

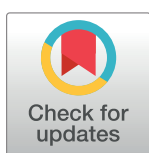
RESEARCH ARTICLE

Modeling the sequence dependence of differential antibody binding in the immune response to infectious disease

Robayet Chowdhury^{1,2}, Alexander T. Taguchi³, Laimonas Kelbauskas¹, Phillip Stafford¹, Chris Diehnelt¹, Zhan-Gong Zhao¹, Phillip C. Williamson⁴, Valerie Green⁴, Neal W. Woodbury^{1,2*}

1 Center for Innovations in Medicine, Biodesign Institute, Arizona State University, Tempe, Arizona, United States of America, **2** School of Molecular Sciences, Arizona State University, Tempe, Arizona, United States of America, **3** iBio Inc., San Diego, California, United States of America, **4** Creative Testing Solutions, Tempe, Arizona, United States of America

* nwoodbury@asu.edu



OPEN ACCESS

Citation: Chowdhury R, Taguchi AT, Kelbauskas L, Stafford P, Diehnelt C, Zhao Z-G, et al. (2023) Modeling the sequence dependence of differential antibody binding in the immune response to infectious disease. *PLoS Comput Biol* 19(6): e1010773. <https://doi.org/10.1371/journal.pcbi.1010773>

Editor: Jinyan Li, University of Technology Sydney, AUSTRALIA

Received: November 29, 2022

Accepted: May 15, 2023

Published: June 20, 2023

Copyright: © 2023 Chowdhury et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: The data used in the study is publicly available at the following address: <https://doi.org/10.5281/zenodo.7752734>.

Funding: This study was supported by the Defense Threat Reduction Agency research grants # HDTRA-11-1-0010 and HDTRA1-12-C-0058 (both to NW). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Abstract

Past studies have shown that incubation of human serum samples on high density peptide arrays followed by measurement of total antibody bound to each peptide sequence allows detection and discrimination of humoral immune responses to a variety of infectious diseases. This is true even though these arrays consist of peptides with near-random amino acid sequences that were not designed to mimic biological antigens. This “immunosignature” approach, is based on a statistical evaluation of the binding pattern for each sample but it ignores the information contained in the amino acid sequences that the antibodies are binding to. Here, similar array-based antibody profiles are instead used to train a neural network to model the sequence dependence of molecular recognition involved in the immune response of each sample. The binding profiles used resulted from incubating serum from 5 infectious disease cohorts (Hepatitis B and C, Dengue Fever, West Nile Virus and Chagas disease) and an uninfected cohort with 122,926 peptide sequences on an array. These sequences were selected quasi-randomly to represent an even but sparse sample of the entire possible combinatorial sequence space ($\sim 10^{12}$). This very sparse sampling of combinatorial sequence space was sufficient to capture a statistically accurate representation of the humoral immune response across the entire space. Processing array data using the neural network not only captures the disease-specific sequence-binding information but aggregates binding information with respect to sequence, removing sequence-independent noise and improving the accuracy of array-based classification of disease compared with the raw binding data. Because the neural network model is trained on all samples simultaneously, a highly condensed representation of the differential information between samples resides in the output layer of the model, and the column vectors from this layer can be used to represent each sample for classification or unsupervised clustering applications.

Competing interests: The authors have declared that no competing interests exist.

Author summary

Previous studies have shown that it is possible to use high density arrays of near-random peptide sequences as a general, disease agnostic approach to diagnosis by analyzing the pattern of antibody binding in serum to the array. The current approach replaces purely statistical pattern recognition with a machine learning-based approach that comprehensively describes the binding of antibodies to specific amino acid sequences. As one use case, this can be implemented to substantially enhance the diagnostic power of these peptide array-based antibody profiles by incorporating the sequence information with the measured antibody binding to better detect and discriminate infectious diseases. This makes the array analysis much more robust to noise and provides a means of condensing the disease differentiating information from the array into a compact form that can be readily used for disease classification, unsupervised clustering or population health monitoring.

Introduction

The humoral immune response to infectious disease involves the production of pathogen-specific antibodies by B cells. While much has been learned about the mechanisms underlying the development of the adaptive humoral immune response upon infection [1–4], comprehensive models describing the dependence of the antibody molecular recognition profile on antigen structure and the variability of the immune response between individuals are lacking [5–7]. B-cell sequencing has started to fill this gap in terms of understanding the repertoire of antibodies produced [8], but the approach is difficult and expensive to apply to very large numbers of samples. Phage display and other display approaches in which amino acid sequences of various lengths are incubated with serum and antibody binding is detected have provided a means of analyzing more directly target molecular recognition [9], but are generally strongly biased towards detecting strong binding interactions. Protein microarrays have also been used to look at the serum antibody binding in specific proteomes, identifying potential targets for either biomarker or drug development [10,11].

Large-scale peptide microarray technology provides an alternate approach and has been used to measure binding of protein targets to large numbers of peptide sequences in the analysis of protein-protein and peptide-protein interactions [5,12–15]. Antibody reactivity profiling (epitope mapping) has been one of the main research fields employing the technology [12,15,16]. As peptide arrays have become more available and affordable, they have become attractive for looking broadly at sequence-binding relationships. Array based approaches have the advantage that one can obtain datasets on large numbers of peptide-target binding interactions simultaneously, measuring binding across a large dynamic range [17,18].

While most of the antibody or immune response applications of peptide arrays have been focused on measuring the interaction of peptides that represent known or suspected antigens with antibodies, the Johnston lab and others have developed the use of high density quasi-random peptide arrays as a tool for generating antibody binding profiles [17–32]. A key feature of these arrays is that the peptide sequences are chosen to cover combinatorial sequence space as evenly as possible, rather than focusing on biological sequences or known epitopes. Due to the random nature of the peptide sequences, the probability that the peptide array contains sequences that exactly match immunogenic regions of a pathogen is low. As a result, this “immunosignature” approach captures mostly low to moderate affinity interactions of antibodies with the array peptides. None-the-less, statistical analysis of serum antibody binding to

the arrays has been shown to enable robust detection or identification of immune responses associated with numerous infectious and chronic diseases [18,22,23,25–27,30].

The immunosignature method involves applying a small amount of diluted serum to a dense array of peptides with nearly random sequences of amino acids, typically with >100,000 distinct peptide sequences of about 10 amino acids in length [17]. In most of the studies done, only 16 of the 20 natural amino acids were used to synthesize the peptides. The level of antibody binding to the peptides on the array is then detected quantitatively using a fluorescently labeled secondary antibody and imaged by an array scanner. Based on a statistical comparison of binding patterns between case and reference samples, classifier models can be built to distinguish one disease response from another [20].

The cognate epitopes of the antibodies involved in an immune response are highly unlikely to appear within a random set of $\sim 10^5$ sequences on a peptide array. For a linear epitope of ~ 10 amino acids in length, there are $\sim 10^{13}$ possible amino acid combinations, yet somehow the interaction of serum antibodies with only $\sim 10^5$ sequences captures sufficient information to both detect and identify disease state with high accuracy [17,18,21–23,25–27,30,33]. If sufficient information can be obtained from a random sparse sampling of antibody binding to 1 out of every 10^8 possible sequences ($\sim 10^{13}/\sim 10^5$), then the antibodies associated with an immune response must recognize millions to billions of different sequences to some extent in a manner that is disease specific. The fundamental question of the current study is whether this amino acid sequence-dependent antibody binding can be modeled using sparse sampling of the combinatorial binding space. If so, such a relationship could potentially be used to more effectively aggregate information from the array or to design new panels of sequences that more effectively differentiate diseases.

Recently, our group modeled the sequence-binding relationships of nine different, well-characterized, isolated proteins to the peptide arrays described above [34]. Binding patterns of each protein were recorded, and a feed-forward neural network model was used to relate the amino acid sequences on the array to the binding values. Remarkably, it was possible to train the network with 90% of the sequence/binding value pairs and predict the binding of the remaining sequences with accuracy equivalent to the noise of the antibody binding measurements (the Pearson correlation coefficients (R) between the observed and predicted binding values were equivalent to that between measured binding values of multiple technical replicates, and in some cases as high as $R = 0.99$). In fact, accurate binding predictions ($R > 0.9$) for some protein targets could be achieved by training on as few as hundreds of randomly chosen sequence/binding value pairs from the array. In addition, the binding predictions were specific; the model captured not only the bulk binding of individual proteins but also the differential binding between proteins. Finally, a neural network trained on weakly binding sequences effectively predicted the binding values of sequences on the array 1–2 orders of magnitude greater. At least in the context of the combinatorial space of possible sequences in this model array-based system (~ 10 residue peptides using 16 different amino acids with the C-terminus bound to the surface of a silica substrate), training on one set of thousands of randomly selected sequences resulted in statistically accurate prediction of the binding to any other randomly selected set of sequences.

Binding to antibodies, in this case IgG in human sera, represents a much more complex system than binding to isolated proteins, and one might expect substantially more complex sequence-binding relationships. Previous studies have developed such relationships for immune responses using various starting datasets. A number of groups have looked at overlapping peptides presented on microarrays or in phage display libraries generated by tiling antigens or entire proteomes [35–38]. Bio-panning of phage or bacterial peptide display libraries coupled with next generation sequencing have provided broader binding profiles [39,40]. The

advantage of tiling and panning approaches is that one is starting with known or suspected binding sequences, and thus the dataset is naturally rich in strong binding information. In one particularly effective study in this regard, a method referred to as Protein-based Immunome Wide Association Study was used to explore sequence binding relationships in 31 systemic lupus erythematosus samples [41]. Here a large bacterial display library (10^{10} 12-mer sequences) was reduced to $\sim 10^6$ sequences found to bind to serum antibodies from the samples and the enrichment of specific 5-mer and 6-mer sequences within the resulting library was determined. These enriched sequences were then used to identify autoantibodies in the human proteome, and the authors were successful at identifying several known autoantigens for the disease within their top candidates. The same group has used similar methods to perform epitope mapping of antibodies to SARS-CoV-2 [42].

Machine learning algorithms have also been used to develop sequence-based models predicting binding of proteins to peptides, antibodies, and DNA [43–49]. For example, machine learning models have been used to model anti-microbial peptides, infectious viral variants that escape protection, potential epitopes on target antigens, high antibody binding regions on target proteins, and optimization of target DNA sequences for transcription factors. To do this, two approaches have primarily been used: 1) introducing single or multiple point mutations on a target site with known function to identify desired leads, and 2) use of proteomes of interest or known antigenic proteins to predict epitopes. For example, epitope prediction tools such as BepiPred-2.0 are generally developed using known antigens derived from crystal structures of antibody-antigen complexes [50]. With regard to modeling of serum binding to random sequences, Greiff *et al.*, applied multivariate regression to serum antibody binding to a library of 255 random peptides [51]. In that study, serum antibody binding from naïve mice was well modeled by relating peptide composition to binding intensity, though binding of serum antibodies from previously infected mice proved more challenging to model.

