation in which the progeny of any viral strain can contain at most one point mutation and multi-mutant hops along the edges of the sequence space hypercube are forbidden. More complex models may allow, for example, for differential probabilities in mutating from a purine to a pyrimidine nucleobase on the level of the genome, or accounting for synonymous and non-synonymous mutations in the three-base amino acid codon on the level of the proteome. Specification of the fitness landscape defined by the fitness vector f is a more involved problem that is the subject of the following section.

2.4 Viral fitness landscapes from experiment and theory

The genotype of a viral strain within a particular environment specifies its phenotype that, in turn, dictates its replicative fitness. The fitness landscape is the convolution of this double mapping from genotype to phenotype to fitness [64, 74, 76, 77]. This mapping

can be conceived as a (complicated) function of the viral sequence, which integrates over the characteristics and performance of the virus within its environment to produce a non-negative real number specifying its replicative competency. Fitness landscapes have typically been defined by one of two methods: sparse experimental fitness measurements of very limited regions of sequence space, or theoretical models designed to reproduce the statistical properties of real landscapes [116, 117]. The former pertain to one particular viral system and therefore can be used to make fitness predictions, but have proven difficult to both define and to generalize. The latter are designed to be statistical abstractions that are generic to some class of viruses, but necessarily sacrifice predictive accuracy of the fitness of any particular virus.

Experimental determination of comprehensive fitness landscapes is rendered extremely challenging by the vast size of viral sequence space. For a viral proteome comprising N amino acids there are 20^N distinct viral sequences, meaning that experimental measurements of the replicative capacity can only probe a vanishingly small fraction of the sequence space [74]. Taking hepatitis C virus as an example, $N \approx 3000$ meaning that there are 10^{3900} distinct proteomes [31]. Fewer than around 1 in 10^{3897} strains would have to be represented in meaningful quantities during infection to define a sequence space sufficiently small to be comprehensively probed by experimental fitness assays. Nevertheless, advances in next-generation sequencing and high-throughput automated assays has led to a growing body of experimental studies designed to probe general features of fitness landscapes by measuring the fitness of a limited subset of tens to hundreds of mutant strains [74, 77, 116, 118–120]. For example, Qi et al. coupled saturation mutagenesis with deep sequencing to measure the fitness of all 1720 possible point mutants within an 86-residue region of the hepatitis C virus protein NS5A known to be a target for replication inhibitors [121]. Nonetheless, experimental determination of comprehensive fitness landscapes remains inherently intractable for even the smallest viruses, offering only a "glimpse... within the immense genotype space" [74].

A more scalable approach is offered by using limited experimental data to fit statistical

or regression fitness models that may then be extrapolated to predict the fitness of strains that were not directly assayed [74, 77, 119, 120, 122–125]. For example, Hayashi et al. combined molecular evolution experiments with infectivity assays to fit the parameters of an $NK-$model [123]. Hinkley et al. fitted a regularized regression model to *in vitro* fitness measurements of 70,081 mutant strains of HIV incorporating pairwise interactions that was capable of explaining 54.8% of the variance in the measured fitness [120]. Segal et al. employed decision tree-based approaches to predict HIV-1 replicative capacity as a function of the amino acid sequence of the protease and reverse transcriptase viral proteins [124]. In all cases, however, the time, labor, and expense associated with *in vitro* fitness measurements means that models are fit to a sparse library of fitness measurements sampling an infinitesimally small fraction of sequence space. The fitted models are therefore subject to significant bias and possess large extrapolation errors for mutant strains dissimilar to those upon which the model was trained [122].

A number of theoretical fitness landscape models have been proposed that seek to reproduce the statistical features and correlations observed in real landscapes. These models have proved extremely useful in gaining insight into viral dynamics and adaptation over archetypal landscapes, and as baseline models containing parameters that may be tuned to experiment. Epistasis – nonlinear effects due to interaction of two or more mutational loci – gives rise to non-additive fitness landscapes, and appears to play a critical role in the function, dynamics, and evolution of viruses [4, 12, 51, 74, 83, 96, 126]. One of the first and most popular fitness models explicitly designed to capture epistasis is Kaufmann's "tuneably rugged" (i.e., "tuneably epistatic") $NK-$model [76, 117, 123, 127–129]. This model specifies fitness as the sum of random variables corresponding to the fitness contributions from N positions, each of which depends on the mutational state of K other positions. In the limit $K = 0$, there is no epistasis (i.e., all positions have independent fitness contributions) and the landscape is unfrustrated and smooth, the so-called "Mount Fiji" model [74]. The limit of $K = (N - 1)$ corresponds to all-to-all epistasis, producing a highly rugged and frustrated

landscape containing many local fitness maxima known as a "house of cards" model that is isomorphic to a random energy spin glass [74, 130, 131]. Generalizations of the $NK-$model accounting for protein structure and binding interactions have subsequently been proposed [62, 86]. The "rough Mount Fiji" model superposes onto the $NK-$model a fitness contribution that falls off with Hamming distance away from a reference (usually the wild-type) strain [74, 82, 132]. An alternative model is provided by Motoo Kimura's neutral theory of evolution, which envisages mutations to be either neutral or lethal [133]. The corresponding fitness landscape is binary, containing viable strains of equal fitness residing on a neutral plateau fissured by valleys of unviable mutant strains [132]. Since the unviable strains are considered to be mutationally inaccessible (i.e., have zero fitness), this has been described as a "holey" fitness landscape that is isomorphic to a percolation problem over the sequence space hypercube [76, 132].

2.5 Data-driven viral fitness landscapes

Very recently, data-driven approaches have emerged as a third way to determine empirical fitness landscapes from databases of viral sequences [2, 12, 81, 111, 134, 135]. These approaches adopt a Bayesian perspective to determine fitness landscapes consistent with observations of viral strains sequenced from infected hosts by assuming a relationship between strain fitness and the relative prevalence of correlated mutational patterns within the sequence database [122]. Although the predictions of such models may be validated against *in vitro* fitness assays, these techniques are distinguished from the fitting of fitness models to experimental measurements in that they require only a library of observed viral sequences and do not appeal to experimental fitness measurements for their construction. This is a critical distinction, since modern low-cost, high-throughput sequencing technologies has made these approaches massively more scalable than those predicated on laborious and time consuming fitness assays [83, 136].

In 2013, Ferguson et al. pioneered a data-driven approach to reconstruct viral fitness landscapes from viral sequence databases [12]. Subsequent works by ourselves and others have produced sophistications, analyses, and validations of the approach, and applications to both hepatitis C virus and HIV [81, 96, 106, 111, 112]. The essence of the approach is to regard the viral strains within a sequence database as observations over an unknown fitness landscape, and perform Bayesian inference to solve the "inverse problem" and reconstruct the most probable fitness landscape from which the strains were drawn. Provided that founder effects are weak such that the virus rapidly anneals to the immune response of a newly infected host, and that the sequence databases of observed viral strains are sufficiently large and represent hosts with diverse immunological genotypes, it has been demonstrated by both theory and empiricism that the fitness landscapes we compute quantify the intrinsic molecular fitness of the virus uncorrupted by "footprints" of adaptive immunity [12, 81, 106, 111, 137]. In effect, the diversity of possible immune responses means that particular positions within the viral genome are subject to mutational pressure from a small subset of the hosts constituting the database and particular immune responses act as a small perturbation when averaged over sufficiently many hosts [106]. The fitness landscapes determined by our approach – which we detail below in section 2.5.2, and then describe its applications to HIV and HCV vaccine design in chapters 3, 4, and 5 – reflect the *intrinsic* replicative capacity of the virus in the absence of immune pressure, and can be straightforwardly adapted to reflect the *effective* fitness landscape experienced by the virus in any particular host by superposing the adaptive immune responses as an external perturbation [12, 81, 106].

2.5.1 Relationship to other work

In 2015, Niko Beerenwinkel and co-workers proposed an elegant framework to estimate intrahost fitness landscapes from next generation sequencing data of the viral quasispecies within an infected host [2, 83, 136]. This approach shares similarities with our own in that it

employs a Bayesian inference approach to reconstruct fitness landscapes from observations of viral strains, but has some important distinguishing differences. First, the approach is based on intrahost sequencing data and therefore calculates fitness landscapes that are the convolution of the intrinsic fitness (i.e., in the absence of immune pressure) with the adaptive immune responses of the particular host. The authors used their approach to recover the fitness landscape for HIV protein p7 in two hosts, and found the landscapes to be quite different in each case. Containing the "footprints" of host immune pressure [137], it is a challenge to generalize these fitness landscape to new environments (e.g., different immune responses, drug pressure, *in vitro* culture). Second, the approach postulates that the ensemble of strains should follow the equilibrium distribution predicted by quasispecies theory to provide a model linking strain fitness to prevalence in the sequencing data. Due to the immense size of sequence space, it is necessary to operate in a reduced subset of sequence space in which low fitness viral strains are neglected [2]. Moreover, the fitness landscape f and mutation matrix Q are not independently identifiable (eqn. 2.2) [2, 138], meaning that the inference problem requires the specification of a particular Q upon which the inferred f then depends. In contrast, our approach is non-parametric in the sense that it does not appeal to quasispecies theory nor assume an *a priori* functional form for the fitness landscape, instead seeking the least biased (i.e., maximum entropy) model consistent with the data. The authors devise a Monte Carlo approach to perform the Bayesian inference, and have made this available for free public download at `http://www.cbg.ethz.ch/software/quasifit`.

We also observe the relationship of our approach to earlier work by Beerenwinkel et al. in 2005 and Deforche et al. in 2008 that coupled cross-sectional sequence data (i.e., sequence data from different infected hosts at different times) from treated and untreated cohorts of HIV patients with *in vivo* evolutionary models to estimate the effect of drug therapy upon viral fitness and mutation [134, 135]. The more recent work by Beerenwinkel and ourselves seeks not differential fitness responses to drug therapy, but rather the complete fitness landscape either in the presence [2] or absence [12] of host immune pressure.

Our methodology also shares commonalities with innovative work by Bill Bialek and co-workers who employed data-driven spin glass models to quantify antibody diversity in zebrafish, the activity of ensembles of neurons in salamander retinae, and the flocking behavior of birds [113, 115, 139–142], and Martin Weigt, Terry Hwa, José Onuchic and co-workers who used data-driven spin glass models to identify coevolving amino acid residues or nucleotides to predict putative protein and RNA tertiary structure contacts [143–147].

2.5.2 Mathematical and computational details

Given a multiple sequence alignment (MSA) of viral sequences drawn from multiple infected hosts, we have developed an approach to determine the least biased model capable of reproducing the relative prevalence of viral strains within the database [12, 81, 106, 111]. We choose to work with viral proteomes, but the approach is equally applicable for viral genomes. Since the viral strains are drawn from amongst the population of infected hosts and not from within a single host, we assume that the viral prevalence landscape on the level of the infected population is a good proxy for the *fitness landscape* on the level of a single host. We also make the simplifying assumption that the sequences within the database are independent and identically distributed (iid). It has been demonstrated by both analytical theory and experimental comparisons that both of these assumptions are valid for sufficiently large and diverse clinical sequence databases containing strains that are phylogenetically proximate (e.g., belong to the same viral subtype) [12, 81, 106, 111]. We seek the maximum entropy model consistent with the two lowest moments of the amino acid distribution [148, 149], namely the frequency with which the 20 amino acids are observed at each single position and each of the $20 \times 20 = 400$ pairs of amino acids are observed in each pair of positions. The resulting model is the simplest non-trivial fitness landscape capable of capturing epistasis at the level of pairs. A natural extension would be to include higher order epistatic effects (e.g., triplets, quads, etc.). It has been shown, however, that the exponential explosion in the model parameters associated with higher order terms quickly

degrades both the numerical stability and predictive performance of the model [120]. By terminating the expansion at second order, we determine not the intrinsic one- and two-body coefficients, but effective coefficients that implicitly capture higher-order correlations and which we have shown to reproduce both the three- and four-body amino acid frequencies [12, 81]. This is consistent with observations that effective pairwise interaction models can reproduce higher order correlations from the emergent interactions of the constitutive pairs [115, 122, 140, 141].

The maximum entropy model that emerges from this analysis is the infinite-range Potts spin glass from statistical physics [81, 111, 150]. Given an N-residue viral proteome in which each position can be occupied by one of 20 naturally occurring amino acids, $\vec{A} = \{A_i\}_{i=1}^{N}$, where A_i specifies the identity of the amino acid residue in position i, the Potts model specifies the probability of observing that sequence as [81, 111],

$$P(\vec{A}) = \frac{1}{Z} e^{-E(\vec{A})}, \qquad E(\vec{A}) = \sum_{i=1}^{N} h_i(A_i) + \sum_{i=1}^{N} \sum_{j=i+1}^{N} J_{ij}(A_i, A_j), \tag{2.4}$$

where E is a fictitious dimensionless "energy", $E(\vec{A})$ is the spin glass Hamiltonian specifying the mapping from sequence to energy, and the normalizing factor Z is the partition function. The Hamiltonian comprises N 20-element vectors $\{h_i\}$ specifying the contribution of each amino acid in each single position to the sequence energy, and $N(N-1)/2$ 20×20 matrices $\{J_{ij}\}$ specifying the contributions of each pair of amino acids in each pair of positions. Since we fit our model against population level sequence databases, $P(\vec{A})$ should be interpreted as the probability of observing a particular strain within the ensemble of infected hosts. We, and others, have demonstrated using analytical theory [106, 109], computer simulations [106], and comparisons against *in vitro* and *in vivo* data [12, 81, 111] that the prevalence of a strain at the population level $P(\vec{A})$ is a good proxy for intrinsic viral replicative capacity within an infected host $f(\vec{A})$. The Potts Hamiltonian therefore defines the (relative) replicative fitness landscape f of the virus as $f(\vec{A}) \propto e^{-E(\vec{A})}$. If required, the constant of proportionality can

be determined by fitting against experimental fitness measurements for a small number of mutant strains [81, 111].

The model parameters $\{h_i, J_{ij}\}$ are fitted such that eqn. 2.4 reproduces the one- and two-body amino acid frequencies observed in the MSA, and may be computed in many ways, including mean field and post-mean field approximations [144, 151–155], message passing [143], adaptive cluster expansions [111, 156], pseudo-likelihood maximization [157], minimum probability flow [158], and Monte Carlo sampling [139, 159]. We have developed an approach employing iterative gradient descent and Monte Carlo evaluation of model probabilities that provides highly accurate parameter fits. We accelerate convergence using a combination of Bayesian regularization, initial mean field parameter estimates, and parallel Monte Carlo chains exploiting CPU or GPU multicore architectures. The full mathematical details of this method are reproduced in appendix A, and an open-source C++ code implementing our approach is available for free public download as detailed in appendix B (also see ref. [12, 111]).

In the most general form of eqn. 2.4, each h_i is a 21-element vector – 20 amino acids plus an unknown/blank amino acid – and each J_{ij} is a 21-by-21 matrix. In practice we restrict z_i to the subset of amino acids actually observed in the MSA, truncating the h_i vectors and J_{ij} matrices and reducing the number of model parameters, thus accelerating model convergence. The penalty for this simplification is that the fitted model is not capable of making fitness predictions for viral strains containing amino acids not observed within the MSA. For an MSA comprising sufficiently many sequences, we expect broad coverage of the mutationally accessible space of viable viral strains, and the observation of a strain containing an unobserved amino acid to be a rare event. As discussed in ref. [111] and section C.2, we can straightforwardly incorporate parameters for unobserved amino acids using pseudo-counts to set a lower bound on their expected frequency and over come this drawback as needed.

2.6 Applications

To illustrate our approach, we here demonstrate its application to a cartoon two-residue virus as the simplest possible model system capable of exhibiting epistatic couplings between residues. In the following chapters we show its application to determine the fitness landscape of the hepatitis C virus RNA-dependent RNA polymerase (protein NS5B) and show how this landscape may be used for *in silico* design of vaccine immunogens (see chapters 3 and 5). We have also employed our approach to determine fitness landscapes for multiple proteins of both hepatitis C virus and HIV, perform *in silico* design cytotoxic T-cell vaccine immunogens, and make the first theoretical prediction of a viral error catastrophe in an empirically-defined viral fitness landscape for the p6 HIV protein [12, 81, 96, 106, 111] (also chapter 4).

2.6.1 A toy model of a two-residue virus

We first consider a toy virus comprising precisely two amino acid residues. This is the simplest possible system capable of exhibiting epistasis and possesses the appealing property that the fitness landscape can be easily visualized in three dimensions. We specify the fitness landscape of this virus by fiat, and demonstrate the capacity of our approach to correctly infer the fitness landscape from sufficiently many observations of viral strains over the *a priori* unknown fitness landscape.

Identifying the particular amino acid residue in position $i = \{1, 2\}$ by an index $A_i = \{1, 2, \ldots, 20\}$ arbitrarily ordering the 20 natural amino acids, we adopt as the "true" fitness landscape for the cartoon virus the Potts spin glass (eqn. 2.4) with parameters specified as,

$$h_1(A_1) = \frac{A_1 - 1}{4}, \quad h_2(A_2) = \frac{A_2 - 1}{8}, J_{12}(A_1, A_2) = 1.5(e^{-((A_1-4)^2+(A_2-4)^2)} - e^{-((A_1-8)^2+(A_2-9)^2)}). \tag{2.5}$$

This fitness landscape was designed to possess a simple but non-trivial topography, pos-

sessing a global maximum at the sequence [1,1], a global minimum at [20,20], a local maximum at [9,8], and a local minimum at [4,4] (figure 2.3). Since the form of the true fitness landscape and that inferred by our inference technique have the form of a Potts spin glass, our approach should be capable of accurately reproducing the one- and two-body amino acid frequencies and predicting strain fitness given sufficiently many observations of viral sequences over the landscape.

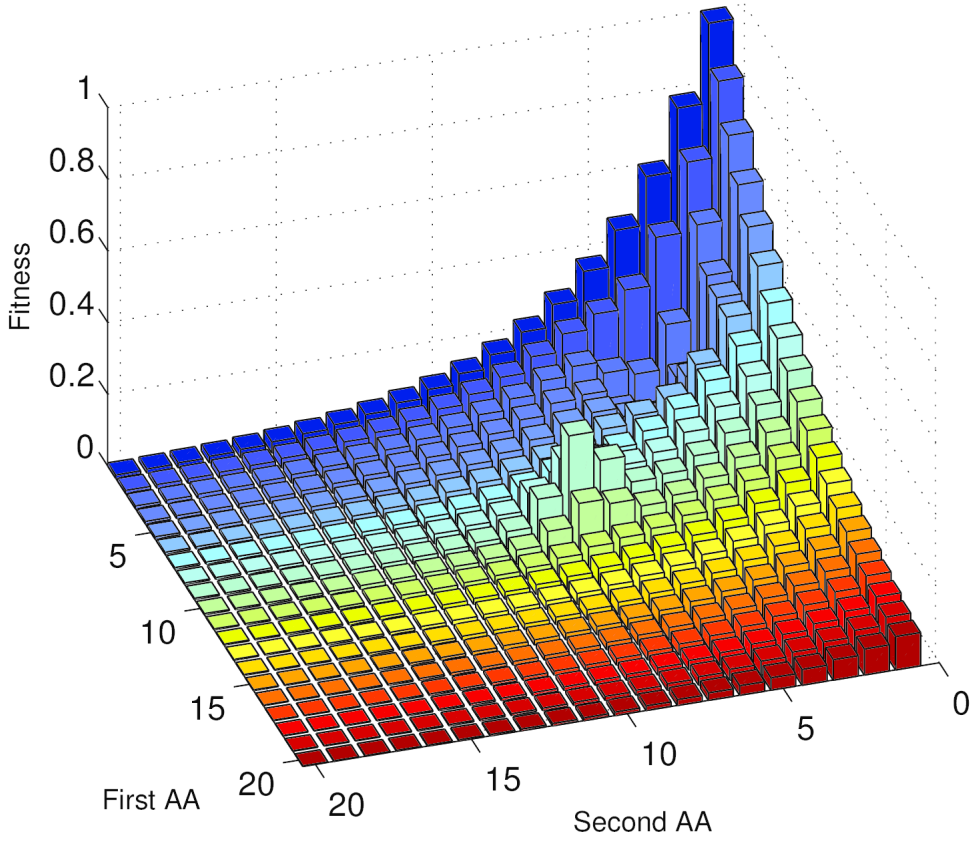


Figure 2.3: Fitness landscape of a cartoon viral protein possessing two amino acid residues specified by eqn. 2.5. We index the amino acid residues available to position i as $A_i = \{1, 2, \ldots, 20\}$, but their ordering is arbitrary. A particular sequence $[A_i, A_j]$ can mutate into any other sequence in the same column by making a point mutation in the first residue, and to any other sequence in the same row by making a point mutation in the second. The fitness landscape was designed to possess a global maximum at the sequence [1,1], a global minimum at [20,20], a local maximum at [9,8], and a local minimum at [4,4].

To test this hypothesis, we generated realizations of viral strains sampled over the land-

MSA size	ρ_{P_1} (p value)	ρ_{P_2} (p value)	RMSE $\{h_i\}$	RMSE $\{J_{ij}\}$
10^0	0.47 (2.2×10^{-3})	0.22 (1.2×10^{-5})	14.7	19.6
10^1	0.52 (6.7×10^{-4})	0.25 (1.2×10^{-7})	17.2	40.9
10^2	0.72 (1.7×10^{-7})	0.37 (3.3×10^{-14})	23.1	93.5
10^3	0.93 (1.9×10^{-18})	0.81 (8.0×10^{-96})	14.7	81.3
10^4	0.99 (1.1×10^{-36})	0.97 (1.1×10^{-238})	2.6	48.5
10^5	1.00 (3.1×10^{-44})	1.00 ($< 1.0 \times 10^{-308}$)	0.7	5.5

Table 2.1: Quality of fitness landscape reconstruction for a cartoon 2-residue virus with a Potts spin glass fitness landscape with parameters specified by eqn. 2.5 for multiple sequence alignments containing various numbers of sequences. We report the Pearson correlation coefficient of the analytically computed one-body (ρ_{P_1}) and two-body (ρ_{P_2}) amino acid frequencies between the true and fitted fitness landscape. We also report the root mean squared error in the inferred $\{h_i\}$ and $\{J_{ij}\}$ parameters relative to their true values.

scape in proportion to their fitness using Markov chain Monte Carlo sampling [81]. We assembled strains generated in this manner into multiple sequence alignments containing different numbers of sequences, then used these alignments and our iterative gradient descent approach to infer the 40 $\{h_i\}$ and 400 $\{J_{ij}\}$ parameters from the data [81, 111]. We report in table 2.1 a quantitative assessment of the capacity of our inferred model to reproduce the one- and two-body amino acid frequencies observed in the MSA, and the agreement of the inferred model parameters with those of the true landscape. Containing just two mutating residues, the one- and two-body mutational frequencies of the true and fitted models can be computed analytically by enumerating over all viral mutants. We present in figure 2.4 parity plots of the true and predicted fitness computed for all possible viral strains.

As anticipated, we find that the fitted model can reproduce the one- and two-body mutational frequencies and predict strain fitness to arbitrarily high fidelity given sufficiently many sequences in the MSA. The agreement between the inferred and true model parameters also improves with the number of sequences in the MSA, but remains non-zero even for near-perfect reproduction of the one- and two-body mutational frequencies. This phenomenon appears to result from the existence of "null spaces" or "sloppy directions" within the $\{h_i, J_{ij}\}$ parameter space such that particular combinations of the parameters can be adjusted in concert without significantly perturbing the energy (i.e., fitness) of a strain or its one- and

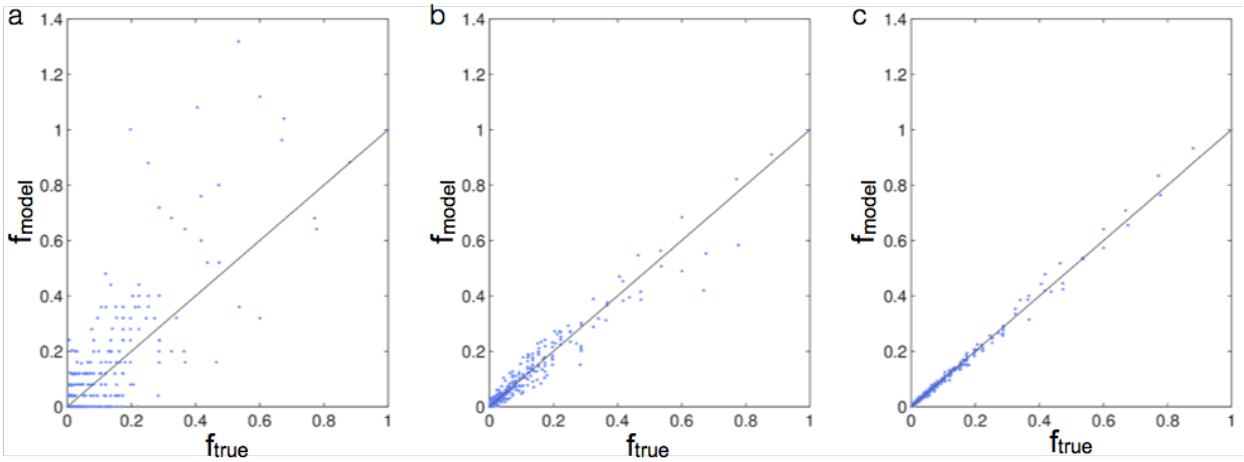


Figure 2.4: Parity plots of true versus reconstructed fitness for the cartoon $M = 2$ residue viral protein possessing the fitness landscape illustrated in figure 2.3. The plots illustrate for the ensemble all $20^M = 400$ viral strains the true strain fitness specified by eqn. 2.5 against the fitness predicted by models reconstructed from multiple sequence alignments containing (a) 10^3, (b) 10^4, and (c) 10^5 viral strains.

two-body mutational probabilities [160]. This simple toy problem indicates that it is possible to reconstruct fitness landscapes to high precision given sufficiently many sequences, but that care should be taken in assigning meaning to individual h_i and J_{ij} values.

Chapter 3

Using Fitness Landscapes in Static Design

3.1 Introduction

In this chapter we detail our development of empirical fitness landscapes for hepatitis C virus (HCV), their validation against experimental and clinical data, and their use in high-throughput screening and computational vaccine design.

Specifically we have inferred the fitness landscape for the RNA-dependent RNA polymerase – nonstructural protein 5B (NS5B) – within HCV that is known to be an important target for natural and therapeutic control [37, 161, 162]. We consider HCV genotype 1a, the most prevalent infecting genotype in the United States, which is responsible for 35-50% of infections domestically [163–165] and ∼60% worldwide [166]. As described below, we validate our fitness landscape in comparisons with experimental measurements and clinical data, and combine our models with population level immunological data to design cytotoxic T lymphocyte (CTL) immunogens predicted to have high efficacy and broad coverage in the United States population. By exhaustively screening the 16.8 million possible T-cell vaccine immunogens targeting epitopes in NS5B, we have identified 86 optimal formulations that we propose for experimental testing. By reducing the search space of immunogen candidates by over five orders of magnitude, our approach can massively reduce the time, expense, and labor of experimental vaccine development, guiding and accelerating the search for a HCV vaccine.

Most of this chapter is an excerpt from ref. [81]: G.R. Hart and A.L. Ferguson "Empirical fitness models for hepatitis C virus immunogen design" Physical Biology 12 066006 (2015)

3.2 Materials and methods

3.2.1 MSA construction and cleaning

In this work, we compute the fitness landscape for the HCV genotype 1a nonstructural protein 5B. NS5B is a 591-residue RNA-dependent RNA polymerase responsible for the replication of the viral genome [162]. Targeting epitopes within this protein has been linked with spontaneous clearance of HCV infection [4, 8, 48]. Inference of the fitness landscape for NS5B can identify vulnerable targets within this protein to guide rational immunogen design.

In principle, our approach is capable of computing fitness landscapes for the entire 3011-residue HCV proteome, but too few sequences currently exist in clinical databases to construct statistically robust models. Accordingly, we are currently constrained to compute landscapes for individual proteins for which the ratio of sequences to residues exceeds approximately 3:1 such that our models are statistically robust. As more HCV sequences become available, we intend to compute full proteome models capable of capturing the impact on viral fitness of mutational couplings within, and between, all HCV proteins.

A multiple sequence alignment (MSA) of NS5B protein sequences was constructed from consensus strains sequenced from hosts infected with HCV genotype 1a. We restricted these sequences to drug naïve hosts so as not to contaminate the models with the impact of drug therapy on the mutational evolution and effective fitness of the virus. Genotype 1a NS5B sequences were downloaded from the Los Alamos National Laboratory HCV database (`http://www.hcv.lanl.gov`) [167–170] (925 sequences), and supplemented by sequences provided from the lab of Dr. Todd Allen at Harvard Medical School (412 sequences; accession numbers are provided in table C.1). We constructed the MSA by aligning these sequences against the H77 reference sequence using the VirAlign tool available at (`http://www.hcv.lanl.gov`) [167]. We then cleaned the MSA by eliminating 361 sequences for which more than 5% of the 591 amino acid residues were unknown (i.e., inconclusively identified by sequencing), leaving

976 sequences containing 1.4% or fewer unknown residues. Finally, we identified 308 of the 591 residues as fully conserved within the MSA, allowing us to eliminate these residues from our fitting procedure. The final MSA contained 976 sequences and 283 mutating positions, leaving us with a ratio of sequences to positions of 3.4:1.

While it is possible that phylogenetic associations between the sequences can be important, based on the geographic and ethnic diversity of the hosts within the database, we make the simplifying assumption that the sequences are independently and identically distributed. We validate this assumption in *post hoc* comparisons of our model against independent experimental measurements and clinical data. It is, however, of continuing interest to us to develop methodologies to deconvolute effects of phylogeny and intrinsic fitness [106], which is expected to be of particular importance in the analysis of sequences drawn from multiple viral subtypes.

We previously showed that the effects of host immune pressure on strain distribution is averaged out if the MSA samples a genetically diverse host population [12, 106, 137]. Thus, the inferred landscape reflects intrinsic viral fitness and contains no "footprints" of adaptive immunity. The geographic and ethnic diversity of the database suggests that the MSA contains sufficient genetic diversity to eliminate signatures of adaptive immunity. We confirmed this by estimating the frequency with which each amino acid position in NS5B is expected to be subject to immune pressure using the approach detailed in ref. [106] (also see section C.1). Direct comparisons of our model predictions against clinical data and experimental measurements described below, provide further support that our inferred NS5B fitness landscape does not contain footprints of adaptive immunity.

3.2.2 Fitness landscape inference

As discussed in chapter 2, there are many ways to solve the inverse problem of fitting the h_i and J_{ij} to reproduce the observed one and two-body amino acid frequencies [139, 143, 144]. In this work we employ our previously developed parallelized semi-analytical iterative gradient

descent algorithm [12, 111, 139]. We use Bayesian regularization employing a Gaussian prior to penalize large J_{ij} parameter values and stabilize and accelerate model convergence, $P(\{J_{ij}\}) = \Pi_{ij} e^{-\gamma J_{ij}^2}$ [171]. We verified that the inferred model parameters were robust to the regularization strength over the range $\gamma = 0.025$-0.20. Full mathematical and algorithmic details of the fitting procedure are provided in appendix A.

We note that a model constructed to reproduce the higher order moments of the amino acid distribution (triplets, quads, etc.) would, in principle, better describe the distribution of viral strains within the evolutionarily accessible sequence space. However, the number of parameters increases exponentially with model order, making inference computationally intractable. The fitted model described by eqn. 2.4 not only accurately reproduces the one- and two-body amino acid distributions from which it was parametrized, it also *predicts* the three-body frequencies (figure C.1 in the appendix), demonstrating that it implicitly captures higher order mutational correlations in the MSA [12, 111]. Furthermore, we have shown that statistically-significant mutational couplings important in dictating viral fitness exist within the MSA, and that our model captures these epistatic effects (figure C.2 in the appendix).

3.3 Results and discussion

Before using our fitness landscapes for computational vaccine design, it is critical that we first test the capacity of our model to predict the fitness of strains not contained in the fitting data. This is done by direct quantitative comparisons of the fitness predictions of our models against independent experimental and clinical data not used in model construction. Specifically, we present five separate validations of the agreement of our NS5B fitness landscape with: (1) experimental measurements of the *in vitro* replicative fitness of mutant viral strains, (2) clinically documented high fitness escape mutations, (3) the location of escape mutations within particular epitopes, (4) the mutational evolution of HCV strains revealed

by longitudinal sequencing of infected hosts, and (5) clinically documented protective HLA types. We then proceed to use the validated landscape for the *in silico* design of CTL vaccine immunogens.

3.3.1 Comparison of model predictions with *in vitro* replicative fitness measurements

The first test of our model is to validate its predictions against *in vitro* fitness measurements. If the fitness landscape truly represents intrinsic viral fitness, then by eqn. 2.4 the relation $E(\vec{z}) \propto -\log\left(f(\vec{z})\right)$ should approximately hold. We tested this by gathering 31 previously published experimental measurements of *in vitro* replicative fitness of engineered NS5B mutants containing up to four polymorphisms [3–5].

Since genotype 1a lacks a robust replicon, two of these studies, refs. [4, 5], used a chimera 1a/1b replicon, with the polymerase coming from the genotype 1a strain H77 (GenBank Accession No. M67463), to enhance the signal. Our MSA did not contain six of the residues in this H77 polymerase. In order to assign an energy (fitness) to this sequence, we augmented our model to incorporate these six additional residues using pseudo-counts as detailed in the appendix (section C.2 also see ref. [111]. The remaining study, ref. [3], used the standard H77S.3 genotype 1a replicon with a mutation (C2466S) that was already present in our model.

A plot of the energy of each of the 31 mutants assigned by our model relative to the wild type, $(E - E_{wt})$, versus the logarithm of the relative fitness, $\log(f/f_{wt})$, exhibits a strong and statistically significant negative correlation (figure 3.1). For each mutant the wild type strain is defined as the replicon used in the experimental assay. If instead we uniformly define the wild type sequence as the H77S.3 reference strain, we can assign energies to each strain using our unaugmented model and the strong and significant correlation remains (figure C.3).

Within a particular individual, the *effective* fitness of the virus is a function of both

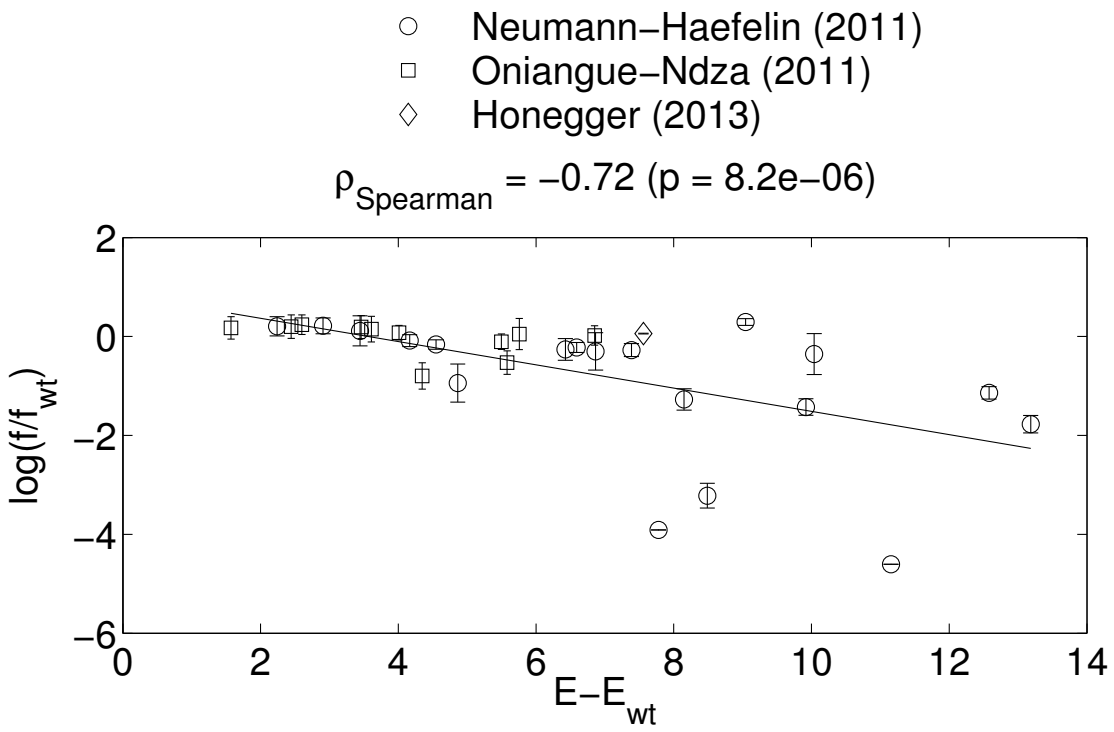


Figure 3.1: Comparison of the *in vitro* replicative fitness relative to wild type, f/f_{wt}, measured for 31 engineered NS5B mutants containing up to four polymorphisms [3–5] against the energy relative to wild type, $(E - E_{wt})$, of each strain predicted by our model. A strong and statistically significant negative correlation, $\rho_{Spearman} = -0.72$ ($p = 8.2 \times 10^{-6}$, two-tailed t-test), validates our fitted model as a good predictor of intrinsic viral fitness. A linear least-squares fit is provided to guide the eye, and error bars delineate estimated uncertainties in the measured relative fitness.

the intrinsic viral fitness and the host immune pressure. Assays of *in vitro* replicative fitness measure intrinsic fitness. The statistically significant correlation between our model predictions and the *in vitro* fitness measurements provides strong empirical evidence that our models capture intrinsic viral fitness. This provides validation of the expectation that the impact of host immunity on the viral fitness have been averaged out over the MSA (cf. section C.1), and that the resulting fitness landscapes are not contaminated by footprints of adaptive immunity [12, 106, 137].

