Pafnucy

A deep neural network for affinity prediction

Marta Stępniewska-Dziubińska

Institute of Biochemistry and Biophysics, PAS

6th RDKit UGM, September 2017, Berlin

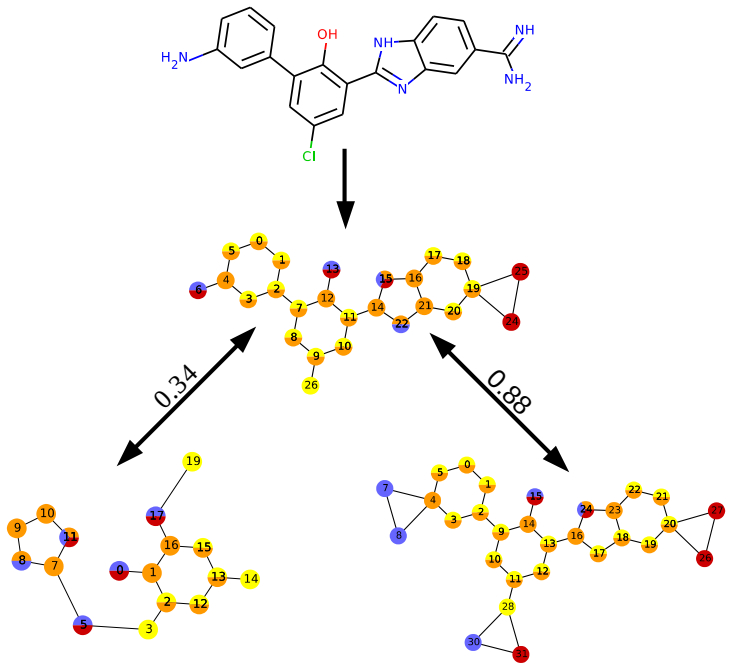
Outline

1. Introduction
2. Meet Pafnucy
3. Lets train it!
4. Results
5. Conclusions

Introduction

We want to find active molecules, as always

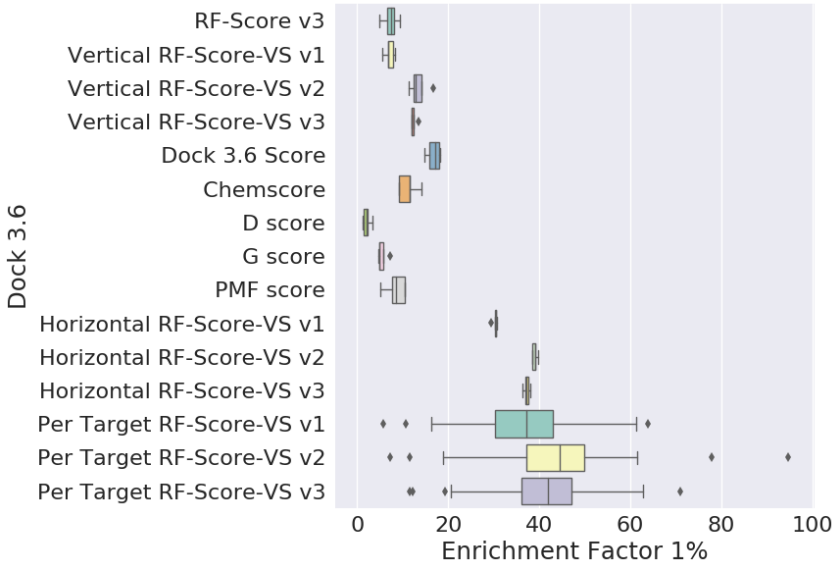
DeCAF (ligand based)



Stepniewska-Dziubinska *et al.* (2017)

doi:10.3390/molecules22071128

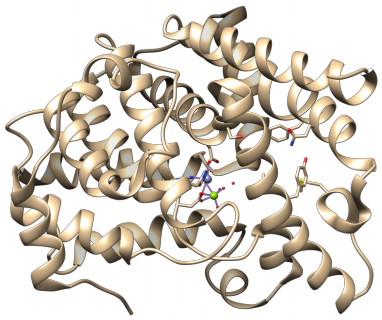
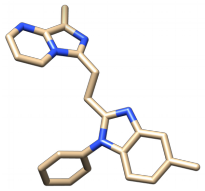
RF-Score-VS (structure based)



Wójcikowski *et al.* (2017)

doi:10.1038/srep46710

Deep learning - new solution to an old problem

f( + ) = binding affinity (e.g. pIC_{50})

Feature engineering

- I understand my data and know which part is important
- I know how to represent the information efficiently
- Don't tell me what to do with my data, I'll process it myself

