Medical & Biological Engineering & Computing (2022) 60:1279–1293
https://doi.org/10.1007/s11517-022-02539-7

ORIGINAL ARTICLE

Predicting residues involved in anti-DNA autoantibodies with limited neural networks

Rachel St. Clair[1] · Michael Teti[1] · Mirjana Pavlovic[2] · William Hahn[1] · Elan Barenholtz[1]

Received: 28 February 2021 / Accepted: 10 January 2022 / Published online: 18 March 2022

Abstract

Computer-aided rational vaccine design (RVD) and synthetic pharmacology are rapidly developing fields that leverage existing datasets for developing compounds of interest. Computational proteomics utilizes algorithms and models to probe proteins for functional prediction. A potentially strong target for computational approach is autoimmune antibodies, which are the result of broken tolerance in the immune system where it cannot distinguish "self" from "non-self" resulting in attack of its own structures (proteins and DNA, mainly). The information on structure, function, and pathogenicity of autoantibodies may assist in engineering RVD against autoimmune diseases. Current computational approaches exploit large datasets curated with extensive domain knowledge, most of which include the need for many resources and have been applied indirectly to problems of interest for DNA, RNA, and monomer protein binding. We present a novel method for discovering potential binding sites. We employed long short-term memory (LSTM) models trained on FASTA primary sequences to predict protein binding in DNA-binding hydrolytic antibodies (abzymes). We also employed CNN models applied to the same dataset for comparison with LSTM. While the CNN model outperformed the LSTM on the primary task of binding prediction, analysis of internal model representations of both models showed that the LSTM models recovered sub-sequences that were strongly correlated with sites known to be involved in binding. These results demonstrate that analysis of internal processes of LSTM models may serve as a powerful tool for primary sequence analysis.

Keywords Auto-immunity · Deep learning · DNA-binding · LSTM · Proteomics · Systemic lupus

1 Introduction

Computational proteomics utilizes algorithms and models to probe proteins for functional prediction. Primary research in this area is often devoted to computer-aided rational vaccine design (RVD) and synthetic pharmacology for effective drug design. A potentially strong target for such a computational approach is autoimmune antibodies, which are the result of broken tolerance in the immune system where it cannot distinguish "self" from "non-self," resulting in attack of its own structures (proteins and DNA, mainly). Despite decades of research, much remains poorly understood about the mechanisms underlying autoantibody function and binding processes.

Considered to be a hallmark of lupus disease, anti-DNA antibody is found in 70–90% of patients with SLE (particularly in those with nephritis), and measurements of its levels in patients' plasma are used to follow the course of disease. However, because anti-DNA antibody has been shown to be both hydrolytic and nephritogenic in a limited number of experimental and clinical studies, and that it also appears before the flare, it is suggested that it may serve as a strong flare predictor [2, 3, 38]. The important role of anti-DNA antibody is supported by studies in mouse models of nephritogenic lupus in which anti-DNA antibodies were found [31] as well as by the findings of [36] and [34]. The chemical structure and processes underlying autoantibodies remain poorly understood. [14, 25, 31] isolated anti-DNA and confirmed their DNA catalytic activities. However, only a small number of anti-DNA binding antibodies' binding sites have been determined. Almost an entire decade of X-ray crystallographic studies

✉ Rachel St. Clair
rstclair2012@fau.edu

[1] Center for Complex Systems and Brain Sciences, Florida Atlantic University, Boca Raton, USA

[2] Department of Electrical Engineering and Computer Science, Florida Atlantic University, Boca Raton, USA

performed by [9] combined with the most recent data generated by [8, 20, 37] observed that tyrosine and tryptophan residues create a hydrophobic pocket within the side chain of the antibody [24, 26]. Thus, oligo-thymidine pentamer enters the hydrophobic pocket between tyrosine and tryptophane from anti-DNA autoantibody in Fab fragment, where they bind to DNA, starting hydrolytic cleavage as a newly known modality of activity in autoimmune pathology (abzyme activity).

Wet-lab sequencing and X-ray crystallography are costly and time consuming, requiring expertise on each particular antibody. Computational approaches, which model existing data to generate novel predictions, can serve to narrow the field of possible candidates that may then be lab tested. Such models have become a standard tool in -omics research, with significant contributions to synthetic protein design and discovery. Recently, deep learning models have far exceeded earlier computational methods in complex feature detection from large datasets, as in the Large Scale Visual Recognition Challenge (ILSVRC) and machine-generated text models like GPT-2 [29, 32]. The unique ability of deep learning networks to define and manipulate important nonlinear features allows the possibility for such models to provide more insightful context than wet-lab and other traditional methods could alone. In recent years, deep learning has been applied to many areas within computational proteomics including protein folding, subcellular localization, and binding motif prediction, classification, and detection [17, 33, 40]. Indeed, nearly all recent computational approaches involve state of the art machine-learning including natural language processing (NLP) techniques, such as encoder-decoder networks and recurrent neural networks (RNNs), support vector machines (SVM), convolutional neural networks (CNNs), and use-case specific optimization algorithms, etc. [1, 21–23, 30, 41].

Most approaches to computational proteomics to date are heavily dependent on hand annotated datasets, supplementary feature input, require extensive background information, and/or are most frequently applied to large generic datasets. It is often the case in novel fields of interest that only limited, smaller datasets, lacking extra domain knowledge (i.e., evolutionary, MSA, tertiary structure data) beyond primary sequence, are available. To date, only a handful of studies have applied deep learning to primary sequence alone to perform protein class — though not binding site — prediction [16, 35]. With respect to binding site prediction, DNA and RNA specificities have been achieved using CNN, RNN, and hand-tailored MSA algorithms to other datasets (namely TFB and RNAB proteins) by [1, 21, 22]. However, these studies used both microarray and sequencing data. Most recently, [15, 44] achieved moderate accuracy in protein-protein interaction interface residue pairs prediction, but used supplementary data and hand-tailored algorithms for inference.

Thus to date, no computational studies have reported successful binding-site prediction from primary sequence alone. As noted, in the case of most novel topics of interest without supplementary domain knowledge, a model capable of analyzing primary sequence alone would be highly useful. Here we introduce a novel approach to achieve this goal based on analysis of hidden activation weights in RNNs, a family of deep learning models that include a form of memory, making them well adapted to analyzing sequential data. In particular, we used a class of RNNs — long short-term memory networks, or LSTMs — which include a memory cell to represent long-term memory allowing for sequential feature detection of position-specific input arrays [10]. Another type of RNN, GRU, combines input and forget gates from the LSTM in a single step, this exposes the whole state at each time step. In the LSTM, the degree to which each states' information is released to the surrounding nodes is more controlled by the separation of these gates. The rationale behind using LSTM is to have a better fine-tuned release of each state, which should capture residue-to-residue interactions better than the GRU. In summary, the GRU mixes each input (i.e., residue) with the rest of the network which could confound the downstream position-specific binding site extraction [6]. LSTMs are well adapted to learning from biological sequences, such as proteins, because of their ability to analyze sequences at multiple levels including whole protein, residue-to-residue interactions, and individual amino acids. More specifically, LSTMs are suited for the binding domain problem because proteins, presented as a primary sequence, can be evaluated beyond their linear representation for features across the primary space that might signify binding in later tertiary structures.

1.1 Current approach

Here, we trained several LSTM-based models to classify antibody primary sequences as DNA-binding or non-binding and then evaluated the model's hidden-states to assess the potential of specific sub-sequences and residues as binding sites. We designed our deep learning model to be fully compatible with the protein data warehouse Uniprot [7]. Since several previous efforts for binding specificities have employed both CNN and RNN models, we compared variants of CNN-based models of similar complexity on the same data. The novel methodology added to the parameter-limited networks used is the internal hidden state analysis. [16] used a similar LSTM model to predict phylogenetically distinct protein families by sequence alone and again in [15] predicted residue specificities but required more than just primary sequence data. Our work takes a similar approach in model architecture, making use of as few parameters as possible from only primary sequences, for the anti-DNA antibody problem set which is lacking in the amount of

biological data available. We directly apply this model to further the implications of the hydrolytic activity exerted on DNA by autoantibodies of various length and phylogenetically distinct protein families (e.g., IGG, IGM). We assessed the applicability of our technique to a small, unascertained problem set to directly elucidate the anti-DNA autoantibody phenomena in a way that allows insight into the model's inference process. To our knowledge, this work is the first application of small LSTM and CNN models to elucidate position-specific residues related to binding function from primary sequence alone and is the first computational model for anti-DNA antibodies.

In both LSTM and CNN cases, we use two models of different-sized trainable parameters to predict binding from primary sequences to better evaluate the use of limited parameters. We evaluated each of the models with regard to binding prediction accuracy. In addition, we evaluated the sub-sequences indicated by the hidden activations in the different models for agreement with previously identified binding sites. The resultant code can be found at https://github.com/mpcrlab/AntibodyBindingPrediction.

