

A publicly available crystallisation data set and its application in machine learning[†]

Received 00th January 20xx,
Accepted 00th January 20xx

DOI: 10.1039/x0xx00000x

www.rsc.org/

Max Pillong,^a Corinne Marx^a, Philippe Piechon^a, Jerome G. P. Wicker^b, Richard I. Cooper^b and Trixie Wagner^a

We present here the crystallisation outcomes for 319 publicly available compounds in up to 18 different solvents spread over 5,710 individual single solvent evaporation trials. The recorded data is part of a much larger, corresponding in-house data base and includes both positive as well as negative crystallisation outcomes. Such data can be used for statistical analyses of solvent performances, machine learning approaches or to investigate the crystallisation behaviour in structurally similar compound classes. The presented data suggests that crystallisation behaviour in different solvents is not correlated to chemical similarity among clusters of highly similar compounds. Further, our machine learning models can be used to guide the solvent choice when crystallising a compound. In a retrospective evaluation, these models proved potent to reduce the workload to a third of our initial protocol, while still guaranteeing crystallisation success rates >92%.

1. Introduction

Machine learning algorithms currently receive an increasing amount of interest throughout multiple fields within chemistry and biology^{1–10}. This includes areas such as the prediction of physicochemical properties², reaction types³, automated image recognition⁴, biological activities of small molecules⁵ or recently even olfactory perception of molecular fragments⁶. In the field of crystallography, recent studies have used machine learning approaches for the prediction of the general crystallinity of compounds^{7,8}, identification of organic porous molecular crystals⁹ or the prediction of suitable solvents for small molecule crystallisation¹⁰. Especially the latter is aggravated by a lack of data available for training, as machine learning algorithms usually require balanced data sets including a sufficient amount of both positive and negative results^{11,12}, the latter of which are rarely reported or documented¹³. In principle, machine learning algorithms try to map numerical or categorical properties to an observable outcome¹⁴. This outcome can be either quantitative or qualitative. Examples for quantitative outcomes are numerical properties such as melting points^{2,15}, logP values¹⁶ or the affinity towards a biological target, commonly expressed as the IC₅₀ value, or the dissociation constant k_d ¹⁷. Qualitative outcomes are applied in classification tasks and include for instance the discrimination between biologically active and inactive compounds in respect to a given activity threshold¹⁸,

reaction types³ or general crystallinity of compounds⁸. In all cases, a variety of numerous mathematical models of varying complexity and customisability exist^{11,14}. These models range from simple linear regression models that try to linearly correlate numerical properties to the outcome, up to highly complex algorithms trying to separate the input data in high-dimensional vector spaces using so-called kernel functions^{11,19}. The numerical or categorical properties used in these models are commonly referred to as descriptors²⁰. For small, drug-like molecules, a wide range of these descriptors exist, also in varying complexity from simple one-dimensional numerical properties such as the molecular weight up to high-dimensional bitvectors describing the presence or absence of molecular features within the molecule²⁰. Regardless of the descriptors or models used, every approach embodies the underlying principle that there is a relationship between the molecular structure and the property captured by the descriptor and thus similar molecules command similar properties²¹.

1.1 Crystal structure determination

Determining the three-dimensional structure of a low molecular weight molecule via x-ray diffraction is an important step in the development of novel small molecule drugs²². Such structures can be used for synthesis verification, as a basis computational modelling in lead design studies and most importantly in the formulation of a drug. The importance of this knowledge in the latter discipline is reflected in the case of Abbot's HIV protease inhibitor Ritonavir, which had to be temporarily withdrawn from the market and reformulated after a second, more stable crystalline salt form was discovered²³ that lowered the bioavailability of the drug. The most important prerequisite for x-ray diffraction are single crystals of suitable size and quality. Depending on the method, suitable sizes range from approximately 10 μm for synchrotron

^a Global Discovery Chemistry Analytics, Novartis Institutes for Biomedical Research, 4002 Basel, Switzerland

^b Chemical Crystallography, Chemistry Research Laboratory, Mansfield Road, Oxford, United Kingdom

[†]Electronic Supplementary Information (ESI) available: Structures, solubility and crystallisation outcomes for all public compounds in respective solvents. CSD analysis of most frequently found solvents. Basic scaffolds for public compound clusters. Example code. See DOI: 10.1039/x0xx00000x

radiation up to 200 μm for a lab-based diffractometer. One commonly used method to grow such crystals is the slow evaporation of solvent²⁴⁻²⁶. Here, the compound is dissolved in a solvent or a mixture of solvents and stored within an area of consistent (temperature, humidity and pressure) atmospheric conditions in a non-sealed vessel. Due to the evaporating solvent, the mixture will go into supersaturation, at which point the dissolved compound will either grow into ordered crystals or precipitate as amorphous material²⁴. As the choice of solvent can be crucial for the formation of crystals, several different solvents or combinations of solvents in different experimental procedures are usually trialled. While a trained crystallographer is able to rationalise these conditions based on experience, there have been only few attempts to use automated algorithms in order to tackle this issue of experimental optimisation¹⁰. In the Global Discovery Chemistry Analytics group at the Novartis Institute for Biomedical Research in Basel, the common protocol is to crystallise each compound in twelve different standard solvents in order to assure the growth of crystalline material suitable for x-ray diffraction from at least one of the experiments. Thus, when using this standardized approach, we require a minimum amount of 20 mg of purified substance to account for the standard protocol as well as potential follow up experiments if the single solvent experiments do not succeed. In order to reduce both the required minimum amount, as well as the workload, we set out to implement machine learning models that are able to predict the tendency of a compound to crystallise in a given solvent. While similar approaches have been made in the past^{7,8,10}, none of these studies have made use of experimentally unified crystallisation data in order to predict the optimal solvent to crystallise a compound in. We thus investigate, whether the collection of such unified crystallisation data is helpful in achieving this goal.

