Discovery of highly polymorphic organic materials: a new machine learning approach

Zied Hosni[1,3]*, Annalisa Riccardi[2], Stephanie Yerdelen[1], Alan R. G. Martin[1], Deborah Bowering[1], and Alastair J. Florence[1]*

Polymorphism is the capacity of a molecule to adopt different conformations or molecular packing arrangements in the solid state. This is a key property to control during pharmaceutical manufacturing because it can impact a range of properties including stability and solubility. In this study, a novel approach based on machine learning classification methods is used to predict the likelihood for an organic compound to crystallise in multiple forms. A training dataset of drug-like molecules was curated from the Cambridge Structural Database (CSD) and filtered according to entries in the Drug Bank database. The number of separate forms in the CSD for each molecule was recorded. A metaclassifier was trained using this dataset to predict the expected number of crystalline forms from the compound descriptors. This approach was used to estimate the number of crystallographic forms for an external validation dataset. These results suggest this novel methodology can be used to predict the extent of polymorphism of new drugs or not-yet experimentally screened molecules. This promising method complements expensive ab initio methods for crystal structure prediction and as integral to experimental physical form screening, may identify systems that with unexplored potential.

Machine Learning methods (ML) are ubiquitous in many areas of modern science and have become a crucial tool where large amounts of data from different sources are available. There is a diverse range of ML algorithms available that have been applied to the modelling and prediction of complex systems and problems. Various factors have an impact on the suitability of ML approaches for different applications. Among those are the size and distribution of the training data in the features space, the correlation of the descriptors, the nature of the problem and its degree of non-linearity. The non-linearity of the problem considered in this study is one of the main drivers for the choice of the ML approach used. Support Vector Machine (SVM) and Random Forest (RF) are ML methods that have already been successfully used for classification and prediction of non-linear chemical processes (*i.e.* the features and the response are not correlated with a linear relationship) and they are suitable for large dimensional problems (*i.e.* many factors affect the response of the phenomena)[1,2].

The k-Nearest Neighbours (k-NN) algorithm joins simplicity and intuitiveness. In the Mitchell group, this method was applied to predict the melting point for 4119 structurally diverse organic molecules and 277 drug-like molecules. The performance of this algorithm was compared with the one from neural networks showing the strengths and the weaknesses to exploit their predictive models. Cross-validation and y-randomisation both proved to be good strategies for prediction validation[3]. Tropsha *et al.* highlighted the importance of validation techniques in Quantitative Structure-Property Relationship (QSPR) models before applying them on real world problems. They enumerated several examples of predictive failure when the validation step was not considered carefully[4].

QSPR modelling is by definition based on the assumption that changes in molecular structure are reflected in variation in the observed macroscopic properties of materials[2]. This approach does not require access to expensive, high-performance computing power and has been shown to deliver scalability, efficiency, robustness and predictability[5]. QSPR has been applied to predict a wide range of material properties such as physicochemical and biological properties of nanomaterials[6], catalytic activity in homogeneous and heterogeneous catalysts[7,8], protein adsorption, cell attachment, cellular proliferation on biomaterial surface[9], glass transition temperature for polymers[10], melting points for ionic liquids and others[11].

Crystal structure prediction is a challenging area and one of the promising applications of QSPR and ML. Philipps *et al.* identified new types of crystalline structures from large data sets of coordinates. They deployed a hierarchy of pattern analysis techniques and applied ML with shape matching algorithms to extract and classify crystals into categories[12]. Clustering and the identification of intrinsic structural features in particle tracking data were also investigated using the Neighborhood Graph Analysis (NGA) method[13]. In inorganic chemistry, the Cluster Resolution Feature Selection (CR-FS) and support vector machine (SVM) classification were applied to predict the crystal structures of ternary equiatomic compositions based only on the constituent elements[14]. Moreover, Principle Components Analysis (PCA) was exploited to render structure maps of spinel nitrides (AB_2N_4)[15].

[1] EPSRC Future Continuous Manufacturing and Advanced Crystallisation Hub, University of Strathclyde, Technology and Innovation Centre, 99 George Street, Glasgow, U.K. G1 1RD. [2] Department of Mechanical and Aerospace Engineering, University of Strathclyde, Glasgow, UK. G1 1XJ. [3] Centre of Computational Chemistry, School of Chemistry, Cantock's Close, Bristol, U.K BS8 1TS. email: zied.hosni@bristol.ac.uk; alastair.florence@strath.ac.uk

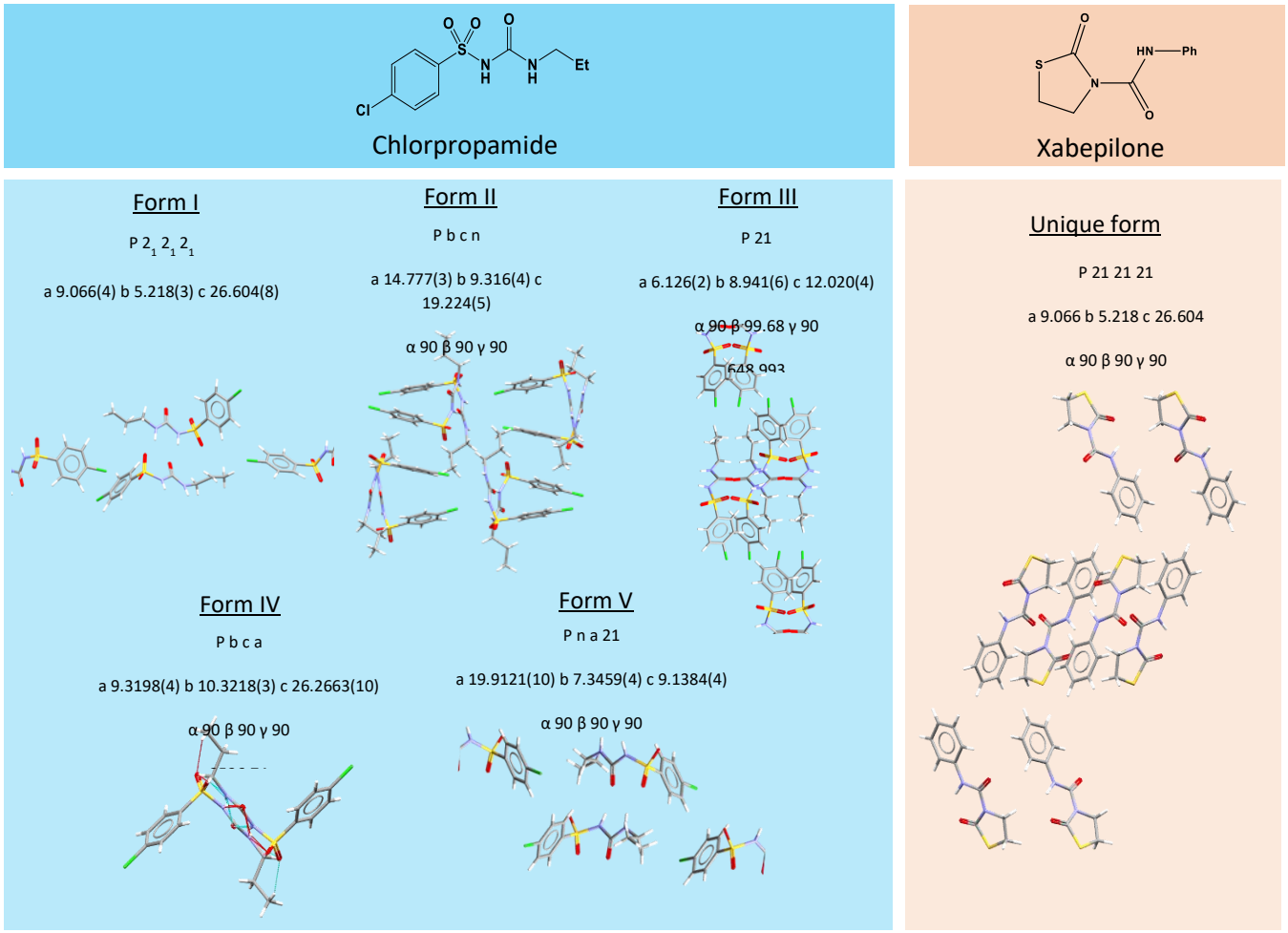

Figure 1 | The 3-dimensional packing and crystallographic information (the distances, the angles and the volume of the unit cell) of two biologically active molecules (Chlorpropamide and Thiazolidinone) showing the five characterised polymorphic forms and the singular reported form in the second.

