FIGURES FOR CHAPTER 4

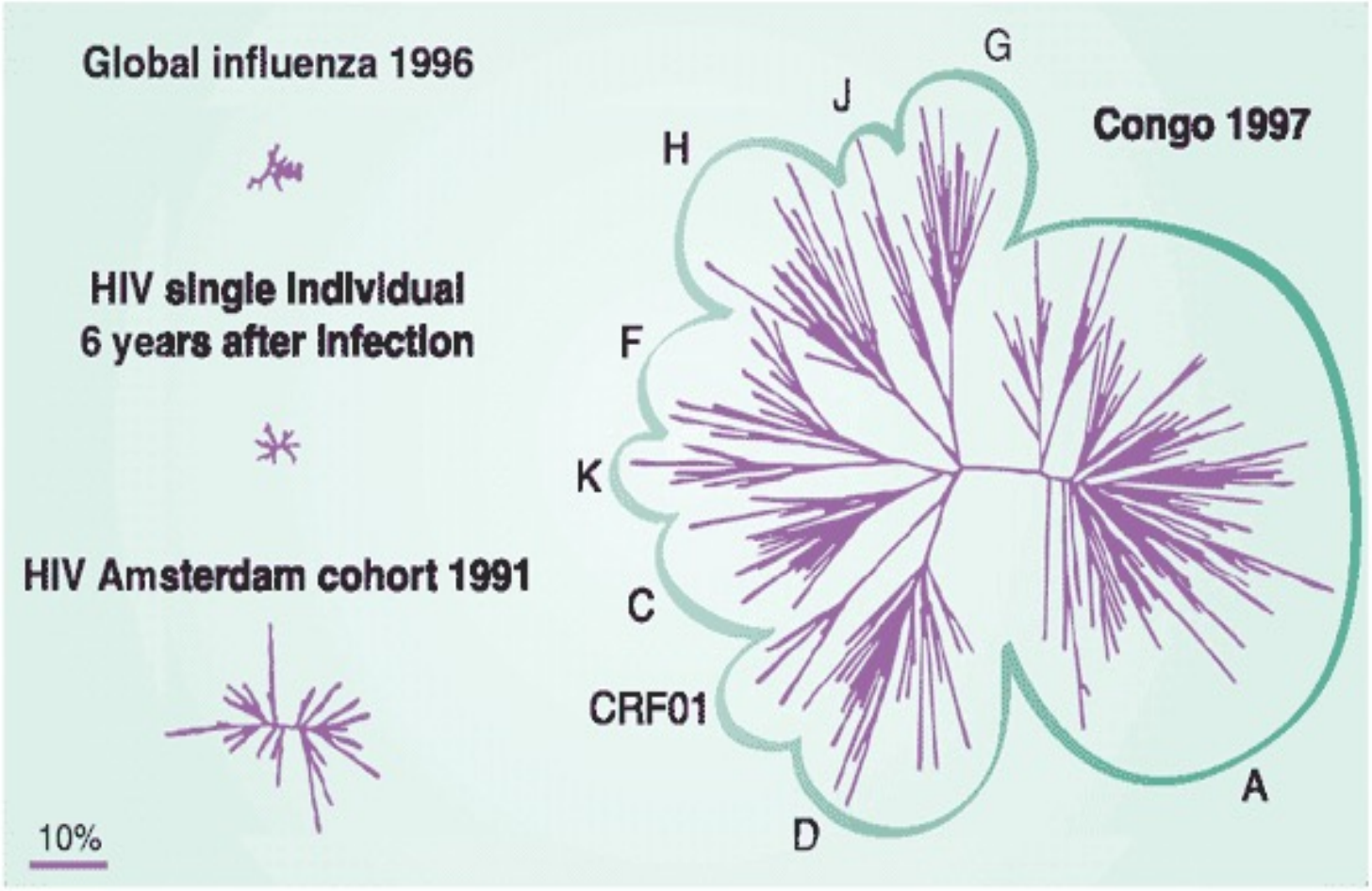


Fig. 4.1: The diversity of the influenza virus in the entire world in a particular year is compared to the diversity of HIV strains in a single region in Africa, in a single HIV-infected person, and a cohort of HIV-infected people. The number of ends in the phylogenetic trees shown reflects the number of circulating strains. This figure is taken from [11].

