CrystEngComm

Accepted Manuscript

Volume 1 | Number 1 | Jan 2013 | Pages 1–100
CrystEngComm
www.rsc.org/crystengcomm
ROYAL SOCIETY OF CHEMISTRY

ROYAL SOCIETY OF CHEMISTRY

www.rsc.org/crystengcomm

Will it crystallise? Predicting crystallinity of molecular materials†

Jerome G. P. Wicker and Richard I. Cooper*

Received Xth XXXXXXXXXX 20XX, Accepted Xth XXXXXXXXX 20XX
First published on the web Xth XXXXXXXXXX 200X
DOI: 10.1039/b000000x

Predicting and controlling crystallinity of molecular materials has applications in a crystal engineering context, as well as process control and formulation in the pharmaceutical industry. Here, we present a machine learning approach to this problem which uses a large input training set which is classified on a single measurable outcome: does a substance have a reasonable probability of forming good quality crystals. While the related problem of crystal structure prediction requires reliable calculation of three dimensional molecular conformations, the method employed here for predicting crystallisation propensity uses only "two dimensional" information consisting of atom types and connectivity. We show that an error rate lower than 10% can be achieved against unseen test data. The predictive model was also tested in a blind screen of a set of compounds which do not have crystal structures reported in the literature, and we found it to have a 79% classification accuracy. Analysis of the most significant descriptors used in the classification shows that the number of rotatable bonds and a molecular connectivity index are key in determining crystallisation propensity and using these two measures alone can give 80% accurate classification of unseen test data.

1 Introduction

Machine learning algorithms have been applied to modelling several quantitative structure-property relationships (QSPRs) of small molecules. These include the prediction of the lower flammability limit,[1] solubility,[2] heat capacity[3] and melting point of organic compounds[4]. The success of machine learning algorithms in these examples suggests that there is some scope for applying this approach to prediction of crystallisation propensity, provided the correct descriptors are used.

Such predictions may be applied to find synthetic modifications that are likely to enhance crystallisation propensity, or to find poorly crystalline materials with large surface areas. The approach taken here is complementary to investigations which rationalise and control aspects of crystallisation by focusing on properties and interactions of individual molecules.[5–8] Instead we have taken a large and diverse input data set, applied a rough and ready method for classification of crystallinity, and developed a predictive model which can then be tested.

There are several reports of the application of machine learning to the prediction of protein crystallinity.[9] These approaches aim to predict crystallinity based on the sequence of amino acids as the descriptors for each sample in the dataset, using various algorithms including: support vector machine[10–12]; random forest[13]; and neural networks[14]. Although the diversity of protein structures is large, the types of interactions that can be formed between the surfaces of two proteins during crystallisation is relatively limited compared to the wide range of chemical interactions in small molecules. This increased complexity requires descriptors of molecular chemistry that are relevant to the problem of crystal formation. Therefore a wide range of descriptors need to be tested, and the most useful ones discovered.

1.1 Crystal Structure Prediction

The prediction of the structures of crystalline organic compounds has made steady progress over the last few decades: A series of crystal structure prediction blind tests[15–19] reveal an improving rate of success, and the most recent indicate that dispersion corrected density functional theory (DFT-D) calculations on small molecules can reliably predict structure, albeit at great computational cost[19].

Structure prediction generally requires two steps: (i) generation of a set of *trial structures* which will hopefully include an approximately correct result; and (ii) the optimisation and ranking of these structures (by energy, density, or some other cost function) in order to identify the most likely crystalline form.[20] In practice some molecules exhibit polymorphism which results in several different crystal structures and successful prediction of these involves mapping out the "crystal energy landscape"[21] by sampling all feasible solid state arrangements and then considering all lattice energy minima within a certain energy of the global minimum, which should correspond to the thermodynamically feasible poly-

† Electronic Supplementary Information (ESI) available: Descriptor definitions.Performance of single variable classifiers. Filtering rules for subset creation. Table of molecules selected for recrystallisation screen. Recrystallisation screen results. Example code. See DOI: 10.1039/b000000x/

Chemical Crystallography, Chemistry Research Laboratory, Mansfield Road, Oxford, UK. E-mail: richard.cooper@chem.ox.ac.uk

morphs. However, in situations where there are many possible crystal structures with comparable lattice energies, it is difficult to distinguish the case where there are multiple accessible polymorphic forms, from the alternative possibility where the molecule does not crystallise well at all, with macroscopic crystal growth being severely inhibited.[22]

1.2 Crystallisation Propensity

From a practical perspective, there are some questions that structure prediction and energy calculations do not answer: (i) will the material crystallise at all? (ii) will chemical modification of a molecule make it more or less crystalline? These questions are of fundamental importance in deciding whether it is worth investing effort in attempting to recrystallise materials for analysis by single crystal X-ray diffraction (SXRD) methods, and when attempting to control crystallinity in formulations of pharmaceutical ingredients, cosmetics, and food products. Molecular glasses[23], which are engineered specifically to be non- crystalline have applications in drug formulations, foods and photo-voltaic cells[24]. Although empirical relationships have been proposed which correlate glass formation with conformational flexibility, use of bulky groups[25] and prevention of directional interactions, the factors which affect glass formation are actually little understood, and there is a long way to go in predicting which molecules will not crystallise.