1.2 Suitable data for machine learning

When attempting to generate successful machine learning models there are several requirements for the data used. The two most common bottlenecks in machine learning studies are usually the quantity and the quality of the data¹¹. While the issue of quantity is usually concerned simply with the number of available data points, the issue of quality addresses several different aspects of the data. In classification tasks like the one presented here for instance, it is highly desirable to have a balanced spread of classes throughout the data set¹¹. If this is not given, machine learning models can be prone to a preference of the overrepresented class. In some models, this issue can be addressed by introducing class weights in such a way that the class instances receive a weight inversely proportional to their overall occurrence¹¹. While this can help to address the issue of an unbalanced dataset, there still is a need for a minimum amount of representatives from both classes that allows the models to distinguish between them. Another crucial criterion is the comparability of the data. Especially in data based on biological or chemical experiments, it is of utmost importance that the data used stems from

unified, comparable experimental setups. For biological data, this implies comparable assay protocols when assessing biological activity of small molecules, in crystallisation trials this calls for consistent conditions with respect to concentration, evaporation speed and temperature as well as pressure. Combining data from non-consistent experimental protocols can introduce considerable amounts of noise in the data, misleading the machine learning approaches.

2. Methods

2.1 Crystallisation experiments

All crystallisation experiments conducted and used in this study were standardised in order to collect comparable data. This included standard conditions for every sample in terms of used solvents, amounts, temperature and vessel. For every compound, 1 mg of solid powder was dissolved in 200 μl of solvent inside a Schott 7x80 mm glass vial. Vials were sealed with one layer of aluminium foil which was incised in two spots and then stored in a hood to allow for slow evaporation of the solvent. For the public data set of compounds presented in this study, this procedure was conducted in our standard set of twelve solvents plus an additional six solvents determined from text mining the Cambridge Structural Database (CSD)²⁷, resulting in a total of 5,710 recorded crystallisation outcomes (excluding compounds with insufficient material and in-house data). The twelve standard solvents used were methanol (MeOH), ethanol (EtOH), *iso*-propanol (iPrOH), acetone (ACT), ethyl-methylketone (EMK), *iso*-butyl-methylketone (iBMK), ethyl acetate (EA), *tert*-butyl-methylether (tBME), tetrahydrofuran (THF), chloroform (CHCl₃), toluene (TOL) and acetonitril (ACN). In addition, we text-mined the CSD for the most common solvents in single solvent experiments excluding salts to test our standard set of solvents against further commonly used ones. Based on the results from the text-mining effort, we included solvents *N,N*-dimethylformamide (DMF), dichloromethane (DCM), diethylether (DEE) and hexane (HEX) into our experiments, due to their high frequency in the CSD. Further, nitromethane (NiMe) and chlorobenzene (MCB) were also included to enrich the chemical space occupied by the solvent selection, based on previously published analyses of solvent chemical space^{25,28}. A full disclosure of the most frequently found solvents in the CSD can be found in the supplement (Fig. S1). For each solvent, we visually characterized the relative solubility into one of four different classes ranging from *readily soluble* (No floating residue after addition of solvent), over *kinetically soluble* (No floating residue after vortexing), *thermally soluble* (No floating residue after heating to 40°C and vortexing) to *partially soluble/insoluble* (Either cloudy solution or floating residue after heating and vortexing or material remains unaffected after addition of solvent, heating and vortexing). After evaporation of the solvents, the resulting residue was examined for crystalline arrangements under a microscope. The crystallinity was classified into one of six different classes, ranging from *crystals* (XX: >30 μm size in the shortest

dimension), over *microcrystals* (YX: <30 µm size or needle-like structures), *crystalline tendencies* (CT: crystalline tendencies indicated by polarized light but no extractable specimen suited for crystallography), *droplets* (DR) and *films* (FI) to *amorphous* (AM). The first two classes (XX and YX) were considered as suitable material for x-ray diffraction.

2.2 Public compound selection

As the majority of the data stored in our in-house crystallisation data base is confidential, a specific set of non-confidential, publicly available compounds was compiled for this study in order to test the crystallisation behaviour of structurally similar compounds and as an evaluation set for this and potentially further machine learning approaches. Several filters were applied in order to isolate a subset of diverse, but drug-like compounds for the analyses. Firstly, we filtered for non-confidential compounds in our in-house library with an annotated UV purity of at least 80%. Further, we excluded all salts from the selection, as well as compounds with less than 100 mg available in the archive. Next, we filtered out compounds with a molecular weight above 500 g/mol and such compounds with less than ten heavy atoms, as well as compounds that featured saturated alkyl chains longer than nine sp³-hybridized carbon atoms. This resulted in a total of 16,949 non-confidential compounds, which were then clustered into similarity clusters based on the Tanimoto distance²⁹ of their respective Morgan fingerprints (RDKit, version 2016.3.01)³⁰. These fingerprints consider both connectivity as well as chemical features of each atom in the molecule, taking into account its direct neighbouring atoms over the course of a given number of bonds. This information is hashed into a bitvector, where set bits correspond to certain fragments within the molecule (In this study, we considered all neighbours within four bonds and folded them into fingerprints of 4,096 bits of length). As similar molecules are comprised of similar fragments, they are thus encoded in similar bitvectors. The similarity of two bitvectors can then be assessed through their Tanimoto similarity by calculating the fraction of all commonly set bits in both vectors (intersection) over all bits set in either of the vectors (union), as shown in Eq. 1²⁹.

