Briefings in Bioinformatics, 2023, **24(5)**, 1–9

https://doi.org/10.1093/bib/bbad290

Advance access publication date 20 August 2023

Problem Solving Protocol

ETLD: an encoder-transformation layer-decoder architecture for protein contact and mutation effects prediction

He Wang, Yongjian Zang, Ying Kang, Jianwen Zhang, Lei Zhang and Shengli Zhang

Corresponding author. S. Zhang, MOE Key Laboratory for Nonequilibrium Synthesis and Modulation of Condensed Matter, School of Physics, Xi'an Jiaotong University, Xi'an 710049, China. Tel.: 029-82660915; Fax: 029-82668559; E-mail: zhangsl@xjtu.edu.cn

Abstract

The latent features extracted from the multiple sequence alignments (MSAs) of homologous protein families are useful for identifying residue–residue contacts, predicting mutation effects, shaping protein evolution, etc. Over the past three decades, a growing body of supervised and unsupervised machine learning methods have been applied to this field, yielding fruitful results. Here, we propose a novel self-supervised model, called encoder-transformation layer-decoder (ETLD) architecture, capable of capturing protein sequence latent features directly from MSAs. Compared to the typical autoencoder model, ETLD introduces a transformation layer with the ability to learn inter-site couplings, which can be used to parse out the two-dimensional residue–residue contacts map after a simple mathematical derivation or an additional supervised neural network. ETLD retains the process of encoding and decoding sequences, and the predicted probabilities of amino acids at each site can be further used to construct the mutation landscapes for mutation effects prediction, outperforming advanced models such as GEMME, DeepSequence and EVmutation in general. Overall, ETLD is a highly interpretable unsupervised model with great potential for improvement and can be further combined with supervised methods for more extensive and accurate predictions.

Keywords: encoder-transformation layer-decoder (ETLD) model, multiple sequence alignments (MSAs), transformation matrix, contact prediction, mutation effects prediction

INTRODUCTION

Long-term environmental selective pressures and protein functional maintenance require inter-site balance during protein sequence evolution [1–11]. This is called co-evolutionary information, which can be obtained from the multiple sequence alignments (MSAs) of homologous protein families. Over the past three decades, a growing number of supervised and unsupervised machine learning methods have been applied to capture the latent features of protein sequences (e.g. co-evolutionary signals), which have been further applied to residue–residue contact determination, mutation effects prediction, near-natural sequences generation, protein evolutionary modeling and protein design [8–10, 12–22]. To date, approximately, 230 million protein sequences have been included in Uniprot and are undergoing exponential growth, but only about 2‰ have been reviewed [23]. In this case, the unsupervised learning-based sequence analysis will have a wide range of applications not only for large numbers of unannotated sequences but also as input features for special-purpose supervised models.

Recently, unsupervised nature language processing models, such as Transformer, BERT and GPT, have rapidly gained popularity, and attention-based protein language models [22] are bringing further breakthroughs in bioinformatics analysis [24]. ESM-1b uses the Transformer model to pre-train 2.5 billion protein sequences and then predicts residue–residue contacts, secondary structures and mutation effects by further linear projections or deep neural networks [11]. MSA transformer employs the axial attention-based Transformer model that learns latent features by training on millions of MSAs and predicts residue–residue contacts by logistic regression [25]. Meier *et al.* proposed ESM-1v based on the zero-shot transfer to predict mutation effects, which does not require further supervision at all [26]. A link of works has emerged to learn deeper features from protein sequences through self-supervised pre-training and to play an increasingly

He Wang is a PhD candidate at the School of Physics, Xi'an Jiaotong University. His research interests are bioinformatics, molecular dynamics and deep learning.
Yongjian Zang is a PhD candidate at the School of Physics, Xi'an Jiaotong University. Her research interests are bioinformatics, molecular dynamics and deep learning.
Ying Kang is a MSc candidate at the School of Physics, Xi'an Jiaotong University. Her research interests are bioinformatics, drug design and screening and molecular dynamics.
Jianwen Zhang is a MSc candidate from the School of Physics, Xi'an Jiaotong University. Her research interests are bioinformatics, drug design and screening and molecular dynamics.
Lei Zhang is a professor at the School of Physics, Xi'an Jiaotong University. His research interests are protein structure and Cryo-EM.
Shengli Zhang is a professor at the School of Physics, Xi'an Jiaotong University. His research interests are bioinformatics, drug design and screening, molecular dynamics and deep learning.

Received: April 12, 2023. **Revised:** June 21, 2023. **Accepted:** July 26, 2023

important role in evolution, structure and function. However, limited interpretability and high computational complexity are the main challenges of protein language models. Therefore, it is necessary to develop low-cost, small-scale, high-accuracy and highly interpretable models for sequence analysis.

Coevolution-based methods have been the state-of-the-art unsupervised predictors of physical contacts before the advent of protein language models. They are based on the idea of co-evolution, i.e. mutation of one residue may cause a chain of mutations in other residues to maintain protein structure and biological function, and contact predictors suggest that these mutations correspond to neighboring residues in the spatial structure. The representative methods are based on direct coupling analysis (DCA), including PSICOV [27] using the sparse inverse of covariance matrix; EVfold [28] and mfDCA [4] using mean-field approximation; and PLMDCA [14], GREMLIN [29] and CCMpred [30] using Potts model by maximizing pseudo-likelihood. A range of supervised machine learning-based or meta-based methods, such as DNCON2 [18], RaptorX-Contact [31], SPOT-Contact [32], MetaPSICOV [15], R2C [16] and MapPred [20], use prediction results from coevolution-based methods, which are important for improving accuracy.

The successful application of DCA in contact prediction has also inspired attempts to construct protein mutation landscapes for mutation effects prediction using co-evolutionary approaches as well. EVmutation [8] built on PLMDCA is one of the representative methods to illustrate epistasis by explicitly modeling the coupling relationships between all residue pairs. However, due to the exponential growth of parameters, PLMDCA can only consider single-site and pairwise constraints. Thus, more methods try to improve the prediction accuracy by considering third-order and higher-order constraints. For example, Riesselman *et al*. developed DeepSequence [9] based on the variational auto-encoders (VAE) model by using latent variable space to capture higher-order interactions between sites; Laine *et al*. proposed GEMME [10], which considers high-order constraints by building a phylogenetic tree of homologous protein families when quantifying mutational effects. It has shown that models considering the sequence co-variation outperformed the site-independent models, such as SIFT [33], PolyPhen-2 [34], CADD [35] and so on [8–10].

In this work, we propose an autoencoder-based architecture referred to as encoder-transformation layer-decoder (ETLD) to capture latent couplings between sites. It is a simple but interpretative unsupervised method with a parameter scale comparable to that of the classical PLMDCA method. ETLD assumes that each residue can be a linear or nonlinear combination of other sites whose weight factors form the so-called transformation matrix (**TM**), which is related to the deeper features in MSAs. We show that ETLD can be used to directly implement predictions of physical contacts and mutation effects with comparable or better performance than models with similar parametric quantities.

