CHAPTER 4

Determining the mutational vulnerabilities of highly mutable viruses for rational design of vaccines

Introduction

We now turn our attention to the main antagonists that the immune system combats, infectious disease-causing pathogens. Infectious diseases have plagued humanity since antiquity. Although there was mention of miasmas and poisons as the cause of disease, the microbial origin of these diseases was unclear in ancient times. But, it was observed that if an individual was afflicted with small pox and recovered, she or he did not fall sick again. Thus, they served as caregivers. The origins of this observation were much debated, and two kinds of ideas (other than mystical ones) were popular at different times. Theories of expulsion argued that humans were born with some "noxious" substances (e.g., traces of menstrual blood), and small pox pustules were a way to expel these substances; one could not be afflicted by the disease again as the noxious substance had been successfully expelled. Theories of depletion were based on the idea that the disease could only thrive on a substrate in the human body, and once that substrate was depleted when an individual was sick, it was not possible for the disease to establish itself again. Beginning with the work of Koch, Pasteur, and many others, we now know unequivocally that microbial pathogens cause infectious diseases.

Vaccination has saved more lives than any other medical procedure. Successful vaccination programs have resulted in the eradication of smallpox, which had caused devastation since antiquity, and the near-eradication of polio. Vaccines for numerous childhood diseases, such as measles, mumps, and rubella, are also a major contributor to the reduction of infant mortality. Modern vaccination roughly follows the paradigm pioneered by Jenner and Pasteur over two centuries ago. A dead or weakened form of the pathogen is injected into humans with the goal of inducing effective memory immune responses. In the twentieth century, additives (called adjuvants) were added to vaccines to help stimulate innate immune responses that are critical for the development of potent adaptive immune responses. The design of adjuvants remains largely an art and many new formulations fail to work. A detailed mechanistic understanding of innate immunity could help make the design of potent adjuvants systematic.

Although the empirical paradigm of vaccine development has been a great success, the traditional approach has not led to successful protective vaccines against some pathogens. Prominent examples are HIV, HCV, tuberculosis, dengue and malaria, many of which are wreaking havoc around the world. We do not have a broadly effective vaccine against influenza either, and attempts to predict the right vaccine for the ensuing year often fail. Many of these pathogens share two features: 1] They present themselves in different guises, thus making them hard to target with specific immune memory responses. 2] They often degrade, or hide from, the immune system.

HIV has characteristics which are extreme examples of both these features. It is a highly mutable virus with a rapid replication rate. Thus, it generates many mutant strains when it infects a person, and an enormous diversity of viral strains is circulating in the human population. Fig. 1 shows a comparison of the diversity of HIV strains in a single infected person, the diversity of circulating influenza strains in the

entire world in a particular year, and that of circulating HIV strains in a single region in Africa during the same time period [11]. The diversity of HIV strains dwarfs that of influenza. The high mutability allows HIV to evade natural or vaccine-induced immune responses [12]. For example, the infecting strain may not be the one for which vaccine-induced memory immune responses exist; even if it is targeted by the memory response, if the virus is not eliminated rapidly, the infecting strain can mutate in the host to escape from this response. Furthermore, HIV principally infects and eventually kills human T helper cells, thus degrading the adaptive immune system. This is the reason why Acquired Immunodeficiency Syndrome (AIDS), the disease associated with HIV infections, results in a severe state of immunodeficiency allowing many normally easy to control infections to afflict patients. Other pathogens listed above also mutate over time (e.g., influenza), and malaria uses a different strategy in that it expresses different interchangeable proteins on the surface. The bacterium that causes malaria hides from the immune system in red blood cells, and tuberculosis suppresses some innate immune functions. Vaccine design against HIV is particularly daunting because, unlike most other pathogens for which vaccines exist, HIV infection is not known to have ever been successfully completely cleared by natural human immune responses. So, there is no model of a human immune response that can eradicate HIV from an infected person which one can aim to mimic with a vaccine. The closest model is provided by a small cohort of patients (introduced in Chapter 3) called elite controllers, whose immune systems can control virus levels to low enough levels that they do not require antiretroviral therapy and do not progress to AIDS.

Successful vaccination against pathogens that have evolved sophisticated strategies to evade human immune responses will benefit from the development of firm scientific principles that can guide rational, rather than empirical, vaccine design [10]. At least two important issues must be studied in this regard: 1] What are the appropriate targets in the pathogen's proteins on which to focus vaccine-induced immune responses such that the ability of the pathogen to evade such responses by mutation while simultaneously maintaining their viability/virulence is severely limited or eliminated? 2] How can such immune responses be induced by vaccination in humans with diverse genotypes?

A convergence of several factors is beginning to enable us to take the first steps toward addressing these questions. Biologists and clinicians can collect enormous amounts of data on sequences of mutant strains of pathogens, and it is also becoming possible to interrogate the immune system on an unprecedented scale. Both the immune system and pathogens function and interact via collective processes that involve myriad individual components, thus making mechanistic interpretation of this data complex. Physicists, especially statistical physicists, have begun to play a role in translating this type of data to mechanistic knowledge that addresses the questions noted above. Engineers and clinician/scientists are beginning to design ways to more effectively deliver vaccines. The goal of developing mechanistic principles that can be harnessed for rational design of vaccines is bringing together physicists, biologists, clinicians, and engineers.

In this chapter, we will focus primarily on defining the mutational vulnerabilities of highly mutable pathogens. For concreteness, we will consider primarily only one virus, HIV, but we will contrast it with influenza, which has a very different evolutionary history. Some topics covered in this chapter are also

pertinent to determining properties of protein families from sequence data, rather than protein structures.

Brief description of the biology of HIV and definition of the key challenges

HIV was transmitted to humans from monkeys, and is estimated to have been circulating in small populations of humans for nearly a century before it was recognized as a new disease causing pathogen [13]. The first well-documented cases were reported in 1981 in the United States. To date, HIV has infected over 70 million people, almost 40 million people have died from complications associated with AIDS, and 1 million people died in 2016. In developed nations, because of wide-spread availability of anti-retroviral (ARV) drugs, HIV infections can be controlled by regular doses of expensive medication, but it cannot be cured. In other parts of the world, ARVs are less easily available, and HIV continues to be a major problem, with sub-Saharan Africa being the epicenter of the disease. For example, although the mortality rates are beginning to stabilize, each day there are approximately 1000 new HIV infections in South Africa alone. A vaccine or cure is needed to eradicate HIV from the planet, but no successes have been reported after more than thirty years of work and very large amounts of money spent since it became known that the causative agent of AIDS was HIV.

HIV is a retrovirus, which carries its genome in the form of RNA. The virus has a membrane through which proteins protrude. These "Envelope" proteins are called gp120 and gp41, which form a non-covalently bonded trimer which constitutes the viral spike (Fig. 2). The density of spikes on the HIV membrane is roughly two orders of magnitude smaller than the spike density of most viruses, a point to which we will return in a later chapter. The outer membrane surrounds a capsid made up of structural proteins which encloses the virus' genome and other key proteins important for viral function (Fig. 2). HIV is a retrovirus, so it carries its genome in the form of RNA.

The life cycle of HIV is depicted schematically in Fig. 3. The trimeric spike binds to host cell surface proteins to initiate infection [14]. For example, the spike can bind to the CD4 co-receptor, which is expressed on the surface of T helper cells [15,16]. This binding event leads to a conformational change in gp120 and the dissociation of gp41, which then forms a six-helix bundle. The conformational change enables gp120 to bind to a second receptor on the surface of the host cell, which can be either CCR5 or CXCR3; viruses that bind to the former receptor are called CCR5-tropic and the ones that bind to the latter are called CXCR3-tropic. These binding events and the free energy gained from gp41 forming an ordered state result in fusion of the virus' membrane with that of the host cell membrane, resulting in release of the capsid in to the cytoplasm. The viral capsid is then uncoated, thus releasing its contents. A viral protein, Reverse Transcriptase, then converts the RNA strands in to DNA. This viral DNA, called a provirus, is transported in to the nucleus of the host cell along with a viral protein called integrase. Integrase inserts the viral DNA in to the genome of the host cell, thus infecting this cell for its lifetime. The HIV genes code for polyproteins (a number of concatenated proteins), and the transcriptional machinery of the host cell is hijacked to express these polyproteins. The polyproteins are chopped up by a HIV protein, called protease, resulting in individual proteins that mediate viral function. In a series of steps, the virus' proteins are properly assembled at the membrane of the host cell [17]. A part of the host cell membrane becomes the membrane for a new virus particle as it buds out.

HIV has nine genes, with four of them, Gag, Pol, Nef, and ENV being the most important. Gag encodes structurally important proteins (such as the ones that make up the capsid), Pol codes for Reverse Transcriptase, Integrase, and Protease, ENV codes for gp120 and gp41, and Nef codes for proteins with many functions, including a role in downregulation of CD4 and MHC proteins. The MHC molecules in humans are called human leukocyte antigen (HLA) proteins.Downregulation of HLA proteins suppresses T cell responses and downregulation of CD4 helps the virus bud out of infected cells because that inhibits binding of the viral spike to the infected cell's CD4 co-receptor.

The two main sources of mutations in HIV's lifecycle arise from reverse transcription not being a very high-fidelity process and mistakes in generating HIV proteins using the human transcriptional machinery. Mutations are introduced at an average rate of 3×10^{-5} per base pair per replication cycle [18]. HIV's genome is about 10^4 base pairs in length, and so this implies that during every replication cycle the probability of evolving a mutant strain is 0.3. Moreover, Reverse Transcriptase can hop from one RNA molecule to the other, thus generating more options for creating diversity. When two different RNA genomes are available, this causes recombination of the genomes of two viral strains. Fitting parameters in ordinary differential equations describing viral dynamics to data from patients treated with drugs revealed that HIV replicates very rapidly [19], producing 10^{10} to 10^{11} virus particles per day in infected humans [20]. Many of the mutant strains that are produced do not grow as they cannot form infective virus particles. But, taken together, the high mutation and replication rates and HIV being a chronic infection are the main reasons underlying the extraordinary diversity of HIV strains circulating in the population and in individuals (Fig. 1).

Upon successful infection, the virus replicates rapidly, and the viral load (that is, the number of viruses circulating in the host) increases (Fig. 4). As immune responses develop, the viral load decreases and then stabilizes at a steady state where the immune system and the virus are in balance [21]. During this phase, there is a dynamic "arms race" between the virus and the immune system. The immune system mounts a response directed at the prevailing viral strains which then mutate to evade these responses, then new responses develop, and the cycle continues. The steady state viral load varies widely across patients [22]. The host cell ultimately dies as new virus particles bud out, and the lifetime of an infected cell is estimated to be about 2 days. Thus, the number of CD4 T cells in an infected patient declines rapidly at first and then increases again to a steady value after the immune response develops to combat the virus (Fig. 4). The period when the viral load and CD4 T cell counts are stable is called the asymptomatic phase as no disease manifests. HIV belongs to a class of viruses called lentiviruses (slow viruses) which cause disease slowly. Without treatment the immune system ultimately loses the battle, viral load goes up, and CD4 T cells decrease to low numbers (Fig. 4). At this point, the individual's immune system is severely compromised, and many opportunistic infections ensue leading ultimately to death.

HIV's ability to generate diverse mutant strains is a major reason why an effective vaccine does not exist [23]. In this book, we will focus on only two aspects of the challenge of creating an effective HIV vaccine. First, we will ask whether the data on thousands of sequences of HIV proteins derived from virus samples extracted from patients can be translated in to knowledge of the mutational vulnerabilities of the virus; i.e., which types of mutations is the virus unable to make to evade immune responses and still

remain viable. This topic will be the focus of this chapter. We will then study how the knowledge thus generated can be used to design the active component of a potentially effective vaccine that elicits potent T cell responses. In a later chapter, we will focus on how to generate potent antibody responses.

Determining the mutational vulnerabilities of HIV

At first glance, determining the mutational vulnerabilities of a virus like HIV appears to be a simple problem given that today we can sequence large numbers of viruses derived from diverse patients. Just lining up these sequences and looking for residues where the most frequent amino acid appears relatively conserved (e.g., measured by low entropy of amino acid variation) should provide the answer. Focusing a vaccine-induced immune response to target such residues should be an effective strategy to control the virus. This is because to evade such an immune response, HIV would have to evolve a mutation at a residue where a particular amino acid is favored, and that should hurt viral function as a particular amino acid is relatively conserved at that residue for a reason. This strategy is blunted because, due to its high replication and mutation rates, HIV can evolve other mutations, so-called compensatory mutations, which can partially restore the fitness cost incurred by making the primary immune-evading mutation [29,30].

Many properties of an evolving virus population can be described by assuming that each residue in the virus' proteins evolves independently. But, if one wishes to determine the mutational vulnerabilities of a virus like HIV, one needs to define the collective mutational pathways that HIV uses to evade human immune responses in order to avoid targeting the involved residues with a vaccine-induced immune response. This is because, even though such compensatory interactions can be relatively rare, the high mutability and replication rate of HIV implies that they can be sampled, especially when immune responses provide a selection force that promotes the evolution of one of the involved mutations. One also needs to determine the combinations of mutations that the virus cannot make and remain viable, so as to target the involved residues with vaccine-induced immune responses and corner the virus between being killed by the immune responses and evolving unfit mutant strains that evade these responses. In short, one needs to determine the fitness landscape [31,32] of the virus – i.e., the ability of the virus to replicate and propagate infection as a function of its sequence, with explicit account for the coupled effects of mutations at different protein residues. Fig. 5 illustrates the concept of a fitness landscape using a 2-dimensional representation of sequence space, such as that of a virus comprised of one protein with two residues. Knowledge of the fitness landscape can be extremely useful for vaccine design. With a vaccine-induced immune response, one wishes to target residues such that combinations of mutations therein correspond to a fitness valley. One also wishes to block the mountain passes corresponding to the collective compensatory mutational pathways that HIV uses to go from one fitness hill to another adjacent one when the first one is under immune attack (Fig. 5).

Vaccines have two key components – the immunogen and the vector. The immunogen is the component of the vaccine that encodes for, or is, the proteome of the pathogen against which one wishes to protect the host. The vector provides a way to deliver the immunogen in a way that results in a strong host immune response to the immunogen. The vector can be a harmless virus, in which case the virulent virus' genome could be inserted into the carrier's genome. There is also a lot of research being directed

these days toward engineering nanoparticles that can serve as vectors, with the immunogen being DNA, proteins, or peptides.

Armed with the fitness landscape of a virus, one could design immunogens that could induce potent T cell responses. Such an immunogen would not be comprised of whole HIV proteomes as is traditional, but would contain only parts of it chosen according to three criteria: 1] Regions that are rife with compensatory pathways are minimized. 2] Regions wherein multiple mutations are especially deleterious are maximized. 3] Regions that can be presented by people with diverse MHC genes are maximized. Knowledge of the fitness landscape can also guide the choice of immunogens that could induce potent antibody responses, a topic that we will consider in a later chapter.

The practical goal of enabling rational design of immunogens motivates efforts to obtain the fitness landscape of viruses. The method that we will consider below relies on translating sequence information alone in to knowledge of the fitness landscape; other approaches have attempted to infer information about HIV fitness by combining sequence information with data from in vitro experiments [33,34], an approach with its own set of potential complications [35]. We will also note that methods based on analyses of protein structures can add an underlying molecular perspective to the knowledge gained from obtaining viral fitness landscapes, but only when structures are available and when coupling between the effects of mutations do not originate from functions that involve multi-protein assembles of proteins or other functions that do not depend on structural stability alone.

Simple calculations can reveal the importance of co-evolution of mutations in a protein or genome

A simple way to determine whether mutations in different residues of a protein (or genome) coevolve is to analyze the covariance matrix defined by the probability of observing coupled pairs of mutations. For illustrative purposes, let us imagine that the amino acids at each residue come in two flavors, wild type or most frequent (denoted by 0) and mutant (denoted by 1). Such a representation is reasonable when the number of types of amino acids observed at each residue is small and there is a dominant mutant amino acid. In such a representation, if z_i (equal to 0 or 1) represents the amino acid at residue, i, the covariance matrix, **C**, is defined as

$$C_{ij} = \frac{\langle z_i z_j \rangle - \langle z_i \rangle \langle z_j \rangle}{\sqrt{V_i V_j}}, \quad (1)$$

where angular brackets represent an average over all sequences and V_i and V_j are the variances of the distribution of mutations at residues i and j, respectively. Normalization of the elements of **C** using the variances allows comparison of the magnitudes of matrix elements corresponding to residues that could differ widely in the extent to which they are mutable. Other variants of the covariance matrix that are normalized differently have been described [], but here we focus on the simplest formulation. A covariance matrix can be diagonalized, and the corresponding eigenvectors are the simplest reflection of collective mutational pathways.

But, the problem is complicated by at least two effects. First, even though we have many samples of the sequences of HIV proteins, it is a finite sample. Even if two variables are completely uncorrelated, given

a finite sample, the covariance matrix will exhibit spurious correlations. Second, some correlations simply reflect the fact that all sequences belong to the same family, or phylogeny, of HIV viruses, and mutational fidelity leads to correlations between sequences. Such correlations are unrelated to those that arise from co-evolution of residues due to fitness constraints.

