OXFORD

Briefings in Bioinformatics, 2022, **23(3)**, 1–12
https://doi.org/10.1093/bib/bbac128
Problem Solving Protocol

Predicting binding affinities of emerging variants of SARS-CoV-2 using spike protein sequencing data: observations, caveats and recommendations

Ruibo Zhang, Souparno Ghosh and Ranadip Pal

Corresponding author. Ranadip Pal, Department of Electrical & Computer Engineering, Texas Tech University, Box 43102, Lubbock, TX 79409-3102, USA. Tel: 806-834-8631; Fax: 806-742-1245; E-mail: ranadip.pal@ttu.edu

Abstract

Predicting protein properties from amino acid sequences is an important problem in biology and pharmacology. Protein–protein interactions among SARS-CoV-2 spike protein, human receptors and antibodies are key determinants of the potency of this virus and its ability to evade the human immune response. As a rapidly evolving virus, SARS-CoV-2 has already developed into many variants with considerable variation in virulence among these variants. Utilizing the proteomic data of SARS-CoV-2 to predict its viral characteristics will, therefore, greatly aid in disease control and prevention. In this paper, we review and compare recent successful prediction methods based on long short-term memory (LSTM), transformer, convolutional neural network (CNN) and a similarity-based topological regression (TR) model and offer recommendations about appropriate predictive methodology depending on the similarity between training and test datasets. We compare the effectiveness of these models in predicting the binding affinity and expression of SARS-CoV-2 spike protein sequences. We also explore how effective these predictive methods are when trained on laboratory-created data and are tasked with predicting the binding affinity of the in-the-wild SARS-CoV-2 spike protein sequences obtained from the GISAID datasets. We observe that TR is a better method when the sample size is small and test protein sequences are sufficiently similar to the training sequence. However, when the training sample size is sufficiently large and prediction requires extrapolation, LSTM embedding and CNN-based predictive model show superior performance.

Keywords: COVID-19, machine learning, biological sequence analysis, protein–protein interaction, topological regression, performance evaluation.

Introduction

COVID-19 is a respiratory disease caused by the novel human coronavirus SARS-CoV-2. The infection is initiated by the binding of the viral spike protein (S-protein) receptor binding domain (RBD) and human angiotensin converting enzyme (ACE2) receptors. The mutations in spike protein, especially near the RBDs, impact the protein expression and viral transmissibility by increasing or decreasing the binding affinity [1]. Since, RBD is also the primary target of most neutralizing antibodies that inhibit the S-protein and ACE2 receptor binding [2], substantial changes in S-protein may prevent antibodies from recognizing the antigen and may diminish the effectiveness of autoimmunity or vaccines [3]. Understanding S-protein mutations and their relationship to viral characteristics are essential for understanding the virulence of SARS-CoV-2, monitoring its dangerous mutations and responding to COVID pandemics with appropriate clinical measures.

Since its discovery in human populations, SARS-CoV-2 has undergone multiple mutations. Periodic emergence of variants of concern (VOCs) and variants of interest (VOIs) have posed renewed threats to public health. The Alpha (B.1.1.7 lineage), Beta (B.1.351 lineage), Gamma (P.1 lineage) and Delta (B.1.617 lineage) variants have become major VOCs showing higher transmissibility and/or virulence as compared with the original strain of SARS-CoV-2. A collaborative genomic surveillance effort has been tracking emerging mutations and the sequencing data of emerging variants are made publicly available through online databases such as National Center for Biotechnology Information (NCBI) [4] and the Global Initiative on Sharing Avian Influenza Data (GISAID)[5]. These open datasets have made SARS-CoV-2 sequences

Ruibo Zhang received his BS degree from Beijing Institute of Technology in 2017. He is currently pursuing his Ph.D. degree in electrical engineering from Texas Tech University. His research interests include the development of machine learning and signal/image processing algorithms for biological applications.
Souparno Ghosh received his M.S. degree in statistics from the University of Calcutta, Kolkata, India, in 2004, and the Ph.D. degree in statistics from Texas A&M University, TX, USA, in 2009. He is currently an Associate Professor with the Department of Statistics, University of Nebraska, Lincoln, NE, USA. His research interests include Bayesian hierarchical model, statistical image analysis, and statistical machine learning.
Ranadip Pal received the B.Tech. degree in electronics and electrical communication engineering from the Indian Institute of Technology, Kharagpur, India, in 2002, and the M.S. and Ph.D. degrees in electrical engineering from Texas University, College Station, TX, USA, in 2004 and 2007, respectively. Since 2007, he has been with Texas Tech University, where he is currently a Professor with the Electrical and Computer Engineering Department. His research interests include machine learning, computational systems biology and stochastic modeling and control.

Received: December 23, 2021. **Revised:** February 13, 2022. **Accepted:** March 16, 2022

readily available and opened up a promising avenue to understand the correlation between sequences and viral pathogenicity.

In addition to experimental efforts, computational methods to predict protein functions based on the sequence characteristics are routinely carried out. Alignment-based homology analysis is perhaps the standard method for analyzing newly identified sequences [6]. This approach uses similarity search tools such as BLAST, PSI-BLAST, SSEARCH, FASTA and HMMER to identify sequences with similar structures or common ancestors by searching for sequences that have obvious matches with significantly higher similarity than random.

The similarity of proteins can be defined using alignment methods—for example, BLAST that approximates Smith–Waterman (SW) algorithm [7] to provide alignment scores as a proxy for similarities among sequences. In the proteomic alignment process, two protein sequences are assumed to share a common ancestor and mismatches/gaps represent the point mutations/insertions and deletions. In this article, the sequences we want to analyze are mutated from the common ancestor wild-type SARS-CoV-2 variant and have a high degree of similarity except for a few mutation loci. Consequently, we use SW algorithm for sequence alignment.

Due to the large variations in amino acid characteristics, different point mutations impact the protein characteristics differently. For example, conservative mutations have minor influences on the protein function as compared with radical mutations [3]. By introducing substitution matrix (PAM, BLOSUM, etc.) protein alignment takes into account differential impact of amino acid mutations as well. Thus, conservative mutations have higher alignment scores than radical mutations reflecting higher protein similarities.

With the similarity measure defined, often shallow data-driven methods are used to predict protein functions after aligning multiple sequences. Customarily, these approaches find the most similar instances to the sequence of interest and assume that high similarity in sequence results in similar responses (protein functions in our case). Models such as k-nearest neighbor (KNN), kernel regression, profile hidden Markov model [8] are popular predictive models that rely on similarity estimates among protein sequences.