$$T_s(A, B) = \frac{\sum_{i=1}^n (A_i \cap B_i)}{\sum_{i=1}^n (A_i \cup B_i)} \quad (1)$$

Due to the fraction shown in Equation 1, the resulting values range between 0 (no similarity) and 1 (full similarity). These Tanimoto similarities were calculated for all compounds in the data set and stored in a similarity matrix. Based on this similarity matrix, we used an unsupervised clustering algorithm published by Butina³¹ in order to cluster the compounds into clusters of highly similar molecular structure. For this, the clustering algorithm first determines the number of neighbours for each compound in the data set. A pair of neighbours is defined as two compounds of or above a certain threshold in their Tanimoto similarity (In this study, we used a

threshold of 0.5). After re-sorting all compounds according to their number of neighbours, the algorithm proceeds by using the compound with the most neighbours as a cluster centroid and then adding each compound with a Tanimoto similarity of or above the given threshold to that cluster. Clustered compounds and the cluster centroids are flagged and thus no longer considered for further clustering. This approach is then iterated until each compound is flagged as either being within a cluster of compounds or forming a singleton with itself. This clustering procedure resulted in a total of 9,641 clusters, out of which 146 clusters contained at least ten compounds and 7,123 clusters were singletons. Based on these clusters, we used a min-max algorithm³² to select 15 structurally diverse clusters based on each cluster's centroid. For a disclosure of the cluster sizes, as well as basic scaffolds for the selected clusters, see supplementary Figures S2 and S3.

2.3 Data bases

In total, two data bases were used in these studies. The first data base (referred to as OXDB) consists of the 319 publicly available vendor compounds clustered into 15 clusters of varying size as described in the previous section. Each compound was trialled for crystallisation propensity in each of the 12 standard solvents plus the additional 6 solvents extracted from the CSD, as described in section 2.1. The results of the solubility assessment and the crystallisation propensity as well as the structure of the compounds can be found in the supplemental material. The second data base (referred to as GUIDEX) incorporated only into the statistics on solubility & crystallisation propensity is the confidential in-house data base including compounds developed, synthesised and crystallised at Novartis over the past five years. This data base features a total of 1,965 samples trialled in varying solvents. An overview of the number of experiments and outcomes found for each solvent within the data bases is given in Table 1. Both the data for OXDB, as well as GUIDEX were stored in an OracleSQL data base for ease of access. The according entity-relationship (ER) diagram can be found in the supplement (Fig. S4).

Table 1 Shown are the solvent individual total numbers of experiments and their respective class quantities. Positive outcomes are crystals and microcrystals, all other crystallisation outcomes are considered negative.

solvent	total	positive	negative	% positive
MeOH	1423	528	895	37.1
EtOH	1026	385	641	37.5
iPrOH	931	343	588	36.8
ACT	1212	435	777	35.9
EMK	889	348	541	39.1
iBMK	904	350	554	38.7
EA	864	380	484	44.0
tBME	552	214	338	38.8
THF	895	185	710	20.7
CHCl ₃	835	219	616	26.2
TOL	828	372	456	44.9
ACN	1105	469	636	42.4

2.4 Machine learning algorithms

In order to predict the crystallisation propensity of the selected compounds, we applied random forest prediction models, as implemented in the SciKit Learn (version 0.17.1, <http://www.scikit-learn.org>)³³ library for python (version 2.7, <https://www.python.org>). Random forests are a collection of randomized, binary decision trees that give a prediction based on a majority vote over all individual trees within the model. In decision trees, the prediction probability p for a certain class c_1 within a leaf i is expressed as the fraction of all instances from class c_1 over the total number n of instances within leaf i during the training step (Equation 2).

$$p(c_1)_i = \frac{|c_1|_i}{n_i} \quad (2)$$

Transferred to random forests, the overall prediction probability p_{RF} is the average prediction probability over all estimators E in the random forest model.

$$p_{RF}(c_1) = \frac{1}{E} * \sum_{i=1}^E p(c_1)_i \quad (3)$$

While the final prediction for an instance is solely based on the class that is primarily predicted over all decision trees, the prediction probability can be seen as an overall confidence of a model's prediction for an instance. In this work, all random forest models were trained using the single solvent experiments stored in GUIDEX, as well as a randomly selected 50% of the OXDB compounds, in order to ensure an appropriate coverage of the chemical space predicted. The remaining 50% of the OXDB compounds were then used to evaluate the models accordingly. To ensure a high quality standard of the data used, we measured the purity of all samples using LCMS and excluded samples with a measured purity below 99%. For each of the twelve standard solvents from our protocol, 100 individual random forest models were implemented applying repeated random sub-sampling and evaluated in order to assess the performance of the yielded models. Each random forest consisted of 100 individual estimators. Splits were evaluated using the *gini* criterion with automated class weights to counter class-imbalanced training sets. No maximum depth or number of features was given and out-of-bags samples were used to estimate the generalization accuracy. All compounds were represented by their respective Morgan fingerprint as elaborated in section 2.2, calculated using the RDKit³⁰.