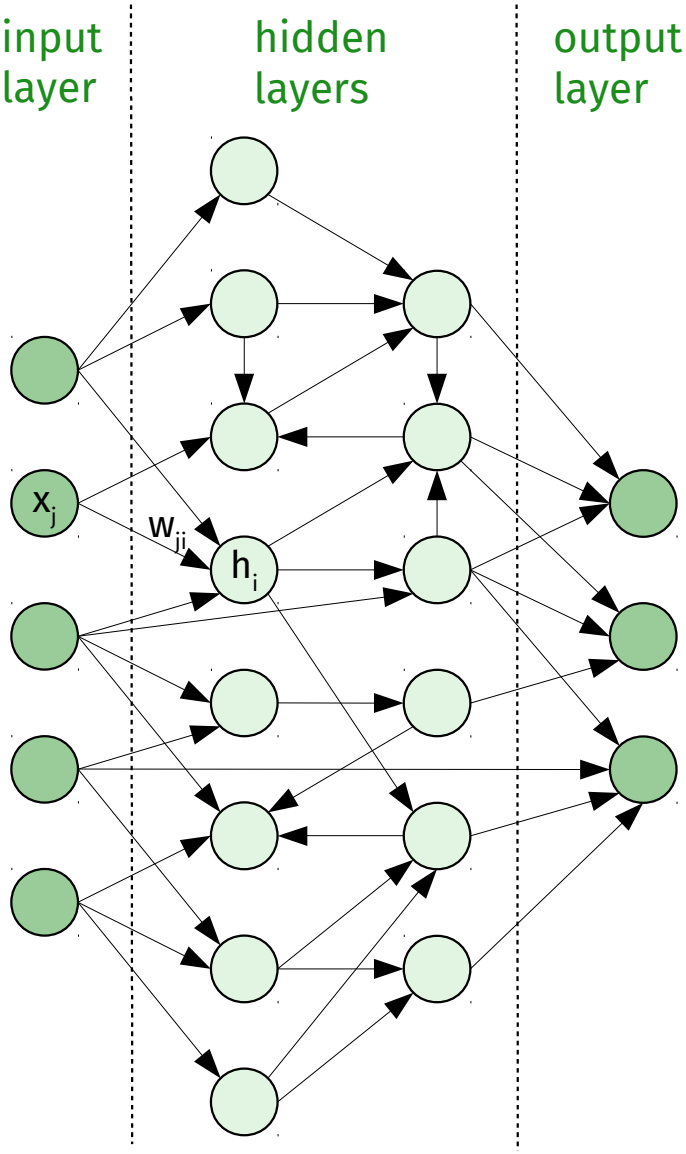
→ PLEC (talk to Maciek!)

Deep model

- I know there is a signal in my data
- I'm not sure how to properly extract it
- I'll give you my raw data and just process it for me, please

→ Pafnucy

How neural networks work



Neural networks consist of multiple layers of non-linear transformations.

