Author/s:
Zhang, H;Bull, RA;Quadeer, AA;McKay, MR

Title:
HCV E1 influences the fitness landscape of E2 and may enhance escape from E2-specific antibodies

Date:
2023-01-01

Citation:
Zhang, H., Bull, R. A., Quadeer, A. A. & McKay, M. R. (2023). HCV E1 influences the fitness landscape of E2 and may enhance escape from E2-specific antibodies. Virus Evolution, 9 (2), https://doi.org/10.1093/ve/vead068.

Persistent Link:
https://hdl.handle.net/11343/344452

VIRUS EVOLUTION

Virus Evolution, 2023, **9(2)**, 1–13

DOI: https://doi.org/10.1093/ve/vead068

Advance Access Publication 18 November 2023

Research Article

HCV E1 influences the fitness landscape of E2 and may enhance escape from E2-specific antibodies

Hang Zhang,[1,†] Rowena A. Bull,[2,3,‡] Ahmed Abdul Quadeer,[1,4,*,§] and Matthew R. McKay[4,5,*,**]

[1]Department of Electronic and Computer Engineering, The Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong, SAR, China, [2]School of Biomedical Sciences, Faculty of Medicine and Health, University of New South Wales, Sydney, NSW 2052, Australia, [3]The Kirby Institute for Infection and Immunity, Sydney, NSW 2052, Australia, [4]Department of Electrical and Electronic Engineering, University of Melbourne, Parkville, VIC 3010, Australia and [5]Department of Microbiology and Immunology, The Peter Doherty Institute for Infection and Immunity, University of Melbourne, Melbourne, VIC 3000, Australia

[†]https://orcid.org/0000-0002-4352-888X

[‡]https://orcid.org/0000-0002-9844-3744

[§]https://orcid.org/0000-0002-5295-9067

[**]https://orcid.org/0000-0002-8086-2545

*Corresponding authors: E-mail: ahmed.quadeer@unimelb.edu.au; matthew.mckay@unimelb.edu.au

Abstract

The Hepatitis C virus (HCV) envelope glycoprotein E1 forms a non-covalent heterodimer with E2, the main target of neutralizing antibodies. How E1–E2 interactions influence viral fitness and contribute to resistance to E2-specific antibodies remain largely unknown. We investigate this problem using a combination of fitness landscape and evolutionary modeling. Our analysis indicates that E1 and E2 proteins collectively mediate viral fitness and suggests that fitness-compensating E1 mutations may accelerate escape from E2-targeting antibodies. Our analysis also identifies a set of E2-specific human monoclonal antibodies that are predicted to be especially resilient to escape via genetic variation in both E1 and E2, providing directions for robust HCV vaccine development.

Keywords: Hepatitis C virus; envelope protein E1; envelope protein E2; fitness landscape; evolutionary model; statistical inference; inter-protein interactions; broadly neutralizing antibodies

1. Introduction

Hepatitis C virus (HCV), a single-stranded RNA virus, is the major cause of liver-associated disease and liver cancer. Currently, an estimated 58 million people are chronically infected with HCV (World Health Organization 2022). Although direct-acting antivirals (DAAs) have been developed and offer promising treatments for chronic HCV infections, their high cost and low rates of HCV diagnosis limit their accessibility to a subset of infected individuals only (Rosenthal and Graham 2016; World Health Organization 2022). Additionally, the efficacy of DAAs is limited by their inability to prevent reinfection and the emergence of drug-resistant viral strains (Wyles and Luetkemeyer 2017; Rossi et al. 2018). Therefore, developing an effective vaccine is crucial for eradication of HCV.

HCV encodes a single polyprotein, which is further cleaved by cellular and viral proteases into three structural proteins (core, E1, and E2) and seven non-structural proteins (NS1, NS2, NS3, NS4A, NS4B, NS5A, and NS5B). The envelope protein E2 is vital for viral entry into liver cells (hepatocytes), and it is a primary target for neutralizing antibodies (Deleersnyder et al. 1997). These antibodies bind to specific E2 regions, blocking the virus's ability to enter host cells and thus inhibiting viral replication and spread. A particular group of these antibodies, termed broadly neutralizing antibodies (Osburn et al. 2014), displays the capacity to neutralize a wide range of HCV genotypes and subtypes (Bankwitz et al. 2021).

Previous studies have indicated that E2 alone can generate a potent humoral immune response on its own and serve as a promising vaccine candidate (Cerino et al. 1997; Li et al. 2016; Yan et al. 2019). Nonetheless, the other envelope protein E1, which forms non-covalent heterodimers with E2 (Deleersnyder et al. 1997), exhibits a functional interdependence with E2 (Wahid et al. 2013; Douam et al. 2014; Li and Modis 2014; Haddad et al. 2017; Moustafa et al. 2018; Tong et al. 2018). For instance, E1 helps E2 maintain its functional conformation and regulates E2's interaction with HCV receptors CD81 and SR-B1, and both E1 and E2 are needed for interaction with Claudin-1, a key factor in HCV entry. E1 has also been shown to modulate the folding of E2 (Cocquerel et al. 2000; Brazzoli et al. 2005). While preliminary experiments suggest that specific mutations in E1 and E2 may jointly modulate viral infectivity (Douam et al. 2014), a comprehensive analysis of the role of E1E2 inter-protein interactions in mediating viral fitness is still lacking. Moreover, fitness of HCV is closely related to its ability to escape from antibody responses (Keck et al. 2011; Vela´zquez-Moctezuma et al. 2021). Therefore, investigating the effect of E1

on escape from E2-specific neutralizing antibodies is of particular interest.

In this work, we develop a computational fitness landscape model that considers interactions between E1 and E2 proteins. A detailed study of this model is performed using the available in vitro infectivity measurements to assess the role of E1E2 inter-protein interactions in mediating viral fitness. We further integrate this model into an in-host evolutionary model and investigate whether E1 may facilitate viral escape from antibodies targeting E2. Our analysis reveals potentially escape-resistant human monoclonal antibodies (HmAbs) against the E1E2 complex, offering directions for the development of an effective vaccine against HCV.

2. Results

2.1. Inference and statistical validation of the joint model for the E1E2 protein

We developed a computational model, termed joint model (JM), for the entire E1E2 protein using the sequence data available for subtype 1a. This model uses a maximum entropy approach to estimate the probability of observing a virus with a specific E1E2 protein sequence. In this model, the probability of any sequence $\mathbf{x} = [x_1, x_2, \ldots, x_N]$ is given by

$$P_{\mathbf{h},\mathbf{J}}(\mathbf{x}) = \frac{e^{-E(\mathbf{x})}}{Z}, \text{where } E(\mathbf{x}) = \sum_{N-1}^{i=1} \sum_{N}^{j=i+1} J_{ij}(x_i, x_j) + \sum_{N}^{i=1} h_i(x_i), \quad (1)$$

where N is the length of the sequence and is a normalization factor, which ensures that the probabilities sum to one. The fields $\mathbf{h}$ and couplings $\mathbf{J}$ represent the effect of mutations on a single residue and interactions between mutations at two different residues, respectively. $E(\mathbf{x})$ denotes the energy of sequence $\mathbf{x}$ (commonly referred to as Hamiltonian in statistical physics), which is inversely related to its prevalence. Inference of a maximum entropy model involves choosing the fields and couplings such that the model can reproduce the single and double mutant probabilities observed in the E1E2 sequence data.

We inferred the E1E2 maximum entropy model using the graphical user interface (GUI)-based software implementation of Minimum Probability Flow–Boltzmann Machine Learning (MPF–BML) (Quadeer et al. 2019) (see Section 5 for details), an efficient inference framework introduced in Louie et al. (2018). The single and double mutant probabilities obtained from the JM matched well with the E1E2 sequence data (Fig. 1A, B). Although not explicitly included in model inference, additional statistics including the connected correlations and the distribution of the number of mutations computed from the model also agreed well with those obtained from the E1E2 sequence data (Fig. 1C, D), demonstrating the predictive power of the inferred model. Overall, these results indicate that the inferred E1E2 JM captures well the statistics of the data.

2.2. E1E2 inter-protein interactions are important in mediating viral fitness

While some studies have considered E2 alone (i.e. independent of E1) (Cerino et al. 1997; Li et al. 2016; Yan et al. 2019), multiple studies have reported that these two proteins are functionally interdependent (Wahid et al. 2013; Douam et al. 2014; Haddad et al. 2017; Moustafa et al. 2018). This suggests that interactions between E1 and E2 may be critical. Previously, we had investigated E2 alone wherein we had inferred a fitness landscape model for E2 and used it to explore HCV escape dynamics from neutralizing antibodies (Quadeer, Louie, and Mckay 2019; Zhang, Quadeer, and McKay 2022). Here, to investigate the importance of E1E2 inter-protein interactions in virus fitness and immune escape, we compared the inferred JM with a model that considers E1 and E2 proteins to be independent (see Section 5 for details). We refer to it as the independent model (IM). In this model, the energy of an E1E2 protein sequence $\mathbf{x} = [\mathbf{x}_{E1}, \mathbf{x}_{E2}]$ is given by the sum of the energies of its E1 and E2 parts, $\mathbf{x}_{E1}$ and $\mathbf{x}_{E2}$, respectively,

$$E(\mathbf{x}) = E(\mathbf{x}_{E1}) + E(\mathbf{x}_{E2}). \quad (2)$$

Here, $E(\mathbf{x}_{E1})$ and $E(\mathbf{x}_{E2})$ are computed separately using inferred E1-only and E2-only maximum entropy models, respectively (see Section 5 for details). Both the E1-only and E2-only models capture well the statistics of the respective sequence data (Supplementary Fig. S1).

Equipped with the JM and IM, we first investigated whether E1 and E2 proteins can be considered statistically independent. This can be quantified by comparing the fraction of the correlated structure (FCS) of the E1E2 protein complex captured by the two models (Mora et al. 2010). FCS captured by a model can be estimated by comparing the entropy of synthetic sequences it generates with the entropy of a site-independent model and the estimated true entropy of the data, which omit and incorporate all correlations among sites, respectively (see Section 5 for details). If FCSs captured by both the JM and IM are similar, it will be suggestive of E1 and E2 to be independent. Based on our analysis, the average FCS of the E1E2 protein complex captured by the JM (63 per cent) was 22 per cent more than that captured by the IM (41 per cent) ($P = 9.1 \times 10^{-5}$; Fig. 2), suggesting that E1 and E2 proteins are not statistically independent. Thus, there seem to be significant inter-protein correlations that are not captured by the IM.