The Random Forest algorithm was able to predict full-Heusler structures and discriminate between Heusler and inverse Heusler structures[16].

Raccuglia *et al.* exploited a successfully failed experiments' database to highlight the factors that control organically templated metal oxides reaction outcome. They applied different algorithms and found that SVM was the most robust one to mine the chemical information rendered from historical reactions[17]. The identification of crystal structure using optimisation of relevant thermodynamic potential in the space of atomic coordinates is gaining significant interest[18]. Various algorithms such as genetic algorithms or simulated annealing were exploited to determine the global energy configurations of crystal structures[19]. A ML model was trained on a dataset of *ab initio* calculation results for 7000 organic molecules. Various molecular descriptors such as nuclear charges and cartesian coordinates were exploited as features for a deep multi-task artificial neural network capable of predicting atomisation energy, ionisation potential and electron affinity simultaneously[20].

Polymorphism is defined as "the existence of a solid crystalline phase of a given compound resulting from the possibility of at least two different arrangements of the molecules of that compound in the solid state"[21,22]. It is difficult to predict *ab initio* whether a specific molecule will adopt more than one crystal structure, how many polymorphs are likely to be observed, or the specific crystal packing arrangements and associated physical properties each polymorph will display[23,24]. Organic crystals are of paramount importance in different industrial sectors including agrochemicals, food, paint, energetic materials, and pharmaceuticals. The polymorphism of these entities dictates their flowability, stability, colour, solubility and mechanical strength[25]. In addition to the challenges related to the production control of a specific solid form, polymorphism is still posing intellectual property conflicts and prolonged legal battles[26].

Considerable progress has been made in the field of crystal energy landscaping (*i.e.* calculating the thermodynamically feasible crystal structures within an energy landscape of possible polymorphs)[27]. This procedure can guide expensive, and time consuming experimental screening approaches for solid forms if thermodynamic and kinetic factors are both taken into consideration[28,29]. Although numerous *ab initio* predictive methodologies have been developed to deal with increasingly complex challenges (flexible conformers, multicomponent crystals), it is not yet possible to rely on such approaches without carrying out experimental investigations. Figure 1 shows the contrast between two biologically active molecules that present completely different behaviours in terms of experimental polymorphism. Indeed, while Chlorpropamide is reported to form at least 5 different crystalline

forms[30], the cytotoxic Thiazolidinone is only reported to have a single polymorph in the Cambridge Structural Database (CSD)[31].

The CSD does not contain a complete record of the full extent of polymorphism for all chemical entries,[32] rather it includes all the entries that have been reported in the literature or reported directly by researchers to the Cambridge Crystallographic Data Centre (CCDC). Thus, it is possible that some molecules, with a single crystal structure entry, are highly polymorphic once studied in an extensive experimental polymorphic screening.

The Random Forest method, in particular, has previously been successfully applied to the design of experimental screens assessing the completeness of experimental screens for solvate formation[33]; prediction of packing types from solvent properties[34] and the crystallisability of organic molecules[35,36]. However, it has not been assessed as a predictive tool for the extent of polymorphism expected from a target molecule. In this work, we exploit curated data from the intersection of the CSD and Drugbank[37] and implement a metaclassifier that enables the discovery of the true extent of polymorphism in organic molecules and their potential to crystallise in with new solid-state forms. This metaclassifier is the combination of various machine learning algorithms. Four types of datasets were exploited to build predictive models. The most robust model was selected to identify an organic molecule susceptible to exist in several solid states. The experimental validation was conducted by the crystallisation screening of this compound in 60 different solvents.

Table 1. Confusion matrix from the fused predictive model of the case study 1

Number of polymorphs	Experimental polymorphism					
Predicted polymorphism	**1**	**2**	**3**	**4**	**5**	**6**
1	55	17	7	6	6	3
2	3	88	0	3	0	0
3	0	0	94	0	0	0
4	1	1	1	91	0	0
5	0	0	0	0	94	0
6	0	0	0	0	0	94

Table 2. Confusion matrix from the fused predictive model of the case study 2

Number of polymorphs	Experimental polymorphism					
Predicted polymorphism	**1**	**2**	**3**	**4**	**5**	**6**
1	90	3	0	1	0	0
2	1	88	4	0	1	0
3	5	23	53	8	4	1
4	0	0	1	93	0	0
5	0	0	0	0	94	0
6	0	0	0	0	0	94

Table 3. Confusion matrix from the fused predictive model of the case study 3

Number of polymorphs	Experimental polymorphism					
Predicted polymorphism	**1**	**2**	**3**	**4**	**5**	**6**
1	334	18	20	3	14	3
2	8	266	52	31	16	19
3	5	13	343	11	11	9
4	2	11	9	359	3	8
5	0	9	12	5	359	7
6	3	6	20	5	10	348

Table 4. Confusion matrix from the fused predictive model of the case study 4

Number of polymorphs	Experimental polymorphism					
Predicted polymorphism	**1**	**2**	**3**	**4**	**5**	**6**
1	366	4	6	11	2	3
2	13	346	8	19	2	4
3	17	13	297	35	20	10
4	12	2	26	345	5	2
5	4	0	5	7	376	0
6	4	0	13	17	6	352

Case study 1: Polymorphism prediction with 2D descriptors and with dimensionality reduction

The nine statistical models (*i.e.* the eight machine learning models and the Prediction Fusion model) generated for the dataset of 2D structures with dimensionality reduction are summarized in Figure 2.A. The comparison between the different models showed that all the algorithms were successful in reaching acceptable accuracy of prediction (>60% for a six classes classification problem where the randomness is $\frac{100}{6}$ % (i.e. This is the probability to predict the correct number of polymorphism by random guessing)) except for the Naïve Bayesian Multinomial and the multilayer Perceptron algorithm. k-Nearest Neighbours and Random Forest were the best methods with an accuracy of 86% and 85%, respectively. The exploitation of the Prediction Fusion enabled an improvement of the predictive capacity and demonstrated a synergetic effect of combining the probability generated from the different algorithms in one unique model. The Prediction Fusion method rendered an accuracy of 91% and a Cohen's kappa of 90%.

The confusion matrix that is depicted in Table 1 explains the high accuracy that characterises the Prediction Fusion model. It is clear that the diagonal of this matrix, explaining the correct prediction, is very rich in samples.

Case study 2: Polymorphism prediction with 2D descriptors and without dimensionality reduction

The nine models generated from the datasets of 2D structures using their corresponding molecular descriptors were plotted in Figure 2.B. They showed very similar performance to the Case study 1 and proved that the dimensionality reduction did not improve dramatically the performance of the obtained models in the Case study 1. It is shown that PCA improved the weakest models (*i.e.* Naive Bayesian and Multilayer Perceptron). Their accuracies increased from 9% and 18% to 55% and 52%, respectively. The comparison of the confusion matrices between the two cases shows the enhancement of prediction for class 1. Therefore, in case 2, only 4 samples were predicted wrong, compared to 39 samples for the class of single form. It is noteworthy that the best models from case 1, where Principle components were used instead of the original molecular descriptors, had very similar results. The worst two predictive models were improved dramatically. For instance, the Multilayer Perceptron achieved an accuracy equal to 52% while it was 18% when dimensionality reduction was considered. This proves that the reduction of the number of dimensions did not help to improve the already robust models and even led to a deterioration of relatively weak models such as the Naïve Bayes multinomial algorithm and the Multilayer Perceptron. The confusion matrix, as depicted in Table 2, explained the good performance of the Prediction Fusion model and showed an enrichment of the matrix's diagonal in samples.