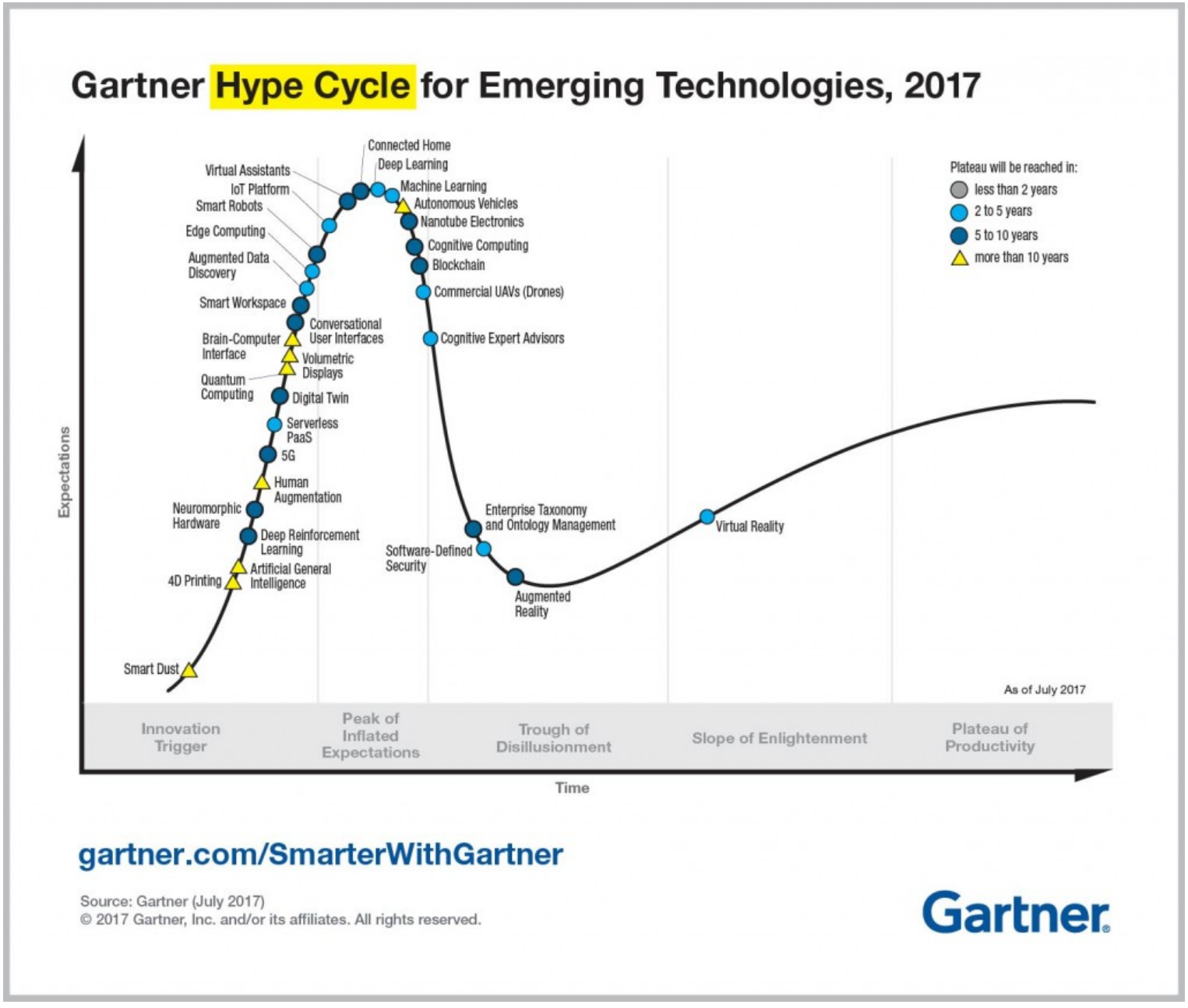
$$h_i = f\left(\sum w_{ji} x_j + b_i\right)$$

During training weight are adjusted to minimize the cost C:

$$w_{ji}(t+1) = w_{ji}(t) + \eta \frac{\partial C}{\partial w_{ji}}$$

Different types of architectures define different constraints on connections between neurons.

