**Supplementary Material**

**Database analysis**

**Ligand preparation**

Datasets on SMILES format were converted to 3D structures using LigPrep[1] software implemented on Maestro Suite[2].LigPrep is a 2D-to-3D conversion tool that includes the addition of hydrogen atoms and options for generating multiple possible tautomers, stereoisomers, ionization at a selected pH range, and ring conformations using molecular mechanics force fields. To carry out our studies, possible ionizations were generated at pH 7.3 in order to obtain the most suitable ionization states of the compounds for that pH range. The ionization states were assigned with Epik module[3]. Also, all the compounds were desalted and no tautomers were generated. In this process, we have restricted the search to obtain just one possible stereoisomer among all that can be found by the program, as well as one low energy ring conformation. The final step of a LigPrep preparation is an energy minimization of the 3D conformers generated using the OPLS 2005 force field[4].

Different conformers and ionization states of the same compounds were reduced in order to keep one 3D structure per initial compound. The selection was made considering the most probable ionization state at physiological pH conditions. This preparation is a crucial step for the following studies and was performed with the aim of obtaining the most suitable 3D structures to further calculate the physicochemical properties of the existing compounds.

**Drug-like properties calculation**

All the prepared compounds were analyzed using Qikprop[5] module of the Small-Molecule Drug Discovery Suite in Schrödinger, an accuratesoftware that predicts structurally significant 2D and 3D descriptors and pharmaceutically relevant properties of organic molecules. Absorption, Distribution, Metabolism, and Excretion (ADME) A total of 44 properties could be predicted. Among all the properties, the program calculates properties like molecular weight, QPlogPo/w (predicted octanol/water partition coefficient), molecular volume, number of H-bond donors or acceptors, polar surface area and violations related to the Lipinski’s Rule of 5 and Jorgensen’s Rule of 3. This allow to filter out compounds with clear cut undesirable properties for drug discovery.

**Model building**

The first step consisted in dividing the dataset into a training (75%) and a test set (25%) following a stratified division. The next step is based on computing the molecular descriptors of all the dataset using DRAGON software[6]. DRAGON is an application tool that provides over 3,000 diverse molecular descriptors (0D, 1D, 2D and 3D) which can be used to evaluate molecular QSAR or QSPR of different databases. To calculate these molecular descriptors, molecules were prepared as it is described in the ligand preparation section. After computing the descriptors, those descriptors which present over 0.95 correlation were deleted. Then, the next stage that takes place is the feature selection procedure and it is performed with DELPHOS[7]. This tool infers multiple alternative selections of molecular descriptors computed by DRAGON for defining a QSAR model. In this work, a total of twenty-five putative subsets have been computed, selecting the best 4 models in terms of MAE and MSE to be further selected to build QSAR models. In the case of the classification models, the target property was discretizated for defining classes using the following threshold: Low activity ≤ 50% > High activity.

Finally, in order to build the models, WEKA software[8] was used. This toolis a collection of machine learning algorithms for data mining tasks. For each inference method, the parameter settings provided by default for WEKA were used in the experiments. The algorithms used in this study are described as follows:

Neural Networks (Multiperceptron): A classifier that uses back-propagation to classify instances. This network can be built by hand, created by an algorithm or both. The network can also be monitored and modified during training time. The nodes in this network are all sigmoid (except for when the class is numeric in which case the output nodes become unthresholded linear units).

Random Forest: class for constructing a forest of random trees. The random trees for constructing a tree that considers K randomly chosen attributes at each node. Also, it has an option to allow estimation of class probabilities (or target mean in the regression case) based on a hold-out set (back fitting).

Random Committee: class for building an ensemble of randomized base classifiers. Each base classifier is built using a different random number seed (but based on the same data). The final prediction is a straight average of the predictions generated by the individual base classifiers.

Metrics of the best models are reported in the work and the descriptors present in the best models are analyzed in terms of correlation using VIDEAN tool.

References

1. LigPrep 3.1; Schrödinger, L., New York: 2014.

2. Maestro 9.9; Schrödinger, L., New York: 2014.

3. Epik, v., Schrödinger, L., New York: 2014.

4. Jorgensen, W.L., Tirado-Rives, J.: The OPLS [optimized potentials for liquid simulations] potential functions for proteins, energy minimizations for crystals of cyclic peptides and crambin. J Am Chem Soc 110,1657-1666 (1988).

5. QikProp, v.S., LLC, New York: 2014.

6. Dragon, V., Talete srl (2007).

7. Soto, A.J., Martínez, M. J., Cecchini, R. L., Vazquez, G. E. & Ponzoni, I.: DELPHOS: Computational Tool for Selection of Relevant Descriptor Subsets in ADMET Prediction. 1st International Meeting of Pharmaceutical Sciences (2010).

8. Eibe Frank, M.A.H., and Ian H. Witten (2016). The WEKA Workbench. Online Appendix for "Data Mining: PracticalMachine Learning Tools and Techniques", Morgan Kaufmann, Fourth Edition, 2016.