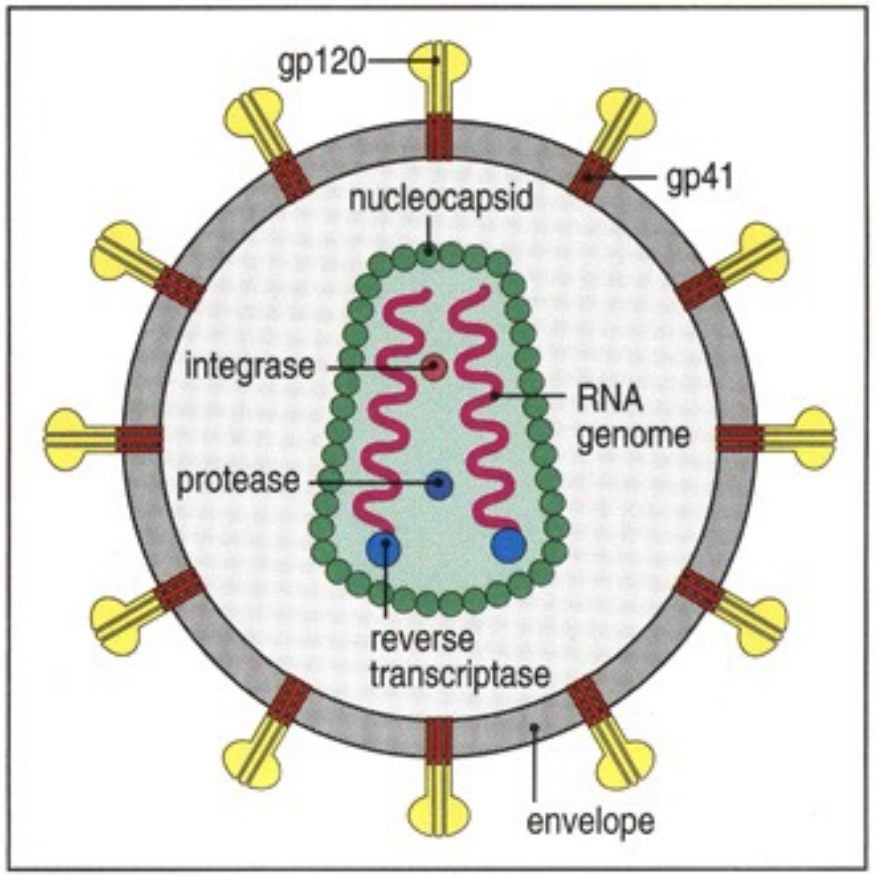


Fig. 4.2: HIV is an enveloped retrovirus. The virus' spike, made up of gp 120 and gp 41 proteins, protrudes through an encapsulating membrane. A nucleocapsid, made up of structural proteins encloses HIV's RNA genome and other key proteins required for virus replication. This figure is taken from Janeway's Immunobiology.

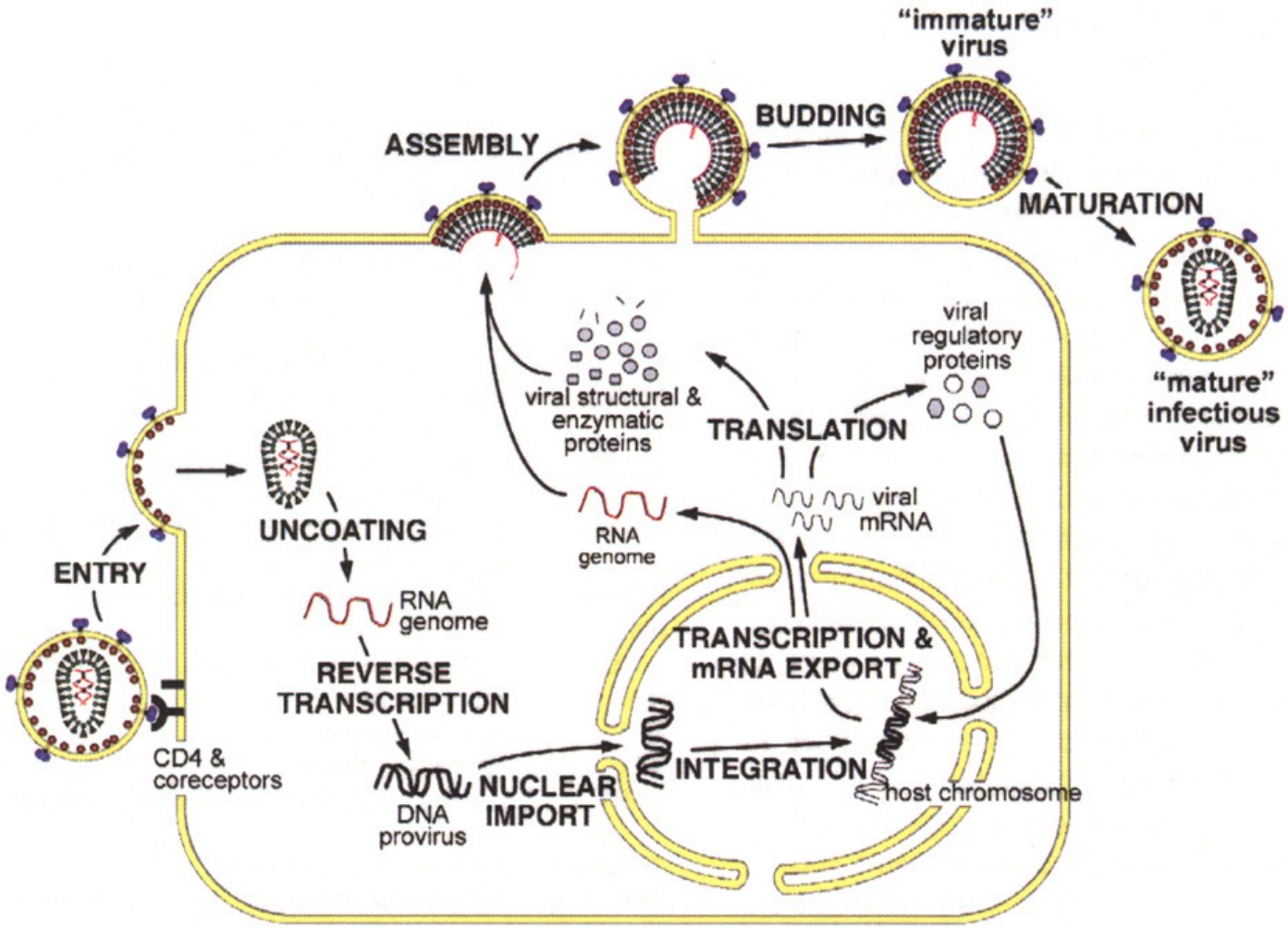


Fig. 4.3: Life cycle of HIV. This figure is taken from Janeway's Immunobiology.