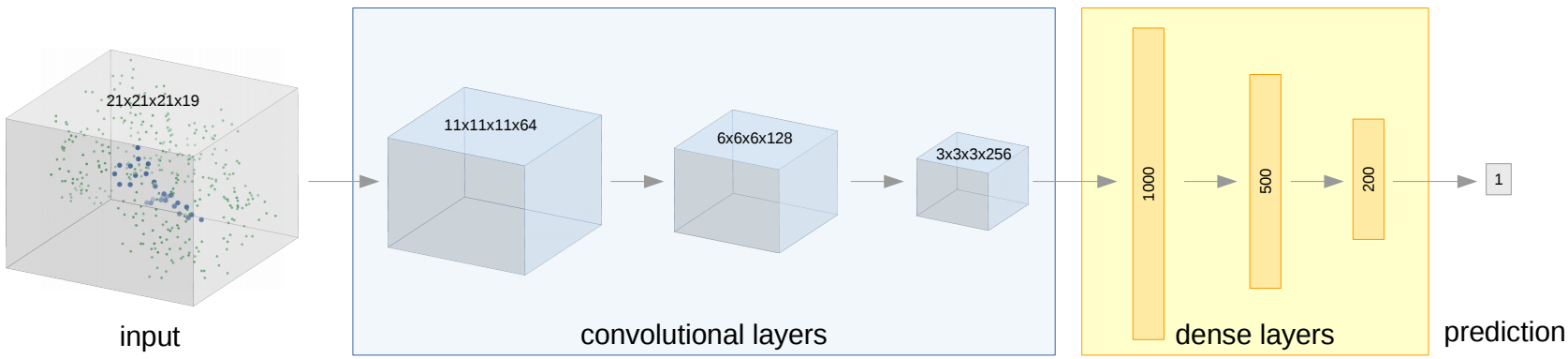
The hype is high



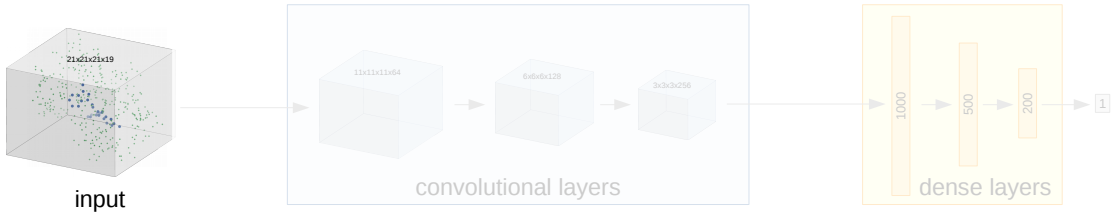
Meet Pafnucy

Architecture

Pafnucy **[pafnutzy]** is a deep convolutional neural network. It predicts binding affinities for 3D structures of molecular complexes. It was built and trained with TensorFlow and you can find it here: gitlab.com/cheminfIBB/pafnucy

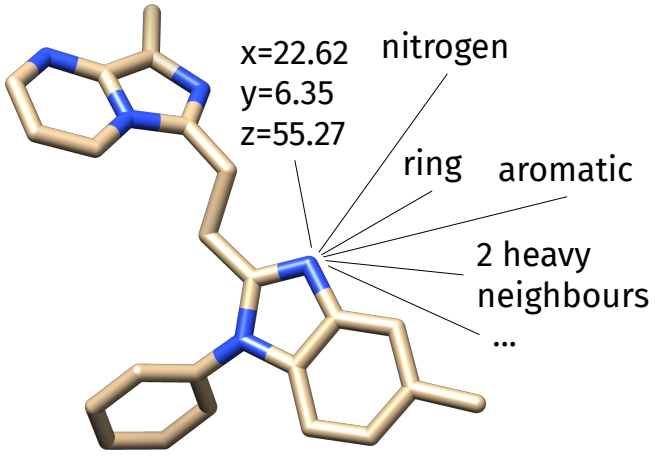


Input

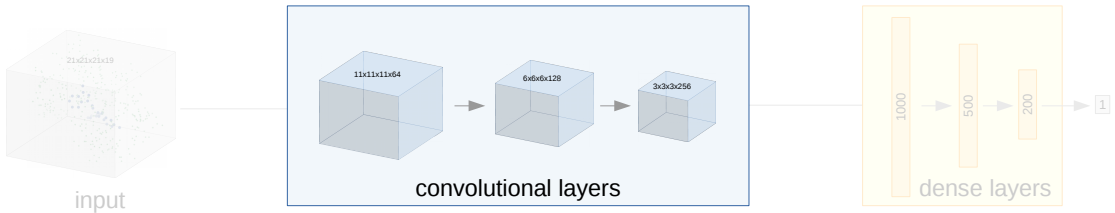


Input data are represented with 4D tensors, in which each point in 3D space is described with 19 atomic properties with:

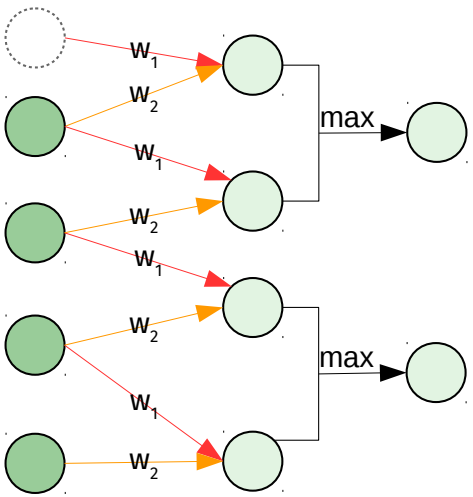
- atom type
- hybridization
- connectivity
- properties defined with SMARTS: *hydrophobic*, *aromatic*, *acceptor*, *donor*, and *ring*
- charge
- molecule it belongs to



