The insight on molecular fingerprint nature – how to enhance the virtual screening performance?

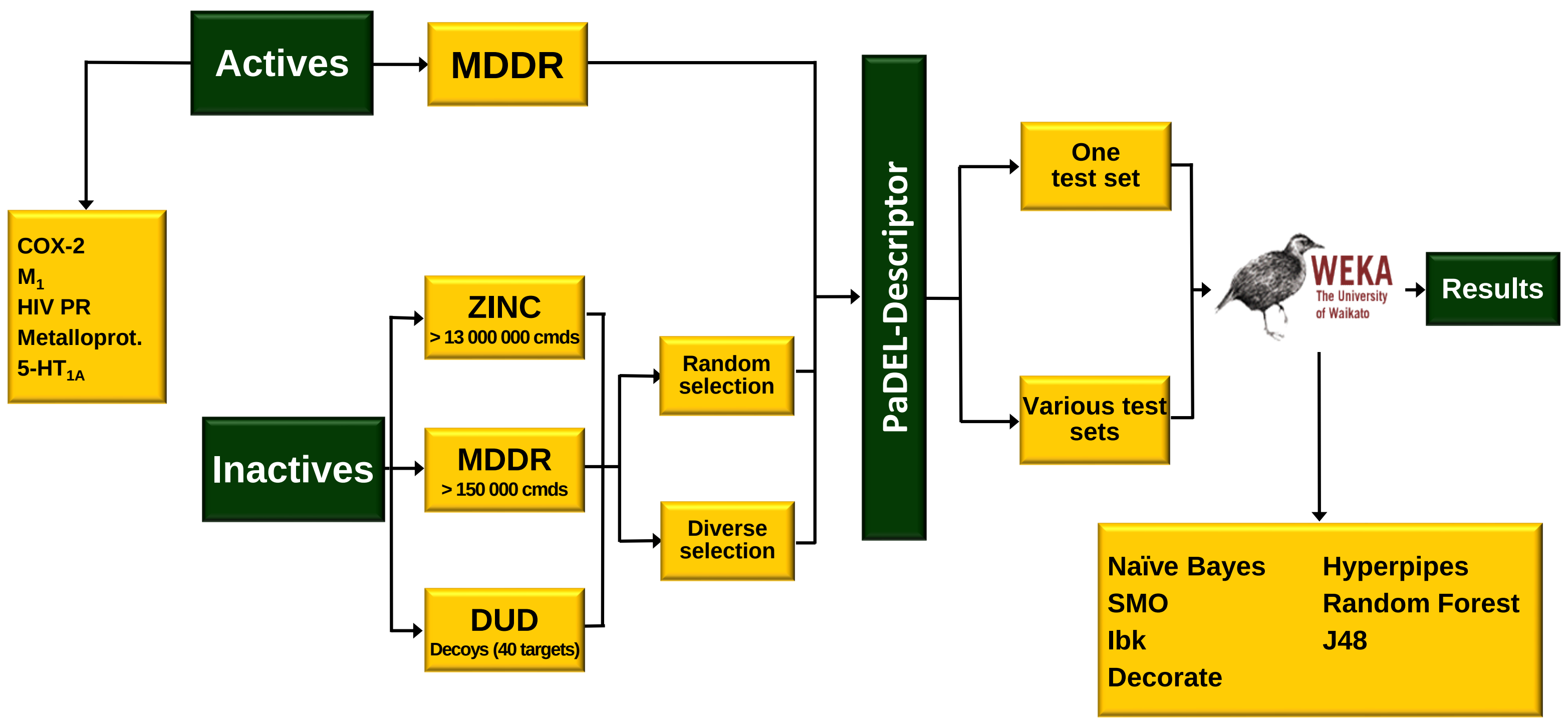
Sabina Smusz[a b], Rafał Kurczab[a], Andrzej J. Bojarski[a]

[a]Institute of Pharmacology Polish Academy of Sciences, 12 Smętna Street, 31-343 Kraków, Poland
[b]Faculty of Chemistry, Jagiellonian University, 3 Ingardena Street, 30-060 Kraków, Poland

Introduction

Molecular fingerprints are gaining more and more popularity in cheminformatic tasks, especially in those connected with application of machine learning (ML). It is a result of relatively low computational expenses connected with their generation and simplicity of making comparisons between two 0-1 strings. The effectiveness of ML methods is strongly dependent on the type of input data and representation used for compounds description [1]. Therefore, an extended study on those relationships was carried out in order to determine optimal conditions for such experiments

