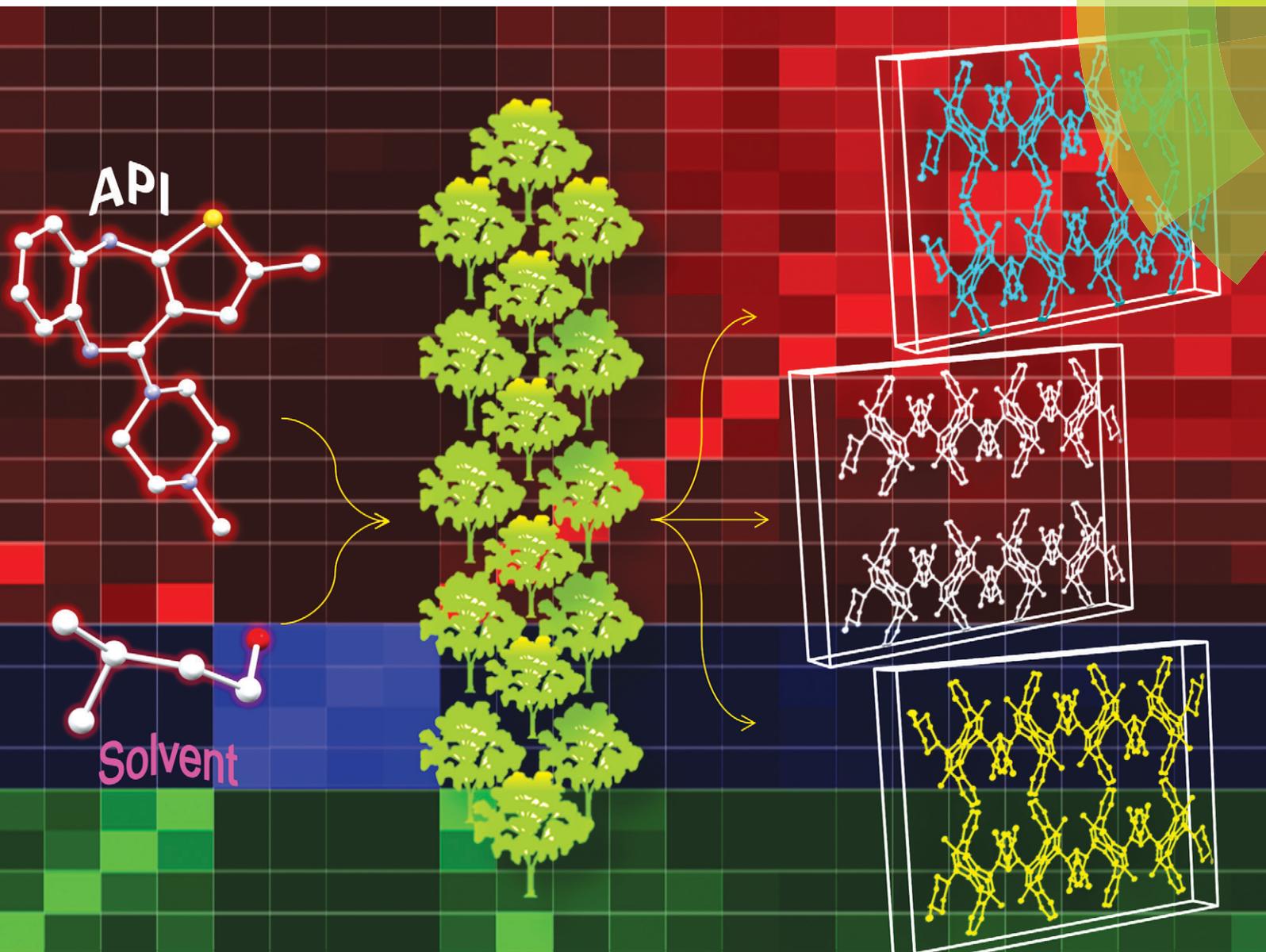


CrystEngComm

rsc.li/crystengcomm





Cite this: *CrystEngComm*, 2018, 20, 3947

Received 16th February 2018,
Accepted 14th March 2018

DOI: 10.1039/c8ce00261d

rsc.li/crystengcomm

A random forest model obtained from calculated physicochemical properties of solvents and observed crystallised structures of olanzapine has for the first time enabled the prediction of different types of 3-dimensional crystal packings of olanzapine solvates. A novel olanzapine solvate was obtained by targeted crystallization from the solvent identified by the random forest classification model. The model identified van der Waals volume, number of covalent bonds and polarisability of the solvent molecules as key contributors to the 3-D crystal packing type of the solvate.