The current work focuses on the feasibility of developing comprehensive sequence-binding relationships that describe the infectious disease specific binding of total IgG to our model library of 122,926 peptides each between 7 and 12 residues in length and composed of 16 of the 20 natural amino acids. While this library is clearly limited in terms of size (only 10^5 of the trillions of possible sequences), composition (16 of 20 natural amino acids) and context (C-terminus affixed to a silica surface), it is capable of distinguishing immune responses to different infectious agents, as described previously [17,18,21,26]. Neural network-based models were used to build quantitative relationships for sequence-antibody binding using sera from cohorts of individuals who are either uninfected (controls) or infected with one of five infectious agents including three closely related members of the family *Flaviviridae* (dengue virus, West Nile virus and hepatitis C virus), a more distantly related member of the family *Hepadnaviridae* (hepatitis B virus) and an extremely complex eukaryotic trypanosome (the agent of Chagas disease, *Trypanosoma cruzi*). This allowed a thorough evaluation of the model's ability to capture the disease-specific information content of the array binding. This study has shown that it is possible to create accurate sequence-binding models, which not only learn the disease specific information, but also effectively capture the binding information on the arrays for applications in noise suppression and disease classification.

Results

Study design and initial analysis

The serum samples shown in Table 1 were incubated on identical peptide microarrays as described in Methods and IgG bound to the array peptides was detected via subsequent incubation with a secondary anti-IgG antibody. The peptide sequence 'QPGGFVDVALSG' is

Table 1. Sample information.

Disease cohort	Sample Source ¹	Samples Collected	Low CV Samples ²	Genome Size(bp)
Hepatitis C Virus (HCV)	CTS	100	78	11,000
Dengue Virus, Serotype 4 (Dengue)	CTS and SeraCare	65	57	9600
West Nile Virus (WNV)	CTS	100	74	11,000
Hepatitis B Virus (HBV)	CTS	100	86	3200
<i>T. cruzi</i>	CTS	96	70	105M
Uninfected (ND)	CTS and ASU	218	177 ³	--

¹CTS is Creative Testing Solutions (Tempe, AZ); ASU is Arizona State University; SeraCare address is Milford, MA

²Arrays passing the data quality metrics used in the initial neural network analysis. The remaining high CV samples were used as a test set for certain classification studies.

³100 randomly selected uninfected samples were used for the bulk of the neural network analysis to remain reasonably balanced with other cohorts.

<https://doi.org/10.1371/journal.pcbi.1010773.t001>

present on the array as a set of replicate features ($n = 276$). This peptide sequence gives a consistently moderate to strong binding value from sample to sample and is used to assess the intra-array spatial uniformity of antibody binding intensities. Median normalized arrays with an intra-array replicate feature coefficient of variation (CV) ≥ 0.3 for this peptide sequence were set aside as well as arrays that showed significant physical defects or overall differences in binding intensity between different regions of the array (collectively these are referred to as “High CV samples”). In all, 20% of the 679 arrays measured were excluded from the initial part of the analysis but considered in the last section which focuses on using the sequence-binding relationship to remove noise from the arrays. Thus, 542 arrays total were considered “Low CV Samples” in Table 1.

Comparison of average binding profiles of peptides to serum IgG. Fig 1 shows the cohort average serum IgG binding intensity distributions of the 122,926 unique peptide sequences. The samples were all median normalized prior to averaging each peptide binding value within the cohort. The \log_{10} of the average binding is displayed on the x-axis as the log distributions are much closer to a normal distribution than are the linear binding values. Sera from individuals infected with HCV, dengue virus or WNV have sharper distributions (smaller full width at half maximum) than the other samples, while sera from individuals infected with HBV show a distribution width similar to those from uninfected donors. Sera from individuals with Chagas disease have a broader binding distribution than the others, with a long tail on the high binding side. Overall, the width of the distribution increases with increasing proteome size. Interestingly, for the viruses with small proteomes, some of the higher binding antibodies are lost compared to uninfected samples. However, it is important to remember that the array peptides have no relationship to the viral proteomes or indeed any biological proteome, except by chance. Thus, what is lost in the small virus samples compared to strong binding in uninfected samples, may well be gained in more specific binding not immediately apparent.

Neural network analysis

The fundamental question of this study is whether it is possible to accurately predict the sequence dependence of the antibody binding associated with an immune response to a given pathogen, both in terms of accurately representing the IgG binding to each peptide sequence in individual serum samples and in terms of the ability of the neural network to capture sequence dependent differences in IgG binding between samples and cohorts. Towards this end, the low CV samples (Table 1) were analyzed using a feed-forward neural network model

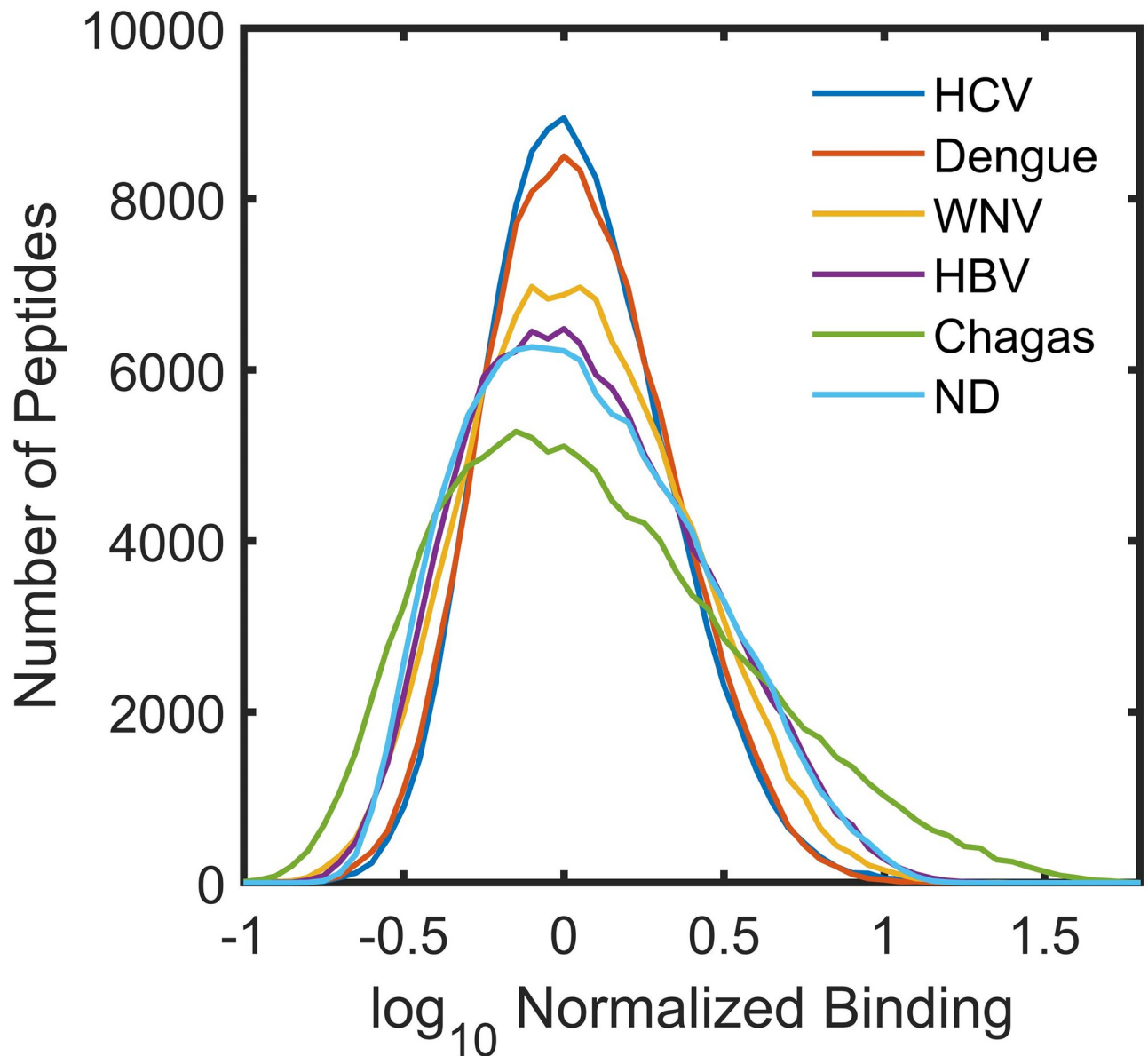


Fig 1. Average Binding Distributions of the Cohorts. Average binding intensity distributions of serum IgG binding to array peptides for the 6 different sample cohorts. For each cohort the \log_{10} of the average binding for each peptide sequence was used to create the distribution.

<https://doi.org/10.1371/journal.pcbi.1010773.g001>

[34]. All samples were fit simultaneously such that a single neural network was trained to predict the binding for all samples for any given sequence. The optimized network involved an input layer, three hidden layers with 542 nodes each and an output layer of width 465 corresponding to the number of target samples. The loss function used was the sum of least squares error based on a comparison of the predicted and measured values for the peptides in the sample.

The neural network uses the sequence information to rapidly converge on a solution. Fig 2A shows the rate at which the loss function drops during training. When the correct sequence is paired with its corresponding binding value (blue and red lines, Fig 2A), the value of the loss function drops rapidly and the values for the training set and test set drop in