We report the development and testing of a model to predict the crystallisation propensity of small organic molecules solely from the atomic connectivity. The aim is to use machine learning algorithms to classify molecules as either crystallisable or non-crystallisable, based on a set of standard computational chemistry descriptors, without the need to consider crystal growth mechanisms or conditions.

2 Methods

All algorithms were executed using Python 2.7.3. Cambridge Structural Database (CSD) molecules were taken from the November 2013 version, while ZINC molecules were taken from a version downloaded in August 2012. The descriptors were calculated using the RDKit cheminformatics toolkit,[26] version Q4 2013. All descriptors are defined in the Supporting Information. Machine learning algorithms and performance metrics were implemented using version 14.1 of the scikit-learn package.[27] An example of the method used to train a model and output a predictive accuracy from a set of training and test data with known crystallinity labels is given in the Supporting Information.

2.1 Selection and classification of training data

If we consider all non-amorphous materials to be "crystalline" to some extent, (with less crystalline materials having their crystal growth restricted due to kinetic or thermodynamic factors) then in single crystal X-ray diffraction experiments, we will only encounter those that can grow to give a minimum crystal dimension of approximately 0.1 mm (or perhaps as small as 10 µm for synchrotron X-ray diffraction experiments), as this will contain enough material to produce a diffraction pattern with sufficient signal for crystal structure determination (Figure 1). A distinction can be made (albeit with a blurred boundary) between molecules which can have their structures determined by SXRD and those which form either amorphous solids or solids which have some periodic arrangement of molecules, but where the domain size is too small to use single crystal diffraction to determine the structure.

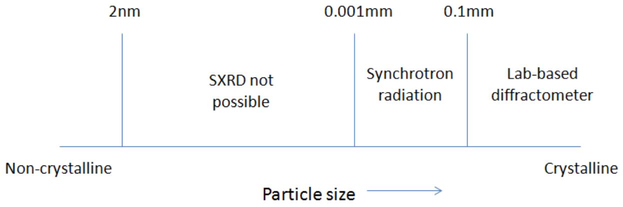


Fig. 1 Crystal structure determination options for various crystallite sizes.

All molecules selected for training and testing this model are contained in the ZINC[28] database, which contains biologically relevant purchasable molecules from many chemical suppliers.

Molecules which were contained within both ZINC and the Cambridge Structural Database (CSD)[29] provided a subset which could be used as a reliable source of crystalline molecular materials. The CSD is almost entirely composed of crystal structures determined by single-crystal diffraction; the few structures determined only by powder diffraction were ignored for this study. Any molecule in this set is known to form crystals of appropriate size and quality for single crystal structure analysis to be carried out, and so this becomes the definition of a crystalline or crystallisable molecule for this study. The remaining ZINC molecules were used as a set of non-crystalline molecular materials.

This approach means that all molecules used are subsets of the commercially available ZINC database and therefore the machine learning does not simply learn to distinguish between commercially available and 'research' compounds.

The classification of molecules using this method has some assumptions: molecules from the CSD are almost certainly crystalline except in rare cases of errors or deliberately falsified data[30], but not all crystalline molecules will have been removed from the remaining ZINC subset. In some cases, recrystallisation of a ZINC molecule may never have been attempted, the molecule may not have been studied by SXRD, or the crystal structure has never been published. This means that while all molecules in the "crystalline" training and test sets can almost certainly form single crystals, there is a much higher probability that some molecules in the "non-crystalline" training and test sets may actually form single crystals. We will show that the bias introduced by this assumption is small by testing the model using a blind recrystallisation trial of commercially available materials in the Crystallisation screening section.

2.2 Standardisation

Salts and organometallic complexes were removed from the datasets, as we are focussed on organic molecules in this application.

Solvents of crystallisation were deleted from input molecules, as these can be considered "crystallisation conditions" and we predicting whether a molecule will crystallise, not how it does so. All molecules were standardised by setting them to standard representation of charges, to enable simple comparisons between sets.

A smaller subset suitable for use with the machine learning algorithms was prepared by filtering the molecules according to rules proposed by Lipinski[31,32] to identify "drug-like" molecules, as detailed in Table 4 of the Supporting Information. These filters halved the size of the organic CSD set to around 100,000 molecules (before cross-referencing with ZINC, which further reduced this subset to roughly 18,000 molecules).

A random subset of the non-crystalline dataset was used since this group was much larger than the crystalline dataset (due to the smaller size of the CSD compared to ZINC) to ensure that similar numbers of molecules were used in each dataset to avoid potential class bias in the trained model.

