Protein-Ligand Extended Connectivity (PLEC) fingerprint and scoring function

Maciej Wójcikowski*[1], Michał Kukiełka[2], Marta M. Stepniewska-Dziubinska[1], Pawel Siedlecki[1,3]

[1] Department of Bioinformatics, Institute of Biochemistry and Biophysics, PAS, Pawinskiego 5a, 02-106 Warsaw, Poland
[2] Faculty of Mathematics, Informatics, and Mechanics, University of Warsaw, Banacha 2, 02-097 Warsaw, Poland
[3] Department of Systems Biology, University of Warsaw, Miecznikowa 1, 02-096 Warsaw, Poland

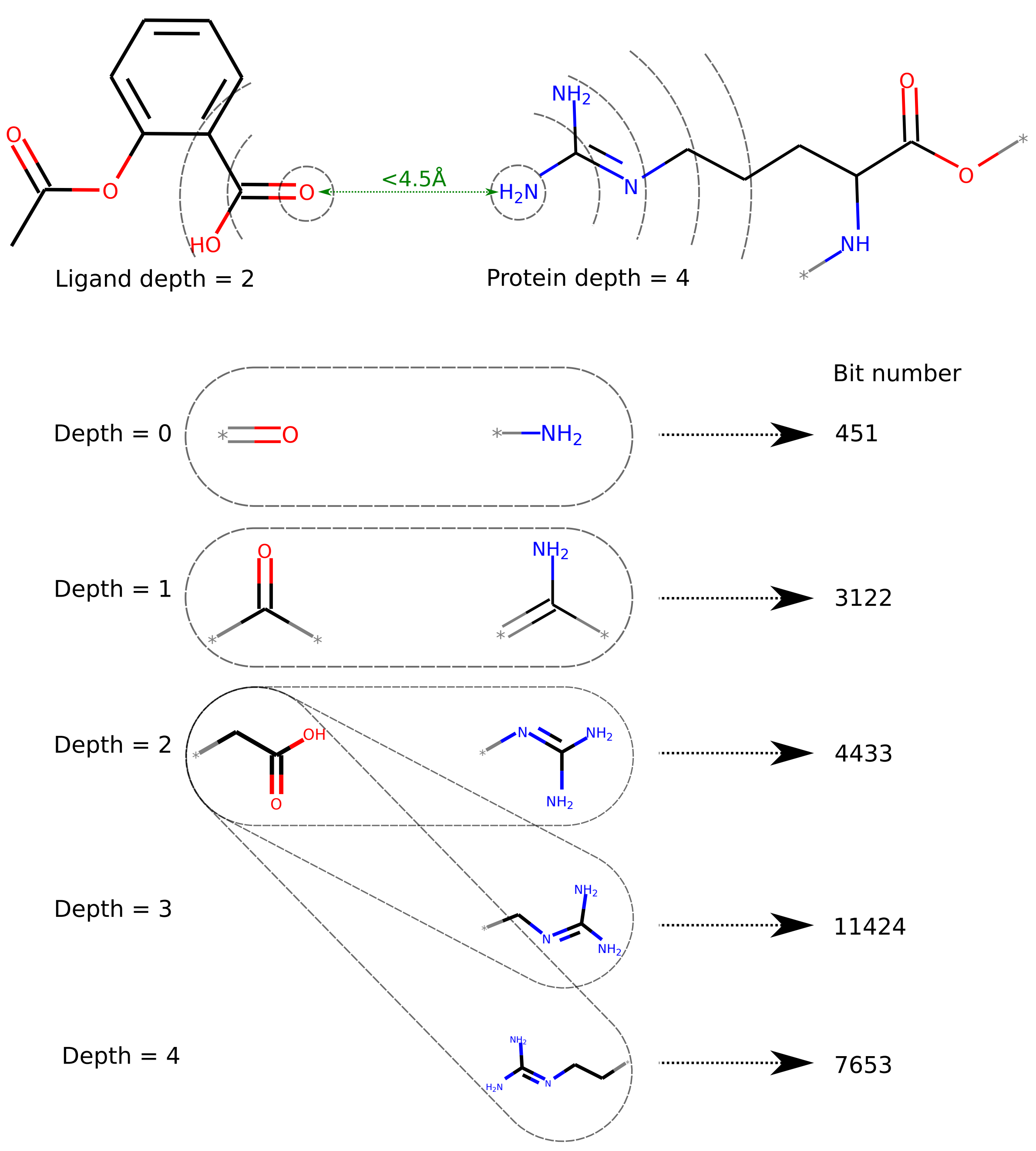
* maciek@wojcikowski.pl

1. Introduction

There are a number of interaction fingerprints (IFPs) that encode protein ligand interaction pattern. The most basic FPs explicitly define well known interaction types like hydrogen bonds, halogen bonds, Pi-stacking, etc. One exception is SPLIF [1], which uses hashed atom types to define implicit interactions. IFPs are mostly used as RMSD replacements, but there are some efforts to make use of them in scoring functions, mainly machine-learning based. Virtual screening and scoring protein-ligand affinity is still an outstanding challenge, thus we saw a niche for novel protein-ligand fingerprint which will provide implicit categorization of interactions and immense feature power for use in scoring functions. With aforementioned in mind we have developed **P**rotein-**L**igand **E**xtended **C**onnectivity (**PLEC**) fingerprint. PLEC builds upon the Extended Connectivity Fingerprints (ECFP) [2], the most versatile FP for most applications [3], to encode atom types and define protein-ligand contacts. The fingerprint is integrated in the latest version of Open Drug Discovery Toolkit (**http://github.com/oddt/oddt**) and is available with both backends (RDKit and OpenBabel).

