Comparison of different techniques for image augmentation.

Ralph Lesch and Joshua Heipel
University of Freiburg, Department of Computer Science, Deep Learning Bachelor Project
joshua.heipel@gmx.de

1. Introduction

Convolutional Neural Networks (CNNs) have become a major aproach for the task of Semantic Segmentation and are nowadays used widely over many different fields of applications. Although for some use cases large datasets with thousands of (manually) labeled images have been published (such as CamVid or CityScape in the context of city traffic), in many situations appropriate training data still remains sparse. As a consequence Deep Neural Networks with lots of trainable parameters tend to overfit small and monotonous datasets while generating poor predictions for new (unseen) observations. In order to improve generalization of such CNNs existing training data can be extended by employing different techniques of image augmentation.

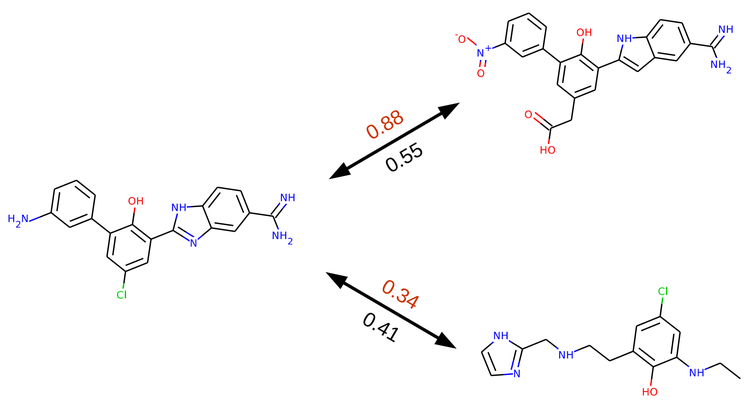
2. General Setting

In this study we use a hierarchical encoder-decoder network with skip connections (CNN of exercise 3) to compare different settings:

- **No Augmentation**
- **Shape Augmentation** with different geometric translations and dropouts (Horizontal Flip, Scaling, Crop and Padding, rectangular Cutouts)
- **Color Augmentation** by varying intensity of pixel values (Adjustment of Brightness and Contrast, Color shifts)
- **Shape and Color Augmentation**