We next investigated if the additional correlations captured by the JM, compared to the IM, make it a better representative of the intrinsic E1E2 fitness landscape. Maximum entropy models have been shown previously to be good representatives of the underlying fitness landscapes for multiple individual viral proteins of HCV (polymerase (Hart and Ferguson 2015), NS3 (Zhang, Quadeer, and McKay 2023) and E2 (Quadeer, Louie, and Mckay 2019; Zhang, Quadeer, and McKay 2022)) and human immunodeficiency virus (HIV) (Ferguson et al. 2013; Mann et al. 2014; Barton et al. 2016; Flynn et al. 2017; Louie et al. 2018). To test this for the JM and the IM, we compared the predictions of both models using the in vitro infectivity measurements available for E1E2. We compiled a total of 156 in vitro infectivity measurements for E1E2 from sixteen studies (Goffard et al. 2005; Drummer et al. 2006; Ciczora et al. 2007; Falkowska et al. 2007; Gal-Tanamy et al. 2008; Rothwangl et al. 2008; Dowd et al. 2009; Keck et al. 2009; Guan et al. 2012; Keck et al. 2012; Urbanowicz et al. 2015; Pierce et al. 2016; El-Diwany et al. 2017; Gopal et al. 2017; Douam et al. 2014; Pfaff-Kilgore et al. 2022). We found that the JM provided a stronger negative Spearman correlation ($r = -0.70$; see Section 5 for details) between the predicted sequence energies (inversely related to prevalence) and experimental fitness values (Fig. 3) than the IM ($r = -0.54$; Fig. 3, inset). This result suggests that the JM is a better representative of the E1E2 fitness landscape. It also indicates the potential importance of E1E2 inter-protein interactions in mediating viral fitness.

2.3. Majority of strong E1E2 inter-protein interactions are compensatory

The couplings of the inferred maximum entropy model (J_{ij} in Equation (1)) are informative of the type of interactions between residues (Butler et al. 2016; Zhang et al. 2020). When the value of

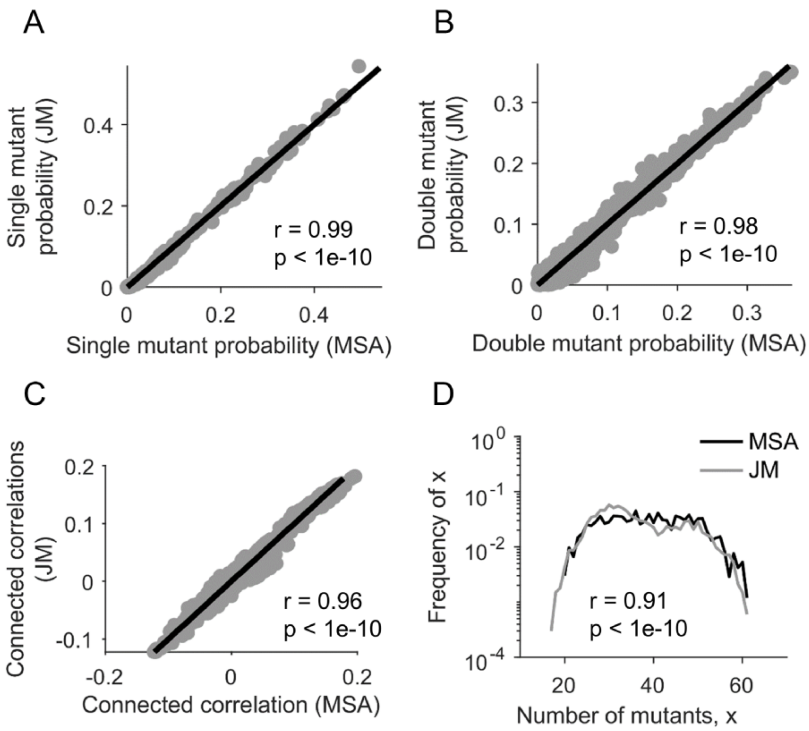

Figure 1. Statistical validation of the inferred E1E2 JM. Comparison of the (A) single mutant probabilities, (B) double mutant probabilities, (C) connected correlations, and (D) distribution of the number of mutants per sequence obtained from the MSA and those predicted by the inferred JM. Samples were generated from the inferred model using the MCMC method (Ferguson et al. 2013).

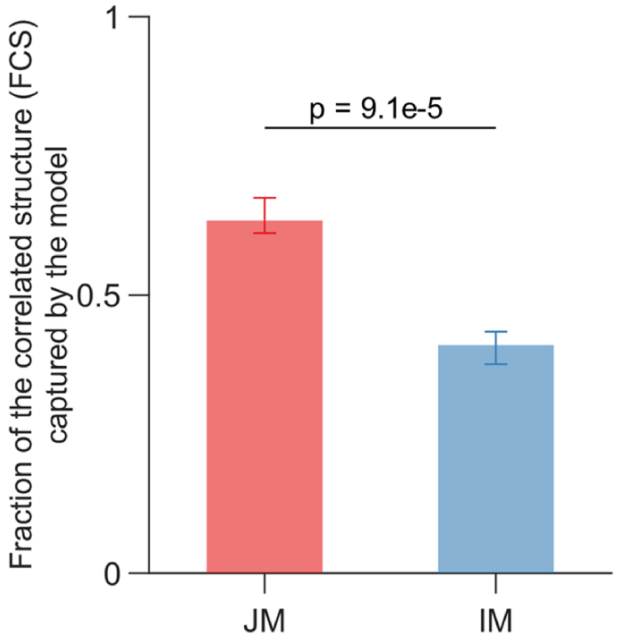

Figure 2. Comparison of the FCS in E1E2 protein captured by the JM and the IM. FCS captured by a model is quantified by I_{model}/I. Here, $I = S_{ind} - S_{true}$ is the multi-information which measures the overall strength of correlations in the system, where S_{ind} denotes the entropy of a site-independent model of E1E2 protein and S_{true} is the true entropy of the E1E2 complex estimated using the approach in Strong et al. (1998) (see Section 5 for details). Similar to I, $I_{model} = S_{ind} - S_{model}$ measures the strength of correlations captured by the JM or IM, where S_{model} is the entropy predicted by the JM or IM based on the data generated using the MCMC method (see Section 5 for details) (Mora et al. 2010). Entropies for the JM and IM were calculated over ten instances of MCMC runs, and the P-value was calculated using the one-sided Mann–Whitney test.

J_{ij} is large and positive, it signifies a strong antagonistic interaction or negative epistasis between residues i and j. This results in a decrease in the fitness of double mutants and makes it harder for new mutations to occur (Bank et al. 2014). On the other hand, when the value of J_{ij} is large and negative in Equation (1), it indicates a strong compensatory interaction or positive epistasis between residues i and j. This signifies improved viral entry or immune evasion capability of double mutants, allowing the virus to acquire diverse mutations.

Analyzing the top 300 pairs of inter-protein couplings (listed in Supplementary Data 1), i.e. with large absolute values of J_{ij}, we found that the majority (70 per cent) were negative (Fig. 4). This suggests that the top inter-protein couplings are largely compensatory and that simultaneous mutations in the two proteins may assist in maintaining a viable virus. This result was robust to the number of top inter-protein couplings considered (Supplementary Fig. S2). A recent study reported E1E2 as a highly fragile complex, with 92 per cent of alanine mutations introduced independently at each residue severely impacting viral infectivity (Pfaff-Kilgore et al. 2022). The strong compensatory interactions identified in our analysis indicate a potential mechanism by which E1 and E2, the most variable HCV proteins, may make multiple mutations while maintaining viral fitness.

We further quantified whether the strongly coupled residues (those associated with top 300 pairs of inter-protein couplings) were enriched in any known functional region of E1 and E2 proteins (see Section 5 for details). Our findings indicate that although no specific region of E1 exhibited statistically significant enrichment with the strongly coupled residues, there was a notable statistical enrichment of such residues within hypervariable Region 1 (HVR1) and hypervariable Region 2 (HVR2) of the E2 protein (Supplementary Table S1), thereby suggesting that these E2 regions may be involved in interactions with E1. This is also consistent

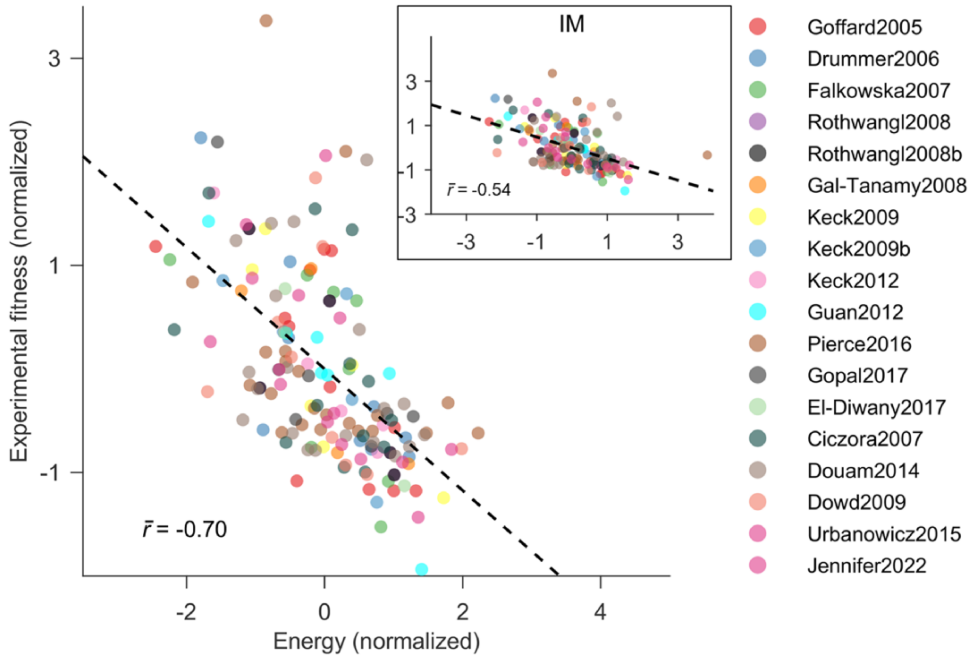

Figure 3. Comparison of the E1E2 fitness prediction by the JM and IM. Normalized energies computed from the inferred JM correlate strongly with the experimental fitness measurements. Conversely, the inferred IM provided a much lower correlation (inset). The legend shows the references from which fitness/infectivity measurements were compiled (Goffard et al. 2005; Drummer et al. 2006; Ciczora et al. 2007; Falkowska et al. 2007; Gal-Tanamy et al. 2008; Rothwangl et al. 2008; Dowd et al. 2009; Keck et al. 2009; Guan et al. 2012; Keck et al. 2012; Douam et al. 2014; Urbanowicz et al. 2015; Pierce et al. 2016; El-Diwany et al. 2017; Gopal et al. 2017; Pfaff-Kilgore et al. 2022).

with the literature that has shown that HVR2 is essential for the formation of the E1E2 heterodimer (McCaffrey et al. 2011), and epistatic interactions exist between E1 and HVR1 of E2 (Vela´zquez-Moctezuma et al. 2019). As the name of these regions suggests, these two regions are highly variable and are known to modulate viral escape from neutralizing antibodies (Alhammad et al. 2015). Hence, the potential compensatory interactions between E1 and these two E2 regions may contribute to viral immune evasion. Structurally, these interacting residues, however, do not demonstrate spatial proximity (Supplementary Fig. S3), suggesting that such interactions may occur through allosteric mechanisms, or alternatively, they may come into contact within the native trimer structure of E1E2 (Falson et al. 2015).