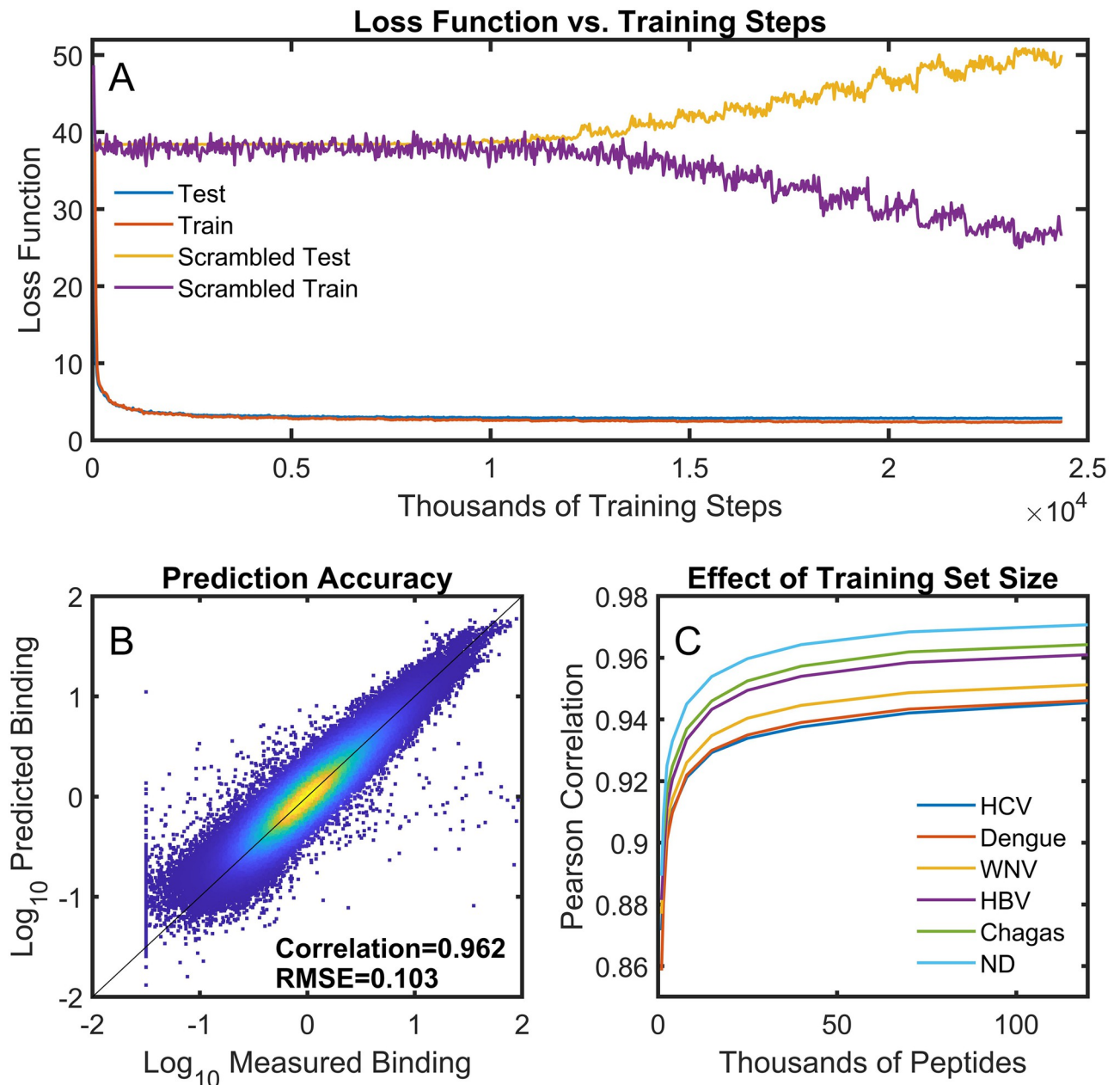


Fig 2. The neural network model accurately represents the sequence-binding relationship. A neural network (3 hidden layers with 250 nodes) was trained on 95% of the sequence/binding data from the 542 low CV samples in Table 1 simultaneously. The remaining 5% of the sequence/binding values (6,146 per sample \times 542 samples = \sim 3.3 million binding values) were held out as the test set. (A) The loss function progression during neural network training. Blue and red traces (overlapping): a neural network trained with properly matched sequences and associated binding values. Purple and yellow traces: training after scrambling the order of the sequences relative to their measured binding values. (B) The scatter plot shows the values predicted by the neural network (y-axis) vs. the corresponding measured values from the array (x-axis) for the test set only. (C) The average predicted vs. measured correlation coefficient for cohort samples as a function of the number of peptide sequences used to train the network.

<https://doi.org/10.1371/journal.pcbi.1010773.g002>

concert; there is almost no overfitting. As a control, the same neural network was used to analyze data in which the order of the peptide sequences was randomized relative to their binding intensities. One would not expect any relationship between sequence and binding under these circumstances. In this case, the loss function value slowly decreases for the training set of

peptides and slowly increases for the test set (yellow trace: test, purple trace: train) indicating overfitting to the training set. This implies that the neural network is capable of rapidly converging on a true relationship between the sequences and their binding values in the context of the array peptide library.

The neural network results in a comprehensive binding model applicable across the model sequence space used. Fig 2B shows a density-colored scatter plot [52] comparing the predicted and measured values from the neural network model. In this case, the model was trained on 95% of the peptide sequence-binding pairs, randomly selected, with the remaining 5% or 6,146 peptide sequences excluded from training and used for model testing (that is 6,146 binding values for each of the 542 low CV samples used = ~3.3 million binding values in the test set). Only the test set values are displayed in Fig 2B. Since the sequences used on the array are nearly random, these sequences should be statistically equivalent to any randomly selected set of sequences from the combinatorial space of possible sequences sampled by the array (peptides of about 10 residues utilizing any of 16 amino acids corresponds to about 10^{12} sequences). The Pearson correlation coefficient (R) between the measured and predicted values for the test sequences shown is 0.956. Repeating the training 100 times with randomly selected train and test sets gives an average correlation of 0.962 ± 0.001 (standard error of the mean) and a root mean square average error of 0.103 ± 0.001 . The correlation coefficient between measured and predicted binding for the 95% of the sequences used to train the neural network was 0.967 ± 0.001 . This implies that there is almost no overfitting associated with the model (the quality of fit between the test and train data is similar), a conclusion also apparent in the loss function data of Fig 2A. S1 Fig shows the correlation coefficient between measured and predicted binding for each individual sample in the test dataset. While some cohorts and some samples were better fit than others, for the vast majority of the samples, the correlation coefficients are greater than 0.9.

10^3 to 10^4 peptides are sufficient to provide a reasonable description of the entire combinatorial peptide sequence space. Neural network models were trained with different numbers of randomly selected peptides, and binding was predicted for the remaining portion of the peptides. Fig 2C explores the dependence of the overall correlation coefficient between measured and predicted binding values for the test set of each of the sample cohorts as a function of the number of peptides used in the training. When at least 10,000 peptide sequences are used to train the neural network, the correlation coefficient is >0.9 for all cohorts, and the correlation is >0.85 when the model is trained using only 2,000 peptides. This implies that even a very sparse sampling of this sequence space provides a reasonably accurate model of the sequence-binding relationship. The correlation coefficients do continue to increase slowly as a function of training set size. Thus, even though a relatively small set of peptides gives a reasonable overall picture, the predictive power of the relationship continues to improve with more data, and if even more peptide sequences were available for training than the entire 122,926 peptides on the array, an improved prediction would be expected.

There are commonalities in the binding of each sample that make simultaneous modeling of all samples as accurate as individual fits. It is possible to either build a single neural network model that fits all samples simultaneously (as done in this work), or to train entirely independent neural network models for each of the samples considered. Fig 3 shows a direct comparison of the measured vs. predicted correlation coefficients of each sample using the simultaneous and individual modeling approaches. In most cases, the simultaneous model is slightly more accurate in terms of the correlation coefficient. This implies that the network learns commonalities between IgG binding from serum across all samples and different

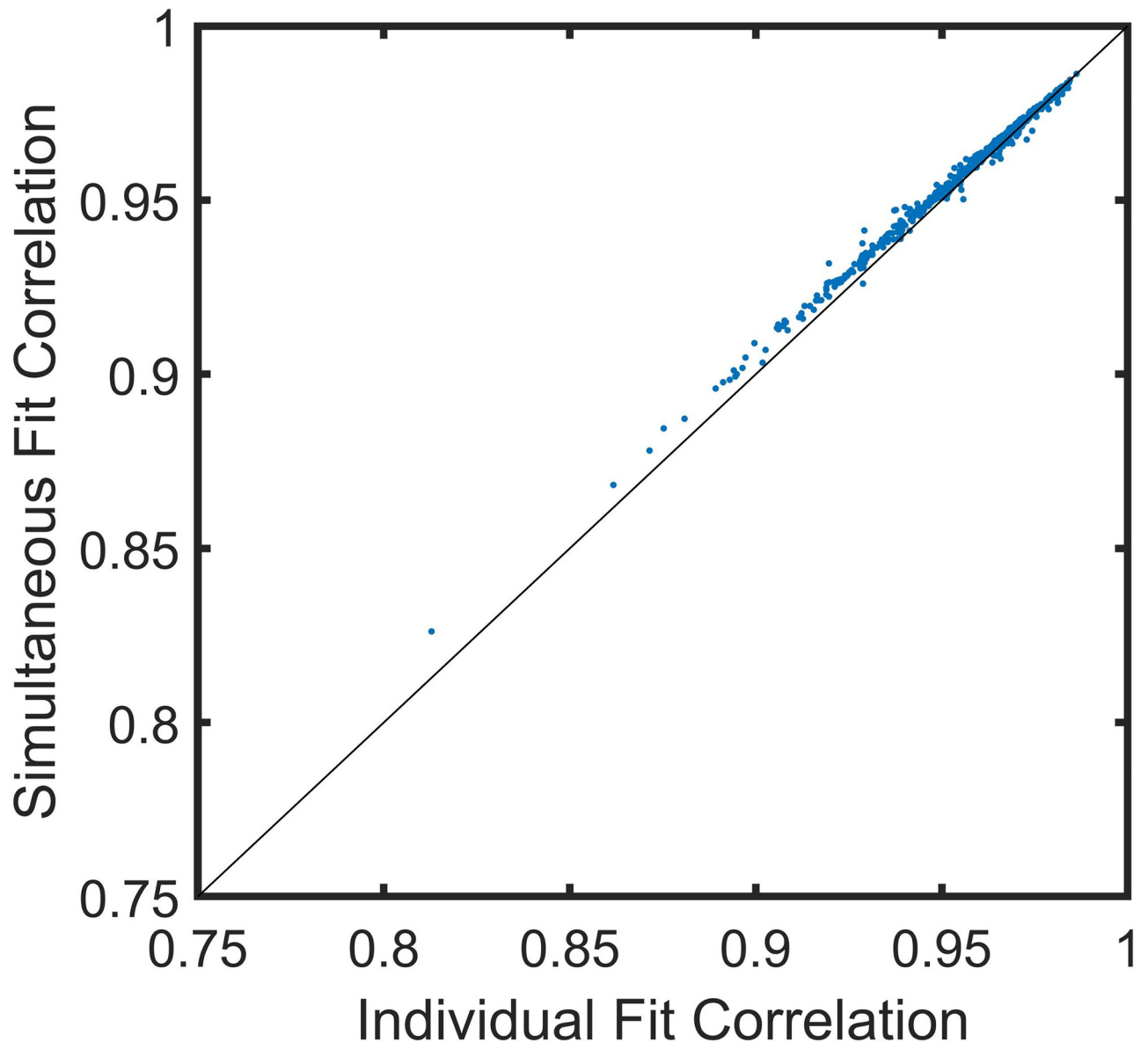


Fig 3. Simultaneous Modeling of All Cohorts is as Accurate as Individual Fits. A comparison of predicted vs. measured correlation coefficients calculated either by fitting samples simultaneously (as in Fig 2) or one at a time.

<https://doi.org/10.1371/journal.pcbi.1010773.g003>

cohorts and takes advantage of those commonalities to refine the weights of the 3 hidden layers. The differences between samples are learned in the output layer (the final weight matrix), with separate columns in that layer giving rise to the predicted binding values for each sample. Additionally, simultaneous modeling is significantly faster than fitting each sample dataset separately. For comparison, a simultaneous training required about 5–6 minutes to complete using a single GPU while the individual modeling required about 10 hours even after optimizing parallel processing. Note that once the network has been trained on sufficient numbers of samples, new samples only require training of the final weight matrix since the column vectors of this matrix are what differentiates the samples.