3. Results and discussion

The relative solubility of the compounds was assessed visually into one of five classes as mentioned in section 2.1. Figure 1a shows a breakdown of the different solubilities for the overall 16,362 experiments recorded from the in-house and the public data. When combining the classes of readily soluble with kinetically and thermally soluble, the highest solubility rates were found in solvents tetrahydrofuran (92.6%), methanol (92.2%) and acetone (90.8%). The lowest solubility rates were

found in solvents *tert*-butylmethether (49.07%), ethyl acetate (79.7%) and toluene (71.9%). However, these results did not reflect upon the crystallisation propensities of the compounds within these given solvents. As shown in Figure 1b, the aforementioned tetrahydrofuran showed the lowest overall growth of crystalline material suitable for measurements (20.9%), whereas toluene (46.1%) and ethyl acetate (46.3%) resulted in the highest total number of crystalline material. In order to compare our standard set of solvents with the additional solvents extracted from the CSD screening, we investigated the uniqueness of the positive crystallisation outcomes (U_s) for each of the standard and extended solvents according to Equation 4.

$$U_s = \frac{\sum_{i=1}^{|N_s|} \frac{1}{C_i}}{|N_s|} * 10, \quad (4)$$

where N_s is the set of all compounds that crystallised in solvent S and C_i denotes the overall number of positive crystallisation outcomes for compound i in all twelve solvents. Thus, U_s gives a numerical measure for how unique a solvent's crystallisation outcomes are on average. Using this score as a weight to scale the overall crystallisation success for the public test set reveals several insights into the solvents' performances (Table 2).

Table 2 Shown are the relative and the scaled crystallisation success ratings for each of the 18 solvents in the public data set. U_s is calculated as denoted in Equation 5.

Solvent	U_s	crystallisation rate [%]	scaled crystallisation rating
MeOH	1.04	39.50	41.11
EtOH	0.98	43.89	42.83
iPrOH	1.07	46.50	49.66
ACT	0.93	48.59	45.33
EMK	0.97	49.52	47.85
iBMK	0.98	42.63	41.87
EA	0.95	50.16	47.71
tBME	0.95	48.26	45.86
THF	0.92	31.43	28.84
CHCl ₃	0.87	38.17	33.24
TOL	1.05	51.10	53.70
ACN	1.14	51.89	59.36
NiMe	1.09	55.17	60.20
DMF	1.29	34.48	44.53
MCB	0.99	36.48	36.19
DCM	0.83	36.19	29.89
DEE	1.04	54.60	56.91
HEX	1.32	48.41	64.00