2.4. Evolutionary simulations suggest that the E1 protein contributes to escape from E2-specific antibody responses

To gain a deeper understanding of the impact of E1 on viral escape dynamics from E2-specific antibody responses, we quantified and compared the average time it takes for E2 residues to escape with and without the influence of E1. To achieve this, we utilized an in-host evolutionary model that takes into account the stochastic dynamics of viral evolution within the host including virus–host interactions, virus–virus competition, and escape pathways that the virus may employ to evade immune pressure. Similar models have been used previously for simulating in-host viral evolution for HIV (Barton et al. 2016) and HCV (Quadeer, Louie, and Mckay 2019; Zhang, Quadeer, and McKay 2022). Here, we incorporated the inferred JM into a population genetics model, similar to the well-established Wright–Fisher model (Ewens 2004). By doing so, we were able to predict the average number of generations, referred to as 'escape time', for each E2 residue to escape selective pressure (see Section 5 for more details). To determine escape times of these residues without the influence of the E1 protein, we utilized the E2-only model developed in our previous work (Quadeer, Louie, and Mckay 2019).

Previously, the E2-only model has been shown to be capable of predicting known escape mutations from multiple E2-specific HmAbs (Kato et al. 1993; Keck et al. 2008; Keck et al. 2009; Morin et al. 2012; Bailey et al. 2015; Keck et al. 2016) (listed in Supplementary Table S2), where these mutations were shown to be associated with lower escape times compared to mutations at other residues (Quadeer, Louie, and Mckay 2019) as they enable the virus to evade the associated antibody pressure. We found that this was also true for the inferred JM ($P = 9.9 \times 10^{-24}$; Fig. 5A). We also analyzed escape times of mutations in buried and exposed residues within E1E2 structures (Section 5). Buried residues, located in the protein core, were expected to have higher escape times compared to exposed ones, and our analysis confirmed this prediction ($P < 10^{-10}$; Supplementary Fig. S4). These results suggest the JM to be capable of distinguishing E2 residues associated with low and high escape times.

We employed the JM to pinpoint E1E2 regions that appear most susceptible to antibody targeting. This was done by overlaying the mutation-associated escape times onto each E1E2 residue, represented as a heat map on the experimentally resolved (Torrents de la Peña et al. 2022) (partial) and AlphaFold-predicted (Mitchell et al. 2019; Mirdita et al. 2022) (complete) E1E2 structure (Fig. 5B). Regions representing CD81 binding sites, HVR1, and HVR2 are also depicted on the structures. An appreciable number of red-colored E1E2 residues on this map indicated a high incidence of surface mutations associated with high escape times. This finding suggests the feasibility of rationally engineering antibodies (Sormanni, Aprile, and Vendruscolo 2015) that specifically target these exposed, difficult-to-escape residues, potentially enabling a potent immune response.

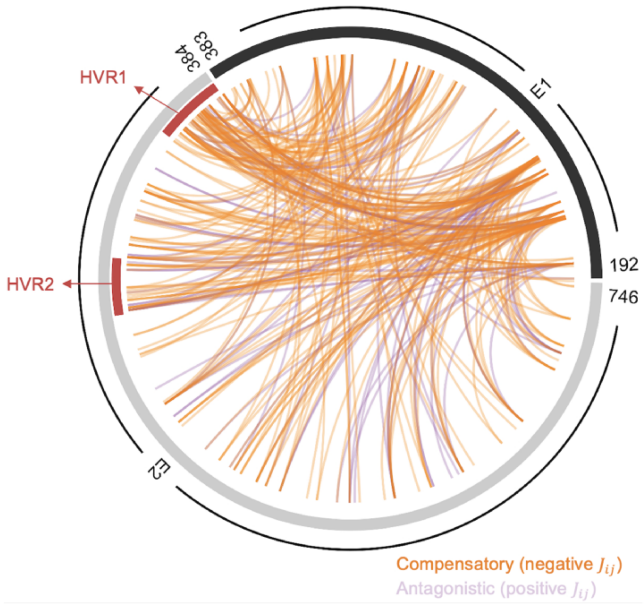

Figure 4. Strong E1E2 inter-protein interactions are largely compensatory. Each pair of mutations between E1 and E2 proteins was ranked by the absolute values of J_{ij} from Equation (1) and top 300 pairs are plotted here. Compensatory interactions (negative values of J_{ij}) and antagonistic interactions (positive values of J_{ij}) are highlighted in different colors. The outer segments of the circle represent E1 (shown in black, encompassing residues 192–383) and E2 (shown in gray, encompassing residues 384–746) proteins. HVR1 and HVR2, which we found to be statistically enriched with strongly coupled residues, are shown as red segments.

Further comparing the escape times of E2 residues inferred from these two models, we found that the escape times of residues associated with escape mutations inferred from the JM were marginally significantly lower ($P = 0.077$; Fig. 5C, left panel) than those from the E2-only model. In contrast, there was not much difference between the escape times of the remaining E2 residues ($P = 0.525$; Fig. 5C, right panel). This suggests that the E1 protein may assist in viral escape from E2-specific antibodies. In addition, we found that the strongly coupled inter-protein residues (Fig. 4) were statistically significantly enriched in escape mutations ($P = 5.4 \times 10^{-19}$; Fig. 5D). This further corroborates the potential role of E1 in mediating viral escape from neutralizing antibodies.

2.5. For multiple E2-specific HmAbs, E1 is predicted to provide accelerated escape dynamics

Previously, we utilized the E2-only model to assess the efficacy of each known E2-specific HmAb based on the minimum escape time predicted for its binding residues (Quadeer, Louie, and Mckay 2019; Zhang, Quadeer, and McKay 2022) (see Section 5 for details). Our aforementioned analysis suggests that E1 may potentially assist E2 in antibody evasion, and hence, we further studied how this would impact the efficacy of known HmAbs predicted by the JM in comparison to the E2-only model. We first employed a binary classifier (Quadeer, Louie, and Mckay 2019) to determine an optimal cut-off value ($\zeta = 96$ generations) for identifying escape-resistant residues based on the JM. This binary classifier utilized known escape mutations (listed in Supplementary Table S2) as true positives and the remaining E1E2 residues as true negatives, as detailed in Section 5. We subsequently evaluated each antibody by comparing the minimum escape time predicted for its binding residues with the corresponding optimal cut-off value ζ for each model. For this analysis, we focused on thirty-two HmAbs for which binding residues have been determined using global alanine scanning experiments (Pierce et al. 2016; Gopal et al. 2017; Keck et al. 2019).

Based on our previous predictions using the E2-only model, we had identified twenty-one E2-specific HmAbs that appear relatively easy for the virus to escape. These predictions were also consistent with the JM (Fig. 6). Among these HmAbs, studies have shown that AR1A, AR1B, AR2A, CBH-4B, CBH-4D, CBH-4G, CBH-20, CBH-21, and CBH-22 were non-neutralizing or isolate-specific (Keck et al. 2005; Law et al. 2008; Kong et al. 2016), which further supports our predictions for both models. The remaining eight E2-specific HmAbs (212.15, 212.25, CBH-7, CBH-23, HC-1, HC33-1, HC84-20 and HCV1) were predicted to be escape resistant by the E2-only model. However, only four (212.15 and 212.25, HC33-1, and HCV1) among these were predicted to be escape resistant by the JM (Fig. 6). The predictions of the JM for these HmAbs align well with literature reports. For instance, HmAbs 212.25 and 212.15, isolated from patients who had spontaneously cleared HCV, were found to be cross-neutralizing (Keck et al. 2019). HC33-1 and HCV1 have also been reported as potentially escape-resistant broadly neutralizing antibodies in multiple studies (Broering et al. 2009; Kong et al. 2012; Keck et al. 2014; Pierce et al. 2016). On the other hand, of the four HmAbs (HC84-20, CBH-23, HC-1, and CBH-7) predicted to be escape resistant by the E2-only model but not by the JM, studies have observed escape for strains isolated from patients who underwent liver transplantation for HmAbs CBH-23 and HC-1, while HmAb CBH-7 was obtained from a patient with chronic HCV infection (Fofana et al. 2012; Keck et al. 2019). These findings suggest that E1 may play a role in facilitating HCV escape from these antibodies. Mapping the epitopes of antibodies on the E1E2 structure suggests the possibility of allosteric interactions between E1 and E2 residues contributing to escape (Supplementary Fig. S5). Notably, the JM enabled identification of one HmAb, IGH526, that targets the E1 protein and may be escape resistant. Multiple studies have reported that IGH526 is cross-neutralizing and can target various HCV isolates from different genotypes (Meunier et al. 2008; Kong et al. 2015).

3. Discussion

E1 and E2 are envelope proteins of HCV that form non-covalent heterodimers. While E2 is the major target of HmAbs and a promising vaccine candidate, E1 is also important for HCV entry and assembly, and it interacts with E2. Comparing a JM that takes into account E1E2 interactions with an IM that does not, we have determined that these interactions are important in mediating virus infectivity and immune escape. The top E1E2 inter-protein interactions are compensatory and enriched in HVR1 and HVR2 of E2. Further using in-host evolutionary modeling, our analysis suggests that E1 may facilitate HCV in escaping E2-specific antibody responses. We have identified potentially escape-resistant HmAbs against the E1E2 complex, which could aid in the development of a robust prophylactic vaccine against HCV.

