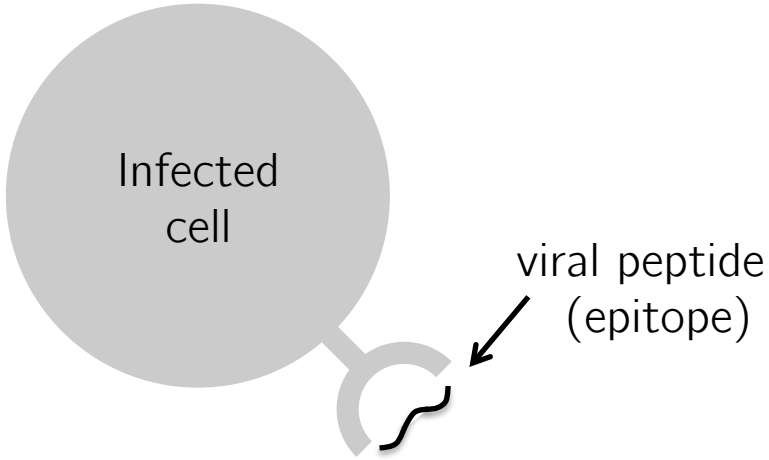
Statistical physics, inference, and the host-pathogen dynamics of HIV

John Barton, MIT

Rutgers Statistical Mechanics Meeting

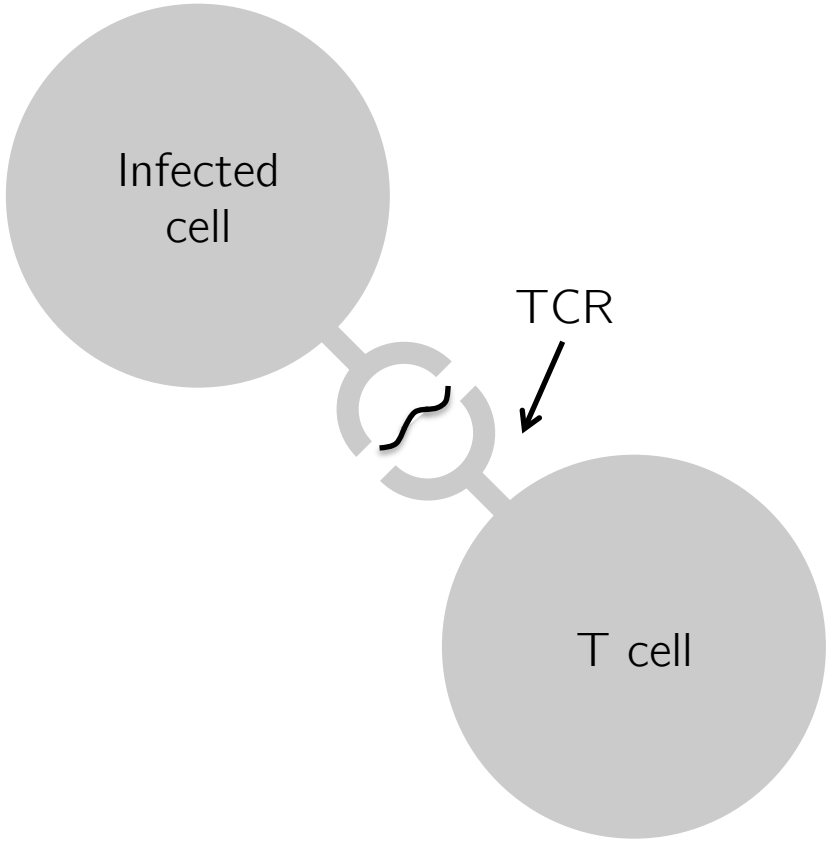
12-20-2016

Rapid evolution of HIV prevents effective immune control of infection



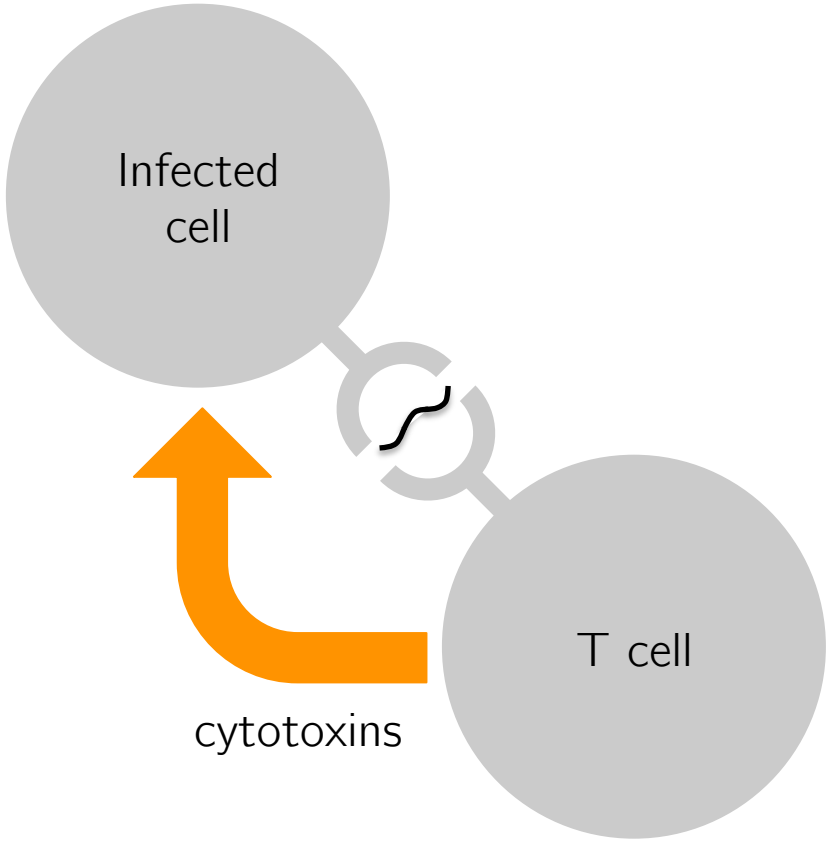
T cells kill infected cells
and help control infection