Influence of the way of the inactives sets generation on machine learning methods performance



Actives

MDDR

COX-2
M_1
HIV PR
Metalloprot.
5-HT_{1A}

Inactives

ZINC
> 13 000 000 cmds

MDDR
> 150 000 cmds

DUD
Decoys (40 targets)

Random selection

Diverse selection

PaDEL-Descriptor

One test set

Various test sets

Results

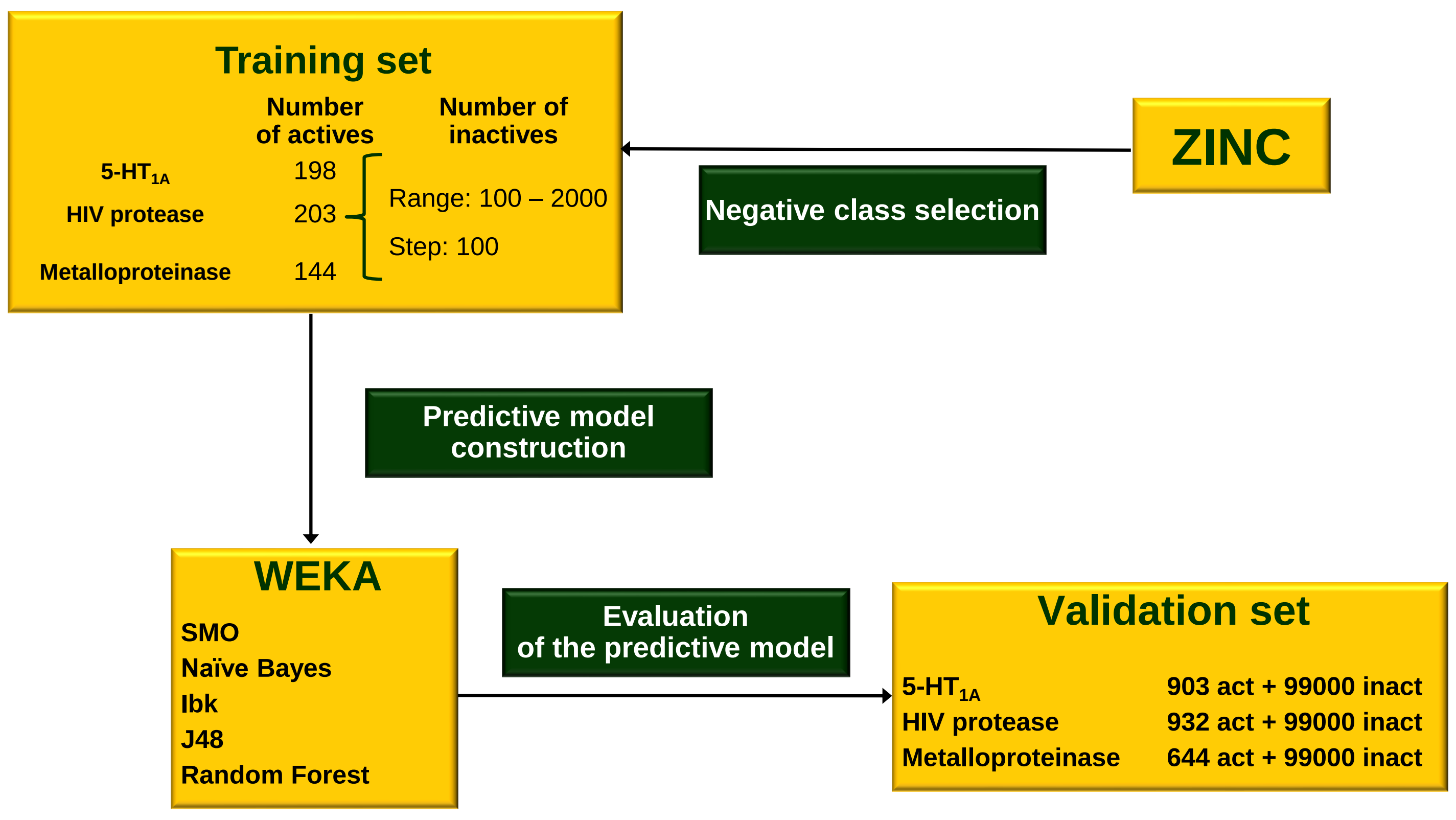
Naïve Bayes
SMO
Ibk
Decorate
Hyperpipes
Random Forest
J48

