Geometric Deep Learning for Protein-Protein Binding Affinity Prediction

- Alen Aliev, aliev.ae@phystech.edu
- Ilya Igashov, dummy@phystech.edu
- Arne Schneuing, dummy@phystech.edu

2022

Contents

Abstract

Proteins are involved in several biological reactions by means of interactions with other proteins or with other molecules such as nucleic acids, carbohydrates, and ligands. Among these interaction types, protein–protein interactions (PPIs) are one of the key factors as they are involved in most of the cellular processes.

In this work we aim to compile a novel benchmark of PPIs with known binding affinity values from refined data and benchmark the resulting deep learning geometry method against existing state-of-the-art approaches.

Introduction

The binding of two proteins is a reversible and rapid process in an equilibrium that is governed by the law of mass action. Binding affinity is the strength of the interaction between two (or more than two) molecules that bind reversibly (interact). It is translated into physico-chemical terms in the dissociation constant Kd, the latter being the concentration of free protein at which half of all binding sites of the second protein type are occupied [2].

Predicting the affinity of protein–protein complexes has been a topic of active research for more than two decades. The availability of experimental data on binding affinity prompted researchers to explore the principles and develop methods for prediction [4]. However, the amount of experimentally observed three-dimensional protein-protein complexes with known binding constants still remains extremely limited, which complicates the application of modern deep learning methods in this task. The most recent computational research on PPI binding affinity prediction is mainly built around the idea of utilizing standard statistical and machine learning methods trained on various handcrafted descriptors such as QSAR features [1], inter-residue contacts and buried surface area , surface tension area and hydrophobicity, and sequence-based descriptors.

Recent advances in adjacent computational problems such as protein structure prediction or protein-protein interaction prediction demonstrate how powerful deep learning methods are as long as enough training data is provided and correct inductive biases are set up.

In this work, we will apply geometric deep learning methods for predicting protein-protein binding affinity. We believe that geometry is a clue for understanding protein-protein interactions, and aim to notably move forward the state of the art in binding affinity prediction with the aid of graph neural networks. To the best of our knowledge, geometric deep learning methods have never been applied to the protein-protein binding affinity prediction problem so far.

Objectives

Three main objectives of this work can be formulated as follows:

Refine PDBbind [3] data and a standard binding affinity dataset, and compile a novel benchmark of PPIs with known binding affinity values

Employ graph-learning toolset to predict binding affinities of PPIs from the new dataset

Benchmark the resulting method against existing state-of-the-art approaches

Problem Statement

Let $\mathcal{D} = \{(X, y)\}$ be the given dataset, where X is referred to the design matrix and y as target vector. Let X be a vector of PDB codes, which represent an atomic coordinate file of a molecular structure. Thus, $X \in \mathbf{X}^n, y \in \mathbb{R}^n_+$. The problem is now defined as optimizing model w^* with respect to following functional: $w^* = \underset{w}{\text{argmin}} = S(w(\mathcal{D}_T), y_T)$, where S is the error function. The set $\mathcal{D}_T \subset \mathcal{D}$ is a training design subset with corresponding target y_T. For the validating purposes we use the quadratic error function: $S(a, b) = ||a - b||_2^2$.

Theory

Consider optimizing model $w = (w_{mod}, w_{pred})$ where $w(\mathcal{D}_T)) = w((X_T, y_T)) = w_{pred}(w_{mod}(X_T), y_t)$. We refer to w_{mod} as modifying function and w_{pred} as predicting function. We narrow the space of all optimizing models W to space of pairs of modifying and predicting functions $W_{mod} \times W_{pred}$.

