Integrative Computational Immunology:

From Molecules to Mortality

by

Eric Andrew Wilson

A Dissertation Presented in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Approved July 2022 by the
Graduate Supervisory Committee:

Abhishek Singharoy, Co-Chair
Karen Anderson, Co-Chair
Neal Woodbury
Petr Šulc

ARIZONA STATE UNIVERSITY

August 2022

ABSTRACT

Computational models have long been used to describe and predict the outcome of complex immunological processes. The dissertation work described here centers on the construction of multiscale computational immunology models that derives biological insights at the population, systems, and atomistic levels. First, SARS-CoV-2 mortality is investigated through the lens of the predicted robustness of CD8+ T cell responses in 23 different populations. The robustness of CD8+ T cell responses in a given population was modeled by predicting the efficiency of endemic MHC-I protein variants to present peptides derived from SARS-CoV-2 proteins to circulating T cells. To accomplish this task, an algorithm, called EnsembleMHC, was developed to predict viral peptides with a high probability of being recognized by CD T cells. It was discovered that there was significant variation in the efficiency of different MHC-I protein variants to present SARS-CoV-2 derived peptides, and countries enriched with variants with high presentation efficiency had significantly lower mortality rates. Second, a biophysics-based MHC-I peptide prediction algorithm was developed. The MHC-I protein is the most polymorphic protein in the human genome with polymorphisms in the peptide binding causing striking changes in the amino acid compositions, or binding motifs, of peptide species capable of stable binding. A deep learning model, coined HLA-Inception, was trained to predict peptide binding using only biophysical properties, namely electrostatic potential. HLA-Inception was shown to be extremely accurate and efficient at predicting peptide binding motifs and was used to determine the peptide binding motifs of 5,821 MHC-I protein variants. Finally, the impact of stalk glycosylations on NL63 protein dynamics was investigated. Previous data has shown that coronavirus crown glycans play an important role in immune evasion and receptor binding, however, little is known about the role of the stalk glycans. Through the integration of computational biology, experimental data,

and physics-based simulations, the stalk glycans were shown to heavily influence the bending angle of spike protein, with a particular emphasis on the glycan at position 1242. Further investigation revealed that removal of the N1242 glycan significantly reduced infectivity, highlighting a new potential therapeutic target. Overall, these investigations and associated innovations in integrative modeling.

DEDICATION

To Terri, Steve, Sean, and Mollie, I dedicate this work to you. For it was only through your unwavering support, that this was made possible.

ACKNOWLEDGMENTS

I would first like to acknowledge my dissertation committee. I was very lucky to be co-advised by two terrific advisors, Karen Anderson and Abhishek Singharoy. Karen, thank you for taking me into your lab as an undergrad, and continuing to mentor me through the last 8 years. Your insight and scientific process had a profound impact on the development of my own. Abhi, I would not be the scientist I am today without your support and guidance. You not only taught me to how to do computational research, but also all the soft skills needed to succeed in academia. Next, I would like to thank my committee members, Petr Šulc and Neal Woodbury. Petr, I have found your insight to be especially thoughtful, and always appreciated. Neal, thank you for showing a genuine interest in my work. You are a role model for interdisciplinary work, and working at the boundary of academia and industry.

This work was also supported by two sets of great lab mates. For the Anderson lab, I would like to thank two different groups of students: the first and second generations. The first generation of students played a significant role in mentoring me early in my research career. These include Diego, Shay, Krishna, Radwa, and Emma. In particular, I would like to thank Diego for inspiring me to pursue computational research, and Krishna for his attempts to make a decent experimentalist out of me. The second generation of graduate students mainly served as colleagues and friends. These include Peaches, Jacquie, Mark, Skyler, Dalton, Siril, and Oliver. I would like to particularly thank Peaches who has been a great friend across Anderson lab generations. Also, I would like to thank the Anderson lab graduate students that started with me, namely Mark and Jacquie, for their friendship and help navigating the early years. Finally, I was greatly assisted by two research professionals, Padhma and Marika, who provided immense support over the years.

For the Singharoy lab, I would like to thank John, Jon, JK, Jake, Jacob, Chun, and

Chitrak. In particular, I would like to thank John for being my partner in the trenches, Jon for the great conversations, and JK for the insights into machine learning.

Good science is collaborative, and I was fortunate enough to work with some of the best. For the work on the NL63 spike protein, I would like to thank David Chmielewski, Jing Jin, Greg Pintilie, and Wah Chiu. For work regarding the Nramp protein, I would like to thank Sam Berry, Gerry Zavala, and Rachelle Gaudet. For work on the B2V2R protein, I would like to thank Anthony Nguyen, Harsh Bansia, and Amedee des Georges. For work on the ChAdOx1 vector, I would like to thank Alex Baker, Ryan Boyd, Daipayan Sarkar, Chun Kit Chan, Taylor Cohen, Josh Vermaas and Mitesh Borad.

None of the research here would be possible without the resources provided by ASU research computing, Oak Ridge National labs, the Open science grid, and XSEDE. In particular, I would like to thank Jason Yalim, Gil Speyer, and Eric Tannehill from ASU research computing for their help.

I would also like to acknowledge my network of friends. From those who I grew up with to those who I meet in college, thank you for providing much needed distractions during times of frustration, and reminding me that there is a world outside the lab.

Finally, I would like to thank my family. To my parents Terri and Steve, thank you for the unconditional love and support. To my brother Sean, thank you for always believing in me. To my fiancé Mollie, your daily support and love is what keeps me going.

TABLE OF CONTENTS

CHAPTER Page

APPENDIX Page

LIST OF TABLES

LIST OF FIGURES

Figure Page

Chapter 1

RESEARCH SUMMARY

The human immune system is a truly remarkable and intricate system, simultaneously operating at various scales and specificities. It is robust enough to passively protect the host against the general milieu of bacteria encountered on a daily basis, while maintaining the specificity to mount highly selective adaptive immune responses against a novel viral infection. Amazingly, the immune system not only has the flexibility to defend the host against the constant barrage of external threats, such as pathogens and microbes, but also the ability to protect the host from internal threats like cancer. Together, these aspects make the immune system a truly fascinating system to study. The burgeoning field of molecular immunology has seen striking advancements in a relatively short amount of time. For example, the field has evolved from the discovery of T cells (Gowans *et al.*, 1962; Miller *et al.*, 1962) to the development of targeted T-cell based immunotherapies capable of driving near-miraculous tumor rejections (Zacharakis *et al.*, 2018; Leidner *et al.*, 2022) in a little over six decades. However, despite such milestones there is still much to discover about the function of the immune system (TW *et al.*, 2019), particularly in the realm of correlating detailed biophysical information to phenotypic outcomes. One method of addressing and predicting the outcomes of complex systems starting from molecular cues is by deploying mathematical and computational models. Early examples of immunology-based modeling, primarily mathematical in nature, included epidemiological models of malaria infections (Ross, 1916), stochastic models of immune induction (Marchalonis and Gledhill, 1968), and population dynamics of persistent viral infections (Nowak and Bangham, 1996). Since, we have seen an explosion of computational modeling to

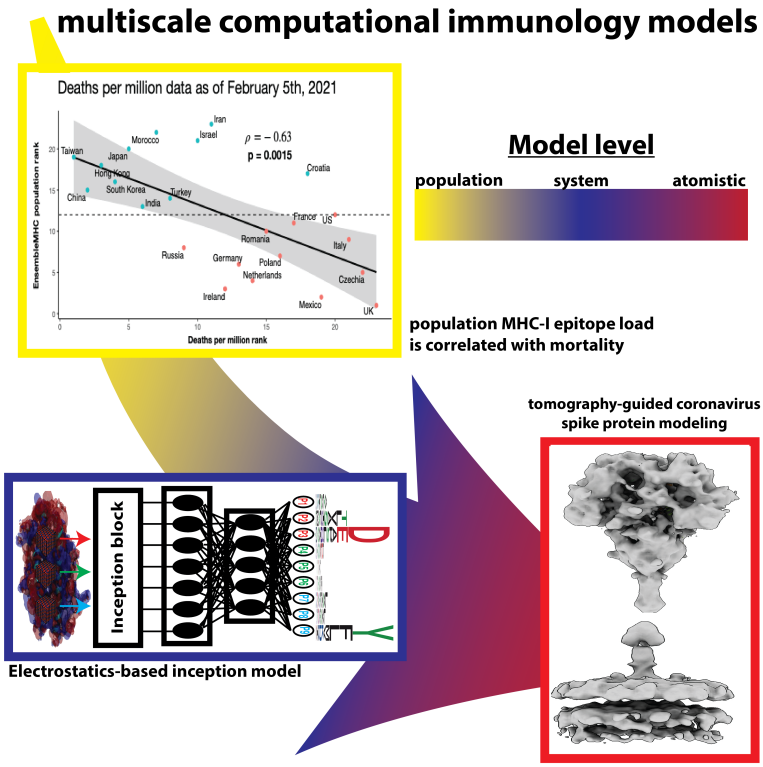

Figure 1.1: Multiscale computational immunology models

describe a diverse range of immunological functions. These models cross paradigms and scales from population-level models, like those investigating large scale immunity to SARS-CoV-2 (Britton *et al.*, 2020; Wilson *et al.*, 2021), to deep natural language models predicting viral escape mutations (Hie *et al.*, 2021). These models often work in tandem with a wide array of omics and experimental data to improve predictions regarding immunological function.

Our work describes integrative computational models that reveal biologically relevant information across three scales: population level, systems level, and atomistic level **(Figure 1.1)**. In each case, population, biophysical, structural, and omics data are leveraged and integrated into the models to answer questions ranging from estimations of population CD8+ T cell response to SARS-CoV-2 viral infections to impact of glycans on coronaviral spike protein dynamics.

In Chapter two, the relationship between observed population mortality to SARS-CoV-2 infections and predicted robustness of CD8+ T cell response is assessed. $CD8^+$ T cell response robustness is modeled by calculating the average magnitude of SARS-CoV-2 viral peptides presented by MHC-I proteins in a population based on the endemic frequencies of MHC-I protein variants. To analyze this relationship, an MHC-I epitope prediction algorithm, coined EnsembleMHC, was developed. Our algorithm was shown to have high accuracy in identifying viral peptides likely to produce a $CD8^+$ T cell based immune response. The application of this algorithm revealed that certain MHC-I protein variants present more virus-derived peptides than others, and countries enriched with these alleles experienced lower SARS-CoV-2 related deaths.

In the third Chapter, the integration of biophysical data and machine learning was explored as a method of improving MHC-I peptide prediction. The identification of MHC-I peptides capable of cell surface presentation is crucial for the further development of antiviral and anticancer immunotherapies. However, the extremely polymorphic nature of the MHC-I protein makes the experimental determination of MHC-I binding motifs difficult, excluding many populations from such therapies. To combat this, a machine learning algorithm, called HLA-Inception, was developed that predicts the peptide binding motifs of 5,821 different MHC-I alleles using only the electrostatic environment of the MHC-I binding pocket. HLA-Inception was found to be both accurate and fast, out competing current state-of-the-art MHC-I prediction algorithms.

Finally, in the fourth chapter, the protein bending dynamics of the NL63 spike protein was investigated. Using a combination of high quality CryoET data and extensive physics-based simulations, a full length protein model recapitulating experimental bending dynamics was built. The model was then used to investigate the impact of

stalk glycan modification on bending profiles. The stalk glycans, and in particular the glycan found at asparagine 1242, produce significant changes on bending after removal. Further experimental analysis revealed that removal of the N1242 glycan significantly reduced infectivity, identifying a new potential therapeutic target for coronavirus vaccines.

In summary, the work in this dissertation describes the development of computational immunology models that transcend multiple paradigms. We show how the application of molecular modeling and biophysical techniques, augmented by innovations in statistical inference can serve as a credible and empirically viable tool for driving biomedical discoveries.

Chapter 2

TOTAL PREDICTED MHC-I EPITOPE LOAD IS INVERSELY ASSOCIATED WITH POPULATION MORTALITY FROM SARS-COV-2

This chapter is published:

Wilson, E.A., Hirneise, G., Singharoy, A. and Anderson, K.S., 2021. Total predicted MHC-I epitope load is inversely associated with population mortality from SARS-CoV-2. Cell Reports Medicine, 2(3), p.100221.

2.1 Abstract

Polymorphisms in MHC-I protein sequences across human populations significantly impacts viral peptide binding capacity and thus alters T cell immunity to infection. Consequently, allelic variants of the MHC-I protein have been found to be associated with patient outcome to various viral infections, including SARS-CoV. In the present study, we assess the relationship between observed SARS-CoV-2 population mortality and the predicted viral binding capacities of 52 common MHC-I alleles. Potential SARS-CoV-2 MHC-I peptides were identified using a consensus MHC-I binding and presentation prediction algorithm, called EnsembleMHC. Starting with nearly 3.5 million candidates, we resolved a few hundred highly probable MHC-I peptides. By weighing individual MHC allele-specific SARS-CoV-2 binding capacity with population frequency in 23 countries, we discover a strong inverse correlation between the predicted population SARS-CoV-2 peptide binding capacity and observed mortality rate. Our computations reveal that peptides derived from the structural proteins of the virus produces a stronger association with observed mortality rate, highlighting the importance of S, N, M, E proteins in driving productive immune responses. The

correlation between epitope binding capacity and population mortality risk remains robust across a range of socioeconomic and epidemiological factors. A combination of binding capacity, number of deaths due to COPD complications, gender demographics. and the proportions of the population that were over the age of 65 and overweight offered the strongest determinant of at-risk populations. These results bring to light how molecular changes in the MHC-I proteins may affect population-level outcomes of viral infection.

2.2 Introduction

In December 2019, the novel coronavirus, SARS-CoV-2 was identified from a cluster of cases of pneumonia in Wuhan, China (Zu *et al.*, 2020; Li *et al.*, 2020). With over 73.1 million cases and over 1.6 million deaths, the viral spread has been declared a global pandemic by the World Health Organization(Guo *et al.*, 2020). Due to its high rate of transmission and unpredictable severity, there is an immediate need for information surrounding the adaptive immune response towards SARS-CoV-2.

A robust T cell response is integral for the clearance of coronaviruses, and generation of lasting immunity(Channappanavar *et al.*, 2014).The potential role of T cells for coronavirus clearance has been supported by the identification of immunogenic $CD8^+$ T cell epitopes in the S (Spike), N (Nucleocapsid), M (Membrane), and E (Envelope) proteins(Janice Oh *et al.*, 2012). Additionally, SARS-CoV specific $CD8^+$ T cells have been shown to provide long lasting immunity with memory $CD8^+$ T cells being detected up to 17 years post infection(Ng *et al.*, 2016; Channappanavar *et al.*, 2014; Le Bert *et al.*, 2020). The specifics of the T cell response to SARS-CoV-2 is still evolving. However, a recent screening of SARS-CoV-2 peptides revealed a majority of the $CD8^+$ T cell immune response is targeted towards viral structural proteins (N, M, S)(Grifoni *et al.*, 2020).

A successful $CD8^+$ T cell response is contingent on the efficient presentation of viral protein fragments by Major Histocompatibility Complex I (MHC-I) proteins. MHC-I molecules bind and present peptides derived from endogenous proteins on the cell surface for $CD8^+$ T cell interrogation. The MHC-I protein is highly polymorphic, with amino acid substitutions within the peptide binding groove drastically altering the composition of presented peptides. Consequently, the influence of MHC genotype to shape patient outcome has been well studied in the context of viral infections(Matzaraki *et al.*, 2017). For coronaviruses, there have been several studies of MHC association with disease susceptibility. A study of a Taiwanese and Hong Kong cohort of patients with SARS-CoV found that the MHC-I alleles HLA-B*07:03 and HLA-B*46:01 were linked to increased susceptibility while HLA-Cw*15:02 was linked to increased resistance (Lin *et al.*, 2003; Wang *et al.*, 2011; Ng *et al.*, 2004). However, some of the reported associations did not remain after statistical correction, and it is still unclear if MHC-outcome associations reported for SARS-CoV are applicable to SARS-CoV-2(Ng *et al.*, 2010; Sanchez-Mazas, 2020). Recently, a comprehensive prediction of SARS-CoV-2 MHC-I peptides indicated a relative depletion of high affinity binding peptides for HLA-B*46:01, hinting at a similar association profile in SARS-CoV-2(Nguyen *et al.*, 2020). More importantly, it remains elusive if such a depletion of putative high affinity peptides will impact patient outcome to SARS-CoV-2 infections.

The lack of large scale genomic data linking individual MHC genotype and outcome from SARS-CoV-2 infections precludes a similar analysis as performed for SARS-CoV(Lin *et al.*, 2003; Wang *et al.*, 2011; Ng *et al.*, 2004). Therefore, we endeavored to assess the relationship between the predicted SARS-CoV-2 binding capacity of a population and the observed SARS-CoV-2 mortality rate. However, current MHC-I prediction algorithms have been characterized by a high false positive rate particularly when predicting peptides that are naturally presented(Zhao and Sher, 2018;

Sarkizova *et al.*, 2020). To mitigate false positives and identify the highest confidence SARS-CoV-2 MHC-I peptides, we developed a consensus prediction algorithm, coined EnsembleMHC, and predicted MHC-I peptides for a panel of 52 common MHC-I alleles(González-Galarza *et al.*, 2015). This prediction workflow integrates seven different algorithms that have been parameterized on high-quality mass spectrometry data and provides a confidence level for each identified peptide(O'Donnell *et al.*, 2020; Jurtz *et al.*, 2017; Andreatta and Nielsen, 2016; Bassani-Sternberg *et al.*, 2017; Zhang *et al.*, 2009; Rasmussen *et al.*, 2016; Sarkizova *et al.*, 2020). The distribution of the number of high-confidence peptides assigned to each allele was used to assess a country-specific SARS-CoV-2 binding capacity, called the EnsembleMHC population score, for 23 countries (for selection criteria, please refer to the Methods). This score was derived by weighing the individual binding capacities of the 52 MHC-I alleles by their endemic frequencies. We observe a strong inverse correlation between the EnsembleMHC population score and observed population SARS-CoV-2 mortality. Furthermore, the correlation is shown to become stronger when considering EnsembleMHC population scores based solely on SARS-CoV-2 structural proteins, underlining their potential importance in driving a robust immune response. Based on their predicted binding affinity, expression, and sequence conservation in viral isolates, we identified 108 peptides derived from SARS-CoV-2 structural proteins that are high-value targets for $CD8^+$ T cell vaccine development.

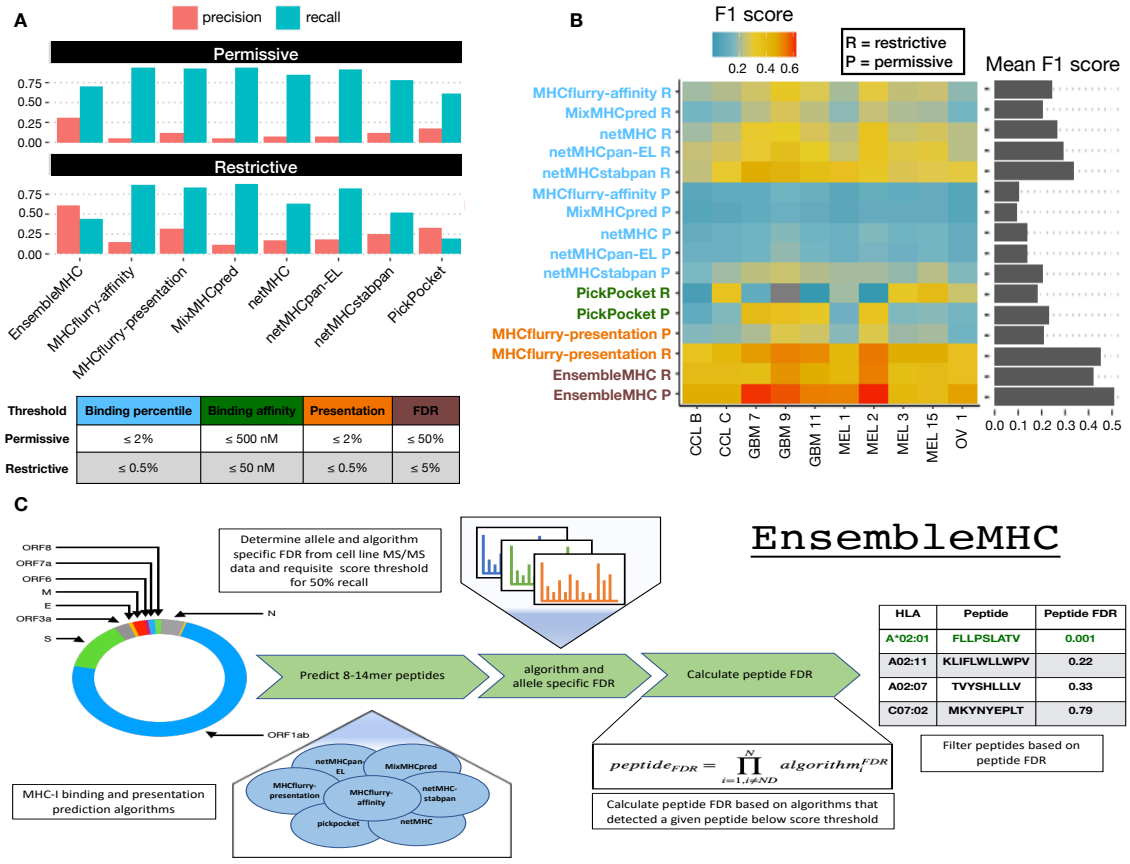

Figure 2.1: Application of the EnsembleMHC prediction algorithm. The EnsembleMHC prediction algorithm was used to recover MHC-I peptides from 10 tumor sample data sets. **A.** The average precision and recall for EnsembleMHC and each component algorithm was calculated across all 10 tumor samples. Peptide identification by each algorithm was based on commonly used restrictive (strong) or permissive (strong and weak) binding affinity thresholds **(inset table)**. **B.** The F1 score of each algorithm was calculated for all tumor samples. Each algorithm is grouped into 1 of 4 categories: binding affinity represented by percentile score (blue), binding affinity represented by predicted peptide IC50 value (green), MHC-I presentation prediction (orange), and EnsembleMHC (brown). The heatmap colors indicate the value of the observed F1 score (color bar) for a given algorithm (y-axis) on a particular data set (x-axis). Warmer colors indicate higher F1 scores, and cooler colors indicate lower F1 scores. The average F1 score for each algorithm across all samples is shown in the marginal bar plot. **C.** The schematic for the application of the EnsembleMHC predication algorithm to identify SARS-CoV-2 MHC-I peptides.

2.3 Results

2.3.1 EnsembleMHC Workflow Offers More Precise MHC-I Presentation Predictions than Individual Algorithms.

The accurate assessment of differences in SARS-CoV-2 binding capacities across MHC-I allelic variants requires the isolation of MHC-I peptides with a high probability of being presented. EnsembleMHC provides the requisite precision through the use of allele and algorithm-specific score thresholds and peptide confidence assignment.

MHC-I alleles substantially vary in both peptide binding repertoire size and median binding affinity(Paul *et al.*, 2013). The EnsembleMHC workflow addresses this inter-allele variation by identifying peptides based on MHC allele and algorithm-specific binding affinity thresholds. These thresholds were set by benchmarking each of the seven component algorithms against 52 single MHC allele peptide data sets(Sarkizova *et al.*, 2020). Each data set consists of mass spectrometry-confirmed MHC-I peptides that have been naturally presented by a model cell line expressing one of the 52 select MHC-I alleles. These experimentally validated peptides, denoted target peptides, were supplemented with a 100-fold excess of decoy peptides. Decoys were generated by randomly sampling peptides that were not detected by mass spectrometry, but were derived from the same protein sources as a detected target peptide. Algorithm and allele-specific binding affinity thresholds were then identified through the independent application of each component algorithm to all MHC allele data sets. For every data set and algorithm combination, the target and decoy peptides were ranked by predicted binding affinity to the MHC allele defined by that data set. Then, an algorithm-specific binding affinity threshold was set to the minimum score needed to isolated the highest affinity peptides commensurate to 50% of the observed allele repertoire size **(Methods, Figure A.1)**. The observed allele repertoire size was

defined as the total number of target peptides within a given single MHC allele data set. Therefore, if a data set had 1000 target peptides, the top 500 highest affinity peptides would be selected, and the algorithm-specific threshold would be set to the predicted binding affinity of the 500^{th} peptide. This parameterization method resulted in the generation of a customized set of allele and algorithm-specific binding affinity thresholds in which an expected quantity of peptides can be recovered.

Consensus MHC-I prediction typically require a method for combining outputs from each individual component algorithm into a composite score. This composite score is then used for peptide selection. EnsembleMHC identifies high-confidence peptides based on filtering by a quantity called $peptide^{FDR}$ **(Methods Eq. 2.1)**. During the identification of allele and algorithm-specific binding affinity thresholds, the empirical false detection rate (FDR) of each algorithm was calculated. This calculation was based on the proportion of target to decoy peptides isolated by the algorithm specific binding affinity threshold. A $peptide^{FDR}$ is then assigned to each individual peptide by taking the product of the empirical FDRs of each algorithm that identified that peptide for the same MHC-I allele. Analysis of the parameterization process revealed that the overall performance of each included algorithms was comparable, and there was diversity in individual peptide calls by each algorithm, supporting an integrated approach to peptide confidence assessment **(Figure A.1)**. Peptide identification by EnsembleMHC was performed by selecting all peptides with a $peptide^{FDR}$ of less than or equal to 5%(Nichols, 2007).

The efficacy of $peptide^{FDR}$ as a filtering metric was determined through the prediction of naturally presented MHC-I peptides derived from ten tumor samples(Sarkizova *et al.*, 2020) **(Figure 2.1)**. Similar to the single MHC allele data sets, each tumor sample data set consisted of mass spectrometry-detected target peptides and a 100-fold excess of decoy peptides. The relative performance of EnsembleMHC was assessed

via comparison with individual component algorithms. Peptide identification by each algorithm was based on a restrictive or permissive binding affinity thresholds **(Figure 2.1A (inset table))**. For the component algorithms, the permissive and restrictive thresholds correspond to commonly used binding affinity cutoffs for the identification of weak and strong binders, respectively(Nielsen *et al.*, 2020). The performance of each algorithm on the ten data sets was evaluated through the calculation of the empirical precision, recall, and F1 score.

The average precision and recall of each algorithm across all tumor samples demonstrated an inverse relationship **(Figure 2.1A)**. In general, restrictive binding affinity thresholds produced higher precision at the cost of poorer recall. When comparing the precision of each algorithm at restrictive thresholds, EnsembleMHC demonstrated a 3.4-fold improvement over the median precision of individual component algorithms. EnsembleMHC also produced the highest F1 score with an average of 0.51 followed by mhcflurry-presentation with an F1 score of 0.45, both of which are 1.5-2 fold higher than the rest of the algorithms **(Figure 2.1B)**. This result was shown to be robust across a range of $peptide^{FDR}$ cutoff thresholds **(Figure A.1)** and alternative performance metrics **(Figure A.1)**. Furthermore, EnsembleMHC demonstrated the ability to more efficiently prioritize peptides with experimentally established immunogenicity from the Hepatitis-C genome polyprotein, the Dengue virus genome polyprotein, and the HIV-1 POL-GAG protein **(Figure A.1)**. Taken together, these results demonstrate the enhanced precision of EnsembleMHC over individual component algorithms when using common binding affinity thresholds.

In summary, the EnsembleMHC workflow offers two desirable features. First, it determines allele-specific binding affinity thresholds for each algorithm at which a known quantity of peptides are expected to be successfully presented on the cell surface. Second, it assigns a confidence level to each peptide call made by each algorithm.

Together, these traits enhance the ability to identify MHC-I peptides with a high probability of successful cell surface presentation.

EnsembleMHC was used to identify MHC-I peptides for the SARS-CoV-2 virus **(Figure 2.1C)**. The resulting identification of high-confidence SARS-CoV-2 peptides allows for the characterization of alleles that are enriched or depleted for predicted MHC-I peptides. The resulting distribution of allele-specific SARS-CoV-2 binding capacities will then be weighed by the normalized frequencies of the 52 alleles **(Figure A.2, Methods Eq. 2.2-2.3)** in 23 countries to determine the population-specific SARS-CoV-2 binding capacity or EnsembleMHC population score **(Methods Eq. 2.4)**. The potential impact of varying population SARS-CoV-2 binding capacities on disease outcome can then be assessed by correlating population SARS-CoV-2 mortality rates with EnsembleMHC population scores. Below, we use EnsembleMHC population scores to stratify countries based on their mortality risks.

2.3.2 The MHC-I Peptide-Allele Distribution for SARS-CoV-2 Structural Proteins Is Especially Disproportionate.

MHC-I peptides derived from the SARS-CoV-2 proteome were predicted and prioritized using EnsembleMHC. A total of 67,207 potential 8-14mer viral peptides were evaluated for each of the considered MHC-I alleles. After filtering the pool of candidate peptides at the 5% $peptide^{FDR}$ threshold, the number of potential peptides was reduced from 3.49 Million to 971 (658 unique peptides) **(Figure A.3, Table A.1)**. Illustrated in **Figure 2.2A**, the viral peptide-MHC allele (or peptide-allele) distribution for high-confidence SARS-CoV-2 peptides was determined by assigning the identified peptides to their predicted MHC-I alleles. There was a median of 16 peptides per allele with a maximum of 47 peptides (HLA-A*24:02), a minimum of 3 peptides (HLA-A*02:05), and an interquartile range (IQR) of 16 peptides. Quality

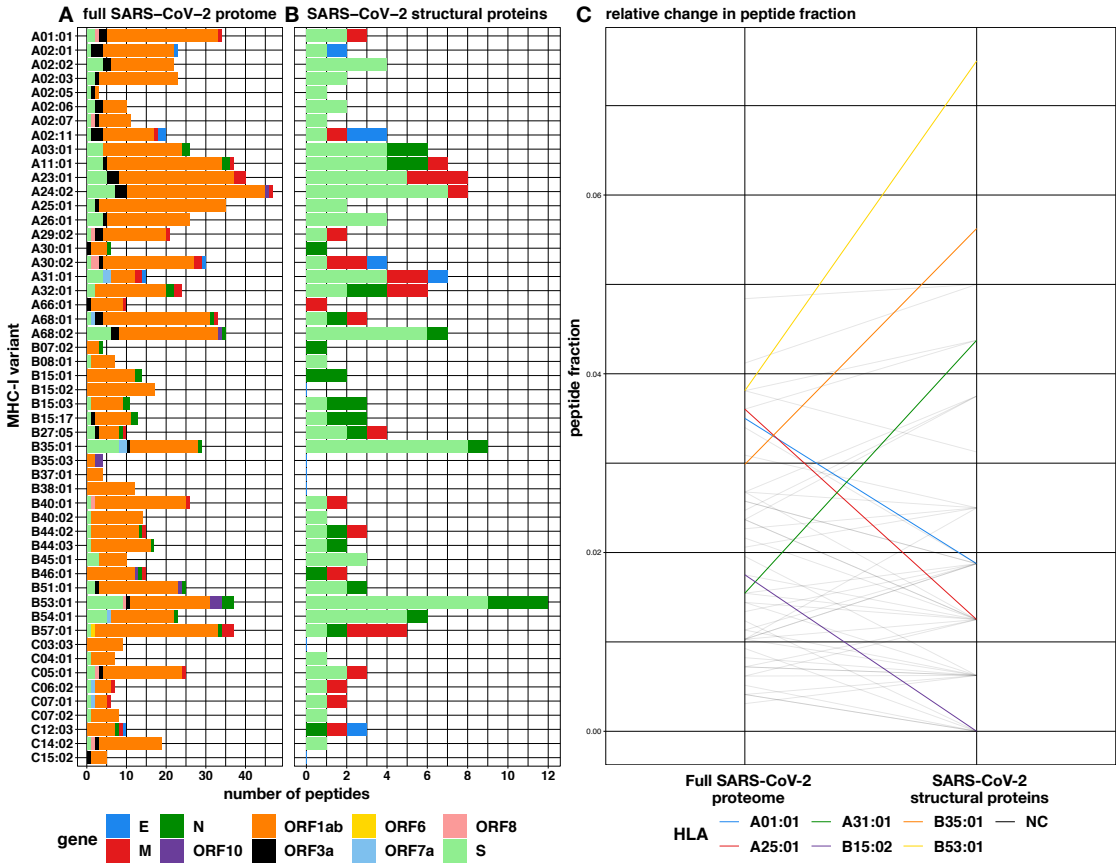

Figure 2.2: Prediction of SARS-CoV-2 peptides across 52 common MHC-I alleles. A-B. The EnsembleMHC workflow was used to predict MHC-I peptides for 52 alleles from the entire SARS-CoV-2 proteome or specifically SARS-CoV-2 structural proteins (envelope, spike, nucleocapsid, and membrane). **C.** The peptide fractions for both protein sets were calculated by dividing the number of peptides assigned to a given allele by the total number of identified peptides for that protein set. Each line indicates the change in peptide fraction observed by a given allele when comparing the viral peptide-MHC allele distribution for the full SARS-CoV-2 proteome proteome or structural proteins. Alleles showing a change of greater than the median peptide fraction, $\tilde{X} = 0.015$, are highlighted in color. For the performance of EnsmebleMHC at a range of different $peptide^{FDR}$ cutoff thresholds, refer to **Figure A.1**

assurance of the predicted peptides was performed by computing the peptide length frequencies and binding motifs. The predicted peptides were found to adhere to expected MHC-I peptide lengths(Trolle *et al.*, 2016) with 78% of the peptides being 9 amino acids in length, 13% being 10 amino acids in length, and 8% of peptides accounting for the remaining lengths **(Figure A.3)**. Similarly, logo plots generated from predicted peptides were found to closely reflect reference peptide binding motifs for considered alleles(Rapin *et al.*, 2010)**(Figure A.3)**. Overall, the EnsembleMHC prediction platform demonstrated the ability to isolate a short list of potential peptides which adhere to expected MHC-I peptide characteristics.

