# Evaluation of Different Machine Learning Methods for Ligand-based Virtual Screening

Rafał Kurczab, Sabina Smusz, Andrzej J. Bojarski

Department of Medicinal Chemistry, Institute of Pharmacology Polish Academy of Sciences, Smętna 12 Street, 31-343 Kraków, Poland

## Introduction

One of the central themes in modern drug design and discovery process is the search for active compounds using various types of computational methods. New active compounds could be found in commercially available databases or in in-house combinatorial libraries using virtual screening methodology. There are two basic approaches to virtual screening: structure-based (which requires the availability of a 3D structure for the biological target of interest) and ligand-based methods (which are used when the 3D structure of a given target is unknown).

In recent years, many different machine learning and data mining methodologies have been successfully applied to identify active compounds, and also evaluation and comparison of these methods in benchmark databases has been done [1,2]. In most cases, however, the performance of only several different combinations of machine learning methods, 2D molecular fingerprints, training and testing sets were used. Due to this fact and lack of a common test for all possible permutations, it is not clear which methods are the most suitable for a particular purpose. Therefore, using a common benchmark we have evaluated over 60 different machine learning methods in combination with six different 2D molecular binary fingerprints and two attributes selection methods.

## Results and discussion

All calculations were performed using a collection of machine learning algorithms for data mining implemented in WEKA package [3], which achieved widespread acceptance within academia and business communities, and is freely available.
Due to the large number of results, in this report we have limited the analysis only to 26 carefully selected machine learning methods.

As a set of known actives, our in-house 5-HT7R antagonists library was used. From this database we have built two training sets: the first containing 38, and the second with 58 known active compounds, alongside with one test set consisting of 90 actives and 90 inactives. The compounds in training and tests sets were selected to be unique. Using 2D binary molecular fingerprints (i.e. Estate, Extended, Graph, MACCS, PubChem, Substructure and Daylight(FP)) calculated in freely available PaDEL software [4], all permutations of fingerprints, machine learning methods and two different training sets were tested. The effectiveness of the given combination was measured as a recall rate of known active compounds from the test set.

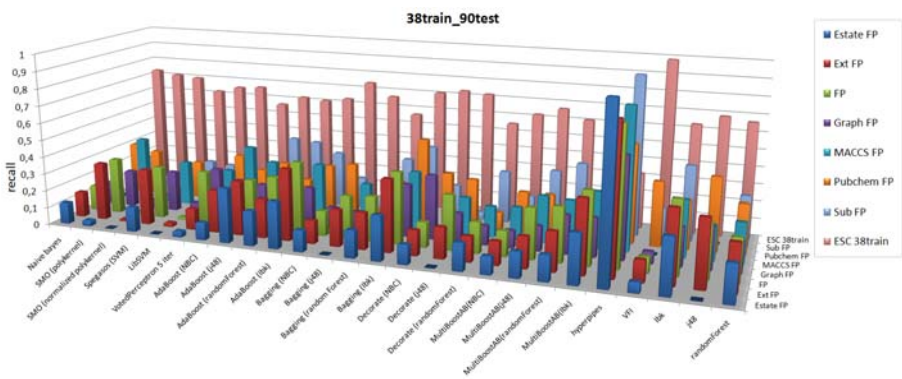


**Figure 1.** The results of performance of different combinations of 2D molecular fingerprints with machine learning methods, obtained for the training set containing 38 known actives and test set built from 90 actives and 90 inactives.

