Crystallisation Thermodynamics and Random Forest Classification for the Prediction of Crystallisation Outcomes

A thesis presented in fulfilment of the requirements for the degree of Doctor of Philosophy in the Faculty of Sciences of the University of Strathclyde

By

Siya Nakapraves

Strathclyde Institute of Pharmacy and Biomedical Sciences

Feb 2023

Declaration of Author's Right

This thesis is the result of the author's original research. It has been composed by the author and has not been previously submitted for examination, which has led to the award of a degree. The copyright of this thesis belongs to the author under the terms of the United Kingdom Copyright Acts as qualified by the University of Strathclyde Regulation 3.50. Due acknowledgement must always be made of the use of any material contained in, or derived from, this thesis

Signed:

Date:

Acknowledgements

I would like to express my deepest appreciation to many people who generously help me along the way with my PhD. Without their support, I could not have undertaken this long journey.

Firstly, I am extremely grateful to my supervisor, Prof. Alastair Florence, for all his invaluable advice, continuous support, and patience during my PhD study. His immense knowledge and plentiful experience have guided me in my research and the process of thesis writing. Also, many thanks for giving me the opportunity to complete my PhD at CMAC.

I would also like to extend my sincere gratitude to Dr. Monika Warzecha, Dr. Chantal Mustoe, and Dr. Vijay Srirambhatla for providing guidance and feedback throughout the research in this PhD. Thanks for teaching me how to start my PhD, how to design the experiments, how to write the thesis, and for all the supportive and encouraging comments you always give me. This thesis would not have been possible without the support from all of you.

Additionally, I would like to acknowledge the help that I received from my colleagues from CMAC, especially Dr. Alan Martin, Dr. Cameron Brown, Dr. Murray Robertson, and Dr. Antony Vassileiou who gave me great advice during this work.

Thanks should also go to my mom and dad, my sister and my brother for being my mental supporters. I am also thankful to my friends and my boyfriend for all the entertainment, emotional support, and lovely party with wonderful meals and karaoke when we were together. Without their tremendous understanding and encouragement in the past few years, it would be impossible for me to complete my PhD.

Lastly, I would be remiss in not mentioning the Government Pharmaceutical Organization (Thailand) for the studentship that allowed me to pursue this PhD.

Abstract

Crystallisation is one of the key unit operations in the pharmaceutical industry. A wide range of crystal attributes affects the bulk particle properties of a crystalline material as well as its downstream manufacturability. Therefore, understanding and controlling the crystallisation process to achieve the desired quality attributes are of significant interest. This thesis investigated the potential of machine learning techniques in terms of the prediction of crystallisation outcomes, focusing on the shapes of mefenamic acid (MFA) crystals from various organic solvents, and solvated structures of small organic molecules considered by Powder X-ray Diffraction (PXRD) patterns. The solubility and nucleation of MFA were also explored in this thesis in an attempt to understand the thermodynamic and kinetic interactions during the crystallisation process of MFA. It was observed that the nucleation of MFA in methanol, ethanol, 2-propanol, 2-butanol, acetone, and tetrahydrofuran (THF) follows a two-step mechanism, in which the crystals nucleate within the metastable clusters. The comparison between surface free energy determined from nucleation rates and that calculated by Turnbull's rule also proposes that the crystals nucleated faster via two-step nucleation compared to classical nucleation theory (CNT), due to the smaller nucleation barrier.

For the machine learning application for predicting the crystallisation outcomes, the result showed that random forest classification models using solvent physical property descriptors can reliably predict crystal morphologies for MFA crystals grown in 20 out of the 28 solvents included in this work. Further characterization of the crystals grown in the remaining 8 solvents with poor model performance also resulted in the discovery of a new THF solvated form of MFA crystals. The ability of machine learning was also investigated to predict the solvated form of small organic molecules from the PXRD patterns derived from Cambridge Structural Database (CSD). The best model in this study showed 68.74% of prediction accuracy. These findings demonstrate the potential role of machine learning and data mining to assist

the decision-making in crystallisation while reducing the uses of materials and time spent during the process development.

Table of Contents

Page

List of Figures

Page

List of Tables

Page

List of abbreviations

API	–	Active Pharmaceutical Ingredient
BCS	–	Biopharmaceutical Classification System
CCDC	–	Cambridge Crystallographic Data Centre
CMAC	–	Continuous Manufacturing and Advanced Crystallisation
CNT	–	Classical Nucleation Theory
CSD	–	The Cambridge Structural Database
DCM	–	Dichloromethane
DMF	–	Dimethylformamide
DSC	–	Differential Scanning Calorimetry
IPA	–	Isopropyl Alcohol
MFA	–	Mefenamic Acid
MIP	–	Metastable Intermediate Phase
MOE	–	Molecular Operating Environment
MSZ	–	Metastable Zone
MSZW	–	Metastable Zone Width
NCE	–	New Chemical Entity
PL	–	Prior Likelihood
PTFE	–	Polytetrafluoroethylene
PXRD	–	Powder X-Ray Diffraction
RF	–	Random Forest
RPM	–	Round Per Minute
SD	–	Standard Deviation
SC-XRD	–	Single-Crystal X-Ray Diffraction
THF	–	Tetrahydrofuran

List of Symbols

A	–	Kinetic parameter
B	–	Thermodynamic parameter
C	–	Concentration
C_0	–	The concentration of nucleation sites
C_e	–	Concentration at Equilibrium
d_{hkl}	–	Interplanar spacing between hkl plane
f^*	–	Attachment frequency
H_{cryst}	–	Enthalpy of crystal
H_{soln}	–	Enthalpy of solution
J	–	Nucleation Rate
k_B	–	Boltzmann's constant
K_{eq}	–	Equilibrium constant
M	–	The total number of experiments
M^+	–	The number of experiments that crystals are detected
m	–	The number of nuclei
N	–	Average number of nuclei forming in a time interval
N_A	–	Avogadro’s number
P_m	–	Probability that m nuclei are formed in a time interval
$P(t)$	–	The probability of detecting crystal at time t
R	–	Universal gas constant
S	–	Degree of Supersaturation
S_{cryst}	–	Entropy of crystal
S_{soln}	–	Entropy of solution
T	–	Temperature
t	–	Induction time
t_g	–	Growth time

t_j	–	The time when the first nuclei were detected
V	–	Solution volume
W*	–	Nucleation work
z	–	Zeldovich factor
ΔH^0_{cryst}	–	Crystallisation Enthalpy
ΔG^0_{cryst}	–	Gibbs' Free Energy
ΔS^0_{cryst}	–	Crystallisation Entropy
γ	–	Surface free energy
Ω	–	Molecular volume

1. Introduction

1 Introduction

1.1 Crystallisation in the Pharmaceutical industry

Crystallisation is used in many applications ranging from purification[1–3] and separation[4–6] to the production of chemicals or active pharmaceutical ingredients [7–9]. Organic molecules can adopt a range of solid forms that have distinct crystal structures that affect their physical properties.[10–13] The corresponding medicinal products produced using different polymorphs can have variable safety and efficacy due to changes in key biopharmaceutical attributes such as solubility and stability.[14] Ritonavir polymorphism is an example of problematic solid form diversity where the more stable polymorph (form-II) with lower solubility was found in the metastable form well after the product had been released to the market. The presence of the more stable polymorph in the consumer product caused slower dissolution and decreased the bioavailability of the formulation. Consequently, ritonavir in an oral capsule formulation was withdrawn from the market in 1998 and reformulated.[15]

Crystallisation has also been used to control crystal size and morphology. These are key attributes impacting the end-product quality, functionality, and downstream manufacturability of a drug candidate.[16–18] For example, crystal morphology influences how easy it is to filter L-glutamic acid. This change in filterability results from the fact that spherulitic crystals (a form of polycrystalline particles with a spheroidal shape, **Figure 1**) of L-glutamic acid have higher cake resistance values than the needle and polyhedral crystals.[19]

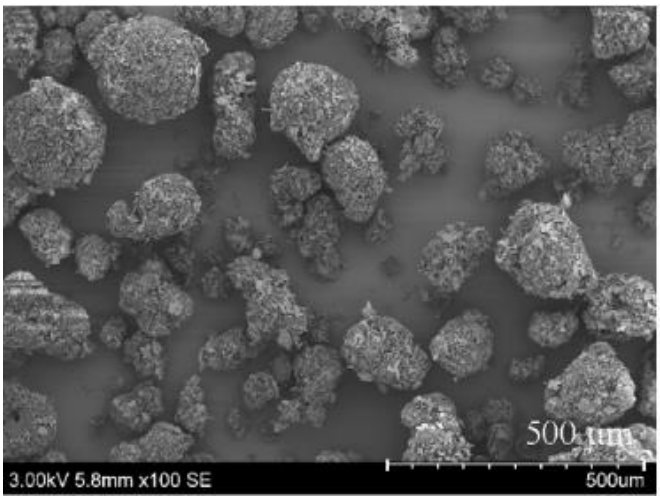


Figure 1. *Spherulitic crystals of L-glutamic acid (figure from reference)* [19]

Previously, the quality of the product was only tested at the end of the manufacturing process (quality-by-testing; QbT). This approach often led to unsatisfactory results as faults are only found after they occur and the analytical testing and sampling lead to additional costs and time delays during pharmaceutical manufacturing.[20] To achieve the desired product quality, several advanced design strategies for controlling crystal size, purity, morphology, and polymorphic form during the crystallisation process have been developed over the last few decades.[21] For example, advanced online sensors in PAT (process analytical technology) have been used to probe real-time crystallisation. Such online sensors provide feedback that can enable control over concentration to allow selectively crystallisation of the desired polymorph[22] in addition to solubility and supersaturation measurements using *in situ* ATR-FTIR (Attenuated Total Reflection – Fourier Transform Infrared Spectroscopy)[23]. Although the application of these control techniques can increase the productivity of the process and/or the quality of the products, controlling crystallisation is still challenging due to the non-linear crystallisation dynamics and high variations in crystal nucleation and growth processes that occur impacting various aspects of the product particle and bulk properties.[24]

To assess the impact of crystallisation solvent and process variables on crystal attributes (purity, polymorphic form, as well as crystal habit), crystallisation screening is used as a preliminary step of the preclinical assessment and solid form selection.[25] In crystallisation screening, experimental variables of interest and associated variable ranges are selected before designing the experiment.[26] The effects of the main process parameters including solvent, supersaturation, cooling rate, and agitation are studied. For instance, the systematic workflow of seeded cooling crystallisation developed by CMAC Future Manufacturing Hub (**Figure 2**) consists of various steps including solvent screening and selection (Stage 2 and 3, **Figure 2**) and process parameters were also studied during Stages 5 and 6 (**Figure 2**). The experiments in this workflow give comprehensive information on the effect of solvents and crystallisation conditions on the observed outcome. The solvent and process parameters which result in the desired crystal attributes can then be selected for further investigation, process development and scale-up.[9] This decision-driven approach provides the first step into the predictive design of crystallisation processes.[9]

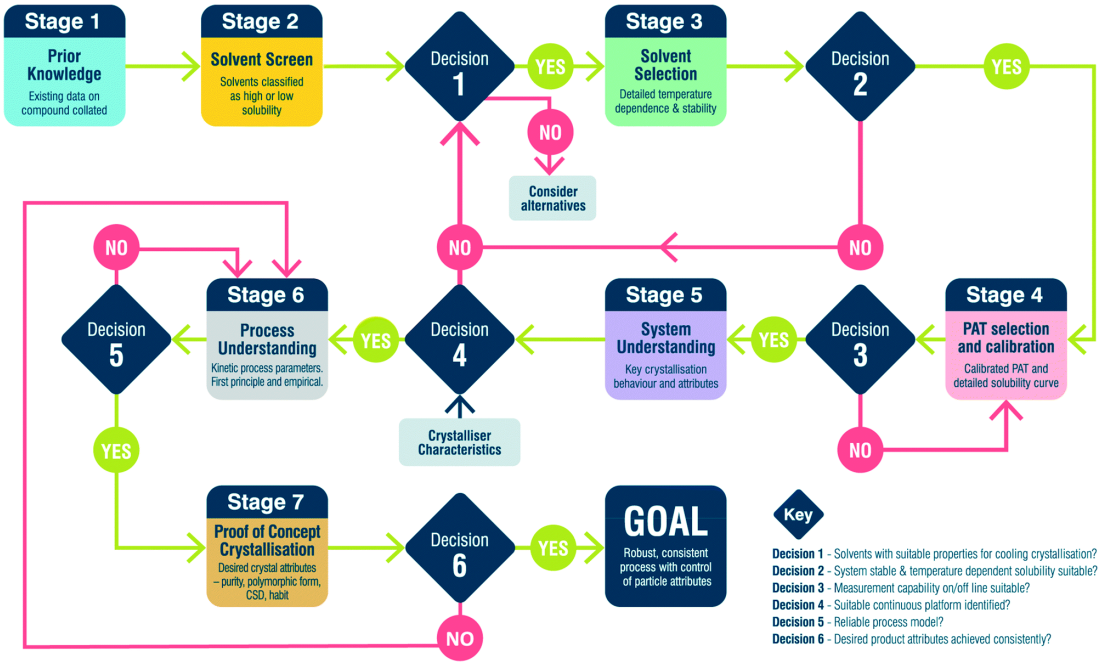


***Figure 2**. The CMAC workflow for seeded cooling crystallisation (figure from reference)* [9]

Advancements in informatics and computational science techniques such as machine learning have shown promise to complement the experimental screening step and introduce useful predictive capability. Using machine learning to predict experimental outcomes can enable informed selection of solvents and minimizes the experiments, cost and time needed in the material and process optimizations.[27] Some examples of machine learning applications in crystallisation include crystal structure prediction (CSP),[28] crystal packing prediction,[29] and the prediction of different crystal outcomes (e.g., solvates and non-solvates,[30,31] different polymorphic forms,[31] or crystalline and non-crystalline[32]). These studies have shown the potential applications of machine learning in the field of crystallisation process design.

1.2 Thermodynamics and kinetics of crystallisation

In crystallisation, both the prevailing thermodynamics of crystallization and the process kinetics influence the observed outcomes. Crystal nucleation and growth rates are affected by the crystallisation driving force or supersaturation, hence determining the yield and polymorphic form, shape, and size of the resultant crystals.[33]

1.2.1 Crystal nucleation

Nucleation is a process where nuclei form from a supersaturated medium. Crystal nucleation is a complex phenomenon that occurs via several different mechanisms. Conventionally, nucleation can occur by primary nucleation, where the nucleation happens spontaneously

from a supersaturated solution, or secondary nucleation, where the presence of crystal seeds and the interaction between the crystalline surfaces and the surrounding environment is required for nucleation to occur.[34]

Figure 3 shows the difference between primary nucleation and secondary nucleation mechanisms.

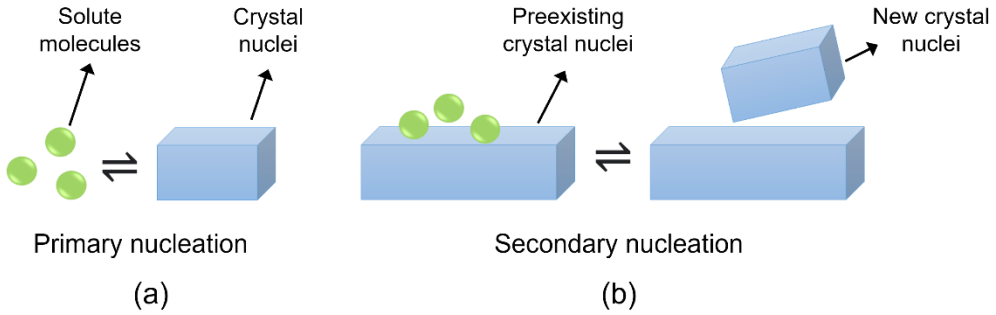


Figure 3. *Different mechanisms of nucleation processes. (a) the monomers of solute molecules spontaneously aggregate into the crystal nuclei during primary nucleation process. (b) the monomers of solute molecules nucleate on the surface of preexisting crystal nuclei in the solution during secondary nucleation process.*[37]

Primary nucleation can be divided into two sub-categories of mechanism: primary homogeneous nucleation and primary heterogeneous nucleation. Homogeneous nucleation occurs in the absence of crystalline surfaces and crystal nuclei precursors. By contrast, heterogeneous nucleation occurs at preferential locations, such as the surfaces of a container or an interphase boundary in liquid or solid.[38] Heterogeneous nucleation can also be induced by the particles of foreign substances, such as dust, impurities, or residues from previous material. Generally, nucleation in an industrial crystallisation is mostly heterogeneous because nucleation can be induced by foreign particles in working stations and it is practically not feasible to remove all particulate contaminants.[34] Heterogeneous nucleation also requires a relatively low free energy barrier to nucleation and, thus, happens more readily at low supersaturation than homogeneous nucleation.[39,40] For the mechanisms of primary nucleation, two main models have been proposed, namely the Classical Nucleation Theory (CNT) and the two-step nucleation theory (see **Figure 4**).

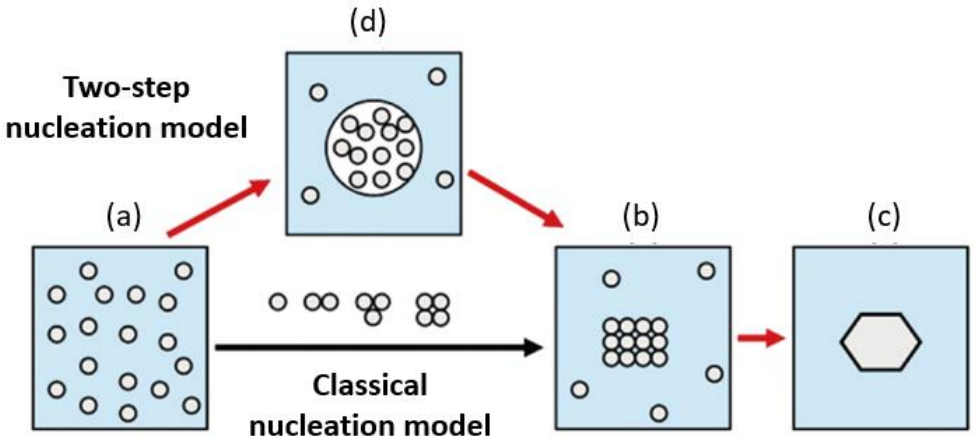


Figure 4. Mechanisms of crystal nucleation from a supersaturated solution. In CNT, monomer associations (a) are formed into a nucleus in a shape with the minimum free energy (b) before gathering into a macro-crystal with distinct facets (c). The two-step nucleation model suggests that the molecules congregate into a disordered precursor (d) before forming a nucleus (figure from reference).[41]

Secondary nucleation is the process by which crystal nuclei are formed within the environment where the nuclei of the crystal in the same species are already presented. These preexisting nuclei are defined as crystal seeds. In general, secondary nucleation occurs at a faster rate than primary nucleation. This is due to the fact that primary nucleation necessitates the spontaneous formation of a nucleus from the supersaturated solution, while secondary nucleation occurs more easily because the crystal seed serves as a starting point for the formation of new crystals.[35] Secondary nucleation plays a major role in many industrial crystallisation processes because the seeds can be designed to control the crystal growth rate, as well as the properties of the resulting crystals, such as polymorphism, crystal shape and crystal size distribution.[36]

1.2.1.1 Classical Nucleation Theory

CNT was introduced by Volmer in 1939. In CNT, a nucleus formed by monomer association has the same structure as the new crystal phase.[42] According to this theory, nucleation rate J ($m^{-3}s^{-1}$), which is defined as the number of crystalline particles formed per unit of time within a specific solution volume, can be expressed by **Equation 1.**[43]

$$J = AS\ exp\left(-\frac{B}{ln^2 S}\right) \qquad \textbf{Equation 1}$$

where S (unitless) represents a degree of supersaturation calculated from the ratio of solution's concentration and equilibrium concentration in a supersaturated system and A ($m^{-3}s^{-1}$) and B (unitless) are the pre-exponential factor and thermodynamic factor, respectively. The pre-exponential factor A is a parameter that describes the molecular kinetics of the nucleation process and reflects the attachment rate for solute molecules moving from the solution to the surface of the nuclei. The thermodynamic factor B reflects the energy barrier of nucleation. As seen in **Equation 1**, the correlation between the thermodynamic factor B and the nucleation rate is inverse and exponential, and thus small changes in the value of supersaturation can have a large impact on the nucleation rate. Although the nucleation of many organic molecules can be explained by CNT, the presence of intermediate metastable species during the nucleation of some protein molecules,[44–46] polymers[47–49] and inorganic structures[50,51] provides evidence of different nucleation mechanisms.

1.2.1.2 Two-step nucleation theory

While classical nucleation theory states that the crystal nuclei are directly formed by the aggregation of solute molecules that are supersaturated in the solution, a two-step nucleation mechanism suggests that the formation of the disordered liquid droplet called a metastable intermediate phase (MIP) occurs in supersaturated solutions before crystal nuclei are produced inside the MIP droplets.[52] This MIP has a thermodynamic stability higher than the parent phase (solution) but lower than the crystal phase.[53] The formation of the MIP conforms to Oswald's rule of stages which states that the least thermodynamically stable form with the closest free energy difference to the initial stage will form first in any crystallisation, followed by the more stable one.[54,55]

The work supporting the two-step nucleation mechanism was carried out using molecular dynamics simulation. Gavezzotti simulated a system consisting of 50 separated molecules of acetic acid within a box of 1,659 solvent units. After increasing the concentration by removing solvent molecules from the system, the aggregation of acetic acid molecules into a microemulsion of liquid-like clusters was observed.[56] Another work using molecular dynamic simulation was done by Shore and Perchak, who studied the nucleation of AgBr in water. Similar to Gavezzotti's work, molecular dynamic simulation was used in this study. The prenucleation clusters of $Ag_{18}Br_{18}$ were found to be disordered.[57] The results from these

studies support the presence of disorder clusters as the initial step of nucleation from solution in a two-step nucleation mechanism.

Evidence of prenucleation clusters in two-step nucleation was also investigated by various analytical techniques. For example, dynamic and static light scattering was used to study the nucleation of lysozyme crystals. The results showed that lysozyme monomers aggregated into the clusters before these clusters restructured into compact structures.[58]

Additionally, The study of protein crystals from aqueous solution by Haas and Drenth proposes that, in the presence of a thin liquid film which covers protein crystals, the surface energy of the crystals will considerably decrease. The high protein concentration in the liquid film enables the protein molecules to adsorb and incorporate into the crystal surface easier than in the process in the absence of the liquid film.[59]

1.2.2 Crystal growth

Crystal growth is a process in which molecules of solute in a supersaturated solution are incorporated onto a crystal surface resulting in an increase in crystal size. The process can be roughly divided into 3 steps: i) solute molecules move through the solution, ii) molecules from the solution are adsorbed to the surface of crystals, and iii) the molecules move to edge and kink positions and orderly arrange on the crystal surface.[60]

1.2.2.1 Crystal growth mechanisms

Two-dimensional layer growth mechanism: This theory is the first atomistic model explored by W. Kossel and I. N. Stranski said that the growth process of crystals occurs in two dimensions. Crystal growth units spread from the crystal nucleus to the directions parallel to the growth layer until the first crystal plane is completely formed. Then two-dimensional nucleation for another growth layer occurs and the growth process continues. Thus, the growth rate of the two-dimensional layer growth mechanism is determined by the nucleation rate. This phenomenon means growth cannot happen below a critical supersaturation (critical supersaturation being the supersaturation in the metastable zone below which nucleation cannot take place).[61] The two-dimensional layer growth mechanism results in the formation of the smooth, flat faces of crystals.[62] **Figure 5** shows the two-dimensional layer growth mechanism.

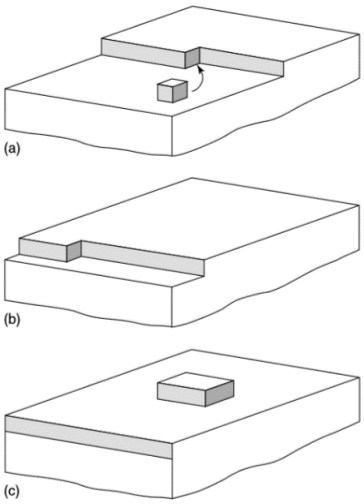


Figure 5. Two-dimensional layer growth mechanism: (a) growth units adsorb to the crystal surface and diffuse to a step, (b) the step continues growing in the direction of crystal edges, (c) a formation of a two-dimensional nucleus occurs after the previous layer is completely formed (figure from reference)[60]

Spiral growth mechanism: Whilst growth in the two-dimensional layer model needs relatively high supersaturations, a spiral growth mechanism has a lower energy barrier thus allowing crystals growth at lower supersaturations. In the spiral growth, dislocations of the crystal molecules create steps on crystal surfaces. In the initial stages of crystal growth, growth units attach to the initial step created from the dislocation. The process continues and a second step is generated, followed by a third step, and so on. Eventually, the spiral pattern forms, as seen in **Figure 6**. If the dislocation creates a curved step at the initial state, the spiral pattern on the crystal surface will be rounded (**Figure 7**). These steps in this growth process are self-perpetuating and do not require additional nucleation events, thus resulting in the lower energy required for this growth mechanism.[60]

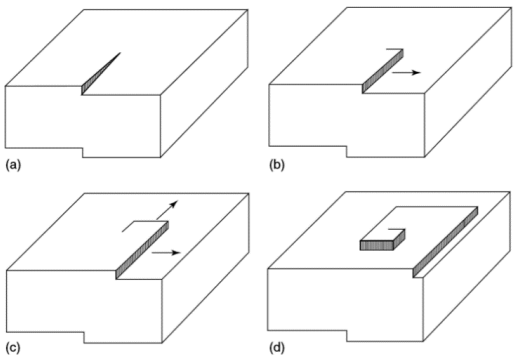


Figure 6. The spiral growth mechanism for crystal growth (figure from reference)[60]

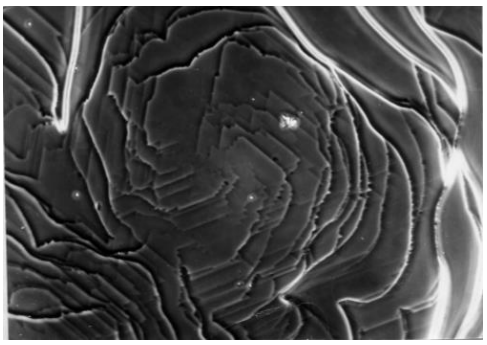


Figure 7. The spiral pattern on the crystal surface(figure from reference)[62]

Adhesive-type growth mechanism: The adhesive-type growth mechanism creates crystals with rough surfaces while other growth mechanisms create those with smooth surfaces. **Figure 8** shows that the adhesive-type growth dominates crystal growth at high supersaturation and hence a transformation of crystal surfaces from smooth to rough takes place as supersaturation increases. In the adhesive-type growth mechanism, growth units are bounded by the crystal surfaces regardless of the crystallographic direction, resulting in spherulitic, fractal, or dendritic crystal morphologies.[60,62]

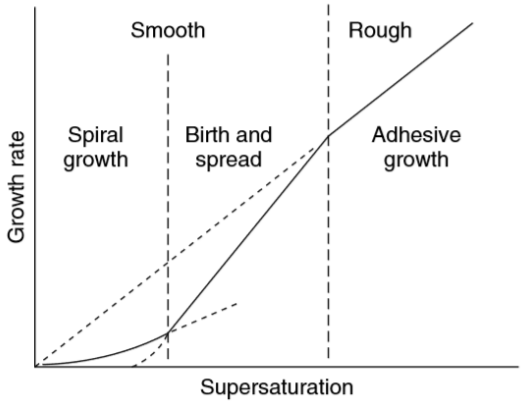


Figure 8. A diagram of the preferred crystal growth mechanism at different supersaturations and crystal growth rates (figure from reference).[60]

1.3 Cooling crystallisation

Crystals can be obtained by various methods, including cooling crystallisation, solvent evaporative crystallisation, precipitation (anti-solvent and pH shift crystallisation), sublimation, vapour diffusion, crystallisation from melts, thermal treatment, and thermal desolvation. Crystallisation from solution is most commonly used for the manufacture of solid bulk APIs in pharmaceutical industries. Here, we will focus on cooling crystallisation as the crystallisation method predominantly used in this thesis.

Cooling crystallisation has a comparatively simple set-up compared to other crystallisation methods. Cooling crystallisation requires that the molecule to be crystallised has a temperature-dependent solubility in the solution of interest.[63] In general, the solubility of organic compounds under elevated temperatures is higher than the solubility at low temperatures. Therefore, crystallisation will take place when the temperature of the solution decreases to a point where the concentration of the solution exceeds its equilibrium solubility. Crystal yield for this method is calculated using the difference between the initial amount of compound and the remaining amount of solute in the solution at the end of the crystallisation process (**Equation 2**).

$$\text{Crystal yield (Y)} = \frac{\text{(Mass of soluted used – Mass of solute remaining in mother liquor)}}{\text{Mass of solute used}} \times 100\ \%$$

Equation 2[64]

Solvent selection should consider the solubility of the crystalline material. Typically, if the initial solubility is too high at high temperatures, the slurry at low temperatures may be too dense for crystallisation to occur. Likewise, if the solubility at low temperature is too low, the precipitation of impurities tends to occur and impact the quality of crystal products.[63] To obtain large enough crystals for crystal shape analysis, the cooling rate should also be taken into account, since rapid cooling results in a large number of small crystals.[65,66] **Figure 9** demonstrates the solubility – supersaturation diagram of cooling crystallisation.

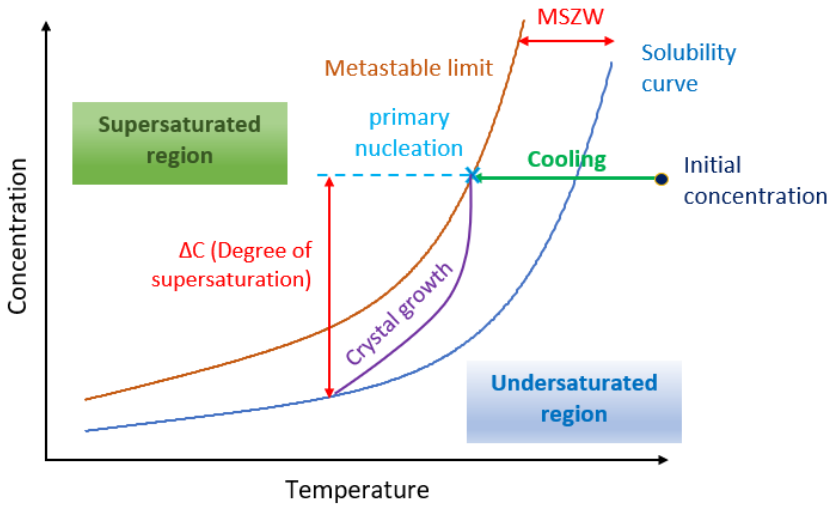


Figure 9. The solubility – supersaturation diagram of cooling crystallisation

From the above diagram, the concentration of a crystallisation solution is initially undersaturated at high temperature (see point labelled 'Initial concentration' in **Figure 9**), which means the concentration is lower than the equilibrium solubility of the compound in the chosen solvent and the compound is completely dissolved. The area between the solubility curve and the metastable limit curve, known as the metastable zone (MSZ), is a supersaturated region where the crystallisation system is in an equilibrium state between the dissolved and crystalline phases. In MSZ, concentration of the solution is higher than the saturation point but the solutes are unable to crystallise because the unstable clusters or nuclei in the solution are consistently formed and dissolved.[67] Once the system reaches the metastable limit, a nucleation occurs spontaneously and the solute molecules in a saturated solution attach to the surface of the nucleus until eventually grow into a crystal.[68] Note that this diagram is for unseeded crystallisation, so there is no crystal growth inside the metastable zone and the nucleation that occurs at the metastable limit is primary nucleation.[69,70]

1.4 Factors affecting the shape of crystals

Crystal shape is one of the important attributes for determining the quality of crystalline materials. Different crystal shapes can impact the physical and chemical properties of crystalline materials in different ways. For example, rod shape crystals of simvastatin have a faster dissolution rate than plate-like crystals. This difference in dissolution rate happens because the different shaped crystals have different surface areas and surfaces with different polarities.[71,72] Crystal shape also has a noticeable impact on mechanical properties that are important for downstream processes, such as flowability[73,74] and filtration.[19,75] Different crystal shapes also cause different tabletting performance (i.e., differences in compressibility and compactibility) even when the same compaction pressure is applied.[76] Undesirable shapes of crystals such as needle-like or plate-like crystals can cause problems in the separation, washing, and drying steps after crystallisation.[77]

The shapes of crystals are governed by chemical composition, the internal structures of a crystal itself and external factors such as solvent effects and the presence of additives or impurities and morphological instabilities. Furthering our understanding of these factors will improve our ability to control crystal shape during crystallisation and attain crystalline products with desirable properties.