3.3.2 Predicted fitness costs of clinically observed escape mutations

HCV establishes chronic infection in 70-80% of patients despite many of them mounting an adaptive immune response [42, 172]. While there does not appear to be a single dominant mechanism for secondary failure of the adaptive immune system, escape mutations in CTL epitopes have been shown to play an important role [16, 173, 174]. Viral strains can escape CTL recognition by establishing one or more point mutations within, or flanking, the target epitope [175]. Clinically observed escape mutations are expected to be those that allow the virus to evade immune recognition with minimal cost to its replicative fitness. If our inferred fitness landscapes accurately represent intrinsic viral fitness, high fitness escape mutations should be assigned low energies by our model, providing the second test of our model.

Using published accounts of escape strains sequenced from HCV infected individuals, and data on proximate HLA associated polymorphisms, we compiled a list of 24 single, eight double, and three triple escape mutations within NS5B [3–6, 11, 16, 176–178]. Taking the H77 reference sequence used in the majority of *in vitro* measurements (cf. section 3.3.1) as the wild type, we computed the energy assigned by our model to each mutant relative to the wild type, $\Delta E = (E - E_{wt})$. Some of the observed polymorphisms were not present in the MSA, and therefore absent in our model. In order to make energy predictions for all escapes we augmented the model with the missing amino acids using the procedure detailed in the appendix (section C.2). The particular mutants, HLA associated epitopes, and computed energies are listed in table C.2. We present the location of each mutant on the energy spectrum of all possible mutants in figure 3.2.

Single mutants. Twenty-one of the 24 single mutants fall in the bottom 31.2% of the spectrum of 944 possible single mutants ($\Delta E < 6.82$). The remaining three – K2471R, Q2467K, and R2937S – fall in the 36^{th}, 74^{th} and 96^{th} percentiles respectively. That these are apparently rather high energy (low fitness) escape mutations can be understood by the fact

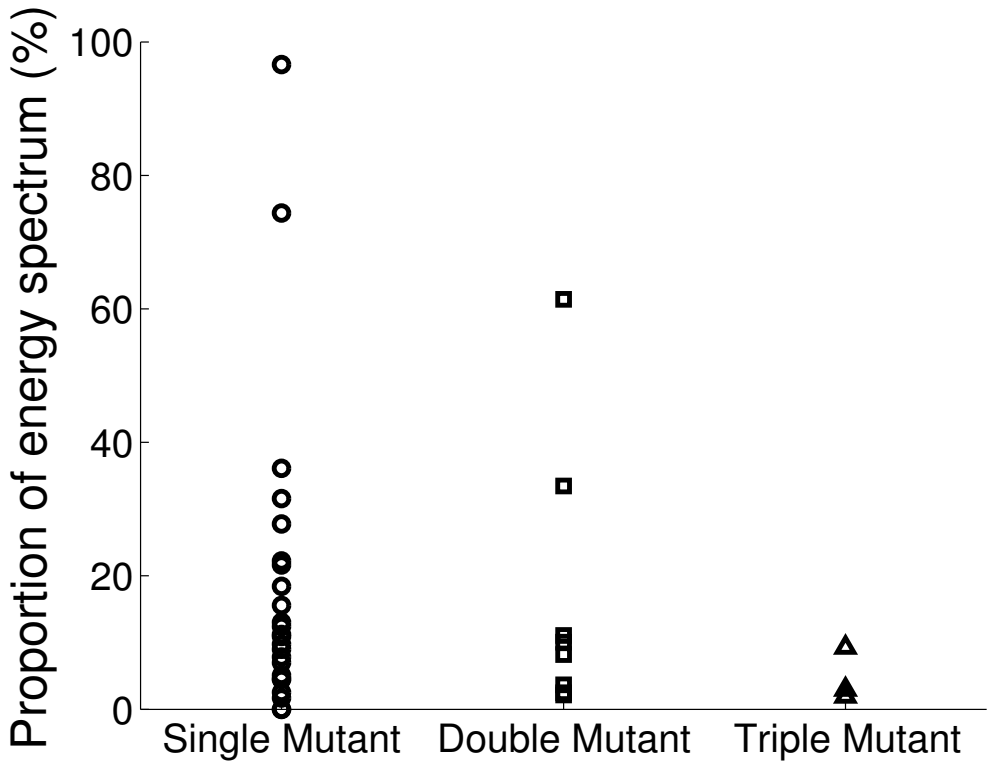


Figure 3.2: Clinically observed escape mutants are low energy (high fitness) strains within our model. The abscissa records the 24 single, 8 double, and 3 triple clinically reported escape mutations within NS5B. The ordinate locates the mutants on the energy spectrum of all possible mutants of the same order as assigned by our model. In all cases residues outside of the epitope were set to the H77 reference sequence. The particular mutants are listed in table C.2.

that they are almost always observed in concert with compensatory mutations that offset the fitness penalty of these polymorphisms [3, 4]. We discuss the details of these compensatory mutational patterns in section C.3 of the appendix.

Double mutants. Six of the eight double mutants fall in the bottom 11.1% of the spectrum of all possible 444,005 doubly mutated strains ($\Delta E < 10.76$). The remaining two – K2937G/I2940T and Q2467K/K2471R – fall in the 33^{rd} and 61^{st} percentiles, respectively. These apparently high energy (low fitness) escape mutations are almost always observed in concert with compensatory mutations that offset their fitness cost [3, 4] (section C.3).

Triple mutants. Of the three clinically observed triple mutants, two fall in the bottom 3% of the energy spectrum of all 138,734,750 triple mutants ($\Delta E < 10.30$), with the third

lying in the 9^{th} percentile ($\Delta E < 16.48$).

That our model predicts low fitness costs to clinically observed escape mutations provides further support that it quantifies intrinsic viral fitness uncontaminated by footprints of adaptive immunity [12, 106, 137].

3.3.3 Predicted location of escape mutations in CTL epitopes

As a third test of our model we assess its ability to predict which positions within an epitope are most likely to support escape mutations. A viral strain containing an epitope recognized by the host immune system can escape by making one or more mutations in the target epitope [175]. Under the simplifying assumption that all mutations are equally effective in abrogating CTL recognition, escape mutations should minimize the fitness penalty and thus maximally preserve viral fitness. If our landscape is a good model of intrinsic viral fitness, we should expect that the clinically documented escape mutations correspond to the highest fitness (lowest energy) mutations within the epitope.

By cross-referencing the list of escape mutations compiled in section 3.3.2 with defined CTL epitopes [3–6, 11, 16, 176–178], we identified nine exact epitopes and three epitope-containing regions with clinically documented escape mutations. We calculated the energy cost, ΔE, associated with each polymorphism observed within our MSA for these epitopes using the H77 reference sequence as the wild type sequence.

Within the *B*27* restricted $\text{GRAAICGKY}_{2936-2944}$ epitope, for example, the two most common escape mutations, R2937K and I2940T, have the lowest, and third lowest, energy costs, respectively, and a less commonly observed escape, K2943R, has the second lowest cost (figure 3.3). All three point mutations are the lowest energy polymorphisms at their respective positions. The R2937S escape has the largest energy cost, but – as mentioned in section 3.3.2 – in addition to being only rarely observed it is always accompanied by the two compensatory mutations E2875K and P2881Q that fall outside of the epitope region [4].

Of the remaining 11 epitopes, nine are reported to escape with a single mutation while

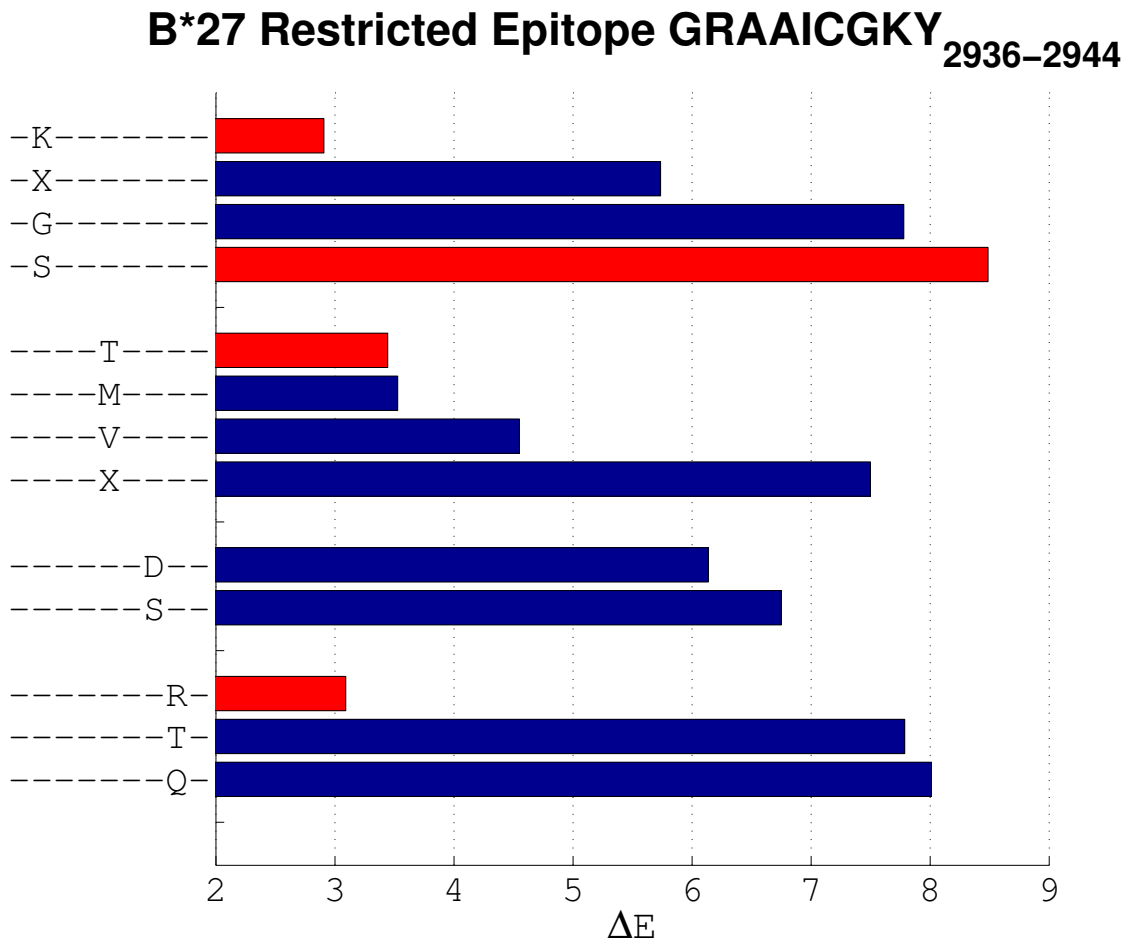


Figure 3.3: Comparison of the energy costs (fitness penalties) relative to the H77 wild type reference sequence predicted by our model for all polymorphisms observed within our MSA occurring within the *B*27* associated $\text{GRAAICGKY}_{2936-2944}$ epitope. The energy cost associated with each single mutation, ΔE, is along the abscissa, and the mutations are shown along the ordinate. Dashes indicate unmutated positions, and letters the mutant amino acid residue. The letter X indicates an unknown amino acid type that was inconclusively identified by experimental sequencing within the ensemble sequences constituting the MSA used to fit our model. The greater the energy cost, the higher the fitness penalty. The bars corresponding to clinically observed escape mutations – R2937K, R2937S, I2940T, and K2943R – are colored in red. Blue bars denote mutations for which no specific clinical information is available. The two most commonly observed clinical escape mutations, R2937K and I2940T, correspond to the two of the three lowest energy (highest fitness) polymorphisms predicted by our model.

the other two require at least two mutations to abrogate recognition [3, 5, 6, 11, 16, 176–178]. Of the nine epitopes requiring only a single mutation, seven possess at least one documented escape mutation corresponding to the first, second, or third lowest energy polymorphism

(figures C.4A-G). The eighth epitope's escape mutation lies at the ninth lowest energy cost position, but there is evidence that a number of mutations in this epitope do not effectively abrogate CTL pressure [8] suggesting that higher fitness cost polymorphisms may be required to effect escape (figure C.4H). The escape mutation reported in the ninth epitope is predicted by our model to be the seventh lowest cost mutation (figure C.4I).

For the two epitopes requiring at least two mutations to effectively escape, each has the lowest or second lowest single mutation as one of the two required mutations. In $\text{ARMILMTHF}_{2841\text{-}2849}$, one of the double mutants known to confer escape has the lowest energy cost of all double mutants in the epitope (figure C.5A). In $\text{THFFSVLIARDQ}_{2847\text{-}2858}$, the double escape falls in the bottom fifth of all double mutants in the epitope (figure C.5B).

Our model predicts that 9 of 12 epitopes have documented escape mutations among the three least costly mutations. In agreement with the expectation that mutational escape should occur among the lowest fitness cost positions, this result provides further support that our model represents intrinsic viral fitness.

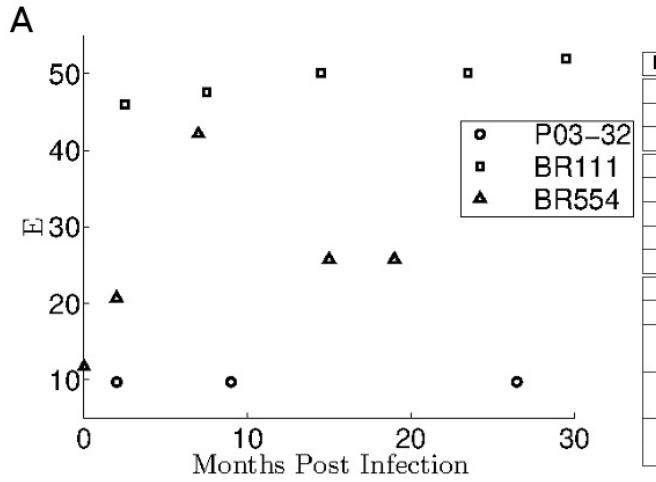
3.3.4 Viral evolution in longitudinal studies of individual hosts

As a fourth test of whether our model reflects intrinsic viral fitness,we compared our predictions to longitudinal studies of four drug naïve hosts over the first 1.5-4 years of infection [3, 6]. Two of these patients (BR554 and M003) possessed viral strains containing residues not present in the MSA we used to fit our model, requiring us to augment our model with pseudo-counts to assign energies to strains containing these polymorphisms using the approach detailed in section C.2. Possessing only a fitness landscape for protein NS5B, our analysis and interpretation of the longitudinal data is necessarily restricted to the viral mutations within, and CTL responses against, a single HCV protein. We shall show that the longitudinal data is consistent with the predictions of our NS5B fitness model, but note that in the absence of full proteome fitness models our analysis necessarily neglects the fitness impact of T-cell and B-cell responses against other proteins, and other factors influencing viral

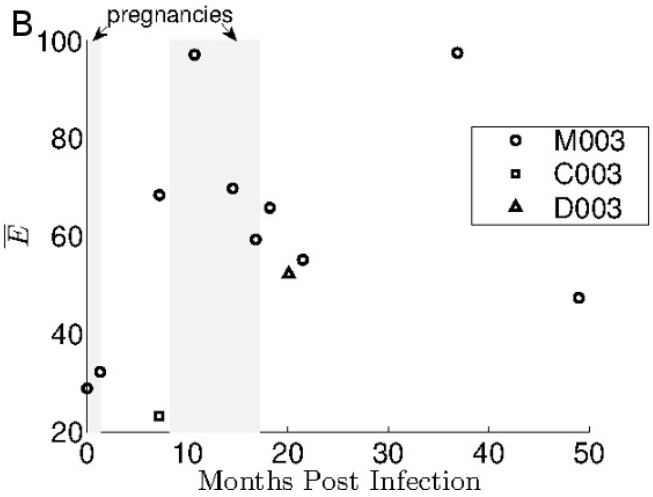
dynamics such as innate immune pressure, interferon production, and hepatocyte apoptosis. Furthermore, incomplete characterization of the full viral quasispecies at discrete time points means that the longitudinal data cannot capture events occurring below the experimental detection threshold, or on time scales below the sampling frequency.

Longitudinal consensus data. We first consider three individuals – 03-32, BR111, and BR554 – for whom consensus sequence data at each time point was reported by Kuntzen *et al.* in ref. [6] and present the energy assigned to the reported consensus strain within each individual as a function of time (figure 3.4A). Patient 03-32 was reported to mount only one detectable CTL response within NS5B, against epitope $\text{LGVPPLRAWR}_{2912-2921}$ (HLA association unknown) [6], and over the course of the 26.5 month study the infecting consensus sequence did not evolve any mutations in the NS5B protein. This situation is consistent with infection by a high fitness founder strain that is subject to low immune pressure and/or for which escape mutations within targeted epitopes carry very large fitness penalties. In good agreement with this scenario, our model assigns a low energy (high fitness) of $E = 9.7$ to this strain, placing it in the 15^{th} percentile of the energy spectrum of the 976 strains in our MSA. Furthermore, as we discuss below (see section 3.3.5), our model predicts escape mutations in this epitope to be among the costliest of all known epitopes in NS5B. This suggests the evolution of polymorphisms in this protein region to carry very high fitness costs and may take years to arise [6], thereby rationalizing the absence of observed escape mutations despite the reported CTL immune response.

Patient BR111 likewise mounted only one detectable CTL response within NS5B, against the *HLA-A*02* restricted $\text{ALYDVVSKL}_{2594-2602}$ epitope, but accumulated three mutations over a period of 29.5 months [6]. Our model assigns a high energy (low fitness) of $E = 46.0$ (96^{th} percentile of the MSA energy spectrum) to the consensus sequence reported at 2.5 months. The first mutation to arise, L2604P at 7.5 months, increased the energy to $E = 47.6$ and is just outside the targeted epitope [6]. The epitope prediction tools on the Immune Epitope Database (`http://www.iedb.org`) [179] indicate that this flanking mutation impairs



Patient	Month	E	Mutations
03-32	2.0	9.7	Original sequence
	9.0	9.7	No mutations
	26.5	9.7	No mutations
BR111	2.5	46.0	Original sequence
	7.5	47.6	L2604P
	14.5	50.1	L2604P E2692Q
	23.5	50.1	L2604P E2692Q
	29.5	52.0	S2482N L2604P E2692Q
BR554	0.0	11.6	Infecting sequence
	2.0	20.6	S2496L K2930R G2963S
	7.0	42.1	H2453N V2704I A2813V I2854M V2905M S2426N K2930R G2963S
	15.0	25.7	H2453N A2813V F2849L I2854M S2426N K2930R G2963S
	19.0	25.7	H2453N A2813V F2849L I2854M S2426N K2930R G2963S



Patient	Month	$\bar{E}$
M003	0.0	28.8
	1.3	32.2
	7.2	68.4
	10.7	97.0
	14.5	69.7
	16.8	59.3
	18.2	65.8
	21.5	55.1
	36.9	97.4
	49.0	47.3
C003	7.2	23.2
D003	20.1	52.2

Figure 3.4: Temporal evolution of viral fitness predicted by our model in longitudinal studies of four drug naïve patients over the first 1.5-4 years of HCV infection. (A) Viral evolution within three individuals – Patients 03-02, BR111, and BR554 – for whom consensus sequence data was available at each time point [6]. On the left we present plots tracking the energy (fitness) of the consensus strain predicted by our model at each time point. Low energy corresponds to high fitness. On the right, we list the particular mutations observed within the NS5B region of the consensus strain, and the energies assigned to the strains by our model. (B) Clonal sequencing results for the viral evolution within an infected mother – Patient M003 – who gave birth to two infected children – Patients C003 and D003 – over the course of the study [3]. The plot on the left tracks as a function of time the average energy over all sequences reported at that time point, $\bar{E}$, for each of the three hosts. The shaded periods in the plot indicate the pregnancies of M003 with C003 and D003. The table on the right lists the numerical values of the data in the plot. A full accounting of the individual sequences, mutations, and energies are provided in the appendix C.4

antigen processing and presentation and thus is selected despite its slightly detrimental impact upon viral fitness. This hypothesis is further supported by the observation that our

model predicts all other point mutations within the epitope to carry higher fitness penalties. The polymorphisms E2692Q – arising at 14.5 months and increasing the energy by $\Delta E = 2.5$ – and S2482N – arising at 29.5 months and increasing the energy by a further $\Delta E = 1.9$ – both elevated the energy (reduced fitness) and are not part of any reported CTL epitope. Both polymorphisms fall within peptides 20 amino acid long that are known to contain HLA Class II epitopes [179–181] and may offer escape from $CD4^+$ T-cells, although the patient's $CD4^+$ response was not characterized.

For Patient BR554, consensus sequences over 19 months and the infecting sequence were available [6]. Our model assigns a low energy (high fitness) to the infecting strain of $E = 11.6$ (23^{th} percentile of the MSA energy spectrum). At 2 months, three mutations, K2930R, G2963S, and S2496L – the latter of which reverted by the next sampling point – arose, with relatively modest impact on fitness, increasing the energy by $\Delta E = 9.0$. At 7 months, six new mutations arose, H2453N, A2813V, I2854M, S2926N, V2704I, and V2905M – the latter two of which reverted by the next sampling point – with a large impact on the fitness, increasing the viral energy by $\Delta E = 21.5$. Two of these mutations – H2453N and I2854M – occurred within the two epitopes against which CTL responses were reported – $\text{SLLRHHNLVYSTTSRSA}_{2449-2465}$ and $\text{THFFSVLIARDQ}_{2847-2858}$ (HLA associations unknown) [6] – offering an explanation for the tolerance of such costly polymorphisms. The final mutation, F2849L, arose at 15 months within the $\text{THFFSVLIARDQ}_{2847-2858}$ epitope. Improving the fitness by $\Delta E = -16.4$, we suggest a role for this polymorphism in compensating for the fitness penalties associated with the escape mutations.

If our model really does capture intrinsic viral fitness, then in individuals who mount weak immune responses, our model should predict the virus to maintain high fitness (low energy) throughout the course of infection. In contrast, in individuals who mount strong and broad immune responses, our model should predict the virus to exhibit a decrease in fitness (increase in energy) over time as it evolves escape mutations in response to host immune pressure. The viral load of Patient 03-32 remains nearly flat over the course of the

study at 10^6-10^7 copies per ml [6]. This observation is consistent with an ineffective CTL response, and in good agreement with our model predictions that the virus should maintain high fitness (figure 3.4A). In contrast, the viral loads of Patients BR111 and BR554 fall from 10^4 copies per ml at 2 months (the first measured time point), to 10^0 copies per ml at 4 months, before rebounding to approximately 10^7 copies per ml at the termination of the study [6]. This response of the viral load is consistent with a robust acute phase CTL response followed by secondary failure, and is in good agreement with our model predictions of an initial reduction in viral fitness (increase in energy) due to the appearance of escape mutations within identified CTL targets followed by mutational progression to a fitness (energy) plateau that may be attributed to viral escape (figure 3.4A).

Longitudinal clonal sequencing data. We now consider Patient M003 and the two children to whom she gave birth during the study and vertically transmitted HCV infection – Patients C003 and D003, delivered at 1.3 months and 17.2 months, respectively – for whom clonal sequencing data were reported by Honegger *et al.* in ref. [3]. For concision we present the mean energy assigned by our model to the reported sequences at each time point, $\bar{E}$, to show that the trends in viral fitness are consistent with the expected trends due to the maternofetal immune tolerance mechanism that suppresses the mother's immune response during pregnancy (figure 3.4B). We provide an in-depth reporting of each clonal sequence observed and its associated energy in the appendix C.4.

M003 became acutely infected with HCV while pregnant with C003. At month 0 of the study, the average viral energy was $\bar{E} = 28.8$, which increased slightly to $\bar{E} = 32.2$ at the time of delivery at 1.3 months. By 7.2 months (25 weeks after delivery of C003) the average energy of the viral strains in M003 had more than doubled to $\bar{E} = 68.4$, and by 10.7 months had further increased to $\bar{E} = 97.0$. Although M003 possesses *HLA-B*0801* and *HLA-B*1501* class I alleles, the maternofetal immune tolerance mechanism impaired the action of HCV-specific CTLs during pregnancy [3], allowing the virus to maintain high fitness (low energy) up until delivery of C003 at 1.3 months, making transient mutations uncorrelated

with host immune pressure (section C.4). After delivery, the maternofetal immune tolerance mechanism disappeared, and M003's strengthened immune response produced mutations in the *B*15* restricted LLRHHNMVY$_{2450-2458}$ and SQRQKKVTF$_{2466-2474}$ epitopes (section C.4). This caused a large decrease in average viral fitness (increase in energy) over the period 1.3-10.7 months.

The pattern of elevated viral fitness during pregnancy followed by reduced fitness after delivery is repeated during M003's second pregnancy with D003, which commences at about about 8.3 months. After the fitness low (energy high) at 10.7 months, the average viral fitness increased during the second and third trimester of pregnancy, with the average energy falling to $\bar{E} = 69.7$ by 14.5 months and $\bar{E} = 59.3$ by 16.8 months. This improvement in fitness is associated with some reversion of escape mutations within the B*15 restricted epitopes (cf. section C.4). After delivery of D003 the fitness falls (energy increases) due to the disappearance of maternofetal immune tolerance, with the average energy increasing to $\bar{E} = 65.8$ by 18.2 months. After a small improvement in fitness at 21.5 months ($\bar{E} = 55.1$) the fitness drops to its lowest value at 36.9 months, ($\bar{E} = 97.4$) 16.8 months after delivery of D003.

At the last measured time point of 49.0 months – four years after infection – the average energy falls to $\bar{E} = 47.3$ corresponding to an increase in viral fitness. This fitness increase is associated with the appearance of a new mutation within the *B*15*-SQRQKKVTF$_{2466-2474}$ epitope (section C.4). We suggest that this observation may correspond to the development of a slowly evolving compensatory mutation that restores viral fitness as the viral quasispecies anneals to the host adaptive immune pressure and progresses to chronicity.

Twenty-five weeks after delivery at 7.2 months, the viral strains within child C003 had an average energy of $\bar{E} = 23.2$, suggesting that the vertical transmission of HCV to the child by the mother has given rise to a fit viral quasispecies within the infant that is experiencing little, if any, immune pressure. C003 inherited the *HLA-B*0801* class I allele from M003 which restricts no known epitopes within NS5B, but *HLA-B*0801* epitopes in NS3 remained

unchanged during the course of the full first year of the child's life. As observed by the authors of the experimental study, vertical transmission of a highly fit HCV strain from the immunosuppressed mother may have "outrun" the nascent immune system of the infant [3].

HCV was also vertically transmitted to child D003. Sequences within D003 12 weeks after birth (at 20.1 months) contain polymorphisms within the HLA-B*15 associated epitopes $\text{LLRHHNLVY}_{2450-2458}$ and $\text{SQRQKKVTF}_{2466-2474}$ that were present in the mother before delivery, constituting a slightly fitter viral population ($\overline{E} = 52.2$) than those in the mother at time of delivery. D003 also inherited the *HLA-B*1501* class I molecule that restricts these epitopes, consistent with the absence of reversion in these epitopes within the first year of the child's life.

The predictions of our model for the temporal evolution of viral fitness correlate with the observed trends in viral load and CTL responses in longitudinal consensus sequencing data for three infected hosts [6], and also track the expected fitness trends due to maternofetal immune tolerance in longitudinal clonal sequencing study of an infected mother over the course of two pregnancies [3]. These findings provide strong support of the fitness predictions of our model.

3.3.5 Clinically documented protective CTL responses

HCV exists within an infected individual as a collection of closely related mutant strains known as a quasispeices [34, 174]. We hypothesized that our fitness landscapes can identify CTL immune responses clinically documented to be particularly effective against HCV as those that preferentially eliminate high-fitness strains within the quasispecies [12]. This constitutes a fifth and final test of the model.

Following the approach in ref. [12], we performed Metropolis Monte-Carlo sampling of our fitness landscapes to generate a population of 99,990 viral strains in which each strain is represented in proportion to its intrinsic fitness specified by our model. This population may be conceived of as a hypothetical quasispecies of viral mutants distributed over the fitness

landscape in the absence of immune pressure [106]. The average energy of a strain in this ensemble, $\langle E \rangle$, provides a measure of the average fitness of the viral population.

We propose that a good immune response will eliminate high fitness (low energy) strains from the population, causing the average fitness (energy) of the population to decrease (increase). We simulate the effect of a particular CTL response upon the population by removing all strains in the viral population possessing wild type residues within the CTL epitope. We then quantify the efficacy of the response as the change in the average energy of a strain in the population, $\Delta\langle E \rangle$, upon eliminating those strains. We make a "worst case" assumption by assuming no cross-reactivity such that single polymorphisms are sufficient to abrogate immune recognition of an epitope [12, 126].

Certain HLA alleles – most notably *B*27* and *B*57* – are associated with natural control and spontaneous clearance of HCV [4, 7, 8, 126, 176, 182, 183]. If our model is a good reflection of the intrinsic viral fitness landscape, naturally protective CTL responses should be associated with large increases in $\Delta\langle E \rangle$, corresponding to large decreases in the fitness of the viral population.

We identified from the Immune Epitope Database (`http://www.iedb.org`) 24 CTL NS5B epitopes that were both exactly defined and the HLA association known (table C.3) [179]. We then calculated $\Delta\langle E \rangle$ for each of these responses (figure 3.5). Of these 24 epitopes, two are reported as immunodominant – *B*57*–KSKKTPMGF$_{2629-2637}$ ($\Delta\langle E \rangle = 6.45$) and *B*27*–ARMILMTHF$_{2841-2849}$ ($\Delta\langle E \rangle = 3.85$) – and a third epitope as high on the dominance hierarchy – *A*02*–ALYDVVTKL$_{2594-2602}$ ($\Delta\langle E \rangle = 3.16$) [7–10]. Dominant epitopes are preferentially targeted among all of the epitopes associated with the same HLA molecules, effectively curtailing the CTL response against these other epitopes. The remaining 21 of the 24 epitopes are categorized as subdominant, meaning that strong CTL immune responses are typically not naturally elicited against these targets.

Of the three dominantly presented epitopes against which strong CTL responses are naturally directed, the two presented by protective HLA alleles – *B*57* and *B*27* – are

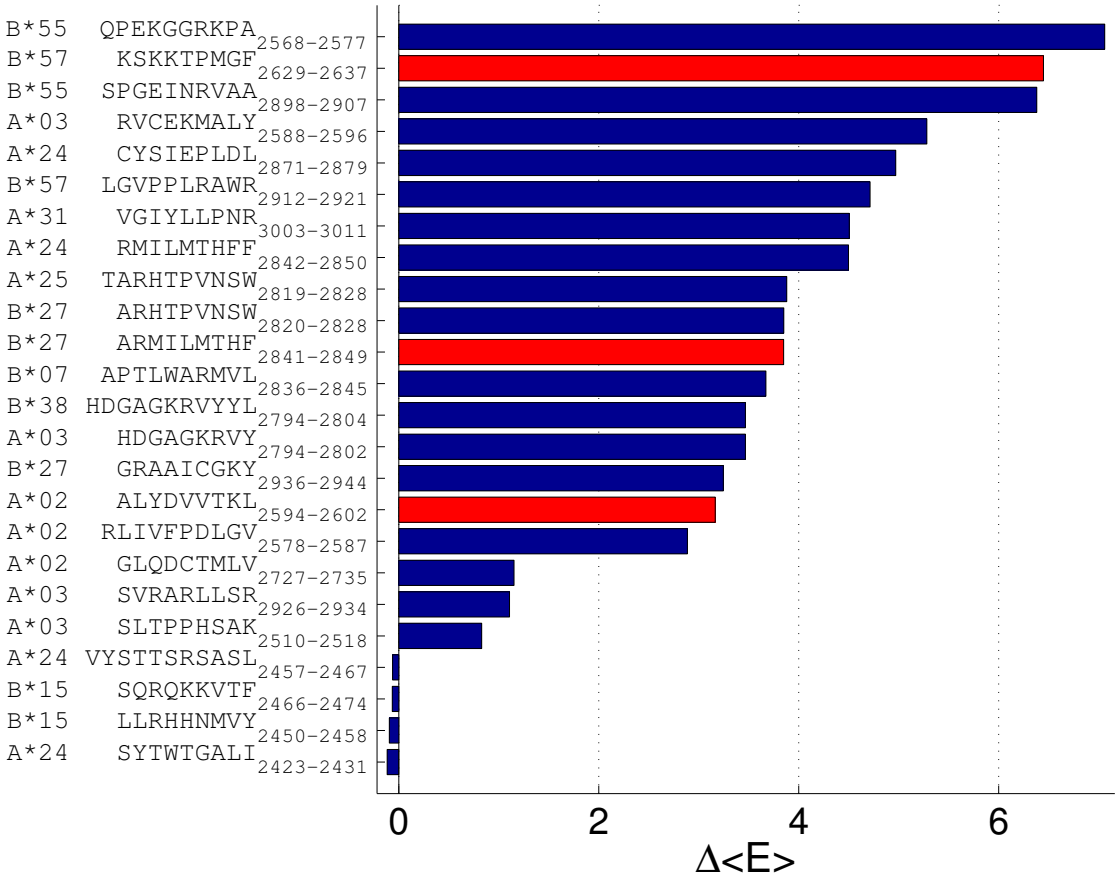


Figure 3.5: Ranking of 24 NS5B class I HLA epitopes according to the computed energy penalty, $\Delta\langle E\rangle$, imposed upon the viral ensemble. Epitopes that are reported as immunodominant are highlighted in red [7–10]. Of the three immunodominant epitopes, the two with the highest penalty are presented by protective HLA alleles associated with spontaneous viral clearance [4, 5, 8, 11]. The negative $\Delta\langle E\rangle$ values associated with the four lowest ranked epitopes results from their preferential elimination of low fitness viral strains from the quasispecies.

assigned higher $\Delta\langle E\rangle$ values by our model than those presented by the non-protective allele – *A*02* [4, 5, 8, 11]. The predictions of our model are consistent with the expectation that the epitopes presented by non-protective alleles have a lower impact on viral fitness than those presented by protective alleles. In other words, our model predicts that persons who naturally control HCV infection preferentially target epitopes that eliminate high fitness viral strains, leaving the viral quasispecies populated with relatively low-fitness escape strains.

Our model predicts that responses against nine subdominant epitopes presented by a

variety of HLA alleles impose fitness penalties upon the viral population that are as high or higher than the immunodominant responses of the protective alleles ($\Delta\langle E\rangle \geq 3.85$). For example, our model predicts that the immune response mounted against the *HLA-B*55* restricted epitope QPEKGGRKPA$_{2568-2577}$ epitope ($\Delta\langle E\rangle = 7.06$) would compromise viral fitness more effectively than the dominantly presented epitope of the protective allele with the strongest correlation with spontaneous clearance (*HLA-B*57* KSKKTPMGF$_{2629-2637}$, $\Delta\langle E\rangle = 6.45$). These findings suggest that vaccine immunogens may be delivered to persons carrying non-protective HLA alleles to elicit potent immune responses against vulnerable viral epitopes identified by our model that are naturally only subdominant. The systematic design of such immunogens to exploit viral vulnerabilities identified by our model is the subject of the next section of the chapter.