Rapid evolution of HIV prevents effective immune control of infection



T cells kill infected cells and help control infection

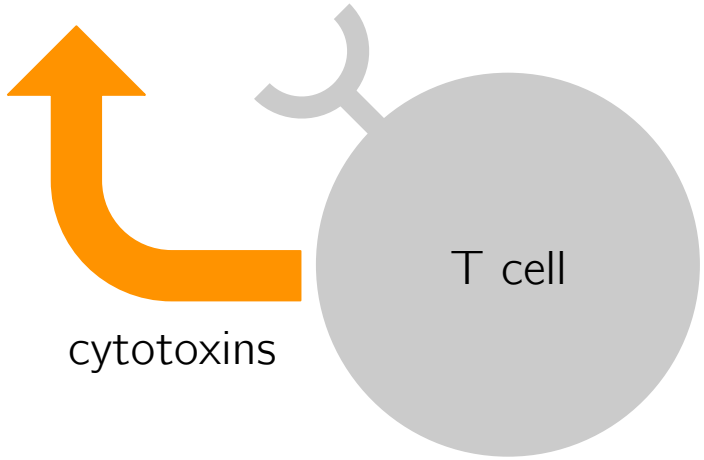
Rapid evolution of HIV prevents effective immune control of infection



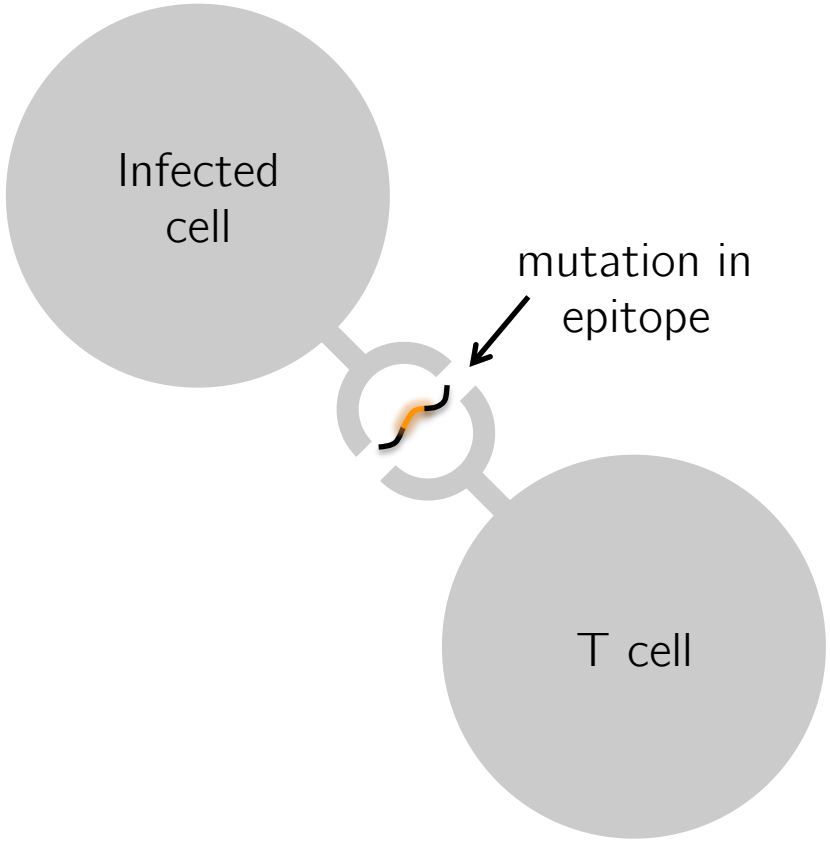
T cells kill infected cells and help control infection

Rapid evolution of HIV
prevents effective immune control of infection

T cells kill infected cells
and help control infection



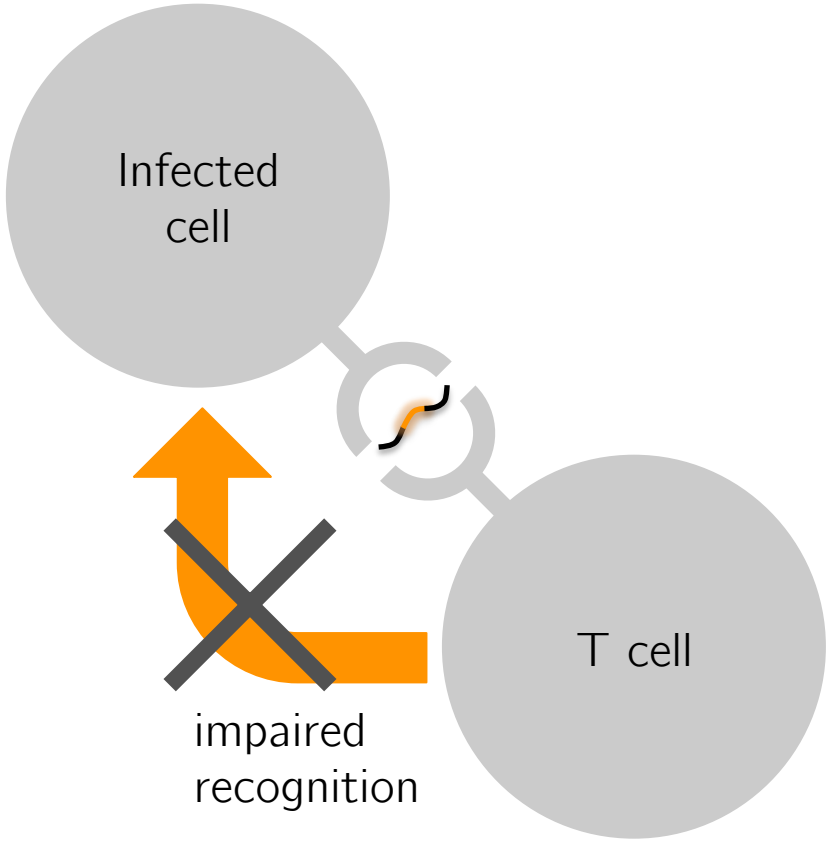
Rapid evolution of HIV prevents effective immune control of infection