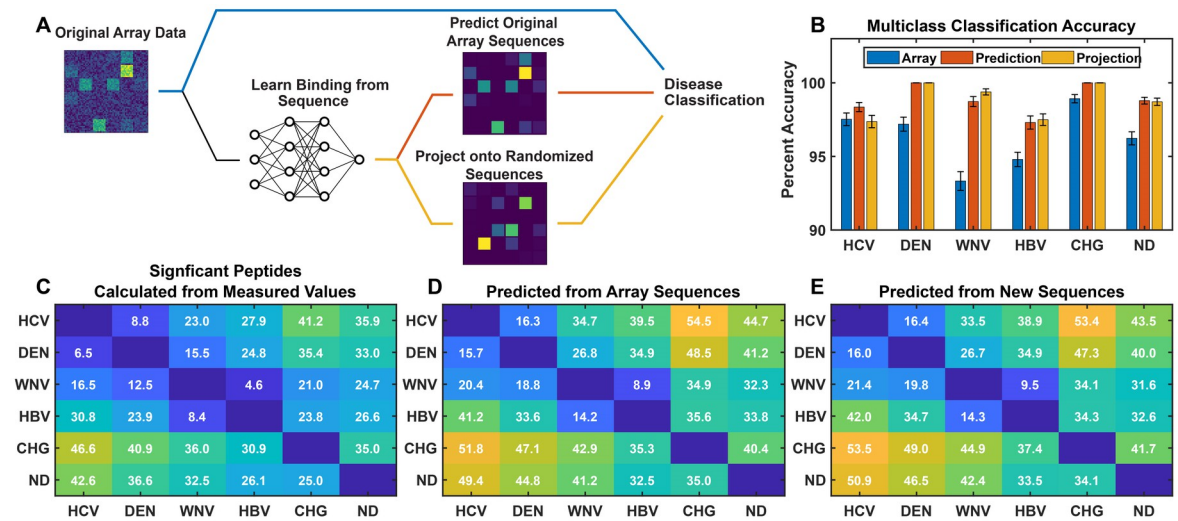


Fig 4. Discriminating between cohorts. (A) The data from the original array was analyzed in three ways: directly (blue line), after training a neural network and predicting the values of the array sequences (red line), and after projecting the trained neural network on a completely new set of sequences (orange line). Disease discrimination was then performed for each approach using multi-class classification or by statistically determining the number of significant peptides distinguishing each cohort comparison. (B) Multi-class classification based on a neural network (see text). Classification was performed 100 times for each dataset leaving out 20% of the samples (randomly chosen) each time. Blue: the original measured array data was used as input to the multiclass classifier. Red: the neural network model prediction of binding values for array peptide sequences were classified. Orange: the neural network projected onto a randomized set of sequences of the same overall size, composition and length distribution as the array sequences. (C) Each array element is the number of array peptides with measured binding values that are significantly higher in the sample cohort on the Y-axis compared to the sample cohort on the X-axis. Significance is defined as a p-value less than $1/N$ in a T-test with 95% confidence ($N = 122,926$ total peptides, thus significant peptides have a p-value $< 0.05/N = 4.1 \times 10^{-7}$). (D) As in (C) except that the neural network predicted binding values of the array peptides were used instead of the measured. The mean of 10 different neural network model training runs is shown; error in the mean is ≤ 0.3 . (E) The same as in (D) except predicted values for an *in silico* generated array of random peptide sequences with the same average composition and length as the peptides in the array were used. The mean of 10 different sequence sets and neural network runs is shown; error of the mean is ≤ 0.4 . For the remaining work a total of 77 uninfected samples were removed to bring the number of samples down to 100, about the same as the other samples and a total number of 465 samples.

<https://doi.org/10.1371/journal.pcbi.1010773.g004>

The neural network learns distinguishing characteristics of cohorts

Fig 4A is a schematic of three approaches to disease classification and discrimination. The blue line is the standard statistical pathway (immunosignaturing). Here, no sequence information is used in the analysis and the binding values are either fed into a multiclass classifier (Fig 4B, the classifier is described in Methods and further in the supplementary information) or used to determine the number of significant peptides that distinguish diseases (Fig 4C), as described below. Alternatively, the neural network can be used to extract information from the sequence/binding relationship for disease classification. This relationship can either be used to recalculate predicted binding values for the array peptide sequences, forcing the data to always be consistent with the sequences (denoising, red line), or it can be projected onto a completely new set of sequences (an *in silico* array, orange line), and those projected binding values used in classification or determining the number of significant distinguishing peptides between disease pairs.

Values predicted by the neural network result in better ability to distinguish cohorts.

In Fig 4C–4E, the number of peptide binding values that are significantly greater in one cohort (on the Y-axis) compared to another (on the X-axis) are shown in each grid. Significance was determined by calculating p-values for each peptide in each comparison using a T-test between cohorts adjusted for multiple hypothesis comparisons using the Bonferroni correction. Significant peptides are those in which the p-value is less than $1/N$ ($N = 122,926$) with $>95\%$

confidence. Fig 4C shows comparisons between cohorts using the measured data from the arrays. As one might expect, the sera from donors infected with the Flaviviridae viruses are most similar to one another in terms of numbers of distinguishing peptides. In general, they are more strongly distinguished from HBV (except for WNV) and very strongly distinguished from Chagas donors. If one follows, for example, the top row of Fig 4C for HCV, moving to the right one sees that the numbers increase as more and more genetically dissimilar comparisons are made. West Nile virus is an exception in this regard. While it is more similar to Dengue virus than it is to Chagas, it is most similar, in terms of numbers of distinguishing peptides, to HBV (Fig 4C).

Fig 4D is the same as Fig 4C except that in this case, the predicted values from the neural network model are used for the array sequences instead of the measured values. Because the network requires that a common relationship between sequence and binding be maintained for all sequences, it increases the signal-to-noise ratio in the system such that significantly more distinguishing peptides are identified in every comparison. The neural network was run 10 times and the results were averaged.

Fig 4E shows results in the same format as panels C and D, but using *in silico* generated sequences with binding values predicted by the neural network model trained on peptide array binding data. These sequences were produced by taking the amino acids at each residue position in the original sequences and randomizing which peptide they were assigned to (considering the sequences as a matrix with rows representing peptides in the array and columns representing residue positions, order of amino acids in each column was randomized separately and at the end any spaces due to varying peptide lengths were removed). This created an *in silico* array with a completely new set of sequences that had the same number, overall amino acid composition and average length as the sequences on the physical array to ensure a consistent comparison. The binding values for each sample were then predicted for this *in silico* array and those values were used in the cohort comparisons. The number of significant peptides identified using the new sequence set (Fig 4E) are identical to within error for each comparison with the predictions from the actual array peptide sequences used in the training (Fig 4D). Note that the result is the average number of significant peptides from analyzing ten different randomized *in silico* arrays.

Another way to understand how well distinguishing information is captured by the neural network model is to compare disease classification performance based on measured values vs. predicted values, where the predicted values are generated from the sequence binding relationship resulting from the neural network analysis of Fig 2 (note that creation of the sequence binding relationship is performed without knowledge of which sample belongs to which cohort so it does not bias the classification). Fig 4B shows the result of applying a multiclass classifier, either to the original measured binding values, the binding values predicted for the array sequences, or binding values predicted for *in silico* generated sequences. A multiclass classifier was built using a neural network with a single hidden layer with 300 nodes (described in the Methods and in the supplementary information). This will be referred to simply as the “multiclass classifier” to avoid confusion with the neural network used to model the sequence-binding relationship. The multiclass classifier cannot effectively use all peptides for each sample. Peptide feature selection was performed using a peptide-by-peptide T-test between the binding values of each cohort vs. all others. Fifteen features were used per cohort, giving a total of 83–84 features selected in total for all six cohorts combined due to some overlap between discriminating features found for different cohorts. The training target is a one-hot representation of the sample cohort identity. 80% of the samples were randomly selected and used to train the multiclass classifier and 20% were used as the test set. Both the feature selection and training were performed only on the training samples. Each test sample was then assigned a

cohort label based on the largest value in the resulting predicted output vector. The process was repeated 100 times, on 100 randomly selected training and test sets, and overall prediction accuracy determined. For every cohort, with the possible exception of HCV, classification was improved relative to direct use of the measured array values (blue bars) when using the predicted values. This was true using either predicted values for the array sequences (red bars) or predicted values resulting from projection of the trained network on the randomized *in silico* array sequences (orange bars).

Understanding the noise reduction properties of neural network modeling

The results presented above show that by using the sequence/binding information to first train a neural network model and then predicting the binding using that model (on the same or a different set of sequences), it is possible to improve the signal-to-noise ratio in the data, at least for the purpose of differentiating between disease cohorts. To understand this in more detail, the effects of noise added to the data was explored.

Gaussian noise is effectively removed by the model. In Fig 5, noise was artificially added to each point in the measured dataset by using a random number generator based on a gaussian distribution that was centered at the measured value:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

In the above equation, mu (μ) is the \log_{10} of the median normalized measured binding value. Sigma (σ) was then varied from 0 to 1 to give different levels of added noise with $\sigma = 1$ corresponding to a 10-fold change, up or down, in the width of the linear binding value distribution (due to the \log_{10} scaling). Fig 5A shows the resulting distribution of peptide binding values after adding noise. The peptide binding values were mean normalized across all cohorts and then plotted as a distribution for each cohort (since this is the \log_{10} of the mean normalized value, the distributions are centered at 0). As sigma is increased, the width of the resulting distribution after adding noise increases dramatically.

Fig 5B plots the multi-class classification accuracy of each dataset for each sample cohort as a function of sigma (this uses the same multiclass classifier as Fig 4). The classification accuracy of the original measured data with increasing amounts of noise added drops rapidly (dashed lines). Since this is a 6-cohort multi-class classifier, random data would give an average accuracy of ~17%. The measured values with added noise approach that accuracy level at the highest noise. However, by running the data with noise added through the neural network and then using predicted values for the array sequences rather than the measured values, the accuracy changes only slightly for sigma values up to 1. Note that because this analysis is on a log scale, a sigma of 1 corresponds to causing the linear measured binding values to randomly vary over a distribution with a width covering between 10% and 1000% of their original values.

To further demonstrate the noise reduction capability of the neural network, S4A Fig shows a scatter plot of the measured data plus noise (sigma = 1) vs. the original measured data, resulting in a correlation coefficient of only 0.356. In contrast, S4B Fig shows a scatter plot of the predicted binding values trained on the measured data plus noise vs. the original measured data, resulting in a correlation coefficient of 0.958, essentially identical to the predictive capability of the neural network when trained on the measured data without noise (Fig 2B).

Neural network predictions of array signals improved classification of high CV samples. As described above, 137 samples were not used in the analyses because they either had high CV values calculated from repeated reference sequences across the array or because there