2 Related work

Although X-ray crystallography is capable of elucidating the DNA binding domain in an antibody, it is typically expensive and time consuming since only one highly reliable protein can be processed at a time. Research in this area, spanning the last several decades, has not achieved a comprehensive understanding of the antibody binding motif involved in DNA recognition and later hydrolytic activity. The most extensive wet-lab work has been completed by [8, 20, 37] for a few proteins, both synthetic and de novo. DNA-binding motif prediction for other target molecules has been achieved in several early works using MSA [27], physics-based simulations [11], and kernel based algorithms [19]. Deep learning–based approaches most often include using RNNs as in [15, 16, 18, 23, 30, 35, 44]. Some other recent works include combinations of CNN and RNN models [1, 21, 22, 43]. It is still unclear in the body of related works whether CNN, RNN, or combination models are more suited for modelucar interaction prediction. All of these studies depended on large datasets, supplementary data, and/or millions of model parameters.

Some recent work suggests that protein primary sequence may not be sufficiently high-dimensional enough for the successful application of deep learning techniques [42], which may account for the lack of sequence-only approaches in the literature. Their approach suggests including features beyond primary sequence in the goal of capturing predictions from a higher-dimensional representation of the underlying molecular biology. Nonetheless, we demonstrate here that primary sequence is high-dimensional enough for deep learning applications to predict binding-site with some degree of accuracy. Similar works in [1] DNA/RNA position-specific sites for TFB proteins were extracted using a brute-force approach by mutating each possible codon in areas of interest, determined by deep learning models, and accessing the respective binding score. This type of approach suffers from combinatorial explosion when applied to protein binding-sites as there are 27 possible residues (instead of four codons). This problem is exacerbated by variable protein length, which can often reach 2000 residues in length and shown to be problematic in the most similar works by [15]. [4, 12, 13, 39, 43] are similary unapplicable since they apply to gentic code which, as mentioned, is a smaller search space. This is demonstrated by these works, with the exception of [43], being classifyable with shallow-learning, or less non-linear to deep learning, approaches (e.g., anomoly detection, clustering, autoencoding, graph networks). Notably, the works of [28] predict DNA binding proteins from primary sequence alone using a LSTM that is given CNN feature vectors created from images of primary sequence, but the authors do not indicate the position-specific residues of the predicted binding, only whether the whole protein will bind or not. In the work presented here, we demonstrate a methodology for predicting which residues create the binding-site in the predicted binding proteins. of antecedent works are their ability to analyze giant datasets and high fidelity in their own applications. However, these works cannot be adapted to the problem presented in this work and others like it due to the limited domain information available, unreliability for sequence-only analysis, and completely forgo hidden state interpretation.

3 Methods

3.1 Dataset

An anti-DNA antibody dataset was curated directly from the protein data warehouse, Uniprot.org, using the query keywords: "Immunoglobuline" and "DNA-binding" in the manually annotated and reviewed records. This method supplied primary sequences of around 780 DNA binding–related antibodies. The counter class was sourced the same way with the exclusionary keyword "NOT+DNA-binding," which resulted in 1,267 antibodies. The data was first inspected for basic discrepancies between binding and non-binding antibodies by computing the amino acid frequency and sequence length between these two classes.

The generated dataset was found to include proteins of MHC and T-cell type that are not antibodies. This reflects the fact that Uniprot pulls all proteins associated with a keyword, but not exclusively; meaning, the sequences originally

retrieved relate to antibody function but may not be antibodies themselves. To create an unambiguous class of antibodies, we queried the generated data removing any proteins associated with MHC and T-cell keywords. After excluding these, only 75 antibody DNA-binding proteins remained. Therefore, we collected extra samples manually from the Protein DataBank website [5] using keywords, "DNA-binding" and "Antibodies." Thirty-three test sequences were hand-selected according to relevance and reliability. After removing duplicates and sequences with lengths less than 50 or greater than 2000 amino acids long, among all datasets, 81 binding antibodies were retrieved. The sequence lengths were constrained due to a two-fold rationale. First, including the few proteins greater than 2000 residues would create a batch of input that would be padded to a much larger max length, increasing the computations necessary for not only the LSTM but subsequent hidden state analysis. Secondly, proteins of residue lengths outside of 50–2000 were few and regarded as outliers since they likely misrepresent the general molecular structure of most anti-DNA binding antibodies. To downsample the much larger non-binding class into a generally representative dataset, we use principal component analysis (PCA) on multiple randomly selected samples of 81 proteins until the PCA more closely resembles the bind sequences' PCA (Fig. 1). Resemblance was determined by the operator matching the general shape of the PCAs as the random samples were selected. The process of sampling was necessary because some random samples occasionally (approximately three out of 20 samples) resulted in PCAs with a different shape, likely due to a small random sample of the data sharing alternate principal components than the main dataset. This dataset was split into training and validation by randomly sampling and checking for sequence length balance. Again, random samples from a small dataset can have the chance of pulling a non-representative sample, so multiple sampling was done to reject any subsamples that by chance, had an obvious skewed representation of the overall sequence lengths (Fig. 2). Finally, this secondary dataset consisted of 61 sequences reserved for training and 20 for validation for the LSTM and CNN binding inference.

3.2 Pre-processing

3.3 Data augmentation

The downsampled dataset of 81 sequences in each class was converted to one-hot images and augmented in two ways. As an LSTM evaluates a sequence, it uses recurrent information to update the hidden state in a way that leads to a correct classification. The hyper-variable domain (HVD), Fab fragment, is most likely to be involved in ligand binding recognition in antibodies and is often written first in FASTA sequences. Since the hidden state is lacking recurrent information in the beginning of each sequence analysis, the hidden state values are often much larger than later time steps in the data image. Therefore, to preserve the hidden state's attention to the important HVD, we reversed all sequences. Data was then augmented with horizontal flips to increase the amount of data available since the one-hot encoding was arbitrarily created from left to right. Augmentation did not significantly increase model performance, but may have increased the robustness of the hidden state evaluation.

3.4 Binding prediction

3.4.1 LSTM prediction

With a total of 244 training sequences and 80 validation sequences, the model was trained with one LSTM layer of 300 hidden nodes, a 50% dropout layer, and a 2-node fully connected layer for 200 epochs with a batch size of one. The preprocessed data, now with each residue position represented as a row and the corresponding residue letter as a column, is given to the LSTM as input. The hidden and cell state vectors are each initialized with zero values and a shape of 1x1x200. The LSTM reads in the protein at each timestep, or row, along with previous hidden state, passing it to four internal gates. Each gate has a corresponding set of weights that are matrix multiplied by the input and hidden vectors. The feature maps output by each gate are then combined according to Fig. 3, to update the hidden and cell state vectors. Each gate creates a feature vector according to Eq. 1, where the activation is sigmoid except for the forget gate which is activated with the hyperbolic tangent and W represents a unique corresponding weight to each gate. After the LSTM computes these operations for each timestep in the input protein, the resultant hidden and cell state vectors are passed through the rest of the model which consists of a rectified linear activation unit, dropout layer, and a sigmoid or softmax activated fully connected layer with 2 outputs. This model has 395,402 trainable parameters, which is considered quite small in the deep learning community. Adam and cross-entropy loss were used as the criterion and optimizers for the model parameters. Due to variability in accuracy caused by random weight initialization and random batch sampling during training, 100 identical models were trained. The same process was repeated for a smaller LSTM with only 200 hidden nodes incurring 183,602 parameters. We later observed an increase in LSTM prediction accuracy with the addition of a final sigmoid activation and thus included it in both small and large LSTM model variants. Model weights were saved according to their best validation accuracy scores for later hidden state extraction.