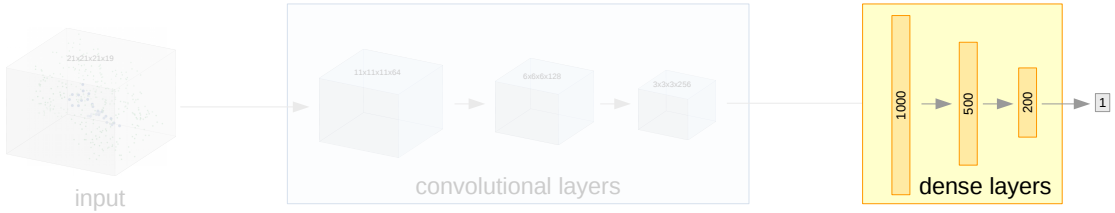
3D convolutional layers



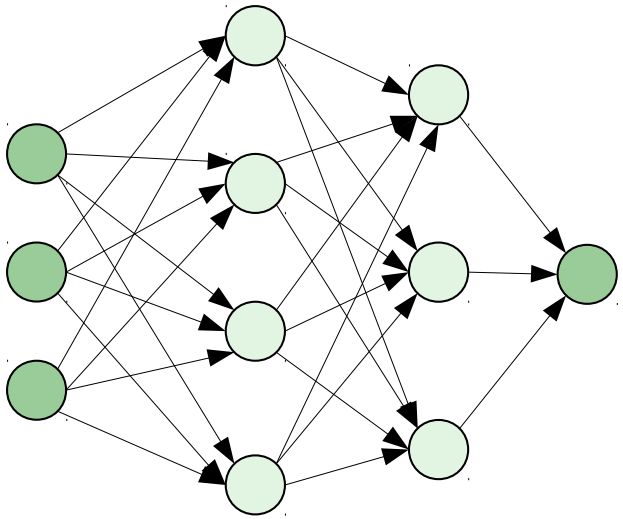
The input is processed with 3 consecutive convolutional layers with 64, 128, and 256 filters, respectively. Each layer has 5Å filter size and is followed by a max pooling layer with 2Å patch size. The result of the last convolutional layer is flattened and used as an input for the first dense layer.



Dense layers



The last block consists of 3 dense layers (also called fully-connected layers) with 1000, 500, and 200 neurons, respectively. In order to improve generalization, dropout with probability of 0.5 was used.



Lets train it!

Training on PDBbind

Pros

- 3D molecular complexes
- experimentally measured binding affinities (pK_i, pK_d, pIC_{50})
- over 13,000 structures
- core set - higher quality, manually curated subset
- used in CASF, so it is easy to compare Pafnucy to other methods

Cons

- crystallographic poses (we do not see them in VS)
- crude pocket preparation

Training details

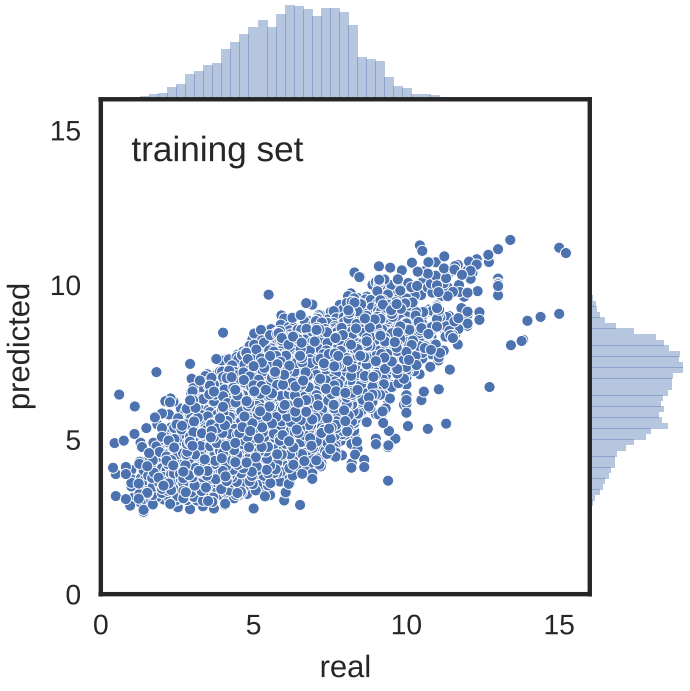
- general and refined sets used for training and validation + core set used as an external test set
- MSE + L_2 as a cost function:

$$C = \frac{1}{N} \sum_{i=1}^{N} (y_i - t_i)^2 + 0.001 \sum_{j=1}^{M} w_j^2$$

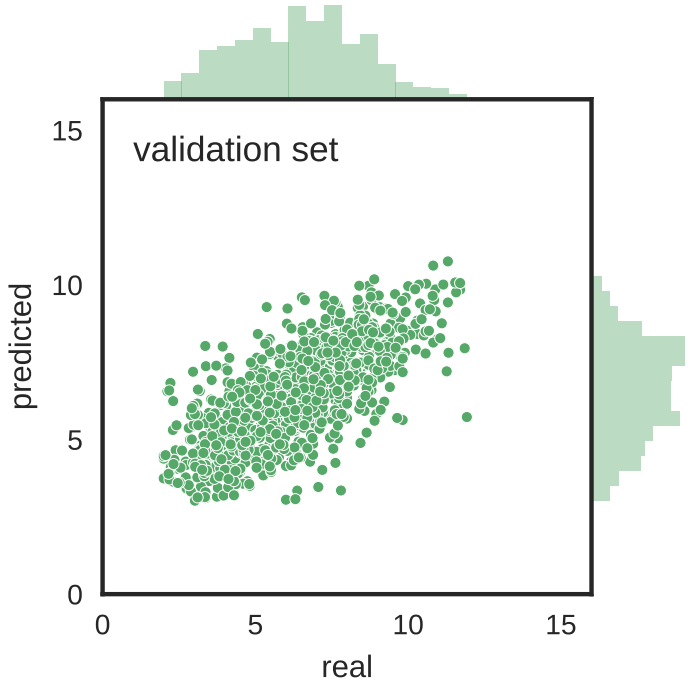
- Adam optimizer ($\eta = 10^{-5}$, 5 examples per minibatch)
- dropout ($p = 0.5$)
- trained for 20 epochs (the best model found after 14 epochs)

