*Protein Engineering, Design and Selection,* 2019, vol. 32 no. 7, pp. 347–354
doi: 10.1093/protein/gzz031
Advance Access Publication Date: 29 August 2019
Original Article

Original Article

# Quantifying the nativeness of antibody sequences using long short-term memory networks

**Andrew M. Wollacott[1,†], Chonghua Xue[2,†], Qiuyuan Qin[2], June Hua[2], Tanggis Bohnuud[1], Karthik Viswanathan[1], and Vijaya B. Kolachalama[2,3,4,5,*]**

[1]Visterra Inc., Waltham, MA 02451, USA [2]Section of Computational Biomedicine, Department of Medicine, Boston University School of Medicine, Boston, MA 02118, USA [3]Hariri Institute of Computing and Computational Science & Engineering, Boston University, Boston, MA 02115, USA [4]Whitaker Cardiovascular Institute, Boston University School of Medicine, Boston, MA 02118, USA [5]Boston University Alzheimer's Disease Center, Boston, MA 02118, USA

*To whom correspondence should be addressed: E-mail: vkola@bu.edu
†These authors contributed equally to this work.

Edited by: Dr. Valerie Daggett

Received 11 June 2019; Revised 11 June 2019; Editorial Decision 7 July 2019; Accepted 7 July 2019

## Abstract

Antibodies often undergo substantial engineering en route to the generation of a therapeutic candidate with good developability properties. Characterization of antibody libraries has shown that retaining native-like sequence improves the overall quality of the library. Motivated by recent advances in deep learning, we developed a bi-directional long short-term memory (LSTM) network model to make use of the large amount of available antibody sequence information, and use this model to quantify the nativeness of antibody sequences. The model scores sequences for their similarity to naturally occurring antibodies, which can be used as a consideration during design and engineering of libraries. We demonstrate the performance of this approach by training a model on human antibody sequences and show that our method outperforms other approaches at distinguishing human antibodies from those of other species. We show the applicability of this method for the evaluation of synthesized antibody libraries and humanization of mouse antibodies.

**Key words:** antibody engineering, antibody humanization, long short-term memory network, machine learning

## Introduction

Antibodies are a preferred treatment modality, particularly in cancer and autoimmune diseases, with more than 50 approved antibodies and more than 500 molecules in various stages of clinical development (Kaplon and Reichert, 2019). Therapeutic antibodies are derived from a variety of approaches, with two of the major sources being natural repertoires (either naïve or immune) (Hust *et al.*, 2012), and synthetic designed libraries (Adams and Sidhu, 2014). Antibodies derived from these sources often undergo further engineering to improve affinity, specificity, and developability profiles. It has been shown that design schemes that more closely resemble the sequence profile and features of natural antibodies lead to better synthetic libraries, with improved rates of expression (Zhai *et al.*, 2011) and stability (Prassler *et al.*, 2011). This concept of antibody 'nativeness' is also applied during humanization, where the similarity to human antibodies, or 'humanness' is a major driver when engineering to improve the safety profile and reduce immunogenicity concerns of sequences derived from non-human sources (Safdari *et al.*, 2013).

With the increased use of antibody engineering, there is a need for improved methods to rapidly and accurately estimate the nativeness of these sequences. A common approach to determine the humanness of an antibody is to assess its proximity to the closest human germline sequence. Indeed, the World Health Organization (WHO) previously categorized monoclonal antibodies based on the percentage of human content in the variable region, requiring ≥85% human germline identity for designation of humanized antibodies (−zumab) (Jones *et al.*, 2016). While straightforward, this approach has several limitations, one of which is that it considers all mutations relative to a germline as equal. However, analyses of natural antibody sequences have shown that somatic hypermutations are not equally distributed (Burkovitz *et al.*, 2014). Alternative metrics, such as the Human String Content (Lazar *et al.*, 2007) and T20 score (Gao *et al.*, 2013), consider similarities to larger sets of reference sequences such as all available human germline sequences or to curated sets of known human antibody sequences. More recent methods, such as the MG score (Clavero-Álvarez *et al.*, 2018), also consider covariation between pairs of amino acids at different positions, better accounting for the context of a particular residue within the sequence.

With the large amounts of antibody sequence data from next generation sequencing (NGS) data of B-cell receptor repertoires that have become available in the last few years, it is possible to analyze antibody sequences and the sequence space they explore in far greater detail (Rouet *et al.*, 2018). This wealth of data enables determination of not just position-specific amino acid propensities, but also coupling between different positions in the sequence. Computationally designed libraries encoded with this additional coupling information can be synthesized given the recent advances in DNA library synthesis, such as the use of oligo-pools, where each variant in the library is a custom and specific design (Chevalier *et al.*, 2017; Rocklin *et al.*, 2017).

Encouraged by recent advances in machine learning, we sought to develop a model that is capable of learning a representation of natural antibodies that captures higher order relationships between positions to provide a more sensitive measure of antibody nativeness. Recurrent neural networks (RNNs) have demonstrated great success for natural language understanding and have previously been applied to biological sequence analysis to predict protein function. Here, we developed a bi-directional long short-term memory (LSTM) network model (Hochreiter and Schmidhuber, 1997), a specialized form of an RNN framework, capable of learning the distribution of antibody sequence data by selectively remembering patterns for long duration of time. We demonstrate the performance of this approach by training a model on human antibody sequences and show that our method outperforms other approaches at sequence classification by distinguishing human antibodies from those of other species. We also show that this method can be applied to evaluate subtle differences in designed libraries. Further, we demonstrate how this method can be applied to antibody engineering, such as humanization, by identifying human frameworks that are predicted to be the most favorable for CDR grafting for a panel of mouse antibody sequences. Lastly, we use the model as an evaluation of antibody humanness and show that it outperforms several other methods when applied to humanization classification of available therapeutic antibody sequences.

