Protein-Protein Binding Affinity

Alen Aliev

MIPT

Research Goals

We analyze protein complexes interactions.

Our goal is to provide a model for predicting binding affinity of complexes.

We suggest using deep-learning methods, i.e. Graph Neural Networks and CNN.

Protein Interaction Problem

Figure 1: Protein Complex Example

$\mathcal{A} = \{Ala, Arg, Asn, \ldots, Val\}$
$a_i \in \mathbf{R}^3$ - atom
$PR = \{a_i\} \in \mathbf{R}^{3^*}$ - single protein
$PC = (PR_0, PR_1)$ - protein complex

Literature

- K Yugandhar and M Michael Gromiha. Computational approaches for predicting binding partners, interface residues, and binding affinity of protein–protein complexes. Springer, 2017
- Anna Vangone and Alexandre MJJ Bonvin. Contacts-based prediction of binding affinity in protein–protein complexes, 2015.
- Iain H Moal, Rudi Agius, and Paul A Bates. Protein–protein binding affinity prediction on a diverse set of structures. Bioinformatics, 2011.

Dataset Description

Dataset consists of protein chains, taken from pdbbind.org

Our hypothesis is that the interacting atoms are the ones selected with KNN method.

Two chains are considered interacting if they maximize number of atoms-neighbours.

Model Criteria

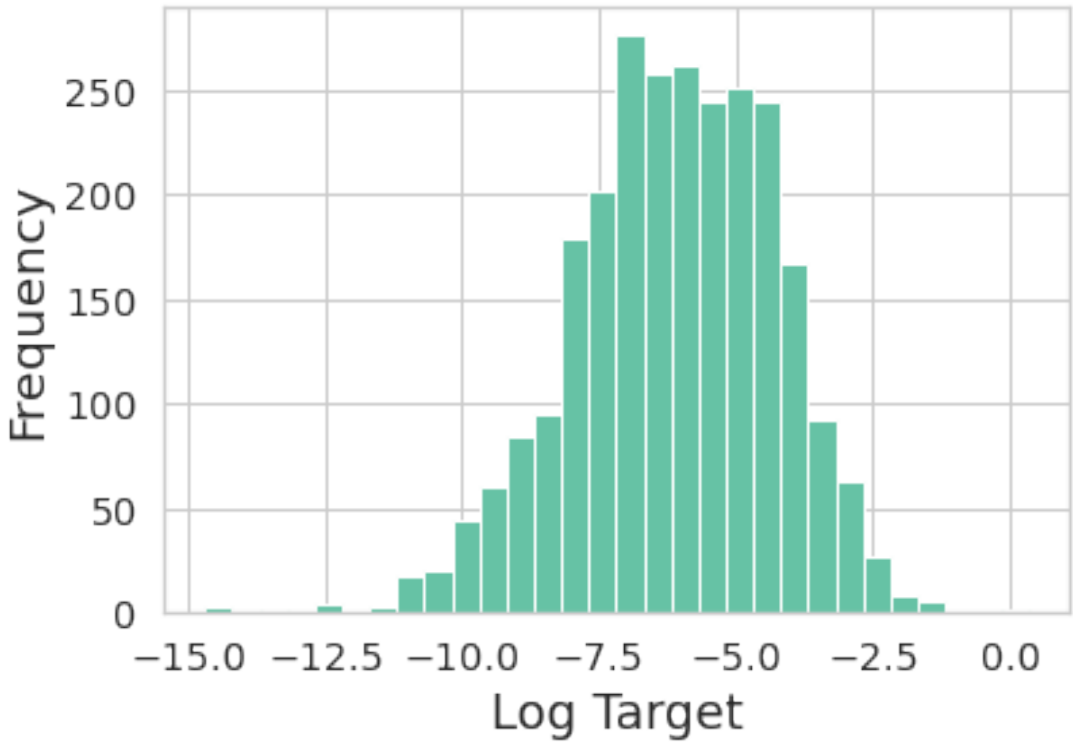

Figure 2: Log of Target Distribution

We predict logarithm of target as it is visually close to normal sample.
We benchmark models against existing state-of-the-art solutions using MSE metric.
Sanity check for us is batch accuracy

Problem Solution: 3D grid

We define C_{PR_0,PR_1} as center of mass of all interacting atoms in chains.

Each complex is now a 3D grid centered around C_{PR_0,PR_1}.

This grid is fitted to 3D CNN.

Problem Solution: Graph

Consider $\mathrm{V} = PR_0 \cup PR_1$

We say that $(a_i, a_j) \in \mathrm{E} \leftrightarrow a_j$ is in K Nearest Neighbours of a_i.