3.3.6 *In silico* design of NS5B CTL immunogens

Having validated our model in comparisons against with five types of experimental and clinical data, we now proceed to employ it in the rational design of CTL vaccine immunogens. CTL vaccines that deliver complete HCV proteins [48–50] may elicit dominant T-cell responses against poor viral targets that drown out potent responses against vulnerable targets [12, 51, 52, 126]. By identifying vulnerable CTL targets that can be targeted dominantly or subdominantly by particular HLA alleles (section 3.3.5), our fitness landscapes can inform the design of epitope-based vaccine immunogens to prime potent CTL responses in a particular individual or population.

Although a single immunological correlate of protection against HCV remains to be defined [34], potent and broad CTL responses against protein NS5 are correlated with spontaneous clearance of infection in persons possessing natural HCV immunity [7, 8, 34, 48]. We illustrate the value of our HCV subtype 1a NS5B fitness landscape in the design of candidate epitope-based vaccine immunogens to elicit CTL immune responses against vulnerable NS5B targets. Subtype 1a is the most prevalent HCV subtype in the United States,

leading us to adopt as our target population the top 66 haplotypes of North Americans, accounting for 40.9% of the North American population (http://www.ncbi.nlm.nih.gov/projects/gv/mhc) [184]. By recovering fitness landscapes for other HCV proteins, and for other subtypes, an analogous approach could be used to design immunogens for other HCV proteins, subtypes, and populations.

We performed computational immunogen design according to the following protocol. We first cross-referenced the list of 66 haplotypes with the CTL NS5B epitopes listed in the Immune Epitope Database (http://www.iedb.org) [179] to identify 24 distinct epitopes restricted by the class I HLA-A and HLA-B molecules present within this population. We then constructed all 16,777,215 possible candidate immunogens containing all combinations of $1, 2, 3, \ldots, 24$ epitopes, and scored each candidate along three criteria:

Criterion 1 - Population averaged fitness impact $\overline{\Delta\langle E\rangle}$. The precise immune responses activated by the immunogen candidate in a particular recipient host depends on which components of the epitope-based immunogen can be restricted by the HLA alleles of the host. Assuming that all vaccine-induced CTL responses are activated, and that each response completely eliminates all viral strains that are fully wild type within their cognate epitope, we use the procedure described in section 3.3.5 to compute the fitness penalty, $\Delta\langle E\rangle_i^j$, exacted upon the viral population by immunogen $i = 1 \ldots 16,777,215$ in haplotype $j = 1 \ldots 66$. We then define the population averaged fitness impact within the North American target population as, $\overline{\Delta\langle E\rangle_i} = \sum_{j=1}^{66} w_j \Delta\langle E\rangle_i^j$, where w_j is the fraction of the population possessing haplotype j.

Criterion 2 - Population coverage. Immunogen candidates possessing large values of $\overline{\Delta\langle E\rangle}$ are predicted to impose large fitness penalties upon the viral quasispecies, but may only prime effective immune responses in a small proportion of the target population. In the extreme case, some fraction of the population may possess haplotypes that prevent them from developing immune responses against any of the epitopes delivered in the immunogen. To balance these considerations, we adopt as our second criterion the fraction of the target

population who respond to at least one epitope in the immunogen candidate. This criterion can be straightforwardly modified to measure the fraction of the population who respond to more than one epitope. Of the 66 haplotypes in the target population, 18 do not restrict any of the 24 NS5B epitopes, meaning that we can cover at most 71.2% of the target population. To attain 100% coverage, we would need to design a vaccine immunogen containing CTL epitopes from other HCV proteins that can be restricted by persons possessing haplotypes that cannot target NS5B. As more sequence data becomes available, we propose to extend our approach to other HCV proteins to design multi-protein epitope-based CTL vaccines with complete population coverage.

Criterion 3 - Immunogen size. Larger immunogens containing more epitopes can provide broader population coverage and larger viral fitness penalties by virtue of their size. It may be beneficial, however, to limit immunogen size due to the increased cost and complexity of the resultant vaccine [185]. Accordingly, we selected as our third criterion the number of epitopes in the candidate immunogen.

We locate each of the 16,777,215 candidate immunogens in the three-dimensional space spanned by the three scoring criterion (figure 3.6). We have identified within this ensemble the 86 candidates lying on the Pareto frontier [186]. Candidates residing on this frontier are optimal in the sense that improvements in any one criterion (increased $\overline{\Delta\langle E\rangle}$, broader population coverage, smaller immunogen) are necessarily accompanied by a deterioration in another (reduced $\overline{\Delta\langle E\rangle}$, narrower population coverage, larger immunogen). Immunogen candidates away from the frontier are non-optimal in that any one criterion can be improved without compromising another. We list the compositions and scores of the 86 optimal immunogens in table C.4.

By massively reducing the search space of all possible 16,777,215 candidate immunogens by over five orders of magnitude to 86 optimal formulations, our computational platform offers an inexpensive pre-screening process to identify promising CTL immunogens for experimental testing. The relative weightings of the three design criteria may be used to inform

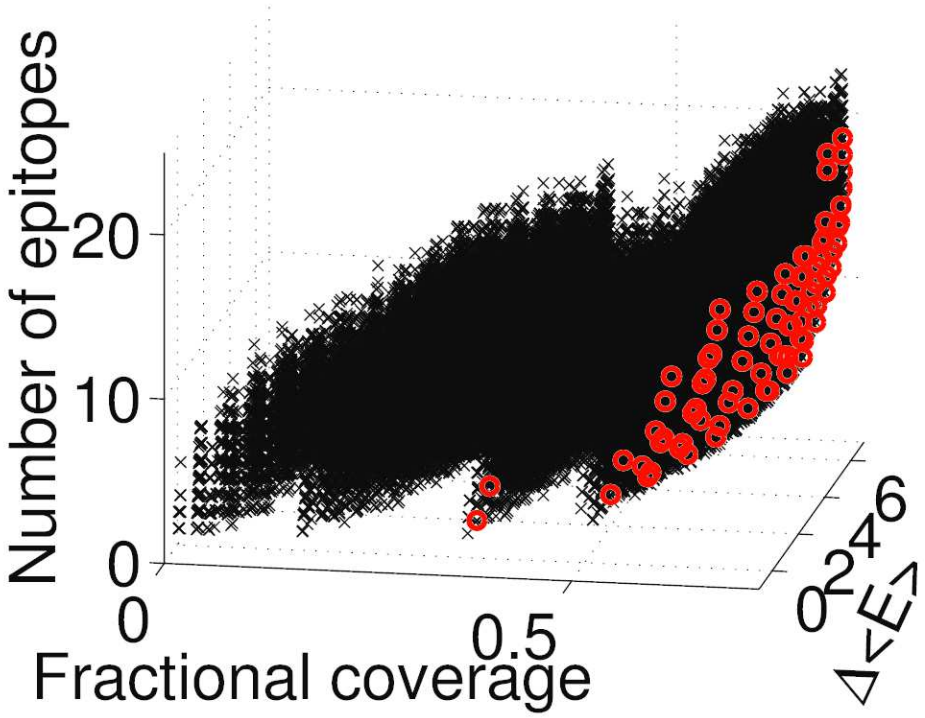


Figure 3.6: Scatter plot of all 16,777,215 NS5B CTL vaccine immunogen candidates (black crosses) in the three-dimensional design space of: (1) the weighted average fitness impact in the target population, $\overline{\Delta\langle E\rangle}$, (2) fraction of the target population that respond to at least one epitope in the immunogen (fractional coverage), and (3) the number of epitopes in the vaccine. The target population consisted of the 66 most prevalent haplotypes in North Americans, accounting for 40.9% of the North American population. Our procedure identifies 86 immunogen candidates residing on the Pareto frontier (red circles), which are optimal formulations in the sense that improvements in any one of the three design criteria are necessarily accompanied by a deterioration in another. The optimal candidates are listed in table C.4.

which region of the optimal frontier is most desirable for further investigation. For example, the smallest optimal candidate with the highest possible coverage of the target population, 71.2%, is 10 epitopes (97 residues) in size and possesses a population averaged fitness impact of $\overline{\Delta\langle E\rangle} = 6.6$. Alternatively, if immunogen size is not considered a restriction, we may select from the optimal frontier a 19-epitope (179-residue) candidate with 71.2% coverage of the target population and $\overline{\Delta\langle E\rangle} = 9.3$. We observe that it is straightforward to modify the scoring criteria, and incorporate additional criteria, without changing the approach.

3.4 Conclusion

Despite 20 years of research, no HCV vaccine is yet available. The development of an effective vaccine has been hampered by the high mutability and rapid replication rate of HCV, producing large diversity in the viral quasispecies and facilitating escape from natural and vaccine-induced immune responses. Effective prophylactic and therapeutic vaccines should prime immune responses against vulnerable regions of the viral proteome in which escape mutations carry a large cost in replicative fitness, without eliciting ineffective responses against poor targets from which viral escape is easy and carries a small fitness cost.

We have demonstrated an approach to infer the empirical fitness landscape for the NS5B protein in HCV subtype 1a from viral sequence databases, and validated the predictions of our model against independent experimental and clinical data. We have used our landscape to computationally design epitope-based NS5B CTL vaccine immunogens for the North American population. The immunogens designed by our approach are predicted to redirect host CTL immune responses to preferentially target vulnerable regions in the virus within which mutations cripple viral fitness. This inexpensive *in silico* design platform can serve as a valuable tool to identify promising immunogen candidates for experimental testing, helping to alleviate the burden of trial-and-error experimentation and accelerating the search for an efficacious HCV vaccine.

We are currently extending our approach to other HCV proteins and polyproteins to capture mutational couplings between viral proteins to enable the design of multi-protein immunogens, and are partnering with experimental collaborators to test our predicted immunogen candidates. We anticipate that with increasing computational power and reducing sequencing costs, it will soon become feasible within the coming years to apply our technology to the complete HCV proteome and perform rational *in silico* design of a complete anti-HCV immunogen.

Chapter 4

Using Fitness Landscapes to Show Treatment Possibilities by Inducing Error Catastrophe

4.1 Introduction

HIV has killed more then 30 million persons worldwide, with another 30 million infected [187]. Antiretroviral drugs have rendered HIV infection a manageable condition [187], but their high cost makes them effectively inaccessible in the developing world [188] and rates of drug resistant mutations in persons on therapy more than 36 months exceed 20% [189]. A vaccine remains unavailable [190, 191]. The high mutation rate and sequence diversity of HIV present significant challenges for therapy, but recent computational advances offer new ways to identify susceptible targets to guide the design of new drugs and vaccines [12, 51, 106, 111].

We recently presented an approach to translate clinical databases of HIV sequences into models of the viral "fitness landscape" that quantify viral replicative capacity as a function of its amino acid sequence [1, 12, 111]. Most viral mutations are deleterious, compromising fitness, but certain patterns of mutations enable the virus to escape immune surveillance while maintaining high fitness [51]. Empirical fitness models can explicitly resolve these high fitness pathways, and reveal vulnerabilities where drug therapy or vaccine-induced immune pressure can cripple viral fitness [12, 106, 111]. Regarding sequences in the database as observations from an underlying probability distribution, the least structured (i.e., maximum entropy) model capable of reproducing the two lowest order moments of the amino acid

Most of this chapter is an excerpt from ref. [96]: G.R. Hart and A.L. Ferguson "Error catastrophe and phase transition in the empirical fitness landscape of HIV" Phys. Rev. E 91 032705 (2015)

frequencies in the clinical database (i.e., frequency with which each single amino acid is observed in each position, and pairs of amino acids in pairs of positions) is the infinite range Potts model [111, 139, 148, 192],

$$P(\vec{z},T) = \frac{e^{-\beta E(\vec{z})}}{Z(T)},\ E(\vec{z}) = \sum_{i=1}^{m} h_i(z_i) + \sum_{i=1}^{m}\sum_{j>i}^{m} J_{ij}(z_i,z_j), \tag{4.1}$$

where $E(\vec{z})$ is the infinite range Potts Hamiltonian, m is the number of positions in the protein, $\vec{z}$ is a m-element vector encoding the protein sequence, and each element of $\vec{z}$ is an integer in the range 1-21 corresponding to the 20 natural amino acids plus the gap or blank [144]. As in statistical thermodynamics we term E the dimensionless "energy", $Z(T) = \sum_{\vec{z}} e^{-\beta E(\vec{z})}$ the partition function, and $\beta = 1/T$ the dimensionless inverse temperature, that we set to unity for the purposes of parameter inference [12, 111, 139, 192]. (In contrast to the standard expressions, our energy and temperature are dimensionless and we have eliminated Boltzmann's constant.) Determination of the model parameters – the one-body external fields $\{h_i\}$ and two-body pairwise couplings $\{J_{ij}\}$ – that reproduce the observed one and two-body amino acid frequencies constitutes solution of a canonical inverse problem that may be tackled in many ways [111, 139].

Under the ansatz that highly prevalent viral strains should also be highly fit, the prevalence of a strain in the population of infected hosts, $P(\vec{z})$, should be a good proxy for its replicative fitness, $f(\vec{z})$, and thus the energy assigned by our model, $E(\vec{z})$, should be negatively correlated with the logarithmic fitness, $\log(f(\vec{z}))$ [12, 106, 111]. This relationship can be derived exactly under restrictive assumptions [109], and we have shown that under mild conditions on the sequences in the clinical databases, this relationship holds generically [106]. As described in chapter 3, we have inferred and validated much models of viral fitness against clinical and experimental data.

Since the model parameters are fitted at a temperature of unity, T_{op}=1 is the effective "biological operating temperature" [139]. Modulating T away from unity corresponds to

a uniform scaling of the $\{h_i\}$ and $\{J_{ij}\}$ parameters [139]. As $T \to 0$, the fitness landscape predicts the lowest energy (highest fitness) strain to dominate the viral population, whereas for $T \to \infty$, all strains become equally represented independent of energy (fitness).

In this chapter we show that the fitness landscape for the p6 protein in HIV-1B predicts that the viral population is poised close to a phase transition at T=1.20 between a high-fitness, low-diversity (i.e., low-energy, low-entropy) and a low-fitness, high-diversity (i.e., high-energy, high-entropy) population, and that the transition may be induced – and viral fitness crippled – by elevating the mutation rate or forcing mutations at particular amino acid positions.

4.2 Fitness landscape

4.2.1 Model inference

An ensemble of 1824 DNA sequences of the HIV-1 p6 protein were downloaded from the LANL HIV database [193]. Sequences were restricted to subtype B – the most prevalent form in Europe and the Americas – treatment-naïve hosts, and not classified as "problematic". Sequences were aligned to the HXB2 reference sequence [194], and translated to the cognate 53 amino acid protein sequence. Ambiguous codons were translated as a blank. We fitted the $\{h_i\}$ and $\{J_{ij}\}$ parameters in eqn. 2.4 using the approach in appendix A.

4.2.2 Model validation

Protein p6 is less well experimentally and clinically studied than the other HIV Gag proteins [12, 111], but analysis by Brumme *et al.* of the IHAC cohort of chronically infected hosts identified two statistically significant p6 escape mutations directly associated with T-cell immune pressure: B40-E34D (mutation from E (Glu) to D (Asp) at p6 position 34 in an epitope presented by the human leukocyte antigen (HLA) B40) and A68-R42K [195].

Particular contiguous groups of amino acids known as *epitopes* are recognized by T-cells when all positions contain wild type (most probable) amino acids [196]. The virus escapes immune recognition by mutating at one or more of the positions in the epitope. Cross referencing the two escape mutations with the LANL T-cell epitope maps [193], we identified the former as associated with the B40-K_{33}ELYPLTSL_{41} cytotoxic T-cell (CTL) epitope, and the latter with the A68-K_{33}ELYPLTSLRS_{43} and A68-E_{29}PIDKELYPLTSLRS_{43} helper T-cell (Th) epitopes. If we assume that all polymorphisms within the targeted epitope are equally efficient at mediating escape, then the virus should make mutations incuring the smallest energy costs (lowest fitness penalty). Our model predicts that of all the p6 polymorphisms in the clinical sequence database (i) the E34D and R42K escapes are the single lowest energy polymorphisms at these two positions, and (ii) the E34D escape is the 2/69 lowest energy polymorphism within the B40-K_{33}ELYPLTSL_{41} epitope, and the R42K escape as the 1/73 and 1/118 lowest energy polymorphism within the A68-K_{33}ELYPLTSLRS_{43} and A68-E_{29}PIDKELYPLTSLRS_{43} epitopes, respectively. That our model predicts the clinically observed escape mutations to incur the smallest energy cost (fitness penalty) provides support that it reflects intrinsic viral fitness. We observe that analogous models for HIV Gag p17 and p24 have been validated against the more comprehensive clinical and experimental data available for these proteins [12, 111].

4.3 Error catastrophe

4.3.1 Density of states

Computing the partition function in eqn. 2.4 "solves" the Potts model in the sense that nearly all thermodynamic quantities are calculable from $Z(T)$ [197]. We may express the partition function as a sum over energy levels, $Z(T) = \sum_{\vec{z}} e^{-\beta E(\vec{z})} = \sum_E g(E) e^{-\beta E}$, where $g(E)$ is the density of states. We developed a parallel implementation of the Wang-Landau algorithm [198, 199] described in refs. [200, 201] to estimate the density of states, $\hat{g}(E) \approx C \times g(E)$,

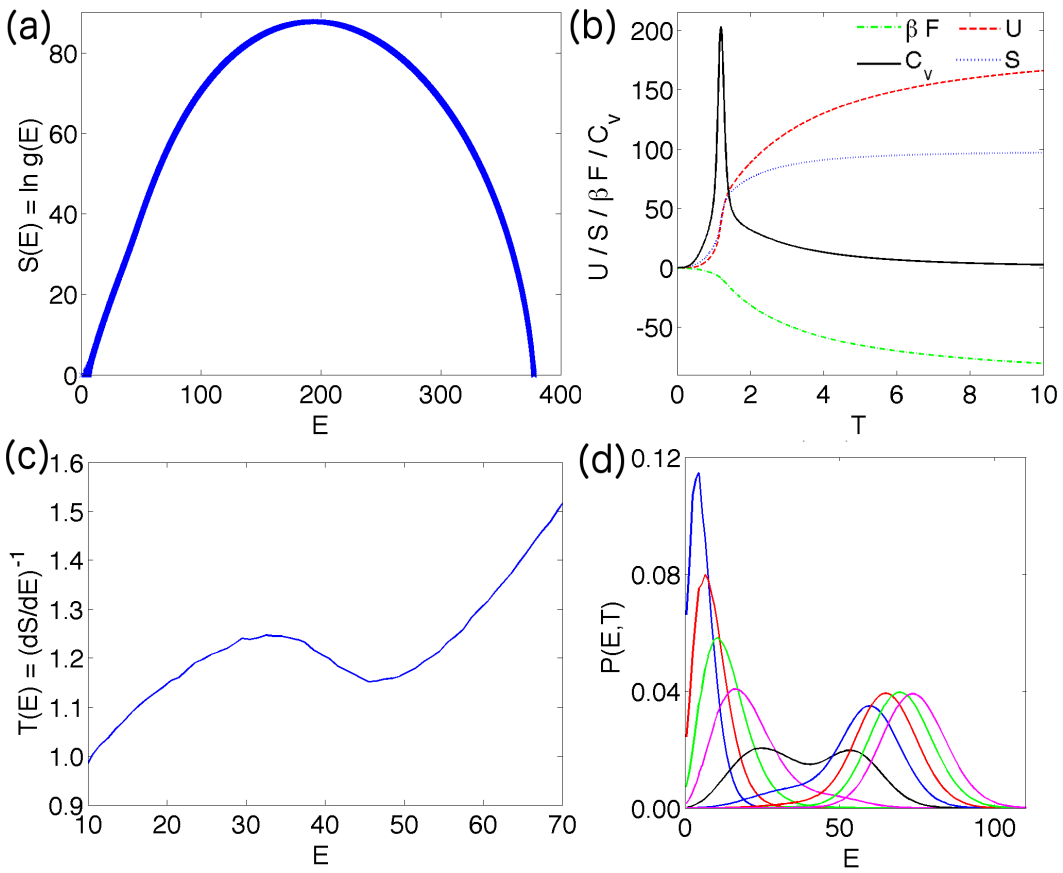


Figure 4.1: Thermodynamics of HIV-1B protein p6. (a) Density of states estimated by Wang-Landau sampling. (b) Dimensionless energy, U, entropy, S, free energy, F, and heat capacity, C_v, as a function of the dimensionless temperature, $T{=}\beta^{-1}$. (c) The microcanonical caloric curve, $T(E)$, which has a negative gradient over the region $E{=}35$–45. (d) The canonical distribution, $P(E,T)$, at (left to right) $T{=}0.8{:}0.1{:}1.6$ exhibits a bimodal distribution at $T_{coex}{=}1.20$.

up to a multiplicative constant, C, that we fix by asserting the uniqueness of the lowest energy (highest fitness) wild type strain (figure 4.1(a)). From $Z(T)$ we computed the energy, $U(T){=}-\partial \ln Z(T)/\partial\beta$, free energy, $\beta F(T){=}-\ln Z(T)$, entropy, $S(T){=}\beta U(T){-}\beta F(T)$, heat capacity, $C_v(T){=}\partial U/\partial T$, microcanonical temperature prescribed by the caloric curve, $T(E){=}[\partial S(E)/\partial E]^{-1}{=}[\partial(\ln g(E))/\partial E]^{-1}$, and canonical distribution, $P(E,T){=}g(E)e^{-\beta E}/Z(T)$ (figure 4.1(b-d)).

Phase transitions are formally defined in infinite systems as non-analyticities in the equation of state, but there have been many studies of finite system analogs of bulk transitions

[92]. We observe a jump in $U(T)$ and $S(T)$, sharp peak in $C_v(T)$, bimodality in $P(E,T)$, and a negative gradient in the microcanonical caloric curve – indicative of positive curvature in S(E) and a negative microcanonical heat capacity, $C_v(E)$=$-(\partial S/\partial E)^2/(\partial^2 S/\partial E^2)$ – at the coexistence temperature, T_{coex}=1.20, defined by equal areas below the peaks of $P(E,T)$ [198] (figure 4.1). These are all (equivalent) indicators of a finite system phase transition in the sequence space of p6 at T_{coex}=1.20 [92, 202].

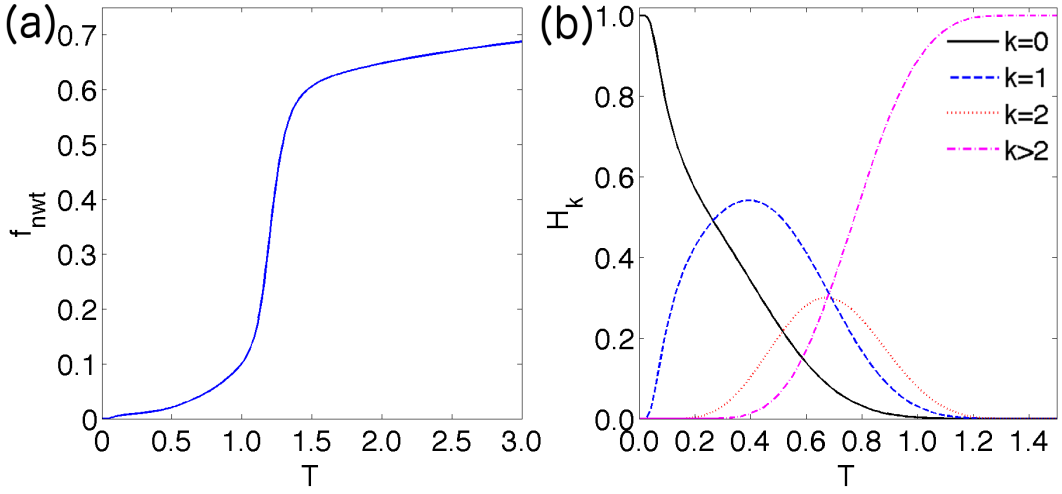


Figure 4.2: Prevalence of mutant strains. (a) The average fraction of non-wild type residues per strain, f_{net}, exhibits a sharp jump at T_{coex}=1.20. (b) The fraction of strains in the population that are a Hamming distance of k=0,1,2 from the wild type strain as a function of temperature.

4.3.2 Prevalence of mutant strains

In performing Wang-Landau sampling, we also collected estimates of the average fraction of the 53 amino acids that were non-wild type for all sequences of a particular energy, $f_{nwt}(E)$, and the fraction of sequences containing precisely k=$\{0,1,2\}$ non-wild type amino acids, $H_k(E)$. Combining these distributions with our density of states, we computed these quantities as a function of temperature using the identity, $X(T) = \sum_E X(E)g(E)e^{-\beta E}/Z(T)$ [203] (figure 4.2). Consistent with the sharp increase in the energy and entropy of the viral population (i.e., decrease in fitness and increase in diversity) at T_{coex}=1.20, we observe a

concomitant jump in the fraction of mutant residues per strain, $f_{nwt}(T)$. At T_{op}=1, the fraction of strains in the viral population that are wild type, contain a single mutation, and contain exactly two mutations are $H_0(T)$=0.34%, $H_1(T)$=3.0%, and $H_2(T)$=7.8%. At T_{coex}=1.20 these fractions are 0.0079%, 0.11%, and 0.45%, with 99.4% of strains containing three or more point mutations.

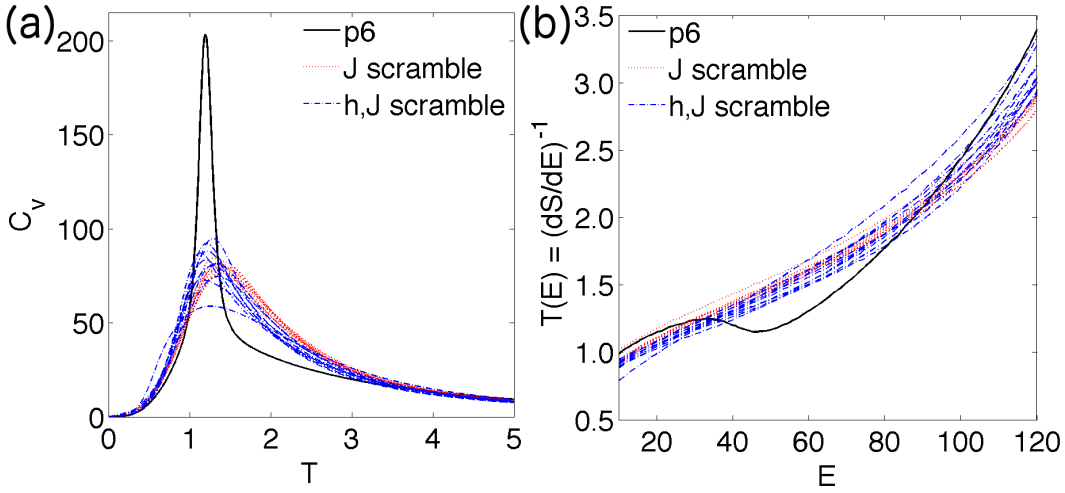


Figure 4.3: Artificial Hamiltonians generated by ten random shuffles of the $\{J_{ij}\}$ parameters (dotted lines), and ten shuffles of the $\{h_i\}$ and $\{J_{ij}\}$ parameters (dot-dashed lines), do not show signatures of the phase transition exhibited by p6. (a) The sharp peak in C_v disappears, and (b) the caloric curve does not possess any region of negative gradient.

4.3.3 Permutation test

To determine whether the observed phase transition is a generic property of our Potts Hamiltonian, we computed the density of states for 10 artificial Hamiltonians constructed by random shuffling of the $\{J_{ij}\}$ parameters leaving the $\{h_i\}$ parameters intact, and 10 in which both the $\{h_i\}$ and $\{J_{ij}\}$ parameters were independently shuffled. In all cases the signatures of the phase transition – jump in $U(T)$ and $S(T)$, bimodality in $P(E,T)$, back bending in $T(E)$, and sharp peak in $C_v(T)$ – vanished (figure 4.3), indicating that the phase transition is contingent on the precise structure of external fields and mutational couplings within the p6 protein.

4.3.4 Interpretation of T

The high replication and mutation rates of HIV cause it to exist as a *quasispecies*, or cloud of closely related mutant strains, whose nonequilibrium evolution is described by Eigen's equation [62, 79, 80, 94]. Leuthäuser mapped Eigen's equation to the equilibrium properties of an 2D lattice model [84, 204]. We have shown that the Ising models inferred using our approach (i.e., eqn. 2.4 with binary $\vec{z}$ elements denoting wild type and mutant amino acids) well approximate the equilibrium distribution of strains in Leuthäuser's formulation [106]. Under this correspondence, $\beta{=}T^{-1}{=}\ln\sqrt{q/(1-q)}$, where q is the per position probability of correctly copying an amino acid in a viral replication event. In the limit of perfect replication fidelity ($q{\to}1$) $T{\to}0$, whereas for random copying ($q{\to}0.5$) $T{\to}\infty$. Extending this correspondence to the Potts model, we may interpret T as an increasing function of the viral mutation rate, and the proximity of the phase transition at $T_{coex}{=}1.20$ to $T_{op}{=}1$ is precisely analogous to the observation that the HIV error rate lies very close to the *error catastrophe* beyond which there is lethal error accumulation, and the quasispecies collapses [62, 89–91, 94, 99]. The existence of the viral error catastrophe is well established in theoretical models of viral mutational dynamics [62, 82, 88], and a recent study of HIV employing theoretical fitness models motivated by experimental data detected an error catastrophe in both quasispecies models and stochastic population genetics-based simulations [82]. To the best of our knowledge, the present work represents the first time that the error catastrophe has been detected in an empirical model of viral fitness. Experimental reports supporting an HIV error catastrophe *in vivo* have motivated clinical trials of an HIV mutagen [100, 101], but confounding factors, such as cellular resource scarcity at high mutation rates and the existence of a competing "extinction threshold", have so far precluded the unambiguous confirmation of this transition [82, 205, 206].

The proximity of the phase transition is thought to convey survival advantage by providing a phenotype reservoir [90], maximizing adaptability [99], and optimizing immune escape

[62]. Sella and Hirsh have proposed that a viral quasispecies should minimize a "free fitness", $G{=}U - T_{op}S$, as an optimal balance between fitness and diversity [109]. We find that $G(T){=}U(T){-}(T_{op}{=}1)S(T)$ exhibits a minimum at $T{=}1.00$, suggesting that the quasispecies is structured in accordance with this principle. Immunologically, we may exploit the proximity of the transition to induce a large increase in the average energy (decrease in fitness) of the viral population, $U(T)$, by a small increase in T. In practice, T_{op} may be pushed beyond $T_{coex}{=}1.20$ by drug therapies that elevate the viral mutation rate [102, 103]. Promising results have emerged in recent years [100, 101], but the clinical translation of mutagens designed to exploit the purported error catastrophe warrants careful further study [38].

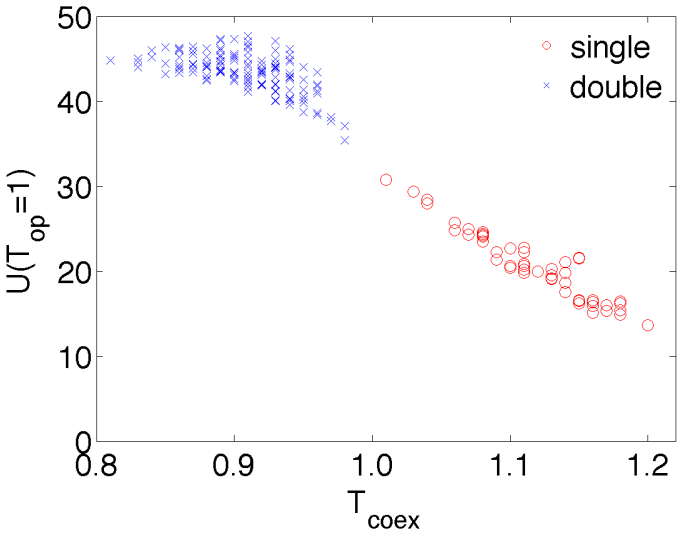


Figure 4.4: Impact on the average energy of the viral population at the operating temperature, $U(T_{op}{=}1)$, by forcing (a) single mutations away from the wild type residue at each of the 53 amino acids (circles), and (b) double mutations at pairs of the 17 positions at which single mutations caused a T_{coex} to fall below $T{=}1.1$ (crosses).

4.4 Inducing the error catastrophe

4.4.1 One and two-point mutations

We hypothesized that it may also be possible to induce the phase transition by forcing the virus to make mutations at particular positions in the protein. To test this conjecture, we recomputed the density of states under the constraint that the virus was forbidden from mutating to strains in which the amino acid at a single position i was wild type for each of the i=1...53 positions. Rejection of trial moves to these forbidden states in the Wang-Landau algorithm were treated according to ref. [207]. As illustrated in figure 4.4, forcing single point mutations depresses T_{coex} with an attendant increase in the average energy (decrease in fitness) of the viral population. Remarkably, forcing a point mutation in the amino acid at position i=11 depresses the coexistence temperature to T_{coex}=1.01, and more than doubles the population energy at the operating temperature from $U(T_{op}=1)$=13.5 (cf. figure 4.1(b)) to 30.7.

Application of mutational pressure to any single position does not, however, depress T_{coex} below T_{op}=1. Accordingly, we identified the 17 positions in the protein at which mutational pressure pushed $T_{coex} < 1.1$, and recomputed the density of states for the (17×16)/2=136 systems constrained such that pairs of amino acids at positions i and j within this set were both forbidden to be wild type. As illustrated in figure 4.4, all 136 instances of two-point mutational pressure induced the phase transition (i.e., $T_{coex} < 1$), with the coexistence temperature maximally depressed to T_{coex}=0.81 by positions (i=11, j=37), and the average energy maximally elevated to $U(T_{op}=1)$=47.7 by positions (i=44, j=53).

4.4.2 CTL immune pressure

As an intracellular viral protein, the T-cells of the adaptive immune system are primarily responsible for recognizing p6 as a pathogenic protein by binding to epitopes comprising a small number of contiguous amino acids. The LANL T-cell epitope maps list five

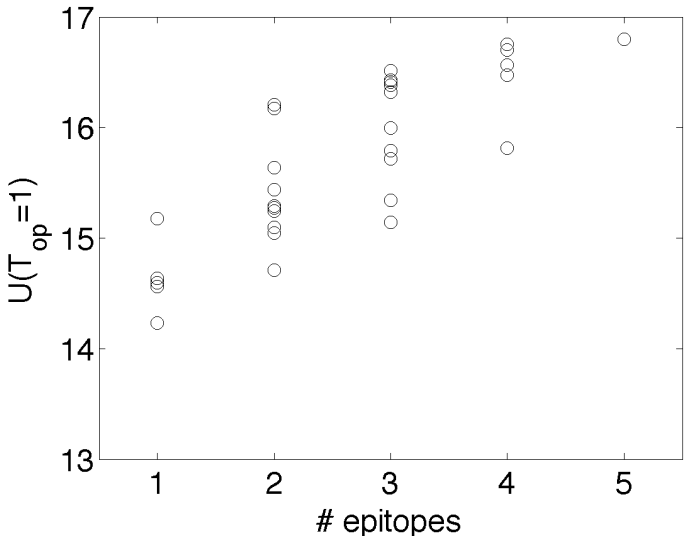


Figure 4.5: Impact on the average energy of the viral population at the operating temperature, $U(T_{op}=1)$, upon applying immune pressure to all possible {1,2,3,4,5}-epitope combinations of the five known CTL epitopes in p6.

known CTL epitopes within p6: T23PSQKQEPI31, S25QKQEQIDK33, E29PKDREPL38, K33ELYPLTSL41, and Y36PLASLRSLF45 [193]. Eight Th epitopes have been mapped to within a region of 18 amino acids or less, but only one was confirmed as an exact epitope. Motivated by interest in the design of CTL-based HIV vaccine immunogens [12, 191, 208], we assessed whether the application of CTL immune pressure can also induce the phase transition. To simulate CTL targeting at each single epitope, we recomputed the density of states under the constraint that the virus was forbidden to contain all wild type amino acids within the epitope [207]. We performed analogous calculations for all possible doubles, triples, quads, and quints of epitopes.

As illustrated in figure 4.5, targeting of any single epitope only raises the average energy of the viral population from $U(T_{op}=1)$=13.5 to 14.2–15.2. Simultaneous targeting of all five epitopes corresponds to $U(T_{op}=1)$=16.8. All possible combinations of epitope targeting result in T_{coex}=1.17-1.19, indicating that CTL immune pressure at any known epitopes is unable to induce the phase transition, and can only affect relatively modest increases in population energy. Indeed, we find that it is not possible to induce the phase transition by

targeting *any* contiguous group of nine amino acids within p6.