MATERIALS AND METHODS

ETLD model

How to extract co-evolutionary information from MSAs has been explored for at least three decades. Except for the state-of-the-art protein language models, the site-independent model, Potts model and VAE model have played important roles in studying the inter-site correlations of protein sequences. From the perspective of sequence generation (Figure 1A), in the site-independent model, the selection of amino acids at each site depends only on the properties of that site, such as amino acid distribution frequencies, physicochemical properties, etc. Potts model will consider the pairwise coupling relationship. The VAE model further considers higher-order inter-site constraints, where each site can be viewed as a nonlinear union of all sites in the latent space. Also based on the auto-encoder, in contrast, ETLD emphasizes the effects of other sites on the target site, and the **TM** is introduced to represent the effect weights rather than the invisible hidden space in the VAE model.

The overall flowchart of ETLD is shown in Figure 1B. ETLD takes individual amino acid sequences as entries, with the sequence in the source protein MSA as the input and the corresponding sequence in the target protein MSA as the target output. Here, we adopt $\upsilon = 22$ amino acid markers, including 20 standard amino acids, a start marker '*' to indicate the beginning of each sequence, and a padding marker '-' to indicate gaps or unknown amino acids. Each amino acid marker is converted into a learnable parameter vector with size d weighted by multiply $\sqrt{d}$ and then the source sequence with length n is embedded into $x \in \mathbb{R}^{n\times d}$ by the embedding layer. The encoder and the decoder use the residual block (Figure S1 available online at http://bib.oxfordjournals.org/), where the input and output have dimension d and the inner layer has dimensionality d_{inner}. The transformation layer consists of a **TM** ($\mathbf{TM} \in \mathbb{R}^{n\times m}$), acting as a coefficient matrix for converting the encoder output $\mathbf{O}^{\mathrm{E}} \in \mathbb{R}^{n\times d}$ to the decoder input $\mathbf{I}^{\mathrm{D}} \in \mathbb{R}^{m\times d}$ (Equation (1)), where m is the length of the target sequence. In other words, each row vector in $\mathbf{I}^{\mathrm{D}}$ can be represented as a weighted sum of row vectors in $\mathbf{O}^{\mathrm{E}}$.

$$\mathbf{I}^{\mathrm{D}} = \mathbf{TM}^{\mathrm{T}} \cdot \mathbf{O}^{\mathrm{E}}$$

$$\text{i.e. } \mathbf{I}^{\mathrm{D}}_{j,:} = \sum_{i} \mathbf{TM}_{i,j} \cdot \mathbf{O}^{\mathrm{E}}_{i,:}$$

$$i = 1, 2, \ldots, n; j = 1, 2, \ldots, m \qquad (1)$$

where $\mathbf{TM}^{\mathrm{T}}$ is the transpose of **TM**. Note that when the source and target MSA are the same, there is $n = m$, and we need to fix the diagonal parameter of **TM** to 0 (i.e. $\mathbf{TM}_{\mathrm{ij}} = 0$) before Equation (1) to avoid the self-impact of each site. Finally, the linear layer together with the 'Log-SoftMax' function will convert the decoder output into the predicted probabilities of amino acid markers, as follows:

$$\mathrm{LogSoftMax}_{\mathrm{i}}^{\mathrm{s}}(\sigma) = \log_e\left[\frac{\exp\left(\mathbf{O}^{\mathrm{L}}_{\mathrm{i},s}\right)}{\sum_s \exp\left(\mathbf{O}^{\mathrm{L}}_{\mathrm{i},s}\right)}\right] = \log_e\left(p_{\mathrm{i}}^{\mathrm{s}}(\sigma)\right) \qquad (2)$$

where $\mathbf{O}^{\mathrm{L}}_{\mathrm{i},s}$ is the value of the linear layer output ($\mathbf{O}^{\mathrm{L}} \in \mathbb{R}^{m\times\upsilon}$) taking amino acid marker s at the site i of sequence σ; $p_{\mathrm{i}}^{\mathrm{s}}(\sigma)$ is the corresponding statistical probability. Finally, the amino acids with the highest probability are strung together into the predicted sequence. The loss function is then constructed by comparing the differences between the predicted and target sequences. Note that when the source and target protein MSAs are the same, the target sequence is the input sequence. And, after several epochs of training, ETLD is able to capture the inter-site couplings through the transformation layer, which will be incorporated into the reconstruction of the target sequence. These can subsequently be used for contact prediction and mutation effects prediction (Figure 1C and D).

The double-channel and multi-head TM

The transformation layer with a one-layer **TM** ($\mathbf{TM} \in \mathbb{R}^{n\times m}$) can theoretically represent only linear relationships between sites. To consider the linear and nonlinear associations, **TM** was designed as a double-channel mode, where one of the channels uses the