More recently, deep learning methods have become popular tools to predict protein functions from their structures. These methods generally rely on quantification of protein sequences either at residual (single amino acid) level or at peptide level. At the residual level, each amino acid is quantified separately. Numeric vectors representing each residual in the protein sequence are then concatenated to represent the entire protein. For instance, principal components scores associated with the Vectors of Hydrophobic, Steric and Electronic properties (VHSE) [9] for individual amino acid descriptors are row concatenated to produce a 2-D array representing a protein sequence which, in turn, forms the input for the predictive model. At peptide level, numerical vectors quantifying the amino acids are further processed to produce an embedding vector representing the entire protein. Two immediate benefits of peptide level embeddings are (a) the configuration of residuals in each protein sequence is taken into account while generating the embedding and (b) the dimension of the embedding vector can be fixed a priori thereby alleviating the problem of dealing with wide variation in the length of the protein sequences.

Popular classes of models for generating the foregoing peptide level embeddings consist of (a) sequence neural network models, such as 1-D convolutional neural network (CNN), recurrent neural network (RNN), (b) k−mer based approaches for sequence analysis [10] and (c) autocovariance (AC)-based approaches that encode the protein sequence in terms of distance among the residuals [11]. Several deep learning methods have been developed for sequence embedding in the context of protein quantification. DeepAffinity [12] uses autoencoder-based Seq2Seq embedding, ProtSolver [13] and ELASPIC [14] use graph convolution-based neural networks to generate embeddings, DeepConv-DTI [15] and DeepDTA [16] apply 1-D convolution on protein sequences and generate convolutional embeddings of the entire sequence. MDeePred exploits multiple aspects of protein characteristics and feed them into the CNN model as a multi-channel input [17]. More recently, natural language processing (NLP) techniques have been brought to bear to encode protein sequences. These methods typically understand residuals as words/tokens and each sequence as a sentence. ProtVec [18] uses Word2Vec (Skip-gram) [19] to generate embeddings of protein sequences, with the properties of the protein being encoded in the semantic embedding. DeepAffinity integrates unsupervised Seq2Seq RNN for embedding and supervised CNN for prediction [12]. In ProtTrans [20], several transformer models (Transformer-XL, XLNet, Bert, Albert) are trained on 2.1 billion protein sequences and can be used to generate the vector embeddings and predict protein structures and characteristics. Turning to viral sequence embedding, [21] used bidirectional long short-term memory (BiLSTM) to analyze the semantic embeddings, viral fitness and immune escapes. The semantic change and grammaticality that were generated from the BiLSTM embedding were subsequently used for predicting the properties of sequences. Reference [22] developed long short-term memory (LSTM) based SPBuild for generating protein embeddings, reference [23] developed SeqVec algorithm that used Embeddings from Language Models to provide a fast model to create protein embeddings.

k−mer based methods, on the other hand, count all the subsequences of length k (k-mers) to produce the set $\{V, F\}$, where V is the set of all possible k-mers and F is the frequency of each appearing in the traversed sequence. A normalized version of fixed-dimensional F represents

the embedding of the protein sequence. The AC approach requires the residuals to be numerically encoded. Then, for each dimension of the encoded vector, AC can be calculated at given lags. The combination of k-mer and AC is also used as a protein-level featurization method [24].

We note that similarity-based approaches can be implemented after obtaining the protein embeddings as well. For example, the Euclidean distance of BiLSTM embeddings can be used to quantify similarity among protein sequences. The pairwise distances between sequences and their corresponding responses can be used to train distance regression models [25]. For a query sequence, such distance regression model predicts the distances in the response space that can be mapped to actual response value using backscoring techniques [26, 27]. These techniques respect the topology of the input and response spaces and allow more flexibilities in the sense that sequences that are similar are not assumed to produce responses that are similar as well. We call this technique topological regression (TR) and offer more details in Section 1 Topological Regression, Supplementary Material.

In the context of predicting viral properties of SARS-CoV-2 sequences, [28] utilized Bi-path Convolutional Neural Networks (BiPathCNN) to compare the infectivity patterns of SARS-CoV-2 and predict host range and potential reservoir species of this virus. Reference [29] utilized Screening for Non-acceptable Polymorphisms [30] to predict the protein function changes induced by single mutations on SARS-COV-2 spike RBD and discovered the critical mutations that influence S-protein stability and binding affinity. Wang *et al.*[31] utilized AA-index and CNN to predict binding affinity, protein expression and antibody escape. Reference [21] predicted the mutation probability and their effect on viral fitness and immune escape of SARS-CoV-2 using BiLSTM. Reference [14] used deep-learning feature extraction and gradient boosting decision tree to predict the effects of single mutations on the binding to ACE2. In [32], Wang *et al.* developed a joint model to predict paired viral–human protein pathogen–host interactions and examined their model on the dataset of the interactions between 26 SARS-CoV-2 proteins and human proteins found in [33]. They used ontology information was embedded into vectors, combined with one-hot encoded sequences and modeled as a classification problem with convolutional and fully connected networks.

In this article, we investigate several feasible data-driven viral characteristics prediction pipelines. We train these machine learning methods on lab-created mutations of SARS-CoV-2 strains and test the ability of the trained models in predicting the binding affinity and expression profiles of the strains observed in the real-world variants. We investigate how the size of the training set and the similarity between the training and test set impact the predictive performance of these empirical models. When the test sequences are close to (far from) the set of sequences in the training sample, we call the prediction as interpolation (extrapolation) task. Our results suggest that machine learning models should be carefully chosen depending on the size of the training dataset and 'location' of the test sequences vis-a-vis their training counterparts.

Methods

We begin by reviewing current computational approaches for predicting viral functions based on their sequences. We then propose a generic predictive pipeline that includes visualization of viral sequences, protein feature extraction, training predictive models. Figure 1 provides a graphical description of the pipeline. We confine ourselves to supervised models only and compare the prediction performance of several shallow and deep learners, including the foregoing TR method in the following sections.