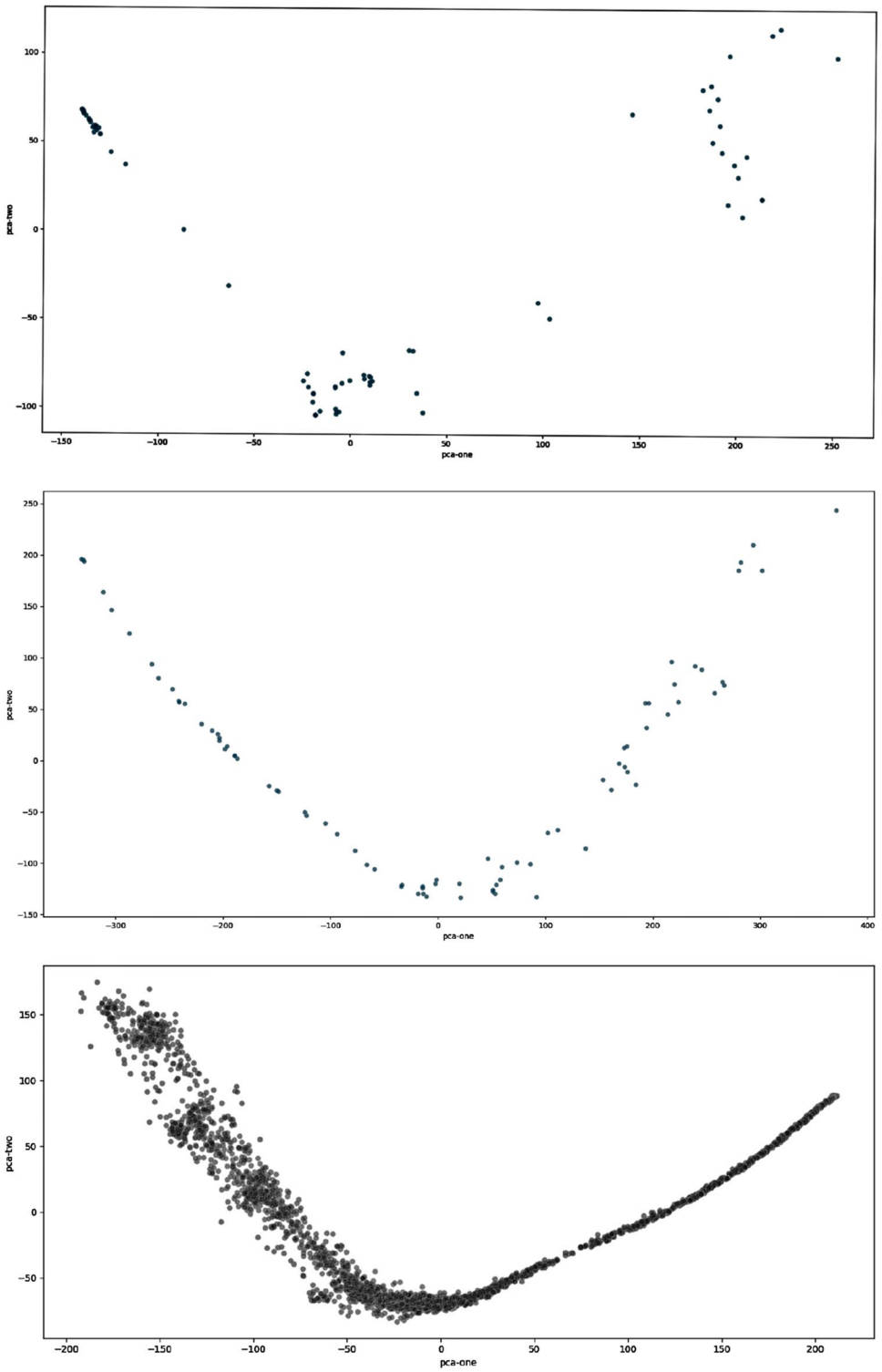

Fig. 1 Sequence PCA. Binding dataset PCA (left) and non-bind dataset PCA after (middle) and before downsampling (right)

$$activation(x(t) \bullet W_1 + h(t-1) \bullet W_2 + W_3) \quad (1)$$

Equation 1 Typical gate equation for LSTM gate, where W is a unique weight for that gate, x(t) is the input at timestep t, and h(t-1) is the previous hidden state vector.

3.4.2 CNN prediction

The CNN was designed to have a similar number of limited parameters (394,425) as the LSTM network in order to equate the models to the greatest extent possible. Sequences

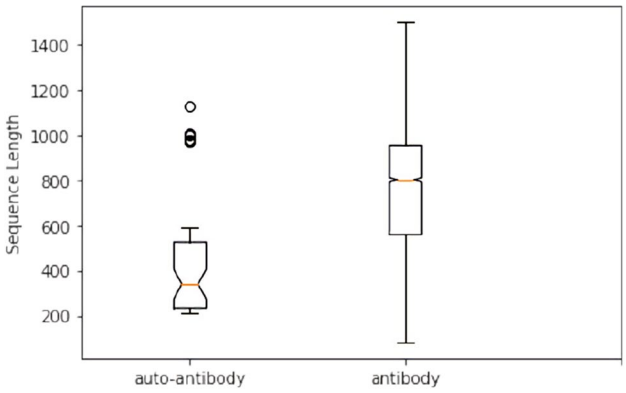

Fig. 2 Sequence length per class

were encoded as one-hot grayscale images, padded with an arbitrary value to the maximum sequence length of 1,750 and sampled with a batch size of one. The model consisted of three convolutional layers each followed by dropout (.5) and rectified linear unit (ReLU). Each convolution kernel was 3x3, 5x5, and 1x1, respectively, per layer and drop-out with rectified linear unit was used after each convolution layer. The network's final linear layer outputted 2 nodes and was evaluated by cross-entropy loss and ADAM optimizer. The last convolutional layer was designed to retain the input sequences' size outputting one feature map of size 1,751 by 28. A smaller variation of the model with only 183,743 parameters was also evaluated (Fig. 4). Hidden states were extracted from the best performing CNNs by summing the last convolutional layer's feature map across all sequences and then across the one-hot encoding dimension. This allowed for total activation for each position to be calculated and processed similarly to the LSTM hidden state analysis. The CNN model variant was trained 100 times separately to account for random weight initialization. Best performing models were selected according to the same procedure as used in the LSTM binding-site analysis method.

3.5 Binding-site analysis

Evaluation of the LSTM states' hidden layer and CNN feature map activations were performed by extracting the respective weights from all correctly predicted, reversed sequences at each time step for the top five performing models from each model variant. Since negative weights do not necessarily mean negation in class prediction, the absolute value of all hidden cell activation weights (LSTM) and last convolutional layer (CNN) feature maps were recovered. Top models were those that most accurately predicted all sequences during testing of all 81 sequences in each class using the previously trained models' learned weights. Once the best performing test models were determined, their original training and validation loss and accuracy trends were evaluated for obvious overfitting (i.e., poor training accuracy, loss in validation lower than loss in accuracy, etc.). All hidden cell weights were reversed so positions now align with FASTA formatting (position zero is the first residue in the sequence and so forth). PCA was performed on the hidden cell weights of all top models combined for LSTM and CNN, respectively. All sequence weights were then summed per class,

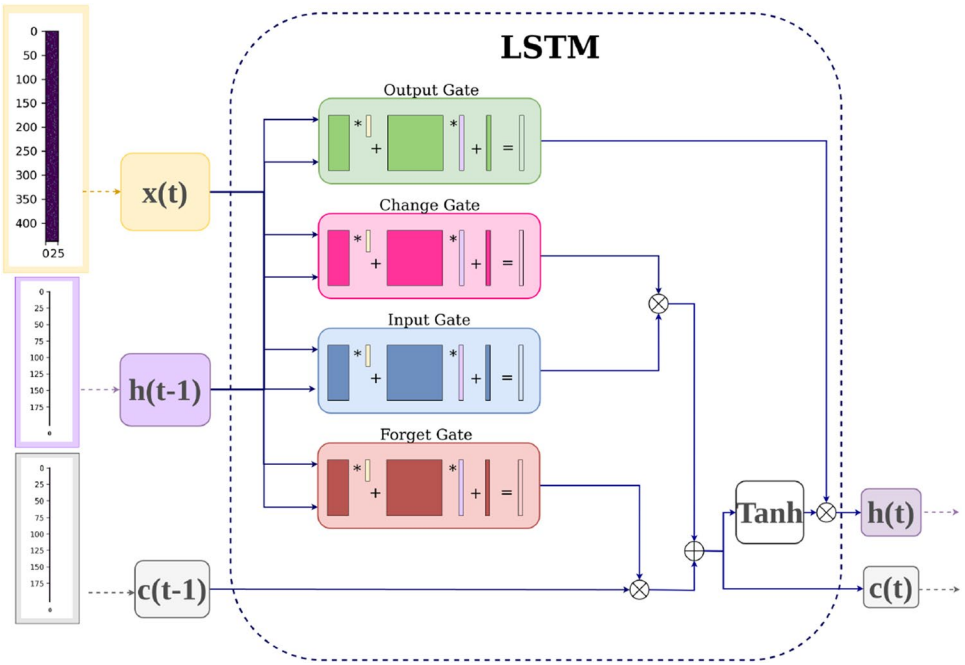

Fig. 3 LSTM model architecture

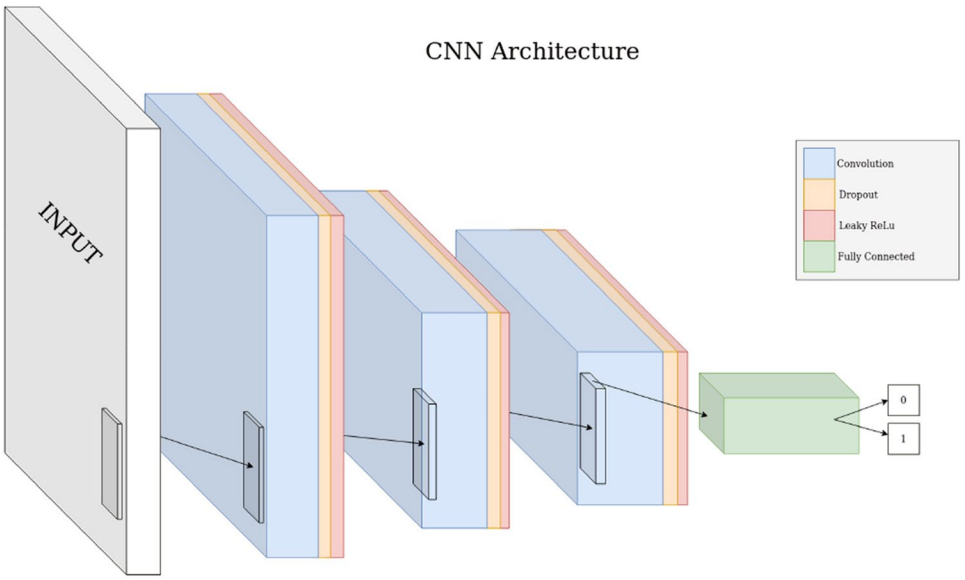

Fig. 4 CNN model architecture

per model. Differences in position-specific areas of interest were first visualized in all weights for each time step across the summed weight matrices. Activation weights were then summed across all nodes per class and scaled between 0 and 1 for comparison. The following tests were performed collectively on the five top models for each of the four model variants.