The high expression, relative conservation, and reduced search space of SARS-CoV-2 structural proteins (S, E, M, and N) makes MHC-I binding peptides derived from these proteins high-value targets for $CD8^+$ T cell-based vaccine development. **Figure 2.2B** describes the peptide-allele distribution for predicted MHC-I peptides originating from the four structural proteins. This analysis markedly reduces the number of considered peptides from 658 to 108 **(Table A.1)**. The median number of predicted SARS-CoV-2 structural peptides assigned to each MHC-I allele was found to be 2 with a maximum of 12 peptides (HLA-B*53:01), a minimum of 0 (HLA-B*15:02, B*35:03,B*38:01,C*03:03,C*15:02), and a IQR of 3 peptides. Analysis of the molecular source of the identified SARS-CoV-2 structural protein peptides revealed that they originate from enriched regions that are highly conserved **(Figure A.4AB)**. This indicates that such peptides would be good candidates for targeted therapies as they are unlikely to be disrupted by mutation, and several peptides can be targeted using minimal stretches of the source protein. Altogether, consideration of MHC-I peptides derived only from SARS-CoV-2 structural proteins reduces the number of potential peptides to a condensed set of high-value targets that is amenable to experimental validation.

Both the peptide-allele distributions, namely the ones derived from the full SARS-CoV-2 proteome and those from the structural proteins, were found to significantly deviate from an even distribution of predicted peptides as apparent in **figure 2.2AB** and reflected in the Kolmogorov–Smirnov test p-values, full proteome = 5.673e-07 and structural proteins = 1.45e-02). These results support a potential allele-specific hierarchy for SARS-CoV-2 peptide presentation.

To determine if the MHC-I binding capacity hierarchy was consistent between the full SARS-CoV-2 proteome and SARS-CoV-2 structural proteins, the relative changes in the observed peptide fraction (number of peptides assigned to an allele / total number of peptides) between the two protein sets was visualized **(Figure 2.2C)**. Six alleles demonstrated changes greater than the median peptide fraction ($\tilde{X} = 0.015$) when comparing the two protein sets. The greatest decrease in peptide fraction was observed for A*25:01 (1.52 times the median peptide fraction), and the greatest increase was seen with B*53:01 (2.38 times the median peptide fraction). Furthermore, the resulting SARS-CoV-2 structural protein peptide-allele distribution was found to be more variable than the distribution derived from the full SARS-CoV-2 proteome with a quartile coefficient of dispersion of 0.6 compared to 0.44, respectively. This indicates that peptides derived from SARS-CoV-2 structural proteins experience larger relative inter-allele binding capacity discrepancies than peptides derived from the the full SARS-CoV-2 proteome. Together, these results indicate a potential MHC-I binding capacity hierarchy that is more pronounced for SARS-CoV-2 structural proteins.

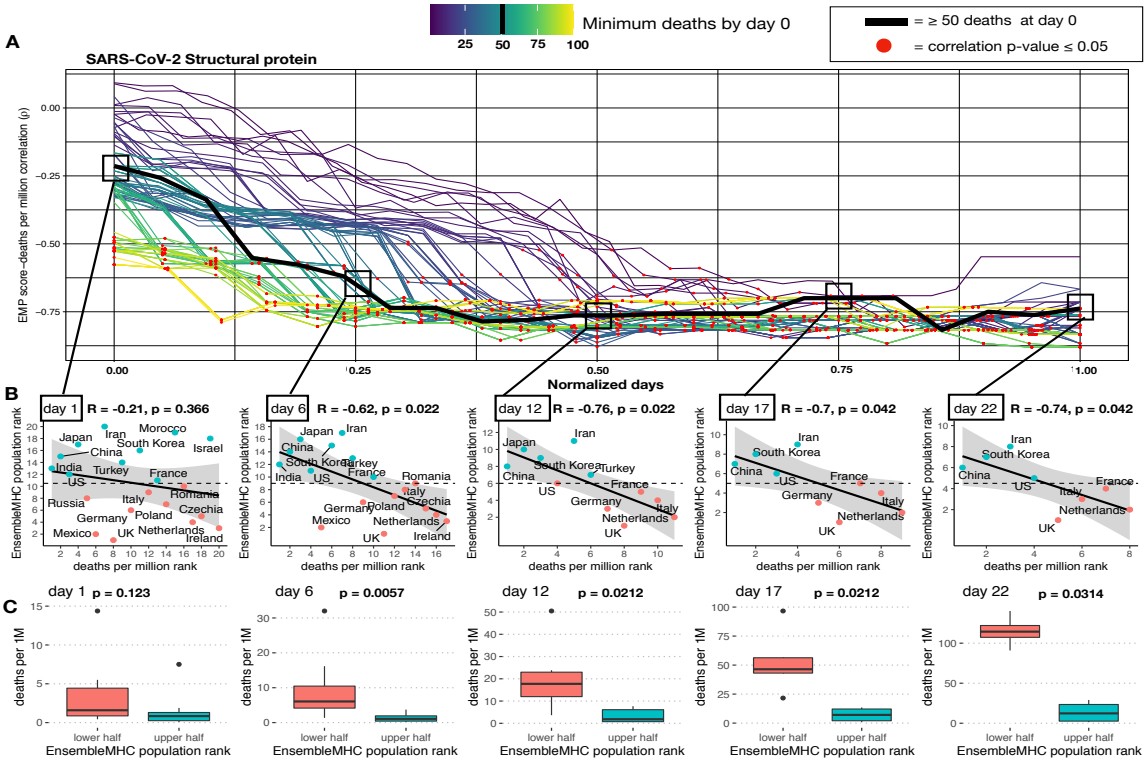

Figure 2.3: Predicted total epitope load within a population inversely correlates with mortality. A. SARS-CoV-2 structural protein-based EnsembleMHC population scores were assigned to 23 countries **(Table A.2)**, and correlated with observed mortality rate (deaths per million). The correlation coefficient is presented as a function of time. The Spearman's rank correlation coefficient between structural protein EMP score and SARS-CoV-2 mortality rate was calculated at every day following day 0 for each of the minimum death thresholds (methods for correlation and temporal control can be seen in Methods section). Correlations that were shown to be statistically significant (p-value ≤ 0.05) are indicated by a red point. **B.** The correlations between the structural protein EnsembleMHC population score (y-axis) and deaths per million (x-axis) were shown for countries meeting the 50 minimum deaths threshold at days 1, 6, 12, 17, and 22. Correlation coefficients and p-values were assigned using Spearman's rank correlation and the shaded region signifies the 95% confidence interval. Red points indicate a country that has an EnsembleMHC population rank less than the median EnsembleMHC population rank of all countries at that day, and blue points indicate a country with an EnsembleMHC population rank greater than the median EnsembleMHC population rank. **C.** The countries at each day were partitioned into a upper or lower half based on the median observed EnsembleMHC population rank. Therefore, countries with an EnsembleMHC population rank greater than the median group EnsembleMHC population score were assigned to the upper half (red), and the remaining countries were assigned to the lower half (blue). p-values were determined by Mann-Whitney U test.

2.3.3 Total Population Epitope Load Inversely Correlates with Reported Death Rates from SARS-CoV-2.

The documented importance of MHC-I peptides derived from SARS-CoV-2 structural proteins(Grifoni *et al.*, 2020), coupled with the observed MHC allele binding capacity hierarchy and the high immunogenicity rate of SARS-CoV-2 structural protein MHC-I peptides identified by EnsembleMHC (95% peptides tested *in vitro*, **Figure A.5**), prompts a potential relationship between MHC-I genotype and infection outcome. However, due to the absence of MHC genotype data for SARS-CoV-2 patients, we assessed this relationship at the population-level by correlating predicted country-specific SARS-CoV-2 binding capacity (or EnsembleMHC population score) with observed SARS-CoV-2 mortality.

EnsembleMHC population scores (EMP) were determined for 23 countries **(Table A.2)** by weighing the individual binding capacities of 52 common MHC-I alleles by their normalized endemic expression(González-Galarza *et al.*, 2015) **(Methods, Figure A.2)**. This results in every country being assigned two separate EMP scores, one calculated with respect to the 108 unique SARS-CoV-2 structural protein peptides (structural protein EMP) and the other with respect to the 658 unique peptides derived from the full SARS-CoV-2 proteome (full proteome EMP). The EMP score corresponds to the average predicted SARS-CoV-2 binding capacity of a population. Therefore, individuals in a country with a high EMP score would be expected, on average, to present more SARS-CoV-2 peptides to $CD8^+$ T cells than individuals from a country with a low EMP score. The resulting EMP scores were then correlated with observed SARS-CoV-2 mortality (deaths per million) as a function of time. Temporal variance in community spread within the cohort of countries was corrected by truncating the SARS-CoV-2 mortality data set for each country to start after a certain minimum

death threshold was met. For example, if the minimum death threshold was 50, then day 0 would be when each country reported at least 50 deaths. The number of countries included in each correlation decreases as the number of days increases due to discrepancies in the length of time that each country met a given minimum death threshold **(Table A.3)**. Therefore, the correlation between EMP score and SARS-CoV-2 mortality was only estimated at time points where there were at least eight countries. The eight country threshold was chosen because it is the minimum sample size needed to maintain sufficient power when detecting large effect sizes (ρ ¿ 0.85). The strength of the relationship between EMP score and SARS-CoV-2 mortality was determined using Spearman's rank-order correlation (for details concerning the choice of statistical tests, please refer to the Methods section). Accordingly, both EMP scores and SARS-CoV-2 mortality data were converted into ascending ranks with the lowest rank indicating the minimum value and the highest rank indicating the maximum value. For instance, a country with an EMP score rank of 1 and death per million rank of 23 would have the lowest predicted SARS-CoV-2 binding capacity and the highest level of SARS-CoV-2-related mortality. Using the described paradigm, the structural protein EMP score and the full proteome EMP score were correlated with SARS-CoV-2-related deaths per million for 23 countries.

Total predicted population SARS-CoV-2 binding capacity exhibited a strong inverse correlation with observed deaths per million. This relationship was found to be true for correlations based on the structural protein EMP **(Figure 2.3A)** and full proteome EMP **(Figure A.5)** scores with a mean effect size of -0.66 and -0.60, respectively. Significance testing of the correlations produced by both EMP scores revealed that the majority of reported correlations are statistically significant with 63% attaining a p-value of ≤ 0.05. Correlations based on the structural protein EMP score demonstrated a 24% higher proportion of statistically significant correlations compared to the full

proteome EMP score (74% vs 51%). Furthermore, correlations for EMP scores based on structural proteins produced narrower 95% confidence intervals **(Figure A.5, table A.3)**. Due to relatively low statistical power of the obtained correlations **(Figure A.6)**, the positive predictive value for each correlation **(Methods, Eq. 2.5)** was calculated. The resulting proportions of correlations with a positive predictive value of $\geq 95\%$ were similar to the observed significant p-value proportions with 62% of all measured correlations, 72% of structural protein EMP score correlations, and 52% full proteome EMP score correlations **(Figure A.5)**. The similar proportions of significant p-values and PPVs supports that an overall true association is being captured. Furthermore, analysis of similar sized peptide sets sampled from the full SARS-CoV-2 proteome revealed that the observed distinction between the correlations produced by the two protein groups are unlikely to be due to differences in peptide set sizes **(Figure A.7)**

Finally, the reported correlations did not remain after randomizing the allele assignment of predicted peptides prior to $peptide^{FDR}$ filtering **(Figure A.7)**, through the use of any individual algorithm **(Figure A.7)**. This indicates that the observed relationship is contingent on the high-confidence peptide-allele distribution produced by the EnsembleMHC prediction algorithm. Altogether, these data demonstrate that the MHC-I allele hierarchy characterized by EnsembleMHC is inversely associated with SARS-CoV-2 population mortality, and that the relationship becomes stronger when considering only the presentation of SARS-CoV-2 structural proteins.

The ability to use structural protein EMP score to identify high and low risk populations was assessed using the median minimum death threshold (50 deaths) at evenly spaced time points **(Figure 2.3A, squares)**. All correlations, with the exception of day 1, were found to be significant with an average effect size of -0.71 **(Figure 2.3B)**. Next, the countries at each day were partitioned into a high or

low group based on whether their assigned EMP score was higher or lower than the median observed EMP score **(Figure 2.3C)**. The resulting grouping demonstrated a statistically significant difference in the median deaths per million between countries with low structural protein EMP score and countries with high structural protein EMP scores. Additionally, it was observed that deaths per million increased much more rapidly in countries with low structural protein EMP scores. Taken together, these results indicate that structural protein EMP score may be useful for assessing population risk from SARS-CoV-2 infections.

In summary, we make several important observations. First, there is a strong inverse correlation between predicted population SARS-CoV-2 binding capacity and observed deaths per million. This finding suggests that outcome to SARS-CoV-2 may be tied to total epitope load. Second, the correlation between predicted epitope load and population mortality is stronger for SARS-CoV-2 structural MHC-I peptides. This suggests that $CD8^+$ T cell-mediated immune response maybe primarily driven by recognition of epitopes derived from these proteins, a finding supported by recent T cell epitope mapping of SARS-CoV-2(Grifoni *et al.*, 2020). Finally, the EnsembleMHC population score can separate countries within the considered cohort into high or low risk populations.

2.3.4 Structural Protein EMP Score Correlates Better with Population Outcome than Identified Individual Risk Factors.

Recent large scale patient studies have identified several socioeconomic and health-related factors associated with increased risk of death from SARS-CoV-2 infections(Williamson *et al.*, 2020; de Lusignan *et al.*, 2020). To delineate the relative importance of the structural protein EMP score as a SARS-CoV-2 severity descriptor, 12 additional risk factors were assessed for their ability to model population level

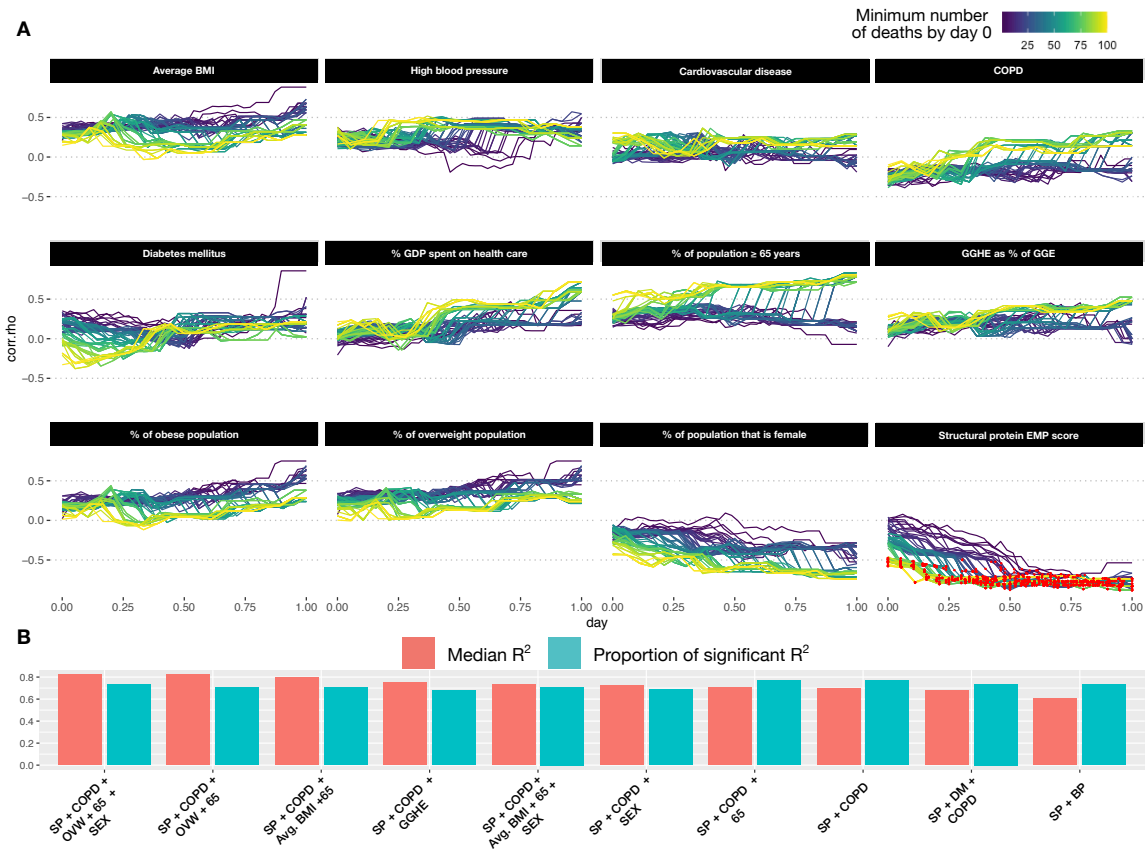

Figure 2.4: Analysis of other SARS-CoV-2 covariates with observed SARS-CoV-2 population mortality and development of an integrative model. A. 12 covariates associated with SARS-CoV-2 mortality on the individual patent level were assessed for correlation with population level mortality **(table A.4)** . The correlation of each country-level covariate was determined at each time point after a minimum death threshold was met (line color). The x-axis represents the number of days (normalized) following when a minimum death threshold was met, and the y-axis indicates the observed effect size for that covariate at a given time point. Correlations achieving statistical significance are colored with a red dot. **B.** All possible combinations of covariates were used to fit a linear model. The top 10 models, ranked by median adjusted R^2 (red bars), were identified (**B**). The proportion of regressions performed by that model that were found to be statistically significant (F-test 0.05) are represented by the blue bars.

SARS-CoV-2 outcome in 21 countries (**Table A.4**).

Overall, the structural protein EMP scores produced a significantly stronger association with population SARS-CoV-2 mortality compared to other 12 descriptors **(Figure 2.4A)**. While various effect size trends were observed, all additional covariates

failed to produce statistically significant correlations. To determine if the modeling of SARS-CoV-2 mortality rate could be improved by the combination of single socioeconomic or health-related risk factors with structural protein EMP scores, a set of linear models consisting of either a single risk factor (single feature model) or that factor combined with structural protein EMP scores (combination model) were generated for every time point across each minimum death threshold **(Methods)**. Following model generation, the adjusted coefficient of determination (R^2) and significance level of each individual model was extracted and aggregated by dependent variable **(Figure A.8)**. Single feature models were characterized by low R^2 ($\tilde{x} = -0.0262$) while combination models showed significant improvement ($\tilde{x} = .496$). Similarly, combination models demonstrated a substantially higher proportion of statistical significance **(Figure A.8B)**. To determine the set of features that produce the best fitting model, all possible combinations of explanatory factors (risk factors and structural protein EMP score) were tested. Subsequently, the top ten performing models, ranked by adjusted R^2 value, were selected for analysis **(Figure 2.4B)**. The identified models were found to be largely significant (average proportion of significant regressions = 72%) and produce strong fits to the data (average $R^2 = 0.7$).

Analysis of the dependent variables included in the top performing models revealed that all models included structural protein EMP scores followed by deaths per million due to complications from COPD (90% of models). The median model size included 3 features with a maximum of 5 features and a minimum of 2 features. The model producing the best fit (median $R^2 = 0.791$) consisted of structural protein EMP scores, gender demographics, number of deaths due to COPD complications, the proportion of the population over the age of 65, and proportion of the population that is overweight **(Figure 2.4B)**. All together, these results further indicate the robustness of the structural protein EMP score as a population level risk descriptor and identifies a

potential candidate model for predicting pandemic severity.

2.4 Discussion

In the present study, we uncover evidence supporting an association between population SARS-CoV-2 infection outcome and MHC-I genotype. In line with related work highlighting the relationship between total epitope load with HIV viral control(Rolland *et al.*, 2008), we arrive at a working model that MHC-I alleles presenting more unique SARS-CoV-2 epitopes will be associated with lower mortality due to a higher number of potential T cell targets. The SARS-CoV-2 binding capacities of 52 common MHC-I alleles were assessed using the EnsembleMHC prediction platform. These predictions identified 971 high-confidence MHC-I peptides out of a candidate pool of nearly 3.5 million. In agreement with other *in silico* studies(Nguyen *et al.*, 2020; Campbell *et al.*, 2020), the assignment of the predicted peptides to their respective MHC-I alleles revealed an uneven distribution in the number of peptides attributed to each allele. We discovered that the MHC-I peptide-allele distribution originating from the full SARS-CoV-2 proteome undergoes a notable rearrangement when considering only peptides derived from viral structural proteins. The structural protein-specific peptide-allele distribution produced a distinct hierarchy of allele binding capacities. This finding has important clinical implications as a majority of SARS-CoV-2 specific $CD8^+$ T cell response is directed towards SARS-CoV-2 structural proteins(Grifoni *et al.*, 2020). Therefore, patients who express MHC-I alleles enriched with a large potential repertoire of SARS-CoV-2 structural proteins peptides may benefit from a broader $CD8^+$ T cell immune response.

The variations in SARS-CoV-2 peptide-allele distributions were analyzed at epidemiological scale to track its impact on country-specific mortality. Each of the 23 countries were assigned a population SARS-CoV-2 binding capacity (or EnsembleMHC

population score) based on the individual binding capacities of the selected 52 MHC-I alleles weighted by their endemic population frequencies. This hierarchization revealed a strong inverse correlation between EnsembleMHC population score and observed population mortality, indicating that populations enriched with high SARS-CoV-2 binding capacity MHC-I alleles may be better protected. The correlation was shown to be stronger when calculating the EnsembleMHC population scores with respect to only structural proteins, reinforcing their relevance to viral immunity. Finally, The molecular origin of the 108 predicted peptides specific to SARS-CoV-2 structural proteins revealed that they are derived from enriched regions with a minimal predicted impact from amino acid sequence polymorphisms.

The utility of structural protein EnsembleMHC population scores was further supported by a multivariate analysis of additional SARS-CoV-2 risk factors. These results emphasized the relative robustness of structural protein EMP scores as a population risk assessment tool. Furthermore, a linear model based on the combination of structural protein EMP scores and select population-level risk factors was identified a potential candidate for a predictive model for pandemic severity. As such, the incorporation of the structural protein EMP score in more sophisticated models will likely improve epidemiological modeling of pandemic severity.

In order to achieve the highest level of accuracy in MHC-I predictions, the most up-to-date versions of each component algorithm were used. However, this meant that several of the algorithms (MHCflurry, netMHCpan-EL-4.0 and MixMHCpred) were benchmarked against subsets of mass spectrometry data that were used in the original training of these MHC-I prediction models. While this could result in an unfair weight applied to these algorithms in $peptide^{FDR}$ calculation, the individual FDRs of MHCflurry, netMHCpan-EL-4.0 and MixMHCpred were comparable to algorithms without this advantage **(Figure A.1)**. Furthermore, the peptide selection of SARS-

CoV-2 peptides was shown to be highly cooperative within EnsembleMHC **(Figure A.3)**, and individual algorithms failed to replicate the strong observed correlations between population binding capacity and observed SARS-CoV-2 mortality **(Figure A.7)**.

In the future, the presented model could be applied to predict individual T cell capacity to mount a robust SARS-CoV-2 immune response. Evolutionary divergence of patient MHC-I genotypes have shown to be predictive of response to immune checkpoint therapy in cancer and HIV(Chowell *et al.*, 2019; Arora *et al.*, 2020). However, confirmation will require large data sets associating individual patient MHC-I genotype and outcome. Additionally, future use of EnsembleMHC to design personalized T cell vaccines will require broad experimental validation of high scoring peptides, since EnsembleMHC predicts MHC-I peptides with a high probability of antigen presentation as opposed to directly predicting peptide immunogenicity. While previous work has determined that a majority of successfully presented viral MHC-I peptides are immunogenic(Croft *et al.*, 2019), there is an expectation that some presented SARS-CoV-2 MHC-I peptides will fail to produce an immune response.

The current work assessed the relative importance of the structural protein EMP score with respect to other population-level risk factors (e.g. population incidence of risk-associated commodities, healthcare infrastructure, age, sex), however, it should be noted that the impacts these risk factors on patient outcome are likely to vary significantly on a individual basis. Furthermore, other genetic determinants of severity were not considered(Cao *et al.*, 2020). Therefore, a complete understanding of the relative importance of MHC genotype and SARS-CoV-2 presentation capacity on patient outcome will require the integration individual patient genetic and clinical data.

The versatility of the proposed model will be improved by the consideration of

additional MHC-I alleles. To reduce the presence of confounding factors, EnsembleMHC was parameterized on only a subset of common MHC-I alleles that had strong existing experimental validation. While the selected MHC-I alleles are among some of the most common, personalized risk assessment will require consideration of the full patient MHC-I genotype. The continued mass spectrometry-based characterization of MHC-I peptide binding motifs will help in this regard. However, due to the large potential sequence space of the MHC-I protein, extension of this model will likely require inference of binding motifs based on MHC variant clustering.

2.5 Methods

EnsembleMHC prediction workflow

EnsembleMHC component binding and processing prediction algorithms. EnsembleMHC incorporates MHC-I binding and processing predictions from 7 publicly available algorithms: MHCflurry-affinity-1.6.0(O'Donnell *et al.*, 2020), MHCflurry-presentation-1.6.0(O'Donnell *et al.*, 2020), netMHC-4.0(Andreatta and Nielsen, 2016), netMHCpan-4.0-EL(Jurtz *et al.*, 2017), netMHCstabpan-1.0(Rasmussen *et al.*, 2016), PickPocket-1.1(Zhang *et al.*, 2009) and, MixMHCpred-2.0.2(Bassani-Sternberg *et al.*, 2017). These algorithms were chosen based on the criteria of providing a free academic license, bash command line integration, and demonstrated accuracy for predicting SARS-CoV-2 MHC-I peptides with experimentally validated binding stability(Prachar *et al.*, 2020).

Each of the selected algorithms cover components of MHC-I binding and antigen processing that roughly fall into two categories: ones based primarily on MHC-I binding affinity predictions and others that model antigen presentation. To this end, MHCflurry-affinity, netMHC, PickPocket, and netMHCstabpan predict binding affinity

based on quantitative peptide binding affinity measurements. netMHCstabpan also incorporates peptide-MHC stability measurements and PickPocket performs prediction based on binding pocket structural extrapolation. To model the effects of antigen presentation, MixMHCpred, netMHCpan-EL, and MHCflurry-presentation are trained on naturally eluted MHC-I ligands. Additionally, MHCflurry-presentation incorporates an antigen processing term.

Parameterization of EnsembleMHC using mass spectrometry data. EnsembleMHC is able to achieve high levels of precision in peptide selection through the use of allele and algorithm-specific binding affinity thresholds. These binding affinity thresholds were identified through the parameterization of each algorithm on high-quality mass spectrometry data sets(Sarkizova *et al.*, 2020). The mass spectrometry data sets used for algorithm parameterization were collected in the largest single laboratory MS-based characterization of MHC-I peptides presented by single MHC allele cell lines. These characteristics significantly reduces the number of artifacts introduced by differences in peptide isolation methods, mass spectrometry acquisition, and convolution of peptides in multiallelic cell lines. An overview of the EnsembleMHC parameterization is provided in supplemental figures **(Figure A.1)**.

Fifty-two common MHC-I alleles were selected for parameterization based on the criteria that they were characterized in Sarkizova *et al.* (2020) data sets and that all 7 component algorithms could perform peptide binding affinity predictions for that allele. Each target peptide (observed in the MS data set) was paired with 100 length-matched randomly sampled decoy peptides (not observed in the MS data set) derived from the same source proteins. If a protein was less than 100 amino acids in length, then every potential peptide from that protein was extracted.

Each of the seven algorithms were independently applied to each of the 52 allele

data sets. For each allele data set, the minimum score threshold was determined for each algorithm that recovered 50% of the allele repertoire size (the total number of target peptides observed in the MS data set for that allele). Additionally, the expected accuracy of each algorithm was assessed by calculating the observed false detection rate (the fraction of identified peptides that were decoy peptides) using the identified algorithm and allele specific scoring threshold. The parameterization process was repeated 1000 times for each allele through bootstrap sampling of half of the peptides in each single MHC allele data set. The final FDR and score threshold for each algorithm at each allele was determined by taking the median value of both quantities reported during bootstrap sampling.

Peptide confidence assessment. Peptide confidence is assigned by calculating the $peptide^{FDR}$. This quantity is defined as the product of the empirical FDRs of each individual algorithm that detected a given peptide. The $peptide^{FDR}$ is calculated using equation 1,

$$peptide^{FDR} = \prod_{i=1, i \neq ND}^{N} algorithm_i^{FDR} \tag{2.1}$$

, where N is the number of MHC-I binding and processing algorithms, ND represents an algorithm that did not detect a given peptide, and $algorithm^{FDR}$ represents the allele specific FDR of the Nth algorithm.

The $peptide^{FDR}$ represents the joint probability that all MHC-I binding and processing algorithms that detected a particular peptide did so in error, and therefore returns a probability of false detection. Unless otherwise stated, EnsembleMHC selected peptides based on the criterion of a $peptide^{FDR} \leq 5\%$.

Application of EnsembleMHC to tumor cell line data

Tumor MHC-I peptide data sets. Ten tumor samples were obtained from the **Sarkizova et al.** data sets. Tumor samples were selected for analyis if at least 50% of the expressed MHC-I alleles for that sample were included in the 52 MHC-I alleles supported by EnsembleMHC. For each data set, decoy peptides were generated in a manner identical to the method used for algorithm parameterization on single MHC allele data.

Tumor MHC-I peptide identification. Peptide identification by each algorithm was based on restrictive or permissive binding affinities thresholds. These thresholds correspond to commonly used score cutoffs for the identification of strong binders (restrictive) or all binders (permissive). These thresholds are 0.5% (percentile rank) or 50nM (IC50 value) for strong binders, and 2% (percentile rank) or 500nM (IC50 value) for all binders. Due to the lack of recommend score thresholds for MHCflurry-presentation, the raw presentation score was converted to a percentile score by histogramming the presentation scores produced by 100,000 randomly generated peptides.

Application of EnsembleMHC for the prediction of SARS-CoV-2 MHC-I peptides

SARS-CoV-2 reference sequence. MHC-I peptide predictions for the SARS-CoV-2 proteome were performed using the Wuhan-Hu-1(MN908947.3) reference sequence (https://www.ncbi.nlm.nih.gov/genbank/sars-cov-2-seqs/). All potential 8-14mer peptides (n= 67,207) were derived from the open reading frames in the reported proteome, and each peptide was evaluated by the EnsembleMHC workflow.

SARS-CoV-2 polymorphism analysis and protein structure visualizations.

Polymorphism analysis of SARS-CoV-2 structural proteins were performed using 102,148 full length protein sequences obtained from the COVIDep database(Ahmed *et al.*, 2020). Solved structures for the E (5X29) and S (6VXX) proteins (Berman *et al.*, 2000) and predicted structures for the M and N proteins(Zhang *et al.*, 2020) were visualized using VMD(Humphrey *et al.*, 1996).

Application of EnsembleMHC to determine population SARS-CoV-2 binding capacity

The peptides identified by the EnsembleMHC workflow were used to assess the SARS-CoV-2 population binding capacity by weighing individual MHC allele SARS-CoV-2 binding capacities by regional expression (for a schematic representation see **Figure A.2**).

Population-wide MHC-I frequency estimates by country. The selection of countries included in the EnsembleMHC population binding capacity assessment was based on several criteria regarding the underlying MHC-I allele data for that country **(Figure A.2)**. The MHC-I allele frequency data used in our model was obtained from the Allele Frequency Net Database (AFND)(González-Galarza *et al.*, 2015), and frequencies were aggregated by country. However, the currently available population-based MHC-I frequency data has specific limitations and variances, which we have addressed as follows:

Quality of MHC data within countries. We define MHC-typing breadth as the diversity of identified MHC-I alleles within a given country, and its depth as the ability to accurately achieve 4-digit MHC-I genotype resolution. High variability was observed in both the MHC-I genotyping breadth and depth **(Figure A.2 inset)**. Consequently, additional filter-measures were introduced to capture potential sources of variance

within the analyzed cohort of countries. The thresholds for filtering the country-wide MHC-I allele data were set based on meeting two inclusion criteria: 1) MHC genotyping of at least 1000 individuals have been performed in that population, avoiding skewing of allele frequencies due to small sample size. 2) MHC-I allele frequency data for at least 51 of the 52 (95%) MHC-I alleles for which the EnsembleMHC was parameterized to predict, ensuring full power of the EnsembleMHC workflow.

Ethnic communities within countries. In instances where the MHC-I allele frequencies would pertain to more than one community, the reported frequencies were counted towards both contributing groups. For example, the MHC-I frequency data pertaining to the Chinese minority in Germany would be factored into the population MHC-I frequencies for both China and Germany. In doing so, this treatment resolves both ancestral and demographic MHC-I allele frequencies.

Normalization of MHC allele frequency data. The focus of this work was to uncover potential differences in SARS-CoV-2 MHC-I peptide presentation dynamics induced by the 52 selected alleles within a population. Accordingly, the MHC-I allele frequency data was carefully processed in order to maintain important differences in the expression of selected alleles, while minimizing the effect of confounding variables.

The MHC-I allele frequency data for a given population was first filtered to the 52 selected alleles. These allele frequencies were then converted to the theoretical total number of copies of that allele within the population (*allele count*) following

$$allele\ count = allele_{freq} \times 2 \times n, \tag{2.2}$$

where $allele_{freq}$ is the observed allele frequency in a population and n is the population sample size for which that allele frequency was measured. The allele count is then

normalized with respect to the total allele count of selected 52 alleles within that population using the following relationship

$$norm\ allele\ count_i = \frac{allele\ count_i}{\sum_{i=1}^{52} allele\ count_i}, \tag{2.3}$$

where i is one of the 52 selected alleles. This normalization is required to overcome the potential bias towards *hidden alleles* (alleles that are either not well characterized or not supported by EnsembleMHC) as would be seen using alternative allele frequency accounting techniques (e.g. sample-weighted mean of selected allele frequencies or normalization with respect to all observed alleles within a population **(Figure A.6)**). The SARS-CoV-2 binding capacity of these *hidden alleles* cannot be accurately determined using the EnsembleMHC workflow, and therefore important potential relationships would be obscured.