By comparing the correlation between in vitro infectivity measurements and predictions of the JM and the IM (Fig. 3), our study highlighted the importance of E1E2 inter-protein interactions in mediating viral fitness. This was further reinforced by comparing the predictions of the JM with those of a site-independent E1E2 fitness landscape model (see Section 5 for details), which showed that the correlation between the JM predictions and in vitro fitness measurements was much higher than that of the site-independent model ($r = -0.54$; Supplementary Fig. S6). These findings are consistent with previous studies that have

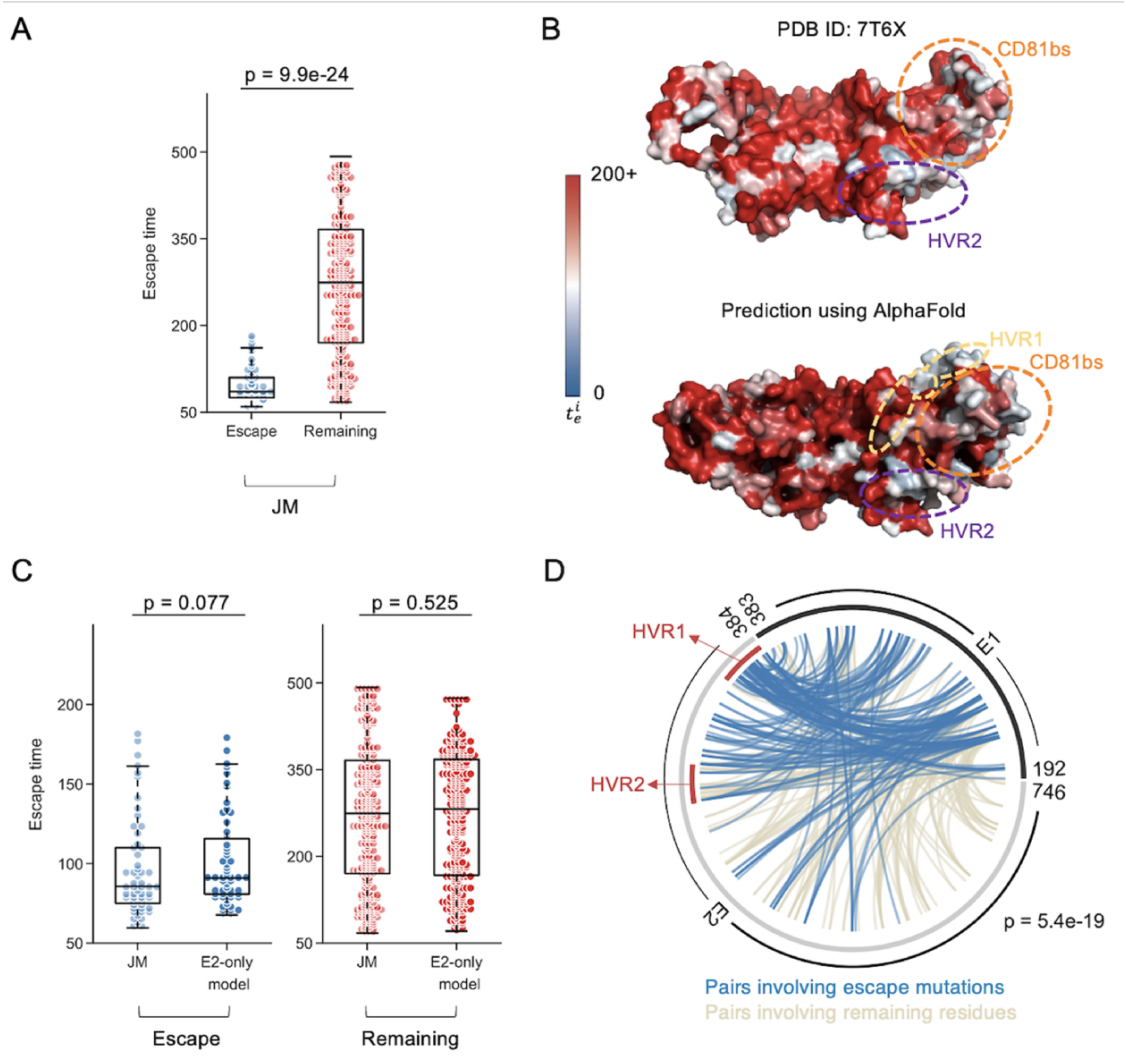

Figure 5. Role of E1 in facilitating viral escape from E2-specific HmAbs. (A) Distribution of escape times of E2 residues using the inferred JM. Residues were divided into two categories: those with known escape mutations from E2-specific HmAbs (listed in Supplementary Table S2) and the remaining E2 residues. *P*-value was calculated using the one-sided Mann–Whitney test, as residues with known escape mutations from E2-specific HmAbs would be expected to have lower escape times compared to the remaining residues. (B) Superimposing relative escape times on the E1E2 crystal structures (top panel: PDB ID: 7T6X (Torrents de la Peña et al. 2022); bottom panel: structure predicted by AlphaFold (Mitchell et al. 2019; Mirdita et al. 2022)). Residues incurring high escape times upon mutation (large values of t^i) are shown in red and those associated with low escape times (small values of t^i)) are shown in blue. Locations of HVR1-, HVR2-, and CD81-binding sites (Pierce et al. 2016) are shown by dashed lines on each crystal structure. Note that HVR1 is not included in the experimentally resolved structure (PDB ID: 7T6X). (C) Comparison of escape times of E2 residues inferred from the JM and the E2-only model for the known E2 escape mutations (left panel) and the remaining E2 residues (right panel). *P*-values were calculated using the one-sided Mann–Whitney test. This choice was motivated by the expectation that E1 residues, which potentially aid the escape of E2 residues, would lead to lower escape times for the latter in the JM. (D) Circos plot distinguishing the interactions between strongly coupled residues (Fig. 4) involving escape mutations and the remaining residues. The reported *P*-value measures the probability of observing by a random chance at least the observed number of E2 escape mutations among strongly coupled residues (see Section 5 for details).

emphasized the importance of considering interactions when inferring protein fitness landscapes (Ferguson et al. 2013; Hart and Ferguson 2015; Barton et al. 2016; Flynn et al. 2017; Louie et al. 2018; Quadeer, Louie, and Mckay 2019; Quadeer et al. 2020; Sohail et al. 2021; Zhang, Quadeer, and McKay 2022; Zhang, Quadeer, and McKay 2023) and for identifying networks of residues that play crucial roles in the protein structure and function of viruses (Dahirel et al. 2011; Quadeer et al. 2014; Quadeer, Morales-Jimenez, and McKay 2018; Ahmed et al. 2019; Gaiha et al. 2019).

A recent experimental study has shown that E1E2 is a fragile protein complex wherein even a single alanine mutation at 92 per cent of positions severely impacts the infectivity of the virus (Pfaff-Kilgore et al. 2022). Therefore, our finding that 70 per cent of the top 300 pairs of mutations (ranked by absolute values of J_{ij}) between E1 and E2 are compensatory suggests that these interactions may play a significant role in mediating viral fitness. To further investigate this experimentally, it would be helpful to conduct assays that quantify the change in replicative fitness by site-directed mutagenesis of the pairs of mutations identified to be associated with strong compensatory interactions (e.g. top 10) individually and simultaneously (Supplementary Data 1).

Comparing the JM and the E2-only model, we found ten residues that were predicted to be escape resistant by the E2-only model but easy to escape according to the JM. Interestingly, four of these (residues 424, 437, 537, and 538) are known antibody

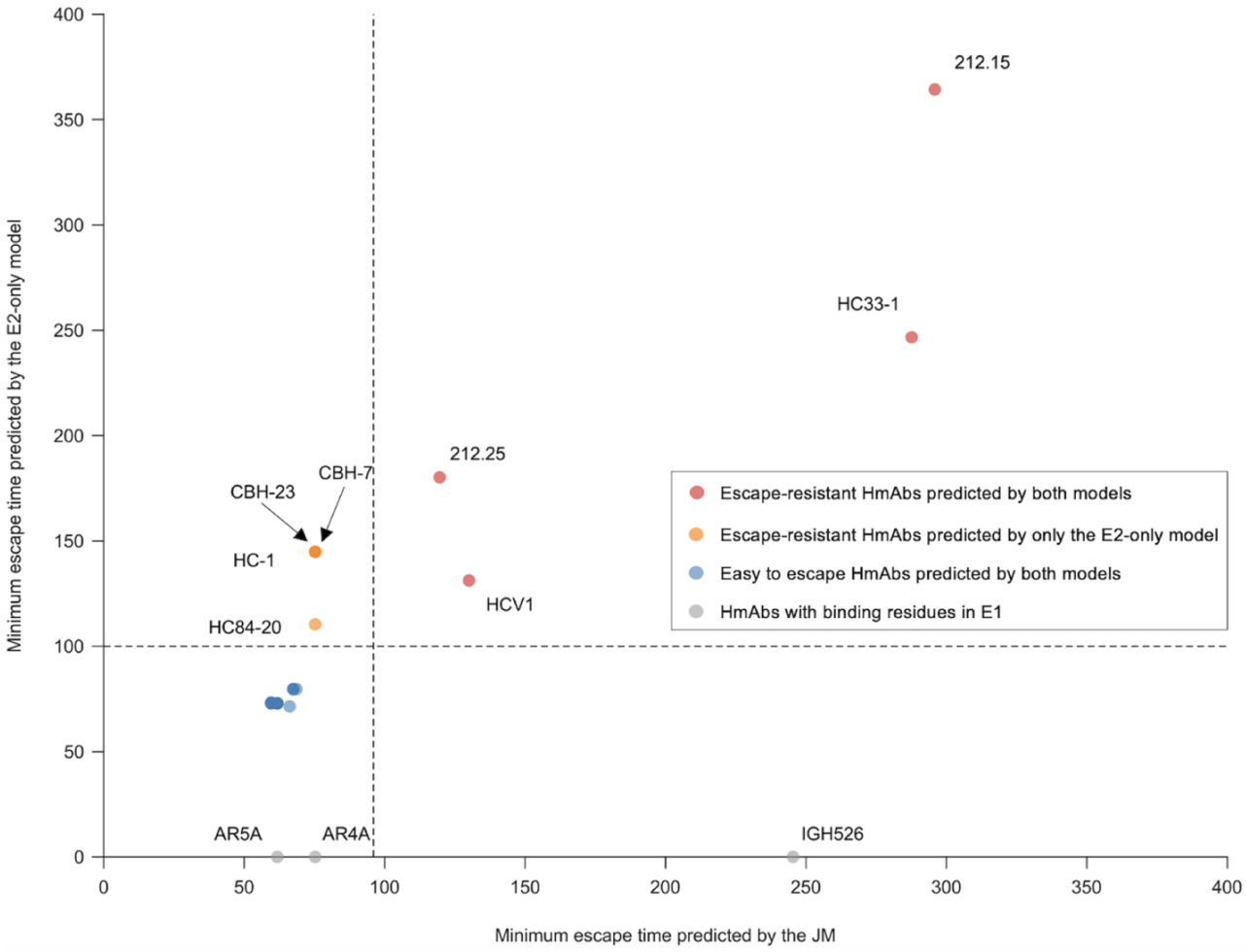

Figure 6. Evaluation of known HmAbs using the escape times inferred from the JM and the E2-only model. For each HmAb, escape time associated with all binding residues was predicted using both models. Each circle in the figure represents the minimum escape time associated with the binding residues of each HmAb predicted by the JM (x-axis) and the E2-only model (y-axis). Global alanine scanning mutagenesis (Pierce et al. 2016; Gopal et al. 2017; Keck et al. 2019) was used to determine the binding residues of each HmAb, where each residue of the wild-type sequence was replaced by alanine (or glycine/serine if the residue in the wild-type was alanine). We defined binding residues of each of these HmAbs as residues with relative binding (the fraction of the mutant sequence's binding compared to the wild-type sequence) less than or equal to 20 per cent. The twenty one HmAbs predicted to be easy to escape by both models are 212.1.1, 212.10, A27, AR1A, AR1B, AR2A, AR3A, AR3B, AR3C, AR3D, CBH-4B, CBH-4D, CBH-4G, CBH-5, CBH-20, CBH-21, CBH-22, HC33-4, HC-11, HC84-24, and HC84-26. The HmAbs having binding residues in E1 are plotted along the x-axis, since the E2-only model could not be used to predict their escape time. The dashed line denotes the optimal cut-off value ζ for each model (see Section 5 for details).

binding residues, which suggests that the E1 protein may interact with these residues during antibody evasion. This motivates experimental studies for investigating the interactions between these four residues in the E2 and the E1 protein. One approach could involve longitudinal experiments (Alhammad et al. 2015), where the virus is allowed to infect cells in the presence of antibodies that specifically target these four residues, and changes in these residues as well as the E1 protein are monitored over time. By doing so, it could be determined if mutations arise at these residues in response to antibody pressure and if simultaneous mutations are also observed in the E1 protein. This would provide important insights into the mechanisms by which the virus evolves to evade immune responses (Frumento, Flyak, and Bailey 2021), which could ultimately inform the design of an effective vaccine against HCV.