Case study 3: Polymorphism prediction with crystallographic descriptors and with dimensionality reduction

Case 3 exploits the information from the dataset of 3D structures. Instead of using the crystallographic descriptors, a Principle Components Analysis was conducted to reduce the dimensionality of the system to 9 components. Comparing to the two previous cases, all

the obtained models in case 3, as illustrated in Figure 2.C, were less robust in predicting the polymorphism of the molecules of interest than the previous models of the 2D structures datasets. K-NN, RF and PF were still very robust to predict the polymorphism. Their accuracies are equal to 84%, 83% and 89%, respectively. The low number of independent variables can explain this robustness compared to the number of samples within the dataset. The corresponding confusion matrix for the Prediction Fusion models that are illustrated in Table 3 explains the high accuracy in estimating the polymorphism from 3D structures and dimensionality reduction.

Case study 4: Polymorphism prediction with crystallographic descriptors and without dimensionality reduction

The last case uses the 3D structures from the intersection of the Drug Bank and the CSD databases and their corresponding crystallographic descriptors such as the unit cell parameters. The performance of the different models was plotted in Figure 2.D. The utilisation of the original crystallographic descriptors improved the performance of all the models slightly, with no exception. Naïve Bayes Multinomial, Simple Logistic and the Multilayer Perceptron were the weakest predictive models. As before, k-NN, RF and PF were at the head of the list to estimate the polymorphism. Support Vector Machine, Ordinal Classic Classifier and the Gradient Boosted Trees had a relatively acceptable accuracy between 60% and 75%. The Prediction Fusion model exploited the probabilities from all the generated models with respect to their individual accuracies. Table 4 explains the robustness of this model through the sample-rich diagonal.

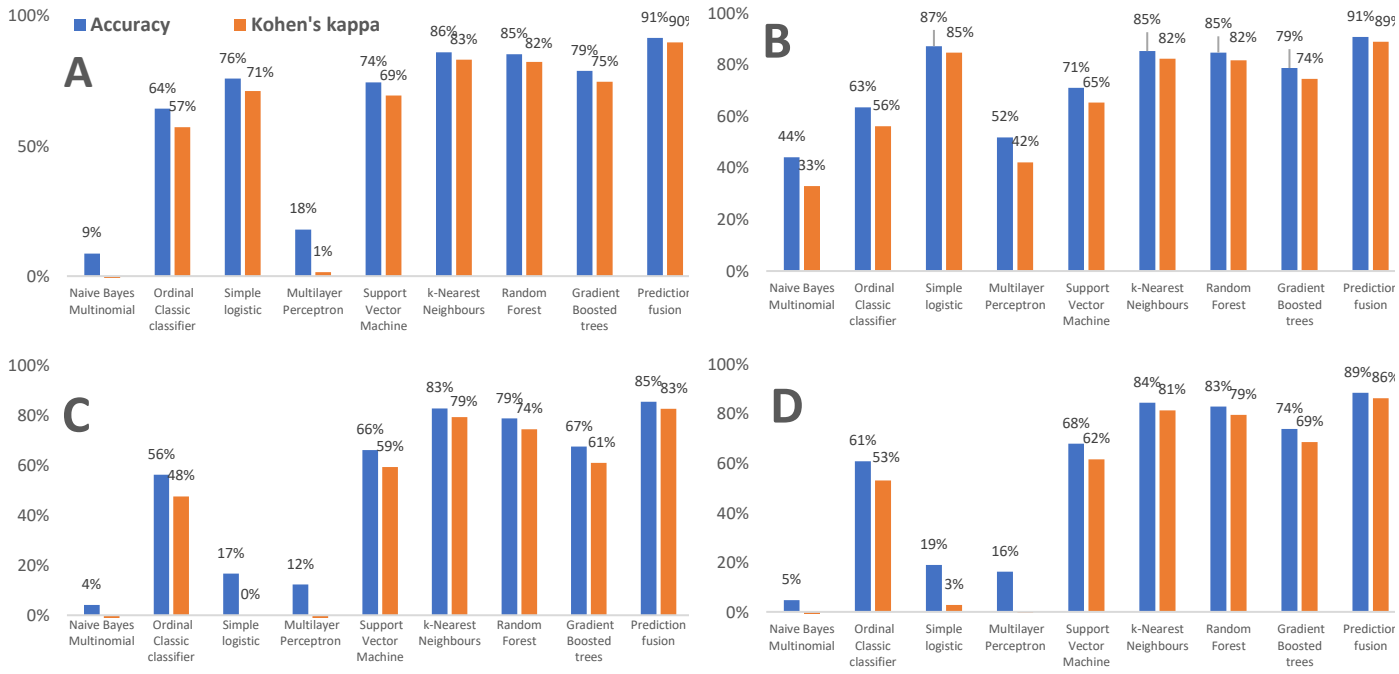

Figure 2 | Performance of the 8 independent machine learning algorithms to generate statistical models in: A- Case study 1, B- Case study 2, C- Case study 3, D- Case study 4.

Descriptors importance

When the Principle Components analysis was not applied for the datasets of 2D and 3D structures (*i.e.* case 2 and 4), it was possible to check the importance of the independent variable and to interpret their contribution to define the accuracy of the designed model. In the case of the 2D structures dataset, molecular descriptors were generated from the MOE and RDKit software packages.

After the pre-processing and filtration step, 169 molecular descriptors were employed to build different models. The backward selection was used in a loop with the k-NN algorithm as assessor of accuracy because it has already demonstrated a good prediction performance and it does not require other predictive models as in the case of the Prediction Fusion. Each descriptor was deleted at each iteration of the loop, and the accuracy was measured. The most important variables were those which significantly deteriorated the accuracy of the model. From the best 2D structure model, the most influential variables on the performance of the predictive models were: Q-VSA-NEG (Total negative van der Waals surface area), Q-VSA-Pol (Total positive van der Waals surface area). These two descriptors belong to partial charge descriptors. There are also the molecular quantum numbers MQN3 (number of chlorines) and MQN26 (number of acyclic single valent nodes), a_ICM (This is the entropy of the element distribution in the molecule). A detailed explanation of all the previous descriptors is included in the manual of MOE and RDKit software[38,39]. In the case of the crystallographic descriptors, the most influencial descriptors were the “a” and “b” parameters of the reduced cell. This was expected because these two parameters define most of the crystal geometry.

In silico Discovery of hidden polymorphism

The ultimate goal of this work was to discover polymorphism in neglected or new molecules. This can be conducted from 2D or 3D structures by applying the suitable model (*i.e.* one of the 4 cases explained above). Finding the real potential of a molecule to give a number of polymorphs has many benefits and can be exploited in

several stages of solid-state research. For instance, this information can be useful for an initial screening from large databases like ZINC[40] and ChEMBL[41]. It is also practical for already investigated polymorphs to see whether other solid forms are missing, and further experimental screening would lead to producing them.

We exploited our predictive models to estimate the polymorphism in a subset (*i.e.* 100 different 2D structures from the CSD that has not been involved in any stage of designing the predictive models). This subset presents the occurrence of polymorphism for molecules that were not included in the Drug Bank like the majority of molecules in the CSD. The comparison of the distribution of the polymorphs in this subset according to their corresponding number of possible forms was summarised in Figure 7. k-NN, Random forest, and Prediction Fusion were selected as they are the most effective predictive models for estimating polymorphism. From the current experimental observations, it was clear that there was a dominance of the structures possessing two different polymorphs. All the different models estimated over 40% of the molecules in the CSD have 2 polymorphs. Currently, 54% of the database structures were classified as possessing just two solid forms. Except for the k-NN, RF and PF models estimated that structures with a single solid form are overestimated. Indeed, RF estimated that only 4% of the structures have a unique crystalline form. Only 3 % of the database was predicted to have 4 different polymorphs. Interestingly, all the statistical models estimated a higher occurrence of structures having 4 forms than the current experimental estimation. For examples, k-NN and PF predicted 7% and 13% of 4-forms occurrence, respectively.

