# VoroCNN: Deep convolutional neural network built on 3D Voronoi tessellation of protein structures

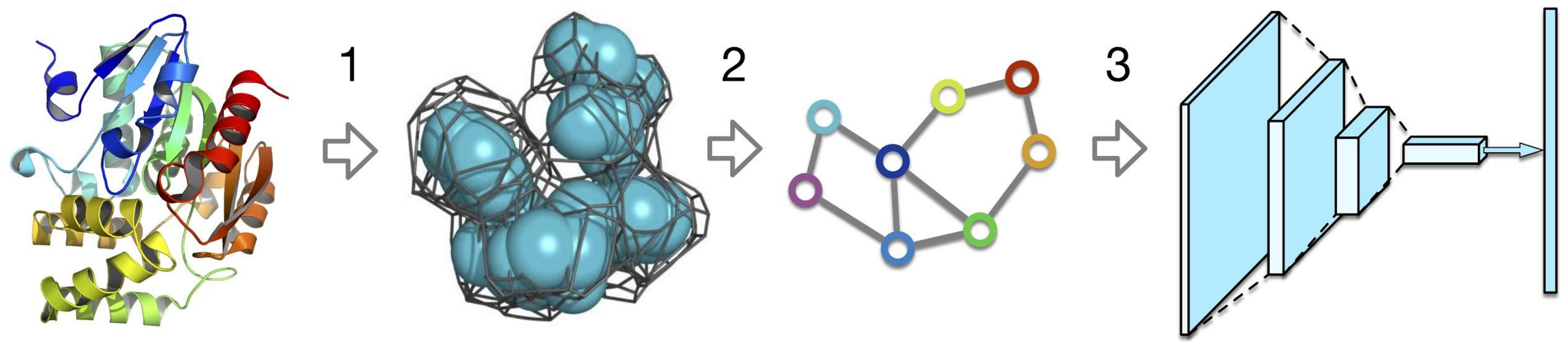
**Ilia Igashov[1,2], Kliment Olechnovic[3], Maria Kadukova[1,2], Česlovas Venclovas[3], Elodie Laine[4], Sergei Grudinin[1]**

[1]Nano-D team, Inria & LJK CNRS, Grenoble, France; [2]Moscow Institute of Physics and Technology, Dolgoprudny, Russia; [3]Institute of Biotechnology, Life Sciences Center, Vilnius University; [4]Sorbonne Université, CNRS, IBPS, Laboratoire de Biologie Computationnelle et Quantitative (LCQB)

## Introduction

For the first time we present a deep convolutional neural network (CNN) constructed on a Voronoi tessellation of 3D molecular structures. Despite the irregular data domain, our data representation allows to efficiently introduce both convolution and pooling operations of the network. We trained our model, called VoroCNN, to predict local qualities of 3D protein folds. The prediction results are competitive to the state of the art and superior to the previous 3D CNN architectures built for the same task. We also discuss practical applications of VoroCNN, for example, in the recognition of protein binding interfaces.

## Workflow



1. Voronoi tessellation of a 3D-model is computed with Voronota [1].
2. The graph is built based on Voronoi tessellation.
3. The graph neural network predicts local CAD-scores of all residues.