This graph representation is fitted to Graph Neural Network.

Convergence

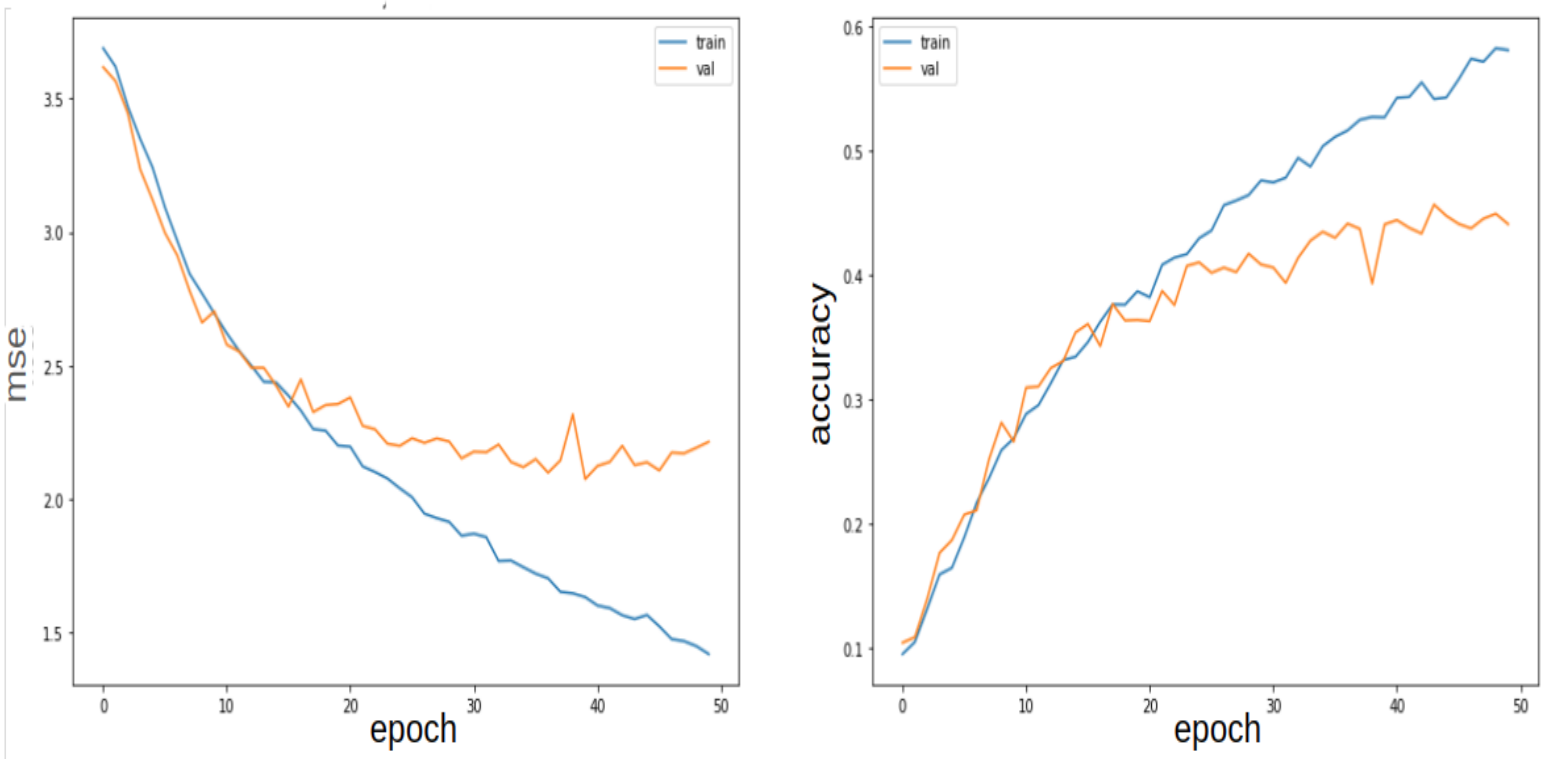

Figure 3: CNN performance

Around 50 epochs is sufficient to fit the model according to stoppage criterion.

Error Analysis

Test metrics		
Model	MSE	Accuracy
3D-CNN, VGG based	0.8715	0.72
Graph NN	0.9	0.71
CatBoost on manual features	1.2	0.53
Basic Model	4.8	0.172

Conclusion

- We provide two different interpretations of protein complex
- We conduct a computational experiment proving our models are better performing than existing solutions according to our metric set.