## Materials and Methods

### Problem formulation

Given a sequence $\left[x^{(1)}, x^{(2)}, \cdots, x^{(T)}\right] \in X^T$, where $X$ denotes the set of all distinct amino acids, the learning task is to estimate the probability $p_t\left(x^{(t)} | x^{(1)}, \cdots, x^{(t-1)}, x^{(t+1)}, \cdots, x^{(T)}\right)$ for $t \in \{1, 2, \cdots, T\}$. The underlying assumption here is that if a sequence is drawn from a target antibody repertoire, a well-trained model would be able to predict any single residue by learning appropriate information from its neighbors with high confidence. For each antibody sequence, we can compute an overall score (which we also refer to as an LSTM score) defined as the averaged sum of negative logarithms (NLS) of all conditional probabilities defined as

$$\mathrm{NLS} = -\frac{1}{T}\sum_{t=1}^{T} \log p_t \ \left(x^{(t)} | x^{(1)}, \cdots, x^{(t-1)}, x^{(t+1)}, \cdots, x^{(T)}\right). \tag{1}$$

A single NLS score can be computed for each antibody sequence, where lower scores indicate a higher degree of nativeness. Here, $p_t$ is defined by the Boltzmann distribution $e^{-\epsilon^{(t)}} / \sum_{x' \in X} e^{-\epsilon^{(t)}_{x'}}$ without temperature, and the energy term $\epsilon^{(t)}$ is a learned parameter.

### Long short-term memory network model

RNNs have been used previously for capturing complex patterns in biological sequences. The LSTM framework was introduced recently to overcome the issues related to traditional RNN frameworks such as vanishing gradients and long-term dependencies (Hochreiter and Schmidhuber, 1997). As a specific sub-class of RNN, the LSTM model still takes the traditional recurrent form of $h^{(t)} = f(x^{(t)}, h^{(t-1)})$, where $f$ denotes the recurrent cell function, $h^{(t)}$ denotes the hidden state (or output) at time $t$, $h^{(t-1)}$ denotes hidden state at the previous time ($t-1$), and $x^{(t)}$ is the input at time ($t$). The structure inside an LSTM cell is defined as follows:

$$i^{(t)} = \sigma\left(W_i \cdot \left[h^{(t-1)}, x^{(t)}\right] + b_i\right), \tag{2}$$

$$f^{(t)} = \sigma\left(W_f \cdot \left[h^{(t-1)}, x^{(t)}\right] + b_f\right), \tag{3}$$

$$\overset{\sim}{c}^{(t)} = \tanh\left(W_{\overset{\sim}{c}} \cdot \left[h^{(t-1)}, x^{(t)}\right] + b_{\overset{\sim}{c}}\right), \tag{4}$$

$$o^{(t)} = \sigma\left(W_o \cdot \left[h^{(t-1)}, x^{(t)}\right] + b_o\right), \tag{5}$$

$$\begin{aligned} c^{(t)} &= f^{(t)} \odot c^{(t-1)} + i^{(t)} \odot \tilde{c}^{(t)}, \text{and} \\ h^{(t)} &= o^{(t)} \odot \tanh\left(c^{(t)}\right). \end{aligned} \tag{6}$$

Here, $\odot$ denotes the Hadamard product operator, $[\cdot\ ,\ \cdot]$ is the vector concatenation and $\sigma(\cdot)$ is the sigmoid function. The LSTM introduces four gates $i^{(t)}, f^{(t)}, \overset{\sim}{c}^{(t)}$, and $o^{(t)}$, which denote input gate, forget gate, modulation gate, and output gate, respectively. For each gate, an affine transformation is applied on $[h^{(t-1)}, x^{(t)}]$ along with an activation function (i.e. $\sigma$ or tanh). By combining them, the LSTM cell is capable of deciding the appropriate set of information that needs to be passed or suppressed in favor of a given task. We used this framework to learn the nativeness of a given antibody sequence (Fig. 1).

The LSTM model can take an antibody sequence as its input by reading a single amino acid residue at a time and learning its context as it contributes to the nativeness of the sequence. As our model is bi-directional, the antibody sequence was scanned from the N-terminus to the C-terminus once and then scanned again in the opposite direction. Each residue in the antibody sequence was converted into a one-hot encoded vector (size: 21 × 1), where each element of the vector is assigned a value of 0, except at the position of the amino acid, where it was assigned a value of 1. Before passing the antibody sequence to the LSTM model, we first embedded all the

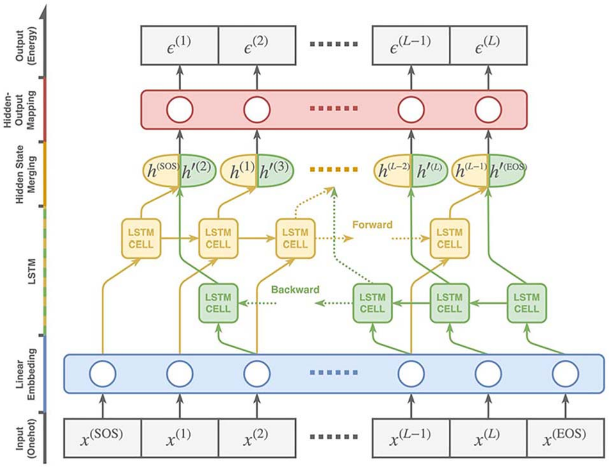

**Fig. 1** Schematic of the LSTM framework. The LSTM model is comprised of a linear embedding layer followed by a bi-directional LSTM layer along with a fully connected layer and an output layer. The bi-directional LSTM layer was padded on both sides.

amino acids into a vector. This step was implemented by introducing a 21 × $M$ matrix, where $M = 32$ is the embedding dimension.