Hypothesis 1 *Consider pair* (w_{mod}, w_{pred}) *with the least loss function on test set. Consider* ϕ *- linear space transformation. Then* $w_{mod}(X_T) = w_{mod}(\phi(X_T))$ *and* $w_{pred}(w_{mod}(X_T), y_t) \neq w_{pred}(w_{mod}(\phi(X_T)), y_t)$ *for almost all* X_T*.*

Hypothesis 2 *Consider pair* (w_{mod}, w_{pred}) *with the least loss function on test set. Consider* ψ *- permutation. Then* $w_{mod}(X_T) = w_{mod}(\psi(X_T))$*.*

Note 1 *Hypothesis 1 states significance of* w_{mod}*. For example, we do not consider identity operator as a possible modifying function.*

Based on these assumptions, our solution is described as follows:

1. Set the space W_{mod} such as $\forall w_{mod} \in W_{mod}$ hypothesis 1 and 2 stands.
2. Set the space W_{pred}
3. Cross-validate error function of each pair $(w_{mod}, w_{pred}) \in W_{mod} \times W_{pred}$ with respective domain and image.
4. Benchmark the best model from step 3 against existing state-of-the-art solutions on the test set.

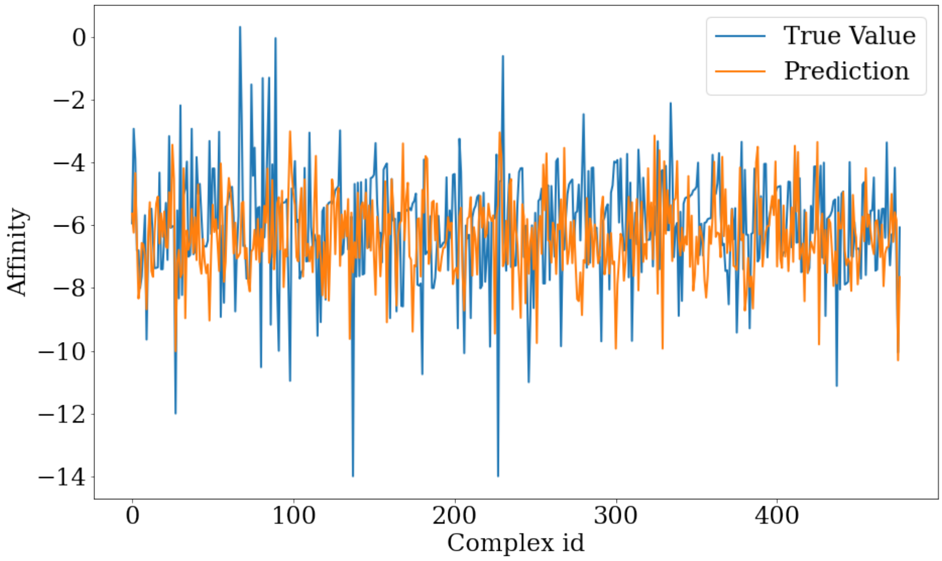
Computational experiment

The goal of this experiment is to select important features and set up baseline patterns for the resulting model.

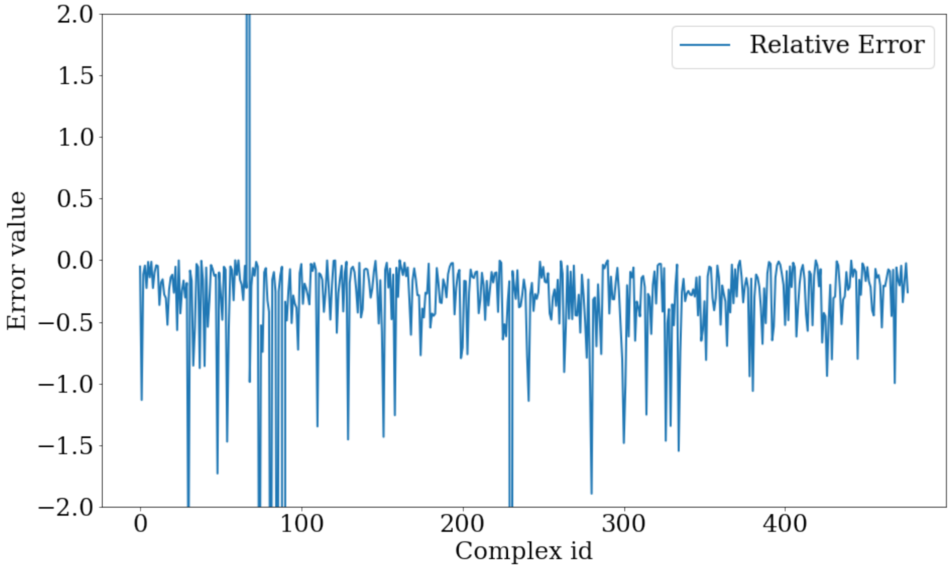
As a basic dataset we use PDBCNN database[3]. For the analysing purposes we retrieve the following information from each experiment - size of the first protein and size of the second protein. Therefore in the basic set we have 647 objects each represented by 2 real features and an affinity target.

