**Issue 15, 2024**

https://doi.org/10.23939/sisn2024.15.460

**UDC 004.8:004.021**

# METHOD FOR RECOGNIZING THE CHARACTERISTIC ELEMENTS OF PROTEIN SECONDARY STRUCTURE FROM THE LLM OF ITS AMINO ACID SEQUENCE

**Yaroslav Teplyi**

Lviv Polytechnic National University, ISN, Lviv, Ukraine
yaroslav.b.teplyi@lpnu.ua, ORCID: 0009-0001-5548-5530



**The spatial structure of a protein determines its biochemical properties and, consequently, its function. The same applies to elements of secondary structure, which adopt shapes of helices, coiled coils, strands, sheets and other formations in three-dimensional space. Automatic detection of such formations based on their corresponding amino acid sequences in the protein will enable the cataloging of these sequence fragments, examining and systematizing their correspondence to spatial protein formations. This, in turn, should simplify the task of searching for complementary and functional similarities among different proteins. For this purpose, a method based on covariance, autocorrelation, and spatial-spectral analysis of embeddings of their amino acid sequences has been developed and tested.**

**Key words: protein structure prediction, ESM-2 model, embeddings, autocorrelation analysis, protein secondary structure.**

## Introduction. Problem statement

The application of a large language model (Large Language Model, LLM) to the decoding results of protein amino acid sequences [1] has enabled the rapid construction of an exceptionally comprehensive database of their spatial structures. This database, named ESM-2 [2], is openly accessible, allowing a wide range of researchers in the fields of biochemistry and bioinformatics to study the functional features of proteins in relation to their spatial structure.

An important stage of such study is the analysis of the impact of individual regions of the protein on its functional features in relation to the structural characteristics of these regions, as well as the influence of a specific sequence of amino acids on the spatial structure of the corresponding region.

The use of LLM has provided a digital representation of the protein by a multidimensional vector of embeddings of its amino acids. This model numerically formulates the probability of spatial proximity of the mutual arrangement of amino acids of a protein folded in a certain way.

The reliability of the models is verified by the CASP project, within which the computed models are compared with experimentally determined structures [3].

## Analysis of recent research

Representing the protein sequence as a multidimensional vector of embeddings that encode contextual information about the sequence and its spatial structure has made it possible to apply analysis methods for detecting patterns in the folding process. Traditional approaches to embedding analysis used models such as SeqVec [4] and Spike2Vec [5], based on the LSTM architecture, to generate sequence

embeddings. Although these methods were quite effective, they rely on the architecture of artificial neural networks, which is becoming less popular due to the widespread use of neural network models known as transformers in structural analysis of proteins.

The method described in [7] is based on a combination of LSTM-based sequence embeddings and structural information obtained from PDB (Protein Data Base [6]). The results showed that combining sequence embeddings and spatial information improves prediction efficiency compared to using these embeddings separately. However, a possible issue with this approach is the use of data from different sources. The employed methods, SeqVec and Spike2Vec, despite providing a sufficiently accurate representation of the sequence, are fundamentally separated from the structural data obtained from PDB files, which could lead to worse outcomes.

The AlphaFold 2 model [8] made a significant breakthrough in the field of structural biology by proposing a solution to the protein folding problem using deep learning algorithms for accurate prediction of protein structures. This model uses evolutionary data obtained from MSA for constructing sequence and pairwise embeddings. The pairwise embeddings are then refined with the information from a database of known protein structures, which increases prediction accuracy. However, a key feature of the ESM-2 model is that it requires only sequence for structure prediction. This highlights the potential of ESM-2 embedding analysis for discovering patterns in the protein folding process. The ESM-2 model directly analyzes amino acid sequences, creating sequence and pairwise embeddings that contain information about the protein structure. Compared to the approach presented in [7], the analysis of ESM-2 model embeddings offers a potentially more efficient approach for studying protein structure, as it suggests consistency of embeddings and a correlation between sequence features and structure.

## Research purpose

This research involves a preliminary analysis of the ESM-2 model, using covariance and autocorrelation methods to study sequence embeddings and evaluate the pairwise embeddings it generates. These embeddings encapsulate both contextual and spatial information about amino acids, providing data for discovering patterns that indicate secondary and tertiary structural elements of the protein. This analytical approach aims to assess the accuracy of the ESM-2 model in predicting elements of protein structure.

