Electronic Supplementary Material (ESI) for CrystEngComm.

Predicting Crystal Packing of Olanzapine Solvates using Random Forest

Rajni M. Bhardwaj,[a] Susan Reutzel-Edens,[b] 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, United Kingdom

[b] Eli Lilly and Company, Indianapolis, Indiana 46285, United States

* To whom correspondence should be addressed. E-mail: alastair.florence@strath.ac.uk

Table of Contents

Contents

1. Schematic Workflow of Random Forest 2
2. Crystal Packing in solvates based on SC_{31}, SC_{32} and SC_{33} 3
3. Preparation of the Dataset 3
4. Model Building 6
5. Important Descriptors Assessment 8
6. References 9

1. Schematic Workflow of Random Forest

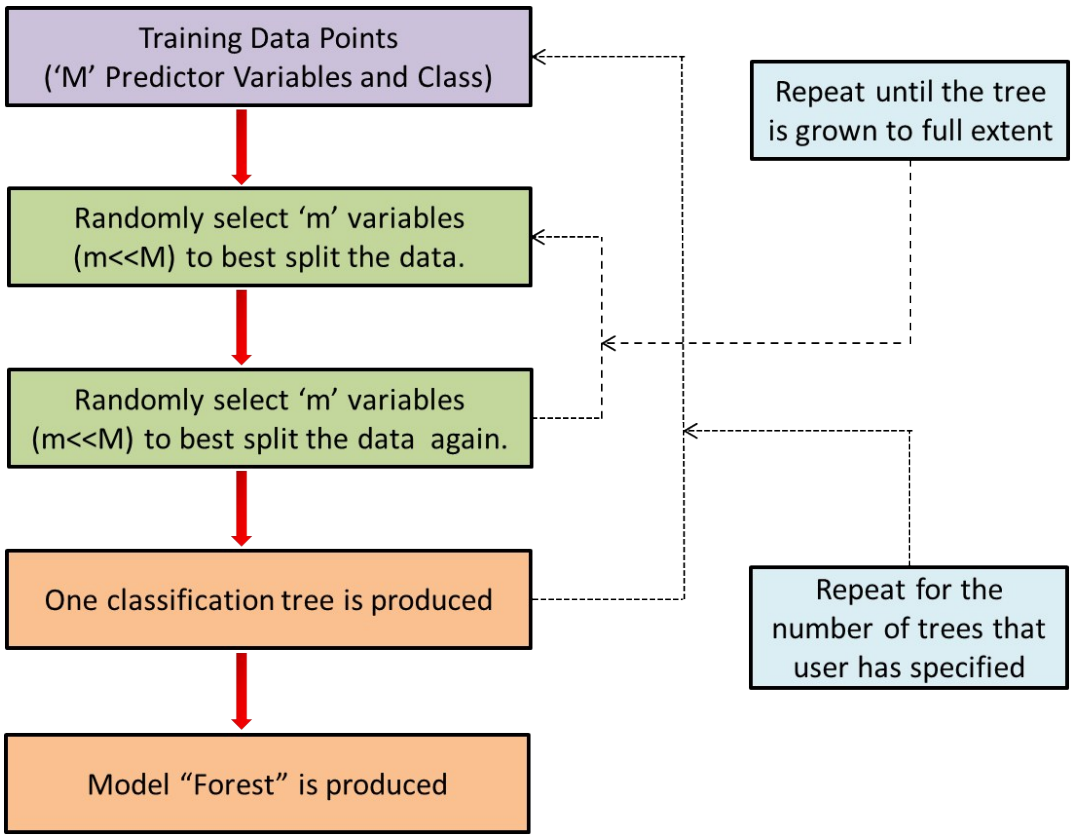

Fig. 1. Schematic workflow of building of Random Forests model.

.