1.4.1 Supersaturation

Supersaturation can influence the final product of crystallisation. In cooling crystallisation, supersaturation can be controlled by changing the cooling profile,[78] the solution temperature,[79,80] and the initial concentration of a solution.[81] Since crystal shape is determined by the different growth rates along each crystallographic direction and these growth rates are dependent on supersaturation, the desired shape of crystals can be obtained by controlling the supersaturation level.[78,82] Two studies on the relationship between crystal shape and supersaturation carried out by Zuozhong Liang and Diana Camacho show that the aspect ratio of triclinic N-docosane crystals grown from dodecane solution (**Figure 10**) and the aspect ratio of benzoic acid crystals grown from water (**Figure 11**) are both supersaturation-dependent. Indeed, these studies showed that aspect ratio decreases with increasing supersaturation levels.[79,81] Another example of supersaturation affecting crystal shape is a study of the crystal growth rate of methyl stearate as a function of supersaturation. In this study, the growth rates of individual crystal faces were measured by observing the increased distance perpendicularly from the centre of the crystal to each face using subsequent crystal micrographs recorded every 5-20 s, and then the mean growth rate of each crystal face was calculated. The results showed that growth rates of the crystal faces differed at different supersaturations.[80] Furthermore, a study of paracetamol crystals grown from an aqueous solution conducted by Finnie et. al. showed that paracetamol crystals exhibited a columnar shape at low supersaturation while plate-like crystals were observed at high supersaturation (**Figure 12**).[83]

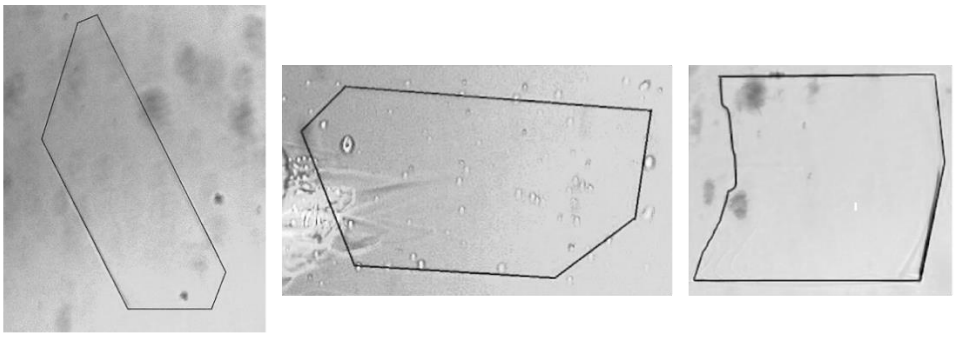


Figure 10. *The shape of N-docosane crystals grown from n-dodecane solution at different supersaturations (a) S = 0.01, (b) S = 0.02, and (c) S = 0.05 (figure from reference). The authors of this work provided these images to illustrate the trend that crystal aspect ratio can decrease as supersaturation increases.*[79]

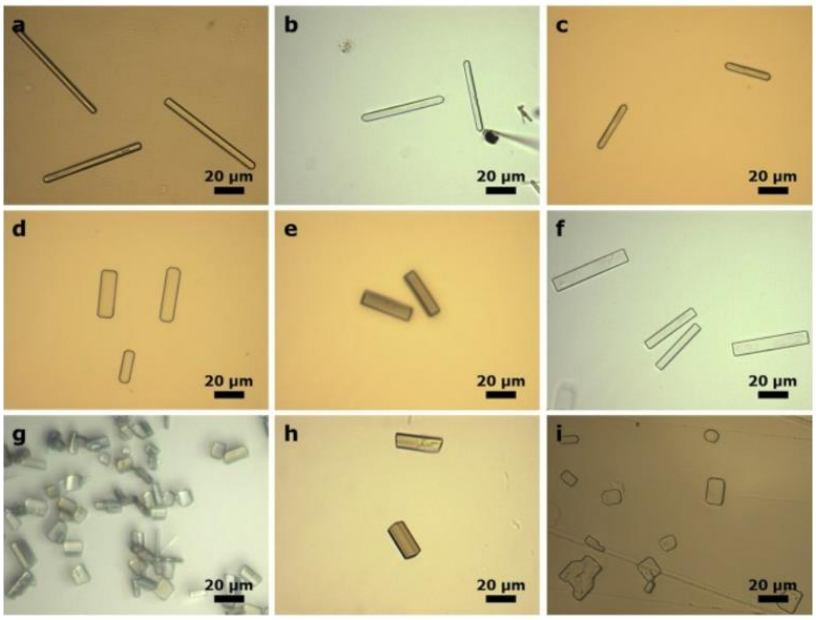


Figure 11. *The shape of benzoic acid crystals grown from aqueous solution at different supersaturations (a) S = 1.029, (b) S = 1.103, (c) S = 1.206, (d) S = 1.353, (e) S = 1.397, (f) S = 1.47, (g) S = 1.618, (h) S = 2.059 and (i) S = 2.941 (figure from reference)*[81]

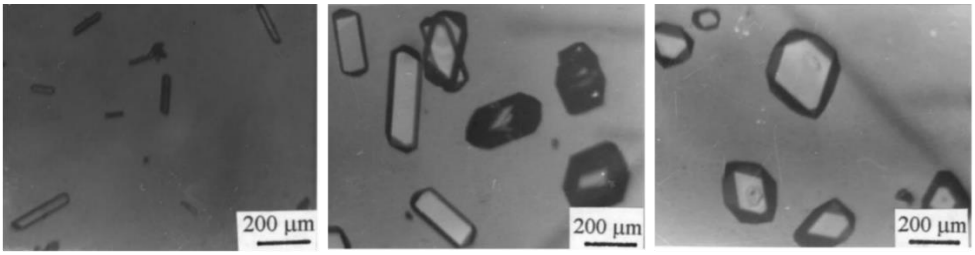


Figure 12. *The shape of paracetamol crystals grown from aqueous solution at different supersaturations (a) low S, (b) moderate S, and (c) high S (figure from reference)*[83]

1.4.2 Solvent effects

Intermolecular interactions between the solvent and solute molecules on the crystal surface, such as hydrogen bonding, can influence the shape of crystals by impeding crystal growth via molecular attachment in specific crystallographic directions.[66] One example of this phenomenon arises in a study on steroid 7αMNa. In this study, crystals of 7αMNa form I grown from acetone solution exhibited a plate-like appearance with distinct (010) faces. However, the crystals showed an unexpected polar shape (a crystal habit that results from two opposite crystal faces having different growth rates, see **Figure 13**) when methanol or ethanol was used as a solvent. The polar-shaped crystals suggest that there were differences in directional growth rates. The authors suggested that the different crystal shapes resulted

from the alcohol solution molecules interacting with the hydroxyl groups of the 7αMNa molecules on the crystal surface, thus inhibiting growth in the relevant direction.[84] Studies on dirithromycin crystals and salicylamide crystals grown from different organic solvents showed similar results. The results from these studies indicate that strong intermolecular interactions between solvent molecules and dirithromycin and salicylamide result in different crystal habits and different dominant crystal faces.[85,86]

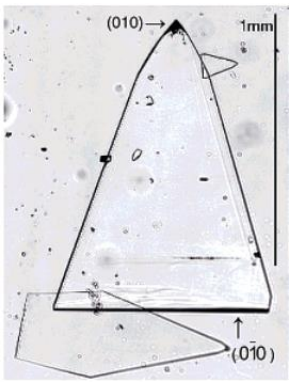


***Figure 13**. The polar-shaped 7αMNa crystal where the growth rate of (010) face was faster than that of (0$\bar{1}$0) face (figure from reference).*[84]

In a study on solvent effects on benzoic acid crystals, the correlation between solvents and aspect ratios of the crystals reveals that some solvent properties can determine how crystals grow. Experimental observations of crystal shapes indicate that the aspect ratio is directly proportional to solvent polarity and inversely proportional to the molecular size of the solvent. Conversely, the intrinsic crystal structure of benzoic acid is not affected by the solvent even for crystals exhibiting different crystal habits.[87]

Additionally, the viscosity of solvents can affect the crystal habit due to the limitation on mass transfer in high viscosity solvents.[88,89] Since the crystal growth rate is controlled by the diffusion rate of the solute in the liquid phase, the growth will be faster in low viscosity solvents and slower in high viscosity solvents.

1.4.3 Additives/Impurities

Similar to solvents, molecules of additives or impurities can be adsorbed at growth sites on crystal faces and impede the growth of specific crystal faces. The incorporation of these compounds causes a difference in growth rates of the crystal surfaces and hence can affect the crystal shape.[90–93] For example, the influence of additives on the morphology of γ-aminobutyric acid (GABA) crystals was studied by Gabas and Lin.[92] In this work, additives

were added to an aqueous solution of GABA, and the face growth rates of GABA crystals in the presence of the additives were compared to the growth rates of GABA crystals in an aqueous solution without additives. Different face growth rates and different crystal morphologies were observed when adding either hydrated chromium nitrate, dodecyltrimethylammonium bromide or hexane-1,2-diol. These three additives slowed down the growth rate of the (001) GABA crystal face and resulted in a flatter crystal shape than the crystals grown from pure aqueous solution.[92]

Generally, the role of tailor-made additives is to selectively hinder the growth of crystal faces. The crystal face whose growth is inhibited will then dominate the resulting crystal morphology. The work of Davey et al., which focused on the effect of tailor-made additives on glycine crystallisation, also discovered that some additives can selectively inhibit the growth of glycine crystals in α-form. In this study, the additives make the growth of γ -glycine crystals, crystals which are normally less likely to nucleate compared to α-glycine, more favourable.[94]

1.4.4 Crystallisation parameters

Experimental parameters, such as crystallisation temperature, cooling rate,[78] and evaporation rate,[95] also influence crystallization driving force and so crystal growth rates, and hence, crystal shape. For instance, the study of α-lactose monohydrate crystallising from an aqueous solution is a good example of crystallisation conditions influencing crystal morphology. The results from this study showed that lactose crystals exhibited more regular shapes and smoother surfaces when they were prepared at 40 °C crystallisation temperature than compared to those prepared at 0 °C.[96] Additionally, external movement or vibration can also affect the crystal shape because vibrations can induce crystal nucleation. A faster nucleation rate can cause large numbers of disordered small crystals to form.[97]

1.4.5 Morphological instability

Morphological instability occurs when a crystal cannot maintain its interfacial form during the growth process. Under conditions with high crystallisation driving forces (i.e. high supersaturation), an adhesive-type growth mechanism dominates the crystal growth. Consequently, the resultant crystal has a rough interface. In such situations, plate or needle crystals may grow very rapidly around a nucleus and form bow-tie or spherulitic crystal aggregates. Another example of morphological instability is the growth of crystals with fractal patterns. These patterns are caused when continuous nucleation at the end of the crystal

results in connected patterns of the crystals. While spherulitic crystals form from polycrystalline aggregates, fractal or dendritic crystals form from a single crystal.[62]

1.5 Machine learning

With advancements in computer technology, many problems can be solved by using a computer operating system. Generally, a set of instructions used to generate an output by assessing input data, also known as an algorithm, is required for solving a problem. There are different algorithms for different tasks, such as algorithms for sorting numbers into ascendant order or algorithms for finding a maximum value from a set of random numbers, etc. Some tasks are more challenging because the correlation between input and output is unresolved or unknown. The use of machine learning can facilitate such tasks by using a training step that does not require a comprehensive understanding of the specific problem.[98] At present, many machine learning algorithms have been developed and applied to various jobs. For example, machine learning can be used for email spam filtering[99,100] and email classification.[101] In healthcare, machine learning has been used for the detection, diagnosis, and monitoring of certain diseases.[102–106]

1.5.1 Learning methods of machine learning

Machine learning can be classified by learning style into 3 groups as followed: supervised learning, unsupervised learning, and reinforcement learning.

1.5.1.1 Supervised learning

Training datasets in supervised machine learning must include a predefined label or a target for each data point. During the training step, the model learns to search for the correlations between independent variables and the predefined labels. To predict the label for new data, the model will use the optimised fit from the training dataset to predict the most likely outcome for the new data. Examples of problem types that can be solved by this method are classification and regression. Supervised machine learning is commonly used in the physical sciences and can be used for tasks such as predicting physical properties of interest from chemical composition.[98,107]

1.5.1.2 Unsupervised learning

Unlike the labelled data in supervised learning, the input data for unsupervised machine learning does not have a known result. This type of machine learning aims to find patterns in

the input data by organizing the data via a mathematical process or similarities in the data. Examples of problem types for which unsupervised learning can be used are clustering and dimensionality reduction.[98,107]

1.5.1.3 Reinforcement learning

Algorithms in reinforcement learning search for a sequence of actions that can achieve the best results or maximize cumulative rewards. Reinforcement learning is similar to learning by trial and error in that the learning process of reinforcement learning iteratively performs a sequence of actions where, in each new sequence, the algorithm can change in response to outcomes from previous sequences. In the other words, reinforcement learning will repeat actions that had positive results (a reward defined by the programmer) and avoid the actions which lead to less reward. When successful, this process will iterate until the algorithm reaches a target goal. Games like chess are similar to reinforcement learning because there are a large number of possible moves the players can make that result in different outcomes.[98]

1.5.2 Machine learning algorithms

Machine learning algorithms have been developed and applied to various tasks in the area of crystallisation.[98] Many studies that investigate crystallisation outcomes employ Random Forest (RF) classification and regression. An explanation of RF algorithms is given below followed by a discussion of the use of RF in the literature for the prediction of crystallisation outcomes. This discussion includes comparisons between the performance of RF and the performance of other algorithms for predicting crystallisation outcomes. As RF often performed as well as or outperformed other algorithms in these studies, RF is the algorithm that this thesis will focus on.

1.5.2.1 RF classification and its applications in Predicting Crystallisation Outcomes

RF classification algorithms consist of a large number of decision trees. The structure of the decision tree is demonstrated in **Figure 14**. Each tree has 3 main components: decision nodes, branches, and leaf nodes. Each decision node corresponds to a test, which is an independent variable in the dataset and has two or more branches to the other variables. Each branch corresponds to a result of the test and indicates which node should be considered next. The topmost decision node of the tree structure containing the best predictor is called a root node. The definition of the best predictor is the feature that can be used to split the dataset

into subgroups with the largest differences from each other. The result of the prediction is indicated by the leaf nodes, which are the end of the tree structure.[108,109]

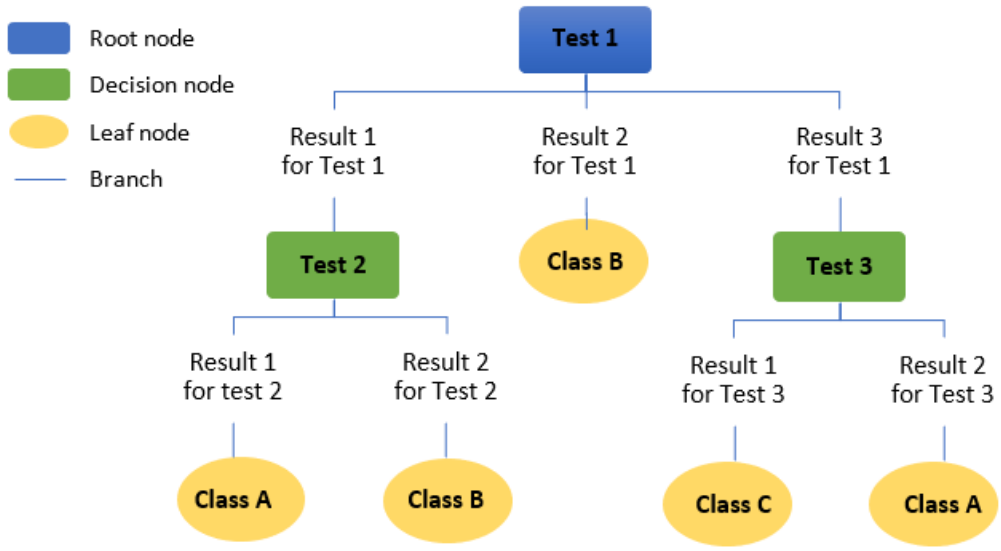


***Figure 14**. The structure of the decision tree (adapted from [108])*

In the case of a dataset with a large number of independent variables, adding too many variables may cause a too complicated decision tree which can lead to overfitting. RF is a methodology developed based on the decision tree but can avoid the overfitting of training data.[110] The key is that RF works by randomly selecting the data from the original dataset to build a set of different decision trees. Each tree in an RF has different structures due to the different subsets of randomly selected variables, which results from a feature randomness technique. Additionally, RF uses the subsampling technique called bagging or bootstrapping to randomly select the observations from the dataset to create multiple decision trees with the same size. This techniques samples the data with replacement in order to make the subsequent selections independent from the previous selections.[111] With the combination of these techniques, an RF has variation among the trees in the model, and each tree has a low correlation to the other trees in the RF. In the classification, the class which is chosen by the majority of the trees in the model will become the model's prediction. As a result, the predictions from the RF model are likely to be more accurate than those from a decision tree because the errors made from one individual tree can be mitigated by the others, so long as all trees do not make mistakes in the same direction.[112] **Figure 15** illustrates how the RF classification model works.

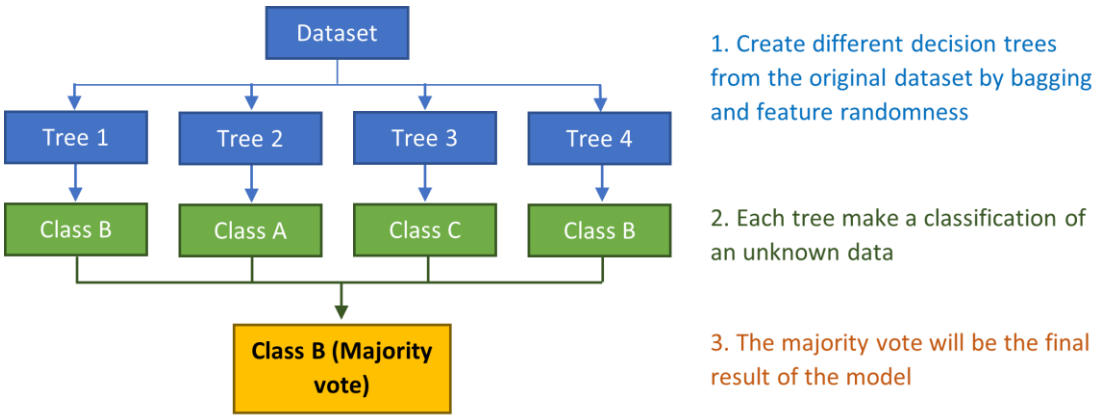


***Figure 15**. Random forest classification model*

Apart from the resistance to overfitting, RF also carries many advantages over other algorithms, including the robustness of the model to outliers and noises. Moreover, RF can provide an important score ranking indicating which variables have a high impact on the target or dependent variable by considering an increase of prediction error when each variable is permuted while all other variables remain the same.[112,113]

RF is previously applied for the prediction of crystallisation outcomes. For example, the prediction of the crystallisation propensity of organic molecules carried out by Bhardwaj *et. al.*,[114] the prediction of the crystallisability of carbamazepine solvates by Johnston *et al.*,[31] and the prediction of propensity to form solvates of pharmaceutical organic molecules by Xin et al.[115]

From the previous studies, RF classification was applied for the prediction of carbamazepine solvate and its three polymorphs. In this study, the classification model was trained by the numerical physicochemical solvent descriptors, crystallisation conditions (vacuum, vortex stirring, temperature), and the crystallisation outcomes obtained from 326 experiments of carbamazepine crystallisations. The prediction results from this model suggested that carbamazepine had the potential to form solvates in three of the solvents that initially yielded non-solvated crystals. Guided by these predictions, the novel carbamazepine solvate was discovered from recrystallisation at lower temperatures. This finding highlights machine learning's potential to predict novel crystal forms.[116]

RF classification has also been used to predict the crystallisability of small organic molecules. The training data was obtained from the crystallisation of 382 different acylanilide compounds. The predictive model was trained with calculated molecular descriptors for each

acylanilide compound and the crystallisation outcomes (i.e. indicating whether a single crystal was observed). This study showed that molecular structure can be used to predict the crystallisability of small organic molecules. The study also indicated some problematic compounds, an outcome that could inform the selection of the crystallisation compound in the early-stage experiments.[114]

Similar studies were carried out using RF classification to predict crystallisation propensity. A molecule's propensity to crystallise was identified using the presence or absence of single-crystal diffractograms in CSD, and molecular descriptors were used to train the predictive model. The results from this study showed that only a few molecular features impact crystallisation propensity and that the prediction accuracy can reach 90.3% by selecting only those features as the model variables.[32]

In another study, an RF classification model was also applied to predict crystallisation propensity in more detail with two positive crystallisation classes (crystals and microcrystals) and four negative crystallisation classes (crystalline tendencies but no crystals, droplets, films, and amorphous). The experimental data were obtained from 5,710 solvent evaporative crystallisation experiments. Molecular weight and relative solubilities of compounds available in the literature were used as the model variables. This approach can guide the decision-making in solvent selection for crystallisation with success rates of over 92%.[117]

Predicting crystal packing for olanzapine solvates was also achieved by applying an RF classification model that was trained with the molecular descriptors of crystallisation solvents. Three classes of crystal packing were observed in the crystallisation experiments of olanzapine solvates. This study revealed that van der Waals volume, number of covalent bonds, and polarizability of the solvent molecules play an important role in olanzapine crystal packing.[29]

RF classification, amongst other ML methods, was also shown to be an effective tool for predicting whether or not a single crystal will grow in given conditions.[118] Another study compared the efficacy of three algorithms (RF, support vector machines, and neural networks), in building models to predict the crystallisation propensity of small organic compounds. The results showed that the RF model had the highest prediction accuracy, especially when a large dataset was used to train the models. Additionally, this study also suggested that the presence of impurities and degradants had a negative impact on the

prediction ability of the model. Consequently, the model was improved by excluding such experiments from the training dataset.[119]

In another study, RF classification also showed the best performance when compared with other model types for predicting three classes of crystallisation outcomes, namely crystalline, polymorphic, and amorphous. By using calculated molecular descriptors and fingerprints for chemical structures of crystallisation compounds, this model can predict the crystallisation outcomes from an external test set with up to 80% accuracy.[120]

As the studies described above have demonstrated various successful applications of RF classification to predicting crystallisation outcomes, RF will be predominantly used in the work presented in this thesis to explore similar challenges in crystallisation experiments.

1.5.2.2 Logistic regression

Logistic regression is an analytic tool commonly used in classification tasks. It is often used for estimating the probability of a binary outcome.[121] Unlike linear regression which searches for the best fit line to the data using the least square method, logistic regression fits an S-curve called "logistic function". This function informs us of the probability between two classes which values between 0 to 1. Values close to 0 indicate low probability of the outcome, and values close to 1 indicate high probability.[122] Logistic function is defined as **Equation 3**.

$$p = \frac{1}{1+e^{-z}} \qquad \textbf{Equation 3}^{122}$$

Where p is the probability of the outcome occurring and z is the linear combination of the input variables (x_i) and the model parameters (b_i) as demonstrated in **Equation 4**.

$$z = b_0 + b_1x_1 + b_2x_2 + ... + b_nx_n \qquad \textbf{Equation 4}^{122}$$

Figure 16 illustrates the differences between linear regression and logistic regression when they are applied to a classification task with binary outcome. When linear regression is applied, the predicted Y values can exceed the 0 – 1 range (infinite range by theory). On the other hand, logistic regression uses sigmoid curve so the predicted Y values are limited in the range of 0 – 1.[123] In this example, Logistic regression is preferrable due to a non-linear correlation of the data and that there are only two possible outcomes that need to be predicted.

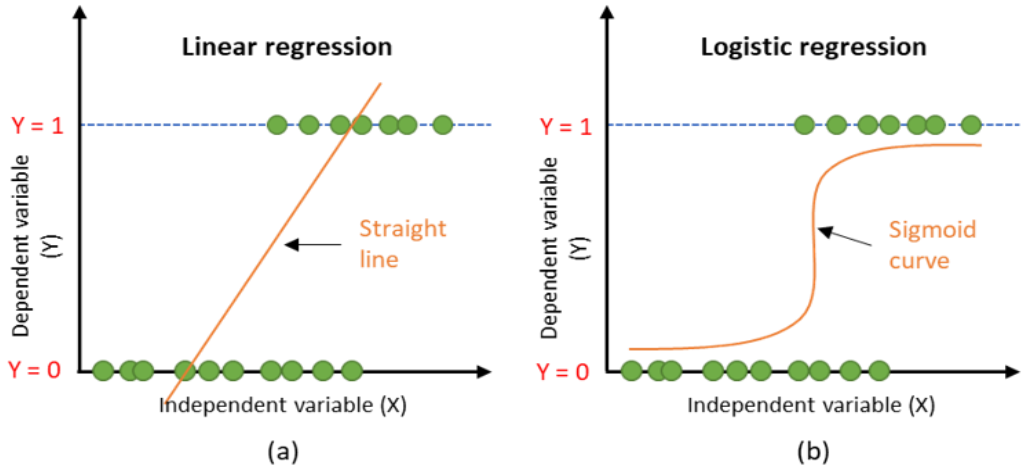


***Figure 16**. Comparison between linear regression and logistic regresstion. (a) Linear regression creates the best-fit straight line used for predicting the continuous y values. The value of y predicted by linear regression can exceed the 0 – 1 range. (b) Logistic regression uses a sigmoid curve (S-curve) to classify the data into two classes (0 or 1, true or false, or any binary outcomes). The value of y predicted by logistic regression can be only in the 0 – 1 range. In the case of classification task with a binary outcome, logistic regression outperforms linear regression. Green points represent the data points with the value of y = 0 and y = 1.*

Logistic regression has been applied in the studies in crystallisation area, such as predicting whether the crystal growth process model with different set of simulated data (process variables) results in monocrystalline or polycrystalline crystals,[124] or predicting the crystallisation propensity of proteins.[125] These studies demonstrated that logistic regression can be used to identify the key variables influencing the process outcomes. In the study carried out by Eri Shimono et. al.,[126] logistic regression analysis also showed the potential to be a useful tool in crystal design. The logistic regression model in this study was applied to the large-scaled crystallographic database and the element group number was used for the prediction of chiral crystal propensity. The results revealed that the crystals containing some specific element groups have a high probability of becoming chiral crystals. This finding supports that logistic regression can advance the design of chiral crystals.[126]

1.6 General challenges with machine learning

Machine learning is a complex method which requires a lot of complicated mathematical and statistical processes. There are many challenges that we have faced in an attempt of building the model using machine learning techniques.

- Poor quality of training data

Data plays an important role in machine learning. The model trained by poor quality data, such as variables unrelated to the outputs, uncleaned data, missing data, and data containing many outliers, cannot produce an accurate and reliable prediction. Therefore, good practice in data collection and data preprocessing is necessary for building the model.[127–129]

- Small dataset

A sufficient amount of data in the training dataset is crucial to achieving an accurate machine learning model, especially for the prediction which requires complicated data. Less amount of data cannot provide machine learning with enough training and results in an inaccurate and biased model. The amount of time for collecting enough data with good quality for machine learning is also one of the challenges that need to be concerned.[127,128,130]

- Class imbalance

Class imbalance can be a problem for some predictive models. The model tends to overpredict the data in the class with a large amount of data and underpredict the data in the class with a small amount of data, resulting in a biased prediction. Accordingly, a balanced distribution of data in all outputs is preferable for training the machine learning model.[127,128]

- Overfitting and underfitting

Overfitting and underfitting are the words used in data sciences to describe a machine learning problem in terms of model flexibility in training data. Overfitting occurs when a model is too flexible and fits too much training data that emphasizes even noises or outliers rather than actual values in a dataset. This usually happens in a complicated model in which the dataset contains too many independent variables compared to the number of samples. The overfitting model has very high accuracy for in-sample data (using data in a training dataset to evaluate the model) but it will poorly perform with new data or untrained data. As a result, the accuracy of the overfitting model will be very low when it comes to an out-of-sample evaluation (using test data to evaluate the model). On the contrary, the underfitting model means the model which is too simple that it cannot represent the actual trend in a dataset. It may happen when the number of variables in a dataset is not enough for making a reasonable prediction, or when the model's function is selected incorrectly. For

example, when a linear function is selected to use in a model with non-linear data. The Underfitting model performs very poorly with both in-sample and out-of-sample data.[131,132]

- Data interpretation

Since machine learning is a process that consisted of many complex steps, it is sometimes difficult to explain how the prediction is made or how the outputs correlated to the model's variables. Some machine learning algorithms which involve a data transformation process, such as a support vector machine that applies kernel function to transform the data into higher dimensional space, can make the data interpretation more complicated.[129]

2. Aims and Objectives

2 Aims and Objectives

This thesis aims to help understand the fundamentals of the crystallisation process, specifically nucleation kinetics, and to investigate the potential and reliability of machine learning in the research field of crystallisation. Predicting crystallisation outcomes can minimize the number of required experiments in the screening step, accelerate development and reduce the cost and time required to develop robust processes. Understanding the correlation between solvent structures and crystallisation thermodynamics and kinetics can also guide decision-making in solvent selection. It is important to note that all of the crystallisation experiments discussed within this thesis employ the cooling from solution technique and are conducted in an absence of a seed crystal (i.e. unseeded cooling crystallisation). A description of each experimental chapter in this thesis with the aims and objectives for each chapter follows.

Chapter 4, How crystallisation thermodynamics affect the nucleation barrier:

Aims

This chapter focused on improving our understanding of the thermodynamics and kinetics of the crystallisation of MFA in thirty-two different organic solvents. Crystallisation enthalpy (ΔH^0_{cryst}), entropy (ΔS^0_{cryst}), and Gibbs' free energy (ΔG^0_{cryst}), were calculated using solubility data and the nucleation rate was determined from the probability distribution of induction time measurements.

Objectives

- Determine the solubility of MFA in various organic solvents from solubility curves obtained from turbidity measurement using a Crystal16 crystallizer.
- Calculate thermodynamic and kinetic parameters of crystallisation of MFA from van 't Hoff coordinate plots of the solubility as a function of temperature.
- Determine the nucleation rates of MFA in different organic solvents by the probability distribution of induction time and maximum likelihood estimation.
- Compare the calculated thermodynamic/kinetic parameters and nucleation rates of MFA between six organic solvents.

- Plot the correlations between thermodynamic factor B and ΔH^0_{cryst}, ΔS^0_{cryst}, and ΔG^0_{cryst} to see if the crystallisation of MFA conforms with Turnbull's empirical rule
- Determine the impact of solvent on the observed solution thermodynamics and observed crystallization outcome.

Chapter 5, Prediction of mefenamic acid crystal shape by random forest classification:

Aims:

This chapter investigates the potential of using machine learning to predict crystal shape. In this chapter, cooling crystallisation screening of MFA in various organic solvents was studied. The shape of the resulting crystals was observed, and, together with the supersaturation and molecular descriptors of solvents used, these observations formed a dataset for RF classification models for crystal shape prediction. A range of advanced analytical techniques, such as Powder X-ray Diffraction (PXRD), single-crystal X-ray diffraction (SD-XRD), and DSC, has also been applied to characterize the MFA crystals.

Objectives:

- Observe the crystal shapes using optical microscope and determine polymorphic forms of MFA using PXRD in a diverse range of organic solvents from small-scale cooling crystallisation screening
- Develop machine learning models for the prediction of MFA crystal shape by applying an RF classification algorithm to the dataset containing experimental supersaturations and calculated molecular descriptors of crystallisation solvents as the model's input variables, and using crystal shape data as the model's output.
- Apply logistic regression to understand the impact of solvent descriptors on the model performance
- Assess the predictive performance of the machine learning model

Chapter 6, Investigating potential correlations between PXRD peaks at low angles and the crystal structures of solvates/non-solvates:

Aims

This chapter explores the powder patterns generated from crystallographic data available in the Cambridge Structural Database (CSD) to determine if there is a correlation between the low-angled peaks of the PXRD patterns and the solvated forms of small organic molecules.

An RF classification algorithm was applied to the models for the prediction of solvated and non-solvated structures using the peak data of the PXRD patterns as the model's variables and the model's performances were evaluated.

Models

- Explore solvate and non-solvate crystal structures of small organic molecules from the CCDC database and extract their PXRD data
- Investigate the statistical data of the solvate and non-solvate structures crystallised from different crystallisation solvents in CCDC and the presence of the peaks in PXRD at low 2-theta angles
- Investigate the solvent properties that may correlate with the probability of solvate formation
- Develop machine learning models for the prediction of solvated and non-solvated crystal structures from PXRD data and assess their performance

3. Materials and Methods

3 Materials and Methods

3.1 Overview

Materials and methods specific to each chapter can be found within the individual research chapters (See Section 4.3 for Chapter 4, 5.2 for Chapter 5, and 6.5 for Chapter 6). Solubility measurements can be found below as these methods are used in Chapter 4 and Chapter 5. X-ray powder diffraction and introduction to the model evaluation method for validating RF classification can be found below as these methods are used in Chapter 5 and Chapter 6.

3.2 Solubility measurements

A known amount of MFA was weighed into a 1.5 ml high-performance liquid chromatography (HPLC) vial. 1.5 ml of a specific solvent selected from the library was then pipetted into this pre-weighed vial containing the solid material and stirrer bar. The vial was then reweighed to determine the exact mass of solvent added and therefore the exact molar composition of the sample. Each vial was capped tightly and the cap was carefully wrapped in parafilm tape to create a seal and prevent solvent loss at high temperatures. The overall weight (mg) of the sealed vial containing the solvent, stirrer, and solid material was recorded to check for weight loss after the solubility measurements in the Crystal16 Multiple Reactor (Technobis Crystallization Systems, The Netherlands). This method uses the transmission of light through the vial as an indication of complete dissolution (100% transmissivity) or precipitation of the crystals (less than 100% transmissivity). To dissolve the particles in the suspension, the bottom stirrer at 700 rpm of stirring rate was applied together with 0.2 K/min of heating rate up to a pre-set high temperature. For the precipitation, the temperature was decreased with 0.4 K/min of cooling rate. The temperature was kept constant for 30 min at both the pre-set low and high temperatures to ensure complete dissolution (0% transmissivity). The selection of the Crystal16 parameters was based on previous published work by Samir A. Kulkarni.[133] The same parameters were employed in our study.

Solubility for each vial was calculated as an average of clear points (0% transmissivity) from all four cycles. For each solvent the solubility temperature was collected at four different concentrations and plotted in the van't Hoff coordinates (**Figure 17**). The Van't Hoff equation was then used for calculating the solubility of MFA in each solvent. Note that the Van't Hoff

equation for solubility calculation is from the linear correlation between ln(C_e) and 1/T(K^{-1}). For example, the solubility of MFA in methanol at 25 °C is calculated from the **Equation 5**.

$$\ln C_e = -3{,}578.48*\left(\frac{1}{25+273}\right) + 8.15 \qquad \textbf{Equation 5}$$

$$C_e = \text{Exp}(-3.86) = 0.02 \text{ mol/L}$$

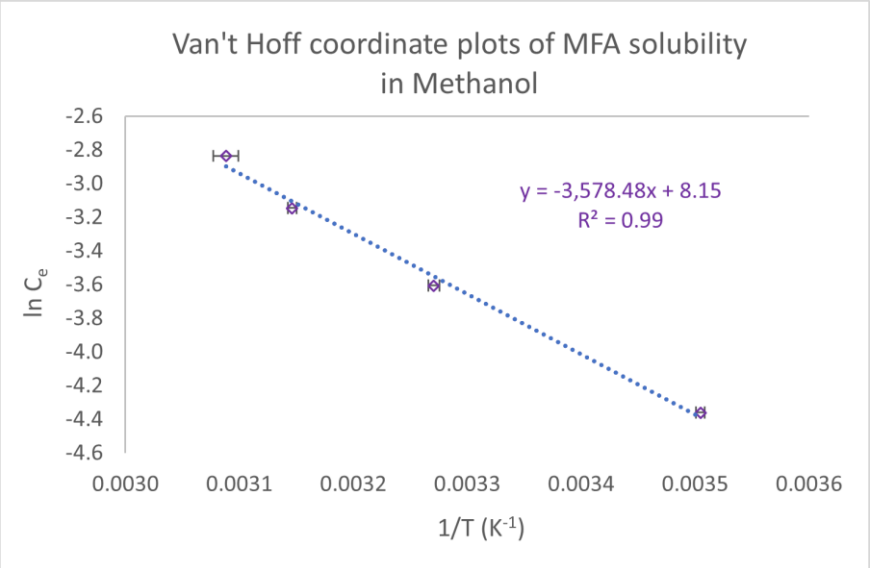


***Figure 17**. Van't Hoff coordinate plot of lnC vs 1/T(K^{-1}) with Van't Hoff equation for the solubility calculation of MFA in methanol*

3.3 Powder X-ray Diffraction (PXRD)

Powder X-ray Diffraction is a non-destructive technique for the analysis of crystalline materials. This technique has been used for the determination of unit cell dimensions and phase identification of crystalline materials. The repetitive arrangement of electron density within the crystalline materials causes the X-rays to diffract in various specific directions depending on crystal planes, resulting in a diffraction pattern which provides information on peak intensities and peak positions at specific 2-theta angles.[134,135]

When the X-ray radiation interacts with crystals, the secondary diffracted radiations from different crystal lattice planes can interfere with each other and cause either constructive or destructive interference. The constructive interference or reinforcement will occur when the path difference between the diffracted radiations is an integral number of the wavelength (**Figure 18**). These diffracted radiations are Bragg reflections. From this condition, the angle of incidence will be equal to the angle of reflection,[135,136] as shown in **Figure 19**.

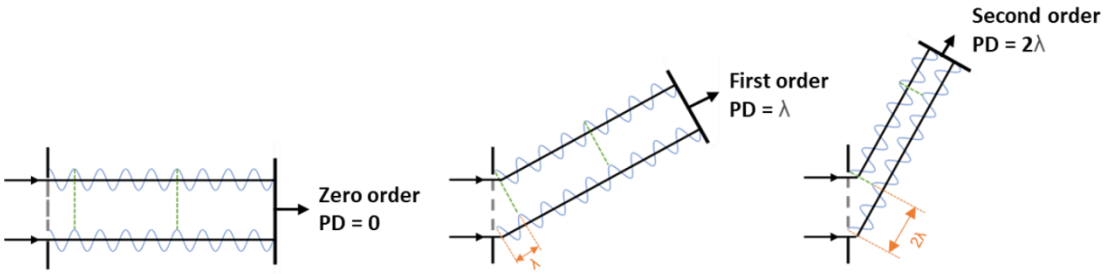


Figure 18. *The path difference (PD) between two diffracted waves shows the condition in which the reinforcement of the waves takes place when the path difference is an integral number of wavelengths.*

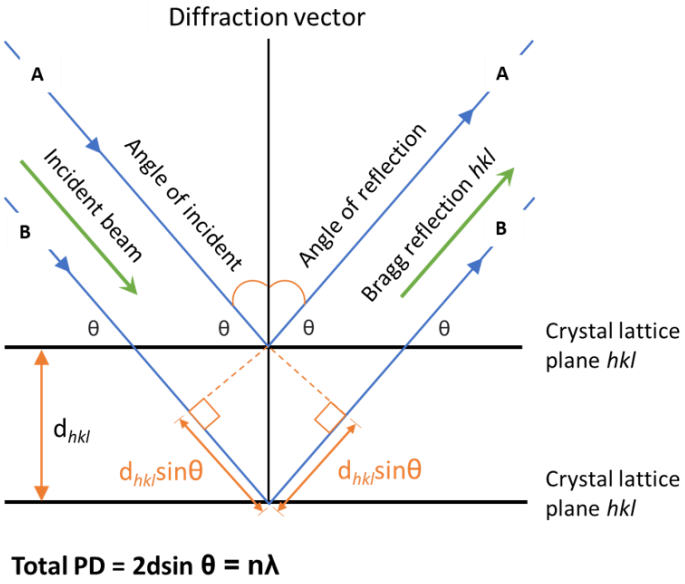


Figure 19. *The diffraction of X-rays by crystal planes in correspondence with Bragg's law*

Bragg's equation describes the relationship between the interplanar spacing in the crystal lattice and the path difference of diffracted radiations, see **Equation 6.**

$$n\lambda = 2d_{hkl}\sin\theta \qquad \textbf{Equation 6}$$

where λ is the wavelength of the incident X-ray, n is the reflection order, d_{hkl} is the interplanar spacing between hkl planes, and θ is the diffraction angle.[137]

Since the unit cell dimensions and Miller indices depend on the interplanar spacing d_{hkl}, Bragg's law is used for the interpretation of diffraction patterns by X-ray. A diffraction pattern is a unique pattern for a particular crystal structure and can be used for the identification of unknown crystal structures.[135,136] Peak positions in PXRD patterns are determined by the wavelength of the radiation, and lattice parameters of the material. Peak intensity is determined by (i) the structural factors of the atoms, (ii) the specimen factors such as grain

size and distribution, the microstructure of the sample, and (iii) instrumental factors such as the properties of the radiation, the type of focusing geometry, properties of the detector, slit and /or monochromator geometry.[135]

3.4 Random Forest Classification

The details of RF models (model structure, independent and dependent variables) will be specifically described in Chapter 5 and Chapter 6.