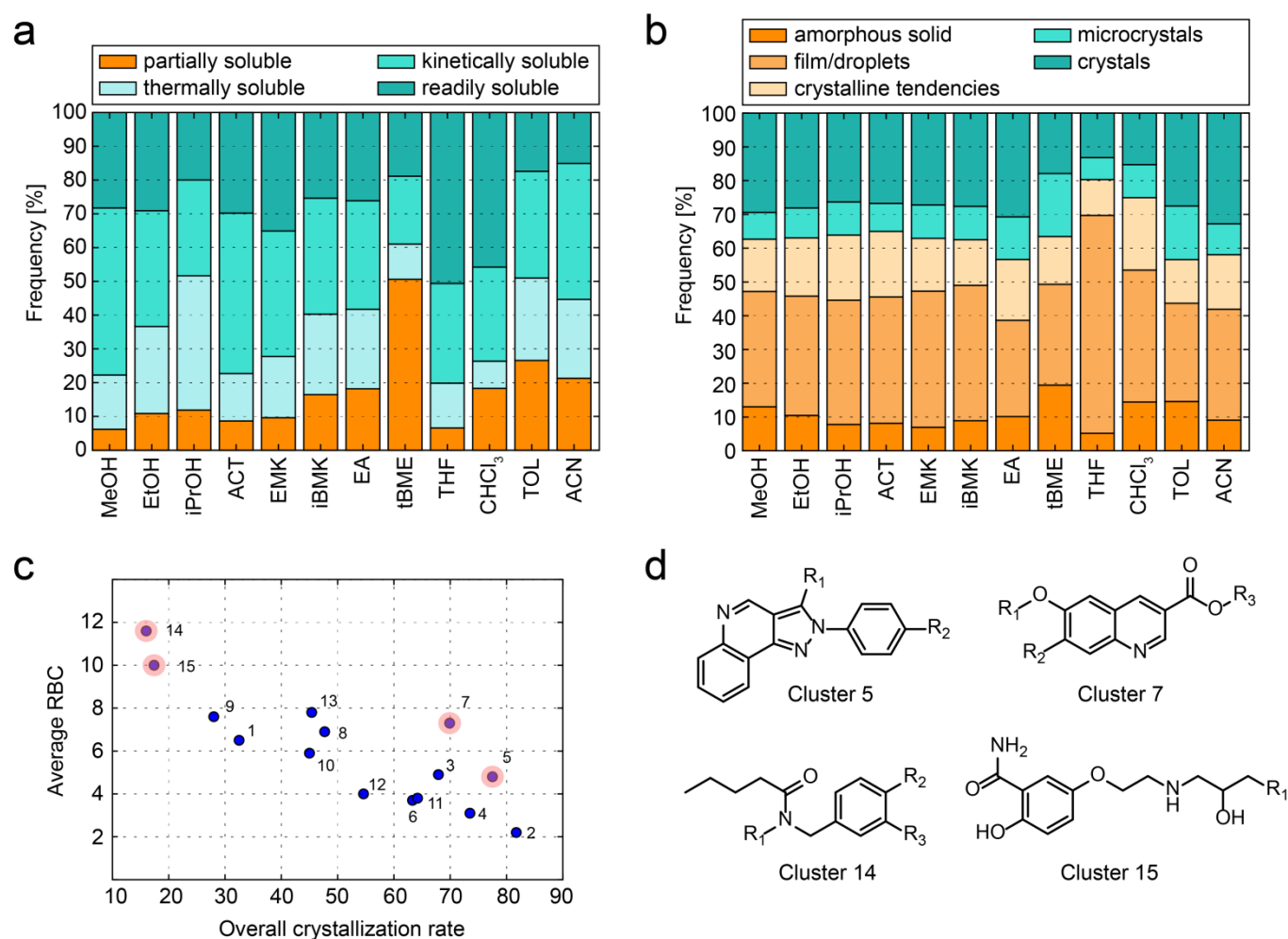


Fig. 1 Shown are the different classifications used for solubility (a) and crystallinity (b) for the twelve standard solvents used c) Overall crystallisation success rate and average rotatable bond count for all clusters. d) Basic scaffold for clusters 5, 7, 14 and 15.

Solvents toluene and acetonitrile not only feature a high crystallisation success rate for this dataset (51.1% and 51.9% respectively), but also U_s values above 1, increasing their scaled crystallisation rating to 53.7 and 59.4. This indicates their unique ability to crystallise compounds that prove more difficult to crystallise within our standard methods. Further, the crystallisation rates of solvents tetrahydrofuran (31.4%) and chloroform (38.1%) are even further diminished by a low U_s , indicating that these solvents tend to crystallise compounds that are easily crystallised within the standard setting anyways. For the additional solvents extracted from the CSD, there is a clear indication that solvents nitromethane, dimethylformamide and hexane could greatly improve the crystallisation outcomes within this dataset, whereas chlorobenzene and dichloromethane do not seem to add relevant crystallisation outcomes to the overall results. Thus, these results would point towards an exchange of solvents tetrahydrofuran and chloroform for any two of the solvents nitromethane, dimethylformamide and hexane in our standard protocol. When inspecting the crystallisation propensities of the public compounds by cluster rather than solvent, there is a clear indication of clusters that tend to crystallise better than

others (Figure 1c). Clusters 2 through 8 as well as cluster 11 generally exhibited high crystallisation propensities (>63%), the opposite case is exhibited by clusters 1, 9, 14 and 15, as these clusters exhibit poor overall crystallisation propensities (<33%). The remaining clusters feature medium overall crystallisation propensities between 33% and 63%. The overall crystallisation propensity over the whole dataset averages to 52.3%. As expected, the overall cluster specific crystallisation propensities correlate well with the rotatable bond count (RBC) found on average within each cluster (Fig. 1c). Clusters 14 and 15 exhibit the lowest crystallisation success rates and the highest average RBC, while cluster two features the highest overall crystallisation rate and the lowest average RBC. Two considerable exceptions to this correlation are clusters 5 and 7, with overall crystallisation propensities of 73.5% and 69.9% despite their average 4.8 and 7.3 rotatable bonds. These exceptions can be explained by the (phenyl-pyrazolo-) quinoline motif present in these clusters (Fig. 1d). In both cases, this considerably large aromatic scaffold creates an overall planar core of the molecules that allows for favourable π -stacking interactions in the crystals, while the rotatable bonds introduce flexibility only in the sidechains. In contrast,

clusters 14 and 15 feature rotatable bonds in the centre of the scaffold, thus greatly increasing the overall flexibility of these molecules. For a full disclosure of each cluster's structural scaffold, see supplementary Figure S3. Even though the compound clusters feature compounds of high chemical similarity (Tanimoto similarity threshold of 0.5), for some clusters, the crystallisation propensities vary greatly. This is reflected by the analysis shown in Figure 2. In this analysis we investigate whether the chemical similarity expressed by the Morgan fingerprint (MFP) introduced in section 2.2 dictates a similarity in crystallisation propensity. In order to do so, we introduce a second fingerprint based on the crystallisation outcomes in the twelve standard solvents used throughout all experiments. This crystallisation fingerprint (CFP) thus consists of a twelve-dimensional bitvector, where each dimension represents the binarised crystallisation outcome in the respective solvent. Bits of solvents that yielded either crystals or microcrystals are set to 1, whereas all other crystallisation outcomes are denoted as 0. Based on the Tanimoto similarity we can thus calculate a value ranging from 0 to 1 that indicates how similar two compounds' crystallisation propensities are. This analysis was conducted for each compound towards all compounds in its own cluster (intra), as well as towards all compounds in the other clusters (inter). For each of the two fingerprints, the average inter- and intra-cluster similarities were calculated as given in equation 5.

$$\sigma_{fp}(i) = \frac{1}{n} \sum_{j=1}^n T_{s(fp)}(i, j) \quad (5)$$

Where fp is one of the two fingerprints (MFP or CFP), n is the total number of comparisons and $T_s(i, j)$ is the Tanimoto similarity of compounds i and j . This equation can be used to calculate both the average inter- as well as the average intra-cluster similarities by adjusting the search space of j respectively. Figure 2 shows the average intra-cluster similarities for each compound as circles, whereas the average inter-cluster similarities are shown as diamonds. The average CFP similarity is given on the x-axis, the average MFP similarity