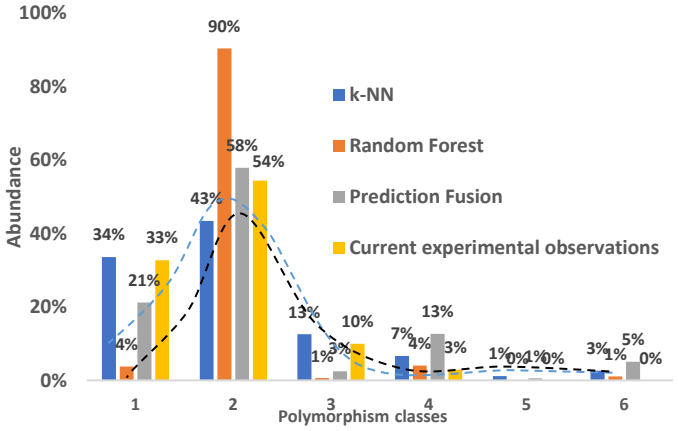

Figure 3 | Comparison of the distribution of the number of polymorphs of molecules in the CSD database if the predictive model of polymorphism is exploited or not. The bar chart presents the occurrence of polymorphism in each of the 6 classes. The blue and the black dashed curves show an approximate trend of this abundance before and after applying the predictive models, respectively.

This overall shape of polymorphism occurrence distribution was kept in the different predictive models. This is illustrated in Figure 3 with black and blue dashed curves for the predicted and the current experimental abundance of polymorphism, respectively. Interestingly, we observed that the predicted area of a high number of polymorphs (*i.e.* 4, 5 or 6 form per 2D structure) is wider than what has been achieved experimentally, thus far. The current statistics of the 100-sample subset showed that there is no structure with six polymorphs. The same trend applies to the structures with five polymorphs. This leads us to conclude that building predictive models based on carefully selected molecules (*i.e.* structures that were heavily screened for polymorphism in the pharmaceutical industry) enabled the discovery of a hidden area of the chemical space.

Experimental screening

The X-Ray powder diffraction patterns of the crystallised samples, depicted in Figure 4.C provided evidence of the presence of new solid forms of pentoxifilline in addition to the already characterised form in the CSD database. The comparison of the Pearson correlation between the patterns identified 4 clusters as depicted in the dendrogram and the clusters plot below. Comparison of the X-Ray diffraction patterns of the samples crystallised in tetrahydrofuran, diethylene glycol, acetic acid and benzylamine shows the presence of additional Bragg reflections, which cannot be explained by the reference pattern. This is indicative of the presence of new solid forms, but the exact nature of these new forms is still not known. The DSC/TGA analysis, represented in Figure 4.D and 4.E, confirms these results by thermal events identified which do not correspond to the crystallisation solvent or the thermal transition of the reference form within the temperature range investigated. ATR-IR spectra were collected as a fingerprint for the new forms and compiled in Figure 4.FMinor differences in the spectra can be rationalised by the presence of different orientation of the pentoxifilline in space, which affects the non-covalent interactions such as the hydrogen bonds.

Two different datasets were extracted from the Drug Bank database and the CSD. They contain 2D and 3D structures or organic molecules and their corresponding polymorphism number. Molecular descriptors were employed as independent variables for 2D structures dataset and crystallographic descriptors were used for 3D structures dataset. PCA was exploited to reduce the dimensions of each of the two datasets, which allow the generation of 4 different datasets in total. 8 different machine learning algorithms were applied to the different dataset, and a metaclassifier was built from the probabilities estimated from each algorithm. 9 statistical models were rendered for each dataset with various capabilities to predict the real number of experimentally achievable polymorphs. It was clear that K-Nearest Neighbours and Random forest were reliably the most robust statistical models. A synergistic effect has also been obtained a metaclassifier called the Prediction Fusion. This latter gave higher accuracy than the RF or the K-NN models. It is also noteworthy to mention that the application of the dimensionality reduction for these systems did not improve the results but slightly deteriorate them.

In addition, the most robust models were exploited to detect the most influential descriptors on the polymorphism capability of each structure. As expected, the reduced unit cell parameters were the most important features in the case of 3D structures approach. A number of molecular descriptors such as the Total negative and positive van der Waals surface area and the number of chlorine atoms in the molecules were among the most influencing molecular descriptors on the models built from 2 structures.

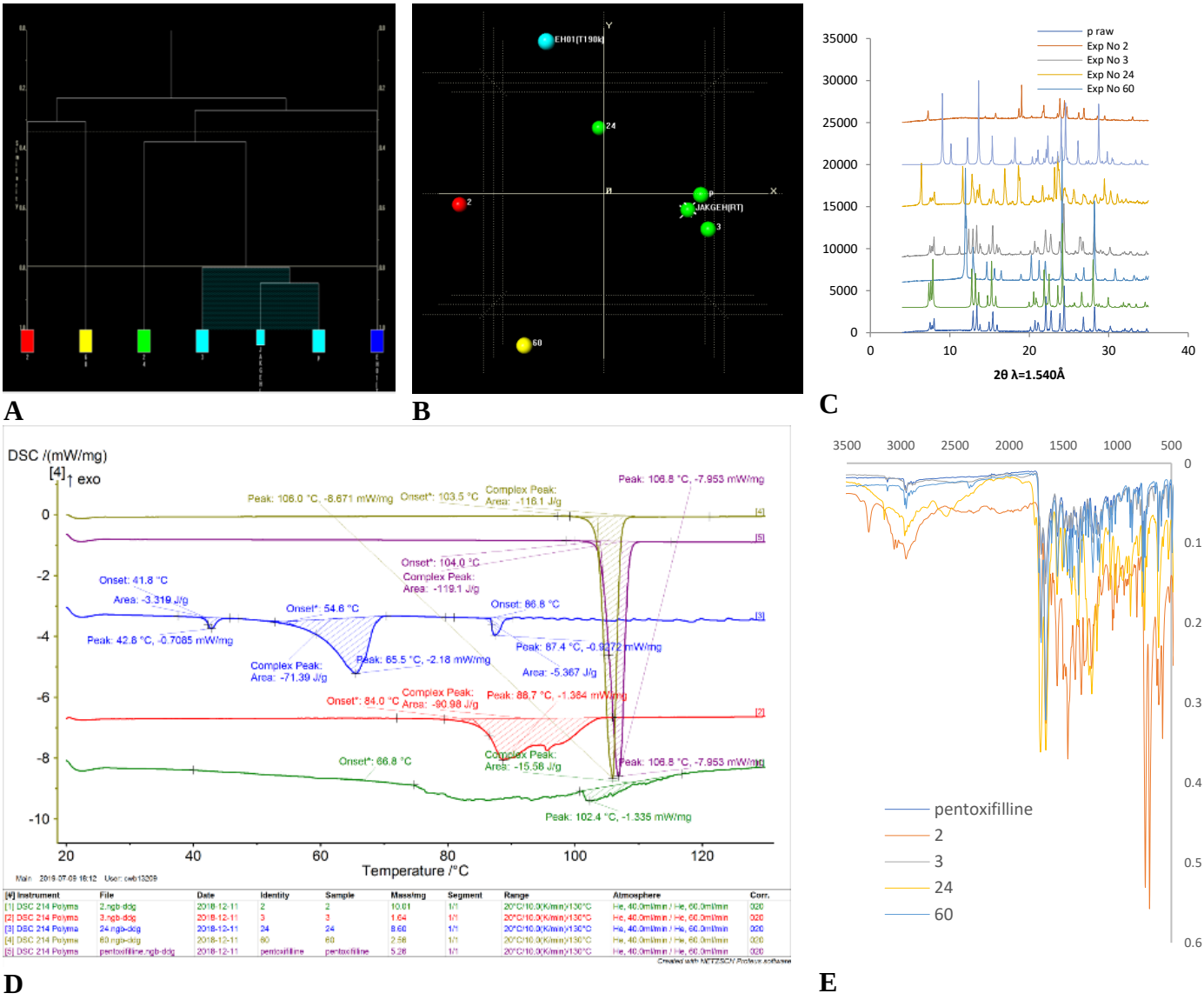

Figure 4 | Analysis of the crystallised pentoxifylline in the different solvents. A- Dendrogram of the most relevant samples screened experimentally and found in the literature, showing the similarity between the solid forms. B- Clustering plot of the selected samples distinguishing between the new form and the mixtures of the existing forms. C- X-Ray powder diffraction patterns of the reference material extracted from the CSD and the selected forms from the experimental screening. D- DSC traces of the selected forms presenting the thermal events occurring during the heating of the samples. E- ATR-IR spectra of the selected new polymorphs

The comparison between the distribution of the abundance of the number of polymorphs in the current CSD with what was predicted from the best-designed models reveals a hidden area of chemical space that was potentially underestimated and under-screened for polymorphism. In other words, these models show the real potential of any known or unknown structure to give a certain number of crystalline forms. In the present work, the most robust model has successfully predicted the number of solid forms that are missing. This was validated experimentally by conducting a solvents screening that revealed the hidden forms. This has paramount practical importance for crystallographers and materials engineers because referring to the best of our knowledge today, this is the first computational tool based on data mining and machine learning that gives experimentalists an initial guideline about the hidden potential of organic molecules to render extra solid forms, not yet discovered and isolated experimentally.