Random forest (RF) is a classification and regression methodology^{1–3} that has been used in various physical chemistry and life science applications.^{4–12} In the area of crystallisation, RF has been used to predict solvate formation for carbamazepine¹³ and more recently, to assess the crystallisability of small organic molecules.^{14,15} A schematic diagram of an RF workflow is shown in the ESI.† Advantages of the RF method over other statistical methods, such as principal component analysis (PCA)¹⁶ and artificial neural networks (ANN),^{17,18} include the relative ranking of descriptor importance, model robustness to data outliers, missing data points and noise, and resistance to over-fitting of training data. This communication reports a novel application of RF classification to solvate crystal packing analysis. Using calculated solvent physicochemical properties and results of an experimental crystal packing analysis based on a previous study of olanzapine^{19–21} (OZPN, Fig. 1), the 3-dimensional (3-D) packing of OZPN solvates and factors responsible for their packing were used to construct a predictive model. The results of this analysis provided a rational basis for the subse-

A random forest model for predicting crystal packing of olanzapine solvates†

Rajni M. Bhardwaj,^{ib} Susan M. Reutzel-Edens,^{ib}
Blair F. Johnston^{ac} and Alastair J. Florence^{*ac}

quent targeted crystallisation which yielded a new OZPN solvate.

OZPN has been crystallized in at least 60 forms, of which 56 are solvated and 35 experimental crystal structures are known. All 35 structures are based on packing of a dispersion stabilised centrosymmetric dimer, SC₀ (Fig. 2a). The acronym SC refers to supramolecular construct, and has been defined as a recurring periodic or discrete arrangement of molecules with unique spatial characteristics.²² Of the structures based on SC₀, 28 feature a common 2-dimensional (2-D) sheet (SC₂₁) (Fig. 2b). Different stacking of the SC₂₁ sheets with different inter-sheet spacing and translations produce three distinct crystal packings based on three 3-D SCs, namely SC₃₁, SC₃₂ and SC₃₃ (Fig. 2c–e). A Hasse plot showing the structural relationships amongst OZPN solvates is shown in Fig. 3.

SC₃₂ and SC₃₃ differ in the separation distance between neighbouring SC₂₁ sheets. In the case of SC₃₁, the distance between SC₂₁ sheets is similar to that of SC₃₂, but the neighbouring SC₂₁ sheets are shifted with respect to one another. In other words, SC₂₁ sheets pack in a staggered and an eclipsed manner in SC₃₁ and SC₃₂, respectively.¹⁹ Solvent molecules occupy the void spaces between the SC₂₁ sheets in solvates based on SC₃₁ and SC₃₃ and within the SC₂₁ sheet in the case of solvates based on SC₃₂ (see ESI†). It was therefore of interest to explore the three observed 3-D crystal packing types in 28 experimental crystal structures in more detail and to confirm the role of the solvent molecules in directing the crystal packings of these 28 OZPN solvates. To achieve this, RF modelling was employed to relate calculated solvent

^a Strathclyde Institute of Pharmacy and Biomedical Sciences, University of Strathclyde, 161 Cathedral Street, Glasgow G4 0RE, UK.

E-mail: alastair.florence@strath.ac.uk

^b Eli Lilly and Company, Indianapolis, Indiana 46285, USA

^c EPSRC Centre for Continuous Manufacturing and Crystallisation (CMAC), University of Strathclyde, Glasgow, G1 1RD, UK

† Electronic supplementary information (ESI) available: Random forest working, list of the descriptors. CCDC 1479180. For ESI and crystallographic data in CIF or other electronic format see DOI: 10.1039/c8ce00261d

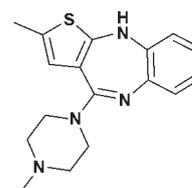


Fig. 1 Molecular structure of OZPN.