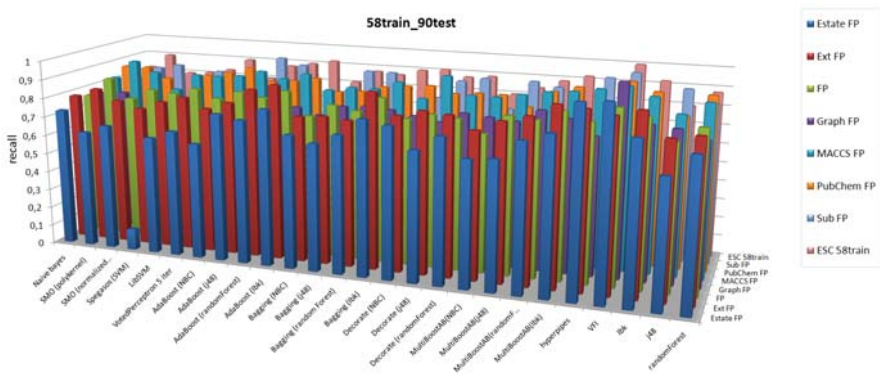


**Figure 2.** The results of performance of different combinations of 2D molecular fingerprints with machine learning methods, obtained for the training set containing 58 known actives and test set built from 90 actives and 90 inactives.

Globally, the results show (Figure 1 and 2) that there is no universal and the best performing combination. It should also be noted that the accuracy of classification strongly depends on the size of a training set. The exception is the hyperpipes method, which performs much better for the smaller training set than for the larger one. Moreover, this algorithm also worked better on shorter fingerprints pattern than other methods.
The second important conclusion is a strong dependence between the type and length of 2D molecular fingerprints and the classification results. On average, the best recall rates were obtained for the longest ones, but this effect is less pronounced for the larger training set.

For one type of fingerprint (FP of 2048 bits length) and two methods of attributes analysis and selection (i.e. the best first, and the genetic search algorithms), influence of their combinations on the classification accuracy of machine learning methods was performed. The results of this test (Figure 3) indicate that there is no clear correlation between the methods of selection of attributes and the size of training set, and moreover, it is hard to assign the best combinations. However, some of the methods (i.e. hyperpipes, VFI, Bagging(Ibk)) improved the classification results (even up to 30%), independently of the rest of parameters used.

In summary, our studies provided an overall view of the possibility of using different configurations of available methods and tools in ligand-based virtual screening. Further studies will be continued to find more useful correlations and general regularity to the most effective use of these tools in virtual screening.

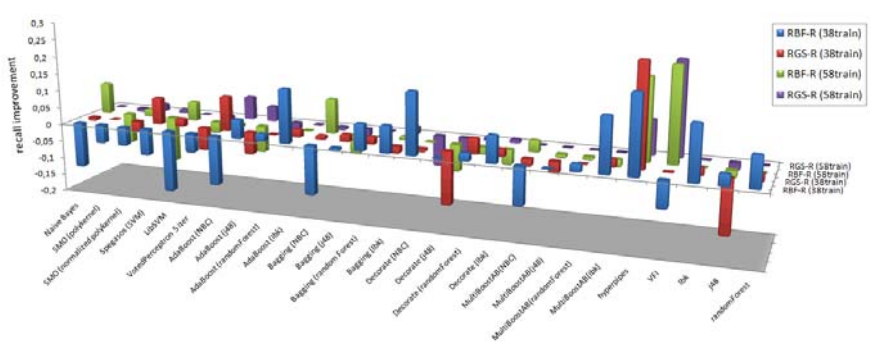


**Figure 3.** The influence of two attribute selection methods (the best first, and the genetic search algorithm) combined with Daylight (2048 bits length) 2D molecular fingerprint and two training sets, on the recall rates for different machine learning methods.

## References

[1]. Plewczynski D, Spieser S, Koch U: Performance of machine learning methods for ligand-based virtual screening. *Comb. Chem. High Throughput Screening* 2009, 12:358-368.
[2]. Ma X, Jia J, Zhu F, Xue Y, Li Z, Chen Y: Comparative analysis of machine learning methods in ligand-based virtual screening of large compound libraries. *Comb. Chem. High Throughput Screening* 2009, 12:344-357.
[3]. Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten I: The WEKA Data Mining Software. *SIGKDD Explorations* 2009, 11:10-18.
[4]. Yap CW, PaDEL-Descriptor: An open source software to calculate molecular descriptors and fingerprints. *Journal of Computational Chemistry. 2010,* In press

## Acknowledgments

This study was partly supported by a grant PNRF-103-AI-1/07 from Norway through the Norwegian Financial Mechanism.

Polish-Norwegian Research Fund

norway grants

www.cns-platform.eu