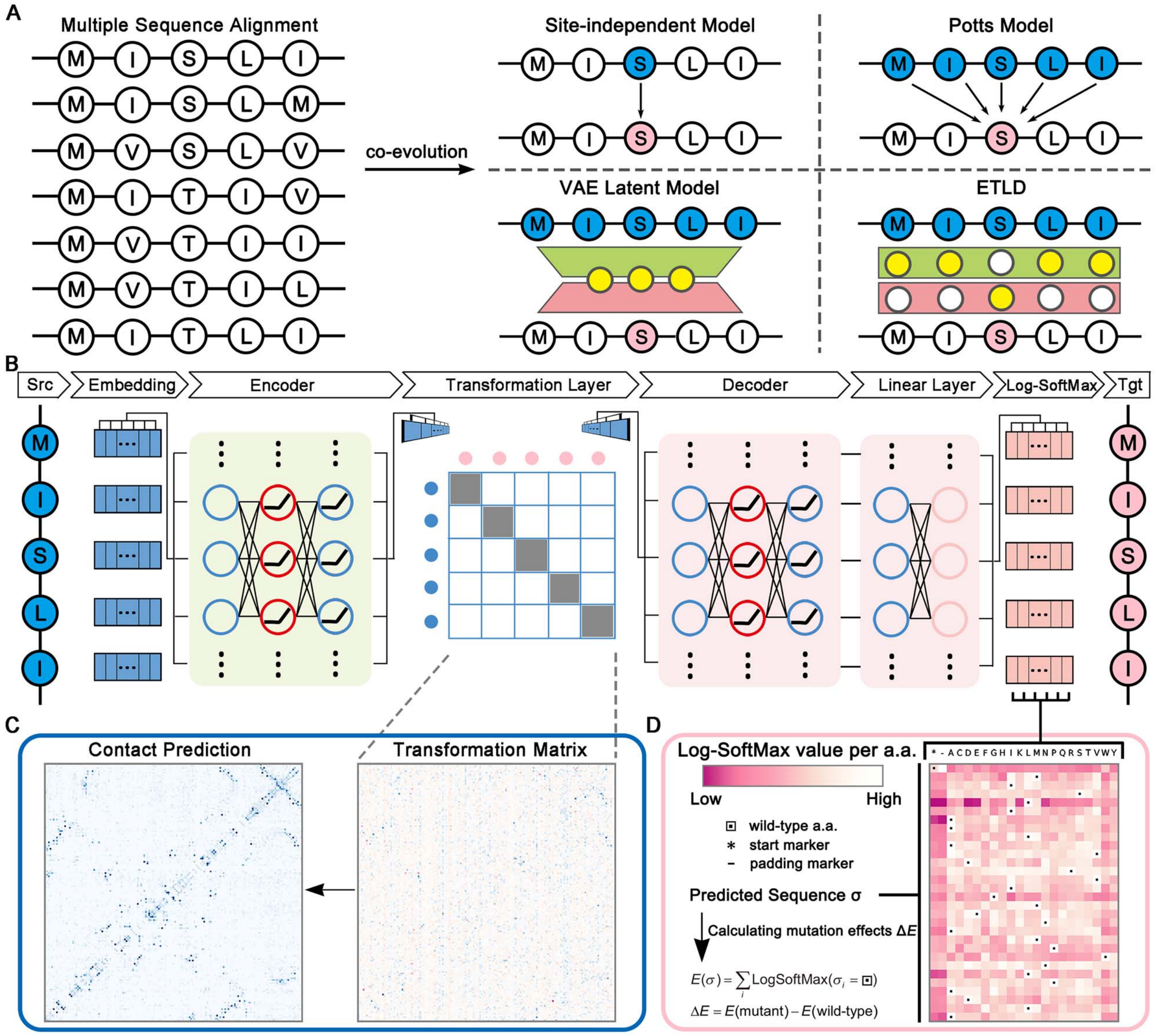

Figure 1. The architecture of ETLD. (**A**) Comparison of the ETLD mechanism with the site-independent model, Potts model and VAE latent model. (**B**) The overall flowchart of ETLD. The source sequence ('Src') is reconstructed into the target sequence ('Tgt') by passing through the embedding layer, encoder, transformation layer, decoder, linear layer and the 'Log-SoftMax' function in turn. (**C**) Contact prediction derived from the transformation layer. (**D**) Mutation effects prediction by using the output of the 'Log-SoftMax' function.

non-linear activation function, and the other does not. Meanwhile, the dimension d of the embedding vector is divided according to the number of heads (h), and the length of each head is $l = d/h$ (note that d must be an integer multiple of h). And, the new **TM** becomes a double-channel and multi-head matrix with the dimension of $2 \times h \times n \times m$ (Figure 2). Then, the encoder output can be expressed as $\mathbf{O}^{\mathrm{E}} \in \mathbb{R}^{h \times l \times n}$ (dimensions have been transposed for easy expression) and the k^{th} head of the decoder input ($\mathbf{I}^{\mathrm{D}} \in \mathbb{R}^{h \times l \times m}$) can be calculated as follows:

$$\mathbf{I}^{\mathrm{D}}_{k,:,j} = \sum_i \left[\mathrm{GELU}\left(\mathbf{O}^{\mathrm{E}}\right)_{k,:,i} \cdot \mathbf{TM}_{0,k,i,j} + \mathbf{O}^{\mathrm{E}}_{k,:,i} \cdot \mathbf{TM}_{1,k,i,j} \right] \tag{3}$$

$$i = 1, 2, \ldots, n; j = 1, 2, \ldots, m; k = 1, 2, \ldots, h$$

In the actual model, the channel hyperparameters can be set to 0 (linear channel alone), 1 (non-linear channel alone) or 2 (both linear and non-linear channels).

In many cases, there is permissible substitutability between amino acids, which does not change the local or overall structural and functional properties. And, residue–residue correlation matrix will be considered to improve the prediction accuracy [9]. In this work, to model the substitutability between amino acids, we introduce a similar transformation layer with a double-channel and single-head amino acid **TM** ($a.a.$**TM** $\in \mathbb{R}^{2 \times 22 \times 22}$) between the linear layer and the 'Log-SoftMax' function.

Sequence weights

To reduce the sampling biases during the construction of MSA, we reweigh the sequence distribution by computing each sequence weight w_σ when calculating the loss function. The w_σ is defined as the reciprocal of the number of sequences within a given Hamming distance cutoff θ:

$$w_\sigma = \frac{\mathrm{scale}}{\sum_t^N I\left[D_H\left(\mathrm{x}^t, \mathrm{x}^\sigma\right) < \theta\right]} \tag{4}$$

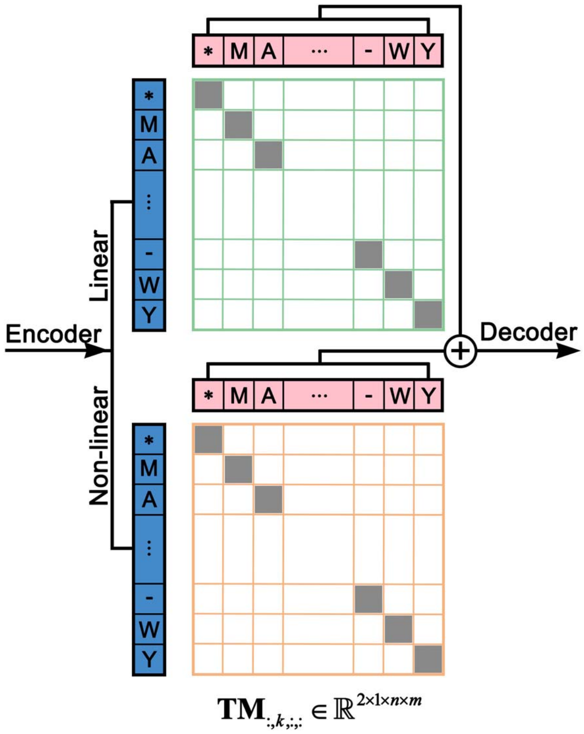

Figure 2. The kth head of the **TM** ($\mathbf{TM}_{:,k,:,:}$). It contains two channels, one linear and the other non-linear. When the source and target MSAs are the same, the diagonal parameters of **TM** are set to 0 (i.e. $\mathbf{TM}_{:,k,i,i} = 0$).

where $D_H\left(x^t, x^\sigma\right)$ is the normalized Hamming distance between the sequence σ and all sequences t; N is the number of sequences in MSA; **scale** is the scale factor and we empirically set **scale** $= N$ in this paper.