3.5 Model evaluation

3.5.1 Train-test split

Train-test split is a fast and easy method to evaluate the performance of a supervised machine learning algorithm. In this procedure, the dataset will be split into two subsets, in which one subset will be used to train the model. Then the prediction will be made for the other subset and the model predicted class will be compared to the actual one. These subsets are referred to as training datasets and test datasets, respectively. Train-test split is suitable when the dataset is of sufficient size to be split into two subsets while the test subset is still a good representation of the model's question. Additionally, the dataset should not contain the data with class-imbalance problem[138].

3.5.2 N-fold Cross-validation

The n-fold cross-validation is an out-of-sample evaluation method that uses all data in a dataset to evaluate a predictive model. A dataset will be split into equal-sized subsets. The number of the subsets can be customized, for example, 10-fold cross-validation splits the data into ten subsets. Then, one subset is selected as a test set and the remaining subsets are used as the training set. The process is repeated until all subsets are used as a test set. Since the whole dataset is used for training and testing the model, the bias from subset selection that might occur in the train-test split can be avoided.[139] **Figure 20** illustrates an example of 4-fold cross-validation.

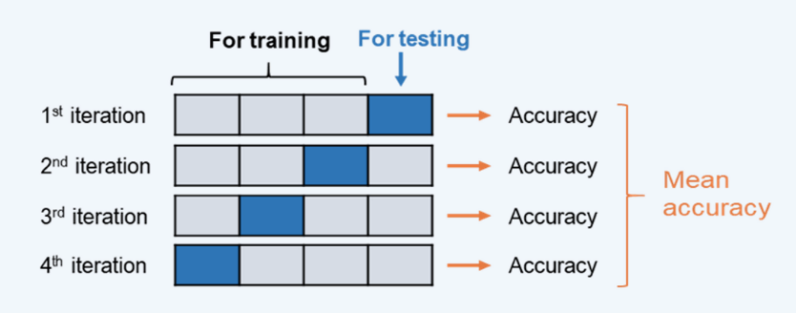


Figure 20. *An explanation of the 4-fold cross-validation method for model evaluation. A dataset was split into four equal subsets. Three subsets were used to train the model and the other subset was used to test the model. The same process was repeated 4 times, but different subsets were selected for testing the model. The model's accuracy was calculated from the average of accuracies from 4 iterations.*

3.5.3 *Accuracy, precision, recall, and F1-score*

Accuracy is a simple measure for evaluating a model. It is a ratio of the number of correctly predicted data and the total amount of data in a dataset, regardless of which class a data belongs to. However, to evaluate a model's performance, accuracy is not the only measure that should be considered. Precision, recall and F1-score are also important, especially in cases where the model prediction is very poor for some classes but better for others.[140]

Precision and recall are the measures for determining the number of correctly predicted data in a particular class. Precision compares the number of correct predictions with the total number of predictions for that class, while recall compares the number of correct predictions with the total number of data points that actually belong to that class.[141–143] **Figure 21** shows an example of a model prediction and **Equation 7-9** demonstrates the calculation of accuracy, precision, and recall, respectively.

		Predicted	
		Positive	**Negative**
Actual	**Positive**	True **positive**	False **negative**
	Negative	False **positive**	True **negative**

Figure 21. *An example of model prediction with two outputs, positive and negative*

$$\text{Accuracy} = \frac{True\ positive + True\ negative}{The\ total\ number\ of\ data} \qquad \textbf{Equation 7}$$

$$\text{Precision} = \frac{True\ positive}{True\ positive + False\ positive} \quad \text{: for class positive, or}$$

$$= \frac{True\ negative}{True\ negative + False\ negative} \quad \text{: for class negative} \qquad \textbf{Equation 8}$$

$$\text{Recall} = \frac{True\ positive}{True\ positive + False\ negative} \quad \text{: for class positive, or}$$

$$= \frac{True\ negative}{True\ negative + False\ positive} \quad \text{: for class negative} \qquad \textbf{Equation 9}$$

The other measure for model evaluation is F1-score, which is a good measure to determine a balance between precision and recall. Unlike an accuracy that does not concern the number of data belonging to each class, F1-score is useful when the model has biased class distribution. **Equation 10** shows how to calculate F1-score.

$$\text{F1-score} = 2\,x\,\frac{Precision\ x\ Recall}{Precision + Recall} \qquad \textbf{Equation 10}$$

The highest possible value of precision, recall, and F1-score is 1, which can occur when the model has 100% accuracy.[141,143]

3.6 Molecular descriptors

Molecular descriptors can be defined as the molecular features in numerical form which represent the structural and chemical characteristics of a molecule. These descriptors are meaningful values obtained from the transformation of chemical information by applying logical and mathematical approaches. The descriptors which are derived from solely the two-dimensional chemical structure, without considering 3D molecular conformation, are 2D molecular descriptors. Some examples of 2D molecular descriptors range from simple attributes such as the number of atoms, the number of internal bonds, or molecular weight, to the properties such as atomic polarizabilities, molecular mass density, and LogP value.[144–146] On the other hand, 3D molecular descriptors can be classified into 2 types, one depends on internal coordinates only and the other also depends on the absolute orientation of molecules.[146]

Molecular descriptors can be calculated from various software. One of those is Molecular Operating Environment (MOE), the integrated computer-aided molecular design platform that operates various useful tools for visualizations, simulations, modelling, and screening, including cheminformatics. Molecular descriptors calculated from MOE consist of two-

dimensional and three-dimensional molecular descriptors. two-dimensional descriptors were divided into 7 categories, namely 1.) physical properties, 2.) subdivided surface areas, 3.) atom counts and bond counts, 4.) Kier & Hall connectivity and Kappa shape indices, 5.) adjacency and distance matrix, 6.) pharmacophore feature, 7.) partial charge.[146,147]

4. How crystallisation thermodynamics affect the nucleation barrier

4 How crystallisation thermodynamics affect the nucleation barrier

4.1 Introduction

Solubility is one of the physical properties measured for the process development of new crystallization processes for pharmaceutical materials. The solubility of a compound across a range of temperatures provides a key baseline for selecting crystallisation conditions by enabling us to calculate the supersaturation at which crystallisation occurs and identify the driving force and ultimately the kinetics of the key crystallisation processes. Solubility data can also via application of van't Hoff analysis provide information on solution thermodynamics and solute-solvent interactions providing a deeper insight into the fundamentals of the solution chemistry and how this may impact crystallisation.[148,149]

Accessing a large number of solid-state forms consumes enormous effort due to the metastable or elusive nature of many forms and the rapid kinetics of the transformation process in solution. Polymorphs stability can often be described using Ostwald's rule of stages[150,151] whereby at high supersaturation, the most soluble (thermodynamically least stable) form nucleates first and goes through several transformations until the least soluble (thermodynamically most stable) form is produced. This implies that the rate of nucleation of the metastable form is always higher than that of the stable form and numerous metastable polymorphs should be observed before the most stable form is isolated. This rule has been a gold standard for chemists seeking previously unknown crystal forms. Limitations of the rule come from the fact that it is rather empirical and has no theoretical foundation[152] and is not universally obeyed.[153] For most compounds, we usually observe direct crystallisation to the most stable form, and kinetic forms can be isolated from specific solvents, e.g. for sulfathiazole, Ostwald's rule was observed during the cooling crystallisation in ethanol, but was not observed in n-propanol[154].

In thermodynamics, the universe is divided into two primary regions: the system and the surroundings. The system refers to the specific region of interest, such as a chemical reaction vessel, while the surroundings comprise the region outside the system and serve as the point of reference for measurements. To establish a thermodynamic model, it is assumed that the system is in thermal equilibrium, meaning that the temperature is constant throughout the

system and there is no net flow of heat. Additionally, the system is considered to be at constant temperature and pressure and is closed, meaning there is no exchange of mass between the system and the surroundings. This approach enables the study of the thermodynamic behavior of a specific system while isolating it from external factors, such as pressure changes, and is a fundamental assumption in thermodynamic analysis.[155] It is essential to note that this is an idealized scenario and in practice, the conditions of the external factors may vary, and mass exchange with the surroundings may occur. However, this simplifies the analysis and enhances our understanding of the system.

Thermodynamic principles can be used to predict the conditions under which a new crystal phase can form, however, they do not provide a comprehensive description of the complex process of crystal nucleation. The thermodynamic model does not take into account the kinetics of the nucleation process, such as the rate at which new crystals form, the size of the nuclei, or the shape of the crystals. Additionally, the thermodynamic model does not consider the behavior of individual molecules or atoms, instead, it assumes that the system is macroscopic and does not take into account the effects of non-thermodynamic forces, such as electric or magnetic fields. These considerations make it difficult to evaluate the role of impurities or defects in the nucleation process, which can greatly affect the kinetics and the final crystal structure.[155]

The crystallisation enthalpy ΔH^0_{cryst}, entropy ΔS^0_{cryst}, and free energy ΔG^0_{cryst} characterize the difference between the crystals and the solution. Since the crystals grown in all tested solvents belong to the same polymorphic form, the disparities of ΔH^0_{cryst}, ΔS^0_{cryst}, and ΔG^0_{cryst} in different solvents distinguish the state of the solute in each solvent. Conventionally, the crystallisation enthalpy ΔH^0_{cryst} can be measured calorimetrically by scaling the heat released during crystallisation (at constant temperature, T and pressure, p) with the crystallised mass.[156] For an alternative method, ΔH^0_{cryst} can also be determined from the solubility C_e of the crystals at different temperatures, using standard thermodynamics relations and $C_e(T)$. The equilibrium constant for the reaction molecule in solution ⇆ molecule in crystals is $K = C_e^{-1}$. Assuming that the respective activity coefficients are close to one due to the low C_e, and that the solution is ideal.[148] In this study, the crystallisation thermodynamics were determined from the solubility, C_e, because it requires less specialized equipment and can be performed using standard laboratory tools and techniques. Additionally, this approach is also more high-throughput, allowing for the

screening of multiple samples simultaneously. More details including the limitations of both approaches, as well as the comparison between the crystallisation enthalpy directly measured by calorimetry and that calculated from the determination of solubility as a function of temperature can be found in Appendix.

From a thermodynamic perspective, the change in Gibbs free energy resulting from crystallisation (ΔG^0_{cryst} or the free energy difference between 1 mole of the compound of interest in solution and 1 mole of the compound of interest in crystalline form) controls the energy barrier for crystallisation and, thus, should be related to nucleation rate (*J*). ΔG^0_{cryst} can be determined from the crystallisation enthalpy (ΔH^0_{cryst}) and crystallisation entropy (ΔS^0_{cryst}). ΔH^0_{cryst} and ΔS^0_{cryst} are derived from the dependence of equilibrium solubility on temperature determined using the van't Hoff equation. ΔH^0_{cryst} and ΔS^0_{cryst} complete the thermodynamic picture of solution thermodynamics and the associated solute-solvent interactions.[148,157] **Figure 22** illustrates the thermodynamic interactions and associated enthalpies and entropies in the crystallization process.

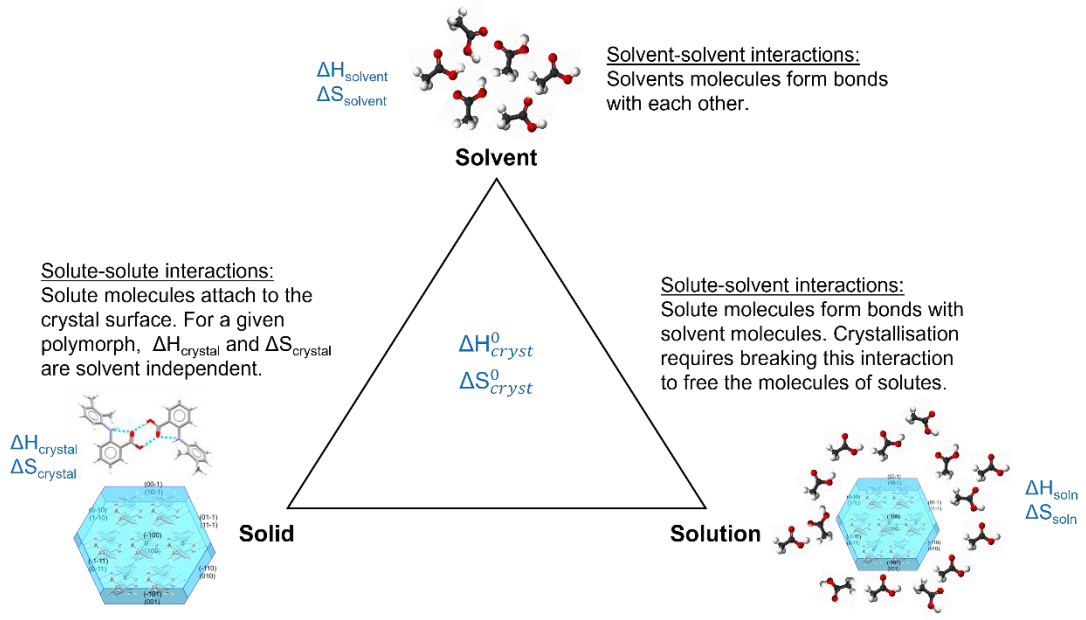


Figure 22. *The different types of molecular interactions occurring during crystallisation (solute-solute, solute-solvent, solvent-solvent) and associated enthalpies and entropies. In this thesis, ΔH^0_{cryst} and ΔS^0_{cryst} represent the enthalpy and entropy of the entire system composed of all of the interactions described in this figure.*

This chapter addresses the thermodynamic interactions that direct the crystallisation of MFA from different organic solvents. MFA form I was crystallised from thirty-two organic solvents

including alcohols (methanol, ethanol, 2-propanol, 1-butanol, 2-butanol, 1-octanol, 2-methoxyethanol), ketones (acetone, 2-butanone), acetates (methyl acetate, ethyl acetates, butyl acetate, isobutyl acetate), halogenated carbons (iodomethane, 1-chlorobutane, 1-bromobutane, 1-iodobutane, dichloromethane, 1,2-dichloroethane, chloroform, trichloroethylene), ethers (anisole, 1,4-dioxane, THF) and others (aniline, acetic acid, toluene, 1-methylnaphthalene, nitromethane, acetonitrile, diethyl sulfide) As an indication of the solute−solvent interactions, we employ the relative solution enthalpies (H_{soln})and solution entropies (S_{soln}) of MFA form-I in the thirty-two solvents, evaluated from ΔH^0_{cryst} and ΔS^0_{cryst}.

4.2 Mefenamic acid

Mefenamic acid (MFA, **Figure 23**) is a nonsteroidal anti-inflammatory that is widely indicated for pain related to menstrual disorders. MFA is classified in BCS class II indicating poor aqueous solubility but high permeability. It shows high hydrophobicity and propensity to stick to surfaces, which pose great problems during granulation and tabletting.[158] The molecule consists of the phenyl ring with the carboxyl group, connected to a twisted dimethyl-substituted phenyl ring by an imino bridge (torsion angle τ, **Figure 23a**) stabilised by a strong intramolecular hydrogen bond (N−H···O). MFA crystallises in three polymorphic forms (I, II, III). In all three known polymorphs, MFA forms a symmetric carboxylic acid dimer (**Figure 23b**). The main difference among the three polymorphs is the torsion angle (τ) of molecular conformation, which is ±120.0° in form I, ±68.2° in form II, and ±80.82° in form III (**Figure 23c**). Stable MFA form I (**Figure 23d**) crystallises in the most common solvents at ambient conditions. The metastable form II (**Figure 23e**) was reported in polymorphic transformation at a temperature above 500K,[159] high-pressure recrystallisation,[160] quenching cooling from DMF,[161] or SAM templates.[162] Forms I and II are enantiotropically related with a transition temperature of ca. 448K.[163] Form II is metastable and the rate of transformation to form I is sensitive to relative humidity (RH) and the solvent system, showing accelerated rates in the least polar mixtures.[164] Form III (**Figure 23f**) was found in the failed co-crystallisation with cytosine in a 1:1 DMF/methanol mixture and converts back to Form I immediately.[165] Form III is the least stable form of MFA at any conditions, hence, it will transform to Form I at ambient conditions or to Form II in the environment with high temperatures.

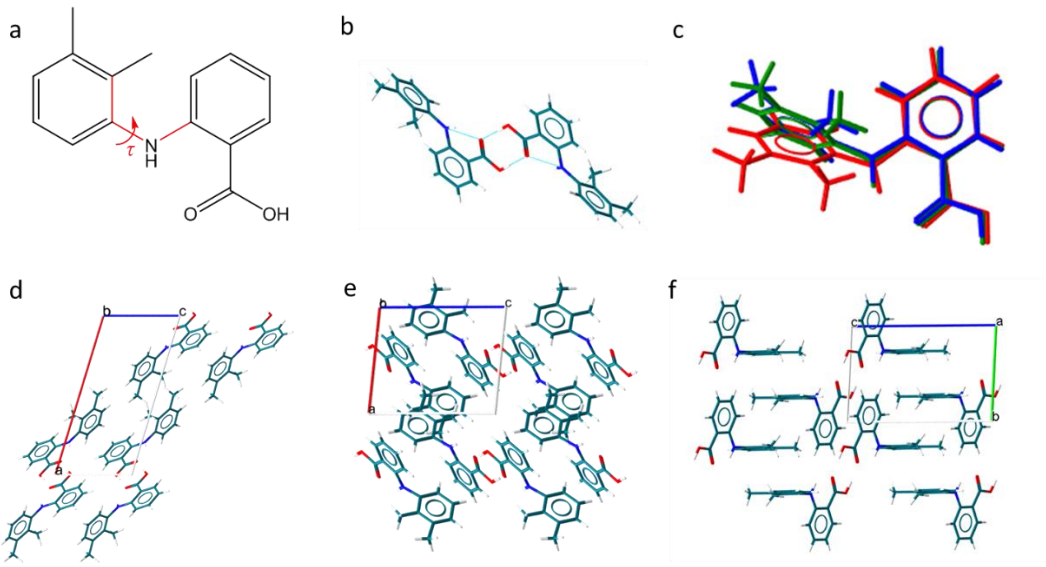


Figure 23 *Structure of MA. a) the molecular structure of MFA, b) MFA carboxylic dimer, c) the overlay of MFA molecular conformation in Form I (red, dihedral angle equal 120.0°, CCDC ref code XYANAC), Form II (blue, dihedral angle equal to 68.2°, CCDC ref code XYANAC07) and Form III (green, dihedral angle equal to 80.82°, CCDC ref code XYANAC03), the crystal structure of MFA d) form I, e) Form II, f) Form III.*

4.3 Methods

4.3.1 Induction Time Measurement

The nucleation rate was determined by the maximum likelihood estimation[166] measured by a Crystal 16 multiple reactor setup (Avantium, Amsterdam). The setup consists of 16 reactors to hold 1.5 mL vials and records the induction time by detecting the variation in light transmission of solution.

A 100 mL of stock solution was prepared by dissolving known amounts of MFA crystals in various solvents. After being dissolved at an elevated temperature, the solution was filtered through a 0.45 µm filter membrane. The 1.5 mL filtrate was then transferred to preheated vials. The vials were incubated at 323 K for 1 hour. Cooling crystallisation was carried out in feedback control mode. In this mode, the next cycle of crystallisation will be started when the % transmissivity of all samples in the same reactor reaches the target. The heating and cooling procedures were repeated five times to achieve a maximum of 80 sets of induction time data. In this experiment, the samples needed to be cooled to the set temperature as quick as possible so that the nucleation was detected at the desired crystallisation

temperature, while the heating rate should be slow to ensure that all the samples were redissolved before the next cooling step started. Similar parameters were selected following the previous work by Samir A. Kulkarni.[133] The heating and cooling rate were set at 0.5 K/min and 5 K/min, respectively (**Figure 24**). The bottom stirring speed was set constantly at 700 rpm throughout the entire process. Induction time was calculated as a difference between the time when the system reached desired temperature (298 K) and the time when the optical transmittance of the individual vials started to drop from 100%. Supersaturation is defined as $S = C/C_e$, where C is the solute concentration and C_e is the temperature-dependence solubility of the solute at 298 K.

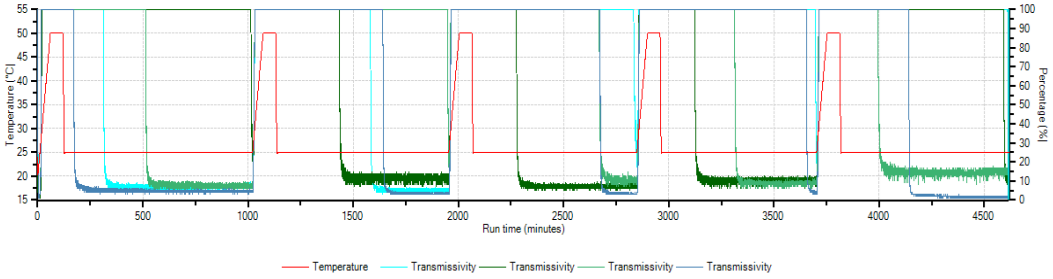


Figure 24 *Heating and cooling profile along with the %transmissivity of 4 individual MFA sample solutions. The transmissivity reached 100% when the solute was completely dissolved and started dropping from 100% when the crystal nuclei were detected. The time interval between the point where the process temperature reached the target (25 °C) and the point where %transmissivity started dropping was considered as an induction time.*

4.3.2 Nucleation rates estimation from probability distributions of induction times

Crystal nucleation is considered a stochastic process. Therefore, a Poisson distribution could be applied to calculate the probability of forming a particular number of nuclei within a specific time interval.[167]

$$P_m = \frac{N^m}{m!} exp(-N) \qquad \textbf{Equation 11}$$

Where P_m is the probability that m nuclei are formed in a time interval, and N is the average number of nuclei forming in the time interval.

The probability of the event where at least one nucleus was formed can be written as **Equation 12**.

$$P_{\geq 1} = 1 - P_0 = 1 - exp(-N)$$ **Equation 12**

Since N is the average number of nuclei forming during a time interval, it can be calculated from the nucleation rate J, solution volume V, and time interval t_j (**Equation 13**).

$$N = J V t_j$$ **Equation 13**

Since the Crystal16 can detect the appearance of crystal only once it grew to a measurable size, the time, t_j, obtained from the Crystal16 is the time since the first nuclei appeared (induction time; t) with the addition of the time until it grew to the detectable size (growth time; t_g). Accordingly, the probability P(t) of detecting crystals at time t can be determined by **Equation 14**.

$$P(t) = 1 - exp\,(-J\,V\,(t - t_g))$$ **Equation 14**

The nucleation rate can be determined from the variation of induction times from the same experimental setup (supersaturation, temperature profile, stirring rate, and solution volume) for small-scale crystallisations. The number of experiments per supersaturation was 80, which, according to the study of Jiang Shanfeng and Joop Ter Horst[43], was proved to be sufficient for determining nucleation rate. In their study, a series of 80 induction times were simulated for a known nucleation rate, growth time, and solution volume. The nucleation rate was then redetermined from the probability distribution of a random induction time series using **Equation 14** and compared to the known nucleation rate. It appeared that there is approximately 80% chance that the nucleation rates measured from this method were within the range of actual nucleation rates with 20% margin of error. This result showed that 80 induction times were sufficient to determine the nucleation rate for a given system.[43]

The induction time probability P(t) can be defined as the ratio of the number of experiments in which crystals are detected at time t ($M^+(t)$) and the total experiments (M) as described in **Equation 15**.

$$P(t) = \frac{M^+(t)}{Ms}$$ **Equation 15**

Figure 25 presents the variation of induction time from eighty crystallisation experiments carried out by Crystal16. The histogram in **Figure 25a** shows a large variation in induction time even though the experiments were done under the same crystallisation conditions (supersaturation, crystallisation temperature, stirring rate, and solution volume). The

induction time probability P(t) was further calculated by **Equation 15** and plotted as a function of the induction times in **Figure 25b**.

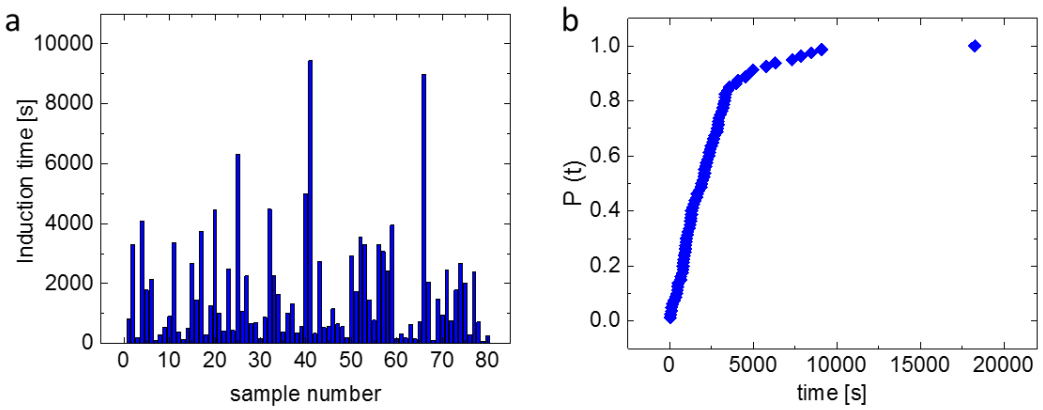


Figure 25 *Determination of the nucleation rate, a) a histogram showing a large variation in induction times from 80 crystallisation experiments of MFA solution under the same condition (supersaturation, crystallisation temperature, stirring rate, and solution volume), b) the induction time probability distribution.*

4.3.3 Powder X-ray diffraction for solid state identification

Powder X-ray diffraction pattern (PXRD) patterns were obtained using a Bruker AXS D8-Advance transmission diffractometer equipped with θ/θ geometry, primary monochromatic radiation (Cu, λ = 1.54056 Å). Data were collected in the 2θ range of 4−35° with a 0.015° 2θ step and 1 s/step speed. Reference powder patterns were produced using the Mercury 3.8 (CCDC) software from single-crystal data (CSD ref code: MFA form I: XYANAC, MFA form II: XYANAC02).

4.4 Results

4.4.1 Solution thermodynamics of MFA crystallisation

From thirty-two solvents selected for this study, MFA crystallised consistently as Form I. The PXRD patterns of MFA in studied solvents are presented in **Figure 26**.

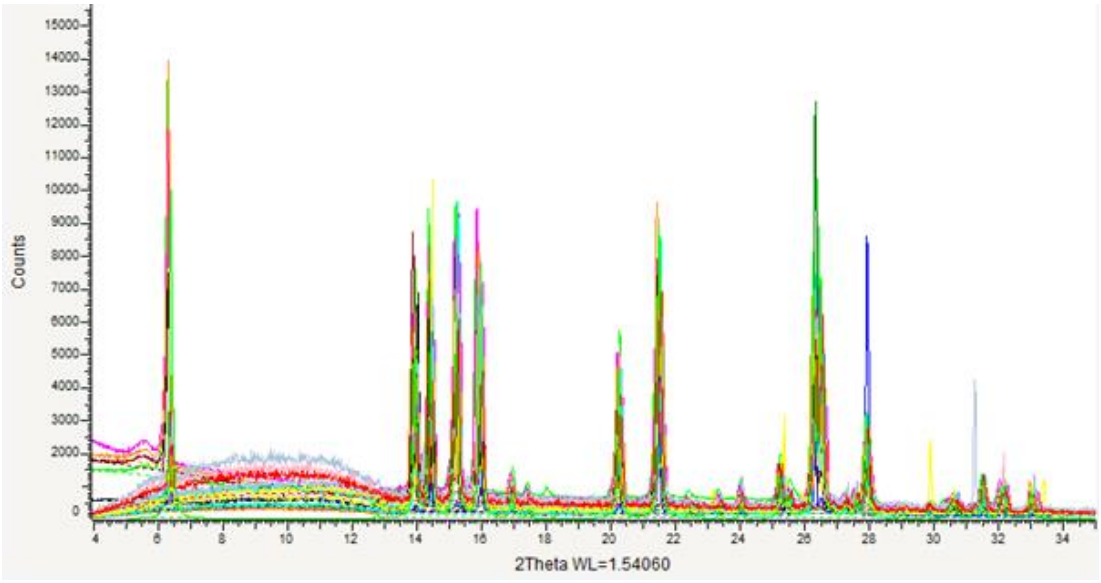
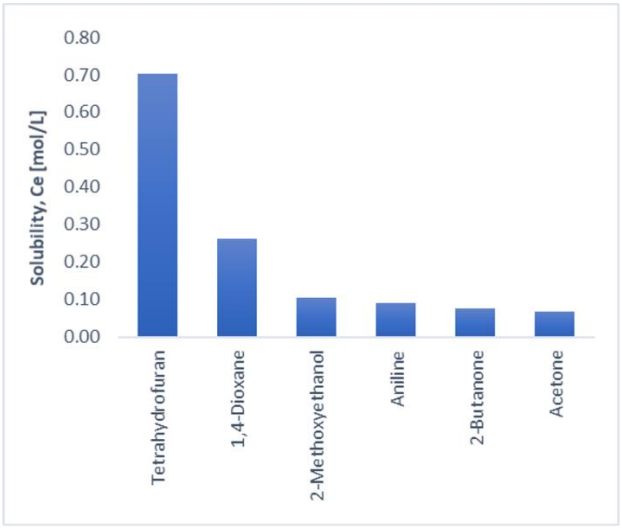
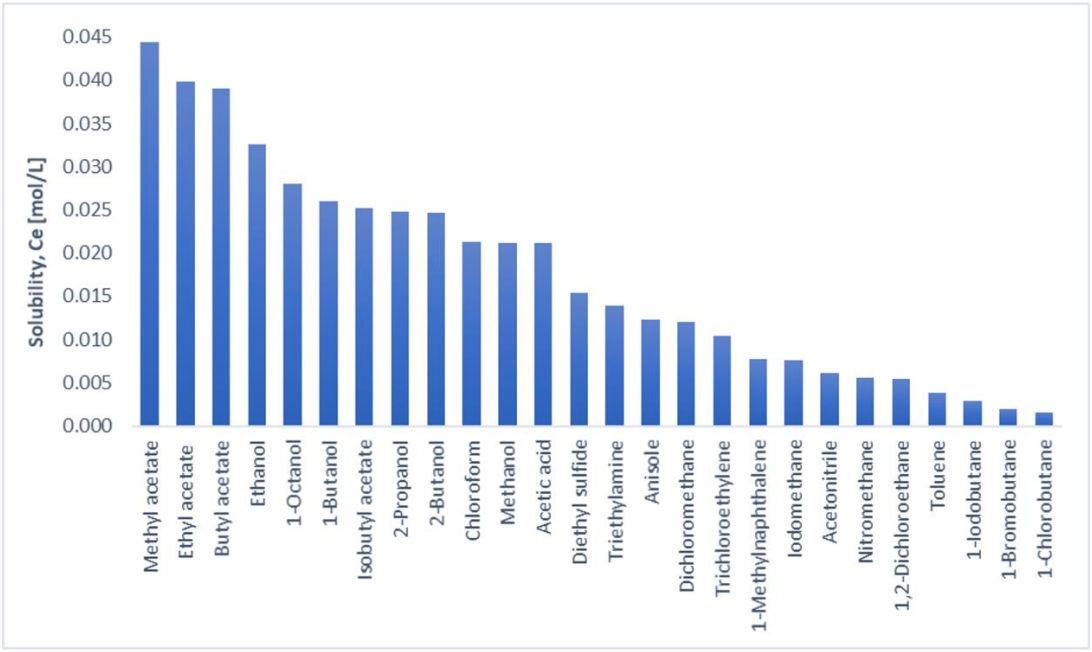


Figure 26. *Powder patterns of mefenamic acid crystals from the solvents studied in this work. All patterns corresponded to mefenamic acid form-I. Different colours represent MFA crystallised from different organic solvents.*

Overall, MFA Form I shows good solubility in the solvents containing the functional groups of ethers (tetrahydrofuran, 1,4-dioxane, and 1-methoxyethanol) and ketones (2-butanone and acetone). The highest solubility at 25°C was measured in THF (0.70 mol/L or 169.33 mg/mL), which can be categorized as freely soluble (100 – 1,000 mg/mL) according to the USP solubility guideline.[168] The solubility of MFA form I in all chlorinated hydrocarbons is low, between 0.0076 mol/L for iodomethane and 0.0015 mol/L for 1-chlorobutane. **Figure 27** presents the compared experimental solubility of MFA in thirty-two organic solvents.



(a)



(b)

Figure 27. *Experimental solubility of MFA form I at 25°C in various organic solvents calculated from Van't Hoff plot of lnC as a function of 1/T(K) (a) the solubility is higher than 0.05 M, (b) the solubility is lower than 0.05 M.*

For all solvents, the solubility curve was plotted to allow the extraction of the thermodynamic parameters crystallisation enthalpy (ΔH^0_{cryst}) and entropy (ΔS^0_{cryst}). **Figure 28** presents the dependence of MFA form I solubility on temperature and corresponding van't Hoff equations.

As with any process in nature at constant temperature and pressure, the transfer of solute molecules from solution to the crystal is governed by the change of Gibbs free energy,

ΔG^0_{cryst}.[148] According to the Gibbs–Helmholtz equation, the change in ΔG^0_{cryst} at constant temperature T can be stated as the net effect of the contributions of the enthalpy ΔH^0_{cryst} and entropy ΔS^0_{cryst} as:[148]

$$\Delta G^0_{cryst} = \Delta H^0_{cryst} - T\Delta S^0_{cryst} \quad \textbf{Equation 16}$$

If ΔG^0_{cryst} is negative, the process is thermodynamically favoured, which means that the system becomes more favourable as a consequence of crystallization (the energy is changed from high to low). To determine the ΔH^0_{cryst} in the solvents, the solubility of MFA at different temperatures $C_e(T)$ was employed. **Figure 28(a-d)** presented the quasi-exponential dependence of the solubility of MFA form-I on temperature T(K).

In the crystallisation equilibrium MFA(solution) ⇆ MFA(crystal), the equilibrium constant can be expressed as $K_{eq} = C_e^{-1}$.[149] While the associated crystallisation constant at equilibrium can be expressed as:[148]

$$K_{eq} = \exp\left(-\Delta G^0_{cryst} \Big/ RT\right) \quad \textbf{Equation 17}$$

Hence, the change in Crystallisation Free Energy ΔG^0_{cryst} can be expressed as:

$$\Delta G^0_{cryst} = -RT \ln K_{eq} = RT \ln C_e \quad \textbf{Equation 18}$$

where R is the universal gas constant, T is the temperature (K), and C_e is the solubility (M). To evaluate the crystallisation enthalpy ΔH^0_{cryst} of MFA form I, the van't Hoff relation plotted between ln C_e and T^{-1}(K) was determined. The equation can be expressed as:[148,149]

$$\frac{\partial ln C_e}{\partial (1/T)} = \frac{\Delta H^0_{cryst}}{R} \quad \textbf{Equation 19}$$

The van 't Hoff relation assumes that the slope of the correlation ln $C_e(T^{-1})$ is proportional to ΔH^0_{cryst}. The data of MFA solubility plotted in van't Hoff coordinates (**Figure 28e-Figure 28h**) indicate that ΔH^0_{cryst} is constant in the studied temperature range. Lastly, $\Delta S^0_{cryst} = (\Delta H^0_{cryst} - \Delta G^0_{cryst})/T$ and is proportional to the intercept of the plot of correlation ln $C_e(T^{-1})$.