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. Crystal Packing in solvates based on SC_{31}, SC_{32} and SC_{33}

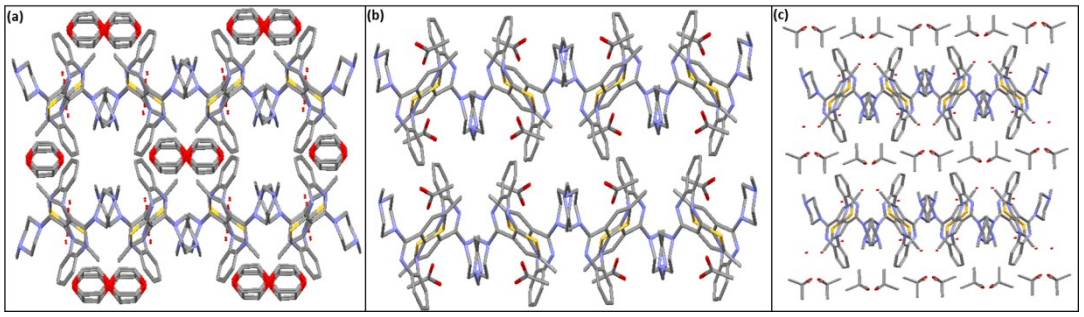

Fig. 2. Representative crystal structures of olanzapine solvates based on (a) SC_{31}- olanzapine-1,4-dioxane solvate hydrate; (b) SC_{32}-Olanzapine-acetic acid solvate and (c) SC_{33}-Olanzapine-TBMe hydrate; showing the position of solvent molecules within the crystal lattice

3. Preparation of the Dataset

All the solvent 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.

4. Model Building

Out of 35 OZPN solvates, 28 are based on a common 2-D SC_{21} and differ only in 3-D originated by different stacking of SC_{21} sheets. A RF model was built to confirm the role of solvent molecules in directing the crystal packings of these 28 OZPN solvates in 3-D.

25 out of 28 solvates were taken for random forest classification as there were no calculated descriptors for 3 solvents (1,4-butanediol, t-butanol and ethylene glycol (EG). OZPN forms two dihydrates i.e. dihydrate B and dihydrate E which fall in class 2 and class 1 respectively. Control over hydrate (dihydrate B vs. dihydrate E) formation can be achieved via appropriate crystallisation conditions. For this classification, hydrate was considered only in one class.

The error plot in Fig. 3 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.

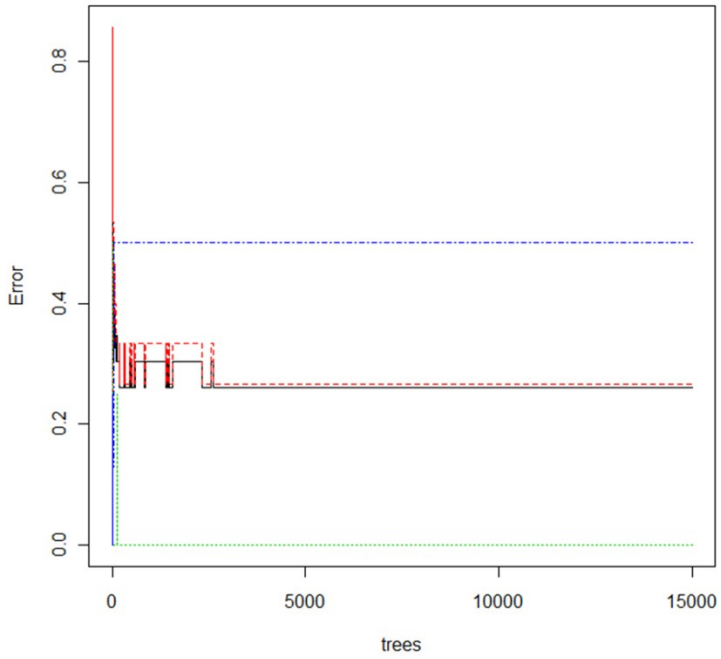

Fig. 3. Error Plot of Random Forest classification error for molecular packing types of olanzapine solvates. The black line represents the error associated with the overall model. Red, blue and green dotted lines represent the error associated with the molecular packing type classes; SC_{31}, SC_{32} and SC_{33}.

Table 2. Confusion matrix generated from Random Forest classification for the three molecular packing types of olanzapine solvates.