Results

Predictions - training and validation sets

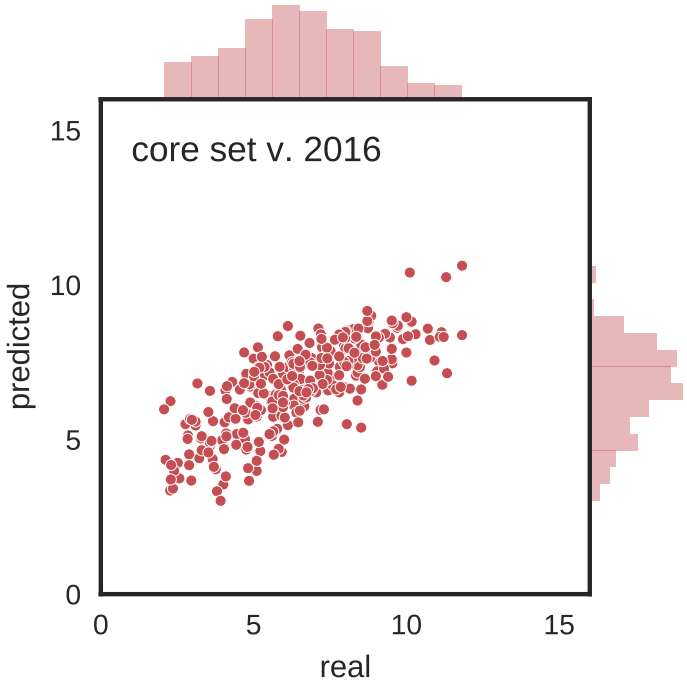


(a) RMSE=1.211, R=0.77

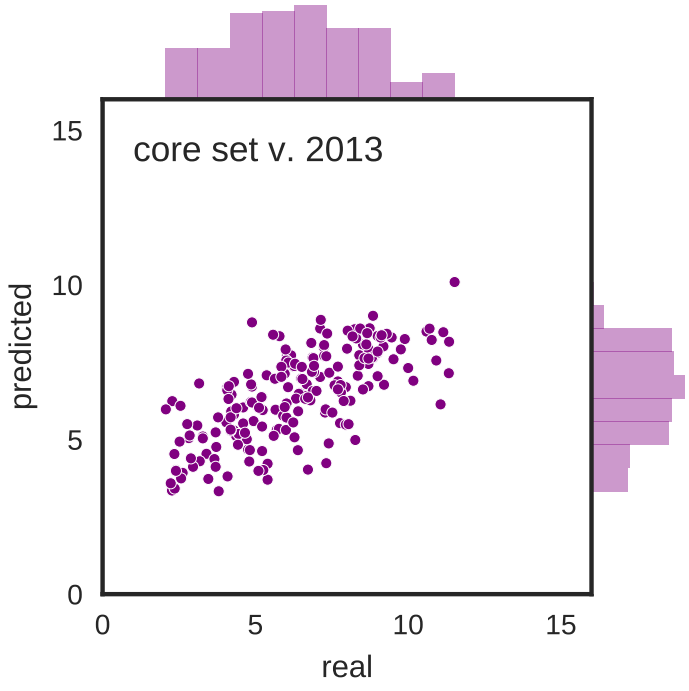


(b) RMSE=1.439, R=0.72

Predictions - external test set



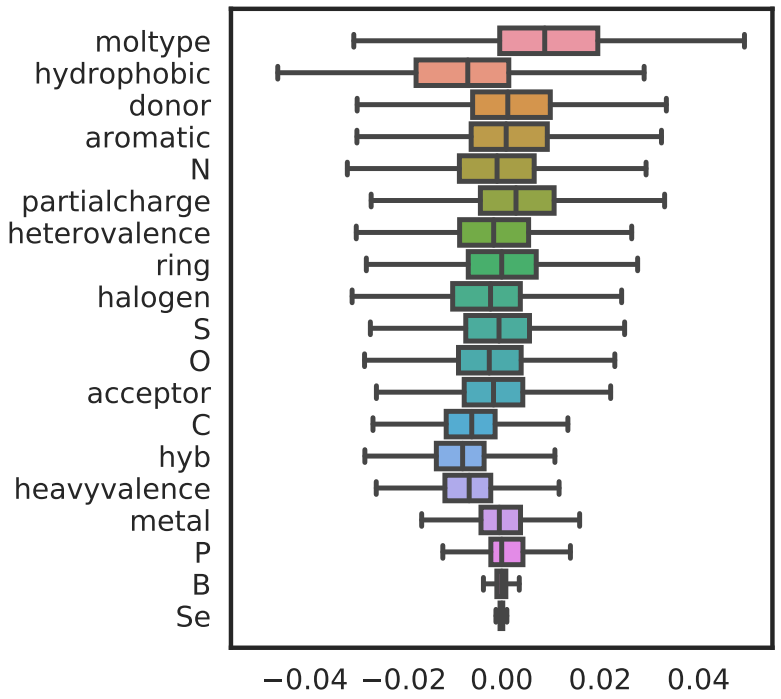
(a) RMSE=1.417, R=0.78



(b) RMSE=1.616, R=0.70

Feature importance

Model was trained with L_2, so unimportant features have outgoing weights close to 0. The wider the range, the more important the feature is.
...and *moltype* is that one bit, that distinguishes the protein from the ligand.



Why it sometimes fails?

For several complexes the model fails. It is related to...

- [ ] number of similar examples in the training set
- [ ] data quality
- [ ] crystallization method or environment
- [ ] particular protein or organism