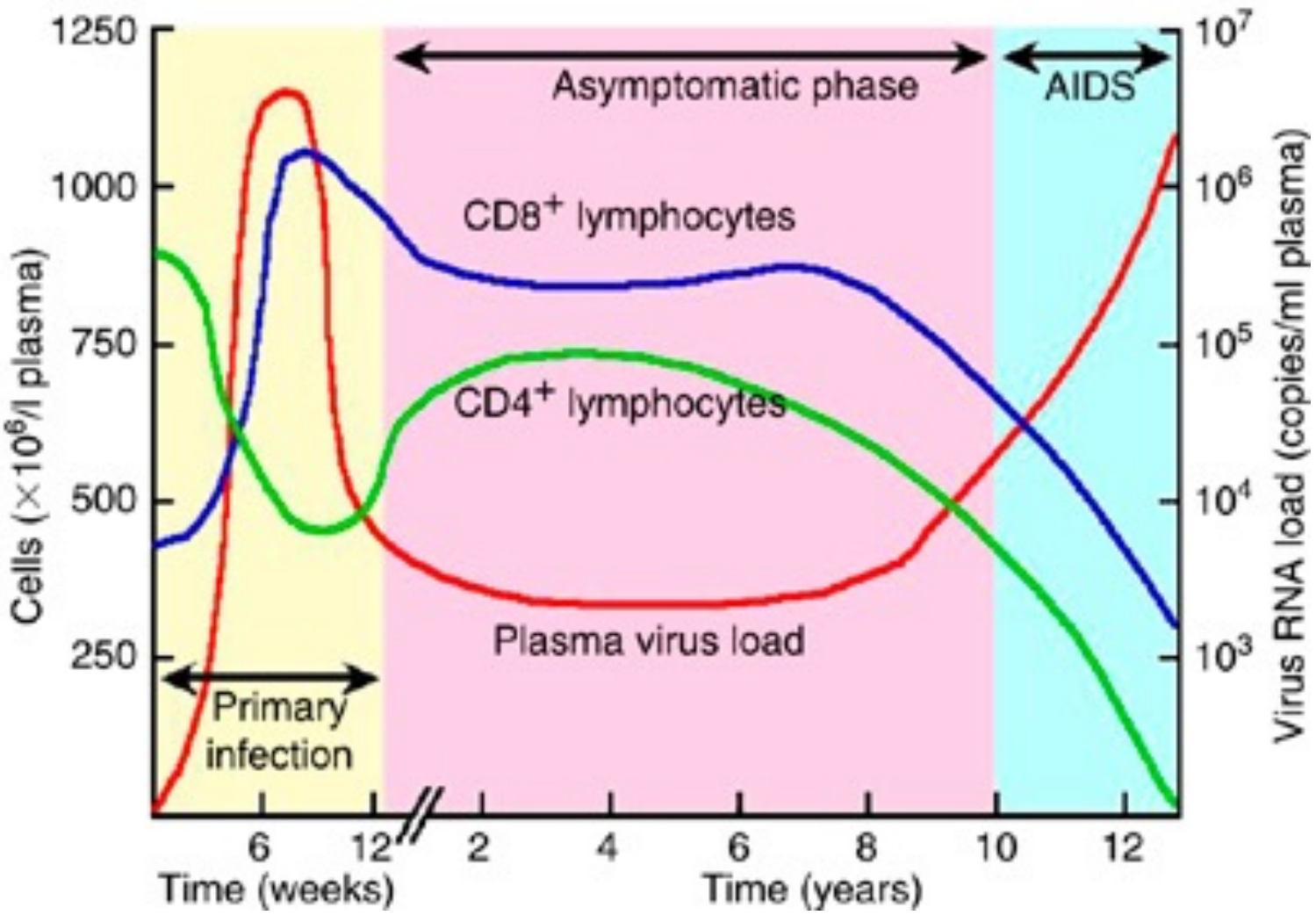


Fig. 4.4: Temporal evolution of viral load and T cells after HIV infection. Antibodies also target HIV, but their evolution is not shown. This figure is taken from Janeway's Immunobiology.