Data preparation

First, we outline the data acquisition protocol. In this article, we used the deep mutation scan (DMS) dataset that was created from experiments in order to understand the relationships between the S-protein amino acid sequence and the viral fitness and antigenicity [34]. More specifically, the dataset was created to investigate RBD mutations' effects on the expression of spike protein and binding affinity to ACE2. Mutated hCoV19 spike protein sequences were generated from wild-type Wuhan-Hu-1 with each variant having one or more amino acid mutations. For the binding affinity data, log binding constants $\Delta log_{10}(K_D, app)$ relative to the wild-type SARS-CoV-2 RBD were provided and used as the responses. Each variant's mean fluorescence intensity (MFI) was measured for the expression data, and the $\Delta log(MFI)$ relative to the unmutated sequence was also supplied. Data were processed following the procedures described in [21]. After discarding the invalid variants, 105 526 and 116 258 variants were retained for training the models for predicting binding affinity and expression profiles, respectively.

For real-world sequences, we collected S-protein sequences from GISAID as of 3 December 2021 and related sample information as the metadata [5]. The metadata, for our purpose, included the dates of sample collection, locations and strains/lineages. We determined the variants according to the Pangolin lineage [35] in the metadata, following the VOC and VOI definition suggested by CDC. Sequences containing 'X', denoting unknown amino acids, were discarded. In total, 131 719 unique sequences with metadata were retained for testing purposes.

Sequence dissimilarity measure

We used SW algorithm with BLOSUM55 as the substitution matrix to measure similarities among protein sequences. We define the alignment distance as the difference between the maximum alignment score and the actual alignment score. These distances were

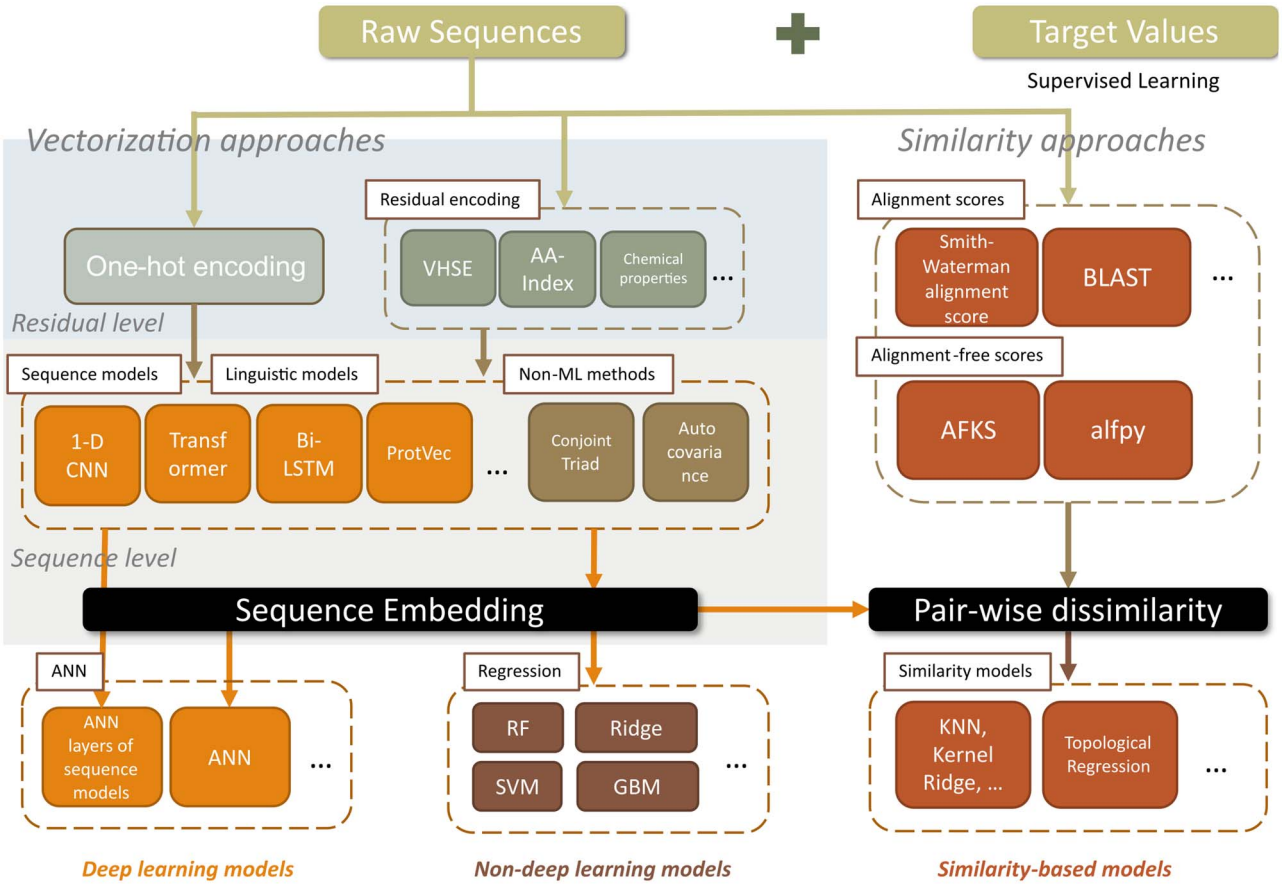

Figure 1. Overview of proposed protocol for viral sequence prediction.

used to offer a high-level visualization of sequence clusters and formed the input in the downstream TR model. Alignment scores and prediction performance are dependent on the substitution matrix. In this paper, we compared the substitution matrices provided in the library Biopython.

Approaches for visualization

After computing the similarities among protein sequences, we used multidimensional scaling (MDS) and t-Distributed Stochastic Neighbor Embedding (t-SNE) on the similarity measures to find the embeddings of sequences in a 2-D space. A scatter plot was then used to display the 2-D coordinates associated with the sequences. The coordinates were color coded according to their variant label.

Approaches for feature extraction

We used VHSE [9] technique for residual level featurization. It uses eight dimensions, corresponding to the dominant eight principal components of 50 physicochemical variables, to describe physicochemical properties of amino acids. Thus, for an S-protein sequence of length l, VHSE produced a matrix of size $l \times 8$.

For protein-level embedding, we used pre-trained ProtBert from [20] and BiLSTM from [21]. We note that, ProtBert was not exclusively trained on the DMS dataset. Rather, we used the transformer layers of the pre-trained ProtBert to obtain the embeddings of the protein sequences in DMS dataset. Admittedly, in our deployment, ProtBert was not optimized for DMS. However, we posit that this deployment would offer deeper insight into its performance in potentially out-of-distribution query points. The BiLSTM model, on the other hand, was indeed pre-trained on the same DMS dataset from [34], we simply ignored the output layer(s) of [34] and only retained the protein-level embeddings generated by the LSTM/transformer layers.

