Electronic Supplementary Material (ESI) for CrystEngComm.

A Random Forest Model for Predicting the Crystallisability of Organic Molecules

Rajni M. Bhardwaj,[a] Andrea Johnston,[a] Blair F. Johnston,[a] and Alastair J. Florence[a] *

[a]Strathclyde Institute of Pharmacy and Biomedical Sciences, University of Strathclyde, 161 Cathedral Street, Glasgow G4 0RE, U.K
* To whom correspondence should be addressed. E-mail: alastair.florence@strath.ac.uk

Table of Contents

1. Schemtaic Workflow of Random Forest

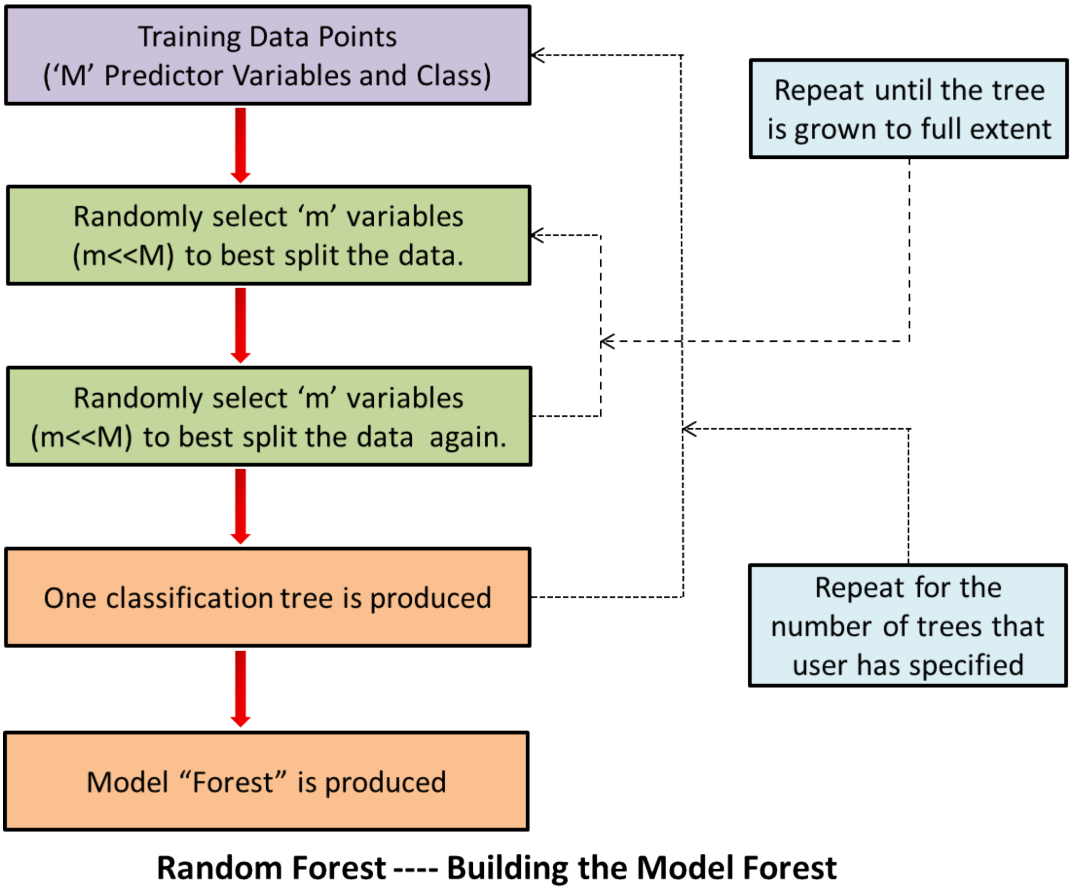
Training Data Points
('M' Predictor Variables and Class)
Randomly select 'm' variables
(m<<M) to best split the data.
Randomly select 'm' variables
(m<<M) to best split the data again.
One classification tree is produced
Model "Forest" is produced
Repeat until the tree is grown to full extent
Repeat for the number of trees that user has specified
Random Forest ---- Building the Model Forest

Fig. 1 Schematic workflow of building of Random Forests model.

Bootstrap sampling and random selection of input descriptors are used to induce randomness in the input data used to develop the RF model. This ensures that the classification trees grown in the forest are dissimilar and not correlated to each other. Using bootstrap sampling, classification trees are grown using 2/3rd of the dataset and remaining 1/3rd of the dataset [Out Of Bag (OOB) data] is employed to obtain unbiased estimates of correct classification rates (internal estimates of error). Compared to a single classification tree, this algorithm yields better prediction rates and is more robust in dealing with noise in the data set because the forest of trees are grown to the full extent. The generalisation error of a forest of trees classifiers depends on the strength of the individual tree in the forest and the correlation between them.