2.3 Machine Learning

Machine learning algorithms use *descriptors* of the entries in a data set to make predictions. An algorithm can be either *supervised*, in which case the algorithm finds a function that attempts to reproduce known outputs values from the input data, or *unsupervised*, where it attempts to find previously unknown patterns in the data. In the approach presented here supervised algorithms are required, where training data has a determined output value: "crystalline" or "non-crystalline". The machine learning algorithm determines a model to give the best predictions for the *training* data and is judged by its success rate predicting the crystalline state of a previously unseen *test* data set. Supervised algorithms are split into two types: regression algorithms and classifiers. Regression algorithms are used to predict continuous properties of the samples, while classifiers attempt to group the samples into categories. To predict whether molecules will form crystalline or non-crystalline materials, a classifier algorithm is appropriate.

The classifiers used in this study were support vector machines,[33] which find the linear separating hyperplane between the two classes which has the maximal margin. For data which is not linearly separable, the "kernel trick"[34] can be used to map the data into a higher dimensional space in which it may be easier to separate the two classes. We compare the success rates of linear and kernel SVMs with a random forest (RF) algorithm[35], which uses an ensemble of decision tree predictors, where each node in each tree is split using a random subset of the descriptors. The probabilistic prediction of the trees is averaged to give an overall classification for each molecule.

The training data comprised a randomly selected 75% of each of the non-crystalline and crystalline datasets and each entry was labelled with the known classification (0 for non-crystalline, 1 for crystalline). The remaining 25% of each set became the test data to which the predictive model attempts to assign labels. The test set gives a measure of the expected accuracy of each model when applied to real data sets by comparing predicted and known classifications for molecules which have not previously been seen by the model. Both SVM and RF algorithms provide continuous probabilistic classification predictions in the range 0.0-1.0, and a score greater than 0.5 results in the molecule being classified as crystalline.

In order to test the repeatability of the use of the model using different splits of test and training data, a further calculation was carried out by splitting the total dataset into 5 equally-sized random subsets, each containing roughly the same number of crystalline and non-crystalline molecules.[4] The training and testing of the model was carried out using each subset in turn as the test set with the other 4 subsets making up the training dataset. This meant that every molecule is tested on, and allows an estimate of the error in the method to be determined.

2.3.1 Descriptor analysis.

Identification of the most important descriptors is useful for practical and scientific reasons: it allows the computational cost of the analysis to be reduced by removing irrelevant descriptors from the learning algorithm; it informs development of new descriptors which could capture the same information in a manner more relevant to the classification problem; and it gives an insight into the most important factors governing the crystallinity of molecules, allowing rationalisation of observed behaviour and development of rational rules for how to modify compounds to decrease or increase their propensity to crystallise.

To identify the most important descriptors of the molecules for making these predictions, an independent feature selection analysis was required since the importance of each feature cannot be computed explicitly from a nonlinear SVM.[36] Several feature selection methods have previously been applied to problems to improve classifier performance, including the F-score statistical test[37] which identifies important features, and recursive feature elimination[36] which removes redundant features. We compared the accuracy of many classifiers built using only a single molecular descriptor[38] in order to find the most important feature. The second most significant descriptor is then found by comparing two-descriptor classifiers which use the most important descriptor paired with every other descriptor in turn. The two descriptors which gave the highest percentage accuracy were chosen.

3 Results and discussion

The breakdown of the training and test molecules for the drug-like data is shown in Table 1. There is almost no size imbalance between the two classes, to prevent bias towards a particular class.

Table 1 Breakdown of training and test molecules for drug-like molecules from ZINC and CSD

	Non-crystalline	Crystalline	Total
Training	13440	13453	22733
Test	4480	4485	8965
Total	17920	17938	35858

The predictive accuracy of three machine learning algorithms trained using all 177 descriptors generated using RDKit are compared below. The SVM algorithm, using an RBF kernel with parameters optimised by a grid search of C and γ values using accuracy as the score function, was found to give the highest percentage accuracy, as shown in Table 2. This model achieves accuracies of 90.3% on the drug-like data sets.

Confusion matrices in Table 2 show the SVM with RBF kernel misclassifies the fewest molecules for each class and is particularly accurate on the crystalline dataset. The confusion matrices for the SVM classifiers show no significant imbalance in the misclassification between the two classes. Table 2 also shows that while the average from 5 tests for each of the algorithms is similar to the result from the single initial test, the variance in the value for the RF model is much greater than for the SVM models, showing that they behave more consistently.