## Problem formulation

Given a language model, denoted as *M*, for predicting the three-dimensional structure of a protein from the sequence *S*. The input data for the model is the sequence of amino acids in the protein $S = \{s_1, s_2, \dots, s_N\}$, where $s_i$ represents an individual element of the sequence (the letter corresponding to the amino acid), and *N* indicates the number of amino acids in the sequence. The model *M* is defined by a set of pre-learned parameters $\theta$.

After processing the sequence *S*, the model outputs a set of parameters $Y = \{y_1, y_2, \dots, y_N\}$, where *N* is the number of parameters generated by the model, and each $y_i$ corresponds to a specific parameter. The mapping of *S* to *Y* can be formally described as the function $f_\theta(S) = Y$, where $f_\theta$ summarizes the computational logic of model *M* with parameters $\theta$.

Among the set of output parameters *Y*, we focus on two specific parameters $S^s$ and $S^z$, which are the subjects of this study. These parameters are directly influenced by changes in the length of the input sequence *S*.

The parameter $S^s$ is a matrix that represents the mapping of the input sequence into a higher-dimensional space *(N, 1024)*. Hence, $S^s \in R^{N\times 1024}$, where each row in $S^s$ corresponds to an element from *S* transformed into a 1024-dimensional vector, which encodes contextual information about the given element, its properties, and its interaction with other elements in the sequence. This matrix is constructed by concatenating parameters from each internal layer of model *M*.

The parameter $S^z$ is a matrix that represents the mapping of the input sequence into a higher-dimensional space *(N, N, 128)*. Therefore, $S^z \in R^{N\times N\times 128}$, where each row in $S^z$ corresponds to an element from *S* transformed into a two-dimensional matrix *(N, 128)* of embeddings, which encode spatial information about the relationships of the given element relative to all other elements in the sequence. Thus, this matrix stores information about pairwise relationships between elements in the sequence and reflects the structure of the protein.

## Methodology

*Autocorrelation Function*

We define the autocorrelation function for a vector $V \in R^N$, where *N* is the length of the vector, using a sliding window of size *w*, and denote it as $ACF_e(V, w)$. By applying the autocorrelation function to each window *w* sliding over the input vector *V*, we obtain a matrix of autocorrelation functions $A \in R^{N-w-1\times w}$. We then normalize the matrix *A* using the normalization function *N(A)*.

$$ACF_e(V, w) = \{R_{XX}(V_{i:j+w})\} \quad (1)$$

where $R_{XX}$ is the autocorrelation function, $i \in \{1,\ldots,N\}, j \in \{1,\ldots,N\text{-}w\text{-}1\}$.

*Normalization Function*

We define the normalization function *N(X)* for the autocorrelation function that takes as input a matrix *X* and normalizes it by the maximum value of the corresponding row, resulting in a matrix *X'*, where each row is divided by its maximum value:

$$N(X) = \frac{X}{max(X_i)} \quad (2)$$

where $i \in \{1, \ldots, N\}$.

*Self-Similarity of Embeddings*

Given the matrix $S^s \in R^{N\times 1024}$ that represents the embeddings of the protein sequence, we transpose this matrix $(S^s)^T \in R^{1024\times N}$, to compute autocorrelation. Thus, computing self-similarity between corresponding dimensions of embeddings across the entire sequence. We define the self-similarity function for the sequence $ACF_p(P, w)$:

$$ACF_p(P, w) = \{ACF_e((S_i^s)^T, w)\} = A^{1024\times N-w-1\times w} \quad (3)$$

where $i \in \{1, \ldots, 1024\}$.

*Similarity of Two Sequences*

For the first sequence, we compute $ACF_p(S_1^s, w) = A_1^{1024\times N-w-1\times w}$, and for the second sequence, accordingly $ACF_p(S_2^s, w) = A_2^{1024\times M-w-1\times w}$. We then calculate the Pearson correlation coefficient between each fragment of ACF of length *w* from the first sequence relative to all fragments of ACF from the second sequence in the given dimension.

Let $v_n$ be the set of ACF fragments from $A_1$ of length *w* where $n = \{1, \ldots, N\text{-}w\text{-}1\}$, and $v_m$ be the set of fragments of length *w* from $A_2$, where $m = \{1, \ldots, M\text{-}w\text{-}1\}$.