Acknowledgments

The authors would like to acknowledge that this work was carried out in the CMAC National Facility, housed within the University of Strathclyde's Technology and Innovation Centre, and funded with a UKRPIF (UK Research Partnership Institute Fund) capital award, SFC ref. H13054, from the Higher Education Funding Council for England (HEFCE).

Keywords: Machine learning, metaclassification, polymorphism, materials discovery, crystal structure, Artificial Intelligence

References and Notes

1. Lowe, R., Glen, R. C. & Mitchell, J. B. O. Predicting phospholipidosis using machine learning. *Mol. Pharm.* (2010).

doi:10.1021/mp100103e
2. Hughes, L. D., Palmer, D. S., Nigsch, F. & Mitchell, J. B. O. Why Are Some Properties More Difficult To Predict than Others? A Study of QSPR Models of Solubility, Melting Point, and Log P. *J. Chem. Inf. Model.* **48**, 220–232 (2008).
3. Nigsch, F. *et al.* Melting Point Prediction Employing k -Nearest Neighbor Algorithms and Genetic Parameter Optimization. *J. Chem. Inf. Model.* **46**, 2412–2422 (2006).
4. Tropsha, A., Gramatica, P. & Gombar, V. The Importance of Being Earnest: Validation is the Absolute Essential for Successful Application and Interpretation of QSPR Models. *QSAR Comb. Sci.* **22**, 69–77 (2003).
5. Le, T., Epa, V. C., Burden, F. R. & Winkler, D. A. Quantitative Structure–Property Relationship Modeling of Diverse Materials Properties. *Chem. Rev.* **112**, 2889–2919 (2012).
6. Lubinski, L. *et al.* Evaluation criteria for the quality of published experimental data on nanomaterials and their usefulness for QSAR modelling. *SAR QSAR Environ. Res.* **24**, 995–1008 (2013).
7. Yao, S., Shoji, T., Iwamoto, Y. & Kamei, E. Consideration of an activity of the metallocene catalyst by using molecular mechanics, molecular dynamics and QSAR. *Comput. Theor. Polym. Sci.* **9**, 41–46 (1999).
8. Cruz, V. L. *et al.* 3D-QSAR study of ansa-metallocene catalytic behavior in ethylene polymerization. *Polymer (Guildf).* (2007). doi:10.1016/j.polymer.2007.05.081
9. Norbert, W., Durgadas, B., L., B. S. & Joachim, K. Small changes in the polymer structure influence the adsorption behavior of fibrinogen on polymer surfaces: Validation of a new rapid screening technique. *J. Biomed. Mater. Res. Part A* **68A**, 496–503
10. K., T. Makromol. Chem. Macromol. Symp. vol. 69 1993. 4th European Polymer Federation Symposia on Polymeric Materials. Symposium Editors C. Bubeck, H. W Spiess. *Acta Polym.* **45**, 57
11. Greaves, T. L. & Drummond, C. J. Protic Ionic Liquids: Properties and Applications. *Chem. Rev.* **108**, 206–237 (2008).
12. Phillips, C. L. & Voth, G. A. Discovering crystals using shape matching and machine learning. *Soft Matter* **9**, 8552 (2013).
13. Reinhart, W. F., Long, A. W., Howard, M. P., Ferguson, A. L. & Panagiotopoulos, A. Z. Machine learning for autonomous crystal structure identification. *Soft Matter* **13**, 4733–4745 (2017).
14. Oliynyk, A. O. *et al.* Disentangling Structural Confusion through Machine Learning: Structure Prediction and Polymorphism of Equiatomic Ternary Phases ABC. *J. Am. Chem. Soc.* **139**, 17870–17881 (2017).
15. Balachandran, P. V, Broderick, S. R. & Rajan, K. Identifying the {\textquoteleft}inorganic gene{\textquoteright} for high-temperature piezoelectric perovskites through statistical learning. *Proc. R. Soc. London A Math. Phys. Eng. Sci.* **467**, 2271–2290 (2011).
16. Oliynyk, A. O. *et al.* High-Throughput Machine-Learning-Driven Synthesis of Full-Heusler Compounds. *Chem. Mater.* **28**, 7324–7331 (2016).
17. Raccuglia, P. *et al.* Machine-learning-assisted materials discovery using failed experiments. *Nature* **533**, 73–76 (2016).
18. Christian, S. J. & Martin, J. First Step Towards Planning of Syntheses in Solid-State Chemistry: Determination of Promising Structure Candidates by Global Optimization. *Angew. Chemie Int. Ed. English* **35**, 1286–1304
19. Abraham, N. L. & Probert, M. I. J. A periodic genetic algorithm with real-space representation for crystal structure and polymorph prediction. *Phys. Rev. B* **73**, 224104 (2006).
20. Montavon, G. *et al.* Machine learning of molecular electronic properties in chemical compound space. *New J. Phys.* (2013). doi:10.1088/1367-2630/15/9/095003
21. Haleblian, J. & McCrone, W. Pharmaceutical applications of polymorphism. *Journal of Pharmaceutical Sciences* (1969). doi:10.1002/jps.2600580802
22. Nangia, A. Conformational polymorphism in organic crystals. *Acc. Chem. Res.* (2008). doi:10.1021/ar700203k
23. Brittain, H. G. Polymorphism and solvatomorphism 2010. *Journal of Pharmaceutical Sciences* (2012). doi:10.1002/jps.22788
24. Hilfiker, R. *Polymorphism: In the Pharmaceutical Industry*. *Polymorphism: In the Pharmaceutical Industry* (2006). doi:10.1002/3527607889
25. Le, T., Epa, V. C., Burden, F. R. & Winkler, D. A. Quantitative Structure–Property Relationship Modeling of Diverse Materials Properties. *Chem. Rev.* **112**, 2889–2919 (2012).
26. Cabri, W., Ghetti, P., Pozzi, G. & Alpegiani, M. Polymorphisms and patent, market, and legal battles: Cefdinir case study. *Organic Process Research and Development* (2007). doi:10.1021/op0601060
27. Braun, D. E., McMahon, J. A., Koztecki, L. H., Price, S. L. & Reutzel-Edens, S. M. Contrasting Polymorphism of Related Small Molecule Drugs Correlated and Guided by the Computed Crystal Energy Landscape. *Cryst. Growth Des.* **14**, 2056–2072 (2014).
28. Hulme, A. T. *et al.* Search for a Predicted Hydrogen Bonding Motif − A Multidisciplinary Investigation into the Polymorphism of 3-Azabicyclo[3.3.1]nonane-2,4-dione. *J. Am. Chem. Soc.* **129**, 3649–3657 (2007).
29. Vasileiadis, M., Pantelides, C. C. & Adjiman, C. S. Prediction of the crystal structures of axitinib, a polymorphic pharmaceutical molecule. *Chem. Eng. Sci.* **121**, 60–76 (2015).
30. Drebushchak, V. A., Drebushchak, T. N., Chukanov, N. V. & Boldyreva, E. V. Transitions among five polymorphs of chlorpropamide near the melting point. *J. Therm. Anal. Calorim.* (2008). doi:10.1007/s10973-007-8822-0
31. Weng, J. Q., Shen, D. L., Tan, C. X. & Liu, H. J. 2-Oxo-N-phenylthiazolidine-3-carboxamide. *Acta Crystallogr. Sect. E Struct. Reports Online* (2004). doi:10.1107/S1600536804008621
32. Allen, F. H. & Motherwell, W. D. S. Applications of the Cambridge Structural Database in organic chemistry and crystal chemistry. *Acta Crystallogr. Sect. B Struct. Sci.* **58**, 407–422 (2002).
33. Johnston, A., Johnston, B. F., Kennedy, A. R. & Florence, A. J. Targeted crystallisation of novel carbamazepine solvates based on a retrospective Random Forest classification. *CrystEngComm* (2008). doi:10.1039/b713373a
34. Bhardwaj, R. M., Reutzel-Edens, S. M., Johnston, B. F. & Florence, A. J. A random forest model for predicting crystal packing of olanzapine solvates. *CrystEngComm* **20**, 3947–3950 (2018).
35. Bhardwaj, R. M., Johnston, A., Johnston, B. F. & Florence, A. J. A random forest model for predicting the crystallisability of organic molecules. *CrystEngComm* (2015). doi:10.1039/c4ce02403f
36. Wicker, J. G. P. & Cooper, R. I. Will it crystallise? Predicting crystallinity of molecular materials. *CrystEngComm* (2015). doi:10.1039/c4ce01912a
37. Wishart, D. S. *et al.* DrugBank 5.0: A major update to the DrugBank database for 2018. *Nucleic Acids Res.* (2018). doi:10.1093/nar/gkx1037
38. Inc., C. C. G. Molecular Operating Environment (MOE), 2016.08. *1010 Sherbooke St.West, Suite #910, Montreal, QC, Canada, H3A 2R7* (2016).
39. Landrum, G. RDKit: Open-source Cheminformatics. *Http://Www.Rdkit.Org/* (2006). doi:10.2307/3592822
40. Irwin, J. J. & Shoichet, B. K. ZINC - A free database of commercially available compounds for virtual screening. *J. Chem. Inf. Model.* (2005). doi:10.1021/ci049714+
41. Gaulton, A. *et al.* The ChEMBL database in 2017. *Nucleic Acids Res.* (2017). doi:10.1093/nar/gkw1074
42. Allen, F. H. The Cambridge Structural Database: A quarter of a million crystal structures and rising. *Acta Crystallogr. Sect. B Struct. Sci.* (2002). doi:10.1107/S0108768102003890