Figure 2 shows the receiver operating characteristic (ROC) curves[39] of the three different models trained using RDKit descriptors of drug-like molecules. These curves are generated by ranking the molecules in descending order of the probability of the molecule being crystalline (as calculated from the algorithm). Taking each molecule in turn, if the actual label is "crystalline", then this is a true positive result, whereas if the actual label is "non-crystalline" this is a false positive. The true positive rate is then plotted against the false positive rate and the area under the curve (AUC) provides a measure of the ability of the model to rank the crystalline molecules relative to the non-crystalline ones. Again, the SVM algorithm with an RBF kernel performs best. It has the highest AUC of 0.96 showing the most effective ranking of the molecules according to crystallinity and also has the steepest curve, showing good classification of molecules strongly predicted to be crystalline.

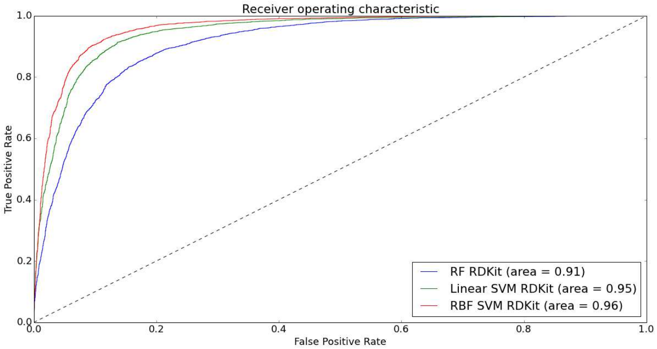


Fig. 2 ROC curves for SVM (linear), SVM (RBF) and RF models trained using drug-like molecules with RDKit descriptors.

Percentage accuracy, confusion matrices, and ROC curves all show that the SVM algorithm using an RBF kernel provides the best predictive model, which is consistent with the fact that this kernel is known to be widely used for its high predictive accuracy. The RDKit descriptors clearly capture much of the important information about which molecules will easily crystallise in a given set of molecules, and must be capturing underlying factors such as which materials exist in the solid state at all.

The SVM method using a linear kernel misclassified roughly 1% more of the molecules than the RBF kernel, due to the constraint on the linear algorithm to use only simple hyperplanes to separate the two classes. The confusion matrices show that the majority of the difference in failed predictions between RBF and linear kernels occur for the crystalline class. However, an advantage of the linear algorithm is that it takes significantly less time to train than the RBF kernel.

The RF algorithm performed least well for all models in this data set. In all cases it was the worst performer on the molecules most confidently predicted to be crystalline, as shown by the shallower gradient of the ROC curves. Despite this, RF still provided a predictive accuracy of 84% which, coupled with its ability to be trained using unscaled data and the relative speed of the training step could make this a useful algorithm for large scale calculations where accuracy can be traded off against speed.

Table 2 Confusion matrices of drug-like molecules using RDKit descriptors for a) Linear SVM b) RBF SVM c) RF. Key: T (NC) = Non-crystalline correctly predicted; F (NC) = non-crystalline but predicted to be crystalline; F (C) = crystalline but predicted to be non-crystalline; T (C) = crystalline correctly predicted

Key		SVM (linear)		SVM (RBF)		RF	
T (NC)	F (NC)	**86.3%**	13.6%	**87.9%**	12.1%	**83.2%**	16.8%
F (C)	T (C)	8.6%	**91.4%**	7.2%	**92.8%**	15.5%	**84.5%**
Overall		88.9%		90.3%		84.4%	
Average of 5 tests		88.7(2)%		90.3(3)%		84.3(5)%	

In order to further validate the use of the approach using another independent test dataset, the CSD update for February 2014 was used to provide a second set of “crystalline” molecules which could be used to test the predictive accuracy of the model. This set was filtered in the same way as the original test set, and any molecules which were already present in the CSD were removed. Only those which were already present in ZINC (but had not been used previously in either the test or training datasets) were included. This gave us a set of 354 new crystalline molecules independent of the initial training and test datasets, which were then classified using the original model. Of these, 312 molecules were successfully predicted to be crystalline, giving a classification accuracy of 88.1%, which is very similar to the accuracy obtained from the original test data.

3.1 Descriptor Analysis

The C and γ parameters for the SVM algorithm were optimised using a grid search with cross-validation using accuracy as the score function. The ${}^{0}\chi^{v}$ index[40] was found to give the highest predictive accuracy for the unseen test data from a single variable classifier with an accuracy of 77.2%. Table 3 shows that rotatable bond count (RBC) with ${}^{0}\chi^{v}$ was found to provide the most successful two-variable classifier with a predictive accuracy of 80%, which indicates that the majority of the accuracy obtained with 177 descriptors can be achieved using just two descriptors. The distribution of crystalline and non-crystalline materials with these two descriptors is shown in Figure 3. As can be seen from Table 3, the increase in accuracy from linear SVM to the RBF kernel SVM is not significant and this can be understood using Figure 3, where the decision boundary is found to be almost linear even when using the kernel trick.