T cells kill infected cells and help control infection

Escape mutations prevent T cell recognition allowing replication to continue unchecked

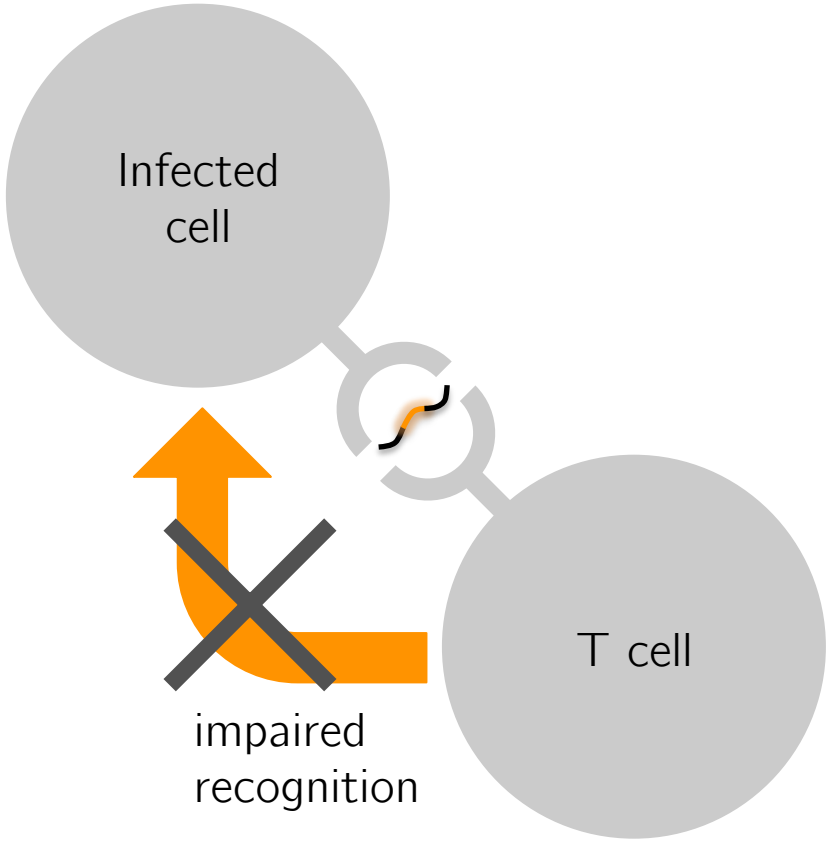
Rapid evolution of HIV prevents effective immune control of infection



T cells kill infected cells and help control infection

Escape mutations prevent T cell recognition allowing replication to continue unchecked

Rapid evolution of HIV prevents effective immune control of infection



T cells kill infected cells and help control infection

Escape mutations prevent T cell recognition allowing replication to continue unchecked

Goal: identify **mutational vulnerabilities** of HIV

Approach: use data to identify parts of viral sequences that are **highly constrained**

Maximum entropy modeling

Statistical inference and model tests

Coevolution

Understanding how HIV escapes from the immune system in vivo

Correlations quantify important constraints on HIV sequences

Consider a set of HIV sequences, isolated from many different patients

We notice that

- some sites are more variable
- mutations are correlated

These properties are quantified by correlations,

frequency of observing amino acid α at site i: $p_i(\alpha)$
frequency of observing amino acids α, β at sites i, j: $p_{ij}(\alpha, \beta)$

What is the simplest (maximum entropy) model consistent with these constraints?

Maximum entropy models allow us to map from sequence to fitness, but inference is challenging

The maximum entropy model reproducing the correlations is a Potts model

$$P(\underline{s}) = \frac{\exp(-E(\underline{s}))}{Z}, \qquad E(\underline{s}) = -\sum_{i=1}^{N} h_i(s_i) - \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} J_{ij}(s_i, s_j)$$

Fields and **couplings** are Lagrange multipliers enforcing the constraints

$$\sum_{\underline{s}} \delta(s_i, \alpha) P(\underline{s}) = p_i(\alpha)$$

$$\sum_{\underline{s}} \delta(s_i, \alpha) \delta(s_j, \beta) P(\underline{s}) = p_{ij}(\alpha, \beta)$$

Maximum entropy models allow us to map from sequence to fitness, but inference is challenging

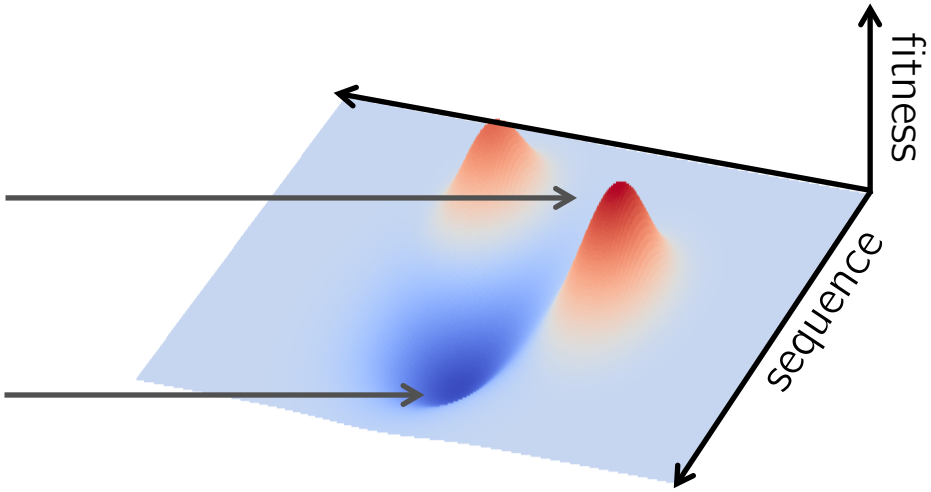
The maximum entropy model reproducing the correlations is a Potts model

$$P(\underline{s}) = \frac{\exp(-E(\underline{s}))}{Z}, \quad \boxed{E(\underline{s})} = -\sum_{i=1}^{N} h_i(s_i) - \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} J_{ij}(s_i, s_j)$$