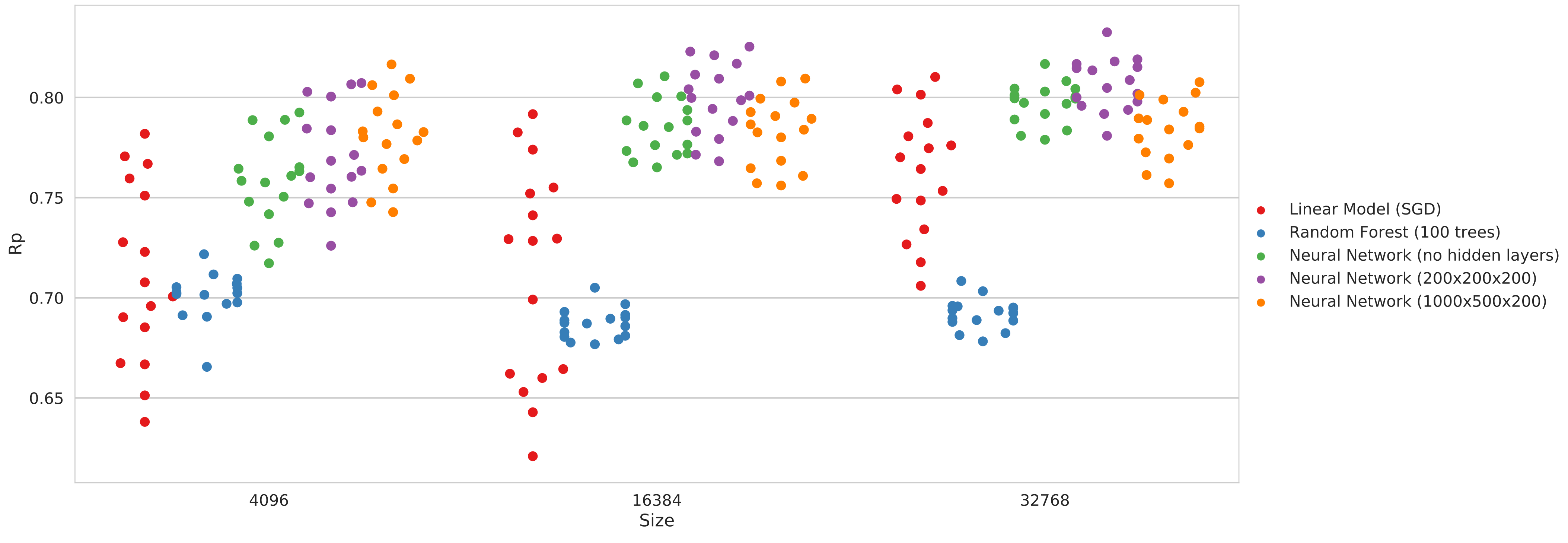
2. Constructing PLEC fingerprint

The PLEC fingerprint is a derivative of the original Extended Connectivity Fingerprint (ECFP) [2]. In contrast to ECFP not all atoms are used for creating PLEC FP for protein-ligand complex, but only pairs involved in **contacts defined by a distance cutoff of 4.5Å**. Consequently, for a pair of atoms in contact, the environments are generated for both atoms as in ECFP fingerprint. PLEC uses **multiple environments of ligand and protein**, and **variable maximum depths**, as opposed to SPLIF, which retains only one depth of 1 and includes 3D coordinates of bits. Each environment of ligand atom is paired with an environment of corresponding depth from protein and hashed to a final bit position in PLEC FP. In case of non-symmetrical depths (e.g. protein has greater depth), extra environments are paired with the largest one from the other molecule. The raw fingerprint is 32-bit long and is folded, as every other fingerprint, to much smaller length. Our analysis uses sizes of 4096, 16384, and 32768.