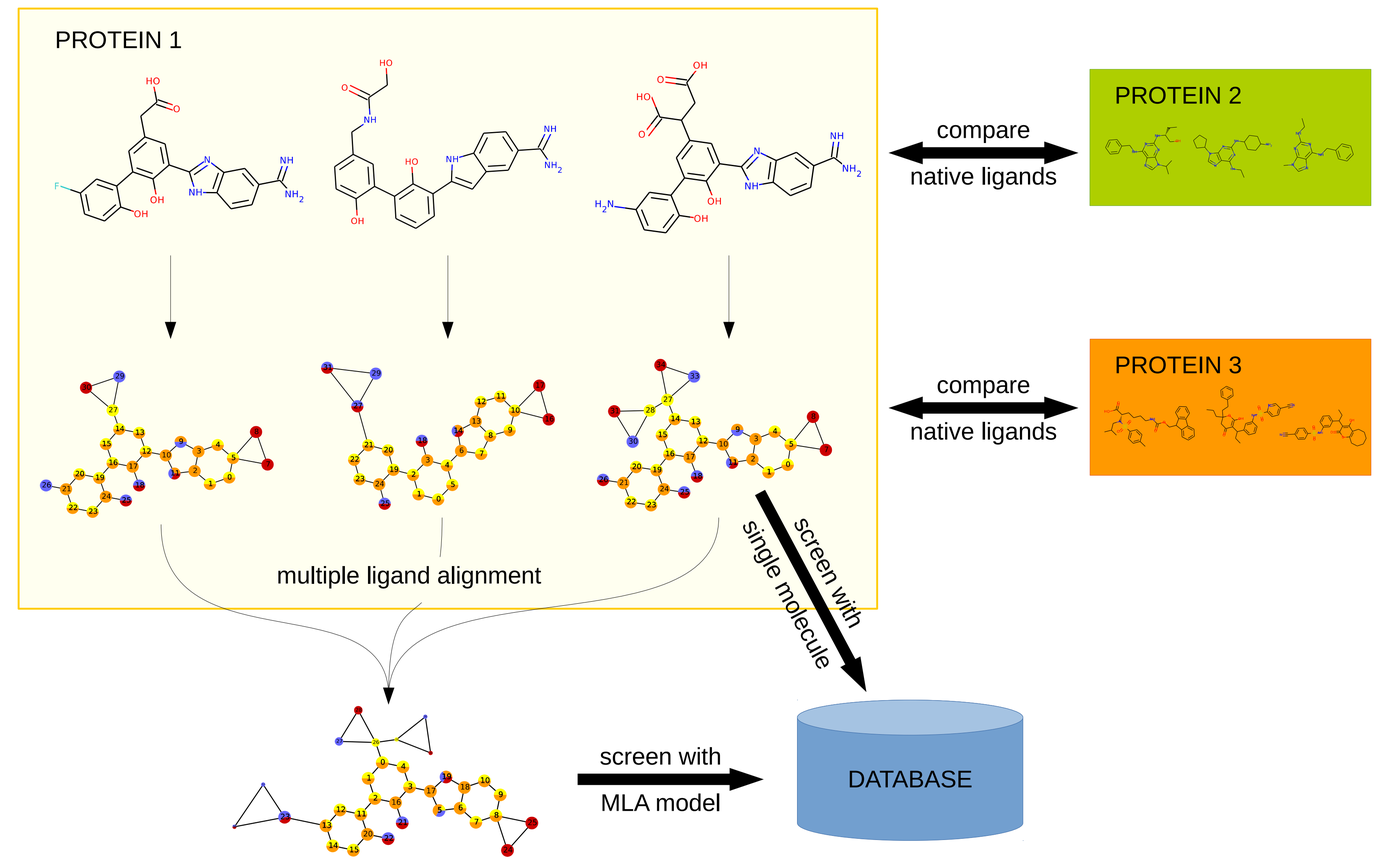
3. Similarity measure

To measure similarity of two molecules or to combine them into one model, DeCAF first finds their **maximum common substructure (MCS)**. To provide fast, but accurate method for solving MCS problem, we combined Generic Match Algorithm (GMA) [1] with backtracking algorithm proposed by Yiqun Cao [2].
Here we present comparison of molecules with similar and with different structures. DeCAF scores and **Tanimoto coefficient (Tc)** values are shown in red and black, respectively.