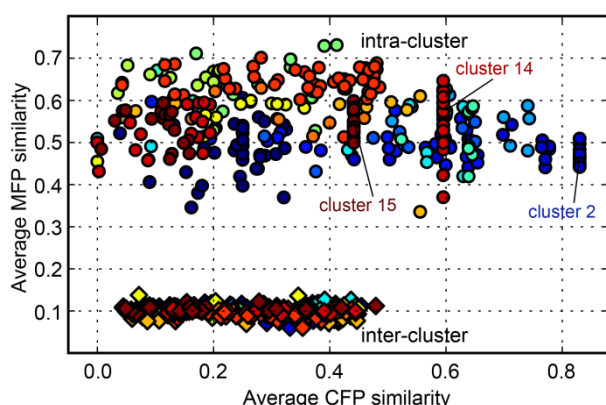


Fig. 2 Shown is a scatter plot of the average distances of each compound towards either its own cluster (intra) or all other clusters (inter). The x-axis gives the distances for the crystallinity fingerprint (CFP), the y-axis gives the distance for the Morgan fingerprint (MFP). Points are coloured according to their cluster.

on the y-axis. As expected, the MFP intra-cluster similarities receive significantly higher values than the MFP inter-cluster similarities, as the same fingerprint was used to generate the clustering. In terms of the CFP however, the CFP intra-cluster similarities span over the complete range of the CFP inter-cluster similarities, indicating that the range of average CFP distances is not bound to chemical similarity, as low CFP similarities are also populated by highly similar compounds. Special cases can be seen in clusters 2, 14 and 15, where the binarisation of the overall either positive (cluster 2) or negative (clusters 14 and 15) crystallisation propensity dictates a highly conserved intra-cluster similarity in the CFP, indicated by those clusters shaping into vertical lines in the plot in Figure 2. From these results, we conclude that a high chemical similarity does not necessarily dictate an expectable similarity in crystallisation propensities.

Based on previously published work by Wicker & Cooper⁸, we set out to base the machine learning models on two distinct numerical descriptors implemented in the RDKit³⁰, namely the rotatable bond count (RBC) and the zero order molecular valence connectivity index ($^0\chi^v$). The latter is a summation over all heavy atoms calculated as shown in Equation 6.

$$^0\chi^v = \sum_{i=1}^n \left(\frac{Z_i^v - h_i}{Z_i - Z_i^v - 1} \right)^{-0.5}, \quad (6)$$

where n is the total number of heavy atoms in the molecule, Z is the atomic number of an atom i , Z^v is the number of valence electrons of an atom i and h is the number of hydrogens attached to an atom i . Based on these descriptors, the random forest predictors for the individual solvents averaged a predictive accuracy of only 60.8% (For individual solvent model outcomes, see Fig. 3). We therefore aimed to improve the description of the molecules by including several other numerical descriptors implemented in the RDKit, namely the number of H-bond donors and acceptors, the Labute and polar accessible surface area^{30,35} and a calculated logP value based on the prediction algorithm of Wildman and Crippen³⁴, as well as the descriptor sets for PEOE_VSA, SMR_VSA and SlogP_VSA taken directly from their respective RDKit modules³⁰. The models based on this extended numerical descriptor set averaged a slightly improved predictive accuracy of 62.3% for the individual solvent models. Thus, we moved away from numerical descriptions of the molecules and re-implemented the models using fingerprint descriptors in order to capture distinct functional and topological features in the molecules. These models averaged a predictive accuracy of 67.8% over all individual solvents. Figure 3a shows the individual performances of the machine learning models for each of the twelve solvents for the fingerprint-based models. The highest predictive accuracy is found in models for ethyl acetate (69.6%), ethanol (69.1%) and acetone (69.0%), whereas the lowest predictive accuracies are found in models for solvents *iso*-propanol (65.9%) and ethyl-methylketone (66.0%). These observations reflect the overall number of training data available for these models, as shown in Table 2. When including the data from the confidential GUIDEX data base into

the training step of the models, these predictive accuracies can be pushed further up (>70.0% predictive accuracy, data not shown due to reasons of confidentiality). This increased predictive accuracy indicates the positive effect of an extended input space available during the training step. From a machine learning point of view, none of the observed predictive accuracies over the different descriptor combinations would be considered sound in order to predict the crystallisation in a given solvent confidently. However, the models used do not only predict one of two classes (crystalline/non-crystalline), but also give a prediction probability, which reflects the models' confidence in the individual prediction as introduced

in section 2.4. In order to assess these prediction probabilities, we conducted the following retrospective analysis. For each compound in the test set that crystallised in at least one of the solvents, we calculated the prediction for each respective solvent and stored the prediction probabilities of all twelve solvent models in an array. We then sorted this array in descending order according to each solvent model's prediction probability. This allowed us to retrospectively prioritize on any given number of experiments for each compound. Evaluating the actual crystallisation of the ranked solvents, we determined an accuracy of 68.2% considering only the most confident model, 83.0% considering the two most confident models and 89.5% considering the three most confident models. The complete assessment for this analysis is shown in Figure 3b. As reflected by this analysis, the crystallisation success rate quickly converges towards 90% when considering three or more experiments, allowing us to reduce the overall experimental effort and required material significantly. As with the initial predictions, these success rates increase when incorporating the additional data from our in-house database. Figure 3c shows a breakdown of the prioritized solvents when considering the three most confident predictions. As can be seen, the most confident predictions range over all twelve solvents, rather than being focused on only a small set of high performing solvents, indicating individual compound-dependant prioritisations.