### Preparation of antibody sequence datasets

NGS sequences from different antibody repertoires were prepared for training and testing the LSTM models. Human heavy and light chain sequences were obtained from the pre-processed antibody repertoire NGS datasets in the Observed Antibody Space database (Kovaltsuk *et al.*, 2018). To reduce sequence errors, only reads observed at least 4 times were selected. These sequences were further clustered at the 97% identity level to avoid sampling highly related sequences between the training and testing sets. Mouse heavy chain antibody sequences were processed based on a high-quality mouse antibody repertoire dataset (Greiff *et al.*, 2017). Chicken antibody heavy chain sequences were processed from a naïve B-cell library, and a set of llama sequences were obtained from two enriched llama immune libraries of VHH antibodies. Due to the smaller library size for the llama set, reads were kept if they were observed at least twice, while still clustering at the 97% level. For human and mouse sequences, 25000 sequences of each dataset were randomly selected for training, 10 000 for validation, and 10 000 for testing. For chicken and llama 10 000 and 4000 sequences, respectively were selected for testing. All mouse, chicken, and llama datasets were obtained by Illumina MiSeq paired-end NGS sequencing and processed using the Repertoire Sequencing Toolkit (pRESTO) (Vander Heiden *et al.*, 2014). ANARCI (Dunbar and Deane, 2015) was used to assign germline annotations and sequence clustering was performed using CD-HIT (Fu *et al.*, 2012). To facilitate comparison with other models, antibody sequences were aligned using the AHo antibody numbering scheme (Honegger and Plückthun, 2001), as implemented by ANARCI.

### Performance metrics

The LSTM score (Eq. (1)) was used to evaluate the nativeness of each antibody sequence. We generated receiver operating characteristic (ROC) curves and computed area under curve (AUC) for each model in order to assess performance. Additionally, we computed the maximum value of the Youden's J-statistic (YJS) and Matthews correlation coefficient (MCC) on each set of model predictions. YJS is defined as:

$$\mathrm{YJS} = \mathrm{sensitivity} + \mathrm{specificity} - 1, \tag{7}$$

for all points of the ROC curve. The maximum value of the YJS can be used as a criterion for selecting the optimum cut-off point, when a diagnostic test gives a numeric result rather than a dichotomous result. MCC is a balanced measure of quality for dataset classes of different sizes of a binary classifier, defined as follows:

$$\mathrm{MCC} = \frac{[(\mathrm{TP} \times \mathrm{TN}) - (\mathrm{FP} \times \mathrm{FN})]}{[(\mathrm{TP} + \mathrm{FP})\,(\mathrm{TP} + \mathrm{FN})\,(\mathrm{TN} + \mathrm{FP})\,(\mathrm{TN} + \mathrm{FN})]^{0.5}} \tag{8}$$

Here, TP denotes true positive values, FP and FN denote false positive and false negative cases, respectively, and TN denotes true negative values.

## Results

### Training the LSTM model

LSTM models were generated by training on a dataset of human antibody VH sequences. Model training took approximately 12 hours for 50 epochs on a 6-core Intel Xeon CPU, with the same task taking about 20 minutes on a GPU (NVIDIA GTX 1060). Query sequences scored rapidly, taking $< 2$ minutes to process 10 000 sequences.

We assessed the number of sequences required for generation of a robust model. Models were generated using training sets of increasing size, and we observed that the performance of the model (as determined using MCC, AUC, and YJS), plateaued for training sets larger than 20 000 sequences (Fig. S1). Based on this observation, the LSTM model was trained on a dataset of 25000 human antibody VH sequences, using a validation set of 10 000 sequences during training to ensure the model was not overfit.

### Classification of human antibody sequences

The performance of the LSTM model was assessed by determining its ability to correctly distinguish natural human antibody sequences from those originating from other species. Test datasets of antibody VH sequences from mouse, llama, chicken, and human were prepared, and LSTM scores were calculated for each sequence. An analysis of the distribution of LSTM scores demonstrated a near complete separation between human sequences and chicken, but with some overlap in scores between mouse and human sequences, and a more substantial overlap between llama sequences and human (Fig. 2a). An ROC analysis showed that in classifying sequences as human-like, the model had an AUC of 0.999 for chicken sequences, 0.995 for mouse sequences, and 0.976 for llama sequences (Fig. 2b).

Given that human antibody sequences in the database, and therefore in our training set, evolved from a limited set of germline genes, one might expect the LSTM model to favor sequences that are more germline-like since the germline sequences should be near central to sequence clusters. As expected, we found the LSTM score to be correlated to the sequence identity of the closest human germline v-gene (Fig. 3), and this correlation appears more pronounced for llama and mouse sequences, and less for human and chicken antibody sequences. Some of the llama and mouse sequences in the test dataset have LSTM scores that compare very favorably with human sequences and would be considered indistinguishable from human sequences by the LSTM model. These low-scoring non-human antibody sequences have higher sequence similarity to human

