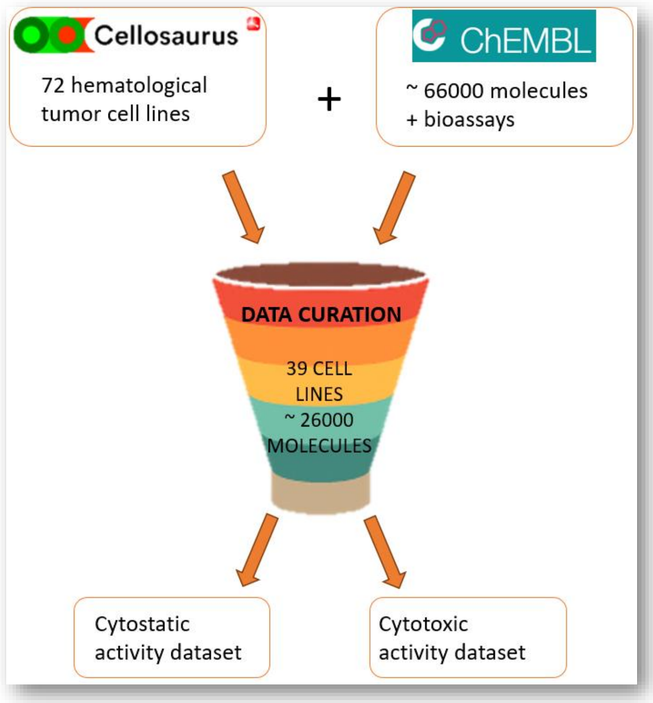
A MULTI-TARGET *IN SILICO* MODEL FOR THE DISCOVERY OF ANTI HEMATOLOGICAL CANCERS DRUGS

Valentina Pellicioni[1], Amit Kumar Halder[2], Gustavo Marfia[1], Maria Natalia Dias Soeiro Cordeiro[2], Carmela Fimognari[1]

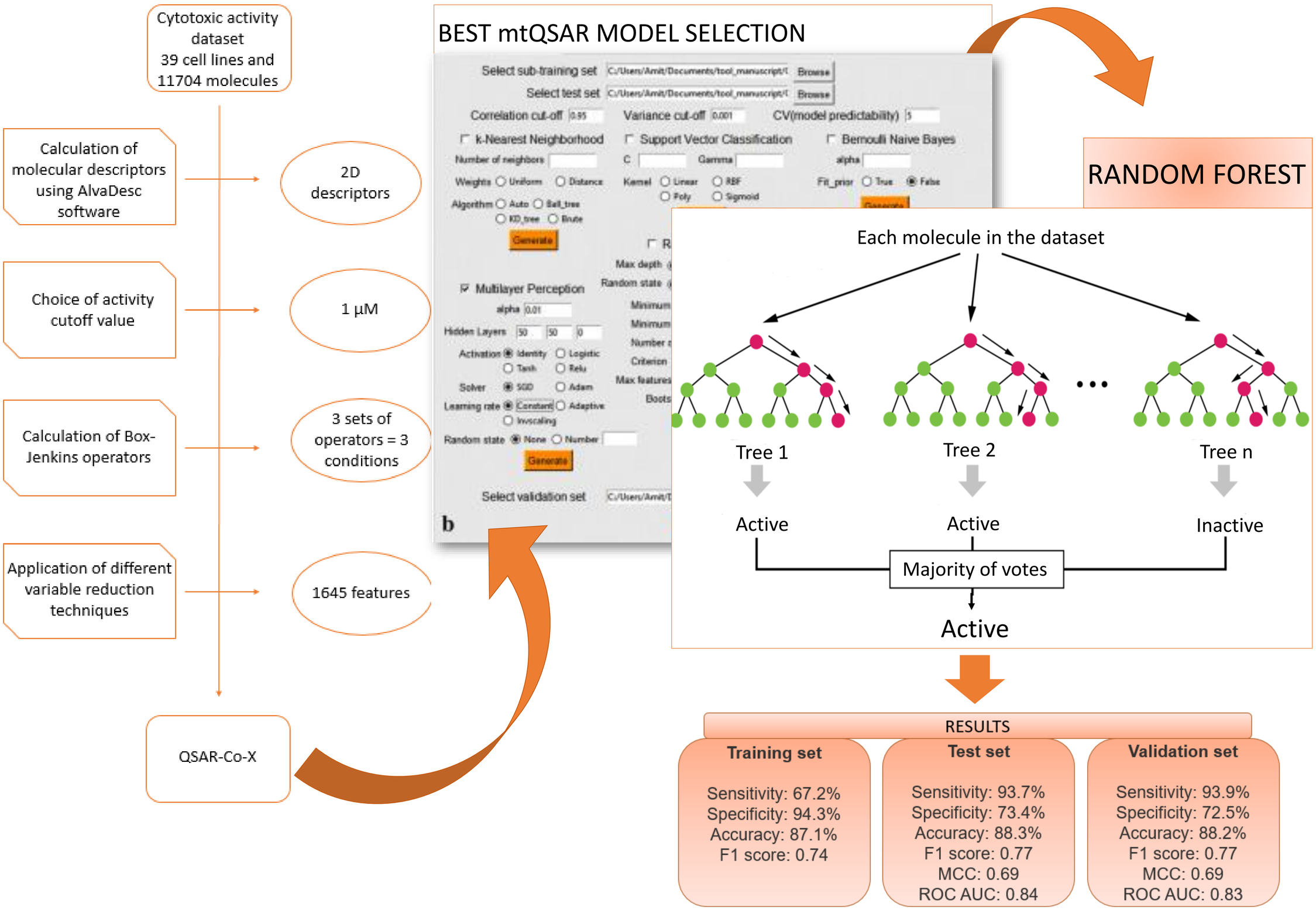
[1] Department for Life Quality Studies, Alma Mater Studiorum - University of Bologna, corso d'Augusto 237, 47921 Rimini, Italy
[2] LAQV@REQUIMTE/Faculty of Sciences, University of Porto, Rua do Campo Alegre, s/n, 4169-007 Porto, Portugal