Random matrix theory can be used to "clean" the correlation matrix of spurious noise and various approaches can be attempted to deconvolute the effects of phylogeny. Indeed, analyses of the "cleaned" correlation matrix have led to interesting insights into correlated mutations, and their biological significance, in various contexts. For example, Ranganathan and co-workers have analyzed some bacterial protein families in this manner and described distinct sets of co-evolving residues within such proteins, as determined from analysis of the eigenvectors. They called each set of coevolving residues a "sector", with each sector being important for a distinct function of the protein. Similar analyses have also shed light on co-evolution of mutations in the HIV polyprotein, Gag, and HCV proteins. For example, it was shown that multiple simultaneous mutations in certain sets of residues in a protein called p24 (in Gag) evolved rarely because they are involved in protein-protein interfaces that are critical for the assembly of the structure of the viral capsid. Thus, viral strains with mutations at multiple such residues were likely to destabilize capsid assembly, and make the virus unviable. Consistent with this interpretation, it was shown that T cell responses in elite controllers of HIV targeted such coupled sets of residues.

The type of analysis outlined above can be helpful in identifying some collective mutational correlations. However, it does not provide a quantitative metric that differentiates relative fitness costs incurred by the virus upon making one set of mutations versus another. The absence of a metric of relative fitness also does not allow for the calculation of in-host evolutionary dynamics in response to different immune pressures. Such calculations can allow one to determine which types of immune responses will either extinguish the virus population or be able to keep the virus cornered for long times before mutational escape can occur. These are the types of immune responses that one would wish to induce by vaccination. The inability to make predictions like the ones noted above also make it difficult to test predictions against *in vitro* and clinical data in order to establish the veracity of the models inferred from sequence data. Another important point is that the methods noted above obtain the structure of correlated mutations by analyzing the population of prevalent or circulating viral strains. But, these virus samples are derived from patients in each of whom a host-pathogen battle had ensued, thus likely forcing certain mutations to evolve that evade the immune response. The resulting prevalent viral strains are not necessarily the intrinsically fittest strains, but those that are fittest in this individual, given her/his immune response. We seek information about intrinsic fitness because we want to learn how to target the virus with immune responses so that escape mutations would have a very low intrinsic ability to replicate and propagate infection (i.e., low intrinsic fitness). The methods noted above do not easily allow for a principled way to analyze and deconvolute the effects of human immune responses to understand how prevalence and fitness of strains are related. For these reasons, while insightful, these methods are insufficient for obtaining the fitness landscape that we seek.

The next section describes methods that can address the pertinent issues for determining the fitness landscapes of HIV proteins. We will also discuss why these methods would require significant

modifications to obtain the fitness landscapes of viruses like influenza, which have a very different evolutionary history.

Inference of prevalence and fitness landscapes of HIV proteins

Inference of the fitness landscape from sequence data can only be done in a statistically meaningful way if a sufficient number of sequences are available. At the time of writing, the number of sequences of whole genomes of HIV strains circulating in the population is insufficient for this purpose. A sufficient number of sequences of the HIV polyproteins are available. Therefore, we will focus on the fitness landscapes of HIV polyproteins. The landscape of the virus thus obtained ignores co-evolution of proteins encoded by different genes. From a biological standpoint, this limitation could be a problem. Fortunately, intergenic epistasis (coupling between mutations) seems to be rare for HIV, but it is not absent. From a methodological standpoint, this is a detail, since the same methods could be applied to whole genomes when a sufficient number become available (although the numerical challenges would increase).

One can begin by seeking to construct a model for the prevalence landscape, or the probability P(**z**) of observing a sequence, **z**, of an HIV polyprotein in the circulating virus population [46]. The sequence data contains this information. It also has information on the probability of observing single mutations at every residue of a protein, double mutations at every pair of residues, triple mutations at every triplet of residues, etc. Any mathematical model for P(**z**) that can recapitulate these mutational correlation functions will accurately describe the prevalence landscape. One way to approach this inference problem is to ask: what is the "least biased" model for P(**z**) that recovers the one- and two-point mutational correlations observed in the available sequence data.

As described in the Appendix to this chapter, exploiting the connection between statistical mechanics and information theory, one may interpret "least biased" to mean the probability distribution P(**z**) which has the maximum entropy subject to the constraints on preserving the observed mutational correlations [47]. A similar approach has been used to infer contacts in protein structures [48,49], correlations between the firing of neurons [50], etc. Related methods have also been employed to study structural properties of HIV protease [51] and inter-protein interactions [52]. Indeed, the development and use of such inference methods is a very active area of research.

For most HIV proteins, a Potts model is appropriate for representing the amino acids at each residue (see details later). This is because a number of amino acids are observed at each residue in the circulating viral strains. For ease of illustration, let us use an Ising representation (amino acids at a residue are either the most frequent or a mutant) to simplify the notation below. The generalization to Potts models is straightforward. With this simplified notation, the quantity we wish to maximize is

$$-\sum_z P(z)\log P(z) - \alpha\left(\sum_z P(z) - 1\right) - \sum_i h_i\left(p_i - \sum_z z_i\, P(z)\right) - \sum_{i<j} J_{ij}\left(p_{ij} - \sum_z z_i\, z_j P(z)\right) \quad (2)$$

where the constraints are enforced through the Lagrange multipliers α, h_i, and J_{ij}. Here p_i and p_{ij} represent the observed one- and two-point mutational correlations, respectively. Maximizing this functional with respect to P (z) yields:

$$P(z) = \frac{e^{-H(z)}}{Z}, \qquad H(z) = -\sum_i h_i z_i - \sum_{i<j} J_{ij} z_i z_j , \tag{3}$$

where the fields h_i and couplings J_{ij} are those that constrain the one- and two-point correlation functions to be the observed ones. The partition function, Z, ensures that the probability distribution is properly normalized. The couplings J_{ij} can have both positive and negative signs and so there are interesting analogies between the form of the Hamiltonian in Eq. 3 and the Hopfield model of neural networks [53], which we will not discuss [54]. Note that positive and negative values of J_{ij} correspond to compensatory and antagonistic effects of double mutations at the i-j pair of residues, respectively.

Once we choose the form of P (**z**) as in Eq. 3, the log-likelihood of the observed sequence data is maximized when the fields and couplings are chosen to be those that predict the observed one and two-point mutational correlations. The likelihood of the observed data is given by the product of the probabilities predicted by the model for observing each sequence in the data. The log of this likelihood divided by the number of sequences, B is thus:

$$\ell = \frac{1}{B} \log\left(\prod_{k=1}^{B} P(z^{(k)}) \right) = -\log Z + \frac{1}{B} \sum_{k=1}^{B} \left(\sum_i h_i z_i^{(k)} + \sum_{i<j} J_{ij} z_i^{(k)} z_j^{(k)} \right), \tag{4}$$

where $\mathbf{z}^{(k)}$ represents the *k*th sequence. Maximizing l with respect to the parameters (the fields and the couplings) yields

$$\frac{\partial \ell}{\partial h_i} = p_i - \sum_z z_i \frac{e^{-H(z)}}{Z} = 0 , \qquad \frac{\partial \ell}{\partial J_{ij}} = p_{ij} - \sum_z z_i z_j \frac{e^{-H(z)}}{Z} = 0 , \tag{5}$$

where $p_i = \Sigma_k z_i^{(k)}/B$, $p_{ij} = \Sigma_k z_i^{(k)} z_j^{(k)}/B$; the sums in the second terms on the right hand side of each part of Eq. 5 are over all sequences, **z**, and so these terms represent model predictions for the one and two-point mutation correlations, respectively. That is, the fields and couplings that maximize the likelihood of the data are also the ones that recapitulate the observed one- and two-point mutational correlation functions.

Inferring the fields and couplings that maximize the likelihood of the data (Eq. 4), given that we know the empirically observed one and two-point mutational correlation functions, is referred to as the inverse Ising (or Potts) problem. The problem is challenging to approach directly through traditional optimization methods because the likelihood depends on the partition function. In principle, as there are 20 possible amino acids (or states) for each residue, the partition function for a Potts model for a protein of length, N, involves summing 20^N terms. So, the computational cost of evaluating the partition function grows exponentially with the length of the protein, and this quantity has to be evaluated many times, once during each iteration of the algorithm that searches for the parameters in Eq. 4 that

maximize l. The conceptually most straightforward way to do this is Boltzmann learning. You start with an initial guess for $\{h_i\}$ and $\{J_{ij}\}$, and generate a large ensemble of sequences according to $p(\mathbf{z}) \sim e^{-H}$, using Monte-Carlo sampling. Briefly, this is done as follows. Attempt to mutate every residue in a given sequence that has been generated with a certain probability (e.g., the mutation rate of HIV). Generate a random number on a computer that is drawn from a uniform distribution and lies between zero and one. The sequence is accepted as part of the sequence ensemble if the random number is less than e^{-H} for this sequence, otherwise it is rejected, and another mutational move is tried. Such a procedure ultimately generates a set of sequences according to the prevalence probabilities predicted by the model. Compute $\{p_i\}$ and $\{p_{ij}\}$ from this ensemble of sequences and compare the values to the empirically observed values in the sequence data. Use any standard gradient descent method to generate a new estimate for $\{h_i\}$ and $\{J_{ij}\}$, and repeat until convergence. Such a method is computationally prohibitive for all but the smallest systems because of the burden of carrying out the Monte-Carlo sampling a large number of times (as it is analogous to computing the partition function many times). For example, with just an Ising representation for p24 sequences (which is 231 amino acids long), such an algorithm required of the order of 10^8 steps in each MC sampling step to obtain converged averages for $\{p_i\}$ and $\{p_{ij}\}$, and 20,000 iterations for parameter convergence.

A number of methods have been developed to address this challenge. The reader is encouraged to peruse descriptions of algorithms like iterative scaling, pseudo-likelihood, perturbation expansions, mean-field (or Gaussian) models, minimum probability flow, etc. Here, we will discuss only one algorithm that has proven to be useful for most HIV proteins.

The Selective Cluster Expansion (SCE) Method

This method is rooted in the rich history of cluster expansions in statistical physics, and takes advantage of the fact that, for most proteins, the matrix of couplings, $\{J_{ij}\}$, is sparse.

Notice that Eq. 4 can be expressed in the following ways:

$$Bl = -B \log Z - \sum_{k=1}^{B} H\left[\overrightarrow{z_k}\right] \qquad (6)$$

The first term on the right-hand side can be interpreted as a Helmholtz Free energy, and the second as the negative of the energy. So, $-Bl$ can be viewed as an entropy, S. Now,

$$-l = \frac{S}{B} = S^* = \log Z - \sum_i h_i p_i - \sum_{ij} J_{ij} p_{ij} \qquad (7)$$

The intensive entropy, S*, is referred to as the intensive cross-entropy between the data and the inferred model. Maximizing the likelihood of the data is tantamount to minimizing this cross entropy.

As we will see shortly, the SCE method solves for the fields and couplings by breaking the problem up into one where one solves for the fields and couplings in disconnected clusters of residues (i.e., with many values of J_{ij} equal to zero). The size of these clusters is progressively grown, starting from independent residues, by adding connections between clusters until the inferred model recapitulates the observed one and two-point mutational correlations. But, before we describe the SCE algorithm, let

us note how in maximizing the likelihood of the data, or minimizing the cross entropy, we must account for the fact that the data itself is usually under sampled because only a finite number of sequences are available.

Suppose that the true probabilities of observing the one and two-point mutational correlations are p_i^t and p_{ij}^t, respectively. If the data consists of B sequences, the probability of observing n mutations at a particular residue is given by the binomial distribution,

$$\binom{B}{n}\left(p_i^t\right)^n\left(1-\ p_i^t\right)^{B-n} \qquad (8)$$

The variance of this binomial distribution, $<(\delta n)^2>$, is $B\ p_i^t\ (1-\ p_i^t)$, and so the error in the data for p_i is

$$\delta p_i = \sqrt{\frac{<(\delta n)^2>}{B^2}} = \sqrt{\frac{p_i^t\,(1-\,p_i^t)}{B}} \sim \sqrt{\frac{p_i\,(1-\,p_i)}{B}} \qquad (9)$$

In replacing p_i^t with p_i in the last approximate equality above, we have ignored higher order terms in 1/B. Now, similarly, $\delta p_{ij} = \sim\sqrt{\frac{p_{ij}\,(1-\,p_{ij})}{B}}$.

For sufficiently small values of B, the error in the data can be significant. In particular, if Bp_i or Bp_{ij} << 1, then the corresponding single or double mutant would not be likely to be observed in the data. This would result in inferring that the magnitude of the corresponding field or coupling is infinite (so, the weight of a sequence with a mutation at i, or the double mutant at the i,j pair, is zero). This inference would be incorrect, and purely a manifestation of under sampling of sequences. For sequences of viral proteins, such as the cases of interest in this chapter, this problem is manifested mostly in estimates of the couplings. This is because, statistically, the errors are greater for higher order mutational correlations (e.g., for uncorrelated mutations, $p_{ij} = p_i p_j$, and Bp_{ij} can be much less than 1 for values of B for which Bp_i is not less than unity).

A principled way to ameliorate the finite sampling problem is obtained by assuming some prior knowledge of the distribution of the couplings. Using Bayes theorem in statistics, we can write:

$$P^L\left[\{\vec{z}\}\parallel\left\{\vec{h},\boldsymbol{J}\right\}\right] *\ P^{prior}\left[\vec{h},\boldsymbol{J}\right] = \mathrm{P}\left[\left\{\vec{h},\boldsymbol{J}\right\}\parallel\{\vec{z}\}\right] * P\left[\{\vec{z}\}\right] \qquad (10a)$$

or equivalently,

$$\mathrm{P}\left[\left\{\vec{h},\boldsymbol{J}\right\}\parallel\{\vec{z}\}\right] = \frac{P^L\left[\{\vec{z}\}\parallel\left\{\vec{h},\boldsymbol{J}\right\}\right] * P^{prior}\left[\vec{h},\boldsymbol{J}\right]}{P\left[\{\vec{z}\}\right]} \qquad (10b)$$

Here, $P^L\left[\{\vec{z}\}\parallel\left\{\vec{h},\boldsymbol{J}\right\}\right]$ is the likelihood of the sequence data, given the fields and couplings, $P^{prior}\left[\vec{h},\boldsymbol{J}\right]$ is the prior estimate for how the fields and couplings are likely to be distributed, $P\left[\{\vec{z}\}\right]$ is the probability of observing the sequence data out of all possible Potts models (appears only as a

normalization), and $P\left[\left\{\vec{h}, \boldsymbol{J}\right\} \parallel \{\vec{z}\}\right]$ is the likelihood of the fields and couplings having specific values given the data.

In we assume that the finite size of the data principally impacts the estimated two-point mutation correlations (see earlier), and in the absence of more precise knowledge, it is reasonable to assume that the couplings, J_{ij}, are distributed as a Gaussian. Therefore,

$$P^{prior}\left[\vec{h}, \boldsymbol{J}\right] \approx P^{prior}\left[\boldsymbol{J}\right] = \frac{1}{(2\pi\sigma^2)^{-\frac{N(N-1)}{2}}} \exp\left[-\frac{1}{2\sigma^2}\sum_{i,j}^{N} J_{ij}^2\right] \qquad (11)$$

N is the number of residues in the protein. Now noting that, by definition, the log likelihood of the data, given the fields and couplings, is $Bl = -BS^*$, the numerator on the right-hand side of Eq. 10 b can be written as:

$$\frac{1}{(2\pi\sigma^2)^{-\frac{N(N-1)}{2}}} \exp\left[-B\left\{S^* + \frac{\gamma}{2}\sum_{i,j}^{N} J_{ij}^2\right\}\right] \qquad (12)$$

with $\gamma = \frac{1}{B\sigma^2}$. Given that $P\left[\{\vec{z}\}\right]$ is just a normalization factor, Eq. 12 is $P\left[\left\{\vec{h}, \boldsymbol{J}\right\} \parallel \{\vec{z}\}\right]$ within such a normalization. We would like the fields and couplings to be those that maximize $P\left[\left\{\vec{h}, \boldsymbol{J}\right\} \parallel \{\vec{z}\}\right]$. This is tantamount to minimizing $S^* + \frac{\gamma}{2}\sum_{i,j}^{N} J_{ij}^2$. That is, to obtain the best estimate for the fields and couplings, given that in a Bayesian sense we apriori expect the couplings to be distributed as a Gaussian centered around zero, we should minimize S* (or maximize the likelihood of the data), with a term added to it. The expression, $S^* + \frac{\gamma}{2}\sum_{i,j}^{N} J_{ij}^2$, is referred to as the regularized cross entropy, and this quantity is minimized in the SCA method. The regularization term effectively penalizes large values of J_{ij} during the optimization procedure, thus partially mitigating the effects of under sampling in the sequence data. Although some ways have been proposed to choose the value of σ^2, the choice really is arbitrary. Often, it is chosen such that the probabilities of observing higher order mutations in the sequence data are also predicted reasonably by the inferred quadratic model.

Eq. 12 also allows us to estimate the errors in the fields and couplings due to finite sampling. Without the regularization term, the probability of inferring particular values of the fields and couplings is proportional to exp [-BS*] as per Eq. 12 (S* depends on $\{h_i\}$ and $\{J_{ij}\}$ as per Eq. 7). If B is very large, this probability distribution will be narrowly distributed around the true values of the fields and couplings. For finite B, the deviations in the fields and couplings from the true values will asymptotically be distributed as a Gaussian with a variance proportional to 1/B, and thus the deviations will scale as $\frac{1}{\sqrt{B}}$.

