Spectral graph theory for quality assessment of protein structures (Protein model quality assessment using Graph Convolutional Networks based on spectral graph theory)

Pavel Severilov[1]***, Vadim Strijov***[1]
severilov.pa@phystech.edu
[1]Moscow Institute of Physics and Technology, Russia

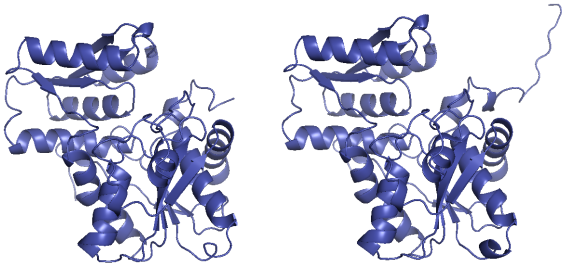
This paper investigates the problem of protein structure quality assessement. The problem is the regression of the modeled protein structures on the value of proximity $\text{CAD}_{\text{score}}$ between it and the target protein structure. This paper introduce a graphical approach to the problem in combination with the convolution transformation. It analyzes the spectrum of graph convolution and apply graph convolutional neural networks to the protein structure quality assessment problem. The authors propose a regression model and carry out experiments to test it on the data from the CASP competitions. Each molecule from the data represents three-dimensional coordinates and chemical properties of protein atoms. Thus, graphical representations of the data are built. The quality of the proposed model is comparable to the alternative models that give the best quality for the problem.

Keywords: *protein structures, graphs, graph convolutions, graph neural networks, convolutional neural networks, spectral analysis.*

1 Introduction

Proteins are the most universal macromolecules in living systems. They perform essential functions in various biological processes [1]. The shape of a protein structure determines functions it performs [1]. Understanding protein structures and their functions is essential for medical, pharmaceutical and genetic research [2]. Solving the problem of determining in which *target structure* will fold a sequence of amino acids in a protein takes a lot of time and resources.

Every two years the Critical Assessment of Protein Structure Prediction (CASP [3]) competitions are held to solve the protein structure prediction problem. The computational methods, which solve it consist of two stages: modeling the structure of a protein from their amino acid sequences and assessing the quality of prediction. This work develops only the second stage. Prediction quality is understood as the value of the proximity of *modeled* and target structures (for example, metrics $\text{CAD}_{\text{score}}$ [4], LDDT [5], GDT [6]). It is computationally expensive calculating these metrics directly, so this problem is considered as an individual task.



(*a*) Target T0861 (*б*) Modeled Atome2_-CBS_TS4

Рис. 1 Example of target and model protein structures

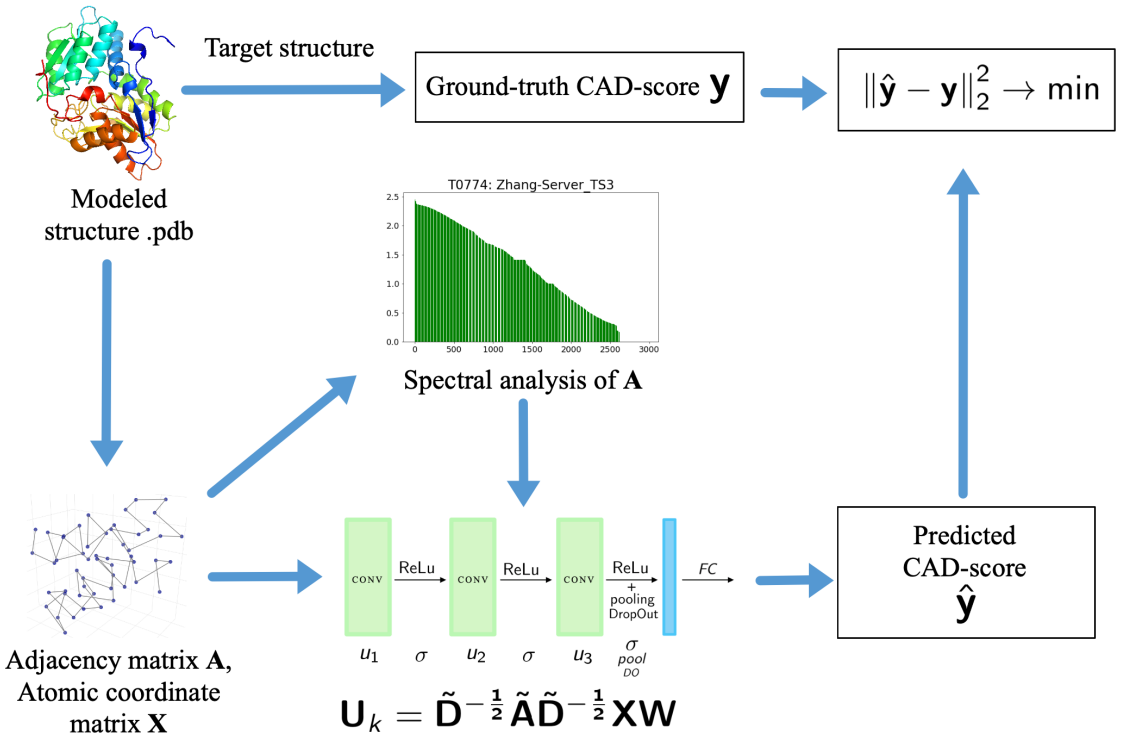


Рис. 2 Protein Structure Quality Assessment diagram

The protein structure consists of one or more chains of smaller molecules – amino acid residues. The sequence of residues S=$\{a_i\}_{i=1}^N$ represents its primary structure, where a_i is one of 22 types of amino acids. The interactions between neighboring residues and the environment determine how the chain folds into complex structures that represent the secondary structure and tertiary structure of the protein [2].