43. López-Mejías, V., Kampf, J. W. & Matzger, A. J. Nonamorphism in Flufenamic Acid and a New Record for a Polymorphic Compound with Solved Structures. *J. Am. Chem. Soc.* **134**, 9872–9875 (2012).
44. Peterson, M. L. *et al.* Iterative High-Throughput Polymorphism Studies on Acetaminophen and an Experimentally Derived Structure for Form III. *J. Am. Chem. Soc.* **124**, 10958–10959 (2002).
45. P. Mazanetz, M., J. Marmon, R., B. T. Reisser, C. & Morao, I. Drug Discovery Applications for KNIME: An Open Source Data Mining Platform. *Curr. Top. Med. Chem.* **12**, 1965–1979 (2012).
46. Jagla, B., Wiswedel, B. & Coppée, J.-Y. Extending KNIME for next-generation sequencing data analysis. *Bioinformatics* **27**, 2907–2909 (2011).
47. Beisken, S. *et al.* KNIME-CDK: Workflow-driven cheminformatics. *BMC Bioinformatics* (2013). doi:10.1186/1471-2105-14-257
48. ChemicalComputingGroupInc. Molecular Operating Environment (MOE). *Sci. Comput. Instrum.* (2004). doi:10.1017/CBO9781107415324.004
49. Chawla, N. V., Bowyer, K. W., Hall, L. O. & Kegelmeyer, W. P. SMOTE: Synthetic minority over-sampling technique. *J. Artif. Intell. Res.* (2002). doi:10.1613/jair.953
50. McCallum, A. & Nigam, K. A Comparison of Event Models for Naive Bayes Text Classification. *AAAI/ICML-98 Work. Learn. Text Categ.* (1998). doi:10.1.1.46.1529
51. Schapire, R. E., Stone, P., McAllester, D., Littman, M. L. & Csirik, J. A. Modeling auction price uncertainty using boosting-based conditional density estimation. *Mach. Learn. Work. Then Conf.* (2002).
52. Landwehr, N., Hall, M. & Frank, E. Logistic model trees. *Mach. Learn.* (2005). doi:10.1007/s10994-005-0466-3
53. Pal, S. K. & Mitra, S. Multilayer Perceptron, Fuzzy Sets, and Classification. *IEEE Trans. Neural Networks* (1992). doi:10.1109/72.159058
54. Ben-Hur, A. & Weston, J. A user's guide to support vector machines. *Methods Mol. Biol.* (2010). doi:10.1007/978-1-60327-241-4_13
55. Aha, D. W., Kibler, D. & Albert, M. K. Instance-Based Learning Algorithms. *Mach. Learn.* (1991). doi:10.1023/A:1022689900470
56. Breiman, L. Random forests. *Mach. Learn.* (2001). doi:10.1023/A:1010933404324
57. Si, S. *et al.* Gradient Boosted Decision Trees for High Dimensional Sparse Output. *Icml* (2017).
58. Hall, M. *et al.* The WEKA data mining software. *SIGKDD Explor. Newsl.* (2009). doi:10.1145/1656274.1656278
59. Song, L., Smola, A., Gretton, A., Bedo, J. & Borgwardt, K. Feature selection via dependence maximization. *J. Mach. Learn. Res.* (2012). doi:10.1145/1273496.1273600

Methods

Database curation

The datasets generated to build predictive models for the polymorphism of organic molecules were curated from two main databases: The Drug Bank [37] (DB) and the Cambridge Structural Database[42] (CSD). The choice of these two databases was based on the complementarity of information that they provide. In the CSD, which is approaching one million crystal structure entries, each molecule entry may have metadata field describing its polymorphism and whether it is available in the DB with the corresponding reference. The DB currently has 9292 entries comprising 3189 small molecule drugs, 926 approved biotech (protein/peptide) drugs, 108 nutraceuticals and over 5,069 experimental drugs. To evaluate the role of ML in predicting polymorphism in the drug like molecules, a subset database was created. Filters were applied to the full CSD database, to identify entries that were organic molecules with recorded polymorphs (*i.e.* >1 distinct crystal structure with the same formula unit) and which also appeared within the DB. The choice of the intersection between the two databases, depicted in Figure 5, was based on the increased likelihood that commercially available drugs will have been screened experimentally for polymorphism and more likely to have a complete record of the number of experimentally achievable crystallographic forms compared with molecules investigated in academia or other industries. While it is still probable that all forms may not have been reported in the public domain or even structurally characterised to allow them to be included in the CSD, this database still provides a useful training set to assess the potential application of ML. From the initial CSD dataset of 15202 3D structures, only 883 structures remained after the filtration process. These three-dimensional structures can be further categorised into 178 two-dimensional structures (smiles). The 15202 3D structures for the 883 molecules also include redeterminations, variable temperature studies, conformational polymorphs, polymorphic forms, co-crystals, salts, hydrates and solvates. Therefore, a manual filter was used to discard co-crystals and redeterminations of each molecule. Although they have chemically different compositions, the salts, hydrates and solvates were kept in our datasets because they have very relevant information that determines the diversity of the polymorphic forms, especially for pharmaceutical applications.

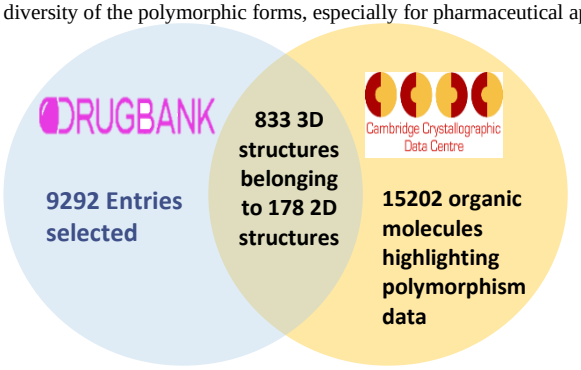

Figure 5 | The intersection between two different databases: CSD and Drug Bank giving drug like molecules that have likely been screened for polymorphism with entries in the CSD.