Approaches for predictive modeling

We compared the performance of models created under various combinations of featurization and predictive methods. The 1st class of models consisted of semantic embedding that encoded protein sequence information and then used shallow learners, trained on NLP embedding layers' output, as the predictive model. As mentioned before BiLSTM with semantic changes and grammaticality encoding [21] and ProtBERT [20] were two prototypical embedding generation models considered in class. Once the embeddings were generated, we also calculated the distances in the (embedded) feature space and trained our TR model.

The 2nd type of model consisted of residual level featurization with deep learning predictive models that can handle the sequential nature of the inputs. VHSE

was used to quantify each amino acid and the protein was encoded via row concatenation of the 8-D vectors representing each amino acid in the sequence. Since the columns of this protein-encoding matrix consist of principal components scores, the interpretation of the variables remains consistent across the rows of this matrix. Hence, a 1-D CNN, with eight channels matching the column dimension of the protein matrix, was used as a predictive model. For the GISAID dataset, sequences were padded with 0s to a maximum length of 1330 before being fed into the CNN model. The architecture of CNN is provided in Table 2, Supplementary Material.

The 3rd type of model relied on similarity-based approaches. We trained the TR model using the pairwise distances (dissimilarities) among sequences as inputs and response distance as the output. In principle, the entire training dataset could be used to calculate the pairwise distances, but the resulting distance matrix is too large for efficient downstream computation. One option is to induce sparsity by imposing some tapering structure [36]. However, in the absence of a biologically justifiable tapering structure, we preferred to use a randomly selected subset of the training dataset to act anchor points. These anchor points can be envisioned as a special case of 'knots' that are inserted in the construction of spatial predictive processes [37]. Selecting the number of anchor points depends on the trade-off between predictive accuracy and computational time. In general, the prediction performance increases with the number of anchor points. However, since the distances between the training data and the anchor points need to be computed and stored, fewer anchor points reduce the computational time and memory requirement considerably. Herein we compared 200 and 500 anchors to demonstrate the performance of TR. More details about TR are provided in Topological Regression Section, Supplementary Material. The list below summarizes various model combinations that were tested in this article:

- Semantic embedding + shallow learners:
  BiLSTM embedding (semantic change, CSCS scores) + Linear regression
  BiLSTM embedding + Ridge regression
  BiLSTM embedding + RF
  BiLSTM embedding + Fully-Connected (FC) ANN
  ProtBert embedding + Ridge regression
  ProtBert embedding + RF
  ProtBert embedding + FC-ANN
- Residual encoding + deep sequence model:
  VHSE encoding + 1-D CNN
- Similarity-based models:
  SW alignment score + Topological regression
  SW alignment score + KNN
  CNN embedding distances + Topological regression

Note that the listed methods are specifically for regression. However, the proposed pipeline also applies to classification tasks. We discuss its application on classification in Section 5, Proposed Pipeline in Classification Scenarios, Supplementary Material.

To illustrate the impact of the size of the training set on the predictive performance of the above set of models, we generated three different training sets by sampling 0.5%, 5% and 100% of the whole DMS dataset, resulting in $\approx$ 500, $\approx$ 5000 and $\approx$ 105 000 samples, respectively. Data were split into 60% training, 20% validation and 20% testing. Tuning of hyperparameters, for all models, were performed on the validation set via grid search. Hyperparameters with the best mean square error were selected to arrive at the final predictive model that was deployed on the test set. Details for hyperparameter tuning can be found in Section 6, Hyperparameters Tuning, Supplementary Material.

We also evaluated the performance gains from simple model stacking, where a linear regression model was trained on the validation set using the outputs of multiple base models as predictors and target values in the validation set as responses. The estimated regression parameters provided weights that were used to linearly combine the predictions generated by various candidate models on the test set.

Results

Visualization based on pairwise similarities

To visually illustrate the distances among the S-protein sequences, especially between DMS training set versus GISAID test data, we perform unsupervised MDS and t-SNE analyses. (Modified here) For visual clarity, among all 131K sequences collected from GISAID, we randomly select 100 samples for each VOC and 200 for others and concatenate them with the sequences from the DMS training set and then using the sequence dissimilarity measure discussed earlier, we project the sequences on 2-D MDS (t-SNE) planes. We display the 700 randomly selected points in Figure 2 and color-code different variants determined from their lineages. We notice that the variants were well separated in both MDS and t-SNE planes, and the main collection dates coincide with the clusters accordingly. Clearly, there exist considerable dissimilarities among the variants to suggest predictive models trained on one type of variant must be deployed with caution when trying to predict the outcomes associated with a different variant. Furthermore, the training set (DMS dataset) formed a cluster well separated from the two major VOCs, Alpha and Delta. Consequently, the prediction problem turned out to be an extrapolation problem with the feature set associated with the test set may be well outside the space spanned by the feature set associated with the training data. This scatter plot forms a part of exploratory data analysis that can help us choose the downstream predictive models. For example, if we want to computationally predict the binding affinity associated with the GISAID samples collected after June 2021 (marking the emerging of Delta variant) using the samples collected until June as the training set, we may

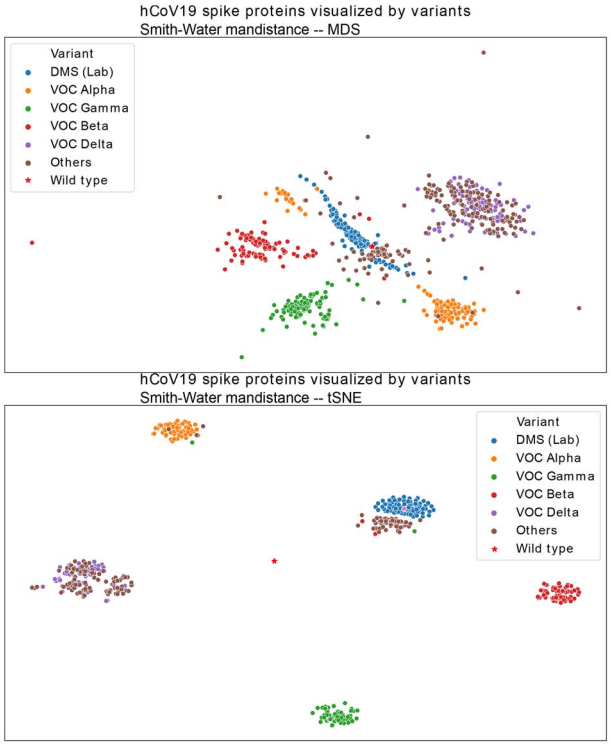

Figure 2. Scatter plot visualization of GISAID samples and DMS training dataset.

need to utilize the time stamps associated with each data point and develop a regime-switching time series model (emergence of new variants marking new regimes) that takes high dimensional temporally varying covariates as inputs and predict the binding affinities over time. The forecast errors that such non-stationary regime switching model [38] generates can potentially reveal higher levels of extrapolation error which, in turn, can be used to assess the reliability of these forecasts. Although, this modeling approach is appealing in its own right, we do not pursue this strategy any further in this article.