4. Conclusions

We report here the solubility and crystallisation propensities for 319 small molecules in up to 18 different. To our current knowledge, the resulting data set of 5,710 individual experimental crystallisation outcomes (both positive and negative) over a standard of solvents and experimental setup is the first of its kind to be made publicly available. As indicated by the experimental results, the crystallisation propensity of a compound in any given solvent is not directly related to its solubility in the solvent. In fact, the solvent with the highest overall solubility (THF) was found to be the worst in terms of crystallisation success rate. Further, while there were certain clusters with higher overall crystallisation propensity than others, there was no overall indication of conserved crystallisation behaviour in individual solvents among clusters due to their compounds' structural similarity. While certain clusters exhibited overall positive or negative crystallisation propensities, no conserved crystallisation behaviours over the set of selected solvents was found. Therefore, in order to successfully crystallise new chemical entities, our laboratories usually apply a standard trial including twelve different single solvent experiments. The presented machine learning approach aims to reduce the amount of these experiments and thus required substance in order to yield crystalline material suitable for x-ray crystallography. This was accomplished by training random forest models on a data set comprised of standardized experimental protocols. Our findings indicate that molecular fingerprints were a more potent descriptor for the task at

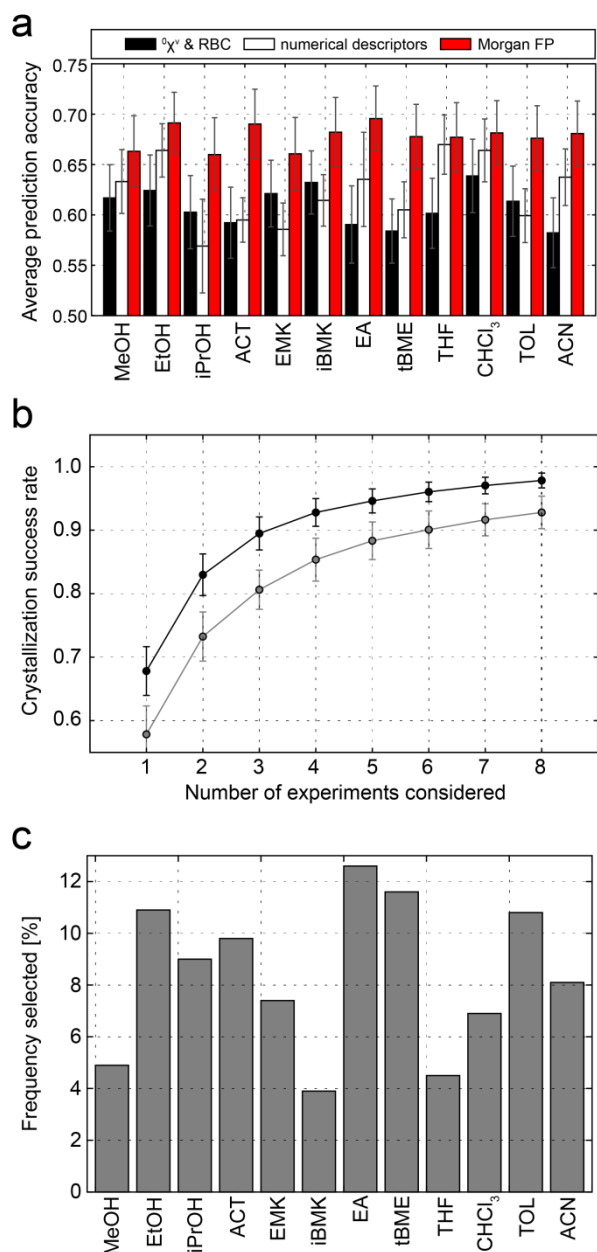


Fig. 3 Shown are the average accuracies for the individual solvent models and the different descriptor combinations (a), as well as the predictive accuracy when considering the n most confident predictions in black and random selection in grey (b). Shown in (c) is the frequency with which each solvent was among the three most confident ones.

hand than previously identified numerical descriptors. A further increase in predictive accuracy could be achieved by including the data from an equivalent but confidential in-house database into the model training. This indicates the potential of such a unified data base as well as the need for extending it accordingly. While the predictive accuracy for predicting the crystallisation state for individual solvents was mediocre at best throughout all models, useful information could be extracted from the prediction probabilities of these individual models. Therefore, rather than predicting the general crystallinity of a molecule, our approach focuses on the prediction of crystallisation propensities in different solvents, allowing for a rational prioritisation of the experimental setup. Considering only the solvent with the highest predicted crystallisation propensity, this approach yielded a retrospective crystallisation success rate of 68.2%. When considering the second and third highest predicted crystallisation propensities, the success rate increases to 83.0% (89.5% respectively). Thus, we are able to induce a crystallisation success rate >92%, while reducing the amount of required experiments to a third of our initial approach, based on the data collected over the past five years.

Acknowledgements

The authors would like to thank the Novartis Institute for Biomedical Research Postdoc Office for funding, as well as Greg Landrum, Nadine Schneider and Nikolas Fechner for discussion and input.