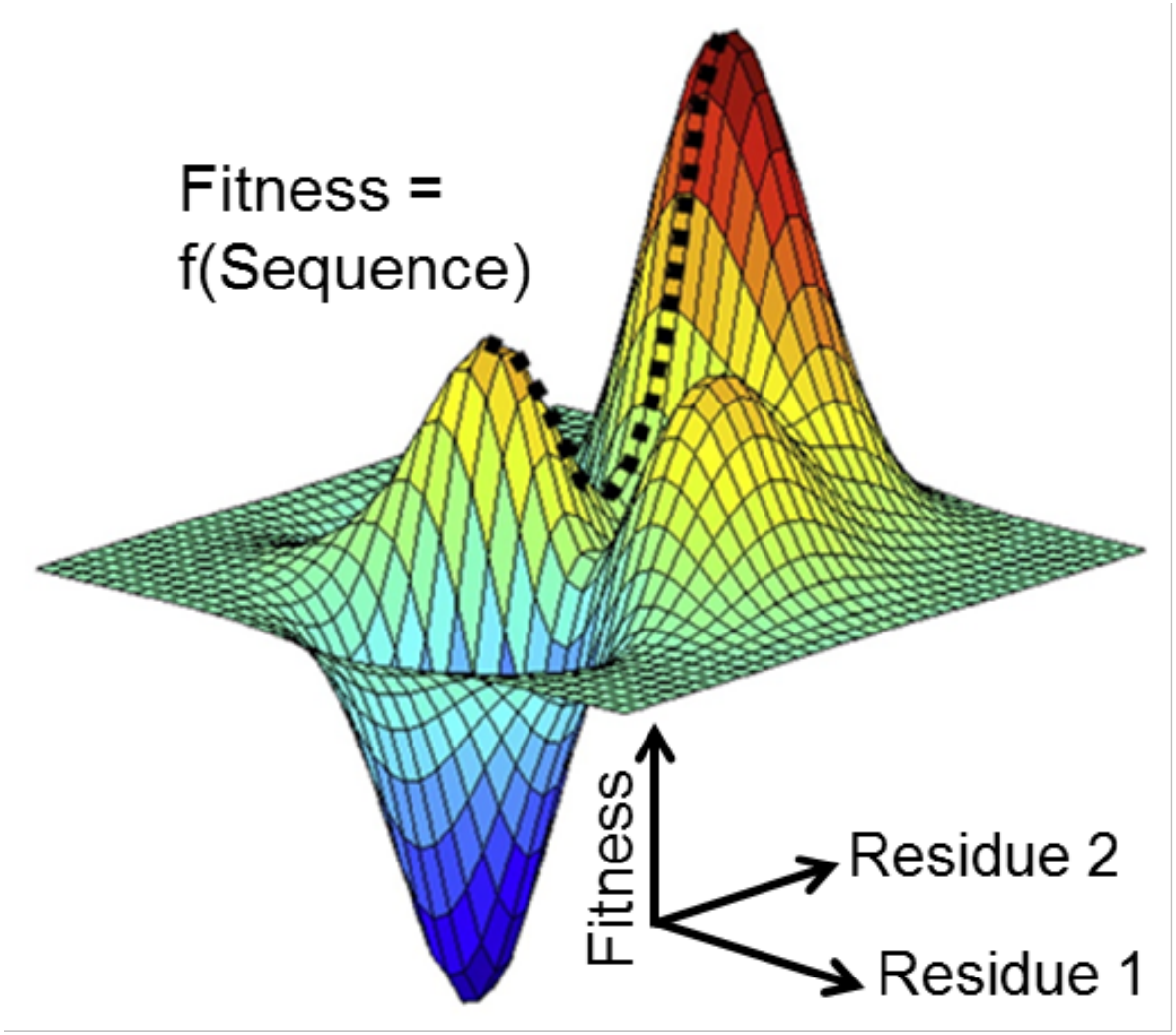


Fig. 4.5: Illustration of a fitness landscape for a 2-dimensional representation of sequence space. Each grid point on the two axes that make up the plane represent variant amino acids at two sites of a cartoon viral protein. Information about the ability of the virus to replicate, function, and propagate infection is shown in the vertical dimension, with hills and valleys representing regions of high and low fitness, respectively. The dotted line represents a compensatory mutational pathway that can evade immune responses.

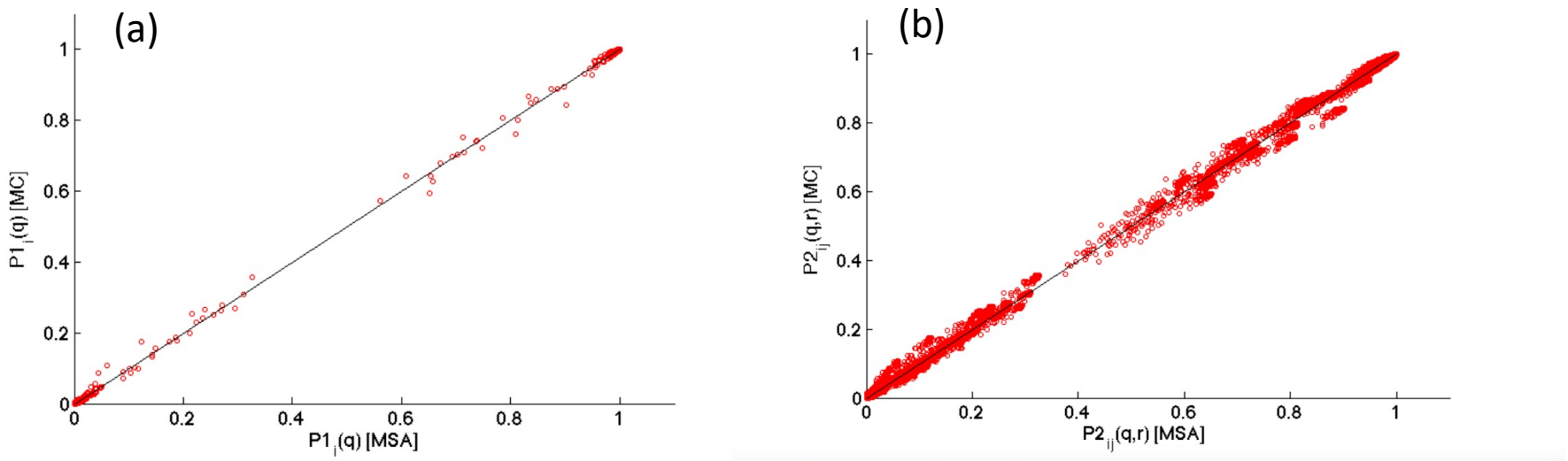


Fig. 4.6: Comparison between the inferred and observed mutational correlations for the HIV protein (p24) that forms the viral nucleocapsid. Panels (a) and (b) show the one and two-point correlations, respectively. The ordinate shows the values obtained by sampling P (**z**) with the inferred fields and couplings using Monte Carlo (MC) simulations and the abscissa shows the corresponding values observed by aligning the available sequences (MSA) of p24.