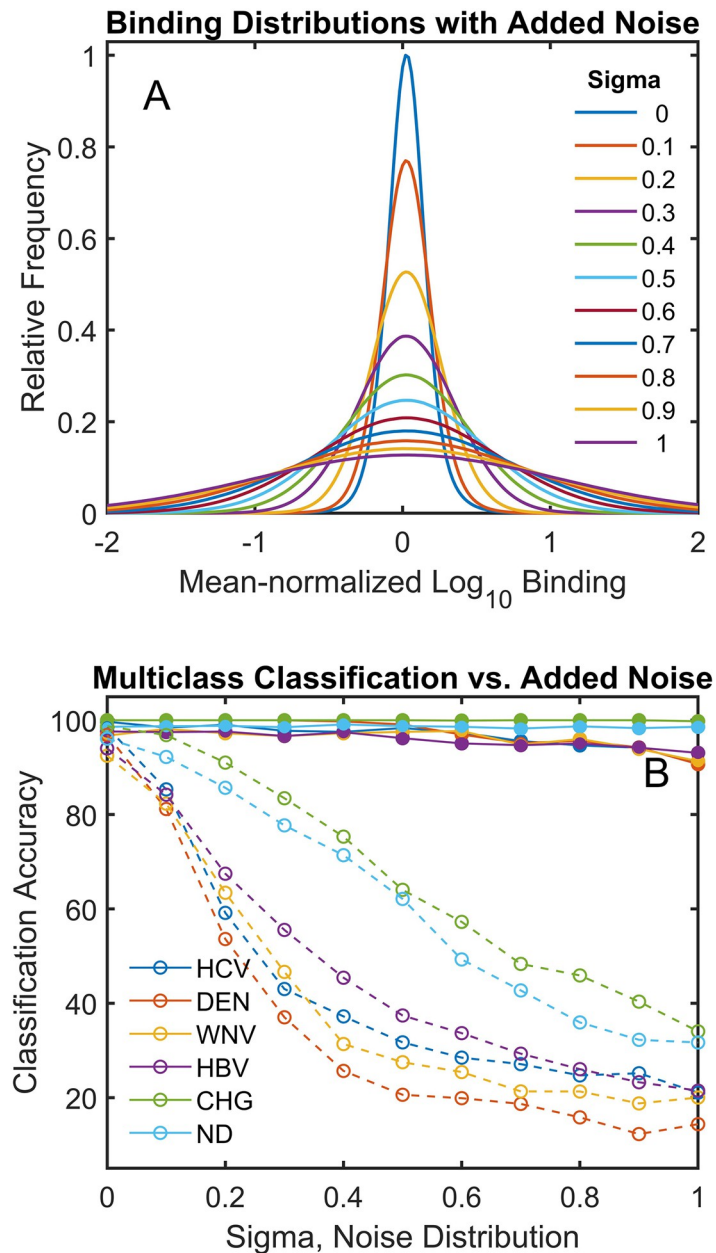


Fig 5. Effect of added noise on multiclass classification. Noise was added to each peptide in the sample using a randomly chosen value from a gaussian distribution centered at the \log_{10} of the measured value. The sigma of the distribution was varied between 0 and 1 (the binding, and thus sigma, is on a log scale). (A) The resulting distributions of binding values for each peptide in a cohort and then including all peptide binding values in the distribution. (B) Results of applying a multi-class classifier (as in Fig 4B) to the data for measured binding values plus noise (dashed lines) and binding values predicted using the measured data plus noise to train the neural network (solid lines) at each value of sigma. Each classification was repeated 100 times (noise at each level was randomly added 10 times and each of these were reclassified 10 times leaving out 20% of the samples as the test set).

<https://doi.org/10.1371/journal.pcbi.1010773.g005>

were visual artifacts such as scratches or strong overall intensity gradients across the array. A neural network model was trained on all 679 samples in [Table 1](#) (all 542 low CV + 137 high CV) simultaneously. Note that the model does not include any information about what cohort each sample belongs to, so modeling does not introduce a cohort bias. The overall predicted vs. measured scatter plots and correlations are given in [Fig 6A and 6B](#) for the low CV and high CV data, respectively. The number of points displayed was randomly selected to be constant between datasets and make the plots comparable. Prediction of the binding values for the high CV data results in more scatter relative to measured values, due to the issues with those particular arrays.

In [Fig 6C](#), the measured and predicted values for the 542 low CV samples were used to train a multiclass classifier which was then used to predict the cohort class of the high CV samples. Three different data sources were used: 1) the measured array data (blue bars), 2) predicted binding values for the array peptide sequences based on the neural network model (red bars) and 3) projected values for *in silico* generated arrays similar to those used in [Fig 4](#) (orange bars). The classifier used was the same as that in [Fig 4](#) and the number of features selected was 15 for each cohort comparison. In each case except for the non-disease samples, the use of predicted values resulted in a significantly better classification outcome.

Discussion

A quantitative relationship between peptide sequences and serum IgG binding

The work described above shows that it is possible to use a relatively simple neural network model to generate a quantitative relationship between amino acid sequence and serum antibody binding over a large amino acid sequence space by training on a very sparse sampling of binding to that sequence space, similar to what was seen previously for isolated proteins binding to the array [34]. Indeed, a reasonably accurate prediction can be obtained with only thousands of sequences ([Fig 2C](#)).

The advantage of using neural networks to describe these sequence-antibody binding relationships is that there is no underlying assumption of a specific physical model. This allows the neural network to describe nonlinear interactions between antibodies and the peptide sequences that would be difficult or impossible to anticipate in an explicit physical description. Immune reactions are complex and polyclonal in nature; there is no single antibody or epitope sequence involved, and thus one would not expect a single binding constant or interaction mechanism. While the neural network does not itself provide a physical interpretation of the antibody-sequence binding, it does provide a rapid means of identifying which sequences are likely involved, for example, within the proteome of a pathogen, for further and more detailed investigation.

The model system used here to explore the relationship between antibody molecular recognition profiles and amino acid sequences has limitations. Only 16 of the 20 natural amino acids were used in this model for technical reasons (see [Materials and methods](#)). The sequences are also bound at one end to an array surface, and the other end has a free amine rather than a peptide bond as would be seen in a protein. In addition, the array peptides are short, linear and largely unstructured. This limits the range of molecular recognition interactions that can be observed, and thus the level of generality of the conclusions, but also suggests that comprehensive and accurate structure/binding relationships for humoral immune responses should be possible to generate given binding data in a broader sequence context. Such relationships would be invaluable for epitope prediction, autoimmune target characterization, vaccine development, effects of therapeutics on immune responses, etc. Even this rather

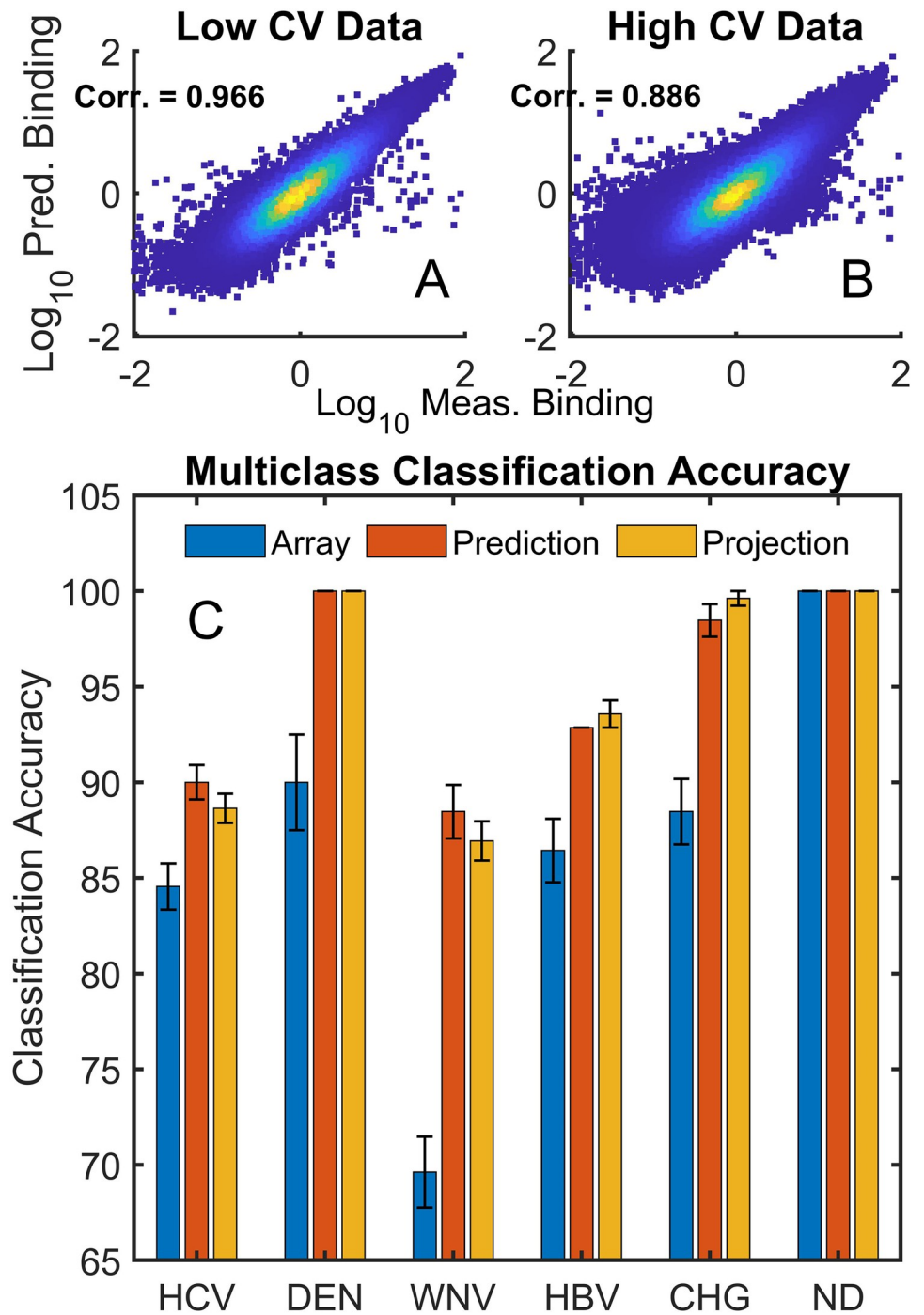


Fig 6. Classification accuracy for high CV samples. (A) Neural network predicted vs. measured values for low CV data and (B) for high CV data. (C) Multiclass classification of the high CV data. Blue, Red and Orange bars represent use of measured, predicted and projected data as in Fig 4.

<https://doi.org/10.1371/journal.pcbi.1010773.g006>

simple model system for sequence space already shows the ability to capture differential binding information between multiple diseases simultaneously, including infectious diseases that involve closely related pathogens (Fig 4).

The fact that one can develop comprehensive sequence/binding relationships within this model sequence space also explains, at least in part, why the immunosignature technology is as effective as it is. Immunosignaturing technology as applied to diagnostics uses the quantitative profile of IgG binding to a chemically diverse set of peptides in an array followed by a statistical analysis and classification of the resulting binding pattern to distinguish diseases. The approach has been successfully used to discriminate between serum samples from many different diseases [17, 18, 21–23, 25–27, 29, 30] and has been particularly effective with infectious disease [17,18,21,31], as exemplified by the robust ability to classify the immune response to the infectious diseases studied here (Fig 4D). This raises the question, why would antibodies that are generated by the immune system to bind tightly and specifically with pathogens show any specificity of interaction to nearly random peptide sequences on an array? The success of the neural network in comprehensive modeling of the sequence/binding interaction provides an answer. The *information* about disease-specific IgG binding is dispersed broadly across peptide sequence space, even in the interaction with sequences that themselves bind weakly and with low specificity, rather than being focused only on a few epitope sequences. It is not necessary to measure binding to the epitope if you have a selection of sequences that are broadly located in the vicinity of the epitope in sequence space.

Note also that by working with sequence/binding relationships, rather than purely statistical comparisons of binding values associated with specific sequences, one can combine information from arrays that contain different peptides. As shown in Fig 2C, when 50% of the array is used to predict the other 50%, the correlation coefficient on average is well over 0.9.

The advantage of analyzing many samples simultaneously

The results of Fig 3 demonstrate that simultaneous neural network analysis of all samples from all cohorts provides a slightly more accurate overall description of binding than does sample by sample analysis, or at least no worse. Conceptually, this suggests that there is enough information in common between the antibody molecular recognition profiles of the various samples that using the same hidden layers to describe all of them, followed by an output layer with a distinct column describing each sample, is sufficient to both describe the general and specific binding interactions. An added practical benefit to this approach is a significant reduction in computation time, as described above.

