Journal Name

RSCPublishing

COMMUNICATION

A Random Forest Model for Predicting the Crystallisability of Organic Molecules

Cite this: DOI: 10.1039/x0xx00000x

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

Received 00th January 2012,
Accepted 00th January 2012

DOI: 10.1039/x0xx00000x

www.rsc.org/

A Random Forest model has for the first time enabled the prediction of the crystallisability (crystals vs. no crystals) of organic molecules with ~70% accuracy. The predictive model is based on calculated molecular descriptors and published experimental crystallisation propensities for a library of substituted acylanilides.

Random forests (RF) is a method for classification and regression [1-3] and has been used in various physical and life science applications such as for predicting aqueous solubility[4], mutagenicity[5], QSAR studies[6], for building drug likeness classification models[7], as well as other applications in life sciences[8-12]. There is only one report of the successful application of RF in the area of crystallisation, where it was used to predict solvate formation of carbamazepine[13]. RF has been described elsewhere[1-3] and offers various advantages over other statistical methods such as principal component analysis (PCA)[14] and artificial neural networks (ANN)[15, 16] that make it well suited for the analysis of complex transformations such as solvate formation, crystal packing[17] and crystallisation. Major advantages include no over-fitting of data, estimation of internal errors, measures for descriptor's importance and robustness to outliers, missing data points and noise. A schematic diagram of an RF workflow is shown in ESI.

Organic compounds can exhibit different crystallisation propensities: some may crystallise well or quickly, while others do so badly or slowly or not at all. Poor crystallisation behaviour can have an impact on processes/industry and can include a collection of outcomes including nano/micro crystal formation, oiling out, poor impurity rejection and/or agglomeration. Despite efforts towards better theoretical understanding of crystal nucleation and growth[18], it is not currently to predict ab initio which molecules are likely to show undesirable crystallisation behaviour. Hence in most practical situations trial-and-error and empirical knowledge are largely relied upon to achieve a desirable outcome when problems are encountered.[19] Various crystallisation propensity predictive models have been developed for proteins (with predictive accuracy ranging from ~70-80%) [20-27] which only require the protein sequence as input to predict crystallisability. There are no reports of crystallisability prediction of small molecules from solution; therefore, it was of interest to use the RF technique in predicting the crystallisation propensities for small molecules as a tool to guide experimental approaches to develop crystallisation processes. This communication reports the prediction of crystallisation propensities ("crystallisability") of small organic molecules[19] using a training set comprising their calculated 2-D and 3-D molecular descriptors and the published experimental crystallisation outcomes. The outcomes used in the model were only 'crystal' or 'no crystal'. The developed model has also provided a list of molecular descriptors that govern the varied crystallisation outcomes which help to rationalise the experimental observations.

The RF classification was carried out using a commercially available package, RandomForests® (Salford Systems). There are very few examples of systematic crystallisation studies on a series of related organic molecules that record the ease of crystallisation, Hursthouse, et al. have published a dataset[19] comprising crystallisation outcomes for 382 acylanilide compounds containing different R and X groups (Fig. 1).[19] This provides a diverse library of molecules that share a common molecular nucleus and displayed different crystallisation outcome forming an excellent basis for the development and testing of a predictive model.

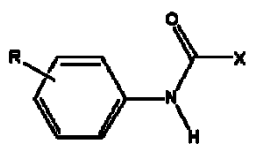

Fig. 1 Basic skeleton of acylanilide molecules. X includes H, CH_3, C_2H_5, C_3H_7, $C(CH_3)_3$, CF_3, OCH_3, $OC_2H_5NH_2$ and Cl. R includes H, CH_3, C_2H_5, $C(CH_3)_3$, OCH_3, OC_2H_5, OCF_3, F, Cl, Br, I, CF_3, OH, NH_2 and COOH.

The training dataset comprised 151 calculated 2- and 3-D descriptors for each molecule (detailed in ESI) alongside the crystallisation outcome from the original report. The outcomes were described as: class 1, where a single crystal was observed and class 2, where no single crystal was observed. The RF classification model was trained using all 151 calculated descriptors and 2 crystallisation outcomes for the 382 molecules using the following parameters: ntree = 20000, mtry = 12, jclasswt = 1 (for class 1) and 1950000 (for class 2),

nodesize =1, seed = 45€. During RF classification model building, the overall error rate converged with an increase in the number of trees (see ESI). The final RF model classified the molecules in two classes with an overall OOB error of prediction of 32.6% and prediction accuracy of 67.4% (see ESI). The RF model has predicted the number of molecules in each class with similar percentage accuracy (see ESI). The RF program computed the proximities between pairs of molecules which were then scaled down into two dimensions using multidimensional scaling (MDS). The MDS plot of scaling coordinates 1 vs. 2 (Fig. 2), obtained from the proximity matrix generated by RF showed two distinct zones belonging to two classes and an overlapped zone which comprises molecules from both classes.