## 3D Graph

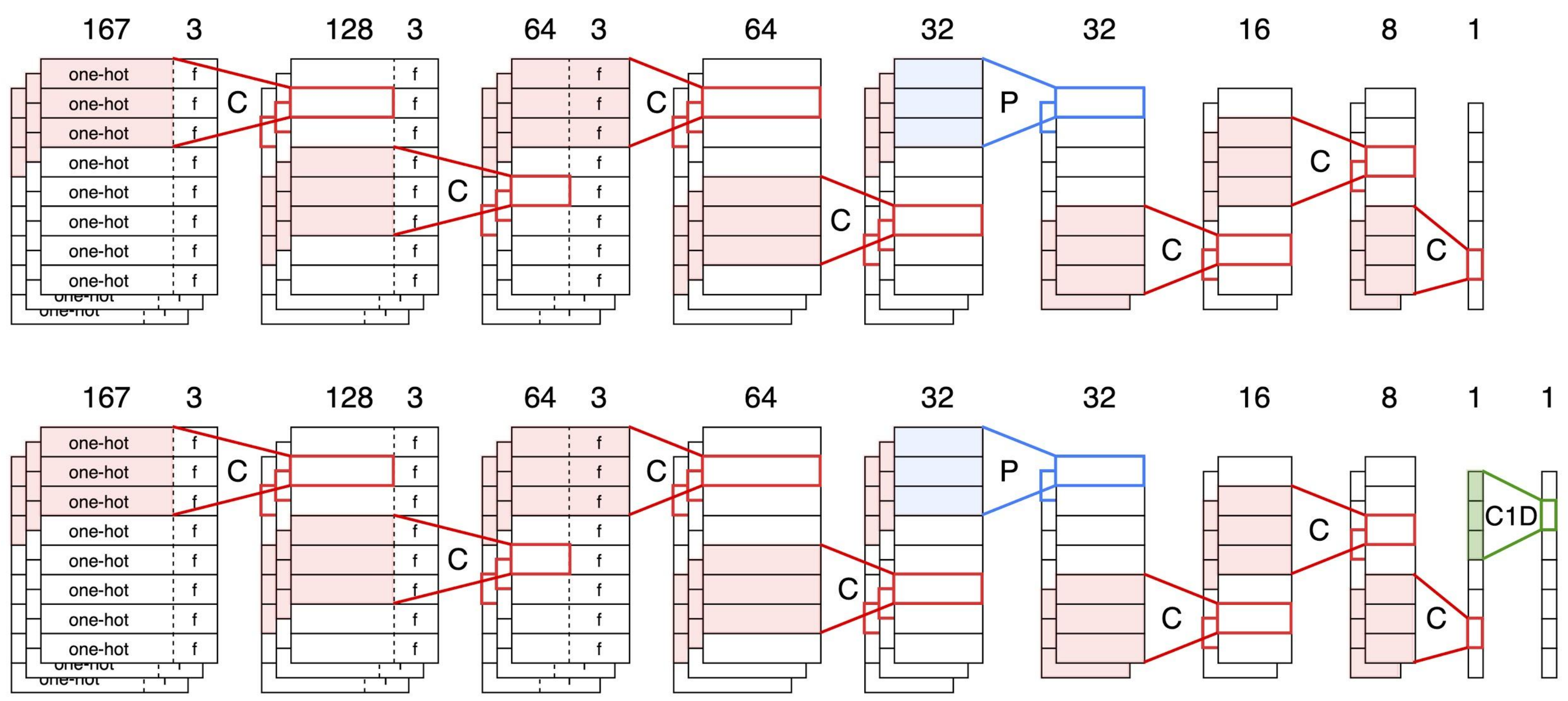
| Nodes | Edges |
|---|---|
| 1. One-hot vector representing atom type (167 in total)<br>2. Geometric features:<br>a. Surface of the contact area<br>b. Volume of the Voronoi voxel<br>c. Solvent-accessible surface area<br>d. Topological distance in the graph to the nearest solvent-accessible atom | 1. Edges corresponding to **covalent bonds** (3 independent types)<br>2. Edges corresponding to **contacts between Voronoi voxels** (6 independent types according to the sequence-separation factor) |

## References

[1] K. Olechnovič & C. Venclovas. *J Comput Chem.*, **2014**, 30;35(8):672-8.
[2] A. Hoffmann & S. Grudinin. *J Chem Theory Comput.*, **2017**, 13 (5), pp.2123-2134.
[3] I. Igashov, K. Olechnovič, M. Kadukova, Č. Venclovas, and S. Grudinin. *bioRxiv*, **2020**, doi: https://doi.org/10.1101/2020.04.27.063586.

CASP14 QA groups: **VoroCNN**, **VoroCNN-GDT**, **VoroCNN-GEMME**

## Graph Neural Networks



Convolutional layer contains trainable vectors:

$$\mathbf{W} \in \mathbb{R}^{d_{\text{in}} \times d_{\text{out}}} \quad \mathbf{W}^{\text{b}} \in \mathbb{R}^{d_{\text{in}} \times d_{\text{out}} \times d_{\text{b}}} \quad \mathbf{W}^{\text{c}} \in \mathbb{R}^{d_{\text{in}} \times d_{\text{out}} \times d_{\text{c}}}$$