Therefore, it is necessary to take into account both spatial information about atoms, tertiary structure, and features in the form of sequences, the primary structure of the protein for the tasks of predicting and assessing the quality of protein structures. In [7, 8] the authors used LSTM or 1D-CNN to assess the quality of protein structures prediction. These methods represent proteins as a sequence with spatial features. In [9,10], the authors predict the quality of protein structure using 3D-CNN, but the primary structure of the protein is not taken into account. These papers do not take into account both the primary and tertiary structures of the proteins. We can take into account both amino acid sequences and spatial, geometric structures of proteins by using the graph representation.

The work [2] is the only one where authors use a graph representation of the protein structure to solve the problem of protein structure model quality assessement. The authors used graph neural networks based on the algorithm described in [11]. These networks show results that are superior to other state of the art methods. The model from [2] does not use convolutions. The main results in the protein quality assessement problem rely on convolutional neural networks (CNNs) [10]. In our paper, we associate the success of convolutional neural networks and graph representation of proteins to solve the protein quality assessment problem.

Figure 2 shows a diagram of solving the protein quality assessment problem using CNNs and graph representation of proteins. We define convolution on graphs by the methods of spectral

graph theory. Then we analize its spectrum. A graph convolutional neural network model for the quality assessing of protein structure prediction was built on the basis of the obtained graph convolution transform. The data for the experiments were taken from the CASP competitions of previous years. Each molecule from the data is presented in the form of information about atoms and their spatial location for target and modeled protein structures. We construct graph representation of each modeled structure based on this data – the coordinate matrix $\mathbf{X}$ and the adjacency matrix $\mathbf{A}$.

2 Problem statement for quality assessment of protein structure

There given a sample set

$$\mathfrak{D} = \{\mathbf{x}_i, y_i\}_{i=1}^{m},$$

where $\mathbf{x}_i \in \mathbb{R}^{n_i \times 3}$ are molecules. Each vector describes a set of three-dimensional coordinates of all its n_i atoms, $y_i \in \mathbb{R}$ is a proximity assessment of modeled and target protein structures. The proximity assessment is measured by various metrics: $\mathrm{CAD}_{\mathrm{score}}$ [4], LDDT [5], GDT [6]. In this paper, $\mathrm{CAD}_{\mathrm{score}}$ is selected.

2.1 Estimation of CAD score [4]

P Denotes the set of all pairs of elements of the amino acid sequence (residues) (i, j) that have a nonzero contact area $N_{(i,j)}$ in the target structure. Then, for each pair of residues $(i, j) \in P$ we calculate the contact area $M_{(i,j)}$ of the modeled structure.

For each pair of residues $(i, j) \in P$, the difference in the contact areas $\mathrm{CAD}_{(i,j)}$ is determined as the absolute difference in the contact areas between the residues i and j in the target N and the modeled structure M:

$$\mathrm{CAD}_{(i,j)} = \left| N_{(i,j)} - M_{(i,j)} \right|.$$

We take a bounded $\mathrm{CAD}_{(i,j)}$ for computational stability: $\mathrm{CAD}_{(i,j)}^{\mathrm{bounded}} = \min\left(\mathrm{CAD}_{(i,j)}, N_{(i,j)}\right)$. Thus, $\mathrm{CAD}_{\mathrm{score}}$ for the entire structure is defined as

$$\mathrm{CAD}_{\mathrm{score}} = 1 - \frac{\sum_{(i,j)\in P} \mathrm{CAD}_{(i,j)}^{\mathrm{bounded}}}{\sum_{(i,j)\in P} N_{(i,j)}}. \tag{1}$$

Figure 3 shows an example of the intersection of the target structure T0861 (yellow) and its modeled structure Atome2_CBS_TS4 (green) with $\mathrm{CAD}_{\mathrm{score}} = 0.829$.

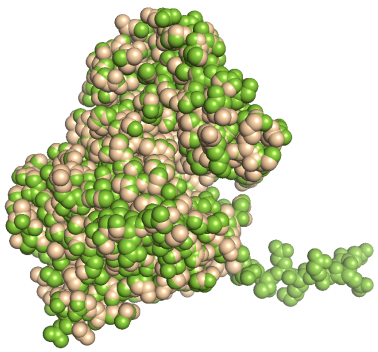


Рис. 3 Intersection of target and modeled structures

2.2 Regression of protein structures on $\mathrm{CAD}_{\mathrm{score}}$

Denote $\mathbf{X} = \bigcup_{i=1}^{m} \mathbf{x}_i$. A set of parametric models which maps $\mathbf{X}$ to a vector of $\mathrm{CAD}_{\mathrm{score}}$ predictions $\hat{\mathbf{y}} = \mathbf{f}(\mathbf{X}, \mathbf{w}) \in \mathbb{R}^m$:

$$\mathfrak{F} = \{\mathbf{f}_k \colon (\mathbf{w}, \mathbf{X}) \to \hat{\mathbf{y}} \mid k \in \mathfrak{K}\}, \tag{2}$$

where $\mathbf{w} \in \mathbb{W}$ are the model parameters. The models are taken from the class of graph convolutional neural networks.

This paper solves the regression problem for predicting the numerical value of $\mathrm{CAD}_{\mathrm{score}}$ y_i from the modeled protein structure $\mathbf{x}_i$.