By applying the JM and the E2-only model to evaluate the efficacy of known HmAbs, we identified twenty-five HmAbs with consistent predictions for both models (Fig. 6). Among these, four HmAbs were predicted to be escape resistant, while the other twenty-one HmAbs were not. This motivates investigating the differences in escape dynamics (Augestad et al. 2020) between these two sets of HmAbs. For instance, experimentally quantifying the average time (number of generations) it takes for the virus to escape from HmAbs 212.15, 212.25, HC33-1, or HCV1 (escape-resistant HmAbs) in comparison to HmAbs AR3A, AR3C, or AR3D (non-escape-resistant HmAbs) would be a helpful follow-up study.

Four HmAbs (HC-1, CBH-23, CBH-7, and HC84-20) were associated with different predictions based on the JM compared to the E2-only model (Fig. 6). We found that the different predictions for these HmAbs were due to the differences in escape times of two specific binding residues 437 and 537 by these two models, which are shared by these HmAbs. Intriguingly, these two residues are also CD81-binding residues (Ströh, Nagarathinam, and Krey 2018). Experiments to study the interactions between E1- and CD81-binding residues may be beneficial for discovering their

potential roles in compensating viral infectivity or mediating viral entry.

4. Limitations of the study

Our study has several limitations. First, our investigation relied on a computational model (JM) to reveal that E1 and E2 interactions are predominantly compensatory. However, the interactions identified by our model mostly occurred within regions known for their high variability, namely, HVR1 and HVR2. This observation could be attributed to the inherent variability of these regions, making them more prone to detection by our model. Therefore, conducting additional experimental studies involving the mutation of residues within these regions, as well as the E1 residues predicted to interact with them, would be valuable for validating this hypothesis. Second, our in-host evolutionary modeling suggests a role for E1 in facilitating viral escape from E2-specific antibodies. However, the statistical significance of these results was only marginal (Fig. 5C). Therefore, conducting experiments to more thoroughly investigate the impact of E1 on viral escape from E2-specific antibodies would be beneficial in order to obtain more conclusive evidence. Third, we focused solely on studying E1E2 interactions within the context of HCV Subtype 1a. To assess the generalizability of our findings, it would be valuable to perform similar analyses using data from other HCV genotypes or subtypes. However, the limited availability of E1E2 sequence data for these other subtypes poses a challenge for accurate model inference. We recognize the significance of exploring this aspect further in future studies.

5. Methods

5.1. Inference of computational models for the E1E2 protein

To explore the role of E1E2 inter-protein interactions, we considered two types of computational models for the E1E2 protein: One takes into account the E1E2 inter-protein interactions, named the JM, and the other without the E1E2 inter-protein interactions, named the IM.

5.1.1. *JM*

To infer a maximum entropy (least-biased) model for the whole E1E2 protein jointly, we downloaded 8,021 aligned E1 Subtype 1a and 6,225 aligned E2 Subtype 1a sequences from the HCV-Genes Linked by Underlying Evolution (GLUE) database (http://hcv.glue.cvr.ac.uk) (Singer et al. 2018, 2019), both with genome coverage of ≥99 per cent. We constructed the multiple seuqence alignment (MSA) of the whole E1E2 protein by stitching together E1 and E2 sequences based on the information in their headers, yielding 6,198 E1E2 sequences. We conducted a principal component analysis on the pair-wise similarity matrix (6198×6198) of the sequences (Strimmer and Haeseler 2009), where the (i, j)th entry of the similarity matrix represents the fraction of residues that are identical in sequences i and j, to remove any outlier sequences. We considered a sequence as an outlier if its corresponding value in the first principal component (PC) was more than three scaled median absolute deviations (Leys et al. 2013) from the median of the first PC. We also excluded 264 sequences for which patients' information was not available. After these filtering procedures, we had $M = 5{,}867$ sequences from $W = 871$ patients. Moreover, we excluded twenty-one fully conserved E1E2 residues to improve the quality of the residues. Hence, the processed MSA was composed of $M = 5{,}867$ sequences (listed in Supplementary Data 2) and $N = 534$ residues. We constructed a least-biased maximum entropy model for the E1E2 protein that can reproduce the single and double mutant probabilities of this processed MSA (Equation (1)).

To infer parameters ($\mathbf{h}$ and $\mathbf{J}$) of the maximum entropy model, we used the GUI realization of MPF-BML (Quadeer et al. 2019), an efficient inference framework introduced in Louie et al. (2018). This software requires an MSA as input and a vector comprising the patient weight of each sequence included in the MSA. Patient weight is computed as the inverse of the number of sequences associated with each patient. The MPF-BML parameters used for inferring the model parameters (fields $\mathbf{h}$ and couplings $\mathbf{J}$) are as follows: (1) L_1 regularization parameters were set to 5×10^{-4} for both fields and couplings; (2) L_2 regularization parameters were set to 0.05 for fields and 125 for couplings; and (3) all other parameters were set to their default values. The first- and second-order statistics of the inferred JM matched well with those of the MSA (Fig. 1).

5.1.2. *IM*

The IM comprised two maximum entropy models, one for the E1 protein and the other for the E2 protein. The maximum entropy models for E1 protein and E2 protein were inferred using the E1 part and the E2 part of the E1E2 processed MSA, respectively. Specifically, the MSA of both E1 and E2 consisted of $M = 5{,}867$ sequences from $W = 871$ patients, where each sequence contains $N = 187$ residues (five fully conserved ones were excluded) for E1 and $N = 347$ residues (sixteen fully conserved ones were excluded) for E2. The MSA and the patient weights were further set as the input of the MPF-BML software using the same parameters as the JM except that both L_1 and L_2 regularization parameters were set to 50 for couplings for E1 and 15 for E2 and 5×10^{-4} for fields for both E1 and E2. Both the statistics of the inferred E1-only model and the E2-only model lined up well with those of the respective MSAs (Supplementary Fig. S1). The final IM was a linear combination of these two models, where the energy of a full E1E2 sequence $\mathbf{x} = [\mathbf{x}_{E1}, \mathbf{x}_{E2}]$ is given by

$$E(\mathbf{x}) = E(\mathbf{x}_{E1}) + E(\mathbf{x}_{E2}). \tag{3}$$

Here, $E(\mathbf{x}_{E1})$ and $E(\mathbf{x}_{E2})$ represent the energy of the E1 part $\mathbf{x}_{E1}$ and E2 part $\mathbf{x}_{E2}$ of sequence $\mathbf{x}$ calculated from each E1-only or E2-only model according to Equation (1).

As we had inferred a maximum entropy E2-only model in a previous study (Quadeer, Louie, and Mckay 2019), we further investigated if our previous E2-only model (inferred from 3,363 sequences of E2 available at that time) was capable of capturing the statistical variations in the E2 MSA we curated in this study (5,867 sequences). Our results support that this is indeed the case (Supplementary Fig. S7), suggesting that both these E2-only models are equally representative of the variations in the E2 protein sequence data. In addition, the correlation of both models with in vitro infectivity measurements was also similar, suggesting that both E2-only models are also equally good representatives of the E2 fitness landscape (Supplementary Fig. S8).

5.2. Calculation of the FCS captured by each model

FCS of the E1E2 protein complex captured by a model is given by I_{model}/I. Here, I reflects the overall strength of correlations in the protein complex (Mora et al. 2010), quantified by the difference between the site-independent model entropy (S_{ind}) and the true entropy of the protein complex (S_{true}), which disregard and include all correlations among sites, respectively. In contrast, I_{model} represents the strength of correlations captured by a model, calculated

by the difference between the site-independent model entropy (S_{ind}) and the inferred model entropy (S_{model}). Below, we describe how we calculated these different entropies.

S_{ind}, the entropy of site-independent model, was computed by considering amino acids at each E1E2 residue independently with the observed frequencies, which is given by

$$S_{ind} = \sum_{a\in\Omega}\sum_{i=1}^{N} f_i(a)\ln f_i(a), \tag{4}$$

where $\Omega = \{A, R, \ldots, V, -\}$ (the 20 amino acids and the gap).

S_{true} was estimated using the procedure described in Mora et al. (2010) and Strong et al. (1998) that involves incrementally subsampling the data and measuring its entropy. Specifically, we first randomly chose M sequences and calculated the 'naive estimate' of the entropy $S_{naive}(M)$ through

$$S_{naive}(M) = \sum_{\mathbf{x}\in M\text{sequences}} -f(\mathbf{x})\ln f(\mathbf{x}), \quad M = 500, 1000, \ldots \tag{5}$$

where $f(\mathbf{x})$ is the frequency of sequence $\mathbf{x}$. We repeated this procedure 100 times with different random seeds for M sequences ($M = 500,\ 1000, \ldots$) and took the mean of $S_{naive}(M)$, denoted by $\langle S_{naive}(M)\rangle$, over these iterations for each given M. As shown in Strong et al. (1998), the naive estimate of the entropy can be well fit by

$$\langle S_{naive}(M)\rangle = S_{true} + \frac{S_1}{M} + \frac{S_2}{M^2}, \tag{6}$$

where S_1 and S_2 are constants that depend on the distribution of the data. They account for the bias and variance that arise due to finite sample size effects. When $M \to +\infty$, these correction terms vanish and the naive estimate converges to the true entropy S_{true}. By plotting $\langle S_{naive}(M)\rangle$ against $\frac{1}{M}$, we can observe the quadratic relationship between the two variables (Supplementary Fig. S9). Extrapolating the y-intercept (when $\frac{1}{M} \to 0$) from this plot provides an estimate for S_{true}.