For each fragment of $v_n$, we compute the Pearson correlation coefficient with each fragment of $v_m$. The result is a correlation matrix $Corr \in R^{N-w-1\times M-w-1}$, where each element $Corr_{nm}$ represents the correlation coefficient between fragments of $v_n$ and fragments of $v_m$.

$$Corr_{nm} = \frac{cov(v_n, v_m)}{\sigma(v_n)\times \sigma(v_m)} \quad (4)$$

By applying this function to all dimensions, we calculate a correlation matrix of two sequences $Corr \in R^{1024\times N-w-1\times M-w-1}$. The resulting matrix will contain information about the mutual similarity between the ACF of the sequences, describing the local similarity of the two protein sequences.

*Processing Pairwise Embeddings*

Given the matrix $S^z \in R^{N\times N\times 128}$, which represents the pairwise embeddings of a given protein sequence. These embeddings contain information about the protein structure, namely the contacts between amino acid residues. A common approach to comparing embeddings is the measure of cosine similarity.

Thus, we define the cosine similarity function *CosSim* for the matrix :

(5)

where and , are corresponding vectors from the matrix . By applying the function *CosSim* to all pairs of embeddings, we obtain a matrix , where *N* is the length of the input sequence.

## Experimental Results

Modern models for predicting protein structure (such as AlphaFold 2 [6]) operate on two representations of the amino acid sequence in their architecture. The first representation is based on evolutionary similarity to other sequences, generated by searching large datasets of protein sequences obtained through DNA sequencing and aligning these sequences against the target, a method also called multiple sequence alignment (MSA). Correlated changes in the positions of two amino acid residues in the MSA sequences can be used to infer which amino acid residues might contact each other. The second representation is the pairwise representation of contacts between the amino acid residues of the target sequence.

Unlike AlphaFold 2, one of the key features of the ESM-2 model is that it does not use the MSA representation, but instead predicts the spatial structure of the protein using only a single sequence.

*Analysis of Sequence Embeddings*

We analyze several pairs of protein sequences by evaluating the correlation between their embeddings using the established methodology. We experimentally verify the ability of the ESM-2 model to learn the dependencies and evolutionary context of sequences. The resulting correlation matrix was visualized as a heatmap and compared with an MSA generated using the T-Coffee program [9] to validate the results of comparing two sequences. We present three cases of protein sequence comparison:

The left part of Fig. 1 displays a heatmap that represents the comparison of two protein sequences, which was obtained by applying formula (4) that calculates the correlation between the ACF of individual regions of the two sequences. The horizontal and vertical axes represent the sequences of two proteins, in this case, denoted as *1R0R_1* and *1MEE_1*. The brighter areas indicate where the sequences are similar or identical. The diagonal line indicates sequence alignment or areas of high similarity, which may suggest similarity in function and structure of the proteins.

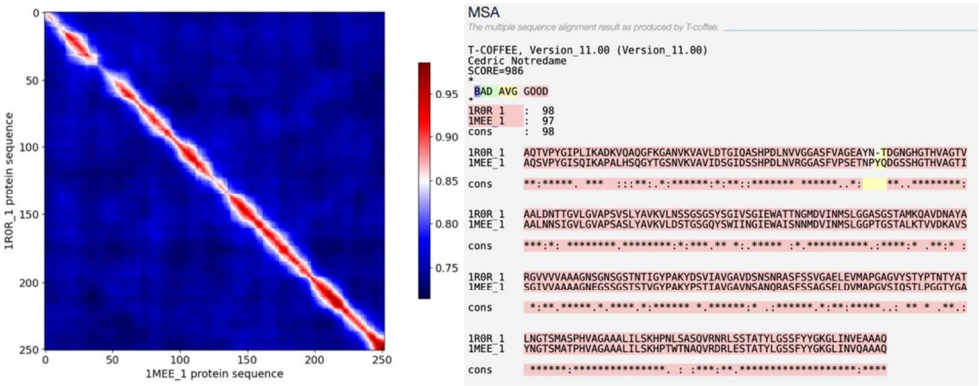


*Fig. 1. Comparison of the correlation heatmap of the 1R0R_1 and 1MEE_1 protein sequence embeddings with the MSA alignment of these sequences*

The right part of Fig. 1 shows the sequence alignment of *1ROR_1* and *1MEE_1*. Under the sequence alignment of each block, a symbolic conservation score is displayed. An asterisk '*' signifies that the amino acid is the same for the sequences at that position. A colon ':' indicates that the sequences contain different amino acids in that position, but the chemical properties are quite similar. When an amino acid is replaced with a very similar amino acid in the alignment position, it is called a conservative substitution. A space ' ' means that the amino acids are very different at that position, i.e., the substitution is non-conservative. A

period '.' indicates a semi-conservative substitution, which is something in between a conservative and a non-conservative substitution. This means that the chemical properties of the amino acids at that position are somewhat similar. The color panel indicates the quality of the alignment from *low* (green) to *high* (red).