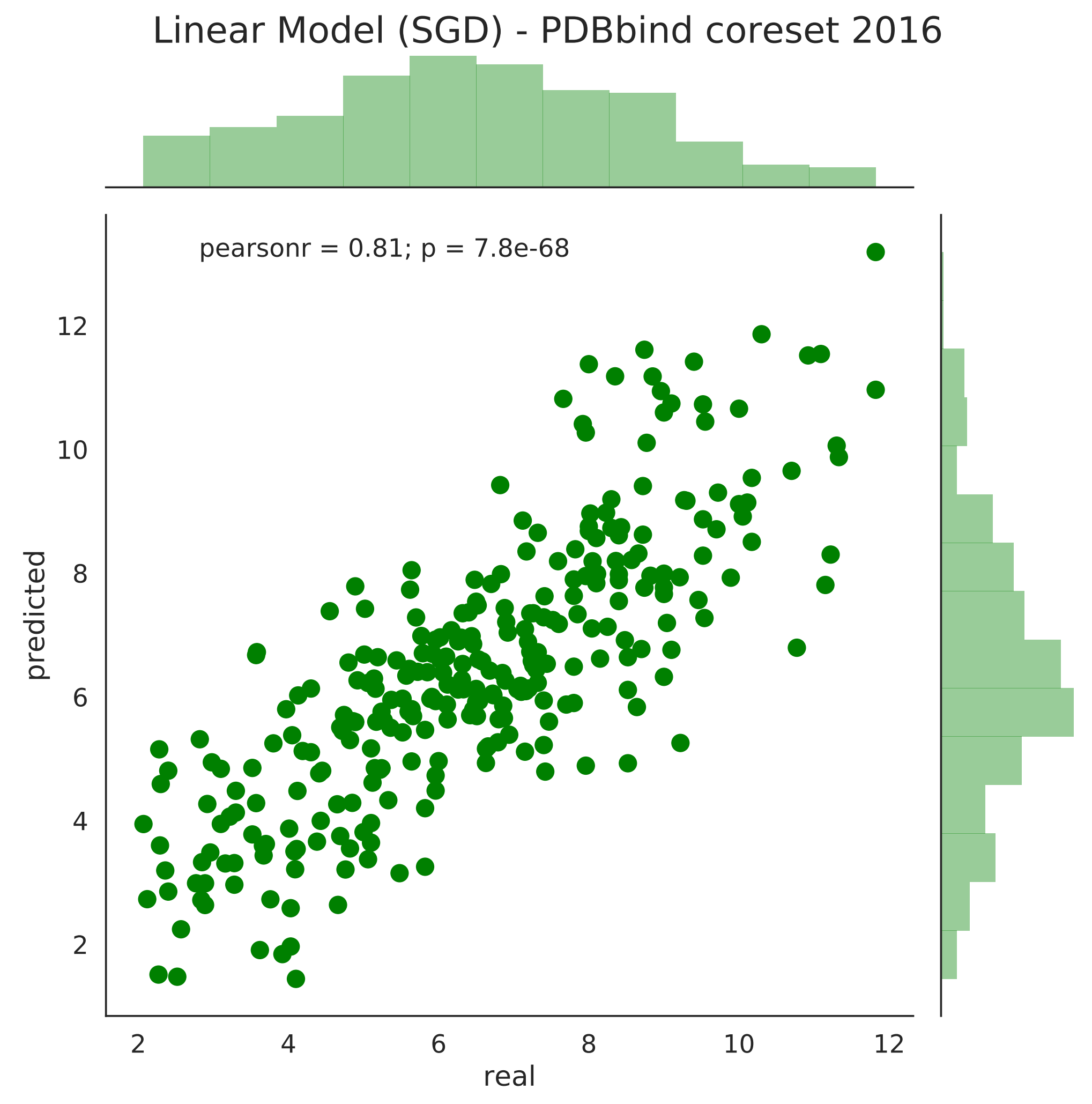
3. Scoring function

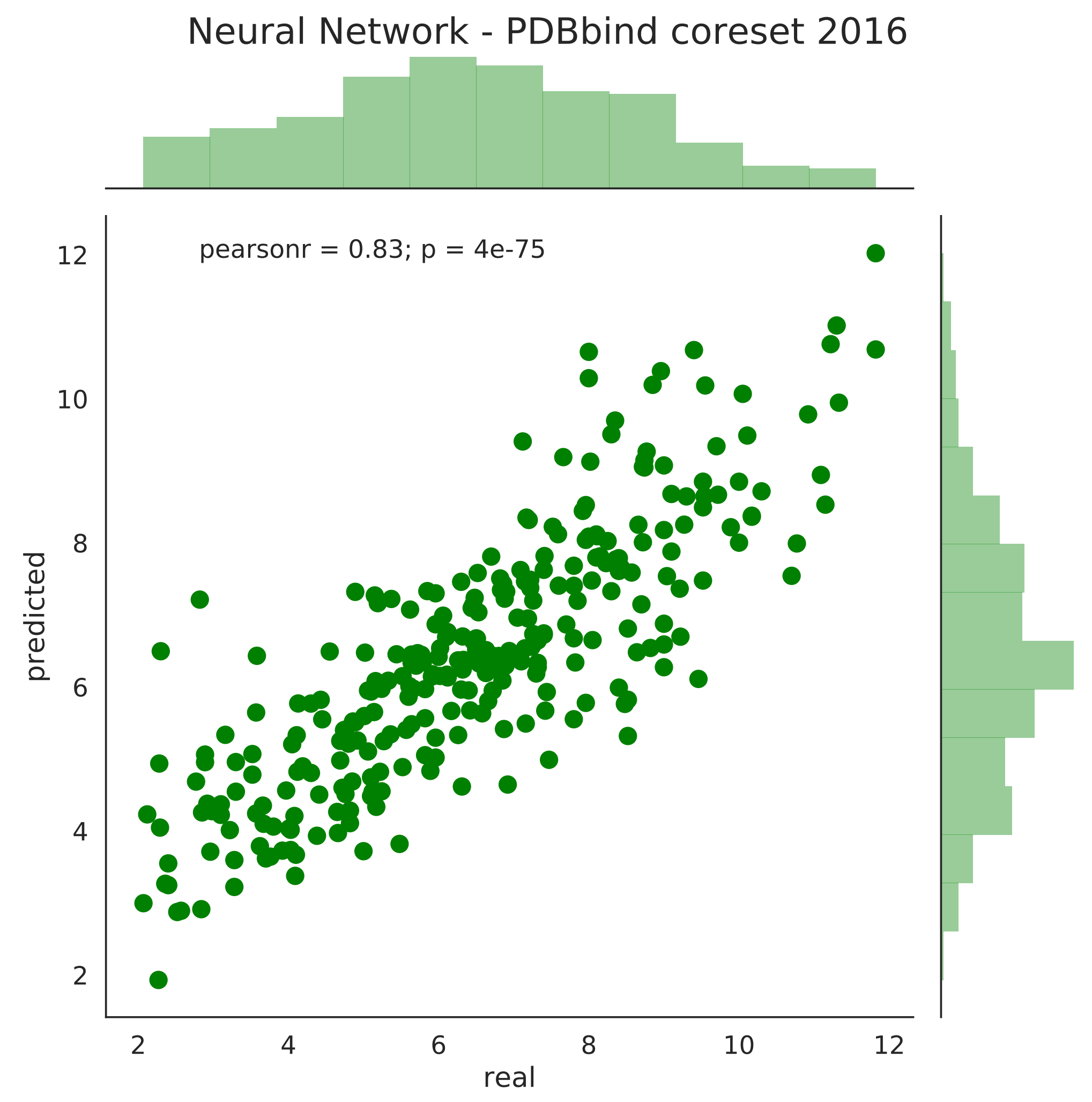
The PLEC fingerprint was used as an input vector for various machine learning models, which constructed novel scoring function. Following figure presents a combination of various models (Linear, Random forest and Neural network) trained on the PLEC fingerprint generated with protein and ligand depths ranging from 1 to 4.