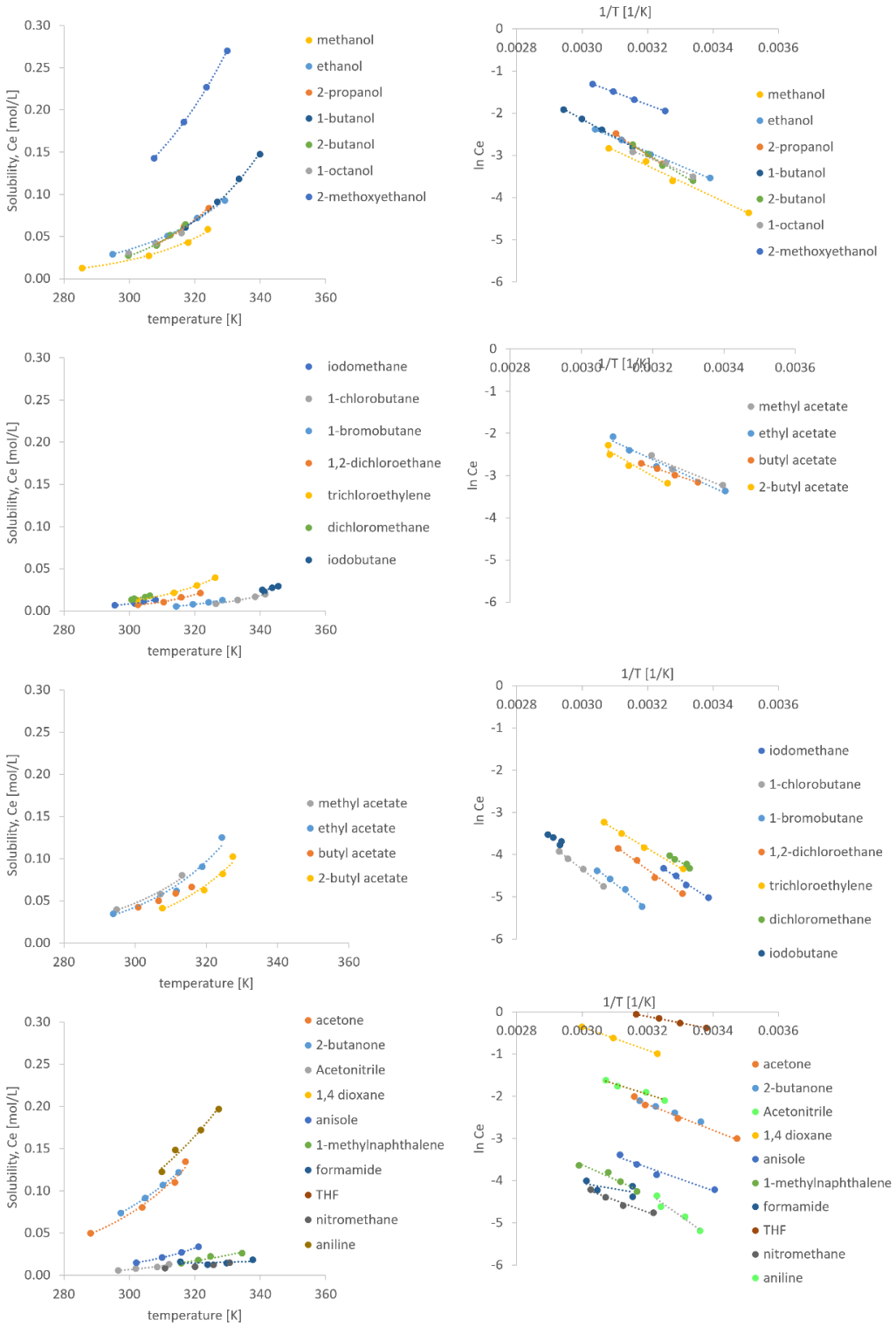


Figure 28 *Solubility of MFA Form I in various organic solvents. The legend shows the solvent in the studies, (a-d) The temperature dependence of the solubility C_e; dashed lines are polynomial fits, (e-h) MFA solubility plotted in van 't Hoff coordinates; dashed lines are linear regression fits. Note that some systems have only 3 points because there was an error during the experiment of 1 concentration, resulting in the high standard deviation of clear points from 4 cycles, so they were excluded.*

During the crystallisation process, the solute molecules in the saturated solution are incorporated into a crystal surface at kinks, which are the end sites of an unfinished crystal layer. In this study, MFA crystallised as the same polymorph (Form-I) in all tested solvents. Therefore, the same interactions were formed and the kinks of MFA crystals are identical, implying that the crystal enthalpy ($H_{crystal}$)and crystal entropy ($S_{crystal}$) are independent of the type of solvent. Subsequently, the differences in crystallisation enthalpy ΔH^0_{cryst} (**Equation 20**) and the differences in crystallisation entropy ΔS^0_{cryst} (**Equation 21**) between different solvents can reflect the solution enthalpy (H_{soln}) and solution entropy (S_{soln}) of MFA in each solvent.[148]

$$\Delta H^0_{cryst} = H_{crystal} - H_{soln} \qquad \textbf{Equation 20}$$

$$\Delta S^0_{cryst} = S_{crystal} - S_{soln} \qquad \textbf{Equation 21}$$

The values of ΔH^0_{cryst} and ΔS^0_{cryst} derived from the solubility data can be used to determine the strength of solute-solvent interactions in each solvent and link these parameters to the rate of nucleation of MFA in different crystallisation solvents.

The results showed that the obtained values of ΔH^0_{cryst} in all tested solvents are negative (**Figure 29a**), indicating the greater values of solution enthalpies (H_{soln}) over the enthalpy of the MFA form-I crystals ($H_{crystal}$). According to the **Equation 20**, $H_{soln} = H_{crystal} - \Delta H^0_{cryst}$, the respective H_{soln} will be low in the solvents with algebraically high ΔH^0_{cryst}, and the low value of H_{soln} suggested the strong solute−solvent interactions. In this study, the strongest interaction was found in THF and the weakest interaction was found in the group of solvents with halogenated structures.

The values of crystallization entropy ΔS^0_{cryst} are also negative in all solvents (**Figure 29b**). This result suggested that the overall degrees of freedom of the system were lost during the crystallisation. Additionally, similar trends between ΔH^0_{cryst} and ΔS^0_{cryst} observed in all tested solvents (**Figure 29** and **Figure 30**) suggest that the thermodynamics of MFA solutions were supported only by solute−solvent interactions but not by solvent-solvent interactions. This conclusion comes from the fact that, if there were strong solvent-solvent bonds formed as the layers of solvent molecules around the molecules of solutes, ΔS^0_{cryst} should notably increase during the process of crystallisation.[169,170] The breakage of strong interactions within the layer of solvent molecules when the solute molecules incorporate into the crystal would

significantly increase the crystallisation entropy while causing minimal impact on ΔH^0_{cryst}. This phenomenon arises because the energy of the solvent (related to the enthalpy) in the solvent layer is, numerically, almost equivalent to the energy of the free solvent.[148] For example, in aqueous solutions where the H-bond contribution increases, the value of ΔH^0_{cryst} and ΔS^0_{cryst} will no longer correlate with each other (**Figure 29** and **Figure 30**).[149] The observed correlation between ΔH^0_{cryst} and ΔS^0_{cryst} in this work is in agreement with the results of the work done by Rajshree and Peter, which focuses on the crystallisation thermodynamics of etioporphyrin I in five organic solvents: Dimethylsulfoxide (DMSO), octanol, hexanol, butanol, and caprylic acid.[148]

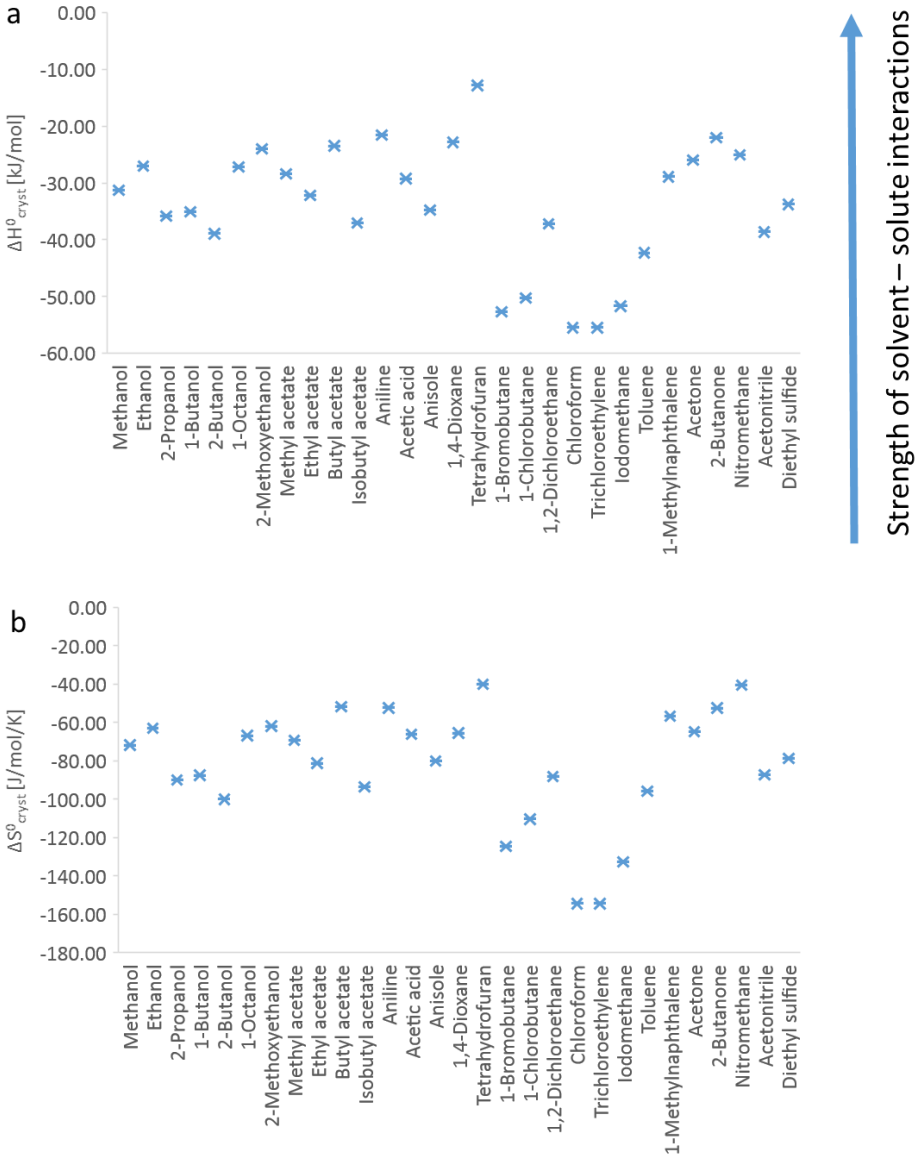


Figure 29 *Thermodynamic parameters of crystallization of MFA in different solvents. a) The crystallisation enthalpy, ΔH^0_{cryst} b) The crystallisation entropy, ΔS^0_{cryst}.*

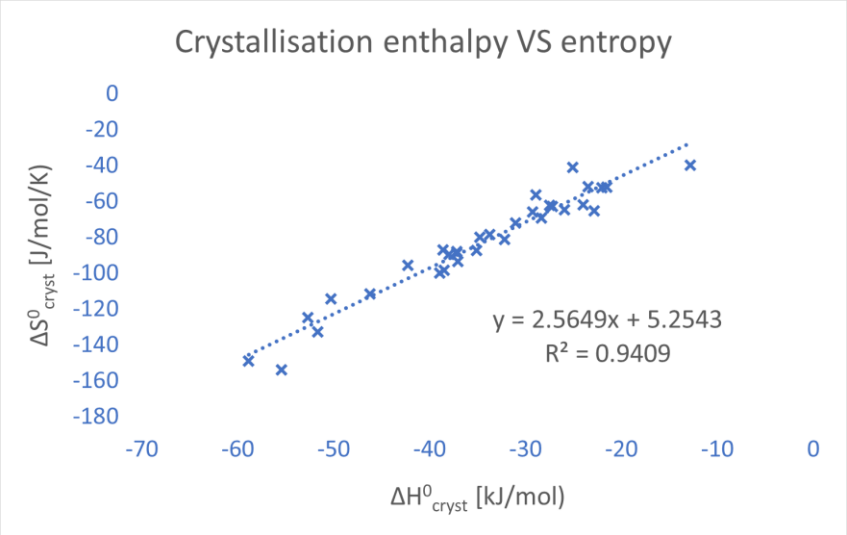


Figure 30. *The plot between ΔH^0_{cryst} and ΔS^0_{cryst} shows linear correlation.*

4.4.2 Nucleation rates of MFA in six different solvents

The crystal nucleation rate (J) is defined as the number of crystalline particles forming in the supersaturated solution per unit of volume and time. Measuring the rate of crystal nucleation is difficult because of the stochastic nature of the nucleation process. Additionally, the small size of the crystal nucleus is also a limitation for direct measurement of the first occurrence of crystal nuclei.[171] In this study, the nucleation rate of MFA was measured using a Crystal16 multiple reactor setup (Avantium, Amsterdam). This method uses transmission of the light that passes through the vial as an indicator for complete dissolution (100% transmissibility) of precipitation of crystals (less than 100% transmissibility). Six organic solvents, namely 2-butanol, 2-propanol, acetone, ethanol, methanol, and THF, were selected to be studied because four of them are alcohol with different numbers of carbon atoms (methanol, ethanol, 2-propanol, and 2-butanol), THF is the solvent in which MFA has the highest solubility, and acetone is a solvent commonly used in the crystallisation besides water and alcohols. The solution of MFA was filtered into 1.5-mL vials. The temperature was initially set up at 323 K and held for 30 minutes before instantly cooling to 298 K (5 K/min cooling rate). Induction time was defined as a time difference between the point where the temperature of the system reached 298 K and the point where the light transmissibility of the individual vials started to reduce from 100%. The total 80 induction times were measured from 16 vials that repeated 5 cycles of the same heating and cooling conditions. The induction time probability (P(t)) was calculated from those measured 80 induction times (see **Equation 15**), and the nucleation rate (J) was calculated from the probability distributions of induction times (see

Equation 14) by applying the maximum likelihood estimation method. **Figure 31** shows a typical induction time probability distribution of MFA crystals at four different supersaturations in six different solvents. After an initial time period in which no crystals were detected in all samples, the probability of induction time (P(t)) quickly increased and levelled off towards a probability of 1, especially in acetone solution. The possible explanation of this finding can be that acetone is a solvent with low nucleation barrier (according to parameter B showed in **Table 2**). Although THF has the lowest B value among six studied solvents, the viscosity of acetone is lower than that of THF (at 273.15 K, the viscosity of acetone and THF are 0.00389[172] and 0.114[173] Pa.s, respectively). Low viscosity promotes mass transfer of the solute in a supersaturated solution to the crystal nuclei, resulting in faster nucleation.[174,175] For all solvents studied, the shape of the curve correlated to the probability model of the single nucleus mechanism[171]. Heterogeneous nucleation cannot be completely ruled out since it is assumed that the crystal nucleation from solution usually occurs through heterogeneous nucleation on foreign particles such as dust particles or on interfaces such as a glass wall and the stirrer surface.

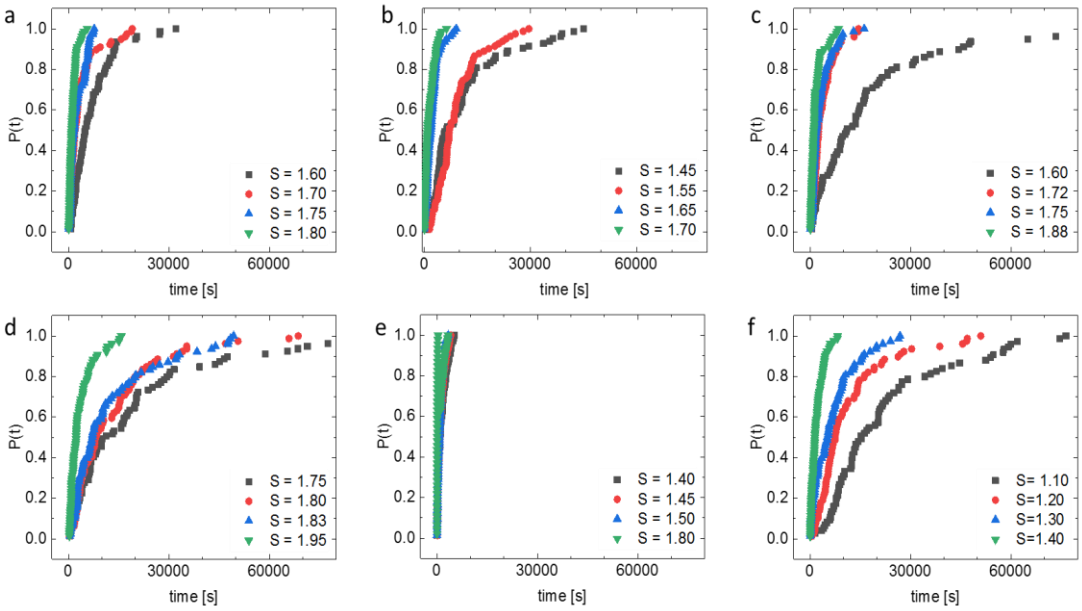


Figure 31. *Determination of nucleation rates at various supersaturations from the probability distributions of induction times (P(t)) of MFA in a solution of a) methanol, b) ethanol, c) 2-propanol, d) 2-butanol, e) acetone, and f) THF.*

4.4.3 Determination of thermodynamic and kinetic constants A and B

According to CNT, the nucleation rate can be described by **Equation 22**:

$$J = ASexp\left(\frac{-B}{ln^2 S}\right) \qquad \textbf{Equation 22}$$

in which A is the kinetic parameter, B is the thermodynamic parameter for nucleation, and S is the supersaturation calculated by the ratio of solution concentration and equilibrium concentration (C/C_e). The equation can be rearranged as:

$$ln\frac{J}{S} = \ln A - \frac{B}{ln^2 S} \qquad \textbf{Equation 23}$$

A plot of ln(J/S) versus $1/ln^2S$ (**Figure 32**) allows access to the kinetic parameter A from the intercept, while the thermodynamic parameter B can be derived from the slope. The values of nucleation rate J and supersaturation in the plot were shown in

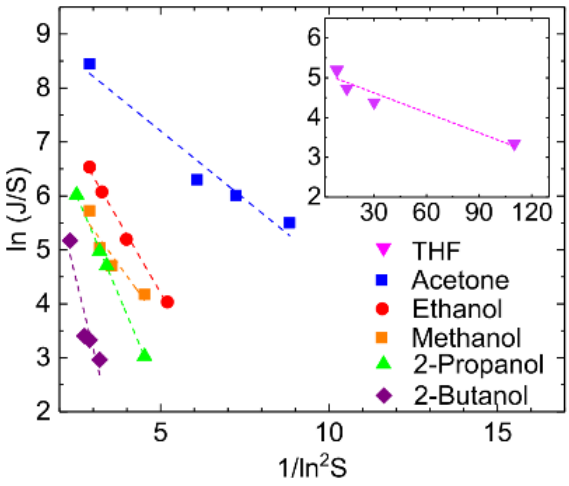


Figure 32 *Determination of A and B parameters from nucleation rates at different supersaturation for methanol, ethanol, 2-propanol, 2-butanol, acetone, and THF (inserted).*

Table 1**. The value of S, $1/ln^2S$, J, J/S, and ln(J/S) used for the plot in **Figure 32

Solvent	S	$1/ln^2S$	J	J/S	ln(J/S)
2-Butanol	1.75	3.19	33.9	19.4	3.0
	1.80	2.89	50.2	27.9	3.3
	1.83	2.74	55.1	30.1	3.4
	1.93	2.31	340.2	176.2	5.2
2-Propanol	1.60	4.53	33.0	20.6	3.0
	1.72	3.40	190.2	110.6	4.7
	1.75	3.19	253.0	144.6	5.0
	1.88	2.51	774.0	411.7	6.0
Methanol	1.60	4.53	104.0	65.0	4.2
	1.70	3.55	188.4	110.8	4.7
	1.75	3.19	270.1	154.3	5.0
	1.80	2.89	550.8	306.0	5.7
Ethanol	1.55	5.21	87.5	56.4	4.0
	1.65	3.99	297.4	180.2	5.2
	1.74	3.26	755.8	434.3	6.1
	1.80	2.89	1239.7	688.7	6.5
Acetone	1.40	8.83	443.3	316.7	5.5
	1.45	7.24	591.2	407.7	6.0
	1.50	6.08	698.5	465.6	6.3
	1.80	2.89	8464.3	4702.4	8.5
THF	1.10	110.08	31.0	28.2	3.3
	1.20	30.08	96.0	80.0	4.4
	1.30	14.53	147.0	113.1	4.7
	1.40	8.83	266.0	190.0	5.2

Parameters A and B for the nucleation of MFA crystals were determined in methanol, ethanol, 2-propanol, 2-butanol, acetone, and THF. The values of these two parameters are summarized in **Table 2**. The values of the obtained parameters are notably different between the six studied solvents.

4.4.3.1 Interpretation of thermodynamic parameter B

The thermodynamic parameter B of MFA nucleated from 2-butanol is highest, while the thermodynamic parameter B of MFA nucleated from THF is lowest. This result suggests that, at the same supersaturation level, the energy barrier for the nucleation of MFA in 2-butanol is much larger than the nucleation in the other solvents in this study, especially THF. The relation between the energy barrier of the nucleation (or nucleation work), W*, and the thermodynamic parameter B is given by **Equation 24**.[171,176]

$$\frac{W*}{k_BT} = \frac{B}{ln^2S} \qquad \textbf{Equation 24}$$

The parameter B reflects the energy barrier for forming the critical cluster which is the cluster in an equilibrium state to grow to become the crystal nucleus or to dissolve in the supersaturated solution. The value of thermodynamic parameter B is affected by different surface free energy (γ) of the crystal nucleus, which establishes the excess energy of the interface between a nucleus and the surrounding solution. This thermodynamic barrier along with the kinetics of the process in which the solute molecules attach to the critical cluster determines the rate of nucleation. In the case of a spherical cluster, the relationship between B and the surface free energy γ can be expressed by **Equation 25**.[171,177]

$$\gamma = \sqrt[3]{\frac{3B(k_B T)^3}{16\pi\Omega^2}} \qquad \textbf{Equation 25}$$

Where γ is a surface free energy of the nuclei, kB is the Boltzmann constant (1.38×10^{-23} J/K), T is a temperature (K), and Ω is a molecular volume, which is 315.83 $Å^3$ for MFA form I. The calculated values for γ are summarized in **Table 2**.

4.4.3.2 Interpretation of kinetic parameter A

While thermodynamic parameter B represents the energy barrier of the nucleation process, the pre-exponential factor A reflects the kinetics of the process in which the attachment and detachment of molecules into the clusters of crystal nuclei take place in the supersaturated solution. Similar to the thermodynamic parameter B, the value of kinetic parameter A in this study varied strongly depending on the crystallisation solvents. Parameter A of MFA nucleated in 2-butanol is two orders of magnitude higher than that in THF (**Table 2**).

In the equilibrium state where the concentration of the clusters is constant, the kinetic factor A can be described as **Equation 26**.

$$AS = zf^*C_0 \qquad \textbf{Equation 26}$$

From **Equation 26**, A is a pre-exponential factor which determines the rate of the attachment of building units onto the crystal at the nucleation site. This equation expresses that the pre-exponential factor A depends on the Zeldovich factor (z), the attachment frequency (f*), and the concentration of nucleation sites (C_0). The Zeldovich factor was first introduced by Becker and Döring, describing the probability of the clusters around the critical nucleus size redissolving rather than growing into macroscopic crystal nuclei.[178] The attachment frequency f* indicates the number of nuclei in the critical size (n*) that can grow further into supernuclei by the attachment of a building unit (n*+1).[171]

Theoretically, the kinetic factor A is estimated to be in the range between $10^{15} – 10^{25}$ $m^{-3}s^{-1}$.[176] However, the kinetic parameter A of the nucleation of MFA in the six studied solvents (**Table 2**) has significantly lower experimental values than the values estimated theoretically. The lower values of A observed in the experiment most likely results either from a lower than expected attachment frequency in the experiment, f*, or a lower than expected concentration of nucleation sites, C_0, in the experiment. A lower than expected concentration of nucleation sites, C_0, may result from the heterogenous particles presenting in the process. On the other hand, the particles may have a higher energy barrier for the attachment of growth units to the crystal nuclei than expected as this high energy barrier would lower the attachment frequency, f*. The work done by Davey et al, also found lower experimental values of A than estimates predict and came to similar conclusions as stated here.[42]

Table 2. *Summary of the kinetic and thermodynamic parameters for MFA in various solvents*

Solvent	B	A [$m^{-3}s^{-1}$]	ΔH^0_{cryst} [kJ/mol]	ΔS^0_{cryst} [J/mol.K]	ΔG^0_{cryst} [kJ/mol]	γ from Turnbull rule [mJ/m^2]	γ by Eq.24 [mJ/m^2]
THF	0.02	163.5	-12.8 (±1.6)	-40.1 (±1.7)	-0.9 (±1.7)	20.0	2.8E-07
Acetone	0.50	16,866.8	-26.0 (±1.7)	-64.8 (±5.4)	-6.7 (±0.2)	27.9	6.9E-06
Ethanol	0.63	5,259.6	-27.2 (±0.5)	-62.9 (±1.6)	-8.5 (±0.1)	29.1	8.7E-06
Methanol	0.86	2845.2	-31.0 (±0.5)	-71.8 (±7.7)	-9.6 (±0.1)	32.0	1.2E-05
2-Propanol	1.48	16,762.5	-35.4 (±0.9)	-89.9 (±2.8)	-8.6 (±0.1)	38.5	2.1E-05
2-Butanol	2.53	47,562.5	-39.0 (±0.7)	-99.9 (±2.1)	-9.2 (±0.1)	41.8	3.5E-05
Toluene*	2.71	361.4	-42.3	-95.7	-13.8	-	-

* Values from literature [179]

4.4.4 Thermodynamic parameters determine the kinetics of crystal nucleation

To understand how thermodynamic parameter relates to the crystal nucleation rate, the correlations between nucleation kinetics ΔH^0_{cryst}, ΔS^0_{cryst} and ΔG^0_{cryst} derived from the van't Hoff relationship and thermodynamic factor B were plotted (**Figure 33**). Initially, the main assumption in this study was that the nucleation rate should correlate to the Gibbs free energy of crystallisation, ΔG^0_{cryst}. Surprisingly, the best fit is obtained from ΔH^0_{cryst}, and ΔS^0_{cryst} (**Figure 33a** and **Figure 33b**, respectively) instead of ΔG^0_{cryst} (**Figure 33c**). This relationship is in an agreement with Turnbull's empirical rule for the estimation of surface free-energy, γ. This rule states that the surface free energy is proportional to ΔH^0_{cryst}, as expressed in **Equation 27**.[180]

$$\gamma = \frac{\alpha|\Delta H^0_{cryst}|}{\Omega^{2/3} \cdot N_A}$$ **Equation 27**

Where α is the scaling parameter accounting for the molecular environment in the crystal which generally values around 0.3 and N_A is Avogadro's number (approx. $6.023*10^{23}$). The surface free energy, γ, of MFA nucleation in six solvents in this study was calculated by Turnbull's rule (**Equation 27**) values between 20-40 mJ/m^2 (**Table 2**).

By combining **Equation 25** with **Equation 27**, the relation between thermodynamic parameter B and ΔH^0_{cryst} can be written as:

$$B = \frac{16\pi}{3} \frac{\alpha^3 |\Delta H^0_{cryst}|^3 \Omega^2}{(k_B T)^3 \cdot N_A^3}$$ **Equation 28**

From **Equation 28**, the thermodynamic parameter B is expected to be linearly proportional to $\left(\Delta H^0_{cryst}\right)^3$, as shown in **Figure 33d** – **Figure 33f**. Since ΔH^0_{cryst} can be extracted from the Van't Hoff plot of solubility of a given compound in a solvent, this linear correlation between $\left(\Delta H^0_{cryst}\right)^3$ and thermodynamic parameter B suggests that the solubility data can be used to reflect the energy barrier of the nucleation.

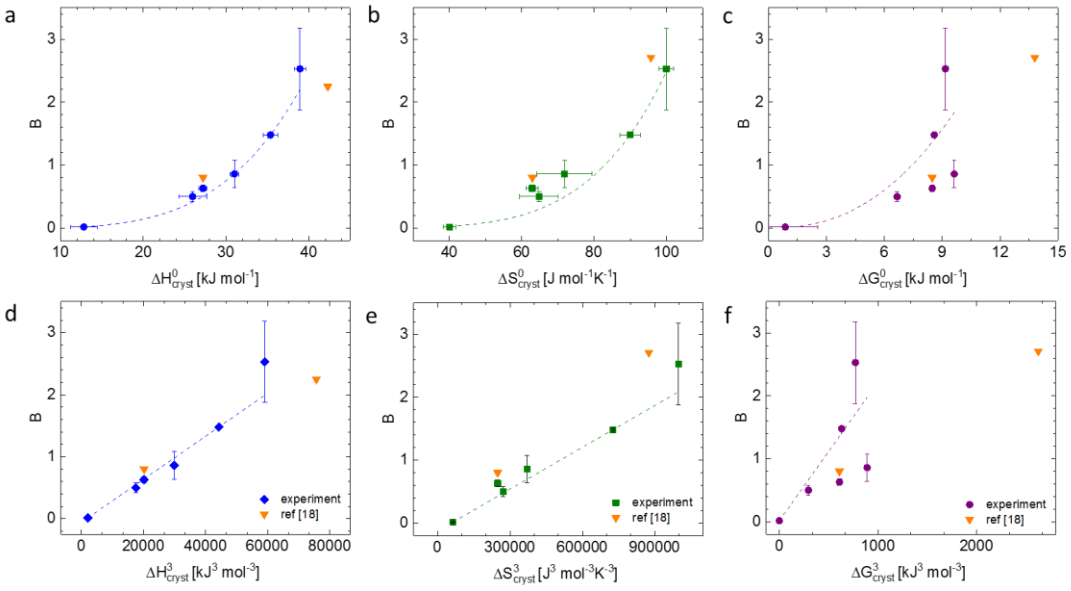


Figure 33 *Relationship between B and a) ΔH^0_{cryst}, b) ΔS^0_{cryst} c) ΔG^0_{cryst} and linear fit between B and a) ΔH^3_{cryst}, b) ΔS^3_{cryst} c) ΔG^3_{cryst}. The error bars on B values were determined from the standard error of the linear fitting (**Figure 32**) using Origin software. The error bars on ΔH^0_{cryst}, ΔS^0_{cryst}, and ΔG^0_{cryst} were determined from 3-4 seperated solubility measurements.*

4.4.5 Nucleation mechanism of MFA: CNT or Two-step Nucleation?

According to CNT, the crystals are formed directly from the individual molecules of solute in the supersaturated solution. Therefore, the crystal nuclei have an identical molecular arrangement to that of the new crystal phase, resulting in an equal surface free energy between the nuclei and the crystal interfaces.[177] CNT suggests that the nucleation barrier is proportional to the surface area of the crystal nucleus, and the systems found to be supported by CNT are mostly in relatively low supersaturation.[181] Whereas the two-step nucleation model assumes that the solute molecules aggregate into an intermediate metastable cluster before assembly into the crystal phase with the most thermodynamically stable conformation.[182] In two-step nucleation, the molecular arrangement in the clusters is different to the crystal phase, and the nucleation barrier is controlled by supersaturation and the surface free energy of the crystal nucleus surrounded by the precursor phase.[53,183–185] The surface free energy also influences the nucleation rate in the clusters.[42]

In the absence of any direct information on the pre-nucleation behaviour of MFA in solution, the nucleation mechanism can be determined by the comparison of the surface free energy γ calculated from the experimental data describing nucleation rate (**Equation 25**) to those calculated based on Turnbull's rule (**Equation 27**). The result shown in **Table 2** presents that the γ values of MFA nucleated in all solvents are much lower when the calculation is based on the observed experimental nucleation rate compared with those calculated from Turnbull's rule which is based on the solution thermodynamics. This finding suggests that the nucleation of MFA in these six studied solvents follows a two-step nucleation mechanism. The assumption is that, if the nucleation process follows CNT, the γ from the experimental nucleation rate should be more than 10^6-fold greater, or similar to the values obtained from the calculation using Turnbull's rule (see the differences between the values of γ calculated by two methods in **Table 2**). An increasing value of γ would dramatically heighten the nucleation barrier, and hence, decrease the nucleation rate by many orders of magnitude.[181]

4.5 Conclusions

The relationship between kinetic and thermodynamic parameters and crystal nucleation of MFA was studied in this chapter.

In this work, we studied ΔH^0_{cryst} which provides a measure of the heat released when one mole of solute crystallised from a supersaturated solution and ΔS^0_{cryst} which describes the degree of disorder or randomness of the crystallisation system. The results show that the

thermodynamic parameter B, which reflects the nucleation barrier, is linearly related to $\left(\Delta H^0_{cryst}\right)^3$ in accordance with Turnbull's rule. This observed correlation can be used to estimate the nucleation kinetics of organic crystals in various solvents based on their solution thermodynamics, and facilitate the rational selection of crystallisation solvents based on the strength of the interactions with the solute molecules expressed by thermodynamic parameters such as ΔH^0_{cryst} and ΔS^0_{cryst}. Additionally, the difference of surface free energy determined by nucleation rates to those calculated from Turnbull's rule suggests that the nucleation of MFA in six studied solvents (methanol, ethanol, 2-propanol, 2-butanol, acetone, and THF) occurs via two-step nucleation. Due to the smaller nucleation barrier, the rate of nucleation via the two-step nucleation mechanism is faster than the nucleation rate in CNT. Further investigations can be done by various analytical techniques to confirm the two-step nucleation in this study. For instance, dynamic and static light scattering techniques, which help us observe the size of solutes or aggregation in the solution, can be applied for the investigation of pre-nucleation clusters forming during the two-step nucleation.[58]

This information helps to understand the thermodynamic properties of the crystal formation process and the solute behavior in different solvents. The thermodynamic approach is important to predict the conditions under which new crystal phases can form.

5. Prediction of mefenamic acid crystal shape by random forest classification

5 Prediction of mefenamic acid crystal shape by random forest classification

5.1 Introduction

There is a considerable drive across the pharmaceutical industry to enhance the agility and productivity of activities involved in the development and manufacture of medicines.[186] Central interests focus on enabling faster, cost-effective drug production whilst improving sustainability and delivering improved security of supply whilst still assuring the quality and safety of medicines to patients.[187,188] Advanced particle formation and control is an area to address as this can also enable the disruptive benefits from more closely associated knowledge across drug substance and drug product manufacturing.[189] Cyber-Physical Systems embed Industry 4.0 principles and industrial digital technologies and realise benefits from digital design,[190] advanced process technology,[191] and data-driven manufacturing and control such as Digital Twins[192] or medicines development and manufacture that encompass the data, models, and knowledge that describe the inter-relationships between materials, products, processes, and performance.

Crystal shape is one of the important attributes dictating the physicochemical and bulk properties of a crystalline material, which can have an impact on the process-related characteristics as well as the quality attributes of the final formulated products.[193] Certain shapes of crystals are problematic during the key unit process used in the production of raw materials and downstream formulated product manufacturing. For example, needles can cause poor flowability of particulate solids and result in problems during various processes including powder flow,[194] filtering,[195] and tabletting[196]. Therefore, the ability to routinely predict the crystal shape yielded from a given solvent could improve efficiencies in process development and medicine manufacturing and reduce the costs of research and development.

Several theoretical models are already available for crystal shape i.e. geometrical morphology based on Bravais-Friedel-Donnay-Harker (BFDH) theory,[197] growth morphology based on an attachment energy calculation, the theory of Hartman-Perdok[198] or periodic bond chain (PBC).[199] Experimental results often vary from theoretical predictions due to the influence of solvent,[87,200] impurities,[201] and additives[193] in the crystallisation medium, and

although progress has been made in the prediction of morphologies,[202,203] there is a need for new models that can provide practically useful, rapid prediction across a wide range of potential crystallisation environments.