3.6 DNA-1 anti-DNA autoantibody

Activation weights for DNA binding antibody DNA-1 were recovered individually and compared to the position-specific residues important for binding given by X-ray crystallography reported in [37]. To remove the models' internal representations of non-binding proteins and make activations more interpretable, convolutions were performed with $v=[\frac{1}{6},\frac{1}{6},\frac{1}{6},\frac{1}{6},\frac{1}{6},\frac{1}{6}]$ without decreasing activation length on the absolute difference between the standardized DNA-1 and standardized non-binding class activation sums (Eq. 2). Here, x and y are activations of DNA-1 and non-bind activations, respectively. In later analysis, the second term including DNA-1 is removed. Peak activations were then determined using operator set thresholds.

$$\sum_{m=-\infty}^{\infty} v\left(\left|\frac{x-\mu}{\sigma}-\frac{y-\mu}{\sigma}\right|\right)-v\left(\left|\frac{d-\mu}{\sigma}-\frac{y-\mu}{\sigma}\right|\right) \quad (2)$$

Equation 2 Standardized convolution operation to process LSTM hidden states for binding site prediction, where x is a binding activation, y a non-binding activation, d is the DNA-1 activation, σ is the standard deviation of the respective activation, μ is the respective average activation, and v is the convolution operation.

3.6.1 Knockout test

To validate the activations provided by the hidden state's analysis on DNA-1, a "knockout" test was performed. We reasoned that if the suspected autoantibody binding site is being used for class prediction by the model, once such information is removed, the model should be more likely to predict the sequence as non-binding. This process is similar to genetic knockout in traditional transgenic mice models. For each non-binding sequence, the residues at the literature binding sites were transplanted into a copy of the DNA-1 sequence at the literature binding site positions. For each transplant, sequences in both classes were paired according to similar lengths. For binding sequences that were paired with a non-binding sequence of shorter length, a random non-binding sequence was chosen to fill in the binding sites exceeding the original sequence length. Only five non-binding sequences were smaller than the last binding site position, 327. These modified DNA-1 sequences were reversed and evaluated by the top trained prediction models. Hidden states were extracted and processed according to equation one between knockout and nonbinding class activations and compared to DNA-1. To reduce noise between major peaks found in both DNA-1 and knockout activation outside the literature binding sites, the difference between DNA-1 and knockout greater than zero provided an alternative bind site prediction for peaks at various operator-set thresholds.

3.6.2 Insertion test

To determine if binding sites are all that is necessary for class prediction, this process was repeated for an "insertion" test. Again, this is similar to transgenic mice models. DNA-1 literature binding sites were transplanted into non-binding

sequences of similar length. If a non-binding sequence was shorter than 444 residues, the length of DNA-1, only the available binding site positions were swapped and lengths were retained. Sequences were reversed and evaluated by the top trained models. Hidden states were extracted and processed according to equation one between insertion sequences' activation and non-binding sequences' activation. The insertion activation was then compared to the DNA-1 activation and literature known binding sites.

3.6.3 Peak knockout test

The knockout test described previously relies on literature binding site knowledge. The work proposed here is attempting to provide viable suggestions for proteomic interactions in cases where domain knowledge is extremely limited, as is often the case for synthetic protein design. As a method of predicting binding sites in such cases, another knockout test was performed on the major peaks in DNA-1 activations (i.e., "peak knockout"). This approach is similar to the original knockout test, but instead of using literature known binding sites, we use the positions predicted by the model's peak activation sites. The literature known binding site for DNA-1 is 66 residues, approximately 15% of the total sequence. Therefore, to make balanced comparisons with the original knockout test, peaks above the 58% threshold resulted in 68 residues to be modified in the subsequent test. Similar thresholds were chosen for the smaller LSTM and CNN model variants. Positions of these peaks were then used as the sites that were replaced by non-binding sequence residues. All sequences were evaluated accordingly with the original knockout test procedure. Comparisons were then made between the peak knockout and DNA-1 activations. To reduce noise in the activations, the difference between DNA-1 and peak knockout activations yield an alternative binding site suggestion. Final binding site sub-sequences were found by overlapping the activations created by previous DNA-1 analysis peaks and peak knockout occluded DNA-1 activations peaks.

4 Results

4.1 Preliminary data processing

Boxplots of antibody and autoantibody sequence lengths indicated that the median sequence length was significantly lower in autoantibody sequences than in antibody sequences, as determined by the non-overlapping notches between the two boxes (Fig. 2). This informed the use of the LSTM network and our specific CNN implementation because these did not require that all sequences were of the same length. In contrast, a model with fixed length inputs would require the shorter sequences to be padded to the maximum length, and this padding could potentially be exploited by a model in predicting whether a given sequence was an antibody or autoantibody.

Table 1 Average validation accuracies across LSTM and CNN models

Model	Sigmoid	Parameters	Avg. validation accuracy
LSTM	No	395,402	53.41% t(99)=5.7345, p .0001
LSTM	Yes	395,402	66.21% t(99)=11.7782, p .0001
LSTM	Yes	183,602	72.64% t(99)=20.5201, p .0001
CNN	Yes	395,425	63.80% t(99)=7.2982, p .0001
CNN	No	395,425	73.34% t(99)=12.0717, p .0001
CNN	No	183,743	87.81% t(99)=90.9673, p .0001

Table 2 Average test accuracies for small LSTM model for each class

Class	Model	Average test accuracy
Binding	LSTM	87.07%
Non-binding	LSTM	88.56%
Binding	CNN	96.56%
Non-binding	CNN	97.81%

4.2 Binding specificity

Table 1 shows the accuracies for each of the four models. Across both LSTMs and CNNs, the smaller variants resulted in better binding prediction accuracy. The smaller CNN model had the best average validation accuracy across 100 models and was statistically significant at 87.81% by one sample t-test, t(99) = 90.9673, p .0001. The smaller LSTM binding prediction achieved average validation accuracy of 72.64% and was also statistically significant, t(99) = 20.5201, p .0001. Notably, the LSTM prediction performance increases with added sigmoid activation, while the same effect is not observed for CNN. During the testing phase on the trained models, average accuracy score was 87.07% for the binding class and 88.56% for the non-binding class by the LSTM and 96.56%, 97.81% by CNN, respectively, for each smaller variant and similarly observed for larger variants (Table 2). The chosen top models for hidden state analysis were all above 95% accurate on all sequences. PCA alone was not enough to separate binding from non-binding proteins (Fig. 5). The percentage of variability for each principal component was 26.2%, 7.2%, and 4.4% respectively, accounting for 37.8% of the total variance in the dataset. Raw activation weight visualization showed distinct horizontal bands at particular time steps (Fig. 6).