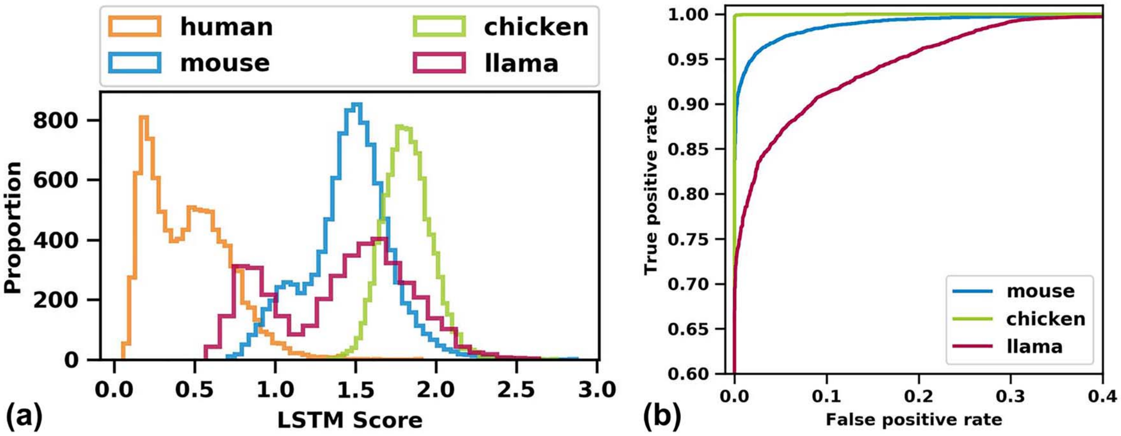

**Fig. 2** Performance of the LSTM model trained on human sequences. **(a)** LSTM scores were calculated based on a model trained on human antibody sequences, and the distribution of scores shown for test datasets of human, mouse, chicken, and llama antibody sequences. **(b)** ROC plot showing the performance of the LSTM model in distinguish human antibody sequences from antibodies from mice, chicken, or llamas.

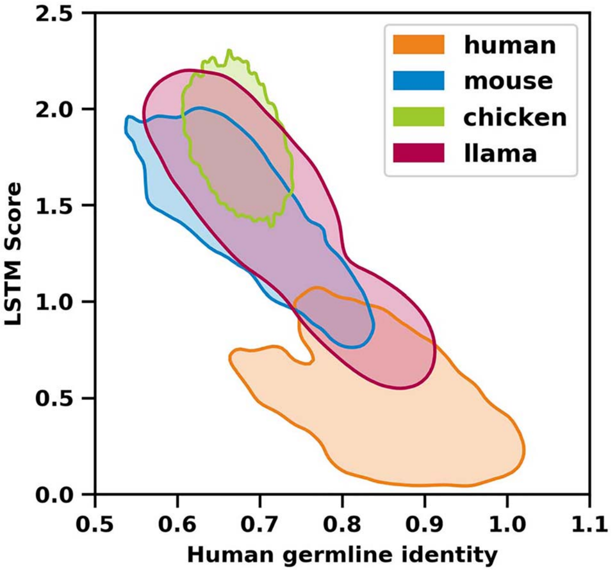

**Fig. 3** Relationship between LSTM score and the identity to human germline sequences shown, for simplicity, as a Kernel Density Estimate (KDE) plot. LSTM scores were calculated for antibody sequences in test datasets for human, mouse, chicken, and llama, and related to the sequence identity to the closest human germline sequence.

germline genes than many native human antibodies (Fig. 3). A closer inspection of these low scoring non-human antibody sequences confirms the striking similarity to human VH sequences (Fig. S2), with most of the differences concentrated in the CDR region. This level of similarity to a human VH germline gene is comparable to that seen for human antibody sequences leading to low LSTM scores for non-human antibody sequences.

Since the LSTM score was found to correlate with germline identity, we investigated whether the LSTM model favored some human germlines over others. The training dataset was randomly selected from a large dataset of human sequences, leading to unequal representation of germlines. Interestingly, there is a general trend where sequences derived from germlines that have lower representation have higher (worse) LSTM scores (Fig. S3), which can be attributed to their relative scarcity in the training set. It should be noted that in this particular application, we intended to train the model on a dataset that was representative of the distribution of antibody sequences found in a native repertoire. However, depending on the application needs, the training set can be modified to represent each germline equally.

We also investigated the long tail in the score distributions for human sequences to determine why some human sequences had very poor LSTM scores. For this analysis, we considered sequences that had an LSTM score $>$ 1.0 as having a poor score, which represents $<$2.5% of sequences in the test set, and explored sequence features, which may be contributing to the poor scores. We found the deviation from human germline to be a strong indicator for sequences having a high score; more than 20% of sequences with human germline identity $<$ 70% had high LSTM scores (Fig. S4a). Sequences that had lengths of HCDR1 and HCDR2 that were rarely observed in the training set also tended to have higher LSTM scores (Fig. S4b), again demonstrating that the model is able to identify sequences with rare attributes as being outliers. Sequences with long HCDR3 portions ($>$24) also tended to be more likely to have a high score (Fig. S4c). This indicates that the LSTM model is learning many of the attributes of human antibodies that make up the bulk of sequences in the training set, and human antibodies that have unfavorable scores tend to have one or more features that are uncommon in the training set.

## Assessing an LSTM model trained on mouse sequences

When assessing the performance of the LSTM model that was trained on human sequences, the model performed very well at distinguishing mouse-derived antibodies from human sequences, but some mouse sequences scored better than actual human antibody sequences. One question that arises is whether this observation is due to a limitation in the LSTM model or due to the breadth of the sequence space covered by human sequences. Mouse sequences are known to have fewer somatic hypermutations relative to their germline genes compared to human sequences, essentially forming tighter sequence clusters. This can be seen in Fig. S5, where over 83.4% of mouse sequences have $>$ 90% sequence identity to the closest germline, whereas only 32.4% of human sequences show $>$ 90% sequence identity.

**Table I.** Comparison of AUC values computed on four different models

| Area under curve | LSTM | MG | Germ | Profile |
|---|---|---|---|---|
| Mouse | 0.9947 | 0.9847 | 0.9646 | 0.8572 |
| Chicken | 0.9998 | 0.9998 | 0.9836 | 0.9842 |
| Llama | 0.9759 | 0.9471 | 0.8887 | 0.7977 |

We trained an LSTM model using a training set of 25000 mouse sequences and assessed its performance in distinguishing a test set of mouse sequences from a test set of human, llama, and chicken antibody sequences. Interestingly, the LSTM model trained on mouse sequences could almost completely distinguish mouse sequences from antibodies from the other species (Fig. S6), showing an AUC of 0.9991, 0.9997, and 0.9999 for human, llama, and chicken sequences, respectively, outperforming the LSTM model that was trained on human antibody sequences. This indicates that the performance of the LSTM model is related to the underlying sequence space used in the training set. We hypothesize that when training on human sequences, which are more sequence-diverse relative to their cluster centers (i.e. germline sequences) compared to mouse, the LSTM model must learn to allow for greater diversity and so antibodies from other species are captured in this search space. In contrast, when trained on mouse sequences, the LSTM model learns that sequences show little deviation from their cluster centers, and so the search space learned by the LSTM model is more restrictive and thus more specific to mouse antibodies.

