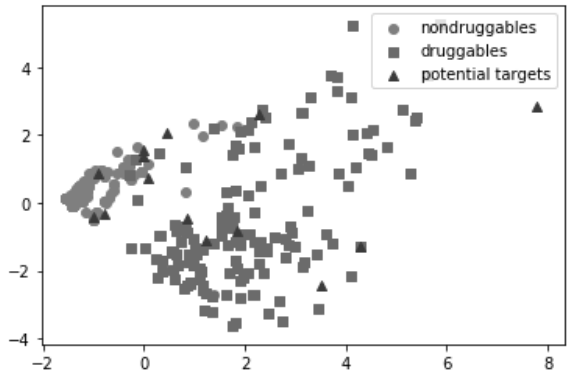
Druggability Prediction using Gene Ontology terms

Marie Detzler, Michael C. Hutter*,

** corresponding author: michael.hutter@bioinformatik.uni-saarland.de*

Center for Bioinformatics, Saarland University, Campus E2.1, 66123 Saarbrücken

Classification of potential drug targets into druggables and nondruggables has gained increased interest in the course of preclinical development and virtual screening. The considerable amount of potential targets in the human genome, as well as in pathogens has emphasized the need for corresponding *in silico* detection of valid targets. Whereas there is a large number of targets for which small molecule inhibitors have been approved as medications, [1] only few cases of targets that turned out to nondruggable were reported. Such disparate data sets where one class is highly overrepresented are a big problem for classification algorithms in general, because predictions for the underrepresented class are far less accurate. For example, simply the chance of accidentally classifying a member of the larger class correctly is much higher. Consequently, we applied several resampling techniques that comprise oversampling of the minority class and under-sampling the majority class in the framework of random forest and ada boost. [2, 3] Whereas other approaches for druggability prediction [4, 5] require 3D structures of the corresponding targets as input, we used terms form Gene Ontology, which describe the protein at hand by its associated biological processes, molecular function, and sub-cellular location. [6]



The first two principal components of the data sets indicate a separation of known druggabble (squares) and nondruggable (circles) targets, whereas the additional potential targets (triangles) are more spread among the two classes.

[1] M. Rask-Andersen, S. Masuram, H. Schiöt, *Ann. Rev. Pharmacol. Toxicol.*, **2014**, *54*, 9-26.
[2] N. Chawla, K. Bowyer, L. Hall, W. Kegelmeyer, *J. Artificial Intell. Res.*, **2002**, *16*, 321-357.
[3] D. Wilson, *IEEE Transactions Systems Man Cybernetics*, **1972**, *2-3*, 408-421.
[4] E. Fauman, B. Rai, E. Huang, *Curr. Opin. Chem. Biol.* **2011**, *15*, 463-468.
[5] A. Volkamer, D. Kuhn, F. Rippmann, M. Rarey, *Bioinformatics*, **2012**, *28*, 2074 -2075.
[6] The Gene Ontology Consortium, *Nucl. Acid. Res. D*, **2007**, *36 suppl.1*, D440-D444.