4.5 Conclusions

In summary, we have translated sequence databases of the HIV-1 clade B p6 protein into an empirical fitness landscape quantifying the replicative capacity of the virus as a function of its amino acid sequence. Using Wang-Landau sampling, we have identified a phase transition in the sequence space of the p6 viral protein corresponding to an error catastrophe. Our model predicts that the transition can be induced by elevating the operating temperature beyond the coexistence temperature, $T_{op} > T_{coex}$, using drug therapies that increase the viral mutation rate [102, 103]. Alternatively, the coexistence temperature may be suppressed below the operating temperature, $T_{coex} < T_{op}$, by forcing particular pairs of mutations, suggesting a means to cripple the HIV virus by applying mutational pressure at carefully selected positions. In principle, this might be achieved by drugs or small molecule inhibitors with localized binding sites on the p6 protein, similar to the anti-HIV drugs that bind protease, integrase, and reverse transcriptase [209, 210]. In practice, development of such molecules is a challenging problem in drug design. Interestingly, CTL immune pressure at any combination of known epitopes in p6 cannot induce the transition, providing an empirical rationalization for why the virus can exist in close proximity to the error catastrophe without sustaining catastrophic fitness costs due to adaptive immune pressure.

Chapter 5

Using Fitness Landscapes in Dynamic Design

5.1 Introduction

A fitness landscape gives an organism's replicative fitness as a function its amino acid sequence [1]. With a fitness landscape we can calculated the fitness for a large number of sequences and determine where maximum and minimum are, allowing us to to rank the effectiveness of different vaccine candidates [81]; however everything is static and the viral population does not respond to the immune pressure. While these predictions could be very good options for a vaccine, it is possible that the virus has escape pathways from a particular immune response. It is these escapes that have made it difficult to develop vaccines and drugs for HCV. In order to address this problem we need to also explicitly model the dynamics of the host-pathogen interaction under the action of the vaccine.

Mathematical models of evolution have been studied for nearly a century with both simple and complex landscapes [211]. Despite this long history these models continue to be useful. For example Tripathi et al. used evolutionary dynamics on a simple fitness landscape to explore the viability of treating HIV by triggering the error catastrophic with mutagenic drugs, in addition to elucidating an additional evolutionary benefit of recombination [82]. Barton et al. were the first to simulate viral dynamics on a detailed empirical fitness landscape. They showed that with simple viral dynamics and a constant bias to the landscape (representing immune pressure) they could predict what escape mutation arose in a patient,

Most of this chapter is an excerpt from a paper currently under preparation: G. R. Hart and A. L. Ferguson "Evolutionary Dynamics Over an Empirical Fitness Landscape for Improved Vaccine Design"

the order the mutations arose in, and the relative time between them [212]. In the approach presented in this chapter, we simulate viral evolution on the fitness landscape which is biased in response to immune pressure, as they do, but we also allow for the immune system to change in response to the evolving viral population. We begin by testing our model by simulating infection in different individuals and comparing it to the actual infections in those individuals. Using this to establish the validity of our model we go on to show how this model can be used to test different immunogen candidates for a vaccine for a specific individual. Finally we compare this dynamic design process to the static one we presented in chapter 3. Full details of the determination and validation of this landscape from clinical sequence databases is presented therein.

5.2 Method

5.2.1 Fitness landscape

We employed the fitness landscape for the hepatitis C virus (HCV) RNA-dependent RNA polymerase (protein NS5B) determined from clinical sequence data using the approach described in chapter 3.

5.2.2 Viral dynamics

The viral side of the dynamics are modeled using Fisher-Wright type dynamics [213, 214]. The algorithm proceeds as follows. Each of the N infected cell produces p free virus particles. The replication process is not prefect, but is subject to a mutation rate μ. For each of the pN free virus particles we specify the intrinsic fitness, f. We also calculate a penalty to fitness, S, based on the sequences susceptibility to immune pressure (see section 5.2.3). With both the intrinsic fitness and the immune susceptibility we can calculate the effective fitness, $F = f/S$. Using the effective fitness as a weight we randomly select N free virus

particles to infect the next generation of cells. Once these N sequences are selected the T-cell dynamics are then updated to reflect their response to the changing viral population (see section 5.2.3), and then the iterative uptade process is repeated.

We can find measurements of the mutation rate, $\mu \approx 1.2 \times 10^{-4}$, from the literature [31]. In additional we tested our model over a range of mutations ($10^{-5} \leq \mu \leq 10^{1}$) and showed it to be robust around the measured value ($10^{-5} \leq \mu \leq 10^{-3}$). The number of progeny, $p \approx 150$ virons/cell/day, can also be gleaned from experiments [31]. We showed that above a certain threshold ($p \approx 5$) our dynamics do not change, allowing us to accelerate our simulation without comprising accuracy by using a smaller value of $p = 10$. The effective population size, N, is not readily found from experiment. We tested this value over several order of magnitude ($10^3 \leq N \leq 10^5$) finding a robust range to run in, $N = 5 \times 10^4$.

5.2.3 Immune dynamics

To model the dynamics of the T-cells we use the ODEs listed from ref. [108], except rather than having an equation for the number of infected cells this number remains fixed (see section 5.2.2), but the susceptibility of the infected cells to the T-cells changes. The equations governing the T-cell dynamics are

$$\frac{dN_i}{dt} = -aN_i \sum_{\{k\}} \chi_{ik} I_k + bN_i - eN_i + w \tag{5.1}$$

$$\frac{dE_i^1}{dt} = (aN_i + a'M_i) \sum_{\{k\}} \chi_{ik} I_k - rE_i^1 - dE_i^1 - gE_i^1 \tag{5.2}$$

$$\frac{dE_i^n}{dt} = -rE_i^n + 2rE_i^{n-1} - dE_i^n - gE_i^n \qquad n = 2 \ldots N_D - 1 \tag{5.3}$$

$$\frac{dE_i^{N_D}}{dt} = 2rE_i^{N_D-1} - d'E_i^{N_D} - gE_i^{N_D} \tag{5.4}$$

$$\frac{dM_i}{dt} = a'M_i \sum_{\{k\}} \chi_{ik} I_k - hM_i + g \sum_{n=1}^{N_D} E_i^n. \tag{5.5}$$

Here N_i is the number of naive cells targeting epitope i, E_i^n is the number of effector cells of generation n, and M_i is the memory cells. The number cells infected with strain k is I_k. Using tools from the Immune Epitope Database ([215]) we can calculate the susceptibility of strain k to immune cells targeting epitope i, χ_{ik} that provides a numerical estimate of the affinity between each epitope and T-cell. As mentioned above the total number of infected cells is fixed, but as mutations arise creating different strains, the strength of the immune pressure can decrease as the effective number of infected cells, $\sum_{\{k\}} \chi_{ik} I_k$, falls.

These dynamics feed back into the viral dynamics through a penalty to the fitness representing T-cells recognition that modulates the intrinsic fitness of a particular viral strain. For viral strain k, the fitness penalty incurred due to recognition, summed over all N_D generations of effector cells for all T-cell types is $S_k = exp(2\bar{h} * \sum_i \chi_{ik} \sum_{n=1}^{N_D} E_i^n)$, where $\bar{h}$ is the average cost for a single mutation in the protein.

The values for all parameters in the model were extracted from experimental studies conducted in refs. [216–222] and are listed in table 5.1.

Symbol	Parameter	Value	Units
a	Naive cell activation	10^{-7}	$cell^{-1}day^{-1}$
a'	Memory cell activation	10^{-6}	$cell^{-1}day^{-1}$
b	Naive cell replication rate	10^{-4}	day^{-1}
d	Effector cell death rate	0.2	day^{-1}
d'	Terminal effector cell death rate	3	day^{-1}
r	Effector cell replication rate	6	day^{-1}
e	Naive cell death rate	3×10^{-4}	day^{-1}
g	Effector cell to memory cell rate	0.03	day^{-1}
h	Memory cell death rate	0.03	day^{-1}
w	Naive cell birth rate	0.1	$cells\ day^{-1}$
N_D	Effector cell division limit	9	

Table 5.1: Parameters implemented in the model of T-cell dynamics.

Due to small copy numbers of T-cells, stochastic effects are expected to be important in governing the system dynamics. To explicitly expose the role of stochasticity, we integrate the coupled T-cell ODEs both deterministically using a 4th order Runge-Kutta algorithm, and also stochastically using the Gillespie algorithm [223]. For every simulation we run it

once using the deterministic integration and 99 times with the stochastic integration and calculate the mean and standard deviation.

5.2.4 Likelihood calculations

To test the predictions of our model, we compare its predictions for the infection time course against longitudinal sequencing data, tracking the viral load within each host as a function of time. To make this comparison quantitative, we compute the likelihood of observing the specific mutational time courses observed in the patient given the parameters of our model. We begin by writing the likelihood of the whole time course as the product of the likelihoods of seeing the mutations at each time point:

$$P(\{n_i^t\}|\{p_i^t\}) = \prod_t P_t(\{n_i\}^t|\{p_i\}^t), \tag{5.6}$$

where n_i^t represents the number of sequences at time t with mutation i, p_i^t is the model probability of see mutation i at time t, $\{n_i\}^1 = \{n_i^1\}$ making $\{n_i^t\} = \{\ \{n_i\}^1, \{n_i\}^2, \ldots \{n_i\}^{t_{max}}\ \}$, and likewise $\{p_i\}^1 = \{p_i^1\}$ making $\{p_i^t\} = \{\ \{p_i\}^1, \{p_i\}^2, \ldots \{p_i\}^{t_{max}}\ \}$. We can write the likelihood of the observed mutations $\{n_i\}^t$ at time t as

$$P_t(\{n_i\}^t|\{p_i\}^t) = \frac{(\sum_i n_i^t)!}{\prod_i n_i^t!} \prod_i \left(p_i^t\right)^{n_i^t}, \tag{5.7}$$

where the prefactor is a multinomial coefficient that accounts for the different permutations for the mutational pattern.

To estimate the statistical significance of our model's ability to predict mutational pathways we calculate p-values for the calculated likelihood by generating 50,000 likelihoods using random sets of parameters. Additionally we calculated the maximum possible likelihood for comparison with our model. The maximum likelihood occurs when the model parameters are equal to the frequency data: $p_i^t = n_i^t / \sum_j n_j^t = n_i^t / N_T^t$. This can be seen by taking the

derivative of the likelihood with respect to p_i^t and setting it to zero. Note that $\sum_i p_i^t = 1$, so one of the parameters, arbitrarily we chose the last one p_N^t, can be written in terms of the others, $p_N^t = 1 - \sum_i^{N-1} p_i^t$. For ease of mathematical manipulation, we maximize the log-likelihood:

$$\frac{\partial}{\partial p_i^t} \log(P^t(\{n_i^t\}|\{p_i^t\})) = \frac{\partial}{\partial p_i^t} \sum_t (\log(\frac{(\sum_j^N n_j^t)!}{\prod_j^N n_j^t!}) + n_N^t \log(1 - \sum_j^{N-1} p_j^t) + \sum_j^{N-1} n_j^t \log(p_j^t)) \tag{5.8}$$

$$= \sum_t \frac{n_i^t}{p_i^t} - \frac{n_N}{1 - \sum_j p_j^t}. \tag{5.9}$$

Now substituting $p_i^t = n_i^t / N_T^t$ in

$$\sum_t \frac{n_i^t N_T^t}{n_i^t} - \frac{n_N N_T^t}{N_T^t - \sum_j n_j^t} = \sum_t \frac{(N_T^t - \sum_j n_j^t) n_i^t N_T^t - n_i^t n_N N_T^t}{n_i^t (N_T^t - \sum_j n_j^t)} \tag{5.10}$$

$$\sum_t \frac{n_i^t {N_T^t}^2 - n_i^t N_T^t N_T^t}{n_i^t (N_T^t - \sum_j n_j^t)} = 0 \tag{5.11}$$

For most of the patients the sequences where taken in a traditional manner and n_i^t is readily defined. However in the case of patients for whom longitudinal sequencing data was reported as relative prevalences and the absolute number of sequences is not available. To handle these situations, we set the total number of counts for these patients to be commensurate with the other patients in the corpus and verified that results were not sensitive to variations over 10-fold changes in this value.

5.3 Results

We first validate our dynamic model in comparisons to longitudinal sequencing data taken from infected hosts. We then proceed to use the validated model to rationally design HCV vaccines predicted to maximally suppress viral fitness.

5.3.1 Validation

To validate our dynamic model we simulate the course of infection in seven individuals who have participated in various clinical studies [3, 6, 224]. These simulations were started from a uniform population initialized with the first sequence taken from the patient, except for patient (BR554) for whom the infecting sequence was know. We allowed any immune response against known epitopes consistent with the patients HLA haplotypes, except for one patient (P03_32) who had a measured response against an epitope not consistent with his HLA type. For this patient we allowed this epitope to be targeted. We used the tools from the immune epitope database [215] to estimate binding affinities between this epitope and the patient's HLA alleles. We assumed the epitope was restricted by the HLA allele with the strongest binding affinity.

In comparing our model to the actual sequence data of the patients we focus on the sites where a mutation arose in the patient or reached a frequency of 50% or more in our model population. For these positions we calculate the likelihood of our model producing the patient's time course, the p-value, and the maximum possible likelihood to provide a sense of scale (see section 5.2.4). We find that for most sites for most of the patients we attain a relatively high likelihood that attains statistical significance (see table 5.2). Only for patient P03_32 do our results not reach statisitcal significance. However, the reported T-cell responses against the virus launched by this host are inconsistent with the assigned immune genotype (i.e., haplotype), suggesting that the patient was immunologically mischaracterized. These results provide strong support for the validity of our model in predicting the dynamical evolution of the host-pathogen dynamics.

5.3.2 Tailoring vaccines

Having validated our dynamic immune model, we employ the model to rationally design vaccines for patient 0684MX. Contrary to the vaccine design procedure presented in chap-

Patient	Mean likelihood	Max likelihood	p value
023	0.75	0.80	$< 10^{-16}$
0684MX	0.35	0.94	$< 10^{-16}$
1086MX	0.36	0.85	$< 10^{-16}$
BR111	0.47	1.00	2.3^{-6}
BR554	0.49	1.00	$< 10^{-16}$
M003	0.11	0.40	7.7^{-8}
P03_32	9.5e-4	1.00	0.86

Table 5.2: Mean likelihood, maximum likelihood, and p-value for the observed longitudinal sequencing profile observed in each of the seven patients considered given the parameters of our dynamical model.

ter 3, this design process explicitly incorporates the coupled host-pathogen dynamics and is expected to present a more powerful and reliable tool for vaccine design. We start by generating a list of T-cell epitope based vaccine candidates. We limit ourselves to immunogens that have at most one epitope restricted by each of the individual HLA alleles (A*02 and B*27, each restricting 3 epitopes in NS5B) to reflect the immunodominance profile of the host [12, 51, 52]. There are 3 single A*02 epitope immunogens, 3 single B*27 epitope immunogens, and 9 immunognes with an A*02 epitope with a B*27 epitope, giving 15 immunogens that could be used in a vaccine of this type. Next we simulate the course of infection in the individual having had each vaccine. Vaccination is represented by seeding the simulation with a preexisting pool of memory cells against the epitopes in the vaccine.

To evaluate the effectiveness of the vaccines we will look at two properties of the time course: (i) the maximum fitness penalty, and (ii) the length of time for which the fitness penalty is sustained (figure 5.1). Since our viral dynamics are based on a fixed population size model (see section 5.2.2) the viral population always survives and in a specific individual will eventually reach the same equilibrium fitness. Thus the depth of the initial fitness drop is a proxy for the likelihood of viral clearance and the width of the drop indicates whether different vaccines control the viral load longer than others. This length of control is one of the main things missing in the static vaccine design process.

We present the results of our vaccine design procedure in figure 5.2. Of the 15 immuno-

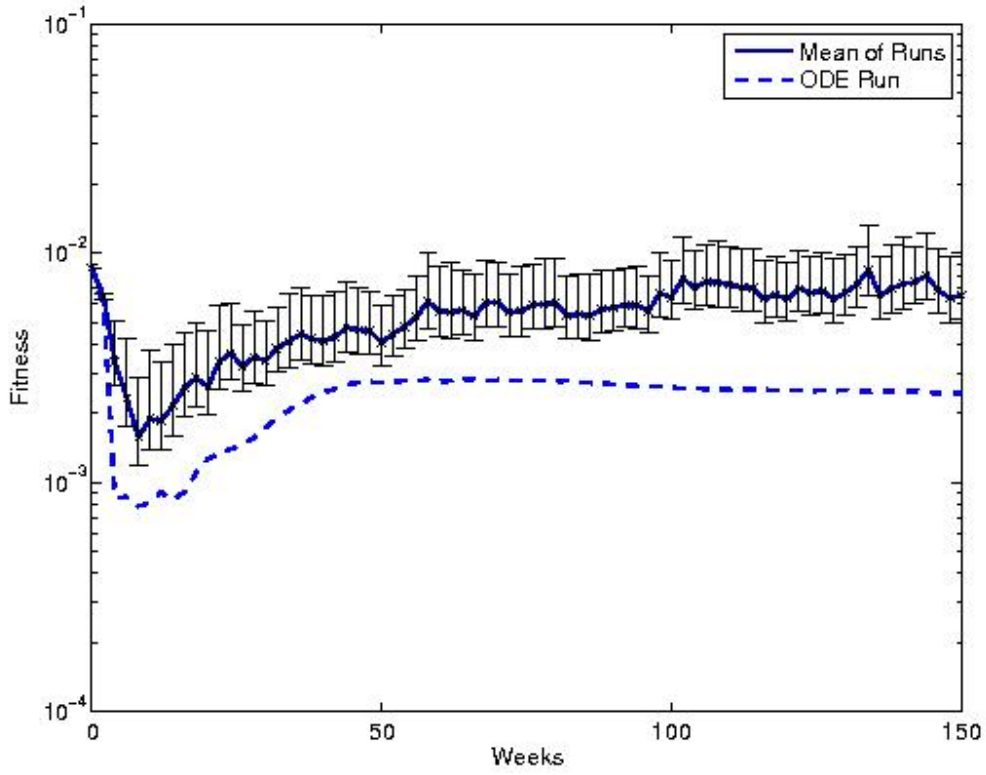


Figure 5.1: The time course for the average fitness of the viral population in patient 0684MX with no vaccination. Since our model uses a fixed population size the virus can never be cleared and will always find the same equilibrium fitness on long time scales. Accordingly, we look at the depth of the fitness dip and its full width half maximum value to estimate how strongly the virus is repressed and for how long. The solid line represents the mean of the 99 simulations employing the Gillespie algorithm to integrate the T-cell dynamics along with the associated standard deviations. The dashed line represents the results of the deterministic numerical integration. The Gillespie results explicitly account for stochastic effects arising from finite populations of T-cells.

gens 1 was worse than no vaccination in both dimensions (fitness depression and length of depression), 5 were worse in one dimension (time), 2 were better in one dimension, and 7 where better in both dimensions. The prediction of a vaccine that is worse than no vaccine is unexpected. A review of the time course of the T-cells for the unvaccinated and bad vaccination simulations reveal that priming against epitopes A*02 GLQDCTMLV and B*27 GRAAICGKY suppresses the respond against A*02 ALYDVVTKL and B*27 ARMILMTHF as well as delaying the response against A*02 RLIVFPDLGV. We will see below that the

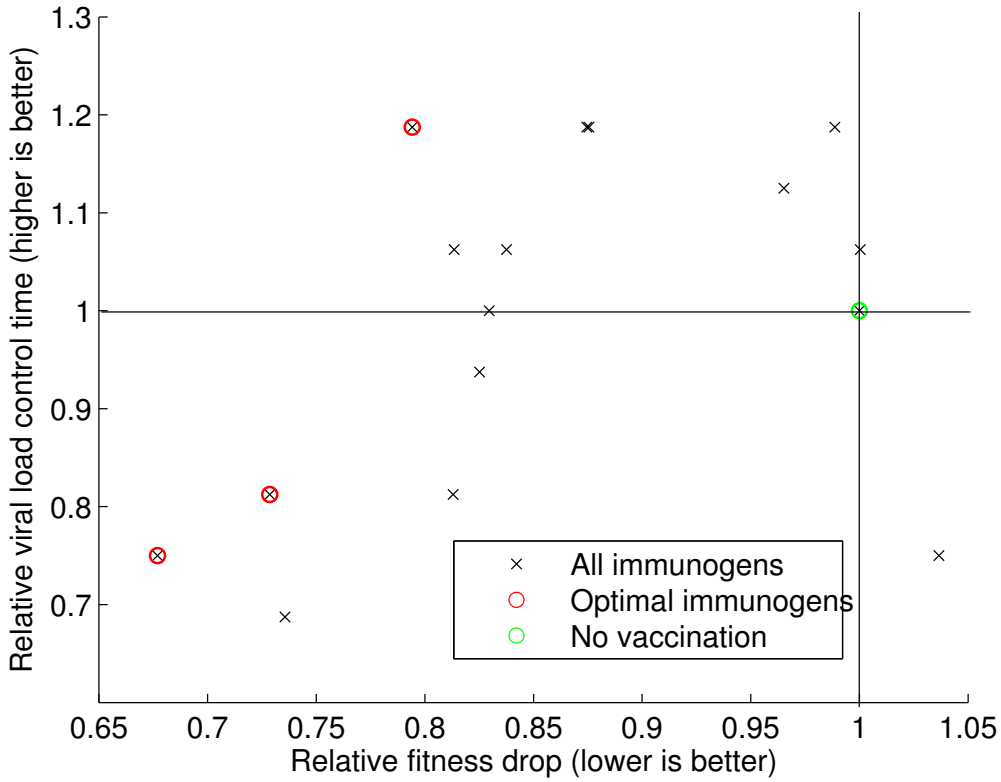


Figure 5.2: The black x's represent the 16 viral infections (15 different vaccines and no vaccines) in the space of maximum fitness depression and length of depression relative to no vaccine. The green circle indicates no vaccine. The red circles indicate the three Pareto optimal vaccines. The figure is divided into four quadrants. The bottom right quadrant represents vaccines that are worse than no vaccine. The top left represents vaccines that are better than no vaccine. The bottom left and top right represent vaccine that are better in one dimension but not the other and so may or may not be better than no vaccine.

suppressed responses are found in the optimal immunogens. This behavior is reminiscent of the "original antigenic sin" phenomenon discussed in the influenza literature [225, 226]. Employing a Pareto analysis over the 15 vaccines, we identified three Pareto optimal vaccines of which two were only better than no vaccine in one deminsion, the last one was better in both dimensions. The latter vaccine candidate – containing A*02 RLIVFPDLGV and B*27 ARMILMTHF – represents the optimal vaccine candidate for this patient designed using our approach.

Comparing our dynamic vaccine design procedure to the optimal vaccine for patient

0684MX predicted by the static design protocol detailed in chapter 3, we find that they do not predict the same optimal immunogens. While the exact composite of these optimal immunogens are different we observe that there are similarities. Both methods predict that optimal immunogens will contain either A*02 ALYDVVTKL or B*27 ARMILMTHF (see table 5.3). Therefore, as expected, the dynamic design process is not producing entirely new predictions, but adjustments or corrections to the predictions of the static design process.

Epitopes	Static optimal	Dynamic optimal
A*02 ALYDVVTKL		
A*02 GLQDCTMLV		Yes
A*02 RLIVFPDLGV		
B*27 ARHTPVNSW		
B*27 ARMILMTHF	Yes	
B*27 GRAAICGKY		
A*02 ALYDVVTKL, B*27 ARHTPVNSW	Yes	
A*02 ALYDVVTKL, B*27 ARMILMTHF		Yes
A*02 ALYDVVTKL, B*27 GRAAICGKY		
A*02 GLQDCTMLV, B*27 ARHTPVNSW		
A*02 GLQDCTMLV, B*27 ARMILMTHF		
A*02 GLQDCTMLV, B*27 GRAAICGKY		
A*02 RLIVFPDLGV, B*27 ARHTPVNSW		
A*02 RLIVFPDLGV , B*27 ARMILMTHF		Yes
A*02 RLIVFPDLGV, B*27 GRAAICGKY		

Table 5.3: The 15 possible T-cell immunogens using no more than one epitope for each HLA.

5.4 Conclusion

In this work we were able to build an immune simulator coupled to an empirical fitness landscape. This coupling allowed for the simulation of the viral population evolving in the presence of the host's immune system. This allows us to explore viral responses to immune pressure, identify potential escape pathways, and perform dynamical vaccine design.

We showed that our model can largely reproduce the mutational patterns that arise in a patient over the course of an infection. This is a strong indicator that our simple model is capturing the important feature of the host-viral interactions. With this confidence we can

gain insight into the specific escape ways the virus uses to evade immune (or drug) pressure and design treatments that either are compromised by the pathways or block them.

To this end, we present a simple example of how this dynamic model can be used in the design of vaccines. By designing optimal vaccine candidates for a single individual using both the fitness landscape alone and dynamics over the fitness landscape, we are able to easily see that the two methods predict different things. This indicates that while there are some epitopes where mutations introduce large fitness penalties, the viral population is able to quickly recover for this. It is possible with our model to look at the specific compensatory mutations that arise; however, we are instead turning our focus onto developing a vaccine for a general population instead of one person.

We plan on predicting optimal vaccine candidates for whole populations by simulating vaccination in many different individuals (different genetic backgrounds). Then for each individual we can calculate the two dimension score for the vaccine (suppression strength and duration). Finally, we can aggregate the individual scores weighted by the frequency of the genetic background as we do for the static vaccine design (see chapter 3). We can optimize on the two criterion or add addition ones as needed (such as immunogen size).

Chapter 6

Conclusion and Future Work

This thesis concerns the data-driven construction of empirical fitness landscapes and their use in the design of vaccines and other viral treatment. The development of vaccines is one of the great successes of medicine. Vaccines have improved the quality of life and lengthened life expectancy. Although the principles behind vaccines are well understood, many of the world's most taxing diseases continue to evade efforts to develop a effective vaccine. The use of an computational vaccine design platform based on empirical fitness landscapes can be used to accelerate the vaccine development process by offering a rapid rational design method for vaccine candidates.

In chapter 2 we discuss fitness landscapes. A fitness landscape for a virus [1, 12] can be conceptualized as a topographical map in which the location of a point on the landscape specifies the amino acid sequence of the virus, and the elevation of the landscape at that point specifies its replicative capacity. We discuss a little of the history of fitness landscapes and their importance in mathematical models of viral evolution. We discuss the basic types of theoretical fitness landscapes and move on to efforts to determine them experimentally or otherwise reconstruct them from data. After this discussion of the efforts of others we present the model we are using as well as its strengths and weaknesses. Finally we demonstrate how our model works by creating a fitness landscapes for toy protein with two amino-acids, which allows the landscapes to be viewable in 3 dimensions.

In chapter 3 we construct a fitness landscape for HCV protein NS5B from clinical sequences. We then compared our fitness landscape with 5 different types of experimental and clinical data. Our model had good agreement with measured *in vitro* fitness and could

predict or explain viral behavior such as which positions in a protein mutate to escape immune pressures. With the validity of the landscape confirmed we than demonstrated how the fitness landscape can be used in the vaccine design process. We generated a list 16.8 million vaccine candidates based on the targets of killer T-cells in NS5B and then ranked them all based on the average fitness cost to evade immune pressure and how much of the human population could respond to the vaccine. This ranking revealed 86 optimal vaccine candidates, reducing the experimental search space by 5 orders of magnitude.

In chapter 4 we demonstrated a precise mathematical correspondence between the viral error catastrophe and a first order phase transition in the sequence space of the virus. It has been suggested that rapidly mutating viruses may be susceptible to treatment with mutagenic drugs to induce an error catastrophe, where the mutation rate raises to the point where not enough genetic material is passed from one generation to another to propagate the population. Using a fitness landscape for a small HIV protein (p6) we showed that the protein resides near the error catastrophe, and we showed that the error catastrophe could be triggered by targeting several pairs of amino acids, opening up the possibility of small molecule inhibitors or other drugs designed to induce the error catastrophe. Furthermore we showed that no known killer T-cell targets in the the protein could induce the effect providing a possible rationalization for why HIV may reside so close to the error catastrophe with impunity.

In chapter 5 we discussed coupling our fitness landscapes with population dynamics models. This allowed us to create an immune simulator in which the fitness landscape represented the "playing field" that the virus evolves over. We showed that our model does a good job of reproducing the mutational pathways the viral population observed in longitudinal sequencing data of HCV infected hosts. We further showed that with our dynamics model we can simulate vaccination and thus screen candidates. This will allow us to follow a similar design procedure to that introduced in chapter 3, but now explicitly incorporating information on the host-pathogen dynamics.

In continuing work on HCV, I have determined fitness landscapes for proteins such has NS3 and NS5A (the other targets of antiviral treatment). For some of the proteins I have inferred landscapes for additional genotypes as well and have constructed one multi-genotype landscape. The next step is to develop collaborations with experimentalists and clinicians to test our predictions. I have also begun preliminary work on other viruses, fitting landscapes for all of influenza A's proteins and have been working on both T-cell and B-cell vaccine design. To fully tackle a non-chronic disease like influenza our methodology needs to extend to handle antibody targets. T-cell epitopes are well defined small continuous strings of amino acids making it straightforward to predict a mutation effecting them and thus calculate the fitness cost of such a mutation. Furthermore, as mentioned in chapter 5, tools have been developed to predict the effect on binding affinity and recognition of mutations within an T-cell epitopes [215]. Antibody epitopes on the other hand are dependent on conformation. This means the amino acids involved are not necessarily next to one another in the sequence. Furthermore amino acids not actually binding to the antibody can effect the conformation and thus it is hard to define the epitope. In the future I hope to learn more about the available tools testing antibodies (supplementing them with molecule dynamics if necessary) and add antibodies to our design process.

In addition to fully utilizing our vaccine design process there are still some open questions about our model for fitness landscapes. One that has been of particular interest to me is the relationship between the fitness of a sequence, $f(\vec{A})$ and its prevalence in the population, $P(\vec{A})$. As mentioned in chapter 2 our model actually finds the prevalence landscape. However under certain conditions it has been shown that the rank ordering for prevalences and fitnesses are preserved [106, 109]. We have shown empirically that our model is a strong predictor *in vitro* fitness [12, 81, 111]. However, Niko Beerenwinkel and co-workers have showed that using the quasispecies model the rank ordering does not have to be preserved when going from prevalence to fitness or visa versa [2]. This opens up the questions: what determines if prevalence is a good proxy for fitness; will it always be a good proxy for biolog-

ically relevant landscapes; and can we improve our prediction by more carefully considering the relationship between prevalence and fitness? I am currently working on these questions, and I am developing a way to transform our prevalence landscape into a fitness landscape in a restricted portion of sequence space.

One of the other questions that lingers in my mind: what landscape structure is necessary for a phase transition to exist? We showed in chapter 4 that the HIV p6 protein is on the edge of a phase transition, but we also showed a phase transition is not inherent to the model. This leaves open the question of what features of the viral fitness landscape dictate whether there is a phase transition.

Appendix A

Mathematical Details

A.1 Maximum entropy

First we will show that the maximum entropy model that reproduces the observed frequencies of animo acids and pairs of amino acids is the Potts model. The Shannon entropy is

$$S = -\sum_{k=1}^{q^m} P(\vec{z}_k) \log P(\vec{z}_k), \tag{A.1}$$

where the protein is m amino acids long and $q = 21$ if we include all natural amino acids and an unknown. The entropy is constrained by the normalization condition

$$\sum_{k=1}^{q^m} P(\vec{z}_k) = 1, \tag{A.2}$$

and reproducing the amino acid frequencies

$$P_i^{obs}(p) = \sum_{k=1}^{q^m} \sigma_{pz_i} P(\vec{z}_k) \tag{A.3}$$

$$P_{ij}^{obs}(p,r) = \sum_{k=1}^{q^m} \sigma_{pz_i}\sigma_{rz_j} P(\vec{z}_k). \tag{A.4}$$

These constraints make maximizing the entropy a Lagrangian maximization problem

$$L = -\sum_{k=1}^{q^m} P(\vec{z}_k) \log P(\vec{z}_k) \tag{A.5}$$

$$+\alpha \left(\sum_{k=1}^{q^m} P(\vec{z}_k) - 1 \right) \tag{A.6}$$

$$+\sum_{i=1}^{m} \sum_{p=1}^{q} \left[h_i(p) \left(P_i^{obs}(p) - \sum_{k=1}^{q^m} \sigma_{pz_i} P(\vec{z}_k) \right) \right] \tag{A.7}$$

$$+\sum_{i=1}^{m} \sum_{j=i+1}^{m} \sum_{p=1}^{q} \sum_{r=1}^{q} \left[J_{ij}(p,r) \left(P_{ij}^{obs}(p,r) - \sum_{k=1}^{q^m} \sigma_{pz_i} \sigma_{rz_j} P(\vec{z}_k) \right) \right], \tag{A.8}$$

where α, $\{h_i\}$, and $\{J_{ij}\}$ are the Lagrangian multipliers.

We proceed by taking all the partial derivatives and simultaneously setting them to zero,

$$\frac{\partial L}{\partial P(\vec{z}_k)} = 0 = -\log P(\vec{k}) - 1 + \alpha - \sum_{i=1}^{m} h_i(z_i) - \sum_{i=1}^{m} \sum_{j=i+1}^{m} J_{ij}(z_i, z_j). \tag{A.9}$$

Solving for $P(\vec{z}_k)$

$$\log P(\vec{z}_k) = -1 + \alpha - \sum_{i=1}^{m} h_i(z_i) - \sum_{i=1}^{m} \sum_{j=i+1}^{m} J_{ij}(z_i, z_j) \tag{A.10}$$

$$P(\vec{z}_k) = e^{-1+\alpha} e^{-\left(\sum_{i=1}^{m} h_i(z_i) + \sum_{i=1}^{m} \sum_{j=i+1}^{m} J_{ij}(z_i, z_j)\right)}. \tag{A.11}$$

The factor of $e^{-1+\alpha}$ is identical for all the $P(\vec{z}_k)$, it is in fact the normalization factor which we call Z

$$Z = \sum_{k=1}^{q^m} P(\vec{z}_k) \tag{A.12}$$

$$= \sum_{k=1}^{q^m} e^{-\left(\sum_{i=1}^{m} h_i(z_i) + \sum_{i=1}^{m} \sum_{j=i+1}^{m} J_{ij}(z_i, z_j)\right)}. \tag{A.13}$$

Therefore

$$P(\vec{z}_k) = \frac{e^{-\left(\sum_{i=1}^{m} h_i(z_i) + \sum_{i=1}^{m}\sum_{j=i+1}^{m} J_{ij}(z_i,z_j)\right)}}{\sum_{k=1}^{q^m} e^{-\left(\sum_{i=1}^{m} h_i(z_i) + \sum_{i=1}^{m}\sum_{j=i+1}^{m} J_{ij}(z_i,z_j)\right)}} \tag{A.14}$$

$$= \frac{1}{Z} e^{-H(\vec{z})}, \tag{A.15}$$

if we call Z the partition function, H the Hamiltonian, $\{h_i\}$ the external fields, and $\{J_{ij}\}$ the interaction couplings, we have an infinite range Potts model.

A.2 Maximum likelihood model

We now show that adjusting the model parameters $\{h_i, J_{ij}\}$ to reproduce the observed frequency of amino acids and amino acid pairs ($\{P_i^{obs}(p), P_{ij}^{obs}(p,r)\}$), not only maximizes the entropy, but also result in a maximum likelihood estimate. Given a multiple sequence alignment (MSA see the beginning of section 2.5.2) containing K sequences, let the observed probability of a particular sequence $\vec{z}_k$ be $P^{obs}(\vec{z}_k)$ and the model prediction of that state be $P(\vec{z}_k)$. This gives us

$$L(model|data) = P(data|model)P(model) \tag{A.16}$$

where $P(model)$ is a Bayesian prior or alternatively it can be thought of as a regularization factor. For this factor we use $\prod_i e^{-\beta\lambda_i\|\theta\|_2}$, where λ_i is the regularization strength, θ_i represents the model parameters, and $\beta = \frac{1}{kT}$ the inverse temperature common in statistical physics. Setting $\lambda_i = 0$, represents a uniform prior. Going forward we will use D instead of writing out *data* and $\vec{\theta}$ in place of *model* representing the vector of adjustable model parameters, $\{h_i, J_{ij}\}$.