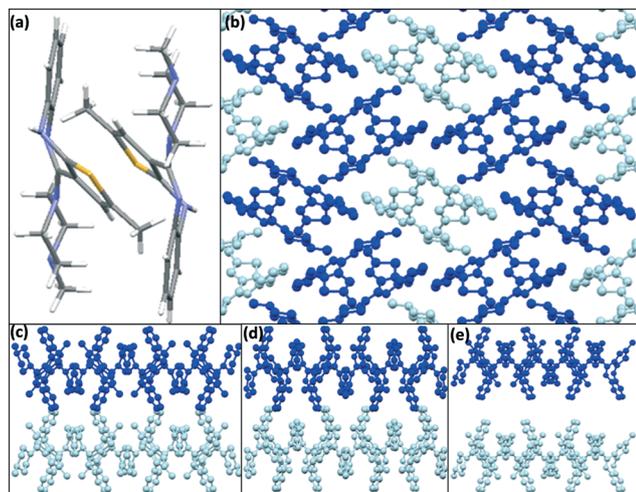


Fig. 2 (a) Centrosymmetric dimers, SC_0 (b) 2-D SC_{21} sheet, (c–e) three 3-D SCs, (c) SC_{31} , (d) SC_{32} and (e) SC_{33} , formed by SC_{21} sheets in OZPN solvates.

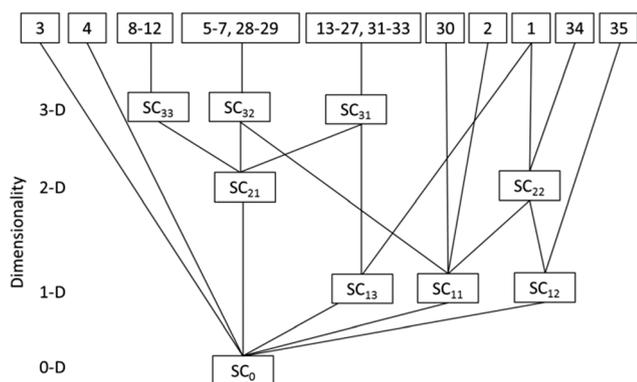


Fig. 3 Diagram showing the similarity relationship between 35 structures of OZPN.† Reproduced with permission from ref. 19.

physicochemical properties with the three 3-D molecular packing types observed in the crystal structures of OZPN solvates.

The RF classification was carried out using a commercially available software package, Random Forests® (Salford Systems) as well as with ‘Random Forest’ library package,² in the statistical computing environment ‘R’ v2.11.0.²³ The training dataset was derived from 250 calculated 2- and 3-D physicochemical descriptors for 25 solvent molecules (detailed in ESI†) alongside the three known experimental olanzapine frameworks: class 1, class 2 and class 3 comprised of SC_{31} , SC_{32} and SC_{33} , respectively. The RF classification model was trained using the following parameters: $n_{tree} = 15000$, $m_{try} = 15$, $nodesize = 1$, $seed = 45$.§ During the RF classification model build, the overall error rate converged with an increase in the number of trees (see ESI†). The final RF model showed an overall error rate of 24.5% with 100% accurate prediction for OZPN solvates present in class 3 (see ESI†). Prediction accuracy of class 1 was 75.5% followed by 50% for class 2 (see ESI†). The RF program computed a ma-

trix of proximities between OZPN solvates which were then scaled to two dimensions using multidimensional scaling (MDS). The MDS plot of scaling coordinates 1 vs. 2 (Fig. 4) showed three distinct zones relating to each of the three classes.

Four OZPN solvates based on SC_{31} (methanol solvate hydrate, dihydrate E, acetonitrile solvate hydrate and nitromethane solvate hydrate) are classified into class 2 and two OZPN solvates based on SC_{32} (acetic acid and ethanol solvates) are classified in class 1. It is worth considering why the RF model failed to correctly classify 6 of OZPN solvates taken in the training dataset. All the data corresponding to these 6 solvents were removed from the training dataset and the model was retrained using the remaining dataset. Subsequent predictions of the crystal packing type of OZPN solvates with these 6 solvents suggested that acetonitrile and nitromethane have a significant probability of forming solvates in both the classes (class1/class2 probabilities for nitromethane and acetonitrile are 54.5%/44.9% and 56.1%/43.6%, respectively). For acetic acid and ethanol, the RF model suggested that it has a higher tendency (with a prediction probability of 72% and 60%, respectively) of forming a solvate with OZPN in class 1.