As we plan to obtain the fields and couplings from the data by minimizing the regularized cross entropy (Eq. 12), it is appropriate to ask if this quantity has a unique minimum. It is easy to show that the regularized cross entropy is convex (i.e., the Hessian of the regularized cross entropy is a positive definite matrix), and so it has a unique minimum. The terms in the Hessian matrix of S* can be obtained by using Eqs 7. They have the following form:

$$\frac{\partial^2 S^*}{\partial h_i h_j} = \langle z_i z_j \rangle - \langle z_i \rangle \langle z_j \rangle \qquad (13a)$$

$$\frac{\partial^2 S^*}{\partial h_i J_{kl}} = \langle z_i z_k z_l \rangle - \langle z_i \rangle \langle z_k z_l \rangle \qquad (13b)$$

$$\frac{\partial^2 S^*}{\partial J_{ij} J_{kl}} = \langle z_i z_j z_k z_l \rangle - \langle z_i z_j \rangle \langle z_k z_l \rangle \qquad (13c)$$

That is, the Hessian of the cross entropy is the covariance matrix of mutational correlations. A covariance matrix is positive definite, which can be most readily illustrated by considering a zero centered two-point covariance matrix, **C**, with individual terms, $\langle z_i z_j \rangle$. For the matrix, **C**, to be positive definite, we require that the scalar, $\overrightarrow{y^T}\ \boldsymbol{C}\ \vec{y}$, for any vector, $\vec{y}$, be positive. This scalar quantity can be re-written as follows to show that it is positive:

$$\overrightarrow{y^T}\ \boldsymbol{C}\ \vec{y} = \sum_{i,k} y_i\, C_{ik} y_k = \langle \sum_{i,k} y_i\, z_i z_k y_k \rangle = \langle (\sum_i y_i z_i)^2 \rangle > 0 \qquad (14)$$

The term added to S* in Eq. 12 in order to regularize it has a Hessian that equals $\frac{\gamma}{2}\boldsymbol{I}$, where $\boldsymbol{I}$ is the identity matrix. So, the regularized S* has a positive definite Hessian, and so has a unique minimum.

The SCE method is a particular algorithm that obtains the fields and couplings by minimizing the regularized cross-entropy. The method reduces the computational burden by taking advantage of the fact that for many biological applications the **J** matrix is relatively sparse. The fields and couplings are inferred for disjoint connected clusters of increasing size using the following recursive algorithm.

The entropy, S, of the system can be written as follows:

$$S = \sum_\Gamma \Delta S_\Gamma \qquad (15)$$

where ΔS_Γ is the change in entropy by including connected clusters of size, Γ, compared to the entropy obtained by considering smaller size clusters; i.e.,

$$\Delta S_\Gamma = S_\Gamma - \sum_{\Gamma' \in \Gamma} \Delta S_{\Gamma'} \qquad (16)$$

where Γ'represents the smaller clusters. We illustrate Eq. 16 explicitly, by writing down formulas for clusters of increasing size below:

$$\Delta S_i = S_i$$

$$\Delta S_{ij} = S_{ij} - S_i - S_j$$

$$\Delta S_{ijk} = S_{ijk} - \Delta S_{ij} - \Delta S_{ik} - \Delta S_{jk} - \Delta S_i - \Delta S_j - \Delta S_k \qquad (17)$$

The recursive SCE algorithm proceeds via the following steps:

1] Start with single residues. In this case, we can directly minimize the cross entropy to obtain an estimate for the fields as the regularization term affects only the values of the couplings. Minimizing the cross entropy (Eq. 7) leads to the following expression:

$$p_i = \frac{\partial \log Z}{\partial h_i} = \frac{\sum_{\{\vec{s}\}} s_i \exp[\sum_i h_i s_i]}{Z} = \frac{\sum_{s_i=0,1} s_i \exp[h_i s_i]}{\sum_{s_i=0,1} \exp[h_i s_i]} = \frac{\exp[h_i]}{1+\exp[h_i]} \qquad (18)$$

Therefore,

$$h_i = \log\left(\frac{p_i}{1-p_i}\right) \quad (19)$$

The entropy, S_i, for each spin is just $-p_i \log p_i - (1-p_i)\log\,(1-p_i)$.

2] Now construct clusters of all pairs of residues, and minimize the cross entropy including the regularization term ($S^* + \frac{\gamma}{2}\sum_{i,j}^{N} J_{ij}^2$) to obtain new estimates for the fields and couplings. The minimization can be carried out using various methods, such as gradient descent, etc. The starting guess for the fields is the one obtained in step 1, and an independent pair approximation can be made to estimate the initial values of the couplings. Upon convergence of the minimization algorithm, compute the entropy, - ($\sum_{\{\vec{z}\}} p\,(\overrightarrow{z}) \ln\, p\,(\overrightarrow{z})$) of the two-residue clusters by sampling configurations of each cluster using the inferred values of the fields and couplings. Then, using Eq. 16 (or 17) compute the change in entropy between the two-residue clusters and that for single residues (obtained in step 1). If the entropy change, ΔS_{ij}, is larger than a threshold value (T), then the particular coupling constant and two-residue cluster is kept. Otherwise, the connection is not included.

3] Now construct three residue clusters starting from the connected pairs, and minimize the cross entropy including the regularization term ($S^* + \frac{\gamma}{2}\sum_{i,j}^{N} J_{ij}^2$) to obtain new estimates for the fields and couplings. The starting guess for the fields and couplings is the one obtained in step 2, and additional ones for the new couplings. Upon convergence of the minimization algorithm, compute the entropy of the three-residue clusters by using the inferred values of the fields and couplings to sample configurations of each cluster, and using the fact that the entropy is ($-\sum_{\{\vec{z}\}} p\,(\overrightarrow{z}) \ln\, p\,(\overrightarrow{z})$). Then, using Eq. 16 (or 17) compute the change in entropy upon considering three-residue clusters compared to the connected residue pairs obtained in step 2. A particular three-residue cluster and corresponding couplings are kept only if the entropy change, ΔS_{ijk}, is larger than a threshold value (T).

4] Keep growing cluster sizes in this way. For a given value of T, the procedure will end with a disjoint set of clusters that cannot be grown further without an entropy change less than T. Using the fields and couplings thus inferred sample configurations using Monte-Carlo simulations to obtain the one and two-point mutational probabilities. If these probabilities compare well with the observed values estimated from the data, then the procedure has converged. If not, repeat steps 1 – 4 with a lower value of T.

Fig. 6 shows a comparison of the mutational correlation functions calculated by sampling sequence configurations according to Eq. 3 using the converged fields and couplings inferred by the SCE algorithm and the sequence data for the p24 protein contained in the Gag polyprotein of HIV. As illustrated therein for the ENV polyprotein, the inferred model can also capture the probabilities of observing higher order mutations, and this is true for other HIV polyproteins.

Here we have provided only a sketch of the SCE algorithm. Many details of the algorithms that are used can be found in reviews co-authored by Cocco and Monason.

<u>Obtaining the sequence data</u>

We downloaded the amino acid multiple sequence alignment (MSA) of HIV-1 Clade B gp160 sequences from the Los Alamos National Laboratory (LANL) HIV sequence database (www.hiv.lanl.gov; accessed 14th May, 2016). To control sequence quality, we excluded sequences (i) labeled by LANL as 'problematic', (ii) with ≥ 2% gaps at non hyper-variable residues

(https://www.hiv.lanl.gov/content/sequence/VAR_REG_CHAR/variable_region_characterization_explanation.html), (iii) with > 10 consecutive insertions at residues where there is an amino acid at every other sequence or (iv) which are outliers as determined by principal component analysis (PCA) (as described in (1)). We then removed sequences labeled or predicted to be CXCR4-tropic, where the prediction was performed using (2). A total of 20043 sequences remained after excluding the above sequences, with these sequences belonging to a total of 1918 patients.

To control residue quality, we excluded residues which are (i) 100% conserved, (ii) contain >90% gaps or are ambiguous, or (iii) are in the hyper-variable regions due to the poor quality of the alignment. We then included eight additional

"artificial" residues whose states reflect the number of residues and N-linked glycans in the four hyper-variable regions we have excluded.

Commented [MOU1]: Has to be generalized

Constructing Potts models

All of the analyses shown above is easily generalized to Potts models, but the numerical cost grows quickly with the number of amino acids that are explicitly considered at each residue of the protein. Therefore, some simplifications are made. The amino acids that arise more frequently at each residue are represented explicitly and the less frequent amino acids are grouped in to a single pseudo amino acid. The number of amino acids represented explicitly at each residue is different.

Let the probability of observing amino acid, a, at residue, i, be denoted by $p_i(a)$. In a representation where the least observed amino acids are lumped in to a single pseudo amino acid, we rewrite the probabilities of occurrence of amino acids ($p_i'(a)$) as follows:

$$\begin{aligned} p_i'(a) &= p_i(a) \quad if\ a < k_i + 1 \\ &= p_i' \quad if\ a = k_i + 1 \\ &= 0 \quad if\ a > k_i + 1 \end{aligned} \tag{20}$$

Here the amino acids labeled 1 through k_i are the k_i most frequently observed amino acids that are represented explicitly, and the others are lumped into one pseudo amino acid which appears with a frequency equal to

$$p_i' = \sum_{a=k_i+1}^{q_i} p_i(a) \tag{21}$$

where q_i is the number of amino acids observed at residue, i.

The choice of k_i determines the fraction of the entropy associated with amino acid variation at residue, i, that is captured in spite of lumping the least frequent amino acids in to a single pseudo amino acid. The total entropy at residue, i, S_i is:

$$S_i = -\sum_{a=1}^{q_i} p_i\,(a) \log p_i\,(a)$$

The entropy upon lumping the least frequent amino acids in to a single pseudo amino acid according to Eq. 20 is:

$$S_i\,(k_i) = -\sum_{a=1}^{k_i} p_i\,(a) \log p_i\,(a) - p_i' \log p_i' \qquad (22)$$

$$S_i\,(k_i) = \varphi\, S_i \qquad (23)$$

where ϕ is the fraction of entropy that is captured by the coarse-grained representation described by Eq. 20.

There are two principal factors that determine the choice of k_i, and hence, ϕ: 1] we wish to capture as much of the observed entropy as possible, which is favored by large values of k_i. 2] we wish to minimize the computational cost which grows with k_i. A way to balance these two factors is obtained by noting that the data for the probability of observing the least frequently observed amino acids is statistically more noisy because of under sampling than that for the more frequently observed amino acids. This can be seen easily by computing the relative error due to under sampling, $\frac{<(\delta p_i)^2>}{p_i^2}$, using Eq. 9. A principled way to determine k_i is to make the error introduced by lumping the less frequent amino acids in to a single pseudo amino acid comparable to the variance in the observed sequence data. The ratio of the error to the variance, given a choice of ϕ (or k_i), is denoted as $\beta_i\,(\phi)$ in the formula below:

$$\beta_i\,(\varphi) = \frac{\sum_{a=1}^{q_i} [p_i(a) - p_i'\,(a)]^2}{\sum_{a=1}^{q_i} \frac{1}{B} [p_i\,(a)\,(1 - p_i\,(a)]} \qquad (24)$$

We can now define a value of β that is averaged over all residues of the protein as follows:

$$\beta\,(\varphi) = \frac{1}{L} \sum_{i=1}^{L} \beta_i\,(\varphi) \qquad (25)$$

The value of ϕ can now be chosen such that $\beta\,(\varphi)$ is roughly unity. Empirically, it is found that, for example, for the large HIV polyprotein, ENV, the resultant value of ϕ is 0.95. Typically, values of ϕ above 0.9 seem to be sufficient for accurate estimates (as measured by tests against experimental data).

Connection between prevalence and fitness of HIV strains

The model that we infer by following the procedure above describes the prevalence of circulating strains of HIV. One could argue that the more prevalent strains are also the ones that are intrinsically more fit, but this relationship can be complicated by several factors. The sequences of strains that are used to infer the prevalence model are derived from patients whose immune systems have battled the virus. Mutations that make the virus intrinsically less fit may allow the virus to evade the immune response,

thus making a strain bearing these mutations effectively more fit and prevalent in a particular host. How does the host-pathogen riposte affect the relationship between the prevalence and intrinsic viral fitness?

Computer simulations aimed to mimic the way in which the virus samples were collected, with explicit consideration of host immunity, may be able to shed some light. One realization of such an effort is described in [68]. As schematically depicted in Fig. 7, an "in silico" person is infected with N copies of HIV, and a host-pathogen battle ensues (described below). At a randomly chosen time, a randomly chosen virus strain from this in silico person is transmitted to another host, and the infection begins anew with N copies of the transmitted strain. HIV is a chronic infection and its proteins are extensively subjected to T cell responses that target peptides bound to MHC molecules. Because the newly infected person is likely to have a different MHC genotype and antibody responses also vary across hosts, in the calculations the virus evolves in response to different immune responses in each in silico person (see below). At a randomly chosen time, a virus sample from this person infects another new host, and this process is continued. At separate randomly chosen times, a virus sample from the current virus population within each host is recorded, and these samples can be thought to represent a sequence database that is later used to infer the prevalence model.

Within each person, the fitness of the virus is modulated by the immune response. Let us assume that the intrinsic fitness is described by the inferred prevalence landscape. As will become clear below, the results of the computer simulations will inform us about the veracity of this assumption. The immune response, which acts on particular residues of viral proteins that are part of the peptides bound to MHC or antibody epitopes, can be modeled as a set of external fields that act at a few points to promote mutations. Again, for simplicity, in an Ising representation, the effective Hamiltonian (or in-host fitness) can be written as

$$H_{eff} = -\sum_{i=1}^{N} h_i z_i - \sum_{i<j} J_{ij} z_i z_j + \sum_{i'=1}^{k} b_i' z_i' \qquad (26)$$

The fields, b_i', act only at residues being targeted by T cells and antibodies in the particular person under consideration. For all other sites b_i' is zero. The superscript prime denotes the set of targeted sites and note that they may not be contiguous. There are short-range correlations between the locations at which these fields act (within peptides or antibody epitopes), which are ignored for simplicity. The number and location of the residues at which immune pressure is applied in each individual comes from a statistical distribution, whose choice can be guided by clinical data. Human immune responses, especially those due to T cells, are extraordinarily diverse [69] [4]. Thus, the same residues are not consistently targeted among different hosts. For example, of the 363 residues in the immunogenic structural proteins, p17 and p24, only 46 are targeted by T cells in more than 10% of humans, none by more than 23%, and 146 residues are not targeted at all [40]. Thus, as a rough approximation, the targeted sites in a HIV protein within each host can be selected from a uniform distribution across the entire protein, mimicking the high diversity of targeted epitopes by humans with diverse genotypes. The number of targeted residues, k, may be chosen randomly between 0 and n_{max} (=6 in the calculations

due to Shekhar et al). At each targeted site, a reasonable way to set the value of the field b_{si} is to choose it from a Gaussian distribution, whose mean and variance are the same as for the inferred h_i. Since we have assumed that the latter correspond to the intrinsic fitness, this choice of the values of b_{si} ensures that some immune responses will promote mutations.

Viral dynamics within each host can be simulated using a model similar in spirit to Wright-Fisher models in evolutionary biology (Fig. 8). Each residue in proteins in each viral strain in an individual can mutate with a certain probability per site (that of HIV). The new viral strains produce progeny (are positively selected) with a probability equal to

$$P_{surv}(z) = \frac{e^{-H_{eff}}}{1+e^{-H_{eff}}} \quad (27)$$

In Eq. 27, the fitness of the strain, z, is compared to the consensus sequence, which is assumed to have the best fitness (equal to 1). The fitness of the strain under consideration can also be compared to the average fitness of all the strains present at that time, <exp (-H (z))>, and the qualitative results turn out to be the same. After this selection step, the number of viruses is then scaled back to a total of N strains that reflect the same probability of occurrence of each strain that survived selection, and the calculation is repeated. The assumption of constant population size may be appropriate as it is expected that most samples of viruses were drawn from patients who were in the chronic stage of infection when viral load does not fluctuate much. Moreover, several studies have shown that, for large enough population sizes, Wright-Fisher like dynamics asymptotically approach results of calculations that do not impose constant population size. Shekhar et al. studied population sizes that ranged from $2\times10^3 - 5\times10^5$.

Commented [MOU2]: Is this consistent with estimated effective population sizes

Upon carrying out simulations following the method described above one obtains a set of sequences derived from the in silico patients that reflects host-pathogen riposte. One can now obtain the one and two-point mutational correlations from this set of sequences and compare them to those corresponding to the sequences derived from the real patients (i.e., the ones used to infer the prevalence landscape). If these sets of mutational correlations are the same, then the prevalence landscape is the same as the fitness landscape. This is because the assumed intrinsic fitness landscape is the prevalence landscape which, by construction, fits the mutational correlation functions observed in the sequences derived from actual patients. So, if the sequences obtained from the in silico patients (post host-pathogen riposte) exhibit the same mutational correlations, the prevalence landscape inferred from these sequences would be statistically the same as the assumed intrinsic fitness landscape. Shekhar et al carried out such calculations for the p17 HIV protein, and found that the two sets of mutational correlations are not the same (Fig. 9). So, the prevalence landscape is not the fitness landscape.