Table 3 Percentage accuracy for the top two features

linear SVM	RBF SVM	RBF SVM
	$(C = 1\ \gamma = 1)$	$(C = 100\ \gamma = 0.001)$
79.22	**79.97**	79.45

${}^{0}\chi^{v}$ is the zero order molecular valence connectivity index calculated from the hydrogen-suppressed skeleton of a molecule. Each non-hydrogen atom has a δ^{v} atomic valence delta value associated with it, which is calculated according to equation 1,

$$\delta^{v} = \frac{Z^{v} - h}{Z - Z^{v} - 1}, \qquad (1)$$

where Z is the atomic number of the atom, Z^{v} is the number of valence electrons of that atom and h is the number of attached hydrogen atoms.

${}^{0}\chi^{v}$ itself is then calculated from these by performing a summation over a function of these atomic delta values for all non-hydrogen atoms (equation 2).

$${}^{0}\chi^{v} = \sum_{i=1}^{n} (\delta^{v})^{-0.5}. \qquad (2)$$

This index has been shown to correlate strongly with the molecular volume[41] and so can be thought of as a simple descriptor of the size of the molecule.

Figure 3 shows that non-crystalline molecules are concentrated in a region with an RBC of 4-7 and a ${}^{0}\chi^{v}$ value of 12-17, while the crystalline molecules mostly occupy a slightly more spread out region of both lower RBC and lower ${}^{0}\chi^{v}$. The line obtained from the SVM algorithm seems to effectively distinguish between the majority of crystalline and non-crystalline molecules, validating the discriminatory importance of these two descriptors. However, there is significant overlap between the classes at the centre of the graph, at ${}^{0}\chi^{v}$ values of 11-13 and RBC values of 3-5. It is in these cases where more descriptors are necessary to improve the predictive accuracy of the calculation.

The influence of the rotatable bond count on the tendency to crystallise can be rationalised by considering that a molecule exists as a mixture of many different conformers in solution, which may all be of similar energies. For crystallisation to occur, the molecule must achieve the “correct” conformation so that it can nucleate and then grow into a crystal. The greater the number of rotatable bonds a molecule has, the more conformationally flexible that molecule is, which means it will have a large number of potential conformers in equilibrium in

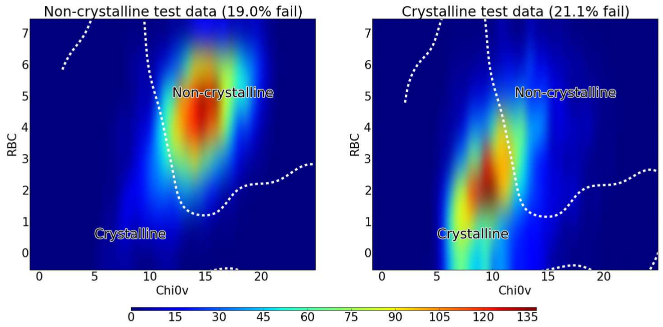


Fig. 3 Distribution of rotatable bond count against ${}^0\chi^v$ for all test molecules colour-coded by density of molecules. The dashed line shows the boundary between the crystalline and non-crystalline regions as predicted by the SVM algorithm using RBF kernel.

solution. This effectively dilutes the concentration of the desired conformer, decreasing the degree of supersaturation and therefore the crystallisation tendency. This effect is further increased when the crystallising conformer is of a relatively high energy, as the beginning of the crystallisation process will deplete this conformer and it will need to be replaced for crystallisation to continue, which occurs at a rate dependent on the energy barrier. Hence a molecule with fewer rotatable bonds, and therefore a higher concentration of the crystallising conformer, will have a greater propensity to crystallise.[42]

There is a slight positive linear correlation between ${}^0\chi^v$ and the number of rotatable bonds of 0.59 which can be seen in the distribution in Figure 3. The support vectors are separating the two classes along a line which runs approximately perpendicular to the direction of the correlation, allowing an improved classification of otherwise overlapping data. Figure 4 shows the distribution of molecules in the crystalline and non-crystalline classes for both of these descriptors, demonstrating the advantage in using both descriptors together to classify the data.

The propensity of molecules with a low ${}^0\chi^v$, and therefore a lower molecular volume, to crystallise could be attributed to the greater ease with which solvent molecules around smaller solute molecules can rearrange to allow access to the surface on crystal growth.[43]

To investigate the effect of removing the ability of the machine learning algorithm to classify using the two principal variables, a test subset was created using molecules where the values of both of these descriptors was constant. Values were chosen which ensured similar numbers of "crystalline" and "non-crystalline" molecules to prevent bias in this subset, which on inspection of the histograms in Figure 4 led to choosing ${}^0\chi^v$ to lie in a narrow range of 12–13 (as it is a continuous variable) and a rotatable bond count of 3.