Hyperparameters

The hyperparameters in ETLD can be divided into two categories, one of which is the model hyperparameters and the other is the training hyperparameters. The model hyperparameters include the number of training epochs (e), the number of repeats training under random seeds (r), the embedding vector dimension (d), the inner-layer dimension (d_{inner}), the amino acid transformation layer channel ($c_{\text{a.a.}}$), the Hamming distance cutoff (θ), the transformation layer channel (c), the number of multi-heads (h), etc. We will adjust the model hyperparameter settings to achieve optimal results for different protein families. Instead, fixed training hyperparameters are used across all training: a cross-entropy loss function is employed as well as an Adam optimizer with a weight decay 0.01 (where $\beta_1 = 0.9$, $\beta_2 = 0.999$ and $\varepsilon = 10^{-9}$); a GELU function is applied to all activation layers and a dropout rate of 0.1 and a batch size of 64 are adopted.

Mathematical derivation for contact prediction

How to determine the residue–residue contacts **C** from **TM** is an interesting issue. To address this issue, we make some additions first. (1) As mentioned earlier, each target site is represented as the joint of the source sites with the weight factors recorded in **TM**. The weight factor can be a positive or negative value corresponding to the positive or negative correlation between two sites. Anyway, a larger absolute (or square) value implies a stronger relationship and therefore a higher probability to form a physical contact. (2) We have additionally appended the start marker for each sequence when constructing the training data, and the marker site should be site-independent, i.e. it has zero effect on the other sites. (3) In the double-channel and multi-head transformation layer, $\mathbf{TM} \in \mathbb{R}^{2\times h\times n\times m}$ contains $2 \times h$ matrices of dimension $n \times m$, whose sum can be considered as the total weight coefficients. Then, the contact prediction can be derived as follows:

$$\mathbf{C} = |\mathbf{TM}|^2 \tag{5}$$

$$\mathbf{C} = \mathbf{C} - \mathbf{C}_{:,:,0,0} \tag{6}$$

$$\mathbf{C}_{\text{sum}} = \text{Sum}\left(\mathbf{C}\right) \tag{7}$$

$$\mathbf{C}_{\text{final}} = \frac{\mathbf{C}_{\text{sum}} - \overline{\mathbf{C}_{\text{sum}}}}{\sigma_{\text{sum}}} \tag{8}$$

where $\mathbf{C}_{:,:,0,0}$ indicates the self-effect of the start marker site. **Sum** $(*)$ is the summation of the $2 \times h$ matrices, and $\mathbf{C}_{\text{sum}} \in \mathbb{R}^{n\times m}$. Equation (8) adopts a standardized process empirically, where $\overline{\mathbf{C}_{\text{sum}}}$ is the mean and σ_{sum} is the standard deviation. Residue–residue contacts are considered symmetric when the source MSA is identical to the target MSA, then:

$$\mathbf{C}_{\text{final\_sym}} = \frac{\mathbf{C}_{\text{final}} + \mathbf{C}_{\text{final}}.\text{T}}{2} \tag{9}$$

where $\mathbf{C}_{\text{final}}.\text{T}$ is the transpose of $\mathbf{C}_{\text{final}}$. Since contact prediction is a multi-solution problem with non-convex optimization, we fit the model with ten repeats from different initial conditions and consider the average performance.

Residual convolutional neural network (ResNet) for supervised contact prediction

In addition to mathematical derivation method for contact prediction, we present here a supervised ResNet that predicts the contact matrix using **TM** as input. The flowchart of ResNet is shown in Figure 3. We compose the $\mathbf{TM} \in \mathbb{R}^{2\times8\times n\times m}$ (double-channel and 8-head matrices) of 10 independent trainings into the input matrix $\mathbf{I} \in \mathbb{R}^{10\times2\times8\times n\times m}$. We first merge the channel dimension and the multi-head dimension to obtain $\mathbf{I} \in \mathbb{R}^{10\times16\times n\times m}$ and then do the following before feeding it into the model:

$$\mathbf{I} = |\mathbf{I}| \tag{10}$$

$$\mathbf{I} = \mathbf{I} - \mathbf{I}_{:,:,0,0} \tag{11}$$

$$\mathbf{I}_{:,:,:,j} = \frac{\mathbf{I}_{:,:,:,j} - \overline{\mathbf{I}_{:,:,:,j}}}{\sigma\left(\mathbf{I}_{:,:,:,j}\right)} \tag{12}$$

The input is passed through an initial convolutional layer with a kernel size of 3 × 3, followed by a two-dimensional batch normalization layer and a GELU activation function layer. Next are two sets of residual convolutional blocks, each containing four residual structures, as shown in Figure S2, available online at http://bib.oxfordjournals.org/. Then, a convolutional layer with a kernel size of 3 × 3 is connected that follows a two-dimensional batch normalization layer and a GELU activation function layer. A linear layer is used to merge the first dimension into 1. After dimensional squeezing and symmetrizing, it is activated using the 'Log-Sigmoid' function. Finally, the contact prediction matrix is obtained after another symmetrization. We also use the average result of five independent training models as the final prediction result.

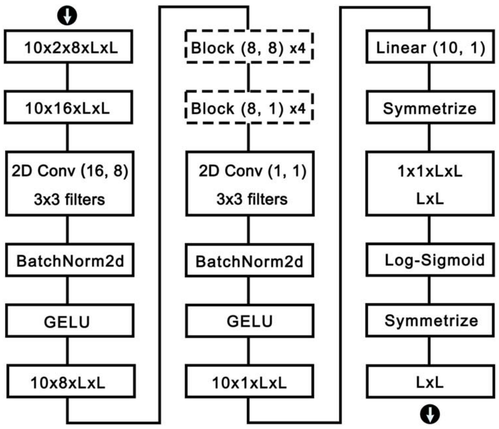

Figure 3. Residual convolutional neural network (ResNet) for supervised contact prediction. 'Conv (16, 8)' means that the convolution channel is converted from 16 to 8. The dashed box are residual convolution blocks (Figure S2 available online at http://bib.oxfordjournals.org/), with each block stacked four times.

Mutation effects prediction

ETLD will take into account the potential relationships between source sites in the reconstruction of the target sequence and further assigns a probability distribution to the selection of amino acids at each target site. Inspired by statistical physics, the evolutionary statistical energy $E(\sigma)$ can be defined based on Equation (2):

$$E(\sigma) = \sum_{i=1}^{N} \log_e \left(p_i^s(\sigma)\right) = \sum_{i=1}^{N} \mathrm{LogSoftMax}_i^s(\sigma) \qquad (13)$$

Furthermore, we can determine the statistical-energy difference ΔE between the wild-type sequence and the mutant sequence as [8, 9]:

$$\Delta E\,(\mathrm{wild-type} \rightarrow \mathrm{mutant}) = E\,(\mathrm{mutant}) - E\,(\mathrm{wild-type}) \qquad (14)$$

If $\Delta E \leq 0$, the mutant sequence has lower statistical energy related to the beneficial mutation prediction; otherwise, it is related to the deleterious mutation prediction.