Now, assuming independent and identically distributed observation we have

$$L(\vec{\theta}|D) = P(D|\vec{\theta})P(\vec{\theta}) \tag{A.17}$$

$$= \left[\prod_{\{\vec{z}\}} P(\vec{z})^{P^{obs}(\vec{z})K}\right]\left[\prod_i e^{-\beta\lambda\|\theta_i\|_2}\right] \tag{A.18}$$

$$\log L(\vec{\theta}|D) = \sum_{\{\vec{z}\}} KP^{obs}(\vec{z})\log P(\vec{z}) - \sum_i \beta\lambda_i\|\theta_i\|_2 \tag{A.19}$$

$$= K\sum_{\{\vec{z}\}} P^{obs}(\vec{z})\log P(\vec{z}) - \sum_i \beta\lambda_i\|\theta_i\|_2 \tag{A.20}$$

$$\frac{1}{K}\log L(\vec{\theta}|D) = \sum_{\{\vec{z}\}} P^{obs}(\vec{z})\log P(\vec{z}) - \sum_i \frac{1}{K}\beta\lambda_i\|\theta_i\|_2 \tag{A.21}$$

$$= \sum_{\{\vec{z}\}}\left[P^{obs}(\vec{z})\log P^{obs}(\vec{z}) - P^{obs}(\vec{z})\log\frac{P^{obs}(\vec{z})}{P(\vec{z})}\right] - \sum_i \frac{1}{K}\beta\lambda_i\|\theta_i\|_2 \tag{A.22}$$

$$= -S^{obs} - D_{KL}(P^{obs}||P) - \sum_i \frac{1}{K}\beta\lambda_i\|\theta_i\|_2 \tag{A.23}$$

$$= -S^{obs} - \left[D_{KL}(P^{obs}||P) + \sum_i \frac{1}{K}\beta\lambda_i\|\theta_i\|_2\right] \tag{A.24}$$

$$= -S^{obs} - D_{KL}^R \tag{A.25}$$

Here, S^{obs} is the entropy of the MSA, D_{KL} is the Kullback-Leibler divergence between the observed and model probability distributions, and D_{KL}^R is the regularized KL divergence. Maximizing this log-likelihood is the same as minimizing D_{KL}^R. We proceed by taking the partial derivatives of D_{KL}^R with respect to h_i^p (where h_i^p is equivalent to $h_i(p)$) and set them

to zero,

$$\frac{\partial D_{KL}^{R}}{\partial h_i^p} = \frac{\partial}{\partial h_i^p} \sum_{\{\vec{z}\}} \left[P^{obs}(\vec{z}) \log P^{obs}(\vec{z}) - P^{obs}(\vec{z}) \log P^{obs}(\vec{z}) \right] + \frac{\partial}{\partial h_i^p} \left[\sum_i \frac{1}{K} \beta \lambda_i \|\theta_i\|_2 \right] \tag{A.26}$$

$$= \frac{\partial}{\partial h_i^p} \sum_{\{\vec{z}\}} \left[-P^{obs}(\vec{z}) \log \frac{e^{-\beta H(\vec{z})}}{Z} \right] + \frac{\beta}{K} 2 h_i^p \lambda_i^p \tag{A.27}$$

$$= \frac{\partial}{\partial h_i^p} \sum_{\{\vec{z}\}} \left[\beta H(\vec{z}) P^{obs}(\vec{z}) + \log Z P^{obs}(\vec{z}) \right] + \frac{\beta}{K} 2 h_i^p \lambda_i^p \tag{A.28}$$

$$= \beta \sum_{\{\vec{z}\}} P^{obs}(\vec{z}) \frac{\partial}{\partial h_i^p} H(\vec{z}) + \frac{\partial}{\partial h_i^p} \log Z \sum_{\{\vec{z}\}} P^{obs}(\vec{z}) + \frac{\beta}{K} 2 h_i^p \lambda_i^p \tag{A.29}$$

$$= \beta \sum_{\{\vec{z}\}} P^{obs}(\vec{z}) \frac{\partial}{\partial h_i^p} H(\vec{z}) + \frac{1}{Z} \frac{\partial}{\partial h_i^p} Z + \frac{\beta}{K} 2 h_i^p \lambda_i^p. \tag{A.30}$$

In going from equation A.26 to A.27 we used the fact that $P^{obs}(\vec{z}) \log P^{obs}(\vec{z})$ is independent of h_i^p and in going from equation A.29 to A.30 we used the fact that $\sum_{\{\vec{z}\}} P^{obs}(\vec{z}) = 1$. Now we need the derivatives of the partition function and Hamiltonian. First the partition function

$$Z = \sum_{\{\vec{z}\}} e^{-\beta H(\vec{z})} = \sum_{\{\vec{z}\}} e^{-\beta \left[\sum_i \sum_p h_i^p \sigma_{z_i p} + \frac{1}{2} \sum_i \sum_{j \neq i} \sum_p \sum_r J_{ij}^{pr} \sigma_{z_i p} \sigma_{z_j r} \right]} \tag{A.31}$$

$$= \sum_{\{\vec{z}\}} e^{-\beta \left[\sum_p h_k^p \sigma_{z_k p} + \sum_{j \neq k} \sum_p \sum_r J_{kj}^{pr} \sigma_{z_k p} \sigma_{z_j r} \right]} e^{-\beta \left[\sum_{i \neq k} \sum_p h_i^p \sigma_{z_i p} + \frac{1}{2} \sum_i \sum_{j \neq i, j \neq k} \sum_p \sum_r J_{ij}^{pr} \sigma_{z_i p} \sigma_{z_j r} \right]} \tag{A.32}$$

$$\frac{\partial Z}{\partial h_k^s} = -\beta \sum_{\{\vec{z}\}} \sigma z_k s E^{-\beta H(\vec{z})} \tag{A.33}$$

$$= -\beta Z(z_k = s) \tag{A.34}$$

where $Z(z_k = s)$ is a partial partition function, summing over only sequences with $z_k = s$.

And now for the Hamiltonian,

$$\frac{\partial H(\vec{z})}{\partial h_k^s} = \frac{\partial}{\partial h_k^s}\left[\sum_p h_k^p \sigma_{z_k p} + \sum_{j\neq k}\sum_p\sum_r J_{kj}^{pr}\sigma_{z_k p}\sigma_{z_j r} + \frac{1}{2}\sum_i\sum_{j\neq i, j\neq k}\sum_p\sum_r J_{ij}^{pr}\sigma_{z_i p}\sigma_{z_j r}\right] \tag{A.35}$$

$$= \sigma_{z_k s} \tag{A.36}$$

These equations (A.34 and A.36) are now substituted back into equation A.30,

$$\frac{\partial D_{KL}^R}{\partial h_i^p} = \beta\sum_{\{\vec{z}\}} P^{obs}(\vec{z})\sigma_{z_i p} - \beta\frac{1}{Z}Z(z_i = p) + \frac{\beta}{K}2h_i^p\lambda_i^p \tag{A.37}$$

$$= \beta\left[P_1^{obs}(z_i = p) - P_i(z_i = p) + \frac{1}{K}2(h_i^p)\lambda_i^p\right] \tag{A.38}$$

$$\tag{A.39}$$

Now we proceed in a similar fashion, taking partial derivatives with respect to $\{J_{ij}^{pr}\}$:

$$\frac{\partial D_{KL}^R}{\partial J_{ij}^{pr}} = \beta\sum_{\{\vec{z}\}} P^{obs}(\vec{z})\frac{\partial}{\partial J_{ij}^{pr}}H(\vec{z}) + \frac{1}{Z}\frac{\partial}{\partial J_{ij}^{pr}}H(\vec{z}) + \frac{\beta}{K}2(J_{ij}^{pr})\lambda_{ij}^{pr}. \tag{A.40}$$

Now for Z and H again:

$$Z = \sum_{\{\vec{z}\}} e^{-\beta H(\vec{z})} = \sum_{\{\vec{z}\}} e^{-\beta\left[\sum_i\sum_p h_i^p\sigma_{z_i p} + \frac{1}{2}\sum_i\sum_{j\neq i}\sum_p\sum_r J_{ij}^{pr}\sigma_{z_i p}\sigma_{z_j r}\right]} \tag{A.41}$$

$$Z = \sum_{\{\vec{z}\}} e^{-\beta\left[\sum_p h_k^p\sigma_{z_k p} + \sum_{j\neq k, j\neq l}\sum_p\sum_r J_{kj}^{pr}\sigma_{z_k p}\sigma_{z_j r}\right]} e^{-\beta\left[\sum_p h_l^p\sigma_{z_l p} + \sum_{j\neq k, j\neq l}\sum_p\sum_r J_{lj}^{pr}\sigma_{z_l p}\sigma_{z_j r}\right]} \\ e^{-\beta\left[\sum_p\sum_r J_{kl}^{pr}\sigma_{z_k p}\sigma_{z_l r}\right]} e^{-\beta\left[\sum_{i\neq k, i\neq l}\sum_p h_i^p\sigma_{z_i p} + \frac{1}{2}\sum_i\sum_{j\neq i, j\neq k, j\neq l}\sum_p\sum_r J_{ij}^{pr}\sigma_{z_i p}\sigma_{z_j r}\right]} \tag{A.42}$$

$$\frac{\partial Z}{\partial J_{kl}^{st}} = -\beta \sum_{\{\vec{z}\}} \sigma_{z_k s}\sigma_{z_l t} e^{-\beta H(\vec{z})} = -\beta Z(z_k = s, z_l = t) \tag{A.43}$$

$$\frac{\partial H(\vec{z})}{\partial J_{kl}^{st}} = \sigma_{z_k s}\sigma_{z_l t}. \tag{A.44}$$

Substituting back,

$$\frac{\partial D_{KL}^{R}}{\partial J_{ij}^{pr}} = \beta \sum_{\{\vec{z}\}} P^{obs}(\vec{z})\sigma_{z_k s}\sigma_{z_l t} + \beta \frac{1}{Z} Z(z_i = p, z_j = r) + \frac{\beta}{K} 2(J_{ij}^{pr})\lambda_{ij}^{pr} \tag{A.45}$$

$$= \beta \left[P_2^{obs}(z_i = p, z_j = r) + P_2(z_i = p, z_j = r) + \frac{\beta}{K} 2(J_{ij}^{pr})\lambda_{ij}^{pr} \right] \tag{A.46}$$

Thus, D_{KL}^{R} is minimized, and log likelihood maximized, when

$$P_1(z_i = p) = P_1^{obs}(z_i = p) + \frac{1}{K} 2(h_i^p)\lambda_i^p \tag{A.47}$$

$$P_2(z_i = p, z_j = q) = P_2^{obs}(z_i = p, z_j = q) + \frac{\beta}{K} 2(J_{ij}^{pr})\lambda_{ij}^{pr}. \tag{A.48}$$

Taking the regularization to zero, $\lambda_i^p = \lambda_{ij}^{pr} = 0$, we see that the maximum likelihood estimate of the parameters are the same as the parameter set that produces the observed one and two amino acid frequencies. Thus we can use Bayesian inference to estimate the model parameters.

A.3 Regularization

In the above section the mathematical details were worked out with a non-uniform prior. This prior can also be thought of as a regularization. We use two forms of regularization which we refer to as offline and online regularization. The offline or a priori regularization addresses the issue of noisy or missing data and the online or Bayesian regularization addresses issues of numeric stability in the algorithm.

The size of sequence space (q^m) is very large much larger than the number of sequences

K that have been sampled. This means that there are many unobserved mutations and pairs of mutations, leading to some $P_i^{obs}(z_i)$ and $P_{ij}^{obs}(z_i, z_j)$ equaling zero. If the probability of seeing these truly are zero than the corresponding parameters $h_i(z_i)$ and $J_{ij}(z_i, z_j)$ need to be infinite. This is where the offline regularization comes in, by adding a pseudo-count to our observed or targeted probabilities. This means adding a (possibly non-integer) number of fictitious observations of an amino acid (or pair of amino acids) within the MSA to reflect the belief that the probability of observing this amino acid (or pair) in this position is not precisely zero, but rather a low-probability event that is not observed within the finite number of strains within the MSA. However many sequences are not viable and some amino acids may not be observed even in arbitrarily large data sets. Therefore we adopt the following middle ground. We assume that K is large enough to observe any single amino acid that is viable, therefore we remove unobserved amino acids from our model (effective setting the corresponding h_i to infinity). We further assume that any single amino acid that appears in the MSA could appear in a pair with any other amino acid in the MSA and use pseudo-counts to assure non-zero P_{ij} values.

These regularized target probabilities are the input for our fitting algorithm. The numerical stability of this algorithm is improved by our second (online) regularization. Without this regularization the algorithm can be slow to converge as often coupled groups of h_i and J_{ij} elements exhibit uncontrolled growth (some positive, some negative) that taken together have little effect on the predicted frequencies. This issue is addressed with the Gaussian Bayes prior from the previous section. In the fitting algorithm this factor acts as a penalty on model parameters that grow large, limiting the uncontrolled growth. This regularization could also be viewed from a frequentist perspective as the addition of pseudo-counts.

When using pseudo-counts it is important to rebalance the target probabilities to assure

they are normalized. This is done as follows,

$$N_1^{tar} = KP_1(z_i = p) = KP_1^{obs}(z_i = p) + 2(h_i^p)\lambda_i^p \tag{A.49}$$

$$= N_1^{obs} + 2(h_i^p)\lambda_i^p = \tilde{N}_1^{obs} \tag{A.50}$$

$$\tilde{P}_1(z_i = \gamma) = \frac{\tilde{N}_1^{obs}(z_i = \gamma)}{\sum_\alpha \tilde{N}_1^{obs}(z_i = \alpha)} \tag{A.51}$$

$$= \frac{KP_1^{obs}(z_i = \gamma) + 2(h_i^\gamma)\lambda_i^\gamma}{\sum_\alpha KP_1^{obs}(z_i = \alpha) + 2(h_i^\alpha)\lambda_i^\alpha} \tag{A.52}$$

$$= \frac{P_1^{obs}(z_i = \gamma) + \frac{1}{K}2(h_i^\gamma)\lambda_i^\gamma}{\sum_\alpha P_1^{obs}(z_i = \alpha) + \frac{1}{K}2(h_i^\alpha)\lambda_i^\alpha} \tag{A.53}$$

$$\tilde{P}_1^{obs}(z_i = \gamma) \leftarrow \frac{\tilde{P}_1^{obs}(z_i = \gamma)}{\sum_\alpha \tilde{P}_1^{obs}(z_i = \alpha)}, \tag{A.54}$$

assuming probabilities over residues, α, at each site, i, sum to unity by renormalizing the elements but retaining relative proportion altered by pseudo-counts.

For P_2, must assume that the marginals are consistent:

$$\exists\sigma \qquad \sum_\alpha P_2^{obs}(z_i = \alpha, z_j = \sigma) = P_1^{obs}(z_j = \sigma) \tag{A.55}$$

$$\exists\sigma \qquad \sum_\gamma P_2^{obs}(z_i = \gamma, z_j = \alpha) = P_1^{obs}(z_i = \gamma) \tag{A.56}$$

Succinctly, we define a factor $\frac{1}{r_k}$ for each row k, and $\frac{1}{c_l}$ for each column l, of the $P_2^{obs}(z_i, z_j)$

matrix which may be adjusted to ensue row and column normalization:

$$\mathbf{R^{-1}PC^{-1}} = \begin{bmatrix} \frac{1}{r_1} & & \\ & \frac{1}{r_2} & \\ & & \ddots \end{bmatrix} \begin{bmatrix} P_2^{obs}(z_i = \alpha_1, z_j = \alpha_1') & P_2^{obs}(z_i = \alpha_1, z_j = \alpha_2') & \dots \\ P_2^{obs}(z_i = \alpha_2, z_j = \alpha_1') & P_2^{obs}(z_i = \alpha_2, z_j = \alpha_2') & \dots \\ \vdots & \vdots & \ddots \end{bmatrix} \begin{bmatrix} \frac{1}{c_1} & & \\ & \frac{1}{c_2} & \\ & & \ddots \end{bmatrix} \tag{A.57}$$

$$= \begin{bmatrix} \frac{P_2^{obs}(z_i=\alpha_1, z_j=\alpha_1')}{r_1 c_1} & \frac{P_2^{obs}(z_i=\alpha_1, z_j=\alpha_2')}{r_1 c_2} & \dots \\ \frac{P_2^{obs}(z_i=\alpha_2, z_j=\alpha_1')}{r_2 c_1} & \frac{P_2^{obs}(z_i=\alpha_2, z_j=\alpha_2')}{r_2 c_2} & \dots \\ \vdots & \vdots & \ddots \end{bmatrix}. \tag{A.58}$$

with this $\mathbf{R^{-1}PC^{-1}}\vec{1} = \tilde{\mathbf{P}}_1^{obs}(z_i = \alpha)$, a column vector of ones sums over rows of $\mathbf{R^{-1}PC^{-1}}$ returning $\tilde{\mathbf{P}}_1^{obs}(z_i = \alpha)$ a column vector of site i marginals, normalized as per above. And $\vec{1}^T\mathbf{R^{-1}PC^{-1}} = \tilde{\mathbf{P}}_1^{obs}(z_j = \alpha')$, a row vector of ones sums over columns of $\mathbf{R^{-1}PC^{-1}}$ returning $\tilde{\mathbf{P}}_1^{obs}(z_j = \alpha')$ a row vector of site j marginals, normalized as per above.

This is a nonlinear optimization problem. We just iterate until converged:

$$\vec{r} = sum(\mathbf{P}, 2) \tag{A.59}$$

$$\mathbf{P} \leftarrow \text{diag}(\frac{P_1}{r})\mathbf{P} \tag{A.60}$$

$$\vec{c} = sum(\mathbf{P}, 1) \tag{A.61}$$

$$\mathbf{P} \leftarrow \mathbf{P}\text{diag}(\frac{P_1}{c}) \tag{A.62}$$

A.4 Gauge fixing

The model we have describe here has a gauge invariance, to use this model we need a gauge fixing scheme. As a reminder the probability assigned to a particular state $\vec{x}$ within the

q-state Potts mode is:

$$P(\vec{x}) = \frac{1}{Z} \exp\left[-\beta(\sum_i h_i(X_i) + \sum_{j>i} J_{ij}(X_i, X_j))\right] \tag{A.63}$$

where $X_i = \{1 \dots q\}$.

The number of parameters $= \left[Nq + \binom{N}{2}q^2\right]$, there is one h for each state (amino acid) at each site and one J for each possible pairing of states at every pair of sites.

$$\begin{aligned} T =& Nq + \binom{N}{2}q^2 \qquad \leftarrow \text{1-body \& 2-body marginals} \\ & - N \qquad \leftarrow \text{1-body normalization}^1 \\ & - \binom{N}{2} \qquad \leftarrow \text{2-body normalization}^2 \\ & - \binom{N}{2}2(q-1) \qquad \leftarrow \text{2-body to 1-body marginalization}^3 \end{aligned} \tag{A.64}$$

$$= N(q-1) + \binom{N}{2}(q^2 - 2(q-1) - 1) \tag{A.65}$$

$$= N(q-1) + \binom{N}{2}(q^2 - 2q + 1) \tag{A.66}$$

$$= N(q-1) + \binom{N}{2}(q-1)^2 \tag{A.67}$$

We make the following observations regarding this calculation:

1. $\sum_{p=1}^{q} P(X_i = p) = 1 \Rightarrow P(X_i = q) = 1 - \sum_{p=1}^{q-1} P(X_i = p)$: one of the one-body marginals at each site is not independent. There are N sites.
2. $\sum_{p=1}^{q}\sum_{r=1}^{q} P(X_i = p, X_j = r) = 1 \Rightarrow P(X_i = q, X_j = r) = 1 - \sum_{p=1}^{q}\sum_{r=1}^{q-1} P(X_i = p, X_j = r) + \sum_{p=1}^{q-1} P(X_i = p, X_j = q)$: one of the two-body marginals at each pair of sites is not independent. There are $\binom{N}{2}$ pairs of sites.
3. $P(X_i = p) = \sum_{r=1}^{q} P(X_i = p, X_j = r)$: one-body marginals are specified by the two-body marginals. q-such equations every ordered pair of sites this gives $N(N-1)q$. But one such

equation must be removed since it is known from the normalization of $P(X_i)$ in 1. (i.e. only $(q-1)$ such equation for every ordered pair) so we don't recount one-body normalization. This gives $N(N-1)(q-1) = \binom{N}{2}2(q-1)$

Therefore, we have more parameters $P = (Nq + \binom{N}{2}q^2)$ than conditions $T = (N(q-1) + \binom{N}{2}(q-1)^2)$ and so we may fix some of the parameters to break the degeneracy of the model. In effect, H is unchanged under certain shifts in h and J, it moves along a null space, or in other words it has a gauge invariance.

Gauge invariance presents interpretability problems, since the same 1 and 2 site marginals can be reproduced by altering contributions between local fields and couplings. (weight).

To fix our model we choose to pin to zero N h parameters and $\binom{N}{2}2(q-1)$ J parameters.

1. Fix at each site: $h(X_i = 1) = 0$
2. Fix for each pair of sites: $J(X_i = 1, X_j) = J(X_j, X_i = 1) = 0$ for $x_j = \{1 \dots q\}$.

This makes all couplings and external fields measured with respect to state 1 (the most frequent amino acid at that position).

We now show that this gauge fixing works for $\{h_i\}$

$$Z = \sum_{A_\alpha=\{1\dots q_\alpha\}} e^{-\left[\sum_i h_i(A_i)+1/2\sum_i\sum_{j\neq i} J_{ij}(A_i,A_j)\right]} \tag{A.68}$$

$$= \sum_{A_\alpha=\{1\dots q_\alpha\}} e^{-\left[\sum_i h_k(A_k)+\sum_{j\neq k} J_{kj}(A_k,A_j)\right]} e^{-\left[\sum_{i\neq k} h_i(A_i)+1/2\sum_{i\neq k}\sum_{j\neq i,j\neq k} J_{ij}(A_i,A_j)\right]} \tag{A.69}$$

$$Z(A_k=a_k) = \sum_{A_\alpha=\{1\dots q_\alpha\},\alpha\neq k} e^{-\left[\sum_i h_k(A_k=a_k)+\sum_{j\neq k} J_{kj}(A_k=a_k,A_j)\right]} e^{-\left[\sum_{i\neq k} h_i(A_i)+1/2\sum_{i\neq k}\sum_{j\neq i,j\neq k} J_{ij}(A_i,A_j)\right]} \tag{A.70}$$

$$= \sum_{A_\alpha=\{1\dots q_\alpha\},\alpha\neq k} e^{-\left[h_k(A_k=a_k)+\sum_{j\neq k} J_{kj}(A_k=a_k,A_j)\right]} P(\vec{(A)}_k) \tag{A.71}$$

$$= e^{-h_k(A_k=a_k)} \sum_{A_\alpha=\{1\dots q_\alpha\},\alpha\neq k} e^{-\left[\sum_{j\neq k} J_{kj}(A_k=a_k,A_j)\right]} P(\vec{(A)}_k) \tag{A.72}$$

$$P(A_k=a_k) = \frac{Z(A_k=a_k)}{Z} = \frac{e^{-h_k(A_k=a_k)}\sum_{A_\alpha=\{1\dots q_\alpha\},\alpha\neq k} e^{-\left[\sum_{j\neq k} J_{kj}(A_k=a_k,A_j)\right]} P(\vec{(A)}_k)}{\sum_{A_\alpha=\{1\dots q_\alpha\}} e^{-h_k(A_k=a_k)} e^{-\left[\sum_{j\neq k} J_{kj}(A_k=a_k,A_j)\right]} P(\vec{(A)}_k)} \tag{A.73}$$

$$= \frac{e^{-h_k(A_k=a_k)}\sum_{A_\alpha=\{1\dots q_\alpha\},\alpha\neq k} e^{-\left[\sum_{j\neq k} J_{kj}(A_k=a_k,A_j)\right]} P(\vec{(A)}_k)}{e^{-h_k(A_k=a_k)}\sum_{A_\alpha=\{1\dots q_\alpha\}} e^{-[h_k(A_k)-h_k(A_k=a_k)]} e^{-\left[\sum_{j\neq k} J_{kj}(A_k=a_k,A_j)\right]} P(\vec{(A)}_k)} \tag{A.74}$$

Therefore $P(A_k=a_k)$ is unchanged upon subtracting a constant from all members of the vector $\vec{h}_k = h_k(A_k=1)\dots h_k(A_k=q_k)$. Showing this for $\{J_{ij}\}$ is similar.

A.5 Newton step

Our algorithm for determining the parameters based on gradient descent. This gradient decent takes the form of,

$$\begin{bmatrix} \vec{P}_1^{obs} \\ \vec{P}_2^{obs} \end{bmatrix} = \begin{bmatrix} \vec{P}_1 \\ \vec{P}_2 \end{bmatrix} + \mathbf{J} \begin{bmatrix} \Delta\vec{h} \\ \Delta\vec{J} \end{bmatrix} \tag{A.75}$$

where P_1 and P_2 are derived from the model and $\mathbf{J}$ is the Jacobian. We sample sequences from the current model using Markov-chain Metropolis Monte-Carlo (MC), and from these sequences calculate P_1 and P_2. By assuming that coupling is small, i.e. that each probability only needs to be expanded in its "own" parameter, the Jacobian becomes diagonal and this system simplifies to

$$P_i^{obs}(p) = P_i(p) + \frac{\partial P_i(p)}{\partial h_i(p)} \Delta h_i(p) \tag{A.76}$$

$$P_{ij}^{obs}(p,r) = P_{ij}(p,r) + \frac{\partial P_{ij}(p,r)}{\partial J_{ij}(p,r)} \Delta J_{ij}(p,r). \tag{A.77}$$

Thus the step is,

$$\Delta h_i(p) = \frac{P_i^{obs}(p) - P_i(p)}{\frac{\partial P_i(p)}{\partial h_i(p)}} = \frac{P_i^{obs}(p) - P_i(p)}{P_i(p)(P_i(p) - 1)} \tag{A.78}$$

$$\Delta J_{ij}(p,r) = \frac{P_{ij}^{obs}(p,r) - P_{ij}(p,r)}{\frac{\partial P_{ij}(p,r)}{\partial J_{ij}(p,r)}} = \frac{P_{ij}^{obs}(p,r) - P_{ij}(p,r)}{P_{ij}(p,r)(P_{ij}(p,r) - 1)}. \tag{A.79}$$

In practice these steps are multiplied by a softening factor, $\gamma < 1$ to improve numeric stability.

Now we will derive the derivatives used in equations A.81 and A.82,

$$Z = \sum_{A_\alpha = \{1 \dots q_\alpha\}} e^{-\left[\sum_i h_i(A_i) + 1/2 \sum_i \sum_{j \neq i} J_{ij}(A_i, A_j)\right]} \tag{A.80}$$

$$= \sum_{A_\alpha = \{1 \dots q_\alpha\}} e^{-\left[\sum_i h_k(A_k) + \sum_{j \neq k} J_{kj}(A_k, A_j)\right]} e^{-\left[\sum_{i \neq k} h_i(A_i) + 1/2 \sum_{i \neq k} \sum_{j \neq i, j \neq k} J_{ij}(A_i, A_j)\right]} \tag{A.81}$$

$$= \sum_{A_\alpha = \{1 \dots q_\alpha\}, \alpha \neq k} \left[\sum_{A_\alpha = \{1 \dots q_k} e^{-\left[\sum_i h_k(A_k) + \sum_{j \neq k} J_{kj}(A_k, A_j)\right]} \right] e^{-\left[\sum_{i \neq k} h_i(A_i) + 1/2 \sum_{i \neq k} \sum_{j \neq i, j \neq k} J_{ij}(A_i, A_j)\right]} \tag{A.82}$$

$$Z(A_k = a_k) = \sum_{A_\alpha = \{1 \ldots q_\alpha\}, \alpha \neq k} e^{-\left[\sum_i h_k(A_k = a_k) + \sum_{j \neq k} J_{kj}(A_k = a_k, A_j)\right]} e^{-\left[\sum_{i \neq k} h_i(A_i) + 1/2 \sum_{i \neq k} \sum_{j \neq i, j \neq k} J_{ij}(A_i, A_j)\right]} \tag{A.83}$$

$$\frac{\partial Z}{\partial h_k(A_k = a_k)} = - Z(A_k = a_k) \tag{A.84}$$

$$\frac{\partial Z(A_k = a_k)}{\partial h_k(A_k = a_k)} = - Z(A_k = a_k) \tag{A.85}$$

$$P(A_k = a_k) = \frac{Z(A_k = a_k)}{Z} \tag{A.86}$$

$$\frac{\partial P}{\partial h_k(A_k = a_k)} = \frac{\frac{\partial Z(A_k = a_k)}{\partial h_k(A_k = a_k)} Z - \frac{\partial Z}{\partial h_k(A_k = a_k)} Z(A_k = a_k)}{Z^2} \tag{A.87}$$

$$= \frac{-Z(A_k = a_k) Z + Z(A_k = a_k)^2}{Z^2} \tag{A.88}$$

$$= - P(A_k = a_k) + P(A_k = a_k)^2 \tag{A.89}$$

$$= P(A_k = a_k)[P(A_k = a_k) - 1]. \tag{A.90}$$

The derivative for P_2 is similar.

By assuming that there are no couplings between amino acids, $\mathbf{J}_{ij} = \mathbf{0}$, we can calculate what the $\{h_i\}$ parameters would be under this independent site approximation to generate an initial guess for iterative fitting procedure.

$$P(A_k = a_k) = \frac{e^{-\beta h_k(A_k = a_k)}}{\sum_{A_k = \{1 \ldots k\}} e^{-\beta h_k(A_k}} = \frac{e^{-\beta h_k(A_k = a_k)}}{Z_k} \tag{A.91}$$

$$\ln P(A_k = a_k) = - \beta h_k(A_k = a_k) - \ln Z_k \tag{A.92}$$

$$h_k(A_k = a_k) = - \frac{1}{\beta} \ln P(A_k = a_k) - \frac{1}{\beta} \ln Z_k \tag{A.93}$$

$$h_k(A_k = a_k) = - \frac{1}{\beta} \ln P(A_k = a_k) + C \tag{A.94}$$

Where C is an additive constant that will be eliminated by gauge fixing $h_k(A_k = a_k) \leftarrow h_k(A_k = a_k) - h(A_k = 1)$. We initialize the $\{J_{ij}\}$ to zero. We observe that more sophisticated

initialization procedures (e.g., mean field approximations, post-mean field approximations such as solving the Thouless-Anderson-Palmer (TAP) equations [227] are also possible.

Appendix B

Code

The code for inferring fittness landscapes can be found here: https://github.com/GregoryRHart/Potts_fitting.git. It is a C++ code that can be run serially or compiled with OpenMP to be run in parallel. There is also an option to compile with CUDA for GPU acceleration.

The C++ code and accompanying MATLAB scripts for the immune simulator can be found here: https://github.com/GregoryRHart/Population_Dynamics.git. This code can be run serially or compiled with OpenMP for parallelization.

Appendix C

Addition Data and Figures From Static Design

C.1 Average immune pressure estimate

It has previously been shown that the effects of host immune pressure on strain distribution is averaged out if the MSA samples a genetically diverse host population [12, 106, 137]. In other words, the fact that hosts with different haplotypes target very different regions of the viral proteome means that if the number of sequences in the ensemble used to fit the model are sufficiently numerous and drawn from hosts with diverse immunological haplotypes, no single position in the viral proteome is subjected to a disproportionate mutational pressure when averaged over the sequence ensemble. The sequences where collected in multiple locales in nine countries on three continents [168–170], and while the majority of patients were Caucasian, there exists a strong representation of Africans as well as Hispanics, Asians, and other ethnic groups. This geographic and ethnic diversity suggests that the MSA contains sufficient genetic diversity to eliminate signatures of adaptive immunity. To quantify this assertion we estimated the frequency with which each amino acid position in NS5B is expected to be subject to immune pressure using the approach detailed in ref. [106]. We compiled from the Immune Epitope Database (`http://www.iedb.org`) the 24 CTL NS5B epitopes that were both exactly defined and the HLA association known [179], and determined the frequency with which each HLA occurs within the North American population as a representative group (`http://www.ncbi.nlm.nih.gov/projects/gv/mhc`) [184]. Finally, we estimated the probability that persons possessing these HLA types recognize and target the cognate CTL as the mean of the reported non-zero recognition frequencies of CTL

epitopes across the HCV proteome [228]. We observe that this provides a conservative estimate higher than any value reported for a NS5B epitope. Using these values, we estimated no position within the NS5B protein to be targeted by more than 8% of the population. This percentage is substantially lower than the values of 17% and 23% estimated for the HIV-1 proteins p17 and p24, for which we have previously computed fitness landscapes and validated in extensive comparisons against experimental and clinical data to demonstrate that they are not contaminated with signatures of adaptive immune pressure [12, 106]. In sum, our targeting frequency estimate for NS5B, prior empirical, numerical, and theoretical studies [106], and direct comparisons against clinical data and experimental measurements described in chapter 3, all provide support that our inferred NS5B fitness landscape does not contain footprints of adaptive immunity.

C.2 Model augmentation

The Potts model fitted to the MSA data detailed in chapter 3 contained parameters describing the fitness impact of each amino acid residue in each single position, and each pair of amino acids in each pair of positions. In comparing our model predictions against clinical and experimental data, we twice encountered a situation in which the experimentally reported viral strains contained amino acid residues absent in our MSA and therefore not contained within our model. Rather than simply discarding these sequences from our comparisons, we constructed two separate augmented models containing parameters for amino acid variants that were unobserved in the MSA.

The first augmentation was required in order to fully compare our model predictions with the measured *in vitro* fitness data in section 3.3.1. Of the 31 *in vitro* measurements we collated from the literature, 30 of them – all from the same lab [4, 5] – used the H77 sequence (GenBank Accession No. M67463) as their wild type baseline sequence, rather than the more commonly used H77S.3 sequence (GenBank Accession No. AF011751). Our fitted

Potts model contained all amino acids in H77S.3, but did not contain parameters for six residues in H77: K2469, A2512, L2637, R2703, R2715, and W2925. To assign energies to all strains considered, we augmented our model with parameters for these six unobserved residues to generate Augmented Model I.

The second augmentation was required to assign energies to all of the clinically observed escape, and compensatory, mutations that we analyzed in section 3.3.2, and the sequences in the longitudinal studies considered in section 3.3.4. The sequences from the longitudinal studies contained 456 residues in particular positions not observed in our MSA. The clinical escape mutations contained two amino acids that were not contained in our MSA, and which were coincident with two of the 456 unobserved amino acids within the longitudinal data. Accordingly, in order to assign energies to all longitudinal strains and escape mutations we augmented our model with parameters for the 456 unobserved residues to generate Augmented Model II.

To perform model augmentation it is necessary to incorporate h_i and J_{ij} parameters for each unobserved residue in position i. To estimate values of these model parameters, we specified the probability with which the unobserved amino acids appear within the MSA to be non-zero using pseudo-counts [144, 171]. This procedure adds a (possibly non-integer) number of fictitious observations of the amino acid within the MSA to reflect the belief that the probability of observing this amino acid in this position is not precisely zero, but rather a low-probability event that is not observed within the finite number of strains within the MSA. From a Bayesian perspective, the use of pseudo-counts may be considered the incorporation of prior knowledge into the model inference procedure [171], in this case the prior knowledge that the probability that these amino acids exist within an HCV strain should be non-zero. In this work, we specify the pseudo-count modified probability of observing amino acid A in position i, $P_i(A)$, as,

$$P_i(A) = \frac{1}{\lambda + N}\left(\frac{\lambda}{q_i} + \sum_{k=1}^{N} \delta_{A,z_i^k}\right), \tag{C.1}$$

where N is the number of sequences in our MSA, q_i is the number of distinct amino acids (including this unobserved amino acids added to the model) at position i, z_i^k is the identity of the amino acid in position i in sequence k of the MSA, δ_{A,z_i^k} is an indicator function that is unity when $A = z_i^k$ and zero otherwise, and λ is a pseudo-count [144, 229]. At positions where we supplement our model with unobserved amino acids, we choose $\lambda = \frac{q_i N}{N+1-q_i}$ such that for an unobserved amino acid A at that position, $P_i(A) = \frac{1}{N+1}$. This quantity may be interpreted as an estimated upper bound on the frequency with which amino acid A is observed in position i corresponding to supplementing the MSA with one (hypothetical) additional sequence containing amino acid A in position i. We then iteratively rescaled the two-position target probabilities, $P_{ij}(A, B)$, such that the marginal probabilities over i and j are consistent with the one-position target probabilities, $P_i(A)$ [12]. At positions where no unobserved amino acids were added, no pseudo-counts were added (i.e., $\lambda = 0$).