The next example is displayed in Fig. 2 with more complex sequences, where the alignment is partial. Again, we see the main diagonal line indicating regions of similarity between these two sequences. However, unlike the first example, the sequence similarity shifts starting from the middle of the map and reappears after a certain interval of the sequence. Comparing the heatmap and the alignment of these sequences, there is a pattern between the bright areas on the map and the corresponding areas in the MSA, characterized by asterisks and colons, then the alignment shifts (the '-' symbol in the second sequence alignment), which corresponds to the shift of the diagonal on the heatmap.

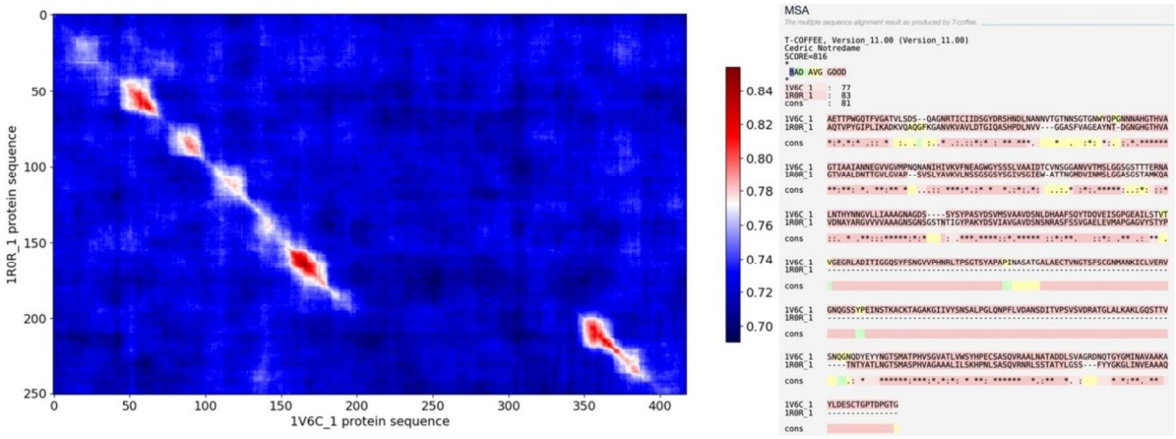


*Fig. 2. Comparison of the correlation heatmap of the 1R0R_1 and 1V6C_1 protein sequence embeddings with the MSA alignment of these sequences*

In Fig. 3, we consider a case where the similarity between sequences is very low. On the heatmap, we will see that the lack of clear patterns corresponds with sparse and less frequent matches in the MSA. This indicates that any similarities between proteins *2WP0_1* and *7AL8_1* are limited and may represent only isolated regions of common structure or function, rather than overall similarity.

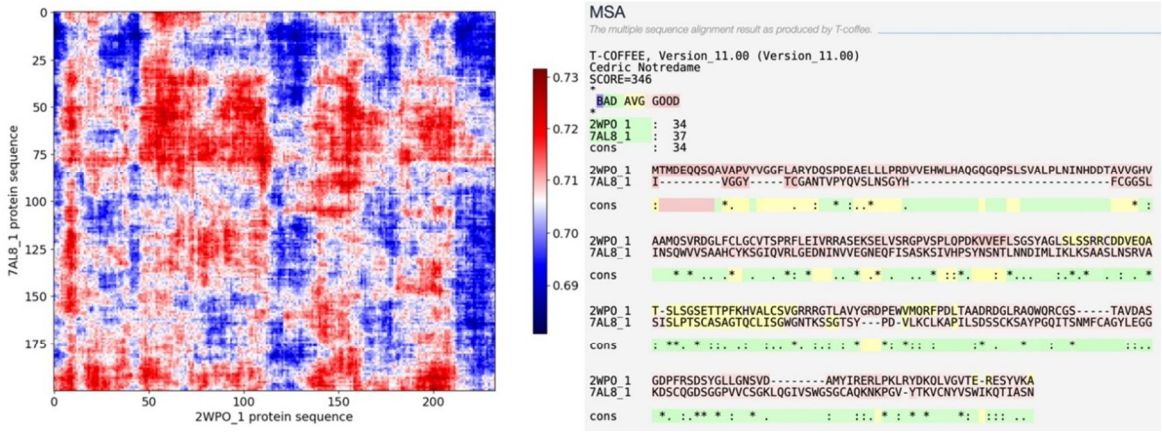


*Fig. 3. Comparison of the correlation heatmap of the 2WPO_1 and 7AL8_1 protein sequence embeddings with the MSA alignment of these sequences*

Therefore, we can conclude the correctness of the applied algorithm for analyzing sequence embeddings and that the ESM-2 model is capable of learning the characteristics and evolutionary context of the sequence without prior training on MSA representations.