We calculated S_{model}, the entropy predicted by the inferred models, using sequence ensemble generated by a Markov Chain Monte Carlo (MCMC) procedure (Mora et al. 2010). For the JM, the sequence ensemble comprised 99,990 full E1E2 sequences, and the model entropy was calculated as $S_{JM} = -\sum_{\mathbf{x}} f(\mathbf{x})\ln f(\mathbf{x})$. For the IM, a sequence ensemble of 99,990 sequences was generated for each of the E1 and E2 proteins separately using their respective individual models. The entropy for IM was calculated as $S_{IM} = -\sum_{\mathbf{x}_{E_1}} f(\mathbf{x}_{E_1})\ln f(\mathbf{x}_{E_1}) - \sum_{\mathbf{x}_{E_2}} f(\mathbf{x}_{E_2})\ln f(\mathbf{x}_{E_2})$, where $\mathbf{x}_{E1}$ and $\mathbf{x}_{E2}$ are sequences from the E1 and E2 sequence ensemble, respectively. Entropies were calculated over ten instances of MCMC runs for both the JM and the IM. All entropies calculated earlier are shown in Supplementary Fig. S10.

5.3. Fitness verification

We used in vitro experimental infectivity measurements compiled from the literature (Goffard et al. 2005; Drummer et al. 2006; Ciczora et al. 2007; Falkowska et al. 2007; Gal-Tanamy et al. 2008; Dowd et al. 2009; Rothwangl et al. 2008; Keck et al. 2009; Keck et al. 2012; Douam et al. 2014; Urbanowicz et al. 2015; Pierce et al. 2016; Pfaff-Kilgore et al. 2022) to investigate if our inferred models for E1E2 (JM and IM) are capable of capturing the infectivity of the virus. The details of the specific fitness measurements (listed in Supplementary Data 3) from each study are presented in Supplementary Table S3. These in vitro infectivity measurements involve assessing the ability of HCV to infect and replicate within cultured cells, and energy is inversely related to prevalence according to our model. Thus, if we observe a negative correlation (as shown in Fig. 3) between model-predicted sequence energies and in vitro infectivity measurements of these sequences, it provides evidence that our inferred prevalence landscape is a reasonably good proxy for the fitness landscape. As experiments were conducted under different laboratory settings, we considered the weighted average of Spearman correlation coefficients from different experiments. This can be written as

$$\bar{r} = \frac{\sum_{i=1}^{q_{exp}} Q_i r_i}{\sum_{i=1}^{q_{exp}} Q_i},$$

where r_i is the Spearman correlation coefficient obtained from experiment i and Q_i is the number of measurements. q_{exp} is the total number of experiments.

5.4. Identification of strongly coupled residues in the E1 and E2 proteins

To identify strongly coupled pairs of mutations (top inter-protein couplings) between the E1 and E2 proteins, we constructed 'null models' to determine a threshold (Quadeer et al. 2020; Barton, Kardar, and Chakraborty 2015). Specifically, to maintain the observed single mutant probabilities but to break any pair-wise correlations in the E1E2 sequence data, we first constructed a 'null MSA' by choosing amino acids at each residue with the observed frequencies while keeping the same number of sequences ($M = 5{,}867$) and number of residues ($N = 534$). We then used the 'null MSA' to infer a maximum entropy model, i.e. a null model. This procedure was repeated ten times, and the threshold was set as the top 0.1 percentile of the absolute mean value of J_{ij} of these ten null models, which corresponds to roughly choosing about top 300 pairs of inter-protein couplings in the JM. The residues that are present in these 300 inter-protein couplings are referred to as 'strongly coupled residues' throughout the manuscript.

5.5. Statistical significance testing

We calculated the statistical significance of the number of strongly coupled residues (identified by our model) in each functional region of E1 and E2 proteins, as well as in known escape mutations (listed in Supplementary Table S2), using a P-value. For a given set of residues in a protein region, this P-value corresponds to the probability of observing at least i residues out of j strongly coupled residues in that region, where there are n total residues in that protein region out of N total residues of a protein (187 for E1 and 347 for E2). These can be written as

$$p = \sum_{q=i}^{min(j,n)} \frac{\binom{j}{q}\binom{N-j}{n-q}}{\binom{N}{n}}. \tag{7}$$

A P-value less than 0.05 for a protein region indicates statistically significant enrichment of residues of that region within the strongly coupled E1E2 inter-protein interactions.

5.6. Visualization of interactions between strongly coupled pairs of mutations

To visualize the interactions between strongly coupled pairs of mutations, we utilized a Circos plot. The E1E2 residues were evenly distributed along the outer edge of the circles in Figs. 4 and 5D. The numbering of the residues was started at 192 (corresponding to the first residue of E1 according to the H77 sequence) at the 3 o'clock position and progressed in a counter-clockwise direction.

Each link within the circle represents a pair of strongly coupled mutations (ranked by the absolute values of J_{ij} from Equation (1)).

5.7. Prediction of complete E1E2 structures using AlphaFold

The experimentally resolved crystal structure of E1E2 has been recently published (Torrents de la Peña et al. 2022) and is available at Protein Data Bank (PDB) with ID: 7T6X. However, this structure is not complete and encompasses only E1 residues 192–256 and 294–346 and E2 residues 420–717. Thus, we predicted the complete E1E2 protein structure by AlphaFold (Mitchell et al. 2019; Mirdita et al. 2022) utilizing the same E1E2 sequence as 7T6X. We also investigated if the structure predicted by AlphaFold is accurate. Based on the resolved structure (PDB ID: 7T6X), we observed eighty-three pairs of residues (thirty E1 residues and thirty-six E2 residues) that exhibit contact (<8 Angstroms distance between carbon-alpha atoms) between E1 and E2 proteins. For the structure predicted by AlphaFold, we identified 150 pairs of residues (fifty-sixE1 residues and fifty-one E2 residues) predicted to be in contact between E1 and E2 proteins. Notably, seventy-one pairs (twenty-six E1 residues and thirty-one E2 residues) were found to be common between the two structures, indicating a reasonable prediction of the E1E2 structure by AlphaFold.

5.8. Evolutionary simulation

To quantify the average time it takes for each residue in E2 to escape with the effect of the E1 protein, we considered a viral intra-host population genetics evolutionary model incorporated with the inferred JM similar to that in Quadeer, Louie, and Mckay (2019). We used the 'escape time' metric to represent the number of generations it takes on average for the virus with a mutation at a given residue to reach majority (frequency > 0.5) in a fixed-sized viral population under targeted immune pressure.

To be specific, we used a well-established Wright–Fisher model (Ewens 2004), where in each generation, the virus population undergoes mutation, selection, and random sampling steps. The virus population size was fixed at $M_e = 2000$, in line with the effective HCV population size in in-host evolution (Bull et al. 2011). For each residue i of the E2 protein, we formed the initial viral population with duplicates of a sequence with the consensus amino acid at residue i. In the mutation step, the nucleotide of each sequence was mutated randomly to another nucleotide at a fixed rate of $\mu = 10^{-4}$, consistent with the known HCV mutation rate (Cuevas et al. 2009; Sanjuan et al. 2010). In the selection step, each sequence was selected based on its fitness predicted from the inferred JM. Specifically, we calculated the survival probability of a virus with sequence $\mathbf{x}$ by

$$f_{\mathbf{h},\mathbf{J}}(\mathbf{x}) = \frac{g_{\mathbf{h},\mathbf{J}}(\mathbf{x})}{\sum_{\mathbf{y}} g_{\mathbf{h},\mathbf{J}}(\mathbf{y})}, \tag{8}$$

where $g_{\mathbf{h},\mathbf{J}}(\mathbf{x})$ is a function that maps the predicted energy of sequence $\mathbf{x}$ smoothly to a value between 0 and 1. This function is defined as

$$g_{\mathbf{h},\mathbf{J}}(\mathbf{x}) = \frac{e^{\beta(\bar{E}-E_{\mathbf{h},\mathbf{J}}(\mathbf{x}))}}{1+e^{\beta(\bar{E}-E_{\mathbf{h},\mathbf{J}}(\mathbf{x}))}}, \tag{9}$$

where $\bar{E}$ is the average energy of the current sequence population, while $\beta \sim 0.1$ was chosen based on the slope between predicted sequence energies and in vitro infectivity measurements (Quadeer, Louie, and Mckay 2019). To model the immune pressure at residue i, the fitness of all sequences having the consensus amino acid at residue i was decreased by a fixed value b, thereby providing a selective advantage to the sequences having a mutation at this residue. The value of b was set according to the largest value of the field parameter in the inferred landscape. Next, the subsequent generation of virus population was generated through a standard multinomial sampling process parameterized by M_e and $f_{\mathbf{h},\mathbf{J}}(\mathbf{x})$. This procedure was continued until the mutations at residue i reached a frequency of >0.5. The number of generations at this iteration was recorded. This process was repeated 100 times with the same initial sequence and twenty-five distinct initial sequences as well. The final escape time t^i of residue i was the mean number of generations over all these runs of simulation.

To perform a fair comparison between the escape times predicted by the JM and those by the E2-only model in Quadeer, Louie, and Mckay (2019), we set the same simulation parameters for both models, including the fitness penalty factor b (10), the number of generations (500), the number of distinct sequences forming the initial population (25) for each residue and the number of runs of simulation (100) for each distinct initial sequence. The mean escape time predicted for each residue by the JM and the E2-only model is provided in Supplementary Data 4.

5.8.1. *Identification of escape-resistant residues*

We ran the evolutionary simulation using the JM for all E1E2 residues following the same procedure described earlier. We employed a binary classifier that utilized known escape mutations (listed in Supplementary Table S2) as true positives and all other residues as true negatives, which achieved an area under curve of 0.92 (Supplementary Fig. S11a). We selected the optimal cut-off value of $\zeta \sim 96$ for determining whether a residue in the E1E2 protein is relatively escape resistant or not based on the maximum F1 score and Matthews correlation coefficient (Supplementary Fig. S11b), commonly used metrics for evaluating binary classifiers.

5.9. Identification of buried and exposed residues from the E1E2 structure

Residues in both E1E2 structures (the experimentally resolved partial structure (PDB ID: 7T6X) and the AlphaFold-predicted complete structure) were classified as buried or exposed based on the standard relative solvent accessibility (RSA) metric. Specifically, we used the get-area() function in the PyMOL software (www.pymol.org) with a 1.4 solvent radius parameter to assign each residue in each structure with a solvent accessible surface area (SASA). We obtained the RSA values of each residue by normalizing the respective SASA values per residue in a Gly-X-Gly tripeptide construct (Miller et al. 1987). As suggested in Jardine et al. (2016), residues with a RSA of >0.2 were considered as exposed, while the remaining residues were considered buried.

5.10. Evaluation of the efficacy of known HmAbs

To evaluate the efficacy of known HmAbs based on the escape times obtained from the JM or the E2-only model, we adopted the following criteria. We compared the minimum escape time $t_e^{\min}$ predicted for a HmAb's binding residues (Pierce et al. 2016; Gopal et al. 2017; Keck et al. 2019) with the cut-off value (ζ) for each model. If $t_e^{\min}$ of a HmAb was greater than ζ for a model, that HmAb was characterized as relatively escape resistant by that model and vice versa.