Having specified the pseudo-count modified probabilities, we re-fitted the parameters of the Potts model using the procedure detailed in section 3.2.2. Constituting a relatively small perturbation to the target probabilities for the fitting procedure, the model parameters changed very little from their unaugmented values, and by initializing the $\{h_i\}$ and $\{J_{ij}\}$ parameters to their unaugmented values, the fitting procedure quickly converged. The energy predictions of the augmented and unaugmented models are in close agreement, as illustrated by the correspondence of strain energies in figure 2 (augmented model) and figure C.3 (unaugmented model).

C.3 Predicted fitness costs of clinical escape mutations

In section 3.3.2 we looked at the energy cost (fitness cost) of documented escape mutations and where these costs fall in the spectrum of possible mutations. Three of the single mutations, K2471R, Q2467K, and R2937S, fall in the 36^{th}, 74^{th} and 96^{th} percentiles, respectively.

Two of the double mutations, R2937G/I2940T and Q2467K/K2471R fall in the 33^{rd} and 61^{st} percentiles, respectively. The relatively high energy (fitness) costs of these mutations can be rationalized by the fact that they are almost always observed in concert with compensatory mutations that place them far lower on the energy (fitness) cost spectrum. The details of those compensatory mutations follow.

K2471R (36^{th} percentile) and Q2467K (74^{th} percentile) are typically observed as the double mutant, Q2467K/K2471R (61^{st} percentile). Furthermore, they are almost always seen in connection with another mutation H2453Y which is compensatory and mediates further immune escape [3]. H2453Y is an escape mutation in a nearby epitope presented by the same HLA molecule. H2453Y/K2471R falls in the 20^{th} percentile of all double mutants with $\Delta E = 12.50$, and H2453Y/Q2467K falls in the 34^{th} percentile of all double mutants with $\Delta E = 14.05$. The triple mutant H2453Y/Q2467K/K2471R falls in the 35^{th} percentile of all triple mutants with $\Delta E = 21.42$.

We observe that the Q2467K/K2471R double mutant may represent a temporary "metastable" escape, since K2471R was observed to revert back to wild type after Q2467K was replaced by the Q2467L polymorphism. This alternative escape mutation has been reported to be less effective at mediating CTL escape, but is of much higher fitness (lower energy), falling in the 4^{th} percentile of the energy spectrum [3].

R2937S (96^{th} percentile) is very rare and is always observed to be accompanied by E2875K and P2881Q [4]. This E2875K/P2881Q/R2937S triple mutant falls in the 4^{th} percentile of all triple mutants and has an energy cost $\Delta E = 13.26$. R2937G/I2940T (33^{rd} percentile) is also very rare and always observed with E2875K and P2881Q [4].

C.4 Longitudinal clonal sequencing study

In section 3.3.4 we analyzed longitudinal sequencing data of HCV progression in Patient M003 and the two children to whom she gave birth during the study and vertically trans-

mitted HCV – Patients C003 and D003 – as reported in ref. [3]. For concision, we considered in chapter 3 only the average energy assigned by our model to the strains at each time point, $\bar{E}$, in our predictions of the fitness of the viral ensemble over the course of the study. Here we present a more detailed analysis of the clonal sequencing data for M003 (and her children C003 and D003), for whom multiple sequences are available for each time point.

Table C.5 presents, for each viral strain identified at each time point, the energy of the strain from by our model, and the sequence of the 6 epitopes B*15-LLRHHNMVY$_{2450-2458}$, B*15-SQRQKKVTF$_{2466-2474}$, A*02-RLIVFPDLGV$_{2578-2587}$, A*02-ALYDVVSKL$_{2594-2602}$, A*02-GLQDCTMVL$_{2727-2735}$, and A*31-VGIYLLPNR$_{3003-3011}$ for which M003 possesses the cognate HLA molecules, and position 2510 which is associated with an A*31 associated polymorphism (S2510N) [230]. As our reference sequence for this analysis we adopt the consensus sequence at the first time point of the study at 0.0 months.

M003 presented with acute HCV during her pregnancy with C003 at 0.0 months. Consistent with a suppressed immune system due to maternofatel immune tolerance, our model assigns high fitness (low energy) to the sequences retrieved at this time and at the time of delivery (1.3 months). The sequences reveal scattered mutations within epitopes, but all are transient and do not appear to be correlated with host immune pressure.

After delivery of C003 at 7.2 months our model predicts a sharp decrease in fitness (increase in energy) in all sequences reported in M003, that appears to be correlated with an increase in host immune pressure due to disappearance of the maternofetal immune tolerance mechanism after delivery of the infant. In particular, the sequence ensemble contains four immune related polymorphisms: H2453Y, Q2467X, K2471R, and S2510N. The H2453Y mutation appears in epitope B*15-LLRHHNLVY$_{2450-2458}$ and is known to abrogate T-cell recognition [3]. The mutations Q2467X and K2471R appear in epitope B*15-SQRQKKVTF$_{2466-2474}$. Although the specific information on all polymorphisms present is not available, Q2476L is reported to decrease T-cell recognition and Q2467K/K2471R to abolish it [3]. Our model predicts the energy cost of Q2467K ($\Delta E = 9.0$) to be more than

three times that of Q2467L ($\Delta E = 2.8$). The elimination of Q2467L in favor of Q2467K by month 10.7 is consistent with a scenario of mounting immune pressure on this epitope, and the fixation of a higher fitness cost escape mutation that more effectively abrogates T-cell recognition. The A*31 associated polymorphism S2510N also becomes fixed in M003 after delivery, consistent with her HLA haplotype.

As M003 enters a second pregnancy with D003 at month 8.3, our model again predicts an increase in the fitness of the viral ensemble (decrease in energy) consistent with suppression of specific HCV host immune pressure by the maternofetal immune tolerance mechanism. In addition to an increase in fitness (decrease in energy) of most sequences, by month 16.8, K2471R has reverted to wild type, and Q2467K/H to Q2467L in all but one sequence, consistent with a decrease in host immune pressure mediating the reversion of high fitness cost escape mutations. These reversions persisted for several months post-delivery of D003, but at 4.3 months after delivery (21.5 months) the rebounding of M003 host immune pressure induced the more costly, but more effective, escape mutations to arise again, with Q2467K/K2471R appearing in half of the clonal population.

The next ensemble of viral sequences from M003 are reported more than a year later, at 36.9 months. At this time all the sequences possess the wild type amino acid at position 2451 and a histidine residue at position 2467. Our model predicts the cost of Q2467H ($\Delta E = 4.0$) to be intermediate to that of Q2467L and Q2467K/K2451R. We suggest that Q2467H offers some immune escape and hence is tolerated over the more fit Q2467L. The final sequence data from M003 come after another year, at 49.0 months. Continuing to evolve under immune pressure a new polymorphism arises in half the sequences, Q2467T. In all cases Q2467T appears with K2571R. Our model predicts that Q2467T/K2451R ($\Delta E = 18.2$) is less fit than Q2467K/K2451R ($\Delta E = 16.3$); however we see a increase in the fitness of the strains with Q2467T/K2451R indicating that compensatory mutations arose to make it more fit.

None of the sequences vertically transmitted to either child C003 and D003 show significant differences from the maternal sequences at (or near) the time of birth. Sequences

within C003 25 weeks after birth (at 7.2 months) contain none of the escape mutations that arose within in the mother after delivery, consistent with transmission of a fit HCV strain from the immunosuppressed mother "outrunning" the nascent immune system of the newborn [3] . Sequences within D003 12 weeks after birth (at 20.1 months) contain the S2510N polymorphism and H2453Y and Q2467L polymorphisms within the HLA-B*15 associated epitopes $\text{LLRHHNLVY}_{2450-2458}$ and $\text{SQRQKKVTF}_{2466-2474}$ present in the mother before delivery, and constitute a slightly more fit population ($\overline{E} = 52.2$) of viral strains than those of the mother at time of delivery. D003 inherited the HLA-B*1501 class I molecule from the mother, and that reversion of the polymorphisms within these unfit strains is not observed is consistent with continued immune pressure at these epitopes by the child's immune system.

C.5 Figures

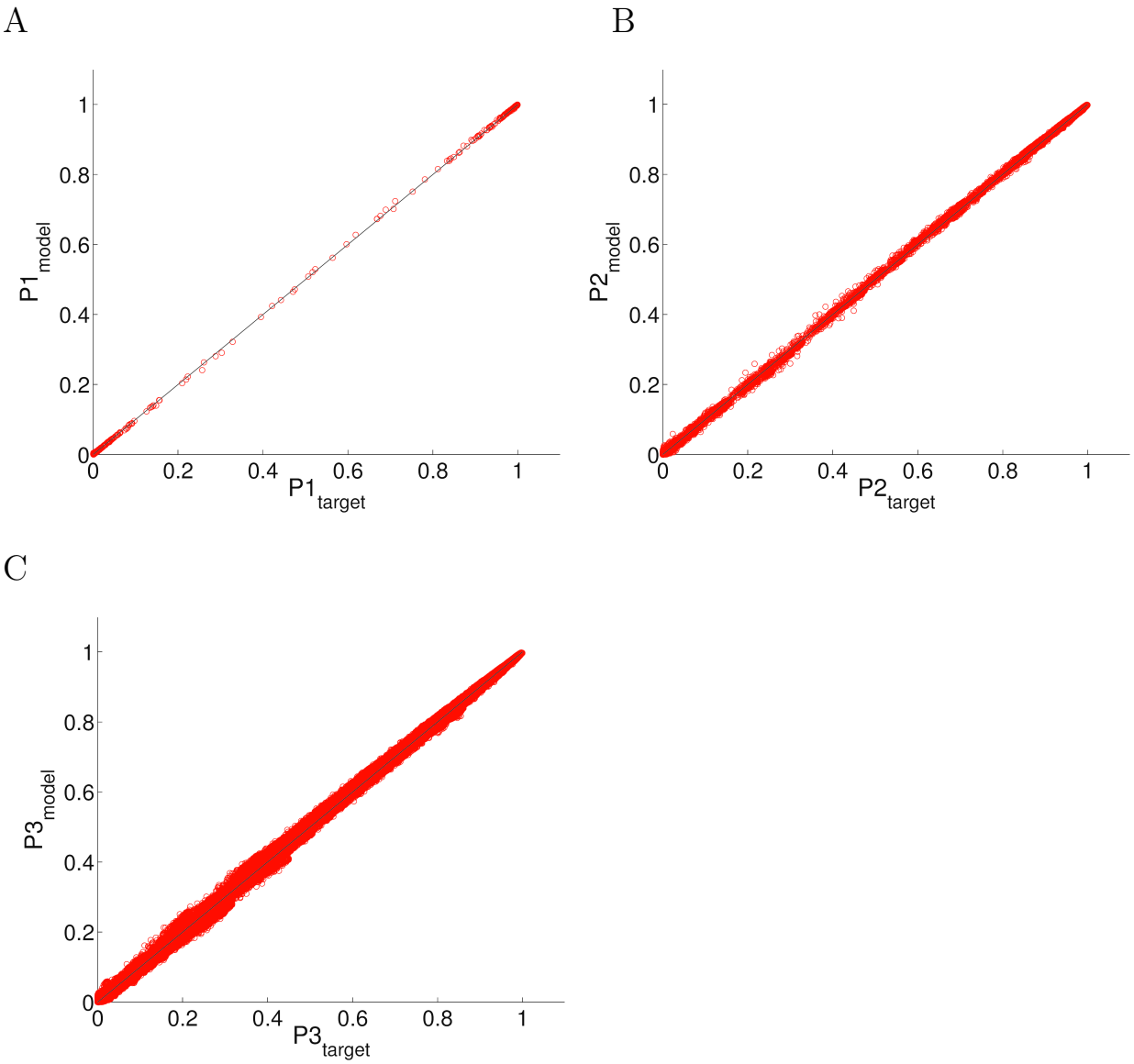


Figure C.1: Comparison for each amino acid, A_i, at each position, i, the (A) one-position, $P1(A_i)$, (B) two-position, $P2(A_i, A_j)$, and (C) three-position, $P3(A_i, A_j, A_k)$, amino acid frequencies observed within the MSA, $P1_{target}$, to those computed by the fitted Potts model, $P1_{model}$, by performing 99,990 rounds of Monte-Carlo sampling from the model (cf. ref. [12]). The parameters of the model were explicitly fitted to reproduce the one and two-position frequencies and so are expected to reproduce the observed mutational frequencies. That the model also predicts the three-position amino acid frequencies observed within the MSA demonstrates that our model *predicts* higher order mutational correlations within its effective one and two-position interaction parameters.

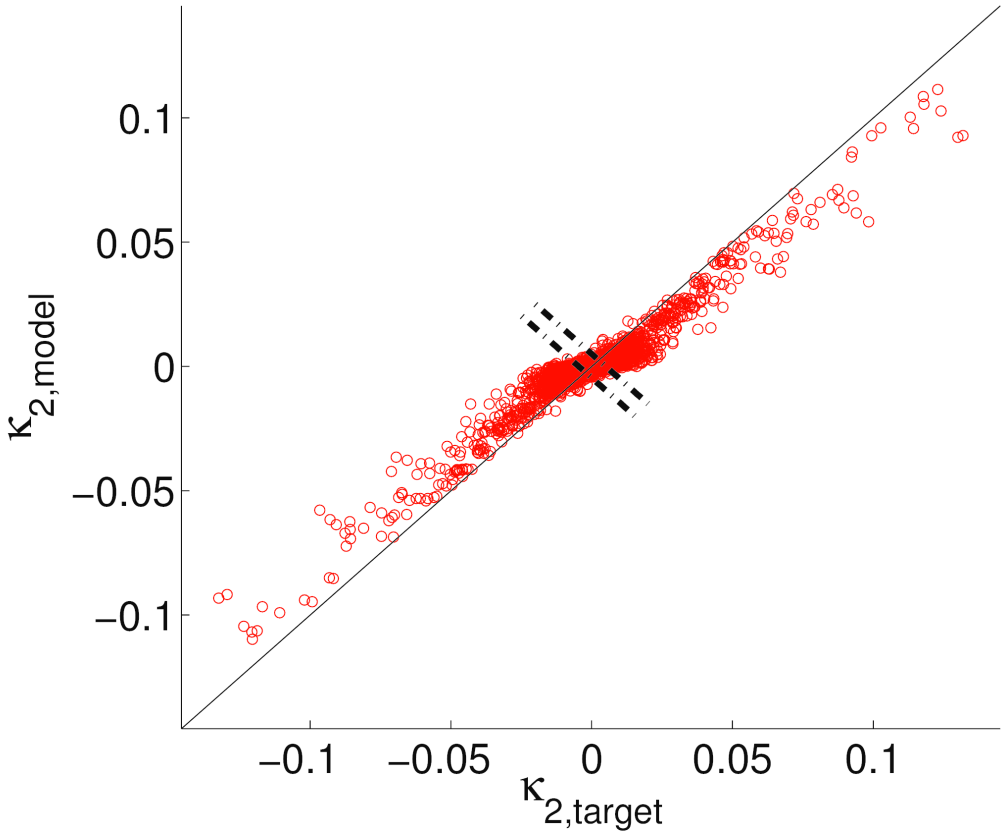


Figure C.2: Comparison of the second order cumulants for each pair of amino acids, A_i and A_j, at each pair of positions, $\kappa_2(A_i, A_j) = P2(A_i, A_j) - P1(A_i)P1(A_j)$ observed within the MSA, $\kappa_{2,\,target}$, to those computed by the fitted Potts model by performing 99,990 rounds of Monte-Carlo sampling from the model, $\kappa_{2,\,model}$, (cf. ref. [12]). κ_2 measures the difference between actual two-position probability of observing a particular pair of amino acids at a particular pair of positions and the two-position probability that would be expected if the two positions were mutationally uncoupled. $\kappa_2 \in [-0.25, 0.25]$, where κ_2 ¿ 0 indicates that the mutations are correlated, $\kappa_2 = 0$ uncorrelated, and κ_2 ¡ 0 anti-correlated. To define a statistically-significant correlation, we performed 10 independent scrambles of the columns of the MSA to randomize the amino acids located in each position of the protein and artificially break mutational correlations. The dashed lines in the plot indicate the 0.5^{th} and 99.5^{th} percentiles of the observed distribution of κ_2 under this permutation test – $\kappa^{0.5\%}_{2,\,model} = -2.3 \times 10^{-3}$ and $\kappa^{99.5\%}_{2,\,model} = 2.5 \times 10^{-3}$ – presenting an empirical measure of the expected range of κ_2 in the absence of mutational correlations and defining a 1% significance level for measured values of κ_2. The distribution of $\kappa_{2,\,target}$ indicates that while most mutational pairs are relatively uncorrelated, there are a significant number of strongly correlated and anti-correlated mutations, reflecting the presence of important epistatic effects within the protein. Furthermore, the clustering of the data around the diagonal indicates that our model captures these epistatic effects.

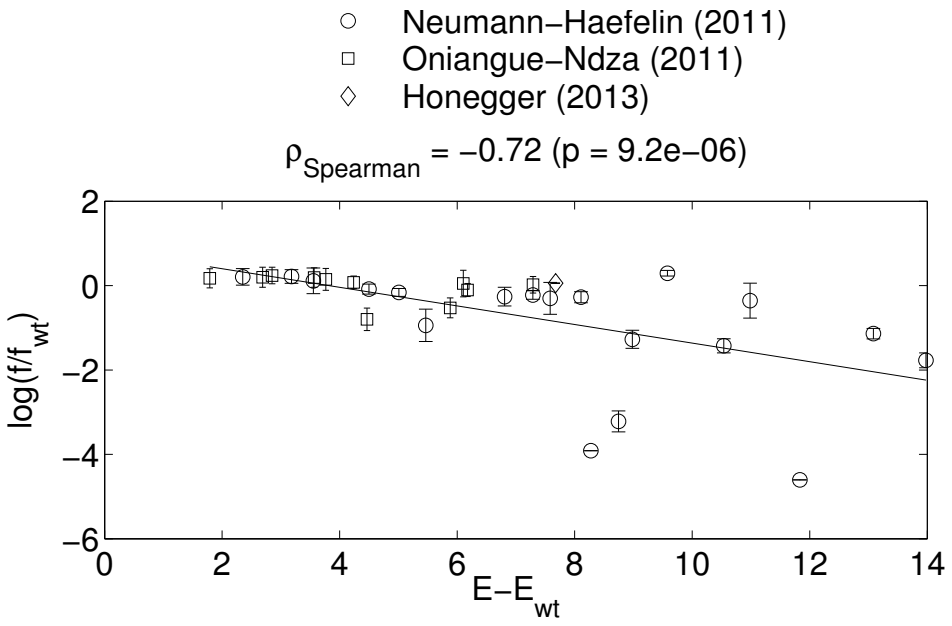
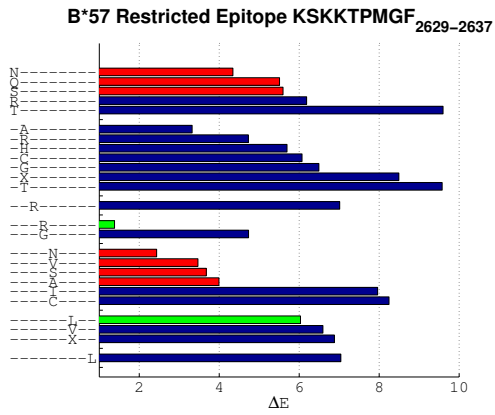
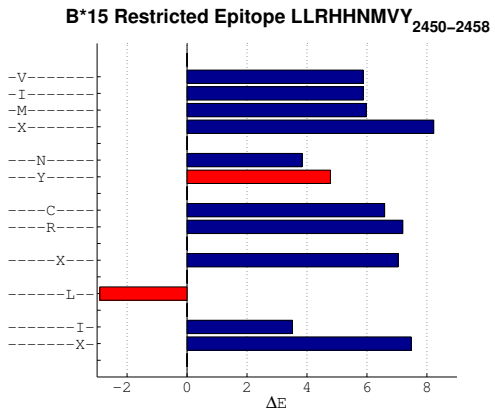


Figure C.3: Comparison of the *in vitro* replicative fitness relative to wild type, f/f_{wt}, measured for 31 engineered NS5B mutants containing up to four polymorphisms [3–5] against the energy relative to H77S.3 reference sequence, $(E - E_{wt})$, of each strain predicted by our unaugmented model. A strong and statistically significant negative correlation, $\rho_{Spearman} = -0.72$ ($p = 9.2 \times 10^{-6}$), validates our fitted model as a good predictor of intrinsic viral fitness. A linear least-squares fit is provided to guide the eye, and error bars delineate estimated uncertainties in the measured relative fitness.

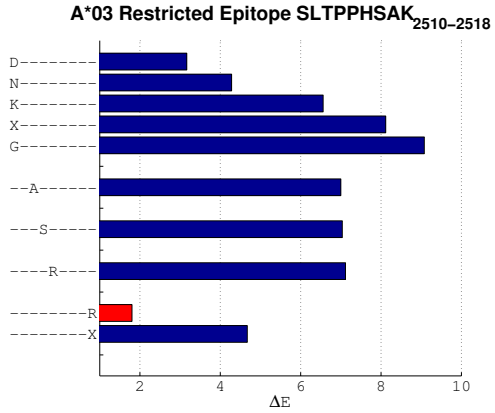
A



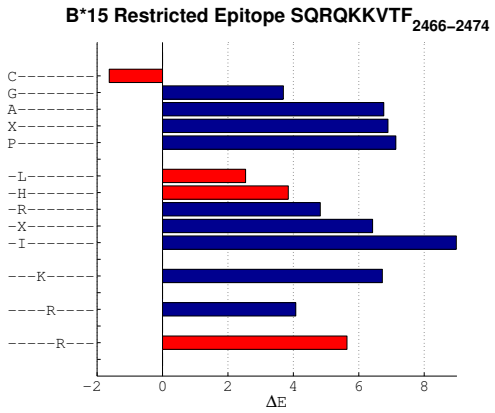
B



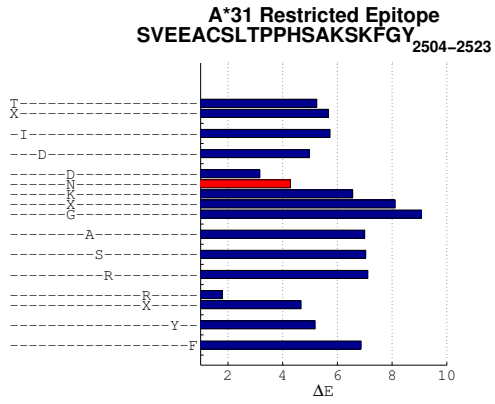
C



D



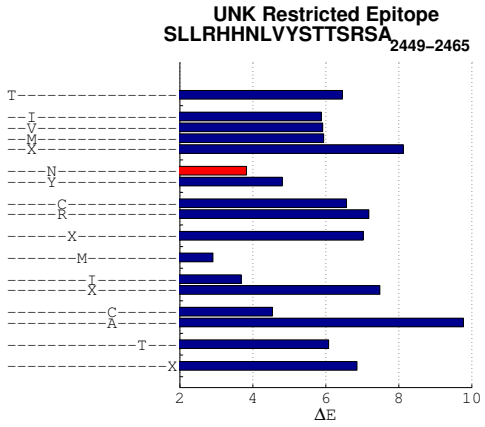
E



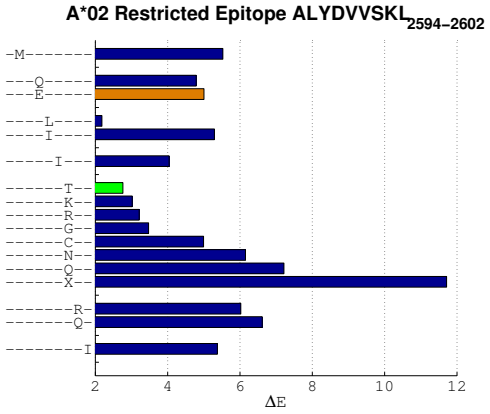
F

Continued on next page

G



H



I

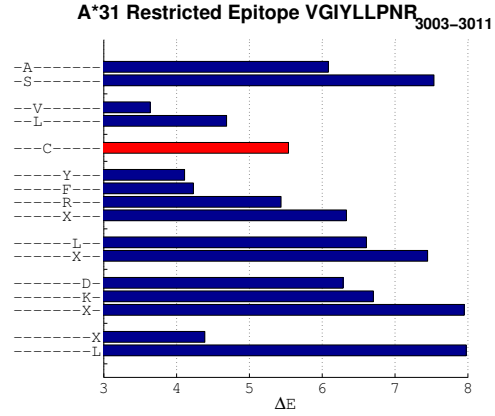
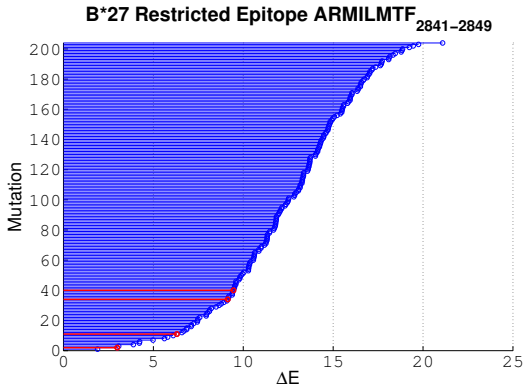


Figure C.4: Comparison of the energy costs (fitness penalties) relative to the H77 wild type reference sequence predicted by our model for all polymorphisms observed within our MSA occurring within the nine indicated CTL epitopes. All nine epitopes possess single amino acid mutations known to confer CTL escape. The energy cost associated with each single mutation, ΔE, is along the abscissa, and the mutations are shown along the ordinate. Dashes indicate unmutated positions, and letters the mutant amino acid residue. The letter X indicates an unknown amino acid type that was inconclusively identified by experimental sequencing within the ensemble sequences constituting the MSA used to fit our model. The greater the energy cost, the higher the fitness penalty. Polymorphisms possessing negative ΔE values are those predicted to *elevate* fitness relative to the H77 reference sequence. Red bars denote documented escape mutations that abrogate CTL recognition, green bars denote cross-reactive mutations that do not mediate escape, brown bars denote polymorphisms that have been reported both as escapes and as cross-reactive, blue bars denote mutations for which no specific clinical information is available. In panels A-G, one or more of the first, second, or third least costly polymorphisms within the epitope corresponds to a documented escape mutation. In panel H the escape mutation has the ninth lowest cost, although we note that it remains disputed as to whether this polymorphism conveys escape or is cross reactive [6, 10, 13–16]. In panel I the escape mutation is the seventh lowest cost polymorphism.

A



B

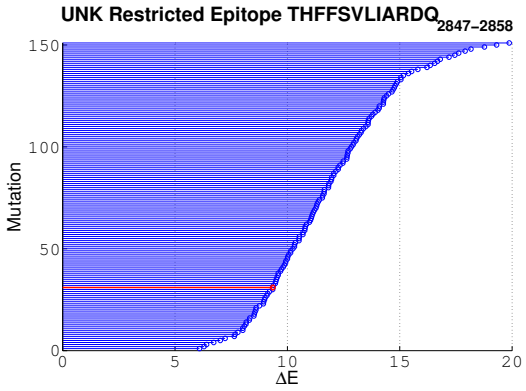


Figure C.5: Plots of the energy (fitness) cost of all double mutations observed within our MSA within the CTL epitopes (a) $\text{ARMILMTHF}_{2841-2849}$ and (b) $\text{THFFSVLI-ARDQ}_{2847-2858}$. These two epitopes require at least two mutations in order to escape CTL pressure. The energy cost associated with each double mutation, ΔE, is along the abscissa, and the double mutations – indexed according to their energy cost – are shown along the ordinate; the greater the energy cost, the higher the fitness penalty. Red bars denote documented escape mutations. In panel A, the least costly of all double mutants predicted by our model corresponds to a clinically documented escape mutation. In panel B, the documented escape mutation lies in the bottom fifth of the energy spectrum of all double mutations, ranked as the 31/151 lowest energy double mutant.

C.6 Tables

EU781776	EU482846	EU781802	EU781779	EU781748	EU781781	EU781747	EU781771	EU256056	EU781749
EU781797	EU256044	EU781796	EU256055	EU781798	EU482840	EU256053	EU781788	EU482841	EU781793
EU781799	GQ149768	EU781751	EU781800	EU781787	EU781794	EU781795	EU781786	EU781782	EU781804
EU155276	EU781789	EU781790	EU781760	EU155277	EU781810	EU781780	EU781803	EU781752	EU781823
EU781783	EU781750	EU781768	EU781777	EU781774	EU482837	EU482847	EU260396	EU256040	EU482878
EU256039	EU862828	EU155247	EU256041	EU482843	EU482848	EU256058	EU256043	EU256051	EU255940
EU569723	EU155265	EU155270	EU862831	EU482882	EU155271	EU482844	EU155278	EU155275	EU155282
EU155248	EU482845	EU155249	EU256067	EU482842	EU155266	EU256052	EU155251	EU256060	EU482838
EU155267	EU256046	EU155252	EU256047	EU155272	EU256057	EU256048	EU256049	EU255941	EU256068
EU155268	EU155273	EU155283	EU155274	EU256050	EU255939	EU155269	EU255942	EU781770	EU781753
EU781772	EU781767	EU781763	EU781773	EU781785	EU862823	EU482831	EU155378	EU482873	EU155379
EU781809	EU256106	EU482832	EU155380	EU256107	EU862834	EF032890	EU255963	EU255964	EU255965
EU155338	EU255966	EU155339	EU255968	EU255969	EU862839	EU255970	EU234063	EU482889	EU255973
EU255974	EU234064	EU255975	EU255976	EU155340	EU482850	EU255977	EU255978	EU255979	EU255980
EU255981	EU155341	EU255982	EU255983	EU155342	EU255984	EU255985	EU255986	EU255987	EU255988
EU255989	EU255990	EU255991	EU234065	EU255992	EU155343	EU569722	EU155344	EU482852	EU862827
EU781821	EU781815	EU781820	EU781812	EU781817	EU781807	EU781814	EU781808	EU781824	EU781822
EU781818	EU781811	EU781764	EU781792	EU781765	EU781754	EU781758	EU781762	EU687193	EU687194
EU687195	EU781775	EU781757	EU781761	EU781766	EU781759	EU781756	EU781791	EU781784	EU781801
EU781778	EU781769	EU781755	EU155311	EU239713	EU155284	EU155312	EU155285	EU155286	EU155287
EU256096	EU155288	EU155289	EU256097	EU155321	EU256105	EU155322	EU155290	EU482834	EU155291
EU595697	EU482836	EU155292	EU250017	EU155319	EU155293	EU155320	EU155323	EU482876	EU155295
EU239716	EU239715	EU482835	EU155296	EU155297	EU155298	EU155309	EU155299	EU660387	EU155313
EU155310	EU155314	EU256104	EU255927	EU255943	EU255944	EU255945	EU255946	EU482853	EU255947
EU256069	EU255948	EU255928	EU256070	EU256071	EU255949	EU155346	EU255930	EU256072	EU155347
EU155348	EU482854	EU255951	EU155349	EU255952	EU255953	EU482855	EU482856	EU529681	EU255954
EU255955	EU155350	EU155351	EU660385	EU255956	EU255957	EU482872	EU482857	EU256074	EU255958
EU155353	EU255959	EU155354	EU155355	EU482858	EU482884	EU529676	EU256002	EU595698	EU256004
EU482865	EU256008	EU862841	EU781816	EU781819	EU256009	EU256010	EU256011	EU529677	EU256013
EU155239	EU660384	EU256014	EU256015	EU256017	EU256018	EU256019	EU256020	EU529678	EU256021
EU256022	EU482887	EU256023	EU155240	EU482868	EU482869	EU256024	EU155242	EU256094	EU482870
EU529679	EU482871	EU595699	EU256025	EU529680	EU256095	EU781813	EU256026	EU256027	EU256028
EU256029	EU155243	EU256030	EU255938	EU256031	EU155244	EU155245	EU256032	EU155246	EU256034
EU256035	EU256036	EU256037	EU256038	EU660383	EU155213	EU255993	EU255994	EU155214	EU155215
EU255995	EU255996	EU255931	EU255997	EU256086	EU255932	EU255998	EU256087	EU155216	EU255999
EU256003	EU155233	EU482861	EU256012	EU155236	EU256005	EU256006	EU255934	EU482862	EU255936
EU482863	EU155237	EU255937	EU482864	EU482866	EU482867	EU255935	EU256007	EU155238	FJ024087
FJ181999	FJ024274	FJ024275	FJ205867	FJ024276	FJ182000	FJ024278	FJ390399	FJ205868	FJ182001
FJ024280	FJ024281	FJ024282	FJ390394	FJ410172	FJ390395				

Table C.1: Los Alamos National Laboratory HCV database (http://www.hcv.lanl.gov) accession numbers of 386 of 412 sequences shared with us by Dr. Todd Allen (Harvard Medical School) which are identical to publicly available sequences that now appear in this database.

Mutations	Epitope	ΔE	%	References
D2597E	A*02 ALYDVVSKL$_{2594-2602}$	5.9	22.2	[16]
Q2729R	A*02 GLQDCTMLV$_{2727-2735}$	2.7	4.5	[177]
K2518R	A*03 SLTPPHSAK$_{2510-2518}$	1.7	2.5	[178]
S2510N	A*31 epitope incompletely defined	4.7	12.4	[3, 230]
Y3006C	A*31 VGIYLLPNR$_{3003-3011}$	6.5	27.8	[3]
H2453Y	B*15 LLRHHNMVY$_{2450-2458}$	5.2	15.6	[3, 176]
M2456L	B*15 LLRHHNMVY$_{2450-2458}$	0.0	0.1	[176]
S2466C	B*15 SQRQKKVTF$_{2466-2474}$	0.0	0.1	[176, 181, 231]
K2471R	B*15 SQRQKKVTF$_{2466-2474}$	7.2	36.1	[3]
Q2467H	B*15 SQRQKKVTF$_{2466-2474}$	4.0	7.8	[3]
Q2467K	B*15 SQRQKKVTF$_{2466-2474}$	9.0	74.4	[3]
Q2467L	B*15 SQRQKKVTF$_{2466-2474}$	2.8	4.6	[3, 176]
R2937S	B*27 GRAAICGKY$_{2936-2944}$	10.2	96.6	[4]
R2937K	B*27 GRAAICGKY$_{2936-2944}$	0.0	0.1	[4]
I2940T	B*27 GRAAICGKY$_{2936-2944}$	5.5	18.4	[4]
K2943R	B*27 GRAAICGKY$_{2936-2944}$	3.6	7.0	[4]
K2629N	B*57 KSKKTPMGF$_{2629-2637}$	4.6	11.2	[5]
K2629Q	B*57 KSKKTPMGF$_{2629-2637}$	6.8	31.6	[5]
K2629S	B*57 KSKKTPMGF$_{2629-2637}$	5.8	21.6	[5]
T2633A	B*57 KSKKTPMGF$_{2629-2637}$	4.8	13.0	[5]
T2633S	B*57 KSKKTPMGF$_{2629-2637}$	4.4	9.8	[5]
T2633V	B*57 KSKKTPMGF$_{2629-2637}$	4.2	9.0	[5]
T2633N	B*57 KSKKTPMGF$_{2629-2637}$	3.1	5.1	[5, 8]
H2453N	unknown SLLRHHNLVYSTTSRSA$_{2449-2465}$	4.6	10.9	[6]
Q2467K/K2471R	B*15 SQRQKKVTF$_{2466-2474}$	16.3	61.4	[3]
A2841V/I2844V	B*27 ARMILMTHF$_{2841-2849}$	7.3	3.7	[11]
A2841V/M2846L	B*27 ARMILMTHF$_{2841-2849}$	4.0	2.1	[11]
M2846L/T2847P	B*27 ARMILMTHF$_{2841-2849}$	10.5	10.1	[11]
I2844V/T2847P	B*27 ARMILMTHF$_{2841-2849}$	10.8	11.1	[11]
R2937K/I2940V	B*27 GRAAICGKY$_{2936-2944}$	5.2	2.5	[4]
R2937G/I2940T	B*27 GRAAICGKY$_{2936-2944}$	14.0	33.5	[4]
F2849L/I2854M	unknown THFFSVLIARDQ$_{2847-2858}$	9.9	8.2	[6]
A2841V/I2844V/M2846L	B*27 ARMILMTHF$_{2841-2849}$	7.8	1.9	[11]
A2841V/I2844V/T2847S	B*27 ARMILMTHF$_{2841-2849}$	16.5	9.3	[11]
A2841V/M2846L/T2847P	B*27 ARMILMTHF$_{2841-2849}$	10.3	3.0	[11]

Table C.2: List of the 35 escape mutations analyzed in section 3.3.2, the associated epitope and HLA allele, energy of the mutant relative to the H77 wild type $\Delta E = (E - E_{wt})$, and the percentile within which the mutant is located on the energy spectrum of all possible mutants of the same order.