The RF model also suggested that methanol and water have a higher tendency (~70%) of forming solvates with OZPN in class 2 than class 1. Experimentally, methanol, ethanol and water form solvates with OZPN in class 1 as well as in class 2 depending upon the crystallisation conditions.¹⁹ The RF classification model does not take into account crystallisation conditions, however and for methanol, ethanol and water, crystallisation conditions *e.g.* residual water in the solvent of crystallisation, % relative humidity (RH), and temperature of crystallization dictate the crystallisation outcome in terms of molecular packing type of resulting OZPN solvates. This model was then further tested to predict the

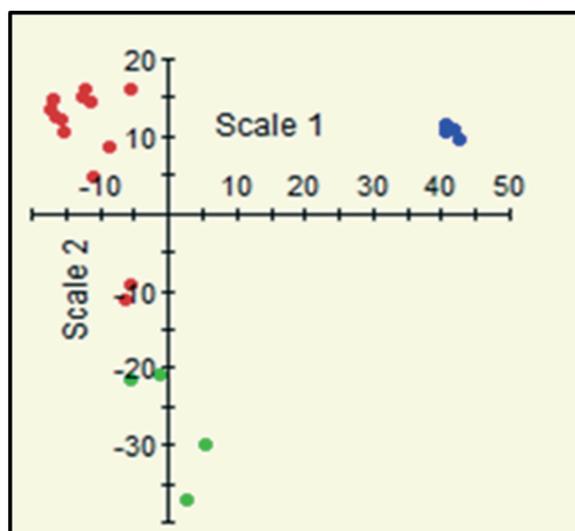


Fig. 4 The MDS plot obtained from the RF classification proximity matrix. Each of the 25 points represents an OZPN solvate and is coloured according to crystal packing type: class 1 (red), class 2 (green) and class 3 (blue).

packing type for other small alkanol solvents. During experimental screening, 1-propanol and 1-butanol formed solvates with OZPN in class 1.¹⁹ Similar probability ratios as that of ethanol were obtained for 1-propanol for both classes 1 and 2. This suggests that in addition to the known class 1, 1-propanol might also form a solvate with OZPN based on SC₃₂ (class 2). The prediction probability for 1-butanol was 90% for class 1 which indicates that it has a significantly higher probability of forming a solvate whose packing falls within class 1 than in class 2.

Based on the model's performance, it can be hypothesised that solvent molecules bigger than 1-propanol cannot be accommodated between eclipsed SC₂₁ sheets to give rise to packing based on SC₃₂. Alternatively, the driving force would be towards forming packing based on SC₃₁, where alternate SC₂₁ sheets are arranged in a staggered manner, leaving more space to accommodate larger solvent molecules. Solvent molecules smaller than 1-propanol can adopt any of the two packing types (SC₃₁ and SC₃₂) depending on the crystallisation conditions. Having significant probabilities for some solvates in two classes might point to the potential for polymorphic solvates of OZPN. However, in this case, true polymorphs were not found under the conditions tested with these solvents instead producing both SCs amongst different solvents.

OZPN solvates based on SC₃₁ and SC₃₃ 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 (see ESI†).

With reference to crystal packing type, the RF classification also provided a rank dependence of OZPN solvate classification on solvent descriptors. The most important descriptors were found to be vdW_vol (van der Waals volume) b_count (number of covalent bonds), SMR (molecular refractivity which is a measure of the total polarisability of the molecule) and apol (sum of the atomic polarisabilities). The variable dependency plots (see ESI†) indicate that the crystal structures in class 3 are favoured by large solvent molecules and these can be easily distinguished from other solvent molecules forming solvates in other two classes (Table 1). Smaller solvent molecules like water and methanol favour the packing based on SC₃₂, although there is an overlap in the descriptors values of the solvents forming solvates based on SC₃₁ and SC₃₂ (Table 1). Depending on the crystallisation conditions (presence of water during crystallisation, RH and temperature) some solvents that form solvates based on SC₃₂ may also form solvates with OZPN based on SC₃₁ and *vice versa*.

Table 1 Numerical values of solvent descriptors obtained from variable dependency plots generated during RF classification of OZPN solvates

Packing type	SMR (m ³)	apol (cm ² V ⁻¹)	vdW_vol (cm ³ mol ⁻¹)	b_count
SC ₃₁	0.8–2.4	5.2–14.5	59.0–132.9	5–15
SC ₃₂	0.3–1.3	<8.3	<72.5	<8
SC ₃₃	>2.6	>17.6	>144.6	>17