Predicted → / Observed ↓	Class 1 (SC_{31})	Class 2 (SC_{32})	Class 3 (SC_{33})	Class Error (%)
Class 1(SC_{31})	13	4	0	23.5
Class 2 (SC_{32})	2	2	0	50
Class 3 (SC_{33})	0	0	4	0

OZPN solvates based on SC_{31} and SC_{33} have water molecules in the asymmetric unit and so the effect of incorporation of water molecules on the predictive capability of the RF

classification model was also investigated. However the same results were obtained after taking account of water as a categorical descriptor (present/not present) and even by taking it as a numerical descriptor whose value was equivalent to the number of water molecules in the asymmetric unit of the respective OZPN solvate.

5. 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). Variable dependency plots obtained for all three molecular packing type classes are shown in Fig. 4. Numerical values of descriptors are provided in Table 1 of the main article.

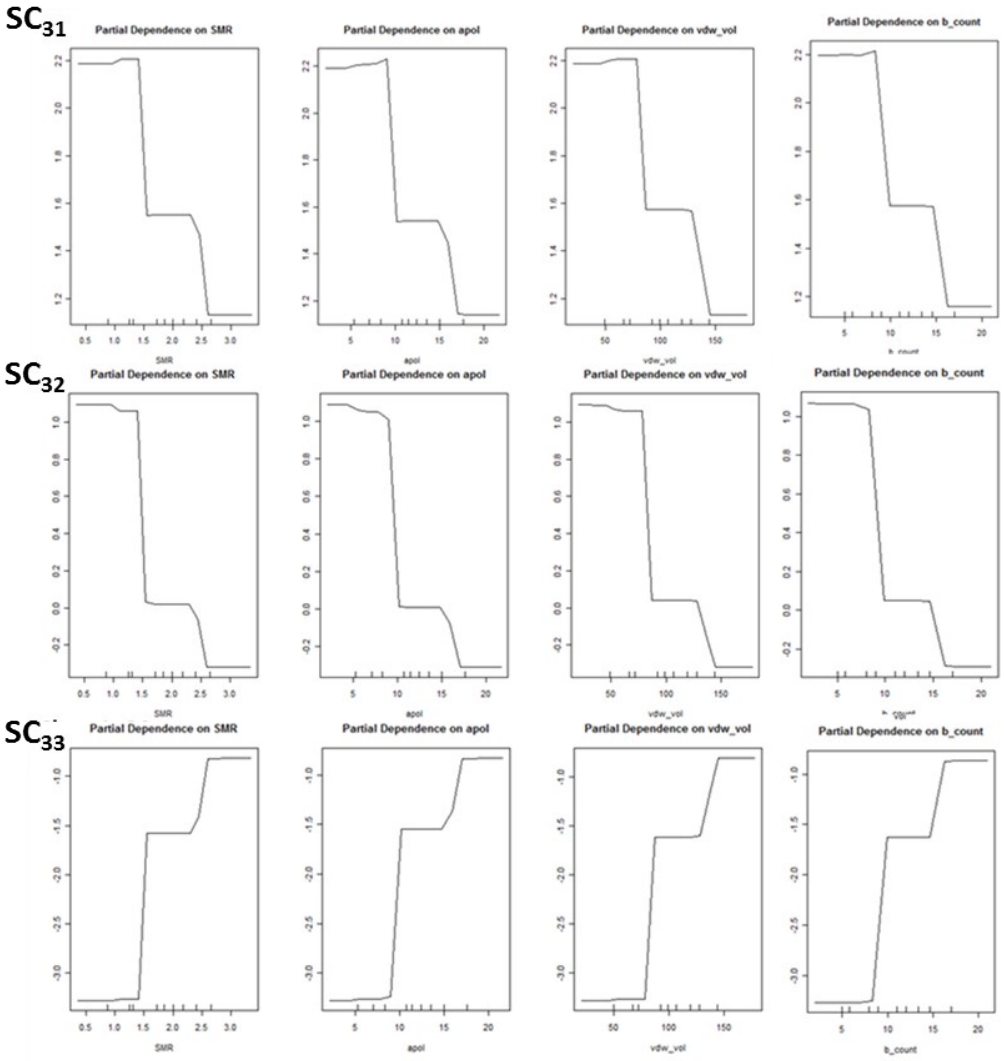

Fig. 4. Variable dependency plots (SMR, apol, VdW_vol and b_count) for molecular packing types, SC_{31}, SC_{32} and SC_{33}.

6. References

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