Improving Blood Brain Barrier Penetration *in silico* Models with a Hybrid Approach for Descriptor Selection.

Ana L. Teixeira,[a, b] Andre O. Falcao,[a]

a) LaSIGE, Departamento de Informática, Faculdade de Ciências, Universidade de Lisboa, Lisboa, Portugal; b) Centro de Química e Bioquímica, Faculdade de Ciências, Universidade de Lisboa, Lisboa, Portugal; e-mail: ateixeira@lasige.di.fc.ul.pt

In silico modeling of Blood-Brain Barrier (BBB) penetration for molecules is a difficult task due to the complexity of the BBB permeation process and also due to the incomplete and biased information available. To face such data issues a previous work used a Bayesian approach for modelling BBB penetration prediction [1]. In this work we present an important extension of the previous model by trying to ascertain which chemical descriptors are more relevant for prediction and, in the process, define the minimal subset of descriptors relevant for BBB penetration prediction modelling. The process followed used a hybrid two-phase approach, where initially the machine learning methodology of random forests is used for defining a hierarchical variable ranking, Then on a second phase, an exhaustive model search is performed based on support vector machines, where a progressively larger number of variables is added according to the predefined rank. The dataset for model training and cross-validation was comprised of 1850 molecules. Preliminary analysis of the model classification results strongly suggests that using a selected smaller number of chemical descriptors is better than using all available information and produces significantly better models. Nonetheless simplistic models with too few descriptors are not enough to produce reliable results. The optimal number of descriptors was found to be around 200, which produced cross-validated results with an expected Mean Squared Contingency Coefficient (MSCC) of 0.687, and overall accuracy of 90.1%, thus definitely superior to the results obtained when using all descriptors: MSCC=0.431 and accuracy = 88.6%.

[1] I. F. Martins, A. L. Teixeira, L. Pinheiro, A. O. Falcao, *J. Chem. Inf. Model.* **2012**, *52*, 1686-1697.

ALT gratefully acknowledges Fundação para a Ciência e a Tecnologia (FCT) for a doctoral grant (SFRH/BD/64487/2009). AOF acknowledges the Multiannual Funding Programme of LaSIGE by FCT.