Comparison of predictive performance on virus binding affinity under homogeneous setting: interpolation task

In this section, we confine ourselves to the DMS dataset only. We compared the predictive performance of foregoing 11 models under three different sample size scenarios. Each model was trained on 316, 3165 and 63 000 samples and tested on 106, 1056 and 21 000 samples, respectively. Models were compared using the Spearman and Pearson correlation coefficients, normalized root mean square error (NRMSE) and normalized Mean Absolute Error (NMAE). Since the training and test samples were obtained from relatively homogeneous DMS set (as revealed in Figure 2), we consider this prediction task as interpolation. Figure 3 shows the foregoing comparison metrics for each model under each scenario.

First, observe that when the size of the training set is small ($\approx$ 100) TR, with each training sample acting as anchor points, outperformed the competing candidate models. Furthermore, TR offered considerable improvement over KNN regression implying that in complex regression problems the assumption of isometry between input and output space is perhaps not tenable. For moderate size training set ($\approx$ 3000), VHSE-CNN showed the best predictive performance. However, TR, with only 500 anchor points, produced fairly similar predictive results, at a much lower computation cost, as compared with the aforesaid deep learner. As dataset size increases, we observe that VHSE-CNN steadily pulls away from the other competing models affirming the reliability of the deep learners when sufficient samples are available to train them.

Turning to model stacking, we select the top two individual models, TR and VHSE-CNN, under all three sample size scenarios and generate stacked predictions obtained using the above discussed linear regression method. For a complete picture, we also report the prediction performance for each candidate model, their average prediction (mean model) in Figure 4. As expected, stacking produces better predictive performance as compared with individual models and the mean model under all sample size scenarios.

We perform similar analyses on the DMS protein expression dataset from [34]. After discarding the sequences containing stops (*) in the middle, 116k sequences were retained. Again, the entire dataset was sampled at 0.5%, 5% and 100% level to assess the performance of models under different sizes of training set. TR again turned out to be the best performer under small sample scenario with VHSE-CNN dominating other models for the remaining two scenarios. Stacking of TR and VHSE-CNN produced consistently superior performance in terms of error. All the results for protein expression were relegated to the Supplementary Materials.

Binding affinity prediction on GISAID sequences: extrapolation task

Ideally, we would like the predictive models, trained using the information available on the existing/experimental strains, to accurately assess the potency/transmissibility of new virus strains. To understand the predictive 'reach' of the models discussed herein, we trained four prototypical models on DMS dataset using binding affinity as the target response and generated predictions for the real-world variants extracted from the GISAID dataset. The prototypical models considered in this section are: (a) RF and Ridge regression with BiLSTM embedding (RF and Ridge were chosen to represent model-free and parametric predictive approaches, respectively.), (b) RF and Ridge regression with ProtBert embedding, (c) 1-D CNN with VHSE encoded sequences and (d) TR with SW distance with BLOSUM55 substitution matrix. Observe that, above mentioned models represent four major classes of the 11 models considered in this article and our intention here is to understand which class of models generate reliable predictions when the features associated with

Figure 3. Comparison of models for viral fitness dataset with 3 different sample sizes. The data were sampled from the DMS fitness dataset which has 105k samples, and three subsets were obtained by using different sampling fractions.

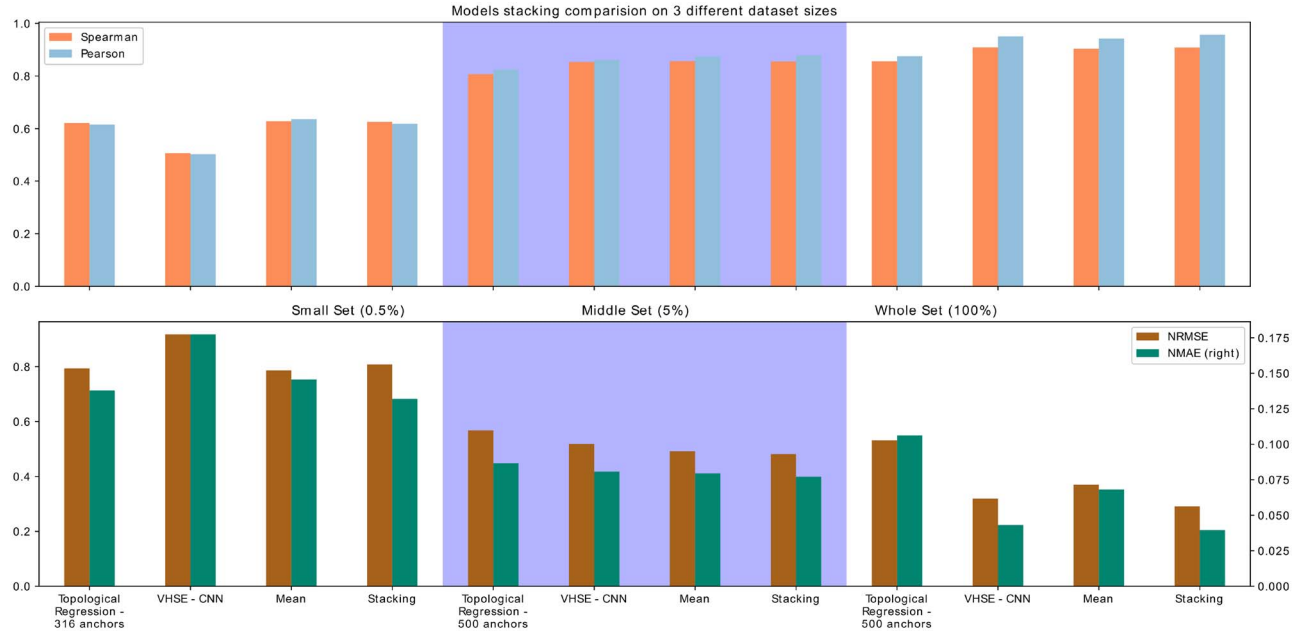

Figure 4. Mean stacking performance of TR and VHSE CNN for fitness dataset at three different sample sizes.

test samples may not belong to the space spanned by the training features (Figure 2), and therefore, the prediction becomes an extrapolation task.