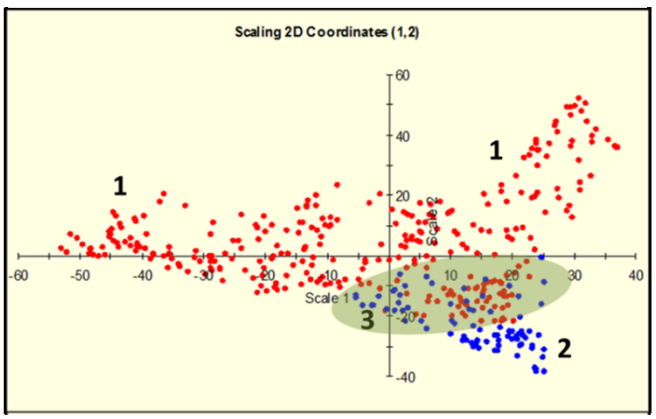

Fig. 2 The MDS plot of the scaling coordinates (1, 2) obtained from the Random Forests classification proximity matrix. Each of the 382 points on the multidimensional scaling plot represents a molecule from the dataset and is coloured according to crystallisation outcome: class 1 (red) and class 2 (blue). 1 and 2 denote the two separate zones correspond to molecules from class1 and class 2. 3 (encircled in green) denotes the overlapped zone with molecules from both classes.

A convex hull plot which is an alternative to the MDS plot and offers a useful representation of large datasets with a considerable overlap of points between them[28] was also generated (Fig. 3). The analysis showed that the molecules in class 2 were confined in a limited space of the plot indicating a common set of calculated descriptors that are consistent with poor crystallisability while molecules in class 1 were present across a larger area (see ESI). The predictive accuracy of this model was tested by removing a crystallisation outcomes for a subset of molecules, followed by rebuilding the model and

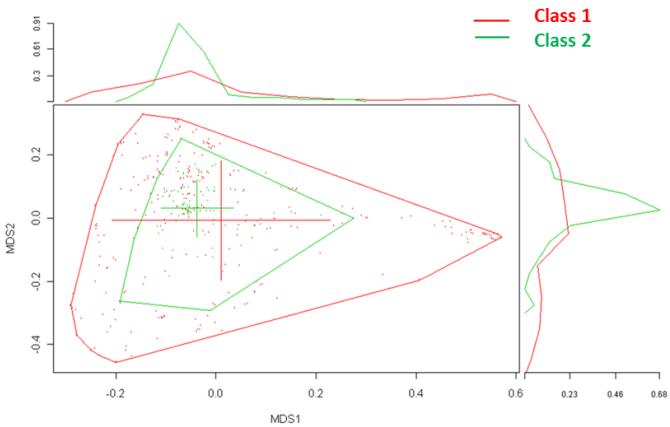

Fig. 3 Convex hull plot of scaling coordinates obtained from RF proximity matrix. Molecules in class 1 and 2 are represented by red and green points respectively. The cross sign is the mean of MDS1 and MDS2 for each group.

subsequent prediction of their crystallisation outcome resulted in a similar predictive accuracy (67%).

The mean decrease in accuracy method was used to assess the relative importance of molecular descriptors responsible for the predictive model. The top 10 most important molecular descriptors responsible for crystallisation behaviours of molecules include those that that describe the relative energies of the different molecules in terms of their torsion energy, van der Waals and steric energy terms as well as atomic connectivity, conformation and number of rotatable bonds (these are all listed with their definitions in the ESI). It is worth noting that the molecular descriptors identified by the mean decrease in accuracy method are consistent with chemical and structural expectations and/or knowledge leading to confidence in this method. For example, a number of the descriptors relate to conformational flexibility in the molecule which is known to play an important role in reducing the crystallisation tendency of molecules.[29] Molecules containing long alkyl chains generally have multiple conformations in the crystallising media which may impact on their integration within the emerging crystal lattice.[29] Similar trends were deduced for molecules in the original report i.e. crystallisation tendency was reduced on increasing the length of the alkyl chain. Propionanilide and butylanilide derivatives had poorer crystallisation tendencies compared to acetanilide and trimethylacetanilide derivatives. The method reported here also is also consistent with the observation from the original report that para-substituted derivatives were easier to crystallise than the ortho-substituted derivatives, which in turn were crystallised more easily than meta-substituted derivatives.[19]

Cheminformatics approaches for the identification and selection of critical molecular descriptors responsible for varied crystallisation outcome is a potentially powerful tool in materials design, crystallisation and process development. Although the importance of specific descriptors cannot be quantified using this method the information provides a potential means to identify crystallisation issues during initial studies and can help in designing improved crystallisation processes and in understanding the role of specific molecular attributes on this important physical transformation.