INTRODUCTION: Haematological cancers are divided into leukaemias, lymphomas and myelomas, each comprising a large number of subtypes. They are often called "liquid tumours" because they do not form nodules or masses, as do "solid tumours". Because of this, they cannot be surgically removed and the mainstay of their treatment is chemotherapy. The development of new anticancer drugs remains a challenging, lengthy and expensive process: the average time to develop them is 7.3 years and the average cost is $648.0 million [1]. Over the last 15 years, biological and chemical research has produced an enormous amount of data that has been digitised, most of which is freely accessible because it is contained within open access databases available online (ChEMBL, etc.). Thanks to this, the modern drug discovery process has embraced the era of Big Data [2]. The use of Big Data has transformed the way data on chemical molecules is extracted and used in research. Central to this change has been the aid of artificial intelligence approaches, such as machine learning and deep learning algorithms, which have been successfully applied within Computer Aided Drug Design (CADD). By combining Big Data, CADD and artificial intelligence, it is possible to create computational models capable of predicting specific biological activity. These models can then be used for virtual screening, which allows to identify new molecules with the desired activity and exclude those without such activity and/or with undesirable side effects. This creates a 'bottleneck' process that leads to the testing of the most promising molecules only, making the drug discovery process faster, cheaper and more sustainable.



OBJECTIVE: The aim of the study was to create a multi-target Quantitative Structure-Activity Relationship (mt-QSAR) classification model, using machine learning techniques, for the prediction of cytotoxic drugs simultaneously active against leukaemia, lymphoma and myeloma cell lines.

MATERIALS AND METHODS: Using data from the Cellosaurus and ChEMBL databases, a dataset of 11704 molecules tested against 39 cell lines was extrapolated, in which cytotoxic activity was reported as IC50 (concentration capable of inhibiting cell viability by 50%). Although bioactivity assays data all come from ChEMBL database, they were obtained from different research groups and thus ensure a wide experimental diversity. For each molecule, a set of 2D descriptors were calculated using AlvaDesc software. To discriminate active from inactive molecules, a cutoff value of 1 µM was chosen. This mt-QSAR model was created using the Box-Jenkins moving average approach, which allows multiple experimental assay conditions to be incorporated into a single model for simultaneous activity prediction. The dataset thus constructed was subjected to several machine learning techniques using QSAR-Co-X software in order to identify the one that would yield the best predictive model. This technique turned out to be Random Forest.



RESULTS: The result is a model with good predictive ability, as shown by the following statistical metrics: 1. accuracy, greater than 88%, 2. Matthews correlation coefficient (MCC), equal to 0.69, and 3. the area under receiver operating characteristic curve (ROC AUC), greater than 0.83 in both the test set and the validation set.

DISCUSSION: In accordance with these results, the above-mentioned mt-QSAR model demonstrates that it is not a random classifier and can be effectively used, through virtual screening, to search for molecules active against different haematological cancer cell lines within laboratory or commercial online databases.

Bibliografia:
1. Prasad V, Mailankody S. Research and Development Spending to Bring a Single Cancer Drug to Market and Revenues After Approval. *JAMA Intern Med*. 2017;177(11):1569–75
2. Zhu H. Big Data and Artificial Intelligence Modeling for Drug Discovery. *Annu Rev Pharmacol Toxicol*. 2020;60:573-589.