The RF analysis highlights the possible solvate forming ability of acetonitrile, nitromethane, 1-propanol and acetic acid with OZPN in both classes SC₃₁ and SC₃₂. To test this OZPN was recrystallized from anhydrous 1-propanol solution using slow cooling. An OZPN 1-propanol solvate was successfully obtained and the single crystal structure[¶] confirmed that it was based on SC₃₂ *i.e.* class 2 as predicted (Fig. 5). This result suggests that the RF model is effectively predicting the potential for new OZPN solvates with specific structural features. Further experiments would be required under a wider range of conditions to confirm whether acetonitrile, nitromethane and acetic acid would also form OZPN solvates based on SC₃₂ and SC₃₁, respectively perhaps by carrying out crystallisation using anhydrous solvents and in the presence of water. It may have been possible to obtain the OZPN 1-propanol solvate in the initial screen if a wider range of experimental conditions had been included. However, this would further increase the number of experiments and analytical efforts and would likely include conditions that would not necessarily yield new forms. In practice when faced with the necessity to shorten development timelines and reduce operating cost, extending crystallisation search space and the associated analytical requirements are less attractive. This computationally assisted approach is an inexpensive and efficient strategy for assessing the completeness of the initial experimental search and demonstrates the potential to target conditions where further forms are more likely to be obtained.

The model built in this work explores the impact of physicochemical properties of solvent molecules on crystal packing and provides a means to understand the underlying factors dictating molecular packing in solvates. Further extension of the model would see the inclusion of details on crystallisation conditions (solvent, rate of evaporation, temperature *etc.*) and possibly properties such as shape, stability,

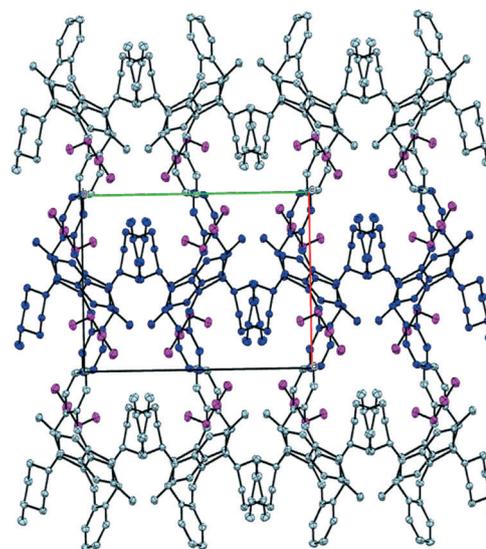


Fig. 5 Crystal packing of OZPN-1-propanol solvate. 2-D SC₂₁ sheets of OZPN molecules are shown in cyan and blue colours and 1-propanol molecules are shown in pink colour.

solubility to molecular properties and crystal structure/packing features to assess the potential to improve the value of the training data set and hence, predictive capabilities.

In summary, the RF classification model showed that the crystal packing of OZPN solvates is dependent on the van der Waals volume, number of covalent bonds and polarisability of the solvent molecules incorporated in the crystal lattice. The largest solvent molecules tend to form solvates based on SC₃₃ and using this classification model, these can be predicted more accurately than the solvents which tend to form solvates based on SC₃₁ and SC₃₂. Using this analysis, a novel OZPN 1-propanol with SC₃₂ was obtained using targeted crystallisation. This study has demonstrated for the first time the ability to predict the crystal packing of novel solvates based on a crystal packing classification model using the physicochemical properties of known solvate forming solvents. There is clear potential to apply this approach more generally to improve the efficiency of experimental crystallisation form screening and target specific crystal packing features.

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

RMB thanks Commonwealth Scholarship Commission for providing the scholarship. The authors thank EPSRC for funding this work through the Basic Technology program Control and Prediction of the Organic Solid State (www.cposs.org.uk) (EP/F03573X/1).

Notes and references

‡ Numbers in Fig. 3 correspond to following structures: form I [1], form II [2], 2-butoxyethanol [3], 2,2,2-trifluoroethanol [4], acetic acid [5], ethylene glycol [6], ethanol [7], methyl *tert*-butyl ether hydrate [8], 2-pentanol hydrate [9], *N,N,N*-triethylamine hydrate [10], *t*-butanol hydrate [11], isoamylalcohol hydrate [12], 1,4-butanediol hydrate [13], dimethylformamide hydrate [14], pyridine hydrate [15], tetrahydrofuran hydrate [16], 1,4-dioxane hydrate [17], acetone hydrate [18], 1-butanol hydrate [19], acetonitrile hydrate [20], nitromethane hydrate [21], 1,2-dimethoxyethane hydrate [22], methoxyethanol hydrate [23], methyl acetate hydrate [24], 1,2-propanediol hydrate [25], 1-propanol hydrate [26], higher hydrate [27], methanol solvate [28], dihydrate B [29], dichloromethane [30], dimethylsulfoxide hydrate [31], methanol hydrate [32], ethanol hydrate [33], dihydrate D [34], 2-butanol hydrate [35].