Calculation of amino acid correlations

We calculate the correlations between amino acids by combining the amino acid representations in the embedding layer and the amino acid transformation layer as follows:

$$a.a.\mathbf{TM}^{\mathrm{A.A.}} = a.a.\mathbf{TM}^{\mathrm{A.A.}} - a.a.\mathbf{TM}_{:,0,0} \qquad (15)$$

$$\mathbf{R} = \left(\mathbf{R}_{\mathrm{emb}}{}^{\mathrm{T}} \cdot \mathrm{Sum}\left(a.a.\mathbf{TM}^{\mathrm{A.A.}}\right)\right)^{\mathrm{T}} \qquad (16)$$

$$\mathbf{C}_{\mathrm{A.A.}} = \mathrm{Spearman}\,(\mathbf{R}) \qquad (17)$$

where $a.a.\mathbf{TM}^{\mathrm{A.A.}} \in \mathbb{R}^{2\times20\times20}$ is the standard amino acids part in $a.a.\mathbf{TM}$, and $a.a.\mathbf{TM}_{:,0,0} \in \mathbb{R}^{2\times1\times1}$ is the autocorrelation of the start marker. $\mathbf{R}_{\mathrm{emb}}$ is the representations of the 20 standard amino acids in the embedding layer. $\mathbf{Sum}\,(*)$ function sums $a.a.\mathbf{TM}^{\mathrm{A.A.}}$ in the channel dimension, and $\mathbf{Spearman}\,(*)$ function calculates the Spearman correlation coefficients that are used to represent the correlations between amino acids ($\mathbf{C}_{\mathrm{A.A.}}$).

Performance evaluation

For comparison with the native residue–residue contacts in 3D structure, according to the standard CASP definition [36], two residues are considered to be in contact if the Euclidean distance between their C_β atoms (C_α atoms for glycine) is less than a specified threshold (8 Å). These physical contacts are divided into short, medium and long range when the sequence distance of two residues falls into [6, 11], [12, 23] and ≥ 24, respectively. The prediction accuracy is defined as the percentage of native contacts among the top L/k-predicted ($k = 1, 5$) long-range contacts, where L is the protein sequence length. Accuracy is the proportion of true positive samples in the total number of predicted positive samples, which is defined by

$$\mathrm{Acc} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \qquad (18)$$

where **TP** is the number of correctly predicted contact pairs, and **FP** is the number of incorrectly predicted contact pairs.

Datasets

PSICOV150 [15, 19, 27, 31, 37] dataset is used to evaluate the contact prediction, which contains 150 protein structures and corresponding MSAs. In addition, we selected a total of 2919 proteins with sequence length ≤ 600 from the top 3000 proteins (sorted by protein name) in the trRosetta dataset [25, 38] to form trRosettaT3000. The first 600 proteins in the trRosettaT3000 are used to train the ResNet model, and the rest are used to evaluate.

In the mutation effects prediction, we access ETLD model on the 41 high-throughput mutational scans of 33 proteins and 1 protein complex from Riesselman *et al.* [9] And, this dataset is named MUTSCAN41 in this paper. The mutation effects prediction of GEMME, DeepSequence, EVmutation and site-independent model on MUTSCAN41 are obtained directly from Riesselman *et al.* [9] and Laine *et al.* [10].

RESULTS AND DISCUSSION

Comparison of methods in contact prediction

We first investigated the effect of model hyperparameters on contact prediction by using the 10 selected proteins that were ranked in the top 10 of CCMpred prediction accuracy in the PSICOV150 dataset as test targets. We used the grid search method to evaluate the effect of each hyperparameter in a given range of values (Figure S3 available online at http://bib.oxfordjournals.org/), and the basic hyperparameter settings are shown in Table S1, available online at http://bib.oxfordjournals.org/: the contact prediction accuracy first increases steeply with the increasing e and r and then gradually converges; hyperparameters d and d_{inner} or c and h in a suitable combination can yield a higher accuracy; however, θ shows no obvious regularity, which may be caused by the fact that these MSAs have different sequence similarities.

In addition to the prediction accuracy, the number of model parameters is also noteworthy, which is related to the training time. The effects of partial hyperparameters on the total number of model parameters are shown in Figure S4, available online at http://bib.oxfordjournals.org/, and the basic hyperparameter settings are listed in Table S2 available online at http://bib.oxfordjournals.org/. It shows that most hyperparameters and the parameter number are linearly related except for the sequence length, which corresponds to a quadratic relationship. In general, the larger the hyperparameter is, the higher the prediction accuracy is, but the longer the model training time

Table 1. Contact predictions on PSICOV150 and trRosettaT3000 (long-range precision)

Method	PSICOV150		trRosettaT3000	
	L	L/5	L	L/5
CCMpred	0.39	0.69	0.28	0.47
ETLD$_{math}$	0.33	0.65	0.24	0.42
ETLD$_{ResNet}$	**0.45**	**0.71**	**0.46**	**0.71**

Note: The highest values are highlighted in bold.

will be. Therefore, we will prefer smaller hyperparameter settings with guaranteed accuracy. We can also see that ETLD and CCMpred have comparable parameter scales (Figure S4A available online at http://bib.oxfordjournals.org/), which is one of the reasons why we choose CCMpred to evaluate the contact prediction.

Table 1 compares the contact prediction performance of ETLD and CCMpred on the PSICOV150 and trRosettaT3000 datasets. The mathematically derived predictions are on average five points lower than those of the CCMpred method, but the ResNet method outperforms the CCMpred method, especially when more contacts are predicted. The accuracy of the mathematical derivation and CCMpred methods decline significantly on the trRosettaT3000 dataset, while ResNet maintains essentially the same prediction performance on both datasets, indicating the better stability of ResNet. Compared to ResNet, direct mathematical derivation methods may require more processing to obtain considerable accuracy. Overall, the transformation layer captures the coupling relationships between sites and contains rich structural information such as protein contacts, which has potential for further applications.

Comparison of methods in mutation effects prediction

We assessed ETLD's predictive power against experimental measures in [9], which contains 41 high-throughput mutational scans of 33 proteins and 1 protein complex. Among them, 38 scans are single-mutation measures, 1 is a double-mutation measure (PABP_YEAST_Fields2013-doubles), and 2 are multiple-mutation measures (HIS7_YEAST_Kondrashov2017 and parEparD_3_Laub2015_all). These scans contain mutants ranging in number from 300 to 500 000. Spearman's rank correlation between predicted and experimental values was used for accuracy scoring [8–10]. And, the related model hyperparameter settings are listed in Table S4 available online at http://bib.oxfordjournals.org/.

