SUPPLEMENT A

Table A List of RDKit descriptors used for the dataset analysis

SlogP	Smarts LogP, Octanol Water Partition Coefficient
LabuteASA	Labute's Approximate Surface Area, approximated surface area of a molecule (Labute, 2000)
TPSA	Total Polar surface area
ExactMW	Molecular weight
NumRotatableBonds	Number of rotatable bonds
NumHBD	Number of hydrogen bond donors
NumHBA	Number of hydrogen bond acceptors
NumAmideBonds	Number of amide bonds
NumHeteroAtoms	Number of hetero atoms
NumHeavyAtoms	Number of heavy atoms
NumAtoms	Number of atoms
NumRings	Number of rings
NumAromaticRings	Number of aromatic rings
NumSaturatedRings	Number of saturated rings
NumAliphaticRings	Number of aliphatic rings
NumAromaticHeterocycles	Number of aromatic heterocycles
NumSaturatedHeterocycles	Number of saturated heterocycles
NumAliphaticHeterocycles	Number of aliphatic heterocycles
NumAromaticCarbocycles	Number of aromatic carbocycles

NumSaturatedCarbocycles	Number of saturated carbocycles
NumAliphaticCarbocycles	Number of aliphatic carbocycles
FractionCSP3	Fraction of sp3 hybridized Carbons
SMR	Molecular refractivity

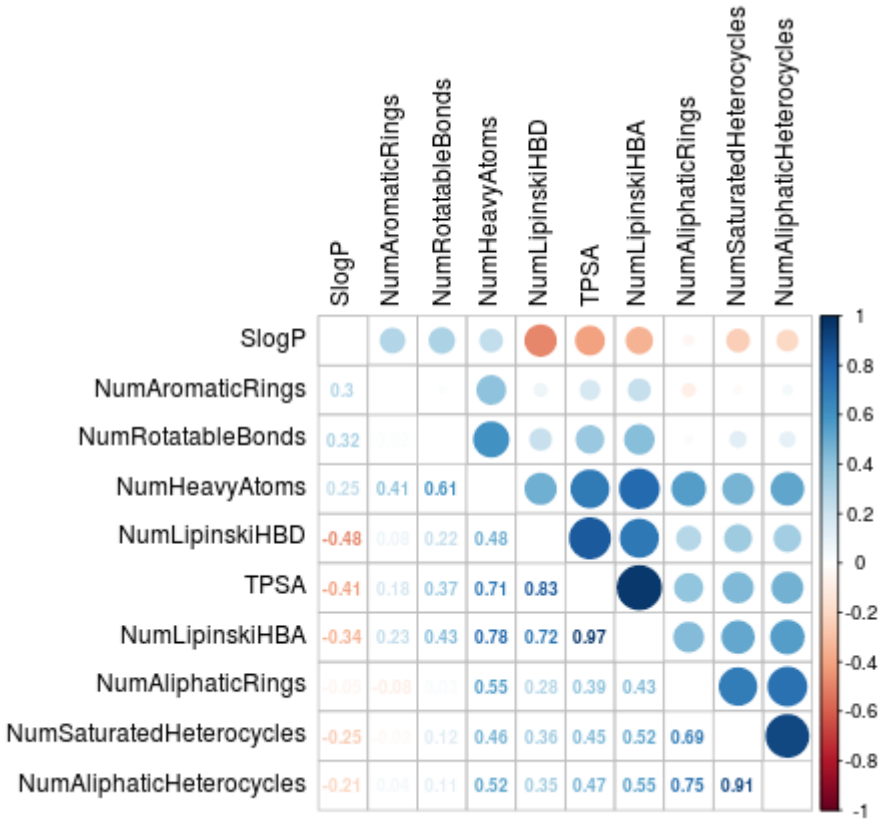

Figure A Correlation matrix of the descriptors used for the random forest

The size and color of the dots in the upper triangle represents the correlation of the respective. Larger dots indicate a higher contribution. The lower triangle represents the correlation as a numeric vale.

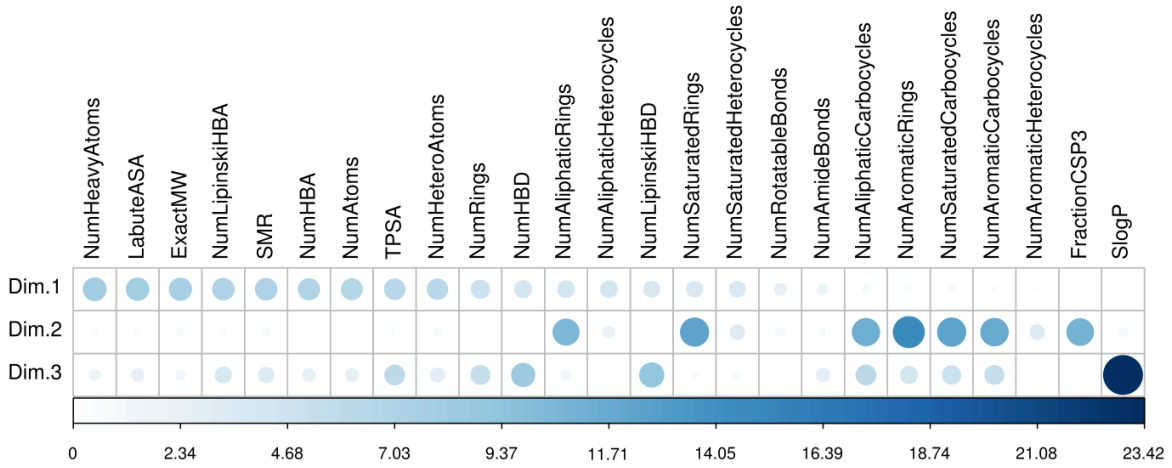

Figure B PCA contribution of variables.
The size and color of the dots represents the contribution of the respective descriptor to the principal component. Larger dots indicate a higher contribution.