Confirmation of polymorphism was provided by checking the original published source article for each selected REFCODE entry. Although some papers claim the existence of molecules with up to 9 different polymorphs such as the flufenamic acid[43], the datasets ignored these molecules as they are a very extreme minority that risks unbalancing the datasets heavily. In the other extreme of the polymorphism spectrum, only molecules that are commercially available as drugs were considered as "true negative" as any drug is expected to be screened for polymorphism before it becomes available in the market. There are several works that claim the presence of 60 different solid forms for Atorvastatin but surprisingly none was reported in the CSD because no single crystal was identified and isolated.[44] The final dataset used in this study considered six classes that describe chemicals in the database with 1, 2, 3, 4, 5 or 6 discrete, crystallographically different polymorphs including solvates and salts but not co-crystals.

Data workflow

The pre-processing of the 833 3D structures dataset "condensed" in 178 2D structures dataset was carried out through a series of data transformation and cleaning available in Knime[3] software and described in Figure 6[45–47]. The initial data transformation was the generation of molecular descriptors for the 2D structures dataset. This step was not performed in the case of the 3D structures dataset because crystallographic descriptors will be used instead of most of the molecular descriptors generated with MOE[48] and RDKit[47] nodes available in Knime. In a second step, structures having descriptors with missing or highly correlated values (*i.e.* Correlation threshold was set to 0.9 as a lower limit of correlation) were also eliminated from both datasets. The descriptors with low variance (*i.e.* Variance upper bound was set to 0.01) were discarded, too. A normalisation between 0 and 1 was additionally applied to unify the scale employed for the different descriptors. The order of samples was randomly chosen via the "Shuffle" node. The Synthetic Minority Over-sampling algorithm to balance the classes of each dataset was used[49]. This step was crucial to get balanced classes in each dataset. Figure 6 illustrates in a pie chart the occurrence of the different classes (*i.e.* Number of polymorphs) before the application of the balancing process through the SMOTE algorithm. This technique oversamples only the minority classes and takes into account the 5 nearest neighbours. In the final pre-processing stage, the Principle Components Analysis was applied to reduce the dimensionality of the dataset especially when hundreds of descriptors were included. Therefore, starting from the two datasets (the 2D structures list and the 3D structures list), two new datasets were generated where the principal components replaced the original molecular or crystallographic descriptors.

A number of principal components were used by keeping 95%, and 90% of the variance were preserved for the 2D and the 3D structures datasets, respectively. These percentages were chosen in the way of keeping the maximum of the variance in the Data and generating in the same time a reasonable number of principal components for the case 2 and 4 that will be described later. At the end of the pre-processing phase, four different datasets were compiled and ready to be trained over machine learning algorithms (*i.e.* the two original datasets and the two obtained with PCA).

Crystallographic and molecular descriptors

The four datasets have different number and type of descriptors as depicted in Table 1. In the first dataset using the 2D structures, only molecular descriptors were included. These descriptors highlight the properties of the whole molecule, and they can be categorised into constitutional, geometrical, topological and electronic descriptors. For instance, one cites the molecular mass, the number of carbon atoms, the polar surface size, the charge, the count of halogen or hydrogen atoms in the molecule of interest. etc. In the second dataset, these descriptors were replaced by the principal components generated from the dimensionality reduction. It is noteworthy that these descriptors have no chemical meaning but were defined to reduce the dimensionality of the feature space and ease the task of machine learning algorithms. In the third dataset, crystallographic descriptors were employed including the unit cell parameters and the volume of the cell. etc. Finally, the fourth dataset is the dimensionally reduced dataset obtained from the application of PCA on the crystallographic descriptors (*i.e.* the third dataset). The table below summarises the number and the type of descriptors used for each dataset.

[3] www.knime.com

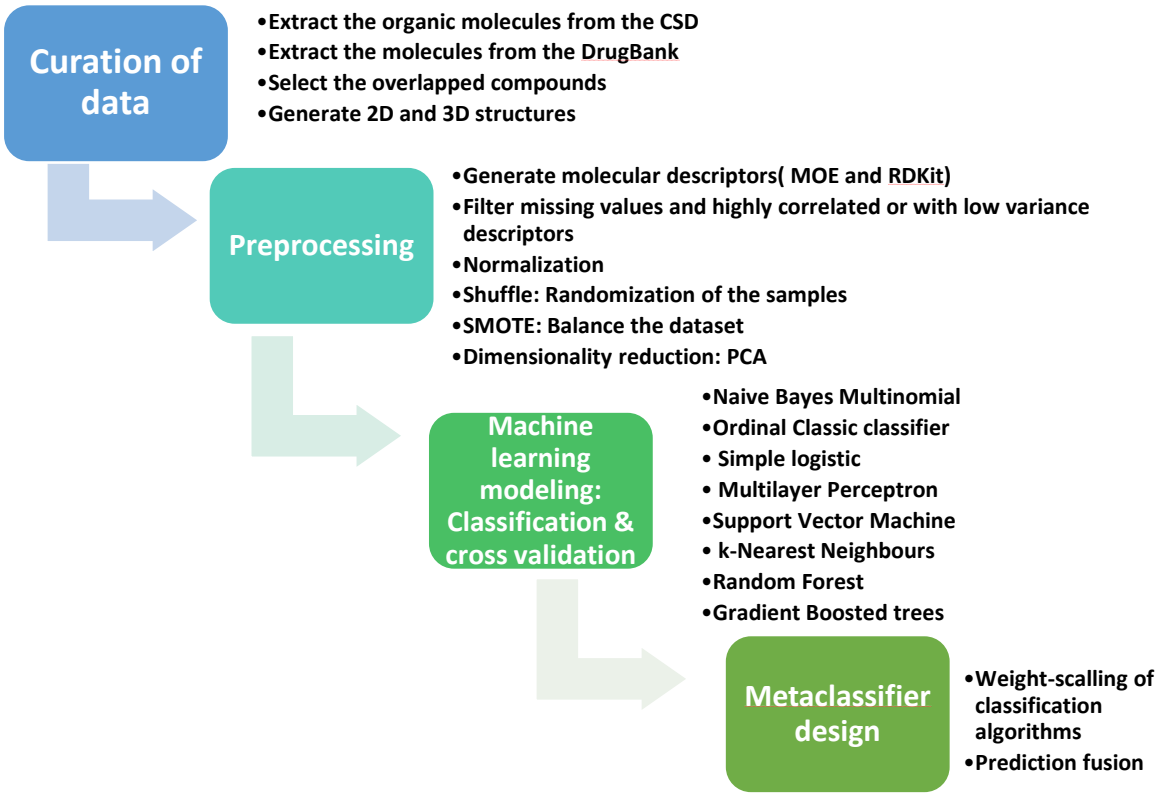

Figure 6 | Data flow and the different steps followed from the data collection, passing through the pre-processing and generation of machine learning models and ending with a meta classifier design to predict the polymorphism of unknown structures.

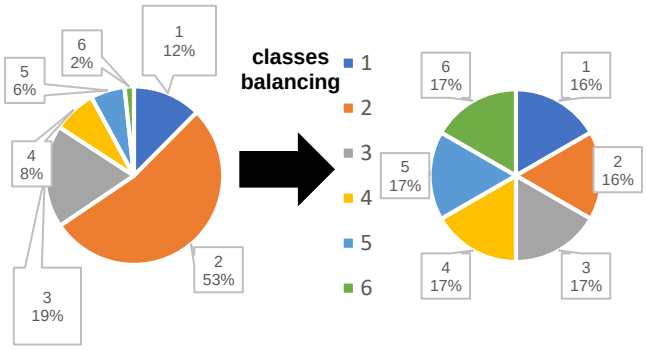

Figure 7 | The occurrence of polymorphism in the datasets and the role of the SMOTE algorithm to balance the classes.