*Analysis of Pairwise Embeddings*

By applying formula (5) to pairwise embeddings, we construct a contact map between the amino acid residues of the target protein, and compare it with a distance map between residues, constructed from the actual three-dimensional structure of the protein. Fig. 4 demonstrates two maps: the contact map (a) and the distance map (b). The contact map is based on the cosine similarity of pairwise embeddings, a high similarity index (close to 1) indicates that the residues are in contact and in close proximity within the protein structure, or they share a similar environment within the structure, such as being on the same surface.

The diagonal line from the top left to the bottom right corner represents the similarity of each residue to itself, which is always equal to 1. Clusters of red regions off the diagonal indicate residues that are in contact with each other, and an analysis of these regions can yield insights into the elements constituting the secondary structure of the protein.

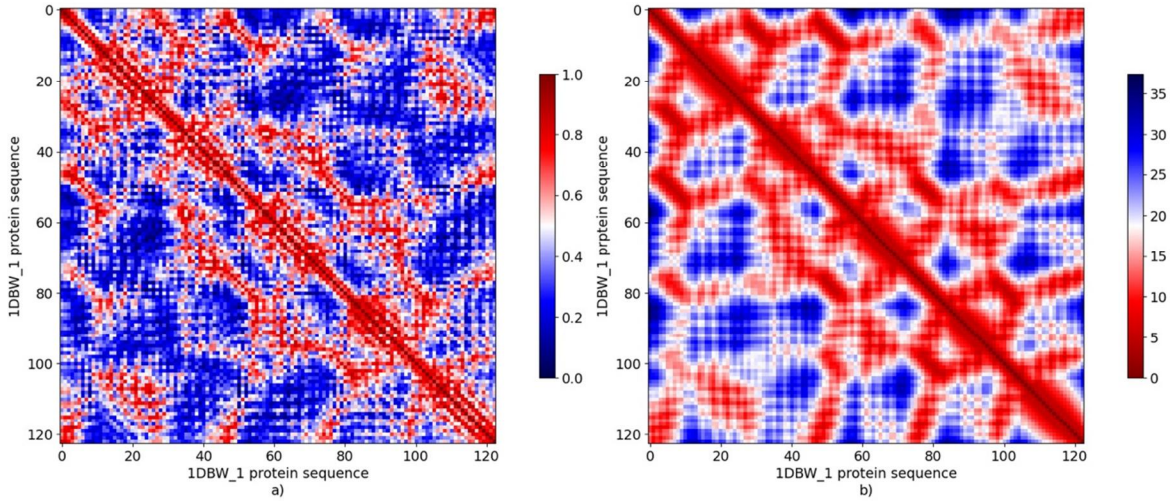


*Fig. 4. Comparison of filtered contact and distance maps of 1DBW_1 protein*

The distance map (Fig. 4. *b*) is constructed based on the actual three-dimensional structure of the protein, measuring distances between the C-alpha atoms of each pair of amino acid residues, where the distance is represented in units of length – angstroms (Å). The C-alpha atom is often used as a representative point for the location of an amino acid in structural models.

Patterns on the two maps are inversely proportional: high similarity off the diagonal on the contact map should coincide with small distances in the corresponding region on the distance map. This confirms that the representations of the ESM-2 model are accurate and identify contacts in the protein structure.

For a better visualization of patterns (regularities) on the maps, a value threshold filter is applied. Applying a threshold value of 0.76 for the contact map, so only the closest predicted contacts are displayed. A threshold of 8Å for the distance map describes the physical proximity between residues, where distances are less than or equal to this value.

On the contact map (Fig. 5. *a*) and the distance map (Fig. 5. *b*), the presence of wide regions along the diagonal line indicates regions where alpha helices can be found. Alpha helices are characterized by hydrogen bonds that lead to a helical structure, and the proximity of such helices is reflected by diagonal or anti-diagonal scattered clusters of contacts.

Conversely, narrow regions along the diagonal line represent beta-sheets or coils. Beta-sheets are a collection of beta-strands that usually run adjacent to one another, resulting in a series of neighboring parallel or antiparallel lines on the distance map. Coils, on the other hand, are less structured areas that connect secondary structure elements and can be represented as shorter diagonal lines or scattered contacts on the map.