Energy is negatively correlated with **fitness**

Most prevalent (low energy) sequences should be fittest

Rarely observed (high energy) sequences should be unfit



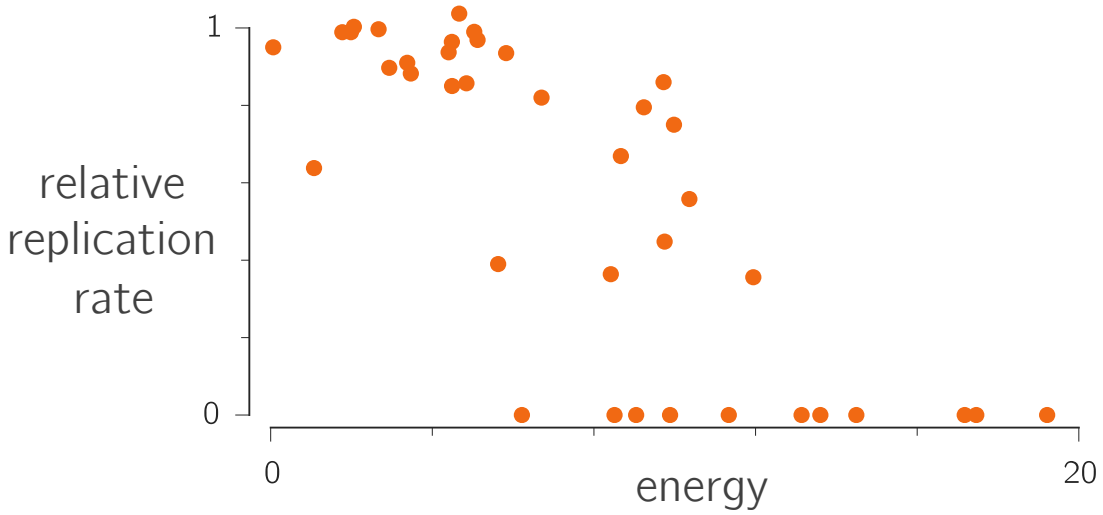
Ferguson et al, Immunity (2013)
Mann=, JPB=, Ferguson= et al, PLOS Comp Biol (2014)

Maximum entropy models allow us to map from sequence to fitness, but inference is challenging

The maximum entropy model reproducing the correlations is a Potts model

$$P(\underline{s}) = \frac{\exp(-E(\underline{s}))}{Z}, \quad \boxed{E(\underline{s})} = -\sum_{i=1}^{N} h_i(s_i) - \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} J_{ij}(s_i, s_j)$$

Energy is negatively correlated with **fitness**



Ferguson et al, Immunity (2013)
Mann=, JPB=, Ferguson= et al, PLOS Comp Biol (2014)

Maximum entropy models allow us to map from sequence to fitness, but inference is challenging

The maximum entropy model reproducing the correlations is a Potts model

$$P(\underline{s}) = \frac{\exp(-E(\underline{s}))}{Z}, \quad E(\underline{s}) = -\sum_{i=1}^{N} h_i(s_i) - \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} J_{ij}(s_i, s_j)$$

Energy is negatively correlated with **fitness**

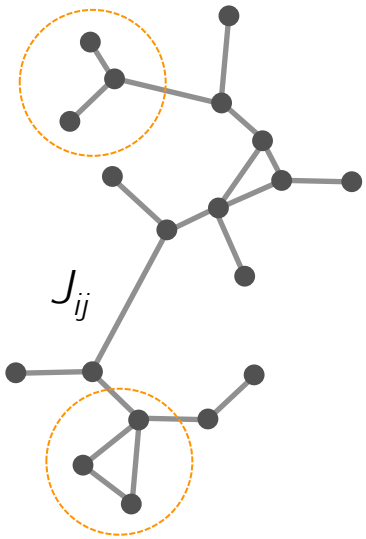
Need to find parameters that maximize the (average log-) likelihood

$$-\mathcal{L} = \log Z - \sum_i \sum_\alpha h_i(\alpha)\, p_i(\alpha) - \sum_{i,j} \sum_{\alpha,\beta} J_{ij}(\alpha, \beta)\, p_{ij}(\alpha, \beta)$$

$$Z = \sum_{\underline{s}} \exp\left(-E(\underline{s})\right)$$

Number of terms in Z is **exponential** in the system size

Adaptive cluster expansion (ACE) builds global solutions from local ones



Assumption: to accurately estimate the parameters, we can use local information

Formally, assume that the inverse susceptibility $\chi^{-1} = dJ/dp$ is short-ranged

Expand the likelihood (entropy) in terms of contributions from **clusters** of sites Γ

$$\mathcal{L} = \sum_{\Gamma} \Delta\mathcal{L}_{\Gamma} \,, \quad \boxed{\Delta\mathcal{L}_{\Gamma}} = \boxed{\mathcal{L}_{\Gamma}} - \sum_{\Gamma' \subset \Gamma} \Delta\mathcal{L}_{\Gamma'}$$