Using the sequence/binding relationship to suppress noise

In Fig 4, both the number of distinguishing peptides between cohorts and the classification accuracy improved when the measured values for each array sequence were replaced by the corresponding predicted values. Effectively, the neural network aggregates information from the entire peptide dataset onto each of the predicted values. In Fig 5, random noise (sequence independent variation) is purposely added to the array. Since the noise is added to the \log_{10} of the binding value, a sigma of 1 corresponds to a 10-fold increase in distribution width, as can be seen in Fig 5A. In other words, a sigma of 1 randomly changes the linear values up or down by roughly an order of magnitude. As a result, multi-class classification of the original data with noise added performs poorly (Fig 5B, dashed lines). However, because the neural network predictions effectively aggregate the combined information from nearly 123,000 sequence/binding values in the generation of the sequence/binding relationship, random noise is dramatically reduced and a sigma of 1 has very little

effect on classification performance (Fig 5B, solid lines). This concept is taken further in Fig 6, where arrays that for technical reasons were rejected because of excessive noise or physical artifacts affecting part of the array are included in the simultaneous analysis of all samples and their excess noise and defects are effectively repaired by requiring that all binding be strictly sequence dependent. This is done without the network that creates the sequence-binding relationship having any information about which cohort is which in the analysis. The implication for array based diagnostic applications is that replacing a purely statistical approach like immunosignaturing with a structure-based approach provides a means of eliminating noise that is unrelated to the binding properties of the sequences while retaining the real patient to patient variance. Note that if this noise reduction approach was used in a diagnostic application, additional samples could be analyzed without retraining the entire model. Only the final layer, which contains all the discriminating information for the new samples, would need to be trained, and this is a linear optimization.

Using the neural network model itself for disease discrimination

As shown in both Figs 4 and 6, predicted binding values for a set of peptide sequences that approximately cover the same model sequence space as the array sequences can be used to discriminate between cohorts of samples just as well as predicted values of the original array sequences. In fact, it is the sequence/binding relationship that contains the discriminating information, and it is not necessary to use predicted binding to real sequences at all. In the neural network used here for simultaneous analysis of all samples, the output layer consists of one column corresponding to each sample. The length of the column is the same as the width of the last hidden layer (250 values in this case). The 250 values associated with each sample in this output layer, combined with a single bias value added at the end, contains all of the distinguishing information for that sample and can effectively be used to replace the ~123,000 sequence/binding values measured with only 251 values. Importantly, the neural network that gives rise to the final weight matrix has no information provided to it about which cohort each sample belongs to. The column vectors representing each sample are simply a mathematical representation of the data which combines sequence and binding information together in a much more compact form than the original ~123,000 sequence/binding pairs. Table 2 shows that this set of 251-element vectors can be used in place of the predicted binding values and perform in a classification analysis just as well. Fig 7 shows an unsupervised clustering using the algorithm UMAP [53,54]. Fig 7A shows the clustering that results from using the ~123,000 measured values as input. Fig 7B shows the clustering that results from using the 251 values of the final weight matrix + bias for each sample as input. In each case, UMAP reduced the dimensionality to 2 components. These component values are plotted for each sample and the different cohorts are color coded. When the original measured values are used for clustering (Fig 7A), Uninfected samples and Chagas samples are well separated from the virus samples, but the virus samples are less well separated from each other. This is reasonable considering the levels of pathogen similarity. However, when the final weight matrix of the neural network

Table 2. Classification Accuracy, Predicted Binding Values vs. Final Weight Matrix.

	HCV	Dengue	WNV	HBV	Chagas	Uninfected
Predicted Peptides	0.984	1.000	0.986	0.981	1.000	0.980
Final Weights	0.992	1.000	0.995	0.987	0.999	0.987

Errors on all values are +/- 0.001

<https://doi.org/10.1371/journal.pcbi.1010773.t002>

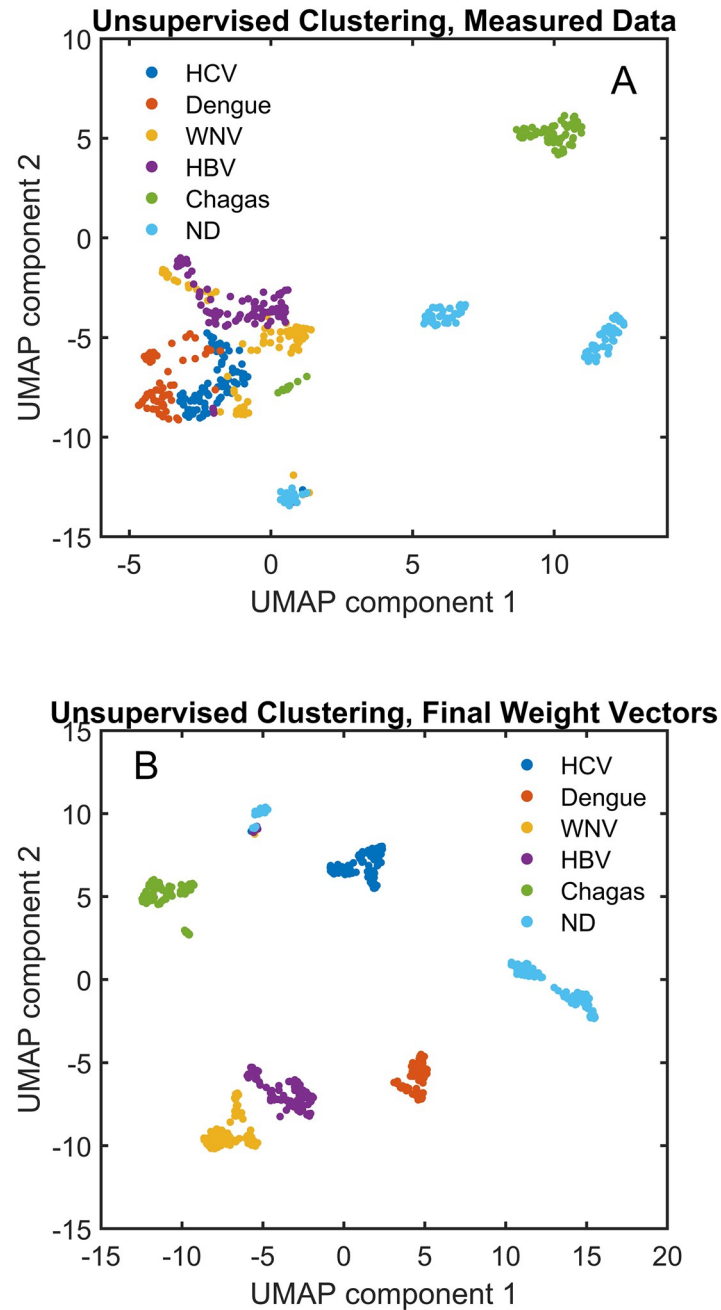


Fig 7. Unsupervised clustering of samples. A Matlab implementation of UMAP (Uniform Manifold Approximation and Projection, see text) was used to reduce (A) ~123,000 measured binding values and (B) 251 values from the final weight matrix of the neural network and the bias for each sample to 2 component values which are plotted. Samples are color coded by cohort.

<https://doi.org/10.1371/journal.pcbi.1010773.g007>

model is fed into the UMAP clustering algorithm, all of the cohorts are well separated. The plots provide a biological interpretation of what the neural network model is doing; the network extracts the discriminating information from the samples and focuses it in a much more concise form (251 element vectors in this case) which excludes sequence-independent noise. As demonstrated by [Table 2](#), once the sequence-binding relationship is created, there is

actually no need to calculate predicted binding values of all the sequences at all in order to distinguish the different cohorts effectively, as the concise final weight matrix contains all the relevant information.

Interestingly, in both panels of Fig 7 there is one small cluster consisting of different kinds of samples separated from the others. The cluster contains mostly uninfected individuals but also a small number of WNV, HBV and HCV samples (upper left, Fig 7B). UMAP is a nonlinear clustering algorithm which looks for the most similar features in samples to determine clustering. Apparently, this cluster of individuals had some other unknown immunological stimulus in common that distinguished them from all others. The ability to detect such clusters could prove useful in public health bio-surveillance applications.

Conclusions

While infectious disease diagnostics and associated classification approaches are one possible use-case of the approaches described above, the key conclusions of the work are more general. First, measured binding of serum antibodies to a very sparse and nearly random subset of all possible 10-mer peptides ($\sim 10^5$ out of $\sim 10^{12}$) is sufficient to capture all the relevant disease-specific information available from the array data (Figs 4, 5 and 6). Secondly, the representation of IgG binding is statistically comprehensive, as demonstrated by the high correlation (~ 0.96) between predicted and measured binding using peptide sequences not involved in the training of the network (Fig 2B). Thirdly, creating a single model to fit all samples simultaneously effectively captures disease specific information, and takes significantly less time (~ 400 seconds on a single GPU) than fitting each of the 465 samples separately (which takes about 12 hours even with parallel processing) (Fig 3). Fourth, the neural network imposes the requirement that all predicted binding values are consistent with their associated sequence, effectively removing any sequence-independent noise. Finally, the final layer of the neural network provides a condensed but complete representation of the disease-specific information in the array, with one column for each sample consisting of a compact set of 250 values plus a bias term instead of the $\sim 123,000$ binding values. Unsupervised clustering using the final weight matrix as input provides an excellent separation of samples into their respective cohorts (Fig 7).

Materials and methods

Peptide array content

The peptide arrays used were produced locally at ASU (see below). The synthesized wafers were cut into microscope slide sized pieces, each slide containing a total of 24 peptide arrays. Each array contained 122,926 unique peptide sequences that were 7–12 amino acids long (average of 10). A 3 amino acid linker consisting of GSG was attached to each peptide and connected the C-terminus to the array surface via amino silane. The peptides were synthesized using 16 of the 20 natural amino acids (A,D,E,F,G,H,K,L,N,P,Q,R,S,V,W,Y) in order to simplify the synthesis process (C and M were excluded due to complications with deprotection and disulfide bond formation and I and T were excluded due to the similarity with V and S and to decrease the overall synthetic complexity and the number of photolithographic steps required [55]. The arrays were created in 64 photolithographic steps (4 rounds through addition of the 16 amino acids) and sequences were chosen from the set to sample all possible sequences as evenly as the synthesis would allow. A detailed description of the amino acid composition of the arrays and peptide length distribution was published previously [34] (these arrays are referred to as CIMw189-S9 in that publication).

Array synthesis methods

Peptide arrays were synthesized photolithographically using the method described previously [17]. Briefly, an oxide-coated silicone wafer functionalized with a monolayer of aminosilane and coupled to Boc-glycine was prepared as array substrate. The wafer was coated with photoresist containing a photolabile acid generator. The binding sites on the wafer were created by photoactivating selected areas on the wafer with UV light using a series of physical photomasks resulting in a patterned deprotection of Boc-protected amines on the substrate. The photoactivation step was followed by the addition of a selected Boc-protected amino acid dissolved in a coupling solution to enable binding to the acid-activated regions. The process was repeated using a pre-determined sequence of amino acids and photolithographic mask pairs to create the peptide library with the desired sequences.