EnsembleMHC population score. The predicted ability of a given population to present SARS-CoV-2 derived peptides was assessed by calculating the EnsembleMHC Population (EMP) score. After the MHC-I allele frequency data filtering steps, 23 countries were included in the analysis. The calculation of the EnsembleMHC population score is as follows

$$EMP\ score = \frac{\sum_{i=1}^{52} peptide_{frac} \times norm\ allele\ count_i}{N_{norm\ allele\ count \neq 0}}, \tag{2.4}$$

where *norm allele count* is the observed normalized allele count for a given allele in a population, $N_{norm\ allele\ count\ \neq 0}$ is the number of the 52 select alleles detected in a given population (range 51-52 alleles), and $peptide_{frac}$ is the peptide fraction or the fraction of total predicted peptides expected to be presented by that allele within the total set of predicted peptides with a $peptide^{FDR} \leq 5\%$.

Death rate-presentation correlation. The correlation between the EMP score and the observed deaths per million within the cohort of selected countries was calculated as a function of time. SARS-Cov-2 data covering the time dependent global evolution of the SARS-CoV-2 pandemic was obtained from Johns Hopkins University Center for Systems Science and Engineering(Dong *et al.*, 2020) covering the time frame of January 22nd to April 9th. The temporal variations in occurrence of community spread observed in different countries were accounted for by rescaling the time series data relative to when a certain minimum death threshold was met in a country. This analysis was performed for minimum death thresholds of 1-100 total deaths by day 0, and correlations were calculated at each day sequentially following day 0 until there were fewer than 8 countries remaining at that time point. The upper-limit of 100-deaths was chosen to ensure availability of death-rate data on at least 50% of the countries for a minimum of 7 days starting following day 0. Additionally, a steep decline in average statistical power is observed with day 1 death thresholds greater than 100 deaths **(Figure A.6)**.

The time death correlation was computed using Spearman's rank correlation coefficient (two-sided). This method was chosen due to the small sample size and non-normality of the underlying data **(Figure A.6)**. The reported correlations of EMP score and deaths per million using other correlation methods can be seen in appendix A **figure A.6**.

The low statistical power for some of the obtained correlations were addressed by calculating the Positive Predictive Value (PPV) of all correlations using the following equation(Button *et al.*, 2013)

$$PPV = \frac{1 - \beta \times R}{1 - \beta \times R + \alpha}, \tag{2.5}$$

where $1-\beta$ is the statistical power of a given correlation, R is the pre-study odds, and α is the significance level. A PPV value of $\geq 95\%$ is analogous to a p value of ≤ 0.05. Due to an unknown pre-study odd (probability that probed effect is truly non-null), R was set to 1 in the reported correlations. The proportion of reported correlations with a PPV of 95% at different R values can be seen in appendix A **figure A.6**. The significance of partitioning high risk and low risk countries based on median EMP score was determined using Mann-Whitney U-test. Significance values were corrected for multiple tests using the Benjamini-Hochberg procedure(Benjamini and Hochberg, 1995).

Sub-sampling of peptides from the Full SARS-CoV-2 proteome. 108 unique peptides, derived from the Full SARS-CoV-2 proteome and passing the 5% $peptide^{FDR}$ filter, were randomly sampled. Then, the time series EMP score - death per million correlation analysis used to generate **Figure 2.3** was applied to each sampled peptide set. The sub-sampling procedure was repeated for 1,000 iterations **(Figure A.7A)**. To quantitatively describe the similarity of the distributions, the Kullback-Leibler divergence (KLD), a measure of divergence between two probability distributions, was calculated for the correlation distribution of each sub-sample iteration relative to either the correlation distribution of the Full SARS-CoV-2 proteome or SARS-CoV-2 structural proteins **(Figure A.7B)**.

Analysis of additional SARS-CoV-2 risk factors

Additional SARS-CoV-2 risk factors. Twelve potential SARS-CoV-2 risk factors **(table A.4)** were selected for analysis. Country-specific data for each risk factor was obtained from the Global Health Observatory data repository provided by the World Health Organization (https://apps.who.int/gho/data/node.main). Countries were

selected for analysis based on the criteria of having reported data in the WHO data sets and inclusion in the set of 23 countries for which EnsembleMHC population scores were assigned **(table A.4A)**. Data regrading the total number of noncommunicable disease-related deaths (Cardiovascular disease, Chronic obstructive pulmonary disease, and Diabetes mellitus) were converted to deaths per million.

Correlation of additional risk factors with observed deaths per million. Correlation analysis of each additional factor was carried out in a similar manner to that of the EnsembleMHC population score. In short, Spearman's correlation coefficient between each individual factor and observed deaths per million was estimated as a function of time from when a specified minimum death threshold was met **(Figure 2.4)**. The significance level was set to $p \leq 0.05$ and significant PPV was set to $PPV \geq 0.95$ (eq 2.5).

Linear models of SARS-CoV-2 mortality. For the single and combination models, individual linear models were constructed for each considered death threshold as a function of time (similar to the univariate correlation analysis). Each model consisted of 1 (a single socioeconomic or health-related risk factor) or 2 (a combination of 1 risk factor and structural protein EMP score) dependent variables and deaths per million as the independent variable. The adjusted R^2 value and statistical significance of the model (F-test) were then extracted from each individual model and aggregated by dependent variable **(Figure 2.4, Figure A.8)**.

The best performing models were determined by assessing all possible combinations of factors including structural protein EMP score. This resulted in the consideration of 4,083 different linear models. The top performing models were then selected by ranking each model by median adjusted R^2.

Code and data availability.

All data analysis and statistical tests were performed using the R Statistical Computing Environment v.3.6.0 (http://www.r-project.org). Data sets and example code are available at https://github.com/eawilson-CompBio/EnsembleMHC-Covid.git

Chapter 3

HLA-INCEPTION: A STRUCTURE-BASED MHC-I BINDING MOTIF PREDICTION ALGORITHM

3.1 Abstract

The ability to accurately identify peptide ligands for a given major histocompatibility complex class I (MHC-I) molecule has immense value for targeted anticancer and antiviral therapeutics. However, the highly polymorphic nature of the MHC-I protein makes universal prediction of peptide ligands challenging due to lack of experimental data describing most MHC-I variants. To address this challenge, we have developed a deep convolutional neural network, HLA-Inception, capable of predicting MHC-I peptide binding motifs using biophysical properties of the MHC-I binding pocket. By approaching this problem from a 3-dimensional perspective, we can fully consider the impact of sidechain arrangement and topology on peptide binding, a feature not inherently captured by the popular protein sequence-based MHC-I prediction methods. Through a combination of molecular modeling and simulation, 5,821 MHC-I alleles were modeled, providing extensive coverage of all human populations. The topology and interaction forces within the MHC-I binding pocket were accounted for by solving the electrostatic potential near the surface of the protein. HLA-Inception was then trained on all MHC-I alleles with known peptide binding motifs and applied to the full set of MHC-I models. Predicted peptide binding motifs fell into distinct and well-defined clusters, which maintained disease associations. We demonstrate that the predicted MHC-I binding motifs can be used for MHC-I ligand prediction, and are more generalizable than sequence-based methods. The scores generated by HLA-Inception

are strongly correlated with quantitative MHC-I binding data, indicating predicted peptides can be ranked. Finally, we show that HLA-inception has a higher precision than the current state-of-the-art models when predicting naturally presented MHC-I ligands.

3.2 Introduction

The major histocompatibility complex I (MHC-I) protein plays an integral role in permitting $CD8^+$ T cell based immune surveillance of host cells. As such, this protein complex, and related pathways, have been directly implicated in the successful viral clearance and tumor rejection. The MHC-I protein drives this process by presenting endogenous protein fragments on the cell surface for interaction with $CD8^+$ T cells via T cell receptors(Rock *et al.*, 2016). The MHC-I processing and presentation pathway has the ability to present peptides from virtually any expressed cytosolic protein, and as such, grants an unprecedented view into intracellular protein production. The MHC-I protein cannocially binds peptides that are 8-14 amino acids(Trolle *et al.*, 2016) in length with binding being mainly driven by key interactions between peptide ligand and binding pockets within MHC-I binding cleft(Garrett *et al.*, 1989; Nguyen *et al.*, 2021). The peptide ligand residues that primarily drive this interaction (typically position 2 and the C-terminal residue of binding peptides) are often referred to anchor positions, and have highly conserved amino acids identities in peptides that are capable of stable binding. Therefore, peptides that bind to a given MHC-I variant will typically exhibit an overall peptide binding motif. Once presented on the cell surface, the peptide-loaded MHC-I complex is able to interact with T cell receptors expressed on circulating $CD8^+$ T cells. Stably bound peptide ligands that are sufficiently diverged from naturally presented host peptides, such as peptides derived from viral or mutated proteins, can trigger an immune reaction(Sundberg *et al.*, 2007). The cell specific

nature of MHC-I driven immune responses is partly responsible for some of the most remarkable anticancer immunotherapies(Leidner *et al.*, 2022; Zacharakis *et al.*, 2018). However, such therapies rely on the ability to identify peptide targets from an antigen of interest. While recent developments of high throughput techniques to rapidly solve tumor-associated immunopeptidomes have allowed for experimental verification of tumor associated MHC-I targets(Chong *et al.*, 2018; Lan Zhang *et al.*, 2020), the costs are still prohibitive for any clinical application. Therefore, *in silico* methods are commonly used to identify potential MHC-I peptides from antigens of interest(O'Donnell *et al.*, 2020; Bassani-Sternberg *et al.*, 2017; Lan Zhang *et al.*, 2020; Reynisson *et al.*, 2020).

The major obstacle to MHC-I prediction is the significant genetic diversity of the MHC-I protein in the human population. The MHC-I protein is one of the most polymorphic proteins in the human genome with over 24,000 known sequence variations(Robinson *et al.*, 2020). This can prove to be a formidable challenge as even single point mutations in the binding pocket can lead to altered MHC-I peptide binding motifs(Parham *et al.*, 2018). *In vitro* binding experiments(Peters *et al.*, 2006) and mass spectrometry-based profiling of cell lines(Lan Zhang *et al.*, 2020) or tumors(Bassani-Sternberg *et al.*, 2016) have provided crucial data for training MHC-I peptide prediction algorithms. However, due to both the size of the MHC allele space and the costs associated with such experiments, there are currently only public datasets describing the binding motifs of 205 alleles(Vita *et al.*, 2019). The sequence diversity of the MHC-I protein was originally addressed through the definition of MHC-I "supertypes" or clusters of MHC-I alleles assumed to produce similar MHC-I binding motifs(Sidney *et al.*, 2008). Recently, this diversity has been tackled through the development of machine learning algorithms that perform pan-allele predictions(O'Donnell *et al.*, 2020; Bassani-Sternberg *et al.*, 2017; Lan Zhang

et al., 2020; Reynisson *et al.*, 2020). In order to extrapolate to all alleles, these methods rely on the amino acid identity of key positions within the MHC-I binding pocket to infer the binding specificity of unresolved alleles(Nielsen *et al.*, 2007). While sequence-based methods are effective, they are sensitive to sequence variations. This sensitivity can be problematic as changes in sequence are not necessarily commensurate to significant changes in the biochemical properties. Attempts to address this issue have involved numerically encoding amino acids with biochecmial properties such as hydrophobicity, or the BLOSUM62 substitution matrix scores(Reynisson *et al.*, 2020; Lan Zhang *et al.*, 2020). While numerically encoding residues better quantify physical changes upon substitution, they are unable to accurately account for the impact of mutations in binding pocket topology and sidechain orientation. Therefore, we hypothesize that developing an MHC-I motif and binding prediction algorithm that learns the underlying biophysics of peptide binding will better map the impacts of polymorphisms to peptide binding motifs, resulting in better generalization of predictions to significantly diverged alleles.

In what follows, we describe a fully structure-based MHC-I peptide binding motif and ligand prediction algorithm. This was accomplished by training an inception-based convolutional neural network (CNN) that is able to predict MHC-I binding motifs based on the electrostatic properties of the MHC-I binding pocket. We found that the generated motifs formed 12 well-defined clusters, and that euclidean distances between motifs strongly correlated variations in binding motifs and immunopeptidome diversity. The distances between predicted motifs could also recapitulate known allele associations with HIV infection control. The predicted binding motifs were then used for peptide ligand prediction, which outperformed sequence-based models when generalizing to unseen data. Amazingly, this was accomplished despite never being explicitly trained on peptide sequence. Finally, we show that HLA-inception is

more precise when identifying naturally presented MHC-I peptides when compared to current state-of-the-art sequence models.

3.3 Results

In what follows, we first demonstrate whether the electrostatic potential of the MHC-I binding pockets can describe MHC-I binding motifs. Using computational models of 5,821 MHC-I binding pockets (4,464 unique), an Inception-based convolutional neural network model was trained to predict MHC-I binding motifs based on electrostatic potentials alone. Second, we find that the predicted MHC-I binding motifs form clusters of alleles with similar biochemical properties, and these motifs distinguish between MHC-I alleles with known disease association. Finally, we show that the predicted MHC-I binding motifs can be applied to perform pan-allele MHC-I peptide prediction from protein targets.

3.3.1 Electrostatic Potentials Track Peptide Binding Motif Variation

Electrostatic features are shown to be predictive of protein function, and have been applied to determine the charge environment of protein structures(Santiveri *et al.*, 2020). In view of the electrostatic interactions between peptide ligands and MHC proteins(Collins *et al.*, 1994), the MHC-I molecule is an ideal target for such analysis. A barrier to performing this analysis is the relatively small number of experimentally solved MHC-I structures. Nonetheless, the high structural homology shared between even significantly diverged MHC-I alleles indicates that MHC-I protein structure could be accurately predicted using informatics-based models. **Figure 3.1** demonstrates an overview of the multiresolution modeling pipeline, i.e. *sequence* $\rightarrow$ *ensemble* $\rightarrow$ *electrostatics*, used to generate structures of 5,821 structures of the MHC-I binding pocket **(Methods)**.

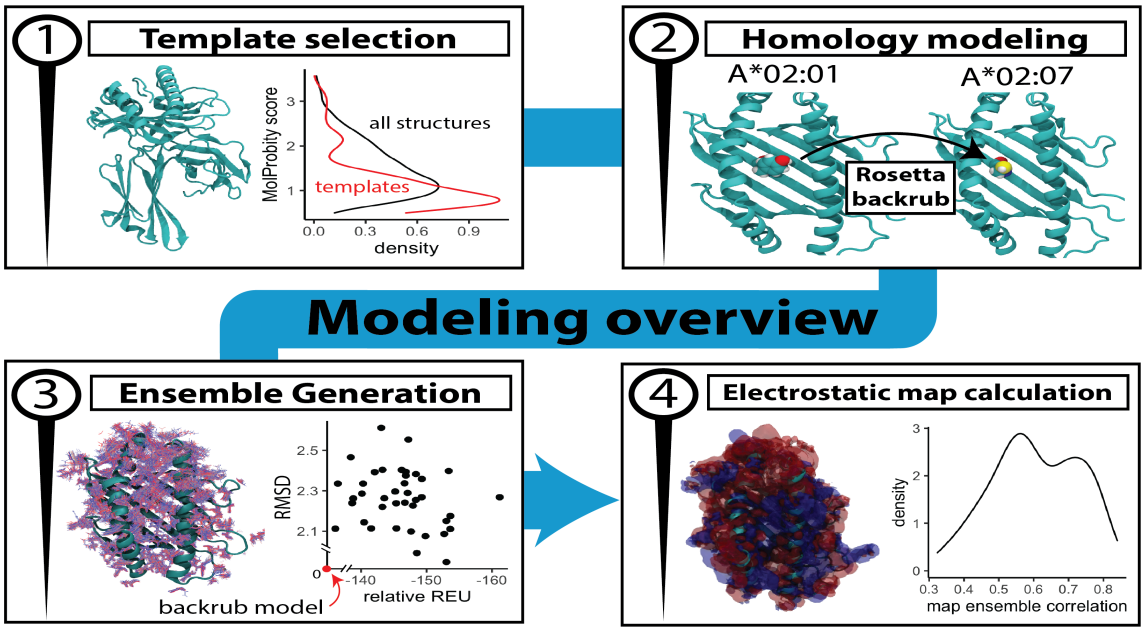

Figure 3.1: Building models of 5,821 MHC-I binding pockets. The above figure illustrates the MHC-I modeling process. **(1.)** MHC-I structural templates are selected based on molprobity score, a measure of structure quality. **(2.)** MHC-I variants without existing structures are a modeled using the best aligning template structure. **(3.)** An ensemble of 40 binding pockets are generated via sidechain rotamer sampling using the Rosetta simulation software. **(4.)** The electrostatic potential of the MHC-I binding pocket is calculated for each ensemble member using APBS.

Following model construction and calculation of binding pocket electrostatic environment, we use a metric to monitor MHC-I allele divergence defined by computing the euclidean distance between two sets of binding pocket electrostatics potentials called the electrostatic potential distance (EPD). Two more sequence-dependent methods, namely hamming distance and BLOSUM80 alignment, were also calculated for inter-allele comparisons **(Figure 3.2A, Methods)**. Using all three metrics, the pairwise inter-alleleic distances were computed for 133 MHC-I molecules with at least 50 known peptide ligands. These distances were then correlated with the pairwise binding motif variations, measured using Kullback–Leibler divergence or KLD **(Methods)**. Illustrated in **figure 3.2A**, Hamming distances produced a correlation coefficient of 0.22, while the more biochemically and structurally aware BLOSUM80 alignment

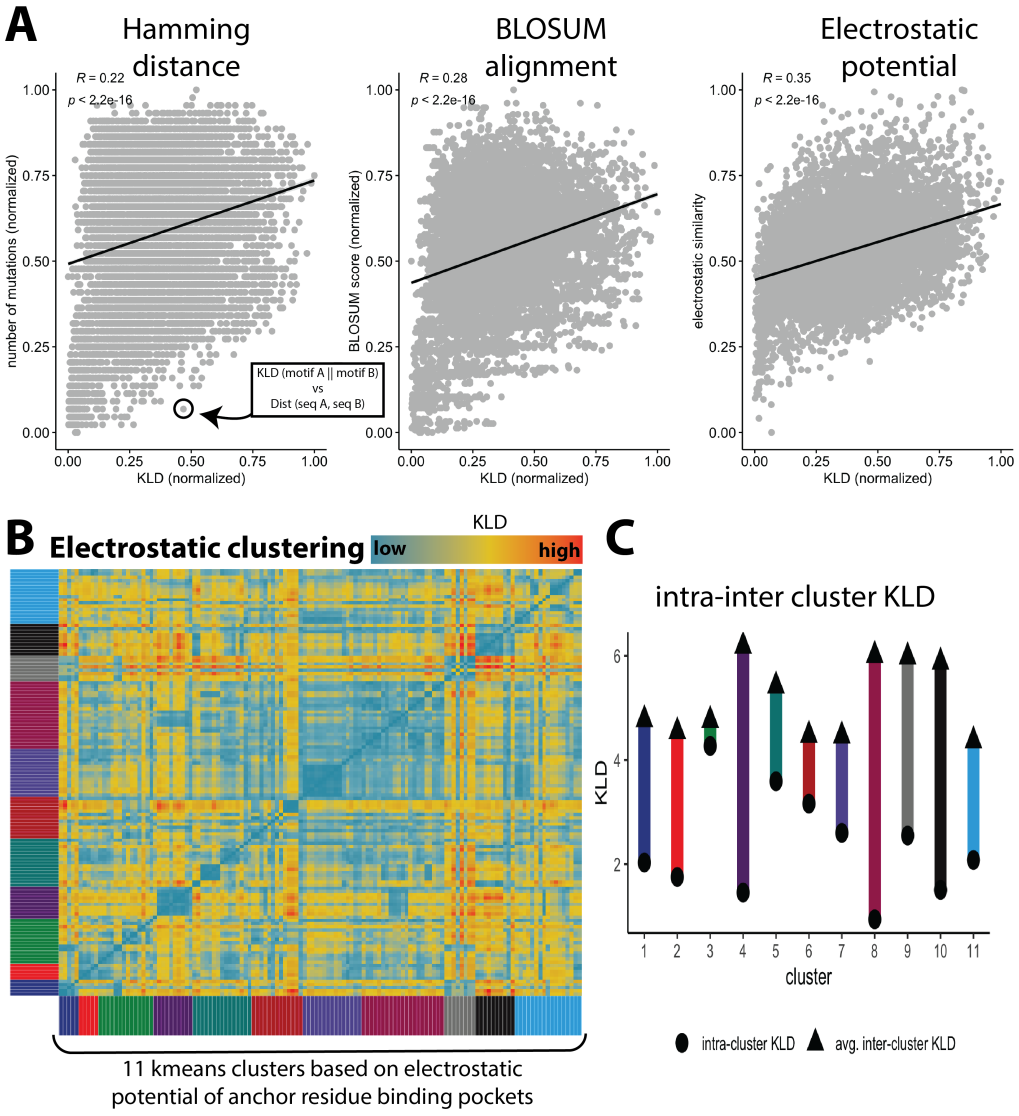

Figure 3.2: Electrostatic potential configurational space better captures MHC-I binding motif variation. A. The correlation between different binding pocket variation measurement methods and binding motif KLD for 133 MHC-I alleles with known binding motifs. Each dot indicates a pairwise comparison between two alleles with the binding pocket distance metric defined by the header of plot. **B.** Spherical regions of electrostatic potential corresponding to N- and C-terminal anchor binding pockets were extracted, and the k-means clustering method was applied to find 11 different groups. Alleles were then listed in order of their respective k-means clusters (x- and y-axis) with each block indicating one of the 133 MHC-I alleles with known binding motifs and colors representing the cluster. The fill color within the plot indicates the pairwise KLD between alleles, with blue indicating a low KLD (more similar binding motif) and red indicating a high KLD (diverged motif). **C.** The KLD within each cluster identified in **B** was compared to the pairwise KLD of every allele outside that cluster.

improved the correlation coefficient to 0.28. Finally, we masked the electrostatic potentials near the N- and C-terminal binding pockets, and calculated the EPD. The correlation between allele KLD and EPD showed an improvement of at least 20% over the sequence-centric methods. Overall, our EPD analysis suggests that the incorporation of information beyond simple sequence may be useful for MHC-I motif prediction.

Next, the structure of the inter-allele electrostatic relationships was investigated. K-means clustering was performed on the extracted electrostatic potentials masks, revealing 11 unique clusters **(Figure 3.2B, Methods)**. We found that MHC-I alleles with similar binding motifs were generally clustered together, with the average KLD within clusters being less than the KLD between clusters **(Figure 3.2C)**. Taken together, this indicates that MHC-I alleles can be grouped by electrostatic potential into clusters of similar binding motifs.

In summary, using a multiresolution modeling approach, we generated 5,281 structures of the MHC-I protein. Next, we show that electrostatic potentials extracted from these predicted structures better predict changes to MHC-I binding motifs. Finally, we show that electrostatic potentials can be used to group MHC-I allele into clusters of similar binding motifs.

3.3.2 Identifying MHC-I Binding Motif Complementarity with Inception Model Trained on Electrostatic Features

While the EPD values successfully tracked with variations in MHC-I binding motifs, it is inherently limited to experimental resolved alleles. Therefore, we developed a deep learning model, named HLA-inception, which predicts MHC-I binding motifs from predicted or experimentally determined MHC-I binding pocket structures **(Figure 3.3, Methods)**. This approach works by segmenting the MHC-I binding pocket

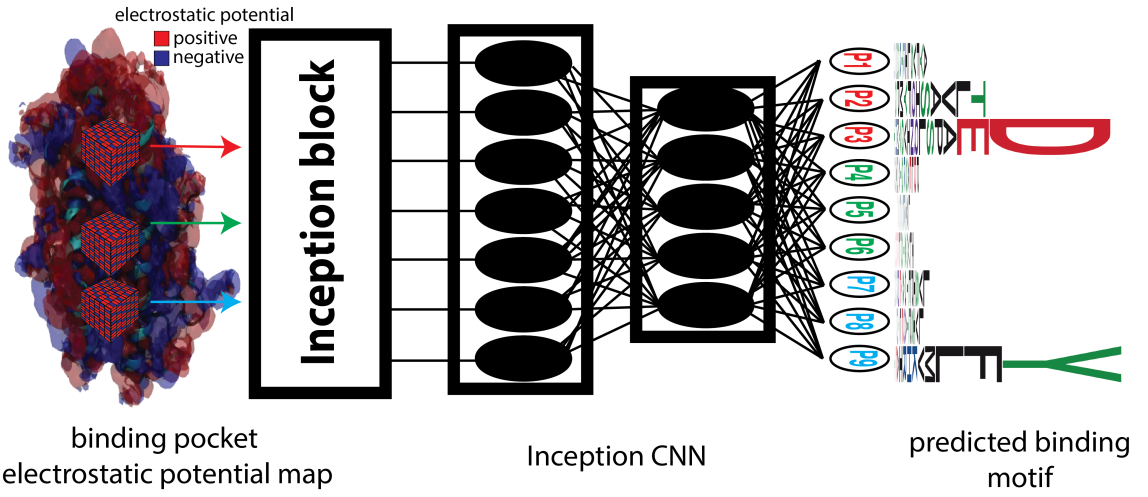

Figure 3.3: Learning peptide-protein complementarity. The above figures demonstrate a schematic of the inception-based CNN that was trained on MHC-I binding pocket electrostatic potentials in order to predict binding motifs.

electrostatic potential grid into a number of sections; here we have chosen three equal sized segments – the region corresponding to N-terminal binding pocket, TCR contact region, and C-terminal binding pocket. As described in methods **(HLA-Inception model)**, the N-terminal region was used to predict amino acid motifs for peptide positions 1 - 3, the TCR contact domain was used to predict peptides positions 4-6, and the C-terminal domain was used to predict peptide positions 7-9. The amino acid residue distribution at each peptide location was predicted individually by passing the corresponding block through HLA-Inception with the output layer being the empirical distribution at that position for a given allele. Hyperparameter tuning was conducted in order to optimize the model. Due to the overall importance of positions 2 and 9 to peptide binding, hyperparameter tuning was focused on these positions. We found that the testing KLD, with optimal parameters, converged at 0.4384 for position 2 and 0.1228 for position 9 **(Figure B.1)**. These parameter values were then used to train the model corresponding to each position.

Following model training with 5,320 unique maps (40 ensemble members x 133 alleles), The HLA-Inception model was applied to the average electrostatic potential

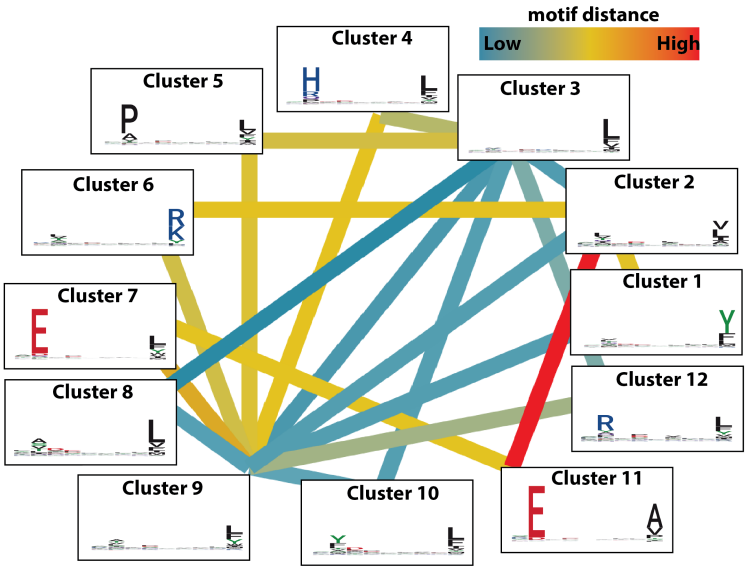

Figure 3.4: k-means clustering of predicted motifs. k-means clustering was applied to the predicted motifs of all 5,821 alleles. This resulted in the identification of 12 clusters. The binding motif of each cluster is represented by logo plots of the average binding motif of all alleles within that cluster. The average of all pairwise euclidean distances between clusters was calculated, and each node in the graph was connected to the two closest (most similar) allele clusters. Edge color indicates the relative magnitude of the average distances.

maps from all generated MHC-I structures. Using the predicted binding motifs, k-means clustering was used to create MHC-I "supertypes". We found that the predicted motifs could be classified into 12 different clusters **(Figure 3.4)**. These clusters were largely defined by biochemical characteristics of the N- and C- terminal anchors. The identified anchor classes were as follows: positively charge (R,K), bulky positive charge (H), negatively charged (E,D), branched hydrophobic(L,V,M), aromatics (Y,F,W), and small hydrophobic (A,P).

In brief, an Inception-based CNN, called HLA-Inception, was trained on electrostatic features of the MHC-I binding pocket. We found that predicted MHC-I binding motifs formed general clusters analogous to traditional MHC-I "supertypes", indicating a potential biochemical basis for observed results from sequence-based clustering.

3.3.3 Exploration of Predicted Binding Motifs

Using the HLA-Inception generated binding motifs, we explored whether important allele-specific peptide binding characteristics were being captured. Success in this instance was defined by the overall correlation between the pairwise L2-norm (euclidean distance) of predicted motifs, a quantity we refer to as 'motif distance', and empirical motif KLDs. L2-norm was chosen due to its numerical robustness and usage in similar approaches(Bassani-Sternberg *et al.*, 2017). We investigated how inter-allele distances, in the context of a full MHC-I genotype (genotype distance), impact the diversity of the immunopeptidome. Finally, we test whether motif distance can be used to recapitulate known MHC-I allele disease associations.

First, the motif distances were calculated for the set of 133 MHC-I alleles with known binding motifs and correlated with pairwise KLD. Pairwise motif distance was found to be strongly correlated with observed KLD **(Figure 3.5A)** with an overall correlation of 0.83. This result indicates that the predicted motifs are capturing empirical binding motif variations. Previous work has demonstrated that of MHC-I genotype diversity improves response to immune checkpoint blockade (Chowell *et al.*, 2019). Therefore, we determined if genotype distance could be predictive of immunopeptidome diversity. The average pairwise distances between all 6 MHC-I alleles (genotype distance) of multiallelic cell lines were compared to the average BLOSUM62 alignment score of anchor residues of peptides recovered from MS-based immunopeptidomic experiments of those cell lines(O'Donnell *et al.*, 2020). We found that the average gentoype distance was significantly correlated with BLOSUM62 alignment score, indicating that genotype distance is predictive of immunopeptidome diversity **(Figure 3.5B)**. Overall, we found that the predicted motifs were correlated empirical KLDs, and that the distance between motifs can be predictive of immunopeptidome

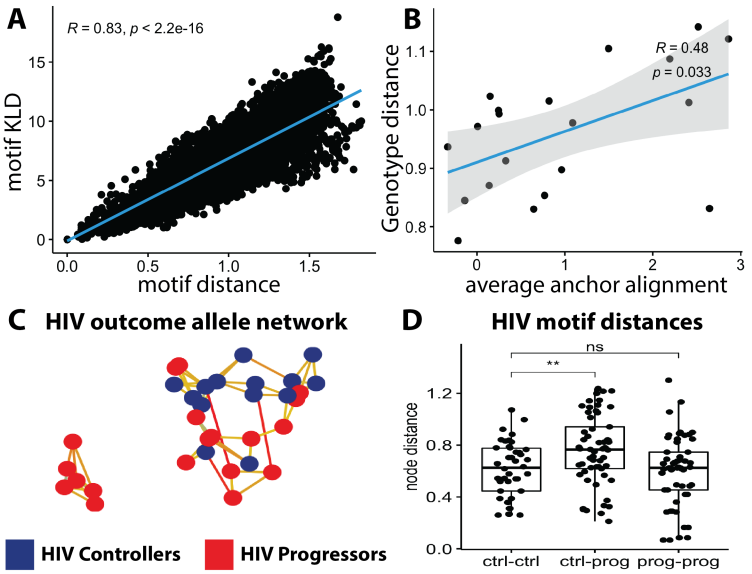

Figure 3.5: Quantifying inter-allele motif distances. A. The pairwise motif distances for all 133 allele were calculated and correlated with inter-allele KLD. both Higher motif distance and KLD indicate greater divergence between alleles. Each point represents a comparison of two motifs, and therefore associated with a motif distance and an empirical binding motif KLD. The blue line is a linear fit to the data. **B.** The average motif distance for cell lines with full MHC-I genotype resolution was correlated with the average BLOSUM62 alignment score of anchor residues of MS-detected MHC-I peptides eluted from each cell lines. **C.** A force-directed graph was generated of alleles associated with HIV outcome. alleles associated with HIV control are colored blue while alleles associated with HIV progress are colored red. Each node was connected to the 3 nearest nodes. Nodes in close proximity to each other indicate more similar predicted peptide binding motifs. **D.** The pairwise motif distances of allele associated with HIV outcome were calulated. The x-axis indicates the relative groups to which distance was measured. ctrl-ctrl indicates the all pairwise distances between alleles associated with HIV control while prog-prog is the distances between all alleles associated with HIV progression. crtl-prog is the distance between control allele and progression alleles.

diversity.

Finally, we assessed whether motif distance captures known trends in MHC-I disease associations, specifically, for alleles associated with HIV viral control (International HIV Controllers Study *et al.*, 2010). The motif distance between alleles associated with higher probability of being an HIV controller (individuals with these alleles progress to AIDS slower than those without) and HIV progresser (individual progress to AIDS faster than those without) was calculated. To visualize potential clusters of HIV associated alleles, a force directed graph(Fruchterman and Reingold, 1991) was generated, where each node is a different HIV associated allele. Edges between the allelic nodes were defined by connecting each allele to the three most similar alleles, as ascertained by motif distance. Illustrated in **figure 3.5C**, the algorithm resolved distinct clusters. When all pairwise motif distances between HIV associated MHC-I alleles were calculated, the distances between controlling and progressing alleles were found to be higher than those within each respective group (Figure 3.5D). In light of these results, It is inferred that the predicted binding motifs are persevering known allele associations with HIV outcome.