However, the two sets of mutational correlations are statistically monotonically correlated. If the Hamiltonian in Eq. 3 was of the ferromagnetic form, Griffiths theorem in statistical physics [70] would lead to the conclusion that the fields and coupling constants are also monotonically correlated. Thus, the fitness and prevalence landscapes would be related by a simple shift. This is sufficient for vaccine design as we only need to know the relative fitness of strains, not the absolute fitnesses, to determine which epitopes would be more efficacious to target with an immune response. However, the

Hamiltonian is not of this form, as the couplings, J_{ij} have both signs, and for such a Hamiltonian, an analog of Griffiths theorem does not exist.

The numerical simulations, although useful, seem to have led us to a dead end with respect to determining the relationship between the prevalence and fitness landscapes. Below, we discuss an approximate analysis which, informed by the simulation results, sheds more light.

Eigen considered the evolution of a swarm of species (e.g., virus strains) in the limit of an infinite total virus population. The resulting deterministic equation for the temporal variation of the frequency of each strain is given by [71]:

$$\frac{dx_\alpha}{dt} = f_\alpha\, x_\alpha - \sum_{\beta\neq\alpha} W_{\beta\alpha}\, x_\alpha + \sum_{\beta\neq\alpha} W_{\alpha\beta}\, x_\beta - x_\alpha \sum_\gamma f_\gamma\, x_\gamma \quad (28)$$

Here, x_α is the frequency of strain α, $W_{\beta\alpha}$ is the rate of mutation from strain α to β, and f_α is the replication rate of strain α (which corresponds to its fitness in our treatment above). The last non-linear term is required to ensure that the strain frequencies, x_α, are normalized. From a biological standpoint, this term embodies competition between strains and so the growth rate of strain α, f_α, relative to the average replication rate of the prevailing strains, $\sum_k f_\gamma x_\gamma$ is pertinent.

Leuthäusser showed that Eigen's equation describing a non-equilibrium process is isomorphic with the equilibrium statistical mechanics of a two-dimensional Ising model [72]. To see how this isomorphism emerges, define the following variable:

$$y_\alpha = x_\alpha\, exp\left[\int_0^t dt' \sum_k f_\gamma x_\gamma\right] \quad (29)$$

Using Eq. 28, the derivative of y_α is given by:

$$\frac{dy_\alpha}{dt} = f_\alpha y_\alpha - \sum_{\beta\neq\alpha} W_{\beta\alpha} y_\alpha + \sum_{\beta\neq\alpha} W_{\alpha\beta} y_\beta \quad (30)$$

Note that the variables, y_α are not normalized anymore. The linear equation above can be written in discrete form as follows:

$$y_\beta(1) = \sum_\alpha W_{\beta\alpha} y_\alpha\,(0); \quad \vec{y}\,(1) = \underline{W}\,\vec{y}\,(0) \quad (31)$$

We may view $y_\beta(1)$ to be proportional to the number of progeny of type β in generation 1, given that the number of strains of each type in generation, 0, is proportionally represented by the vector, **y** (0). The strain frequencies can be obtained from the vector **y** (1) as follows:

$$x_\alpha\,(1) = \frac{y_\alpha\,(1)}{\sum_\alpha y_\alpha\,(1)} \quad (32)$$

Each element of the matrix, W, is proportional to a transition probability for going from one strain to another in a single generation, as is made clear by the following argument. Suppose that our starting swarm of viruses is composed of a single strain, α, as is often the case for HIV infection since clinical evidence suggests that usually a single strain succeeds in establishing infection [REF]. If the number of copies of the initial strain, α, is N_0, then the numbers of copies of the other strains, β, in generation, 1, is given by:

$$y_\beta(1) = W_{\beta\alpha} N_0 \qquad (33)$$

and the frequency or probability of observing strain β in generation 1 is:

$$P_\beta(1) = x_\beta(1) = \frac{y_\beta(1)}{\sum_\beta y_\beta(1)} = \frac{W_{\beta\alpha} N_0}{\sum_\beta W_{\beta\alpha} N_0} = \frac{W_{\beta\alpha}}{\sum_\beta W_{\beta\alpha}} \propto W_{\beta\alpha} \qquad (34)$$

The matrix, W, also has a Markovian character as it describes (within normalization factors) the transition probabilities between strains in one evolutionary generation, regardless of past history. This Markovian property of the process implies that we can represent the probability of observing the vector **y** (n) after n generations, given the vector, **y** (0), P [**y** (n); **y** (0)], as follows:

$$P[\boldsymbol{y}(n); \boldsymbol{y}(0)] = \sum_{\boldsymbol{y}(n-1)} P[\boldsymbol{y}(n); \boldsymbol{y}(n-1)]\, P[\boldsymbol{y}(n-1); \boldsymbol{y}(0)] \qquad (35)$$

Where the sum runs over all possible realizations of the elements of the vector **y** after (n-1) generations. This process can be repeated as follows:

$$P[\boldsymbol{y}(n); \boldsymbol{y}(0)] = \sum_{\boldsymbol{y}(n-1)} \sum_{\boldsymbol{y}(n-2)} P[\boldsymbol{y}(n); \boldsymbol{y}(n-1)]\, P[\boldsymbol{y}(n-1); \boldsymbol{y}(n-2)]\, P[\boldsymbol{y}(n-2); \boldsymbol{y}(0)]$$

$$= \sum_{\boldsymbol{y}(n-1)} \sum_{\boldsymbol{y}(n-2)} \sum_{\boldsymbol{y}(n-3)} P[\boldsymbol{y}(n); \boldsymbol{y}(n-1)\, P[\boldsymbol{y}(n-1); \boldsymbol{y}(n-2)]\, P[\boldsymbol{y}(n-2); \boldsymbol{y}(n-3)]\, P[\boldsymbol{y}(n-3); \boldsymbol{y}(0)]$$

$$= \sum_{\boldsymbol{y}(n-1)} \sum_{\boldsymbol{y}(n-2)} \cdots \cdots \sum_{\boldsymbol{y}(2)} \sum_{\boldsymbol{y}(1)} P[\boldsymbol{y}(n); \boldsymbol{y}(n-1)]\, P[\boldsymbol{y}(n-1); \boldsymbol{y}(n-2)] \cdots \cdots \cdots P[\boldsymbol{y}(2); \boldsymbol{y}(1)]\, P[\boldsymbol{y}(1); \boldsymbol{y}(0)] \qquad (36)$$

Within normalization factors, the transition probabilities across one generation in the last line of Eq. 36 is the matrix, W. So, P [**y** (n); **y** (0)] can be calculated as follows. Repeated operation of the matrix W on a given **y** (0) (i.e., repeated matrix multiplication) yields the probability of a particular set of intermediate vectors, **y** (n), **y** (n-1).....**y** (2), **y** (1), or the probability of a particular evolutionary trajectory. To obtain P [**y** (n); **y** (0)] we now need to sum over all such evolutionary trajectories that are n generations long by summing over all possible intermediate states. In other words, we have a path integral formulation, wherein the probability of any given evolutionary trajectory that starts from some founder strain is given by the nth power of the matrix W operating on this sequence, and the probability of observing various strains at generation n, P [**y** (n); **y** (0)], is given by summing over all possible evolutionary trajectories. If the matrix, W, contains information about the intrinsic fitness of strains and

the immune pressures that are in effect in different generations, P [**y** (n); **y** (0)] can be interpreted to be the prevalence landscape.

Fig. 7 shows a pictorial depiction of an evolutionary trajectory. The spins (up or down) represent the identity of the amino acid at a particular residue in a sequence. The sequence depicted in a particular row is the progeny of the sequence above it. Thus, a particular configuration of this 2-dimensional Ising magnet represents an evolutionary trajectory. The weight of an evolutionary trajectory is the summand in Eq. 36. If this weight was known, we could sum over evolutionary trajectories of length n generations by sampling configurations of the evolutionary trajectories according to this weight. Keeping track of the frequency with which sequences were sampled in the n^{th} generation would provide us with P [**y** (n); **y** (0)]; i.e., the prevalence landscape. Given that each term in Eq. 36 can be identified with the matrix $\underline{W}$, to obtain the weights of each configuration of the Ising representation of evolutionary trajectories, we have to obtain the elements of $\underline{W}$.

How can we compute the elements of the matrix, $\underline{W}$? If μ is the mutation rate per replication cycle, $d_{\alpha\beta}$ is the Hamming distance (i.e., the number of residues wherein strains α and β have different amino acids), and λ is the number of possible amino acids (in the Ising representation we are using for this illustrative calculation, $\lambda = 2$), then $W_{\beta\alpha}$ can be written as:

$$W_{\beta\alpha} \;=\; f_\alpha \,(1 - \mu)^{N - d_{\alpha\beta}} \left(\frac{\mu}{\lambda - 1}\right)^{d_{\alpha\beta}} \qquad (37)$$

Where, like earlier, N is the length of the protein (i.e., the number of residues).

In the formulas defining the prevalence landscape in the Ising representation (Eq. 3), the two types of possible amino acids (wild type and mutant) at residue, i, corresponded to values of z_i equal to 0 and 1. Some of the expressions to follow are more transparent in the z_i equal to $\pm\, 1$ representation, and thus it is employed below; this will be made consistent later by a simple transformation when we explicitly incorporate considerations of strain fitness. With the {1, -1} representation, we can express $d_{\alpha\beta}$ as:

$$d_{\alpha\beta} \;=\; \sum_{i=1}^{N} \frac{(z_i^\alpha \;-\; z_i^\beta)^2}{4} \;=\; \frac{1}{2}\left[N - \sum_{i\,=\,1}^{N} z_i^\alpha z_i^\beta\right] \quad (38)$$

Using the fact that $a^x = \exp [x \ln a]$, and Eq. 38, allows us to express $W_{\beta\alpha}$ as:

$$W_{\beta\alpha} \;=\; f_\alpha \, h \, exp\left[M \sum_i z_i^\alpha z_i^\beta\right] \qquad (39)$$

where h and M are defined as follows:

$$h \;=\; exp\left[\frac{N}{2}\, ln\,(\mu\,(1-\mu))\right];\;\; M \;=\; \frac{1}{2}\, ln\, \frac{(1 - \mu)}{\mu} \quad (40)$$

Notice that M is positive for a mutation rate of less than half, which is true for any realistic virus.

Let $\overrightarrow{z^0} = \{z_i^0\}$ be a vector of length N that specifies the values of the spin states corresponding to the amino acids at each residue, i, for the strain in generation 0; i.e., the founder strain. Eqs 39 and 40 state that the probability of occurrence of the strain, $\overrightarrow{z^1} = \{z_i^1\}$, in generation 1, is proportional to $W_{10} = f_0\ h\ exp\left[M\ \sum_i z_i^1 z_i^0\right]$. Generalizing this for any two strains in consecutive generations, q and q + 1, we can write $W_{q+1\,q} = f_q\ h\ exp\left[M\ \sum_i z_i^{q+1} z_i^q\right]$.

As per our previous considerations (Eq. 36, the discussion immediately following it, and Fig. 7), given a particular founder strain, the probability of observing a particular evolutionary trajectory, T, is proportional to the product:

$$\prod_{q=0}^{n-1} W_{q+1\,q} = \prod_{q=0}^{n-1} f_q\ h\ exp\left[M \sum_i z_i^{q+1} z_i^q\right]$$

$$= h^n\ exp\left[\sum_q ln\ f_q + M \sum_q \sum_i z_i^{q+1} z_i^q\right] \quad (41)$$

The normalized probability of observing a particular trajectory, T, is:

$$P\ (T) = \frac{1}{Z} exp\left[\sum_q ln\ f_q + M \sum_q \sum_i z_i^{q+1} z_i^q\right] = \frac{1}{Z} exp\ (-H\ (T))$$

$$Z = \sum_T exp\ (-H\ (T)) \quad (42)$$

where H (T) is the "Hamiltonian" corresponding to a particular trajectory, and Z is the "partition function" obtained by summing over trajectories. In this path integral formulation, the non-equilibrium dynamical problem under consideration is transformed in to studying the equilibrium statistical mechanics of a 2-dimensional ising model (Fig. 7). Each configuration of the Ising magnet represents an evolutionary trajectory. Each row z_i^q of this Ising model corresponds to a particular strain, and the next row z_i^{q+1} is its progeny. Let us now assume that the fitness of a strain under immune pressure is as noted in Eq. 26. Then, Eq 42 can be rewritten as:

$$P\ (T) = \frac{1}{Z} exp\left[\sum_q \sum_i h_i z_i^q + \sum_q \sum_{i \neq j} J_{ij}\ z_i^q z_j^q - \sum_q \sum_{i'} b_{i'}^q z_{i'}^q + M \sum_q \sum_i (1 - 2\, z_i^{q+1})(1 - 2\, z_i^q)\right]$$
$$= \frac{1}{Z} \exp[-H\ (T)] \quad (43)$$

In writing Eq. 43, we transformed the spin variables to the {0,1} Ising spin representation from the ± 1 representation for consistency among all the terms. The in-row couplings (Fig. 7) in Eq. 43 reflect the intrinsic fitness of a viral strain modulated by the immune response, and there is a nearest-neighbor ferromagnetic (recall that M is positive) coupling across rows that reflects mutational fidelity originating

from the fact that the mutation rate is less than half. The immune pressure represented by the fields $\{b_i^{'}\}$ act only on a few selected residues in any one human, and different humans can impose immune pressure at different residues because they have different MHC genes and antibody responses.

Note that only the probability of observing a strain in the last generation, n, has biological meaning. Just as in the Wright-Fisher simulations, we could assume that the intrinsic fitness landscape is the one inferred from the prevalence data and numerically sample trajectories of varying length, n (to mimic that the virus evolved for different times in each individual) in each of whom a different realization of the immune fields applies (drawn from the same statistical distribution as for the simulations). The probability of obtaining different strains in generation, n, thus obtained could then be compared with predictions from the assumed fitness landscape (inferred prevalence landscape) to obtain insights in to the relationship between the prevalence and intrinsic fitness landscapes.

The formulation of Eigen's equation as an isomorphic problem in equilibrium statistical mechanics, however, enables us to obtain insights via a standard approximate analysis; viz., to carry out Variational calculations using the Feynman-Bogoulibov bound. The partition function, Z (Eq 42), can be rewritten as:

$$Z = \sum_T exp\,(-(H\,(T) - H_v\,(T)))\,exp\,(-Hv\,(T));\ \frac{Z}{Z_v} = \langle exp\,(-(H\,(T) - Hv\,(T))\rangle_v \quad (44)$$

where H_v (T) is a different Hamiltonian, $\langle exp\,(-(H\,(T))\rangle_v$ indicates the average of the quantity in angular brackets using the Hamiltonian, H_v, and Z_v is the partition function corresponding to H_v (T). Jensen's inequality now tells us that the average of the exponential of a quantity (a convex function) is greater than or equal to the exponential of the average of the quantity, which leads us to:

$$ln\,Z \geq ln\,Z_v - \langle (H\,(T) - H_v\,(T))\rangle_v \quad (45)$$

If H_v is chosen properly, the right-hand side of Eq. 45 can be determined exactly. The resulting expression can then be maximized with respect to parameters in H_v to obtain an approximate Hamiltonian that variationally bounds the real Hamiltonian. Analysis of the properties of this variationally optimized Hamiltonian can lead to insights.

The Hamiltonian used to infer the prevalence landscape is of quadratic form (Eq. 3), and so we choose H_v to be quadratic form:

$$H_v(T) = \sum_q \left[-\sum_i a_i^q s_i^q - \sum_{\alpha \neq \beta} K_{ij}^q\, s_i^q s_j^q \right] \quad (46)$$

where K_{ij}^q and a_i^q are in-row couplings and fields, respectively, the optimal values of which are obtained using the Variational approximation. The resulting equations are:

$$\frac{\partial \ln Z_v}{\partial\, a_i^q} - \frac{\partial < H\,(T) - H_v\,(T) >}{\partial\, a_i^q} = 0$$

$$\frac{\partial \ln Z_v}{\partial K_{ij}^q} - \frac{\partial < H(T) - H_v(T) >}{\partial K_{ij}^q} = 0 \qquad (47)$$

Solving these equations while taking care to separate the surface layer from the others (see earlier comment about only the probabilities of observing the surface layer being interpretable) obtains the following formulas for the Variationally optimal couplings and fields in the nth (surface) generation:

$$a_i^n = h_i - 2M(1 - 2 < z_i^n >_v) + b_i'^n$$

$$K_{ij}^n = J_{ij} \qquad (48)$$

In obtaining the first line in Eq. 48, we have assumed that $< z_i^n >_v = < z_i^{n-1} >_v$. Since $< z_i^n >_v$ is the average of whether a mutant or wild-type amino acid is present at residue i in generation, n, this approximation should be reasonably good for $\mu < 1$ and sufficiently large, n. The details of the algebra in going from Eq 47 to Eq. 48 is left as an exercise for the reader, and is available in reference [].

Interestingly, the Variational approximation predicts that the effects of the immune fields and mutational fidelity does not alter the couplings observed in the prevalence landscape from the intrinsic values. The expression for a_i^n, however, shows that the fields inferred from sequence data on prevailing strains should be different from the ones that describe the intrinsic fitness landscape. Notice also, that a_i^n is not determined explicitly by Eq. 48 because its right-hand side depends on H_v through $< z_i^n >_v$. H_v cannot be determined without knowing a_i^n. So, Eq. 48 must be solved self-consistently to obtain a_i^n, as is typical for such variational, or mean-field, calculations.