This model is based solely on 2-D and 3-D calculated molecular descriptors of a number of specific organic molecules in a relatively small range of crystallisation conditions and does not take into account of effects of impurities[19], solvent effects and variation in the crystallisation conditions in individual experiments (e.g. %RH, rate of solvent evaporation, slight variations in temperature, or other disturbances). The solubility of the compound in different solvents can vary significantly and affects the nucleation process and consequent appearance of crystals. Very high solubility may lead to increased viscosity possibly leading to gums/oils being produced whilst inadequate solubility may lead to extremely dilute solutions which will rarely give large crystals. A limited range of solvents and crystallisation conditions were used for the reported crystallisation study on this library of molecules. As this model is trained on data from the molecules and conditions described, predictive application is only justified for similar chemical and experimental conditions. However there is clearly opportunity to exploit data from other sources and in house experimental programmes to develop the tool further as means of providing a means to identify where the formation of crystals is likely to be facile or problematic.

This dataset is taken from the literature and given the aim of the original study was a structural systematics investigation; the systematic effort towards crystallisation may have been limited. The crystallisation experiments were done under similar conditions but may not have been tightly controlled leading to changes in concentration, supersaturation due to temperature or evaporation rate fluctuations. All these factors might have an effect on the

crystallisation outcome. Impurities often play a role in inhibiting crystal growth and phase transformation.[30] The crystallisation behaviour may be different under different sets of conditions used for synthesis and crystallisation. These molecules are likely to have major/minor impurities which may have an effect on the crystallisation outcome.

The next step towards achieving this kind of statistical modelling approach would be to incorporate information about crystallisation conditions such as solvent, rate of solvent evaporation, RH, temperature etc. which would certainly improve the value of the training data set and hence, predictive capability. The extension and application of this kind of statistical model to salts and co-crystal systems may also provide additional insights into the ease to crystallisation of multi-component systems.

To make robust crystallisability predictive models, systematic studies are required to obtain sufficiently comprehensive datasets and associated crystallisation data[31] which are not commonly done. However with advancements in instrumentation and automation, it is now possible to generate huge datasets of crystallisation properties.[32, 33] In addition, it is also important to store all the relevant data in accessible electronic database formats. These databases with suitable statistical modelling techniques would open the avenues for researchers to study relationships between solute, solvent, physical form and crystallisation conditions.[13]

In conclusion, the RF classification model built in this work explores the impact of molecular structure on crystallisability and provides a convenient, automatic means to highlight and understand the molecular factors that inhibit or promote crystallisation. This is the first study on crystallisability prediction for small molecules using statistical modelling techniques and provides a reasonable opportunity to highlight problematic compounds (e.g. those exhibiting nano/micro crystal formation, agglomeration, oiling out, slow nucleation etc.) at early stages so that resource planning can be accommodated to obtain effective crystallisation processes. Although, this model does not provide a mechanistic understanding of the crystallisation process, it still represents a rational and pragmatic approach which enables crystallisability prediction with a reasonable degree of confidence and can inform further mechanistic investigation.

Acknowledgements

RMB thanks Commonwealth Scholarship Commission for providing scholarship.

Notes and references

[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

€ 'ntree' refers to the number of trees grown during model building and was increased incrementally until no further improvement was observed in the model (see ESI). 'mtry' is the number of different molecular descriptors tried at each split and the default value is the square root of the total number of input descriptors. 'jclasswt' allows weightings to adjust error rates between classes that have very different number of observations. 'nodesize' refers to the minimum nodesize below which leaves are not further subdivided and the default value is 1. 'Seed' refers to any non-zero integer number which controls the random number generator. It was arbitrarily set to 45 to provide reproducibility in the random numbers required by the RF. OOB error of estimate was used as a guide during model training process.

The RF model reports the crystallisation prediction as probabilities, which correspond to the percentage votes across all trees for a molecule as each crystallisation outcome (class 1 vs. class 2). For each molecule, RF prediction provides a distribution of percentage votes for each defined outcome, totalling 100%.