We ranked the accuracy of ETLD in predicting mutation effects from highest to lowest (Figure 4A). The average Spearman's rank

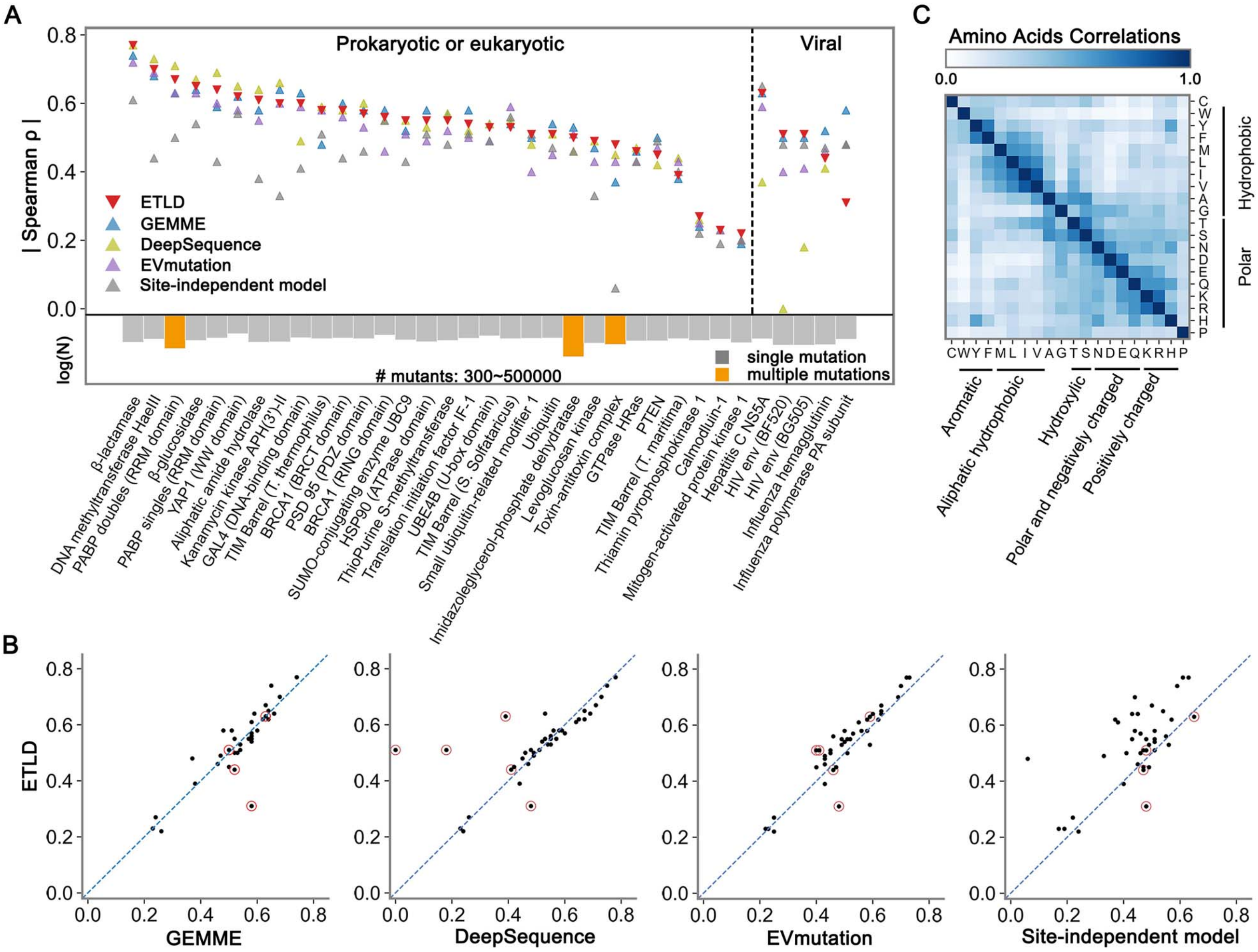

Figure 4. Comparison of mutation effects prediction performance between ETLD and other representative methods. (**A**) Spearman's rank correlation coefficient ρ between predicted and experimental measures for 35 selected high-throughput experiments. The bar plot shows the number of mutants (log value) in each scan. Non-viral proteins are on the left, and viral proteins are on the right. (**B**) Comparison of prediction accuracy of ETLD to that of the GEMME, DeepSequence, EVmutation and site-independent model (from left to right). Viral scans are emphasized with circles. (**C**) Amino acid correlations by ETLD. The closer the value is to 1, the more closely related to the two amino acids.

correlation of ETLD is $\overline{\rho} = 0.53 \pm 0.13$ compared to GEMME ($\overline{\rho} = 0.53 \pm 0.12$), DeepSequence ($\overline{\rho} = 0.51 \pm 0.17$), EVmutation ($\overline{\rho} = 0.50 \pm 0.12$) and site-independent model ($\overline{\rho} = 0.45 \pm 0.12$). In detail, the prediction of ETLD ensemble correlated equally as well or better with experimental mutation effects across a majority of the datasets (Figure 4B) compared with predictions from GEMME ($\Delta\rho_{\mathrm{ETLD-GEMME}} \geq 0$ in 23/41 datasets), DeepSequence ($\Delta\rho_{\mathrm{ETLD-DEEP}} \geq 0$ in 22/41 datasets), EVmutation ($\Delta\rho_{\mathrm{ETLD-EVmut}} \geq 0$ in 34/41 datasets) and site-independent model ($\Delta\rho_{\mathrm{ETLD-SiteInd}} \geq 0$ in 33/41 datasets). When predictions were made using optimized parameters for each dataset, ETLD can achieve an overall accuracy of $\overline{\rho} = 0.56 \pm 0.13$. More details are given in Table S5 available online at http://bib.oxfordjournals.org/, and these results suggest that ETLD obtains comparable results to currently used models in mutation effects prediction.

Correlations between amino acids

Amino acids with similar physicochemical properties are more likely to undergo substitution mutations, and this information is also implied in MSAs. GEMME and DeepSequence have shown their ability to classify amino acids based on their physicochemical properties. In this work, we calculated the amino acid correlations by combining the amino acid's representation vectors in the embedding layer and the amino acid transformation layer (see 'Materials and Methods' section).

The amino acid correlations are shown in Figure 4C. Amino acids can be directly classified into hydrophobic and polar groups. In the hydrophobic group, tyrosine (Y) and phenylalanine (F) are aromatic amino acids, and tryptophan (W) is heterocyclic one; methionine (M), leucine (L), isoleucine (I) and valine (V) are aliphatic ones. In the polar group, threonine (T) and serine (S) are hydroxyl-containing amino acids; aspartic acid (D) and glutamic acid (E) are negatively charged ones, while asparagine (N) and glutamine (Q) are their polar-derived ones, respectively; lysine (K), arginine (R) and histidine (H) are positively charged ones. In addition, histidine (H) is not only a basic amino acid, but has an imidazole structure, and is also associated with aromatic amino acids, especially tyrosine. Proline (P) is in a separate class, which does not have a prominent connection with other amino acids. Cysteine (C) contains a hydroxyl group and has a strong association with serine. Although cysteine is a polar amino acid, it shows a strong correlation with hydrophobic amino acids and a weak association with charged ones. Additionally, the Spearman correlation of our calculations with the well-known substitution matrix BLOSUM62 is 0.90, indicating that ETLD can learn the intrinsic properties of amino acids from MSAs.