## Comparison of the LSTM model to other models for prediction of humanness

The performance of the LSTM model trained on human sequences was compared to that of several competing models in its ability to correctly distinguish human antibody sequences from those from mouse, chicken, and llama. The models evaluated included the sequence identity to the closest human V-gene, a profile score (using log-odds ratios of the observed frequency of each amino acid at each position), and the MG score (which uses the corresponding negative log probability density obtained from a multivariate Gaussian distribution). Table I shows AUC values for model performance, which demonstrates that the performance of competing models was inferior to the LSTM model. Overall, LSTM performed the best, followed by MG score, sequence identity to germline, and finally profile-scores. All models performed well in distinguishing chicken sequences from human, with the LSTM and MG score models both having near perfect performance.

## Application to evaluation of synthetic libraries

In our initial evaluation of the LSTM model, we demonstrated the ability to outperform other models at distinguishing native human antibody sequences from those of other species. While some non-human sequences show a high degree of similarity to human sequences, in most cases, these differences between species are in the relatively conserved framework regions, making this classification easier. We set out to assess the performance of the LSTM model on a more difficult scenario, where sequence differences were limited to the CDR regions of the antibody, which are known to be much more sequence diverse. The purpose of this test was two-fold: (1) to assess the LSTM model as a means to computationally evaluate designed synthetic libraries for their nativeness, and (2) to more methodically evaluate the performance of the LSTM across a range of CDR hypervariable sequences where the degree of nativeness is more systematically varied.

To evaluate the performance of the LSTM in learning the representation of the hypervariable CDR regions, we computationally generated several sets of designs, which varied the origin and degree of variability in the CDR regions (Fig. 4a). For this test, we generated four sets of designed sequence; for SetG (graft) entire intact CDR fragments from native human sequences were grafted into a human framework sequence derived from a different germline; for SetR (random) CDR sequences were designed by randomly choosing any of the 20 amino acids at each position; for SetPA (profile-all) CDR sequences were designed by randomly choosing an amino acid at each position, weighted by its frequency of occurrence observed in the training set of all human antibodies; SetPG (profile-germline) was similar to SetPA, but the profile used is germline-specific (i.e. derived only from sequences that have the same germline as the host framework). A set of native antibody sequences (SetN) was included to serve as a baseline for comparison. In all cases, mutations at non-CDR positions were reverted back to the amino acid found in the appropriate germline, to ensure sequence differences were limited to CDR regions.

These designed sets enable us to probe several different aspects of the LSTM model. SetG provides sequences that have native CDR sequences in the wrong framework context; SetR provides sequences that are essentially scrambled and thus do not look like natural antibodies; SetPA provides sequences where CDR amino acid profiles match all antibodies in the training set, but where the profiles may not be optimal for a given framework; lastly, SetPG provides a very stringent test of the model since the amino acid profile at each CDR position should be indistinguishable from native antibodies with the same germline. Importantly, for both SetPA and SetPG since residues in a CDR are selected independently, the coupling between residues that is present in native sequences will be lost.

To demonstrate the stringency of these datasets, profile scores were calculated for each of the sets of designed antibody sequences as well as the native set. As expected, the distribution of profile scores for each of the sets is almost identical except for SetR, which contains randomized residues in the CDRs. In contrast, the LSTM model is able to discriminate between the different designed sets including between stringent SetPG and native (Fig. 4b). Unsurprisingly, SetR had the highest (worst) LSTM scores, with a median score of 1.6 compared to a median score of 0.3 for SetN. SetG had higher LSTM scores than native sequences, with a median score of 0.6. SetPG, the most stringent test dataset, also scored higher than native, with a median score of 0.4.

An analysis of the performance on these designs shows that the LSTM has learned couplings between residues and is aware of the sequence context when considering CDR residues. Scores for SetPG, the best scoring designed set, were still higher (worse) than native sequences, despite having amino acids at the same frequency as