In the field of crystallisation, data-driven approaches using machine learning can be powerful tools for finding relevant patterns in high-dimensional data. During the past few years, several machine learning studies showed great promise and lead to the successful discovery of novel crystal forms[116] and the successful prediction of the small molecule crystallisability,[204] crystal packing,[29] polymorphism, and co-crystallisation.[205]

In this work, the crystal shape prediction of MFA in different solvents was investigated. MFA (2-[(2,3-Dimethylphenyl)amino]benzoic acid, $C_{15}H_{15}NO_2$) is a high-dose analgesic drug in the non-steroidal anti-inflammatory (NSAIDs) group. It is widely used for the treatment of mild to moderate pain due to menstruation (primary dysmenorrhea).[206–208] It is classified as a compound in class II based on the biopharmaceutical classification system (BCS) which indicates low aqueous solubility with high permeability.[209,210] Apart from the solvated form, MFA has 3 different solid-state forms, which are forms I, II, and III.[211] During manufacturing, MFA often causes problems in processes such as granulation and tabletting because of its hydrophobicity and tendency to stick to surfaces that result from the specific crystal surface chemistry expressed. MFA is therefore a useful example to illustrate the impact of crystal shape during drug manufacturing[212,213] and to explore the prediction of solvent effects on crystal shape to inform subsequent process development and engineer the bulk properties of active pharmaceutical ingredients. Control of shape through appropriate particle engineering strategies can also allow the avoidance of additional downstream processing steps such as milling.

A variety of crystal shapes have been reported from prior experimental studies for MFA, ranging from plate-like to needle-like crystals.[159,214,215] Plates or elongated crystals of MFA were observed when crystallised from tetrahydrofuran,[159] ethanol,[216] ethyl acetate,[159,212] dimethylacetamide (DMA),[215,217] and isopropanol,[218] while needle-like crystals were often observed when MFA was crystallised from acetone.[218,219] However, many studies of the crystallisation of MFA have yielded different results for crystal shape despite using the same crystallisation solvent. For example, the crystallisation of MFA from ethyl acetate carried out by Mudalip et al. produced needle-like crystals,[215] while the SEM pictures of MFA crystallised from ethyl acetate showed plate-like crystals in the study of Panchagnula et al.[159] The latter

study has also shown that the shape of MFA crystal grown from tetrahydrofuran and ethyl acetate changed as supersaturation levels changed.[159]

Previously, a random forest (RF) algorithm has been applied to predict the crystallisation outcomes.[119,120] From these studies, RF performed as well as or better than other algorithms, such as support vector machines (SVM),[119,120] neural networks,[119] and deep learning multilayer perceptron networks.[120]

RF has advantages over other algorithms including SVM or k-nearest neighbours which generally are more sensitive to data outliers. On the other hand, RF is robust to the outliers since its prediction relies on the averaged output from multiple independent decision trees.[112] This attribute of RF algorithm also provides the low risk of over-fitting to training data.[220] Additionally, RF also provides us with the relative ranking of variable importance which can be used to guide a feature selection and support model interpretability.[221] Therefore, in this work, we applied RF classification to predict the crystal shape of MFA as a function of recrystallisation solvent. MOE molecular descriptors were used for 30 solvents and three different sets of variables (one set that contained all available 2D descriptors, a second set that focused on molecular structure and a third set that focused on physical properties) were tested to optimise model performance. To identify which solvent descriptors were associated with RF model performance, logistic regression was applied, and variable coefficients, as well as recursive feature elimination, were considered. Powder X-ray Diffraction (PXRD) for solid-state determination and Differential Scanning Calorimetry (DSC) for thermal analysis was carried out for crystallisation from solvents which resulted in poor model performance.

5.2 Materials and methods

Materials. MFA (>98% purity) was purchased from Merck (UK). All solvents were purchased from Fisher Scientific (UK).

5.2.1 Cooling crystallisation

Small-scale crystallisation was carried out in 20-mL scintillating vials. Appropriate amounts of MFA powder and organic solvent, as determined by the solubility experiments, were transferred into the vials. The vials were capped and covered with parafilm to avoid solvent evaporation. Vials were heated using a hot plate until all solid had visibly dissolved. To ensure no solid remained, the solution was then filtered through 0.45 µl PTFE filter discs into a clean

vial. The vials were capped and placed in an incubator at 25 °C without disturbance for 5 days. All samples were prepared in different solvents at various supersaturations for comparison.

5.2.2 *Optical microscopy*

An optical microscope (Leica M165C, supplied by Leica Microsystems (UK) Ltd.) was used for capturing two-dimensional images of the resulting crystals. Without removing the remaining solvent, the crystal samples obtained from cooling crystallisation were observed in the same container under the bright-field mode of optical microscope. The crystal shapes were manually classified into two classes: polyhedral and needle. Polyhedral crystals were comprised of any crystals with regular bounding facets including shapes such as prisms, plates and elongated crystals. Needles were defined by any sample with elongated crystals with no discernable edges or faces. Note that polyhedral and needle crystals were introduced with broad definitions for the practical implications for downstream pharmaceutical manufacturing processes as needle-shaped crystals are more likely to cause issues during manufacturing than crystal shapes with aspect ratios closer to 1, and so are generally undesirable. Any spherulitic crystals were classed as needle crystals as they were a form of needle crystal aggregates.[62] Example images of different crystal shapes from our dataset can be seen in **Figure 34**.

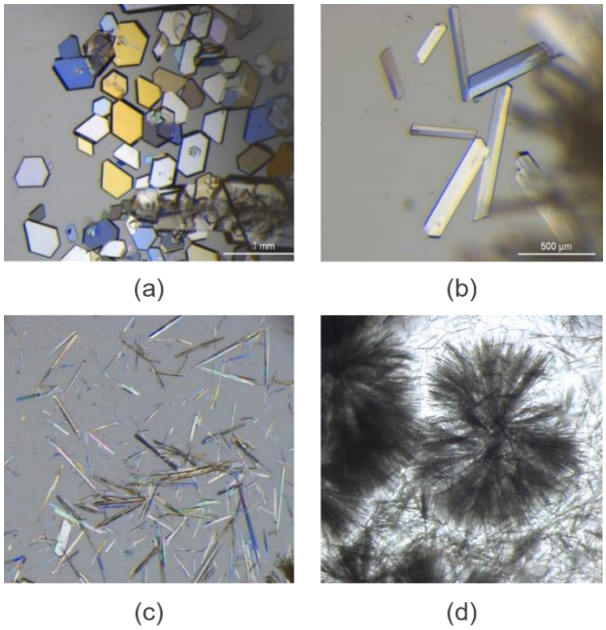


Figure 34. *Examples of crystal shapes: (a) plates, (b) elongated plates, (c) needles, and (d) spherulites. Plate and elongated plate crystals were assigned to the polyhedral class while needle and spherulitic crystals were both assigned to needle crystals.*

5.2.3 *X-ray diffraction data*

X-ray diffraction aims to determine the solid-state form (polymorphism) of MFA crystallised in each organic solvent. For face indexing, single crystal X-ray diffraction (SC-XRD) was performed using D8 Venture (Bruker UK Limited), equipped with Photon III CCD detector and Cu (Copper) Kα X-ray energy source which corresponds to an x-ray wavelength of 1.5406 Å. A single crystal was prepared and fixed onto a low diffraction loop connected to a three-circle fixed Chi goniometer. APEX3 software was used to specify the faces of a single crystal. For powder X-ray diffraction (PXRD), the data were collected from 4° to 35° 2-theta (step size 0.017°) for all samples at ambient temperature. PXRD on triethylamine samples was repeated in a capillary set up and the data was collected from 3° to 40° 2-theta.

5.2.4 *Random forest predictions*

Random Forest (RF) classification (Random Forest Classifier in Scikit-learn 1.0.2, Python 3.10) was applied to all models as RF have been shown to be effective for the prediction of crystallisation outcomes in previous works.[31,114,204] The number of decision trees was set at 100 by setting parameter n_estimators = 100 and the random state was set at 0. Other parameters were used as default values (bootstrap = True, max_depth = None, max_features = auto, max_leaf_nodes = None, min_samples_leaf = 1, min_samples_split = 2).

The justifications for selecting each parameters are provided as follows:

- n_estimators: This parameter controls the number of trees in the random forest model. The higher the number of trees, the more complex the model is.[222] Increasing n_estimators can improve the model performance but also increase the computational time. Moreover, too complicated models may end up causing overfitting, resulting in the model being less efficient to external data. Setting n_estimators = 100 is considered a good trade-off between the model performance and computational time.[223]
- random_state: This parameter is useful for the reproducibility of running the random forest model as it specifies a particular random subset of the input data to the model. 0 is the common choice for this parameter.[223]
- bootstrap: By setting bootstrap = True, bootstrap sampling technique will be applied. Bootstrap sampling randomly selects the input data from the original dataset with replacement for building each tree in the random forest, promoting randomness to the model and cause the trees to be more independent from each other.[223]

- max_depth: By setting max_depth = None, the depth of the tree structure is unrestricted, which makes the splitting process continue until the number of samples at a leaf node = 1 (as min_samples_leaf is also set as 1).[223]
- max_features: By setting max_features = auto, the number of features (input variables) used for considering the best split will be determined by the square root of the total number of features in the dataset. This setting is commonly used as it is a default value of Random Forest Classifier in scikit-learn that can work well in practice.[223]
- max_leaf_nodes: By setting max_leaf_nodes = None, the number of leaf nodes in each decision tree is unrestricted, which makes the trees can grow as large as possible to fit the data or until the number of samples at a leaf node = 1 (as min_samples_leaf is also set as 1).[223]
- min_samples_leaf and min_samples_split: By setting min_samples_leaf = 1 and min_samples_split = 2, the node will be splitted until the number of samples at a leaf node = 1 (impurity = 0).[223]

5.2.4.1 Building models

Experimental solubility of MFA at 25°C, supersaturation levels, 2D MOE solvent molecular descriptors [224], solvent boiling point and melting point, and solvent density[225] were included in the dataset as input for training predictive models. Each experiment in the dataset was labelled with the crystal shape outcome. MOE descriptors used in this work were calculated from molecular structures using SMILE codes. After data cleaning by removing the descriptors with NaN value (missing data) and the descriptors which contain the same value for all solvents, 206 descriptors were left in the dataset (see **Table 3** for details of descriptors).

Table 3. *2-D molecular descriptors calculated from MOE*

Descriptors	**Category**	**Description**
2-D descriptors		
apol, bpol, Fcharge, mr, SMR, Weight, logP (o/w), SlogP, vdw_vol, density, vdw-area	physical properties	Physical properties are calculated from the connection table of a molecule
SlogP_VSA0-SlogP_VSA9, SMR_VSA0 - SMR_VSA7	subdivided surface areas	The Subdivided Surface Areas are descriptors based on an approximate accessible van der Waals surface area calculation for each atom, v_i along with

		some other atomic property, p_i.
a_aro, a_count, a_heavy, a_ICM, a_IC, a_nH, a_nB, a_nC, a_nN, a_nO, a_nF, a_nP, a_nS, a_nCl, a_nBr, a_nI, b_1rotN, b_1rotR, b_ar, b_count, b_double, b_heavy, b-rotN, b_rotR, b_single, b_triple, VAdjMa, VAdjEq	atom count and bond count	The atom count and bond count descriptors are functions of the counts of atoms and bonds
chi0, chi0_C, chi1, chi1_C, chi0v, chi0v_C, chi1v, chi1v_C, Kier1 - Kier3, KierA1 - KierA3, KierFlex, zagreb	Kier&Hall Connectivity and Kappa Shape Indices	The Kier and Hall kappa molecular shape indices compare the molecular graph with minimal and maximal molecular graphs and are intended to capture different aspects of molecular shape.
balabanJ, diameter, petitjean, radius, VDistEq, VDistMa, weinerPath, weinerPol	Adjacency and Distance Matrix Descriptors	The adjacency matrix, M, of a chemical structure is defined by the elements [Mij] where Mij is 1 if atoms i and j are bonded and zero otherwise. The distance matrix, D, for a chemical structure is defined by the elements [Dij] where Dij is the length of the shortest path from atoms i to j; zero is used if atoms i and j are not part of the same connected component.
a_acc, a_acid, a_base, a_don, a_hyd, vsa_acc, vsa_acid, vsa_base, vsa_don, vsa_hyd, vsa_other, vsa_pol	Pharmacophore Feature Descriptors	The Pharmacophore Atom Type descriptors consider only the heavy atoms of a molecule and assign a type to each atom
Q_PC+ PEOE_PC+, Q_PC- PEOE_PC-, Q_RPC+ PEOE_RPC+, Q_RPC- PEOE_RPC-, Q_VSA_POS PEOE_VSA_POS, PEOE_VSA_NEG, PEOE_VSA_PPOS, PEOE_VSA_PNEG, PEOE_VSA_HYD, PEOE_VSA_POL, PEOE_VSA_FPOS, PEOE_VSA_FNEG, Q_VSA_FPPOS PEOE_VSA_FPPOS, Q_VSA_FPNEG PEOE_VSA_FPNEG, Q_VSA_FHYD	Partial Charge Descriptors	Descriptors that depend on the partial charge of each atom of a chemical structure require calculation of those partial charges.

PEOE_VSA_FHYD, Q_VSA_FPOL PEOE_VSA_FPOL, PEOE_VSA+6 - PEOE_VSA+0, PEOE_VSA-0 - PEOE_VSA-6		

From this dataset, 87 models were built to assess the optimum performance for predicting crystal shape.

Model 1 used the entire dataset for 3-class prediction as follows: polyhedral (134 observations), needle (83 observations), and no crystal (44 observations). The class of no crystal was then removed from the datasets for all remaining models due to the relatively low occurrence of this outcome. As class imbalance was present in the dataset used for Model 2, some observations in the polyhedral class were removed from the dataset used in Model 3. In this step, some observations were removed to maintain the spread of original data (i.e. data points for solvents with low numbers of observations were kept in the dataset while some data points for the solvents with higher numbers of observations were removed) rather than random selection. The numbers of observations in the dataset used for Models 1 – 3 are shown in **Table 4**.

Table 4. *Numbers of observations in the dataset used for training and testing each predictive model*

Model	**Number of observations**			
	Polyhedral class	**Needle class**	**No crystal class**	**Total**
Model 1	134 (51.3 %)	83 (31.8 %)	44 (16.9 %)	261
Model 2	134 (62.0 %)	82 (38.0 %)	-	216
Model 3	82 (50.0 %)	82 (50.0 %)	-	164

From 206 descriptors, feature selection was applied to the final 84 models to investigate if different sets of solvent molecular descriptors (one set that contained all available 2D descriptors, a second set that focused on molecular structures and a third set that focused on physical properties) would affect model performance. The details of the selected descriptors were listed in **Table 5**.

Table 5. *The list of physical properties, atom counts and bond counts, and pharmacophore feature solvent descriptors with codes and descriptions*

Category	Descriptors	Descriptions
Molecular structure and connectivity (second set of descriptors)	a_aro	Number of aromatic atoms
	a_count	Number of atoms
	a_heavy	Number of heavy atoms
	a_nH, a_nC, a_nN, a_nO, a_nS, a_nCl, a_nBr, a_nI	Number of hydrogen, carbon, nitrogen, oxygen, sulfur, chlorine, bromine, iodine atoms
	b_ar	Number of aromatic bonds
	b_count	Number of bonds
	b_heavy	Number of bonds between heavy atoms
	b_rotN	Number of rotatable bonds
	b_single, b_double, b_triple	Number of single, double, and triple bonds
	chiral	Number of chiral centres
	opr_brigid	Number of rigid bonds
	rings	Number of rings.
	a_acc	Number of hydrogen bond acceptor atoms
	a_acid	Number of acidic atoms
	a_base	Number of basic atoms
	a_don	Number of hydrogen bond donor atoms
	a_hyd	Number of hydrophobic atoms
Physical properties (third set of descriptors)	apol	Sum of the atomic polarizabilities
	bpol	Sum of the absolute value of the difference between atomic polarizabilities of all bonded atoms in the molecule
	density	Molecular mass density
	mr, SMR	Molecular refractivity
	weight	Molecular weight
	logP(o/w), SlogP	Log of the octanol/water partition coefficient
	logS	Log of the aqueous solubility (mol/L)
	reactive	Indicator of the presence of reactive groups
	TPSA	Polar surface area
	vdw_vol	van der Waals volume
	vdw_area	Area of van der Waals surface

The different considerations and test criteria used for all models are shown in **Figure 35**.

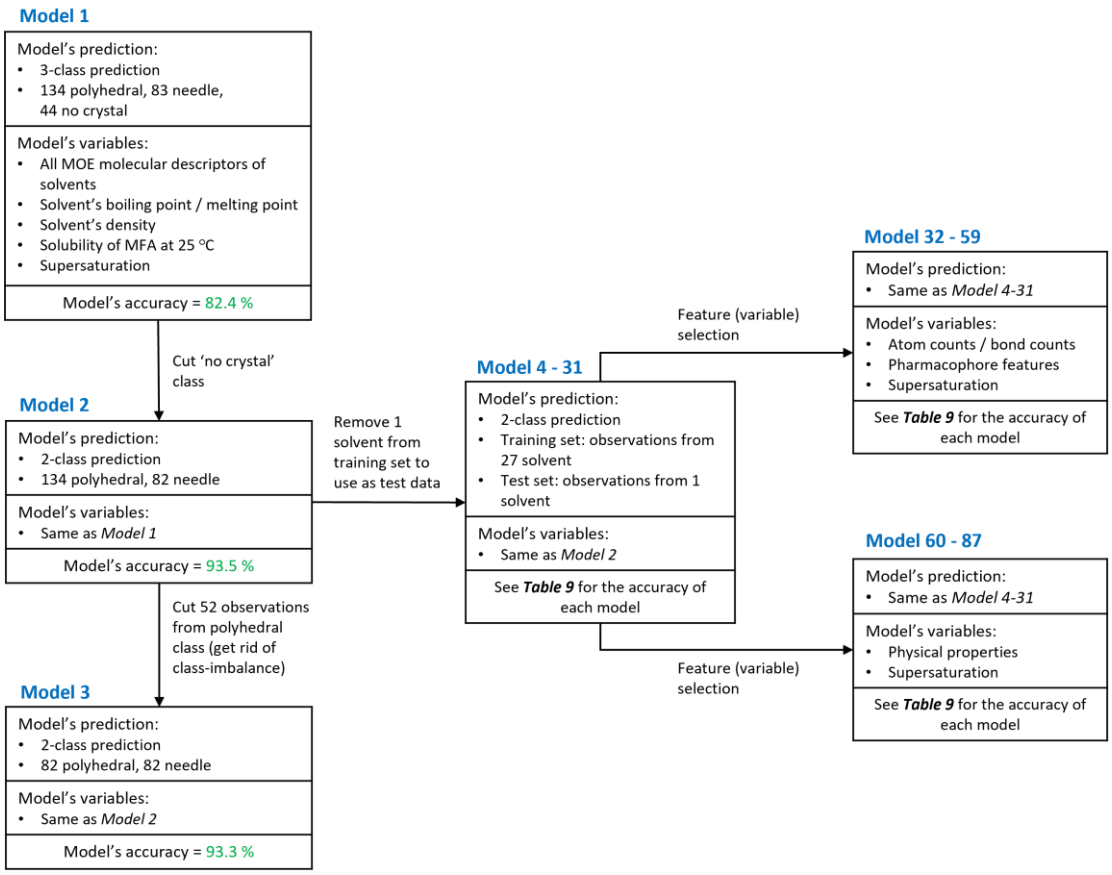


Figure 35. *Diagram showing the dataset, variable and accuracies of all models*

5.2.4.2 Model evaluation:

Train-test split and n-fold cross-validation [226] were used to evaluate the prediction accuracy of the RF classification models. **Table 6** shows the prediction accuracy of the models evaluated with different ratios of training and test data. Ratios of 75:25, 80:20, and 90:10 were used in the train-test split method, comparable to 4-fold, 5-fold, and 10-fold cross-validation, respectively.

Table 6. *Model evaluation by train-test split and cross-validation of Models 1, 2 and 3. SD = standard deviation*

Prediction	**Accuracy by train-test split (train:test)**			**Accuracy by cross-validation**		
	75:25	**80:20**	**90:10**	**4-fold**	**5-fold**	**10-fold**
Model 1 (3 classes)	84.4% (SD = 3.6%)	84.2% (SD = 4.5%)	85.0% (SD = 6.2%)	82.4% (SD = 3.1%)	84.7% (SD = 2.1%)	83.1% (SD = 4.6%)
Model 2 (2 classes w/ class-imbalance)	91.8% (SD = 3.3%)	92.1% (SD = 3.6%)	93.7% (SD = 4.6%)	93.5% (SD = 2.1%)	94.4% (SD = 2.4%)	93.5% (SD = 4.7%)
Model 3 (2 classes w/o class-imbalance)	93.8% (SD = 3.8%)	93.6% (SD = 4.3%)	95.5% (SD = 4.7%)	93.3% (SD = 5.3%)	92.6% (SD = 6.4%)	95.7% (SD = 4.7%)

Overall, the different accuracies as calculated by either train-test split or cross-validation varied by no more than 3%. This consistency shows the RF approach to be robust to different methods of validation. The lowest ratio was used to save computational time and reduce standard deviation in the model.[226] Between the two evaluation methods, the variance of the accuracy calculated from n-fold cross-validation was lower than those from the train-test split. As a result, 4-fold cross-validation was used for evaluating the model performance in this work.

5.3 Results and Discussion

5.3.1 Crystallisation:

MFA was crystallised from 30 solvents over 5 days at a range of supersaturations (261 observations in total). Crystallisation was observed in all solvents except isobutyl acetate and 1-butanol during the 5-day experimental period. **Table 7** presents crystal shapes and corresponding solvents. Four crystal morphologies were observed: plates, elongated plates, needles, and spherulites (**Figure 34**). Plates (**Figure 34a**) and elongated plates (**Figure 34b**) were classified as polyhedral crystals while needle (**Figure 34c**) and spherulitic (**Figure 34d**) crystals were both classified as needle crystals. Based on face-indexing data, the biggest face which dominated the polyhedral crystal is [100] (**Figure 36**). This observed crystal shape corresponded to the BFDH morphology of MFA crystal form-I (**Figure 37**).

Table 7. *The list of organic solvents categorized by the shape of MFA crystals they can produce*

Polyhedral	**Needle**	**Supersaturation dependent** (polyhedral supersaturation range, needle supersaturation range)
1,2 dichloroethane	1-bromobutane	1,4 dioxane (1.18 – 1.28, 1.39 – 1.91)
1-chlorobutane	1-methylnaphtalene	2-butanol (1.51 – 1.83, 1.94 – 2.03)
1-octanol	aniline	2-butanone (1.10 – 1.50, 1.60 – 2.01)
2-methoxyethanol	anisole	2-propanol (1.14 – 1.41, 1.49 – 1.99)
acetic acid	methyl acetate	butyl acetate (1.32, 1.42 – 2.00)
acetone	nitromethane	diethyl sulfide (1.06 – 1.57, 1.76 – 1.94)
acetonitrile	toluene	Methanol (1.13 – 1.22, 1.30 – 1.98)
chloroform		
ethanol		
DMF		
ethyl acetate		
iodomethane		
triethylamine		
trichloroethylene		

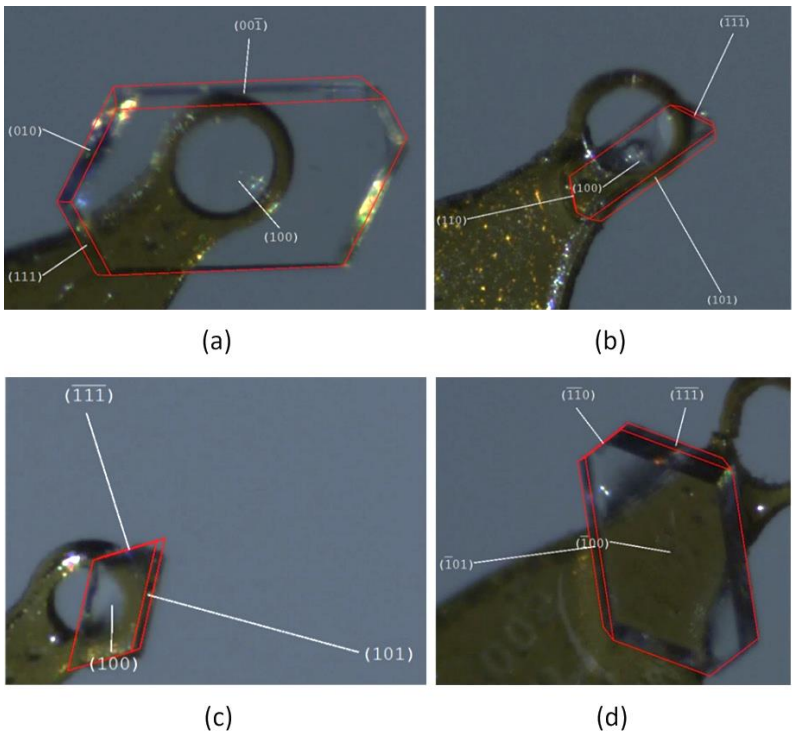


Figure 36. *Face indexing of single crystal of MFA crystallised from (a) methanol, (b) ethyl acetate, (c) acetonitrile, and (d) 2-butanol. The face that dominates crystal morphology is (100).*

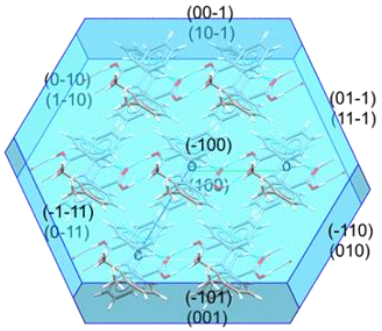


Figure 37. *BFDH morphology of MFA crystal form-I shows plate-like crystal morphology generated with Mercury software (version 2021.2.0)*

Polyhedral crystals were always found at all supersaturation levels (in the range of 1.1 – 2.7) when using the following solvents: 1,2 dichloroethane, 1-chlorobutane, 1-octanol, 2-methoxyethanol, acetic acid, acetone, acetonitrile, chloroform, ethanol, DMF, ethyl acetate, iodomethane, triethylamine, trichloroethylene. At a supersaturation range of 1.1 – 3.0, the crystals of MFA exhibited needle shape when crystallised from the following solvents: 1-bromobutane, 1-methylnaphtalene, aniline, anisole, methyl acetate, nitromethane, toluene. As for crystals grown from 1,4 dioxane, 2-butanol, 2-butanone, 2-propanol, butyl acetate, diethyl sulphide and methanol, the crystal shape was supersaturation dependent. For these

solvents, polyhedral crystals were observed at low supersaturation and needles were observed at higher supersaturations.

5.3.2 *Model performance using crystal shape observations from all solvents in the training set*

Three RF classification models were built initially to determine the efficacy of this method and understand the extent to which the class imbalance present in the dataset would affect prediction accuracies. In Model 1 the full dataset was separated into the following 3 classes: polyhedral (134 data points), needle (83 data points), and 'no crystal' (44 data points). In Model 2, the 'no crystal' class was removed resulting in a 2-class prediction model. The class-imbalance present in Model 2 was removed for the dataset used in Model 3 by removing observations in the polyhedral class until the needle and polyhedral classes were equally populated. For 4-fold cross-validation, Model 1 had the lowest prediction accuracy (82.4%) while Models 2 and 3 had prediction accuracies of 93.5% and 93.3%, respectively. Additionally, the values of accuracy, precision, recall, and F1-score of these three models also agreed with the model accuracies (**Table 8**). As these results indicate that the class imbalance observed in Model 2 did not noticeably affect the model performance, the dataset used in Model 2 was used for further models with the modifications discussed below.

Table 8. *The models' precision, recall, and F1-score. The 'Support' column indicates the number of test data in each crystal class.*

Model prediction	Precision	Recall	F1-score	Support
Model 1 (3 crystal outcomes with class imbalance)				
Polyhedral	0.83	0.94	0.88	31
Needle	0.89	0.80	0.84	20
No crystal	0.85	0.73	0.79	15
Model 2 (2 crystal outcomes with class imbalance)				
Polyhedral	0.91	1.00	0.95	31
Needle	1.00	0.87	0.93	23
Model 3 (2 crystal outcomes without class imbalance)				
Polyhedral	1.00	0.84	0.91	19
Needle	0.88	1.00	0.94	22

Confusion matrixes of *Models 1*, *2*, and *3* are presented in **Figure 38a**, **Figure 38b**, and **Figure 38c**, respectively. The number in each column represents the number of each class predicted by the model, while the number in each row represents the number of experimental results or actual classification in the dataset. The sum of the numbers in each column is the total number of the data in each predicted class, while the sum of the numbers in each row is the

total number of actual data in each class of the test set. The numbers on the diagonal axis of the matrix represent correct predictions. On the other hand, the numbers in the remaining fields represent incorrect predictions.

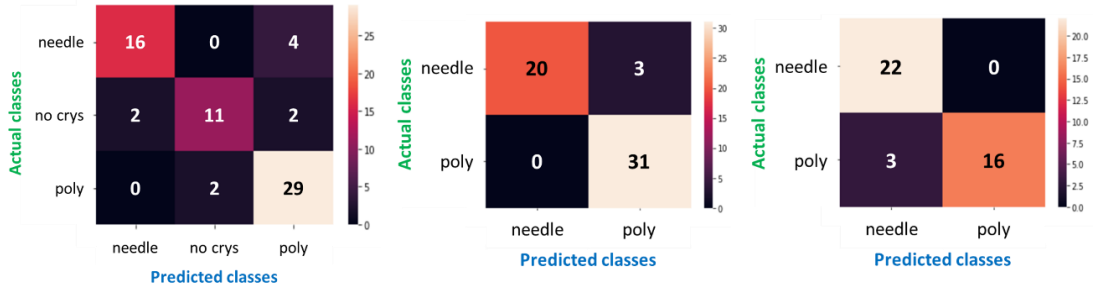


Figure 38. *The confusion matrix of the RF classification model for the prediction of MFA crystal shapes (a) 3-class prediction, (b) 2-class prediction with class imbalance, and (c) 2-class prediction without class-imbalance*

5.3.3 *Prediction of crystal shape from solvents not included in the training set*

To determine the ability of this methodology to predict crystal morphology from solvents for which no data was present in the training set, we built 84 additional models that each had all observations for a single solvent removed from the training data. The performance accuracy for each model was then assessed using the crystal morphologies for the solvent excluded from the training data. Additionally, three different feature sets were tested to determine if model performance accuracy was affected by the inclusion of different variables in the training sets (see **Figure 35** and **Table 9** for more details). The three feature sets were (i) all features present in the original dataset, (ii) atom count, bond count, pharmacophore descriptors for the solvents and supersaturations of the crystallisation experiments, and (iii) solvent physical properties and supersaturations of crystallisation experiments.

In total, 32 out of 84 models predicted the shape of MFA crystals with 100% accuracy, and the models trained with the third feature set resulted in the best overall prediction accuracy for morphologies across all solvents. When including only physical property descriptors and supersaturations in the model features, 12 solvent models had 100% prediction accuracy, 8 solvent models had accuracies from 50-100%, and the remaining 8 models had prediction accuracies below 50%. When using atom count, bond count, and pharmacophore descriptors as variables, 10 models had 100% prediction accuracy, 7 models had accuracies from 50-100%, and 11 models had accuracies below 50%. For the models using all solvent molecular

descriptors as variables, 10 models had 100% prediction accuracy, 6 models had accuracies from 50-100%, and 12 models had accuracies below 50%. Thus, using all descriptors in the feature set resulted in the lowest performance across all solvents while using only solvent physical properties and supersaturations as the feature set had the highest accuracies across all solvents. These results suggest that some of the variables in the atom count, bound count and pharmacophore descriptor feature set had a confounding effect on model performance.

Accuracy trends were also observed for solvent type. All models had high prediction accuracies for morphologies of crystals grown in chlorinated solvents (1,2 chloroethane, chloroform, and trichloroethylene), aniline, anisole, ethanol, and toluene. By contrast, the models performed poorly when predicting morphologies for crystals grown from 1-octanol, triethylamine, methyl acetate, and nitromethane. To understand why RF classification consistently performed well for some solvents and badly for others, these results were explored via logistic regression. Crystal form characterisation was also investigated for crystals grown in solvents where morphology was poorly predicted.

Table 9. *The prediction accuracy of the models testing the prediction of crystal shape from individual solvents. poly = polyhedral crystals, nd = needle. All training set and test set data included the relevant solvent descriptors and experimental supersaturation.*

Solvent in which test set data was collected	Number of samples in test set	Experimental crystal shape	Solvent descriptors					
			Variable group 1: All solvent descriptors		Variable group 2: Atom counts / bond counts + pharmacophore features		Variable group 3: Physical properties	
			Predicted shape	Prediction accuracy	Predicted shape	Prediction accuracy	Predicted shape	Prediction accuracy
1,2-dichloroethane	7	Polyhedral	Polyhedral	100 %	Polyhedral	100 %	Polyhedral	100 %
Chloroform	5	Polyhedral	Polyhedral	100 %	Polyhedral	100 %	Polyhedral	100 %
Trichloroethylene	4	Polyhedral	Polyhedral	100 %	Polyhedral	100 %	Polyhedral	100 %
Ethanol	9	Polyhedral	Polyhedral	100 %	Polyhedral	100 %	Polyhedral	100 %
Aniline	7	Needle	Needle	100 %	Needle	100 %	Needle	100 %
Anisole	10	Needle	Needle	100 %	Needle	100 %	Needle	100 %
Toluene	6	Needle	Needle	100 %	Needle	100 %	Needle	100 %
Acetonitrile	12	Polyhedral	Polyhedral	100 %	Polyhedral	100 %	10 poly, 2 nd	83.3 %
Acetone	9	Polyhedral	7 poly, 2 nd	77.8 %	Polyhedral	100 %	Polyhedral	100 %
Iodomethane	3	Polyhedral	Polyhedral	100 %	1 poly, 2 nd	33.3 %	Polyhedral	100 %
2-propanol	10	6 poly, 4 nd	polyhedral	60.0 %	polyhedral	60.0 %	7 poly, 3 nd	90.0 %
2-methoxyethanol	10	Polyhedral	4 poly, 6 nd	40.0 %	6 poly, 4 nd	60.0 %	Polyhedral	100 %
2-butanol	6	3 poly, 3 nd	1 poly, 5 nd	66.7 %	1 poly, 5 nd	66.7 %	1 poly, 5 nd	66.7 %
2-butanone	9	5 poly, 4 nd	polyhedral	55.6 %	polyhedral	55.6 %	6 poly, 3 nd	88.9 %
1-methylnaphthalene	8	Needle	needle	100 %	needle	100 %	polyhedral	0 %
Methanol	10	6 poly, 4 nd	polyhedral	60.0 %	polyhedral	60.0 %	polyhedral	60.0 %
Diethyl sulfide	7	5 poly, 2 nd	polyhedral	71.4 %	polyhedral	71.4 %	needle	28.6 %
1,4-dioxane	8	2 poly, 6 nd	6 poly, 2 nd	50.0 %	needle	75.0 %	polyhedral	25.0 %

Table 9 (Cont.) *The prediction accuracy of the models testing the prediction of crystal shape from individual solvents. poly = polyhedral crystals, nd = needle. All training set and test set data included the relevant solvent descriptors and experimental supersaturation.*

Solvent in which test set data was collected	Number of samples in test set	Experimental crystal shape	Solvent descriptors					
			Variable group 1: All solvent descriptors		Variable group 2: Atom counts / bond counts + pharmacophore features		Variable group 3: Physical properties	
			Predicted shape	Prediction accuracy	Predicted shape	Prediction accuracy	Predicted shape	Prediction accuracy
DMF	9	Polyhedral	3 poly, 5 nd	33.3 %	needle	0 %	polyhedral	100 %
Ethyl acetate	6	Polyhedral	needle	0 %	3 poly, 3 nd	50.0 %	4 poly, 2 nd	66.7 %
Acetic acid	10	Polyhedral	1 poly, 9 nd	10.0 %	needle	0 %	polyhedral	100 %
Butyl acetate	7	1 poly, 6 nd	polyhedral	14.3 %	polyhedral	14.3 %	3 poly, 4 nd	71.4 %
1-bromobutane	7	Needle	polyhedral	0 %	polyhedral	0 %	2 poly, 5 nd	71.4 %
1-chlorobutane	6	Polyhedral	needle	0 %	1 poly, 5 nd	16.7 %	2 poly, 4 nd	33.3 %
Triethylamine	8	Polyhedral	2 poly, 6 nd	25.0 %	2 poly, 6 nd	25.0 %	needle	0 %
1-Octanol	7	Polyhedral	needle	0 %	needle	0 %	needle	0 %
Methyl acetate	11	Needle	polyhedral	0 %	polyhedral	0 %	polyhedral	0 %
Nitromethane	5	Needle	polyhedral	0 %	polyhedral	0 %	polyhedral	0 %

5.3.4 *Variable Importance in the RF Classification for crystal morphology prediction*

Table 10 shows the two most important variables for each model for solvents with the highest and lowest prediction accuracies. For the first two variable sets, the most important feature focus on the structure of the molecule, mainly the number of rings, number of rigid or single bonds, atom count and adjacency matrix. There is no clear difference between the most important descriptors identified for the models that performed poorly or well. Across all models, using these two sets of variables performed similarly in terms of the number of correct and incorrect predictions. Models using the third variable set (13 physical properties MOE descriptors) performed much better and identified the most important variables including aqueous solubility and molecular refractivity.