Average differences between sequences in MSA

The number of effective sequences, termed as to $\mathbf{N_{eff}}$, is used to describe the depth of MSAs, which is defined as follows [39]:

$$\mathrm{N_{eff}} = \frac{1}{\sqrt{L}} \sum_{N} \frac{1}{\sum_{t}^{N} I\left[D_H\left(x^t, x^\sigma\right) < \theta\right]} \tag{19}$$

where N is the number of sequences, L is the sequence length and θ is the Hamming distance cutoff. And, we further defined $\mathbf{Diff_{MSA}}$ to characterize the average differences between sequences in MSA:

$$\mathrm{Diff_{MSA}} = {}^{\mathrm{N_{eff}}}\Big/_{\sqrt{N}} \tag{20}$$

And, we next investigate the effects of $\mathbf{Diff_{MSA}}$ and $\mathbf{N_{eff}}$ on model training.

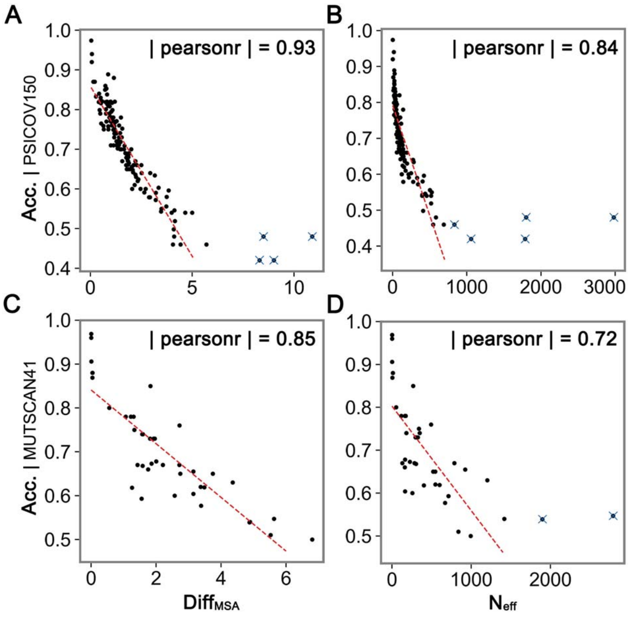

Figure 5. Pearson correlations between the training accuracy and $\mathbf{Diff_{MSA}}$ or $\mathbf{N_{eff}}$. (**A**, **C**) Pearson correlations between training accuracy and $\mathbf{Diff_{MSA}}$ on PSICOV150 and MUTSCAN41 datasets, $\theta = 0.35$. (**B**, **D**) Pearson correlations between training accuracy and $\mathbf{N_{eff}}$ on PSICOV150 and MUTSCAN41 datasets, $\theta = 0.2$. Each point represents an MSA and the points marked with crosses are excluded in the calculation of the Pearson correlation.

Model training accuracy is the average similarity between the predicted and target sequences during the training process and is used to supervise the learning process. We obtained the training accuracy for the last training epoch of each MSA and calculated its Pearson correlation with $\mathbf{Diff_{MSA}}$ and $\mathbf{N_{eff}}$ (Figure 5). It shows a more significant linear relationship between the training accuracy and $\mathbf{Diff_{MSA}}$, reaching a Pearson correlation of 0.93 on the PSICOV150 dataset and 0.85 on the MUTSCAN41 dataset. In contrast, the Pearson correlations between training accuracy and parameter $\mathbf{N_{eff}}$ are smaller, just 0.84 and 0.72, respectively. This indicates that $\mathbf{Diff_{MSA}}$ better describes the training process of the model. We also investigated the Pearson correlations between $\mathbf{Diff_{MSA}}$, $\mathbf{N_{eff}}$ and the contact prediction accuracy or mutation effects prediction accuracy, but we did not find any significant correlation (Figure S5 available online at http://bib.oxfordjournals.org/).

CONCLUSIONS

How to use machine learning or deep learning methods to analyze and apply the latent features in MSAs has attracted researchers to explore at present and for a long time to come. In this work, we proposed ETLD model to learn the potential relationships between sequence sites from MSAs. ETLD introduced a transformation layer in the original auto-encoder architecture to implement the assumption that the residue selection for each site can be derived from the sequence context. The transformation layer was further designed as a double-channel and multi-head **TM** to improve the prediction performance. Also, to consider the replaceability between amino acids, an amino acid transformation layer was introduced. We then show that the **TM** can be used to derive two-dimensional contact maps that are comparable to or exceed the CCMpred method. ETLD can also be used to model the mutational landscape and predict mutation effects. By comparing with GEMME, DeepSequence, EVmutation and site-independent model, our model has a slight advantage in terms of prediction performance. All of these results suggested that ETLD can learn the potential features of sequences from MSAs.

Although ETLD lacks in accuracy compared to the attention-based protein language models that have emerged in the past few years, it requires very few training resources and has more apparent interpretability than the attention mechanism. ETLD is a completely unsupervised model that can be used without additional supervised training for either contact prediction or mutation effects prediction. Further, the **TM** can also be used as input features for supervised learning models for a broader range of prediction tasks.

Key Points

- We proposed a novel autoencoder-based architecture, called ETLD, to learn the potential features from the MSA of homologous protein family. It was the first time to introduce a double-channel and multi-head transformation layer between the encoder and decoder, which can directly capture the coupling relationships between sites.
- A mathematical derivation procedure and a supervised residual convolutional network for predicting two-dimensional residue contact maps using the **TM** in the transformation layer were given.
- We presented how to build the mutational landscape of wild-type and its mutant sequences by ETLD, which in turn gave a theoretical basis for mutation effects prediction.
- ETLD was capable to learn the correlation between amino acids from MSAs, which was closely related to the physicochemical properties of amino acids.

SUPPLEMENTARY DATA

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

FUNDING

National Natural Science Foundation of China (No. 11774279, 11774280).

DATA AVAILABILITY

Code is available at https://github.com/xjtu-xsbsy/ETLD. And, Table S5, available online at http://bib.oxfordjournals.org/, can be found at https://github.com/xjtu-xsbsy/ETLD/tree/main/paper.