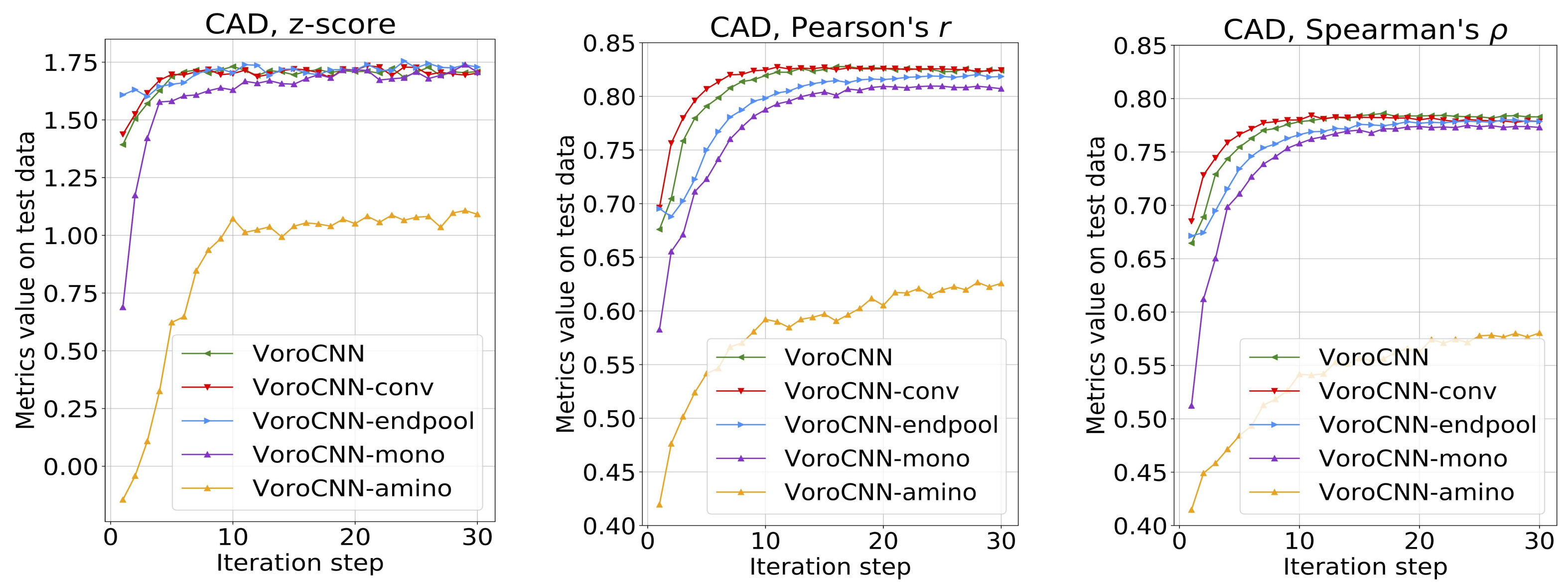
Each layer transforms input feature matrix $\mathbf{Z} \in \mathbb{R}^{N \times d_{\text{in}}}$ into $\mathbf{Z}' \in \mathbb{R}^{N \times d_{\text{out}}}$ according the following formula:

$$\mathbf{Z}' = \sigma \left[ \mathbf{Z}\mathbf{W} + \sigma_{\Sigma}(\hat{\mathbf{A}}^{\text{b}} \diamond \mathbf{Z} \diamond \mathbf{W}^{\text{b}}) + \sigma_{\Sigma}(\hat{\mathbf{A}}^{\text{c}} \diamond \mathbf{Z} \diamond \mathbf{W}^{\text{c}}) \right]$$

$$[\mathbf{X} \diamond \mathbf{Y}]_{ijk} = \begin{cases} \sum_l \mathbf{X}_{ilk} \mathbf{Y}_{lj}, & \text{if } \mathbf{Y} \text{ is an order-2 tensor (matrix)} \\ \sum_l \mathbf{X}_{ilk} \mathbf{Y}_{ljk}, & \text{if } \mathbf{Y} \text{ is an order-3 tensor,} \end{cases}$$