Databases of compounds with reported activity towards particular target usually contain only a few molecules which are proved to be inactive. Therefore, during the preparation for machine learning experiments, the need of generating sets of compounds assumed as inactives arises. In this study, six approaches of inactive molecules set formation were examined: random and diverse selection from ZINC, MDDR and DUD database. ML algorithms were tested in two separate experiments: with the use of one test set, the same for each method of inactive molecules generation, and with the use of test sets with inactives prepared in the analogous way as for training. All studies were performed for 5 different protein targets, with the use of 3 fingerprints for molecules representation and 7 ML algorithms with varying parameters.

It was proved that for virtual screening purposes, ML methods should be trained on sets with inactive molecules randomly selected out of the ZINC database, as it covers the chemical space to the highest extent out of all tested sources of assumed inactives. Strong limitations of chemical space both in MDDR and DUD databases, may be a source of difficulties for ML algorithms to properly identify active compounds, out of datasets with molecules of various structures and properties [2].

Determining the optimal size of the training set for the purpose of machine learning experiments

In parallel to the studies connected with determining an optimal way of inactive molecules generation, another set of experiments was designed in order to examine the influence of increasing number of inactives in the trainining set and determination of a ratio of actives to inactives providing the best performance of ML methods.



Training set

	Number of actives	Number of inactives
5-HT_{1A}	198	Range: 100 – 2000 Step: 100
HIV protease	203	
Metalloproteinase	144	

ZINC

Negative class selection

Predictive model construction

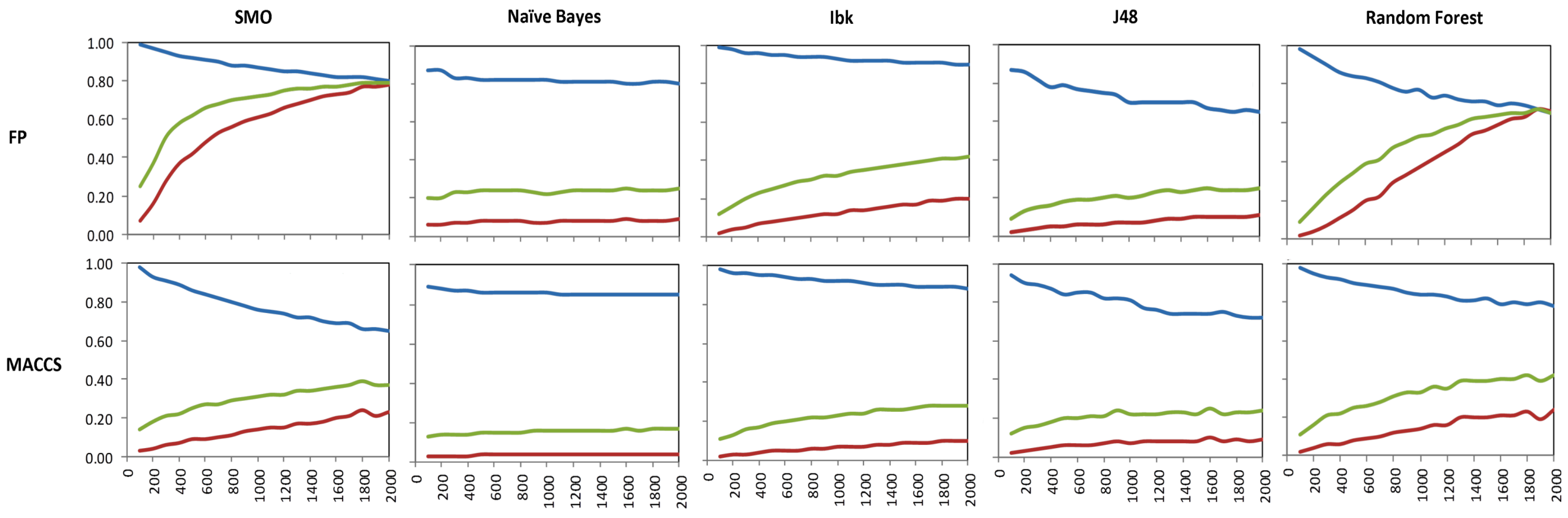
WEKA
SMO
Naïve Bayes
Ibk
J48
Random Forest