- [ ] AA composition of the pocket
- [ ] ligand properties
- [ ] combination of different factors
- [✗] none of the above

Conclusions

Summary

Pafnucy:

- is a deep 3D convolutional neural network
- performs regression
- was trained on PDBbind dataset
- performs better than any other scoring function tested on CASF
- uses data in a reasonable way

Now we need to:

- create a better dataset:
  - more inactive compounds
  - prepared pockets
  - docked ligands
- use `deepchem` as a backend to make Pafnucy more accessible and easier to reuse

Who we are and where to find us

Our team:

Paweł Siedlecki

Maciek Wójcikowski (`mwojcikowski`)

Marta Stępniewska-Dziubińska (`marta-sd`, martasd@ibb.waw.pl)

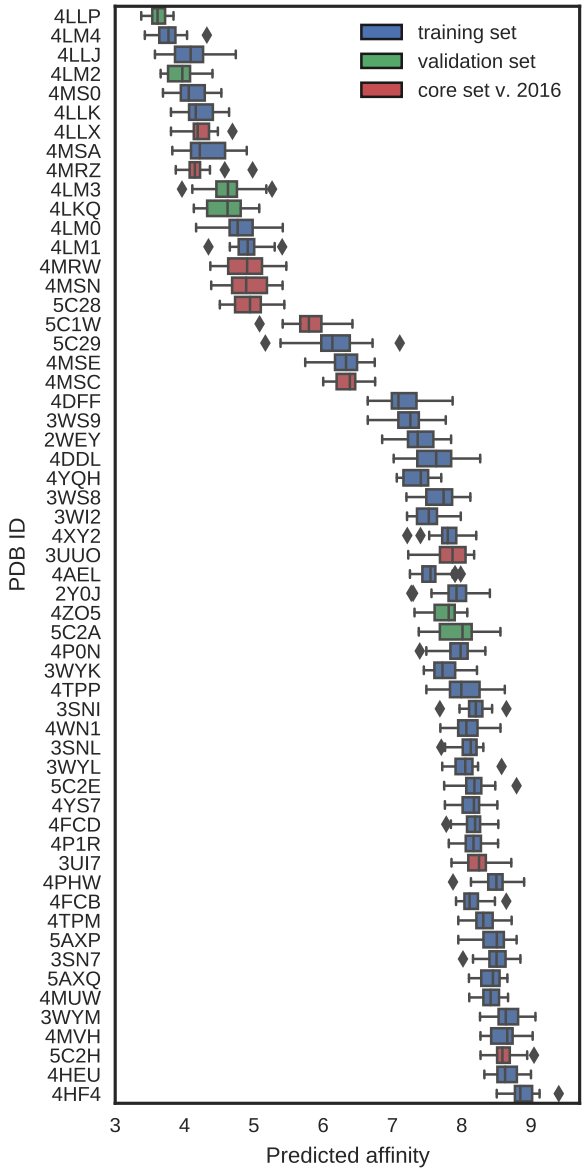
You can find our work at:

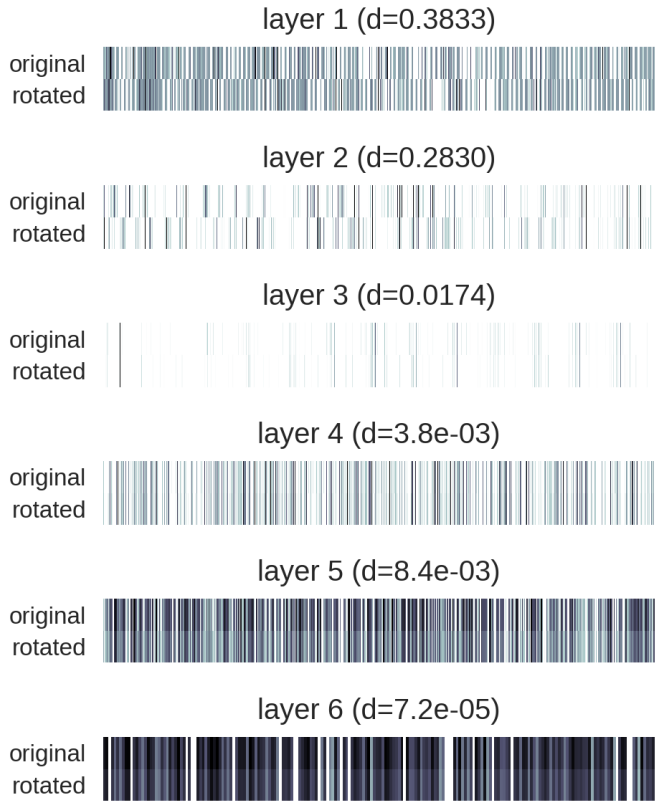
 github.com/cheminfIBB

 gitlab.com/cheminfIBB

 anaconda.org/cheminfIBB

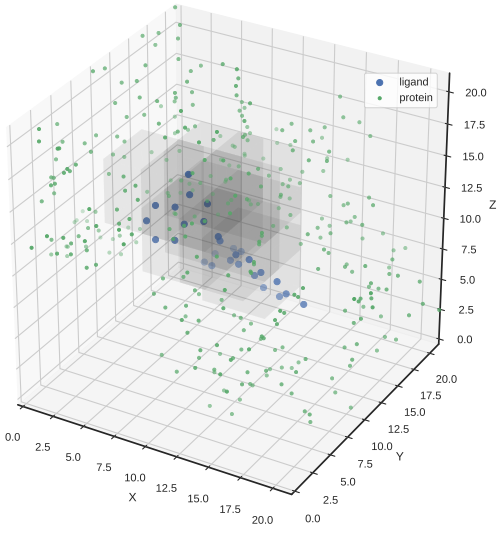
What if we rotate the input?



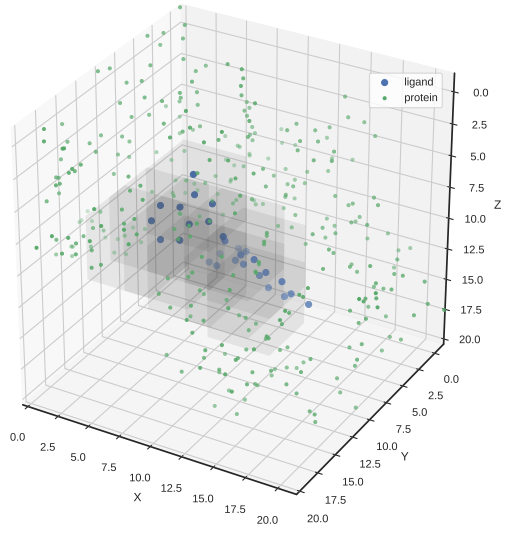


Results are stable. Pafnucy extracts the same information using different filters.

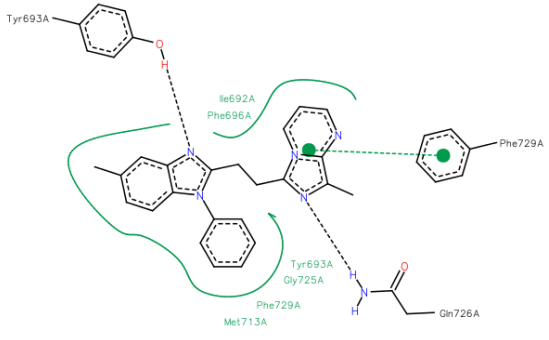
The most important parts of an input



(a) Original orientation



(b) Rotated by 180°



(c) Protein-ligand interactions.

Why we called it Pafucy?

”Shivering trunks” (2010) by Natalia Brożyńska