§ '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 ES†). '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. '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%.

¶ Crystal data for OZPN 1-propanol solvate is C₁₇H₂₀N₄S. C₃H₈O, *M* = 372.52; plate crystallised from anhydrous 1-propanol solution, 0.21 × 0.1 × 0.1 mm; monoclinic, *P*₂₁/*c*, *a* = 10.217(3) Å, *b* = 13.069(4) Å, *c* = 14.918(5) Å, β = 95.342(7)°, *V* = 1983.2(10) Å³, *Z* = 4; *T* = 100 K; μ(Mo-Kα₁) = 0.180 mm⁻¹; λ = 0.71073 Å; 13958 reflections measured; 4767 unique reflections used in refinements with 243 refineable parameters and 0 restraints; final *wR*(*F*) = 0.0927 (all data), *R*(*F*) = 0.0531 (*F*² > 2σ*F*²). CCDC reference number for OZPN 1-propanol solvate is CCDC 1479180.

- 1 L. Breiman, *Mach. Learn.*, 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, *J. Chem. Inf. Comput. Sci.*, 2003, 43, 1947–1958.
- 4 D. S. Palmer, N. M. O'Boyle, R. C. Glen and J. B. O. Mitchell, *J. Chem. Inf. Model.*, 2006, 47, 150–158.
- 5 Q.-Y. Zhang and J. Aires-de-Sousa, *J. Chem. Inf. Model.*, 2006, 47, 1–8.
- 6 Ž. Debeljak, A. Škrbo, I. Jasprica, A. Mornar, V. Plečko, M. Banjanac and M. Medić-Šarić, *J. Chem. Inf. Model.*, 2007, 47, 918–926.
- 7 A. C. Good and M. A. Hermsmeier, *J. Chem. Inf. Model.*, 2006, 47, 110–114.
- 8 K. Lunetta, L. B. Hayward, J. Segal and P. Van Eerdewegh, *BMC Genet.*, 2004, 5, 32.
- 9 X. Huang, W. Pan, S. Grindle, X. Han, Y. Chen, S. Park, L. Miller and J. Hall, *BMC Bioinf.*, 2005, 6, 1–15.
- 10 A. Bureau, J. Dupuis, K. Falls, K. L. Lunetta, B. Hayward, T. P. Keith and P. Van Eerdewegh, *Genet. Epidemiol.*, 2005, 28, 171–182.
- 11 Y. Qi, Z. Bar-Joseph and J. Klein-Seetharaman, *Proteins: Struct., Funct., Bioinf.*, 2006, 63, 490–500.
- 12 S. Li, A. Fedorowicz, H. Singh and S. C. Soderholm, *J. Chem. Inf. Model.*, 2005, 45, 952–964.
- 13 A. Johnston, B. F. Johnston, A. R. Kennedy and A. J. Florence, *CrystEngComm*, 2008, 10, 23–25.
- 14 J. G. P. Wicker and R. I. Cooper, *CrystEngComm*, 2015, 17, 1927–1934.
- 15 R. M. Bhardwaj, A. Johnston, B. F. Johnston and A. J. Florence, *CrystEngComm*, 2015, 17, 4272–4275.
- 16 S. Wold, K. Esbensen and P. Geladi, *Chemom. Intell. Lab. Syst.*, 1987, 2, 37–52.
- 17 G. W. Kauffman and P. C. Jurs, *J. Chem. Inf. Comput. Sci.*, 2001, 41, 1553–1560.
- 18 S. Doniger, T. Hofmann and J. Yeh, *J. Comput. Biol.*, 2004, 9, 849–864.
- 19 R. M. Bhardwaj, L. S. Price, S. L. Price, S. M. Reutzel-Edens, G. J. Miller, I. D. H. Oswald, B. Johnston and A. J. Florence, *Cryst. Growth Des.*, 2013, 13, 1602–1617.
- 20 S. M. Reutzel-Edens, J. K. Bush, P. A. Magee, G. A. Stephenson and S. R. Byrn, *Cryst. Growth Des.*, 2003, 3, 897–907.
- 21 R. M. Bhardwaj and A. J. Florence, *Acta Crystallogr., Sect. E: Struct. Rep. Online*, 2013, 69, o752–o753.
- 22 T. Gelbrich and M. B. Hursthouse, *CrystEngComm*, 2005, 7, 324–336.
- 23 R Development Core Team, *R Foundation for Statistical Computing*, Vienna, Austria, 2006, www.R-project.org.