The model parameters $\mathbf{w} \in \mathbb{W}$ minimize the error function. The error function is defined as

$$\mathfrak{L}(\mathbf{y}, \mathbf{X}, \mathbf{w}) = \|\hat{\mathbf{y}} - \mathbf{y}\|_2^2,$$

where $\hat{\mathbf{y}} = \mathbf{f}(\mathbf{X}, \mathbf{w})$ is the $\mathrm{CAD}_{\mathrm{score}}$ predicted by the model $\mathbf{f}$, $\mathbf{y}$ is the ground-truth $\mathrm{CAD}_{\mathrm{score}}$ given in the sample set $\mathfrak{D}$. Thus, one has to solve the optimization problem

$$\mathbf{w}^* = \operatorname*{argmin}_{\mathbf{w} \in \mathbb{W}} (\mathfrak{L}(\mathbf{w})).$$

We analyze Pearson (R) and Spearman (ρ) correlation coefficients [2, 9, 10]. It helps to understand how close the modeled structure is to the target structure based on the predictions of the regression model and the ground truth values. We calculate the Pearson (R^{target}) and Spearman (ρ^{target}) correlation coefficients for each target structure between the ground truth and predicted $\mathrm{CAD}_{\mathrm{score}}$ for modeled protein structure. Then the correlation coefficients are averaged over all T target structures. Denote $\mathbf{y}_i \in \mathbf{R}^{m_i}$ and $\hat{\mathbf{y}}_i \in \mathbf{R}^{m_i}$, respectively, the vector of ground truth values and the vector of $\mathrm{CAD}_{\mathrm{score}}$ predictions for modeled protein structures corresponding to the target structure i. Here m_i is the amount of modeled structures for the i-th target structure. Then the correlation coefficients are presented as

$$R = R\left(\mathbf{y}, \hat{\mathbf{y}}\right) = \frac{1}{T} \sum_{i=1}^{T} R_i^{\mathrm{target}} = \frac{1}{T} \sum_{i=1}^{T} \mathrm{PEARSON}\left(\mathbf{y}_i, \hat{\mathbf{y}}_i\right) \tag{3}$$

$$\rho = \rho\left(\mathbf{y}, \hat{\mathbf{y}}\right) = \frac{1}{T} \sum_{i=1}^{T} \rho_i^{\mathrm{target}} = \frac{1}{T} \sum_{i=1}^{T} \mathrm{SPEARMAN}\left(\mathbf{y}_i, \hat{\mathbf{y}}_i\right) \tag{4}$$

Here $\mathrm{PEARSON}(\cdot, \cdot)$ and $\mathrm{SPEARMAN}(\cdot, \cdot)$ are Pearson and Spearman correlations, respectively:

$$\mathrm{PEARSON}\left(\mathbf{y}_i, \hat{\mathbf{y}}_i\right) = \frac{\sum_{l=1}^{m_i} \left(\mathbf{y}_{il} - \bar{\mathbf{y}}_i\right) \left(\hat{\mathbf{y}}_{il} - \bar{\hat{\mathbf{y}}}_i\right)}{\sqrt{\sum_{l=1}^{m_i} \left(\mathbf{y}_{il} - \bar{\mathbf{y}}_i\right)^2 \sum_{l=1}^{m_i} \left(\hat{\mathbf{y}}_{il} - \bar{\hat{\mathbf{y}}}_i\right)^2}}$$

$$\mathrm{SPEARMAN}\left(\mathbf{y}_i, \hat{\mathbf{y}}_i\right) = \frac{\sum_{l=1}^{m_i} \left(\mathrm{rank}\left(\mathbf{y}_{il}\right) - \frac{m_i+1}{2}\right) \left(\mathrm{rank}\left(\hat{\mathbf{y}}_{il}\right) - \frac{m_i+1}{2}\right)}{\frac{1}{12}\left(m_i^3 - m_i\right)}$$

2.3 Making the adjacency matrices

Protein data does not contain information about the connections between atoms. Therefore, we calculate adjacency matrices $\mathbf{A}$ for modeled protein structures from the CASP data according to the following rules:

– hydrogen does not connect with hydrogen,
– an atom does not connect with hydrogen if the distance between them is at least 1.21Å,
– atoms that are far away in the sequence do not connect (residue numbers differ by more than 1),
– atoms that create disulfide bonds do not connect,
– atoms are connected, if the distance between them is $r \in (r_{\min}, r_{\max}]$, where $r_{\min} = 0.01$Å, $r_{\max} = (0.6 \cdot (\rho_{\text{atom1}} + \rho_{\text{atom2}}))^2$, ρ_{atom} – atomic radius (maximum possible $r_{\max} = 5.76$ – at $\rho_{\text{atom1}} = \rho_{\text{atom2}} = 2.0$).

The maximum possible distance between atoms, at which they can have a connection according to the presented rules for composing the adjacency matrix, is 5.76 Å. Pairwise distances between atoms in figure 4 shows that atoms marked only with the lightest yellow may have connections. Hence, the adjacency matrix is highly sparse.