However, several key insights in to the biological problem at hand can be obtained without carrying out such a calculation. In order to study the relationship between the inferred prevalence landscape and the fitness landscape, the term representing the immune pressure in Eq. 48 must be averaged over that corresponding to people with diverse genotypes. This is because the sequence data was collected from diverse patients, or equivalently, the circulating strains have evolved in response to diverse immune pressures. One convenient way to visualize and analyze the problem at hand (see earlier comments following Eq. 43) is to consider a single long trajectory that started with a founder strain and wherein the imposed immune fields changed location after a varying number of generations. The switch in locations of the immune fields corresponds to a new infected person. The sequences were collected from diverse patients (i.e., after different numbers of evolutionary generations, n. So, to estimate a_i^n, we must average over various values of n. Let us denote the value of $< z_i^n >_v$ averaged over various generations, n, as $< z_i >_v$, and the value of $b_i'^n$ averaged over many generations and human genotypes (or types of immune pressure) as $< b_i >$. After these steps of averaging, we obtain:

$$a_i = h_i - 2M(1 - 2 < z_i >_v) + < b_i > \qquad (49)$$

Therefore, the Variationally optimized Hamiltonian that incorporates the effects of immune pressure, mutational fidelity, and intrinsic fitness (i.e, the Hamiltonian corresponding to the prevalence landscape) can be written as:

$$H_v = -\sum_{i \neq j} J_{ij}\, z_i\, z_j - \sum_i [h_i - 2M\,(1 - 2 < z_i >_v) + < b_i >]\, z_i \qquad (50)$$

Eqs. 49 and 50 provide insights in to the relationship between the prevalence and fitness landscapes as well as the underlying biological factors that determine this relationship for HIV.

Perhaps, the reader has wondered why the external fields corresponding to the immune pressure are assumed to depend only upon the current generation (or time). Why can a person that was originally infected be infected later along the evolutionary trajectory, and mount effective memory responses elicited earlier. This would correlate the immune fields across time, thus significantly complicating the analysis. The simplification originates in the fact that, although a few HLA-epitope combinations have been associated with better outcome in infected persons [24], HIV is not known to have been cleared in any infected person. So, the global population of HIV has largely not been persistently subjected to a few effective classes of natural or vaccine-induced immune responses. Therefore, the global HIV population has not evolved in narrowly directed ways to avoid past effective herd memory responses in the human population. Contrast this situation with that for influenza, which has been subjected to effective vaccine-induced or natural antibody responses, which continuously drive the evolution of the global influenza population in specific directions to evade such responses. The simplification wherein the immune fields do not depend on past times for HIV cannot be made for studying influenza.

Note that because h_i is negative and $< b_i >$ is positive, Eq. 50 implies that the effect of the immune pressure is to make the virus more mutable than the intrinsic fitness of the virus would suggest. This reflects that the immune pressure imposed at certain residues promotes mutations because the resulting mutant viral strain is no longer subject to immune attack, thus making the mutant effectively more prevalent. Thus, human immune pressures promote the exploration of sequence space by the virus population. Because of the great diversity of immune (especially T cell) responses in the human population, most regions of the viral proteome are targeted by a small fraction of people; so, b_i' acts only for a small number of evolutionary generations corresponding to the fraction of individuals who target residue, i, with their immune responses. Thus, we are led to the conclusion that $< b_i >$ has a relatively small value.

The term, $< z_i >_v$, is expected to be less than one half. This is because the mutation probability is less than half, and also because HIV is a chronic infection that is transmitted from one host to another. If the virus is forced to make an immune evading mutation in a particular individual and this mutation has an intrinsic fitness cost, such deleterious escape mutations can revert over time when the virus is transmitted to a new host (whose immune response likely will not target this residue) [73]. Therefore, in Eq. 49, the term, $-2M\,(1 - 2 < z_i >_v)$, adds a negative value to h_i, making it more difficult to observe a mutation at residue, i, in the prevalence landscape. If we compare two strains, this term is added to the Hamiltonian for every residue for which the amino acids in the two strains are different. The prevalence of two strains that are equally intrinsically fit could thus be different. The strain that has to many more mutations compared to some reference strain will be less prevalent. However, this confounding effect of mutational fidelity is likely to be significant only if we compare strains that differ

from each other by many mutations. For strains that differ by just a few mutations from each other, as those that evolve in a single patient over time, this effect is likely to be small. Note also that this effect is further mitigated by recombination of different viral strains during infection and replication.

The above arguments and analyses suggest that while human immunity is an important driver of HIV evolution, its overall effect on the relationship between prevalence and fitness is likely to be perturbative on the relationship between prevalence and fitness. The effects of mutational fidelity should be significant only if we compare strains that differ from each other by large mutational distances. A way to summarize our conclusions is that, within some range of mutational distances, the HIV population is at steady state. Thus, for purposes of comparing the fitness of strains within this range of mutational distances, the probabilities of prevailing viral strains using maximum entropy models may reflect their rank order of intrinsic fitness. This is decidedly not the case for influenza as it is strongly and continuously driven out of equilibrium for reasons that were noted earlier. Using very different methods than that described above, recent work has also shown that patterns of HIV diversity over long times in single infected patients are mirrored by those across different infected individuals, supporting the claim that universal information about HIV fitness can be derived from prevalence data [74].

We can test the arguments developed above in an approximate manner without carrying out the self-consistent calculation for $< z_i >_v$. We can estimate the value of this quantity from the Wright-Fisher like simulations described earlier as these results include the effects of immune pressure and mutational fidelity. This approximation is likely to be accurate under two assumptions: 1] The Variationally determined Hamiltonian is a good approximation to H (T), and therefore, to Eigen's equation. 2] Eigen's equation (which holds for an infinite population size) is a good approximation to the simulations which were carried out with a finite population size. The latter is expected to be true when the mutation rate multiplied by the finite population size is greater than unity, which is the case in the simulations. The value of $< b_i >$ can also be obtained from the simulations, and they reflect clinical information on the characteristics of the immune pressure (see discussion following Eq. 26).

Using the values of $< z_i >_v$ and $< b_i >$ estimated from the simulations it was found that, for the p17 HIV protein, with high statistical accuracy, the Hamiltonians corresponding to the intrinsic fitness and prevalence are monotonically correlated (Fig. 9). This result suggests that the relative intrinsic fitness of HIV strains can potentially be adequately described by inferring a prevalence landscape using a maximum entropy formulation.

Of course, in individual hosts the virus evolves to evade host immunity, forcing HIV to adapt and explore the sequence space. If a mutation that evades host immunity comes at a substantial fitness cost to the virus, compensatory mutations often arise to restore lost fitness, and so mutations at these combinations of residues are observed more frequently than by chance in the circulating virus population. Similarly, some combinations of mutations that are especially deleterious for the virus are observed less frequently than by chance. These correlations, which reflect intrinsic viral fitness effects observed because host-pathogen riposte forces the virus to sample sequence space, are reflected in our

inferred landscape. Thus, the inferred landscape describes the collective mutational pathways that HIV can uses to evade host immunity and those that it cannot.

Tests against in vitro measurements and clinical data

The approximate calculations and biological arguments noted above seem reasonable. But, they pertain to a very complex problem, and could also be incorrect. Therefore, the veracity of our conclusion that the inferred prevalence landscape for HIV proteins is an adequate proxy for intrinsic fitness can only be established by testing predictions against *in vitro* experiments and clinical data. This is consistent with the scientific method, and common sense. Consider *in vitro* experiments first.

Tests against in vitro measurements of fitness

The inferred fitness landscape can be used to calculate the value of the Hamiltonian (or "energy") corresponding to particular mutant sequences relative to a reference sequence. The model would predict that the replicative fitness of the mutant strain relative to that of the reference sequence should correlate negatively with the energy difference between the mutant strain and the reference sequence (Eq. 3). The mutant sequences for which predictions are made can be generated through site-directed mutagenesis, and then their relative fitness can be measured by assaying their growth rates when placed in culture with human cells that HIV can infect. Notice that such experiments are carried out in the absence of immune pressure, and so reflect the intrinsic fitness of mutant strains. Fig. 10a shows such a comparison between experiments and model predictions for 43 strains of HIV with mutations in the Gag polyprotein [46,79]. As is evident, while not perfect, the comparison is reasonably good. The ENV polyproteins comprise the spike on the surface of HIV particles. The fitness landscape of ENV is the hardest to infer because of its long length and much higher variability compared to other HIV polyproteins. Fig. 10b shows a comparison between predictions based on the inferred landscape and roughly 100 *in vitro* measurements of infectivity. Again, the comparison is reasonable.

We can ask how whether it is important to infer a landscape that includes the effects of epistatic interactions. Would the quality of the predictions be worse if we inferred a model with the fields only? Fig. 10c shows results comparing predictions made with such a landscape for the Gag mutants that were tested in Fig. 10a. As can be seen, the comparison is worse. But, the study with Gag mutants was done as a part of a collaboration between immunologists and the physical scientists who inferred the landscape. They were interested in testing whether the predictions made for epistatic interactions were correct. So, some of the mutant strains that were tested included strains with mutations that were predicted by the inferred landscape to be strongly coupled (as per the magnitude of the corresponding J_{ij} values). The fitness measurements for the ENV mutant strains were conducted independently, and before, the predictions using the fitness landscape. As has been noted before, the matrix, **J**, is sparse because mutations at many residues in a protein are not coupled. As a result, if you take a random set of mutant strains (as for ENV) and compare fitness predictions using landscapes inferred with or without the couplings, the differences are not substantial as many strains do not have mutations at residues that are coupled (Fig. 10d). But, as we have noted earlier, because of the high mutation and replication rates of HIV *in vivo*, rare sets of mutations in coupled residues could be sampled to affect the virus' ability to

evade human immunity. We will shortly discuss comparisons with in vivo clinical data that will make this point vivid. However, the importance of these effects can be seen in other ways as discussed below.

Tests against structural data

We anticipate that mutations in residues that are in contact, or in close spatial proximity, in the functional structure of a protein or protein complex are more likely to be coupled. Thus, the values of the elements of the matrix **J** corresponding to these residues should be relatively high. Evidence for this for HIV was provided even by the simple analyses of covariance matrices using random matrix theory outlined in an earlier section of this chapter. For example, structural reasons were noted for why set of residues in the p24 protein appeared to co-evolve and disfavor multiple simultaneous mutations (corresponding to negative elements of the "cleaned" correlation matrix). These residues are shown superimposed on the structure of p24 in Fig. 11a. This visualization does not provide much insight in to why the identified residues might be negatively coupled or co-evolve. Six p24 proteins form hexamers that tile the viral capsid. In Fig. 11b, the p24 residues marked in Fig. 11a are superimposed on the structure of this hexamer and that of the interface between hexamers. Now, one can see vividly that a majority of these residues are involved in intra-hexamer and inter-hexamer contacts between p24 proteins. Thus, they are likely to co-evolve and simultaneous mutations in pairs of such residues are likely to destabilize the capsid and make the virus unviable.

More generally, a large number of studies have been conducted to study bacterial protein families using methods similar to the maximum entropy models that we are considering. Methods have been developed to determine whether two residues are directly coupled (or in contact) based on the **J** matrix. One such method is the Direct Coupling Analysis (DCA), which we do not discuss here. The interested reader is directed to [xx].

Let us explore whether applying the DCA method to the inferred fitness landscape of the ENV protein can predict residues that are in contact in the viral spike of HIV. This is a prediction that strictly relies on the importance of the **J** matrix. Note also that since the **J** matrix was inferred from viral sequences *in vivo*, we expect that it reflects important contacts in the functionally relevant trimeric spike of HIV, not monomers of the constituent ENV proteins. The native HIV spike is very unstable and so has proven to be difficult to crystallize. A mimic has been prepared by stabilizing the trimer with a few disulfide bonds. The structure of this mimic, called SOSIP, is available. So, it may be appropriate to test predicted contacts against this structure. Fig. 12a shows a comparison of the predicted contacts (function of **J**) to the crystal structure. As is evident, the top twenty predicted values are true positives with high probability. A large fraction of the false positives is in the V2 loop of the trimeric spike or in CD4 contact residues. As noted earlier, upon binding to CD4 on host cells, the ENV proteins that make up the HIV spike undergo conformational changes. Indeed, conformational changes in the V2 loop upon CD4 binding are well-documented. So, it is possible that the predictions from our inferred landscape reflect this *in vivo* effect that is obviously not captured in the SOSIP crystal structure.

Another study highlights the ability of the fitness model to capture the effects of interactions between mutations [80]. HIV protease, which plays an important role in viral replication, has been the target of

antiretroviral drug therapy through a class of drugs known as protease inhibitors. The virus is able to evolve mutations that increase its resistance to protease inhibitors, but these may be associated with substantial fitness costs to the virus. In such cases, the drug resistance mutations that are most likely to be relevant should be the ones whose fitness costs can be compensated by other mutations. One may then ask whether the inferred fitness landscape can identify potential drug resistance mutations based on the values of the coupling constants in the Hamiltonian. Specifically, when coupling constants corresponding to compensatory interactions exceed a certain value, the strain bearing both the mutation that confers drug resistance and the additional compensatory mutation becomes sufficiently fit. This condition $J_{ij}=-h_i-h_j$, can be thought of as a level-crossing phenomenon in statistical mechanics. Fig. 12 b shows that model predictions compare increasingly well with observed drug-resistance mutations as the cut-off value for the coupling constants increases [80]. These results again suggest that the inferred landscape can capture effects of coupling between mutations on the virus' intrinsic fitness. It could be argued, however, that predictions emerging from the inferred could be good simply because it reflects the presence of drug-resistance mutations in the population of circulating viruses today. Therefore, the landscape for protease was inferred using only sequences that were obtained prior to 1996 (the year that antiretroviral therapy was introduced).

Tests against clinical data

One can test predictions emerging from the fitness landscape against clinical data on infected persons in a number of different ways. There exist cohorts of patients, called elite controllers, whose immune systems can control HIV infections without any therapy by maintaining very low viral loads. This small fraction of individuals have not progressed to AIDS, and are less likely to infect other individuals. A number of factors have been posited to explain how elite controllers achieve viral control. In genome-wide studies, the strongest correlation is observed with the MHC (HLA for humans) genes that they possess. Some HLAs, such as HLA B57 and HLA B27 are strongly overrepresented in elite controllers. As we have discussed in Chapter 3, a part of the reason is likely the statistically more cross-reactive TCR repertoire of individuals with these genes. Another reason that has been noted by comparing cohorts of controllers and progressors who have the same HLAs, is that the T cells of controllers exhibit more effective polyfunctional responses. Some studies have also implicated the antibody response of these patients. However, the factor that is implicated most strongly is the HLA genes they possess and the peptides presented by them. Predictions using the HIV fitness landscape inferred using methods described in this chapter are consistent with the idea that mutations at residues in the peptides targeted by the T cell response in elite controllers is associated with relatively large fitness penalties.

As noted earlier, analysis of the covariance matrix associated with sequences of Gag proteins showed that certain sets of residues in p24 that are involved in creating key protein-protein interfaces in the viral capsid co-evolve, and simultaneous mutations in several pairs of residues at these sites are observed less frequently than by chance. These results suggest that multiple mutations at these residues are especially deleterious for the virus. Consistent with this finding, elite controllers disproportionately target peptides that contain these residues. One can use the quantitative fitness landscape inferred using the Potts model to make more comprehensive predictions. For example, one can estimate the fitness cost incurred by mutations at any residue, averaged over all possible sequence backgrounds.

Averaging over the sequence backgrounds thus takes in to account epistatic effects between mutations. One possible way to compute the average fitness penalty is as follows:

$$\delta E\left(\overrightarrow{z_p'}, \overrightarrow{z_p}\right) = \sum_{\vec{z}_r}\left(E\left(\overrightarrow{z_p'}, \overrightarrow{z_r}\right) - E\left(\overrightarrow{z_p}, \overrightarrow{z_r}\right)\right)\frac{e^{-E\left(\overrightarrow{z_p'}, \overrightarrow{z_r}\right)}}{\sum_{\overrightarrow{z_r}} e^{-E\left(\overrightarrow{z_p'}, \overrightarrow{z_r}\right)}} \quad (51a)$$

$$< \Delta E > = \frac{1}{\sum_{\overrightarrow{z_p'}} e^{-\delta E\left(\overrightarrow{z_p'}, \overrightarrow{z_p}\right)}} \sum_{\overrightarrow{z_p'}} \delta E\left(\overrightarrow{z_p'}, \overrightarrow{z_p}\right)\ e^{-\delta E\left(\overrightarrow{z_p'}, \overrightarrow{z_p}\right)} \quad (51b)$$

Here $\delta E\left(\overrightarrow{z_p'}, \overrightarrow{z_p}\right)$ is the average energy cost of evolving a non-synonymous mutation in the targeted peptide, z_p is the epitope sequence being targeted, z_p' is the epitope sequence with a single non-synonymous mutation, and z_r is the rest of the protein sequence. The average is computed over all possible sequence backgrounds, z_r. This is carried out using standard Monte-Carlo simulations with the fitness landscape (Eq. 3). Eq. 51a therefore reflects the fitness cost of evolving a particular escape mutation in all possible sequence backgrounds, and includes the coupled effects of mutations. Eq. 51b then computes an average cost of evolving a mutation in the epitope, averaged over the possible mutations. This way of calculating the latter average emphasizes the contributions of mutations with the lowest fitness costs, as they are the most likely to evolve. Using this metric of the fitness cost, it can be shown that elite controllers target peptides associated with high fitness costs for evolving mutations in all possible sequence backgrounds. Thus, such escape mutations do not emerge rapidly and the individual's T cell response can control the virus to low levels, regardless of other mutations that the virus may evolve. With a vaccine-induced T cell response, one would like to elicit such responses that can corner the virus between being killed by T cells or evolving mutations that are likely to make the virus unviable.