However, the binding affinities of the variants extracted from GISAID were not available, and hence, customary metrics to compare prediction accuracy (NRMSE, NMAE etc.) could not be computed. Instead, we looked at the distribution of predicted binding affinities for each VOC and generated a stochastic ordering of these VOCs in terms of the predicted binding affinities. We then compared this ordering with the results of microscale thermophoresis (MST) experiments that estimated the binding affinities between human ACE2 and RBD of SARS-CoV-2 for the Alpha, Beta, Gamma and Delta variants [39]. It appeared that both MST and molecular dynamics simulations indicated that the Alpha variant had significantly higher binding affinity with ACE2 receptors as compared with Beta, Gamma and Delta variants.

Figure 5 shows the distributions of binding affinity distributions predicted using the above models for VOC Alpha, Beta, Gamma, Delta and sequences not identified as VOCs or VOI (defined as 'Others'). To formally assess whether our predictions are in agreement with the MST results [39], we performed a Kruskal–Wallis test using variants as the factors and predicted binding affinities from representatives of four classes of models as responses. Rejection of Kruskal–Wallis test was followed up with post-hoc Dunn test (with Sidak adjustment) to establish the stochastic ordering among the variants (A random variable Y is said to be stochastically larger than a random variable X if $P(Y > t) \geq P(X > t)$.). We used the existing MST results for hypotheses generation and posited, as alternative hypotheses in the foregoing Dunn test, that predicted binding affinities of Alpha variant will be stochastically larger than the predicted binding affinities of (i) Beta, (ii) Gamma and (iii) Delta variants. Both BiLSTM and TR provided statistical evidence that binding affinities of Alpha were stochastically larger than that of Delta, Beta and Gamma variants (P-values ≈ 0), thereby supporting the results from MST experiments (Table 4, Supplementary Material). However, results from ProtBert and VHSE-CNN indicated different stochastic ordering. We relegate the detailed statistical test results to the supplementary materials Exploration on GISAID dataset section.

Conclusion

In this paper, we briefly reviewed predictive modeling approaches for virus sequences and proposed a general protocol for predicting virus characteristics from raw sequences. We tested three ways of generating predictions: (a) residual encoding followed by deep sequence learners model, (b) semantic sequential embedding followed by shallow learners and (c) similarity-based models for TR, under two different scenarios—(i) interpolation task, when input space of training set closely matches with that of the test set and (ii) extrapolation task, when input space of training set does not overlap with that of the test set. Our results suggest that both sample size and distance of query point(s) from the training set must be taken into account before determining appropriate predictive strategy. Consequently, we strongly recommend performing an initial unsupervised visualization of the input feature cluster associated with training and test samples using MDS or t-SNE projections. If the distance between the input spaces associated with training and test samples is close, we recommend similarity-based TR methods when the size of the

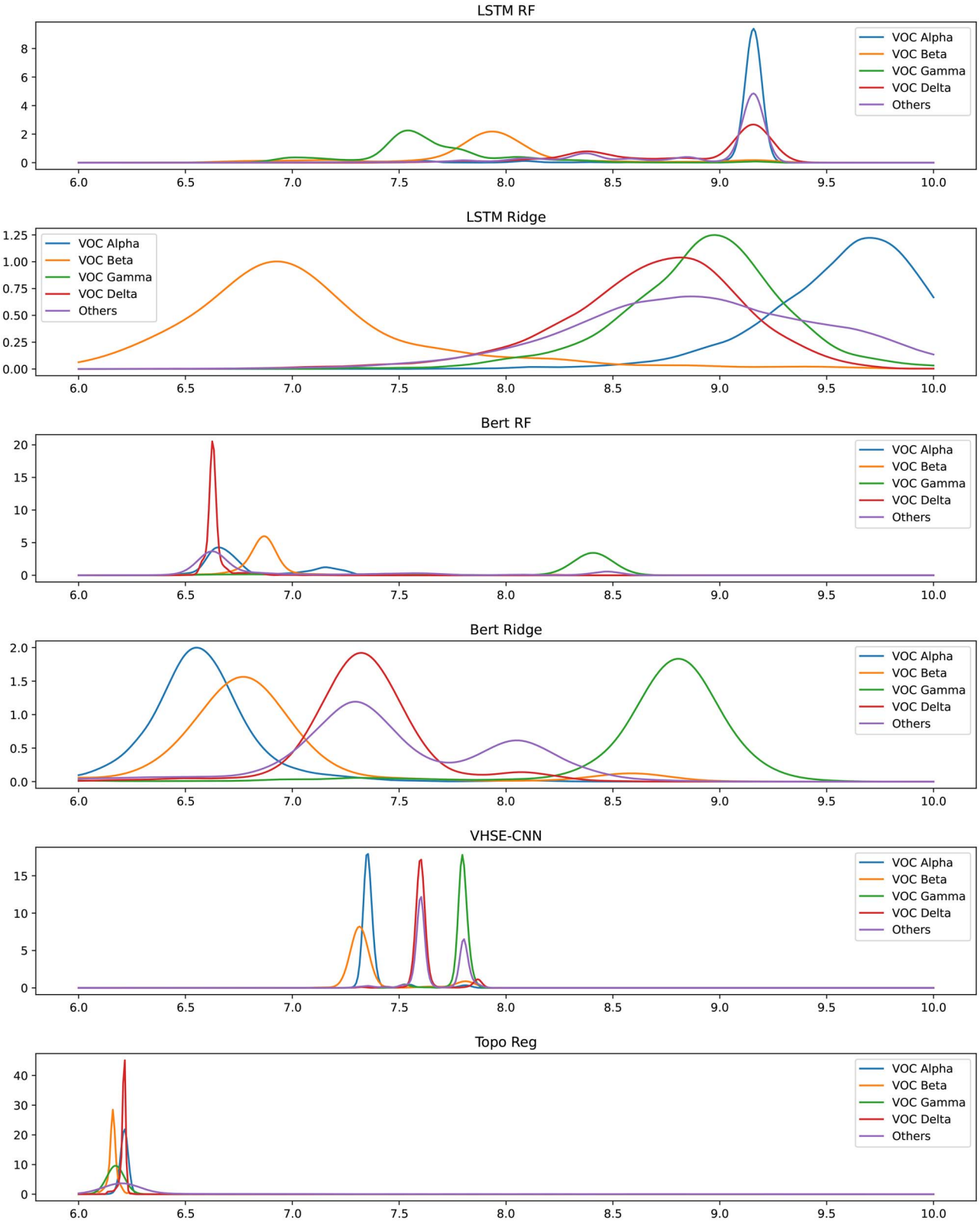

Figure 5. Affinity prediction of GISAID sequences by different models. The distributions of predictions are shown in the figure grouped by VOC variants.

training set is not sufficient to train deep learning models. However, under the same scenario, when sufficient training samples are available, we recommend deep sequence learners (e.g. 1-D CNNs).