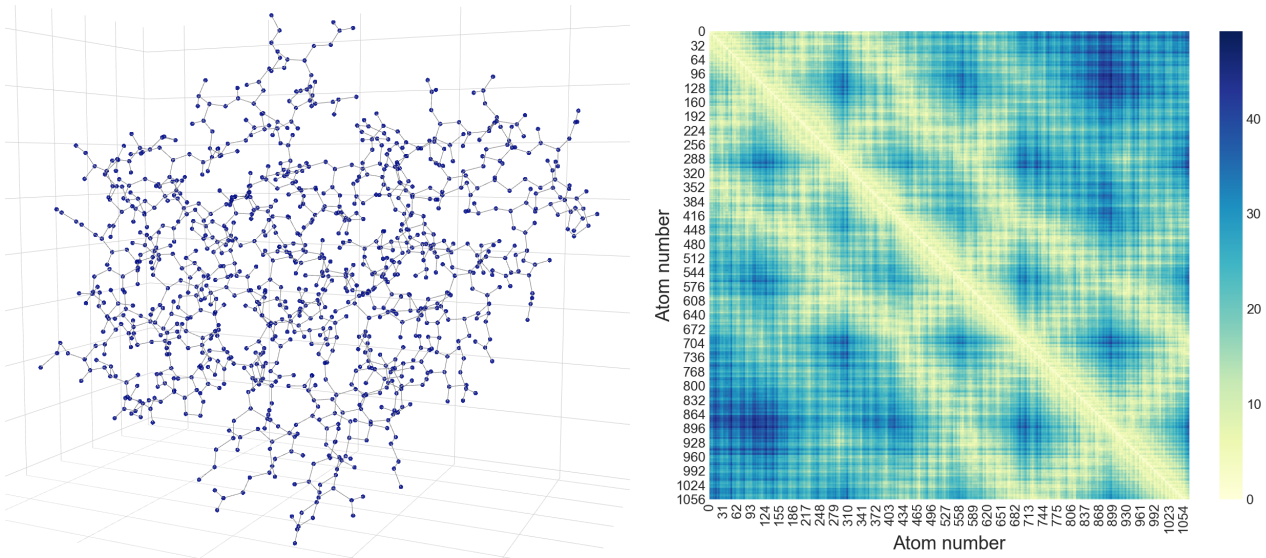


Рис. 4 Three-dimensional representation using coordinates $\mathbf{X}$ and the resulting adjacency matrix $\mathbf{A}$; pairwise distances between atoms of the modeled structure BAKER-ROSETTASERVER_TS3 of the target structure T0870 from the CASP12 dataset

3 Spectral analysis

It is required to define convolutional filters on graphs to generalize convolutional neural networks to graphs. There are two approaches: spatial and spectral [12, 13]. As shown in [14], the spatial approach does not have a general mathematical definition of translation on graphs, while the spectral method has a good mathematical foundation. Therefore, the spectral graph theory is considered.

The elements of the amino acid sequence are correspond to nodes, whose connections (edges) describe the spatial relationship between them.

In general case, the graph $\mathbf{G}$ is defined by the set $(\mathbf{V}, \mathbf{A})$, where $\mathbf{V} \in \mathbb{R}^{n \times c}$ defines the nodes of the graph. The adjacency matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ defines connections between n nodes (edges), where $\mathbf{A}_{ij}$ is the presence of a connection between nodes i and j. We define protein structure as graph using this definition of a graph. The features of amino acid sequence elements are correspond to the elements of $\mathbf{V}$. The spatial proximity between elements is corresponds to the adjacency matrix $\mathbf{A}$.

3.1 Graph convolution transform

Определение 1. *Graph Laplacian [15] is a matrix* $\mathbf{L} = \mathbf{I}_n - \mathbf{D}^{-\frac{1}{2}}\mathbf{A}\mathbf{D}^{-\frac{1}{2}}$, *where* $\mathbf{A}$ *is a graph adjacency matrix* $\mathbf{G}$, $\mathbf{D}$ *is a diagonal matrix of node degrees,* $\mathbf{D}_{ii} = \sum_j (\mathbf{A}_{ij})$ *and* $\mathbf{I}_n$ *is an identity matrix.*

The matrix $\mathbf{L}$ is real-valued symmetric positive semidefinite, therefore it can be represented as $\mathbf{L} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^\mathsf{T}$, where $\mathbf{U} = [\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_n] \in \mathbb{R}^{n\times n}$ is the matrix of eigenvectors ordered by eigenvalues, $\mathbf{\Lambda} \in \mathbb{R}^{n\times n}$ is the diagonal eigenvalue matrix (spectrum), $\mathbf{\Lambda}_{ii} = \lambda_i$. The Laplacian spectral decomposition allows one to determine the Fourier transform on graphs: its eigenvectors correspond to the Fourier decomposition modes and its eigenvalues correspond to frequencies.

Определение 2. *The Graph Fourier Transform [16] for a signal* $\mathbf{x} \in \mathbb{R}^n$ *is defined(given) by*

$$\mathscr{F}(\mathbf{x}) = \mathbf{U}^\mathsf{T}\mathbf{x} \equiv \hat{\mathbf{x}} \in \mathbb{R}^n,$$

where $\mathbf{x}$ *– vector of features of all vertices. Inverse graph Fourier transform:* $\mathscr{F}^{-1}(\hat{\mathbf{x}}) = \mathbf{U}\hat{\mathbf{x}}$.

This transformation is key in defining the graph convolution. It projects the input graph signal onto an orthonormal space, where the basis is formed by the eigenvectors of the graph Laplacian. The elements of the transformed signal $\hat{\mathbf{x}}$ are coordinates of the signal in the new space, so the input signal can be represented as $\mathbf{x} = \sum_i \hat{x}_i \mathbf{u}_i$. It is the inverse graph Fourier transform.

Theorem 1. (Convolution theorem) [17] The Fourier transform of the convolution of two signals is the component-wise product of their Fourier transforms:

$$\mathscr{F}(\mathbf{f} * \mathbf{g}) = \mathscr{F}(\mathbf{f}) \odot \mathscr{F}(\mathbf{g}).$$

Following from the theorem 1, the spectral convolution on graphs is defined for the signal $\mathbf{x}$ and the filter $\mathbf{g} \in \mathbb{R}^n$ as

$$\mathbf{x} * \mathbf{g} = \mathscr{F}^{-1}(\mathscr{F}(\mathbf{x}) \odot \mathscr{F}(\mathbf{g})) = \mathbf{U}\left(\mathbf{U}^\mathsf{T}\mathbf{x} \odot \mathbf{U}^\mathsf{T}\mathbf{g}\right) = \mathbf{U}\mathbf{g}_\theta\mathbf{U}^\mathsf{T}\mathbf{x}, \tag{5}$$

where $\mathbf{g}_\theta = \operatorname{diag}\left(\mathbf{U}^\mathsf{T}\mathbf{g}\right)$ are the spectral filter coefficients.