Aqueous solubility can be linked with the ability of the molecules to form H-bonds while molecular refractivity is related to London dispersive forces.[227] The anisotropy of the rate of incorporation of growth units from solution to individual crystal faces determines crystal shape.[193,228] In solution, both the crystal surface and solute growth units are solvated, and the relative growth rates of faces depend on the strengths of intermolecular interactions between the solute-solvent and solvent-crystal surfaces.[229,230] It was demonstrated previously that the crystallisation from organic solvents is dominated by weak interactions between permanent dipoles and London dispersion forces between the nonpolar groups of the solute and solvent and these interactions are responsible for different crystal shapes obtained from various solvents.[231] Our machine-learning model also identified these interactions as the most important distinguishers between models for solvents that show very good prediction accuracy (100%). We identified that if the model does not identify the two most important variables as aqueous solubility or molecular refractivity, the accuracy of the predictions is low. Note that this assumption refers to the models with the third set of variables (physical properties). We are interested in physical property variables because the overall performance of the models using this variable set is the best among all models. For the models using the first variable set, there is no clear difference between the most important descriptors identified for the models that performed poorly or well. A possible reason why these models are not able to identify aqueous solubility and molecular refractivity as the most important variables may result from too many variables leading to

over-complexity of the model, making it harder to understand which variables are relevant or irrelevant, and which variables are most important.

Table 10. *List of first and second most important variables of the models for predicting the shape of crystals crystallised from individual solvents*

Crystallisation solvents	The most important variables of each model		
	Variable group 1: All solvent descriptors	**Variable group 2: Atom counts / bond counts + pharmacophore features**	**Variable group 3: Physical properties**
Solvents where the crystals were 100% accurately predicted by the models			
1,2-Dichloroethane	1. number of rings 2. adjacency matrix	1. no. of rigid bonds 2. atom count	1. aqueous solubility 2. molecular refractivity
Chloroform	1. adjacency matrix 2. number of rings	1. no. of rigid bonds 2. no. of single bonds	1. aqueous solubility 2. molecular refractivity
Ethanol	1. adjacency matrix 2. number of rings	1. no. of rigid bonds 2. no. of single bonds	1. aqueous solubility 2. molecular refractivity
Trichloroethylene	1. adjacency matrix 2. number of rings	1. no. of rigid bonds 2. no. of single bonds	1. aqueous solubility 2. molecular refractivity
Aniline	1. adjacency matrix 2. number of rings	1. no. of single bonds 2. no. of rigid bonds	1. aqueous solubility 2. bpol[#]
Anisole	1. number of rings 2. distance Matrix	1. no. of rigid bonds 2. no. of single bonds	1. aqueous solubility 2. molecular refractivity
Toluene	1. adjacency matrix 2. number of rings	1. no. of single bonds 2. no. of rigid bonds	1. aqueous solubility 2. molecular refractivity
Solvents where the crystals were incorrectly predicted by the models			
1-Chlorobutane	1. number of rings 2. adjacency matrix	1. no. of rigid bonds 2. no. of single bonds	1. aqueous solubility 2. bpol[#]
1-Octanol	1. chi1_C* 2. zagreb[$]	1. no. of heavy atoms 2. no. of rigid bonds	1. aqueous solubility 2. VDW volume
Triethylamine	1. distance matrix 2. molecular refractivity	1. no. of single bonds 2. no. of rigid bonds	1. molecular refractivity 2. VDW volume
Methyl acetate	1. distance matrix 2. adjacency matrix	1. no. of rigid bonds 2. no. of rings	1. VDW volume 2. molecular refractivity
Nitromethane	1. adjacency matrix 2. number of rings	1. no. of rigid bonds 2. atom count	1. bpol[#] 2. aqueous solubility

5.3.5 Using Logistic Regression to Understand Model Performance

Logistic regression was also used to probe why the RF models consistently performed well for some solvents and poorly for others even when the solvent feature sets were changed. For this analysis, models 60-87 were used (i.e. solvent-exclusion models that used solvent physical properties and supersaturation as training variables), and models with prediction accuracy greater than 50% were labelled as 1 while models with prediction accuracies less

than 50% were labelled as 0. This set of models was chosen as the feature set for these models resulted in the highest overall prediction accuracy across solvents. The most important features in logistic regression can be determined by the highest absolute values of the variable coefficients and/or recursive feature elimination until only the most relevant features remain. The details are presented in **Table 11**.

Table 11. *The MOE descriptors included as variables in the RF classification Models 60-87 listed according to importance scores in the logistic regression analysis of the performance of these models. RF model accuracies above 50% were labelled as 1 in the logistic regression analysis while RF model accuracies below 50% were labelled as 0. Recursive feature elimination was done until the 6 most relevant features/variables remained (these 6 features are ranked as 1 in the table below).*

MOE Descriptor	Summary of MOE Descriptor	Logistic Regression Coefficients	Ranking by Recursive Feature Elimination
bpol	sum of the absolute value of the difference between atomic polarizabilities of all bonded atoms in the molecule	-0.7288	1
apol	sum of the atomic polarizabilities	-0.4332	1
logS	log of the aqueous solubility (mol/L)	0.3232	1
SMR	molecular refractivity	-0.2926	1
vdw_area	Area of van der Waals surface	-0.2872	1
vdw_volume	van der Waals volume	-0.2594	1
mr	molecular refractivity	-0.2587	2
logP(o/w)	log of the octanol/water partition coefficient	-0.2248	3
density	molecular mass density	0.1845	4
reactive	indicator of the presence of reactive groups	0.1039	5
TPSA	polar surface area	-0.1081	6
SlogP	log of the octanol/water partition coefficient	-0.0897	7
Weight	molecular weight	-0.0209	8

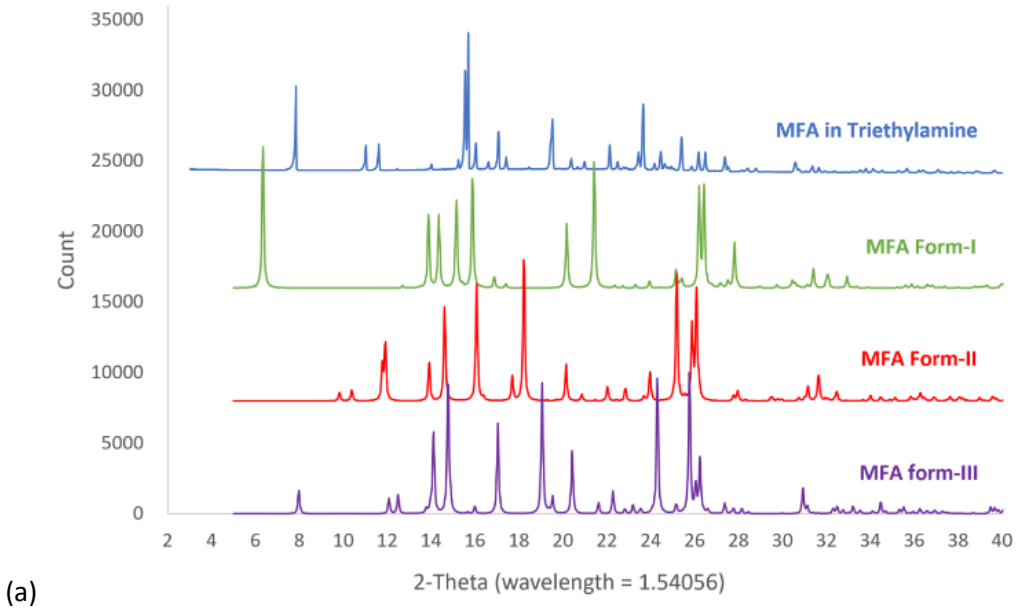
From the relative importance of different variables in the logistic regression analysis, we see that polarizability (apol, bpol) and solubility (logS) play an important role in determining whether the RF classification model performed well for a given solvent. While the polar surface area variable (TPSA) was deemed a relatively unimportant feature, this rating may be due to this variable being redundant after the inclusion of apol and bpol into the models. Variables pertaining to van der Waals interactions (vdw_area and vdw_volume) were also

amongst the more relevant features in determining whether the RF classification models performed well for observations in a given solvent. As we would expect crystal morphologies to be strongly influenced by intermolecular interactions between the MFA and the crystallisation solvent, the importance of variables pertaining to solubility, polarity and van der Waals interactions corresponds with the important physical parameters in a crystallisation experiment.

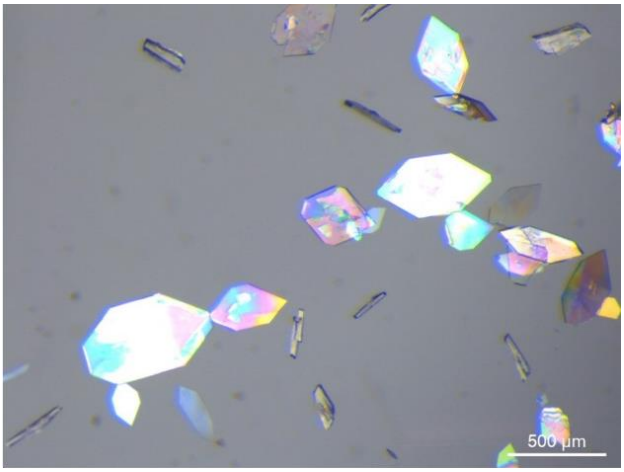
Logistic regression also suggests a possible reason why the model using physical property variables poorly performed when predicting the crystal shape from some solvents. By increasing the value of the variables with negative logistic regression coefficient (i.e. bpol and apol), the probability that a model performs well decreases and the probability that a model poorly performs increases. On the other hand, increasing the value of the variables with positive logistic regression coefficient (i.e. logS) will increase the probability that a model performs well and decrease the probability that a model poorly performs.[122] The results from **Table 11** could explain why the models cannot correctly predict the shape of crystals grown from 1-octanol, since 1-octanol has high values of bpol and apol variables and low value of logS variable. This result suggests a limitation of this predictive model in which there are ranges of the values of some variables that might negatively affect the model performance.

5.3.6 Characterisation of MFA crystals grown in triethylamine

Further crystal characterisation was done for the crystals grown in solvents with the models showing low prediction accuracy. All samples were consistent with MFA form I except the sample crystallised from trimethylamine which exhibited a notably distinct PXRD pattern (**Figure 39a**). Characterisation of the MFA grown in triethylamine was of particular interest as results revealed these crystals to be a previously unidentified solvate of MFA. Additionally, the shape of the crystals grown in triethylamine had thinner flat plates as observed under a microscope when compared to the plate crystals of MFA form-I crystallised from the other solvents (see **Figure 39b**).



(a)



(b)

Figure 39. a) Experimental powder X-ray diffraction pattern of MFA crystallised from triethylamine, compared to the simulated powder patterns of MFA form-I (refcode: XYANAC), II (refcode: XYANAC02), and III (refcode: XYANAC03) calculated from Mercury, b) MFA crystals crystallised from triethylamine at supersaturation = 1.4

Characterisation of these crystals by differential scanning calorimetry (DSC) also suggested that MFA crystals grown from triethylamine were a previously unidentified solvate. According to the DSC results from the work carried out by Adam, et al. (2000),[213] the onset temperature of the first endothermic peak of MFA ranges from 187°C to 205°C and corresponds to the transition temperature from MFA form-I to form-II. A second sharper peak has an onset temperature of around 230°C and is known to correspond to the melting point of MFA form-II.

For our MFA crystals crystallised from triethylamine, the DSC thermogram (**Figure 40**) showed different onset temperatures for both peaks when compared to the literature values. In our results, the first peak and the second peak have onset temperatures of 110.5 °C and 214.1 °C, respectively.

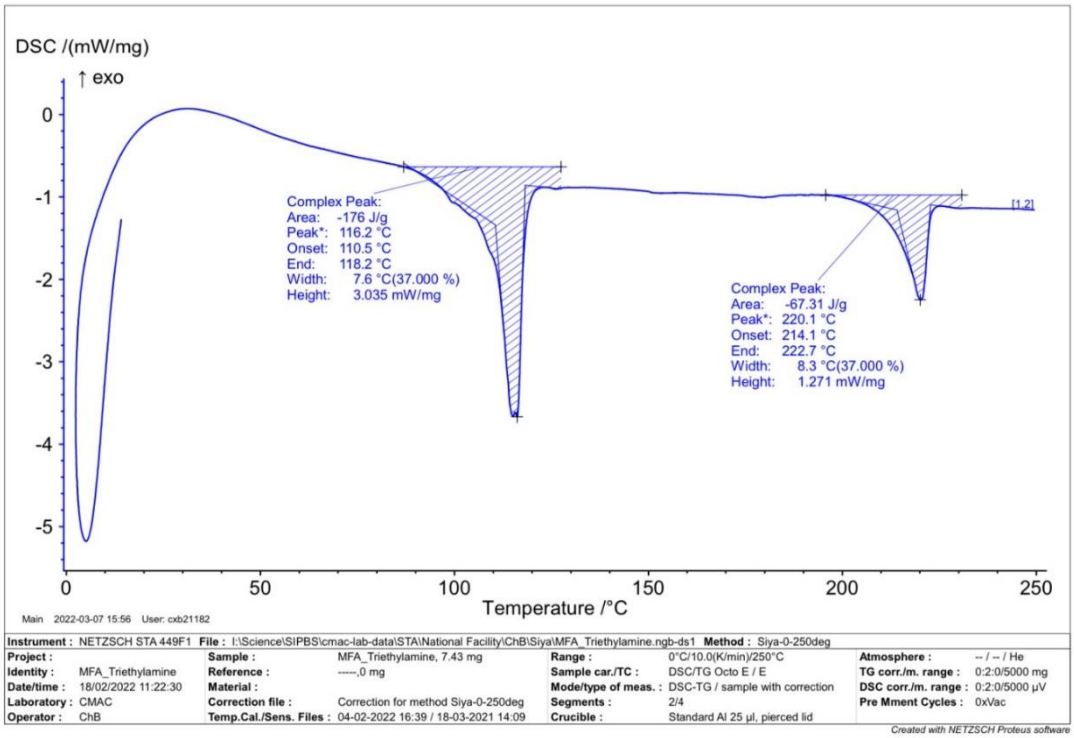


Figure 40. *DSC curve for MFA crystallised from triethylamine by cooling crystallisation*

While all other crystals grown in other solvents for which morphologies were poorly predicted by RF models were shown to be form I, the results shown here suggest that the poor prediction accuracy for the trimethylamine crystal morphology may be due to the distinctiveness of these crystals from the four previously document forms of MFA crystals rather than an innate flaw in the RF classification approach.

5.4 Conclusions

The choice of solvent in crystallisation is a critical design decision and can affect the crystal morphology with further implications for downstream manufacturability. For this work, we generated 261 experimental observations of MFA crystal shape in 30 various organic solvents at the range of supersaturation levels between S = 1.0 – 3.0. RF classification models can predict the shape of MFA crystals observed from different solvents experimentally. Thus, the results illustrate that RF classification can be a useful tool to predict the experimental crystal

shape of MFA. Our two-class RF prediction model with polyhedral and needle classes resulted in a prediction accuracy of 93%. This model was further modified to explore prediction accuracies for crystals grown in specific solvents. For solvents that were excluded from the training set at all supersaturation levels, the prediction accuracy depended on the solvent. The most important variables for the correctly predicted solvents relate to H-bonds and London dispersion forces identifying this interaction as key for the determination of a crystal shape. Whilst demonstrated only for MFA it is expected that with the appropriate data, the application of this tool can be broadened to cover a wider range of active ingredient molecular and crystal attributes. In order to enhance the predictive capabilities of our model for a broader variety of APIs, it is proposed that data from crystallisation experiments on a diverse set of materials should be collected. This data could encompass a range of crystallisation process parameters, including temperature, cooling rate, agitation, the presence of seed crystals, and the use of additives. By incorporating this data, the model will be better equipped to predict the resultant crystal shapes under a variety of crystallisation conditions. Such data are already often collected during physical form selection, solubility and early development studies. The adequacy of the training data for a model is not only dependent on the data quality, but also on the sufficient number of experiments for training the model. To determine whether there is an adequate amount of data for training the model, various techniques such as cross-validation, assessment of out-of-sample performance, and examination of learning curves should be employed. These methods can provide insight into the sufficiency of the data and inform any necessary adjustments to the model. This study also highlights the potential role of machine learning and data-driven predictive tools to support decision-making during pharmaceutical process development, e.g. informing solvent selection, reducing experimental time and material consumption and enabling the selection of conditions that deliver materials engineered to achieve desirable attributes.

6. Investigating potential correlations between PXRD peaks at low angles and the crystal structures of solvates/non-solvates

6 Investigating potential correlations between PXRD peaks at low angles and the crystal structures of solvates/non-solvates

6.1 Introduction

The crystallisation processing steps involved during the drug development of pharmaceutical drugs can result in unforeseen changes to the solid forms of crystalline active pharmaceutical ingredients (APIs) or new chemical ingredients (NCEs). One such change is the incorporation of water or other solvent molecules into a drug's crystal structure, resulting in hydrate or solvate formation, respectively.[232] It has been estimated that around 33% of organic compounds are susceptible to hydrate formation whilst approximately 10% are capable of forming solvates robustly with organic solvents.[233] Examples of the marketed pharmaceutical products in hydrate form are chloral hydrate (Noctec),[234] ciprofloxacin monohydrate (Cipro),[235] and levofloxacin hemihydrate (Levaquin).[236] The phenomenon of solvate formation can affect the physicochemical properties of drugs, and the unexpected formation of solvates can impact the manufacturability and pharmacokinetic properties of drug candidates. A survey of solvates in the literature suggests that some compounds form solvates in a range of solvents[237] while other compounds only form selected solvates.[238,239]

Although progress has been made to understand the structural features that lead to solvate formation, a comprehensive understanding of why some molecules form solvates more readily than others remains elusive. Whilst the discovery of novel solvatomorphs of APIs and NCEs provides an opportunity to alter the physical properties of drug substances, the possibility of forming these solvatomorphs can make it challenging to control the solid form during the drug development process.[240–243] The most commonly used techniques to differentiate between non-solvate and solvate forms of crystalline materials are X-ray diffraction methods (single crystal[244,245] and powder X-ray diffraction[246–248]), thermal techniques such as differential scanning calorimetry (DSC),[246–250] thermogravimetric analysis (TGA),[246–250] Solid State nuclear magnetic resonance spectroscopy (SSNMR),[250] and other spectroscopic techniques such as Raman spectroscopy.[251,252] As solvates and hydrates significantly affect the safety, quality, and efficiency of the crystalline products, the ability to predict whether the crystal structures exhibit solvate or non-solvate forms will be highly useful.

6.2 Solvent incorporation into solvates

The solvent molecules are incorporated into the crystal lattice by the following two mechanisms. When the solute-solvent molecular interaction dominates solute-solute molecular interaction, solvent molecules incorporate into the crystal structure by forming hydrogen bonds with the molecules of APIs. This incorporation of solvates results in the formation of stoichiometric solvates in which the molecules of solvent modify the crystal structure of the host molecule.[241,253,254] Alternatively, when void volume in the crystal structure is sufficient for the inclusion of solvent molecules, the solvates form by occupying the void space through weak interactions, resulting in the formation of channel solvates or non-stoichiometric solvates. In non-stoichiometric solvates, the solvent's molecules have weaker interaction with the molecules of solute and can interact with the surrounding molecules outside the solvate structures. This interaction between solvent molecules inside and outside of the solvate structure can affect the stability and quality of the formulated products.[241,253]

6.3 Powder patterns of solvated and non-solvated structures

PXRD patterns are determined by the crystal structures of crystalline compounds. Since the unit cell parameters (cell lengths, cell angles, and cell volume) of the same compound in solvate form and non-solvate form are different, their powder patterns are also dissimilar.[255] In general, due to the incorporation of solvent molecules in the crystal lattice of the API, the unit cell lengths of solvate forms are longer than those of non-solvate forms which results in a higher cell volume for the structures that belong to the same space group and have the equal number of molecules in an asymmetric unit. Moreover, according to Bragg's equation (**Equation 6**), when interplanar spacing expands by increasing unit cell lengths, as a result, the PXRD peaks are expected at lower 2-theta.[135,136] **Table 12** and **Figure 41** show the differences in the unit cell parameters and powder patterns between ciprofloxacin in non-solvate form and solvate/hydrate forms derived from CSD, respectively.

***Table 12**. Unit cell parameters of ciprofloxacin in non-solvate form, ciprofloxacin hexahydrate, and ciprofloxacin difluoroethanol solvate*

Compounds	Space group	Cell lengths [Å]			Cell angles [°]			Cell volume [Å^3]
		a	b	c	α	β	γ	
Ciprofloxacin (CSD ref code: UHITOV)	P-1	7.96	8.58	10.77	87.87	85.15	88.21	732.43
Ciprofloxacin hexahydrate (CSD ref code: COVPIN01)	P-1	9.51	9.94	11.04	94.23	100.21	91.33	1023.67
Ciprofloxacin difluoroethanol solvate (CSD ref code: ENODOB)	P-1	10.98	13.98	13.98	105.47	90.35	93.36	2063.22

The differences between the powder patterns of non-solvate, hydrate, and solvate crystals of the same compound (**Figure 41**), which is ciprofloxacin in this example, indicate that X-ray powder diffraction can be used to distinguish solvate and non-solvate crystal structures.

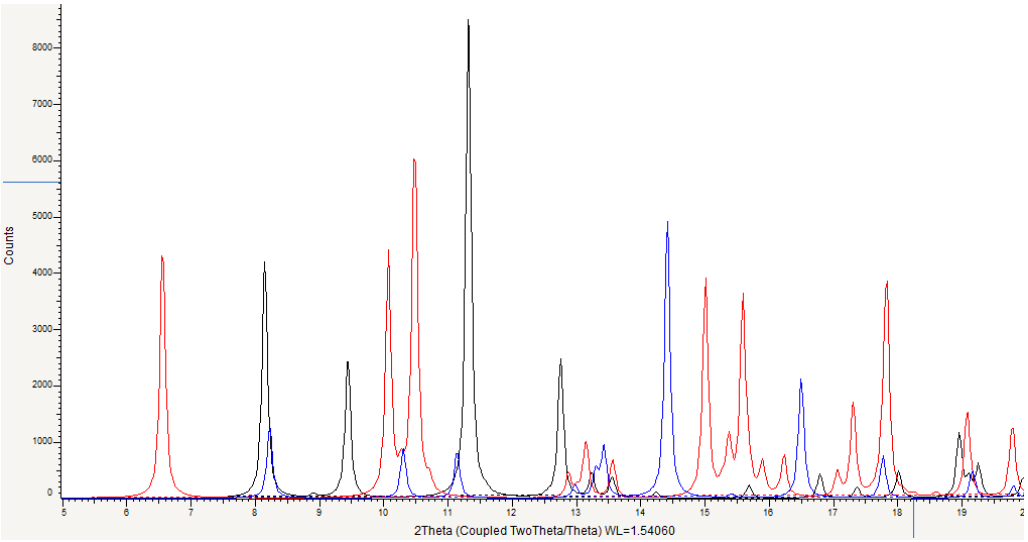


***Figure 41.** Simulated PXRD patterns of ciprofloxacin: non-solvate form (blue), ciprofloxacin hexahydrate (black), and ciprofloxacin difluoroethanol solvates (red). Solvated and hydrated forms show lower angle PXRD peaks compared to the non-solvated form.*

6.4 Methodology

6.4.1 Creation of the dataset of solvate and non-solvate crystal structures

The database of crystal structures was prepared by Laura Straughair, University of Strathclyde (2021). The metadata of organic molecules (containing only the atoms H, O, C, N, S, P, F, Cl, Br, and I) were extracted from the Cambridge Structural Database (CSD) by using CSD Python API (version 5.41, March 2020). The database contains small organic molecules

between 100 and 1,000 Da in molecular weight, with an atom-to-bond ratio of less than 1.3. Duplicate crystal structures have been retained in case there are powder pattern differences between polymorphs.

The database consists of reference codes as the identifier of crystal structures in 4 classes based on the existence of solvent molecules or water molecules in their crystal structures, namely solvate class, non-solvate class, hydrate class, and non-hydrate class. The solvate class was further classified into 4 subclasses, which are heterosolvate subclass, solvate hydrate subclass, ionic solvate subclass, and regular solvate subclass. Each subclass of the solvate class and the non-solvate class was further classified into 11 categories depending on the recrystallisation solvents (**Figure 42**).

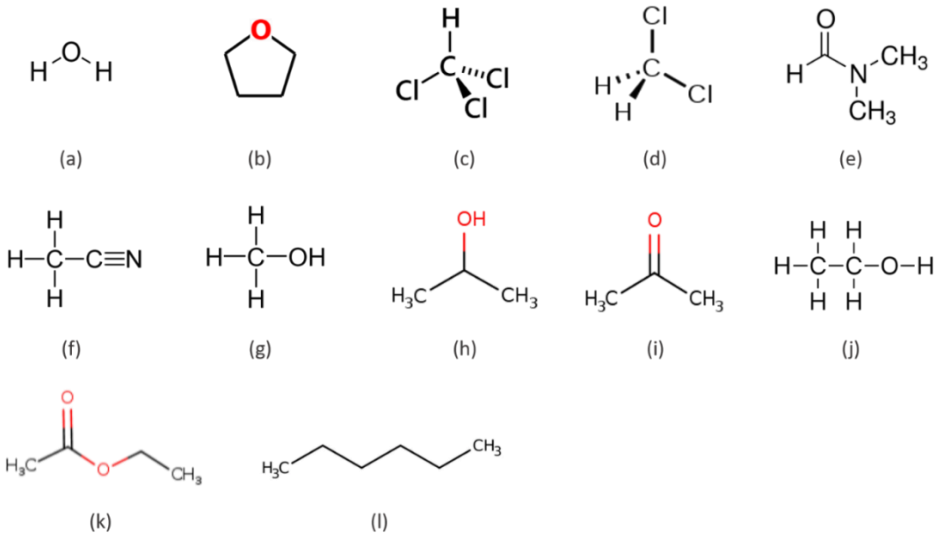


***Figure 42**. Molecular structures of recrystallisation solvents. (a) water, (b) THF, (c) chloroform, (d) DCM, (e) DMF, (f) acetonitrile, (g) methanol, (h) IPA, (i) acetone, (j) ethanol, (k) ethyl acetate, and (l) hexane*

The solvents in this database are acetone, acetonitrile, chloroform, dichloromethane (DCM), dimethylformamide (DMF), ethanol, ethyl acetate, hexane, isopropyl alcohol (IPA), methanol, and tetrahydrofuran (THF). Hydrate and non-hydrate classes consist of the structures crystallised from water, and the structures in the hydrate class were classified into 3 subclasses, namely solvate hydrates, ionic hydrates, and regular hydrates. **Figure 43** shows the structure of the database.

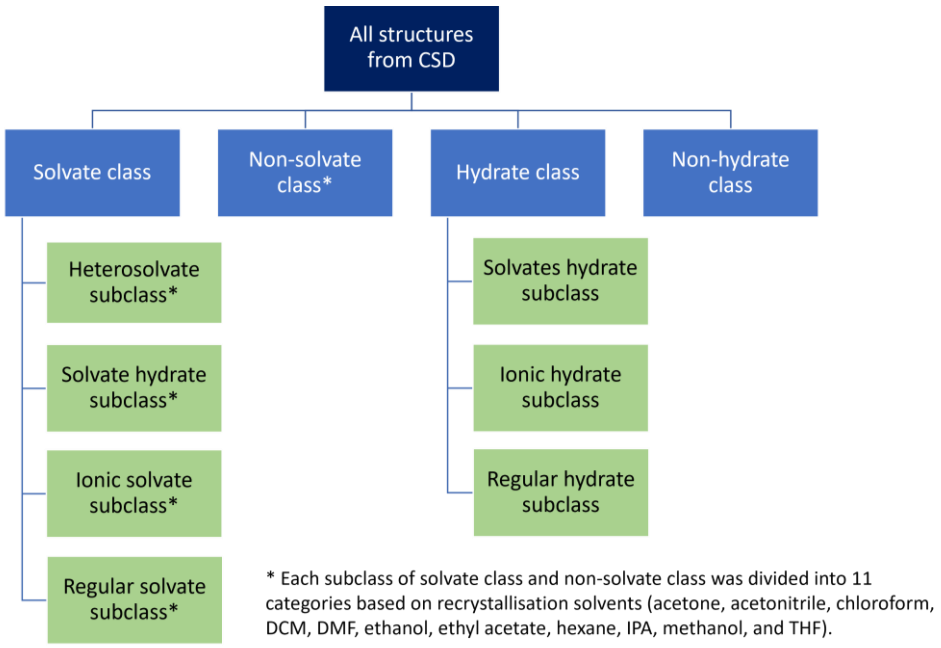


Figure 43. *Classification of the structures extracted from CSD*

In this work, the structures in each class and subclass are defined as follows:

- The non-solvate class represents the structures that do not contain any solvent molecules in their unit cell.
- The solvate class represents the structures that contain (an) organic molecule(s) and at least 1 solvent molecule in the unit cell. The solvate class is subdivided into subclasses as follows.
    - The heterosolvate subclass consists of structures containing multiple unique solvent molecules (excluding water) in their unit cell.
    - The ionic solvate subclass consists of solvates containing at least 1 cation or 1 anion.
    - The solvate hydrate subclass consists of structures containing at least 1 water molecule and at least 1 unique solvent molecule in the unit cell.
    - The solvate subclass includes structures containing only 1 unique solvent molecule.

The process for extracting the crystal structures of small organic molecules from the CSD is described in **Figure 44**.

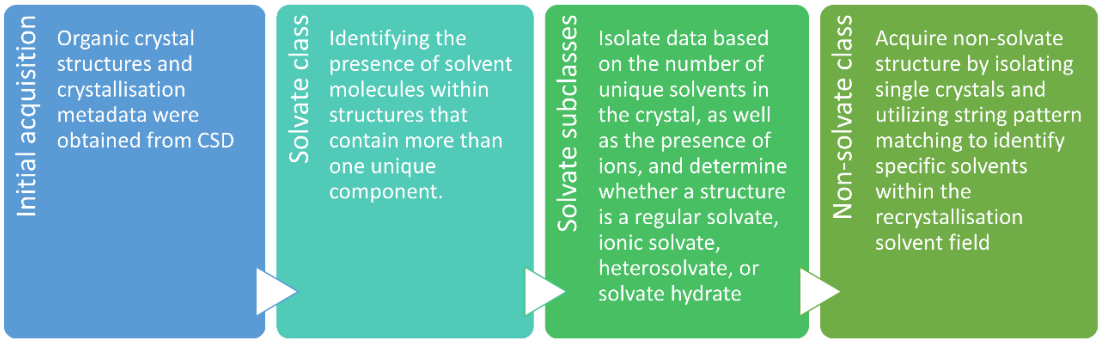


Figure 44. *Diagram presenting the process for extracting the crystal structures of small organic molecules from the CSD and classifying the structures into two classes, solvates and non-solvates. Solvate class can be specified into four solvate subclasses, namely: regular solvates; ionic solvates; heterosolvates, and solvate hydrates.*

The solvates were obtained by identifying the presence of solvent molecules within the crystal structures that contain more than 1 unique component. Then the structures were differentiated whether they are heterosolvates, solvate hydrates, ionic solvates, or regular solvates based on the number of unique solvent molecules and the presence of ions in their crystal structures. As for non-solvate structures, the structures containing only 1 unique component were isolated and string pattern matching was used to identify specific solvents within the recrystallisation solvent field.

6.4.2 *PXRD patterns from CSD and peak search*

The powder patterns were simulated and saved as .xye files consisting of the peak intensity at 2-theta ranging between 5° and 50° with an 0.02° step size. Function 'find_peaks' on Python (version 3.9.7) was applied to the xye files for searching peak positions in PXRD patterns between specific 2-theta ranges.

The datasets for peak counts at 2-theta ranging from 5° to 10° and those from 5° to 7.5° were generated via Python by determining how many peaks were in the range of interest.

6.4.3 *Random Forest Classification algorithms for the predictive design of crystal structures*

The total number of structures in all solvate and non-solvate classes is 37,304 structures (**Figure 45**). The Heterosolvate subclass has the lowest number of structures (336 structures), followed by the solvate hydrate subclass (952 structures), the ionic solvate subclass (2,706

structures), and the regular solvate subclass (9,167 structures). The number of structures in the solvate class and non-solvate class are 13,161 and 24,143 structures, respectively.

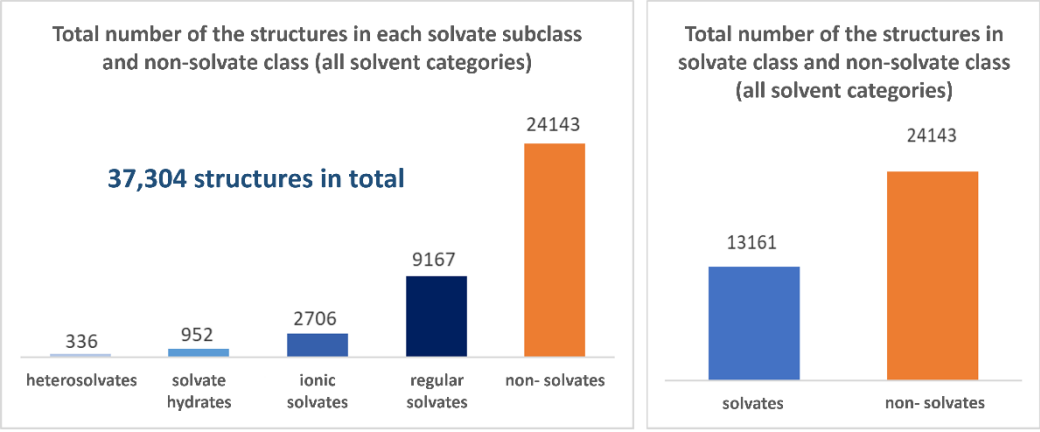


Figure 45. *Comparison of the total number of structures in all solvent categories. (a) individual solvate subclass compared to non-solvate class. (b) solvate class compared to non-solvate class*

In total, eleven models were built for testing and improving the performance of machine learning predictions. *Models 1-8* related to the prediction of solvates while *Models 9-11* were used to predict hydrate structures. *Model 1* included all available structures. The class imbalance present in *Model 1* resulted in the overprediction of the class with the largest number of structures and the underprediction of the class with the smallest number of structures. In *Model 2,* the number of predicted classes was limited to two, and all solvate subclasses were combined into one solvate class. This approach reduced the class imbalance in *Model 2* compared to *Model 1*. **Table 13** shows the details of *Model 1* and *Model 2* for the prediction of solvate structures.