4. Validation

The best performing scoring functions were built on a PLEC fingerprint with protein depth of 4, ligand depth of 1, and final size of 32,768 bits. Noteworthy the linear scoring function is almost as good at predicting pKi as neural network, which highlights the feature power of PLEC fingerprint. PLEC linear SF tested on latest PDBbind v2016 core set scored $\mathbf{R_p=0.81}$ and **SD=1.29**, which to our knowledge is the best linear model. PLEC neural network SF did even better with $\mathbf{R_p=0.83}$ and **SD=1.21**.





PLEC scoring functions with linear model and neural network outperform all 20 different scoring functions tested on the CASF-2013 "scoring power" benchmark [4]. PLEC linear scored $\mathbf{R_p=0.78}$ and **SD=1.41**, while PLEC neural network achieved $\mathbf{R_p=0.78}$ and **SD=1.40**. Comparing to the best X-Score, which got R_p=0.614 and SD=1.78, this is significant improvement. Even with the latest best machine learning scoring function – RF–Score v3, which scored R_p=0.803 and SD=1.42, the PLEC linear SF is on par with much simpler model.

5. Conclusions

- The PLEC FP does implicit enumeration of protein-ligand contacts which performs remarkably better than any other method used in scoring functions, models automatically classify the impact of each contact at the compound's affinity.
- Consistent performance across different machine learning models proves, that the results are the consequence of great feature power, rather than sophisticated model. Although some performance gain for neural network is still present, the linear model is just a little bit worse.

References

[1] C Da and D Kireev. Structural Protein–Ligand interaction fingerprints (SPLIF) for Structure-Based virtual screening: Method and benchmark study. *J. Chem. Inf. Model.*, 54(9):2555–2561, 2014.
[2] David Rogers and Mathew Hahn. Extended-connectivity fingerprints. *J. Chem. Inf. Model.*, 50(5):742–754, 24 May 2010.
[3] Noel M O'Boyle and Roger A Sayle. Comparing structural fingerprints using a literature-based similarity benchmark. *J. Cheminform.*, 8:36, 5 July 2016.
[4] Yan Li, et al. Comparative assessment of scoring functions on an updated benchmark: 2. evaluation methods and general results. *J Chem Inf Model*, 54(6):1717–1736, 2014.