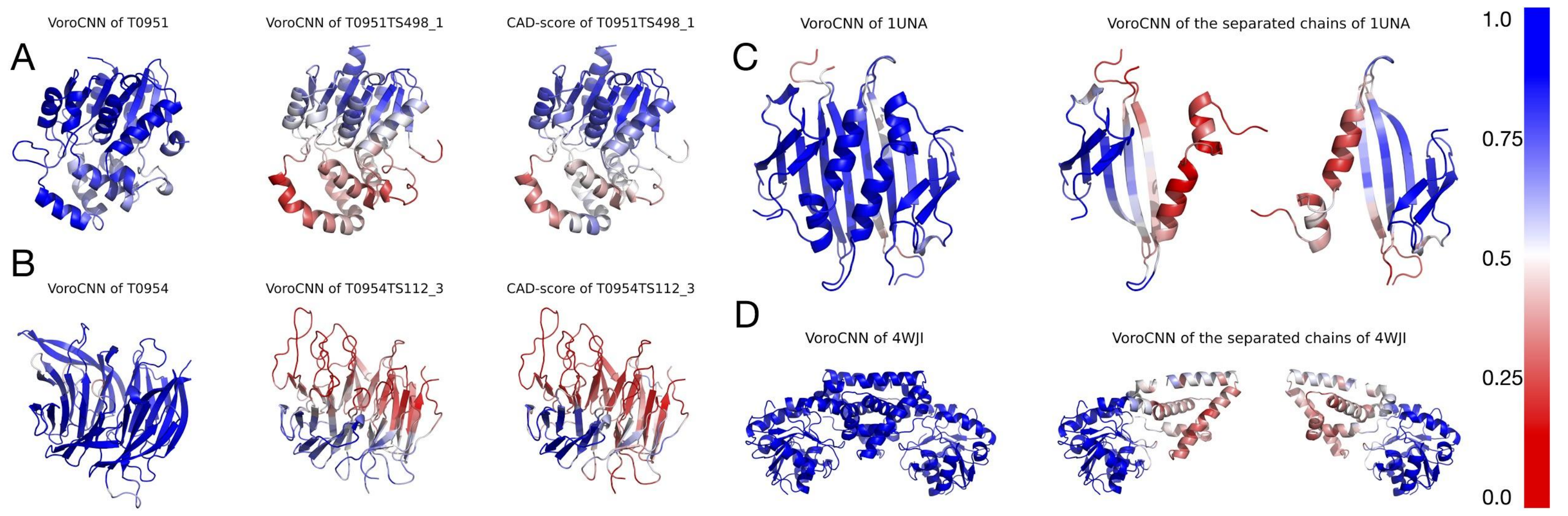
$$[\sigma_{\Sigma}(\mathbf{X})]_{ij} = \sum_k \sigma(\mathbf{X}_{ijk})$$

## Training



Training was performed on CASP[8-11], the data was augmented with near-native conformations generated with NOLB [2]. We stored and processed the adjacency matrices in the sparse format, and the whole training process was conducted in 15 parallel CPU threads.