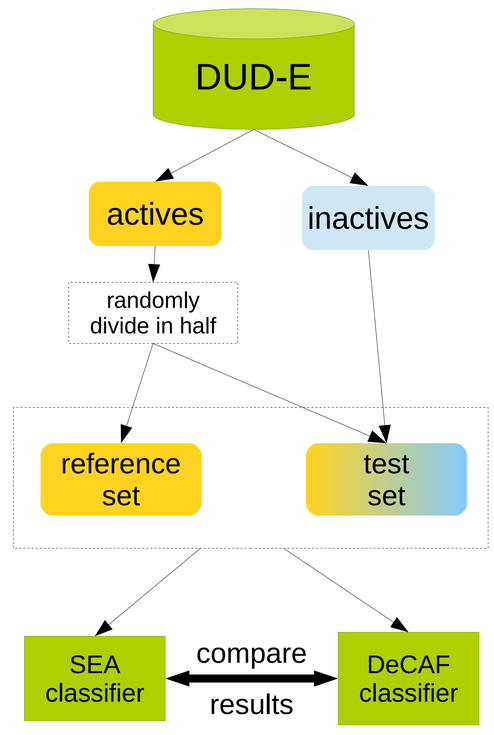


4. Applications

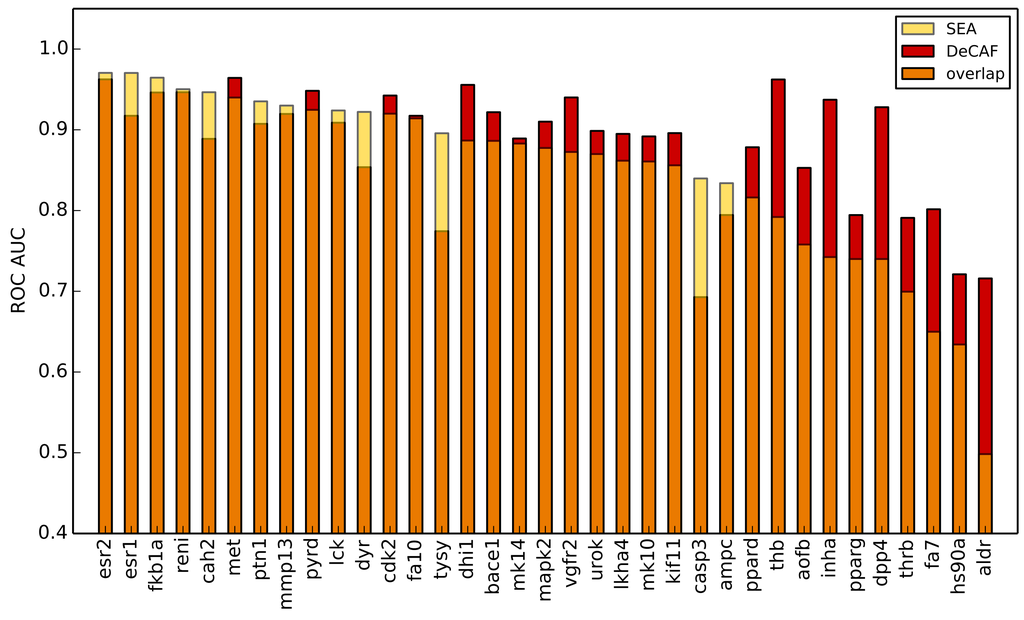
DeCAF is a versatile tool with many possible applications. It allows to compare two molecules or more complex models created from sets of ligands. Our method can be used to align multiple ligands and find crucial pharmacophoric features in a set of active compounds. Pharmacophore models can help in database screening for molecules with desired properties. DeCAF is also suitable for comparing entire sets of ligands, e.g. to analyse properties of proteins in drug repositioning process.



5. DeCAF vs. SEA



We tested DeCAF in 35 diverse targets taken from the DUD-E database, to evaluate its power to classify molecules as active or inactive. We compared DeCAF to the renowned **SEA (Similarity Ensemble Approach)** algorithm [3], which uses Tc as a similarity measure. Dataset preparation steps are shown on the left diagram. Comparison results (**ROC AUC** values for each receptor) are shown below.



6. Conclusions

We proved that DeCAF is a significant improvement over the SEA algorithm, a popular method for comparing sets of ligands.

1. DeCAF gives better results for 23 out of 35 receptors.
2. For targets with easily separable active and inactive datasets, SEA and DeCAF give similar results.
3. In cases in which SEA fails to identify active molecules, our method performs substantially better.

7. References

[1] Jun Xu. Gma: a generic match algorithm for structural homomorphism, isomorphism, and maximal common substructure match and its applications. *J. Chem. Inf. Comput. Sci.*, 36(1):25–34, 1996.

[2] Yiqun Cao, Tao Jiang, and Thomas Girke. A maximum common substructure-based algorithm for searching and predicting drug-like compounds. *Bioinformatics*, 24(13):i366–i374, 2008.

[3] Michael J Keiser, Bryan L Roth, Blaine N Armbruster, Paul Ernsberger, John J Irwin, and Brian K Shoichet. Relating protein pharmacology by ligand chemistry. *Nat. Biotechnol.*, 25(2):197–206, 2007.