3.3.4 Electrostatics-driven Pan-Allele MHC-I Peptide Ligand Prediction

The prediction of the peptide ligands for a target MHC-I protein has significant clinical use for T cell-based immunotherapies. MHC-I binding motifs have previously been used to predict the ligands(Rammensee *et al.*, 1999). In a similar effort, the binding motifs generated by HLA-Inception were used to identify potential MHC-I ligands via peptide scoring by position-weighted matrix (PWM) scores **(Methods)**.

Typically, pan-allele prediction algorithms are validated by performing 'leave-one-out' cross validation analysis in which a target allele is removed from the training set, and then the remaining data is utilized for testing prediction accuracy. However, such

validation approaches fail to account for the existence of highly homologous MHC-I alleles still contained within the training set. The inclusion of homologous alleles has the potential to artificially boost algorithm performance. A more rigorous test of pan-allele predictive properties can be performed by ensuring that highly homologous alleles are removed prior to training and collectively tested, hence gaining a better assessment of algorithm generalizability. To that end, the binding pocket sequences for the 133 allele set were clustered using BLOSUM62 alignments, where each allele was assigned to a cluster of alleles with similar amino acid sequences **(Methods)**. In order to appropriately benchmark the performance of HLA-Inception peptide predictions with a sequence-based approach, a deep and densely connected neural network trained on BLOSUM62 encoded peptide and key binding pocket resides was built **(Methods)**. Using the MHC-I binding pocket sequence clusters, 'leave-one-cluster-out' analysis **(Methods)** was performed by using each algorithm to predict MHC-I peptides. We found that the MCC of the HLA-Inception produced a median MCC of 0.72 (IQR: 0.59-0.79), while the sequence-based model produced a median MCC of 0.52 (range 0.38-0.68). Remarkably, the HLA-Inception model produced an 38% improvement over a sequence-based prediction method.

Next, peptides with known quantitative values, namely peptide binding affinity (IC50) and MHC-I stability (minutes), were ranked based on the aforementioned PWM score. We found that PWM scores were strongly associated with quantitative values **(Figure 3.6)**. PWM scores had an absolute correlation coefficient of 0.62 with MHC-I stability data, and a 0.65 correlation with MHC-I affinity. This result is particularly notable as it suggests that the most probable binders determined from our algorithm are also found to be strong binders, even though no peptide-protein interaction data was used to train the model. The 60-65% correlation indicates that the inception network has learnt to capture the strengths of their molecular interaction

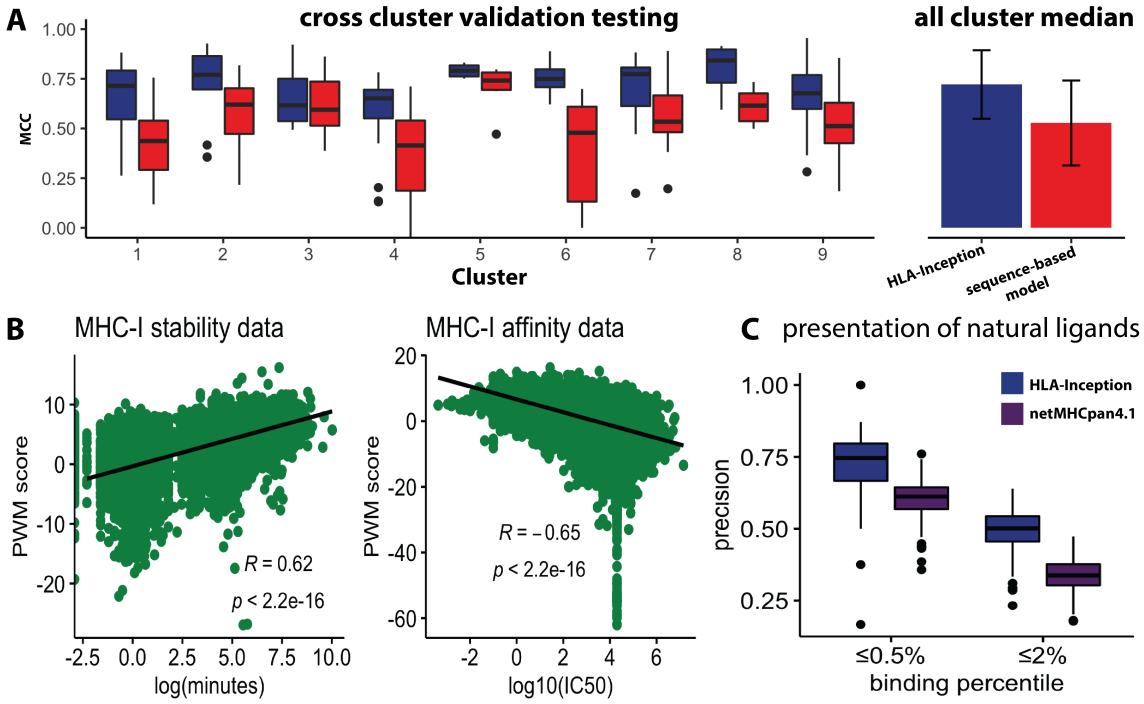

Figure 3.6: Pan-allele peptide prediction with HLA-Inception. A. Matthew's correlation coefficient (mcc) was calculated with respect to peptide prediction using HLA-Inception (blue) and a sequence-based model (red). The x-axis for the box plot (left) indicates the allele cluster that was left out during algorithm training, and then subsequently used for testing. The bar blot (right) shows the median mcc value across all clusters with the lines indicating ±1 standard deviation. **B.** The scatters plots show the correlation between PWM score (y-axis) and MHC-I stability (x-axis;left) or MHC-I binding affinity (x-axis;left). Each point indicates a different peptide that was tested. **C.** The precision (y-axis) of HLA-Inception(blue) or netMHCpan-4.1(purple) in the recovery of naturally presented MHC-I peptides is shown by a box plot. The x-axis indicates the binding score percentile threshold used to select peptides (all peptides at that threshold or better were classified as binders).

with the MHC-I binding pocket, leveraging only information of the MHC- binding pocket environment and overall binding motif of the peptides. Such a training in binding complementarity, gives a biophysical confirmation that the primary peptide anchor positions are the most interactive, and therefore most stabilizing, to peptide binding. Overall, we discover that PWM score offers a physical basis for seeking interaction signatures of the peptides, and designing crucial features for peptide target selection.

Finally, we determined the precision of HLA-Inception on the prediction of naturally presented peptide ligands. The peptides used for this analysis were extracted from a dataset of mass spectrometry-detected MHC-I peptides eluted from monoalleleic cell lines(O'Donnell *et al.*, 2020). The single allele nature of these cell lines simplifies analysis, as each peptide can confidently be linked to a single MHC-I allele. The prediction of naturally presented ligands is particularly important for the identification of T cell targets for cancer immunotherapies. However, this necessitates the consideration of many proteins, which can lead to high false positive rates.The performance of HLA-Inception-based predictions were compared to the current state-of-the-art for pan-allele peptide prediction, netMHCpan-4.1(Reynisson *et al.*, 2020). Peptide selection was based on two commonly used score thresholds, 0.5% and 2%, where peptides with this score or lower can be considered to bind stronger than 99.5% and 98% of all possible peptides respectively. We found that HLA-Inception-based prediction achieved a median precision of 0.74 and 0.5 for the 0.5% and 2% threshold, respectively **(Figure 3.6)**. While netMHCpan achieved significantly lower median precision values (median 0.5% = 0.6; 2% = 0.33). This result is significant, as netMHCpan was explicitly trained on data within this testing set, whereas, HLA-Inception was only trained on an overall motif.

3.4 Discussion

We discovered that simple pairwise euclidean distance calculations of voxelized electrostatic potential regions near the MHC-I B and F binding pockets produced a stronger correlation with observed variations in MHC-I binding motifs than a purely sequence-centric analysis. Motivated by these results, an inception-based convolutional neural network, coined HLA-Inception, was trained on segments of the binding pocket electrostatic potential grid to predict MHC-I binding motifs. The predicted motifs

were found to cluster analogously to previously identified MHC-I "supertypes". More importantly, the generated motifs were predictive of immunopeptidome diversity when applied in multiallelic context, capable of recapitulating known allele associations to HIV outcome, and precise when used for MHC-I ligand prediction.

There are several profound advantages to approaching the prediction of MHC-I binding motifs, and subsequently peptide ligands, from an electrostatic lens. First, we are able to learn one of the biophysical rules that dictates peptide binding. Sequence-centric MHC-I prediction methods do not explicitly state the underlying forces that drive peptide binding, but rather a sequence configuration that leads to a particular binding motif. This understandably makes such approaches highly sensitive to sequence variation, which is problematic given the polymorphic nature of the MHC-I protein. In contrast, by training our model directly on the underlying forces, HLA-Inception is able to learn a measurable quantity which drives peptide binding. This physics formulation enhances interpretability of binding predictions, which is immediately evident by the results of the 'leave-one-cluster-out' analysis. Another advantage of the shift to electrostatic modeling is a striking reduction of the experimental search space. Because electrostatic potential is a degenerative property, as many different sequence configurations can produce similar local electrostatic environments, the number of MHC-I alleles with unknown binding motifs that require experimental validation is significantly diminished. This makes universal coverage of all human MHC-I binding motifs an experimentally tractable goal, and therefore, opens new doors to broadened applications of T cell based immunotherapies. Finally, HLA-inception embodies a methodological advance in computational biology. Arguably, we are the first to determine that molecular interaction strength can be hierarchized by using knowledge of only the electrostatic environment and binding sequence, without any additional geometric information.

There are some caveats to our approach, namely the compositional bias of the training set, the use of nonameric peptide binding motifs, and the use of predicted MHC-I models. Like most machine learning models, predictions are biased by the composition of the training set. In cases where the training set provides a good sampling of the total input space, predictions have a high likelihood of accuracy. Conversely, in cases where isolated populations, not captured by the training set, exist then predictions are unlikely to be accurate for these groups. The immense number of MHC-I variants make this a valid concern for any machine learning approach to MHC-I ligand prediction, and is not specific to our model. However, as outlined above, we expect that our approach will be less affected by this problem due to the learning of the underlying physical nature of peptide binding. To combat this problem, future work can be focused on the experimental resolution of MHC-I alleles with predicted electrostatic environments that fall outside the currently known distribution. Next, predictions were only done with respect to nonameric peptide binding motifs. This decision was made mainly due to the major of observed peptides being 9 amino acids in length. This translated into high resolution binding motifs. However, there is a small but relevant population of peptides at different lengths. We used an approach analogous to NN-align(Reynisson *et al.*, 2020) to extend the 9mer motifs to peptides of length 8-11. We observed reasonably high accuracy for peptides of these lengths, which cover 95% of all observed MHC-I peptides **(Figure B.2)**. Finally, the PWM method used for peptide prediction does not explicitly take into account the effects of neighboring residues with a given peptide. While such effects have been shown to occur, they are generally rare.

In future work, the method of learning the electrostatic environment to perform motif prediction is readily applicable to numerous application, including MHC-II, antibody-antigen binding, and TCR-MHC binding.

3.5 Methods

Data

MHC-I protein sequences. MHC-I sequences were obtained from the IGMT-HLA database (Robinson *et al.*, 2000)(accessed 7/2021). Only MHC-I alleles with resolution of the full canonical lengths were considered (HLA-A: 265 amino acids, HLA-B: 362 amino acids, HLA-C: 366 amino acids), resulting in the consideration of 5,821 sequences. After selection of full length sequences, the binding pocket was extracted (residues 25-210).

MHC-I peptide data. MHC-I peptide data was extracted from the IEDB database(Vita *et al.*, 2019). The data was initially filtered to select only peptides that were 9 amino acids in length and had four digit resolution. From this dataset, a separate quantitative data set was generated by filtering peptides labeled with experimental IC50 values or complex stability values.

Monoalleleic and multiallelic immunopeptidome data. For analysis centered on the recovery of naturally presented MHC-I ligands, the training datasets generated by O'Donnell *et al.* (2020) were used. The mutialleleic dataset was further filtered for data with full genotype (i.e all 6 MHC-I allele) information.

MHC-I binding pocket modeling

MHC-I structures were identified using the IEDB database (Vita *et al.*, 2019) and downloaded from the RCSB Protein Data Bank(Rose *et al.*, 2017). Molprobity(Chen *et al.*, 2010) was then used to score 606 MHC-I crystal structures, with lower scores indicating higher resolution models. Out of the 606, the best structures for each

MHC-I allele were selected, resulting in 50 unique MHC-I template structures. The peptide was removed from each template model, and the structure was truncated to only contain the binding pocket (reside 25-210 of the amino acid sequence). The templates were then minimized using the default Rosetta score function(Park *et al.*, 2016). Models were then generated for all 5,821 MHC-I alleles. First, the binding pocket region (reside 25-210) of each allele was aligned to all 50 template sequences, and the best aligning template was selected. The template was then mutated to match the target allele sequence using the Rosetta backrub application with default parameters(Smith and Kortemme, 2008). Following the backrub simulation, an ensemble of 40 structures were generated by selecting the lowest energy structures from 40 iterations of the rosetta relax application(Nivón *et al.*, 2013). The final result was the generation of 232,840 unique structures (5,281 alleles x 40 ensemble members). Finally, the electrostatic environment of the binding pocket was calculated using the APBS software(Jurrus *et al.*, 2018). First, each of the 40 ensemble members were converted into PQR files using the **pdb2pqr30** function. The electrostatic potential was then determined using the default parameters of APBS. The grid dimensions were set to 129Åx 161Åx 129Åwith the fine grid extending to 24Åbeyond the boundaries of the binding pocket and the coarse grid extending to 12Åbeyond the fine grid. This produced a voxel size of approximately $1Å^3$ Electrostatics were calculated using the linearized Poisson-Boltzmann equation with a protein dielectric of 2, a solvent dielectric of 78.54, and an ion concentration of 0.15M.

MHC molecule distance metrics

Three distance functions were used to calculate the divergence of 133 MHC binding pockets with at least 50 known experimental binders.

Hamming distance. This is a metric that determines the distance between two equal length strings as the number of mismatches between sequences. For example, when comparing the peptides "YMLDLQPET" and "YMLAAQPET", the number of mismatches (colored in red) are two. This means that the hamming distance between these peptides is 2. Higher scores indicate more divergent alleles.

BLOSUM alignment. BLOSUM alignment is the sum of the log-odd ratios of a particular amino acid substitution given the background frequency of that amino acid(Henikoff and Henikoff, 1993). For the calculations in this paper, the BLOSUM80 was used, meaning that the probability matrix was determined from sequences with at least 80% sequence homology. BLOSUM80 alignments were calculated with respect to the binding pocket residues within 6Åof the MHC-I N- and C-terminal anchor residues using the *stringDist* function in the *Biostrings* R package(Pagès *et al.*, 2019). Higher scores indicate more divergent alleles.

Electrostatic Potential Distance (EPD). Voxels that fell into a spherical region (r = 6Å), which originated from the average sidechain center of mass of the position 2 and position 9 of crystallized 9mer peptide ligands, were extracted and concatenated into a one-dimensional vector. The electrostatic potential distance between pairwise combinations of the 133 electrostatic vectors (one for each allele) was defined as the L2-norm or euclidean distance. EPD is defined as follows,

$$EPD = \sum_{i}^{n} \sqrt{(x_i - y_j)^2}, \tag{3.1}$$

where i is a single voxel out of n total voxels and x_i and y_i are the equivalent voxels in two different electrostatic vectors corresponding to allele y and allele x. Higher EPDs indicate more divergent binding pocket electrostatic environments.

MHC motif distance metrics

Similar to MHC molecule distance, the pairwise distance of MHC-I binding motifs was measured using 2 different metrics.

Kullback–Leibler divergence. This metric is a statistical distance measurement that quantifies the divergence of two probability distributions. Kullback–Leibler divergence is calculated using the following equation:

$$D_{KL}(P||Q) = \sum_{i}^{n} P_i \log \frac{P_i}{Q_i}. \tag{3.2}$$

In the above equation, P and Q are discrete probability distributions of length n. Due to the fact that KLD is not symmetric, i.e $D_{KL}(P||Q) \neq D_{KL}(Q||P)$, the KLDs reported in this paper are the average KLD or

$$KLD = \frac{D_{KL}(P||Q) + D_{KL}(Q||P)}{2}. \tag{3.3}$$

For the correlations in **figure 3.2**, KLD was defined as the sum of KLDs for position 2 and position 9. For the correlations in **figure 3.5**, KLD was defined as the sum of the observed KLD at all positions (P1-P9).

Motif distance. Motif distance is defined as the L2-norm between two peptide binding motifs. This distance is calculated as follows:

$$motif\ distance = \sum_{i}^{9} \sum_{j}^{20} \sqrt{(q_{ij} - p_{ij})^2}. \tag{3.4}$$

In the above equation, i represents a position of the peptide binding motif which ranges from 1 to 9, and j represents one of the 20 amino acids. q_{ij} and p_{ij} represent the probabilities of amino acid i at position j for two different MHC-I binding motifs, p and q.

Genotype distance. This metric is a specific application of motif distance which is defined as the average distance between a full MHC genotype. Therefore, high values indicate more diversity in binding motifs for a given genotype.

K-means clustering

K-means clustering was performed using the *kmeans* function from the stats R package(Team, 2020). Electrostatic potentials were clustered using the electrostatics potential vector extracted for EPD distance calculations. Predicted MHC-I binding motifs were clustered by converting the 2D binding motif probability matrix(the numerical representation of a peptide binding motif) into a one-dimensional vector of length 180. MHC-I sequences were clustered with respect to BLOSUM-encoded key positions from the MHC-I binding pocket (described in Nielsen *et al.* (2007)). The optimal number of clusters for each application were determined using the average silhouette width method implemented in the *fviz_nbclust* function in the *factoextra* R package(Kassambara and Mundt, 2017).

HLA-Inception model

Training data preparation. The HLA inception model was trained on the electrostatic potential maps corresponding to MHC-I alleles with at least 50 known binders, resulting in a training set of 5,320 maps (133 alleles x 40 maps). To increase attention on important features, 3 segments of each electrostatic map were extracted, an N-terminal binding pocket region, a TCR contact region, and C-terminal binding pocket region, and used to predict the amino acid distribution of residues most likely to interact with it, i.e. the N-terminal binding pocket region was used to predicted the position 2 amino acid distributions. Each segment had the dimensions of 12 x

6 x 12 voxels with the center of each segment falling on an evenly spaced vector that ran the length of the binding pocket. The N-terminal region covered the area that would likely interact with positions 1-3 of binding peptides; the TCR contact region covered the region approximately below positions 4-6 of binding peptide; the C-terminal regions covered the region that would likely interact with positions 7-9 of binding peptides. The segments were then transformed into 3 tensors with each tensor having the dimensions of 5320 x 12 x 6 x 12. These tensors were then linked to a response tensor containing the amino acid distributions for residues likely to interact with that region (5320 x 20). This resulted in 9 training data set, with a specific training set for each peptide position. For example, when training the model to predict the amino acid distribution of position 2 of a binding peptide, the training set would correspond to the tensor of all of the N-terminal segments (5320 x 12 x 6 x 12 using the N-terminal segment) with a response tensor all position 2 amino acid distributions (5,320 x 20).

Model architecture and training. HLA-inception is modeled after the inception v1 model developed by google(Szegedy *et al.*, 2015). The architecture of HLA-inception consists of one AB inception module followed by 4 densely connected layers which were separated by dropout layers. The output layer returns a one-dimensional vector of length 20 with loss being calculated using KLD, and optimized using the ADAM algorithm. Overall, HLA-inception consists of an ensemble of 9 individual models, each corresponding to a different position of the peptide binding motif. A hyperparameter search was performed to identify the best number of epochs and learning rate. Due to the general importance of position 2 and position 9, a grid search was performed on these positions covering epochs 50, 75, and 100 and learning rates 1e-2, 1e-3, and 1e-4. 100 Epochs and a learning rate of 1e-3 were identified as the most optimal and were

used to train all 9 models when performing 10-fold cross validationB.1.

Motif prediction. Using the optimal parameters, HLA-Inception was trained on all available data and used to predict binding motifs for across all 5,821 alleles. In order to better generalize predictions, maps corresponding to each ensemble member for a given allele were averaged to produce one average electrostatic potential map per allele. The averaged maps were then segmented, as previously described, and used as inputs to the trained model. Full binding motifs were then generated by combined the predictions from all nine HLA-Inception models.

Sequence-based model

A deep sequence-based model, analogous to Nielsen *et al.* (2007), was constructed for comparison to HLA-Inception. The input to this model of was a BLOSUM-encoded vector of key position within the MHC-I binding pocket, and the model was trained on a balanced data set consisting of 315,512 experimentally resolved MHC-I peptides paired with randomly generated decoy peptides. The model consisted of 3 densely connected layers separated by dropouts. The output of the model was the probability of the given peptide being a binder.

Position-weighted matrix score

Position-weighted matrix (PWM) score is a measure of how well a peptide fits a given binding motif. The PWM score is calculated by the sum of the log-odd ratios of observing an amino acid at a particular position, given the background frequency of that amino acid. The equation to calculate PWM is as follows,

$$PWM\ score(pep) = \sum_{i=1}^{9} \log_2 \frac{p_{ij}}{q_j}, \tag{3.5}$$

where *pep* is a peptide being scored, i is the residue number being considered, p_{ij} is the probability of the i-th residue of *pep* at the i-th position according to the binding motif, and q_j is the background frequency of the i-th residue of *pep*. Higher PWM scores indicates a higher probability of binding to a target allele. Allele-specific score threshold were determined by calculating the PWM scores for all 9mer peptides in the human protein and generating a *cdf* of that distribution.

Leave-one-cluster-out analysis

Leave-one-cluster-out analysis is defined as the process of using a cluster of alleles, defined by similar binding pocket sequences, to test models, either HLA-Inception or the sequence-based model, that were trained using a data set that did not contain the withheld cluster of alleles. HLA sequences were clustered using the k-means cluster on BLOSUM-encoded binding pocket sequences as previously described in methods(K-means clustering). This produced 11 distinct clusters. The Leave-one-cluster-out analysis was performed on all clusters, and performance was reported as the individual matthew's correlation coefficients(Chicco and Jurman, 2020) of each allele within the cluster.

Chapter 4

INTEGRATIVE MODELING AND DYNAMICS OF THE NL63 SPIKE PROTEIN

This chapter is part of a collaborative work in preparation for publication:

Bending of coronavirus spike regulated by a hinge glycan

David Chmielewski[1,†], Eric A. Wilson[2,†], Peng Zhao[3], Muyuan Chen[4], Greg Pintilie[5], Michael F. Schmid[4], Steven Ludtke[6], Graham Simmons[6,7], Lance Wells[3], Jing Jin[6,7], Abhishek Singharoy[2], Wah Chiu[1,4,5]

Affiliations:

[1]Biophysics Graduate Program, Stanford University, Stanford, CA 94305, USA

[2]School of Molecular Sciences, Arizona State University, Tempe, AZ USA

[3]Department of Biochemistry and Molecular Biology, University of Georgia, Atlanta, GA, USA

[4]Division of CryoEM and Bioimaging, SSRL, SLAC National Accelerator Laboratory, Stanford University, Menlo Park, CA 94025, USA

[5]Department of Bioengineering, and of Microbiology and Immunology, Stanford University, Stanford, CA 94305, USA

[6]Verna and Marrs Mclean Department of Biochemistry and Molecular Biology, Baylor College of Medicine, Houston, TX 77030, USA

[6]Vitalant Research Institute, San Francisco, CA, 94118, USA

[7]Department of Laboratory Medicine, University of California, San Francisco, San Francisco, CA, 94143, USA

[†]These two authors contributed equally to the work.

4.1 Abstract

The Coronavirus spike glycoprotein is an important mediator of receptor binding and subsequent viral entry. The spike protein is heavily glycosylated and recent research has indicated that glycans located on the crown domain may be important receptor binding and immune evasion. However, it remains unknown if glycans located on the stalk domain contribute to these mechanisms, as the flexibility of this region precludes high resolution structures. Using an integrative appraoch of molecular modeling techniques and high quality Cryo-electron tomography (CryoET) data, we built a complete model of the extracellular region of the NL63 spike protein. Molecular dynamics simulations were then used to capture the conformational landscape of the fully glycosylated spike protein. We show a single glycosylation site (N1242) at the upper portion of the stalk domain is responsible for modulating most of the orientational freedom of the NL63 spike protein. The importance of the N1242 glycan was further supported by functional assays showing that infectivity is reduced by 50% when is this glycan is removed. Overall, we show that the integration of conformational landscape information derived from cryoET with molecular modeling and simulations can be valuable for the future development of coronavirus therapeutics and vaccines.

4.2 Introduction

Coronavirus infection begins with binding of the spike protein to specific cellular receptor(s) to initiate entry and membrane fusion (Belouzard *et al.*, 2012). The spike-receptor interaction determines virus pathogenicity, and mutations in spike are responsible for coronaviruses crossing the species barrier and infecting humans(Yang *et al.*, 2015; Cosar *et al.*, 2022). Spike undergoes a large conformational transition from pre-fusion to post-fusion states to achieve membrane fusion during entry(Belouzard

et al., 2012). Current mRNA vaccines for SARS-CoV-2 encode spike proteins stabilized in the pre-fusion state(Vogel *et al.*, 2021) while antibody- and protein-therapeutics are designed to bind the spike receptor-binding domain (RBD), preventing entry via the human ACE2 receptor(Kim *et al.*, 2021).

Structural analysis and molecular dynamics (MD) simulations of spike protein crown domains reveal that they are highly glycosylated, forming a "glycan shield" that is believed to aid in evasion of the host immune response(Casalino *et al.*, 2020; Grant *et al.*, 2020; Walls *et al.*, 2016). However, the extreme flexibility of the stalk domain, largely due to predicted intrinsically-disordered segments, have made it difficult to resolve this region at a high resolution(Walls *et al.*, 2016). Therefore, it is currently unknown if stalk glycans serve a similar functional role as crown domain glycans. Recent structural studies of SARS-CoV-2 virions revealed that the stalk region is able to accommodate large bending angles(Turoňová *et al.*, 2020). Such observations have led to the hypothesis of a "three hinge" model, where the conformational flexibility of the spike protein is granted by 3 disordered segments found within the stalk region. Interestingly, mass spectrometry-based glycan analysis of the NL63 stalk shows that there are two large high-mannose glycans (N1242 and N1247) positioned directly below the longest intrinsically disordered segment(Walls *et al.*, 2016). Given the impact of glycosylations on local protein flexibility (More *et al.*, 2018) and ability to provide order to intrinsically disordered loops (Prates *et al.*, 2018), we hypothesize that these glycans, coined the hinge glycans, modulate the flexibility of NL63 stalk, an aspect important viral entry(Wu and Nemerow, 2004).

In the following work, high quality CryoET data is combined with molecular modeling and physics-based simulations to generate a full length spike protein model that accurately replicates experimentally observed spike bending profiles. Next, the impact of several stalk glycan modifications are investigated, revealing the importance of

hinge glycans in modulating spike protein dynamics. The biological impact of the stalk glycan modifications are then assessed using functional infectivity assays. Together, these data suggest that N1242 glycan is particularly important for maintaining stalk flexibility and viral function.

4.3 Results

4.3.1 Integrative Modeling of NL63 Spike

The following section summarizes the process of integrating experimental data with computational modeling techniques to build the complete extracellular region of the NL63 spike protein. The overall modeling process is summarized in **figure 4.1**.

Initial modeling of the stalk. The extracellular stalk region of the NL63 spike protein was preliminarily modeled using the I-TASSER protein folding software (Yang *et al.*, 2015) to fold a monomeric subunit of the extracellular region NL63 spike protein stalk (residues 1216 to 1297). I-TASSER was chosen because of its high demonstrated accuracy in the CASP competitions. Furthermore, a template-based modeling approach is ideal due to the NL63 stalk consisting of a well studied structural motif, namely a trimeric coiled-coil. The highest confidence model **(Methods)** indicated that a monomeric NL63 stalk subunit likely consisted of a short alpha helix (1216 to 1228) followed by a disordered region (residues 1229 to 1245) that then transitioned back into a longer alpha helix (residues 1246-1297), similar to other coronavirus spike protein models (Turoňová *et al.*, 2020; Woo *et al.*, 2020).

Construction of a whole extracellular spike model. The predicted stalk subunit model was initially in a folded conformation. However, this was likely due to

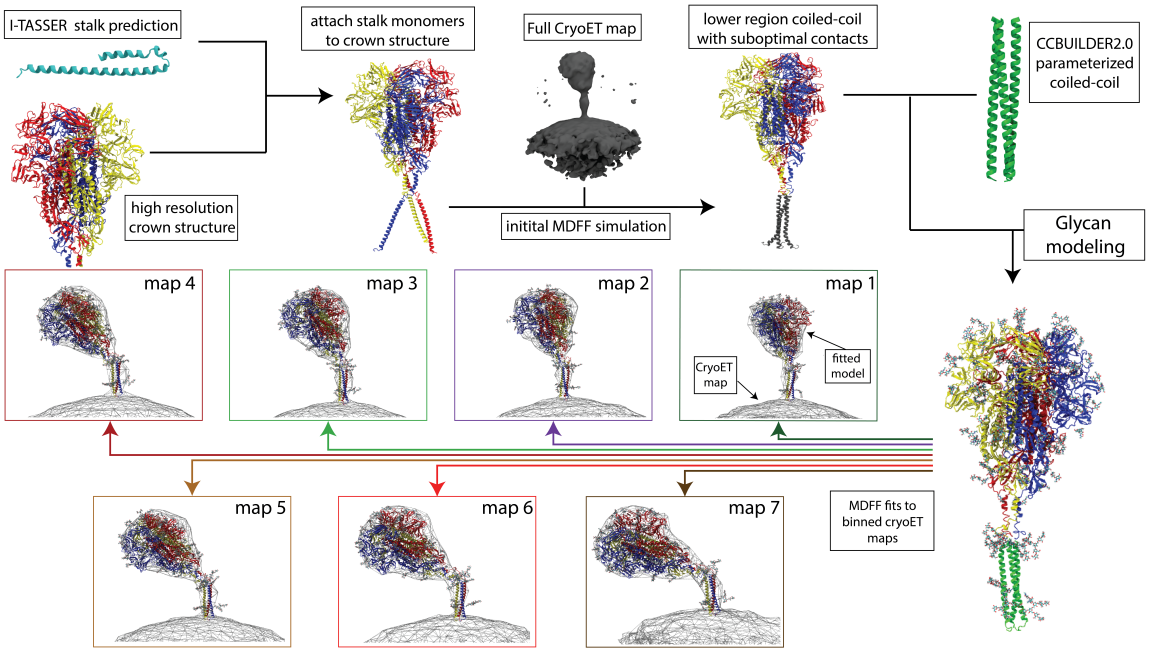

Figure 4.1: Overview of spike model construction. The above figure shows a graphical representation of the NL63 spike modeling procedure.

the modeling process only considering the stalk region as a monomer, a constraint common to most structure prediction methods. Therefore, the monomer was manually manipulated into an elongated confirmation using VMD's interactive MD module. Three elongated monomers were then attached to the high resolution crown model by connecting one elongated monomer to each of the three crown monomers, resulting in a complete, but distorted, extracellular spike protein. Energy minimization and a short 10ns MDFF simulation were then used to reorient the newly attached stalk monomers to improve inter-helix contacts and orientations. The initial MDFF simulation reduced the average amount of solvent exposed hydrophobic residues of the NL63 stalk region (residues 1224 - 1297) by 36% ($47\text{Å}^2 \rightarrow 30\text{Å}^2$). While the simulation successfully oriented the stalk monomers, the poor resolution of the lower coiled-coil segment (residue 1245-1297) precluded direct modeling by MDFF alone. Therefore, an optimized structure of the lower coiled-coil region was generated using the CCbuilder 2.0 program **(Methods)**. The parameterized coiled-coil was

then re-attached to the complete model at position 1245, taking the place of the region modeled by the initial MDFF run. A short MD simulation was performed to relieve any bond strain introduced during the final reattachment process. Stalk glycans were modeled onto the complete model using the CHARMM-GUI interface(Jo *et al.*, 2008), combined with the fully glycosylated crown protein model provided by our collaborators. The glycans for the stalk were chosen based on the combination of the glycans identified in (Walls *et al.*, 2016) and our own mass spectrometry analysis.

Modeling of bent spikes. The different spike protein bending angles observed in the CryoET experimental data **(Figure C.1 (appedix C))** were recreated by fitting the complete model to seven different density maps, each capturing an observed bending angle (ranging from 10° to 70° by an increment of 10°). First, the chimeraX visualization tool(Goddard *et al.*, 2018) was used to align the lower stalk region of the NL63 protein model so that the base of the lower coiled-coil region was flush with the virion surface. The complete model was then fit to the experimental density maps using the Molecular Dynamics Flexible Fitting (MDFF) method(Singharoy *et al.*, 2016). All MDFF simulations were performed using the molecular dynamics simulation (MD) software, NAMD 2.13(Phillips *et al.*, 2020), and the CHARMM36 force field(Brooks *et al.*, 2009). During the 10ns MDFF simulations, a potential energy function (U_{EM}), obtained from the CryoET density map, was applied to all C-α outside of the unstructured region (residues 1224-1241) in order to bias the protein into adopting the experimentally resolved bending angle. Due to the large amount of movement necessary to fit the starting model to the experimental density, especially at large bending angles, a high g scale **(Methods)** was used. To avoid unwanted perturbation of the model, restraints were added to maintain cis-peptide, secondary structure, and chirality of the protein. While fitting to the cryo-ET density data, the

map	starting model	MDFF (round 1)	MDFF (round 2)
1	0.33	0.41	0.36
2	0.25	0.42	0.37
3	0.14	0.42	0.38
4	0.04	0.44	0.38
5	0	0.42	0.38
6	-0.01	0.44	0.39
7	0.01	0.29	0.39

Table 4.1: Cross correlation of spike protein models to CryoET maps. The above table indicates the cross correlation coefficient (ccc) between a protein model and the experimental map used to generate that model. Higher ccc values indicate a better fit between the model and the experimental density. Starting model is the initial ccc before any fitting. MDFF (round 1) indicates the ccc after fitting the starting model to the experimental density using a strong coupling coefficient (g scale = 1). MDFF (round 2) is the ccc after performing the successive MDFF simulations with a decaying g scale ($0.3 \rightarrow 0.1 \rightarrow 0$) with the final structure generated from MDFF (round 1).

overall structure of the crown and the coiled-coil regions were preserved using a domain restraint in the form of a RMSD bias being applied to the C-α of all residues not in the unstructured region using NAMD's Targeted Molecular Dynamics module(Phillips *et al.*, 2020). Finally, excessive flexibility of the lower stalk region due to the lack of a transmembrane domain was avoided by placing a positional restraint on the C-terminal of the protein. Following the initial MDFF simulations, the unstructured region was further refined by performing 3 sequential MDFF simulations. In each simulation, the g scale of the MDFF potential applied to all protein C-α atoms was reduced. The simulations started with a g scale of 0.3 (10ns), transitioned to a g scale of 0.1 (5ns), and ended with a g scale of 0 (5ns). The cross correlation values of the spike protein model at different steps of the modeling process can be seen in **table**

4.1. Overall, the described procedure resulted in the creation of seven NL63 spike protein models, each representing a different experimentally resolved bending angle.