References

1. R. Guha, K. Gilbert, G. Fox, M. Pierce, D. Wild and H. Yuan, *Curr Comput Aided Drug Des*, 2010, 6, 50-67.
2. C. A. S. Bergstrom, U. Norinder, K. Luthman and P. Artursson, *J Chem Inf Comp Sci*, 2003, 43, 1177-1185.
3. N. Schneider, D. M. Lowe, R. A. Sayle and G. A. Landrum, *J Chem Inf Model*, 2015, 55, 39-53.
4. W. J. Godinez, I. Hossain, S. E. Lazic, J. W. Davies and X. Zhang, *Bioinformatics*, 2017, DOI: 10.1093/bioinformatics/btx069.
5. E. Lounkine, M. J. Keiser, S. Whitebread, D. Mikhailov, J. Hamon, J. L. Jenkins, P. Lavan, E. Weber, A. K. Doak, S. Cote, B. K. Shoichet and L. Urban, *Nature*, 2012, 486, 361-367.
6. A. Keller, R. C. Gerkin, Y. Guan, A. Dhurandhar, G. Turu, B. Szalai, J. D. Mainland, Y. Ihara, C. W. Yu, R. Wolfinger, C. Vens, L. Schietgat, K. De Grave, R. Norel, D. O. P. Consortium, G. Stolovitzky, G. A. Cecchi, L. B. Vossall and P. Meyer, *Science*, 2017, 355, 820-826.
7. R. M. Bhardwaj, A. Johnston, B. F. Johnston and A. J. Florence, *Crystengcomm*, 2015, 17, 4272-4275.
8. J. G. P. Wicker and R. I. Cooper, *Crystengcomm*, 2015, 17, 1927-1934.
9. J. D. Evans, D. M. Huang, M. Haranczyk, A. W. Thornton, C. J. Sumby and C. J. Doonan, *Crystengcomm*, 2016, 18, 4133-4141.
10. K. Hosokawa, J. Goto and N. Hirayama, *Chem Pharm Bull (Tokyo)*, 2005, 53, 1296-1299.
11. I. H. Witten, E. Frank and M. A. Hall, *Data Mining: Practical Machine Learning Tools and Techniques*, Morgan Kaufmann, Burlington, 2010.
12. R. Kurczab, S. Smusz and A. J. Bojarski, *J Cheminform*, 2014, 6, 32.
13. N. Matosin, E. Frank, M. Engel, J. S. Lum and K. A. Newell, *Dis Model Mech*, 2014, 7, 171-173.
14. J. B. O. Mitchell, *Wires Comput Mol Sci*, 2014, 4, 468-481.
15. M. Karthikeyan, R. C. Glen and A. Bender, *J Chem Inf Model*, 2005, 45, 581-590.
16. J. L. McDonagh, T. van Mourik and J. B. O. Mitchell, *Mol Inform*, 2015, 34, 715-724.
17. S. Das, M. P. Krein and C. M. Breneman, *J Chem Inf Model*, 2010, 50, 298-308.
18. D. Reker, A. M. Perna, T. Rodrigues, P. Schneider, M. Reutlinger, B. Monch, A. Koeberle, C. Lamers, M. Gabler, H. Steinmetz, R. Muller, M. Schubert-Zsilavecz, O. Werz and G. Schneider, *Nat Chem*, 2014, 6.
19. J. Shawe-Taylor, N. Christianini, *Kernel Methods for Pattern Analysis*, Cambridge University Press, Cambridge, UK, 2004.
20. L. Xue and J. Bajorath, *Comb Chem High Throughput Screen*, 2000, 3, 363-372.
21. C. Hansch, *J Med Chem*, 1968, 11, 920-924.
22. B. Y. Shekunov and P. York, *J Cryst Growth*, 2000, 211, 122-136.
23. J. Bauer, S. Spanton, R. Henry, J. Quick, W. Dziki, W. Porter and J. Morris, *Pharmaceut Res*, 2001, 18, 859-866.
24. J. R. Deschamps, *Life Sci*, 2010, 86, 585-589.
25. B. Spingler, S. Schnidrig, T. Todorova and F. Wild, *Crystengcomm*, 2012, 14, 751-757.
26. H. H. Tung, E. L. Paul, M. Midler and J. A. McCauley, *Crystallization of Organic Compounds: An Industrial Perspective*, Wiley, New York, 2009.
27. C. R. Groom, I. J. Bruno, M. P. Lightfoot and S. C. Ward, *Acta Crystallogr B Struct Sci Cryst Eng Mater*, 2016, 72, 171-179.
28. M. Alleso, J. Rantanen, J. Aaltonen, C. Cornett and F. van den Berg, *J Chemometr*, 2008, 22, 621-631.
29. D. J. Rogers and T. T. Tanimoto, *Science*, 1960, 132, 1115-1118.
30. G. Landrum, RDKit: Open-source cheminformatics, <http://www.rdkit.org>
31. D. Butina, *J Chem Inf Comp Sci*, 1999, 39, 747-750.
32. M. Ashton, J. Barnard, F. Casset, M. Charlton, G. Downs, D. Gorse, J. Holliday, R. Lahana and P. Willett, *Quant Struct-Act Rel*, 2002, 21, 598-604.
33. 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.
34. S. A. Wildman and G. M. Crippen, *J Chem Inf Comp Sci*, 1999, 39, 868-873.
35. P. Labute, *J Mol Graph Model*, 2000, 18, 464-477.