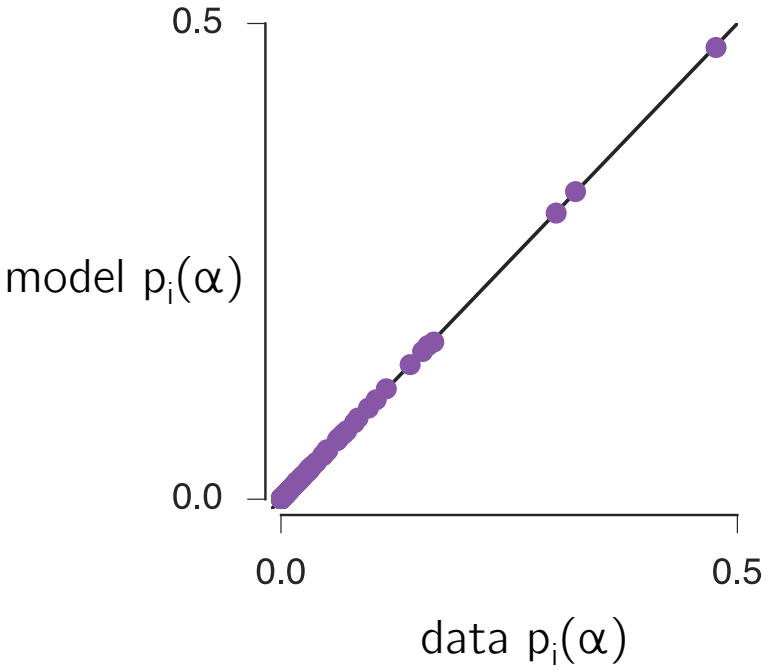
Likelihood of the model restricted to the sites in Γ alone

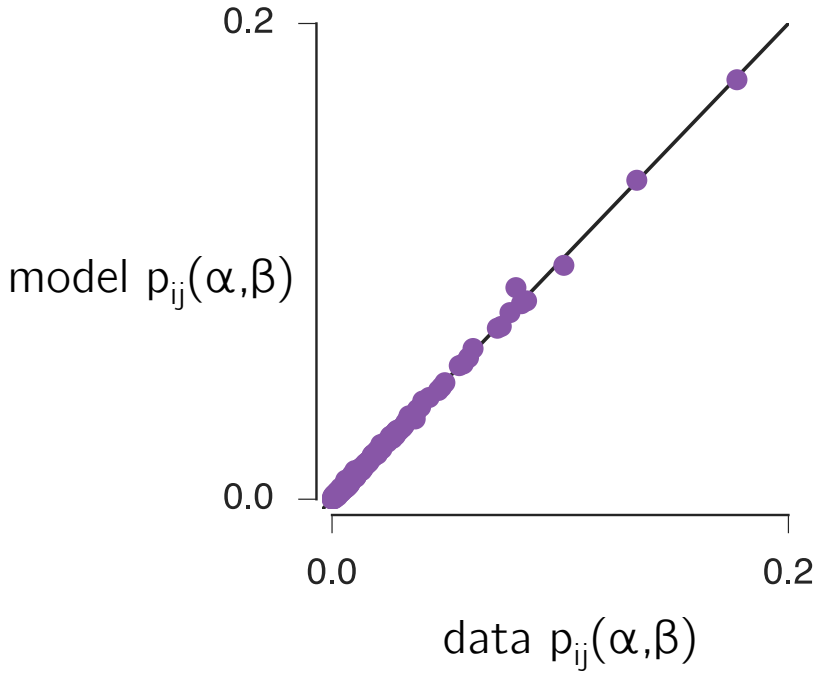
Contribution to the likelihood from the cluster Γ which is not captured by any subsets of Γ

This **cluster expansion** may rapidly converge, can computationally select out largest terms (**adaptive**)

Cocco & Monasson, PRL (2011)
Cocco & Monasson, JSP (2012)
JPB & Cocco, JSM (2013)

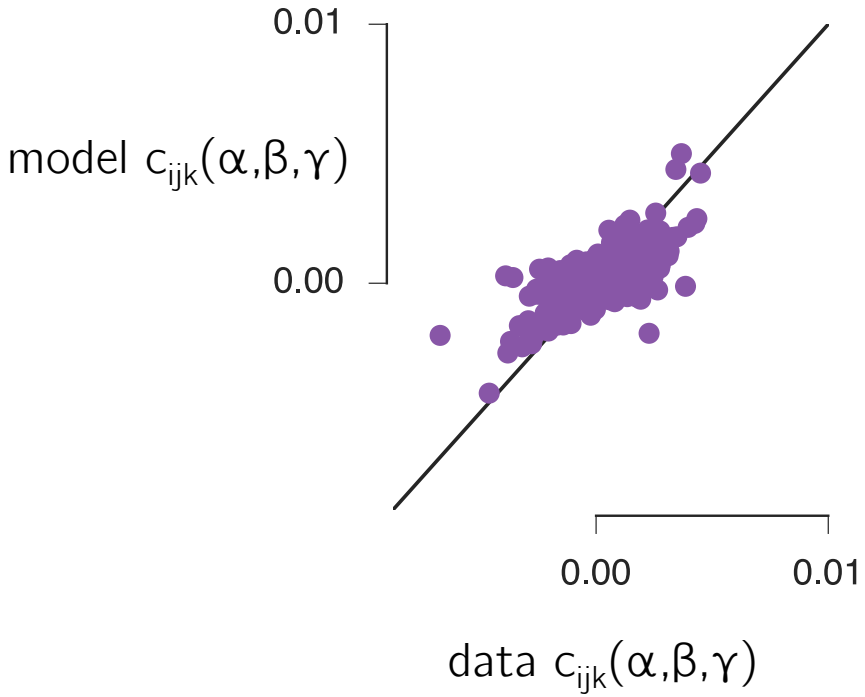
ACE accurately recovers statistics of the data,
including some higher order correlations