Evaluation of the predictive model

Validation set

5-HT_{1A}	903 act + 99000 inact
HIV protease	932 act + 99000 inact
Metalloproteinase	644 act + 99000 inact

In general, the increasing number of negative instances (with constant number of actives) led to improvement in classification efficiency – although, there was a slight decrease in recall, the values of precision and MCC describing the global performance of ML methods were increasing until the ratio of actives to inactives was ~1:9 and remained contant with further addition of inactive molecules to the training set. An exemplary graphs illustrating those dependencies for 5-HT_{1A} ligands are presented below (blue line corresponds to recall, red to precision and green to MCC values).

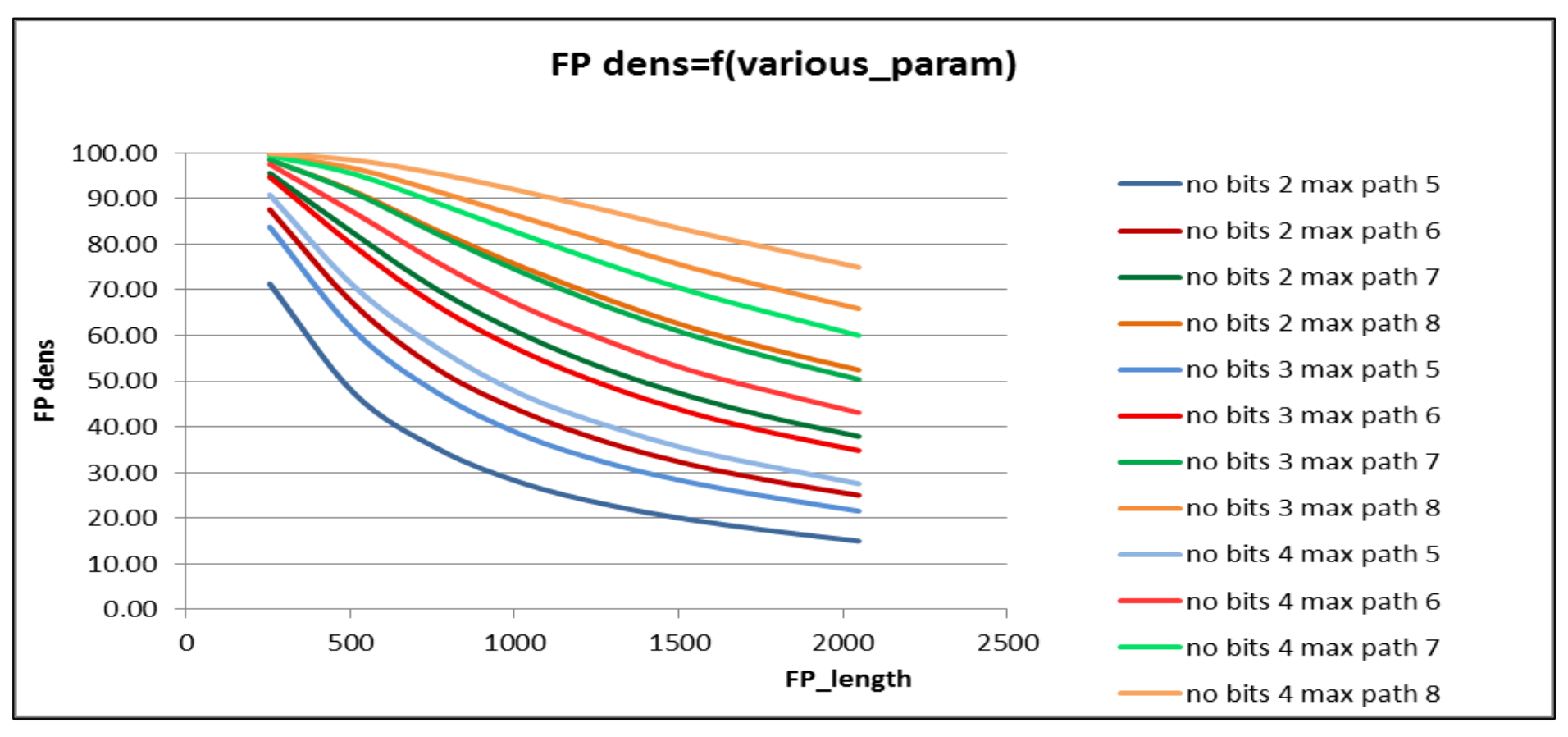


Examining the influence of fingerprint density on the effectiveness of classification carried out by machine learning methods

Fingerprint density (FP_dens) is a parameter that describes the percentage of ’1’ bits in the string. The aim of this part of the research was to study the impact of FP_dens on the performance of machine learning methods.