Table 13. Detail of the models for the prediction of solvate structures (Model 1 – 2)

Model	Cut off	Dataset						Model's variables
		Solvate class				NS	Total	
		HS	SH	IS	RS			
1	None	336	952	2,706	9,167	24,143	37,304	25 peak data variables: Peak counts in every 2θ range of 0.2° step size, started from 5.0 and ended at 10.0 (5.0-5.2, 5.2-5.4, …, 9.6-9.8, 9.8-10.0).
2	None	13,161				24,143	37,304	

* HS = heterosolvate subclass, SH = solvate hydrate subclass, IS = ionic solvate subclass, RS = regular solvate subclass, NS = non-solvate class

For *Models 3-11*, the number of structures was reduced to balance the number of structures in each class or subclass, i.e. *Model 3* (

Table 14) which involved the prediction of all classes (5 outputs: heterosolvate subclass, solvate hydrate subclass, ionic solvate subclass, regular solvate subclass, and non-solvate class) consisted of only 336 structures in each subclass, giving the dataset with 1,680 structures in total. In *Model 4A* (

Table 14), which predicted solvate class and non-solvate class (2 outputs), each solvate class and the non-solvate class consists of 13,161 structures so the whole dataset contained 26,322 structures. The structures that exceeded the required numbers were randomly removed using the 'random' function, rand(), in Microsoft Excel. By applying this function, random numbers were created for all structures. Then the data was re-ordered in ascending order of the created random numbers. The structures at the top of the list after re-ordering were kept in the new dataset and were used for training the models. The detail of the models for predicting solvate structures is described in

Table 14 and **Table 16**. *Model 3* has 5 outputs (4 subclasses in solvate class and 1 non-solvate class) while *Model 4* and *Model 5* have 2 outputs (solvate class and non-solvate class). In *Model 4A*, all solvate subclasses (heterosolvate subclass, solvate hydrate subclass, ionic solvate subclass, regular solvate subclass) are combined to maximize the number of data in the solvate class, while for *Model 4B*, only regular solvate subclass is considered. *Model 5* is a variation of *Model 4A* with the introduction of the cut point on peak intensity. All peaks with intensity below 100 were excluded to determine if including low-intensity peaks was important for solvate prediction.

Table 14. *Detail of the models for the prediction of solvate structures (Model 3 – 5)*

Model	Cut off	Dataset						Model's variables
		Solvate class				NS	Total	
		HS	SH	IS	RS			
3	None	336	336	336	336	336	1,680	25 peak data variables: Peak counts in every 2θ range of 0.2° step size, started from 5.0 and ended at 10.0 (5.0-5.2, 5.2-5.4, …, 9.6-9.8, 9.8-10.0).
4A	None	13,161				13,161	26,322	
4B	None	-	-	-	9,167	9,167	18,334	
5	100	13,161				13,161	26,322	

* HS = heterosolvate subclass, SH = solvate hydrate subclass, IS = ionic solvate subclass, RS = regular solvate subclass, NS = non-solvate class

To examine if there is a more indicative region of 2-theta that would better differentiate solvated and non-solvated structures, the 2-theta range was expanded from the range of 5° – 10° 2-theta (for *Model 1-5)* to 5° – 15° (*Model 6*) and 5° – 20° 2-theta (*Model 7*). Detail for *Models 6* and 7 are described in **Table 15**.

Table 15. *Detail of the models for the prediction of solvate structures (model 6 – 7)*

Model	Cut off	Dataset						Model's variables
		Solvate class				NS	Total	
		HS	SH	IS	RS			
6	None	13,161				13,161	26,322	50 peak data variables: Peak counts in every 2θ range of 0.2° step size, started from 5.0 and ended at 15.0 (5.0-5.2, 5.2-5.4, …, 14.6-14.8, 14.8-15.0).
7	None	13,161				13,161	26,322	75 peak data variables: Peak counts in every 2θ range of 0.2° step size, started from 5.0 and ended at 20.0 (5.0-5.2, 5.2-5.4, …, 19.6-19.8, 19.8-20.0).

* HS = heterosolvate subclass, SH = solvate hydrate subclass, IS = ionic solvate subclass, RS = regular solvate subclass, NS = non-solvate class

In *Model 8A – 8D* (**Table 16**), the variables in the dataset were changed. The peak density was represented by introducing new variables:

- The peak positions of the 1st, 5th and 10th peaks were considered.
- The distance between the 1st and 5th and 1st and 10th peak positions was considered.
- Peak count was calculated in 2-theta intervals: 5-7.5°, 5-10°, 5-15° and 5-20°.

Model 8A used all 9 variables as mentioned above for training the model, while the variables relevant to the 10th peak and 5th peak were removed in *Model 8B* and Model 8C, respectively. For Model 8D, peak position variables were excluded.

Table 16. *Detail of the models for the prediction of solvate structures (Model 8A-8D)*

Model	Cut off	Dataset			Model's variables								
					Peak positions			Peak distance from 1st peak		#Peaks in 2θ range from 5° to n°			
		S	NS	Total	1st	5th	10th	5th	10th	7.5°	10°	15°	20°
8A	None	13,161	13,161	26,322	✓	✓	✓	✓	✓	✓	✓	✓	✓
8B					✓	✓	-	✓	-	✓	✓	✓	✓

8C					✓	-	✓	-	✓	✓	✓	✓	✓
8D					-	-	-	✓	✓	✓	✓	✓	✓

* S = solvate class, NS = non-solvate class

For the prediction of hydrate structures, *Models 9 –11* (

Table 17*)* were constructed with the same variables as *Models 1 – 5*. The total number of possible structures for each class is presented in **Figure 46**. Due to the class imbalance, some of the data were removed from the dataset using the same method as used for *Models 3 – 6*.

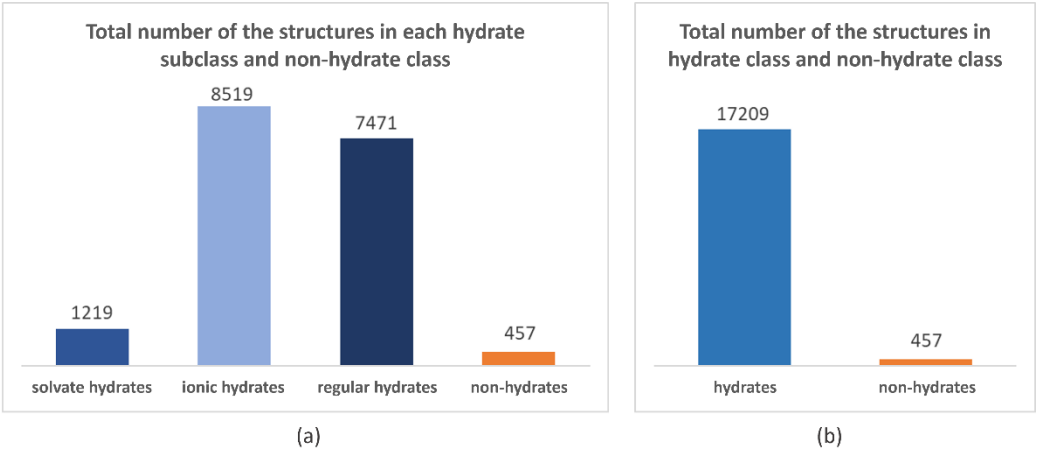


Figure 46. *Comparison of the total number of structures crystallised from water (a) individual hydrate subclass compared to non-hydrate class. (b) hydrate class compared to non-hydrate class*

Table 17. *Detail of the models for the prediction of hydrate structures (Model 9 – 11)*

Model	Cut off	Dataset					Model's variables
		Hydrate class			NH	Total	
		SH	IH	RH			
9	None	457	457	457	457	1,828	25 peak data variables: Peak counts in every 2θ range of 0.2° step size from 5° to 10° (5.0°-5.2°, 5.2°-5.4°, …, 9.6°-9.8°, 9.8°-10.0°).
10A	None	457			457	914	
10B	None	-	-	457	457	914	
11	100	457			457	914	

* SH = solvate hydrate subclass, IH = ionic hydrate subclass, RH = regular hydrate subclass, NH = non-hydrate class

6.4.4 *Model evaluation*

The n-fold cross-validation and train-test split were used to evaluate the predictive models. 4-fold, 5-fold, and 10-fold were applied to evaluate *Model 3 (the smallest dataset)* and the prediction accuracies calculated by these different folds were compared. For the train-test split, the dataset was divided into different ratios of training sets and test sets, namely 75:25,

80:20, and 90:10, which are comparable to 4-fold, 5-fold, and 10-fold cross-validation, respectively. The model evaluation by train-test split was repeated 100 times and the average accuracy was used. **Table 18** presents the model's accuracies from different evaluation methods.

Table 18. *Comparison of different n-fold cross-validation and different ratio of training and test set in train-test split method, applied to Model 3*

n-fold	**4-fold**	**5-fold**	**10-fold**
Cross-validation	35.5 % (SD = 0.02%)	36.1 % (SD = 0.03 %)	37.0 % (SD = 0.03 %)

Train/test ratio	**75% train : 25% test**	**80% train : 20% test**	**90% train : 10% test**
Train-test split	35.9 % (SD = 0.02 %)	36.3 % (SD = 0.02 %)	36.7 % (SD = 0.03 %)

Although the increasing folds and the numbers of test data improve the model accuracy, the model's accuracies from 4-fold, 5-fold, and 10-fold cross-validation, as well as from 75:25, 80:20, and 90:10 of training and test set in train-test split were not significantly different. Therefore, 4-fold cross-validation and train/test split with a 75:25 train/test ratio were selected as model evaluation methods in this work to save computational time. Additionally, the model may overfit when the ratio of training data to test data is too high. Therefore, applying the model evaluation method in which the ratio of training data to test data is relatively low while the accuracy is still comparable is considered appropriate.

6.5 Results

6.5.1 Statistical analysis of the presence of the PXRD peak at low 2-theta

From the fact that the presence of the additional atoms in solvation or hydration layer can increase the unit cell size of the crystals and cause shift in the diffraction peaks to lower 2-theta values compared to the non-solvated crystals,[256] this study was proposed from Pfizer Inc. to explore whether there is any noticeable difference between the PXRD patterns of solvated and non-solvated crystal structures of small organic molecules. The presence of low 2-theta peaks in PXRD patterns of solvated and non-solvated crystal structures was investigated.

The number of structures in each category based on recrystallisation solvents and solvate subclasses was presented in **Table 19**. The number of the structures in the solvate class is the sum of the number of structures in heterosolvate, solvate hydrate, ionic solvate, and regular solvate subclasses. The number of the structures in hydrate and non-hydrate classes (used

water as recrystallisation solvent) are highlighted in blue, and the total number of the structures in each subclass is highlighted in green (this number excludes the structures in hydrate and non-hydrate classes). These data were illustrated as pie charts presented in **Figure 47**. The total number of structures analysed is equal to 37,254. From there, 24,143 are non-solvated structures and 13,161 are solvates. Solvates are further divided into 336 heterosolvates, 952 solvate hydrates, 2706 ionic solvates and 9167 solvates that just consist of one solvent molecule in the crystal structure.

Table 19. *Summary of the number of structures based on solvate subclasses and recrystallisation solvents ordered from the category with the lowest to the highest number of the structures.*

Recrystallisation solvents	Number of structures						
	HS	SH	IS	RS	S	NS	Total
IPA	11	16	72	131	230	403	633
THF	25	18	96	340	479	316	795
DMF	25	78	106	711	920	799	1719
Hexane	12	6	9	194	221	1940	2161
Acetone	12	65	135	662	874	1725	2599
Acetonitrile	27	91	475	744	1337	1448	2785
Ethyl acetate	8	25	20	399	452	2690	3142
Chloroform	54	61	283	1553	1951	1515	3466
DCM	52	69	440	1697	2258	1931	4189
Methanol	77	389	783	1964	3213	4544	7757
Ethanol	33	134	287	772	1226	6832	8058
Water	-	1219	8519	7471	17209	457	17666
Total (excluding water)	**336**	**952**	**2706**	**9167**	**13161**	**24143**	**37304**

* HS = heterosolvate subclass, SH = solvate hydrate subclass, IS = ionic solvate subclass, RS = regular solvate subclass, S = solvate class, NS = non-solvate class

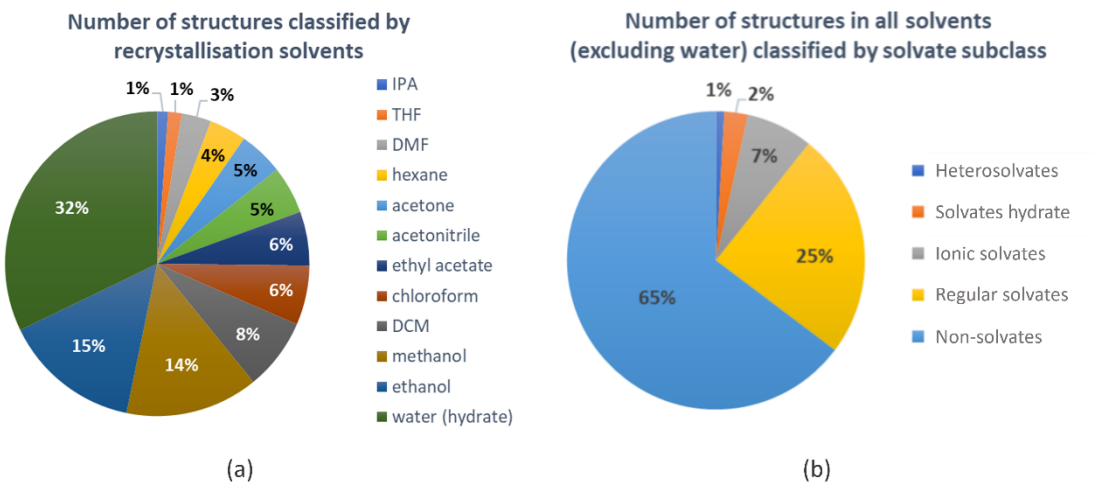


***Figure 47**. (a) Pie chart illustrating the percentage of crystal structures in 12 categories classified by recrystallisation solvents, (b) Pie chart illustrating the percentage of crystal structures in 4 solvate subclasses and non-solvate class*

Figure 47a demonstrates that the structures crystallised from water are the largest group in CSD, followed by those crystallised from ethanol and methanol, with 32%, 15%, and 14%, respectively. Non-solvate structures show the largest group compared to the crystal structures in the solvate class, as presented in **Figure 47b**.

Exported PXRD patterns of solvates were analyzed for the presence of low 2-theta peaks. The 2-theta range was defined in two ways. First, the range 5° - 10° was considered and then the 2-theta value was narrowed to 5° – 7.5°. When examining the powder patterns of solvated and non-solvated crystal structures within a wider range of 5° – 10°, it was found that the non-solvated class exhibited a lack of peaks within this range more frequently compared to the solvated class. (**Figure 48b and c**). The overall difference in peak distribution between the two classes is more pronounced when considering a narrower range of 5° – 7.5°. (**Figure 48d and e**). According to **Figure 48d** and **Figure 48e**, the pie charts show that the majority of powder patterns of solvate structures have at least one peak at a low 2-theta position, while those of non-solvate structures do not show any peak in the same 2-theta range. The same trend in the presence/absence of peaks was also found for each solvent category.

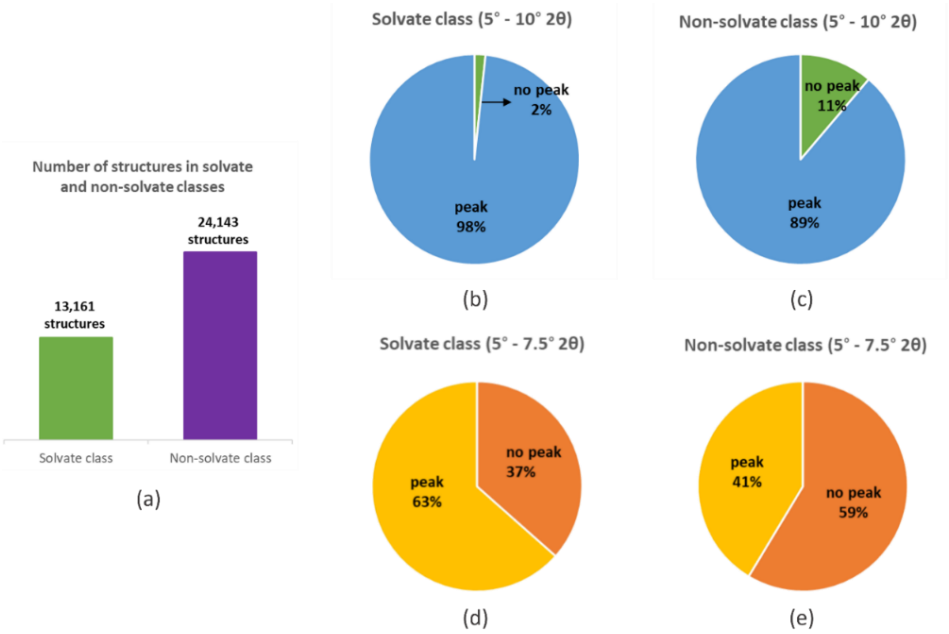


***Figure 48**. (a) The comparison between the percentage of the structures in solvate class and non-solvate class with and without PXRD peak in 2-theta ranges from 5 to 10, (b) for solvate class, (c) for non-solvate class, and 2-theta ranges from 5 to 7.5, (d) for solvate class, and (e) for non-solvate class.*

The bar chart in **Figure 49** shows the number of peaks with high, medium, and low intensity between 5° and 20° 2-theta. This chart shows that solvated structures have a higher number of peaks at all intensity levels compared to non-solvated structures.

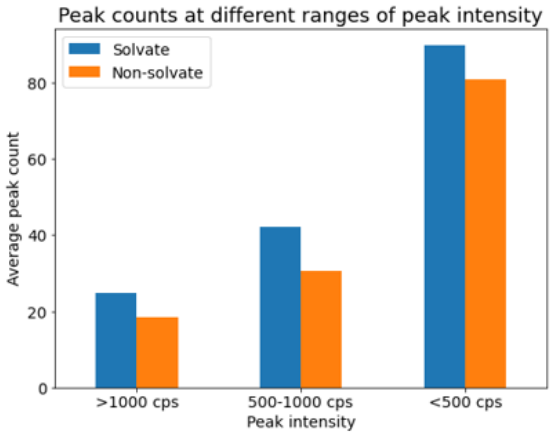


***Figure 49.** The number of peaks with high, medium, and low intensity between 5 and 20 ° 2-theta in the powder patterns of solvate (blue) and non-solvate (orange) structures*

Figure 50 and **Figure 51** compare the number of solvated and non-solvated structures for individual solvents. **Figure 50** shows solvents that form a higher percentage of solvated structures compared to non-solvated structures. **Figure 51** shows solvents that form non-solvate structures in a higher percentage compared to solvated structures. The highest

fraction of solvated structures was observed for water and THF followed by chloroform and dichloromethane. The lowest fraction of solvated structures was observed in hexane.

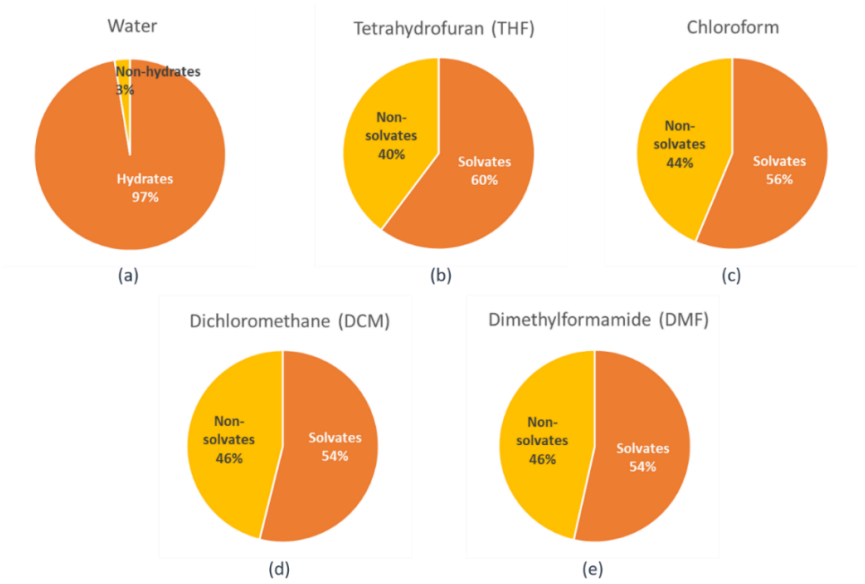


***Figure 50**. Recrystallisation solvents that form solvated structures with higher percentages compared to non-solvated structures. (a) water, (b) THF, (c) chloroform, (d) DCM, and (e) DMF (Group 1 solvents)*

When using acetonitrile, methanol, isopropyl alcohol, acetone, ethanol, ethyl acetate, and hexane, the compounds preferentially crystallised as non-solvates rather than solvates, especially hexane which only 10% of the structures form solvates. This trend could result from the relatively long hydrocarbon chain and the non-polar nature of hexane. Ethyl acetate with only 14% of solvates also has a bulky structure compared to the structures of the other solvents. However, the fraction of solvated structures observed for ethanol is lower than that of isopropyl alcohol. As isopropyl alcohol is a larger molecule, the bulkiness of the solvent molecule does not alone explain the tendency of molecules to crystallise as solvates vs. non-solvates. This suggests a complex interaction between organic molecules and recrystallisation solvents is also involved in determining this trend.

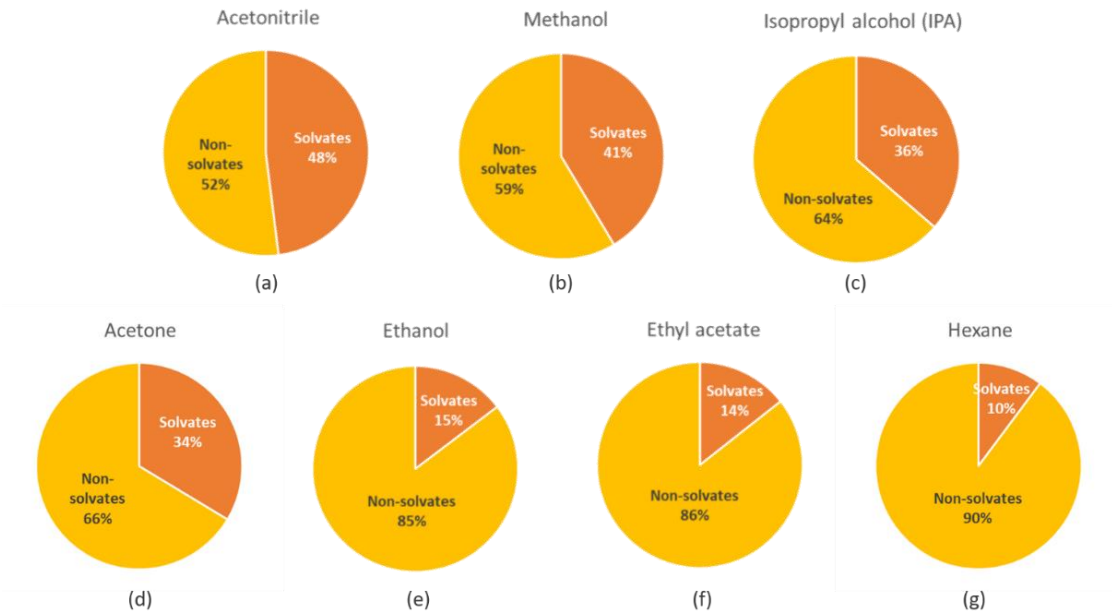


Figure 51. *Recrystallisation solvents that form non-solvated structures with higher percentages compared to solvated structures. (a) acetonitrile, (b) methanol, (c) IPA, (d) acetone, (e) ethanol, (f) ethyl acetate, and (g) hexane (Group 2 solvents)*

Hydrates are the biggest class of solvates available in the studied dataset (32%, **Figure 47**). Based on this data, molecules crystallised in water have a high propensity to form hydrates. This tendency is likely because water molecules are small, easily form hydrogen bonds, and thus are easily incorporated into crystal structures.

For the hydrate class, 47% of structures show peaks in their powder patterns at low 2-theta (5°-7.5°) and 53% show no peak (**Figure 52a**). For non-hydrated structures, 21% show peaks and 79% show no peak at the same 2-theta range (**Figure 52b**). Hydrates can be further divided into 3 subclasses **Figure 52c-e**. Of the total 17,209 hydrate structures in the dataset, 8,519 structures are ionic hydrates (48.2%), in which 48% of their PXRD pattern have peaks at low 2-theta. The solvate hydrate subclass consists of 1,219 structures (6.9%), and 72% of their PXRD patterns have peaks at low 2-theta. For the regular hydrate subclass (structures that only consist of the organic molecule and the molecule of water), the dataset is built from 7,471 structures (42.3%) where 41% have peaks at low 2-theta on the PXRD patterns.

On average, more powder patterns of hydrate and non-hydrate have no peaks between 5° and 7.5° 2-theta. Similar to solvates, the higher percentage of hydrates have peaks in the low 2-theta range compared to non-hydrates. This observation suggests that the higher density of solvate or hydrate crystal structure resulting from the incorporation of solvent or water

molecules may lead to the higher peak density in their powder patterns. Additionally, the solvate hydrate subclass has the highest percentage of powder patterns with peaks. The presence of low 2-theta peaks in this subclass may be due to the incorporation of both water and an additional solvent molecule into the crystal structure resulting in a higher density crystal.

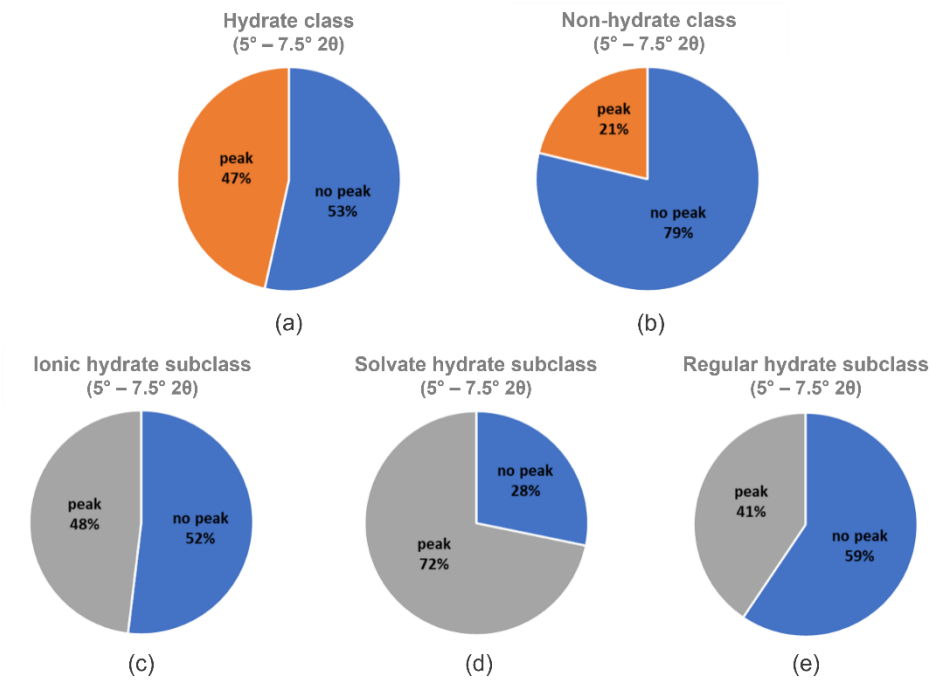


***Figure 52**. Comparison between the percentage of structures with and without PXRD peak in 2-theta ranges from 5 to 7.5: (a) hydrate class, (b) non-hydrate class, (c) ionic hydrate subclass, (d) solvate hydrate subclass, and (e) regular hydrate subclass.*

6.5.2 *Space group preferences of solvates*

The total number of space groups that the compound can crystallise in is 230. The analysis of the prefered space group from the available dataset for solvates and non-solvates suggests that six space groups (P-1, P21/c, P21/n, P212121, P21, C2/c) are prefered over the remaining 224 space groups, which agrees with the work of C. Cabeza, et al. (2007).[257] Space group P-1 is the majority of the solvate structures (28.3%, **Figure 53a**) followed by P21/c (17.0%) and P21/m (12.5%). For non-solvated structures, P21/c is the most popular space group (23.9%) followed by P-1 (20.1%) and P21/m (14.1%, **Figure 53b**). Considering the subclasses of the solvates, the majority of heterosolvates crystallised in space group P-1 (43.3%, **Figure 53c**), which is the same in solvate hydrate (25.7%, **Figure 53d**), ionic solvates (28.7%, **Figure 53e**), and non-solvates (27.9%, **Figure 53f**).

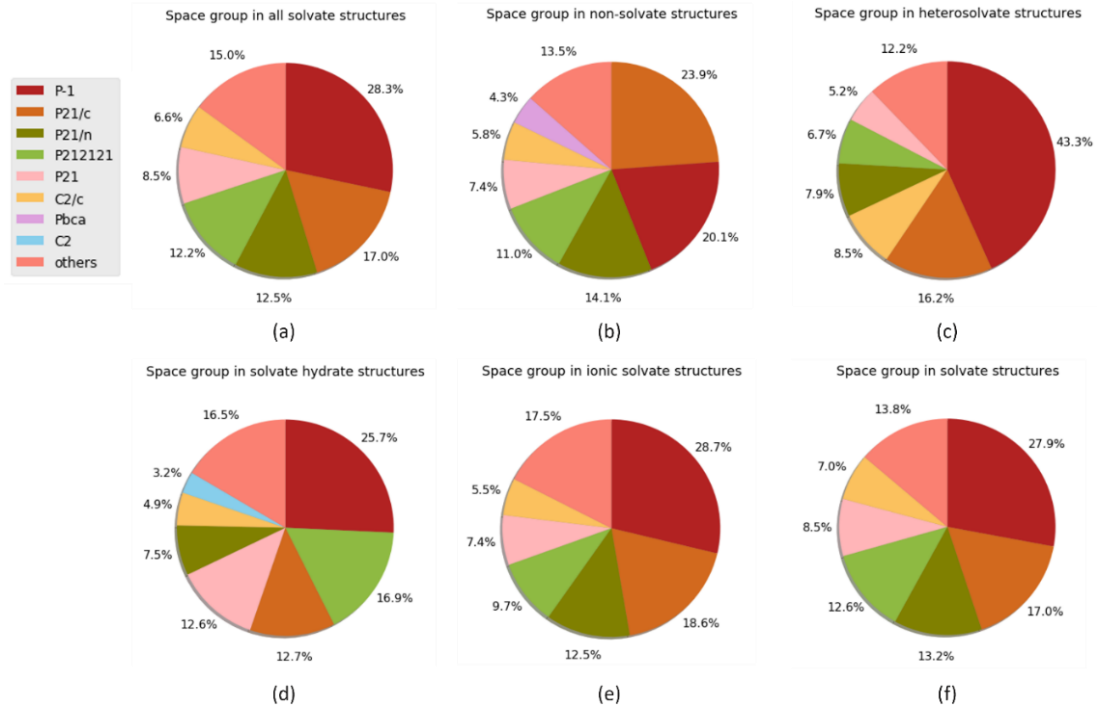


Figure 53. *Pie chart representing the percentage of space groups of crystal structures in (a) solvate class (all solvate subclasses), (b) non-solvate class (c) heterosolvate subclass, (d) solvate hydrate subclass, (e) ionic solvate subclass, and (f) regular solvate subclass. The space groups accounting for less than 3% were considered "others" and their percentages were summed up together.*

Considering individual recrystallisation solvents (**Figure 54**), THF is the only solvent in this study in which the majority of the solvated structures do not belong to the P-1 space group (**Figure 54b**). For THF solvates, most of the structures are in space group P21/c. Thus, using space groups is not, in isolation, a good predictor of solvate formation because both solvates and non-solvates crystallise in the same sets of space groups.

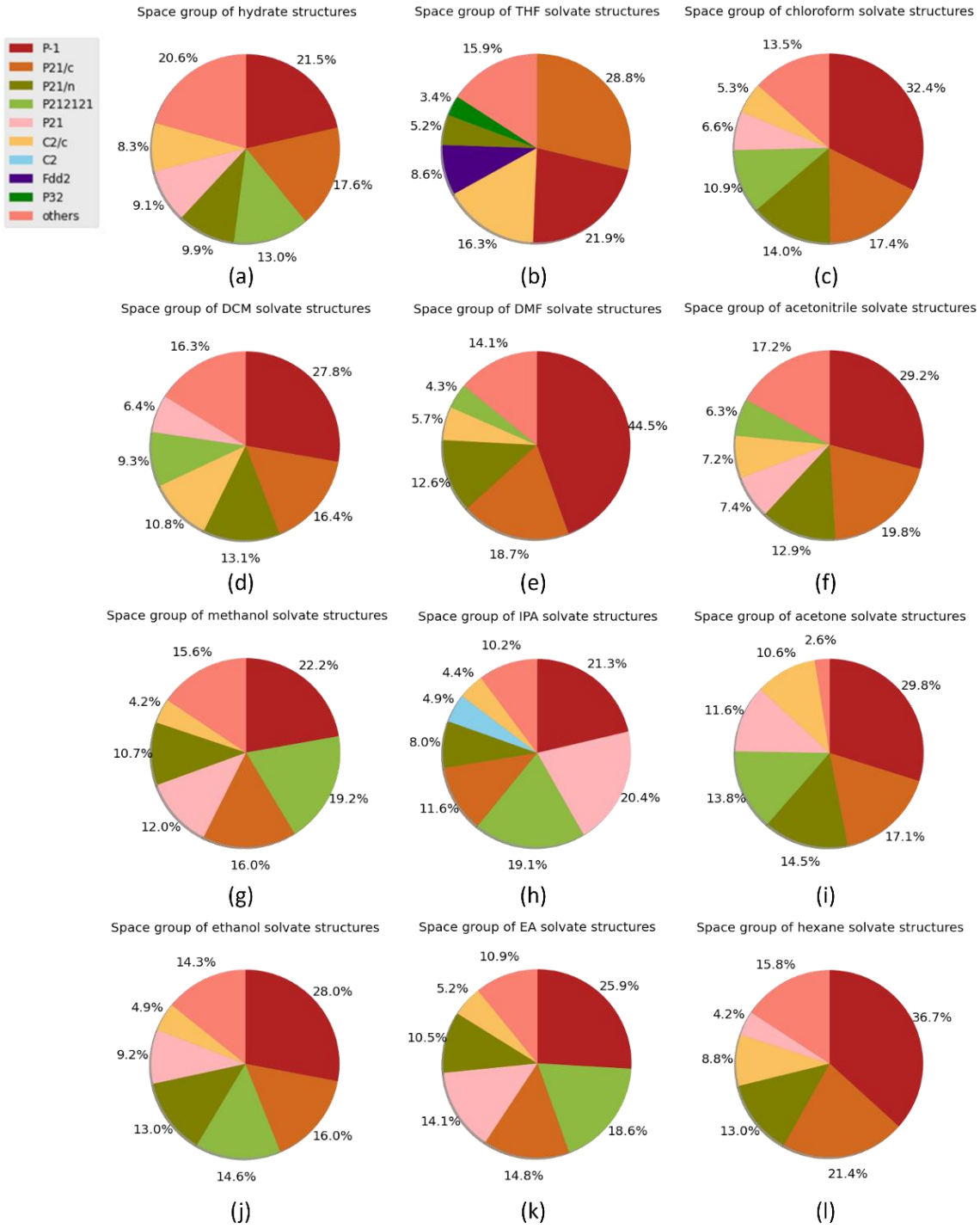


Figure 54. *Overview of the percentage of space groups of crystal structures crystallised from different solvents. (a) water, (b) THF, (c) chloroform, (d) DCM, (e) DMF, (f) acetonitrile, (g) methanol, (h) IPA, (i) acetone, (j) ethanol, (k) ethyl acetate, and (l) hexane. The space groups accounting for less than 3% were considered "others" and their percentages were summed up together.*

6.5.3 *Prior likelihoods of solvents in forming solvates*

For comparing the propensity of organic solvents in forming solvates, prior likelihood (PL), which is the probability of solvate formation from a given solvent before the nature of the solute is considered[258], was used. Prior likelihood can be calculated by **Equation 29**.

$$PL = n_{solvate} / n_{solvent} \qquad \textbf{Equation 29}$$

where: $n_{solvate}$ is the number of solvated structures recrystallized from the particular solvent, and $n_{solvent}$ is the total number of structures recrystallized from the particular solvent

Figure 50, **Figure 51,** and **Table 20** represent the prior likelihood of solvents forming solvate or hydrate structures, ordered from the highest to lowest value. Among the twelve recrystallisation solvents in this work, water has the highest prior likelihood (PL = 0.97) and hexane has the lowest likelihood with only 10% of the structures crystallised as solvates.