Electronic Supplementary Information (ESI) available: Random forest working, List of the descriptors, See DOI: 10.1039/c000000x/

1. L. Breiman, *Machine Learning*, 2001, 45, 5-32.
2. A. Liaw and M. Wiener, *R News*, 2002, 2, 18-22.
3. V. Svetnik, A. Liaw, C. Tong, J. C. Culberson, R. P. Sheridan and B. P. Feuston, *Journal of Chemical Information and Computer Sciences*, 2003, 43, 1947-1958.
4. D. S. Palmer, N. M. O'Boyle, R. C. Glen and J. B. O. Mitchell, *Journal of Chemical Information and Modeling*, 2006, 47, 150-158.
5. Q.-Y. Zhang and J. Aires-de-Sousa, *Journal of Chemical Information and Modeling*, 2006, 47, 1-8.
6. Ž. Debeljak, A. Škrbo, I. Jasprica, A. Mornar, V. Plečko, M. Banjanac and M. Medić-Šarić, *Journal of Chemical Information and Modeling*, 2007, 47, 918-926.
7. A. C. Good and M. A. Hermsmeier, *Journal of Chemical Information and Modeling*, 2006, 47, 110-114.
8. K. Lunetta, L. B. Hayward, J. Segal and P. Van Eerdewegh, *BMC Genetics*, 2004, 5, 32.
9. X. Huang, W. Pan, S. Grindle, X. Han, Y. Chen, S. Park, L. Miller and J. Hall, *BMC Bioinformatics*, 2005, 6, 1-15.
10. A. Bureau, J. Dupuis, K. Falls, K. L. Lunetta, B. Hayward, T. P. Keith and P. Van Eerdewegh, *Genetic Epidemiology*, 2005, 28, 171-182.
11. Y. Qi, Z. Bar-Joseph and J. Klein-Seetharaman, *Proteins: Structure, Function, and Bioinformatics*, 2006, 63, 490-500.
12. S. Li, A. Fedorowicz, H. Singh and S. C. Soderholm, *Journal of Chemical Information and Modeling*, 2005, 45, 952-964.
13. A. Johnston, B. F. Johnston, A. R. Kennedy and A. J. Florence, *Crystengcomm*, 2008, 10, 23-25.
14. S. Wold, K. Esbensen and P. Geladi, *Chemometrics and Intelligent Laboratory Systems*, 1987, 2, 37-52.
15. G. W. Kauffman and P. C. Jurs, *Journal of Chemical Information and Computer Sciences*, 2001, 41, 1553-1560.
16. S. Doniger, T. Hofmann and J. Yeh, *Journal of Computational Biology*, 2004, 9, 849-864.
17. R. M. Bhardwaj, S. M. Reutzel-Edens, B. F. Johnston and A. J. Florence, *CrystEngComm*, Article in preparartion.
18. R. J. Davey, S. L. M. Schroeder and J. H. ter Horst, *Angewandte Chemie International Edition*, 2013, 52, 2166-2179.
19. M. B. Hursthouse, L. S. Huth and T. L. Threlfall, *Organic Process Research & Development*, 2009, 13, 1231-1240.
20. I. M. Overton and G. J. Barton, *FEBS Letters*, 2006, 580, 4005-4009.
21. I. M. Overton, G. Padovani, M. A. Girolami and G. J. Barton, *Bioinformatics*, 2008, 24, 901-907.
22. L. Slabinski, L. Jaroszewski, L. Rychlewski, I. A. Wilson, S. A. Lesley and A. Godzik, *Bioinformatics*, 2007, 23, 3403-3405.
23. L. Kurgan, A. Razib, S. Aghakhani, S. Dick, M. Mizianty and S. Jahandideh, *BMC Structural Biology*, 2009, 9, 50.
24. M. J. Mizianty and L. Kurgan, *Biochemical and Biophysical Research Communications*, 2009, 390, 10-15.
25. M. J. Mizianty and L. Kurgan, *Bioinformatics*, 2011, 27, i24-i33.
26. Nuria Sanchez-Puig, Claude Sauter, Bernard Lorber, Richard Giege and A. Moreno, *Protein & Peptide Letters*, 2012, 19, 725-731.
27. K. K. Kandaswamy, G. Pugalenthi, P. N. Suganthan and R. Gangal, *Protein and Peptide Letters*, 2010, 17, 423-430.
28. G. Vidmar and M. Pohar, *Computer Methods and Programs in Biomedicine*, 2005, 78, 69-74.
29. L. Yu, S. M. Reutzel-Edens and C. A. Mitchell, *Organic Process Research & Development*, 2000, 4, 396-402.
30. N. Blagden, R. J. Davey, R. Rowe and R. Roberts, *International Journal of Pharmaceutics*, 1998, 172, 169-177.
31. A. J. Florence, A. Johnston, S. L. Price, H. Nowell, A. R. Kennedy and N. Shankland, *Journal of Pharmaceutical Sciences*, 2006, 95, 1918-1930.
32. M. Hursthouse, *Crystallography Reviews*, 2004, 10, 85-96.
33. R. Storey, R. Docherty, P. Higginson, C. Dallman, C. Gilmore, G. Barr and W. Dong, *Crystallography Reviews*, 2004, 10, 45-56.