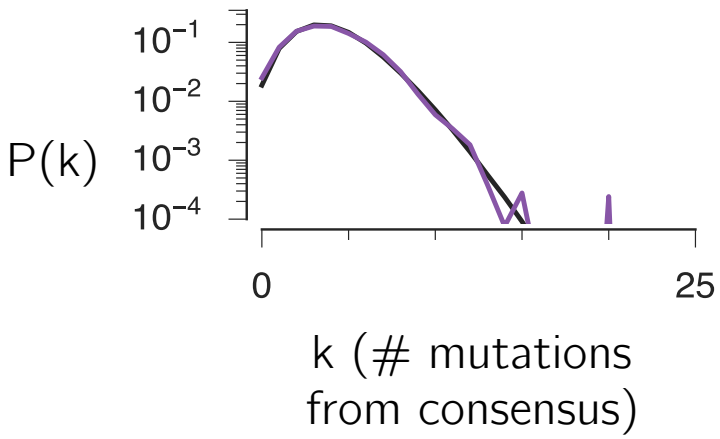




JPB, De Leonardis, Coucke, & Cocco, Bioinformatics (2016)
code @ github.com/johnbarton/ACE

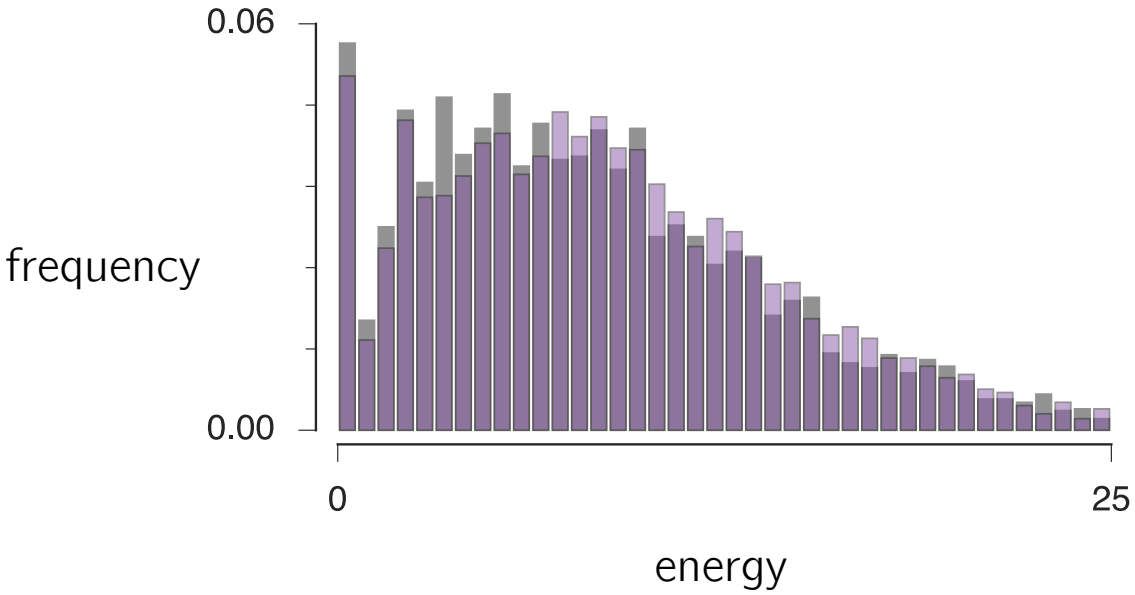
ACE accurately recovers statistics of the data, including some higher order correlations





JPB, De Leonardis, Coucke, & Cocco, Bioinformatics (2016)
code @ github.com/johnbarton/ACE

ACE accurately recovers statistics of the data, including some higher order correlations



Distribution of energies in the **model** matches with the **data**, samples are consistent

JPB, De Leonardis, Coucke, & Cocco, Bioinformatics (2016)
code @ github.com/johnbarton/ACE

Maximum entropy modeling

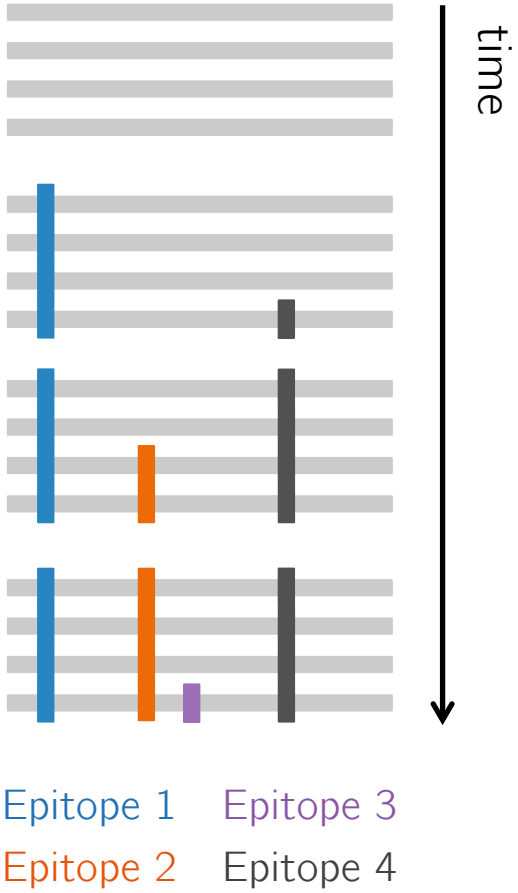
Statistical inference and model tests

Coevolution

Understanding how HIV escapes from the immune system in vivo

JPB et al., Nature Communications (2016)

In vivo evolution of viral populations can be studied in clinical data



CHAVI longitudinal study of 17 individuals, starting from initial infection with follow-up over 1-3 years