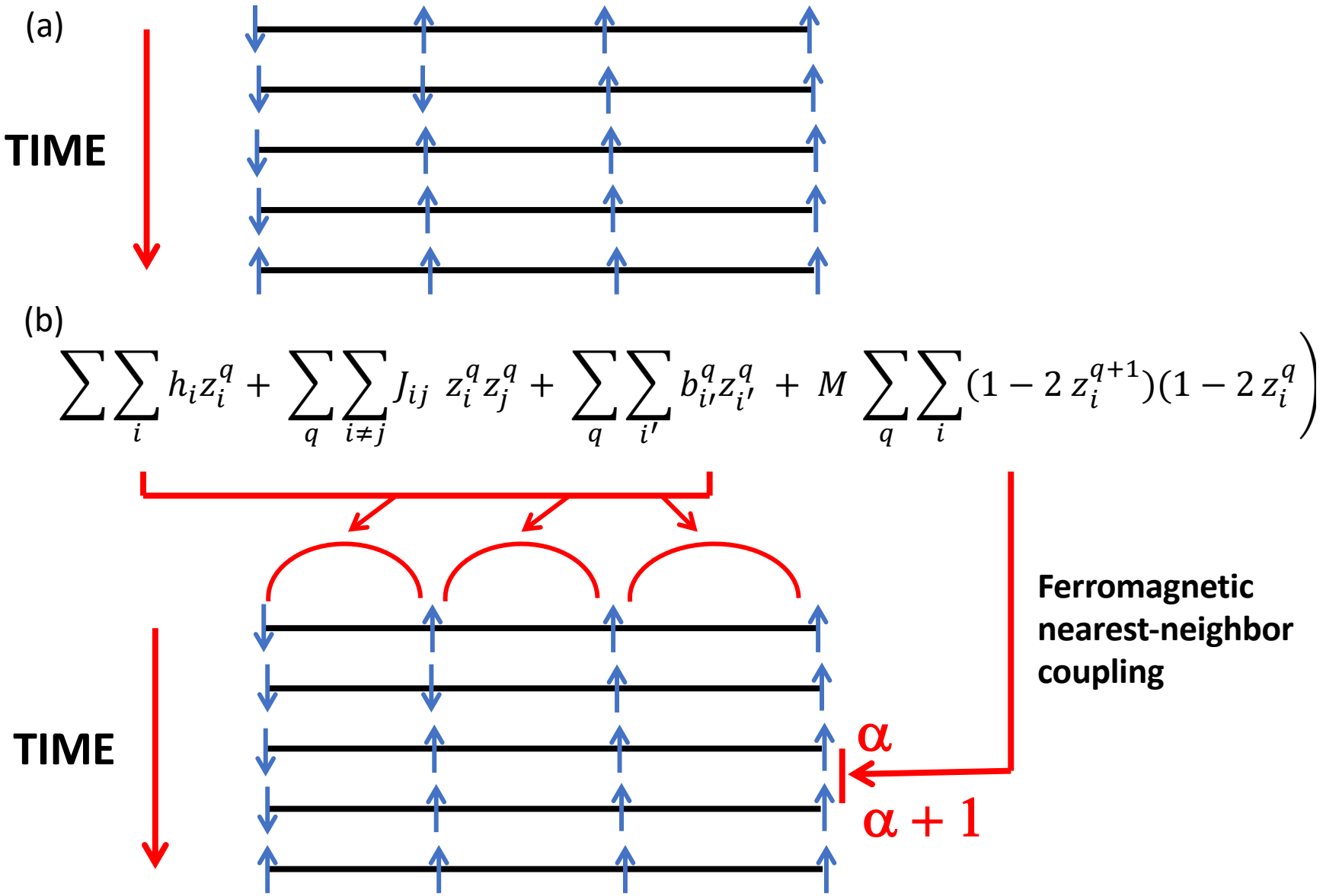


Fig. 4.7: (a) Depiction of an evolutionary trajectory. Each row is a viral strain, with the arrows representing amino acids in the Ising representation (up is wild type and down is a mutant). (b) Depiction of the equations that describe the weight of a particular evolutionary trajectory.