4.3.2 Spike Ensemble Generation

Equilibrium MD simulation. Following construction of all seven bent models, all atom explicit solvent simulations were performed using NAMD 2.13 and CHARMM36m force fields(Phillips *et al.*, 2020; Brooks *et al.*, 2009). Following equilibration, the structures were simulated in triplicates for approximately 50ns. To capture large movements that would be otherwise inaccessible in explicit simulations, approximately 100ns of implicit solvent simulations were performed on the final frames of each explicit simulation. This resulted in a total sampling time of 450ns per map for a total of $3.15\mu s$. For simulations investigating the effects of glycan modulation, glycans were removed prior to the implicit solvent simulation step. An overview of the MD simulation can be seen in **figure 4.2**, and a summary of simulation times can be found in the appendix **(table C.1-C.2)**.

4.3.3 Modification of Hinge Glycans Produces Deviations in Bending Profile

CryoET data indicated that the NL63 spike protein was most commonly found in a bent confirmation ranging from 30-60 degrees with a minimum at 56° **(Figure 4.3A (black line), Figure C.1B)**. Therefore, the integrative model was initially tested to see if a similar bending angle distribution was recovered in MD simulations **(Methods)**. Indeed, the bending profile derived from MD was found to be in agreement with the experimental data **(Figure 4.3A, red line)**, identifying the same most probable bending angle (56°). This match of the simulated spike protein dynamics to the experimental data prompted an investigation into the dynamical effects of stalk gylcan modification.

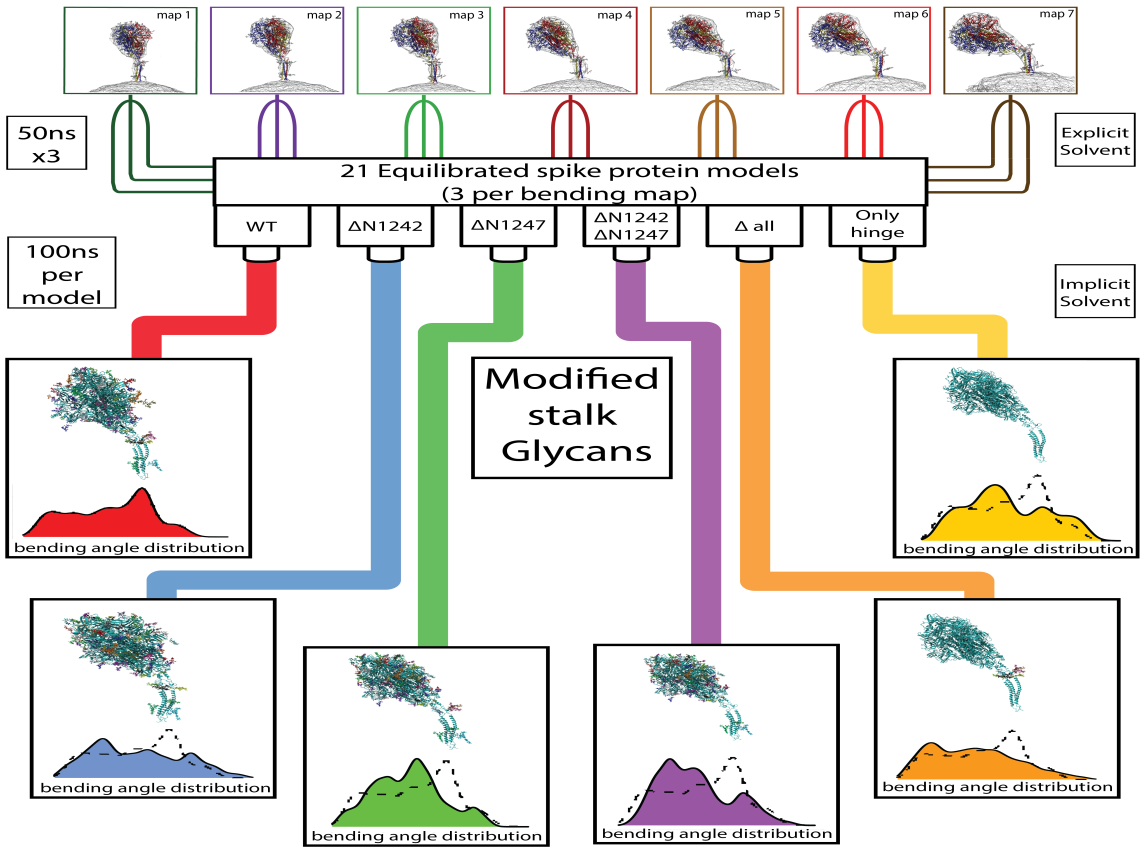

Figure 4.2: Spike simulation Summary. The above figure shows a graphical summary of the spike molecular dynamics simulations. Δ indicates removal of a specified glycan, with "Δ all" indicating the removal of all glycans.

Guided by the proximity and conservation of glycosylated residues to the primary unstructured hinge region**(Figure 4.3B)**, the bending dynamics of the spike protein after removal of the glycans from these positions was accessed with 2.1 microsecond implicit solvent simulations. Illistrated in **(figure 4.3C)**, the removal of any hinge glycan(s) biased the simulations to sampling shallower bending angles. This tendency to sample smaller bending angles was particularly evident for any simulations in which the N1242 glycan removed (one-sided mann-whitney U test: p ¡ 2.2e-16). Altogether, we found that the general removal of glycans near the unstructured hinge region biased simulations to sample more shallow bending angles, with the most pronounced effects being observed upon removal of the N1242 glycan.

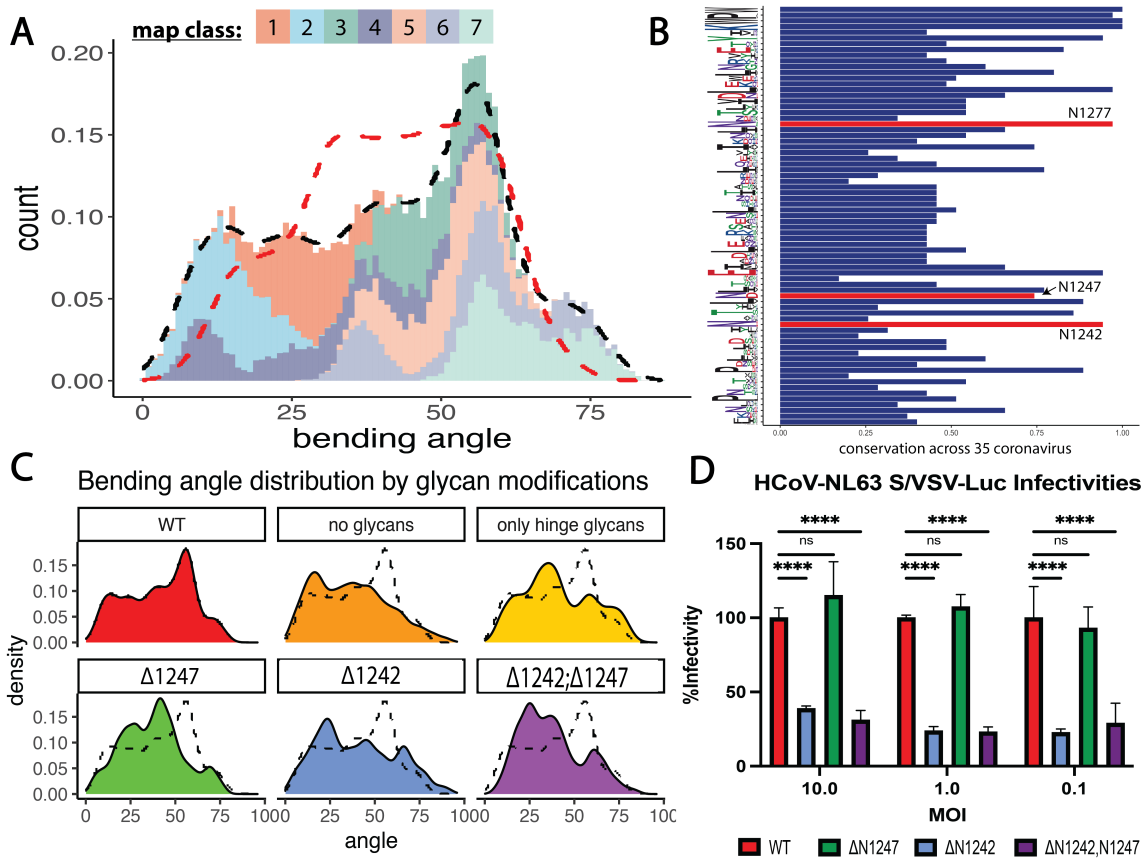

Figure 4.3: Stalk glyan modifications modulate bending dynamics. A. The simulated bending profile of the NL63 spike protein. Different colors represent different starting maps, as indicated by the legend. The black line shows the overall bending angle distribution calculated from the MD simulations, while the red line shows the experimental distribution. **B.** A multiple sequence alignment was preformed using 35 coronaviruses. The x-axis indicates the conservation across all species, and the y-axis indicates a different residue of the NL63 stalk (represented as sequence logos with large letter sizes meaning stronger conservation). The red bars indicate stalk glycan positions. The hinge glycans are defined as the glycans at position 1242 and 1245. **C.** The bending angle distributions for 5 different glycan modifications. The black dotted line indicates the WT bending angle distribution. **D.** Functional infectivity assays for various stalk glycan modifications. Glycan deletion was accomplished by mutating the glycosylated asparagine to alanine. 4 glycan configurations were tested: WT, N1242 or N1247 mutant, and N1242,N1247 double mutant

The striking shift in bending angles after stalk glycan modification motivated functional assays to ascertain the existence of commensurate biological changes. To this end, infectivity assays using VSV-Luc reporter viruses pseudotyped with wild type (WT) NL63 spike or NL63 spike bearing alanine mutations at N-linked glycosylation sites were performed**(Figure 4.3D)**. These experiments indicated that spike proteins without glycans at N1242 had significantly decreased infectivity, providing biological support for the observed dynamic changes.

In summary, we show that our NL63 model is able to faithfully replicated experimental bending angle profiles. Simulations of the fully glycosylated model show that removal of stalk glycans significantly impact bending dynamics with the most significant changes being associated with the removal of the N1242 glycan, a strongly conserved glycan across coronavirus spike proteins. Finally, biological assays support the importance of the N1242 glycan to spike function.

4.3.4 NL63 Stalk Bending is Modulated by N1242 Glycan

In light of the computational and biological support for the importance of hinge glycans in spike protein structure and function, the interactions of these glycans with the spike protein were analyzed. **Figure 4.4** presents the overlaid positions of the hinge glycans when WT simulations are binned by different bending angle ranges. We found that the positions of the two hinge glycans were well mixed at low bending angles, while a more noticeable separation between glycans was observed at higher bending angles**(Figure 4.4A)**. Analysis of the median minimum distance of hinge glycans to protein residue **(Methods)** showed that the simulations containing the N1242 glycan made more significant contacts with the unstructured region and upper coiled-coil **(Figure 4.4B)**. Additionally, this analysis shows that the N1242 glycan is in the closest contact with the unstructured region and upper coiled-coil regions

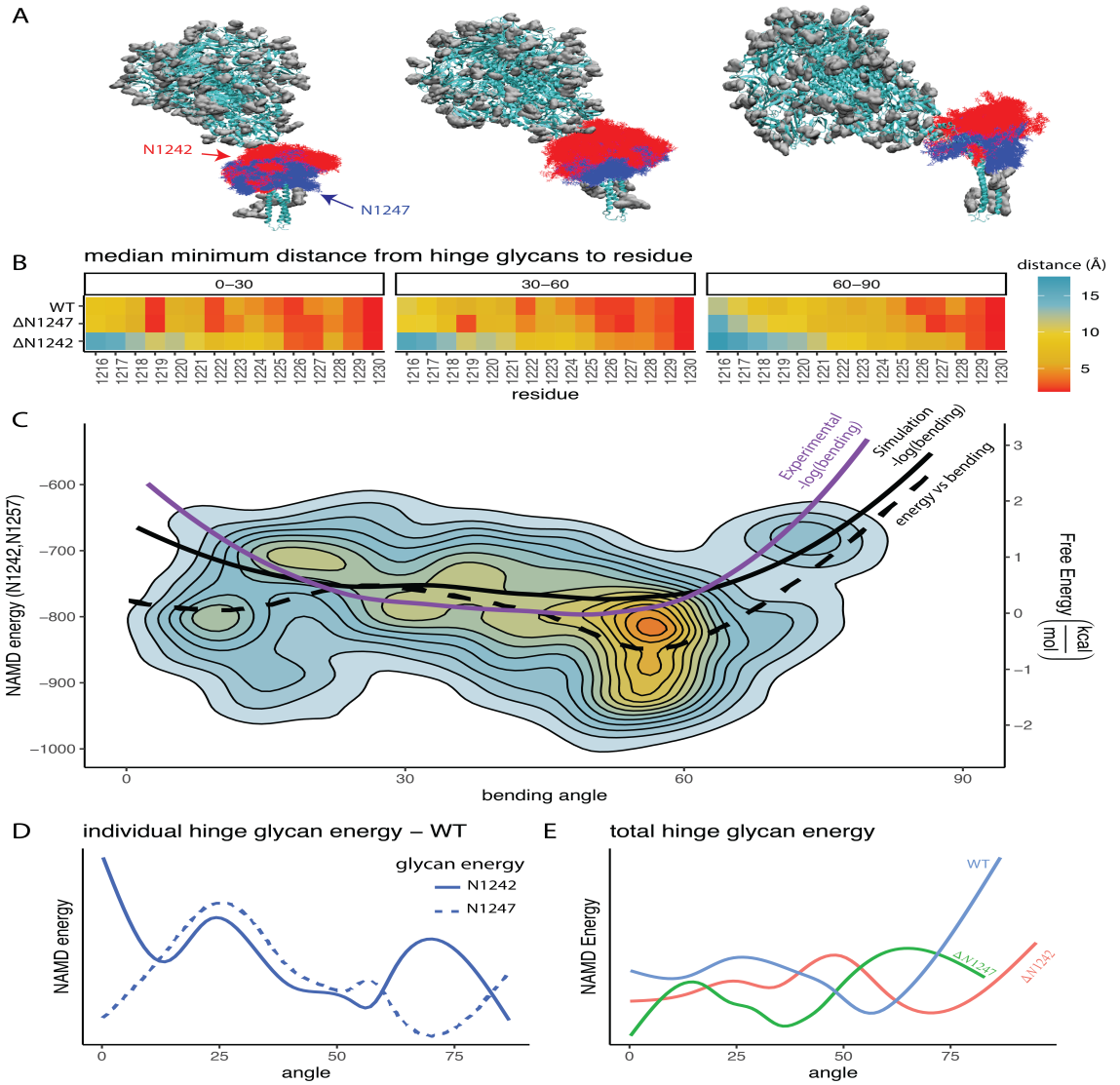

Figure 4.4: Hinge glycan-protein interactions. **A.** MD trajectories of WT spike simulations were grouped into 3 bins based on bending angle (0-30;30-69;60-90). Following binning, the position of the hinge glycans were superimposed, with red and blue represent the N1242 and N1245 glycan, respectively. **B.** For each frame and single glycan deletions, the minimum distance between hinge glycan(s) were calculated to every residue in the unstructured and upper coiled-coiled region. The heatmap indicates the median distance for all frames in that bin. **C.** The relationship between combined hinge glycan-protein interaction energy (y-axis) and bending angle (x-axis) was visualized with a contour map with warmer colors indicating higher density of sampling. The dotted line is a polynomial function fit to the data. The solid purple and black line indicate the $-log$ of the experimental and simulated bending angles, respectively. **D.** The individual energy contributions of the hinge glycans (N1242 and N1245). **E.** Hinge glycan-protein interaction energy for stalk glycan modification simulations.

during the most probable bending angles. Taken together, these data indicate that the N1242 glycan is the primary intermediary between the unstructured region and the hinge glycans.

Based on the contacts observed between the hinge glycans and the unstructured/upper coiled-coil region, we determined the interaction energy using the NAMD energy score function **(Figure 4.4C)**. Interestingly, we found that when a line was fit to the NAMD energy profile of hinge glycans as a function of bending angle, there was a minimum that coincided with the most probable angle. This indicates that the most probable bending angles produced favorable hinge protein-glycan interactions. To further delineate the impact of each hinge glycan, we determined their individual energetic contributions**(Figure 4.4D)**. We found that, while both glycans had similar interaction profiles for moderate bending angles (30-60), there were stark differences in the energies associated with extremes angles. The N1242 glycan showed less favorable interactions at bending angles greater than 60° and less than 20°, while the N1247 had interaction energy minima in both regions. A similar trend was observed in the energy analysis of the hinge glycan-protein interaction energies for the glycan modification simulations**(Figure 4.4E)**. In simulations where the N1242 glycan is deleted, there are poorer glycan-protein interactions at moderate bending angles(30°-60°). Overall, we see that the hinge glycans have favorable energetic interactions at high probability bending angles and that the N1242 glycan may play a role in stabilizing moderate bending angles.

In summary, the nature of the hinge glycan-protein interactions were investigated. We found that the hinge glycan contact, and associated interaction energies, with the unstructured/upper coiled-coil region was dependent on bending angle, with the N1242 glycan making the most significant interactions at the most probable bending angle. When separating the individual energetic contributions of each hinge glycan,

the N1245 glycan was found to have more favorable interactions at extreme bending angle relative to the 1242 glycan. This observation was further supported by analysis showing poorer hinge glycan-protein interactions at moderate bending angles after N1242 glycan removal.

4.4 Discussion

CryoET is a powerful method for resolving heterogeneous structures at the single particle level. For this reason, cryoET has been used to study viral and pathogen entry mechanisms(Prasad *et al.*, 2022; Sun *et al.*, 2022; Quemin *et al.*, 2020). As presented here, we leverage the dynamical data gained from CryoET to create a protein model ensemble, representing key milestones of the native protein dynamics. This synergistic integration of experiments and modeling allowed for the recovery of the full NL63 spike bending profile without the use of any external steering forces. By forgoing the use of such forces, it is possible to derive equilibrium properties and free energies associated with bending(Ovchinnikov and Karplus, 2012) directly from experimental data. It should be noted that the integrative modeling approach describe here is not completely free from all bias, particularly with regard to the modeling of the lower coiled-coil region. Future advances in structure determination methods that can better resolve highly mobile regions will improve such modeling endeavors. Overall, we show that CryoET can be synergistically combined with molecular modeling to create native dynamical ensembles of proteins.

N-linked glycosylation of the coronavirus crown domain has been hypothesized to be important for the viral entry and immune evasion(Casalino *et al.*, 2020; Grant *et al.*, 2020; Walls *et al.*, 2016), however, the role of stalk glycans has remained unresolved. Extensive molecular dynamics simulations of the NL63 spike protein show that the hinge glycans are likely responsible for modulating bending dynamics.

These simulations reveal that the NL63 glycan makes extensive and energetically favorable contacts with the unstructured/upper coiled-coiled regions of the stalk. The importance of these contacts are further demonstrated by the significant changes in bending profiles and viral function upon the removal of key hinge glycans. However, while indirectly supported by both experiments and simulations, future CryoET experiments of NL63 spike proteins with modified glycans will be necessary to confirm the impact of hinge glycan modifications on bending profile. Going forward, the modulating relationship between large glycan moieties and flexible protein regions described in this work may be useful for elucidating the structure-function dynamics of other immunologically relevant glycosylated flexible protein structures(Vaitaitis and Wagner Jr, 2010; Shore *et al.*, 2005; Lee *et al.*, 1992).

In an effort to achieve maximal neutralization, current coronavirus vaccines target the RBD domains of the virus(Vogel *et al.*, 2021). However, such vaccines are susceptible to immune escape given the high mutational rate of RBD domain(Greaney *et al.*, 2021). Our data shows that stalk glycans may be important for viral infectivity, and the highly conserved nature of this region implies a lower rate of immune escape of antibodies targeting this region(Shah *et al.*, 2021). Furthermore, it has been shown that antibodies targeting this region can inhibit viral entry and can be cross reactive across variants and species(Wang *et al.*, 2021; Wu *et al.*, 2022). Therefore, antibodies that target this glycan or restrict flexibility of the stalk region may be useful for the next-generation of universal coronavirus vaccines.

4.5 Methods

Molecular dynamics theory

Molecular Dynamics. Molecular dynamics (MD) is a computational simulation method in which a molecular system is evolved through time according to Newtonian mechanics,

$$m_\alpha \ddot{\vec{r}_a} = -\frac{\partial}{\partial \vec{r_\alpha}} U_{total}(\vec{r_1}, \vec{r_2} \ldots, \vec{r_N}), \quad \alpha = 1, 2 \ldots N. \tag{4.1}$$

In the above equation, m_a represents the mass of atom a, $\vec{r_a}$ represents the position of atom a, and U_{total} represents the potential energy of atom a from the pairwise interactions with all other atoms in the system (Phillips *et al.*, 2005).

MD force fields and potential energy. The potential energy of a simulated system is calculated using a "Force field". A force field is a collection of empirically determined parameters that quantify the pairwise energies associated with bonded and non-bonded interactions between atoms(Brooks *et al.*, 2009). The total potential energy is calculated as the sum of component potential energies

$$U_{total} = U_{bond} + U_{angle} + U_{dihedral} + U_{vdW} + U_{elec}. \tag{4.2}$$

All of the bonded, or covalent, potential energy contributions are captured by $U_{bond} + U_{angle} + U_{dihedral}$. These describe the potential energies associated with the stretching (U_{bond}), bending (U_{angle}) and rotation ($U_{dihedral}$) of all covalent bonds within the system. The final two terms $U_{vdW} + U_{elec}$ make up the non-bonded interactions. U_{vdW} encodes the van der Waals or volume exclusion interaction potential energies. In NAMD, van der Waals interactions are solved using the lennard-jones 12-6 potential equation,

$$U_{vdW} = (-E_{min}) \left[\left(\frac{R_{min}}{r_{ij}} \right)^{12} - 2 \left(\frac{R_{min}}{r_{ij}} \right)^{6} \right], \tag{4.3}$$

where r_{ij} indicates the distance between atoms, E_{min} is the energy well depth, and R_{min} is the distance at which the potential energy is minimized (i.e. E_{min})(Phillips *et al.*, 2005). On the other hand, U_{elec} represents Coulomb's Law of electrostatic interactions. This term is calculated using the following,

$$U_{elec} = \epsilon_{14} \frac{C q_i q_j}{\epsilon_0 r_{ij}}, \tag{4.4}$$

where $q_i q_j$ are the respective charges of atoms i and j, ϵ_0 is the dielectric constant, r_{ij} is the distance between atoms, C is coulomb's constants, and $\epsilon_{14}$is a unitless scaling factor. In theory, non-bonded interactions exist between all pairwise combinations of atoms in a system. However, consideration of all possible combinations is computational expensive, especially as system size scales. Therefore, both the van der Waals and electrostatics potential energy functions are truncated at a user-defined distance. This is accomplished via a switching function which smoothly transitions the interaction energy to 0 at the desired cutoff distance(Phillips *et al.*, 2005).

In systems with periodic boundaries (i.e. Explicit solvent simulations), the user-defined electrostatic interaction cutoff is used to determine the short range component of the electrostatic potential energy function while the long range component is calculated periodically using the particle mesh ewald method or PME. The PME method captures long range electrostatic interactions by summing over interactions in the fourier space where it quickly converges, and thus can be efficiently truncated(Phillips et al. 2005). The efficiency of this method can not be overstated as it reduces the computational complexity of electrostatic calculations from $\mathcal{O}(n^2)$ (pairwise electrostatic calculations for all atoms) to $\mathcal{O}(n \log n)$ where n is the number of atoms.

In systems without periodic boundaries (i.e implicit solvent simulations), both long and short range electrostatic interactions are calculated as pairwise interactions. In this case, the use of only pairwise calculations is computationally tractable as

the lack of solvent causes a marked reduction in the total number of atoms in the system. However, implicit solvent simulations typically require significantly larger cutoff distances to recover an accurate representation of electrostatic forces.

Creating a trajectory. Following the calculation of atom interaction forces, the positions and velocities of all atoms in the system are updated using the velocity verlet integration method. In simple terms, this algorithm first solves the velocities of each atom based on the inter-atom forces as determined from the force field. The atom velocities are then used to update the positions of each atom after a short period of time has passed. Following positional updates, the force field is then used to recalculate the inter-atom forces. This process is repeated for the desired amount of total simulation time. The short amount of time that is allowed to pass when determining the new positions of the atoms is called the time step. The time step is usually on the order of femtoseconds (fs). Larger time steps allow for faster simulations, however, they also introduce instabilities into the simulation. This is particularly true when calculating the movement of very light atoms,s such as hydrogen atoms. When using large time steps, the position of hydrogen atoms can rapidly change, producing unphysical characteristics (i.e. impossible bond lengths or angles) to manifest. This results in extremely high energies which ultimately cause the simulation to fail. Such instabilities are combatted using two methods, the use of which is dependent on the length of the time step. For time step less than or equal to 2 fs, Hydrogen positions and bond lengths can be constrained. This is accomplished via the SHAKE and RATTLE algorithms(Ryckaert *et al.*, 1977; Andersen, 1983) which numerically update the bond lengths and velocities of hydrogen atoms so that they satisfy idealized values. Both algorithms have a user defined tolerance for deviation from idealized values that must be reached by all atoms. If the threshold is not met after the maximum

allowed iterations, the simulation will fail. When simulating time steps ranging from 3-5 fs, Hydrogen mass reparation (HMR) (Hopkins *et al.*, 2015) is required to ensure convergence of the SHAKE and RATTLE algorithm. In HMR-based simulations, the mass of hydrogen atoms are increased by a given factor (usually 3-fold). To conserve mass of the entire system, the additional mass of the hydrogen is subtracted from the heavy atom bonded to each hydrogen. For example, the mass of hydrogens attached to a methyl group would be increased to 3.024 amu while the mass of the carbon atom would be reduced to 5.963 amu.

Thermodynamic ensemble. Molecular dynamics simulations can be defined by its thermodynamic ensemble. Use of the verlet algorithm alone will generate what is known as the NVE or microcanonical ensemble. In this ensemble, the number of atoms (N), the volume(V) of the simulation, and the energy(E) are held constant. While many fundamental statistical mechanics properties are easily calculated from NVE ensembles, most real world applications fail to meet such assumptions. Therefore, most production simulations are run using the NVT (canonical) or NPT (isobaric-isothermal) ensembles. The NVT ensemble is defined as having a constant number of atoms, constant volume, and constant temperature (T). Constant temperature across a simulation can be maintained through the use of a thermostat. In the case of the simulation performed in this work, langevin dynamics was used. The langevin equation is as follows,

$$M\dot{v} = F(r) - \gamma v + \sqrt{\frac{2\gamma k_b T}{M}} R(t), \tag{4.5}$$

where M is the mass, v is the velocity, r is the position, F is the force, γ is the friction coefficient, k_b is the Boltzmann constant, T is the temperature, and $R(t)$ is a random variable following a gaussian distribution. The last two terms of the langevin

equation couple the simulation temperature to the specified constant temperature. In order to integrate the forces when using langevin dynamics, the langevin equation must be incorporated into the velocity verlet algorithm as seen in the BBK method(Brünger *et al.*, 1984). An example of this can be seen in the following equation pertaining to the positional update using the BBK method:

$$r_{n+1} = r_n + \frac{1-\gamma\Delta t/2}{1+\gamma\Delta t/2}(r_n - r_{n-1}) + \frac{1}{1+\gamma\Delta t/2}\Delta t^2 \left[M^{-1}F(r_n) + \sqrt{\frac{2\gamma k_b T}{\Delta M}} z_n \right]. \quad (4.6)$$

The NPT ensemble is defined as having a constant number of atoms, constant pressure (P), and constant temperature. For simulations contained in this work, the Hoover Langevin piston method was used to maintain constant pressure (Phillips *et al.*, 2005). The Hoover langevin piston method is a Hoover-style extension of langevin dynamics to provide pressure control. The equations of motions are as follows:

$$\dot{r} = p/m + \dot{e}r \quad (4.7)$$

$$\dot{p} = F - \dot{e}p - \gamma p + R(t) \quad (4.8)$$

$$\dot{V} \;=\; 3V\dot{e} \quad (4.9)$$

$$\ddot{e} = 3V/W(P - P_0) - \gamma_e \dot{e} + R_e/W \quad (4.10)$$

$$W = 3N\tau^2 kT \quad (4.11)$$

$$\left\langle R_e^2 \right\rangle = 2W g_e kT/h \quad (4.12)$$

In the above equations, $\dot{e}$ is strain rate (with $\dot{e}W$ being the piston momentum); is the volume; W is the weight of the piston; γ_e is the friction of the piston; R_e is

the random noise of the piston; τ is the oscillation period of the piston; h is the length of the piston. Similar to what was previously shown for BBK integration, the following equation governs the positional update when using the Hoover langevin piston method(Quigley and Probert 2004):

$$r_{n+1} = r_n + \Delta t \dot{r}\left[r, p^{n+\Delta t2}, p_e^{n+\Delta t2}\right] \tag{4.13}$$

Solvation models. All human proteins exist with an aqueous environment. Therefore, it is important to represent the effects of solvent in protein dynamics. In simulations, solvent can either be explicitly or implicitly modeled. In explicit solvent simulations, fully atomistic representations of water molecules and ions are added to the system. The interactions of these molecules with the environment are calculated using the TIP3, or 3-point, water molecule force field. While these simulations are the most accurate in terms of capturing protein dynamics, the number of water molecules necessary to surround a given protein, in which periodic boundary conditions are avoided, is significant. This makes explicit solvent simulations computationally expensive for large protein systems. In implicit solvent simulations, the effect of water on electrostatics interactions is mathematically approximated. While implicit solvent simulations are less technically accurate, they are much more computationally efficient, allowing for faster simulations. Furthermore, they engender enhanced conformational sampling due to the reduction of the viscosity imposed by explicit water molecules. The implicit water simulations described in this work were performed using generalized born implicit solvent (GBIS). GBIS is a linear approximation of the Poisson-Boltzmann equation in which each atom is modeled as a charged sphere. An important feature of the implicit solvent model is the assignment of differential dielectric constants between the interior of atoms and the larger environment, with the former typically being much lower than the latter. This tempers the strength of long range electrostatic

interactions, similar to the effect of explicit solvent. The magnitude of screening for each atom is dependent on the immediate environment, with atoms in more crowded environments experiencing less screening.

Protein modeling

In what follows, first the individual tools employed for the modeling of the NL63 stalk are outlined. Thereafter, an integrative modeling scheme is described, wherein all these tools are combined in a pipeline to study both the structure and dynamics of the NL63 spike protein.

I-TASSER structure prediction. I-TASSER is a hierarchical template-based structure modeling software. The I-TASSER algorithm performs structure prediction in two steps. First, structural templates are selected from a PDB library using LOMETS(Wu and Zhang, 2007), an ensemble of protein threading algorithms. Threading or fold recognition is the process of predicting a protein structure by matching stretches of amino acids from a target sequence to fragments of existing PDB structures. The full structure is then predicted by assembling the selected fragments into a complete structure. Regions of the target protein without corresponding fragments from the PDB database are modeled **de novo** using a monte carlo(MC) based method called Touchstone II(Zhang *et al.*, 2003). Ensembles of potential structures are extracted from MC trajectories generated by the Touchstone II algorithm using the SPICKER algorithm(Zhang and Skolnick, 2004). A reassembly process is then performed using the identified clusters. The lowest energy structures from each cluster are then refined using all atom simulations. The final output of I-TASSER is 5 predicted models, each with a predicted confidence score which is a metric based on the statistical significance of template alignments and convergence of assembly simulations.

CCbuilder 2.0. The CCbuilder2.0 is a web-based application to build parametric coiled-coils (Wood and Woolfson, 2018). Parametric protein modeling typically involves building a target structure mathematically based on user defined parameters. This method can be very accurate in cases where the target protein structure adheres to a regular structure with well-defined characteristics. The highly regular structure of α-helical coiled-coil motifs represent an ideal use case for parametric structure prediction. The CCbuilder2.0 model building processing is carried out using the ISAMBARD package(Wood *et al.*, 2017). This package assembles a coiled-coil motif using the amino acid sequence of the target protein, and 3 parameters controlling the inter-coil interface: radius, pitch, and interface. For the modeling of the lower NL63 stalk region, the default parameters were used which were 5.1, 226, and 24 for the radius, pitch, and interface parameters respectively. Coiled-coil motifs are usually defined by a heptad repeat register (which range from a to g). The most common register for coiled-coiled motifs will have hydrophobic residues at positions a and d. We assessed all potential heptad repeat registers, and selected the register that produced the lowest energy structure. Finally, parameterized coils are scored and optimized using the all-atom BUDE and Rosetta force fields (McIntosh-Smith *et al.*, 2015; Das and Baker, 2008).