At each visit, sample sequences from the viral population, map T cell responses

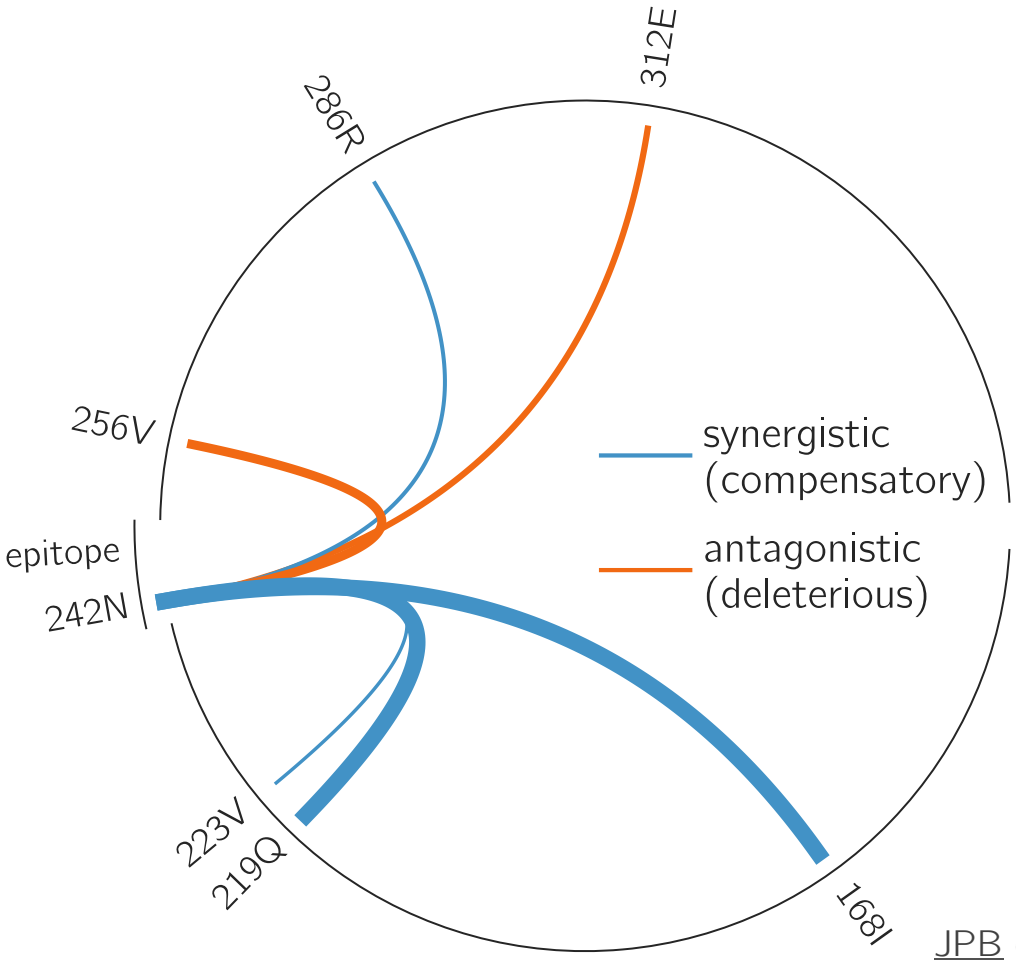
We want to predict **timing** and **location** of escape mutations

Escape should be slower when mutation significantly increases E

Liu et al., J Clin Invest (2013)

The model captures effects of the viral sequence background on escape

Gag TW10 epitope is usually **protective** (high fitness cost of escape), but escaped rapidly in one patient



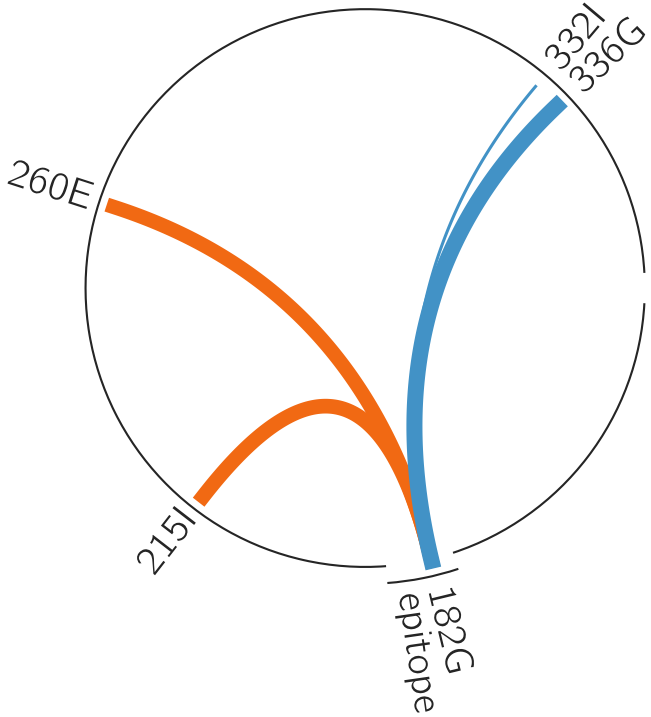
Plot of all strong ($|J|>0.1$) interactions between escape mutation and the sequence background