Spectral methods differ in selection of a filter $\mathbf{g}_\theta$. The expression (5) is computationally expensive because the spectral decomposition requires $O(n^3)$ operations and multiplication with eigenvector matrix $\mathbf{U}$ requires $O(n^2)$ operations. Chebyshev Spectral CNN (ChebNet) [18] solves these problems by approximating $\mathbf{g}_\theta$ using Chebyshev polynomials $\mathbf{T}_k(\mathbf{x})$, removing the need to calculate the Laplacian eigenvectors $\mathbf{L}$.

Определение 3. *The k-th order Chebyshev polynomials* $\mathbf{T}_k(\mathbf{x})$ *are given by the recurrence relation* $\mathbf{T}_k(\mathbf{x}) = 2\mathbf{x} \cdot \mathbf{T}_{k-1}(\mathbf{x}) - \mathbf{T}_{k-2}(\mathbf{x}), \mathbf{T}_0(\mathbf{x}) = 1, \mathbf{T}_1(\mathbf{x}) = \mathbf{x}$. *The polynomials form an orthogonal basis in* $L^2\left([-1,1], \dfrac{dx}{\sqrt{1-x^2}}\right)$.

Represent $\mathbf{g}_\theta$ as

$$\mathbf{g}_\theta = \sum_{k=0}^{K} \theta_k \mathbf{T}_k(\tilde{\mathbf{\Lambda}}),$$

where $\tilde{\mathbf{\Lambda}} = 2\mathbf{\Lambda}/\lambda_{\max} - \mathbf{I}_n \in [-1, 1]$, $\lambda_{\max}$ is the maximum eigenvalue $\mathbf{L}$, and notice that

$$\left(\mathbf{U}\mathbf{\Lambda}\mathbf{U}^\mathsf{T}\right)^k = \mathbf{U}\mathbf{\Lambda}^k\mathbf{U}^\mathsf{T}$$

(the eigenvectors form an orthonormal basis $\mathbf{U}^\mathsf{T}\mathbf{U} = \mathbf{I}$). Then get:

$$\mathbf{U}\mathbf{g}_\theta\mathbf{U}^\mathsf{T}\mathbf{x} = \mathbf{U}\left(\sum_{i=0}^{K}\theta_k\mathbf{T}_k(\tilde{\mathbf{\Lambda}})\right)\mathbf{U}^\mathsf{T}\mathbf{x} = \sum_{k=0}^{K}\theta_k\mathbf{T}_k(\tilde{\mathbf{L}})\mathbf{x}, \tag{6}$$

where $\tilde{\mathbf{L}} = 2\mathbf{L}/\lambda_{\max} - \mathbf{I}_n$.

Graph Convolutional Network (GCN) [19] use the first ChebNet approximation. Assuming $\lambda_{\max} \approx 2$ and taking the first 2 terms in the sum ($K = 1$), the expression (6) simplifies to

$$\mathbf{x} * \mathbf{g} \approx \tilde{\theta}_0\mathbf{x} + \tilde{\theta}_1\left(\mathbf{L} - \mathbf{I}_n\right)\mathbf{x} = \tilde{\theta}_0\mathbf{x} - \tilde{\theta}_1\mathbf{D}^{-\frac{1}{2}}\mathbf{A}\mathbf{D}^{-\frac{1}{2}}\mathbf{x}. \tag{7}$$

Taking $\theta = \tilde{\theta}_0 = -\tilde{\theta}_1$, obtain:

$$\mathbf{x} * \mathbf{g} \approx \theta\left(\mathbf{I}_n + \mathbf{D}^{-\frac{1}{2}}\mathbf{A}\mathbf{D}^{-\frac{1}{2}}\right)\mathbf{x}. \tag{8}$$

The operator $\left(\mathbf{I}_n + \mathbf{D}^{-\frac{1}{2}}\mathbf{A}\mathbf{D}^{-\frac{1}{2}}\right)$ can lead to computational instability and gradient explosion or vanishing, because the eigenvalues of this operator are $\in [0, 2]$. The authors of [19] propose a *renormalization trick* to solve this problem:

$$\mathbf{I}_n + \mathbf{D}^{-\frac{1}{2}}\mathbf{A}\mathbf{D}^{-\frac{1}{2}} \to \tilde{\mathbf{D}}^{-\frac{1}{2}}\tilde{\mathbf{A}}\tilde{\mathbf{D}}^{-\frac{1}{2}}, \text{ где } \tilde{\mathbf{A}} = \mathbf{A} + \mathbf{I}_n,\ \tilde{\mathbf{D}}_{ii} = \sum_j \tilde{\mathbf{A}}_{ij}.$$

There given a graph $\mathbf{G}$ and a matrix $\mathbf{X} \in \mathbb{R}^{n\times c}$ with information about nodes (n is the number of nodes and c is the number of features in each node). Based on (8) and applying the renormalization trick, the graph convolution layer is determined:

$$\mathbf{U} = \tilde{\mathbf{D}}^{-\frac{1}{2}}\tilde{\mathbf{A}}\tilde{\mathbf{D}}^{-\frac{1}{2}}\mathbf{X}\mathbf{W}, \tag{9}$$

where $\mathbf{W} \in \mathbb{R}^{c\times t}$ is the convolution parameter matrix with t filters, and $\mathbf{U} \in \mathbb{R}^{n\times t}$ is the output matrix. Figure 5 shows the scheme of a graph convolutional layer.

