Electronic Supplementary Material (ESI) for CrystEngComm.

Increasing the performance, trustworthiness and practical value of machine learning models: a case study predicting hydrogen bond network dimensionalities from molecular diagrams

Supplementary Information

A. List of molecular descriptors calculated using RDkit package. Marked in bold are the two descriptors that were discarded during the data pre-processing stage, because their values were constant across all dataset examples.

MolWt
ExactMolWt
HeavyAtomMolWt
NumHAcceptors
NumHDonors
NumRotatableBonds
FractionCSP3
NHOHCount
NOCount
NumHeteroatoms
HeavyAtomCount
NumAliphaticCarbocycles
NumAromaticCarbocycles
NumSaturatedCarbocycles
NumAliphaticHeterocycles
NumAromaticHeterocycles
NumSaturatedHeterocycles
RingCount
NumAliphaticRings
NumAromaticRings
NumSaturatedRings
NumValenceElectrons
NumRadicalElectrons
qed
MolLogP
MolMR
BalabanJ
MaxAbsPartialCharge
MinAbsPartialCharge
MaxPartialCharge
MinPartialCharge
MaxAbsEStateIndex
MinAbsEStateIndex
MaxEStateIndex
MinEStateIndex
PEOE_VSA1
PEOE_VSA10
PEOE_VSA11
PEOE_VSA12
PEOE_VSA13
PEOE_VSA14
PEOE_VSA2
PEOE_VSA3
PEOE_VSA4
PEOE_VSA5
PEOE_VSA6
PEOE_VSA7
PEOE_VSA8
PEOE_VSA9
SMR_VSA1
SMR_VSA10
SMR_VSA2
SMR_VSA3
SMR_VSA4
SMR_VSA5
SMR_VSA6
SMR_VSA7
SMR_VSA8
SMR_VSA9
SlogP_VSA1
SlogP_VSA10
SlogP_VSA11
SlogP_VSA12
SlogP_VSA2
SlogP_VSA3
SlogP_VSA4
SlogP_VSA5
SlogP_VSA6
SlogP_VSA7
SlogP_VSA8
SlogP_VSA9
EState_VSA1
EState_VSA2
EState_VSA3
EState_VSA4
EState_VSA5
EState_VSA6
EState_VSA7
EState_VSA8
EState_VSA9
EState_VSA10
EState_VSA11
VSA_EState1
VSA_EState2
VSA_EState3
VSA_EState4
VSA_EState5
VSA_EState6
VSA_EState7
VSA_EState8
VSA_EState9
VSA_EState10
LabuteASA
Chi0
Chi0n
Chi0v
Chi1
Chi1n
Chi1v
Chi2n
Chi2v
Chi3n
Chi3v
Chi4n
Chi4v
HallKierAlpha
Kappa1
Kappa2
Kappa3
BertzCT
TPSA
Ipc
FpDensityMorgan1
FpDensityMorgan2
FpDensityMorgan3

B. Statistical methods considered for model generation, along with optimised parameters and mean 5-fold cross validation accuracy.

Method	Optimal Parameters	Accuracy
Random Forest	max_depth: 30, n_estimators: 100	0.57
SVM Linear kernel	C: 50	0.57
SVM RBF kernel	C: 50, gamma: 0.001	0.58
SVM Sigmoid kernel	gamma: 0.001	0.53
Logistic Regression (one *vs* rest)	C: 200	0.54
Logistic Regression multinomial	C: 100	0.56
K Nearest Neighbours	n_neighbors: 50	0.51
Gaussian Naive Bayes	var_smoothing: 0.01	0.43
Gaussian Process (one *vs* rest)	random_state: 0	0.47
Gradient Boosting Classifier	learning_rate: 0.1, max_depth: 10, n_estimators: 100	0.57