The accuracy of the model on this subset of 254 molecules was 85.4%. This is a decrease in accuracy compared to the full test dataset, as expected when preventing the algorithm from classifying using the two most important descriptors. However, this does show that there is appreciable predictive accuracy to be obtained from some of the 175 other descriptors, either because they provide similar information to ${}^0\chi^v$ and the rotatable bond count, or because the different information they provide is in fact useful for classification in the region of chemical space where the values of the top two descriptors overlap for both "crystalline" and "non-crystalline" molecules, as shown in Figure 3.

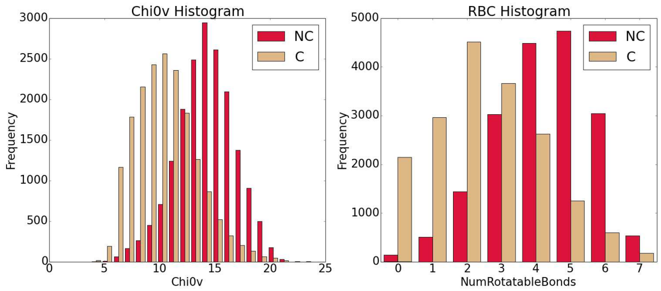


Fig. 4 Distributions of ${}^0\chi^v$ and rotatable bond count for all molecules. The subset was chosen to include molecules with a ${}^0\chi^v$ between 12 and 13 and a rotatable bond count of 3. Note that ${}^0\chi^v$ is a continuous variable and molecules are divided into bins of width 1.0 to produce this plot.

3.2 Crystallisation screening

The accuracies provided in Table 2 are based on "unseen" testing data, but may be subject to biases resulting from the possibility that some molecules classified as non-crystalline may actually be crystalline. To validate these accuracies, a blind recrystallisation screen was carried out using a set of 20 test molecules not present in the CSD (and therefore nominally non-crystalline). Diversity of the set was achieved by clustering based on molecular fingerprints, which were calculated using a variation of the Morgan fingerprint algorithm[44,45]. 20 clusters were created using the Ward hierarchical clustering algorithm[46] and one molecule was chosen from each cluster. Selection was also limited by pragmatics: a maximum cost of GBP20 per sample was enforced. This led to choosing 12 molecules which the algorithm predicted to be crystalline (class F(NC)) and 8 predicted to be non-crystalline (class T(NC)), which were then obtained from well-known commercial chemical suppliers. Recrystallisations of samples of these molecules were attempted by slow evaporation from a single solvent to try to form crystals of sufficient size and quality for SXRD to be used to determine the structure of the molecule. A range of solvents of varying polarity and volatility were chosen in order to provide a broad range of conditions so that each molecule had the maximum possible opportunity to form crystals.

One of the samples predicted to be crystalline was rejected from the study as mass spectrometry indicated that it was not the compound that was ordered, leaving only 11 molecules predicted to be crystalline in the test group.

7 of the 11 molecules predicted to be crystalline were successfully recrystallised with crystals large enough to be used for SXRD, while the other 4 remained as powders. Seven previously unreported crystal structures were obtained. No crystals of sufficient size and quality were obtained for any of the molecules predicted to be non-crystalline, although 3 of them did form needle-like crystals unsuitable for use in SXRD. This was confirmed by attempting to obtain a diffraction pattern using a crystal of one of these samples, which provided a low quality pattern with broad streaks as opposed to well defined diffraction spots, indicating large amounts of strain or modulation of the structure and making the crystals unsuitable for use in structure determination. Overall, the predictive accuracy was 79% which, when taking into account the error of 9.7 % in the accuracy of the predictions, gives a similar accuracy to the theoretical test. This indicates that while some of the molecules in the 'non-crystalline' training and test sets are actually crystalline, this has only a minor impact on the effectiveness of the model.

4 Conclusions

This approach can be used to train a classification algorithm to predict whether a molecule may be apt to form crystals with an accuracy of 90.3%. The comparison of the machine learning algorithms shows that the SVM algorithm with RBF kernel provided the best model when trained.

Only a few features of a molecule dominate its propensity to crystallise. Feature selection followed by retraining of the algorithm using the two most important features gave an accuracy of 80%. The single variable classifier approach found the most important features of the molecules to be ${}^0\chi^v$ and the rotatable bond count. The importance of these descriptors has been rationalised using thermodynamic arguments.

Experimental testing of the predictions validated the use of the CSD as the crystalline dataset and ZINC as the non-crystalline dataset once CSD molecules are removed. The crystallisation screen of 19 test molecules gave a prediction accuracy of 79%. A carefully curated set of crystallisation conditions and outcomes would provide a much better target for application of machine learning, but due to the large number of unknown factors influencing crytallisation, we anticipate that such a database would have to be very large in order to make useful predictions.

As the SVM algorithm outputs a score which ranks molecules according to predicted propensity to crystallise, it can potentially be applied to guide synthetic derivatisation of target molecules in order to increase or decrease crystallinity as desired.