CHARMGUI glycan builder. Stalk glycans were added to the complete model using the web based CHARMMGUI glycan builder interface(Park *et al.*, 2019). This interface operates in two steps: a modeling step and confirmation sampling step. First, it builds a model of the user specified glycans using PDB structures of template glycan subunits. The glycosylation torsion angles for the modeled glycan are generated by finding the average angles from clusters of angles sampled in the selected templates.

Second, Clashes and glycan positions are optimized using rigid body rotations of each glycan using the CHARMM force field. The optimization protocol begins by fixing the atoms of the protein structure with the expectation of the asparagine molecule bearing the glycosylation. The torsion angles of each rotatable glycosidic bond is then sampled. If a given orientation has fewer than 5 bad contacts, as determined by heavy atoms being closer than 2.5Å, a short minimization is performed on that orientation. The protein-glycan interaction is then measured; if the interaction energy is lower than the previous interaction, that orientation is stored. Following the initial orientation search, a second search is conducted on the stored orientations identified in the previous step. If any sampled orientation falls below the specified cutoff energy threshold (i.e. 60 kcal/mol), the orientation search is terminated and that orientation is selected for the final model. If the energy threshold is not reached, the orientation search will restart with a different cluster identified in step 1 of the CHARMMGUI glycan builder process. In the event that the minimum energy threshold is never reached, the lowest energy structure is selected. After glycans have been modeled at all positions, A final minimization procedure with improper dihedral restraints is then applied to all glycosylation sites.

Molecular dynamic simulations of NL63 spike protein

All-atom simulations. A combination of explicit followed by implicit solvent molecular dynamics simulations was used to determine the bending dynamics of the NL63 spike protein. Similar combined simulation schemes have been employed in the past (Kleinjung and Fraternali, 2014), where the explicit solvent MD simulation is used first to thermalize the initial model and subsequently, the implicit solvent MD is performed to allow for enhanced sampling of the conformations. This scheme is particularly useful for modeling the soluble proteins(Mishra and Koča, 2018), and has recently

been extended to study glycan systems(Roy *et al.*, 4113), showing that implicit solvent simulations can sample the conformational space akin to accelerated sampling schemes.

Explicit system setup. All MD simulations were initiated using the NAMD 2.13 (Phillips *et al.*, 2020) simulation software. The stalk model was solvated with 3 points (TIP3P) water molecules, representing a unit cell of initial dimensions equal to 429Åx 319Åx 429Å. Na^+ and Cl^- ions were then added to ensure electronic neutrality. The completely solvated system, including the protein, glycans, water molecules, and ions, amounted to nearly 5.8 million atoms. After 11,000 steps of conjugate-gradient energy minimization, the system was equilibrated for 10ns at 310K. All MD simulations were performed in the isobaric-isothermal ensemble with the MD program NAMD 2.13 and the CHARMM36m all-atom force fields for proteins. The temperature was maintained at 300 K using Langevin dynamics with a damping constant of 1 ps^{-1}. The pressure was fixed at 1 atm using the Langevin piston method. Van der Waals and electrostatic short-range interactions were smoothly truncated with a 12 Å cutoff, and a switching function was applied at 10 Å. Long-range electrostatic forces were computed with the particle mesh Ewald algorithm. Simulations were performed using 4 fs time steps via hydrogen mass repartitioning and constraining covalent hydrogen bonds. For production run simulations, a harmonic positional restraint was placed on the C-α atoms of C-terminal residues (1296-1297). This was done to simulate the effect of being attached to a membrane, as the transmembrane region was not included in our model.

Implicit system setup. Implicit solvent simulations were performed using the Generalized Born Implicit solvent method implemented in NAMD. Simulations were performed using a solvent dielectric of 80 and an ion concentration of 0.1 M. The Born

radius cutoff parameter was set to 14Å with the switch distance and cutoff set to 15Å and 16Å respectively. Constant temperature was maintained at 300K using langevin dynamics with a damping coefficient of 5 ps^{-1}. Simulations were carried out with a similar time step size as in the explicit solvent simulations.

Fitting protein structure to Cryo-ET maps. Molecular Dynamics Flexible Fitting or MDFF is MD-based method to bias a simulation to adopt conformations identified from an electron density map(Singharoy *et al.*, 2016). This is a specialized implementation of a GBIS simulation where user-defined atoms within the starting structure are driven to align with high density regions of the electron density map. As such, this method is popular for refining lower resolution structure determination methods like CryoEM or CryoET. The following equation will describe the MDFF bias potential. The MDFF potential can be determined by

$$V_{EM}(r) = \begin{cases} \zeta\left(1 - \frac{\Phi(r)-\Phi_{thr}}{\Phi_{max}-\Phi_{thr}}\right) & \text{if } \Phi(r) \geq \Phi_{thr}\zeta \\ \text{if } \Phi(r) < \Phi_{thr} \end{cases}. \tag{4.14}$$

In the previous equation, $\Phi(r)$ is the coulomb potential associated with the EM map, Φ_{thr} is the noise threshold, and Φ_{max} is the maximum calculated $\Phi(r)$. ζ, also known as the g scale, is a scaling factor that modulates the coupling of strength between atoms and the EM map. The total potential energy of MDFF potential can be solved by $U_{EM} = \sum_i w_i V_{EM}(r_i)$ where i is an atom coupled to the EM map and w_i is the mass of that atom. Total energy during an MDFF simulation can be calculated using the following equation

$$U_{total} = U_{MD} + U_{EM} + U_{SS} \tag{4.15}$$

where U_{MD} is the potential energy determined from the MD force field, U_{EM} is the potential derived from the electron density map, and U_{SS} is the potential energy

that is added to preserve the secondary structure of the protein. The U_{SS} term is necessary as U_{EM} can be strong enough to warp the secondary structure if unchecked.

MD analysis.

Bending angle calculation. The bending of the crown relative to the stalk was determined by finding a vector that passed through the center of the lower stalk region while remaining normal to the virion surface and a second vector that was defined as running through the center of the NL63 crown. The bending angle was then defined as the arccosine of the dot product between unit norms of these two vectors. The angle was calculated with the following equation,

$$\theta = \arccos \|A\| \cdot \|B\|, \tag{4.16}$$

where A is the vector passing through the center of the stalk, and B is the vector passing though the lower coiled-coiled region. Convergence of the simulations was determined by a bootstrapping analysis of the sampled bending angles.

NAMD energy. The interaction energy of the hinge glycans (N1242,1247) were measured using the NAMD energy plugin(Phillips *et al.*, 2020). Energy was either measured from both hinge glycans to the rest of the system or with respect to individual hinge glycans to the protein.

Median minimum glycan to protein distance. The median minimum glycan to protein distance is defined as the shortest distance between any two atoms within the stalk and the glycan, respectively.

Ahmed, S. F., A. A. Quadeer and M. R. McKay, "Covidep: A web-based platform for real-time reporting of vaccine target recommendations for sars-cov-2", Nature reviews microbiology **15**, 2141–2142 (2020).

Andersen, H. C., "Rattle: A "velocity" version of the shake algorithm for molecular dynamics calculations", J. Comput. Phys. **52**, 1, 24–34 (1983).

Andreatta, M. and M. Nielsen, "Gapped sequence alignment using artificial neural networks: application to the mhc class i system", Bioinformatics **32**, 4, 511–517 (2016).

Arora, J., F. Pierini, P. J. McLaren, M. Carrington, J. Fellay and T. L. Lenz, "Hla heterozygote advantage against hiv-1 is driven by quantitative and qualitative differences in hla allele-specific peptide presentation", Molecular biology and evolution **37**, 3, 639–650 (2020).

Bassani-Sternberg, M., E. Bräunlein, R. Klar, T. Engleitner, P. Sinitcyn, S. Audehm, M. Straub, J. Weber, J. Slotta-Huspenina, K. Specht, M. E. Martignoni, A. Werner, R. Hein, D. H Busch, C. Peschel, R. Rad, J. Cox, M. Mann and A. M. Krackhardt, "Direct identification of clinically relevant neoepitopes presented on native human melanoma tissue by mass spectrometry", Nat. Commun. **7**, 13404 (2016).

Bassani-Sternberg, M., C. Chong, P. Guillaume, M. Solleder, H. Pak, P. O. Gannon, L. E. Kandalaft, G. Coukos and D. Gfeller, "Deciphering hla-i motifs across hla peptidomes improves neo-antigen predictions and identifies allostery regulating hla specificity", PLoS computational biology **13**, 8, e1005725 (2017).

Belouzard, S., J. K. Millet, B. N. Licitra and G. R. Whittaker, "Mechanisms of coronavirus cell entry mediated by the viral spike protein", Viruses **4**, 6, 1011–1033 (2012).

Benjamini, Y. and Y. Hochberg, "Controlling the false discovery rate: a practical and powerful approach to multiple testing", Journal of the Royal statistical society: series B (Methodological) **57**, 1, 289–300 (1995).

Berman, H. M., J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov and P. E. Bourne, "The protein data bank", Nucleic acids research **28**, 1, 235–242 (2000).

Britton, T., F. Ball and P. Trapman, "A mathematical model reveals the influence of population heterogeneity on herd immunity to SARS-CoV-2", Science **369**, 6505, 846–849 (2020).

Brooks, B. R., C. L. Brooks, 3rd, A. D. Mackerell, Jr, L. Nilsson, R. J. Petrella, B. Roux, Y. Won, G. Archontis, C. Bartels, S. Boresch, A. Caflisch, L. Caves, Q. Cui, A. R. Dinner, M. Feig, S. Fischer, J. Gao, M. Hodoscek, W. Im, K. Kuczera, T. Lazaridis, J. Ma, V. Ovchinnikov, E. Paci, R. W. Pastor, C. B. Post, J. Z.

Pu, M. Schaefer, B. Tidor, R. M. Venable, H. L. Woodcock, X. Wu, W. Yang, D. M. York and M. Karplus, "CHARMM: the biomolecular simulation program", J. Comput. Chem. **30**, 10, 1545–1614 (2009).

Brünger, A., C. L. Brooks and M. Karplus, "Stochastic boundary conditions for molecular dynamics simulations of ST2 water", Chem. Phys. Lett. **105**, 5, 495–500 (1984).

Button, K. S., J. P. Ioannidis, C. Mokrysz, B. A. Nosek, J. Flint, E. S. Robinson and M. R. Munafò, "Power failure: why small sample size undermines the reliability of neuroscience", Nature Reviews Neuroscience **14**, 5, 365–376 (2013).

Campbell, K. M., G. Steiner, D. K. Wells, A. Ribas and A. Kalbasi, "Prediction of sars-cov-2 epitopes across 9360 hla class i alleles", bioRxiv (2020).

Cao, Y., L. Li, Z. Feng, S. Wan, P. Huang, X. Sun, F. Wen, X. Huang, G. Ning and W. Wang, "Comparative genetic analysis of the novel coronavirus (2019-ncov/sars-cov-2) receptor ace2 in different populations", Cell discovery **6**, 1, 1–4 (2020).

Casalino, L., Z. Gaieb, J. A. Goldsmith, C. K. Hjorth, A. C. Dommer, A. M. Harbison, C. A. Fogarty, E. P. Barros, B. C. Taylor, J. S. McLellan, E. Fadda and R. E. Amaro, "Beyond shielding: The roles of glycans in the SARS-CoV-2 spike protein", ACS Cent. Sci. **6**, 10, 1722–1734 (2020).

Channappanavar, R., J. Zhao and S. Perlman, "T cell-mediated immune response to respiratory coronaviruses", Immunologic research **59**, 1-3, 118–128 (2014).

Chen, V. B., W. B. Arendall, 3rd, J. J. Headd, D. A. Keedy, R. M. Immormino, G. J. Kapral, L. W. Murray, J. S. Richardson and D. C. Richardson, "MolProbity: all-atom structure validation for macromolecular crystallography", Acta Crystallogr. D Biol. Crystallogr. **66**, Pt 1, 12–21 (2010).

Chicco, D. and G. Jurman, "The advantages of the matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation", BMC Genomics **21**, 1, 6 (2020).

Chong, C., F. Marino, H. Pak, J. Racle, R. T. Daniel, M. Müller, D. Gfeller, G. Coukos and M. Bassani-Sternberg, "High-throughput and sensitive immunopeptidomics platform reveals profound Interferonγ-Mediated remodeling of the human leukocyte antigen (HLA) ligandome", Mol. Cell. Proteomics **17**, 3, 533–548 (2018).

Chowell, D., C. Krishna, F. Pierini, V. Makarov, N. A. Rizvi, F. Kuo, L. G. Morris, N. Riaz, T. L. Lenz and T. A. Chan, "Evolutionary divergence of hla class i genotype impacts efficacy of cancer immunotherapy", Nature medicine **25**, 11, 1715–1720 (2019).

Collins, E. J., D. N. Garboczi and D. C. Wiley, "Three-dimensional structure of a peptide extending from one end of a class I MHC binding site", Nature **371**, 6498, 626–629 (1994).

Cosar, B., Z. Y. Karagulleoglu, S. Unal, A. T. Ince, D. B. Uncuoglu, G. Tuncer, B. R. Kilinc, Y. E. Ozkan, H. C. Ozkoc, I. N. Demir, A. Eker, F. Karagoz, S. Y. Simsek, B. Yasar, M. Pala, A. Demir, I. N. Atak, A. H. Mendi, V. U. Bengi, G. Cengiz Seval, E. Gunes Altuntas, P. Kilic and D. Demir-Dora, "SARS-CoV-2 mutations and their viral variants", Cytokine Growth Factor Rev. **63**, 10–22 (2022).

Croft, N. P., S. A. Smith, J. Pickering, J. Sidney, B. Peters, P. Faridi, M. J. Witney, P. Sebastian, I. E. Flesch, S. L. Heading *et al.*, "Most viral peptides displayed by class i mhc on infected cells are immunogenic", Proceedings of the National Academy of Sciences **116**, 8, 3112–3117 (2019).

Das, R. and D. Baker, "Macromolecular modeling with rosetta", Annu. Rev. Biochem. **77**, 363–382 (2008).

de Lusignan, S., J. Dorward, A. Correa, N. Jones, O. Akinyemi, G. Amirthalingam, N. Andrews, R. Byford, G. Dabrera, A. Elliot *et al.*, "Risk factors for sars-cov-2 among patients in the oxford royal college of general practitioners research and surveillance centre primary care network: a cross-sectional study", The Lancet Infectious Diseases (2020).

Dong, E., H. Du and L. Gardner, "An interactive web-based dashboard to track covid-19 in real time", The Lancet infectious diseases (2020).

Fruchterman, T. M. J. and E. M. Reingold, "Graph drawing by force-directed placement", Software: Practice and (1991).

Garrett, T. P., M. A. Saper, P. J. Bjorkman, J. L. Strominger and D. C. Wiley, "Specificity pockets for the side chains of peptide antigens in HLA-Aw68", Nature **342**, 6250, 692–696 (1989).

Goddard, T. D., C. C. Huang, E. C. Meng, E. F. Pettersen, G. S. Couch, J. H. Morris and T. E. Ferrin, "UCSF ChimeraX: Meeting modern challenges in visualization and analysis", Protein Sci. **27**, 1, 14–25 (2018).

González-Galarza, F. F., L. Y. Takeshita, E. J. Santos, F. Kempson, M. H. T. Maia, A. L. S. d. Silva, A. L. T. e. Silva, G. S. Ghattaoraya, A. Alfirevic, A. R. Jones *et al.*, "Allele frequency net 2015 update: new features for hla epitopes, kir and disease and hla adverse drug reaction associations", Nucleic acids research **43**, D1, D784–D788 (2015).

Gowans, J. L., D. D. McGREGOR and D. M. Cowen, "Initiation of immune responses by small lymphocytes", Nature **196**, 651–655 (1962).

Grant, O. C., D. Montgomery, K. Ito and R. J. Woods, "Analysis of the SARS-CoV-2 spike protein glycan shield reveals implications for immune recognition", Sci. Rep. **10**, 1, 1–11 (2020).

Greaney, A. J., T. N. Starr, C. O. Barnes, Y. Weisblum, F. Schmidt, M. Caskey, C. Gaebler, A. Cho, M. Agudelo, S. Finkin *et al.*, "Mapping mutations to the sars-cov-2 rbd that escape binding by different classes of antibodies", Nature communications **12**, 1, 1–14 (2021).

Grifoni, A., D. Weiskopf, S. I. Ramirez, J. Mateus, J. M. Dan, C. R. Moderbacher, S. A. Rawlings, A. Sutherland, L. Premkumar, R. S. Jadi *et al.*, "Targets of t cell responses to sars-cov-2 coronavirus in humans with covid-19 disease and unexposed individuals", Cell (2020).

Guo, Y.-R., Q.-D. Cao, Z.-S. Hong, Y.-Y. Tan, S.-D. Chen, H.-J. Jin, K.-S. Tan, D.-Y. Wang and Y. Yan, "The origin, transmission and clinical therapies on coronavirus disease 2019 (covid-19) outbreak–an update on the status", Military Medical Research **7**, 1, 1–10 (2020).

Henikoff, S. and J. G. Henikoff, "Performance evaluation of amino acid substitution matrices", Proteins **17**, 1, 49–61 (1993).

Hie, B., E. D. Zhong, B. Berger and B. Bryson, "Learning the language of viral evolution and escape", Science **371**, 6526, 284–288 (2021).

Hopkins, C. W., S. Le Grand, R. C. Walker and A. E. Roitberg, "Long-Time-Step molecular dynamics through hydrogen mass repartitioning", J. Chem. Theory Comput. **11**, 4, 1864–1874 (2015).

Humphrey, W., A. Dalke, K. Schulten *et al.*, "Vmd: visual molecular dynamics", Journal of molecular graphics **14**, 1, 33–38 (1996).

International HIV Controllers Study, F. Pereyra, X. Jia, P. J. McLaren, A. Telenti, P. I. W. de Bakker, B. D. Walker, S. Ripke, C. J. Brumme, S. L. Pulit, M. Carrington, C. M. Kadie, J. M. Carlson, D. Heckerman, R. R. Graham, R. M. Plenge, S. G. Deeks, L. Gianniny, G. Crawford, J. Sullivan, E. Gonzalez, L. Davies, A. Camargo, J. M. Moore, N. Beattie, S. Gupta, A. Crenshaw, N. P. Burtt, C. Guiducci, N. Gupta, X. Gao, Y. Qi, Y. Yuki, A. Piechocka-Trocha, E. Cutrell, R. Rosenberg, K. L. Moss, P. Lemay, J. O'Leary, T. Schaefer, P. Verma, I. Toth, B. Block, B. Baker, A. Rothchild, J. Lian, J. Proudfoot, D. M. L. Alvino, S. Vine, M. M. Addo, T. M. Allen, M. Altfeld, M. R. Henn, S. Le Gall, H. Streeck, D. W. Haas, D. R. Kuritzkes, G. K. Robbins, R. W. Shafer, R. M. Gulick, C. M. Shikuma, R. Haubrich, S. Riddler, P. E. Sax, E. S. Daar, H. J. Ribaudo, B. Agan, S. Agarwal, R. L. Ahern, B. L. Allen, S. Altidor, E. L. Altschuler, S. Ambardar, K. Anastos, B. Anderson, V. Anderson, U. Andrady, D. Antoniskis, D. Bangsberg, D. Barbaro, W. Barrie, J. Bartczak, S. Barton, P. Basden, N. Basgoz, S. Bazner, N. C. Bellos, A. M. Benson, J. Berger, N. F. Bernard, A. M. Bernard, C. Birch, S. J. Bodner, R. K. Bolan, E. T. Boudreaux, M. Bradley, J. F. Braun, J. E. Brndjar, S. J. Brown, K. Brown, S. T. Brown, J. Burack, L. M. Bush, V. Cafaro, O. Campbell, J. Campbell, R. H. Carlson, J. K. Carmichael, K. K. Casey, C. Cavacuiti, G. Celestin, S. T. Chambers, N. Chez, L. M. Chirch, P. J. Cimoch, D. Cohen, L. E. Cohn, B. Conway, D. A. Cooper, B. Cornelson, D. T. Cox, M. V. Cristofano, G. Cuchural, Jr, J. L. Czartoski, J. M. Dahman, J. S. Daly, B. T. Davis, K. Davis, S. M. Davod, E. DeJesus, C. A. Dietz, E. Dunham, M. E. Dunn, T. B. Ellerin, J. J. Eron, J. J. W. Fangman, C. E. Farel, H. Ferlazzo, S. Fidler, A. Fleenor-Ford, R. Frankel, K. A. Freedberg, N. K. French, J. D. Fuchs, J. D. Fuller, J. Gaberman, J. E. Gallant, R. T. Gandhi, E. Garcia, D. Garmon, J. C. Gathe, Jr, C. R. Gaultier, W. Gebre, F. D. Gilman, I. Gilson, P. A. Goepfert, M. S. Gottlieb, C. Goulston, R. K. Groger, T. D. Gurley,

S. Haber, R. Hardwicke, W. D. Hardy, P. R. Harrigan, T. N. Hawkins, S. Heath, F. M. Hecht, W. K. Henry, M. Hladek, R. P. Hoffman, J. M. Horton, R. K. Hsu, G. D. Huhn, P. Hunt, M. J. Hupert, M. L. Illeman, H. Jaeger, R. M. Jellinger, M. John, J. A. Johnson, K. L. Johnson, H. Johnson, K. Johnson, J. Joly, W. C. Jordan, C. A. Kauffman, H. Khanlou, R. K. Killian, A. Y. Kim, D. D. Kim, C. A. Kinder, J. T. Kirchner, L. Kogelman, E. M. Kojic, P. T. Korthuis, W. Kurisu, D. S. Kwon, M. LaMar, H. Lampiris, M. Lanzafame, M. M. Lederman, D. M. Lee, J. M. L. Lee, M. J. Lee, E. T. Y. Lee, J. Lemoine, J. A. Levy, J. M. Llibre, M. A. Liguori, S. J. Little, A. Y. Liu, A. J. Lopez, M. R. Loutfy, D. Loy, D. Y. Mohammed, A. Man, M. K. Mansour, V. C. Marconi, M. Markowitz, R. Marques, J. N. Martin, H. L. Martin, Jr, K. H. Mayer, M. J. McElrath, T. A. McGhee, B. H. McGovern, K. McGowan, D. McIntyre, G. X. Mcleod, P. Menezes, G. Mesa, C. E. Metroka, D. Meyer-Olson, A. O. Miller, K. Montgomery, K. C. Mounzer, E. H. Nagami, I. Nagin, R. G. Nahass, M. O. Nelson, C. Nielsen, D. L. Norene, D. H. O'Connor, B. O. Ojikutu, J. Okulicz, O. O. Oladehin, E. C. Oldfield, 3rd, S. A. Olender, M. Ostrowski, W. F. Owen, Jr, E. Pae, J. Parsonnet, A. M. Pavlatos, A. M. Perlmutter, M. N. Pierce, J. M. Pincus, L. Pisani, L. J. Price, L. Proia, R. C. Prokesch, H. C. Pujet, M. Ramgopal, A. Rathod, M. Rausch, J. Ravishankar, F. S. Rhame, C. S. Richards, D. D. Richman, B. Rodes, M. Rodriguez, R. C. Rose, 3rd, E. S. Rosenberg, D. Rosenthal, P. E. Ross, D. S. Rubin, E. Rumbaugh, L. Saenz, M. R. Salvaggio, W. C. Sanchez, V. M. Sanjana, S. Santiago, W. Schmidt, H. Schuitemaker, P. M. Sestak, P. Shalit, W. Shay, V. N. Shirvani, V. I. Silebi, J. M. Sizemore, Jr, P. R. Skolnik, M. Sokol-Anderson, J. M. Sosman, P. Stabile, J. T. Stapleton, S. Starrett, F. Stein, H.-J. Stellbrink, F. L. Sterman, V. E. Stone, D. R. Stone, G. Tambussi, R. A. Taplitz, E. M. Tedaldi, A. Telenti, W. Theisen, R. Torres, L. Tosiello, C. Tremblay, M. A. Tribble, P. D. Trinh, A. Tsao, P. Ueda, A. Vaccaro, E. Valadas, T. J. Vanig, I. Vecino, V. M. Vega, W. Veikley, B. H. Wade, C. Walworth, C. Wanidworanun, D. J. Ward, D. A. Warner, R. D. Weber, D. Webster, S. Weis, D. A. Wheeler, D. J. White, E. Wilkins, A. Winston, C. G. Wlodaver, A. van't Wout, D. P. Wright, O. O. Yang, D. L. Yurdin, B. W. Zabukovic, K. C. Zachary, B. Zeeman and M. Zhao, "The major genetic determinants of HIV-1 control affect HLA class I peptide presentation", Science **330**, 6010, 1551–1557 (2010).

Janice Oh, H.-L., S. Ken-En Gan, A. Bertoletti and Y.-J. Tan, "Understanding the t cell immune response in sars coronavirus infection", Emerging microbes & infections **1**, 1, 1–6 (2012).

Jo, S., T. Kim, V. G. Iyer and W. Im, "CHARMM-GUI: a web-based graphical user interface for CHARMM", J. Comput. Chem. **29**, 11, 1859–1865 (2008).

Jurrus, E., D. Engel, K. Star, K. Monson, J. Brandi, L. E. Felberg, D. H. Brookes, L. Wilson, J. Chen, K. Liles, M. Chun, P. Li, D. W. Gohara, T. Dolinsky, R. Konecny, D. R. Koes, J. E. Nielsen, T. Head-Gordon, W. Geng, R. Krasny, G.-W. Wei, M. J. Holst, J. A. McCammon and N. A. Baker, "Improvements to the APBS biomolecular solvation software suite", Protein Sci. **27**, 1, 112–128 (2018).

Jurtz, V., S. Paul, M. Andreatta, P. Marcatili, B. Peters and M. Nielsen, "Netmhcpan-4.0: improved peptide–mhc class i interaction predictions integrating eluted ligand

and peptide binding affinity data", The Journal of Immunology **199**, 9, 3360–3368 (2017).

Kassambara and Mundt, "Package 'factoextra"', R stats (2017).

Kim, C., D.-K. Ryu, J. Lee, Y.-I. Kim, J.-M. Seo, Y.-G. Kim, J.-H. Jeong, M. Kim, J.-I. Kim, P. Kim, J. S. Bae, E. Y. Shim, M. S. Lee, M. S. Kim, H. Noh, G.-S. Park, J. S. Park, D. Son, Y. An, J. N. Lee, K.-S. Kwon, J.-Y. Lee, H. Lee, J.-S. Yang, K.-C. Kim, S. S. Kim, H.-M. Woo, J.-W. Kim, M.-S. Park, K.-M. Yu, S.-M. Kim, E.-H. Kim, S.-J. Park, S. T. Jeong, C. H. Yu, Y. Song, S. H. Gu, H. Oh, B.-S. Koo, J. J. Hong, C.-M. Ryu, W. B. Park, M.-D. Oh, Y. K. Choi and S.-Y. Lee, "A therapeutic neutralizing antibody targeting receptor binding domain of SARS-CoV-2 spike protein", Nat. Commun. **12**, 1, 288 (2021).

Kleinjung, J. and F. Fraternali, "Design and application of implicit solvent models in biomolecular simulations", Curr. Opin. Struct. Biol. **25**, 126–134 (2014).

Lan Zhang, Clauser, Hacohen, Carr and others, "A large peptidome dataset improves HLA class I epitope prediction across most of the human population", Nature (2020).

Le Bert, N., A. T. Tan, K. Kunasegaran, C. Y. Tham, M. Hafezi, A. Chia, M. H. Y. Chng, M. Lin, N. Tan, M. Linster *et al.*, "Sars-cov-2-specific t cell immunity in cases of covid-19 and sars, and uninfected controls", Nature **584**, 7821, 457–462 (2020).

Lee, W.-R., W.-J. Syu, B. Du, M. Matsuda, S. Tan, A. Wolf, M. Essex and T.-H. Lee, "Nonrandom distribution of gp120 n-linked glycosylation sites important for infectivity of human immunodeficiency virus type 1.", Proceedings of the National Academy of Sciences **89**, 6, 2213–2217 (1992).

Leidner, R., N. Sanjuan Silva, H. Huang, D. Sprott, C. Zheng, Y.-P. Shih, A. Leung, R. Payne, K. Sutcliffe, J. Cramer, S. A. Rosenberg, B. A. Fox, W. J. Urba and E. Tran, "Neoantigen T-Cell receptor gene therapy in pancreatic cancer", N. Engl. J. Med. **386**, 22, 2112–2119 (2022).

Li, Q., X. Guan, P. Wu, X. Wang, L. Zhou, Y. Tong, R. Ren, K. S. Leung, E. H. Lau, J. Y. Wong *et al.*, "Early transmission dynamics in wuhan, china, of novel coronavirus–infected pneumonia", New England Journal of Medicine (2020).

Lin, M., H.-K. Tseng, J. A. Trejaut, H.-L. Lee, J.-H. Loo, C.-C. Chu, P.-J. Chen, Y.-W. Su, K. H. Lim, Z.-U. Tsai *et al.*, "Association of hla class i with severe acute respiratory syndrome coronavirus infection", BMC Medical Genetics **4**, 1, 9 (2003).

Marchalonis, J. J. and V. X. Gledhill, "Elementary stochastic model for the induction of immunity and tolerance", Nature **220**, 5167, 608–611 (1968).

Matzaraki, V., V. Kumar, C. Wijmenga and A. Zhernakova, "The mhc locus and genetic susceptibility to autoimmune and infectious diseases", Genome biology **18**, 1, 76 (2017).

McIntosh-Smith, S., J. Price, R. B. Sessions and A. A. Ibarra, "High performance in silico virtual drug screening on many-core processors", Int. J. High Perform. Comput. Appl. **29**, 2, 119–134 (2015).

Miller, J., A. H. E. Marshall and R. G. White, "The immunological significance of thymus. advances in immunology", (1962).

Mishra, S. K. and J. Koča, "Assessing the performance of MM/PBSA, MM/GBSA, and QM–MM/GBSA approaches on Protein/Carbohydrate complexes: Effect of implicit solvent models, QM methods, and entropic contributions", J. Phys. Chem. B **122**, 34, 8113–8121 (2018).

More, A. S., R. T. Toth, 4th, S. Z. Okbazghi, C. R. Middaugh, S. B. Joshi, T. J. Tolbert, D. B. Volkin and D. D. Weis, "Impact of glycosylation on the local backbone flexibility of Well-Defined IgG1-Fc glycoforms using hydrogen Exchange-Mass spectrometry", J. Pharm. Sci. **107**, 9, 2315–2324 (2018).

Ng, M., S. Cheng, K. Lau, G. Leung, U. Khoo, B. C.-y. Zee and J. Sung, "Immunogenetics in sars: a case-control study.", Hong Kong medical journal= Xianggang yi xue za zhi **16**, 5 Suppl 4, 29 (2010).

Ng, M. H., K.-M. Lau, L. Li, S.-H. Cheng, W. Y. Chan, P. K. Hui, B. Zee, C.-B. Leung and J. J. Sung, "Association of human-leukocyte-antigen class i (b* 0703) and class ii (drb1* 0301) genotypes with susceptibility and resistance to the development of severe acute respiratory syndrome", Journal of Infectious Diseases **190**, 3, 515–518 (2004).

Ng, O.-W., A. Chia, A. T. Tan, R. S. Jadi, H. N. Leong, A. Bertoletti and Y.-J. Tan, "Memory t cell responses targeting the sars coronavirus persist up to 11 years post-infection", Vaccine **34**, 17, 2008–2014 (2016).

Nguyen, A., J. K. David, S. K. Maden, M. A. Wood, B. R. Weeder, A. Nellore and R. F. Thompson, "Human leukocyte antigen susceptibility map for sars-cov-2", Journal of Virology (2020).

Nguyen, A. T., C. Szeto and S. Gras, "The pockets guide to HLA class I molecules", Biochem. Soc. Trans. **49**, 5, 2319–2331 (2021).

Nichols, K., "False discovery rate procedures", in "Statistical Parametric Mapping", pp. 246–252 (Elsevier, 2007).

Nielsen, M., M. Andreatta, B. Peters and S. Buus, "Immunoinformatics: Predicting peptide–mhc binding", Annual Review of Biomedical Data Science **3** (2020).

Nielsen, M., C. Lundegaard, T. Blicher, K. Lamberth, M. Harndahl, S. Justesen, G. Røder, B. Peters, A. Sette, O. Lund and S. Buus, "NetMHCpan, a method for quantitative predictions of peptide binding to any HLA-A and -B locus protein of known sequence", PLoS One **2**, 8, e796 (2007).

Nivón, L. G., R. Moretti and D. Baker, "A pareto-optimal refinement method for protein design scaffolds", PLoS One **8**, 4, e59004 (2013).

Nowak, M. A. and C. R. Bangham, "Population dynamics of immune responses to persistent viruses", Science **272**, 5258, 74–79 (1996).

O'Donnell, T. J., A. Rubinsteyn and U. Laserson, "MHCflurry 2.0: Improved Pan-Allele prediction of MHC class I-Presented peptides by incorporating antigen processing", Cell Syst **11**, 4, 418–419 (2020).

Ovchinnikov, V. and M. Karplus, "Analysis and elimination of a bias in targeted molecular dynamics simulations of conformational transitions: application to calmodulin", The Journal of Physical Chemistry B **116**, 29, 8584–8603 (2012).

O'Donnell, T. J., A. Rubinsteyn and U. Laserson, "Mhcflurry 2.0: Improved pan-allele prediction of mhc class i-presented peptides by incorporating antigen processing", Cell Systems (2020).

Pagès, Aboyoun, Gentleman and DebRoy, "Biostrings: Efficient manipulation of biological strings", R package version (2019).

Parham, Rossjohn, Vivian and Purcell, "HLA-B57 micropolymorphism defines the sequence and conformational breadth of the immunopeptidome", Nature (2018).