If, however, the input space associated with the training set is sufficiently far from that of the test samples, CNN-based deep learners may not be the best predictive strategy. This caveat is important because

CNN-based deep learners can show excellent k−fold cross validation performance within the training set and therefore can offer an objective justification for selecting this type of models. However, under extrapolation scenario, k-fold cross validation can potentially overestimate the extrapolative potential of these predictive models [40]. Our results suggest that, for extrapolation tasks, pre-processing input features using transformer-based architectures that capture semantic information coupled with a standard shallow predictive model can produce more reliable predictions. This finding agrees with [41] who demonstrated that transformers, trained on sufficiently large datasets, can produce robust predictions over a spectrum of covariate shifts in test samples.

We also note that, for regression tasks, although there may exist considerable variations in the predicted binding affinity values for the test samples obtained under different models, there may exist agreement in terms of ranking the variants according to the predicted binding affinity. It appears that, at least in the context of this article, ranking the target properties of new variants apropos of existing observed variants is more reliable. Furthermore, we caution against using a single model, which may turn out to be the best performing one during the training phase, for predicting the viral properties of new variants from their S-protein sequence. Instead, we recommend using multiple transformer-based algorithms to encode the protein sequence information and then use these extracted features to train a set of shallow learners as the predictive bag-of-models. Each combination of embedding and shallow predictive model should be used to predict the properties of multiple known variants and generate a stochastic ordering of these variants. The level of agreement among the candidate model combinations in terms of ranking the variants should be used to quantify the reliability of the prediction exercise. If the candidate models do not produce statistically similar ordering (recall, ProtBert and BiLSTM produced very different ordering of variants), we caution against making any conclusive statements about the ordering of the variants.

In summary, we submit that, in the absence of readily available experimental information on focal viral characteristics of emerging variants of SARS-CoV-2, machine learning models can be used to provide a relatively quick assessment of the characteristics of interest. However, customary predictive accuracy metrics associated with empirical models cannot be computed in this context due to the inavailability of ground-truthing experimental information. Hence, in the absence of well-established formulae for extrapolation penalties associated with complex machine learning models, we recommend that outputs obtained from these models should also be accompanied by distance metrics quantifying how far the input space of training sample is compared with the input space of the query variants. We caution against using conventional deep learning predictive algorithms when the input space of the training samples does not intersect with that of the query samples. We also caution against using a single model for extrapolation tasks regardless of how well the said model performs in the training phase.

Key Points

- Utilizing the proteomic data of SARS-CoV-2 to predict its viral characteristics will greatly aid in disease control and prevention for this rapidly evolving virus.
- We review and compare recent successful prediction methods based on long short-term memory (LSTM), transformer, convolutional neural network (CNN) and a similarity-based topological regression model and offer recommendations about appropriate predictive methodology depending on the similarity between training and test datasets.
- We also explore how effective these predictive methods are when trained on laboratory-created data and are tasked with predicting the binding affinity of the in-the-wild SARSCoV-2 spike protein sequences obtained from the GISAID datasets.
- In the absence of well-established formulae for extrapolation penalties associated with complex machine learning models, we recommend that outputs obtained from these models should also be accompanied by distance metrics quantifying how far the input space of training sample is compared with the input space of the query variants.
- We caution against using conventional deep learning predictive algorithms when the input space of the training samples does not intersect with that of the query samples. We also caution against using a single model for extrapolation tasks regardless of how well the said model performs in the training phase.

Data and Code Availability

The DMS data are available in [34]. The GISAID data are available at https://www.gisaid.org/ [5]. Codes to reproduce the results in this paper are available at https://github.com/Ribosome25/cov_seqs.

Supplementary Data

Supplementary data are available online at https://academic.oup.com/bib.

Author Contributions Statement

R.Z., S.G. and R.P. formulated the problem and associated analysis. R.Z. conducted the experiments. R.Z., S.G. and R.P. wrote and reviewed the manuscript.

Acknowledgements

The authors thank the anonymous reviewers for their valuable suggestions. The authors acknowledge the High Performance Computing Center (HPCC) at Texas Tech

University for providing computational resources that have contributed to the research results reported within this paper. URL: http://www.hpcc.ttu.edu

Funding

National Science Foundation (Grants Nos 2007903, 2007418).