Serum samples

Deidentified serum samples were collected from three different sources: 1) Blood donors' samples from Creative Testing Solutions (CTS), Tempe, AZ, 2) LGC SeraCare, Milford, MA, and 3) Arizona State University (ASU) (Table 1). The dengue serotype 4 serum samples were collected from 2 of the above sources: 30 samples were provided by CTS and 35 samples were purchased by Lawrence Livermore National Labs (LLNL) from SeraCare before they were donated to the Center for Innovations in Medicine (CIM) in the Biodesign Institute at ASU. Uninfected/control samples consisted of 200 CTS samples and 18 samples from healthy volunteers at ASU. All deidentified infectious case samples came from CTS. All samples provided by CTS were residual samples collected from blood donors who were asymptomatic at the time of blood donation and were identified as test-reactive for infectious disease markers during blood screening at CTS. At the time of donation, blood donors agreed to the use of their samples in research. Serum samples were frozen shortly after collection and not thawed before being received as aliquots. ASU samples were collected under IRB protocol STUDY00002876: DHS Immunosignaturing—A Platform for Detecting and Identifying Multiple Infectious Diseases—July 2015. Serum samples were frozen at the time of collection and not thawed before being received as aliquots. Further sample description and in-house validation of disease state is described in the supplementary materials.

Sample processing and serum IgG binding measurements

Serum from the 6 sample cohorts (5 disease cohorts and uninfected) were diluted (1:1) in glycerol and stored at -20°C . Before incubation, each serum sample was prepared as 1:625 dilution in 625 μL incubation buffer (phosphate buffered saline with 0.05 Tween 20, pH 7.2). The slides, each containing 24 separate peptide arrays were loaded into an ArrayIt microarray cassette (ArrayIt, San Mateo, CA). Then, 20 μL of the diluted serum (1:625) was added on a Whatman 903T Protein Saver Card. From the center (12 mm circle) of the protein card, a 6 mm circle was punched, and put on the top of each well in the cassette, and covered with an adhesive plate seal (3M, catalogue number: 55003076). Incubation of the diluted serum samples on the arrays was performed for 90 minutes at 37°C with rotation at 6 RPM in an Agilent Rotary incubator. Then, the arrays were washed 3 times in distilled water and dried under nitrogen. A goat anti-human IgG(H+L) secondary antibody conjugated with either AlexaFluor 555 (Life Technol.) or AlexaFluor 647 (Life Technol.) was prepared in 1x PBST pH 7.2 to a final concentration of 4 nM. Following incubation with primary antibodies, secondary antibodies were added to the array, sealed with a 3M cover and incubated at 37°C for 1 hour. Then the slides were washed 3 times with PBST (137 mM NaCl, 2.7 mM KCl, 10 mM Na_2HPO_4 , and 1.8 mM KH_2PO_4 , 0.1% Tween (w/v)), followed by distilled water, removed from the cassette, sprayed

with isopropanol and centrifuged, dried under nitrogen, and scanned at 0.5um resolution in an Innopsys Innoscan 910 0.5 um laser scanner (Innopsys, Carbonne, Fr), excitation 547 nm, emission 590 nm. Each image was analyzed (GenePix Pro 6.0, Molecular Devices, San Jose, CA) and the raw fluorescence intensity data was exported as a GenePix Results ('gpr') file.

Binding analysis using neural networks

The neural network used to relate peptide sequences on the array to the measured binding of total serum IgG is similar to that described previously [34]. The amino acid sequences are input as one-hot representations and fed into a fully connected, feed-forward neural network (3 hidden layers, 250 nodes each), with each layer including a rectified linear unit activation function. This was followed by a linear final output layer (whose width is the same as the number of samples) used to predict total serum IgG binding. The neural network is trained on the peptide sequence/binding value pairs by optimizing an L2 loss function (sum of squared errors) between the measured and predicted binding values. The model performance is assessed by calculating the Pearson correlation coefficient and the RMSE between the measured and predicted binding values for a test dataset not involved in the training. Except where otherwise stated, the neural networks used in this work are trained on all samples simultaneously, where all weights of the neural network are shared across cohorts except for the final layer of the neural network. The neural network was trained using the \log_{10} of the median-normalized binding values from the peptide array (normalized by the binding values of all peptides in a given sample). Any zeros in the dataset were replaced by 0.03x the median prior to taking the logarithm. A diagram and description is given in the supplementary information (S2 Fig).

Classification algorithm

A second neural network was used as a multiclass classifier. The input to the network was a subset of binding values for each sample. In some cases, the binding values were measured values and in other cases predicted values. In Table 2, the columns of the final output weight matrix of the sequence-binding neural network above was used to represent each sample. In each of these cases, feature selection was performed with one vs. all T-tests. The binding values of each peptide in a cohort were compared to the binding values of all the other cohorts and 15 peptides from each of the six one vs. all comparisons were selected. These binding values (typically 83–84 because of some duplication between comparisons) were fed into the input layer of a neural network with 1 fully connected hidden layer with 300 nodes followed by a rectified linear unit activation function. This was followed by a linear output layer that was 300 x 6. The target values were one-hot vectors, 6 long, which represented the cohort that sample was associated with. A diagram and description is given in the supplementary information (S3 Fig).

Supporting information

S1 Fig. The correlation coefficient between the predicted and measured values for each of the 465 samples used in the analysis of Fig 2.

(TIF)

S2 Fig. Neural Network architecture used for generating sequence-binding relationship.

(TIF)

S3 Fig. Neural Network architecture used for multiclass classification.

(TIF)

S4 Fig. Noise reduction capability of the neural network. (A) a scatter plot of the \log_{10} of the measured data values with $\sigma = 1$ gaussian noise added vs. the \log_{10} measured data values without noise added. (B) in this scatter plot, the Y-axis is \log_{10} of the binding values predicted from a sequence-binding neural network model generated using the measured data values with $\sigma = 1$ gaussian noise added and the X-axis is again the \log_{10} of the measured values without noise. The function `dscatter` was used to generate a density-colored scatter plot (see the description of Fig 2B ref. 52 in the text). The neural network sequence-binding relationship effectively removed the sequence independent noise that was added to the data. (TIF)

Acknowledgments

The principal investigator on the DTRA grant that supported the collection of data used in this analysis was Professor Stephen Johnston; he provided valuable concepts and insights that underlie the background work upon which this analysis rests.

Author Contributions

Conceptualization: Robayet Chowdhury, Phillip Stafford, Chris Diehnelt, Neal W. Woodbury.

Data curation: Robayet Chowdhury, Laimonas Kelbauskas, Zhan-Gong Zhao.

Formal analysis: Robayet Chowdhury, Alexander T. Taguchi, Chris Diehnelt, Zhan-Gong Zhao.

Funding acquisition: Neal W. Woodbury.

Investigation: Laimonas Kelbauskas, Chris Diehnelt, Neal W. Woodbury.

Methodology: Alexander T. Taguchi, Laimonas Kelbauskas, Phillip Stafford, Neal W. Woodbury.

Project administration: Neal W. Woodbury.

Resources: Phillip C. Williamson, Valerie Green.

Software: Alexander T. Taguchi.

Supervision: Laimonas Kelbauskas, Neal W. Woodbury.

Visualization: Robayet Chowdhury.

Writing – original draft: Robayet Chowdhury, Neal W. Woodbury.

Writing – review & editing: Laimonas Kelbauskas, Neal W. Woodbury.

References

1. Akkaya M, Kwak K, Pierce SK. B cell memory: building two walls of protection against pathogens. *Nat Rev Immunol.* 2020; 20(4):229–38. <https://doi.org/10.1038/s41577-019-0244-2> PMID: 31836872
2. Cyster JG, Allen CDC. B Cell Responses: Cell Interaction Dynamics and Decisions. *Cell.* 2019; 177(3):524–40. <https://doi.org/10.1016/j.cell.2019.03.016> PMID: 31002794
3. Kotaki R, Moriyama S, Takahashi Y. Humoral immunity for durable control of SARS-CoV-2 and its variants. *Inflamm Regen.* 2023; 43(1):4. <https://doi.org/10.1186/s41232-023-00255-9> PMID: 36631890
4. Murin CD, Wilson IA, Ward AB. Antibody responses to viral infections: a structural perspective across three different enveloped viruses. *Nat Microbiol.* 2019; 4(5):734–47. <https://doi.org/10.1038/s41564-019-0392-y> PMID: 30886356

5. Geysen HM, Meloen RH, Barteling SJ. Use of peptide synthesis to probe viral antigens for epitopes to a resolution of a single amino acid. *Proc Natl Acad Sci U S A*. 1984; 81(13):3998–4002. <https://doi.org/10.1073/pnas.81.13.3998> PMID: 6204335
6. Desai N, Neyaz A, Szabolcs A, Shih AR, Chen JH, Thapar V, et al. Temporal and spatial heterogeneity of host response to SARS-CoV-2 pulmonary infection. *Nat Commun*. 2020; 11(1):6319. <https://doi.org/10.1038/s41467-020-20139-7> PMID: 33298930
7. Liston A, Humblet-Baron S, Duffy D, Goris A. Human immune diversity: from evolution to modernity. *Nat Immunol*. 2021; 22(12):1479–89. <https://doi.org/10.1038/s41590-021-01058-1> PMID: 34795445
8. Zheng B, Yang Y, Chen L, Wu M, Zhou S. B-cell receptor repertoire sequencing: Deeper digging into the mechanisms and clinical aspects of immune-mediated diseases. *iScience*. 2022; 25(10):105002. <https://doi.org/10.1016/j.isci.2022.105002> PMID: 36157582
9. Jaroszewicz W, Morcinek-Orłowska J, Pierzynowska K, Gaffke L, Węgrzyn G. Phage display and other peptide display technologies. *FEMS Microbiology Reviews*. 2022; 46(2):fuab052. <https://doi.org/10.1093/femsre/fuab052> PMID: 34673942
10. Liu BC, Dijohnson DA, O'Rourke DJ. Antibody profiling with protein antigen microarrays in early stage cancer. *Expert opinion on medical diagnostics*. 2012; 6(3):187–96. <https://doi.org/10.1517/17530059.2012.672969> PMID: 23480685
11. Takulapalli BR, Qiu J, Magee DM, Kahn P, Brunner A, Barker K, et al. High density diffusion-free nano-well arrays. *J Proteome Res*. 2012; 11(8):4382–91. <https://doi.org/10.1021/pr300467q> PMID: 22742968
12. Grötzinger C. Applications of Peptide Microarrays in Autoantibody, Infection, and Cancer Detection. *Methods Mol Biol*. 2023; 2578:1–15. https://doi.org/10.1007/978-1-0716-2732-7_1 PMID: 36152276
13. Fodor SP, Read JL, Pirrung MC, Stryer L, Lu AT, Solas D. Light-directed, spatially addressable parallel chemical synthesis. *Science*. 1991; 251(4995):767–73. <https://doi.org/10.1126/science.1990438> PMID: 1990438
14. Scott JK, Smith GP. Searching for peptide ligands with an epitope library. *Science*. 1990; 249(4967):386–90. <https://doi.org/10.1126/science.1696028> PMID: 1696028
15. Szymczak LC, Kuo HY, Mrksich M. Peptide Arrays: Development and Application. *Anal Chem*. 2018; 90(1):266–82. <https://doi.org/10.1021/acs.analchem.7b04380> PMID: 29135227
16. Weber LK, Palermo A, Kügler J, Armant O, Isse A, Rentschler S, et al. Single amino acid fingerprinting of the human antibody repertoire with high density peptide arrays. *J Immunol Methods*. 2017; 443:45–54. <https://doi.org/10.1016/j.jim.2017.01.012> PMID: 28167275
17. Legutki JB, Zhao ZG, Greving M, Woodbury N, Johnston SA, Stafford P. Scalable High-Density Peptide Arrays for Comprehensive Health Monitoring. *Nat Commun*. 2014; 5:4785. <https://doi.org/10.1038/ncomms5785> PMID: 25183057
18. Navalkar KA, Johnston SA, Woodbury N, Galgiani JN, Magee DM, Chicacz Z, et al. Application of immunosignatures for diagnosis of valley Fever. *Clin Vaccine Immunol*. 2014; 21(8):1169–77. <https://doi.org/10.1128/CVI.00228-14> PMID: 24964807
19. Brown JR, Stafford P, Johnston SA, Dinu V. Statistical methods for analyzing immunosignatures. *Bmc Bioinformatics*. 2011; 12. <https://doi.org/10.1186/1471-2105-12-349> PMID: 21854615
20. Kukreja M, Johnston SA, Stafford P. Comparative study of classification algorithms for immunosignaturing data. *Bmc Bioinformatics*. 2012; 13. <https://doi.org/10.1186/1471-2105-13-139> PMID: 22720696
21. Legutki JB, Magee DM, Stafford P, Johnston SA. A general method for characterization of humoral immunity induced by a vaccine or infection. *Vaccine*. 2010; 28(28):4529–37. <https://doi.org/10.1016/j.vaccine.2010.04.061> PMID: 20450869
22. Nayak BP, Putterman C, Gerwien R, Sykes K, Tarasow TM. IMMUNOSIGNATURE TECHNOLOGY IDENTIFIES SYSTEMIC LUPUS ERYTHEMATOSUS FROM A DROP OF SERUM. *Annals of the Rheumatic Diseases*. 2016; 75:1056-.
23. Restrepo L, Stafford P, Johnston SA. Feasibility of an early Alzheimer's disease immunosignature diagnostic test. *J Neuroimmunol*. 2013; 254(1–2):154–60. <https://doi.org/10.1016/j.jneuroim.2012.09.014> PMID: 23084373
24. Richer J, Johnston SA, Stafford P. Epitope identification from fixed-complexity random-sequence peptide microarrays. *Mol Cell Proteomics*. 2014. <https://doi.org/10.1074/mcp.M114.043513> PMID: 25368412
25. Scheck AC, Stafford P, Hughes A, Cichacz Z, Coons SW, Johnston SA. Immunosignaturing for the Diagnosis and Characterization of Human Brain Tumors. *Neuro-Oncology*. 2012; 14:100-.
26. Singh S, Stafford P, Schlauch KA, Tillett RR, Gollery M, Johnston SA, et al. Humoral Immunity Profiling of Subjects with Myalgic Encephalomyelitis Using a Random Peptide Microarray Differentiates Cases