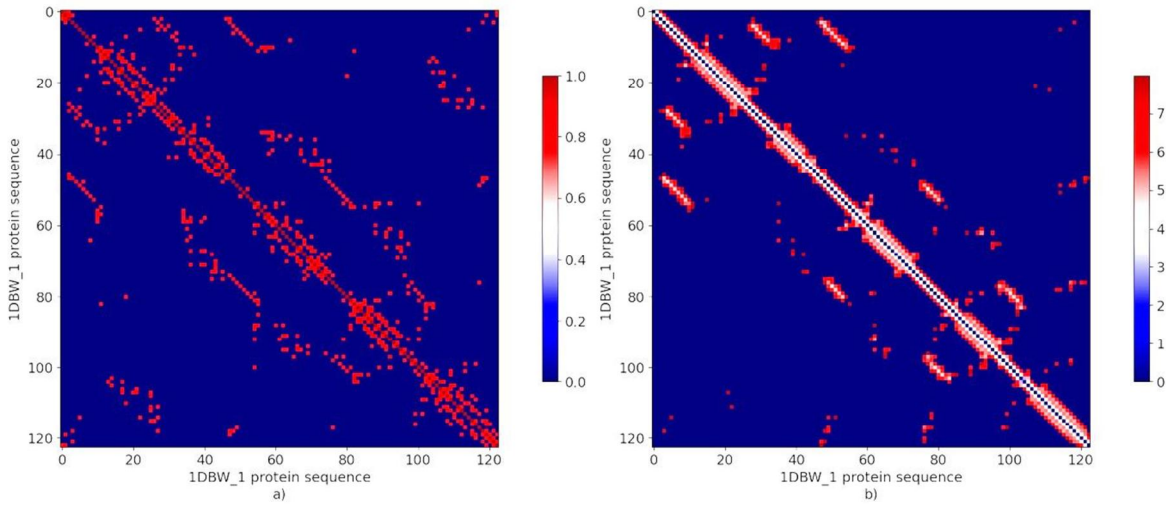


*Fig. 5. Comparison of filtered contact and distance maps of 1DBW_1 protein with highlighted patterns*

The distance map (Fig. 5. *b*) empirically confirms the predicted contacts, showing the actual proximity of each pair of residues. Here, a similar pattern of a wide diagonal line confirms the presence of alpha helices. Lines parallel to the diagonal indicate the presence of beta-strands.

Thus, the ESM2 model is capable of learning the connections between amino acid residues using pairwise sequence representation, as observed by comparing the contact map obtained from embeddings and the distance map, between which structural similarity is clearly visible.

*Spatial visualization of sequence contacts*

For a clear visualization of contacts on the distance map, we used the CMView program [10], which allows us to construct a map and display the contacts on the three-dimensional structure of the protein using the PyMOL [11] molecular visualization system.