References

1. Shang J, Ye G, Shi K, *et al.* Structural basis of receptor recognition by sars-cov-2. *Nature* 2020;**581**(7807):221–4.
2. Piccoli L, Park Y-J, Tortorici MA, *et al.* Mapping neutralizing and immunodominant sites on the sars-cov-2 spike receptor-binding domain by structure-guided high-resolution serology. *Cell* 2020;**183**(4):1024–42.
3. Kupferschmidt K. New mutations raise specter of 'immune escape'. *Science* 2021;**371**(6527):329–30.
4. Sayers EW, Beck J, Bolton EE, *et al.* Database resources of the national center for biotechnology information. *Nucleic Acids Res* 2021;**49**(D1):D10.
5. Shu Y, McCauley J. Gisaid: global initiative on sharing all influenza data–from vision to reality. *Eurosurveillance* 2017; **22**(13):30494.
6. Pearson WR. An introduction to sequence similarity ("homology") searching. *Curr Protoc Bioinformatics* 2013;**42**(1):3–1.
7. Smith TF, Waterman MS. Identification of common molecular subsequences. *J Mol Biol* 1981;**147**(1):195–7.
8. Sinha S, Eisenhaber B, Lynn AM. Predicting protein function using homology-based methods. In: *Bioinformatics: Sequences, Structures, Phylogeny*. Springer, Singapore, 2018, 289–305.
9. Mei H, Liao ZH, Zhou Y, *et al.* A new set of amino acid descriptors and its application in peptide qsars. *Peptide Sci* 2005;**80**(6): 775–86.
10. Manekar SC, Sathe SR. A benchmark study of k-mer counting methods for high-throughput sequencing. *GigaScience* 2018;**7**(12):giy125.
11. Guo Y, Lezheng Y, Wen Z, *et al.* Using support vector machine combined with auto covariance to predict protein–protein interactions from protein sequences. *Nucleic Acids Res* 2008;**36**(9):3025–30.
12. Karimi M, Di W, Wang Z, *et al.* Deepaffinity: interpretable deep learning of compound-protein affinity through unified recurrent and convolutional neural networks. *Bioinformatics* 2019;**35**(18):3329–38.
13. Strokach A, Becerra D, Corbi-Verge C, *et al.* Fast and flexible protein design using deep graph neural networks. *Cell Syst* 2020;**11**(4):402–411.e4.
14. Strokach A, Lu TY, Kim PM. Elaspic2 (el2): Combining contextualized language models and graph neural networks to predict effects of mutations. *J Mol Biol* 2021;**433**(11):166810.
15. Lee I, Keum J, Nam H. Deepconv-dti: prediction of drug-target interactions via deep learning with convolution on protein sequences. *PLoS Comput Biol* 2019;**15**(6):e1007129.
16. Öztürk H, Özgür A, Ozkirimli E. Deepdta: deep drug-target binding affinity prediction. *Bioinformatics* 2018;**34**(17):i821–9.
17. Rifaioglu AS, Cetin Atalay R, Cansen Kahraman D, *et al.* Mdeepred: novel multi-channel protein featurization for deep learning based binding affinity prediction in drug discovery. *Bioinformatics* 2020;**37**(5):693–704.
18. Asgari E, Mofrad MRK. Continuous distributed representation of biological sequences for deep proteomics and genomics. *PLoS One* 2015;**10**(11):e0141287.
19. Mikolov T, Chen K, Corrado G, *et al.* Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781. 2013.
20. Elnaggar A, Heinzinger M, Dallago C, *et al.* Prottrans: towards cracking the language of life's code through self-supervised deep learning and high performance computing. arXiv preprint arXiv:2007.06225. 2020.
21. Hie B, Zhong ED, Berger B, *et al.* Learning the language of viral evolution and escape. *Science* 2021;**371**(6526):284–8.
22. Yamada KD, Kinoshita K. De novo profile generation based on sequence context specificity with the long short-term memory network. *BMC Bioinformatics* 2018;**19**(1):1–11.
23. Heinzinger M, Ahmed Elnaggar Y, Wang CD, *et al.* Modeling aspects of the language of life through transfer-learning protein sequences. *BMC Bioinformatics* 2019;**20**(1):1–17.
24. Wang X, Wang R, Wei Y, *et al.* A novel conjoint triad auto covariance (ctac) coding method for predicting protein-protein interaction based on amino acid sequence. *Math Biosci* 2019;**313**: 41–7.
25. Sim A, Tsagkrasoulis D, Montana G. Random forests on distance matrices for imaging genetics studies. *Stat Appl Genet Mol Biol* 2013;**12**(6):757–86.
26. Tsagkrasoulis D, Montana G. Random forest regression for manifold-valued responses. *Pattern Recognit Lett* 2018;**101**:6–13.
27. Bengio Y, Paiement J-F, Vincent P, *et al.* Out-of-sample extensions for lle, isomap, mds, eigenmaps, and spectral clustering. *Adv Neural Inform Proc Syst* 2003;**16**:177–84.
28. Guo Q, Li M, Wang C, *et al.* Host and infectivity prediction of Wuhan 2019 novel coronavirus using deep learning algorithm. bioRxiv. 2020.01.21.9140442020.
29. Teng S, Sobitan A, Rhoades R, *et al.* Systemic effects of missense mutations on sars-cov-2 spike glycoprotein stability and receptor-binding affinity. *Brief Bioinform* 2020;**22**(2):1239–53.
30. Bromberg Y, Yachdav G, Rost B. Snap predicts effect of mutations on protein function. *Bioinformatics* 2008;**24**(20):2397–8.
31. Wang B, Gamazon ER. Modeling mutational effects on biochemical phenotypes using convolutional neural networks: application to sars-cov-2bioRxiv. 2021.01.28.428521. 2021.
32. Liu-Wei W, Kafkas S, Chen J, *et al.* Deepviral: prediction of novel virus-host interactions from protein sequences and infectious disease phenotypes. *Bioinformatics* 2021;**37**(17):2722–9.
33. Gordon DE, Jang GM, Bouhaddou M, *et al.* A SARS-cov-2 protein interaction map reveals targets for drug repurposing. *Nature* 2020;**583**(7816):459–68.
34. Starr TN, Greaney AJ, Hilton SK, *et al.* Deep mutational scanning of SARS-cov-2 receptor binding domain reveals constraints on folding and ace2 binding. *Cell* 2020;**182**(5): 1295–1310. e20.
35. Rambaut A, Holmes EC, O'Toole Á, *et al.* A dynamic nomenclature proposal for sars-cov-2 lineages to assist genomic epidemiology. *Nat Microbiol* 2020;**5**(11):1403–7.
36. Kaufman CG, Schervish MJ, Nychka DW. Covariance tapering for likelihood-based estimation in large spatial data sets. *J Am Stat Assoc* 2008;**103**(484):1545–55.
37. Banerjee S, Gelfand AE, Finley AO, *et al.* Gaussian predictive process models for large spatial data sets. *J R Stat Soc Series B Stat Methodology* 2008;**70**(4):825–48.
38. Hamilton JD. Regime switching models. In: *Macroeconometrics and Time Series Analysis*. Palgrave Macmillan, London, 2010, 202–9.

39. Kim S, Liu Y, Lei Z, *et al*. Differential interactions between human ace2 and spike rbd of SARS-cov-2 variants of concern. bioRxiv. 2021.
40. Xiong Z, Cui Y, Liu Z, *et al*. Evaluating explorative prediction power of machine learning algorithms for materials discovery using k-fold forward cross-validation. *Comput Mater Sci* 2020;**171**:109203.
41. Bhojanapalli S, Chakrabarti A, Glasner D, *et al*. Understanding robustness of transformers for image classification. arXiv preprint arXiv:2103.14586. 2021.