Park, H., P. Bradley, P. Greisen, Jr, Y. Liu, V. K. Mulligan, D. E. Kim, D. Baker and F. DiMaio, "Simultaneous optimization of biomolecular energy functions on features from small molecules and macromolecules", J. Chem. Theory Comput. **12**, 12, 6201–6212 (2016).

Park, S.-J., J. Lee, Y. Qi, N. R. Kern, H. S. Lee, S. Jo, I. Joung, K. Joo, J. Lee and W. Im, "CHARMM-GUI glycan modeler for modeling and simulation of carbohydrates and glycoconjugates", Glycobiology **29**, 4, 320–331 (2019).

Paul, S., D. Weiskopf, M. A. Angelo, J. Sidney, B. Peters and A. Sette, "Hla class i alleles are associated with peptide-binding repertoires of different size, affinity, and immunogenicity", The Journal of Immunology **191**, 12, 5831–5839 (2013).

Peters, B., H.-H. Bui, S. Frankild, M. Nielson, C. Lundegaard, E. Kostem, D. Basch, K. Lamberth, M. Harndahl, W. Fleri, S. S. Wilson, J. Sidney, O. Lund, S. Buus and A. Sette, "A community resource benchmarking predictions of peptide binding to MHC-I molecules", PLoS Comput. Biol. **2**, 6, e65 (2006).

Phillips, J. C., R. Braun, W. Wang, J. Gumbart, E. Tajkhorshid, E. Villa, C. Chipot, R. D. Skeel, L. Kalé and K. Schulten, "Scalable molecular dynamics with NAMD", J. Comput. Chem. **26**, 16, 1781–1802 (2005).

Phillips, J. C., D. J. Hardy, J. D. C. Maia, J. E. Stone, J. V. Ribeiro, R. C. Bernardi, R. Buch, G. Fiorin, J. Hénin, W. Jiang, R. McGreevy, M. C. R. Melo, B. K. Radak, R. D. Skeel, A. Singharoy, Y. Wang, B. Roux, A. Aksimentiev, Z. Luthey-Schulten, L. V. Kalé, K. Schulten, C. Chipot and E. Tajkhorshid, "Scalable molecular dynamics on CPU and GPU architectures with NAMD", J. Chem. Phys. **153**, 4, 044130 (2020).

Prachar, M., S. Justesen, D. B. Steen-Jensen, S. P. Thorgrimsen, E. Jurgons, O. Winther and F. O. Bagger, "Covid-19 vaccine candidates: Prediction and validation of 174 sars-cov-2 epitopes", bioRxiv (2020).

Prasad, V. M., D. P. Leaman, K. N. Lovendahl, J. T. Croft, M. A. Benhaim, E. A. Hodge, M. B. Zwick and K. K. Lee, "Cryo-et of env on intact hiv virions reveals structural variation and positioning on the gag lattice", Cell **185**, 4, 641–653 (2022).

Prates, E. T., X. Guan, Y. Li, X. Wang, P. K. Chaffey, M. S. Skaf, M. F. Crowley, Z. Tan and G. T. Beckham, "The impact of o-glycan chemistry on the stability of intrinsically disordered proteins", Chem. Sci. **9**, 15, 3710–3715 (2018).

Quemin, E. R., E. A. Machala, B. Vollmer, V. Pražák, D. Vasishtan, R. Rosch, M. Grange, L. E. Franken, L. A. Baker and K. Grünewald, "Cellular electron cryo-tomography to study virus-host interactions", Annual Review of Virology **7**, 239–262 (2020).

Rammensee, H., J. Bachmann, N. P. Emmerich, O. A. Bachor and S. Stevanović, "SYFPEITHI: database for MHC ligands and peptide motifs", Immunogenetics **50**, 3-4, 213–219 (1999).

Rapin, N., I. Hoof, O. Lund and M. Nielsen, "The mhc motif viewer: a visualization tool for mhc binding motifs", Current protocols in immunology **88**, 1, 18–17 (2010).

Rasmussen, M., E. Fenoy, M. Harndahl, A. B. Kristensen, I. K. Nielsen, M. Nielsen and S. Buus, "Pan-specific prediction of peptide–mhc class i complex stability, a correlate of t cell immunogenicity", The Journal of Immunology **197**, 4, 1517–1524 (2016).

Reynisson, B., B. Alvarez, S. Paul, B. Peters and others, "NetMHCpan-4.1 and NetMHCIIpan-4.0: improved predictions of MHC antigen presentation by concurrent motif deconvolution and integration of MS MHC eluted ligand . . . ", Nucleic acids (2020).

Robinson, J., D. J. Barker, X. Georgiou, M. A. Cooper, P. Flicek and S. G. E. Marsh, "IPD-IMGT/HLA database", Nucleic Acids Res. **48**, D1, D948–D955 (2020).

Robinson, J., A. Malik, P. Parham, J. Bodmer and S. Marsh, "Imgt/hla database–a sequence database for the human major histocompatibility complex", Tissue antigens **55**, 3, 280–287 (2000).

Rock, K. L., E. Reits and J. Neefjes, "Present yourself! by MHC class I and MHC class II molecules", Trends Immunol. **37**, 11, 724–737 (2016).

Rolland, M., D. Heckerman, W. Deng, C. M. Rousseau, H. Coovadia, K. Bishop, P. J. Goulder, B. D. Walker, C. Brander and J. I. Mullins, "Broad and gag-biased hiv-1 epitope repertoires are associated with lower viral loads", PloS one **3**, 1 (2008).

Rose, P. W., A. Prlić, A. Altunkaya, C. Bi, A. R. Bradley, C. H. Christie, L. D. Costanzo, J. M. Duarte, S. Dutta, Z. Feng, R. K. Green, D. S. Goodsell, B. Hudson, T. Kalro, R. Lowe, E. Peisach, C. Randle, A. S. Rose, C. Shao, Y.-P. Tao, Y. Valasatava, M. Voigt, J. D. Westbrook, J. Woo, H. Yang, J. Y. Young, C. Zardecki, H. M. Berman and S. K. Burley, "The RCSB protein data bank: integrative view of protein, gene and 3D structural information", Nucleic Acids Res. **45**, D1, D271–D281 (2017).

Ross, R., "An application of the theory of probabilities to the study of a priori pathometry.—part I", Proc. R. Soc. Lond. A Math. Phys. Sci. **92**, 638, 204–230 (1916).

Roy, Poddar and Kar, "Conformational preferences of an N-Glycan in aqueous solution: A case study for evaluating the accuracy of generalized born model", Available at SSRN 4113756 (4113).

Ryckaert, J.-P., G. Ciccotti and H. J. C. Berendsen, "Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes", J. Comput. Phys. **23**, 3, 327–341 (1977).

Sanchez-Mazas, A., "Hla studies in the context of coronavirus outbreaks", Swiss Medical Weekly **150**, 1516 (2020).

Santiveri, M., A. Roa-Eguiara, C. Kühne, N. Wadhwa, H. Hu, H. C. Berg, M. Erhardt and N. M. I. Taylor, "Structure and function of stator units of the bacterial flagellar motor", Cell **183**, 1, 244–257.e16 (2020).

Sarkizova, S., S. Klaeger, P. M. Le, L. W. Li, G. Oliveira, H. Keshishian, C. R. Hartigan, W. Zhang, D. A. Braun, K. L. Ligon *et al.*, "A large peptidome dataset improves hla class i epitope prediction across most of the human population", Nature Biotechnology **38**, 2, 199–209 (2020).

Shah, P., G. A. Canziani, E. P. Carter and I. Chaiken, "The case for s2: the potential benefits of the s2 subunit of the sars-cov-2 spike protein as an immunogen in fighting the covid-19 pandemic", Frontiers in immunology **12**, 637651 (2021).

Shore, D. A., I. A. Wilson, R. A. Dwek and P. M. Rudd, "Glycosylation and the function of the t cell co-receptor cd8", pp. 71–84 (2005).

Sidney, J., B. Peters, N. Frahm, C. Brander and A. Sette, "HLA class I supertypes: a revised and updated classification", BMC Immunol. **9**, 1 (2008).

Singharoy, A., I. Teo, R. McGreevy, J. E. Stone, J. Zhao and K. Schulten, "Molecular dynamics-based refinement and validation for sub-5 å cryo-electron microscopy maps", Elife **5** (2016).

Smith, C. A. and T. Kortemme, "Backrub-like backbone simulation recapitulates natural protein conformational variability and improves mutant side-chain prediction", J. Mol. Biol. **380**, 4, 742–756 (2008).

Sun, S. Y., L.-a. Segev-Zarko, M. Chen, G. D. Pintilie, M. F. Schmid, S. J. Ludtke, J. C. Boothroyd and W. Chiu, "Cryo-et of toxoplasma parasites gives subnanometer insight into tubulin-based structures", Proceedings of the National Academy of Sciences **119**, 6, e2111661119 (2022).

Sundberg, E. J., L. Deng and R. A. Mariuzza, "TCR recognition of peptide/MHC class II complexes and superantigens", Semin. Immunol. **19**, 4, 262–271 (2007).

Szegedy, Liu, Jia, Sermanet and others, "Going deeper with convolutions", Proc. Estonian Acad. Sci. Biol. Ecol. (2015).

Team, R. C., "R: A language and environment for statistical computing (version 4.0. 2)", R Foundation for Statistical Computing (2020).

Trolle, T., C. P. McMurtrey, J. Sidney, W. Bardet, S. C. Osborn, T. Kaever, A. Sette, W. H. Hildebrand, M. Nielsen and B. Peters, "The length distribution of class i–restricted t cell epitopes is determined by both peptide supply and mhc allele–specific binding preference", The Journal of Immunology **196**, 4, 1480–1487 (2016).

Turoňová, B., M. Sikora, C. Schürmann, W. J. H. Hagen, S. Welsch, F. E. C. Blanc, S. von Bülow, M. Gecht, K. Bagola, C. Hörner, G. van Zandbergen, J. Landry, N. T. D. de Azevedo, S. Mosalaganti, A. Schwarz, R. Covino, M. D. Mühlebach, G. Hummer, J. Krijnse Locker and M. Beck, "In situ structural analysis of SARS-CoV-2 spike reveals flexibility mediated by three hinges", Science **370**, 6513, 203–208 (2020).

TW, T., Y. A. E. IM M and O. LO, "The next quarter century", Immunity **50** (2019).

Vaitaitis, G. M. and D. H. Wagner Jr, "Cd40 glycoforms and tnf-receptors 1 and 2 in the formation of cd40 receptor (s) in autoimmunity", Molecular immunology **47**, 14, 2303–2313 (2010).

Vita, R., S. Mahajan, J. A. Overton, S. K. Dhanda, S. Martini, J. R. Cantrell, D. K. Wheeler, A. Sette and B. Peters, "The immune epitope database (IEDB): 2018 update", Nucleic Acids Res. **47**, D1, D339–D343 (2019).

Vogel, A. B., I. Kanevsky, Y. Che, K. A. Swanson, A. Muik, M. Vormehr, L. M. Kranz, K. C. Walzer, S. Hein, A. Güler, J. Loschko, M. S. Maddur, A. Ota-Setlik, K. Tompkins, J. Cole, B. G. Lui, T. Ziegenhals, A. Plaschke, D. Eisel, S. C. Dany, S. Fesser, S. Erbar, F. Bates, D. Schneider, B. Jesionek, B. Sänger, A.-K. Wallisch, Y. Feuchter, H. Junginger, S. A. Krumm, A. P. Heinen, P. Adams-Quack, J. Schlereth, S. Schille, C. Kröner, R. de la Caridad Güimil Garcia, T. Hiller, L. Fischer, R. S. Sellers, S. Choudhary, O. Gonzalez, F. Vascotto, M. R. Gutman, J. A. Fontenot, S. Hall-Ursone, K. Brasky, M. C. Griffor, S. Han, A. A. H. Su, J. A. Lees, N. L. Nedoma, E. H. Mashalidis, P. V. Sahasrabudhe, C. Y. Tan, D. Pavliakova, G. Singh, C. Fontes-Garfias, M. Pride, I. L. Scully, T. Ciolino, J. Obregon, M. Gazi, R. Carrion, Jr, K. J. Alfson, W. V. Kalina, D. Kaushal, P.-Y. Shi, T. Klamp, C. Rosenbaum, A. N. Kuhn, Ö. Türeci, P. R. Dormitzer, K. U. Jansen and U. Sahin, "BNT162b vaccines protect rhesus macaques from SARS-CoV-2", Nature **592**, 7853, 283–289 (2021).

Walls, A. C., M. A. Tortorici, B. Frenz, J. Snijder, W. Li, F. A. Rey, F. DiMaio, B.-J. Bosch and D. Veesler, "Glycan shield and epitope masking of a coronavirus spike protein observed by cryo-electron microscopy", Nat. Struct. Mol. Biol. **23**, 10, 899–905 (2016).

Wang, C., R. van Haperen, J. Gutiérrez-Álvarez, W. Li, N. Okba, I. Albulescu, I. Widjaja, B. van Dieren, R. Fernandez-Delgado, I. Sola *et al.*, "A conserved immunogenic and vulnerable site on the coronavirus spike protein delineated by cross-reactive monoclonal antibodies", Nature communications **12**, 1, 1–15 (2021).

Wang, S.-F., K.-H. Chen, M. Chen, W.-Y. Li, Y.-J. Chen, C.-H. Tsao, M.-y. Yen, J. C. Huang and Y.-M. A. Chen, "Human-leukocyte antigen class i cw 1502 and class ii dr 0301 genotypes are associated with resistance to severe acute respiratory syndrome (sars) infection", Viral immunology **24**, 5, 421–426 (2011).

Williamson, E. J., A. J. Walker, K. Bhaskaran, S. Bacon, C. Bates, C. E. Morton, H. J. Curtis, A. Mehrkar, D. Evans, P. Inglesby *et al.*, "Opensafely: factors associated with covid-19 death in 17 million patients", Nature pp. 1–11 (2020).

Wilson, E. A., G. Hirneise, A. Singharoy and K. S. Anderson, "Total predicted MHC-I epitope load is inversely associated with population mortality from SARS-CoV-2", Cell Rep Med **2**, 3, 100221 (2021).

Woo, H., S.-J. Park, Y. K. Choi, T. Park, M. Tanveer, Y. Cao, N. R. Kern, J. Lee, M. S. Yeom, T. I. Croll, C. Seok and W. Im, "Developing a fully glycosylated Full-Length SARS-CoV-2 spike protein model in a viral membrane", J. Phys. Chem. B **124**, 33, 7128–7137 (2020).

Wood, C. W., J. W. Heal, A. R. Thomson, G. J. Bartlett, A. Á. Ibarra, R. L. Brady, R. B. Sessions and D. N. Woolfson, "ISAMBARD: an open-source computational environment for biomolecular analysis, modelling and design", Bioinformatics **33**, 19, 3043–3050 (2017).

Wood, C. W. and D. N. Woolfson, "CCBuilder 2.0: Powerful and accessible coiled-coil modeling", Protein Sci. **27**, 1, 103–111 (2018).

Wu, E. and G. R. Nemerow, "Virus yoga: the role of flexibility in virus host cell recognition", Trends in microbiology **12**, 4, 162–169 (2004).

Wu, S. and Y. Zhang, "LOMETS: a local meta-threading-server for protein structure prediction", Nucleic Acids Res. **35**, 10, 3375–3382 (2007).

Wu, W.-L., C.-Y. Chiang, S.-C. Lai, C.-Y. Yu, Y.-L. Huang, H.-C. Liao, C.-L. Liao, H.-W. Chen and S.-J. Liu, "Monoclonal antibody targeting the conserved region of the sars-cov-2 spike protein to overcome viral variants", JCI insight **7**, 8 (2022).

Yang, Y., C. Liu, L. Du, S. Jiang, Z. Shi, R. S. Baric and F. Li, "Two mutations were critical for bat-to-human transmission of middle east respiratory syndrome coronavirus", J. Virol. **89**, 17, 9119–9123 (2015).

Zacharakis, N., H. Chinnasamy, M. Black, H. Xu, Y.-C. Lu, Z. Zheng, A. Pasetto, M. Langhan, T. Shelton, T. Prickett, J. Gartner, L. Jia, K. Trebska-McGowan, R. P. Somerville, P. F. Robbins, S. A. Rosenberg, S. L. Goff and S. A. Feldman, "Immune recognition of somatic mutations leading to complete durable regression in metastatic breast cancer", Nat. Med. **24**, 6, 724–730 (2018).

Zhang, C., W. Zheng, X. Huang, E. W. Bell, X. Zhou and Y. Zhang, "Protein structure and sequence re-analysis of 2019-ncov genome refutes snakes as its intermediate host or the unique similarity between its spike protein insertions and hiv-1", Journal of proteome research (2020).

Zhang, H., O. Lund and M. Nielsen, "The pickpocket method for predicting binding specificities for receptors based on receptor pocket similarities: application to mhc-peptide binding", Bioinformatics **25**, 10, 1293–1299 (2009).

Zhang, Y., A. Kolinski and J. Skolnick, "TOUCHSTONE II: a new approach to ab initio protein structure prediction", Biophys. J. **85**, 2, 1145–1164 (2003).

Zhang, Y. and J. Skolnick, "SPICKER: A clustering approach to identify near-native protein folds", J. Comput. Chem. (2004).

Zhao, W. and X. Sher, "Systematically benchmarking peptide-mhc binding predictors: From synthetic to naturally processed epitopes", PLoS computational biology **14**, 11 (2018).

Zu, Z. Y., M. D. Jiang, P. P. Xu, W. Chen, Q. Q. Ni, G. M. Lu and L. J. Zhang, "Coronavirus disease 2019 (covid-19): a perspective from china", Radiology p. 200490 (2020).

APPENDIX A

TOTAL PREDICTED MHC-I EPITOPE LOAD IS INVERSELY ASSOCIATED WITH POPULATION MORTALITY FROM SARS-COV-2: SUPPLEMENTAL MATERIAL

Permission for use

Table A.1: MHC-I peptides identified by EnsembleMHC. All predicted peptides with a peptideFDR 0.05.(Full table can be found at https://doi.org/10.1016/j.xcrm.2021.100221)

A

Countries
China
Japan
South Korea
Taiwan
US
Hong Kong
France
Germany
India
Italy
Russia
UK
Iran
Israel
Croatia
Romania
Netherlands
Mexico
Ireland
Czechia
Morocco

B

Minimum death threshold	Normalized day: 0.25	Normalized day: 0.5	Normalized day: 0.75	Normalized day: 1
5	8	15	23	30
10	7	14	20	27
15	7	13	20	26
20	7	13	20	26
25	7	12	18	24
30	7	12	17	23
35	7	12	17	23
40	7	12	17	23
45	6	11	17	22
50	6	11	17	22
55	6	11	17	22
60	6	11	16	21
65	6	11	16	21
70	6	11	16	21
75	6	11	16	21
80	6	11	15	20
85	6	11	15	20
90	6	11	15	20
95	5	10	14	19
100	5	10	14	19

non-normalized days

Table A.2: Countries included in analysis and normalized day to real day mapping. A. The 23 countries for which SARS-CoV-2 population binding capacities were calculated. **B.** The mapping of normalized days to real days for normalized day quartiles (0.25, 0.5, 0.75, 1) at select minimum death thresholds.

Table A.3: EMP score correlation data. All data pertaining to the correlations between EMP score and deaths per million. This includes rho estimates, 95% CI, non-normalized day values, and sample size for each correlation.(Full table can be found at https://doi.org/10.1016/j.xcrm.2021.100221)

A

Countries
China
Japan
South Korea
US
France
Germany
India
Italy
Russia
UK
Iran
Israel
Croatia
Romania
Netherlands
Mexico
Ireland
Czechia
Morocco

B

Factor	Abbreviation	Description
% of population ≥ 65 years	65	Percentage of the population that is 65 years of age or older (2020).
Average BMI	Avg. BMI	The age-standardized average population body mass index (2016).
Cardiovascular disease	CD	The deaths per million due to cardiovascular disease (2016).
Chronic obstructive pulmonary disease	COPD	The deaths per million due to complications from chronic obstructive pulmonary disease (2016).
Diabetes mellitus	DM	The deaths per million due to complications from diabetes mellitus (2016).
High blood pressure	BP	The age-standardized percentage of the population with a systolic blood pressure ≥ 140 or diastolic blood pressure ≥ 90 (2015).
Obesity prevalence	OBS	The age-standardized percentage of the population with a BMI ≥ 30 (2016).
Overweight prevalence	OVW	The age-standardized percentage of the population with a BMI ≥ 25 (2016).
Structural protein EMP score	SP	The SARS-CoV-2 structural protein presentation score.
% of GDP spent on health care	GDP	Current health expenditure (CHE) as percentage of gross domestic product (2017).
% of total gov. expenditure on health care	GGHE	General government expenditure on health as a percentage of total (2014).
% of population that is female	SEX	The proportion of the total population that is female (2020).

Table A.4: Socioeconomic and health-related risk factors A. 21 countries were selected for analysis based on the existence of data in the Global Health Observatory data repository and inclusion in the 23 country set used for EMP score analysis. **B.** Descriptions and abbreviations for the selected risk factors. Each factor is labeled with the year that the data was collected. In every case, the most recent data was selected for analysis.

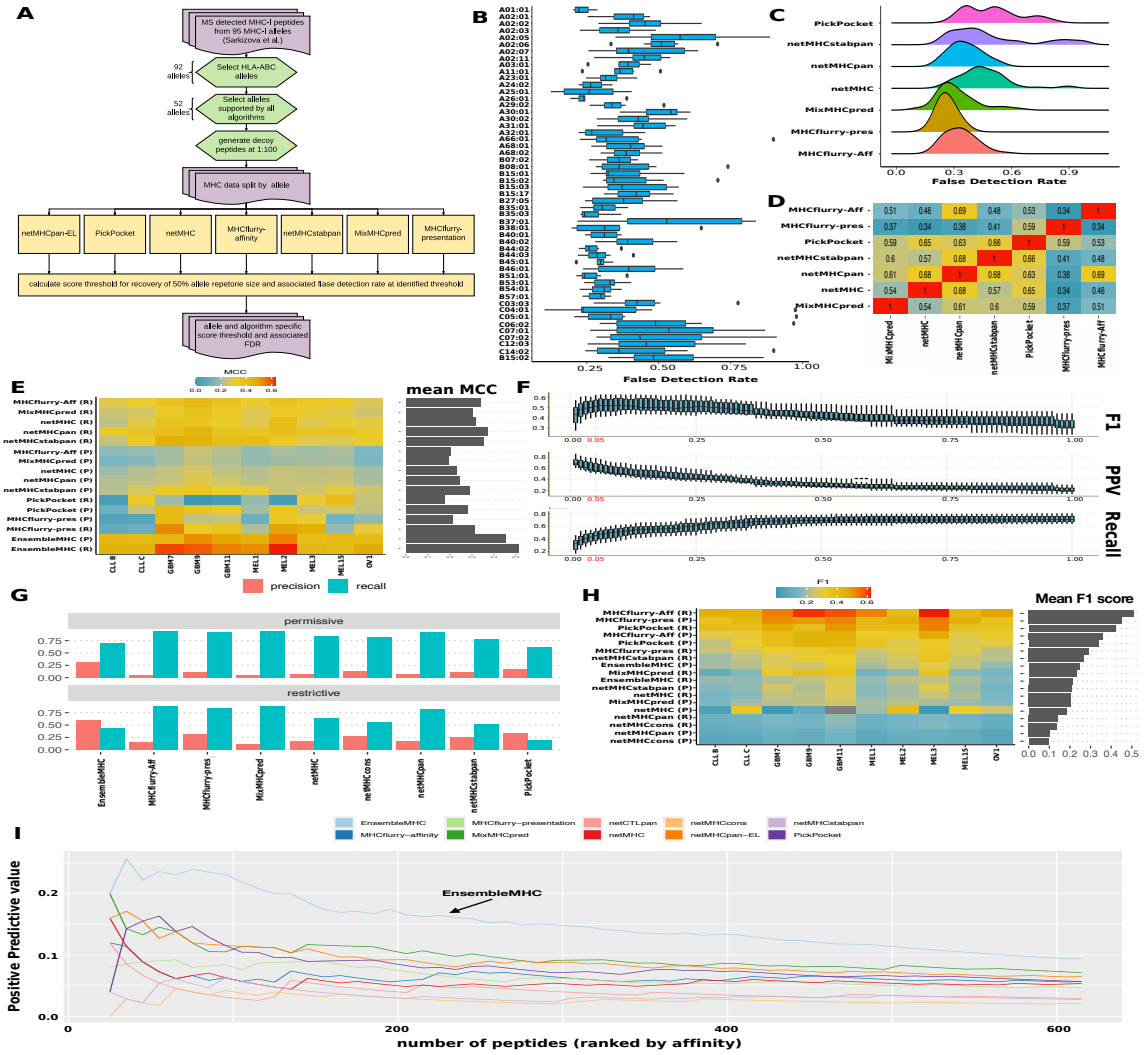

Figure A.1: EnsembleMHC Parameterization overview and viral peptide analysis. A. EnsembleMHC Parameterization workflow. **B.** The allele and algorithm score distrubtion at each allele (n=7) **C.** A density plot of the observed FDRs for each algorithm across all alleles (n = 52). **D.** The correlation between individual peptide scores for each algorithm across all alleles was calculated using Pearson correlation. Warmer colors indicate a higher level of correlation while cooler colors indicate lower correlation. **E.** Matthew's correlation coefficient was calculated for each algorithm. The average MCC for each algorithm is represented by the bar plot on the right margin. **F.** The effect of different $peptide^{FDR}$ cutoff thresholds on the results reported in **figure 1**. **G-H.** The analysis reported in **figure 1 (A-B)** were repeated with additional comparisons to consensus-based MHC-I prediction algorithms, namely netMHCconskarosiene2012netmhccons and netCTLpanstranzl2010netctlpan. **I.** The positive predictive value of each algorithm was calculated with respect to ability to identify immunogenic peptides derived from Hepatitis-C genome polyprotein, Dengue virus genome polyprotein, and the HIV-1 POL-GAG protein when selecting n number of top scoring peptides **(METHODS).**

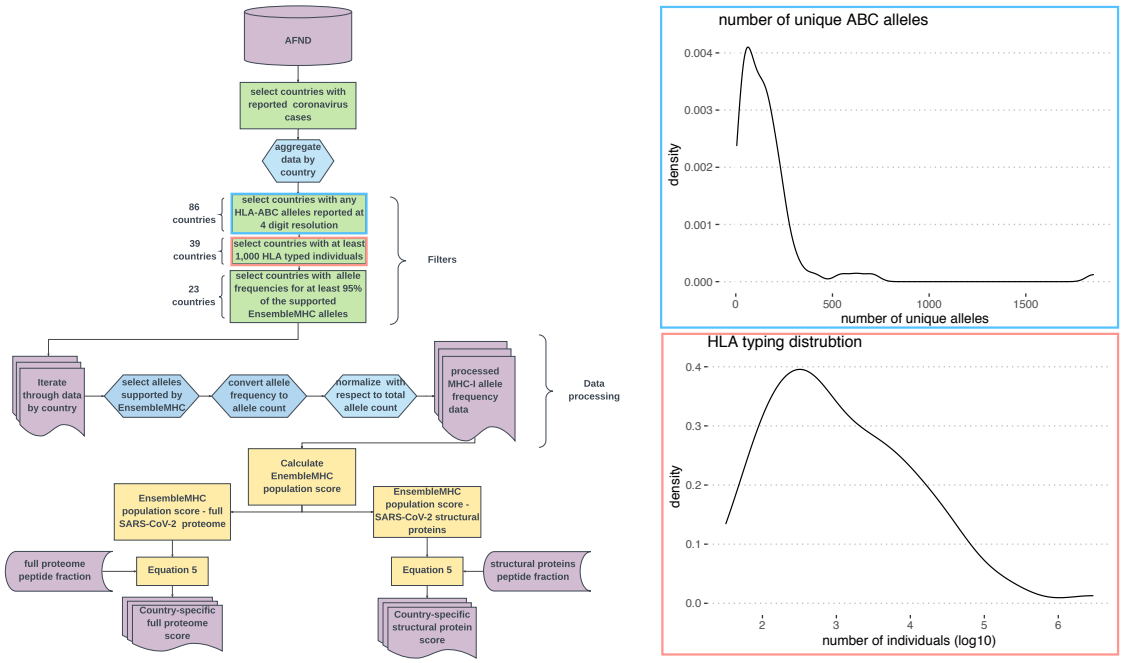

Figure A.2: Data processing and EnsembleMHC population score calculation workflow. The overview of the data processing steps used on the global MHC-I allele frequency data and the calculation of the EnsembleMHC population score with respect to the full SARS-CoV-2 proteome and SARS-CoV-2 structural proteins. **(inset plots)**, The blue inset plot illustrates MHC-typing breadth and depth variation by showing the distribution of the total number of MHC-I alleles reported at 4-digit resolution in 86 countries. The red inset plot shows the distribution of the number of MHC-genotyped individuals in the set of countries with at least 1 reported coronavirus case. **AFND = Allele Frequency Net Database**

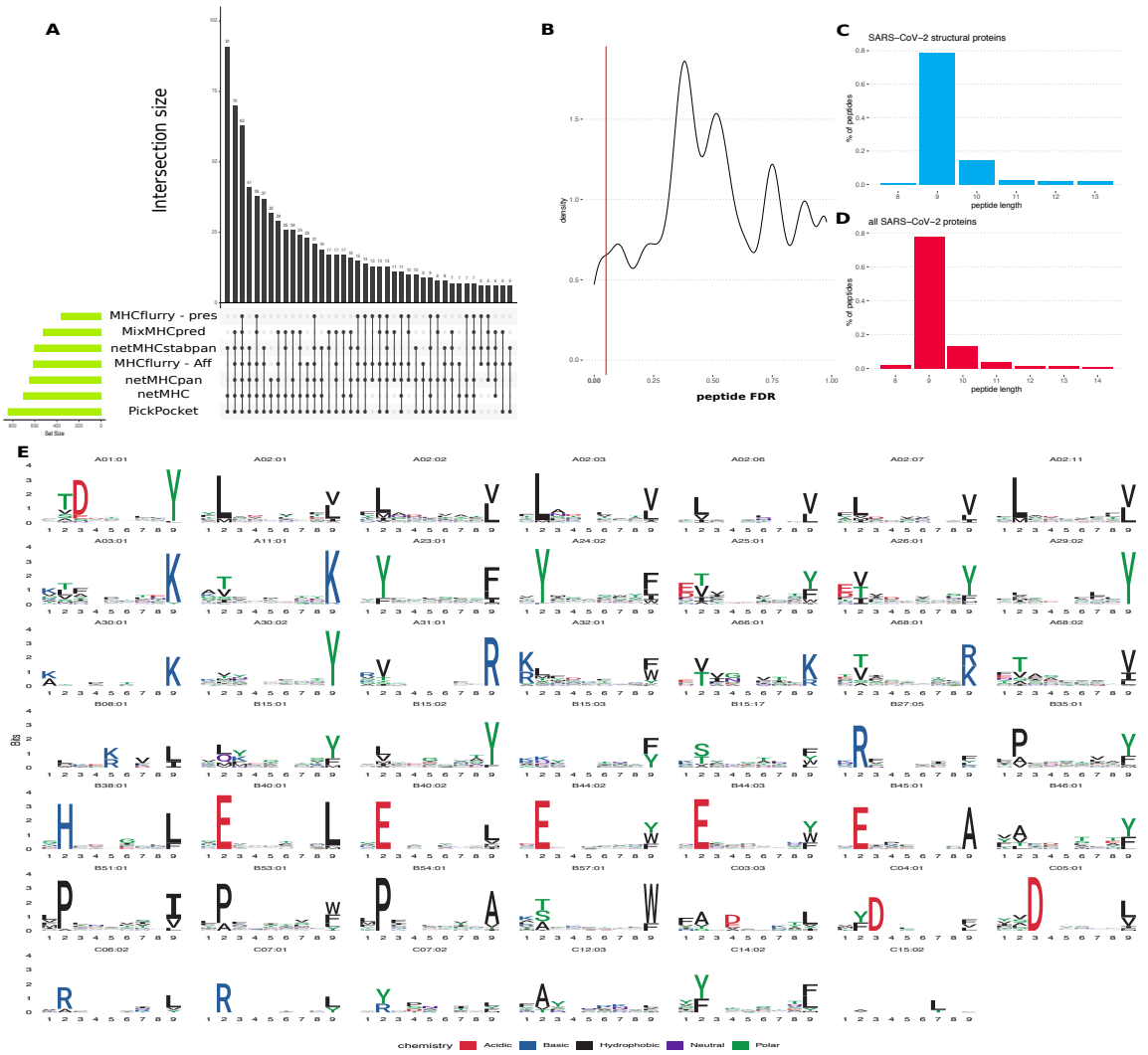

Figure A.3: Characteristics of peptides predicted by EnsembleMHC. A. The UpSet plot shows the contribution of each individual component algorithm to the 658 unique SARS-CoV-2 peptides identified by EnsembleMHC. The top bar plot indicates the number of unique peptides identified by the combination of algorithms shown by the points and segments located under each bar. The bar plot on the left-hand side of the plot indicates the total number of peptides identified by each algorithm. **B.** The $peptide^{FDR}$ distribution of the 9,712 SARS-CoV-2 peptides that fell with the score threshold of at least one component algorithm. The red line indicates an $peptide^{FDR}$ level of $\leq 5\%$. **C.** The length distribution of the 108 high-confidence peptides identified from SARS-CoV-2 structural proteins. **D.** The length distribution of the 658 high-confidence peptides identified from full SARS-CoV-2 proteome. **E.** Logo plots were generated for MHC alleles with at least 5 peptides identified by EnsembleMHC. Peptides shorter than 9 amino acids had random amino acid inserted into a non-anchor position while peptides longer than 9 amino acids had a random non-anchor position deleted. Large amino acid character height indicates a high frequency of that amino acid at that position. Amino acids are colored residue type.