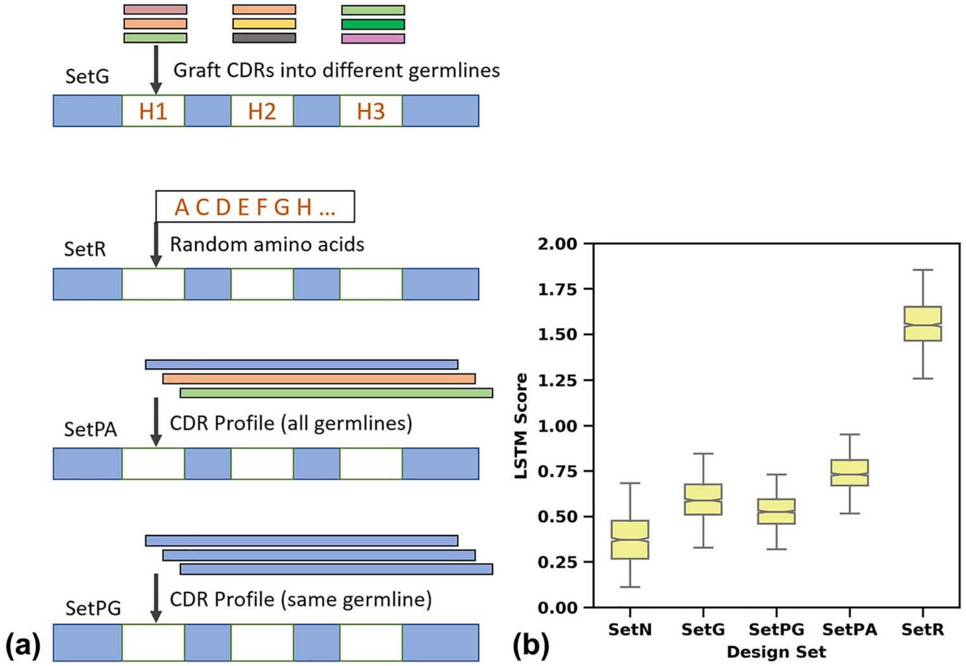

**Fig. 4** Distribution of LSTM scores for several designed synthetic libraries. **(a)** Schematic describing the approach for generating each design set. **(b)** Distribution of LSTM scores for each designed set.

those found in the native sequences. This can be attributed to the lack of couplings between residues in the CDRs; since the LSTM scores native sequences more favorably, this demonstrates that the LSTM has learned to identify relationships between residues in the native sequences, favoring sequences that contain these couplings. Designs from SetG, where native CDRs were grafted into non-matching germlines, scored more poorly than native sequences. Here, even though coupling has been retained within the CDR, couplings between the framework and CDR have been lost. In addition, the frequency of amino acids found in the grafted CDRs can differ from that expected for the receptor framework when originating from different germlines, demonstrating that the LSTM model is able to identify that these CDR sequences are in the wrong context. Scores for SetPA were higher than SetPG, indicating that the LSTM has learned positional preferences for amino acids that are specific to the germline of the antibody.

## Application to humanization

We also assessed the performance of the LSTM model for distinguishing human from humanized and mouse sequences, using a dataset of 46 pharmaceutical antibodies whose sequences have been previously compiled and used for assessment of humanization scoring methods (Clavero-Álvarez *et al*., 2018). We compared the LSTM model to the MG score model, and to a widely used T20 score, which is used for scoring of humanized constructs. For the LSTM and MG score, two models each were trained using a dataset of 25000 VH human sequences or 25000 VK human sequences, and the T20 score was obtained using an online tool (https://dm.lakepharma.com/bioinformatics/). As can be seen in (Fig. 5), the LSTM model outperforms the other models in distinguishing human from humanized and mouse antibodies. The LSTM model for the VH sequences shows a very clear separation between the human and humanized sequences, and also between humanized and mouse antibodies; humanized antibodies score well in general, but do not score as well as the fully human antibodies. In contrast, the MG score model has a large overlap between human and humanized, with more substantial overlap between humanized and mouse antibodies. The T20 model also shows substantial overlap between human and humanized. The same trend is seen for the VK sequences, where the LSTM model outperforms the other two models.

In addition to scoring humanized sequences to assess their humanness, the LSTM model can be used during the humanization procedure to select germline sequences that would serve as better receptors for CDR grafting from non-human sources. As demonstrated with the analysis of designed sequences, the LSTM model is able to assess whether CDR loops are in the right context. Here, we investigated the utility of the LSTM model to select frameworks that would be expected to be most compatible with a given CDR (Fig. 6).

Three mouse antibodies (mab1, mab7, and mab8) were selected for humanization, and all three HCDRs were grafted into 9 of the most commonly used human frameworks for therapeutic applications. Assessment of the LSTM scores of the grafted sequences indicated that each of the mouse antibody CDR loops are more compatible (lower LSTM scores) with specific human germlines. For example, for mab7, the chimeric antibody sequences generated by grafting its CDRs into the human germlines IGHV3–23, IGHV3–30, and IGHV3–48 score much more favorably than when grafted into the other germlines. In fact, grafting the mouse CDRs from mab7 into these three germlines results in a humanized sequence that scores favorably compared to other natural human sequences derived from these germlines. These three germlines, however, are not identified as the most compatible germline for grafting all mouse CDRs—these frameworks appear to be less favorable than others, for example, for grafting CDRs from mab1. This analysis can be extended to humanization of light chains by training on an appropriate set of human light chain sequences. Given the LSTM's demonstrated ability to capture coupling between CDR regions and framework regions, we