2. Preparation of the Dataset

All the molecules were drawn using Chemdraw Ultra (version 11.0) and reliable 3-D conformations generated in Discovery Studio using the Pipeline Pilot interface (Accelrys, 2010). 2-D (185) and internal 3-D (i3-D) (123) molecular descriptors were calculated using MOE.[1] 2-D molecular descriptors are defined to be numerical properties and calculated from the atoms and connection table of the molecule. 3-D molecular descriptors can be classified in two categories: one that depend on internal coordinates only and 2nd that depend on absolute orientation of molecule. A brief explanation of the list of calculated 2- and 3-D molecular descriptors which were used to model the solvent library are given in Table 1 (Source: http://www.chemcomp.com/journal/descr.htm).

Table 1 Molecular descriptors and brief explanation that were calculated for solvent molecules.

Descriptors	**Category**	
2-D descriptors		
apol, bpol, Fcharge, mr, SMR, Weight, logP (o/w), SlogP, vdw_vol, density, vdw-area	physical properties	Physical properties are calculated from the connection table of a molecule
SlogP_VSA0-SlogP_VSA9, SMR_VSA0 - SMR_VSA7	subdivided surface areas	The Subdivided Surface Areas are descriptors based on an approximate accessible van der Waals surface area calculation for each atom, v_i along with some other atomic property, p_i.
a_aro, a_count, a_heavy, a_ICM, a_IC, a_nH, a_nB, a_nC, a_nN, a_nO, a_nF, a_nP, a_nS, a_nCl, a_nBr, a_nI, b_1rotN, b_1rotR, b_ar, b_count, b_double, b_heavy, b-rotN, b_rotR, b_single, b_triple, VAdjMa, VAdjEq	atom count and bond count	The atom count and bond count descriptors are functions of the counts of atoms and bonds
chi0, chi0_C, chi1, chi1_C, chi0v, chi0v_C, chi1v, chi1v_C, Kier1 - Kier3, KierA1 - KierA3, KierFlex, zagreb	Kier&Hall Connectivity and Kappa Shape Indices	The Kier and Hall kappa molecular shape indices compare the molecular graph with minimal and maximal molecular graphs, and are intended to capture different aspects of molecular shape.
balabanJ, diameter, petitjean, radius, VDistEq, VDistMa, weinerPath, weinerPol	Adjacency and Distance Matrix Descriptors	The adjacency matrix, M, of a chemical structure is defined by the elements [Mij] where Mij is 1 if atoms i and j are bonded and zero otherwise. The distance matrix, D, of a chemical structure is defined by the elements [Dij] where Dij is the length of the shortest path from atoms i to j; zero is used if atoms i and j are not part of the same connected component.
a_acc, a_acid, a_base, a_don, a_hyd, vsa_acc, vsa_acid, vsa_base, vsa_don, vsa_hyd, vsa_other, vsa_pol	Pharmacophore Feature Descriptors	The Pharmacophore Atom Type descriptors consider only the heavy atoms of a molecule and assign a type to each atom
Q_PC+ PEOE_PC+, Q_PC- PEOE_PC-, Q_RPC+ PEOE_RPC+, Q_RPC- PEOE_RPC-, Q_VSA_POS PEOE_VSA_POS, Q_VSA_NEG, PEOE_VSA_NEG, Q_VSA_PPOS, PEOE_VSA_PPOS, Q_VSA_PNEG PEOE_VSA_PNEG, Q_VSA_HYD PEOE_VSA_HYD, Q_VSA_POL PEOE_VSA_POL, Q_VSA_FPOS PEOE_VSA_FPOS, Q_VSA_FNEG PEOE_VSA_FNEG, Q_VSA_FPPOS PEOE_VSA_FPPOS, Q_VSA_FPNEG PEOE_VSA_FPNEG, Q_VSA_FHYD PEOE_VSA_FHYD, Q_VSA_FPOL PEOE_VSA_FPOL, PEOE_VSA+6 - PEOE_VSA+0, PEOE_VSA-0 - PEOE_VSA-6	Partial Charge Descriptors	Descriptors that depend on the partial charge of each atom of a chemical structure require calculation of those partial charges.
3-D Descriptors		

E, E_ang, E_ele, E_nb, E_oop, E_sol, E_stb, E_str, E_strain, E_tor, E_vdw, E_rele, E_rsol, E_rvdw	Potential Energy Descriptors	The energy descriptors use the MOE potential energy model to calculate energetic quantities from stored 3D conformations.
ASA, dens, glob, pmi, pmiX, pmiY, pmiZ, rgyr, std_dim1 –std_dim3, vol, VSA	Surface Area, Volume and Shape Descriptors	Descriptors depend on the structure connectivity and conformation
ASA+, ASA-, ASA_H, ASA_P, DASA, CASA+, CASA-, DCASA, dipole, diploeX, dipole, dipoleZ, FASA+, FASA-, FCASA+, FCASA-, FCASA_H, FCASA_P	Conformation Dependent Charge Descriptors	Descriptors depend upon the stored partial charges of the molecules and their conformations.