Index	Epitope	Position	HLA	$\Delta < E >$	Dominant	Protective	References
1	QPEKGGRKPA	2568-2577	B*55	7.06	N	N	[17, 18, 22]
2	KSKKTPMGF*	2629-2637	B*57	6.45	Y	Y	[11, 17, 22, 28, 30]
3	SPGEINRVAA	2898-2907	B*55	6.38	N	N	[17, 18]
4	RVCEKMALY	2588-2596	A*03	5.28	N	N	[30–32]
5	CYSIEPLDL	2871-2879	A*24:02	4.97	N	N	[33]
6	LGVPPLRAWR	2912-2921	B*57	4.71	N	Y	[17, 21]
7	VGIYLLPNR	3003-3011	A*31	4.51	N	N	[30]
8	RMILMTHFF	2842-2850	A*24:02	4.50	N	N	[33]
9	TARHTPVNSW*	2819-2828	A*25	3.88	N	N	[17, 34]
10	ARHTPVNSW	2820-2828	B*27, B*27:02	3.85	N	Y	[35, 36]
11	ARMILMTHF	2841-2849	B*27, B*27:01, B*27:02	3.85	Y	Y	[19, 22, 29, 35, 36]
12	APTLWARMVL*	2836-2845	B*07	3.67	N	N	[22, 37]
13	HDGAGKRVYYL	2794-2804	B*38	3.47	N	N	[30]
14	HDGAGKRVY	2794-2802	A*03	3.47	N	N	[32]
15	GRAAICGKY	2936-2944	B*27, B*27:02	3.24	N	Y	[10, 35, 36]
16	ALYDVVTKL*	2594-2602	A*02, A*02:01	3.16	Y	N	[17–22]
17	RLIVFPDLGV	2578-2587	A*02	2.89	N	N	[38, 39]
18	GLQDCTMLV	2727-2735	A*02, A*02:01	1.15	N	N	[18, 23, 32, 39]
19	SVRARLLSR	2926-2934	A*03	1.11	N	N	[22]
20	SLTPPHSAK	2510-2518	A*03	0.83	N	N	[30]
21	VYSTTSRSASL	2457-2467	A*24:02	-0.06	N	N	[40]
22	SQRQKKVTF	2466-2474	B*15	-0.07	N	N	[15, 25, 37]
23	LLRHHNMVY	2450-2458	B*15	-0.09	N	N	[15, 25, 37]
24	SYTWTGALI	2423-2431	A*24:02	-0.12	N	N	[41]

Table C.3: List of the 24 NS5B CTL epitopes discussed in section 3.3.5 which are precisely mapped and for which the restricting allele is known. The index reported in the first column corresponds to the indices reported in the "Epitopes" column in table C.4. The asterisk character in the second column indicates those epitopes for which some mutations are known to be cross reactive. In calculating the cost of escape for these epitopes we eliminated under our simulated CTL targeting procedure detailed in section 3.3.5 all strains reported to be antigenic. We also report the change in the average energy of a strain in the population, $\Delta\langle E\rangle$, upon eliminating those strains with wild type epitopes, whether the epitope is reported to be immunodominant, and if the associated HLA allele is correlated with protection.

Index	# Components	$\overline{\Delta\langle E\rangle}$	Population Coverage	Epitopes
1	1	1.1	35.0%	16
2	2	2.1	35.0%	16 17
3	2	1.9	50.1%	4 16
4	3	2.9	50.1%	4 16 17
5	3	2.1	54.4%	4 9 16
6	3	2.6	52.9%	4 12 16
7	3	2.3	54.4%	2 4 16
8	4	3.0	54.4%	4 9 16 17
9	4	3.6	52.9%	4 12 16 17
10	4	3.3	54.4%	2 4 16 17
11	4	2.8	57.2%	4 9 12 16
12	4	2.5	58.7%	2 4 9 16
13	4	3.0	57.2%	2 4 12 16
14	5	3.8	57.2%	4 9 12 16 17
15	5	3.5	58.7%	2 4 9 16 17
16	5	4.4	52.9%	4 12 14 16 17
17	5	4.0	57.2%	2 4 12 16 17
18	5	3.2	61.4%	2 4 9 12 16
19	5	2.6	61.9%	2 4 9 13 16
20	6	4.8	52.9%	4 12 14 16 17 18
21	6	4.5	57.2%	4 9 12 14 16 17
22	6	4.2	61.4%	2 4 9 12 16 17
23	6	3.6	61.9%	2 4 9 13 16 17
24	6	4.8	57.2%	2 4 12 14 16 17
25	6	3.3	64.7%	2 4 9 12 13 16
26	7	5.0	57.2%	4 9 12 14 16 17 18
27	7	5.2	57.2%	2 4 12 14 16 17 18
28	7	3.4	66.7%	2 4 5 9 12 13 16
29	7	4.9	61.4%	2 4 9 12 14 16 17
30	7	4.3	64.7%	2 4 9 12 13 16 17
31	7	3.4	67.1%	2 4 7 9 12 13 16
32	8	5.4	61.4%	2 4 9 12 14 16 17 18
33	8	5.6	57.2%	2 4 6 12 14 16 17 18
34	8	4.4	66.7%	2 4 5 9 12 13 16 17
35	8	3.5	69.2%	2 4 5 7 9 12 13 16
36	8	5.0	64.7%	2 4 9 12 13 14 16 17
37	8	4.4	67.1%	2 4 7 9 12 13 16 17

Continued on next page

38	9	5.5	64.7%	2 4 9 12 13 14 16 17 18
39	9	5.8	61.4%	2 4 6 9 12 14 16 17 18
40	9	5.9	57.2%	2 4 6 12 14 16 17 18 19
41	9	5.2	66.7%	2 4 5 9 12 13 14 16 17
42	9	4.5	69.2%	2 4 5 7 9 12 13 16 17
43	9	3.5	71.0%	2 4 5 7 9 12 13 16 22
44	9	3.5	69.3%	1 2 4 5 7 9 12 13 16
45	9	5.2	67.1%	2 4 7 9 12 13 14 16 17
46	10	5.6	66.7%	2 4 5 9 12 13 14 16 17 18
47	10	6.0	61.4%	2 4 6 9 12 14 16 17 18 19
48	10	5.6	67.1%	2 4 7 9 12 13 14 16 17 18
49	10	5.9	64.7%	2 4 6 9 12 13 14 16 17 18
50	10	5.3	69.2%	2 4 5 7 9 12 13 14 16 17
51	10	4.5	71.0%	2 4 5 7 9 12 13 16 17 22
52	10	4.6	69.3%	1 2 4 5 7 9 12 13 16 17
53	10	3.5	71.2%	1 2 4 5 7 9 12 13 16 22
54	11	5.7	69.2%	2 4 5 7 9 12 13 14 16 17 18
55	11	6.1	66.7%	2 4 5 6 9 12 13 14 16 17 18
56	11	6.2	64.7%	2 4 6 9 12 13 14 16 17 18 19
57	11	6.0	67.1%	2 4 6 7 9 12 13 14 16 17 18
58	11	5.3	71.0%	2 4 5 7 9 12 13 14 16 17 22
59	11	5.3	69.3%	1 2 4 5 7 9 12 13 14 16 17
60	11	4.6	71.2%	1 2 4 5 7 9 12 13 16 17 22
61	12	6.3	66.7%	2 4 5 6 9 12 13 14 16 17 18 19
62	12	5.7	71.0%	2 4 5 7 9 12 13 14 16 17 18 22
63	12	5.7	69.3%	1 2 4 5 7 9 12 13 14 16 17 18
64	12	6.2	69.2%	2 4 5 6 7 9 12 13 14 16 17 18
65	12	6.3	67.1%	2 4 6 7 9 12 13 14 16 17 18 19
66	12	5.3	71.2%	1 2 4 5 7 9 12 13 14 16 17 22
67	13	6.4	69.2%	2 4 5 6 7 9 12 13 14 16 17 18 19
68	13	5.7	71.2%	1 2 4 5 7 9 12 13 14 16 17 18 22
69	13	6.2	71.0%	2 4 5 6 7 9 12 13 14 16 17 18 22
70	13	6.2	69.3%	1 2 4 5 6 7 9 12 13 14 16 17 18
71	14	6.5	69.2%	2 4 5 6 7 8 9 12 13 14 16 17 18 19
72	14	6.4	71.0%	2 4 5 6 7 9 12 13 14 16 17 18 19 22
73	14	6.4	69.3%	1 2 4 5 6 7 9 12 13 14 16 17 18 19
74	14	6.2	71.2%	1 2 4 5 6 7 9 12 13 14 16 17 18 22

Continued on next page

75	15	6.5	69.2%	2 4 5 6 7 8 9 10 12 13 14 16 17 18 19
76	15	6.5	71.0%	2 4 5 6 7 8 9 12 13 14 16 17 18 19 22
77	15	6.5	69.3%	1 2 4 5 6 7 8 9 12 13 14 16 17 18 19
78	15	6.4	71.2%	1 2 4 5 6 7 9 12 13 14 16 17 18 19 22
79	16	6.5	71.0%	2 4 5 6 7 8 9 10 12 13 14 16 17 18 19 22
80	16	6.5	71.2%	1 2 4 5 6 7 8 9 12 13 14 16 17 18 19 22
81	16	6.5	69.3%	1 2 3 4 5 6 7 8 9 12 13 14 16 17 18 19
82	17	6.5	69.3%	1 2 3 4 5 6 7 8 9 10 12 13 14 16 17 18 19
83	17	6.5	71.2%	1 2 3 4 5 6 7 8 9 12 13 14 16 17 18 19 22
84	18	6.5	69.3%	1 2 3 4 5 6 7 8 9 10 11 12 13 14 16 17 18 19
85	18	6.5	71.2%	1 2 3 4 5 6 7 8 9 10 12 13 14 16 17 18 19 22
86	19	6.5	71.2%	1 2 3 4 5 6 7 8 9 10 11 12 13 14 16 17 18 19 22

Table C.4: List of the 86 optimal immunogen candidates residing on the Pareto frontier of figure 3.6. The number of components corresponds to the number of epitopes in the immunogen candidate, the population coverage is the fraction of the target population of the the top 66 haplotypes of North Americans who respond to at least one epitope in the immunogen candidate, and $\overline{\Delta\langle E\rangle}$ is the weighted averaged impact of the immunogen upon the fitness of the viral ensemble within the target population as defined in section 3.3.6. The particular epitopes in the immunogen candidate reported in the last column correspond to the indices in the "Index" column in table C.3.

Months	E	2450 2458 LLRHHNLVY	...	2466 2474 SQRQKKVTF	...	2510 S	...	2578 2587 RLIVFPDLGV	...	2594 2602 ALYDVVSKL	...	2727 2735 GLQDCTMLV	...	3003 3011 VGIXXXXXX
0.0	10.6	....		....		.		....		....		....		....
	10.6	....		....		.		....		....		....		....
	15.7	....		....		.		....		....		....		....
	18.9	....		....		.		....		....		....		....
	22.1	....		....		.		....		....		....		....
	23.1	....		....		N		....		....		....		....
	24.9	....		....		.		....		....		....		....
	25.3	....		....		.		....		....		....		....
	27.4	....		....		.		....		....		....		....
	33.8	....		....		.		....		....		....		....
	36.8	....		....		.		....		..H......		....		....
	39.9	....		....		.		....		....		....		....
	40.5	....		....R....		.		....		....		....		....
	49.7	....		....		.		....		....		....		....
	53.4	..P......		.......S.		.		....		....		....		....
1.3	10.9	....		....		.		....		....		....		....
	19.5	....		....		.		.....H....		....		....		....
	19.7	....		....		.		....		....		....		....
	23.7	....		....		.		....		....		....		....
	27.5	....		....		.		....		....		....		....
	35.5	....		....		.		....		....		....		....
	43.1	....		....		.		....S.....		....		....		....
	52.0	..P......		....		.		....		....		....		....
	58.3	........H		....		.		....		......G.P		....		....
7.2	51.4	...Y.....		.H...R...		N		....		....		....		....
	57.0	...Y.....		.L...R...		.		....		....		....		....
	63.1	...Y.....		.K...R...		N		....		....		.......P.		....
	65.1	...Y.....		.H...R...		N		....		....		....		....
	68.9	...Y.....		.L...R..L		N		....		....		....		....
	104.6	...Y.....		.K...R...		N		....		....		....		....
10.7	69.7	...Y.....		.K...R...		N		....		....		....		....
	77.7	...Y.....		.K...R...		N		....		....		....		....
	84.0	...Y.....		.K..ER...		N		....		....		....		....
	84.0	...Y.....		.K..ER...		N		....		....		....		....
	87.4	...Y.....		.H...R...		N		....		....		....		E........
	101.2	...Y.....		.H...R...		N		....		....		....		..V......
	115.9	...Y.....		.H...R...		N		.......P..		....		......T..		....
	120.0	...Y.....		.K...R...		N		....		....		....		....
	133.5	...Y.....		.K...R...		N		....		....		....		..V......

Continued on next page

Months	E	2450 2458 LLRHHNLVY	...	2466 2474 SQRQKKVTF	...	2510 S	...	2578 2587 RLIVFPDLGV	...	2594 2602 ALYDVVSKL	...	2727 2735 GLQDCTMLV	...	3003 3011 VGIXXXXXX
14.5	62.0	...Y.....		.K...R...		N		....		....		....		....
	65.6	...Y.....		.H...R...		N		....		....		....		....
	65.6	...Y.....		.K...R...		N		....		....		....		....
	69.3	...Y.....		.K...R...		N		....		....		....		....
	70.0	...Y....H		.K...R...		N		....		....		....		....
	70.9	...Y.....		.R...R...		N		....		....		....		....
	73.4	...Y.....		.H...R...		N		....		....		....		....
	80.1	...Y.....		.K...R...		N		....		....		....		....
16.8	31.2	...Y.....		.L.......		N		....		....		....		....
	38.0	...Y.....		.L.......		N		....		....		....		....
	46.6	...Y.....		.L.......		N		...A......		.P.......		....		....
	46.6	...Y.....		.L.......		N		...A......		.P.......		....		....
	55.1	...Y.....		.L.......		N		....		....		....		....
	57.1	...Y.....		.L..E....		N		....		....		....		....
	61.1	...Y.....		.K...R...		N		..L.......		....		....		....
	70.7	...Y.....		.L.......		N		....		....		....		..V......
	87.4	...Y.....		.L.......		N		....		....		....		....
	99.1	...Y.....		.L.......		N		....		....		....		....
18.2	29.1	...Y.....		.L.......		N		....		....		....		....
	36.0	...Y.....		.K...R...		N		....		....		....		....
	45.7	...Y.....		.L.......		N		....		.......R.		....		....
	58.4	...Y.....		.L.......		N		....		....		....		....
	66.2	...Y.....		.L.......		N		....		....		....		....
	80.6	...Y.....		.L.......		N		....		..H.A....		....		....
	81.1	...Y.....		.L.R...A.		N		....		....		....		....
	82.5	...Y.....		.L.......		N		....		....		....		....
	84.4	...Y.....		.L.......		N		....		....		....		....
	93.7	...Y.....		.L.......		N		....		....A....		....		....
21.5	28.0	...Y.....		.H.......		N		....		....		....		....
	31.1	...Y.....		.R.......		N		....		....		....		....
	31.8	...Y.....		.H.......		N		....		....		....		....
	47.2	...Y.....		.H.......		N		....		....		....		....
	48.0	...Y.....		.K...R...		N		....		....		....		....
	64.6	...Y.....		.K...R...		N		....		....		....		....
	71.2	...Y....H		.K...R...		N		....		....		....		....
	76.3	...Y.....		.R.......		N		....		....		....		....
	97.9	...Y.....		.KG..R...		N		....		....		....		....
36.9	76.6	...Y.....		.H.......		N		....V.....		....		....		....
	83.7	...Y.....		.H.......		N		....		....		....		....
	102.2	...Y.....		.H.......		N		....		..H......		.P.......		....
	102.6	...Y.....		.H.......		N		....		....		....		....
	122.1	...Y.....		.H.......		N		....		....		....		....

Continued on next page

Months	E	2450 2458 LLRHHNLVY	···	2466 2474 SQRQKKVTF	···	2510 S	···	2578 2587 RLIVFPDLGV	···	2594 2602 ALYDVVSKL	···	2727 2735 GLQDCTMLV	···	3003 3011 VGIXXXXXX
49.0	36.6	...Y.....		.H.....		N		.....		.....		.....		.....
	46.0	...Y.....		.T...R...		N		.....		.....		.....		.....
	46.5	...Y.....		.H.....		N		.....		.....		.....		.....
	46.6	...Y.....		.H.....		N		.....		.....		.....		.....
	46.7	...Y.....		.T...R...		N		.....		.....		.....		.....
	54.5	...Y.....		.T...R...		N		.....		.....		.....		.....
	54.5	...Y.....		.T...R...		N		.....		.....		.....		I.....
Child C003														
7.2	10.6	.....		.....		.		.....		.....		.....		.....
	10.6	.....		.....		.		.....		.....		.....		.....
	10.6	.....		.....		.		.....		.....		.....		.....
	19.4	.....		.....		.		.P.....		.....		.....		.....
	23.8	.....		.....		.		.....		.....		.....		.....
	32.0	.....		.....		.		.....		.....		.....A		.....
	34.2	.....		.....		.		.....		.....		.....		.....
	44.6	.....		.....		.		.....		.....		.....		.....
Child D003														
20.1	31.3	...Y.....		.L.....		N		.....		.....		.....		.....
	39.2	...Y.....		.L.....		N		.....		.....		.....		.....
	78.3	...Y.....		.L.....		N		....V.....		.....		.....		.V.....
	85.3	...Y.....		.L.....		N		.....		.....		.....		.....
	98.1	...Y.....		.L.....		N		.....		.....		.....		.....
	112.2	...Y.....		.L.....		N		.....		.....		.....		.X.....
	153.3	...Y.....		.L.....		N		.....		.....		.....		.V.....

Table C.5: The sequence of six epitopes (B*15-LLRHHNMVY$_{2450-2458}$, B*15-SQRQKKVTF$_{2466\text{-}2474}$, A*02-RLIVFPDLGV$_{2578-2587}$, A*02-ALYDVVSKL$_{2594\text{-}2602}$, A*02-GLQDCTMVL$_{2727-2735}$, and A*31-VGIYLLPNR$_{3003-3011}$) and one HLA associated polymorphism (S2510N) from M003 and her children (C003 and D003) with the energies assigned to the complete NS5B sequence by our model. The shaded regions of the table indicate periods of time during which M003 was pregnant with C003 and, subsequently, D003.

Bibliography

[1] S. Wright, "The roles of mutation, inbreeding, crossbreeding and selection in evolution," in *Proceedings of the sixth international congress on genetics*, 6 (1932) pp. 356–366.

[2] D. Seifert, F. Di Giallonardo, K. J. Metzner, H. F. Günthard, and N. Beerenwinkel, "A framework for inferring fitness landscapes of patient-derived viruses using quasispecies theory," Genetics **199**, 191 (2015).

[3] J. R. Honegger, S. Kim, A. A. Price, J. A. Kohout, K. L. McKnight, M. R. Prasad, S. M. Lemon, A. Grakoui, and C. M. Walker, "Loss of immune escape mutations during persistent HCV infection in pregnancy enhances replication of vertically transmitted viruses," Nature Medicine **19**, 1529 (2013), letter.

[4] C. Neumann-Haefelin, C. Oniangue-Ndza, T. Kuntzen, J. Schmidt, K. Nitschke, J. Sidney, C. Caillet-Saguy, M. Binder, N. Kersting, M. W. Kemper, K. A. Power, S. Ingber, L. L. Reyor, K. Hills-Evans, A. Y. Kim, G. M. Lauer, V. Lohmann, A. Sette, M. R. Henn, S. Bressanelli, R. Thimme, and T. M. Allen, "Human leukocyte antigen B27 selects for rare escape mutations that significantly impair hepatitis C virus replication and require compensatory mutations," Hepatology **54**, 1157 (2011).

[5] C. Oniague-Ndza, T. Kuntzen, M. Kemper, A. Berical, Y. E. Wang, C. Neumann-Haefelin, P. K. Foote, K. Hills-Evans, L. L. Reyor, K. Kane, A. D. Gladden, A. K. Bloom, K. A. Power, R. Thimme, G. M. Lauer, M. R. Henn, A. Y. Kim, and T. M. Allen, "Compensatory mutations restore the replication defects caused by cytotoxic T lymphocyte escape mutations in hepatitis C virus polymerase," Journal of Virology **85**, 11883 (2011).

[6] T. Kuntzen, J. Timm, A. Berical, L. L. Lewis-Ximenez, A. Jones, B. Nolan, J. S. zur Wiesch, B. Li, A. Schneidwind, A. Y. Kim, R. T. Chung, G. M. Lauer, and T. M. Allen, "Viral sequence evolution in acute hepatitis C virus infection," Journal of Virology **81**, 11658 (2007).

[7] C. Neumann-Haefelin, S. McKiernan, S. Ward, S. Viazov, H. C. Spangenberg, T. Killinger, T. F. Baumert, N. Nazarova, I. Sheridan, O. Pybus, F. von Weizsäcker, M. Roggendorf, D. Kelleher, P. Klenerman, H. E. Blum, and R. Thimme, "Dominant influence of an HLA-B27 restricted CD8+ T cell response in mediating HCV clearance and evolution," Hepatology **43**, 563 (2006).

[8] A. Y. Kim, T. Kuntzen, J. Timm, B. E. Nolan, M. A. Baca, L. L. Reyor, A. C. Berical, A. J. Feller, K. L. Johnson, J. S. Z. Wiesch, G. K. Robbins, R. T. Chung, W. B. D., M. Carrington, T. M. Allen, and G. M. Lauer, "Spontaneous control of HCV is associated with expression of HLA-B* 57 and preservation of targeted epitopes," Gastroenterology **140**, 686 (2011).

[9] S. Urbani, J. Uggeri, Y. Matsuura, T. Miyamura, A. Penna, C. Boni, and C. Ferrari, "Identification of immunodominant hepatitis C virus (HCV)-specific cytotoxic T-cell epitopes by stimulation with endogenously synthesized HCV antigens," Hepatology **33**, 1533 (2001).

[10] J. Schmidt, A. K. N. Iversen, S. Tenzer, E. Gostick, D. A. Price, V. Lohmann, U. Distler, P. Bowness, H. Schild, H. E. Blum, P. Klenerman, C. Neumann-Haefelin, and R. Thimme, "Rapid antigen processing and presentation of a protective and immunodominant HLA-B*27-restricted hepatitis C virus-specific CD8+ T-cell epitope," Public Library of Science Pathogens **8**, E1003042 (2012).

[11] E. Dazert, C. Neumann-Haefelin, S. Bressanelli, K. Fitzmaurice, J. Kort, J. Timm, S. McKiernan, D. Kelleher, N. Gruener, J. E. Tavis, H. R. Rosen, J. Shaw, P. Bowness, H. E. Blum, P. Klenerman, R. Bartenschlager, and R. Thimme, "Loss of viral fitness and cross-recognition by CD8+ T cells limit HCV escape from a protective HLA-B27–restricted human immune response," Journal of Clinical Investigation **119**, 376 (2009), hCV fitness measurements.

[12] A. L. Ferguson, J. K. Mann, S. Omarjee, T. Ndung'u, B. D. Walker, and A. K. Chakraborty, "Translating HIV sequences into quantitative fitness landscapes predicts viral vulnerabilities for rational immunogen design," Immunity **38**, 606 (2013).

[13] G. M. Lauer, E. Barnes, M. Lucas, J. Timm, K. Ouchi, A. Y. Kim, C. L. Day, G. K. Robbins, D. R. Casson, M. Reiser, G. Dusheiko, T. M. Allen, R. T. Chung, B. D. Walker, and P. Klenerman, "High resolution analysis of cellular immune responses in resolved and persistent hepatitis C virus infection," Gastroenterology **127**, 924 (2004).

[14] A. L. Cox, T. Mosbruger, G. M. Lauer, D. Pardoll, D. L. Thomas, and S. C. Ray, "Comprehensive analyses of CD8+ T cell responses during longitudinal study of acute human hepatitis C," Hepatology **42**, 104 (2005).

[15] H. C. Spangenberg, S. Viazov, N. Kersting, C. Neumann-Haefelin, D. McKinney, M. Roggendorf, F. von Weizsäcker, H. E. Blum, and R. Thimme, "Intrahepatic CD8+ T-cell failure during chronic hepatitis C virus infection." Hepatology **42**, 828 (2005).

[16] C. Neumann-Haefelin, J. Timm, H. C. Spangenberg, N. Wischniowski, N. Nazarova, N. Kersting, M. Roggendorf, T. M. Allen, H. E. Blum, and R. Thimme, "Virological and immunological determinants of intrahepatic virus-specific CD8+ T-cell failure in chronic hepatitis C virus infection," Hepatology **47**, 1824 (2008).

[17] R. A. Lamb, R. Krug, and D. Knipe, "Fields virology," Fields Virology **1** (2001).

[18] R. A. Edwards and F. Rohwer, "Viral metagenomics," Nature Reviews Microbiology **3**, 504 (2005).

[19] C. M. Lawrence, S. Menon, B. J. Eilers, B. Bothner, R. Khayat, T. Douglas, and M. J. Young, "Structural and functional studies of archaeal viruses," Journal of Biological Chemistry **284**, 12599 (2009).

[20] K. M. Stedman, "Encyclopedia of astrobiology," (Springer Berlin Heidelberg, Berlin, Heidelberg, 2011) Chap. Virus, pp. 1745–1748.

[21] V. Kaminskyy and B. Zhivotovsky, "To kill or be killed: how viruses interact with the cell death machinery," Journal of internal medicine **267**, 473 (2010).

[22] F. Fenner, D. A. Henderson, I. Arita, Z. Jezek, and I. D. Ladnyi, *Smallpox and its Eradication* (World Health Organization, Geneva, 1988).

[23] D. A. Koplow, *Smallpox: the fight to eradicate a global scourge* (University of California Press, Berkeley, 2003).

[24] P. Ghosh, "UN 'confident' disease has been wiped out," News release (2010).

[25] P. G. E. Initiative, "Polio eradication," Website (2017).

[26] H. of Vaccines, "Vaccine development, testing, and regulation," Webpage (2017).

[27] I. P. H. Association, "Vaccine development cycle," Webpage (2017).

[28] D. A. Steinhauer and J. J. Holland, "Rapid evolution of RNA viruses," Annual Review of Microbiology **41**, 409 (1987).

[29] R. Sanjuan, M. R. Nebot, N. Chirico, L. M. Mansky, and R. Belshaw, "Viral mutation rates," Journal of virology **84**, 9733 (2010).

[30] M. Cubero, J. I. Esteban, T. Otero, S. Sauleda, M. Bes, R. Esteban, J. Guardia, and J. Quer, "Naturally occurring ns3-protease-inhibitor resistant mutant a156t in the liver of an untreated chronic hepatitis c patient," Virology **370**, 237 (2008).

[31] C. W. Kim and K.-M. Chang, "Hepatitis c virus: virology and life cycle," Clinical and molecular hepatology **19**, 17 (2013).

[32] World Heath Organization, "World Heath Organization Hepatitis C Fact Sheet," `http://www.who.int/mediacentre/factsheets/fs164/en/` (2016).

[33] M. J. Alter, "Epidemiology of hepatitis C virus infection," World Journal of Gastroenterology **13**, 2436 (2007).

[34] J. Halliday, P. Klenerman, and E. Barnes, "Vaccination for hepatitis C virus: Closing in on an evasive target," Expert Review of Vaccines **10**, 659 (2011).

[35] J. Pawlotsky, "New hepatitis c therapies: The toolbox, strategies, and challenges," Gastroenterology **146**, 11761192 (2014).

[36] J. Bukh, "The history of hepatitis C virus (HCV): Basic research reveals unique features in phylogeny, evolution and the viral life cycle with new perspectives for epidemic control," Journal of Hepatology **65**, S2S21 (2016).

[37] M. P. Manns and T. von Hahn, "Novel therapies for hepatitis C — one pill fits all?" Nature Reviews Drug Discovery **12**, 595 (2013).

[38] E. Callaway, "Hepatitis C drugs not reaching poor." Nature **508**, 295 (2014).

[39] K. M. Hanafiah, J. Groeger, A. D. Flaxman, and S. T. Wiersma, "Global epidemiology of hepatitis C virus infection: New estimates of age-specific antibody to HCV seroprevalence," Hepatology **57**, 1333 (2013).

[40] B. Hajarizadeh, J. Grebely, and G. J. Dore, "Epidemiology and natural history of HCV infection," Nature Reviews Gastroenterology and Hepatology **10**, 553 (2013).

[41] "Quarterwatch: Monitoring FDA medwatch reports," (2007).

[42] L. Swadling, P. Klenerman, and E. Barnes, "Ever closer to a prophylactic vaccine for HCV," Expert Opinion on Biological Therapy **13**, 1109 (2013).

[43] H. Dahari, S. Feinstone, and M. Major, "Meta-analysis of hepatitis C virus vaccine efficacy in chimpanzees indicates an importance for structural proteins," Gastroenterology **139**, 965 (2010).

[44] S. H. Mehta, A. Cox, D. R. Hoover, X.-H. Wang, Q. Mao, S. Ray, S. A. Strathdee, D. Vlahov, and D. L. Thomas, "Protection against persistence of hepatitis C," Lancet **359**, 1478 (2002).

[45] W. O. Osburn, B. E. Fisher, K. A. Dowd, G. Urban, L. Liu, S. C. Ray, D. L. Thomas, and A. L. Cox, "Spontaneous control of primary hepatitis C virus infection and immunity against persistent reinfection," Gastroenterology **138**, 315 (2010).

[46] A. Folgori, S. Capone, L. Ruggeri, A. Meola, E. Sporeno, B. Ercole, M. Pezzanera, R. Tafi, M. Arcuri, E. Fattori, A. Lahm, A. Luzzago, A. Vitelli, S. Colloca, R. Cortese, and A. Nicosia, "A T-cell HCV vaccine eliciting effective immunity against heterologous virus challenge in chimpanzees," Nature Medicine **12**, 190 (2006).

[47] J. Grebely, M. Prins, M. Hellard, A. L. Cox, W. O. Osburn, G. Lauer, K. Page, A. R. Lloyd, and G. J. Dore, "Hepatitis C virus clearance, reinfection, and persistence, with insights from studies of injecting drug users: Towards a vaccine," The Lancet Infectious Diseases **12**, 408 (2012).

[48] G. M. Lauer, "Immune responses to hepatitis C virus (HCV) infection and the prospects for an effective HCV vaccine or immunotherapies," Journal of Infectious Diseases **207**, S7 (2013).

[49] K. Page, J. A. Hahn, J. Evans, S. Shiboski, P. Lum, E. Delwart, L. Tobler, W. Andrews, L. Avanesyan, S. Cooper, and M. P. Busch, "Acute hepatitis C virus infection in young adult injection drug users: A prospective study of incident infection, resolution, and reinfection," Journal of Infectious Diseases **200**, 1216 (2009).

[50] A. L. Cox, D. M. Netski, T. Mosbruger, S. G. Sherman, S. Strathdee, D. Ompad, D. Vlahov, D. Chien, V. Shyamala, S. C. Ray, and D. L. Thomas, "Prospective evaluation of community-acquired acute-phase hepatitis C virus infection," Clinical Infectious Diseases **40**, 951 (2005).

[51] V. Dahirel, K. Shekhar, F. Pereyra, T. Miura, M. Artyomov, S. Talsania, T. M. Allen, M. Altfeld, M. Carrington, D. J. Irvine, B. D. Walker, and A. K. Chakraborty, "Coordinate linkage of HIV evolution reveals regions of immunological vulnerability," Proceedings of the National Academy of Sciences **108**, 11530 (2011).

[52] H. Streeck, J. S. Jolin, Y. Qi, B. Yassine-Diab, R. C. Johnson, D. S. Kwon, M. M. Addo, C. Brumme, J.-P. Routy, S. Little, H. Jessen, A. D. Kelleher, F. M. Hecht, R.-P. Sekaly, E. S. Rosenberg, W. B. D., M. Carrington, and M. Altfeld, "Human immunodeficiency virus type 1-specific CD8+ T-cell responses during primary infection are major determinants of the viral set point and loss of CD4+ T cells," Journal of Virology **83**, 7641 (2009).

[53] H. Wedemeyer, E. Schuller, V. Schlaphoff, R. E. Stauber, J. Wiegand, I. Schiefke, C. Firbas, B. Jilma, M. Thursz, S. Zeuzem, W. P. Hofmann, H. Hinrichsen, E. Tauber, M. P. Manns, and C. S. Klade, "Therapeutic vaccine IC41 as late add-on to standard treatment in patients with chronic hepatitis C," Vaccine **27**, 5142 (2009).

[54] D. Drane, E. Maraskovsky, R. Gibson, S. Mitchell, M. Barnden, A. Moskwa, D. Shaw, B. Gervase, S. Coates, M. Houghton, and R. Basser, "Priming of CD4+ and CD8+ T cell responses using a HCV core ISCOMATRIX vaccine: A phase I study in healthy volunteers," Human Vaccines **5**, 151 (2009).

[55] E. Barnes, A. Folgori, S. Capone, L. Swadling, S. Aston, A. Kurioka, J. Meyer, R. Huddart, K. Smith, R. Townsend, A. Brown, R. Antrobus, V. Ammendola, M. Naddeo, G. O'Hara, C. Willberg, A. Harrison, F. Grazioli, M. L. Esposito, L. Siani, C. Traboni, Y. Oo, D. Adams, A. Hill, S. Colloca, A. Nicosia, R. Cortese, and P. Klenerman, "Novel adenovirus-based vaccines induce broad and sustained T cell responses to HCV in man," Science Translational Medicine **4**, 115ra1 (2012).

[56] F. Habersetzer, G. Honnet, C. Bain, M. Maynard-Muet, V. Leroy, J.-P. Zarski, C. Feray, T. F. Baumert, J.-P. Bronowicki, M. Doffoël, C. Trépo, D. Agathon, M.-L. Toh, M. Baudin, J.-Y. Bonnefoy, J.-M. Limacher, and G. Inchauspé, "A poxvirus vaccine is safe, induces T-cell responses, and decreases viral load in patients with chronic hepatitis C," Gastroenterology **141**, 890 (2011).

[57] E. J. Gowans, S. Roberts, K. Jones, I. Dinatale, P. A. Latour, B. Chua, E. M. Y. Eriksson, R. Chin, S. Li, D. M. Wall, R. L. Sparrow, J. Moloney, M. Loudovaris,

R. French, H. M. Prince, D. Hart, W. Zeng, J. Torresi, L. E. Brown, and D. C. Jackson, "A phase I clinical trial of dendritic cell immunotherapy in HCV-infected individuals." Journal of Hepatology **53**, 599 (2010).

[58] O. Weiland, G. Ahlén, H. Diepolder, M.-C. Jung, S. Levander, M. Fons, I. Mathiesen, N. Y. Sardesai, A. Vahlne, L. Frelin, and M. Sällberg, "Therapeutic DNA vaccination using in vivo electroporation followed by standard of care therapy in patients with genotype 1 chronic hepatitis C," Molecular Therapy **21**, 1796 (2013).

[59] G. R. Hart and A. L. Ferguson, "Systems immunology: An introduction to modeling methods for scientists," (Taylor and Francis, (in press, 2017)) Chap. Viral fitness landscapes: A physical sciences perspective.

[60] ""virus, n.". oed online. march 2016. oxford university press. http://www.oed.com/view/entry/223861 (accessed march 11, 2016)." .

[61] E. Rybicki, "The classification of organisms at the edge of life or problems with virus systematics," South African Journal of Science **86**, 182 (1990).

[62] M. Nowak and R. M. May, *Virus Dynamics: Mathematical Principles of Immunology and Virology* (Oxford University Press, 2000).

[63] S. B. Biswas and A. Biswas, *An introduction to viruses* (Vani Educational Books, 1984).

[64] M. A. Nowak, *Evolutionary dynamics* (Harvard University Press, 2006).

[65] C. K. Biebricher and M. Eigen, "The error threshold," Virus research **107**, 117 (2005).

[66] A. R. Wargo and G. Kurath, "Viral fitness: definitions, measurement, and current insights," Current opinion in virology **2**, 538 (2012).