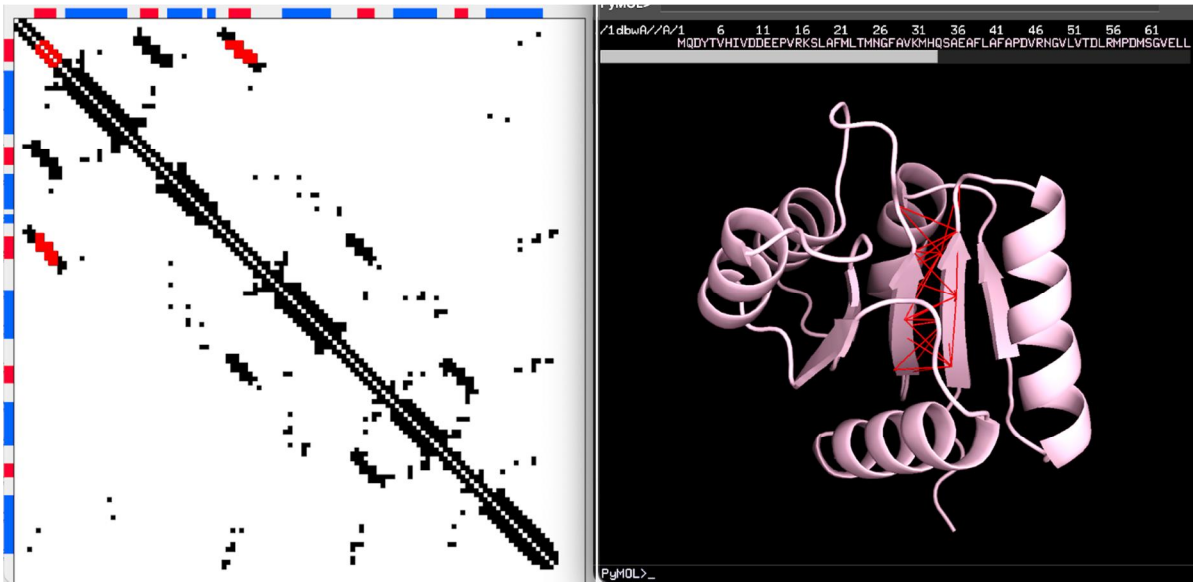


*Fig. 6. Projection of beta strands contacts on the spatial structure of the 1DBW_1 protein*

In Fig. 6, it can be seen that starting from the 6th to the 12 th amino acid, two parallel line patterns extend from the diagonal. The highlighted region on the main diagonal, and the second pattern highlighted

by the nature of contacts correspond to elements of a beta-sheet, and we can see that the representation on the three-dimensional structure of the protein corresponds to this.

Consider another example where contacts are formed by alpha helices. Contacts of this secondary structure are characterized on the diagonal as a wide area of consecutive contacts, and off the diagonal, the proximity to an alpha helix is usually reflected as a scattered cluster of contacts, due to its spatial structure.

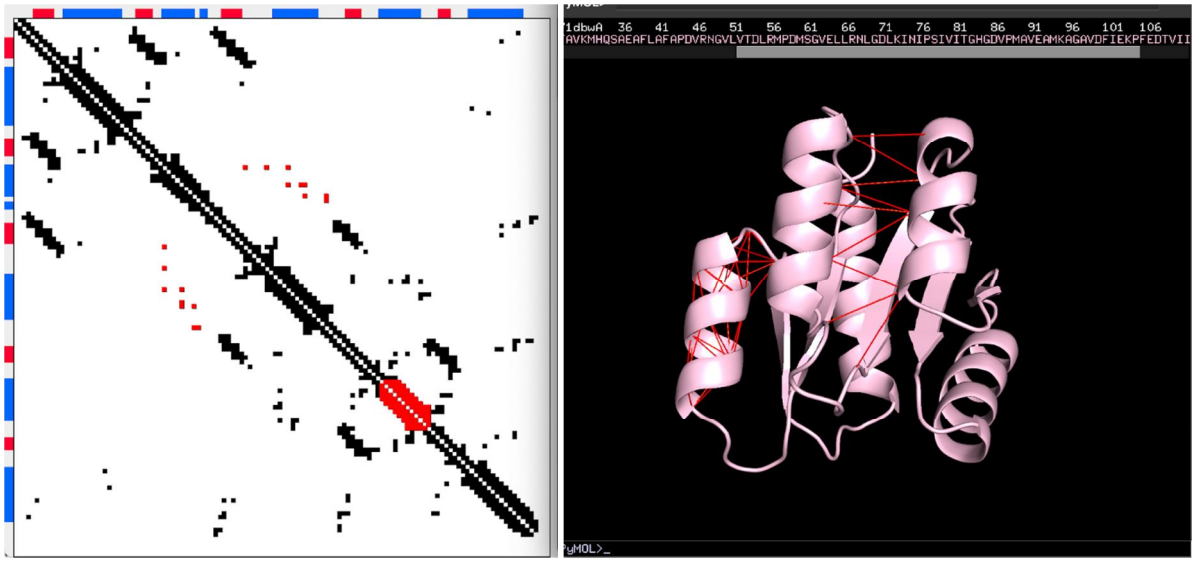


*Fig. 7. Projection of alpha helix contacts on the spatial structure of the 1DBW_1 protein*

In Fig. 7, both cases are displayed, where the highlighted region appears as a scattered cluster of contacts and corresponds to the proximity of two alpha helices. Then, a wide area of contacts on the main diagonal forms an alpha helix itself.

## Conclusions

The results of this study confirm the ability of the ESM-2 model to predict protein structures using only individual sequences. This underscores the utility of the ESM-2 model approach, which simplifies and accelerates the prediction process while maintaining high accuracy.

The application of covariance and autocorrelation analysis to ESM-2 sequence embeddings showed that the model can learn the evolutionary character of sequence development and sequence interrelationships. By evaluating the correlation between embeddings of different protein pairs, we observed clear patterns, and experimentally verified the results with MSA alignment of these sequences, which further confirmed the proposed analysis method.