## Experiments

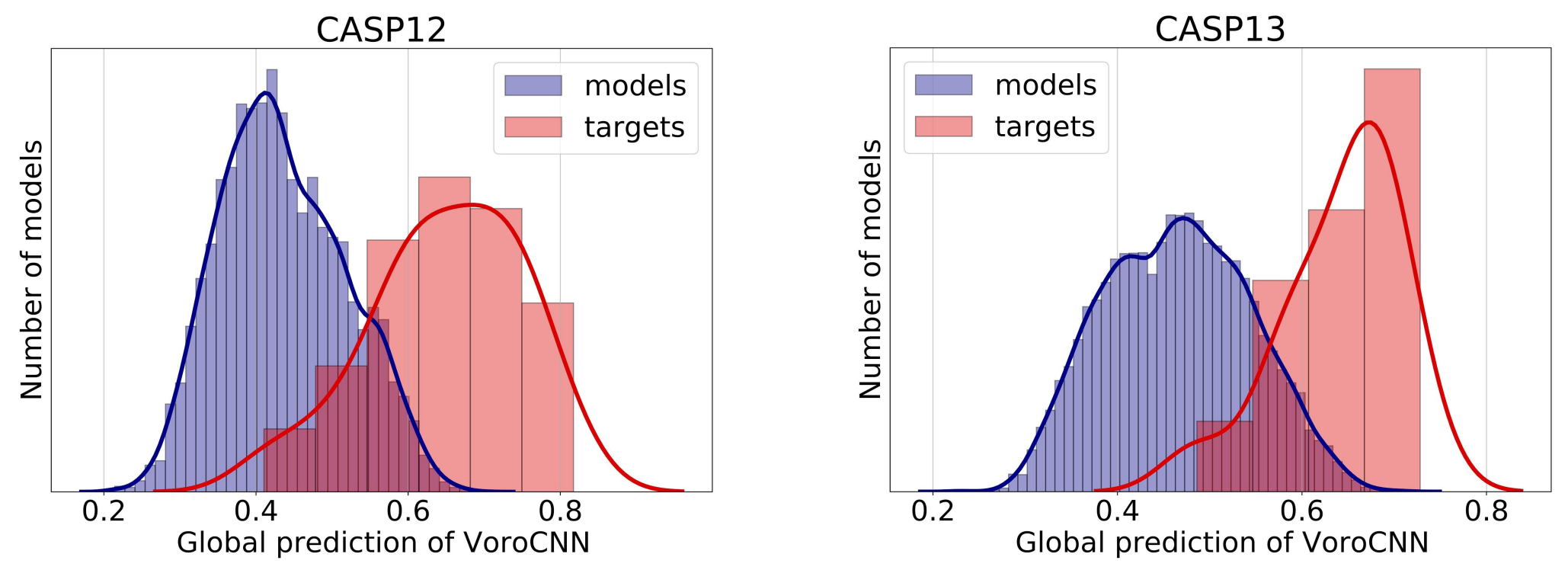


**CASP12 stage-2**

| Method | z-score | | | rank | | | Pearson's r | | | Spearman's ρ | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | CAD | IDDT | GDT-TS | CAD | IDDT | GDT-TS | CAD | IDDT | GDT-TS | CAD | IDDT | GDT-TS |
| ProQ3 | 1.670 | 1.441 | 1.141 | 11.961 | 13.500 | 25.961 | 0.801 | **0.775** | 0.692 | 0.750 | **0.734** | **0.615** |
| SBROD | 1.282 | 1.234 | 1.034 | 23.579 | 18.842 | 27.329 | 0.762 | 0.726 | **0.694** | 0.685 | 0.670 | 0.581 |
| VoroMQA | 1.410 | 1.178 | 0.761 | 17.171 | 18.158 | 36.421 | 0.803 | 0.759 | 0.638 | 0.766 | 0.725 | 0.561 |
| Ornate | 1.780 | 1.440 | 1.180 | 10.776 | 13.355 | 24.539 | **0.828** | 0.729 | 0.573 | 0.781 | 0.686 | 0.499 |
| VoroCNN | **1.871** | **1.518** | **1.191** | **9.276** | **12.000** | 25.829 | 0.817 | 0.704 | 0.565 | 0.774 | 0.682 | 0.509 |
| VoroCNN-conv | **1.857** | **1.480** | **1.271** | **8.197** | **11.447** | **19.408** | 0.823 | 0.700 | 0.563 | **0.783** | 0.679 | 0.508 |

**CASP13 stage-2**

| Method | z-score | | | rank | | | Pearson's r | | | Spearman's ρ | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | CAD | IDDT | GDT-TS | CAD | IDDT | GDT-TS | CAD | IDDT | GDT-TS | CAD | IDDT | GDT-TS |
| ProQ3 | 1.457 | 1.210 | 0.980 | 18.494 | 22.804 | 33.715 | 0.771 | 0.731 | 0.640 | 0.732 | 0.717 | 0.595 |
| ProQ3-IDDT | 1.495 | **1.257** | **1.023** | 15.620 | **20.873** | 30.044 | **0.832** | **0.782** | **0.712** | 0.792 | **0.773** | 0.666 |
| SBROD | 1.052 | 0.894 | 0.734 | 18.321 | 21.850 | **28.979** | 0.772 | 0.724 | 0.673 | 0.762 | 0.753 | **0.670** |
| VoroMQA-B | 1.363 | 1.185 | 0.976 | 21.513 | 23.696 | 33.816 | 0.802 | 0.776 | 0.661 | 0.762 | 0.731 | 0.608 |
| Ornate | 1.410 | 1.134 | 0.843 | 19.051 | 26.525 | 40.127 | 0.814 | 0.752 | 0.606 | 0.781 | 0.714 | 0.577 |
| VoroCNN | **1.581** | 1.149 | 0.948 | **13.625** | 27.042 | 36.333 | 0.763 | 0.671 | 0.541 | 0.728 | 0.630 | 0.531 |
| VoroCNN-conv | **1.508** | 1.219 | 1.004 | **15.297** | 21.810 | 29.778 | **0.832** | 0.756 | 0.648 | **0.798** | 0.734 | 0.610 |



Distribution of VoroCNN scores on target structures and models from CASP12 (left) and CASP13 (right).

## Conclusion

We demonstrate the applicability of learning on 3D Voronoi tessellations using graph convolutional networks. Our results confirm a high potential of using 3D tessellation and graph representation in general in various learning tasks in structural bioinformatics. This work also illustrates a potential of methods that predict local folding accuracies for various structural bioinformatics applications. Indeed, we have demonstrated that VoroCNN can highlight structural inaccuracies in protein models, and can also distinguish protein binding interfaces.