REFERENCES

1. Gobel U, Sander C, Schneider R, *et al.* Correlated mutations and residue contacts in proteins. *Proteins Struct Funct Genet* 1994;**18**: 309–17.
2. Gloor GB, Martin LC, Wahl LM, *et al.* Mutual information in protein multiple sequence alignments reveals two classes of coevolving positions. *Biochemistry* 2005;**44**:7156–65.
3. Dunn SD, Wahl LM, Gloor GB. Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction. *Bioinformatics* 2008;**24**:333–40.
4. Morcos F, Pagnani A, Lunt B, *et al.* Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proc Natl Acad Sci U S A* 2011;**108**: E1293–301.
5. Kajan L, Hopf TA, Kalas M, *et al.* FreeContact: fast and free software for protein contact prediction from residue co-evolution. *BMC Bioinform* 2014;**15**:85.
6. Liu Y, Palmedo P, Ye Q, *et al.* Enhancing evolutionary couplings with deep convolutional neural networks. *Cell Syst* 2018;**6**:65–74.e3.
7. Alexander LT, Lepore R, Kryshtafovych A, *et al.* Target highlights in CASP14: analysis of models by structure providers. *Proteins Struct Funct Bioinform* 2021;**89**:1647–72.
8. Hopf TA, Ingraham JB, Poelwijk FJ, *et al.* Mutation effects predicted from sequence co-variation. *Nat Biotechnol* 2017;**35**: 128–35.
9. Riesselman AJ, Ingraham JB, Marks DS. Deep generative models of genetic variation capture the effects of mutations. *Nat Methods* 2018;**15**:816–22.
10. Laine E, Karami Y, Carbone A. GEMME: a simple and fast global epistatic model predicting mutational effects. *Mol Biol Evol* 2019;**36**:2604–19.
11. Rives A, Meier J, Sercu T, *et al.* Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proc Natl Acad Sci U S A* 2021;**118**:e2016239118.
12. Figliuzzi M, Jacquier H, Schug A, *et al.* Coevolutionary landscape inference and the context-dependence of mutations in Beta-lactamase TEM-1. *Mol Biol Evol* 2016;**33**:268–80.
13. Gouveia-Oliveira R, Pedersen AG. Finding coevolving amino acid residues using row and column weighting of mutual information and multi-dimensional amino acid representation. *Algorithms Mol Biol* 2007;**2**:12.
14. Ekeberg M, Lovkvist C, Lan YH, *et al.* Improved contact prediction in proteins: using pseudolikelihoods to infer Potts models. *Phys Rev E* 2013;**87**:012707.
15. Jones DT, Singh T, Kosciolek T, *et al.* MetaPSICOV: combining coevolution methods for accurate prediction of contacts and long range hydrogen bonding in proteins. *Bioinformatics* 2015;**31**: 999–1006.
16. Yang J, Jin QY, Zhang B, *et al.* R2C: improving ab initio residue contact map prediction using dynamic fusion strategy and Gaussian noise filter. *Bioinformatics* 2016;**32**:2435–43.
17. He B, Mortuza SM, Wang Y, *et al.* NeBcon: protein contact map prediction using neural network training coupled with naive Bayes classifiers. *Bioinformatics* 2017;**33**:2296–306.
18. Adhikari B, Hou J, Cheng J. DNCON2: improved protein contact prediction using two-level deep convolutional neural networks. *Bioinformatics* 2018;**34**:1466–72.
19. Jones DT, Kandathil SM. High precision in protein contact prediction using fully convolutional neural networks and minimal sequence features. *Bioinformatics* 2018;**34**:3308–15.
20. Wu Q, Peng ZL, Anishchenko I, *et al.* Protein contact prediction using metagenome sequence data and residual neural networks. *Bioinformatics* 2020;**36**:41–8.
21. Pereira J, Simpkin AJ, Hartmann MD, *et al.* High-accuracy protein structure prediction in CASP14. *Proteins Struct Funct Bioinform* 2021;**89**:1687–99.
22. Brandes N, Ofer D, Peleg Y, *et al.* ProteinBERT: a universal deep-learning model of protein sequence and function. *Bioinformatics* 2022;**38**:2102–10.
23. Coudert E, Gehant S, de Castro E, *et al.* Annotation of biologically relevant ligands in UniProtKB using ChEBI. *Bioinformatics* 2022;**39**:btac793.

24. Bhattacharya N, Thomas N, Rao R, *et al.* Single layers of attention suffice to predict protein contacts. *bioRxiv* 2020; **2020**:2020.12.21.423882.
25. Rao R, Liu J, Verkuil R, *et al.* MSA transformer. *bioRxiv* 2021; **2021**:2021.02.12.430858.
26. Meier J, Rao R, Verkuil R, *et al.* Language models enable zero-shot prediction of the effects of mutations on protein function. *bioRxiv* 2021;**2021**:2021.07.09.450648.
27. Jones DT, Buchan DWA, Cozzetto D, *et al.* PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. *Bioinformatics* 2012;**28**:184–90.
28. Marks DS, Colwell LJ, Sheridan R, *et al.* Protein 3D structure computed from evolutionary sequence variation. *PloS One* 2011;**6**:e28766.
29. Kamisetty H, Ovchinnikov S, Baker D. Assessing the utility of coevolution-based residue-residue contact predictions in a sequence- and structure-rich era. *Proc Natl Acad Sci U S A* 2013;**110**:18734–4.
30. Seemayer S, Gruber M, Soding J. CCMpred-fast and precise prediction of protein residue-residue contacts from correlated mutations. *Bioinformatics* 2014;**30**:3128–30.
31. Wang S, Sun SQ, Li Z, *et al.* Accurate de novo prediction of protein contact map by ultra-deep learning model. *PLoS Comput Biol* 2017;**13**:e1005324.
32. Hanson J, Paliwal K, Litfin T, *et al.* Accurate prediction of protein contact maps by coupling residual two-dimensional bidirectional long short-term memory with convolutional neural networks. *Bioinformatics* 2018;**34**:4039–45.
33. Sim NL, Kumar P, Hu J, *et al.* SIFT web server: predicting effects of amino acid substitutions on proteins. *Nucleic Acids Res* 2012;**40**:W452–7.
34. Adzhubei I, Jordan DM, Sunyaev SR. Predicting functional effect of human missense mutations using PolyPhen-2. *Curr Protoc Hum Genet* 2013;**Chapter 7**:Unit7.20.
35. Kircher M, Witten DM, Jain P, *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet* 2014;**46**:310–5.
36. Ruiz-Serra V, Pontes C, Milanetti E, *et al.* Assessing the accuracy of contact and distance predictions in CASP14. *Proteins Struct Funct Bioinform* 2021;**89**:1888–900.
37. Michel M, Skwark MJ, Hurtado DM, *et al.* Predicting accurate contacts in thousands of Pfam domain families using PconsC3. *Bioinformatics* 2017;**33**:2859–66.
38. Su H, Wang WK, Du ZY, *et al.* Improved protein structure prediction using a new multi-scale network and homologous templates, advanced. *Science* 2021;**8**:e2102592.
39. Chen MC, Li Y, Zhu YH, *et al.* SSCpred: single-sequence-based protein contact prediction using deep fully convolutional network. *J Chem Inf Model* 2020;**60**:3295–303.