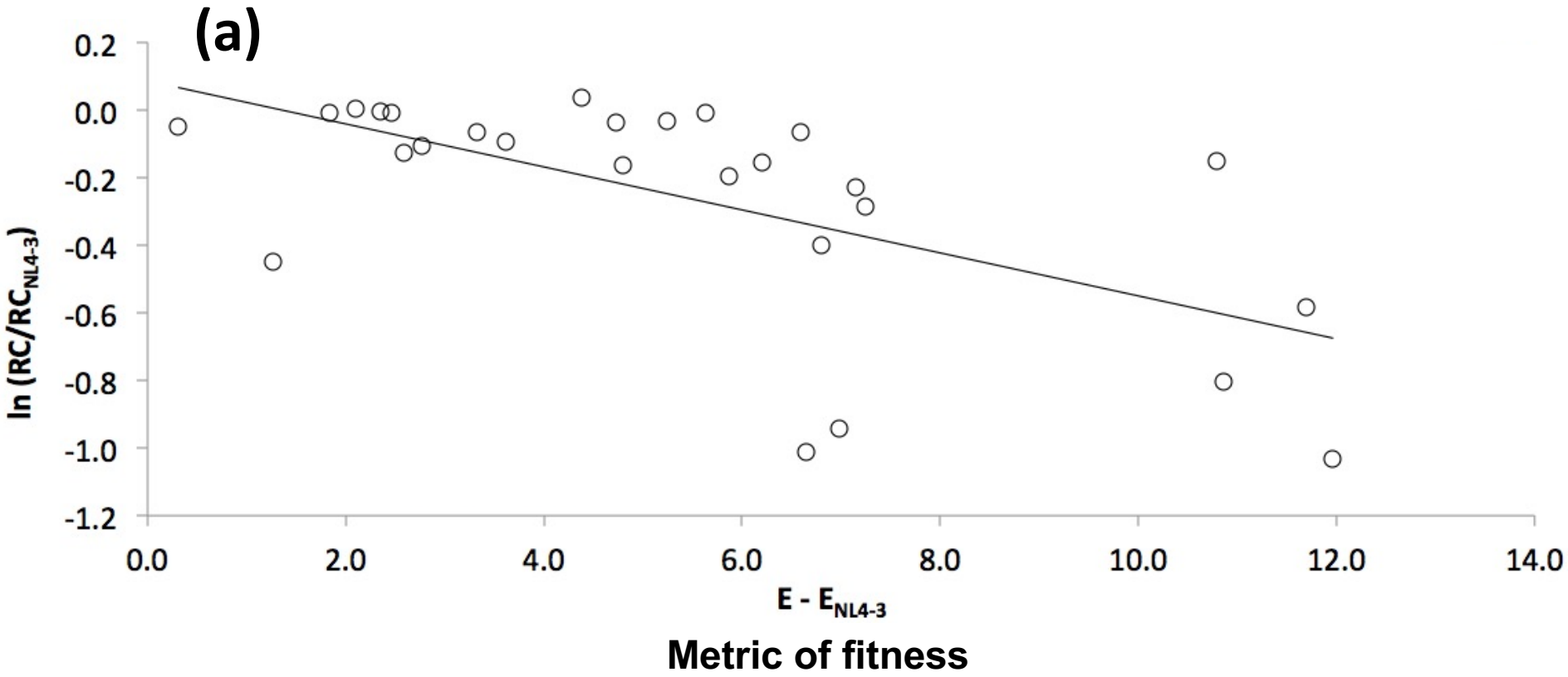
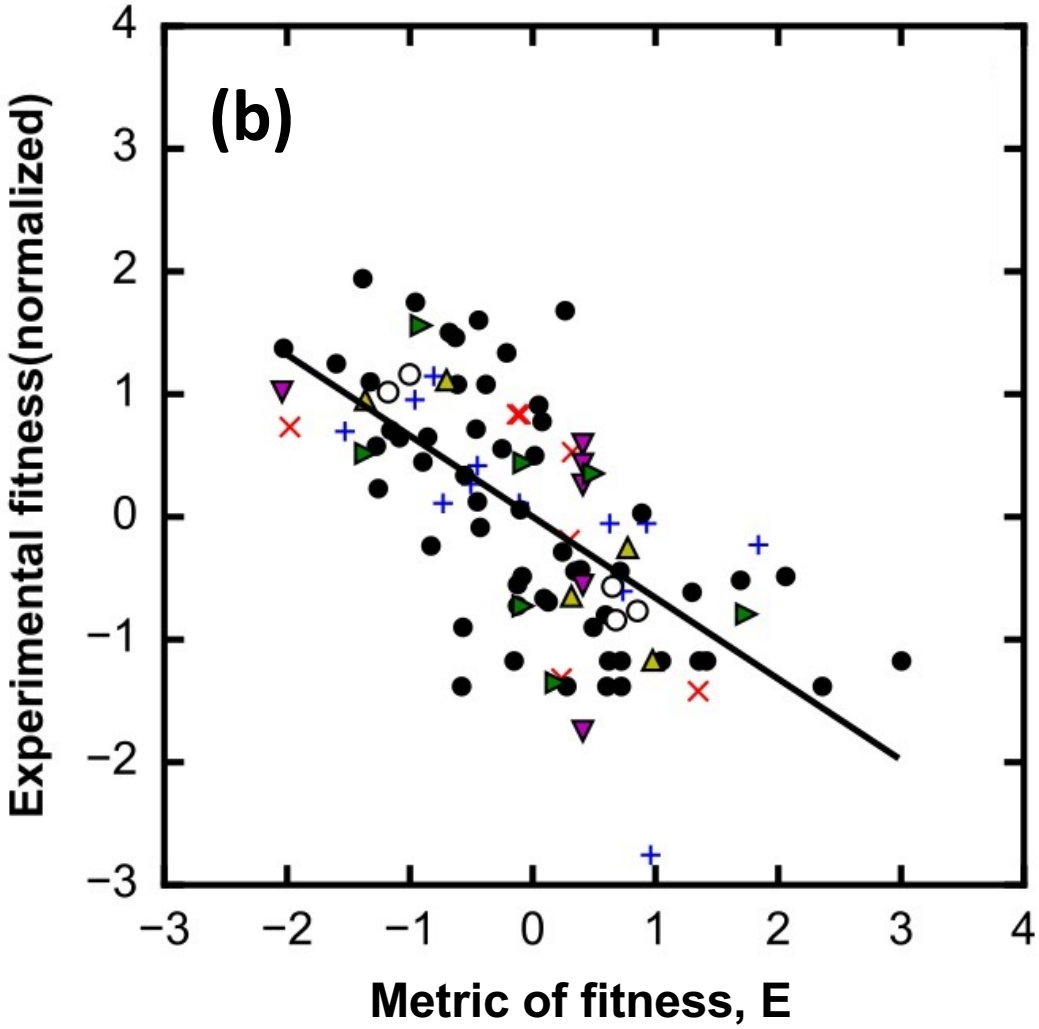


Fig. 4.8: Comparison of predictions of relative fitness of HIV mutant strains with in vitro experiments that measure the ability of mutant strains to infect human cells and grow out. (a) Comparison for 36 strains bearing mutations in the Gag polyprotein. The reference strain used in the experiments is NL4-3. The abscissa shows values of the "energy", E (or Hamiltonian) for the mutants relative to NL4-3 obtained from Eq. 2 and the inferred parameters. The ordinate shows values of the replicative capacity (RC) corresponding to the growth rate of mutant strains relative to NL4-3. The circles represent the measurements and the line corresponds to Eq. 2 with the best fit slope. Data for 36 strains were compared, but only 27 circles are shown because 9 strains predicted to have very low fitness did not grow out. Overall, with high statistical significance (p value = 10^{-8}, the correlation between data and experiments = - 0.8. (b) A similar comparison for mutant strains bearing mutations in the ENV polyprotein. The correlation between experiments and predictions is 0.73.