In a cohort of patients from the US, Malawi, and South Africa, a comprehensive analysis of CTL responses in the early stages of infection was carried out [81]. The epitopes that were considered were those targeted by the individual patients during a time frame spanning a range from the first detection of virus to shortly after the viral load peaked (see Fig. 4). In these patients, the virus evaded the predominant T cell responses directed toward these epitopes via escape mutations. The time required for these mutations to evolve and their locations were recorded.

The peptides, or epitopes, targeted by the T cell responses are usually detected using an experimental assay called ELISPOT. Peptides comprised of overlapping residues of peptides derived from viral proteins are presented on human APCs and displayed in different wells. Blood samples from patients contain T cells. These T cells secrete cytokines, in particular IFNγ, when they interact with their cognate peptide. One counts the number of "spots" of these cytokines that are secreted in individual wells. The number of spots thus detected corresponds to the extent to which a particular epitope is being targeted in the patient. In this manner, one can determine which epitopes are being targeted by a patient's T cells, as well as the relative immunodominance of the epitopes being targeted. Epitopes targeted more dominantly are under stronger selection pressure to evolve mutations.

The time taken for a particular escape mutation to take over the population of viruses (escape time) in a patient can also be estimated if blood samples are collected at different times. Often, the samples are not collected at regularly spaced intervals, and the escape time could correspond to a time point at which a sample was not collected. One way to estimate the escape time from such data has been proposed by Perelson and co-workers []. The data consists of two time ordered vectors:

$$\vec{n} = \{n_1, n_2, \ldots\ldots\ldots n_i \ldots\ldots n_T\} \quad and \quad \vec{k} = \{k_1, k_2, \ldots\ldots\ldots k_i \ldots\ldots k_T\} \qquad (52)$$

where n_i is the number of viral sequences and k_i is the number of these sequences with an escape mutation at a targeted epitope observed at time point, t_i. One can then use a logistic form for the fraction of sequences, f (t), with an escape mutation at time, t:

$$f\,(t) = \frac{f_0}{f_0 + (1 - f_0)e^{-\varepsilon t}} \qquad (53)$$

Here f_0 and ε are parameters that reflect the fraction of sequences with an escape mutation at time zero and the rate at which the logistic equation saturates to unity, respectively. One may interpret f (t) to be a probability of observing an escape mutation at time, t. Therefore, one can write the following expression for the likelihood (L) of the observed data (Eq. 52):

$$L = \prod_{i=1}^{T} C_{k_i}^{n_i} f\,(t_i)^{k_i}\,(1 - f\,(t_i)^{n_i - k_i} \qquad (54)$$

Maximizing L with respect to the parameters, f_0 and ε, provides an estimate for them based on the data. These values of f_0 and ε can then be used in Eq. 53 to determine the time point at which half the viral population is comprised of a sequence with an escape mutation in the targeted epitope ($f\,(t_{escape}) = 0.5$). Data sets like these, which provide information on the epitopes targeted by T cells in individual humans and estimates of the locations of escape mutations and the escape times offer several opportunities to further test inferred fitness landscapes.

For example, the data provide vivid examples of how the time to escape is influenced by the sequence background of the virus (or the **J** matrix). This is best done by comparing pairs of patients that target the same peptide, but the escape mutation evolves over very different time scales. Fig. 13 illustrates one such example. Two patients, labeled CH185 and CH159, target the same epitope in Gag. In CH185, the escape mutation evolved in 122 days after the T cell response was first detected, while in CH159 the escape mutation had not evolved in over 1100 days. Fig. 13 shows that this is because of differences in the sequences of the viruses that infected these individuals. The circles in the figure depict the rest of the protein that contains the targeted epitope. The marked residues are those where mutations existed in the virus that infected the patient. Blue curves indicate that the fitness landscape predicts that the element of the **J** matrix corresponding to the preexisting mutation and the ultimate escape mutation is positive (compensatory). Red curves indicate negative values of the corresponding element of the **J** matrix (antagonistic interactions). The thicknesses of the lines reflect the relative magnitudes of the predicted J-couplings. Patient CH159 was infected with a viral strain that, compared to the infecting

strain in CH185, contained more mutations that coupled negatively (and strongly) to the putative escape mutation. Therefore, it was much more difficult for the escape mutation to evolve and grow out in patient CH159. Table 1 shows other examples of such a situation in this cohort of patients, further highlighting the importance of epistatic couplings between mutations for host-pathogen riposte in individual humans. The values of $< \Delta E >$ reported in Table 1 were computed using Eq. 51.

Eq. 51 reflects the shortest path to evolving an escape mutation and does not account for the dynamics of the evolution of the virus in response to T cell pressure. Thus, it does not properly account for the evolution of complex evolutionary trajectories. As one example of phenomena not captured by Eq. 51, if more feasible trajectories are available for an escape mutation to emerge at a particular residue or epitope, escape is facilitated because the "entropy" of escape trajectories is higher. The importance of such effects and the veracity of fitness landscapes can be tested by attempting to predict the dynamics of virus evolution in individual patients in the cohort described above. Specifically, one can ask if it possible to predict the residues at which escape mutations emerged, and the relative times required for this to happen in these patients by combining Wright-Fisher like evolutionary dynamics with the inferred fitness landscapes and knowledge of the targeted epitopes (and their relative immunodominance) [82].

The Wright-Fisher simulations can be carried out as described in the section on connecting the prevalence and fitness landscapes. A constant population size is not a very good approximation in some of the cases in the data being considered now because the virus population size is still rising if escape occurs before peak viremia. However, let us examine the results that emerge from such approximate calculations. Future studies may be able to improve on these studies by not making the approximation of constant population size. The population size that was used in the studies described in [82] is 10^4 virus particles, which corresponds to the effective population size for HIV populations in individual patients. This has been estimated as ADD THIS FROM PAPERS.

The inputs to the Wright-Fisher simulations from the data are the fraction of each sequence in the virus population when the T cell response was first detected, the targeted epitope, and the relative immunodominance of the targeted epitope. If sequence data was not available at the time point when the T cell response was first detected, the most recent recorded sequences are used. Mutations are carried out at the nucleotide level in the simulations. This is important because mutation rates are known for nucleotides, not amino acids, and because this is the only way that allows proper treatment of mutational paths (transitions between certain amino acids are simply not possible in one step).

Recombination of viral genomes can also occur if two different RNA strands (or viral genomes) are simultaneously present. The rate of recombination for HIV has been estimated by Neher and Leitner []. Recombination can be simulated during the evolutionary dynamics in a simple way. Let r be the rate of recombination events per base pair per replication cycle. During replication of each sequence in the Wright-Fisher simulations, we pick the number of recombination sites (n) with the binomial probability distribution, $p\ (n) = C_n^N\ r^n\ (1-r)^{N-n}$, where N is the length of the protein or genome under consideration, and C_n^N is the number of ways of choosing n locations out of N. If the number of recombination sites thus picked is greater than zero, the specific locations of the recombination sites are picked from a random uniform distribution. A partner sequence is then picked from those in the

evolving swarm from a uniform random distribution. As an example, consider two recombination sites, i and j. Then the recombined sequence has the original sequence between residues 1 to i, the sequence of the randomly picked other sequence from i+1 to j, and the original sequence from j+1 to L. This process mimics Reverse Transcriptase falling off of one RNA strand, hopping to another, and then returning. This is possible because RT is not a very processive enzyme.

Because the Potts models are inferred at the amino acid level, the fitness estimates used during the selection step are done at the amino acid level. The survival probability or probability of being positively selected can be written in a manner analogous to Eq. 27 as follows:

$$P_{surv} = \frac{e^{-\beta\, E(\vec{z})}}{< e^{-\beta E} > + e^{-\beta\, E\,(\vec{z})}}\,; \qquad (55)$$

where the average is computed over the prevailing virus population.

The main difference between Eq. 27 (and accompanying comments) and Eq. 55 is that we have introduced a quantity analogous to "inverse temperature" in statistical physics ($\beta = 1/k_BT$, where k_B is Boltzmann's constant and T is temperature). This parameter can be obtained by graphing the logarithm of in vitro replicative capacity measurements against values of E (**z**). The value of β thus obtained for Gag mutants, for example, equals 0.07 [79]. It is worth reflecting on why $\beta < 1$. The intrinsic fitness of a strain, without time to evolve mutations, is measured in the in vitro experiments. Let us denote the energies corresponding to the intrinsic fitness of a strain and its prevalence by E' (**z**) and E (**z**), respectively. So, the prevalence, P (**z**), and the intrinsic fitness, f (**z**), are given by:

$$f\,(\vec{z}) \;\propto\; e^{-E'\,(\vec{z})}; \quad p\,(z) \;\propto\; e^{-E\,(\vec{z})} \qquad (56)$$

Now, E (**z**) contains the effects of immune pressure, mutational fidelity, reversion of the mutations induced by immune pressure, and other effects discussed in the section on connecting prevalence and fitness of the circulating HIV population. The immune pressure promotes sampling of sequence space by promoting mutations at targeted residues in a single host. But as discussed earlier, averaged across the population of hosts, the immune fields acting at HIV protein residues are expected to be small because of the diversity of human immune responses. Reversion of mutations induced by a host's immune response when infection propagates to hosts that do not target the same epitopes, the contributions due to mutational fidelity, and recombination are all forces that drive the circulating HIV population to remain close to the fittest strains. If these forces are strong, as indicated by the data following virus evolution in single patients over long times [], then one might conjecture that the probability of observing a mutant strain in circulation is likely to be less than what one might expect from intrinsic fitness considerations alone. For example, a strain that is a mutant compared to the consensus (or reference) strain may have evolved to avoid someone's immune response and may be quite fit. But, if it is less fit than consensus and is not targeted further by another host's immune response, then this mutation will revert. So, the mutant may be less prevalent than intrinsic fitness alone might predict. Thus, one might posit that P (**z**) samples a distribution that is at a lower temperature (it samples lower fitness strains less) compared to that expected based on intrinsic fitness considerations alone. In other

words, if k_BT, the natural energy scale for intrinsic fitness equals unity, $\beta E(\mathbf{z}) = E'(\mathbf{z}) = \ln f(\mathbf{z})$, with $\beta < 1$. Note that for the Wright-Fisher like simulations, it is the value of the intrinsic fitness that we need since immune pressure, mutational fidelity, etc are explicitly treated. This is the reason underlying the use of Eq. 55 for the survival probability of a strain.

Note that because the T cell responses were measured experimentally, it is possible to include the effects of each individual's immune system in the model. To mimic the killing of infected cells by epitope-specific T cells, all virus strains that contain a targeted epitope have their fitness decreased by a fixed amount. One way to do this is to multiply $f(\vec{z}) = e^{-\beta E(\vec{z})}$ by a factor of e^{-b}, which is tantamount to adding a positive constant, b, to $\beta E(\mathbf{z})$ to account for the decreased fitness of a strain with a targeted epitope. As noted earlier, the available data also contains information on the relative immunodominance of the epitopes. This information can be incorporated in to simulations by writing the following expression for the fitness penalty of targeted epitopes, b:

Commented [MOU3]: Across epitopes targeted by all in the cohort or within patients. CHECK

$$b = (1 - \%I)b_{min} + \%I\, b_{max} \qquad (57)$$

where %I is the relative immunodominance (varies between 0 and 1), b_{min} and b_{max} are the minimum and maximum values of the penalty in fitness due to immune pressure, respectively. The choice of b_{min} made by Barton and co-workers was sufficiently high that all targeted epitopes had a selection force that would force escape. In particular, b_{min}/β was somewhat larger than the highest value of $< \Delta E >$ among targeted epitopes.

One can conservatively assume that any nonsynonymous mutation within a targeted epitope is sufficient to allow the virus to avoid detection by epitope-specific T cells. Though this assumption is not always correct, there is experimental evidence that most mutations within an epitope tend to substantially impair T cell recognition as we have discussed before [83].

In the evolutionary dynamics, the steps of replication, mutation, and selection occur in a single step. However, biologically these steps are separate and involve different time scales. Mutation occurs rapidly during the reverse transcription of the viral RNA into DNA and during replication, while selection effectively operates at the level of infected cells (which may or may not successfully produce new viruses, and which may be killed during the replication process by cytotoxic T cells); a typical lifetime of infected cells is around 2 days when productively infected [20]. Also, as noted above, the simulations under consideration operate in the regime of very strong selection for escape, so that escape is favored even at epitopes where the fitness cost of mutation is high. For these reasons, it is difficult to precisely connect generations of evolution in the simulations of virus evolution to real time. The predicted generations of evolution required for escape mutations to emerge reflect relative rates at which escape occurs at different epitopes, which is the best that one can hope for from such coarse-grained simulations.

The evolutionary dynamics described above are carried out with many approximations, and efforts to carry out more realistic simulations are necessary. Overall, however, when compared to the specific clinical data being considered, the effect of these assumptions appears to be relatively mild. For example, model predictions for the most likely and second most likely locations for escape mutations

matched the clinical data in roughly 86% of the cases. Figure 14 shows that the clinically measured escape times compare reasonably well with the predicted evolutionary generations (Spearman correlation of 0.73). The latter were predicted by averaging over many simulations for the same epitope, and the error bars show the variance in the results. There is considerable scatter, especially for cases where escape is neither very fast or slow; but, the error bars show that the rough order is correctly predicted by the simulations. Two other points are noteworthy. The first is that if we only use Shannon entropy of the epitopes to predict relative escape times, then the Spearman correlation with the clinical data is only -0.22 (epitopes associated with higher entropy escape faster). This shows that the effects of coupling between mutations is important. The second point to note is made clear by considering simulations where we do not account for relative immunodominance (values of %I in Eq. 57). Then the predicted times for escape correlate with the clinical data with a lower Spearman correlation of 0.53. The nature of the epitope targeted, as characterized by the fitness costs of evolving mutations therein in diverse sequence backgrounds and the number of possible escape paths with low fitness cost that are available for an epitope, are important determinants of escape time. The other important determinant is the selection pressure imposed by the T cell response to evolve mutations. The higher the immunodominance (value of %I), the higher the selection pressure. In fact, if only the values of %I are used to predict escape times, a Spearman correlation of -0.53 is observed [x]. The results of evolutionary dynamics taking both the fitness characteristics of the targeted epitopes and immunodominance into account lead to the highest correlation between computational and clinical results (Spearman correlation of 0.73).

An important question for future exploration is to determine the extent to which the discrepancies between clinical data and predictions can be ascribed to different approximations made in the simulations of evolutionary dynamics or errors in statistical inference of the fitness landscape. For example, is the comparison with data better if we do not make the approximation of constant population size, or do not encapsulate mutation/replication and selection in one effective time step (by treating infection of new cells as a separate step)? Also, in the studies described, the evolutionary dynamics of escape was considered only one epitope at a time. In reality, multiple epitopes are simultaneously targeted in a single individual. Addressing these deficiencies are topics for future research.

Based on the progress made so far and on positive correlations between predictions and data, immunogens can be designed for the T cell component of a vaccine which contain only parts of the HIV proteome as per considerations noted early in this chapter. A major engineering challenge here is devising carriers and adjuvants that can efficiently deliver long peptide immunogens, an issue that also confronts the development of cancer vaccines. This topic is beyond the scope of this book.

Appendix

Connection between Statistical mechanics and information theory

A primer on elementary statistical mechanics

In this appendix, we briefly describe the connection between statistical mechanics and information theory that led us to define the entropy of circulating sequences of HIV proteins as shown in Eq. 2.

We begin with some brief reminders of elementary statistical mechanics. Consider a macroscopic system with a fixed number of particles, volume, and energy (denoted by N, V, and E, respectively). The use of N for the number of particles is standard in most books on statistical mechanics, and is not to be confused with the protein sequence length for which the symbol N was used in the main text of this chapter. There is an ensemble of states consistent with the macroscopic state with fixed N, V, and E, and this ensemble is referred to as the microcanonical ensemble. All microscopic states in this ensemble are equally likely, and so, the probability, P, of being in any one of the microstates is:

$$P = \frac{1}{\Omega} \qquad \text{(A1)}$$

where Ω is the total number of microstates in the ensemble. A basic tenet of statistical mechanics proposed by Boltzamann is that the entropy, which is a measure of disorder or uncertainty in the system, equals $k_B \ln \Omega$ in the microcanonical ensemble.

As it is difficult to control and measure energy, for experimental systems of interest, it is more convenient to control temperature (T) instead of energy. Note that temperature is an intensive variable, which is not a function of system size, while energy grows with system size. The ensemble of microscopic states consistent with the macroscopic constraints of N, V, and T is referred to as the canonical ensemble. There is no reason to believe that the microstates in the canonical ensemble are equally likely as in the microcanonical ensemble, and indeed their probabilities are different as shown by the simple calculation described below.