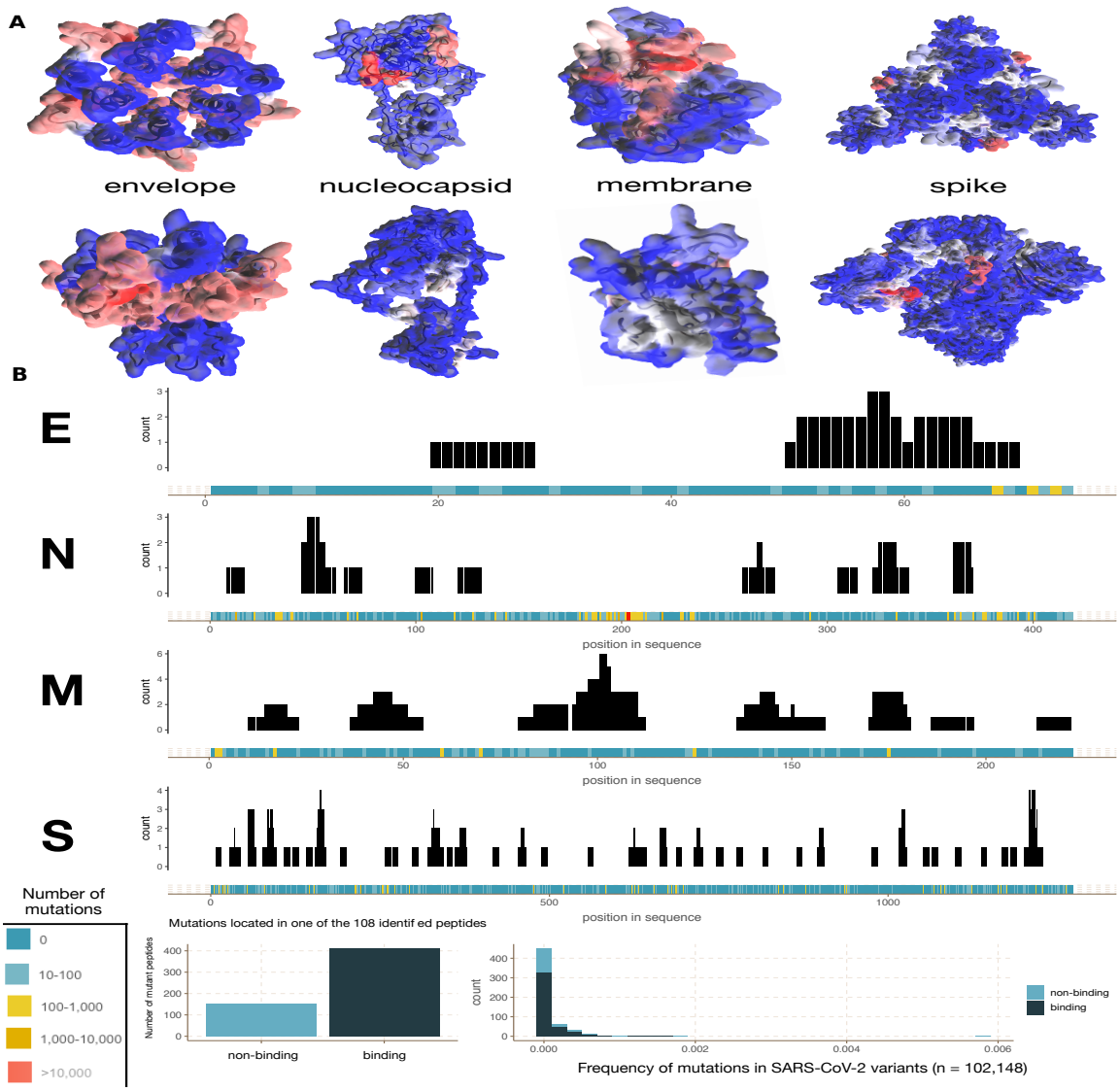

Figure A.4: Molecular origin of predicted SARS-CoV-2 structural protein MHC-I peptides and impact of sequence polymorphism. **A.** The predicted SARS-CoV-2 structural protein MHC-I peptides were mapped onto the solved structures for the envelope and spike proteins, and the predicted structures for the nucleocapsid and membrane proteins. Red highlighted regions indicate an enrichment of predicted peptides while blue regions indicate a depletion of predicted peptides. **B.** The incidence of protein sequence mutations (colored bar) and the frequency of that position in one of the 108 SARS-CoV-2 structural protein peptides (black bars) were calculated for 102,148 SARS-CoV-2 sequence variants. **Lower left panel,** all potential mutations arising in one of the 108 peptides identified by EnsembleMHC were evaluated for changes in binding affinity ($peptide^{FDR} > 0.05$). **Lower right panel,** The overall frequency of mutations impacting EnsmebleMHC-predicted peptides with light blue indicating deleterious mutations, and dark blue indicating neutral mutations.

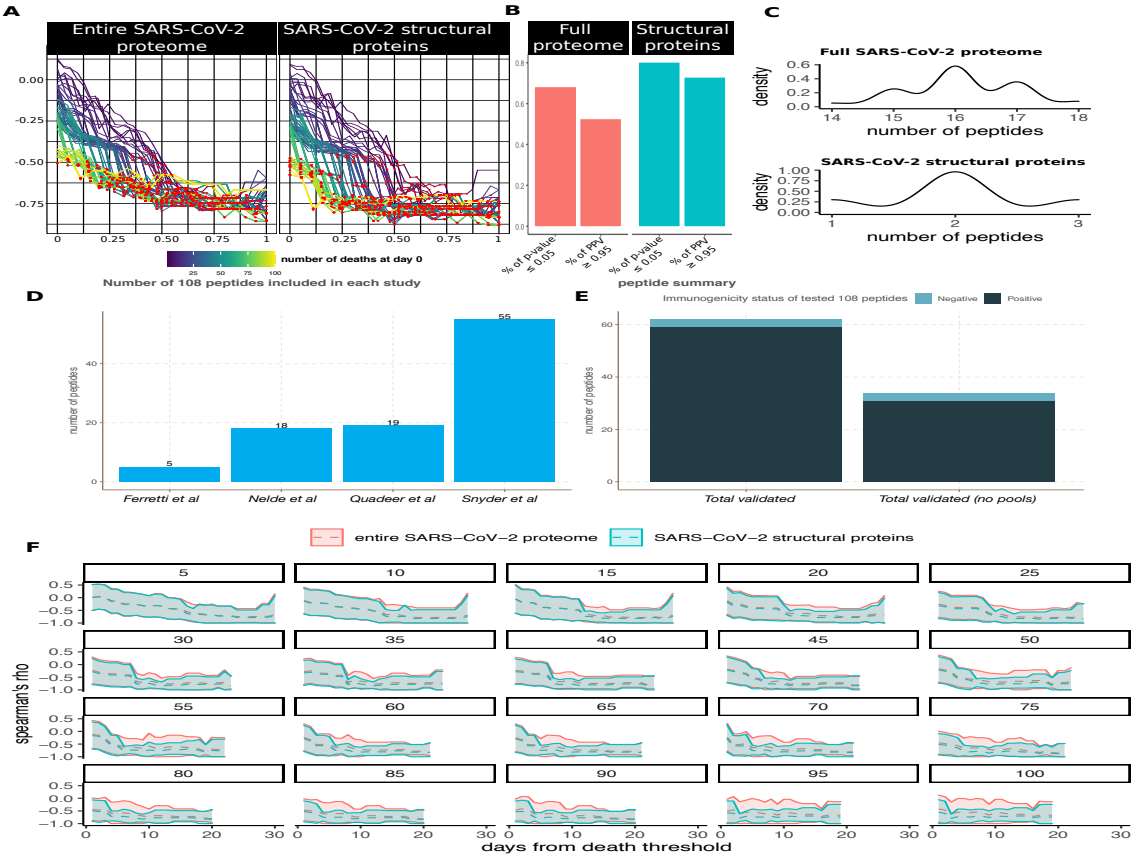

Figure A.5: Comparison of entire SARS-CoV-2 EnsembleMHC population score and structural protein EnsembleMHC population score. A. The correlations between EnsembleMHC population score based on the full SARS-CoV-2 proteome **(left)** or only SARS-CoV-2 structural proteins **(right)**. **B.** The difference in the proportions of significant p-values and PPV between the full SARS-CoV-2 proteome **(left)** and SARS-CoV-2 structural proteins **(right)** (not corrected for multiple testing). **C.** The SARS-CoV-2 peptide-MHC allele distribution resulting from uniform allele sampling. These distribution were used as the partner distributions for the Kolmogorov-smirnov test described in the results. **D.** 62 (57%) EnsembleMHC-identified SARS-CoV-2 structural protein peptides were included for testing in 4 different studies. **E.** The summary of immunogenicity status of tested EnsembleMHC peptides across all studies. These summaries were split into two groups. *Total validated* indicates the total number of experimentally validated peptides while *total validated (no pools)* indicates the number of experimentally validated peptides excluding those only tested in peptide pools. This distinction was made due to the potential of peptide pools to obscure which tested peptide is truly responsible for the observed immune response. **F.** Each individual plot shows the 95% confidence interval (shaded region) for the correlations between EMP scores based on the entire SARS-CoV-2 proteome (red) or SARS-CoV-2 structural proteins (blue) and observed deaths per million for different starting minimum death thresholds (indicated by number above plot).

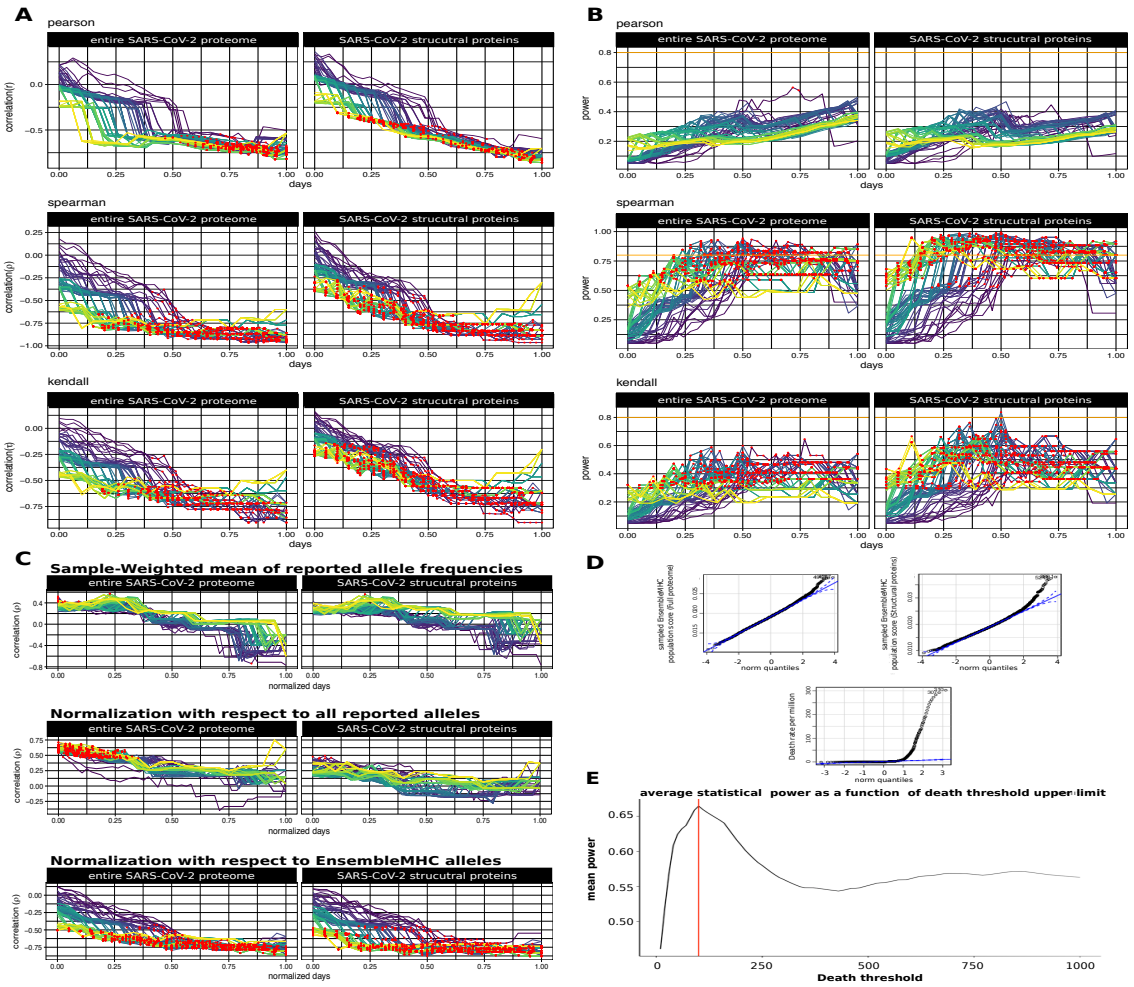

Figure A.6: Justification of statistical tests. A. The correlation between EnsembleMHC population score with respect to all SARS-CoV-2 proteins **(left column)** or SARS-CoV-2 structural proteins **(right columns)** and deaths per million using Pearson's r **(top)**, Spearman's rho **(middle)**, and Kendall's tau **(bottom)**. Correlations that were shown to be statistically significant are colored with a red point. **B.** The statistical power of each reported correlation. Correlations that were shown to be statistically significant are colored with a red point. The orange line indicates a power threshold of 80%. **C.** The effect of different allele frequency normalization techniques on the reported correlations between SARS-CoV-2 mortality and EMP scores based on the full SARS-CoV-2 proteome (left column) or SARS-CoV-2 structural proteins (right column). Definitions of nomralization methods can be seen in the methods **D.** QQ plots were generated from the respective distributions of the full proteome EnsembleMHC population scores, structural protein EnsembleMHC population scores, and deaths per million. **E.** The mean statistical power of all resulting correlations between EnsembleMHC population scores and observed deaths per million at different minimum reported death thresholds. The red line indicates a minimum death threshold of 100 deaths by day 0, the selected upper limit for analysis.

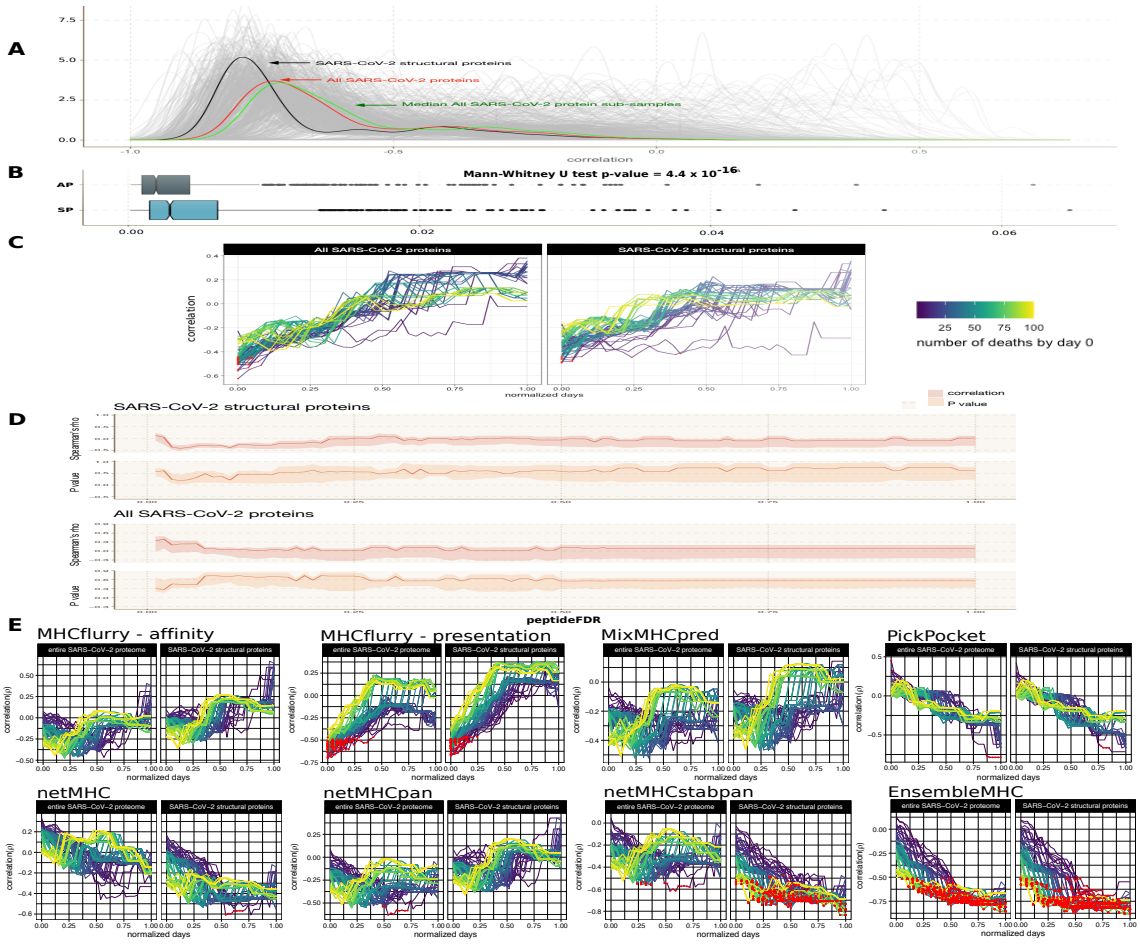

Figure A.7: Robustness of EMP score correlation analysis. A. 1,000 sub-sampling iterations were performed by randomly selecting 108 peptides from the full SARS-CoV-2 proteome that passed the 5% $peptide^{FDR}$ filter. The correlation between the population EMP score produced by each sub-sampled set of peptides and observed deaths per million were plotted (grey lines). The correlation distribution observed for identified SARS-CoV-2 structural protein peptides (black line), all SARS-CoV-2 proteins (red line), and the median correlation distribution across all subsampling iterations (green line) were plotted for comparison. **B.** Kullback-Leibler divergence was calculated for the correlation distribution of each down sample iteration relative to either the correlation distribution of the all peptide group (AP) or the structural peptide group (SP). **C.** The MHC-I allele assessment of peptides that passed an individual algorithm binding affinity thresholds were shuffled prior to $peptide^{FDR}$ filtering. The red points indicate correlations with a p-value $\leq$ 5%. **D.** The impact of varying $peptide^{FDR}$ cutoff threshold on the shuffled MHC data set. For each $peptide^{FDR}$ cutoff threshold (x-axis), the upper bound of the shaded region indicates the 75^{th} percentile, the lower bound indicates the 25^{th} percentile, and the solid line indicates the median. **E.** Population SARS-CoV-2 binding capacities using only single algorithms were correlated to observed deaths per million. Red points indicate a PPV $\geq$ 95%.

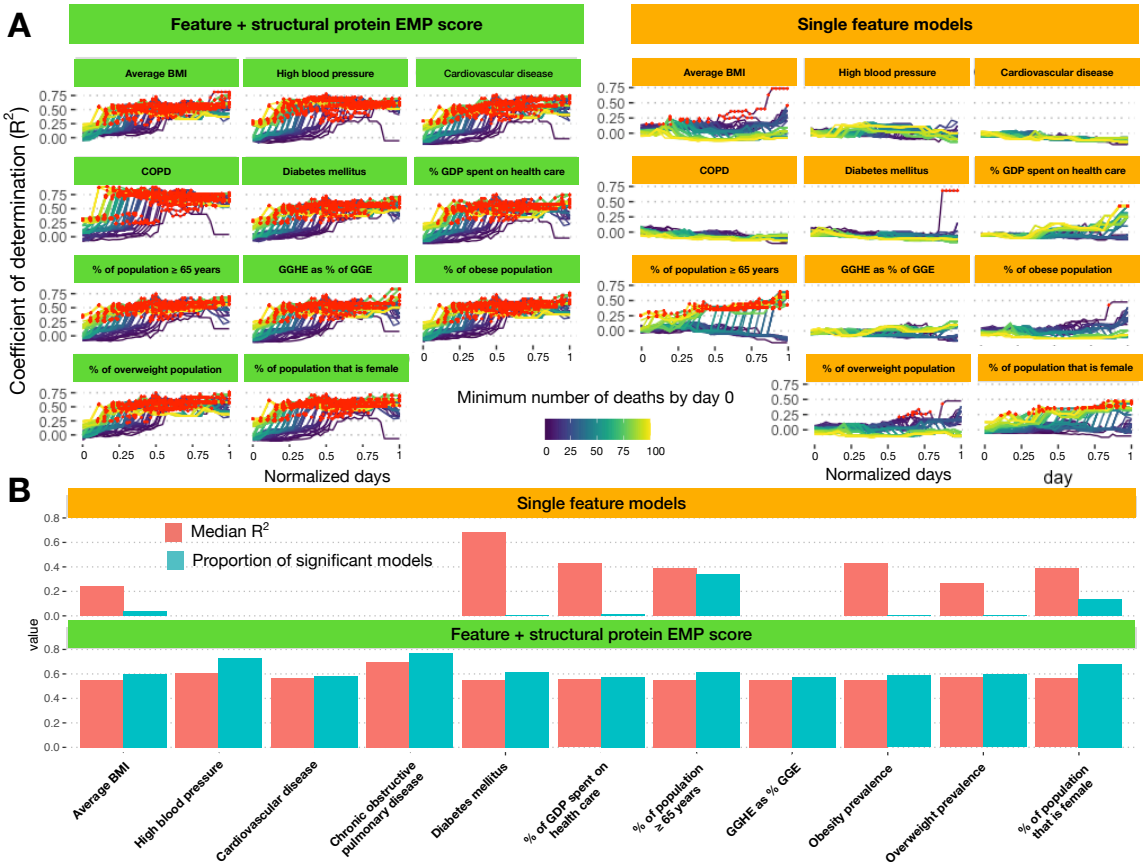

Figure A.8: Addition of structural protein EMP score significantly improves linear model fit to observed deaths per million. A. Linear models were constructed using either a single risk factor (yellow) or a combination of a risk factor and structural protein EMP scores (green). The x-axis indicates the number of normalized days from when a minimum death threshold was met (line color), and the y-axis indicates the observed adjusted R^2 value. **B.** A summary of results obtained from single feature linear models (top panel, yellow) or the combination models (bottom panel, green). The red bars indicate the median R^2 value achieved by that model and the blue bars indicate the proportion of regressions that were found to be significant (F-test ≤ 0.05).

APPENDIX B

HLA-INCEPTION: A STRUCTURE-BASED MHC-I BINDING MOTIF PREDICTION ALGORITHM: SUPPLEMENTAL MATERIAL

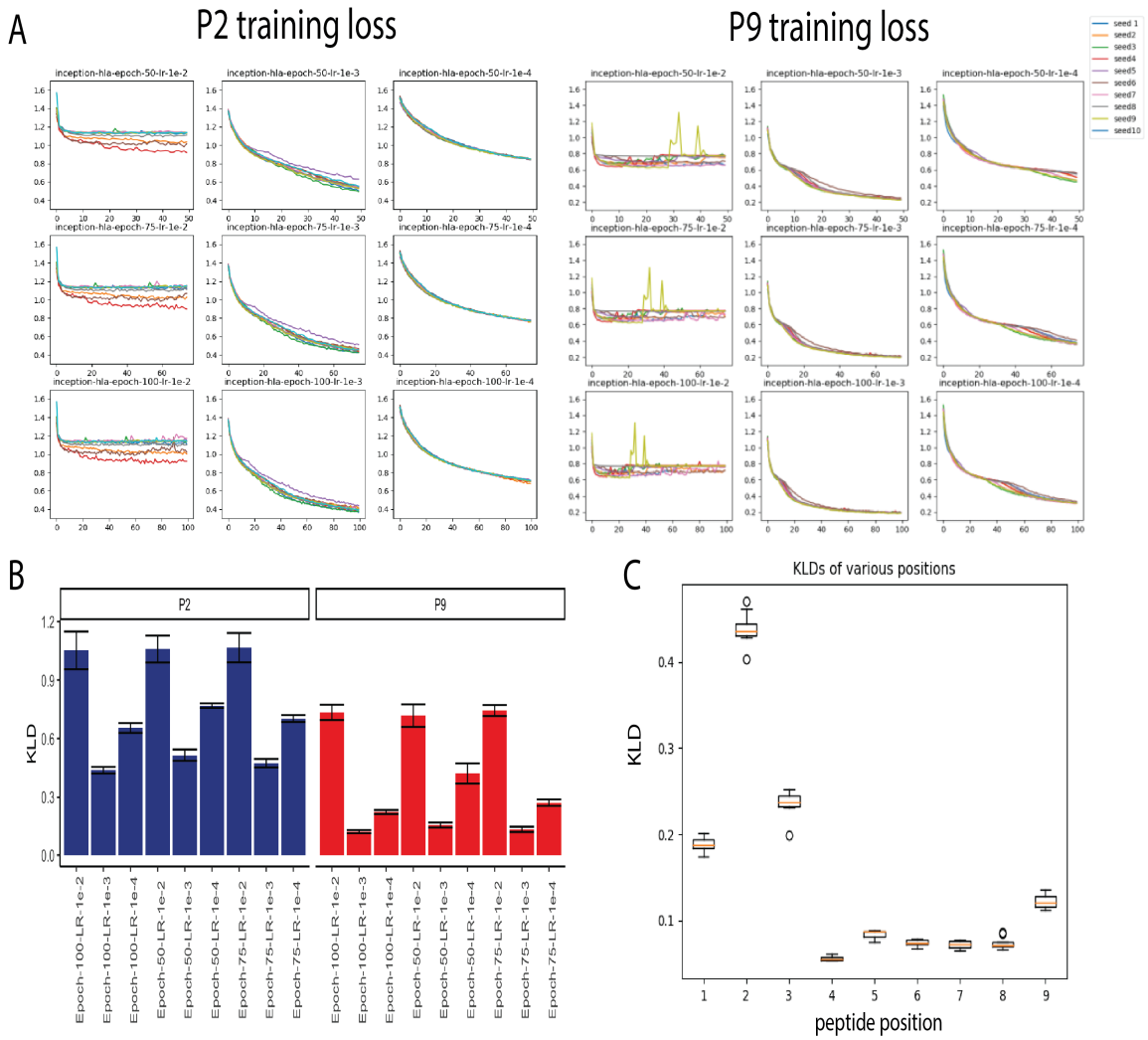

Figure B.1: Hyperparameter tuning with respect to N- and C-terminal anchor binding pockets. A. The training loss curves for tested parameters for position 2 (P2) and position 9 (P9). **B** The median KLD for each parameter set after 10-fold cross validation for P2 and P9. **C.** The position KLDs for 10-fold cross validation when using the optimal parameters (learning rate = 1e-3 ; epochs = 100)

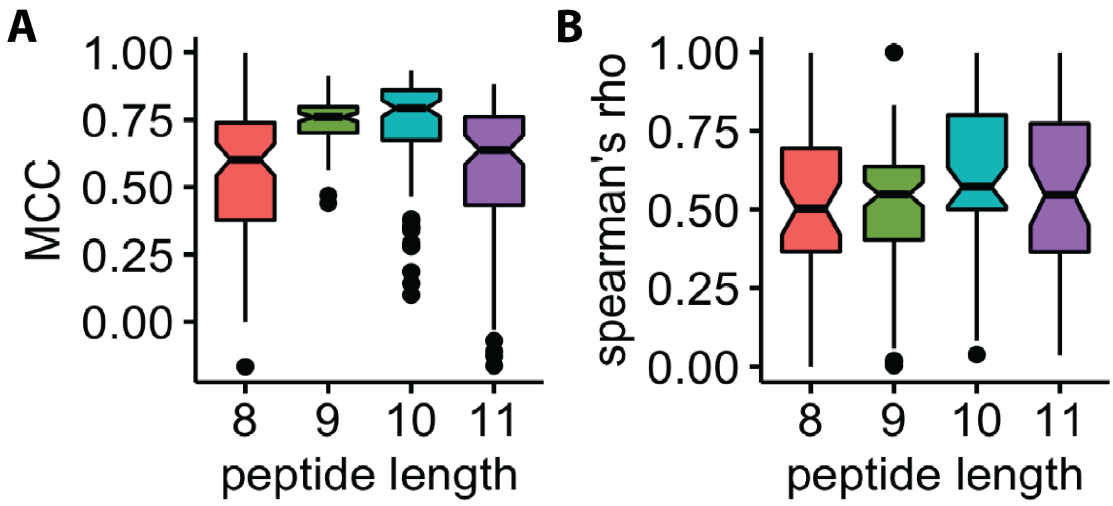

Figure B.2: HLA-inception performance at other lengths. A. The performance of HLA-Inception at different lengths. **B.** The correlation of PWM score with quantitative peptide values for peptides at lengths other than 9

APPENDIX C

INTEGRATIVE MODELING AND DYNAMICS OF THE NL63 SPIKE PROTEIN: SUPPLEMENTAL MATERIAL

Explicit solvent									
Glycan configuration		map 1	map 2	map 3	map 4	map 5	map 6	map 7	total (ns)
wild type	replica 1	45	45	51	45	72	47	62	1081
	replica 2	51	45	49	45	52	55	56	
	replica 3	60	39	72	47	52	50	41	

Table C.1: The above table shows the amount of explicit simulation performed on the spike model.

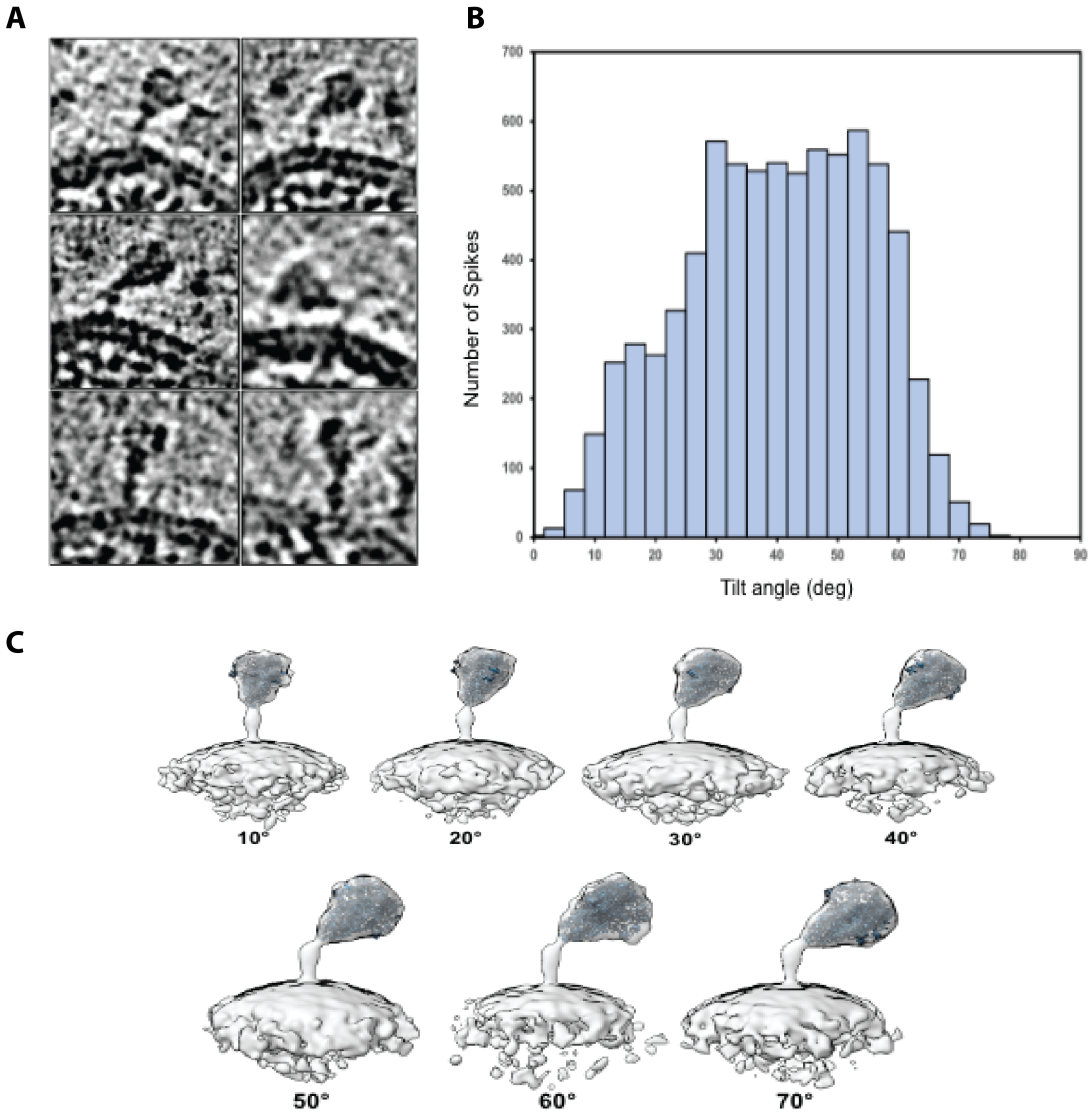

Figure C.1: Spike bending observations. A. single particle images of NL63 spike protein. **B.** The experimental distribution of single particle spike bending angles. **C.** Individual bending maps

Implicit solvent									
Glycan configuration		map 1	map 2	map 3	map 4	map 5	map 6	map 7	total (ns)
wild type	replica 1	100	100	100	95.1	100	100	100	2085.7
	replica 2	99.9	100	100	100	92.4	100	100	
	replica 3	100	100	100	100	98.3	100	100	
del1242; del1247	replica 1	97.8	100	100	100	98.7	100	100	2091.4
	replica 2	100	100	100	100	100	100	100	
	replica 3	100	100	94.9	100	100	100	100	
del1242	replica 1	100	100	100	100	100	100	100	2100
	replica 2	100	100	100	100	100	100	100	
	replica 3	100	100	100	100	100	100	100	
del1247	replica 1	93.4	100	100	100	95.5	100	100	2087.2
	replica 2	100	100	100	100	100	100	100	
	replica 3	100	100	100	100	100	98.3	100	

Table C.2: The above table shows the amount of implicit simulation performed on the spike model.