5.11. Site-independent model

In order to compare the JM with a model that ignores all interactions between residues, we defined a site-independent E1E2 fitness landscape model that is characterized solely by the 'fields' $\mathbf{h}$ as

follows:

$$h_i(a) = \ln \frac{1 - f_i(a)}{f_i(a)}, i = 1, 2, \ldots, N, \quad (10)$$

where $f_i(a)$ is the frequency of observing amino acid a at residue i.

Data availability

All data used in this work are publicly available. Top 300 pairs of inter-protein couplings obtained from the JM are listed in Supplementary Data 1. Accession numbers of E1E2 sequences used for inferring the JM and the IM are listed in Supplementary Data 2. The E1E2 infectivity measurements, used for correlating with predictions obtained from the inferred JM and IM, are included in Supplementary Data 3. The mean escape time predicted for each residue by the JM and E2-only model is provided in Supplementary Data 4. The GUI-based software implementation of the MPF-BML method [22], used for inferring the fitness landscape model, is available at https://github.com/ahmedaq/MPF-BML-GUI [21]. Data and scripts for reproducing the results of this manuscript are available at https://github.com/hangzhangust/HCVE1E2. Any additional information related to the data reported in this paper is available from the lead contact upon request.

Supplementary data

Supplementary data is available at *Virus Evolution* online.

Acknowledgements

H.Z. and A.A.Q. were supported by the Hong Kong Research Grants Council (grant numbers 16204519 and 16204121). A.A.Q. and M.R.M. were supported by the Australian Research Council through Discovery Project (DP 230102850). A.A.Q., R.A.B., and M.R.M. were supported by Australia's National Health and Medical Research Council (NHMRC) through Ideas project (2020192). M.R.M. is the recipient of an Australian Research Council Future Fellowship (project number FT200100928). R.A.B. is a fellow funded by NHMRC.

Conflict of interest: The authors declare no conflict of interest.

References

Ahmed, S. F. et al. (2019) 'Sub-dominant Principal Components Inform New Vaccine Targets for HIV Gag', *Bioinformatics*, 35: 3884–9.

Alhammad, Y. et al. (2015) 'Monoclonal Antibodies Directed toward the Hepatitis C Virus Glycoprotein E2 Detect Antigenic Differences Modulated by the N-terminal Hypervariable Region 1 (HVR1), HVR2, and Intergenotypic Variable Region', *Journal of Virology*, 89: 12245–61.

Alhammad, Y. M. O. et al. (2015) 'Longitudinal Sequence and Functional Evolution within Glycoprotein E2 in Hepatitis C Virus Genotype 3a Infection', *PLoS One*, 10: 1–19.

Augestad, E. H. et al. (2020) 'Global and Local Envelope Protein Dynamics of Hepatitis C Virus Determine Broad Antibody Sensitivity', *Science Advances*, 6: eabb5938.

Bailey, J. R. et al. (2015) 'Naturally Selected Hepatitis C Virus Polymorphisms Confer Broad Neutralizing Antibody Resistance', *Journal of Clinical Investigation*, 125: 437–47.

Bank, C. et al. (2014) 'A Systematic Survey of an Intragenic Epistatic Landscape', *Molecular Biology and Evolution*, 32: 229–38.

Bankwitz, D. et al. (2021) 'Hepatitis C Reference Viruses Highlight Potent Antibody Responses and Diverse Viral Functional Interactions with Neutralising Antibodies', *Gut*, 70: 1734–45.

Barton, J. P. et al. (2016) 'Relative Rate and Location of Intra-host HIV Evolution to Evade Cellular Immunity Are Predictable', *Nature Communications*, 7: 11660.

Barton, J. P., Kardar, M., and Chakraborty, A. K. (2015) 'Scaling Laws Describe Memories of Host-pathogen Riposte in the HIV Population', *Proceedings of the National Academy of Sciences*, 112: 1965–70.

Brazzoli, M. et al. (2005) 'Folding and Dimerization of Hepatitis C Virus E1 and E2 Glycoproteins in Stably Transfected CHO Cells', *Virology*, 332: 438–53.

Broering, T. J. et al. (2009) 'Identification and Characterization of Broadly Neutralizing Human Monoclonal Antibodies Directed against the E2 Envelope Glycoprotein of Hepatitis C Virus', *Journal of Virology*, 83: 12 473–12 482.

Bull, R. A. et al. (2011) 'Sequential Bottlenecks Drive Viral Evolution in Early Acute Hepatitis C Virus Infection', *PLoS Pathogens*, 7: 1–14.

Butler, T. C. et al. (2016) 'Identification of Drug Resistance Mutations in HIV from Constraints on Natural Evolution', *Physical Review E*, 93: 022412.

Cerino, A. et al. (1997) 'Antibody Responses to the Hepatitis C Virus E2 Protein: Relationship to Viraemia and Prevalence in Anti-HCV Seronegative Subjects', *Journal of Medical Virology*, 51: 1–5.

Ciczora, Y. et al. (2007) 'Transmembrane Domains of Hepatitis C Virus Envelope Glycoproteins: Residues Involved in E1E2 Heterodimerization and Involvement of These Domains in Virus Entry', *Journal of Virology*, 81: 2372–81.

Cocquerel, L. et al. (2000) 'Charged Residues in the Transmembrane Domains of Hepatitis C Virus Glycoproteins Play a Major Role in the Processing, Subcellular Localization, and Assembly of These Envelope Proteins', *Journal of Virology*, 74: 3623–33.

Cuevas, J. M. et al. (2009) 'Effect of Ribavirin on the Mutation Rate and Spectrum of Hepatitis C Virus in Vivo', *Journal of Virology*, 83: 5760–4.

Dahirel, V. et al. (2011) 'Coordinate Linkage of HIV Evolution Reveals Regions of Immunological Vulnerability', *Proceedings of the National Academy of Sciences*, 108: 11530–5.

Deleersnyder, V. et al. (1997) 'Formation of Native Hepatitis C Virus Glycoprotein Complexes', *Journal of Virology*, 71: 697–704.

Douam, F. et al. (2014) 'Critical Interaction between E1 and E2 Glycoproteins Determines Binding and Fusion Properties of Hepatitis C Virus during Cell Entry', *Hepatology*, 59: 776–88.

Dowd, K. A. et al. (2009) 'Selection Pressure from Neutralizing Antibodies Drives Sequence Evolution during Acute Infection with Hepatitis C Virus', *Gastroenterology*, 136: 2377–86.

Drummer, H. E. et al. (2006) 'A Conserved Gly436-Trp-Leu-Ala-Gly-Leu-Phe-Tyr Motif in Hepatitis C Virus Glycoprotein E2 Is a Determinant of CD81 Binding and Viral Entry', *Journal of Virology*, 80: 7844–53.

El-Diwany, R. et al. (2017) 'Extra-epitopic Hepatitis C Virus Polymorphisms Confer Resistance to Broadly Neutralizing Antibodies by Modulating Binding to Scavenger Receptor B1', *PLoS Pathogens*, 13: e1006235.

Ewens, W. J. (2004) *Mathematical Population Genetics*. Interdisciplinary Applied Mathematics: New York, US.

Falkowska, E. et al. (2007) 'Hepatitis C Virus Envelope Glycoprotein E2 Glycans Modulate Entry, Cd81 Binding, and Neutralization', *Journal of Virology*, 81: 8072–9.

Falson, P. et al. (2015) 'Hepatitis C Virus Envelope Glycoprotein E1 Forms Trimers at the Surface of the Virion', *Journal of Virology*, 89: 10333–46.

Ferguson, A. L. et al. (2013) 'Translating HIV Sequences into Quantitative Fitness Landscapes Predicts Viral Vulnerabilities for Rational Immunogen Design', *Immunity*, 38: 606–17.

Flynn, W. F. et al. (2017) 'Inference of Epistatic Effects Leading to Entrenchment and Drug Resistance in HIV-1 Protease', *Molecular Biology and Evolution*, 34: 1291–306.

Fofana, I. et al. (2012) 'Mutations that Alter Use of Hepatitis C Virus Cell Entry Factors Mediate Escape from Neutralizing Antibodies', *Gastroenterology*, 143: 223–33.e9.

Frumento, N., Flyak, A. I., and Bailey, J. R. (2021) 'Mechanisms of HCV Resistance to Broadly Neutralizing Antibodies', *Current Opinion in Virology*, 50: 23–9.

Gaiha, G. D. et al. (2019) 'Structural Topology Defines Protective CD8+ T Cell Epitopes in the HIV Proteome', *Science*, 364: 480–4.

Gal-Tanamy, M. et al. (2008) 'In Vitro Selection of a Neutralization-resistant Hepatitis C Virus Escape Mutant', *Proceedings of the National Academy of Sciences*, 105: 19450–5.

Goffard, A. et al. (2005) 'Role of N-linked Glycans in the Functions of Hepatitis C Virus Envelope Glycoproteins', *Journal of Virology*, 79: 8400–9.

Gopal, R. et al. (2017) 'Probing the Antigenicity of Hepatitis C Virus Envelope Glycoprotein Complex by High-throughput Mutagenesis', *PLoS Pathogens*, 13: e1006735.

Guan, M. et al. (2012) 'Three Different Functional Microdomains in the Hepatitis C Virus Hypervariable Region 1 (HVR1) Mediate Entry and Immune Evasion', *Journal of Biological Chemistry*, 287: 35631–45.

Haddad, J. G. et al. (2017) 'Identification of Novel Functions for Hepatitis C Virus Envelope Glycoprotein E1 in Virus Entry and Assembly', *Journal of Virology*, 91: e00048–17.

Hart, G. R., and Ferguson, A. L. (2015) 'Empirical Fitness Models for Hepatitis C Virus Immunogen Design', *Physical Biology*, 12: 066006.

Jardine, J. G. et al. (2016) 'Minimally Mutated HIV-1 Broadly Neutralizing Antibodies to Guide Reductionist Vaccine Design', *PLoS Pathogens*, 12: 1–33.

Kato, N. et al. (1993) 'Humoral Immune Response to Hypervariable Region 1 of the Putative Envelope Glycoprotein (Gp70) of Hepatitis C Virus', *Journal of Virology*, 67: 3923–30.

Keck, Z.-Y. et al. (2014) 'Non-random Escape Pathways from a Broadly Neutralizing Human Monoclonal Antibody Map to a Highly Conserved Region on the Hepatitis C Virus E2 Glycoprotein Encompassing Amino Acids 412-423', *PLoS Pathogens*, 10: 1–13.

——— et al. (2016) 'Antibody Response to Hypervariable Region 1 Interferes with Broadly Neutralizing Antibodies to Hepatitis C Virus', *Journal of Virology*, 90: 3112–22.

——— et al. (2009) 'Mutations in Hepatitis C Virus E2 Located outside the CD81 Binding Sites Lead to Escape from Broadly Neutralizing Antibodies but Compromise Virus Infectivity', *Journal of Virology*, 83: 6149–60.

——— et al. (2005) 'Analysis of a Highly Flexible Conformational Immunogenic Domain a in Hepatitis C Virus E2', *Journal of Virology*, 79: 13199–208.