(b)
Experimental fitness(normalized)
Metric of fitness, E
4
3
2
1
0
−1
−2
−3
−3
−2
−1
0
1
2
3
4

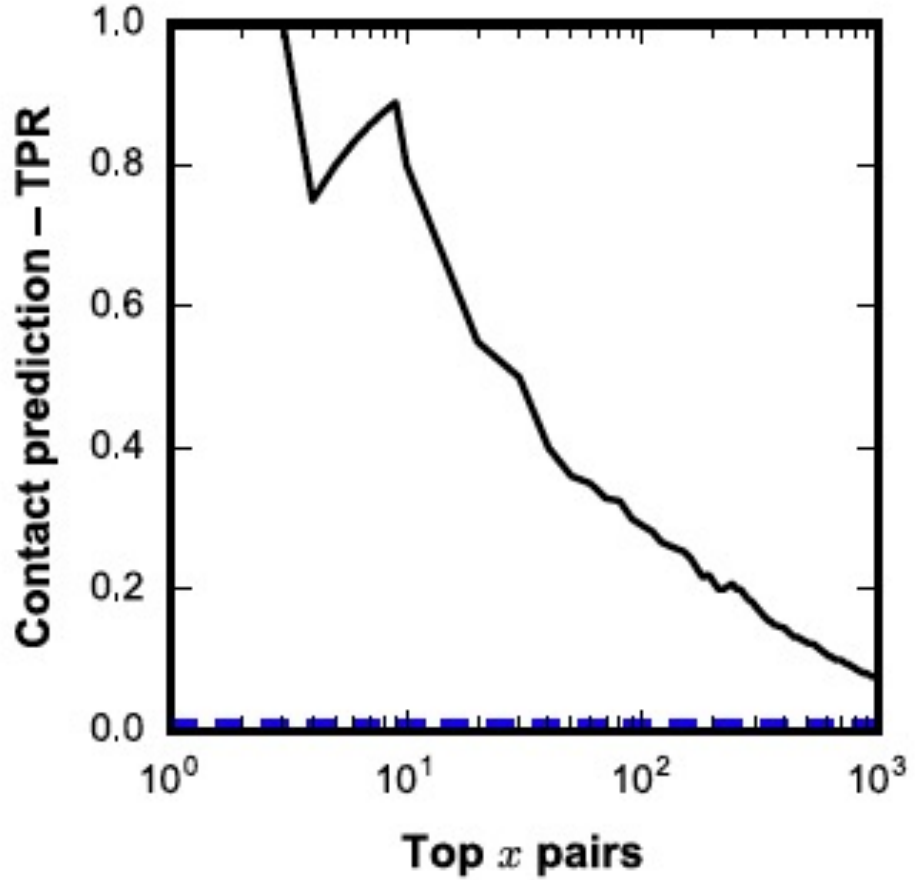
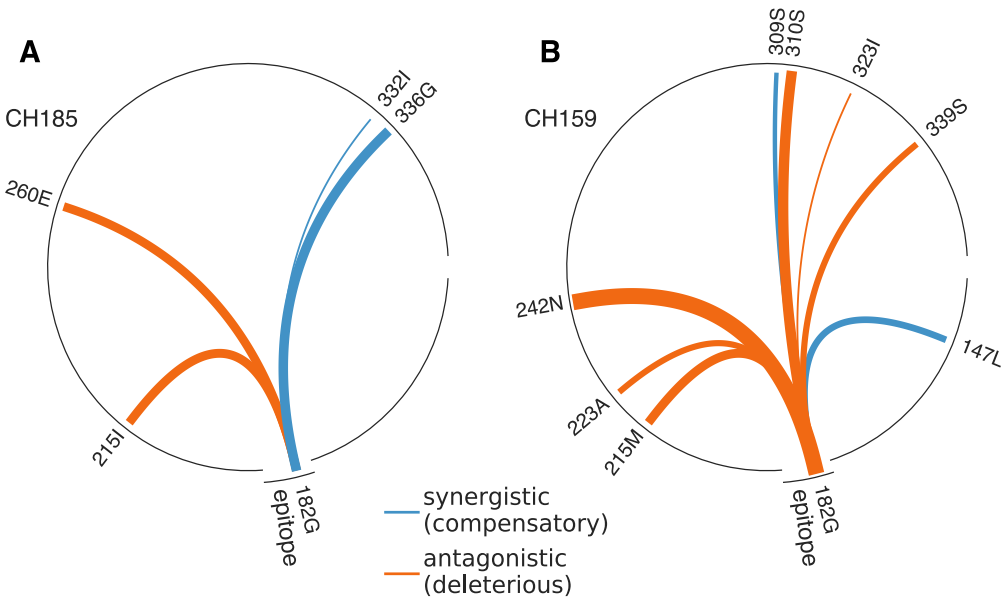


Fig. 4.9: Fraction of contacts in the top "x" predicted pairs that are true contacts (TPR) in the SOSIP crystal structure graphed against the top "x" predicted pairs based on a high value of the corresponding element (J_{ij}) of the **J** matrix (black line). Two sites are predicted to be in contact in the crystal structure if they are within a threshold distance from each other. The blue dotted line corresponds to the number of contacts observed in the crystal structure divided by the number of pairs of sites, and so reflects the chance that a pair of sites would be predicted to be in contact by chance.