[67] E. Domingo and J. Holland, "Rna virus mutations and fitness for survival," Annual Reviews in Microbiology **51**, 151 (1997).

[68] E. Domingo, "Mechanisms of viral emergence," Veterinary research **41**, 38 (2010).

[69] H. A. Orr, "Fitness and its role in evolutionary genetics," Nature Reviews Genetics **10**, 531 (2009).

[70] H. Wu, Y. Huang, C. Dykes, D. Liu, J. Ma, A. S. Perelson, and L. M. Demeter, "Modeling and estimation of replication fitness of human immunodeficiency virus type 1 in vitro experiments by using a growth competition assay," Journal of virology **80**, 2380 (2006).

[71] D. Hughes and D. I. Andersson, "Evolutionary consequences of drug resistance: shared principles across diverse targets and organisms," Nature Reviews Genetics **16**, 459 (2015).

[72] A. F. Marée, W. Keulen, C. A. Boucher, and R. J. De Boer, "Estimating relative fitness in viral competition experiments," Journal of virology **74**, 11067 (2000).

[73] S. Bonhoeffer, A. D. Barbour, and R. J. De Boer, "Procedures for reliable estimation of viral fitness from time-series data," Proceedings of the Royal Society of London B: Biological Sciences **269**, 1887 (2002).

[74] I. G. Szendro, M. F. Schenk, J. Franke, J. Krug, and J. A. G. De Visser, "Quantitative analyses of empirical fitness landscapes," Journal of Statistical Mechanics: Theory and Experiment **2013**, P01005 (2013).

[75] Z. D. Stephens, S. Y. Lee, F. Faghri, R. H. Campbell, C. Zhai, M. J. Efron, R. Iyer, M. C. Schatz, S. Sinha, and G. E. Robinson, "Big data: astronomical or genomical?" PLoS Biol **13**, e1002195 (2015).

[76] H. Richter, "Fitness landscapes: From evolutionary biology to evolutionary computation," Emergence, Complexity and Computation , 331 (2014).

[77] J. A. G. de Visser and J. Krug, "Empirical fitness landscapes and the predictability of evolution," Nature Reviews Genetics **15**, 480 (2014).

[78] J. M. Smith, "Natural selection and the concept of a protein space," (1970).

[79] M. Eigen, "Selforganization of matter and the evolution of biological macromolecules," Naturwissenschaften **58**, 465 (1971).

[80] M. Eigen and P. Schuster, "A principle of natural self-organization," Naturwissenschaften **64**, 541 (1977).

[81] G. R. Hart and A. L. Ferguson, "Empirical fitness models for hepatitis C virus immunogen design," Physical Biology **12**, 066006 (2015).

[82] K. Tripathi, R. Balagam, N. K. Vishnoi, and N. M. Dixit, "Stochastic simulations suggest that HIV-1 survives close to its error threshold." Public Library of Science Computational Biology **8**, e1002684 (2012).

[83] D. Seifert and N. Beerenwinkel, "Estimating fitness of viral quasispecies from next-generation sequencing data," in *Current Topics in Microbiology and Immunology* (Springer, 2015).

[84] I. Leuthäusser, "Statistical mechanics of eigen's evolution model," Journal of statistical physics **48**, 343 (1987).

[85] D. B. Saakian, E. Munoz, C.-K. Hu, and M. Deem, "Quasispecies theory for multiple-peak fitness landscapes," Physical Review E **73**, 041913 (2006).

[86] C. O. Wilke, J. L. Wang, C. Ofria, R. E. Lenski, and C. Adami, "Evolution of digital organisms at high mutation rates leads to survival of the flattest," Nature **412**, 331 (2001).

[87] S. Elena, P. Agudelo-Romero, P. Carrasco, F. Codoner, S. Martin, C. Torres-Barcelo, and R. Sanjuán, "Experimental evolution of plant rna viruses," Heredity **100**, 478 (2008).

[88] M. Nowak and P. Schuster, "Error thresholds of replication in finite populations mutation frequencies and the onset of muller's ratchet," Journal of theoretical Biology **137**, 375 (1989).

[89] J. W. Drake and J. J. Holland, "Mutation rates among rna viruses," Proceedings of the National Academy of Sciences **96**, 13910 (1999).

[90] J. Alonso and H. Fort, "Error catastrophe for viruses infecting cells: analysis of the phase transition in terms of error classes," Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences **368**, 5569 (2010).

[91] M. Eigen, "Error catastrophe and antiviral strategy," Proceedings of the National Academy of Sciences **99**, 13374 (2002).

[92] D. Wales, *Energy Landscapes: Applications to Clusters, Biomolecules and Glasses* (Cambridge University Press, 2003).

[93] J. Swetina and P. Schuster, "Self-replication with errors: A model for polvnucleotide replication," Biophysical chemistry **16**, 329 (1982).

[94] P. Tarazona, "Error thresholds for molecular quasispecies as phase transitions: From simple landscapes to spin-glass models," Physical Review A **45**, 6038 (1992).

[95] M. Eigen, "Natural selection: a phase transition?" Biophysical chemistry **85**, 101 (2000).

[96] G. R. Hart and A. L. Ferguson, "Error catastrophe and phase transition in the empirical fitness landscape of HIV," Physical Review E: Statistical, Nonlinear, and Soft Matter Physics **91**, 032705 (2015).

[97] N. M. Dixit, P. Srivastava, and N. K. Vishnoi, "A finite population model of molecular evolution: Theory and computation," Journal of Computational Biology **19**, 1176 (2012).

[98] E. C. Holmes, "Error thresholds and the constraints to rna virus evolution," Trends in microbiology **11**, 543 (2003).

[99] R. V. Solé, "Phase transitions in unstable cancer cell populations," The European Physical Journal B-Condensed Matter and Complex Systems **35**, 117 (2003).

[100] M. J. Dapp, C. L. Clouser, S. Patterson, and L. M. Mansky, "5-azacytidine can induce lethal mutagenesis in human immunodeficiency virus type 1," Journal of virology **83**, 11950 (2009).

[101] J. I. Mullins, L. Heath, J. P. Hughes, J. Kicha, S. Styrchak, K. G. Wong, U. Rao, A. Hansen, K. S. Harris, J.-P. Laurent, *et al.*, "Mutation of hiv-1 genomes in a clinical population treated with the mutagenic nucleoside kp1461," PloS one **6**, e15135 (2011).

[102] R. A. Smith, L. A. Loeb, and B. D. Preston, "Lethal mutagenesis of hiv," Virus research **107**, 215 (2005).

[103] J. Summers and S. Litwin, "Examining the theory of error catastrophe," Journal of virology **80**, 20 (2006).

[104] J. P. Anderson, R. Daifuku, and L. A. Loeb, "Viral error catastrophe by mutagenic nucleosides," Annu. Rev. Microbiol. **58**, 183 (2004).

[105] A. L. Bauer, C. A. Beauchemin, and A. S. Perelson, "Agent-based modeling of host–pathogen systems: the successes and challenges," Information sciences **179**, 1379 (2009).

[106] K. Shekhar, C. F. Ruberman, A. L. Ferguson, J. P. Barton, M. Kardar, and A. K. Chakraborty, "Spin models inferred from patient-derived viral sequence data faithfully describe HIV fitness landscapes," Physical Review E **88**, 062705 (2013).

[107] F. Balloux, "Easypop (version 1.7): a computer program for population genetics simulations," Journal of heredity **92**, 301 (2001).

[108] E. L. Read, A. A. Tovo-Dwyer, and A. K. Chakraborty, "Stochastic effects are important in intrahost HIV evolution even when viral loads are high." Proceedings of the National Academy of Sciences of the United States of America **109**, 19727 (2012).

[109] G. Sella and A. E. Hirsh, "The application of statistical physics to evolutionary biology," Proceedings of the National Academy of Sciences **102**, 9541 (2005).

[110] M. W. Deem and P. Hejazi, "Theoretical aspects of immunity," Annual review of chemical and biomolecular engineering **1**, 247 (2010).

[111] J. K. Mann, J. P. Barton, A. L. Ferguson, S. Omarjee, B. D. Walker, A. K. Chakraborty, and T. Ndung'u, "The fitness landscape of HIV-1 Gag: Advanced modeling approaches and validation of model predictions by in vitro testing," Public Library of Science Computational Biology **10**, e1003776 (2014).

[112] J. P. Barton, M. Kardar, and A. K. Chakraborty, "Scaling laws describe memories of host-pathogen riposte in the HIV population." Proceedings of the National Academy of Sciences of the United States of America **112**, 1965 (2015).

[113] M. Castellana and W. Bialek, "Inverse spin glass and related maximum entropy problems," Physical review letters **113**, 117204 (2014).

[114] G. Tkačik and W. Bialek, "Information processing in living systems," arXiv preprint arXiv:1412.8752 (2014).

[115] T. Mora, A. M. Walczak, W. Bialek, and C. G. Callan, "Maximum entropy models for antibody diversity," Proceedings of the National Academy of Sciences **107**, 5405 (2010).

[116] W. Rowe, M. Platt, D. C. Wedge, P. J. Day, D. B. Kell, and J. Knowles, "Analysis of a complete dna–protein affinity landscape," Journal of The Royal Society Interface **7**, 397 (2010).

[117] S. Manrubia and E. Lázaro, "Getting to know viral evolutionary strategies: Towards the next generation of quasispecies models," in *Current Topics in Microbiology and Immunology* (Springer, 2015).

[118] A. Acevedo, L. Brodsky, and R. Andino, "Mutational and fitness landscapes of an rna virus revealed through population sequencing," Nature **505**, 686 (2014).

[119] R. D. Kouyos, G. E. Leventhal, T. Hinkley, M. Haddad, J. M. Whitcomb, C. J. Petropoulos, and S. Bonhoeffer, "Exploring the complexity of the HIV-1 fitness landscape." Public Library of Science Genetics **8**, e1002551 (2012).

[120] T. Hinkley, J. Martins, C. Chappey, M. Haddad, E. Stawiski, J. M. Whitcomb, C. J. Petropoulos, and S. Bonhoeffer, "A systems analysis of mutational effects in hiv-1 protease and reverse transcriptase," Nature genetics **43**, 487 (2011).

[121] H. Qi, C. A. Olson, N. C. Wu, R. Ke, C. Loverdo, V. Chu, S. Truong, R. Remenyi, Z. Chen, Y. Du, S.-Y. Su, L. Q. Al-Mawsawi, T.-T. Wu, S.-H. Chen, C.-Y. Lin, W. Zhong, J. O. Lloyd-Smith, and R. Sun, "A quantitative high-resolution genetic profile rapidly identifies sequence determinants of hepatitis C viral fitness and drug sensitivity." Public Library of Science Pathogens **10**, e1004064 (2014).

[122] J. Otwinowski and I. Nemenman, "Genotype to phenotype mapping and the fitness landscape of the e. coli lac promoter," PloS one **8**, e61570 (2013).

[123] Y. Hayashi, T. Aita, H. Toyota, Y. Husimi, I. Urabe, and T. Yomo, "Experimental rugged fitness landscape in protein sequence space," PLoS One **1**, e96 (2006).

[124] M. R. Segal, J. D. Barbour, and R. M. Grant, "Relating hiv-1 sequence variation to replication capacity via trees and forests," Statistical Applications in Genetics and Molecular Biology **3**, 1 (2004).

[125] J. Ma, C. Dykes, T. Wu, Y. Huang, L. Demeter, and H. Wu, "vfitness: a web-based computing tool for improving estimation of in vitro hiv-1 fitness experiments," BMC bioinformatics **11**, 1 (2010).

[126] A. A. Quadeer, R. H. Y. Louie, K. Shekhar, A. K. Chakraborty, I.-M. Hsing, and M. R. McKay, "Statistical linkage analysis of substitutions in patient-derived sequences of genotype 1a hepatitis C virus non-structural protein 3 exposes targets for immunogen design." Journal of Virology **88**, 7628 (2014).

[127] S. A. Kauffman and E. D. Weinberger, "The nk model of rugged fitness landscapes and its application to maturation of the immune response," Journal of theoretical biology **141**, 211 (1989).

[128] S. A. Kauffman, *The origins of order: Self organization and selection in evolution* (Oxford University Press, USA, 1993).

[129] E. Weinberger, "Np completeness of kauffman's n-k model, a tuneable rugged fitness landscape," (Santa Fe Institute, 1996).

[130] J. Kingman, "A simple model for the balance between selection and mutation," Journal of Applied Probability , 1 (1978).

[131] B. Derrida, "Random-energy model: An exactly solvable model of disordered systems," Physical Review B **24**, 2613 (1981).

[132] J. Franke, A. Klözer, J. A. G. M. de Visser, and J. Krug, "Evolutionary accessibility of mutational pathways." Public Library of Science Computational Biology **7**, e1002134 (2011).

[133] M. Kimura, *The neutral theory of molecular evolution* (Cambridge University Press, 1984).

[134] K. Deforche, R. Camacho, K. Van Laethem, P. Lemey, A. Rambaut, Y. Moreau, and A.-M. Vandamme, "Estimation of an in vivo fitness landscape experienced by hiv-1 under drug selective pressure useful for prediction of drug resistance evolution during treatment," Bioinformatics **24**, 34 (2008).

[135] N. Beerenwinkel, M. Däumer, T. Sing, J. Rahnenführer, T. Lengauer, J. Selbig, D. Hoffmann, and R. Kaiser, "Estimating hiv evolutionary pathways and the genetic barrier to drug resistance," Journal of Infectious Diseases **191**, 1953 (2005).

[136] N. Beerenwinkel, H. Gunthard, V. Roth, and K. J. Metzner, "Challenges and opportunities in estimating viral genetic diversity from next-generation sequencing data," Front Microbiol **3**, 329 (2012).

[137] P. C. Matthews, A. J. Leslie, A. Katzourakis, H. Crawford, R. Payne, A. Prendergast, K. Power, A. D. Kelleher, P. Klenerman, J. Carlson, D. Heckerman, T. Ndung'u, B. D. Walker, T. M. Allen, O. G. Pybus, and P. J. R. Goulder, "HLA footprints on human immunodeficiency virus type 1 are associated with interclade polymorphisms and intraclade phylogenetic clustering," Journal of Virology **83**, 4605 (2009).

[138] P. Falugi and L. Giarré, "Identification and validation of quasispecies models for biological systems," Systems & Control Letters **58**, 529 (2009).

[139] T. Mora and W. Bialek, "Are biological systems poised at criticality?" Journal of Statistical Physics **144**, 268 (2011).

[140] G. Tkacik, E. Schneidman, M. J. I. Berry, and W. Bialek, "Ising models for networks of real neurons," Arxiv **preprint**, arXiv:q (2006).

[141] G. Tkacik, E. Schneidman, M. J. I. Berry, and W. Bialek, "Spin glass models for a network of real neurons," Arxiv **preprint**, arXiv:0912.5409 (2009).

[142] W. Bialek, A. Cavagna, I. Giardina, T. Mora, E. Silvestri, M. Viale, and A. M. Walczak, "Statistical mechanics for natural flocks of birds," Proceedings of the National Academy of Sciences **109**, 4786 (2012).

[143] M. Weigt, R. A. White, H. Szurmant, J. A. Hoch, and T. Hwa, "Identification of direct residue contacts in protein–protein interaction by message passing," Proceedings of the National Academy of Sciences **106**, 67 (2009).

[144] F. Morcos, A. Pagnani, B. Lunt, A. Bertolino, D. S. Marks, C. Sander, R. Zecchina, J. N. Onuchic, T. Hwa, and M. Weigt, "Direct-coupling analysis of residue coevolution captures native contacts across many protein families," Proceedings of the National Academy of Sciences **108**, E1293 (2011).

[145] J. I. Sułkowska, F. Morcos, M. Weigt, T. Hwa, and J. N. Onuchic, "Genomics-aided structure prediction," Proceedings of the National Academy of Sciences **109**, 10340 (2012).

[146] S. Cocco, R. Monasson, and M. Weigt, "From principal component to direct coupling analysis of coevolution in proteins: Low-eigenvalue modes are needed for structure prediction," PLoS Comput Biol **9**, e1003176 (2013).

[147] B. Lunt, H. Szurmant, A. Procaccini, J. A. Hoch, T. Hwa, and M. Weigt, "Chapter two-inference of direct residue contacts in two-component signaling," Methods in enzymology **471**, 17 (2010).

[148] E. T. Jaynes, "Information theory and statistical mechanics," Physical Review **106**, 620 (1957).

[149] E. T. Jaynes, "Information theory and statistical mechanics. II," Phys. Rev. **108**, 171 (1957).

[150] K. Binder and A. P. Young, "Spin glasses: experimental facts, theoretical concepts, and open questions," Rev. Mod. Phys. **58**, 801 (1986).

[151] H. J. Kappen and F. d. B. Rodríguez, "Efficient learning in boltzmann machines using linear response theory," Neural Computation **10**, 1137 (1998).

[152] D. J. Thouless, P. W. Anderson, and R. G. Palmer, "Solution of'solvable model of a spin glass'," Philosophical Magazine **35**, 593 (1977).

[153] Y. Roudi, E. Aurell, and J. A. Hertz, "Statistical physics of pairwise probability models," FRONTIERS IN COMPUTATIONAL NEUROSCIENCE **3**, 22 (2009).

[154] Y. Roudi, J. Tyrcha, and J. Hertz, "Ising model for neural data: model quality and approximate methods for extracting functional connectivity," Physical Review E **79**, 051915 (2009).

[155] T. R. Lezon, J. R. Banavar, M. Cieplak, A. Maritan, and N. V. Fedoroff, "Using the principle of entropy maximization to infer genetic interaction networks from gene expression patterns," Proceedings of the National Academy of Sciences **103**, 19033 (2006).

[156] S. Cocco and R. Monasson, "Adaptive cluster expansion for inferring boltzmann machines with noisy data," Physical review letters **106**, 090601 (2011).

[157] E. Aurell and M. Ekeberg, "Inverse ising inference using all the data," Physical review letters **108**, 090201 (2012).

[158] J. S. Dickstein, P. B. Battaglino, and M. R. DeWeese, "New method for parameter estimation in probabilistic models: minimum probability flow," Physical Review Letters **107**, 220601 (2011).

[159] M. Habeck, "Bayesian approach to inverse statistical mechanics," Physical Review E **89**, 052113 (2014).

[160] B. B. Machta, R. Chachra, M. K. Transtrum, and J. P. Sethna, "Parameter space compression underlies emergent theories and predictive models," Science **342**, 604607 (2013).

[161] M. P. Manns, G. R. Foster, J. K. Rockstroh, S. Zeuzem, F. Zoulim, and M. Houghton, "The way forward in HCV treatment — finding the right path," Nature Reviews Drug Discovery **6**, 991 (2007).

[162] F. Penin, J. Dubuisson, F. A. Rey, D. Moradpour, and J.-M. Pawlotsky, "Structural biology of hepatitis C virus," Hepatology **39**, 5 (2004).

[163] M. M. Manos, V. A. Shvachko, R. C. Murphy, J. M. Arduino, and N. J. Shire, "Distribution of hepatitis C virus genotypes in a diverse us integrated health care population," Journal of Medical Virology **84**, 1744 (2012).

[164] L. M. Blatt, M. G. Mutchnick, M. J. Tong, F. M. Klion, E. Lebovics, B. Freilich, N. Bach, C. Smith, J. Herrera, H. Tobias, A. Conrad, P. Schmid, and J. G. McHutchison, "Assessment of hepatitis C virus RNA and genotype from 6807 patients with chronic hepatitis C in the United States," Journal of Viral Hepatitis **7**, 196 (2000).

[165] O. V. Nainan, M. J. Alter, D. Kruszon-Moran, F.-X. Gao, G. Xia, G. McQuillan, and H. S. Margolis, "Hepatitis C virus genotypes and viral concentrations in participants of a general population survey in the United States," Gastroenterology **131**, 478 (2006).

[166] World Heath Organization, "World Heath Organization Hepatitis C Fact Sheet," `http://www.who.int/csr/disease/hepatitis/whocdscsrlyo2003/en/index2.html` (2013).

[167] C. Kuiken, K. Yusim, L. Boykin, and R. Richardson, "The Los Alamos HCV sequence database," Bioinformatics **21**, 376 (2005).

[168] M. J. Donlin, N. A. Cannon, E. Yao, J. Li, A. Wahed, M. W. Taylor, S. H. Belle, A. M. Di Bisceglie, R. Aurora, and J. E. Tavis, "Pretreatment sequence diversity differences in the full-length hepatitis C virus open reading frame correlate with early response to therapy," Journal of Virology **81**, 8211 (2007).

[169] T. Kuntzen, J. Timm, A. Berical, N. Lennon, A. M. Berlin, S. K. Young, B. Lee, D. Heckerman, J. Carlson, L. L. Reyor, M. Kleyman, C. M. McMhon, C. Birch, J. Schulze zur Wiesch, T. Ledlie, M. Koehrsen, G. M. Lauer, H. R. Rosen, F. Bihl, A. Cerny, U. Spengler, Z. Liu, A. Y. Kim, Y. Xing, A. Schneidwind, J. F. Madey, Margaret A. Fleckenstein, V. M. Park, J. E. Galagan, C. Nusbaum, B. D. Walker, E. S. Lake-Bakaar, Gerond V. Daar, I. M. Jacobson, B. R. Gomperts, Edward D. Edlin, S. M. Donfield, R. T. Chung, A. H. Talal, T. Marion, B. W. Birren, M. R. Henn, and T. M. Allen, "Naturally occurring dominant resistance mutations to hepatitis C virus protease and polymerase inhibitors in treatment-naïve patients," Hepatology **48**, 1769 (2008).

[170] D. J. Bartels, J. C. Sullivan, E. Z. Zhang, A. M. Tigges, J. L. Borrian, S. De Meyer, D. Takemoto, E. Dondero, A. D. Kwong, G. Picchio, and T. L. Kieffer, "Hepatitis C virus variants with decreased sensitivity to direct-acting antivirals (DAAs) were rarely observed in DAA-naive patients prior to treatment." Journal of Virology **87**, 1544 (2013).

[171] D. S. Sivia and J. Skilling, *Data Analysis: A Bayesian Tutorial*, 2nd ed. (Oxford University Press, Great Clarendon Street, Oxford OX2 6DP, 2006).

[172] E. H. Sklan, P. Charuworn, P. S. Pang, and J. S. Glenn, "Mechanisms of HCV survival in the host," Nature Reviews Gastroenterology and Hepatology **6**, 217 (2009).

[173] A. L. Cox, T. Mosbruger, Q. Mao, Z. Liu, X.-H. Wang, H.-C. Yang, J. Sidney, A. Sette, D. Pardoll, D. L. Thomas, and S. C. Ray, "Cellular immune selection with hepatitis C virus persistence in humans," The Journal of Experimental Medicine **201**, 1741 (2005).

[174] A. Plauzolles, M. Lucas, and S. Gaudieri, "Hepatitis C virus adaptation to T-cell immune pressure," The Scientific World Journal **2013**, 673240 (2013), article ID 673240, 7 Pages.

[175] K. Murphy, *Janeway's Immunobiology*, 8th ed. (Garland Science, New York, 2012).

[176] M. Ruhl, P. Chhatwal, H. Strathmann, T. Kuntzen, D. Bankwitz, K. Skibbe, A. Walker, F. M. Heinemann, P. A. Horn, D. Allen, Todd M. Hoffmann, T. Pietschmann, and J. Timm, "Escape from a dominant HLA-B*15-restricted CD8+ T cell response against hepatitis C virus requires compensatory mutations outside the epitope," Journal of Virology **86**, 991 (2012).

[177] K. M. Chang, B. Rehermann, J. G. McHutchison, C. Pasquinelli, S. Southwood, A. Sette, and F. V. Chisari, "Immunological significance of cytotoxic T lymphocyte epitope variants in patients chronically infected by the hepatitis C virus," The Journal of Clinical Investigation **100**, 2376 (1997).

[178] S. Merani, D. Petrovic, I. James, A. Chopra, D. Cooper, E. Freitas, A. Rauch, J. di Iulio, M. John, M. Lucas, K. Fitzmaurice, S. McKiernan, S. Norris, D. Kelleher, P. Klenerman, and S. Gaudieri, "Effect of immune pressure on hepatitis C virus evolution: Insights from a single-source outbreak," Hepatology **53**, 396 (2011).

[179] R. Vita, L. Zarebski, J. A. Greenbaum, H. Emami, I. Hoof, N. Salimi, R. Damle, A. Sette, and B. Peters, "The Immune Epitope Database 2.0," Nucleic Acids Research **38**, D854 (2010).

[180] C. L. Day, G. M. Lauer, G. K. Robbins, B. McGovern, A. G. Wurcel, R. T. Gandhi, R. T. Chung, and B. D. Walker, "Broad specificity of virus-specific $CD4^+$ T-helper-cell responses in resolved hepatitis C virus infection," Journal of Virology **76**, 12584 (2002).

[181] J. S. zur Wiesch, G. M. Lauer, C. L. Day, A. Y. Kim, K. Ouchi, J. E. Duncan, A. G. Wurcel, J. Timm, A. M. Jones, B. Mothe, T. M. Allen, B. McGovern, L. Lewis-Ximenez, J. Sidney, A. Sette, R. T. Chung, and B. D. Walker, "Broad repertoire of the CD4+ Th cell response in spontaneously controlled hepatitis C virus infection includes dominant and highly promiscuous epitopes." Journal of Immunology **175**, 3603 (2005).

[182] C. L. Thio, X. Gao, J. J. Goedert, D. Vlahov, K. E. Nelson, M. W. Hilgartner, S. J. O'Brien, P. Karacki, J. Astemborski, M. Carrington, and D. L. Thomas, "HLA-Cw*04 and hepatitis C virus persistence," Journal of Virology **76**, 4792 (2002).

[183] P. Hraber, C. Kuiken, and K. Yusim, "Evidence for human leukocyte antigen heterozygote advantage against hepatitis c virus infection." Hepatology **46**, 1713 (2007).

[184] K. Cao and M. Fernández-Viña, "Native American from the United States; Caucasian from the United States; Asian Pacific Islander from the United States; African American from the United States; Hispanic from the United States. Anthropology/human genetic diversity population reports." in *Immunobiology of the Human MHC: Proceedings of the 13th International Histocompatibility Workshop and Conference*, Vol. 1, edited by J. A. Hansen (IHWG Press, Seattle, WA, 2007) pp. 648–651.

[185] W. Fischer, H. X. Liao, B. F. Haynes, N. L. Letvin, and B. Korber, "Coping with viral diversity in HIV vaccine design: A response to Nickle et al." Public Library of Science Computational Biology **4**, e15; author reply e25 (2008).

[186] J. Arora, *Introduction to Optimum Design* (Academic Press, Oxford, UK, 2011).

[187] K. M. De Cock, H. W. Jaffe, and J. W. Curran, "Reflections on 30 years of AIDS," Emerg. Infect. Dis. **17**, 1044 (2011).

[188] K. A. Gebo, J. A. Fleishman, R. Conviser, J. Hellinger, F. J. Hellinger, J. S. Josephs, P. Keiser, P. Gaist, R. D. Moore, *et al.*, "Contemporary costs of HIV healthcare in the HAART era," AIDS **24**, 2705 (2010).

[189] K. M. Stadeli and D. D. Richman, "Rates of emergence of hiv drug resistance in resource-limited settings: a systematic review," Antivir. Ther. **18**, 115 (2013).

[190] B. M. Baker, B. L. Block, A. C. Rothchild, and B. D. Walker, "Elite control of HIV infection: implications for vaccine design." Expert Opin. Biol. Th. **9**, 55 (2009).

[191] B. D. Walker and D. R. Burton, "Toward an AIDS vaccine," Science **320**, 760 (2008).

[192] F. Y. Wu, "The Potts model," Reviews of Modern Physics **54**, 235 (1982).

[193] Los Alamos National Laboratory HIV Database, http://www.hiv.lanl.gov (Accessed: 12 July 2014).

[194] T. Leitner, B. Korber, M. Daniels, C. Calef, and B. Foley, "Hiv-1 subtype and circulating recombinant form (crf) reference sequences," in *HIV Sequence Compendium 2005 (Report No. LA-UR 06-0680)*, edited by T. Leitner, B. Foley, B. Hahn, P. Marx, F. McCutchan, J. W. Mellors, S. Wolinsky, and B. Korber (Theoretical Biology and Biophysics Group, Los Alamos National Laboratory, 2005) pp. 41–48.

[195] Z. L. Brumme, M. John, J. M. Carlson, C. J. Brumme, D. Chan, M. A. Brockman, L. C. Swenson, I. Tao, S. Szeto, P. Rosato, J. Sela, C. M. Kadie, N. Frahm, C. Brander, D. W. Haas, S. A. Riddler, R. Haubrich, B. D. Walker, P. R. Harrigan, D. Heckerman, and S. Mallal, "HLA-associated immune escape pathways in HIV-1 subtype B gag, pol and nef proteins," Public Library of Science ONE **4**, e6687 (2009).

[196] K. Murphy, P. Travers, and M. Walport, *Janeway's Immunobiology*, 8th ed. (Taylor & Francis, 2011).

[197] D. P. Landau, S.-H. Tsai, and M. Exler, "A new approach to Monte Carlo simulations in statistical physics: Wang-Landau sampling," American Journal of Physics **72**, 1294 (2004).

[198] F. Wang and D. P. Landau, "Determining the density of states for classical statistical models: A random walk algorithm to produce a flat histogram," Phys. Rev. E **64**, 056101 (2001).

[199] F. Wang and D. P. Landau, "Efficient, multiple-range random walk algorithm to calculate the density of states," Phys. Rev. Lett. **86**, 2050 (2001).

[200] M. O. Khan, G. Kennedy, and D. Y. C. Chan, "A scalable parallel monte carlo method for free energy simulations of molecular systems," J. Comput. Chem. **26**, 72 (2005).

[201] L. Zhan, "A parallel implementation of the wang–landau algorithm," Comput. Phys. Commun. **179**, 339 (2008).

[202] P. Chomaz and F. Gulminelli, "Phase transitions in finite systems using information theory," arXiv preprint arXiv:0712.1002 (2007).

[203] M. S. Shell, P. G. Debenedetti, and A. Z. Panagiotopoulos, "Computational characterization of the sequence landscape in simple protein alphabets," Proteins: Struct., Funct., Bioinf. **62**, 232 (2006).

[204] I. Leuthäusser, "An exact correspondence between eigen's evolution model and a two-dimensional ising system," J. Chem. Phys. **84**, 1884 (1986).

[205] A. Grande-Pérez, E. Lázaro, P. Lowenstein, E. Domingo, and S. C. Manrubia, "Suppression of viral infectivity through lethal defection," Proc. Natl. Acad. Sci. U.S.A. **102**, 4448 (2005).

[206] J. J. Bull, R. Sanjuan, and C. O. Wilke, "Theory of lethal mutagenesis for viruses," J. Virol. **81**, 2930 (2007).

[207] B. J. Schulz, K. Binder, M. Müller, and D. P. Landau, "Avoiding boundary effects in wang-landau sampling," Phys. Rev. E **67**, 067102 (2003).

[208] M. J. McElrath and B. F. Haynes, "Induction of immunity to human immunodeficiency virus type-1 by vaccination," Immunity **33**, 542 (2010).

[209] E. J. Arts and D. J. Hazuda, "Hiv-1 antiretroviral drug therapy," Cold Spring Harb. Perspect. Med. **2**, a007161 (2012).

[210] E. De Clercq, "The design of drugs for hiv and hcv," Nat. Rev. Drug Discov. **6**, 1001 (2007).

[211] K. Jain and J. Krug, "Structural approaches to sequence evolution: Molecules, networks, populations," (Springer Berlin Heidelberg, 2007) Chap. Adaptation in Simple and Complex Fitness Landscapes, pp. 299–339.

[212] J. P. Barton, N. Goonetilleke, T. C. Butler, B. D. Walker, A. J. McMichael, and A. K. Chakraborty, "Relative rate and location of intra-host HIV evolution to evade cellular immunity are predictable," Nature Communications **7**, 11660 (2016).

[213] R. Fisher, *The genetical theory of natural selection* (The Clarendon Press, 1930).

[214] S. Wright, "Evolution in Menelian populations," Genetics **16**, 97 (1931).

[215] Y. Kim, J. Ponomarenko, Z. Zhu, D. Tamang, P. Wang, J. Greenbaum, C. Lundegaard, A. Sette, O. Lund, P. E. Bourne, M. Nielsen, and B. Peters, "Immune epitope database analysis resource." Nucleic Acids Research **40**, W525 (2012).

[216] N. Vrisekoop, I. den Braber, A. B. de Boer, A. F. C. Ruiter, M. T. Ackermans, S. N. van der Crabben, E. H. R. Schrijver, G. Spierenburg, H. P. Sauerwein, M. D. Hazenberg, R. J. de Boer, F. Miedema, J. A. M. Borghans, and K. Tesselaar, "Sparse

production but preferential incorporation of recently produced naive T cells in the human peripheral pool." Proceedings of the National Academy of Sciences of the United States of America **105**, 6115 (2008).

[217] C. L. Althaus and R. J. De Boer, "Dynamics of immune escape during HIV/SIV infection." Public Library of Science Computational Biology **4**, e1000103 (2008).

[218] M. K. Jenkins, H. H. Chu, J. B. McLachlan, and J. J. Moon, "On the composition of the preimmune repertoire of T cells specific for peptide-major histocompatibility complex ligands." Annual Review of Immunology **28**, 275 (2010).

[219] M. K. Jenkins and J. J. Moon, "The role of naive T cell precursor frequency and recruitment in dictating immune response magnitude." Journal of Immunology **188**, 4135 (2012).

[220] M. J. van Stipdonk, E. E. Lemmens, and S. P. Schoenberger, "Naïve CTLs require a single brief period of antigenic stimulation for clonal expansion and differentiation." Nature Immunology **2**, 423 (2001).

[221] R. J. De Boer, V. V. Ganusov, D. Milutinović, P. D. Hodgkin, and A. S. Perelson, "Estimating lymphocyte division and death rates from CFSE data." Bulletin of Mathematical Biology **68**, 1011 (2006).

[222] S. M. Kaech and R. Ahmed, "Memory CD8+ T cell differentiation: Initial antigen encounter triggers a developmental program in naïve cells." Nature Immunology **2**, 415 (2001).

[223] D. T. Gillespie, "Exact stochastic simulation of coupled chemical reactions," The Journal of Physical Chemistry **81**, 2340 (1977).

[224] R. A. Bull, F. Luciani, K. McElroy, S. Gaudieri, S. T. Pham, A. Chopra, B. Cameron, L. Maher, G. J. Dore, P. A. White, and A. R. Lloyd, "Sequential bottlenecks drive viral evolution in early acute hepatitis C virus infection," Public Library of Science Pathogens **7**, e1002243 (2011).

[225] S. F. de St. Groth and R. G. Webster, "Disquisitions on original antigenic sin I. evidence in man," The Journal of Experimental Medicine **124**, 331 (1966).

[226] M. W. Deem and H. Y. Lee, "Sequence space localization in the immune system response to vaccination and disease," Physical Review Letters **91**, 229902 (2003).

[227] D. J. Thouless, P. W. Anderson, and R. G. Palmer, "Solution of 'solvable model of a spin glass'," Philosophical Magazine **35**, 593 (1976).

[228] K. P. Burke, S. Munshaw, W. O. Osburn, J. Levine, L. Liu, J. Sidney, A. Sette, S. C. Ray, and A. L. Cox, "Immunogenicity and cross-reactivity of a representative ancestral sequence in hepatitis C virus infection." Journal of Immunology **188**, 5177 (2012).

[229] R. Durbin, S. R. Eddy, A. Kroph, and G. Mitchison, *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*, 1st ed. (Cambridge University Press, 1998).

[230] A. Rauch, I. James, K. Pfafferott, D. Nolan, P. Klenerman, W. Cheng, L. Mollison, G. McCaughan, N. Shackel, G. P. Jeffrey, R. Baker, E. Freitas, I. Humphreys, H. Furrer, H. F. Günthard, B. Hirschel, S. Mallal, M. John, M. Lucas, E. Barnes, and S. Gaudieri, "Divergent adaptation of hepatitis C virus genotypes 1 and 3 to human leukocyte antigen–restricted immune pressure," Hepatology **50**, 1017 (2009).

[231] I. Tester, S. Smyk-Pearson, P. Wang, A. Wertheimer, E. Yao, D. M. Lewinsohn, J. E. Tavis, and H. R. Rosen, "Immune evasion versus recovery after acute hepatitis C virus infection from a shared source." Journal of Experimental Medicine **201**, 1725 (2005).