Known compensatory mutations 219Q, 223V, reduce fitness cost of escape

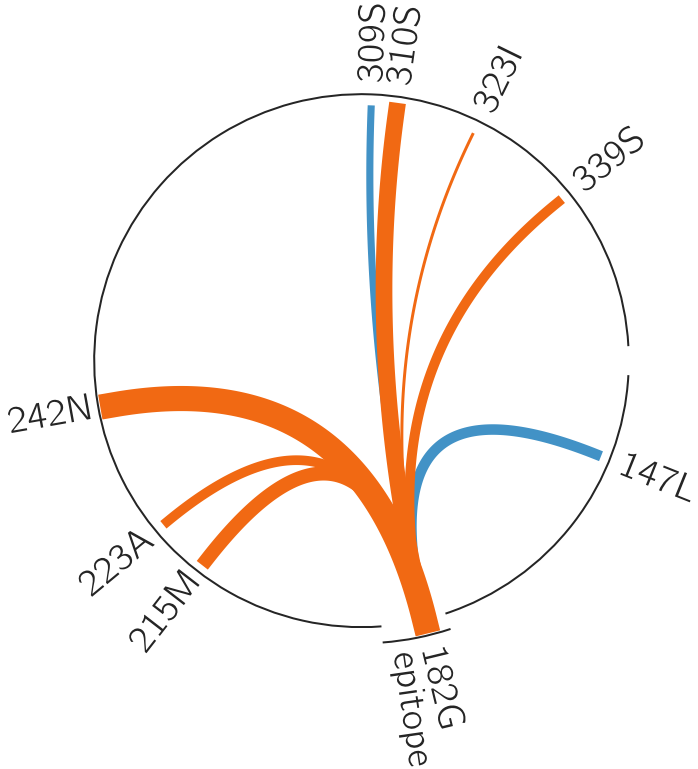
JPB et al., Nature Communications (2016)

The model captures effects of the viral sequence background on escape

Identical epitopes targeted by different patients:



Escape after 120 days

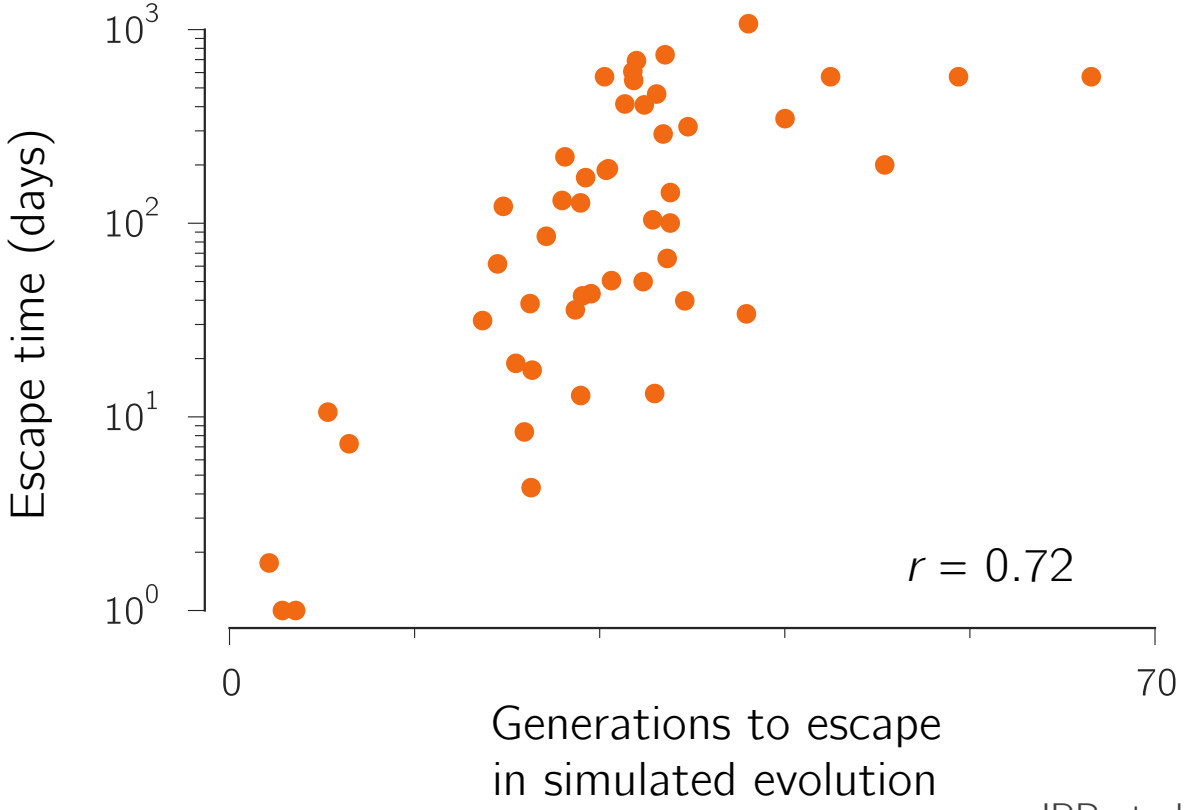


Escape never observed (>1103 days)

JPB et al., Nature Communications (2016)

Simulate viral evolution by coupling the fitness landscape to a Wright-Fisher-like model

Simulate evolution of virus population, with selective advantage of escape **based on strength of immune response**
Measure number of generations to escape in simulation



True locations of escape mutations also typically match simulations (≈85%)

JPB et al., Nature Communications (2016)

Acknowledgments

MIT	Arup Chakraborty Mehran Kardar Thomas Butler Kevin Kaczorowski Dariusz Murakowski Karthik Shekhar
UKZN	Thumbi Ndung'u Jaclyn Mann Erasha Rajkoomar
Ragon Institute	Bruce Walker
Oxford	Andrew McMichael
UNC	Nilu Goonetilleke
UIUC	Andrew Ferguson
ENS	Rémi Monasson Simona Cocco Eleonora De Leonardis Alice Coucke

Ragon Institute of MGH, MIT and Harvard

MIT

imes

Summary

- ACE infers excellent generative models from input HIV sequence data (and others, see github.com/johnbarton/ACE)

- HIV sequence data reveals information about the fitness landscape of the virus

- Given information about fitness, sequences in the viral population, and the immune response, escape from T cell pressure is predictable (probabilistically)

- Signals of host-pathogen coevolution imprinted in viral sequences

  JPB, Kardar, & Chakraborty, PNAS (2015)

- HIV protease drug resistance

  Butler[=], JPB[=], Kardar, & Chakraborty, Phys Rev E (2016)