We split the set into train and validation uniformly from Bernoulli distribution with probability equal to 0.7. Using cross-validation with 3 splits we fit a **CatBoostRegressor** model on train data.

Base CatBoost model performance



Base CatBoost model errors



CatBoost model with default parameters is our base model with MSE of 4.8. On the figures above one can find a relative error on the test data set and a prediction itself. We plan to benchmark our models against existing state-of-the art solutions and this one. Our current suggestions are Graph Neural Network and 3D-CNN.

Experiment Flow

We fit a model on the train dataset. Optimal parameters are chosen through cross-validation if needed. We analyze model robustness and calculate Mean Squared Error of prediction.

Expected tables and figures

1. Graph of model prediction and true target against complex index
2. Histogram of relative error distribution
3. Table with MSE depending on model parameters selection

Visualization

The source dataset contains 2327 protein complexes. We provide following plots for better understanding of given input.

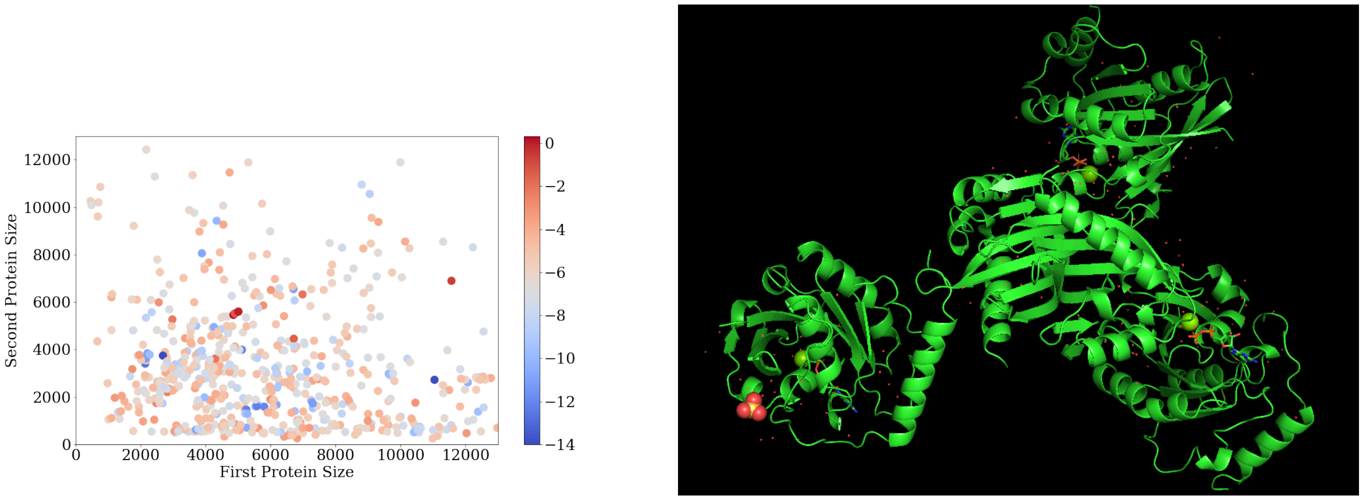


Figure 1: Source data target visualization / Example of source protein

References

[1] Yonggang Lv Feifei Tian and Li Yang. Structure-based prediction of protein–protein binding affinity with consideration of allosteric effect, 2012.

[2] Panagiotis L Kastritis and Alexandre MJJ Bonvin. On the binding affinity of macromolecular interactions: daring to ask why proteins interact, 2013.

[3] Yipin Lu Chao-Yie Yang Renxiao Wang, Xueliang Fang and Shaomeng Wang. The pdbbind database: methodologies and updates, 2005.

[4] K Yugandhar and M Michael Gromiha. Computational approaches for predicting binding partners, interface residues, and binding affinity of protein–protein complexes, 2017.