5 Acknowledgements

We would like to thank Greg Landrum (RDKit) and the authors of Scikit-learn. JGPW is currently supported by a joint EPSRC doctoral training partnership grant and an STFC postgraduate studentship.

References

1 Y. Pan, J. Jiang, R. Wang, H. Cao and Y. Cui, *J. Hazard. Mater.*, 2009, **168**, 962–9.

2 B. Louis, J. Singh, B. Shaik, V. K. Agrawal and P. V. Khadikar, *Chem. Biol. Drug Des.*, 2009, **74**, 190–5.

3 C. X. Xue, R. S. Zhang, H. X. Liu, M. C. Liu, Z. D. Hu and B. T. Fan, *J. Chem. Inf. Comp. Sci*, 2004, **44**, 1267–74.

4 A. Varnek, N. Kireeva, I. V. Tetko, I. I. Baskin and V. P. Solov'ev, *J. Chem. Inf. Model.*, 2007, **47**, 1111–22.

5 D. Murnane, C. Marriott and G. P. Martin, *Crystal Growth & Design*, 2008, **8**, 2753–2764.

6 X. He, U. J. Griesser, J. G. Stowell, T. B. Borchardt and S. R. Byrn, *Journal of pharmaceutical sciences*, 2001, **90**, 371–88.

7 M. A. Deij, T. Vissers, H. Meekes and E. Vlieg, *Cryst. Growth Des*, 2007, **7**, 778–786.

8 D. L. Schruben, J. a. Stephanus and M. E. Gonzalez, *Crystal Growth & Design*, 2009, **9**, 2794–2800.

9 B. Rupp and J. Wang, *Methods*, 2004, **34**, 390–407.

10 P. Smialowski, T. Schmidt, J. Cox, A. Kirschner and D. Frishman, *Proteins*, 2006, **62**, 343–55.

11 G. Babnigg and A. Joachimiak, *Journal of structural and functional genomics*, 2010, **11**, 71–80.

12 M. J. Mizianty and L. Kurgan, *Bioinformatics*, 2011, **27**, 24–33.

13 S. Jahandideh and A. Mahdavi, *J. Theor. Biol.*, 2012, **306**, 115–9.

14 I. M. Overton, C. A. J. van Niekerk and G. J. Barton, *Proteins*, 2011, **79**, 1027–33.

15 J. P. M. Lommerse, W. D. Motherwell, H. L. Ammon, J. D. Dunitz, A. Gavezzotti, D. W. M. Hofmann, F. J. J. Leusen, W. T. M. Mooij, S. L. Price, B. Schweizer, M. U. Schmidt, B. P. van Eijck, P. Verwer and D. E. Williams, *Acta Crystallogr. B*, 2000, **56**, 697–714.

16 W. D. S. Motherwell, H. L. Ammon, J. D. Dunitz, A. Dzyabchenko, P. Erk, A. Gavezzotti, D. W. M. Hofmann, F. J. J. Leusen, J. P. M. Lommerse, W. T. M. Mooij, S. L. Price, H. Scheraga, B. Schweizer, M. U. Schmidt, B. P. van Eijck, P. Verwer and D. E. Williams, *Acta Crystallogr. B*, 2002, **58**, 647–661.

17 G. M. Day, W. D. S. Motherwell, H. L. Ammon, S. X. M. Boerrigter, R. G. Della Valle, E. Venuti, A. Dzyabchenko, J. D. Dunitz, B. Schweizer, B. P. van Eijck, P. Erk, J. C. Facelli, V. E. Bazterra, M. B. Ferraro, D. W. M. Hofmann, F. J. J. Leusen, C. Liang, C. C. Pantelides, P. G. Karamertzanis, S. L. Price, T. C. Lewis, H. Nowell, A. Torrisi, H. a. Scheraga, Y. a. Arnautova, M. U. Schmidt and P. Verwer, *Acta Crystallogr. B*, 2005, **61**, 511–27.

18 G. M. Day, T. G. Cooper, A. J. Cruz-Cabeza, K. E. Hejczyk, H. L. Ammon, S. X. M. Boerrigter, J. S. Tan, R. G. Della Valle, E. Venuti, J. Jose, S. R. Gadre, G. R. Desiraju, T. S. Thakur, B. P. van Eijck, J. C. Facelli, V. E. Bazterra, M. B. Ferraro, D. W. M. Hofmann, M. A. Neumann, F. J. J. Leusen, J. Kendrick, S. L. Price, A. J. Misquitta, P. G. Karamertzanis, G. W. A. Welch, H. A. Scheraga, Y. A. Arnautova, M. U. Schmidt,