——— et al. (2008) 'A Point Mutation Leading to Hepatitis C Virus Escape from Neutralization by A Monoclonal Antibody to a Conserved Conformational Epitope', *Journal of Virology*, 82: 6067–72.

——— et al. (2019) 'Broadly Neutralizing Antibodies from an Individual that Naturally Cleared Multiple Hepatitis C Virus Infections Uncover Molecular Determinants for E2 Targeting and Vaccine Design', *PLoS Pathogens*, 15: e1007772.

——— et al. (2011) 'Mapping a Region of Hepatitis C Virus E2 that Is Responsible for Escape from Neutralizing Antibodies and a Core CD81-binding Region that Does Not Tolerate Neutralization Escape Mutations', *Journal of Virology*, 85: 10451–63.

——— et al. (2012) 'Human Monoclonal Antibodies to a Novel Cluster of Conformational Epitopes on HCV E2 with Resistance to Neutralization Escape in a Genotype 2a Isolate', *PLoS Pathogens*, 8: 1–21.

Kong, L. et al. (2012) 'Structural Basis of Hepatitis C Virus Neutralization by Broadly Neutralizing Antibody HCV1', *Proceedings of the National Academy of Sciences*, 109: 9499–504.

——— et al. (2015) 'Structure of Hepatitis C Virus Envelope Glycoprotein E1 Antigenic Site 314-324 in Complex with Antibody IGH526', *Journal of Molecular Biology*, 427: 2617–28.

——— et al. (2016) 'Structural Flexibility at a Major Conserved Antibody Target on Hepatitis C Virus E2 Antigen', *Proceedings of the National Academy of Sciences*, 113: 12768–73.

Law, M. et al. (2008) 'Broadly Neutralizing Antibodies Protect against Hepatitis C Virus Quasispecies Challenge', *Nature Medicine*, 14: 25–7.

Leys, C. et al. (2013) 'Detecting Outliers: Do Not Use Standard Deviation around the Mean, Use Absolute Deviation around the Median', *Journal of Experimental Social Psychology*, 49: 764–6.

Li, D. et al. (2016) 'Altered Glycosylation Patterns Increase Immunogenicity of a Subunit Hepatitis C Virus Vaccine, Inducing Neutralizing Antibodies Which Confer Protection in Mice', *Journal of Virology*, 90: 10486–98.

Li, Y., and Modis, Y. (2014) 'A Novel Membrane Fusion Protein Family in Flaviviridae?', *Trends in Microbiology*, 22: 176–82.

Louie, R. H. Y. et al. (2018) 'Fitness Landscape of the Human Immunodeficiency Virus Envelope Protein that Is Targeted by Antibodies', *Proceedings of the National Academy of Sciences*, 115: E564–73.

Mann, J. K. et al. (2014) 'The Fitness Landscape of HIV-1 Gag: Advanced Modeling Approaches and Validation of Model Predictions by in Vitro Testing', *PLoS Computational Biology*, 10: e1003776.

McCaffrey, K. et al. (2011) 'The Variable Regions of Hepatitis C Virus Glycoprotein E2 Have an Essential Structural Role in Glycoprotein Assembly and Virion Infectivity', *Journal of General Virology*, 92: 112–21.

Meunier, J.-C. et al. (2008) 'Isolation and Characterization of Broadly Neutralizing Human Monoclonal Antibodies to the E1 Glycoprotein of Hepatitis C Virus', *Journal of Virology*, 82: 966–73.

Miller, S. et al. (1987) 'Interior and Surface of Monomeric Proteins', *Journal of Molecular Biology*, 196: 641–56.

Mirdita, M. et al. (2022) 'ColabFold: Making Protein Folding Accessible to All', *Nature Methods*, 19: 679–82.

Mitchell, A. L. et al. (2019) 'MGnify: The Microbiome Analysis Resource in 2020', *Nucleic Acids Research*, 48: D570–8.

Mora, T. et al. (2010) 'Maximum Entropy Models for Antibody Diversity', *Proceedings of the National Academy of Sciences*, 107: 5405–10.

Morin, T. J. et al. (2012) 'Human Monoclonal Antibody HCV1 Effectively Prevents and Treats HCV Infection in Chimpanzees', *PLoS Pathogens*, 8: e1002895.

Moustafa, R. et al. (2018) 'Functional Study of the C-terminal Part of Hepatitis C Virus E1 Ectodomain', *Journal of Virology*, 92: e00939–18.

Osburn, W. O. et al. (2014) 'Clearance of Hepatitis C Infection Is Associated with the Early Appearance of Broad Neutralizing Antibody Responses', *Hepatology*, 59: 2140–51.

Pfaff-Kilgore, J. M. et al. (2022) 'Sites of Vulnerability in HCV E1E2 Identified by Comprehensive Functional Screening', *Cell Reports*, 39: 110859.

Pierce, B. G. et al. (2016) 'Global Mapping of Antibody Recognition of the Hepatitis C Virus E2 Glycoprotein: Implications for Vaccine Design', *Proceedings of the National Academy of Sciences*, 113: E6946–54.

Quadeer, A. A. et al. (2019) 'MPF-BML: A Standalone GUI-based Package for Maximum Entropy Model Inference', *Bioinformatics*, 36: 2278–9.

Quadeer, A. A. et al. (2020) 'Deconvolving Mutational Patterns of Poliovirus Outbreaks Reveals Its Intrinsic Fitness Landscape', *Nature Communications*, 11: 377.

——— et al. (2014) 'Statistical Linkage Analysis of Substitutions in Patient-derived Sequences of Genotype 1a Hepatitis C Virus Nonstructural Protein 3 Exposes Targets for Immunogen Design', *Journal of Virology*, 88: 7628–44.

Quadeer, A. A., Louie, R. H. Y., and Mckay, M. R. (2019) 'Identifying Immunologically-vulnerable Regions of the HCV E2 Glycoprotein and Broadly Neutralizing Antibodies that Target Them', *Nature Communications*, 10: 2073.

Quadeer, A. A., Morales-Jimenez, D., and McKay, M. R. (2018) 'Coevolution Networks of HIV/HCV are Modular with Direct Association to Structure and Function', *PLoS Computational Biology*, 14: e1006409.

Rosenthal, E. S., and Graham, C. S. (2016) 'Price and Affordability of Direct-acting Antiviral Regimens for Hepatitis C Virus in the United States', *Infectious Agents and Cancer*, 11: 24.

Rossi, C. et al. (2018) 'Hepatitis C Virus Reinfection after Successful Treatment with Direct-acting Antiviral Therapy in a Large Population-based Cohort', *Journal of Hepatology*, 69: 1007–14.

Rothwangl, K. B. et al. (2008) 'Dissecting the Role of Putative CD81 Binding Regions of E2 in Mediating Hcv Entry: Putative CD81 Binding Region 1 Is Not Involved in CD81 Binding', *Virology Journal*, 5: 46.

Sanjuan, R. et al. (2010) 'Viral Mutation Rates', *Journal of Virology*, 84: 9733–48.

Singer, J. B. et al. (2018) 'GLUE: A Flexible Software System for Virus Sequence Data', *BMC Bioinformatics*, 19: 532.

Singer, J. et al. (2019) 'Interpreting Viral Deep Sequencing Data with GLUE', *Viruses*, 11: 323.

Sohail, M. S. et al. (2021) 'MPL Resolves Genetic Linkage in Fitness Inference from Complex Evolutionary Histories', *Nature Biotechnology*, 39: 472–9.

Sormanni, P., Aprile, F. A., and Vendruscolo, M. (2015) 'Rational Design of Antibodies Targeting Specific Epitopes within Intrinsically Disordered Proteins', *Proceedings of the National Academy of Sciences*, 112: 9902–7.

Strimmer, K., and Haeseler, A. V. (2009) 'Genetic Distances and Nucleotide Substitution Models', in Lemey, P., Salemi, M. and Vandamme, A.-M. (eds) *The Phylogenetic Handbook: A Practical Approach to DNA and Protein Phylogeny*, pp. 112–3. Cambridge University Press: Cambridge, England.

Ströh, L. J., Nagarathinam, K., and Krey, T. (2018) 'Conformational Flexibility in the CD81-binding Site of the Hepatitis C Virus Glycoprotein E2', *Frontiers in Immunology*, 9: 1396.

Strong, S. P. et al. (1998) 'Entropy and Information in Neural Spike Trains', *Physical Review Letters*, 80: 197–200.

Tong, Y. et al. (2018) 'Role of Hepatitis C Virus Envelope Glycoprotein E1 in Virus Entry and Assembly', *Frontiers in Immunology*, 9: 1411.

Torrents de la Peña, A. et al. (2022) 'Structure of the Hepatitis C Virus E1E2 Glycoprotein Complex', *Science*, 378: 263–9.

Urbanowicz, R. A. et al. (2015) 'A Diverse Panel of Hepatitis C Virus Glycoproteins for Use in Vaccine Research Reveals Extremes of Monoclonal Antibody Neutralization Resistance', *Journal of Virology*, 90: 3288–301.

Velázquez-Moctezuma, R. et al. (2021) 'Mechanisms of Hepatitis C Virus Escape from Vaccine-relevant Neutralizing Antibodies', *Vaccines*, 9: 291.

——— et al. (2019) 'Hepatitis C Virus Escape Studies of Human Antibody Ar3A Reveal a High Barrier to Resistance and Novel Insights on Viral Antibody Evasion Mechanisms', *Journal of Virology*, 93: e01909–18.

Wahid, A. et al. (2013) 'Disulfide Bonds in Hepatitis C Virus Glycoprotein E1 Control the Assembly and Entry Functions of E2 Glycoprotein', *Journal of Virology*, 87: 1605–17.

World Health Organization. (2022), *Hepatitis C, Fact Sheet* <https://www.who.int/news-room/fact-sheets/detail/hepatitis-c> accessed 1 Nov 2022.

Wyles, D. L., and Luetkemeyer, A. F. (2017) 'Understanding Hepatitis C Virus Drug Resistance: Clinical Implications for Current and Future Regimens', *Topics in Antiviral Medicine*, 25: 103–9.

Yan, Y. et al. (2019) 'A Nanoparticle-based Hepatitis C Virus Vaccine with Enhanced Potency', *The Journal of Infectious Diseases*, 221: 1304–14.

Zhang, T.-H. et al. (2020) 'Predominance of Positive Epistasis among Drug Resistance-associated Mutations in HIV-1 Protease', *PLoS Genetics*, 16: 1–22.

Zhang, H., Quadeer, A. A., and McKay, M. R. (2022) 'Evolutionary Modeling Reveals Enhanced Mutational Flexibility of HCV Subtype 1b Compared with 1a', *iScience*, 25: 103569.

Zhang, H., Quadeer, A. A, and McKay, M. R. (2023) 'Direct-acting antiviral resistance of Hepatitis C virus is promoted by epistasis', *Nat Commun*, 14.