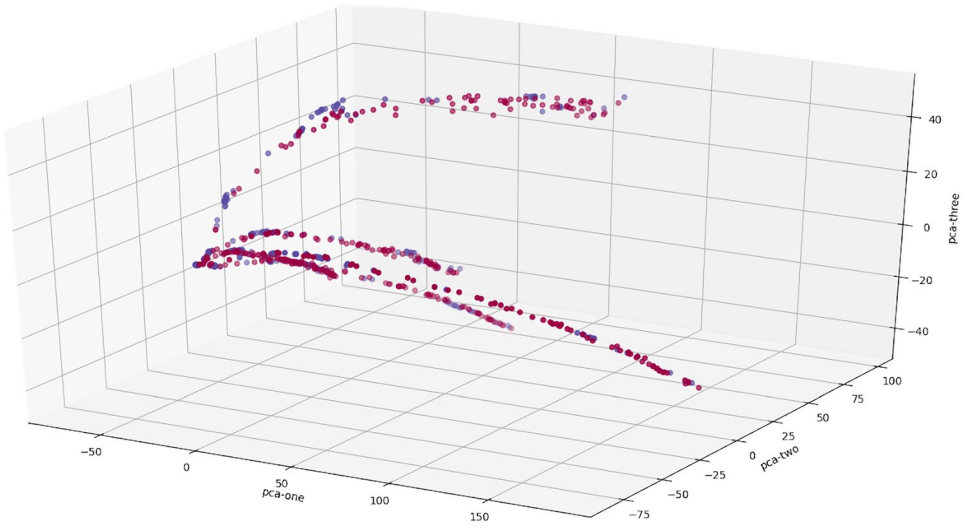

Fig. 5 PCA on hidden cell weights. Each hidden cell activation matrix is encoded per sequence sample for binding (red) and non-binding (purple) classes

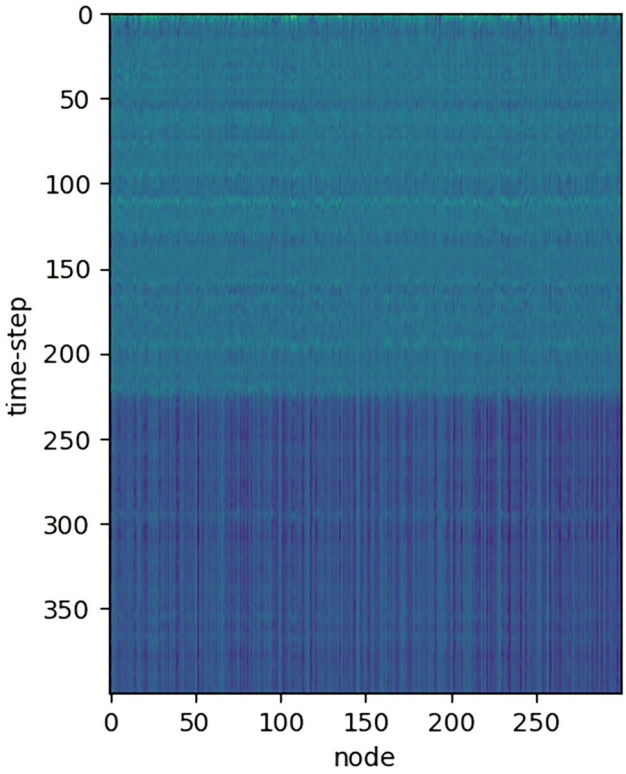

Fig. 6 Raw activation differences between classes. Difference in bind and non-bind hidden cell weights from LSTM model variant

Activation sums showed distinct peaks at different positions per class (Table 3).

4.3 Binding-site analysis

Table 2 shows correlation and significance between the suggested binding sites at each stage of processing for each model. The larger LSTM achieved the best binding site recovery and is reported in the proceeding results. Similar results and figures were observed in other model variants. DNA-1, binding, and non-binding activations had different unique peaks across different positions (Fig. 7). DNA-1 literature-defined binding site and hidden state activation sums overlapped in several major peaks (Fig. 8) but were not significantly correlated, r(798) .0627, $p < .1$. Processed DNA-1 activations according to equation one, r(798) = .1870, $P < .001$ and subsequent peaks at threshold 85%, r(798) = .2476, $p < .0001$, and threshold 58%, r(798) = .1433, $p < .01$, were significantly correlated. Knockout testing results showed on average 86.17% reversal from binding to non-binding class prediction. Hidden state analysis showed similar general trends between the knockout and the DNA-1 activations and was not significant, r(798) = -.0243, $p < .1$. At threshold 58%, however, trending significance was observed between knockout and literature known sites, r(798) = -.0903, $p < .1$. Discrepancies in

Table 3 Pearson correlation coefficients and significance values for binding site activation's at various processing steps compared to literature bind site for DNA-1, df=798

	LSTM	LSTM	CNN	CNN
	Small variant	Large variant	Large variant	Small variant
DNA-1	-0.0799, 0.1106	0.1433, 0.0041	0.0134, 0.7900	0.1071, 0.0322
Occluded	-0.1013, 0.0428	0.2566, 1.9567e-7	-0.1380, 0.0057	0.0750, 0.1341
Overlap	-0.1094, 0.0286	0.3674, 3.1232e-14	-0.0986, 0.04883	0.0711, 0.1555

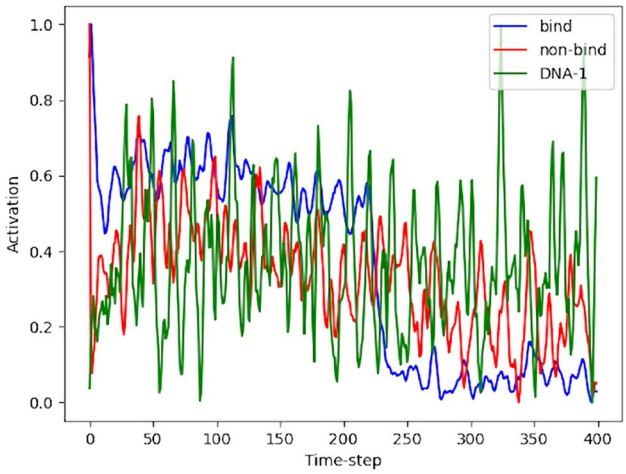

Fig. 7 Non-binding, binding, and DNA-1 activations from LSTM hidden cell standardized between 0 and 1

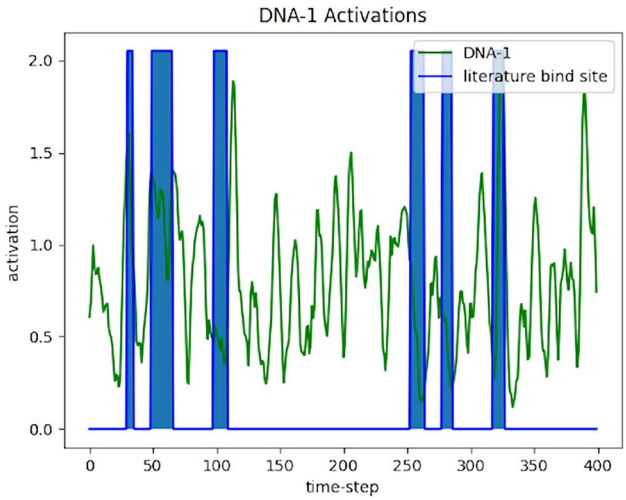

Fig. 8 DNA-1 standardized activations before equation one processing, r(798) = 0.053

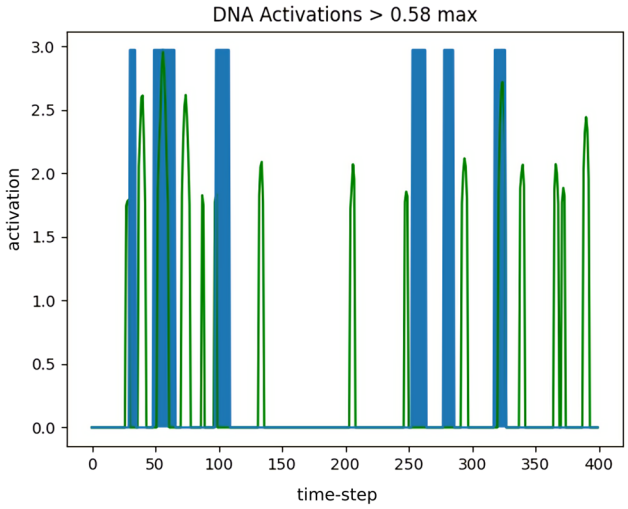

Fig. 9 DNA-1 activations processed according to equation one for peaks at 58% threshold, r(798) = 0.143), p .01

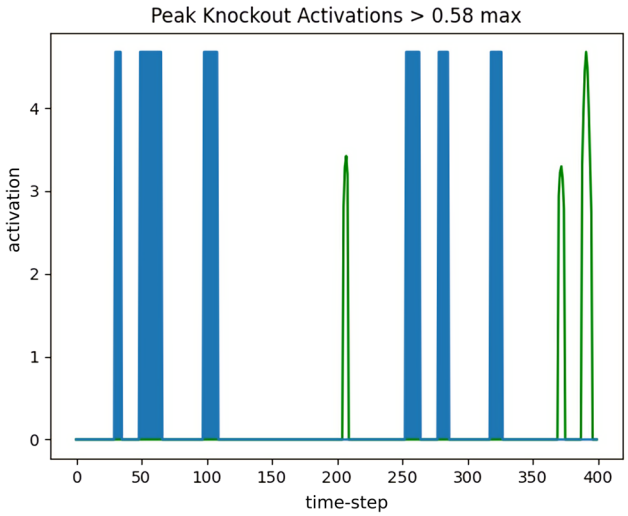

Fig. 10 Knockout of DNA-1 peaks processed according to equation one for peaks at 58% threshold, r(798) = -0.088, p=.013

activation difference between knockout and DNA-1 were least in areas outside the literature binding sites. Thus, occlusion of knockout activations from DNA-1 activations shows significant noise reduction between literature known binding sites, r(798) = .2140, $p < .0001$. Subsequent peaks for threshold 58% were not significant, r(798) = .0328, $p <$ .1. However, lowering this threshold to 50% resulted again in significance between the knockout occluded DNA-1 activations and literature binding sites, r(798) = .1122, $p = .02$. Insertion test was less effective, showing only 6.67% reversal on average with upper bounds at 9.87%. Hidden state analysis showed insertion activations very closely following DNA-1 activations, but without significant correlation to the binding sites, r(798) = .0500, $p <$ .1. However, only looking at peaks above 58% threshold did show significance, r(798) = .1018, p =.04. Knockout of all DNA-1 peaks greater than 58% the max peak (Fig. 9) showed on average 80.74% prediction reversal, with three models correctly predicting 80 or more sequences. Hidden states analysis was similar to that of previous knockout, with activations closely following DNA-1 activations with most discrepancies inside literature binding regions, resulting in non-significance, r(798) = -.0415, $p < .1$. Peaks above 58% threshold were trending towards significant, r(798) = -.0880, $p < .1$ (Fig. 10). Noise reduction by difference between DNA-1 and knockout peak activations showed significant overlap between the literature known