J. van de Streek, A. K. Wolf and B. Schweizer, *Acta Crystallogr. B*, 2009, **65**, 107–25.

19 D. A. Bardwell, C. S. Adjiman, Y. A. Arnautova, E. Bartashevich, S. X. M. Boerrigter, D. E. Braun, A. J. Cruz-Cabeza, G. M. Day, R. G. Della Valle, G. R. Desiraju, B. P. van Eijck, J. C. Facelli, M. B. Ferraro, D. Grillo, M. Habgood, D. W. M. Hofmann, F. Hofmann, K. V. J. Jose, P. G. Karamertzanis, A. V. Kazantsev, J. Kendrick, L. N. Kuleshova, F. J. J. Leusen, A. V. Maleev, A. J. Misquitta, S. Mohamed, R. J. Needs, M. A. Neumann, D. Nikylov, A. M. Orendt, R. Pal, C. C. Pantelides, C. J. Pickard, L. S. Price, S. L. Price, H. A. Scheraga, J. van de Streek, T. S. Thakur, S. Tiwari, E. Venuti and I. K. Zhitkov, *Acta Crystallogr. B*, 2011, **67**, 535–51.

20 H. P. G. Thompson and G. M. Day, *Chemical Science*, 2014, **5**, 3173.

21 S. L. Price, *Phys. Chem. Chem. Phys.*, 2008, **10**, 1996–2009.

22 S. L. Price, *Accounts Chem. Res.*, 2009, **42**, 117–26.

23 P. Strohriegl and J. V. Grazulevicius, *Adv. Mater.*, 2010, **14**, 1439–1452.

24 J. D. Wuest and O. Lebel, *Tetrahedron*, 2009, **65**, 7393–7402.

25 E. Gagnon, T. Maris, P.-M. Arseneault, K. E. Maly and J. D. Wuest, *Cryst. Growth Des*, 2010, **10**, 648–657.

26 G. Landrum, *RDKit: Open-source cheminformatics*, `http://www.rdkit.org/`.

27 F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot and E. Duchesnay, *J. Mach. Learn. Res.*, 2011, **12**, 2825–2830.

28 J. J. Irwin and B. K. Shoichet, *J. Chem. Inf. Model.*, 2005, **45**, 177–82.

29 F. H. Allen, *Acta Crystallogr. B*, 2002, **58**, 380–388.

30 W. T. A. Harrison, J. Simpson and M. Weil, *Acta crystallogr. E*, 2009, **66**, e1–2.

31 C. A. Lipinski, F. Lombardo, B. W. Dominy and P. J. Feeney, *Adv. Drug Deliver. Rev.*, 1997, **23**, 3–25.

32 C. A. Lipinski, *Drug Discov. Today: Technologies*, 2004, **1**, 337–341.

33 C. J. C. Burges, *Data Min. Knowl. Disc.*, 1998, **2**, 121–167.

34 B. Schölkopf and A. Smola, *Learning with kernels: support vector machines, regularization, optimization, and beyond*, 2002.

35 L. Breiman, *Machine Learning*, 2001, **45**, 5–32.

36 Q. Liu, C. Chen, Y. Zhang and Z. Hu, *Artificial Intelligence Review*, 2011, **36**, 99–115.

37 B. Sarojini, N. Ramaraj and S. Nickolas, *CCIS 40*, 2009, pp. 533–543.

38 I. Guyon and A. Elisseeff, *Journal of Machine Learning Research*, 2003, **3**, 1157–1182.

39 A. P. Bradley, *Pattern Recognition*, 1997, **30**, 1145–1159.

40 L. Kier, *Molecular Connectivity In Chemistry And Drug Research*, 1977, pp. 50–60.

41 M. Protic and A. Sabljic, *Aquatic Toxicology*, 1989, **14**, 47–64.

42 L. Yu, S. M. Reutzel-Edens and C. A. Mitchell, *Organic Process Research & Development*, 2000, **4**, 396402.

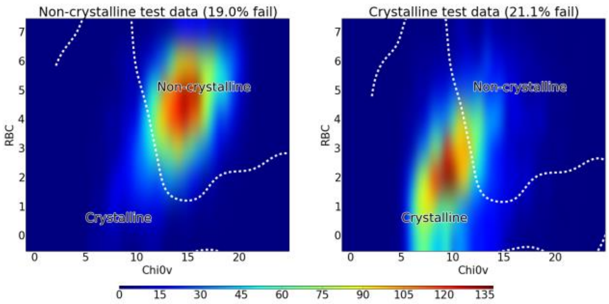
43 J. J. De Yoreo and P. G. Vekilov, *Reviews in mineralogy and geochemistry*, 2003, **54**, 57–93.

44 H. Morgan, *J. Chem. Doc.*, 1965, **5**, 107–112.

45 D. Rogers and M. Hahn, *J. Chem. Inf. Model.*, 2010, **50**, 742–54.

46 J. H. Ward, *J. Am. Stat. Assoc.*, 1963, **58**, 236–244.

Graphical Abstract



Machine learning algorithms can be used to create models which separate molecular materials which will form good-quality crystals from those that will not, and predict how synthetic modifications will change the crystallinity.