**P-46**

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

*Sabina Smusz, Rafał Kurczab, Andrzej J. Bojarski*

*Department of Medicinal Chemistry, Institute of Pharmacology Polish Academy of Sciences*
*Smętna 12, 31-343 Krakow*
*e-mail: smusz@if-pan.krakow.pl*

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.

Several aspects of parameters influencing the performance of ML methods were thoroughly examined. We tried to solve the problem of insufficient data on inactive compounds, by specifying the ways of generation of set of molecules that are assumed inactive and by analysing their impact on efficiency in classification of ML algorithms *(2)*. Another part of the study was connected with the composition of the training set (ratio of actives to inactives) *(3)* and with determination of the optimal fingerprint density (percentage of 1's) *(4)* in terms of their influence on ML methods performance. In order to provide the completeness of the study, we also examined the correlation between some physicochemical features of compounds, such as molecular weight, logP, number of particular atoms, solubility, molecular volume, etc.

**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

[3] Kurczab, R.; Smusz, S.; Bojarski, A. J. The influence of training actives/inactives ratio on machine learning performance. *Journal of Cheminformatics* **2013**, *5*, P30

[4] Smusz, S.; Kurczab, R.; Bojarski, A. J. The influence of hashed fingerprints density on the machine learning methods performance. *Journal of Cheminformatics* **2013**, *5*, P25

**Acknowledgements**

The study was supported by grant PRELUDIUM 011/03/N/NZ2/02478 financed by the National Science Center (www.ncn.gov.pl).