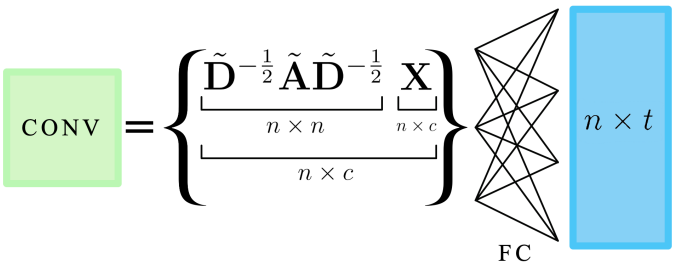


Рис. 5 A graph convolution scheme with the matrix $\mathbf{X}$ of size $n \times c$, where t is the number of filters in the convolution, FC is a fully connected layer. The blue rectangle corresponds to the output matrix of size $n \times t$

3.2 Regression model transformation function

The network architecture is constructed similar to the GCN [19] model. The expression (9) defines convolutional layers (Fig. 5). ReLu is selected as the nonlinear function σ.

The network SpectralQA (2) consists of three convolutional layers, pool max-pooling over the nodes and a fully connected FC layer. Last FC layer is the scalar multiplication with the vector $\mathbf{w}_4$. The convolution parameters t are taken equal to 64, 64, 64, respectively, for the first, second and third convolutional layers. Fig. 6 shows the scheme of the neural network tested in this paper.

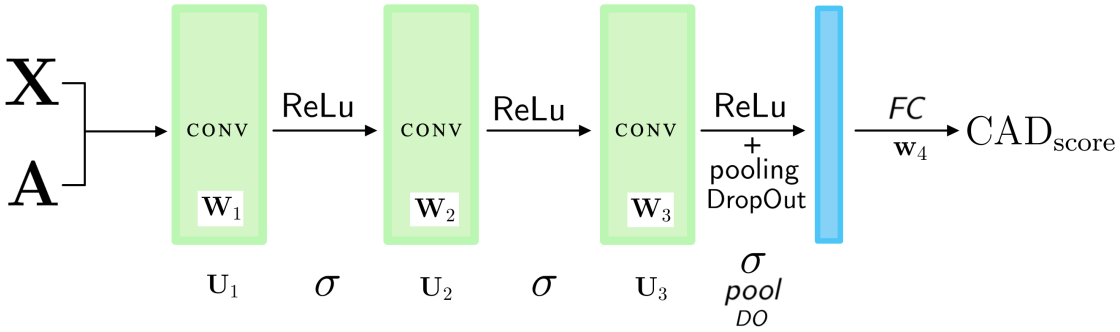


Рис. 6 Schematic representation of the architecture of the SpectralQA graph convolutional neural network used in this paper

Thus, the transformation $f : \mathbf{X} \to \mathrm{CAD}_{\mathrm{score}}$ of the resulting neural network is

$$f = \langle \mathbf{w}_4, \mathrm{DO} \circ \mathrm{pool} \circ \sigma\left(\mathbf{U}_3\right) \circ \sigma\left(\mathbf{U}_2\right) \circ \sigma\left(\mathbf{U}_1\right) \rangle, \tag{10}$$

where $\mathbf{U}_k = \tilde{\mathbf{D}}^{-\frac{1}{2}} \tilde{\mathbf{A}} \tilde{\mathbf{D}}^{-\frac{1}{2}} \mathbf{X} \mathbf{W}_k$, DO is dropout, pool is the maximum over all nodes of the graph.

4 Results

4.1 Protein datasets

The data for the experiment are taken from CASP competitions of the past years: CASP9-CASP12 (table 1). The data consist of pairs of target and modeled structures. Each structure is described by the coordinates and chemical properties of the atoms in the structure. The regerssion model (10) is trained on the data CASP9-CASP11, tested on CASP12. For training and testing procedures, we compute $\mathrm{CAD}_{\mathrm{score}}$ using (1) for all modeled structures based on target structures.

Таблица 1 Datasets of protein structures

Dataset	Target structures	Modeled structures	Split
CASP 9	117	35963	Train, Validation
CASP 10	103	15450	
CASP 11	84	12291	
CASP 12	37	5501	Test
Total	341	69205	

4.2 Eigenspace of the adjacency matrices

We perform a singular value decomposition for each adjacency matrix $\mathbf{A}$ and the matrix after the convolution $\mathbf{U}_k$ to obtain the eigenvalues of the matrices. Fig 7 and 8 show the eigenvalues for the modeled STRINGS_TS3 structure of the target T0759.

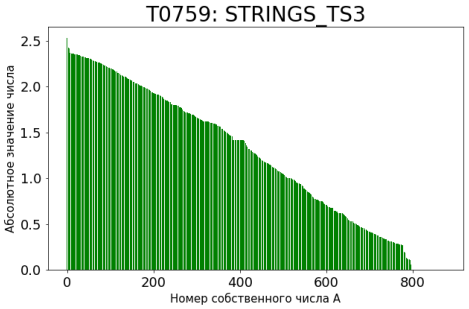


Рис. 7 The eigenvalues of the $\mathbf{A}$

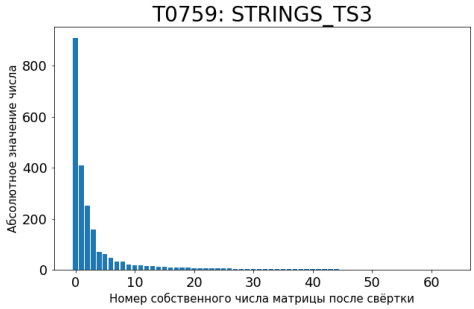


Рис. 8 The eigenvalues of the $\mathbf{U}_k$

The Broken stick model [20] is used to estimate the dimension of the matrix's eigenspace. The set of eigenvalues is compared with the thresholds: with the threshold A for a matrix $\mathbf{A}$, with the threshold U for a matrix $\mathbf{U}_k$. The j-th eigenvector $\mathbf{A}$ is saved in the list of principal components if $\lambda_j > A$. The same is done for a matrix $\mathbf{U}_k$ with the threshold U.