Choice of machine learning classifiers

8 different algorithms were selected from Knime nodes to train the models over the 4 datasets: Naive Bayes Multinomial[50], Ordinal Classic classifier[51], Simple logistic[52], Multilayer Perceptron[53], Support Vector Machine[54], k-Nearest Neighbours[55], Gradient Boosted Trees and Random Forests. Figure 8 demonstrates how the curated data were classified into 2D and 3D structures and how these two datasets were treated differently by applying the original features (*i.e.* molecular descriptors or crystallographic parameters). 4 datasets were finally obtained according to the pre-processing method. Random Forest,[56] Gradient Boosted trees.[57] All these classifiers were implemented in Weka nodes available within Knime software.[58] All these algorithms were used with their default settings except the followings. In SVM, the "Hyper Tangent" was chosen as a kernel with a kappa equals to 0.1 and delta equals to 0.5. In the K-Nearest Neighbours classifier, the number of neighbours was chosen to be 6 as the number of classes in the datasets. In Random Forests classier, the number of trees was adjusted to 500. 10-fold cross-validation was applied for each classifier. The accuracy and Cohen's kappa were used as metrics to evaluate the performance of the models. Also, Confusion matrices, Recall, Precision, sensitivity, specificity and F-measure were provided for each class of each classifier.

Meta-classifier design

Once the eight classifiers were trained over the four datasets, probabilities of classification (*i.e.* the probability for each of the 6 classes of the response) were generated. The "Prediction fusion" node was employed to combine the probabilities from the different classifiers and weigh them according to the robustness of the obtained model. These weights W_i correspond to the accuracy obtained from each ML model. This is translated by the formula below:

$$P_{fusion} = \frac{\sum_{i=1}^{k=Mi} P_i \times W_i}{8} \quad (1)$$

where P_{fusion} is the overall probability of that particular class, M_i is the model among the eight selected model, P_i is the individual probability per model, and W_i is the scaling factor that is equal to the accuracy of the model M_i. Like for individual classifiers, the same metrics were employed to characterise the overall model from the prediction fusion.

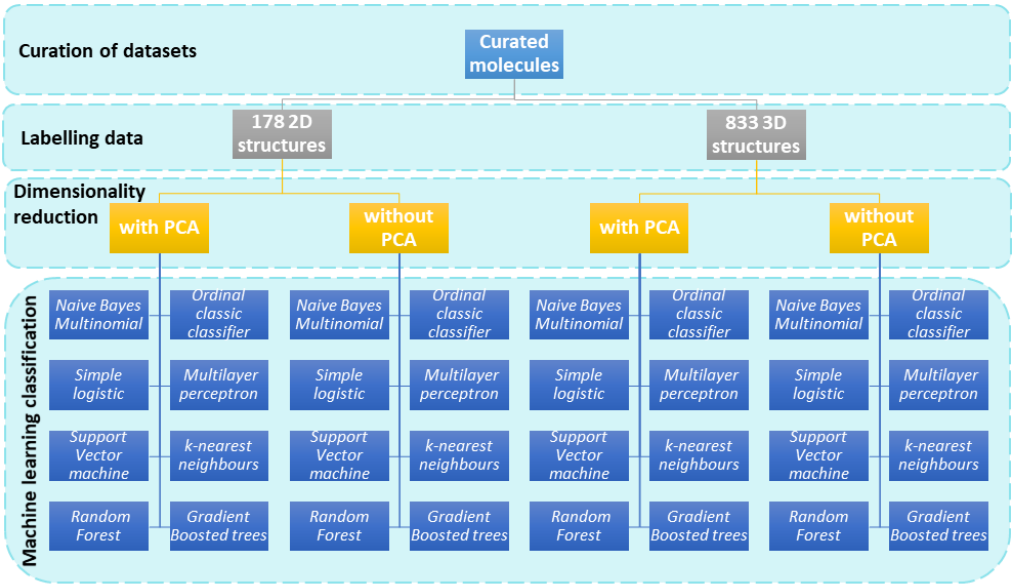

Figure 8 | The partition of the curated datasets into 4 categories where 2D and 3D structures are generated, then 8 different machine learning algorithms were applied to generate statistical models from molecular/crystallographic descriptors or from principal components of PCA.

Table 5. Number and nature of the descriptors used as independent variables to build the predictive model

	2D structures		3D structures	
	Case 1 [a]	Case 2 [b]	Case 3 [c]	Case 4 [d]
Number of dimensions	44	169	9	12
Number samples	564	564	2352	2352

[a] 2D structures treated without PCA , [b] 2D structures traded with PCA, [c] 3D structures treated without PCA (crystallographic descriptors) , [d] 3D structures treated with PCA

Features selection

The backward features selection[59] was carried out in the last stage of the data analysis and the machine learning model design. The classifier that renders the best predictive model was incorporated in the loop of features selection. This technique eliminates each descriptor consecutively and builds a predictive model with the classifier of choice (*i.e.* One feature was removed in each iteration of the elimination loop until all features are eliminated at the end of the loop). In each iteration, all the features are tested by discarding them once then, the feature that has the highest impact on the accuracy is deleted for the next iteration. The models are ranked according to their accuracy, and the most important features are the ones that affect the most the robustness and the accuracy of the model. Therefore, the most important feature is the one that gives the lowest accuracy for the model when it has been eliminated before the training step.

Solvent screening

Pentoxifylline (CAS Number 6493-05-6) was purchased from Sigma Aldrich. Samples of 100mg each were recrystallised in 66 different solvents listed in the supporting information with their available physical properties. Each solution was heated near the boiling point of the corresponding solvent to ensure the full dissolution of the starting material. Once clear solutions are obtained, they were left to cool to ambient temperature. They were then kept in atmospheric conditions to evaporate the solvents. The pure polymorphic forms and the percentage compositions of all the sample mixtures were characterized by differential scanning calorimetry–thermogravimetry (DSC-TG), XRPD, and ATR-IR spectroscopy.

Powder X-Ray Diffraction

For crystalline form identification, a small quantity (10-50 mg) of the sample was analysed using transmission XRPD data collected on a Bruker AXS D8 Advance transmission diffractometer equipped with θ/θ geometry, with primary monochromated radiation (Cu $K_{\alpha 1}$ λ= 1.54056 Å), a Vantec PSD and an automated multiposition x-y sample stage. Samples were mounted on a 28-position sample plate supported on a polyimide (Kapton, 7.5 µm thickness) film. Data were collected from each sample in the range 4-35° 2θ with a 0.015°2θ step size and 1 s per step count time. Samples were oscillated in the x-y plane at a speed of 0.3 mm s^{-1} throughout data collection to maximise particle sampling and minimise preferred orientation effects. **Thermogravimetry analysis**

Netzsch STA 449 F1 Jupiter® performed DSC and TGA simultaneously allowing the monitoring of remaining solvent evaporation.

Differential scanning calorimeter analysis

Differential scanning calorimetry–thermogravimetric experiments were performed on a Netzsch DSC214 Polyma differential scanning calorimeter. The heating rate for all polymorphs was kept constant at 20°C/min and all runs were carried out from 25 °C to 250 °C. The measurements were performed in aluminium crucibles, nitrogen was used as the purge gas in ambient mode, and calibration was performed using indium metal. The cooling of the samples was conducted for all the samples after a temperature plateau at 250 °C.

Attenuated Total Reflectance–Infrared Spectroscopy

Attenuated total reflectance–infrared spectra were collected on a Bruker TENSOR II FT-IR spectrometer with Opus v7.5 software. The spectrometer is fitted with a KBr beamsplitter, which operates in the range 8000–10 cm^{-1} with a universal ATR accessory ('PLATINUM' diamond ATR-accessory), and HYPERION (IR microscope). Spectra were collected in the 4000–650 cm^{-1} range with a resolution of 4.00 cm^{-1} and scan number of 4.

GRAPHICAL ABSTRACT

Polymorphism is a physical feature that characterises solid crystalline compounds Its regulation is crucial for the control of other properties such as the solubility or mechanical resistance. Machine learning is a modern statistical tool exploited to learn from the existent data and to generate models that predict how the molecule is able to give a number of solid forms. This computational tool provides a guideline to experimentalists in order to spot molecules with potentially underestimated polymorphism and to ease the discovery of novel materials.

GRAPHICAL ABSTRACT FIGURE