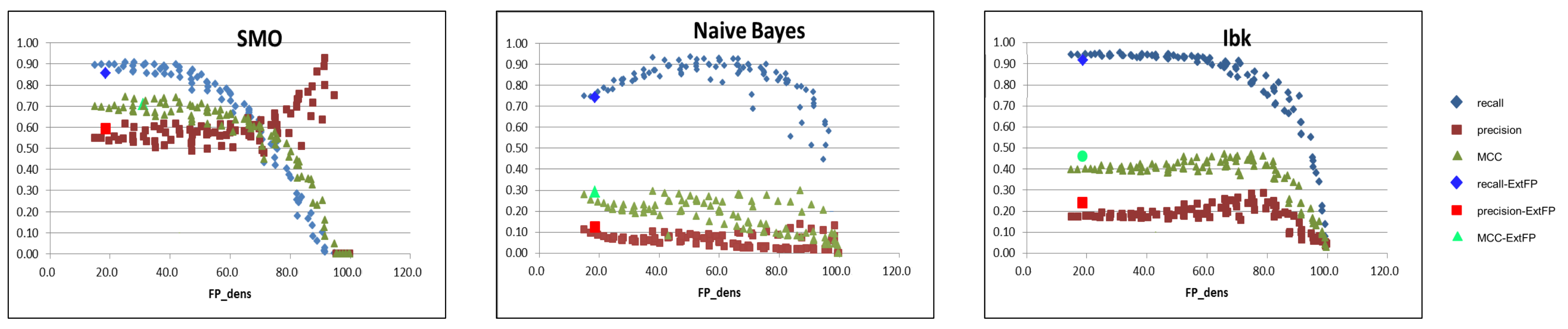
For this purpose, a series of fingerprints with different density values was generated by means of the RDKit software (for 5 protein targets, for sets with inactives exceeding in number ten times active compounds) [3]. The following parameters were changed during this process: fingerprint length, number of bits per hash and maximum path length, which eventually resulted in obtaining 96 different combinations of them.

FP length	Number of bits/hash	Maximum path length
256	2	5
512		
768	3	6
1024		
1280		7
1536	4	
1792		8
2048		



For each of the generated fingerprints, its density values were computed. It turned out that shortening of the fingerprint, increasing the number of bits per hash and extending the maximum path length lead to higher density values.