One modeled structure was selected at random for each target structure from the CASP11 and CASP12 datasets. We calculated eigenvalues for the matrices $\mathbf{A}$ and $\mathbf{U}_k$ for each of the selected modeled structures. The dimension of the eigenspaces of matrices is the number of eigenvalues larger than the threshold. The thresholds $U = 10$ and $A \in \{0.5, 1.0, 2.0\}$ were considered.

The results are shown in Fig. 9. Each point in the figure corresponds to one modeled structure. The dimension of the eigenspace of the matrix after passing through the convolution is compressed 50-100 times. This can be explained by the strong sparseness of the adjacency matrices of protein structures.

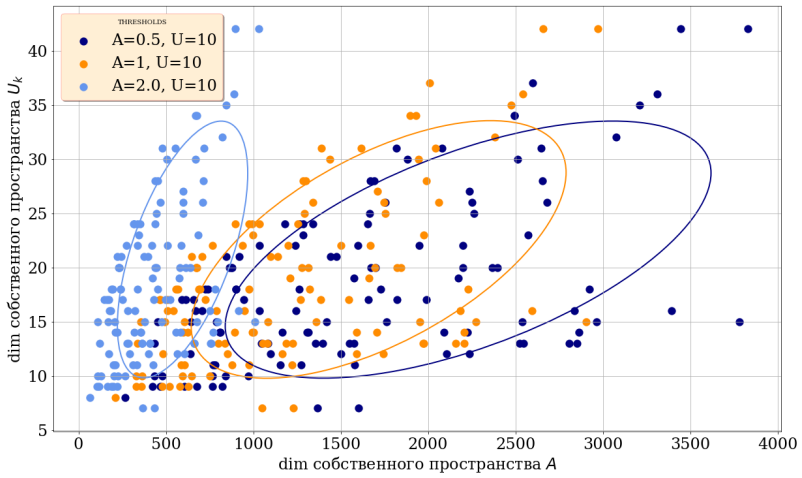


Рис. 9 Eigenspaces for the thresholds $U = 10$ and $A \in \{0.5, 1.0, 2.0\}$.

4.3 Pearson and Spearman Correlation Analysis

The averaged over T target structures Pearson (3) and Spearman (4) correlation coefficients are analyzed when training a neural network. The learning process is presented in the figures 10 и 11

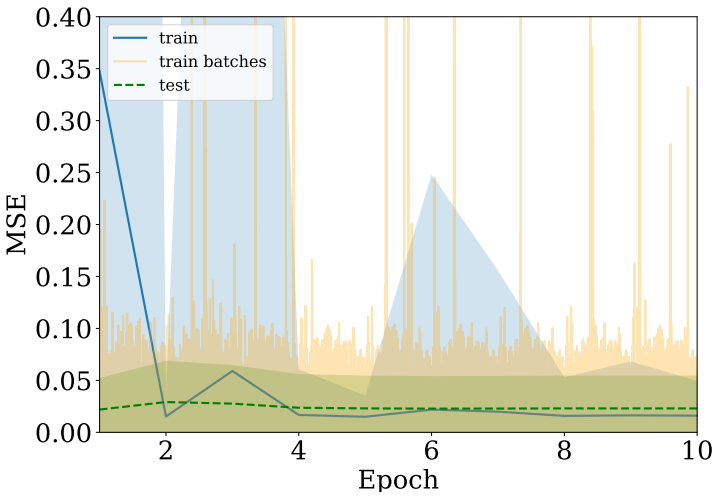


Рис. 10 MSE error plot of SpectralQA on training and test sets

The Pearson and Spearman correlation plots stabilize around one value (Figure 11). The large variance is explained by the fact that there are a lot of modeled structures with very low $\mathrm{CAD}_{\mathrm{score}}$ for some target structures in the data. The $\mathrm{CAD}_{\mathrm{score}}$ value will be equal or very close to 0 for poorly modeled structures due to the expression (1). This also explains the low value of the averaged correlation.

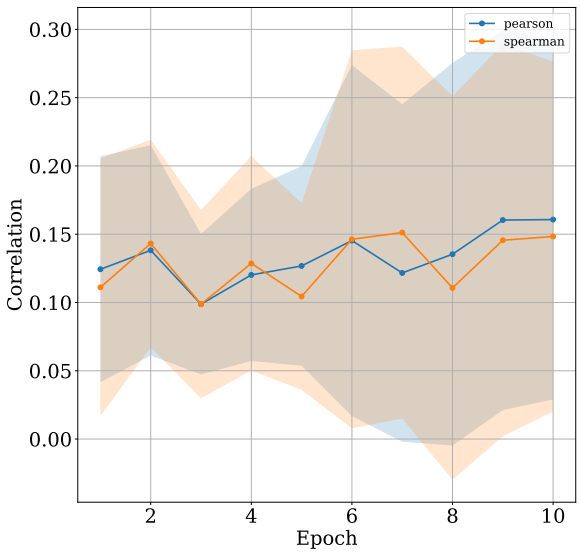


Рис. 11 Pearson and Spearman correlation during training procedure

Table 2 shows the results of testing the model on the data of the CASP12 competition. The correlation here is taken between all predictions and ground truth values. It can be seen from the table that the model from this paper gives a quality comparable to the quality of alternative models that give the best quality in the task.

Таблица 2 Comparison of Pearson and Spearman correlations on CASP12 dataset of existing modern algorithms with SpectralQA model

Method	Spearmann ρ	Pearson R
ProQ3D	0.801	0.750
VoroMQA	0.803	0.766
SBROD	0.685	0.762
Ornate	**0.828**	**0.781**
SpectralQA (this paper)	0.746	0.647

5 Conclusion

In this work, we proposed a solution to the quality assessment of protein structure prediction problem using graph convolutions. The exhaustive analysis of graph convolutions on this problem and analysis of the Pearson and Spearman correlations were carried out.The quality achieved by the proposed model is comparable to the quality of alternative models that do not use graph representation of protein structures. In further studies, we plan to use other existing improvements in spectral convolutions (CayleyNet, Adaptive Graph Convolution Network) as the basis of the network architectures. It is also proposed to take into account in the data the additional chemical properties of atoms and to take into account in the adjacency matrix not only the presence of a bond, but also the distance between atoms in the presence of a bond.