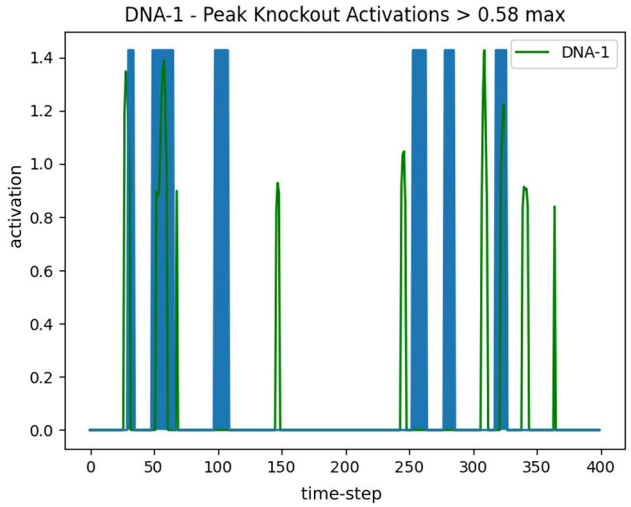

Fig. 11 Peak knockout occluded DNA-1 activations processed according to equation on for peaks at 58% threshold, r(798) = 0.257, p .01

binding sites and the model suggested sites peaks at 58% threshold, r(798) = .2566, $p < .0001$ (Fig. 11).

4.4 Binding sub-sequences

Recovered sub-sequences found by the 58% threshold peaks of DNA-1, peak knockout procedure (without overlap) recovered significant partial residues in literature binding regions, r(798) = .256, $p < .0001$. In Figs. 12 and 13, red letters indicate model suggested regions of interest, bold as literature suggested binding residues, and underlined regions indicating correctly suggested residues. Six other residues and sub-sequences outside the reported binding sites were also suggested. There is no apparent trend in the nature of residues suggested in blatant misses. However, some near-hits are often only a few residues premature of reported binding site sub-sequences. Overlap between DNA-1 peaks and peak knockout occluded DNA-1 peak activations resulted in highest sub-sequence fidelity in significance, r(798) = .367, $p < .0001$. This final method resulted in shorter and less frequent misses. Only one region that was originally a hit was missed; however, a near miss suggestion was directly next to the target residues in this site. Agreement between CNN and LSTM binding sub-sequences is shown in Fig. 14.

5 Discussion

5.1 Binding classification

As expected, model architecture configurations with less trainable parameters performed better. Since the optimizable gradients are less complex, the smaller dataset used in this application is more thoroughly integrated during back propagation. Overall, CNN models perform better on prediction tasks. CNN models have an spatio-temporal inductive bias while LSTM are temporally biased, resulting in features that reflect different types of patterns across the protein's residues. As demonstrated by works aforementioned, a combination of both models has been effective at increasing class prediction from sequence alone. Thus, a combination of both

Fig. 12 Sub-sequences found in knockout peak analysis. Peaks according to non-bind occluded DNA-1 activation peaks at 58% threshold and model suggested peaks via knockout peak occluded DNA-1 activations where red indicates model suggested sites, bold is literature binding sites, and underline is the overlap of the two

QVKLLESGPELVKPGASVKMSCKASGYTFTSY**VMH**WVKQKPGQGLEWIG
Y**IN**PYNDGTKYN**EKFKG**KATLTSDKSSSTAYMELSSLTSEDSAVYYCVR**GG**
YRPYYAMDYWGQGTSVTVSSAKTTPPSVYPLAPGSAAQTNSMVTLGCLV
KGYFPEPVTVTWNSGSLSSGVHTFPAVLQSDLYTLSSSVTVPSSTWPSET
VTCNVAHPASSTKVDKKIVPRDCTSHHHHHHELQMTQSPASLSASVGETV
TITC**RASENIYSYLA**WYQQKQGKSPQLLV**YNAKTLAE**GVPSRFSGSGSGTQ
FSLKINSLQPEDFGSYYC**QHHYGTPLT**FGAGTKLELKRADAAPTVSIFPPS
SEQLTSGGASVVCFLNNFYPKDINVKWKIDGSERQNGVLNSWTDQDSKD
STYSMSSTLTLTKDEYERHNSYTCEATHKTSTSPIVKSFNRNEC

Fig. 13 Overlap of DNA-1 and difference between DNA-1 with peak knockout peaks at 58% and sub-sequence comparison with literature binding site where red indicates model suggested sites, bold is literature binding sites, and underline is the overlap of the two

QVKLLESGPELVKPGASVKMSCKASGYTFT**SYVMH**WVKQKPGQGLEWIG
Y**IN**PYNDGTKYN**EKFKG**KATLTSDKSSSTAYMELSSLTSEDSAVYYCVR**GG**
YRPYYAMDYWGQGTSVTVSSAKTTPPSVYPLAPGSAAQTNSMVTLGCLV
KGYFPEPVTVTWNSGSLSSGVHTFPAVLQSDLYTLSSSVTVPSSTWPSET
VTCNVAHPASSTKVDKKIVPRDCTSHHHHHHELQMTQSPASLSASVGETV
TITC**RASENIYSYLA**WYQQKQGKSPQLLV**YNAKTLAE**GVPSRFSGSGSGTQ
FSLKINSLQPEDFGSYYC**QHHYGTPLT**FGAGTKLELKRADAAPTVSIFPPS
SEQLTSGGASVVCFLNNFYPKDINVKWKIDGSERQNGVLNSWTDQDSKD
STYSMSSTLTLTKDEYERHNSYTCEATHKTSTSPIVKSFNRNEC

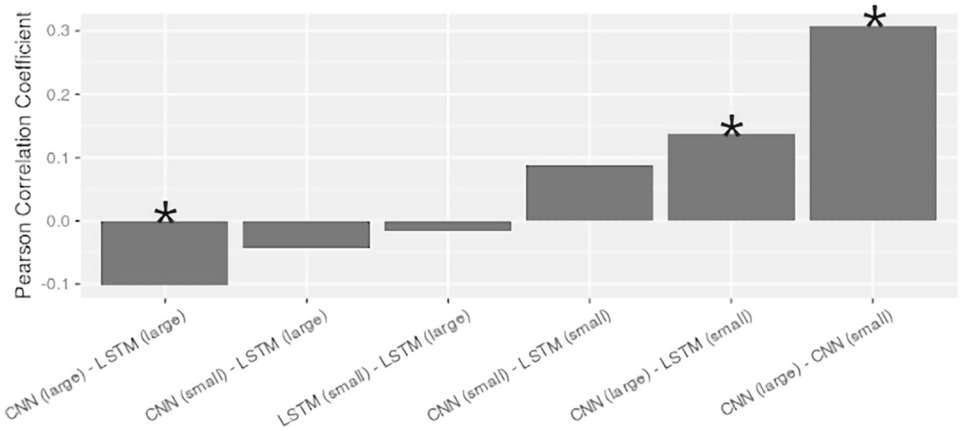

Fig. 14 Agreement between model variants on sub-sequence prediction. From left to right: r(798) = -0.1018, $p < .05$; r(798) = -.0442, $p < .1$; r(798) = -.0155, $p < .1$; r(798) = .0892, $p < .1$; r(798) = .1369, $p < .01$; r(798) = .3070, p .001

types of models could best aid in a more vigorous approach to vaccine and drug design. Future work could clarify how a combined approach might affect the prediction of the binding-site residues from the hidden state analysis.