- from Controls with High Specificity and Sensitivity. *Mol Neurobiol*. 2016. <https://doi.org/10.1007/s12035-016-0334-0> PMID: 27981498
27. Stafford P, Cichacz Z, Woodbury NW, Johnston SA. Immunosignature system for diagnosis of cancer. *Proc Natl Acad Sci U S A*. 2014; 111(30):E3072–80. <https://doi.org/10.1073/pnas.1409432111> PMID: 25024171
 28. Stafford P, Johnston SA, Kantarci OH, Zare-Shahabadi A, Warrington A, Rodriguez M. Antibody characterization using immunosignatures. *Plos One*. 2020; 15(3):e0229080. <https://doi.org/10.1371/journal.pone.0229080> PMID: 32196507
 29. Sykes KF, Legutki JB, Stafford P. Immunosignaturing: a critical review. *Trends Biotechnol*. 2013; 31(1):45–51. <https://doi.org/10.1016/j.tibtech.2012.10.012> PMID: 23219199
 30. Tarasow TM, Rowe MW, Haddad M, Sykes K. Immunosignature technology detects stage I lung cancer from a drop of serum. *Cancer Research*. 2015; 75.
 31. Rowe M, Melnick J, Gerwien R, Legutki JB, Pfeilsticker J, Tarasow TM, et al. An ImmunoSignature test distinguishes *Trypanosoma cruzi*, hepatitis B, hepatitis C and West Nile virus seropositivity among asymptomatic blood donors. *PLoS Negl Trop Dis*. 2017; 11(9):e0005882. <https://doi.org/10.1371/journal.pntd.0005882> PMID: 28873423
 32. Maeda D, Batista MT, Pereira LR, de Jesus Cintra M, Amorim JH, Mathias-Santos C, et al. Adjuvant-Mediated Epitope Specificity and Enhanced Neutralizing Activity of Antibodies Targeting Dengue Virus Envelope Protein. *Front Immunol*. 2017; 8:1175. <https://doi.org/10.3389/fimmu.2017.01175> PMID: 28993770
 33. Hughes AK, Cichacz Z, Scheck A, Coons SW, Johnston SA, Stafford P. Immunosignaturing Can Detect Products from Molecular Markers in Brain Cancer. *Plos One*. 2012; 7(7). <https://doi.org/10.1371/journal.pone.0040201> PMID: 22815729
 34. Taguchi AT, Boyd J, Diehnelt CW, Legutki JB, Zhao ZG, Woodbury NW. Comprehensive Prediction of Molecular Recognition in a Combinatorial Chemical Space Using Machine Learning. *ACS Comb Sci*. 2020; 22(10):500–8. <https://doi.org/10.1021/acscombsci.0c00003> PMID: 32786325
 35. Hecker M, Fitzner B, Wendt M, Lorenz P, Flechtner K, Steinbeck F, et al. High-Density Peptide Microarray Analysis of IgG Autoantibody Reactivities in Serum and Cerebrospinal Fluid of Multiple Sclerosis Patients. *Mol Cell Proteomics*. 2016; 15(4):1360–80. <https://doi.org/10.1074/mcp.M115.051664> PMID: 26831522
 36. Tokarz R, Mishra N, Tagliaferro T, Sameroff S, Caciula A, Chauhan L, et al. A multiplex serologic platform for diagnosis of tick-borne diseases. *Scientific Reports*. 2018; 8(1):3158. <https://doi.org/10.1038/s41598-018-21349-2> PMID: 29453420
 37. Xu GJ, Kula T, Xu Q, Li MZ, Vernon SD, Ndung'u T, et al. Viral immunology. Comprehensive serological profiling of human populations using a synthetic human virome. *Science*. 2015; 348(6239):aaa0698. <https://doi.org/10.1126/science.aaa0698> PMID: 26045439
 38. Xu GJ, Kula T, Xu Q, Li MZ, Vernon SD, Ndung'u T, et al. Comprehensive serological profiling of human populations using a synthetic human virome. *Science*. 2015; 348(6239).
 39. Ionov Y, Rogovskyy AS. Comparison of motif-based and whole-unique-sequence-based analyses of phage display library datasets generated by biopanning of anti-*Borrelia burgdorferi* immune sera. *Plos One*. 2020; 15(1):e0226378. <https://doi.org/10.1371/journal.pone.0226378> PMID: 31940357
 40. Pashov A, Shivarov V, Hadzhieva M, Kostov V, Ferdinandov D, Heintz KM, et al. Diagnostic Profiling of the Human Public IgM Repertoire With Scalable Mimotope Libraries. *Front Immunol*. 2019; 10:2796. <https://doi.org/10.3389/fimmu.2019.02796> PMID: 31849974
 41. Haynes WA, Kamath K, Waitz R, Daugherty PS, Shon JC. Protein-Based Immunome Wide Association Studies (PIWAS) for the Discovery of Significant Disease-Associated Antigens. *Front Immunol*. 2021; 12:625311. <https://doi.org/10.3389/fimmu.2021.625311> PMID: 33986742
 42. Haynes WA, Kamath K, Bozekowski J, Baum-Jones E, Campbell M, Casanovas-Massana A, et al. High-resolution epitope mapping and characterization of SARS-CoV-2 antibodies in large cohorts of subjects with COVID-19. *Communications Biology*. 2021; 4(1):1317. <https://doi.org/10.1038/s42003-021-02835-2> PMID: 34811480
 43. Asif M, Orenstein Y. DeepSELEX: inferring DNA-binding preferences from HT-SELEX data using multi-class CNNs. *Bioinformatics*. 2020; 36(Suppl_2):i634–i42. <https://doi.org/10.1093/bioinformatics/btaa789> PMID: 33381817
 44. Hare J, Morrison D, Nielsen M. Sampling SARS-CoV-2 Proteomes for Predicted CD8 T-Cell Epitopes as a Tool for Understanding Immunogenic Breadth and Rational Vaccine Design. *Frontiers in Bioinformatics*. 2021; 1. <https://doi.org/10.3389/fbinf.2021.622992> PMID: 36303758
 45. Hie B, Zhong ED, Berger B, Bryson B. Learning the language of viral evolution and escape. *Science*. 2021; 371(6526):284–8. <https://doi.org/10.1126/science.abd7331> PMID: 33446556

46. Shrock E, Fujimura E, Kula T, Timms RT, Lee IH, Leng Y, et al. Viral epitope profiling of COVID-19 patients reveals cross-reactivity and correlates of severity. *Science*. 2020; 370(6520). <https://doi.org/10.1126/science.abd4250> PMID: 32994364
47. Wu Z, Kan SBJ, Lewis RD, Wittmann BJ, Arnold FH. Machine learning-assisted directed protein evolution with combinatorial libraries. *Proc Natl Acad Sci U S A*. 2019; 116(18):8852–8. <https://doi.org/10.1073/pnas.1901979116> PMID: 30979809
48. Yoshida M, Hinkley T, Tsuda S, Abul-Hajja YM, McBurney RT, Kulikov V, et al. Using Evolutionary Algorithms and Machine Learning to Explore Sequence Space for the Discovery of Antimicrobial Peptides. *Chem*. 2018; 4(3):533–43.
49. Hare J, Morrison D, Nielsen M. Sampling SARS-CoV-2 proteomes for predicted CD8 T-cell epitopes as a tool for understanding immunogenic breadth and rational vaccine design. *Frontiers in Bioinformatics*. 2021; 1:1. <https://doi.org/10.3389/fbinf.2021.622992> PMID: 36303758
50. Jespersen MC, Peters B, Nielsen M, Marcatili P. BepiPred-2.0: improving sequence-based B-cell epitope prediction using conformational epitopes. *Nucleic Acids Res*. 2017; 45(W1):W24–w9. <https://doi.org/10.1093/nar/gkx346> PMID: 28472356
51. Greiff V, Redestig H, Lück J, Bruni N, Valai A, Hartmann S, et al. A minimal model of peptide binding predicts ensemble properties of serum antibodies. *BMC Genomics*. 2012; 13(1):79. <https://doi.org/10.1186/1471-2164-13-79> PMID: 22353141
52. Eilers PHC, Goeman JJ. Enhancing scatterplots with smoothed densities. *Bioinformatics*. 2004; 20(5):623–8. <https://doi.org/10.1093/bioinformatics/btg454> PMID: 15033868
53. McInnes L, Healy J, Saul N, Großberger L. UMAP: Uniform Manifold Approximation and Projection. *Journal of Open Source Software*. 2018; 3(29):861.
54. Meehan C, Ebrahimian J, Moore W, Meeha S. Uniform Manifold Approximation and Projection (UMAP) MATLAB Central File Exchange. 2022.
55. Stafford P. Pseudorandom vs. Random Polymers—How to Improve the Efficiency of Lithography-Based Synthesis. 2019; 1.