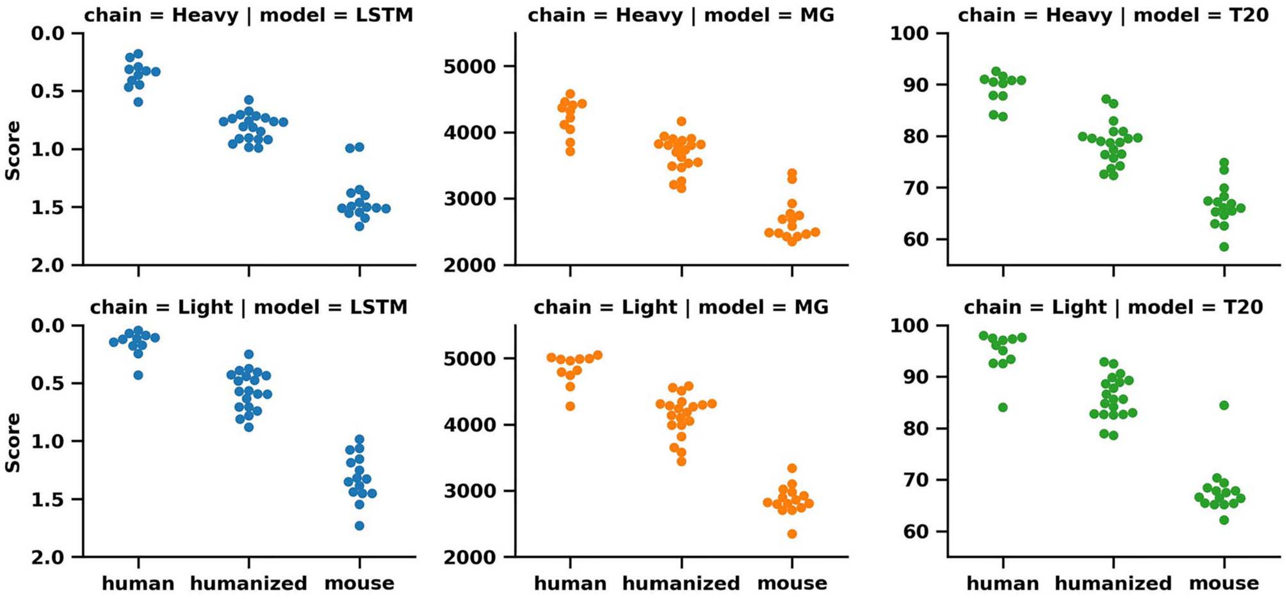

**Fig. 5** Performance of several models at scoring known therapeutic antibody sequences. LSTM scores, MG scores, and T20 scores were evaluated on therapeutic antibodies derived from human, mouse, and those that were humanized.

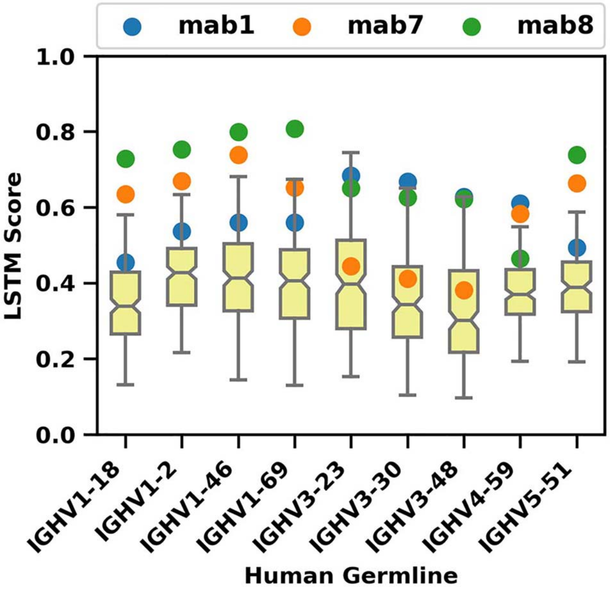

**Fig. 6** LSTM scores of CDR-grafted variants derived from three mouse antibodies. CDR loops from each mouse antibody VH sequence were grafted into each of nine different human germline frameworks, and LSTM scores calculated (color dots). Box plots represent the distribution of LSTM scores in a test dataset of human sequences matching the corresponding human germline framework.

posit that the LSTM model is a sensitive approach for selection of an appropriate template for humanization.

## Discussion

We have developed an LSTM network approach that is capable of learning the nativeness of antibody sequences. The approach makes use of the increasing amount of sequence data available from naturally occurring antibody repertoires, and given the large number of sequences in the training set, is able to learn signature features of native antibodies. When trained only on human antibody sequences, for example, this model outperforms state-of-the-art methods at distinguishing human antibody sequences from antibodies derived from other species. The ability to identify these key features, which may include higher-order coupling between residues, enabled the LSTM model to also outperform other methods at successfully classifying humanized antibodies in a set of therapeutic antibodies.

The LSTM framework is an efficient variant of the traditional RNN. The gates of an LSTM cell facilitate decisions such as allowing the data to enter the cell, leave or be deleted through an iterative process, preserving the error that can be back propagated, and adjusting weights via gradient descent. This characteristic allows for the LSTM model to capture diverse interactions in a computationally efficient fashion along with capturing remote dependencies. In the case of evaluating nativeness of an antibody sequence, even though each amino acid residue is passed sequentially to the LSTM cell, the architecture was able to capture long-range effects (Fig. S7).

The LSTM model that we have developed has several favorable properties that enable it to be applied to antibody engineering. The method is computationally efficient and can assess a large number of sequences rapidly; in our benchmarks, a library of 10000 antibody sequences can be evaluated in a few minutes, making it computationally tractable to apply to synthetic libraries which can contain >1e7 sequences. An additional benefit of using an LSTM approach is that sequences do not have to be aligned, which can be time-consuming and ambiguous when the diversity of the underlying sequences is high (such as in HCDR3). In contrast, other approaches, such as the MG score (Clavero-Álvarez *et al.*, 2018), are dependent on the training and query sets being pre-aligned. Also, the ability for the LSTM framework to process a single amino acid residue at a time enables the analysis of thousands of sequences with variable lengths.

Recent advances in DNA library synthesis, such as the use of oligo-pools, has enabled design of libraries where each variant in the library is a custom and specific design (Chevalier *et al.*, 2017;

Rocklin *et al*., 2017). Information from the LSTM model can then be used to improve these designed libraries by accounting for residue coupling information to make more native-like designs. Even for more traditional library synthesis approaches, an LSTM model can aid design by learning amino acid preferences in CDRs that best match the germline scaffolds being used.

In this application, we have used a curated random sample of antibody sequences in the training set to learn the representation of nativeness of human antibodies. This approach can be extended to learn representations of a variety of sub-populations of antibody sequences that may be enriched in properties of interest. For example, it has been observed that the majority of glycan-binding antibodies are derived from select germlines. The LSTM approach we have developed could be applied to a subset of known glycan-binding antibodies to learn signature features of these antibodies; this model could then provide an assessment on whether query antibody sequences resemble glycan-binding antibodies. Such an approach could be used with other specific populations of antibodies such as DNA-binding or membrane-protein binding antibodies. By using this approach, synthetic libraries could be designed that retain not only native-like sequences but are also enriched in sequence features found in antibodies known to bind a specific class of antigens.