Model variability in training is due to differences in random weight initialization and order of batch sampling creating hurdles in overcoming local minima during parameter optimization. Model prediction had a slight bias for the non-binding class, supporting the absence of overfitting to the binding dataset. Models with higher test accuracy, even if they did not perform as well during the training phase, show accurately learned weights for class recognition, which can be extended to novel or synthetic proteins outside the training and validation dataset. Activation differences across all nodes at specific positions suggest model decisions are portrayed differently per class in the hidden weights. Furthermore, the raw activation across nodes shows the model relies most on sequence positions up to 225 while following positions are less important for binding class prediction. Since most sequences are around this length or greater than, this supports the HVD region, at the beginning of the FASTA, being the most common region of binding. These areas of interest are further shown in the activation sums which lead to the sub-sequence distribution.

5.2 Binding site analysis

Hidden state analysis showed greater correspondence to previously established binding sites in the LSTM vs. CNN models. This is likely because LSTM encodes position-specific information rather than CNN which detects spatial-temporal invariant features. CNN top models suffered in their ability to correctly predict the sequence of interest, DNA-1. This likely resulted in poorly interpreted hidden state analysis, supported by Fig. 14. Furthermore, the overall increased Pearson's coefficient (PC) for larger models suggests that while binding prediction is increased with less parameters, the learned features for extracting binding sites are more interpretable in models with more parameters. Difference between DNA-1 and the overall bind activation weights suggests the binding motif is not only position specific but also sub-sequence dependent, otherwise drop-off for later position indices would have been observed in DNA-1 outside the HVD region. All sites had distinguishable overlapping activation peaks. However, there were major extraneous activation peaks at non-binding sites primarily in late downstream regions, which can be explained by the LSTM's implicit higher activation for beginning sequences due to lack of recurrent information. The low PC value in raw activations suggests activation alone is not sufficient enough for high-fidelity binding site suggestions. This is supported by increased PCs from equation one and noise occlusion processing.

5.2.1 Knockout

Majority reversal of binding prediction by knockout test suggests the prediction model relies heavily on literature binding site positions as features for class prediction. Furthermore, the PC dropped dramatically for the overall trend and thresholded peaks, suggesting removal of the binding sites impares the model's ability to find correct binding sub-sequences, as expected. Discrepancies in peaks between DNA-1 and knockout activations were mostly in binding regions. Peaks outside of the literature binding sites, noise, were reduced in the occlusion processing step. This intermediate noise is likely caused by high variation amongst training sequences. That is, the model is looking in those positions for learned features it has expected from other sequences during training (i.e., feature of proteins in general or of antibody class, etc.). Differences between DNA-1 and knockout suggest these areas are less likely to be true binding site predictions and their removal generally increases PC in lower operator-set thresholds.

5.2.2 Insertion

These assertions are further supported by the weak reversal shown by the insertion of binding sites into non-binding sequences. Insertion test activation trends were most similar to DNA-1 trends and low reversal rates were observed, the model is relying, in part, on other areas of interest due to its training on sequences of various lengths and antibody families. Peaks between the first and last groups of reported binding sites are located where all nodes had activation drop off in the raw visualization. These regions are likely the end of the HVD, proposing the model is also looking for features in this HVD region (residue 0-225) for binding prediction previously learned in the training phase. While this effect does confound the binding site prediction, we propose it strengthens the overall prediction mechanism's ability to generalize.

5.2.3 Peak knockout

Knockout of DNA-1 peaks further support this conjecture as the reversal rate was retained and occlusion of these activations from DNA-1 resulted in the highest significant correlations with literature binding sites. Remarkably, the recovery of binding site information corresponding with literature known binding sites from the peak knockout poses this methodology as reliable for binding site suggestions without extensive domain knowledge. Making it a unique and helpful technique in synthetic design.

5.2.4 Binding sites

Sub-sequence recovery, while somewhat significant without overlap between original DNA-1 peak and peak knockout occluded DNA-1 activations, suggested sites unrelated to binding which could delay research and development. Therefore, the final overlap process shows a highly significant method of computationally predicted residue binding sites and sub-sequences with limited domain knowledge, limited data infrastructure, and low computing resource requirements. Operators can leverage precision and recall in the binding site suggestion methodology by altering the threshold for peak identification and smoothing operator during convolutions throughout the procedure according to specific use-case needs.

5.2.5 External validation

Traditionally, external validation is performed on datasets similar to the one used in this work to verify if the model can generalize out of the training and validation distributions. Unfortunately in the case of anti-DNA antibodies, such biological data is extremely limited. This work presents a collection of such limited data where the primary sequence was available. Thus, there was no further external data to validate one. Future works in molecular biology could produce sequencing data to further validate this model's predictive analysis. Other works could use the same methodology proposed here on larger existing datasets where external assay data is readily available.

6 Conclusion

The current work establishes that the limited (in the number of parameters) deep learning models applied to primary sequences can predict whether a novel sequence will bind to DNA and that the hidden activations of these models yielded significant agreement with the binding site reported in previous studies, $r(798) = 0.3674$, $p = 3.1232e\text{-}14$. These recovered areas allow researchers to closely examine the network's internal state, gaining insight into position-specific residues involved in antibody:DNA-binding. We also show that while CNN is better suited for binding prediction in smaller models, larger LSTM hidden states allow for a more accurate binding site interpretation. The proposed methodology can be extended to other domains of interest that may have limited datasets available. Future work should focus on reducing noise in the hidden state activations and compiling residue investigations/predictions in a comprehensive manner to inform binding site prediction with end-use researchers in mind. Other approaches combining CNN and LSTM may be of use, but configuring the CNN to pass position-specific information to the LSTM hidden states is unlikely using this methodology. A new mathematical analysis of the later LSTM activations would be needed. Findings implicate suggestions for RVD and possible synthetic components. Collective implications of this research will further the rapidly developing field of applied deep learning, which in turn will allow for more efficient applications and directly enhance protein data processing. Additionally, we expect the proposed model to be versatile at evaluating other proteomic datasets and user friendly for researchers without extensive computational background knowledge and computing resources. At the same time, the prospective sequence specificities allow experts in wet-lab approaches, like X-ray crystallography, to make more informed decisions.

Acknowledgements The authors thank PhD candidate Paul Morris at the Center for Complex Systems and Brain Sciences for their insightful discussions on natural language processing models and data analysis.

Funding Research was supported by the Graduate Neuroscientist Training Program and Center Complex Systems and Brain Sciences at Florida Atlantic University.

Declarations

Conflict of interest The authors declare no competing interests.