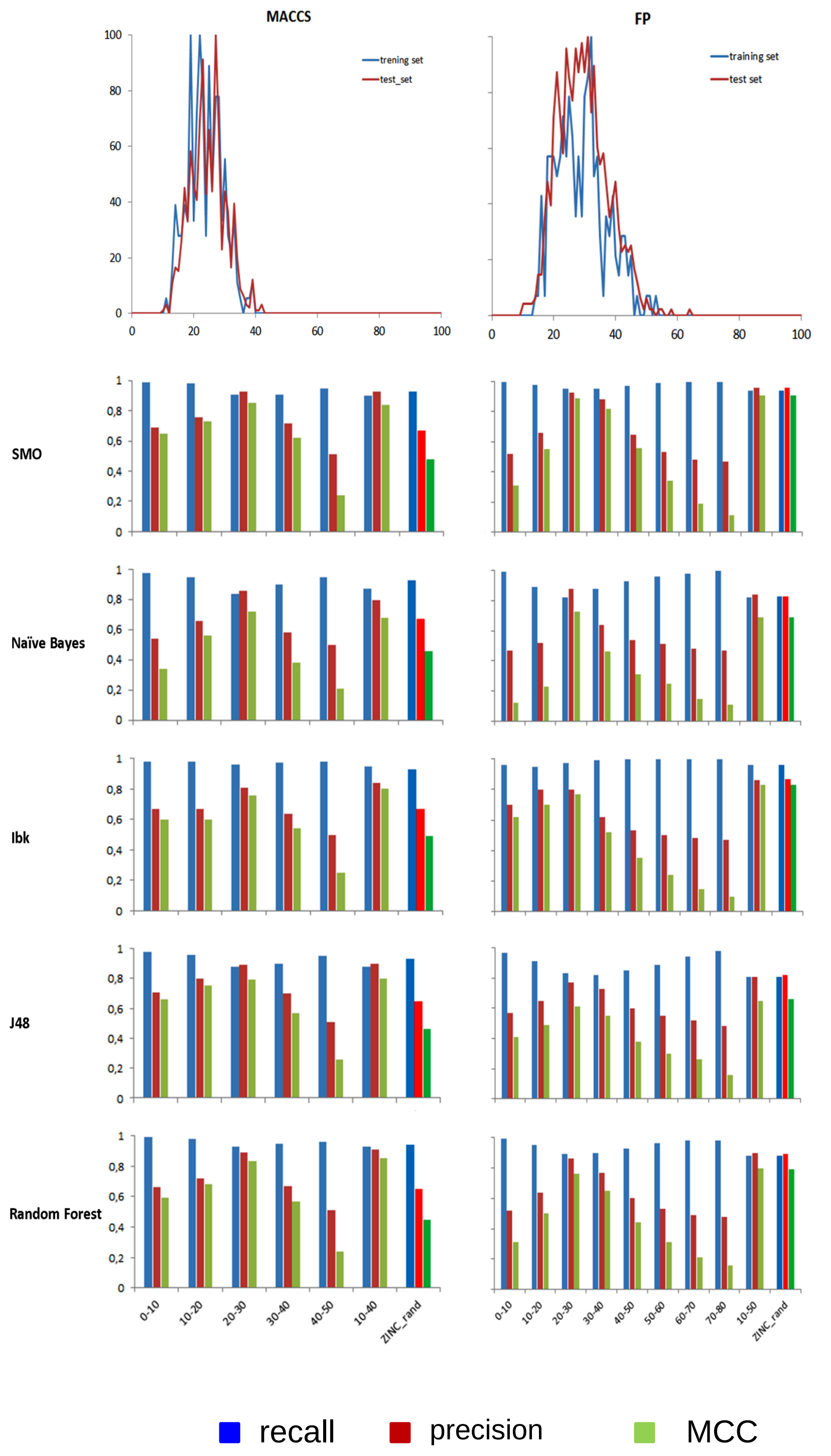
The impact of FP_dens on the effectiveness of the classification of 5-HT1A ligands is presented below. Results were also compared with those obtained for Extended Fingerprint, computed with the use of PaDEL-Descriptor, with fixed parameters (length: 1024, no bits/hash: s3, max path length: 6).



For Naïve Bayes, the most optimal FP_dens is ~60%, whereas for SMO, recall values were lower for high densities (especially when FP_dens was > 70%), whereas precision was improved by increasing number of ’1’ bits in the string. A global classification effectiveness expressed by MCC was the highest for FP_dens in range of ~30-50%. For Ibk, evaluating parameters values remained on quite stable level when FP_dens did not exceed 70%, significantly falling for higher FP_dens values.

The second part of the study of fingerprint density was performed on MACCSFP and FP. For the set of available active molecules towards 5 protein targets and whole ZINC database those fingerprints were calculated together with density values for each instance in the dataset. A series of training and test sets was generated on the basis of FP_dens: for particular experiment only those compounds were chosen that fall within the given FP_dens range. For comparison, all machine learning experiments were repeated for sets with randomly selected compounds and for sets with molecules described by strings of FP_dens in the a range where fall ~95% of them, keeping the size of training and test sets.

The distribution of FP_dens for particular fingerprint, as well as its influence on the performance of learning algorithms is presented in the figure (for 5-HT1A as an example). As it is shown, selecting sets of various FP_dens values influenced the classification effectiveness of learning algorithms. The best performance was provided by sets with FP_dens with the range of 20-30% and after removal from the dataset compounds with extreme percentage of ‘1’ bits.



References

[1] Melville, J. L.; Burke, E. K.; Hirst, J. D. Machine learning in virtual screening. *Combinatorial Chemistry & High Throughput Screening* **2009**, *12*, 332–43.

[2] Smusz, S.; Kurczab, R.; Bojarski, A. J. The influence of the inactives subset generation on the performance of machine learning methods”, Journal of Cheminformatics, **2013**, 5, 17–25

[3] [http://rdkit.org/]

Acknowledgements

The study was supported by a grant PRELUDIUM 2011/03/N/NZ2/02478 financed by the National Science Centre.