The LSTM model can also be used in concert with computational protein design approaches which, due to continued improvements in the field, have begun to find more routine application in antibody engineering. Several recent approaches have used observed sequence profiles of antibodies to bias toward more native-like designs (Adolf-Bryfogle *et al*., 2018). These implementations use a one-body energy term to favor appropriate amino acids at each position, but typically ignore explicit pairwise and higher order couplings between positions which can impact structural stability. Since the LSTM model we have implemented can rapidly evaluate entire antibody sequences, it can be incorporated into computational design approaches as a whole-body energy term to favor sequences that are more representative of the training set. A similar approach has been successfully applied for protein deimmunization (King *et al*., 2014).

## Conclusion

We have developed an LSTM model that is able to learn the representation of protein sequences. This approach is able to learn higher-order linkages between positions in the sequence and is aware of its surrounding sequence context. We applied this model to the analysis of antibody sequences and showed that this model outperforms other published approaches at assessing antibody humanness. The LSTM model described here can help design synthetic libraries that better represent native-like sequences and can also be used for selection of appropriate scaffolds and mutations to guide humanization of antibodies.

## Supplementary Data

Supplementary data are available at *Protein Engineering, Design and Selection* online.

## Data Availability

Python scripts and data are made available on GitHub (https://github.com/vkola-lab/peds2019).

## Author Contributions

A.M.W., C.X., K.V., and V.B.K. designed the study; A.M.W., C.X., Q.Q., J.H., and T.B. performed the experiments; A.M.W., K.V., and V.B.K. wrote the manuscript with assistance from other authors; A.M.W. and V.B.K. provided overall supervision.

## Funding

This work was supported in part by funding from Visterra Inc. VBK also acknowledges support from the National Center for Advancing Translational Sciences, National Institutes of Health, through BU-CTSI (grant 1UL1TR001430), National Institutes of Health (grants 1R01AR070139 and 1R56AG062109), a Scientist Development Grant (17SDG33670323) from the American Heart Association, and a Digital Health Initiative Research Award from Boston University.

## Acknowledgements

We thank Dr. Pierpaolo Bruscolini at University of Zaragoza, Spain for sharing data. We thank Dr. Gregory Babcock and Dr. Zachary Shriver for helpful discussions and critical reading of the manuscript.

## References

Adams, J.J. and Sidhu, S.S. (2014) *Curr. Opin. Struct. Biol*., **24**, 1–9.

Adolf-Bryfogle, J., Kalyuzhniy, O., Kubitz, M., Weitzner, B.D., Hu, X., Adachi, Y., Schief, W.R., Dunbrack, R.L.Jr. (2018) *PLoS Comput. Biol.* **14**, e1006112.

Burkovitz, A., Sela-Culang, I., Ofran, Y. (2014) *FEBS J*., **281**, 306–319.

Chevalier, A., Silva, D.A., Rocklin, G.J. *et al.* (2017) *Nature*, **550**, 74–79.

Clavero-Álvarez, A., Di Mambro, T., Perez-Gaviro, S., Magnani, M., Bruscolini, P. (2018) *Sci. Rep*., **8**, 14820.

Dunbar, J. and Deane, C.M. (2015) *Bioinformatics*, **32**, 298–300.

Fu, L., Niu, B., Zhu, Z., Wu, S., Li, W. (2012) *Bioinformatics*, **28**, 3150–3152.

Gao, S.H., Huang, K., Tu, H., Adler, A.S. (2013) *BMC Biotechnol*., **13**, 55.

Greiff, V., Weber, C.R., Palme, J., Bodenhofer, U., Miho, E., Menzel, U., Reddy, S.T. (2017) *J. Immunol*., **199**, 2985–2997.

Vander Heiden, J.A., Yaari, G., Uduman, M., Stern, J.N.H., O'Connor, K.C., Hafler, D.A., Vigneault, F., Kleinstein, S.H. (2014) *Bioinformatics*, **30**, 1930–1932.

Hochreiter, S. and Schmidhuber, J. (1997) *Neural Comput*., **9**, 1735–1780.

Honegger, A. and Plückthun, A. (2001) *J. Mol. Biol*., **309**, 657–670.

Hust, M., Frenzel, A., Meyer, T., Schirrmann, T., Dübel, S. (2012) *Antibody Engineering: Methods and Protocols*, 2nd edn. Humana Press, Totowa, NJ, pp. 85–107.

Jones, T.D., Holgate, R.G.E., Baker, M.P. *et al.* (2016) *MAbs*, **8**, 1–9.

Kaplon, H. and Reichert, J.M. (2019) *MAbs*, **11**, 219–238.

King, C., Garza, E.N., Mazor, R., Linehan, J.L., Pastan, I., Pepper, M., Baker, D. (2014) *Proc. Natl. Acad. Sci*., **111**, 8577–8582.

Kovaltsuk, A., Leem, J., Kelm, S., Snowden, J., Deane, C.M., Krawczyk, K. (2018) *J. Immunol*., **201**, 2502–2509.

Lazar, G.A., Desjarlais, J.R., Jacinto, J., Karki, S., Hammond, P.W. (2007) *Mol. Immunol*., **44**, 1986–1998.

Prassler, J., Thiel, S., Pracht, C. *et al.* (2011) *J. Mol. Biol*., **413**, 261–278.

Rocklin, G.J., Chidyausiku, T.M., Goreshnik, I. *et al.* (2017) *Science*, **357**, 168–175.

Rouet, R., Jackson, K.J.L., Langley, D.B., Christ, D. (2018) *Front. Immunol*., **9**, 118.

Safdari, Y., Farajnia, S., Asgharzadeh, M., Khalili, M. (2013) *Biotechnol. Genet. Eng. Rev*., **29**, 175–186.

Zhai, W., Glanville, J., Fuhrmann, M. *et al.* (2011) *J. Mol. Biol*., **412**, 55–71.