TESTS AGAINST CLINICAL DATA (w/McMichael lab)

- Most likely or second most likely sites of escape mutations predict observed escape mutation with ~ 86 % accuracy.

- Our results highlight the importance of collective compensatory pathways and sequence background



C

Epitope	Patient	Fitness cost ΔE	Escape time (days)
TPQDLNTML (TL9)	CH185	4.3	120
	CH159	6.1	>1103*
TSTLQEQVAW (TW10)	CAP239	-1.4	1
	CH198	0.1	220
WHLGHGVSI (WI9)	CAP210	2.8	129
	CAP45	4.7	406†
EEVGFPVRPQV (EV11)	CH164	2.6	31
	CAP45	4.0	42

*No escape observed (final sequencing time)

Fig. 4.10: (A and B) Two patients (labeled CH185 and CH159) that target the same peptide epitope (the epitope is called TL9). Schematic depiction of the couplings between the escape mutation and mutations that existed in the rest of the protein from which the peptide is derived in viruses that infected the patients (CH185, A; CH159, B). Compensatory interactions are shown in blue and antagonistic ones in red. Thicker lines correspond to larger values of J_{ij}. (C) Three other examples showing how fitness cost calculated using Eq. 45 correlates with escape times in pairs of other patients targeting the same epitope.

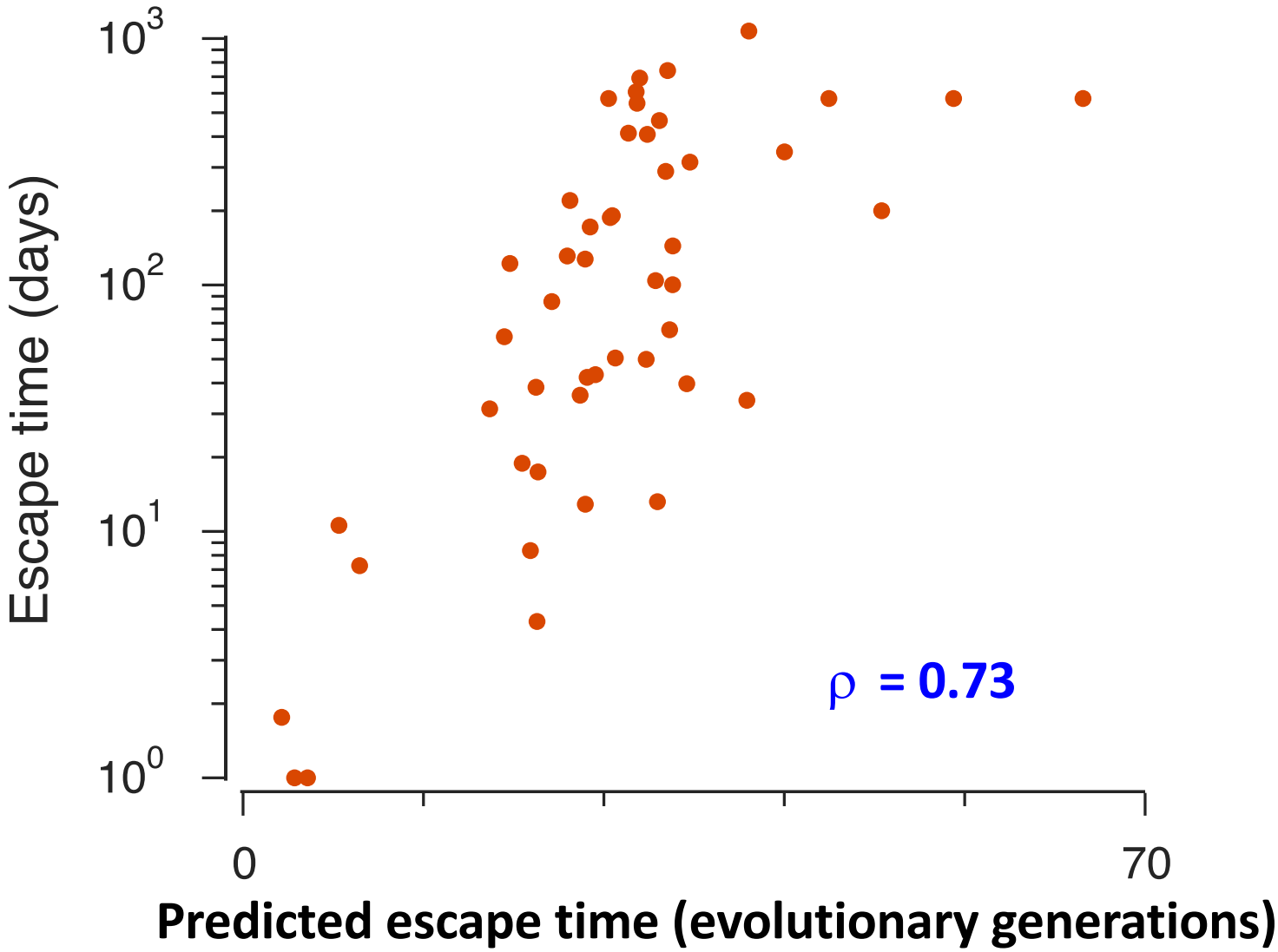


Fig. 4.11: Comparison of the measured escape time for various epitopes targeted by T cells in patients with the predicted escape time from the simulations (measured in generations of evolution. ρ is the Spearman rank correlation between measured and predicted values.