

# Quality prediction of proteins models with spherical convolutions on three-dimensional graphs.

*Nikita Pavlichenko, Sergei Grudin, Ilia Igashov.*

pavlichenko.nv@phystech.ru, sergei.grudin@inria.fr, igashov.i@yandex.ru

<sup>1</sup> Moscow Institute of Physics and Technology, Moscow, Russia

Convolutional neural networks have become very popular in recent years, and, in particular, have found widespread application in computer vision. Recently, active work has also begun on graph convolutional networks. In general, the graphs, unlike the pictures, are irregular structures, and in many tasks of learning on graphs sample objects also do not have unified topology. Therefore, the existing operations of convolution on the graphs are very much simplified, and the task of pulling on the graphs remain open in general. The purpose of this work is to research new operations of convolution on three-dimensional graphs within the framework of solving the problem of quality estimation of three-dimensional models of proteins (the problem of regression on the graph nodes).

**Key words:** *graph convolutional networks, spherical convolutions, three-dimensional graphs learning.*

## 1 Introduction

Protein molecules are an important part of any biological form. They determine cellular functions and behavior of various biological and chemical structures. It makes the discovery and prediction of proteins structure one of the most important points of medical, chemical and genetic science researches.

Molecules of proteins consist of smaller molecules called amino acids. These amino acids form a chain that is folded and placed in space. Thus, protein functions are determined by their positions in a 3D space. So, having this chain of amino acids we need to identify how they are located. There are ways to do this experimentally, but it can be time-consuming, expensive and not always possible. To solve these disadvantages, computational algorithms [1] [6] [11] were developed that generate different chain foldings. The problem is that no algorithm is the best one. Some of proteins are better modeled by one algorithm, others by others. Therefore, we are facing the problem of quality assessment (QA) of these protein models.

This problem has recently got attention from the machine learning community. Various artificial intelligence methods were applied such as neural networks [10] and support vector machines [8] [9]. More recent approaches mostly include deep learning methods [5] [4] [7] [3]. The newest approach is to use graph machine learning methods such as Graph Convolutional Networks (GCN) [2], where the protein is in some way represented as a graph. This work brings the new idea of capturing the 3D structure of this graph to improve the quality of GCN.

## 2 Problem statement

Consider a 3D model of a protein in space. The protein represents a chain of amino acids rolled up in space. Through dividing space around a protein into cells, for example, by the Voronoi method, we can get a 3D-graph, the vertexes of which are amino acids of protein and edges are carried out between those amino acids that are in adjacent cells. Denote the resulting graph by  $G = (V, E)$ , where vertexes  $V = (v_1, \dots, v_n)$  are a set of amino acids,  $E$  are edges. For the  $i$ -th vertex we denote for  $\mathcal{N}(v_i)$  the set of its neighbors in a graph  $G$  and for  $G$  an

adjacency matrix  $\mathbf{A}$ :

$$A_{ij} = \begin{cases} 1, & (v_i, v_j) \in E \\ 0, & \text{otherwise} \end{cases}.$$

Consider that each vertex  $v_i$  is described by some real  $d$ -dimensional vector of attributes  $x(v_i) = \mathbf{x}_i$ . In the simplest case, it can be a one-hot representation of an amino acid type. Using these data we will solve a regression problem: to predict for each vertex  $v_i$  a real number - its "score". In other words, how correctly it is placed in the given 3D-model in comparison with the actual conformation of this protein.

## References

- [1] Konstantin Arnold, Lorenza Bordoli, Jürgen Kopp, and Torsten Schwede. The SWISS-MODEL workspace: a web-based environment for protein structure homology modelling. *Bioinformatics*, 22(2):195–201, nov 2005.
- [2] Federico Baldassarre, Hossein Azizpour, David Menéndez Hurtado, and Arne Elofsson. Graphqa: Protein model quality assessment using graph convolutional network. 2020.
- [3] Matthew Conover, Max Staples, Dong Si, Miao Sun, and Renzhi Cao. AngularQA: Protein model quality assessment with LSTM networks. feb 2019.
- [4] Georgy Derevyanko, Sergei Grudinin, Yoshua Bengio, and Guillaume Lamoureux. Deep convolutional networks for quality assessment of protein folds. *Bioinformatics*, 34(23):4046–4053, jun 2018.
- [5] David Menéndez Hurtado, Karolis Uziela, and Arne Elofsson. Deep transfer learning in the assessment of the quality of protein models.
- [6] Jesper Lundström, Leszek Rychlewski, Janusz Bujnicki, and Arne Elofsson. Pcons: A neural-network-based consensus predictor that improves fold recognition. *Protein Science*, 10(11):2354–2362, dec 2008.
- [7] Guillaume Pagès, Benoit Charmettant, and Sergei Grudinin. Protein model quality assessment using 3d oriented convolutional neural networks. *Bioinformatics*, 35(18):3313–3319, feb 2019.
- [8] Arjun Ray, Erik Lindahl, and Björn Wallner. Improved model quality assessment using ProQ2. *BMC Bioinformatics*, 13(1):224, 2012.
- [9] Karolis Uziela, Nanjiang Shu, Björn Wallner, and Arne Elofsson. ProQ3: Improved model quality assessments using rosetta energy terms. *Scientific Reports*, 6(1), oct 2016.
- [10] Björn Wallner and Arne Elofsson. Can correct protein models be identified? *Protein Science*, 12(5):1073–1086, may 2003.
- [11] Jinbo Xu. Distance-based protein folding powered by deep learning. *Proceedings of the National Academy of Sciences*, 116(34):16856–16865, aug 2019.