A convenient way to keep a system at constant temperature is to surround it with a conducting wall and immerse it in a large bath. The system exchanges energy with the bath, and so its energy fluctuates, but this keeps the temperature of the system constant and equal to that of the bath. This situation is depicted in Fig. A1. As the large bath is insulated from its surroundings, together the bath and the system are characterized by constant values of N, V, and E. Denoting the bath by B, the system with Ξ, and the system and bath together as Ξ', the following relationship holds:

$$E_{\Xi'} = E_{\Xi} + E_B \qquad \text{(A2)}$$

where E_i is the energy of the i^{th} component, and $E_{\Xi'}$ is a constant. Now, suppose the system is in a particular microstate, E_ν, then the energy of the bath is given by $E_{\Xi'} - E_\nu$. When the system is in a particular microstate, ν, the total number of microstates for the bath and system equals 1 * $\Omega_B\,(E_{\Xi'} - E_\nu)$, where $\Omega_B\,(E_{\Xi'} - E_\nu)$ is the number of microstates available to the bath when the system is in the microstate, ν. Since the system and bath together is characterized by constant N, V, and E (equal to $E_{\Xi'}$), the microstates of the system and bath taken together are equally likely. Therefore, the probability of the system being in a particular microstate, ν, is given by dividing the number of microstates available to the system in this situation divided by the total number of possible microstates of Ξ'; i.e.,

$$P_\upsilon^{NVT} = \frac{\Omega_B\,(E_{\Xi'} - E_\nu)}{\Omega_{\Xi'}\,(E_{\Xi'})} \qquad \text{(A3)}$$

The denominator on the right-hand side of Eq. A3 is a constant, and so $P_\upsilon^{NVT} \sim \Omega_B\,(E_{\Xi'} - E_\nu)$. To keep the temperature of the system constant, it must be much smaller than the bath (or the bath and system taken together). So, we can carry out a Taylor expansion in powers of E_ν, and rewrite P_υ^{NVT} as follows:

$$P_\upsilon^{NVT} \sim \exp[\ln \Omega_B\,(E_{\Xi'} - E_\nu)] = \exp\left[\ln \Omega_B\,(E_{\Xi'}) - E_\nu \left(\frac{\partial \ln \Omega_B}{\partial E_B}\right)_{E_{\Xi'}}\right] \sim \exp\left[-\frac{E_\nu}{k_b}\left(\frac{\partial S_B}{\partial E_B}\right)\right] \qquad \text{(A4)}$$

Where in the last equality in A4, we have used Boltzmann's definition of entropy in the microcanonical ensemble. Now standard classical thermodynamics tells is that $\left(\frac{\partial S_B}{\partial E_B}\right) = \frac{1}{T_B}$, where T_B is the temperature of the bath, which is equal to the temperature of the system, T. Therefore, we conclude that

$$P_\upsilon^{NVT} = \frac{e^{-\frac{E_\nu}{k_B T}}}{\sum_\upsilon e^{-\frac{E_\nu}{k_B T}}} = \frac{e^{-\frac{E_\nu}{k_B T}}}{Z} \qquad \text{(A5)}$$

where Z is the partition function that normalizes the probability distribution.

What is the entropy in the canonical ensemble? Classical thermodynamics tells us that a system at equilibrium with N,V and E fixed, corresponds to the minimum energy state. For a system with fixed, N,V and T (canonical ensemble), it is the Helmholtz energy (A) that is minimal at equilibrium. The Helmholtz free energy is given by $A = E - TS$. Note that the energy of a system in the canonical ensemble is the value of the energy averaged over microstates, $\sum_\nu P_\upsilon^{NVT}\, E_\nu$. Using the definitions of E and A, and Eq. A5, it is easy to show that

$$E = \frac{\partial \ln Z}{\partial \beta} \qquad \text{(A6)}$$

where β is $\frac{1}{k_B T}$, an inverse temperature. Simple algebra using Eq. A6 and the definition of A yields that $A = -k_B T \ln Z$. Now substituting formulas obtained above for A and E into the relationship, $S = \frac{E-A}{T}$, obtains

$$S = -k_B \sum_\nu P^\nu \ln P^\nu = -\sum_\nu P^\nu \ln P^\nu \text{ in "entropy units"} \qquad \text{(A7)}$$

Notice that we obtained expressions for the pertinent quantities in the canonical ensemble by considering a system that overall (bath plus system) looked like a system in the microcanonical ensemble.

Information and entropy

The connection between the concepts of information and entropy was first described by Shannon [], and many fine expositions, such as that by Jaynes [], are available. A clear description in the context of biophysics can be found in the book by Bialek, and the much shorter description below is inspired by it.

Consider a situation where the answer to a question could be one of several possibilities. When we learn the answer, we gain information as there is no longer any uncertainty. The larger the number of possible answers the greater our uncertainty, and thus more information is gained by knowing the answer. Notice that entropy in statistical mechanics reflects the uncertainty in our knowledge of the microstate that a system occupies, given the macroscopic constraints. For example, if there is only one possible microstate available, there is no uncertainty, and the entropy is zero (Eq. A7). So, intuition suggests that there may be a connection between information gained upon knowing the answer to a question with several possible answers and the concept of entropy in statistical mechanics. Shannon made this connection precise.

As noted above, the information gained should increase monotonically with the number of possible answers, N, to a question. Consider the following question: what is your name, and where do you live? This question has two independent parts. The information gained by answering a question which has independent parts must be the sum of the information gained by knowing the answer to each independent part. If we associate a probability, p_i, with each possible answer, i, Shannon showed that the only function of the set, $\{p_i\}$, that is consistent with the two constraints on information gained noted above is the entropy. The proof is sketched below.

First, let us consider a situation analogous to the microcanonical ensemble in statistical mechanics – viz., one where there are N equally likely possible answers to a question. Furthermore, consider the general situation where m independent questions need to be answered, each with k possible answers. Therefore, $N = k^m$. Since the information gained by answering each independent part must add, and it must be a function of N, the information gained, I (N), must obey the following relationship:

$$I\ (N) = m\ f\ (k) \qquad \text{(A8)}$$

Now consider a situation where n independent questions need to be answered, each with l possible answers, with the values of l and n such that the following condition holds:

$$k^m\ \leq\ l^n\ \leq\ k^{m+1} \qquad \text{(A9)}$$

Since the information gained must be a monotonically increasing function of the total number of possible answers,

$$I\ (k^m)\ \leq I\ (l^n)\ \leq I\ (k^{m+1}) \qquad \text{(A10)}$$

The fact that the answer is obtained in terms of independent questions further implies that

$$m\ f\ (k) \leq n\ f\ (l)\ \leq (m+1)\ f\ (k) \qquad \text{(A11a)}$$

$$\frac{m}{n} \leq \frac{f\ (l)}{f\ (k)} \leq \frac{m}{n} + \frac{1}{n} \qquad \text{(A11b)}$$

Note that if the function, f, was the logarithm, Eq. A11 would be obeyed because Eqs A8 and A9 would imply that $m \log k\ \leq n \log l\ \leq (m+1) \log k$. This suggests that, if there are N equally likely answers to a question, the information gained by knowing the answer is ~ log N. A mathematically more rigorous

proof of this result can be obtained. Notice that the entropy of a system in the microcanonical ensemble is also the logarithm of the number of possible microstates.

Now consider the more general situation, where there are still N possible answers, but they are not equally likely. The probability of obtaining the i^{th} answer is p_i. To determine the properties of a system, including its entropy, characterized by the canonical ensemble of microstates, we considered the system to be immersed in a large bath and together the system and bath comprised a microcanonical ensemble of states. We could then use the known properties of the microcanonical ensemble to help us determine the properties of the canonical ensemble. In a similar spirit, we can express p_i such that each possible answer has a few equally likely possibilities; i.e.,

$$p_i = \frac{k_i}{\sum_m k_m} \qquad \text{(A12)}$$

where there are a group of k_i equally likely answers from which the i^{th} answer is drawn, and the total number of possible answers is $N = \sum_m k_m$.

The quantity we seek is the information gained, $I(p_i)$, when the answer is the one with probability p_i. In the way that we have formulated the problem, the total information gained by knowing which of the N answers is realized $I(N)$ must be equal to the sum of $I(p_i)$ and the information gained by knowing which of the k_i equally likely answers in group i was obtained. Furthermore, since each of the latter groupings has a probability, p_i, associated with it:

$$I(p_i) + \sum_i p_i \log k_i = I(N) = \log \sum_m k_m \qquad \text{(A13)}$$

Therefore,

$$I(p_i) = \sum_n p_i \log \sum_m k_m - \sum_i p_i \log k_i = -\sum_i p_i \log \frac{k_i}{\sum_m k_m} = -\sum_i p_i \log p_i = S(\{p_i\}) \qquad \text{(A14)}$$

where $\{p_i\}$ is the set of probabilities for the possible answers. The expression for the information gained by knowing that the answer with probability, p_i, is realized is exactly the same as the entropy in the canonical ensemble where different microstates occur with different probabilities.

In the main text, we seek the least biased model for the probability distribution characterizing HIV protein strains in circulation. The least biased model is one where the answer is most uncertain, and so the information gained by determining the probability distribution must be maximal. Eq. A14 says that this is tantamount to seeking the probability distribution characterized by the maximal entropy.

References

[1] Janeway C A, Travers P, Walport M and Capra J D 2005 Immunobiology: the immune system in health and disease

[2] Hozumi N and Tonegawa S 1976 Evidence for somatic rearrangement of immunoglobulin

genes coding for variable and constant regions *P Natl Acad Sci USA* **73** 3628–32

[3] Eisen H N and Siskind G W 1964 Variations in Affinities of Antibodies during the Immune Response *Biochemistry* **3** 996–1008

[4] Robinson J, Halliwell J A, Hayhurst J D, Flicek P, Parham P and Marsh S G 2014 The IPD and IMGT/HLA database: allele variant databases *Nucleic Acids Res* gku1161

[5] Chakraborty A K and Weiss A 2014 Insights into the initiation of TCR signaling *Nat Immunol* **15** 798–807

[6] Victora G D and Nussenzweig M C 2012 Germinal centers *Annu Rev Immunol* **30** 429–57

[7] Victora G D, Schwickert T A, Fooksman D R, Kamphorst A O, Meyer-Hermann M, Dustin M L and Nussenzweig M C 2010 Germinal center dynamics revealed by multiphoton microscopy with a photoactivatable fluorescent reporter *Cell* **143** 592–605

[8] Kepler T B and Perelson A S 1993 Cyclic re-entry of germinal center B cells and the efficiency of affinity maturation *Immunol Today* **14** 412–5

[9] Oprea M and Perelson A S 1997 Somatic mutation leads to efficient affinity maturation when centrocytes recycle back to centroblasts. *J Immunol* **158** 5155–62

[10] De Gregorio E and Rappuoli R 2014 From empiricism to rational design: a personal perspective of the evolution of vaccine development *Nat Rev Immunol* **14** 505–14

[11] Korber B, Gaschen B, Yusim K, Thakallapally R, Keşmir C and Detours V 2001 Evolutionary and immunological implications of contemporary HIV-1 variation *Brit Med Bull* **58** 19–42

[12] Phillips R E, Rowland-Jones S, Nixon D F, Gotch F M, Edwards J P, Ogunlesi A O, Elvin J G, Rothbard J A, Bangham C R M, Rizza C R and McMichael A J 1991 Human immunodeficiency virus genetic variation that can escape cytotoxic T cell recognition *Nature* **354** 453–9

[13] Faria N R, Rambaut A, Suchard M A, Baele G, Bedford T, Ward M J, Tatem A J, Sousa J D, Arinaminpathy N, Pepin J, Posada D, Peeters M, Pybus O G and Lemey P 2014 The early spread and epidemic ignition of HIV-1 in human populations *Science* **346** 56–61

[14] Allan J, Lee T H, McLane M F, Sodroski J, Haseltine W and Essex M 1983 Identification of the major envelope glycoprotein product of HTLV-III *Science* **228** 1091–4

[15] Dalgleish A G, Beverley P C, Clapham P R, Crawford D H, Greaves M F and Weiss R A 1983 The CD4 (T4) antigen is an essential component of the receptor for the AIDS retrovirus. *Nature* **312** 763–7

[16] Klatzmann D, Champagne E, CHAMARET S, Gruest J, Guetard D, Hercend T, Gluckman J-C and

Montagnier L 1983 T-lymphocyte T4 molecule behaves as the receptor for human retrovirus LAV. *Nature* **312** 767–8

[17] Freed E O 2015 HIV-1 assembly, release and maturation *Nat Rev Microbiol* **13** 484–96

[18] Sanjuán R, Nebot M R, Chirico N, Mansky L M and Belshaw R 2010 Viral mutation rates *J Virol* **84** 9733–48

[19] Wei X, Ghosh S K, Taylor M E, Johnson V A, Emini E A, Deutsch P, Lifson J D, Bonhoeffer S, Nowak M A, Hahn B H, Saag M S and Shaw G M 1995 Viral dynamics in human immunodeficiency virus type 1 infection *Nature* **373** 117–22

[20] Perelson A S, Neumann A U, Markowitz M, Leonard J M and Ho D D 1996 HIV-1 dynamics in vivo: virion clearance rate, infected cell life-span, and viral generation time *Science* **271** 1582–6

[21] McMichael A J, Borrow P, Tomaras G D, Goonetilleke N and Haynes B F 2009 The immune response during acute HIV-1 infection: clues for vaccine development *Nat Rev Immunol* **10** 11–23

[22] Fraser C, Lythgoe K, Leventhal G E, Shirreff G, Hollingsworth T D, Alizon S and Bonhoeffer S 2014 Virulence and Pathogenesis of HIV-1 Infection: An Evolutionary Perspective *Science* **343** 1243727

[23] Barouch D H 2008 Challenges in the development of an HIV-1 vaccine *Nature* **455** 613–9

[24] Goulder P J R and Walker B D 2012 HIV and HLA Class I: An Evolving Relationship *Immunity* **37** 426–40

[25] Fellay J, Shianna K V, Ge D, Colombo S, Ledergerber B, Weale M, Zhang K, Gumbs C, Castagna A, Cossarizza A, Cozzi-Lepri A, De Luca A, Easterbrook P, Francioli P, Mallal S, Martinez-Picado J, Miro J M, Obel N, Smith J P, Wyniger J, Descombes P, Antonarakis S E, Letvin N L, McMichael A J, Haynes B F, Telenti A and Goldstein D B 2007 A Whole-Genome Association Study of Major Determinants for Host Control of HIV-1 *Science* **317** 944–7

[26] Pereyra F, Addo M M, Kaufmann D E, Liu Y, Miura T, Rathod A, Baker B, Trocha A, Rosenberg R, Mackey E, Ueda P, Lu Z, Cohen D, Wrin T, Petropoulos C J, Rosenberg E S and Walker B D 2008 Genetic and Immunologic Heterogeneity among Persons Who Control HIV Infection in the Absence of Therapy *J Infect Dis* **197** 563–71

[27] The International HIV Controllers Study 2010 The Major Genetic Determinants of HIV-1 Control Affect HLA Class I Peptide Presentation *Science* **330** 1551–7

[28] Łuksza M and Lässig M 2014 A predictive fitness model for influenza *Nature* **507** 57–61

[29] Martinez-Picado J, Prado J G, Fry E E, Pfafferott K, Leslie A, Chetty S, Thobakgale C, Honeyborne I, Crawford H, Matthews P, Pillay T, Rousseau C, Mullins J I, Brander C, Walker B D, Stuart D I, Kiepiela P and Goulder P 2006 Fitness Cost of Escape Mutations in p24 Gag in Association with Control of Human Immunodeficiency Virus Type 1 *J Virol* **80** 3617–23

[30] Brockman M A, Schneidewind A, Lahaie M, Schmidt A, Miura T, DeSouza I, Ryvkin F, Derdeyn C A, Allen S, Hunter E, Mulenga J, Goepfert P A, Walker B D and Allen T M 2007 Escape and Compensation from Early HLA-B57-Mediated Cytotoxic T-Lymphocyte Pressure on Human Immunodeficiency Virus Type 1 Gag Alter Capsid Interactions with Cyclophilin A *J Virol* **81** 12608–18

[31] Wright S 1932 The roles of mutation, inbreeding, crossbreeding, and selection in evolution Proceedings of the Sixth International Congress of Genetics vol 1, pp 356–66

[32] de Visser J A G M and Krug J 2014 Empirical fitness landscapes and the predictability of evolution *Nat Rev Genet* **15** 480–90

[33] Hinkley T, Martins J, Chappey C, Haddad M, Stawiski E, Whitcomb J M, Petropoulos C J and Bonhoeffer S 2011 A systems analysis of mutational effects in HIV-1 protease and reverse transcriptase. *Nat Genet* **43** 487–9

[34] Kouyos R D, Wyl von V, Hinkley T, Petropoulos C J, Haddad M, Whitcomb J M, Böni J, Yerly S, Cellerai C, Klimkait T, Günthard H F, Bonhoeffer Sthe Swiss HIV Cohort Study 2011 Assessing Predicted HIV-1 Replicative Capacity in a Clinical Setting *PLoS Pathog* **7** e1002321

[35] Otwinowski J and Plotkin J B 2014 Inferring fitness landscapes by regression produces biased estimates of epistasis *P Natl Acad Sci USA* **111** E2301–9

[36] Plerou V, Gopikrishnan P, Rosenow B, Amaral L A N and Stanley H E 1999 Universal and nonuniversal properties of cross correlations in financial time series *Phys Rev Lett* **83** 1471

[37] Laloux L, Cizeau P, Bouchaud J-P and Potters M 1999 Noise Dressing of Financial Correlation Matrices *Phys Rev Lett* **83** 1467–70

[38] Plerou V, Gopikrishnan P, Rosenow B, Amaral L, Guhr T and Stanley H 2002 Random matrix approach to cross correlations in financial data *Phys Rev E* **65** 066126