The analysis of pairwise embeddings and the application of the cosine similarity measure allowed for the construction of a distance map that represents the protein's structural information. Comparing the contact map with the distance map, constructed from the actual spatial structure of the protein, confirmed the ESM-2 model's ability to accurately identify contacts between amino acid residues of a given protein sequence.

The correspondence of certain patterns on the distance map with the local three-dimensional structures, such as alpha helices and beta sheets, was visually demonstrated.

## Acknowledgment

We express our gratitude to Professor Krzysztof Fidelis from the University of California, Davis (USA), whose experience in the field of protein structures was invaluable for our research. The protein

sequence samples provided for analysis greatly contributed to the research outcomes. We are grateful for his support and collaboration.

**Literature**

1. *Wang, C., Fan, H., Quan, R., & Yang, Y. (2024). ProtChatGPT: Towards Understanding Proteins with Large Language Models. arXiv preprint arXiv:2402.09649.*

2. *Rives, A., Meier, J., Sercu, T., Goyal, S., Lin, Z., Liu, J., ... & Fergus, R. (2021). Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. Proceedings of the National Academy of Sciences, 118(15), e2016239118.*

3. *Kryshtafovych, A., Schwede, T., Topf, M., Fidelis, K., & Moult, J. (2023). Critical assessment of methods of protein structure prediction (CASP)—Round XV. Proteins: Structure, Function, and Bioinformatics, 91(12), 1539–1549.*

4. *Heinzinger, M., Elnaggar, A., Wang, Y. et al. Modeling aspects of the language of life through transfer-learning protein sequences. BMC Bioinformatics 20, 723 (2019). https://doi.org/10.1186/s12859-019-3220-8.*

5. *Ali, S., & Patterson, M., "Spike2Vec: An efficient and scalable embedding approach for COVID-19 spike sequences", IEEE International Conference on Big Data, 2021.*

6. *H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, P.E. Bourne, The Protein Data Bank (2000) Nucleic Acids Research 28: 235–242 https://doi.org/10.1093/nar/28.1.235.*

7. *Ali, S., Chourasia, P., & Patterson, M. (2023). When Protein Structure Embedding Meets Large Language Models. Genes, 15(1), 25.*

8. *Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., ... & Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. Nature, 596(7873), 583–589.*

9. *Di Tommaso, P., Moretti, S., Xenarios, I., Orobitg, M., Montanyola, A., Chang, J. M., ... & Notredame, C. (2011). T-Coffee: a web server for the multiple sequence alignment of protein and RNA sequences using structural information and homology extension. Nucleic acids research, 39(suppl_2), W13-W17.*

10. *Vehlow, C., Stehr, H., Winkelmann, M., Duarte, J. M., Petzold, L., Dinse, J., & Lappe, M. (2011). CMView: interactive contact map visualization and analysis. Bioinformatics, 27(11), 1573–1574.*

11. *The PyMOL Molecular Graphics System, Version 2.5 Schrödinger, LLC.*

## МЕТОД РОЗПІЗНАВАННЯ ХАРАКТЕРНИХ ВТОРИННИХ ЕЛЕМЕНТІВ ПРОСТОРОВОЇ СТРУКТУРИ БІЛКА ЗА LLM ЙОГО АМІНОКИСЛОТНОЇ ПОСЛІДОВНОСТІ

**Ярослав Теплий**[1]

Національний університет "Львівська політехніка",
кафедра інформаційних систем та мереж, Львів, Україна
yaroslav.b.teplyi@lpnu.ua, ORCID: 0009-0001-5548-5530



**Просторова структура білка визначає його біохімічні властивості, а отже і функцію. Те ж стосується його вторинних елементів, що набувають у трьох-вимірному просторі форми альфа-спіралей, бета-ланцюгів, петель та інших утворень. Автоматичне виявлення таких утворень за відповідними їм послідовностями амінокислот у білку дасть змогу каталогізувати ці фрагменти послідовностей, дослідити та систематизувати їх відповідність просторовим білковим утворенням, що у свою чергу має спростити задачу пошуку комплементарної і функціональної подібності різних білків. З цією метою розроблено та випробувано метод, що базується на коваріаційному, автокореляційному та просторово-спектральному аналізі ембедінгів їх амінокислотних послідовностей.**

**Ключові слова: прогнозування структури білка, модель ESM-2, ембедінги, автокореляційний аналіз, вторинна структура білка.**