A correlation matrix was prepared using a Pearson correlation coefficient by using a Pipeline Pilot interface. Molecular descriptors which showed zero variance and covariance (threshold of Pearson correlation coefficient >90%) were removed from the dataset. The resultant dataset comprised of 151 calculated molecular descriptors.

3. Error Plot

The error plot in Fig. 2 provides an overall OOB error of prediction and prediction accuracy for the classification model and a confusion matrix in Table 2 provides information on the prediction accuracy and OOB error rate associated with each class.

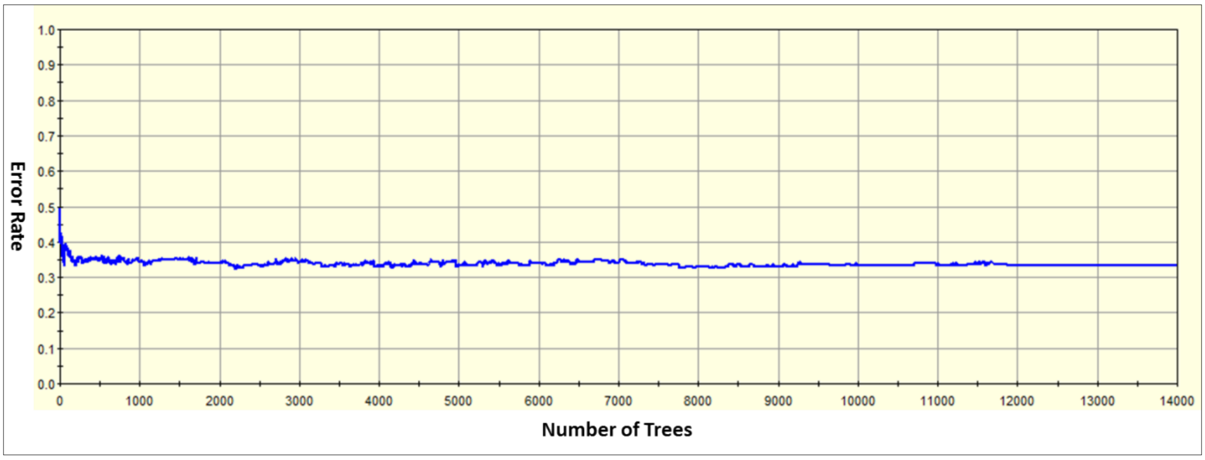

Fig. 2 Error plot for the Random Forests classification model trained using 2-D and 3-D molecular descriptors of the molecules present in the dataset. Blue line shows the evolvement of overall OOB error of prediction with the addition of number of trees.

Table 2 Confusion matrix generated by Random Forests for classification of dataset of 382 molecules. Class 1 represents the molecules which crystallised and class 2 represents the molecules which did not crystallise.

Actual Class	Total Cases	% Correct	1, N=231	2, N=151	Class Error
1	303	67.7	205	98	32.34
2	79	67.1	26	53	32.91

4. Important Descriptors Assessment

The RF algorithm also assesses the importance of descriptors used in building of the classification model. It is assessed by replacing each descriptor in turn by random noise and the resulting deterioration in the model quality is a measure of descriptor importance. The deterioration in the RF model quality is assessed by mean decrease in accuracy (based on OOB data).

Table 2. Ten most important descriptors selected by Random Forests for classification model. Reproduced from web reference http://www.juit.ac.in/attachments/podophyllotoxin/introduction.htm

Descriptor	**Rank**	**Descriptor Definition**	**Further Details of Descriptors**
E_TOR	1	Torsion (proper and improper) potential energy	The energy descriptors use the MOE potential energy model to calculate energetic quantities from stored 3D conformations
E_VDW	2	van der Waals component of the potential energy	
GCUT_SMR_0	3	Atomic contribution to molar refractivity	The GCUT descriptors using atomic contribution to molar refractivity instead of partial charge
AM1_EELE	4	Electronic energy calculated using the AM1 Hamiltonian	Can be calculated from the connection table (with no dependence on conformation) of a molecule
BCUT_PEOE_3	5	Calculated from Adjacency Matrix of a chemical structure	It is calculated from the eigenvalues of a modified adjacency matrix.
B_1ROTR	6	Fraction of rotatable single bonds: b_1rotN divided by b_count.	The atom count and bond count descriptors are functions of the counts of atoms and bonds
VSURF_CW2	7	Capacity factor	Depend on the structure connectivity and conformation
E	8	Value of the potential energy	Same description as for descriptors ranked from 1-2
B_ROTR	9	Fraction of rotatable bonds: b_rotN divided by b_count.	Same description as for descriptors ranked at 6
GCUT_SLOGP_1	10	Atomic contribution to logP	The GCUT descriptors using atomic contribution to logP instead of partial charge.

5. References

1) MOE, 2002, Chemical Computing Group, 1010 Sherbrooke St. W, Montreal, Quebec, H3A 2R7.