References

1. Alipanahi B, Delong A, Weirauch MT, Frey BJ (2015) Predicting the sequence specificities of DNA-and RNA-binding proteins by deep learning. Nature Biotechnology 33(8):831–838
2. Aotsuka S (1988) A kit for the simultaneous estimation of IgG-class antibodies to double-stranded and single-stranded DNA for clinical purposes. The Ryumachi 28:96–101
3. Beckingham JA, Cleary J, Bobeck M, Glick GD (2003) Kinetic analysis of sequence-specific recognition of ssDNA by an autoantibody. Biochemistry 42(14):4118–4126
4. Berikov V (2020) Autoencoder-based low-rank spectral ensemble clustering of biological data. In: 2020 Cognitive sciences, genomics and bioinformatics (CSGB). IEEE, pp 43–46
5. Berman H, Westbrook J, Feng Z, Gilliland G, Bhat T, Weissig H (2000) I. 443 n. Shindyalov, and PE Bourne, 235–242
6. Chung J, Gülçehre Ç, Cho K, Bengio Y (2014) Empirical evaluation of gated recurrent neural networks on sequence modeling. arxiv:1412.3555
7. Consortium M, Consortium (2019) Uniprot: a worldwide hub of protein knowledge. Nucleic Acids Research 47(D1):D506–D515
8. Gu D, Zhou Y, Kallhoff V, Baban B, Tanner JJ, Becker DF (2004) Identification and characterization of the DNA-binding domain of the multifunctional PutA flavoenzyme. Journal of Biological Chemistry 279(30):31171–31176
9. Herron JN, He X, Ballard D, Blier P, Pace P, Bothwell A, Voss E Jr, Edmundson A (1991) An autoantibody to single-stranded DNA: comparison of the three-dimensional structures of the unliganded Fab and a deoxynucleotide–Fab complex. Proteins: Structure, Function, and Bioinformatics 11(3):159–175
10. Hochreiter S, Schmidhuber J (1997) LSTM can solve hard long time lag problems. In: Advances in neural information processing systems, pp 473–479
11. Hou T, Chen K, McLaughlin WA, Lu B, Wang W (2006) Computational analysis and prediction of the binding motif and protein interacting partners of the Abl SH3 domain. PLoS Comput Biol 2(1):e1
12. Kaufmann J, Asalone K, Corizzo R, Saldanha C, Bracht J, Japkowicz N (2020) One-class ensembles for rare genomic sequences identification. In: International conference on discovery science. Springer, pp 340–354
13. Kong Y, Yu T (2020) forgeNet: a graph deep neural network model using tree-based ensemble classifiers for feature graph construction. Bioinformatics 36(11):3507–3515
14. Kozyr A (1996) A novel method for purification of catalytic antibodies toward DNA from sera of patients with lymphoproliferative diseases. IUBMB Life 39(2):403–413
15. Liu J, Gong X (2019) Attention mechanism enhanced LSTM with residual architecture and its application for protein-protein interaction residue pairs prediction. BMC Bioinformatics 20(1):609
16. Liu X (2017) Deep recurrent neural network for protein function prediction from sequence. arXiv:1701.08318
17. Min S, Lee B, Yoon S (2017) Deep learning in bioinformatics. Briefings in Bioinformatics 18(5):851–869
18. Mooney C, Pollastri G, Shields DC, Haslam NJ (2012) Prediction of short linear protein binding regions. Journal of Molecular Biology 415(1):193–204
19. Nielsen M, Lundegaard C, Lund O (2007) Prediction of MHC class II binding affinity using SMM-align, a novel stabilization matrix alignment method. BMC Bioinformatics 8(1):238
20. Ou Z, Bottoms CA, Henzl MT, Tanner JJ (2007) Impact of DNA hairpin folding energetics on antibody-ssDNA association. Journal of Molecular Biology 374(4):1029–1040
21. Pan X, Rijnbeek P, Yan J, Shen H-B (2018) Prediction of RNA-protein sequence and structure binding preferences using deep convolutional and recurrent neural networks. BMC Genomics 19(1):511
22. Pan X, Shen H-B (2017) RNA-protein binding motifs mining with a new hybrid deep learning based cross-domain knowledge integration approach. BMC Bioinformatics 18(1):136
23. Paul M, Rachel SC, William EH, Elan B (2020) Predicting binding from screening assays with transformer network embeddings. Journal of Chemical Information and Modeling
24. Pavlovic M (2009) The role of anti-DNA antibodies in systemic lupus erythematosus (SLE): ranges and perspectives. Rheumatic Disease Clinics of North America
25. Pavlovic M, Chen R, Kats AM, Cavallo MF, Saccocio S, Keating P, Hartmann JX (2007) Highly specific novel method for isolation and purification of lupus anti-DNA antibody via oligo-(dT) magnetic beads. Annals of the New York Academy of Sciences 1108(1):203–217
26. Pavlovic M, Kats A, Cavallo M, Shoenfeld Y (2010) Clinical and molecular evidence for association of SLE with parvovirus B19. Lupus 19:7
27. Pietrokovski S, Henikoff S (1997) A helix-turn-helix DNA-binding motif predicted for transposases of DNA transposons. Molecular and General Genetics MGG 254(6):689–695
28. Qu Y-H, Yu H, Gong X-J, Xu J-H, Lee H-S (2017) On the prediction of DNA-binding proteins only from primary sequences: a deep learning approach. Plos One 12(12):1–18
29. Radford A, Narasimhan K, Salimans T, Sutskever I (2018) Improving language understanding by generative pre-training
30. Rives A, Goyal S, Meier J, Guo D, Ott M, Zitnick CL, Ma J, Fergus R (2019) Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. aRxiv:622803
31. Rodkey L, Gololobov G, Rumbley C, Rumbley J, Schourov D, Makarevich O, Gabibov A, Voss E (2000) DNA hydrolysis by monoclonal autoantibody BV 04-01. Applied Biochemistry and Biotechnology 83(1–3):95–105
32. Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, Huang Z, Karpathy A, Khosla A, Bernstein M et al (2015) ImageNet large scale visual recognition challenge. International journal of Computer Vision 115(3):211–252
33. Senior AW, Evans R, Jumper J, Kirkpatrick J, Sifre L, Green T, Qin C, Žídek A, Nelson AW, Bridgland A et al (2020) Improved protein structure prediction using potentials from deep learning. Nature 577(7792):706–710
34. Spatz L, Iliev A, Saenko V, Jones L, Irigoyen M, Manheimer-Lory A, Gaynor B, Putterman C, Bynoe M, Kowal C et al (1997) Studies on the structure, regulation, and pathogenic potential of anti-dsDNA antibodies. Methods 11(1):70–78
35. Sun T, Zhou B, Lai L, Pei J (2017) Sequence-based prediction of protein protein interaction using a deep-learning algorithm. BMC Bioinformatics 18(1):1–8
36. Swanson PC, Ackroyd C, Glick GD (1996) Ligand recognition by anti-DNA autoantibodies. affinity, specificity, and mode of binding. Biochemistry 35(5):1624–1633

37. Tanner JJ, Komissarov AA, Deutscher SL (2001) Crystal structure of an antigen-binding fragment bound to single-stranded DNA. Journal of molecular biology 314(4):807–822
38. Teodorescu M (2002) Clinical value of anti-ssDNA (denatured DNA) autoantibody test: beauty is in the eyes of the beholder. Clinical and Applied Immunology Reviews 2(2):115–128
39. Tonkovic P, Kalajdziski S, Zdravevski E, Lameski P, Corizzo R, Pires IM, Garcia NM, Loncar-Turukalo T, Trajkovik V (2020) Literature on applied machine learning in metagenomic classification: a scoping review. Biology 9(12):453
40. Trabelsi A, Chaabane M, Ben-Hur A (2019) Comprehensive evaluation of deep learning architectures for prediction of DNA/RNA sequence binding specificities. Bioinformatics 35(14):i269–i277
41. Wang S, Guo Y, Wang Y, Sun H, Huang J (2019) SMILES-BERT: large scale unsupervised pre-training for molecular property prediction. In: Proceedings of the 10th ACM international conference on bioinformatics, computational biology and health informatics, pp 429–436
42. Yoon S-H, Ha S-M, Kwon S, Lim J, Kim Y, Seo H, Chun J (2017) Introducing EzBioCloud: a taxonomically united database of 16S rRNA gene sequences and whole-genome assemblies. International Journal of Systematic and Evolutionary Microbiology 67(5):1613
43. Zhang P, Meng J, Luan Y, Liu C (2020) Plant miRNA-lncRNA interaction prediction with the ensemble of CNN and IndRNN. Interdisciplinary Sciences: Computational Life Sciences 12(1):82–89
44. Zhao Z, Gong X (2017) Protein-protein interaction interface residue pair prediction based on deep learning architecture. IEEE/ACM transactions on computational biology and bioinformatics

Rachel St. Clair PhD candidate, Center for Complex Systems and Brain Sciences & Graduate Neuroscience Training Program. Rachel has been studying artificial intelligence research since 2017. Her work has focused on computer vision, natural language processing (NLP), reinforcement learning, complex systems, and agent-based modelling. She has publications on NLP, taught courses on generative adversarial learning, and is proficient in pytorch, python and Unity3D. Her current research is focused on hierarchical implementations of brain theory in simulated computational environments for higher cognition and learned intelligence. Her main goal is to work on architectures that produce generally intelligent agents.

Michael Teti is a PhD. student at Florida Atlantic University in the Center for Complex Systems and Brain Sciences. He has been studying and working with deep neural networks and neuro-inspired learning algorithms since 2014, most recently at Los Alamos National Lab in the A-4 group while pursuing his doctoral degree. His dissertation research involves using state-of-the-art deep learning models to help understand neuronal networks and mechanisms, while conversely using insights from human perception to help improve deep learning models.

Mirjana Pavlovic Instructor/Research Professor at CEECS Department, FAU. Dr Pavlovic is involved in Bioengineering/Biomedical Engineering Research and Development at FAU for more than ten years with fundamental courses in Introduction to Bioengineering, Stem Cell Engineering, Tissue Engineering and Innovations and Applications in Biomedical Engineering, as the part of Graduate MS program at CEECS Department at FAU. Her research encompasses integration of Biochemistry, Cell Biology, Immunology, and Stem Cell research with over 130 publications, 5 books and numerous chapters published mostly by Springer. She published papers on Rational Vaccine Design (RVD)for Ebola virus and is working now on computational model of Covid-19 RVD. She is the author and co-author of a number of papers on anti- DNA autoantibodies, combining wet and dry lab approaches in order to integrate s of each of them.

William Hahn graduated from Guilford College in 2008 with a B.S. degree in Mathematics/Physics and research focus in neural networks and swarm optimization. After Guilford, William studied artificial intelligence and immune systems at the University of North Carolina at Greensboro before joining the Center for Complex Systems and Brain Sciences at Florida Atlantic University in 2011. William's research focuses on computational modeling, signal processing, and Deep Learning AI. In 2016, he received a Ph.D. for his work in Sparse Coding and Compressed Sensing, and is now the co-director of the Machine Perception and Cognitive Robotics Laboratory and Assistant Professor of Mathematics at Florida Atlantic University.

Elan Barenholtz is Associate professor in FAU'a Department of Psychology and Complex Systems and Brain Sciences and co-directs the Machine Perception and Cognitive Robotics Lab, an interdisciplinary artificial intelligence research facility. He is also Associate Director of FAU's Center for the Future of Mind. His research is concerned with deep learning applications and architectures with an emphasis on modeling biological processes and intelligence. He has published 50+ articles in scientific journals and conference proceedings and his research has received funding from federal, state and private agencies.