Литература

[1] Berg J.M., Tymoczko J.L., Stryer L. Biochemistry, Fifth Edition. — W.H. Freeman, 2002. — ISBN: 9780716730514. — URL: `https://books.google.ru/books?id=uDFqAAAAMAAJ`.

[2] GraphQA: Protein Model Quality Assessment using Graph Convolutional Network / Federico Baldassarre, David Menéndez Hurtado, Arne Elofsson, Hossein Azizpour. — 2019.

[3] Protein Structure Prediction Center. — `http://predictioncenter.org/`.

[4] Olechnovic Kliment, Kulberkytė Eleonora, Venclovas Ceslovas. CAD-score: a new contact area difference-based function for evaluation of protein structural models. // Proteins. — 2013. — Vol. 81 1. — P. 149–62.

[5] lDDT: a local superposition-free score for comparing protein structures and models using distance difference tests / Valerio Mariani, Marco Biasini, Alessandro Barbato, Torsten Schwede // Bioinformatics. — 2013. — Vol. 29. — P. 2722 – 2728.

[6] LGA: A method for finding 3D similarities in protein structures.

[7] Hurtado David, Uziela Karolis, Elofsson Arne. Deep transfer learning in the assessment of the quality of protein models. — 2018. — 04.

[8] AngularQA: Protein Model Quality Assessment with LSTM Networks / Matthew Conover, Max Staples, Dong Si et al. // Computational and Mathematical Biophysics. — 2019. — 01. — Vol. 7. — P. 1–9.

[9] Deep convolutional networks for quality assessment of protein folds / Georgy Derevyanko, Sergei Grudinin, Y. Bengio, Guillaume Lamoureux // Bioinformatics (Oxford, England). — 2018. — 01. — Vol. 34.

[10] Pagès Guillaume, Charmettant Benoit, Grudinin Sergei. Protein model quality assessment using 3D oriented convolutional neural networks // Bioinformatics. — 2019. — 02. — Vol. 35, no. 18. — P. 3313–3319. — http://oup.prod.sis.lan/bioinformatics/article-pdf/35/18/3313/30024731/btz122.pdf.

[11] Relational inductive biases, deep learning, and graph networks / Peter W. Battaglia, Jessica B. Hamrick, Victor Bapst et al. // ArXiv. — 2018. — Vol. abs/1806.01261.

[12] A Comprehensive Survey on Graph Neural Networks / Zonghan Wu, Shirui Pan, Fengwen Chen et al. // CoRR. — 2019. — Vol. abs/1901.00596. — 1901.00596.

[13] Graph Neural Networks: A Review of Methods and Applications / Jie Zhou, Ganqu Cui, Zhengyan Zhang et al. // CoRR. — 2018. — Vol. abs/1812.08434. — 1812.08434.

[14] Spectral networks and locally connected networks on graphs / Joan Bruna, Wojciech Zaremba, Arthur Szlam, Yann Lecun // International Conference on Learning Representations (ICLR2014), CBLS, April 2014. — 2014.

[15] Chung F. R. K. Spectral Graph Theory. — American Mathematical Society, 1997.

[16] The Emerging Field of Signal Processing on Graphs: Extending High-Dimensional Data Analysis to Networks and Other Irregular Domains. / David I. Shuman, Sunil K. Narang, Pascal Frossard et al. // IEEE Signal Process. Mag. — 2013. — Vol. 30, no. 3. — P. 83–98.

[17] Mallat Stphane. A Wavelet Tour of Signal Processing, Third Edition: The Sparse Way. — 3rd edition. — USA : Academic Press, Inc., 2008. — ISBN: 0123743702.

[18] Defferrard Michaël, Bresson Xavier, Van gheynst Pierre. Convolutional Neural Networks on Graphs with Fast Localized Spectral Filtering // Advances in Neural Information Processing Systems 29 / Ed. by D. D. Lee, M. Sugiyama, U. V. Luxburg et al. — Curran Associates, Inc., 2016. — P. 3844–3852. — URL: http://papers.nips.cc/paper/6081-convolutional-neural-networks-on-graphs-with-fast-localized-spectral-filtering.pdf.

[19] Kipf Thomas N., Welling Max. Semi-Supervised Classification with Graph Convolutional Networks // arXiv:1609.02907 [cs, stat]. — 2017. — Feb. — arXiv: 1609.02907. URL: http://arxiv.org/abs/1609.02907 (online; accessed: 2019-12-10).

[20] Cangelosi Richard, Goriely Alain. Component retention in principal component analysis with application to cDNA microarray data // Biology direct. — 2007. — 02. — Vol. 2. — P. 2.

[21] An End-to-End Deep Learning Architecture for Graph Classification / Muhan Zhang, Zhicheng Cui, Marion Neumann, Yixin Chen. — 2018.

[22] R.Evans J.Jumper J.Kirkpatrick L.Sifre T.F.G.Green C.Qin A.Zidek A.Nelson A.Bridgland H.Penedones S.Petersen K.Simonyan S.Crossan D.T.Jones D.Silver K.Kavukcuoglu D.Hassabis A.W.Senior. De novo structure prediction with deep-learning based scoring // Thirteenth Critical Assessment of Techniques for Protein Structure Prediction (Abstracts) 1-4. — 2018. — Dec. — URL: https://deepmind.com/blog/article/alphafold.