[39] Halabi N, Rivoire O, Leibler S and Ranganathan R 2009 Protein sectors: evolutionary units of three-dimensional structure *Cell* **138** 774–86

[40] Dahirel V, Shekhar K, Pereyra F, Miura T, Artyomov M, Talsania S, Allen T M, Altfeld M, Carrington M, Irvine D J, Walker B D and Chakraborty A K 2011 Coordinate linkage of HIV evolution reveals regions of immunological vulnerability *P Natl Acad Sci USA* **108** 11530–5

[41] Quadeer A A, Louie R H Y, Shekhar K, Chakraborty A K, Hsing I M and McKay M R 2014 Statistical Linkage Analysis of Substitutions in Patient-Derived Sequences of Genotype 1a Hepatitis C Virus Nonstructural Protein 3 Exposes Targets for Immunogen Design *J Virol* **88** 7628–44

[42] Sengupta A M and Mitra P P 1999 Distributions of singular values for some random matrices *Phys Rev E* **60** 3389–92

[43] Cocco S, Monasson R and Weigt M 2013 From Principal Component to Direct Coupling Analysis of Coevolution in Proteins: Low-Eigenvalue Modes are Needed for Structure Prediction *PLoS Comput Biol* **9** e1003176

[44] Ganser-Pornillos B K, Yeager M and Sundquist W I 2008 The structural biology of HIV assembly *Curr Opin Struc Biol* **18** 203–17

[45] Deeks S G and Walker B D 2007 Human Immunodeficiency Virus Controllers: Mechanisms of Durable Virus Control in the Absence of Antiretroviral Therapy *Immunity* **27** 406–16

[46] Ferguson A L, Mann J K, Omarjee S, Ndung'u T, Walker B D and Chakraborty A K 2013 Translating HIV Sequences into Quantitative Fitness Landscapes Predicts Viral Vulnerabilities for Rational Immunogen Design *Immunity* **38** 606–17

[47] Jaynes E T 1982 On the rationale of maximum-entropy methods *P IEEE* **70** 939–52

[48] Lapedes A, Giraud B and Jarzynski C 2012 Using Sequence Alignments to Predict Protein Structure and Stability With High Accuracy *arXiv* **q-bio.QM**

[49] Weigt M, White R A, Szurmant H, Hoch J A and Hwa T 2009 Identification of direct residue contacts in protein–protein interaction by message passing *P Natl Acad Sci USA* **106** 67–72

[50] Schneidman E, Berry M J II, Segev R and Bialek W 2006 Weak pairwise correlations imply strongly correlated network states in a neural population *Nature* **440** 1007–12

[51] Haq O, Andrec M, Morozov A V and Levy R M 2012 Correlated Electrostatic Mutations Provide a Reservoir of Stability in HIV Protease *PLoS Comput Biol* **8** e1002675

[52] Flynn W F, Chang M W, Tan Z, Oliveira G, Yuan J, Okulicz J F, Torbett B E and Levy R M 2015 Deep Sequencing of Protease Inhibitor Resistant HIV Patient Isolates Reveals Patterns of Correlated Mutations in Gag and Protease *PLoS Comput Biol* **11** e1004249

[53] Hopfield J J 1982 Neural networks and physical systems with emergent collective computational abilities *P Natl Acad Sci USA* **79** 2554–8

[54] Barton J P, Kardar M and Chakraborty A K 2015 Scaling laws describe memories of host–

pathogen riposte in the HIV population *P Natl Acad Sci USA* **112** 1965–70

[55] Ackley D H, Hinton G E and Sejnowski T J 1985 A learning algorithm for Boltzmann machines *Cognitive Sci* **9** 147–69

[56] Plefka T 1982 Convergence condition of the TAP equation for the infinite-ranged Ising spin glass model *J Phys A: Math Gen* **15** 1971

[57] Georges A and Yedidia J S 1991 How to expand around mean-field theory using high-temperature expansions *J Phys A: Math Gen* **24** 2173

[58] Morcos F, Pagnani A, Lunt B, Bertolino A, Marks D S, Sander C, Zecchina R, Onuchic J N, Hwa T and Weigt M 2011 Direct-coupling analysis of residue coevolution captures native contacts across many protein families *P Natl Acad Sci USA* **108** E1293–301

[59] Ravikumar P, Wainwright M J and Lafferty J D 2010 High-dimensional Ising model selection using ℓ1-regularized logistic regression *Ann Stat* **38** 1287–319

[60] Aurell E and Ekeberg M 2012 Inverse Ising inference using all the data *Phys Rev Lett* **108** 090201

[61] Barton J P, Cocco S, De Leonardis E and Monasson R 2014 Large pseudocounts and L2-norm penalties are necessary for the mean-field inference of Ising and Potts models *Phys Rev E* **90** 012132

[62] Barton J P, De Leonardis E, Coucke A and Cocco S 2016 *ACE: adaptive cluster expansion for maximum entropy graphical model inference* (Cold Spring Harbor Labs Journals)

[63] Cocco S and Monasson R 2011 Adaptive Cluster Expansion for Inferring Boltzmann Machines with Noisy Data *Phys Rev Lett* **106** 090601

[64] Cocco S and Monasson R 2012 Adaptive Cluster Expansion for the Inverse Ising Problem: Convergence, Algorithm and Tests *J Stat Phys* **147** 252–314

[65] Berg J, Willmann S and Lässig M 2004 Adaptive evolution of transcription factor binding sites *BMC Evol Biol* **4** 42

[66] Sella G and Hirsh A E 2005 The application of statistical physics to evolutionary biology *P Natl Acad Sci USA* **102** 9541–6

[67] Brotto T, Bunin G and Kurchan J 2014 Population aging through survival of the fit and stable *arXiv* **physics.bio-ph**

[68] Shekhar K, Ruberman C F, Ferguson A L, Barton J P, Kardar M and Chakraborty A K 2013 Spin models inferred from patient-derived viral sequence data faithfully describe HIV fitness

landscapes *Phys Rev E* **88** 062705

[69] Goldrath A W and Bevan M J 1999 Selecting and maintaining a diverse T-cell repertoire *Nature* **402** 255–62

[70] Griffiths R B 1967 Correlations in Ising ferromagnets. II. External magnetic fields *J Math Phys* **8** 484–9

[71] Eigen M 1971 Selforganization of matter and the evolution of biological macromolecules *Naturwissenschaften* **58** 465–523

[72] Leuthäusser I 1986 An exact correspondence between Eigen's evolution model and a two-dimensional Ising system *J Chem Phys* **84** 1884

[73] Friedrich T C, Dodds E J, Yant L J, Vojnov L, Rudersdorf R, Cullen C, Evans D T, Desrosiers R C, Mothé B R, Sidney J, Sette A, Kunstman K, Wolinsky S, Piatak M, Lifson J, Hughes A L, Wilson N, O'Connor D H and Watkins D I 2004 Reversion of CTL escape–variant immunodeficiency viruses in vivo *Nat Med* **10** 275–81

[74] Zanini F, Brodin J, Thebo L, Lanz C, Bratt G, Albert J and Neher R A 2015 Population genomics of intrapatient HIV-1 evolution *eLife* **4** 13239

[75] Carlson J M, Du V Y, Pfeifer N, Bansal A, Tan V Y F, Power K, Brumme C J, Kreimer A, DeZiel C E, Fusi N, Schaefer M, Brockman M A, Gilmour J, Price M A, Kilembe W, Haubrich R, John M, Mallal S, Shapiro R, Frater J, Harrigan P R, Ndung'u T, Allen S, Heckerman D, Sidney J, Allen T M, Goulder P J R, Brumme Z L, Hunter E and Goepfert P A 2016 Impact of pre-adapted HIV transmission *Nat Med* **22** 606–13

[76] Brumme Z L, John M, Carlson J M, Brumme C J, Chan D, Brockman M A, Swenson L C, Tao I, Szeto S, Rosato P, Sela J, Kadie C M, Frahm N, Brander C, Haas D W, Riddler S A, Haubrich R, Walker B D, Harrigan P R, Heckerman D and Mallal S 2009 HLA-Associated Immune Escape Pathways in HIV-1 Subtype B Gag, Pol and Nef Proteins *PLoS One* **4** e6687

[77] Neher R A and Leitner T 2010 Recombination rate and selection strength in HIV intra-patient evolution *PLoS Comput Biol* **6** e1000660

[78] Batorsky R, Kearney M F, Palmer S E, Maldarelli F, Rouzine I M and Coffin J M 2011 Estimate of effective recombination rate and average selection coefficient for HIV in chronic infection *P Natl Acad Sci USA* **108** 5661–6

[79] Mann J K, Barton J P, Ferguson A L, Omarjee S, Walker B D, Chakraborty A K and Ndung'u T 2014 The fitness landscape of HIV-1 Gag: Advanced modeling approaches and validation of model predictions by in vitro testing *PLoS Comput Biol* **10** e1003776

[80] Butler T C, Barton J P, Kardar M and Chakraborty A K 2016 Identification of drug resistance mutations in HIV from constraints on natural evolution *Phys Rev E* **93** 022412EP–

[81] Liu M K P, Hawkins N, Ritchie A J, Ganusov V V, Whale V, Brackenridge S, Li H, Pavlicek J W, Cai F, Rose-Abrahams M, Treurnicht F, Hraber P, Riou C, Gray C, Ferrari G, Tanner R, Ping L-H, Anderson J A, Swanstrom R, Cohen M, Karim S S A, Haynes B, Borrow P, Perelson A S, Shaw G M, Hahn B H, Williamson C, Korber B T, Gao F, Self S, McMichael A and Goonetilleke N 2013 Vertical T cell immunodominance and epitope entropy determine HIV-1 escape *J Clin Invest* **123** 380–93

[82] Barton J P, Goonetilleke N, Butler T C, Walker B D, McMichael A J and Chakraborty A K 2016 Relative rate and location of intra-host HIV evolution to evade cellular immunity are predictable *Nat Commun* **7** 11660

[83] Lee J K, Stewart-Jones G, Dong T, Harlos K, Di Gleria K, Dorrell L, Douek D C, van der Merwe P A, Jones E Y and McMichael A J 2004 T Cell Cross-Reactivity and Conformational Changes during TCR Engagement *J Exp Med* **200** 1455–66

[84] Binley J M, Wrin T, Korber B, Zwick M B, Wang M, Chappey C, Stiegler G, Kunert R, Zolla-Pazner S, Katinger H, Petropoulos C J and Burton D R 2004 Comprehensive Cross-Clade Neutralization Analysis of a Panel of Anti-Human Immunodeficiency Virus Type 1 Monoclonal Antibodies *J Virol* **78** 13232–52

[85] Li Y, Migueles S A, Welcher B, Svehla K, Phogat A, Louder M K, Wu X, Shaw G M, Connors M, Wyatt R T and Mascola J R 2007 Broad HIV-1 neutralization mediated by CD4-binding site antibodies *Nat Med* **13** 1032–4

[86] Scheid J F, Mouquet H, Feldhahn N, Seaman M S, Velinzon K, Pietzsch J, Ott R G, Anthony R M, Zebroski H, Hurley A, Phogat A, Chakrabarti B, Li Y, Connors M, Pereyra F, Walker B D, Wardemann H, Ho D, Wyatt R T, Mascola J R, Ravetch J V and Nussenzweig M C 2009 Broad diversity of neutralizing antibodies isolated from memory B cells in HIV-infected individuals *Nature* **458** 636–40

[87] Kwong P D, Mascola J R and Nabel G J 2013 Broadly neutralizing antibodies and the search for an HIV-1 vaccine: the end of the beginning *Nat Rev Immunol* **13** 693–701

[88] Wyatt R, Kwong P D, Desjardins E, Sweet R W, Robinson J, Hendrickson W A and Sodroski J G 1998 The antigenic structure of the HIV gp120 envelope glycoprotein *Nature* **393** 705–11

[89] Kwong P D, Doyle M L, Casper D J, Cicala C, Leavitt S A, Majeed S, Steenbeke T D, Venturi M, Chaiken I, Fung M, Katinger H, Parren P W I H, Robinson J, Van Ryk D, Wang L, Burton D R, Freire E, Wyatt R, Sodroski J, Hendrickson W A and Arthos J 2002 HIV-1 evades antibody-mediated neutralization through conformational masking of receptor-binding sites *Nature* **420** 678–82

[90] Sanders R W, Derking R, Cupo A, Julien J-P, Yasmeen A, de Val N, Kim H J, Blattner C, la Peña de A T, Korzun J, Golabek M, de los Reyes K, Ketas T J, van Gils M J, King C R, Wilson I A, Ward A B, Klasse P J and Moore J P 2013 A Next-Generation Cleaved, Soluble HIV-1 Env Trimer, BG505 SOSIP.664 gp140, Expresses Multiple Epitopes for Broadly Neutralizing but Not Non-Neutralizing Antibodies *PLoS Pathog* **9** e1003618

[91] West A P Jr, Scharf L, Scheid J F, Klein F, Bjorkman P J and Nussenzweig M C 2014 Structural Insights on the Role of Antibodies in HIV-1 Vaccine and Therapy *Cell* **156** 633–48

[92] Jardine J G, Kulp D W, Havenar-Daughton C, Sarkar A, Briney B, Sok D, Sesterhenn F, Ereno-Orbea J, Kalyuzhniy O, Deresa I, Hu X, Spencer S, Jones M, Georgeson E, Adachi Y, Kubitz M, deCamp A C, Julien J P, Wilson I A, Burton D R, Crotty S and Schief W R 2016 HIV-1 broadly neutralizing antibody precursor B cells revealed by germline-targeting immunogen *Science* **351** 1458–63

[93] Kepler T B, Munshaw S, Wiehe K, Zhang R, Yu J-S, Woods C W, Denny T N, Tomaras G D, Alam S M, Moody M A, Kelsoe G, Liao H-X and Haynes B F 2014 Reconstructing a B-Cell Clonal Lineage. II. Mutation, Selection, and Affinity Maturation *Front Immunol* **5** 117

[94] Jardine J, Julien J P, Menis S, Ota T, Kalyuzhniy O, McGuire A, Sok D, Huang P S, MacPherson S, Jones M, Nieusma T, Mathison J, Baker D, Ward A B, Burton D R, Stamatatos L, Nemazee D, Wilson I A and Schief W R 2013 Rational HIV Immunogen Design to Target Specific Germline B Cell Receptors *Science* **340** 711–6

[95] Kepler T B and Perelson A S 1993 Somatic Hypermutation in B Cells: An Optimal Control Treatment *J Theor Biol* **164** 37–64

[96] Zhang J and Shakhnovich E I 2010 Optimality of mutation and selection in germinal centers *PLoS Comput Biol* **6** e1000800

[97] Keşmir C and De Boer R J 2003 A spatial model of germinal center reactions: cellular adhesion based sorting of B cells results in efficient affinity maturation *J Theor Biol* **222** 9–22

[98] Swerdlin N, Cohen I R and Harel D 2008 The lymph node B cell immune response: Dynamic analysis in-silico *P IEEE* **96** 1421–43

[99] Meyer-Hermann M E, Maini P K and Iber D 2006 An analysis of B cell selection mechanisms in germinal centers *Math Med Biol* **23** 255–77

[100] Meyer-Hermann M, Mohr E, Pelletier N, Zhang Y, Victora G D and Toellner K-M 2012 A Theory of Germinal Center B Cell Selection, Division, and Exit *Cell Rep* **2** 162–74

[101] Deem M and Lee H 2003 Sequence Space Localization in the Immune System Response to Vaccination and Disease *Phys Rev Lett* **91** 068101

[102] Deem M W and Hejazi P 2010 Theoretical Aspects of Immunity *Annu Rev Chem Biomol Eng* **1** 247–76

[103] Wang S, Mata-Fink J, Kriegsman B, Hanson M, Irvine D J, Eisen H N, Burton D R, Wittrup K D, Kardar M and Chakraborty A K 2015 Manipulating the Selection Forces during Affinity Maturation to Generate Cross-Reactive HIV Antibodies *Cell* **160** 785–97

[104] Luo S and Perelson A S 2015 Competitive exclusion by autologous antibodies can prevent broad HIV-1 antibodies from arising *P Natl Acad Sci USA* **112** 11654–9

[105] Childs L M, Baskerville E B and Cobey S 2015 Trade-offs in antibody repertoires to complex antigens. *Philos T R Soc B* **370** 20140245

[106] Chaudhury S, Reifman J and Wallqvist A 2014 Simulation of B Cell Affinity Maturation Explains Enhanced Antibody Cross-Reactivity Induced by the Polyvalent Malaria Vaccine AMA1 *J Immunol* **193** 2073–86

[107] Perelson A S and Oster G F 1979 Theoretical studies of clonal selection: Minimal antibody repertoire size and reliability of self-non-self discrimination *J Theor Biol* **81** 645–70

[108] Julien J-P, Cupo A, Sok D, Stanfield R L, Lyumkis D, Deller M C, Klasse P-J, Burton D R, Sanders R W, Moore J P, Ward A B and Wilson I A 2013 Crystal Structure of a Soluble Cleaved HIV-1 Envelope Trimer *Science* **342** 1477–83

[109] Lyumkis D, Julien J-P, de Val N, Cupo A, Potter C S, Klasse P-J, Burton D R, Sanders R W, Moore J P, Carragher B, Wilson I A and Ward A B 2013 Cryo-EM Structure of a Fully Glycosylated Soluble Cleaved HIV-1 Envelope Trimer *Science* **342** 1484–90