***Table 20**. Prior likelihood of the solvents forming solvate/hydrate structures*

Recrystallisation solvents	PL
Water	0.97
THF	0.60
Chloroform	0.56
DCM	0.54
DMF	0.54
Acetonitrile	0.48
Methanol	0.41
IPA	0.36
Acetone	0.34
Ethanol	0.15
Ethyl acetate	0.14
Hexane	0.10

As stated previously, the low likelihood of forming solvates of hexane could be because hexane has a relatively long hydrocarbon chain and is more non-polar than compared to the other solvents. To search for the correlation between the prior likelihood and solvent's properties, molecular weight, dielectric constant, density, and boiling point of the solvents were observed (**Figure 55**).

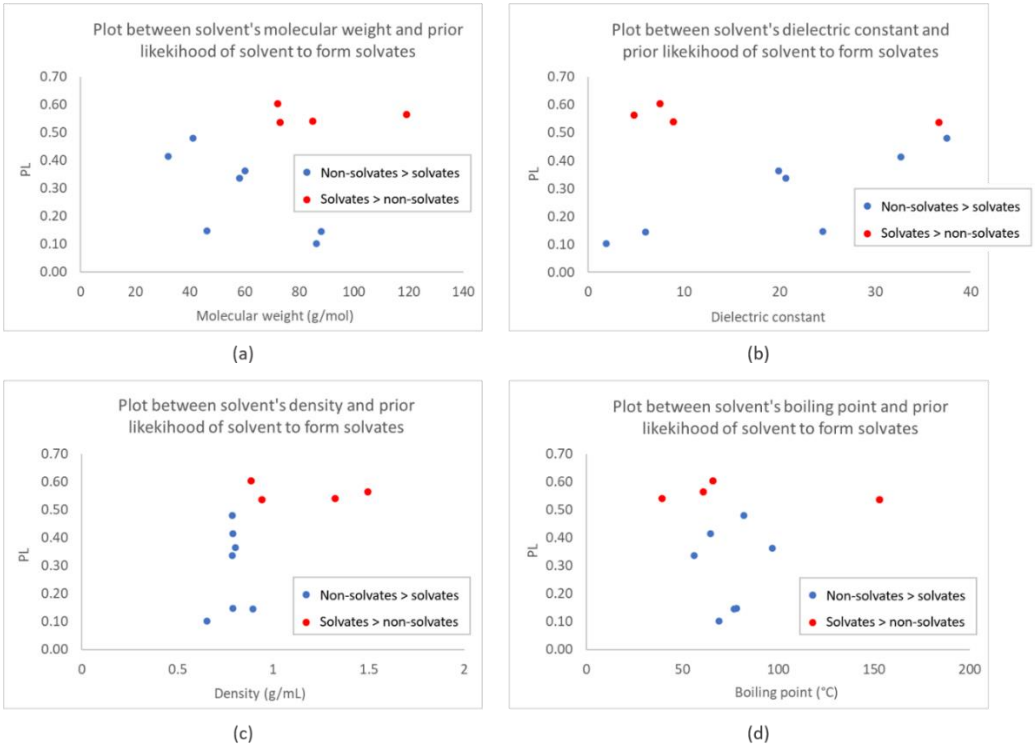


Figure 55. *The prior likelihood of solvent to form solvate vs (a) solvent's molecular weight, (b) dielectric constant, (c) density, and (d) boiling point. Blue points represent group 1 solvents and red points represent group 2 solvents (solvent groups 1 and 2 refer to the groups in* ***Figure 50*** *and* ***Figure 51****, respectively).*

Figure 55c shows that solvents with relatively high density seem to have a higher likelihood to form solvates than those with relatively low density. Most compounds crystallised from the solvents with a density above 0.9 g/mL tend to crystallise as solvates. However, the absence of correlation in all plots (**Figure 55a-d**) suggests that the likelihood of solvate formation cannot be predicted with only the solvent properties investigated here.

6.5.4 Machine learning for the prediction of solvate classes

Conducted statistical analysis did not show a significant preference for low 2-theta position or space group for solvates over non-solvated crystal structures. To increase the prediction capabilities for solvate formation, machine learning was applied. In total, eleven models were built for testing and improving the performance of machine learning predictions. *Models 1-8* relate to the prediction of solvates while *Models 9-11* were used to predict hydrate structures.

6.5.4.1 The model with class-imbalance

To test if class imbalance would affect the prediction accuracies of the model, the following two RF classification models were considered using all available data.

1. Prediction of four solvate subclasses (heterosolvate, ionic solvate, solvate hydrate and regular solvates) and non-solvate class. – *Model 1*
2. Prediction of solvate class and non-solvate class – *Model 2*

The accuracies of *Model 1* and *Model 2* are shown in **Table 21**. The detail of the models for the prediction of solvate structures for *Model 1* and *Model 2* are summarized in the methodology section (**Table 13**)

Table 21. *The prediction accuracies of Model 1 and Model 2 as calculated via train-test split and 4-fold cross-validation*

Model's accuracy	Model 1	Model 2
Average from 100 iterations of train-test split (75/25)	63.80 % (SD = 0.39 %)	68.73 % (SD = 0.34 %)
4-fold cross-validation	62.86 % (SD = 0.22 %)	67.75 % (SD = 1.15 %)

To evaluate *Model 1* and *Model 2*, the following values were considered: accuracy, precision, recall, and F-1 score. Accuracy is a ratio of the number of correctly predicted data and the total number of data in a dataset, regardless of which class a data belongs to (**Equation 7**). According to **Table 21**, Model 1 which consists of 5 possible outputs (with class imbalance) has 63.8% accuracy by train-test split method and 62.8% accuracy by 4-fold cross-validation. These values are considered high compared to a random guess of five classes which would have only approximately 20% accuracy. However, the values of precision and recall in **Table 22** show that the accuracy is falsely inflated by the class imbalance. In **Table 22**, the precision and recall of ionic solvate, solvate hydrate, and regular solvate subclasses are low, which indicates an inaccurate prediction of these subclasses. Additionally, the difference between the values of precision and recall also suggests the model's bias. In the case of *Model 1*, the value of precision is higher than recall in all solvate subclasses, which suggests that the model underpredicts the data in these subclasses. In the non-solvate class, the lower value of precision compared to recall indicates overprediction. The low value of F1-scores of each solvate subclass, especially ionic solvate and solvate hydrate subclasses, in which the value of F1-score is only 0.07, also indicate that *Model 1* cannot correctly predict the structures in all solvate subclasses. The high value of the F1-score in the non-solvate class (0.77) also

suggests that the accuracy of the model comes from the overprediction of the non-solvate class. This bias most likely results from the class imbalance presented in the training dataset.

Table 22. *Solvate subclass and non-solvate class prediction accuracy for Model 1 as represented by precision, recall, and F1-score*

Model prediction	Precision	Recall	F1-score	Support
Heterosolvate	0.69 (SD = 0.06)	0.40 (SD = 0.05)	0.51 (SD = 0.05)	84.8 (SD = 7.1)
Ionic solvate	0.19 (SD = 0.02)	0.06 (SD = 0.01)	0.09 (SD = 0.01)	676.8 (SD = 23.5)
Solvate hydrate	0.18 (SD = 0.05)	0.04 (SD = 0.01)	0.07 (SD = 0.02)	237.8 (SD = 10.8)
Solvate	0.38 (SD = 0.01)	0.23 (SD = 0.01)	0.29 (SD = 0.01)	2295.8 (SD = 38.4)
Non-solvate	0.70 (SD = 0.01)	0.88 (SD = 0.01)	0.78 (SD = 0.00)	6030.9 (SD = 46.4)
Average of all classes	0.43 (SD = 0.02)	0.32 (SD = 0.01)	0.35 (SD = 0.01)	9326.0 (SD = 0.0)

Model 2 has 2 outputs, the solvate class and the non-solvate class, in which the solvate class consists of all subclasses combined. Accuracies of 68.7% by train-test split validation and 67.8% by 4-fold cross-validation did not reflect the overall performance of the model. As seen by the precision, recall, and F1-scores in **Table 23**, although *Model 2* demonstrates superior performance in terms of overall precision, recall, and F1-score, in comparison to *Model 1* due to the lower number of class predictions, *Model 2* still exhibits an overprediction of the non-solvate class, which contains a larger number of structures, and an underprediction of the structures within the solvate class.

Table 23. *Solvate class and non-solvate class prediction accuracy for Model 2 as represented by precision, recall, and F1-score*

Model prediction	Precision	Recall	F1-score	Support
Solvate class	0.58 (SD = 0.01)	0.40 (SD = 0.01)	0.47 (SD = 0.01)	3284.4 (SD = 48.1)
Non-solvate class	0.72 (SD = 0.01)	0.84 (SD = 0.00)	0.78 (SD = 0.00)	6041.7 (SD = 48.1)
Average of all classes	0.65 (SD = 0.00)	0.62 (SD = 0.00)	0.63 (SD = 0.00)	9326.0 (SD = 0.0)

The confusion matrices for *Model 1* and *Model 2* are shown in **Figure 56(a)** and **Figure 56(b)**, respectively.

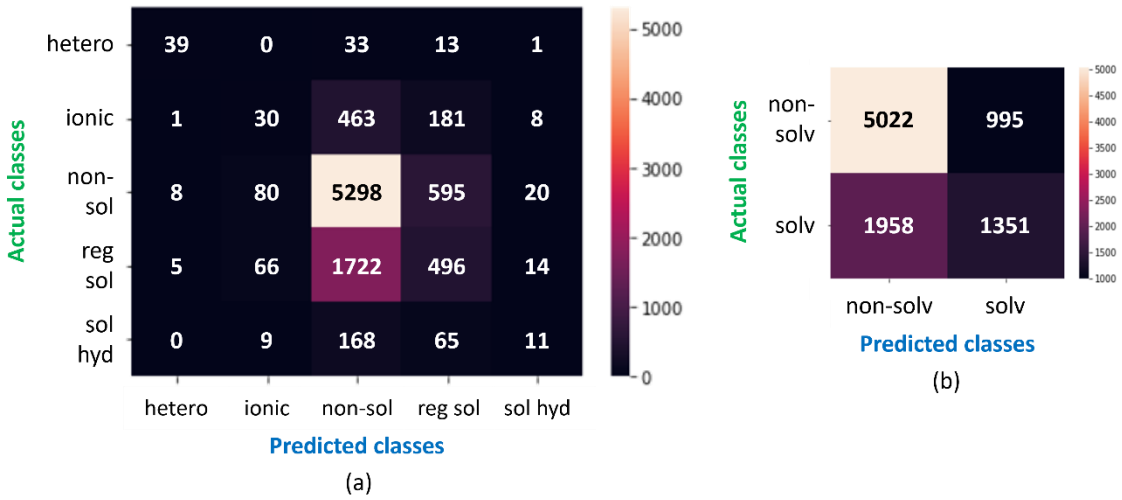


Figure 56. *The confusion matrix of the RF classification model with class imbalance for the prediction of solvate structures (a) prediction of heterosolvate subclass, ionic solvate subclass, solvate hydrate subclass, regular solvate subclass, and non-solvate class (Model 1), (b) prediction between solvate class and non-solvate class (Model 2).*

Since class imbalance highly affected the prediction of the model, some data points were randomly removed to create new datasets without class imbalance for the remaining models (*Model 3 – 11).*

6.5.4.2 The models using datasets without class-imbalance

To solve the problem of class imbalance observed in *Model 1* and *Model 2*, some data in the dataset were randomly excluded. The new dataset containing an equal number of structures in each class was used for training the models.

Model 3 and *Model 4* were used for predicting four solvate subclasses and non-solvate classes (five prediction outputs in total) and for predicting solvate class and non-solvate class (two prediction outputs), respectively. The prediction accuracies of the models are shown in **Table 24**.

Table 24. *The prediction accuracies of Model 3 and Model 4 via train-test split and 4-fold cross-validation*

Accuracies of Models 3&4	Four solvate subclasses and non-solvate class (Model 3)	Solvate class VS Non-solvate class	
		Solvate class consists of all subclasses (Model 4A)	Solvate class consists of only regular solvate subclass (Model 4B)
Average from 100 iterations of train-test split (75/25)	36.03 % (SD = 2.13 %)	65.22 % (SD = 0.53 %)	63.22 % (SD = 0.63 %)
4-fold cross-validation	35.54 % (SD = 1.88 %)	64.29 % (SD = 2.21 %)	62.43 % (SD = 2.76 %)

The accuracy of *Model 3* is around 36% (compared to 20% of random guessing accuracy. The performance of *Model 3* for the prediction of each solvate subclasses and non-solvate class is presented in **Table 25**.

Table 25. *Prediction accuracies of each solvate subclass and non-solvate class of Model 3, as represented by precision, recall, and F1-score*

Model 3	Precision	Recall	F1-score	Support
Heterosolvate	0.58 (SD = 0.05)	0.59 (SD = 0.06)	0.58 (SD = 0.04)	84.8 (SD = 7.0)
Ionic solvate	0.27 (SD = 0.04)	0.24 (SD = 0.04)	0.26 (SD = 0.03)	84.0 (SD = 7.0)
Solvate hydrate	0.27 (SD = 0.05)	0.24 (SD = 0.05)	0.26 (SD = 0.04)	84.0 (SD = 7.2)
Solvate	0.22 (SD = 0.04)	0.20 (SD = 0.04)	0.21 (SD = 0.04)	83.3 (SD = 8.0)
Non-solvate	0.40 (SD = 0.04)	0.51 (SD = 0.06)	0.44 (SD = 0.04)	83.9 (SD = 6.5)
Average of all classes	0.35 (SD = 0.02)	0.36 (SD = 0.02)	0.35 (SD = 0.02)	420.0 (SD = 0.0)

For the models predicting only 2 outputs, solvate and non-solvate classes, the accuracy of *Model 4A* is 65.2% by train-test split and 64.3% by 4-fold cross-validation (compared to 50% of the accuracy from random guessing). In *Model 4B*, only the regular solvate subclass was considered (9,167 structures), in contrast to *Model 4A* where the other subclasses (heterosolvate, hydrate solvate, and ionic solvate - 13,161 structures in total) were also combined into one class. The accuracy of *Model 4B* was lower than *Model 4A*. The decrease in accuracy of *Model 4B* may be due to the smaller dataset used for training the model.

As demonstrated by the F1-score metrics presented in **Table 26** and **Table 27**, the performance of Model 4A and Model 4B in predicting the solvate and non-solvate classes is comparable. Furthermore, these accuracies surpass the baseline of random guessing (50% as per probability rule), indicating that the models possess a significant level of predictive power. These results are particularly notable given that there is no class imbalance effect, as evidenced by the similar values of the F1-score for the solvate and non-solvate classes. These findings reflect the true performance of the models and demonstrate their efficacy in accurately predicting the solvate and non-solvate classes. It is notable that the higher value of recall for the solvate class, in comparison to the non-solvate class, may be attributed to the more complex structure present in the solvate class. This complexity may make it more challenging for the model to accurately classify samples within the solvate class. Additionally, the comparable precision of both classes may suggest that the model is generating a similar number of false predictions for both the solvate and non-solvate classes.

Table 26. *Prediction accuracies for the solvate class and non-solvate class of Model 4A, as indicated by precision, recall, and F1-score*

Model 4A	Precision	Recall	F1-score	Support
Solvate class	0.67 (SD = 0.01)	0.59 (SD = 0.01)	0.63 (SD = 0.01)	3294.2 (SD = 38.6)
Non-solvate class	0.64 (SD = 0.01)	0.71 (SD = 0.01)	0.67 (SD = 0.01)	3286.8 (SD = 38.6)
Average of all classes	0.66 (SD = 0.00)	0.65 (SD = 0.00)	0.65 (SD = 0.00)	6581.0 (SD = 0.0)

Table 27. *Prediction accuracies for the solvate class and non-solvate class of Model 4B, as represented by precision, recall, and F1-score*

Model 4B	Precision	Recall	F1-score	Support
Solvate class	0.65 (SD = 0.01)	0.58 (SD = 0.01)	0.61 (SD = 0.01)	2286.5 (SD = 34.5)
Non-solvate class	0.62 (SD = 0.01)	0.69 (SD = 0.01)	0.65 (SD = 0.01)	2297.5 (SD = 34.5)
Average of all classes	0.63 (SD = 0.01)	0.63 (SD = 0.01)	0.63 (SD = 0.01)	4584.0 (SD = 0.0)

The confusion matrices of *Models 3*, *4A*, and *4B* are shown in **Figure 57(a)**, **Figure 57(b)**, and **Figure 57(c)**, respectively.

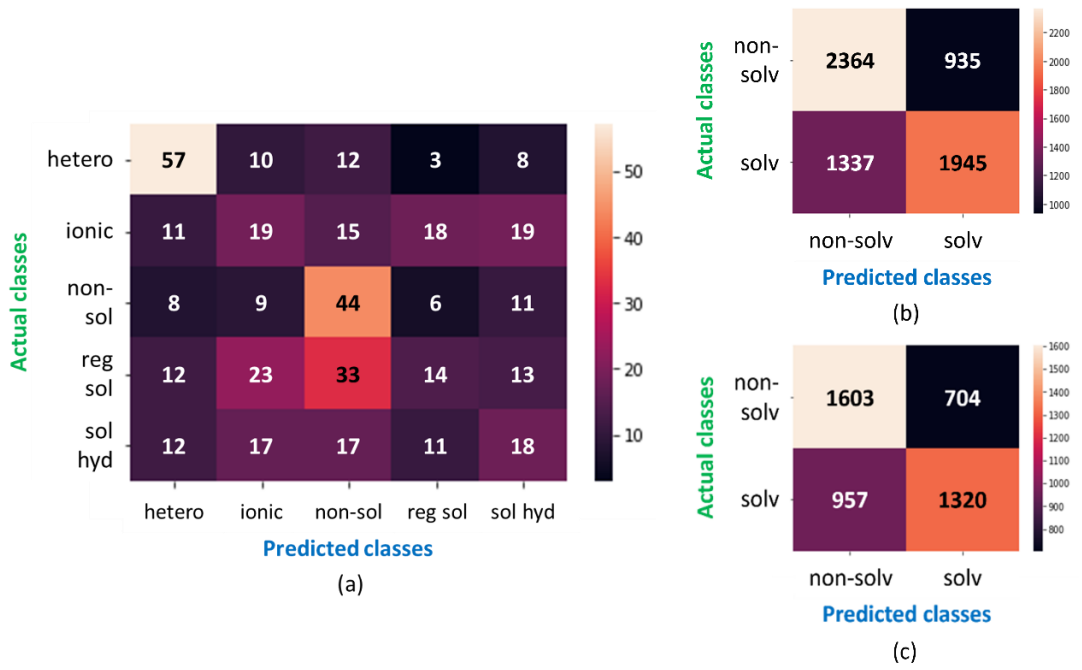


Figure 57. *The confusion matrix of the RF classification model for the prediction of solvate structures (a) prediction of heterosolvate subclass, ionic solvate subclass, solvate hydrate subclass, regular solvate subclass, and non-solvate class (Model 3), (b) prediction of solvate class and non-solvate class, in which solvate class consists of all subclasses (Model 4A), (c) prediction of solvate class and non-solvate class, in which solvate class consists of only the regular solvate subclass (Model 4B)*

RF classification allows the users to look into the importance of different variables by introducing importance scores. The importance scores of *Model 4A* (**Figure 58**) show that the 2-theta region 9.6-9.8° is the most important variable followed by lower 2-theta regions. This trend in importance scores may suggest that including data for peaks in higher 2-theta ranges may improve the prediction accuracy. However, as the peak number also increases with the higher 2-theta (**Figure 59a** and **Figure 59b**), higher feature values may falsely inflate the importance of these variables in the model's decision-making. Both of these possibilities are explored in the following sections.

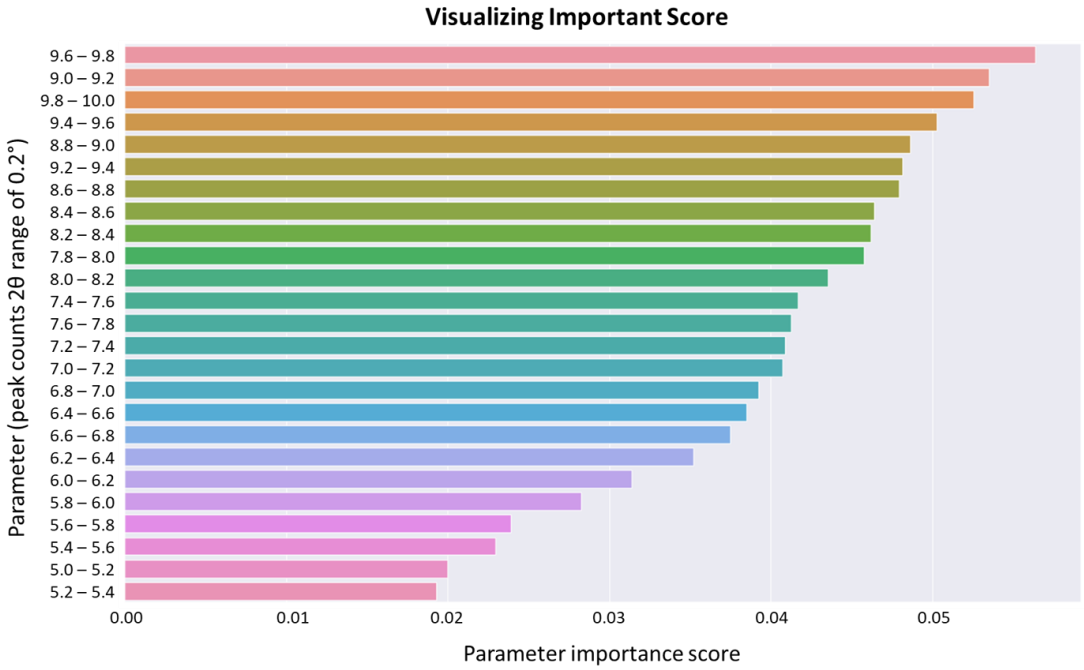


Figure 58. *Important scores of the prediction between solvate class and non-solvate class (Model 4A)*

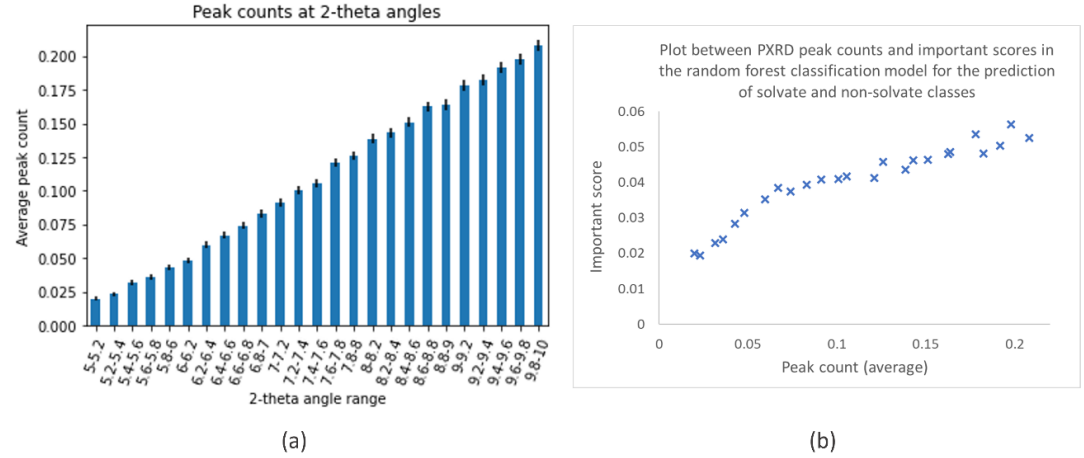
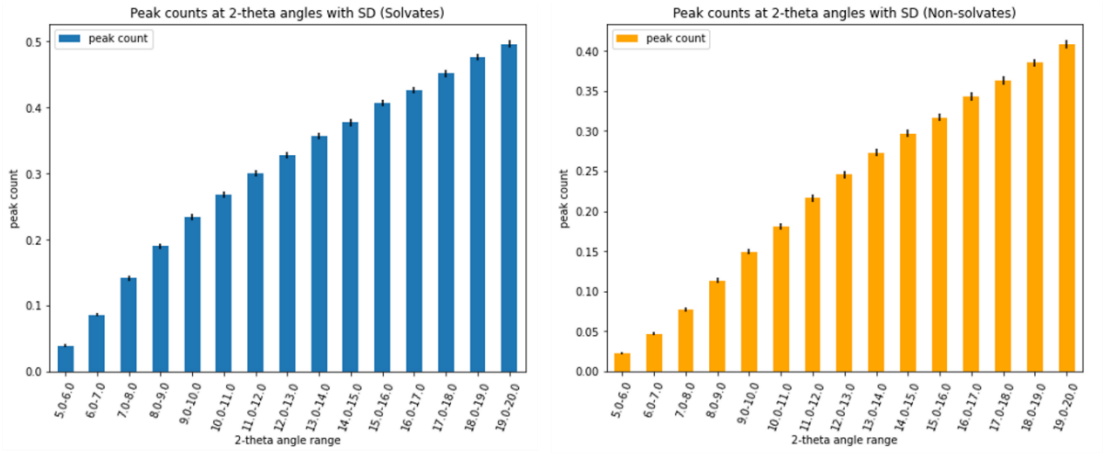


Figure 59. *(a) The average number of peaks in the PXRD patterns between 5° and 10° 2-theta. Each bar is 0.2° 2-theta range. Standard deviations (SD) are represented as the black lines on top of the bars, (b) Scattering plot between PXRD peak counts and important scores in the RF classification model for the prediction of solvate and non-solvate classes (Model 4A)*

6.5.4.3 The models using the dataset containing peak data in different 2-theta ranges

To explore including higher regions of 2-theta in the model data, Models 6 and 7 were built. Peak density for solvates and non-solvates (**Figure 60)** increases as the 2-theta values increase.



(a) (b)

Figure 60. *The average number of peaks in the PXRD pattern between 5° and 20° 2-theta. Each bar is 1° 2-theta range. Standard deviations (SD) are represented as the black lines on top of the bars. (a) solvate class, (b) non-solvate class*

The model's variables were expanded from the range of 5° – 10° 2-theta (*Model 4A*) to 5° – 15° (*Model 6*) and 5° – 20° 2-theta (*Model 7*).

Table 28. *The comparison of the accuracies of the models using the dataset containing peak data in different 2-theta ranges (Model 4A, Model 6, and Model 7)*

Model's accuracy	**2-theta ranges**		
	5° – 10° (Model 4A)	**5° – 15° (Model 6)**	**5° – 20° (Model 7)**
Average from 100 iterations of train-test split (75/25)	65.22% (SD = 5.29 %)	68.65 % (SD = 4.72 %)	69.21 % (SD = 4.93 %)
4-fold cross-validation	64.29 % (SD = 2.21 %)	67.00 % (SD = 1.86 %)	67.62 % (SD = 2.16 %)

Table 29. Comparison of the performance of Model 4A, Model 6, and Model 7 for predicting solvate class and non-solvate class, represented by precision, recall, and F1-score

Model	Model prediction	Precision	Recall	F1-score	Support
4A	Solvate class	0.67 (SD = 0.01)	0.59 (SD = 0.01)	0.63 (SD = 0.01)	3294.2 (SD = 38.6)
	Non-solvate class	0.64 (SD = 0.01)	0.71 (SD = 0.01)	0.67 (SD = 0.01)	3286.8 (SD = 38.6)
	Average of all classes	0.66 (SD = 0.00)	0.65 (SD = 0.00)	0.65 (SD = 0.00)	6581.0 (SD = 0.0)
6	Solvate class	0.68 (SD = 0.01)	0.69 (SD = 0.01)	0.69 (SD = 0.00)	3295.8 (SD = 31.9)
	Non-solvate class	0.69 (SD = 0.01)	0.68 (SD = 0.01)	0.68 (SD = 0.01)	3285.2 (SD = 31.9)
	Average of all classes	0.69 (SD = 0.01)	0.69 (SD = 0.01)	0.69 (SD = 0.01)	6581.0 (SD = 0.0)
7	Solvate class	0.69 (SD = 0.01)	0.70 (SD = 0.01)	0.70 (SD = 0.01)	3290.2 (SD = 36.6)
	Non-solvate class	0.70 (SD = 0.01)	0.69 (SD = 0.01)	0.69 (SD = 0.01)	3290.8 (SD = 36.6)
	Average of all classes	0.69 (SD = 0.00)	0.69 (SD = 0.00)	0.69 (SD = 0.00)	6581.0 (SD = 0.0)

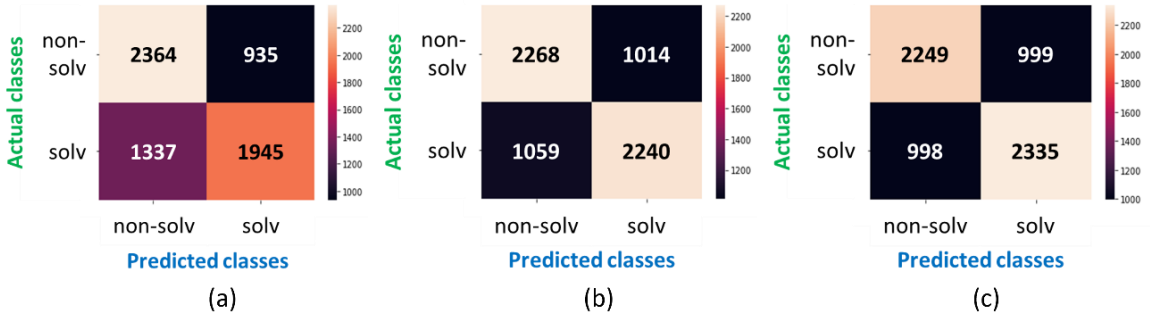


Figure 61. *Confusion matrix of (a) Model 4A, (b) Model 6, and (c) Model 7. The number of structures correctly predicted in the solvate class increased when the model has more peak data variables.*

According to the accuracy in **Table 28**, the model accuracy increased when the 2-theta range was expanded (accuracy increased from 64.29 % to 67.62 % when the 2-theta range increased from 5° – 10° to 5° – 20°). A thorough examination of the confusion matrices, precision, recall, and F1-score in **Table 29** for Model *4A (***Figure 61a**), *Model 6 (***Figure 61b***)*, and *Model 7* (**Figure 61c**) shows an improvement in model performance. This improvement is particularly evident in the F1-score of the solvate class, which increased from 0.63 in *Model 4A* to 0.69 in *Model 6* and further to 0.70 in *Model 7*. These results suggest that the inclusion of additional peak data within the range of 10° to 20° 2-theta has a positive impact on the

model's ability to predict solvate structures. These findings highlight the potential value of incorporating a broader range of X-ray diffraction data in the training process to improve the performance of models in predicting solvate class structures. As unit cell packing density is likely higher in solvated crystals, packing density differences could affect peak density throughout the spectra (not only limited to low 2-theta ranges). Thus, a higher prediction accuracy would be expected as more peak ranges are included.

6.5.4.4 A machine learning model focusing on peak density

To eliminate any bias resulting from higher feature values having more influence in model decision-making, an additional dataset was built. In this dataset, only 9 variables relevant to peak density in powder pattern were considered:

- peak positions of 1st, 5th, and 10th peak [3 variables]
- peak location's differences between 1st and 5th, and between 1st and 10th peaks [2 variables]
- peak counts in 4 different 2-theta ranges (5-7.5°, 5-10°, 5-15°, and 5-20° 2-theta) [4 variables]

This set of variables represents peak density while decreasing biases related to the tendency of peak counts to increase with higher 2-theta.

Model 8A was used for predicting solvate and non-solvate classes via RF classification. According to **Table 30**, the average accuracy of *Model 8A* from 100 iterations of train-test split (75% training data and 25% testing data) is 68.81 % (SD = 0.41 %), and the accuracy calculated by 4-fold cross-validation method is 68.74% (SD = 0.34%). **Figure 62.** shows the confusion matrix of *Model 8A*. By comparing the accuracy of *Model 8A* to the accuracy of *Model 7* (**Table 28**), the higher accuracy of *Model 8A* suggests that representing the data differently was more important to the model than increasing the 2-theta range.

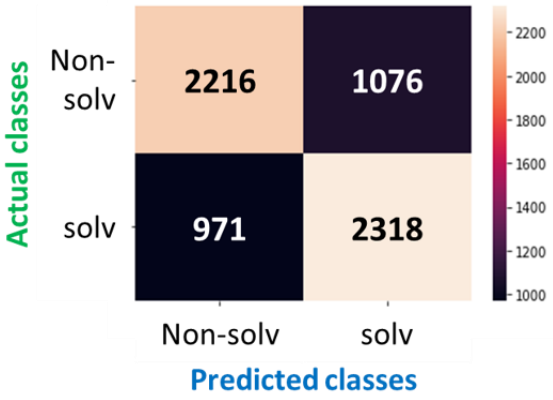


Figure 62. *The confusion matrix of the RF classification Model 8A*

For determining the importance of the variables, some variables were removed from the dataset used for training the model. The variables relevant to the 10th peak (2-theta position of 10th peak and the distance between 1st peak and 10th peak) were removed in *Model 8B*, and the variables relevant to the 5th peak (2-theta position of 5th peak and the distance between 1st peak and 5th peak) were removed in *Model 8C.* In *Model 8D*, peak position was not included in the dataset. **Table 30** presents the accuracy of each model in relation to *Model 8A*, which utilized the dataset incorporating all variables.

Table 30. *A comparison of the model accuracies for Models 8A, 8B, 8C and 8D*

Model's accuracy	**All variables (Model 8A)**	**Remove 10th peak-relevant variables (Model 8B)**	**Remove 5th peak-relevant variables (Model 8C)**	**Remove peak position variables (Model 8D)**
Average from 100 iterations of train-test split (75/25)	68.81 % (SD = 4.06 %)	68.25 % (SD = 5.02 %)	67.86 % (SD = 4.73 %)	67.62 % (SD = 5.22 %)
4-fold cross-validation	68.74 % (SD = 3.44 %)	68.43 % (SD = 3.29 %)	67.96 % (SD = 5.93 %)	67.72 % (SD = 4.90 %)

As evidenced by the prediction accuracies presented in **Table 30**, the removal of certain variables from the training dataset resulted in a very slightly decrease in the model performance, indicating that the selection of only the 2-theta positions corresponding to the 5th or 10th peak is sufficient for the model to achieve a high level of accuracy in its predictions.

To further validate this observation, Error! Not a valid bookmark self-reference. presents a comprehensive analysis of the model performance through the examination of precision, recall and F1-score values for each model. The values in the table demonstrate the comparable performance of Model 8A, 8B, 8C and 8D, further supporting the conclusion that only the 2-theta positions corresponding to the 5th or 10th peak are sufficient for the model to achieve a satisfactory level of accuracy in its predictions.

Table 31. *Comparison of the performance of Model 4A, Model 6, and Model 7 for predicting solvate class and non-solvate class, represented by precision, recall, and F1-score*

Model	Model prediction	Precision	Recall	F1-score	Support
8A	Solvate class	0.68 (SD = 0.01)	0.70 (SD = 0.01)	0.69 (SD = 0.01)	3290.2 (SD = 35.1)
	Non-solvate class	0.69 (SD = 0.01)	0.68 (SD = 0.01)	0.68 (SD = 0.01)	3290.8 (SD = 35.1)
	Average of all classes	0.69 (SD = 0.01)	0.69 (SD = 0.00)	0.69 (SD = 0.00)	6581.0 (SD = 0.0)
8B	Solvate class	0.68 (SD = 0.01)	0.70 (SD = 0.01)	0.69 (SD = 0.00)	3285.8 (SD = 29.7)
	Non-solvate class	0.69 (SD = 0.01)	0.67 (SD = 0.01)	0.68 (SD = 0.01)	3295.2 (SD = 29.7)
	Average of all classes	0.68 (SD = 0.00)	0.68 (SD = 0.00)	0.68 (SD = 0.00)	6581.0 (SD = 0.0)
8C	Solvate class	0.67 (SD = 0.01)	0.69 (SD = 0.01)	0.68 (SD = 0.01)	3290.4 (SD = 42.6)
	Non-solvate class	0.68 (SD = 0.01)	0.67 (SD = 0.01)	0.67 (SD = 0.01)	3290.6 (SD = 42.6)
	Average of all classes	0.68 (SD = 0.01)	0.68 (SD = 0.01)	0.68 (SD = 0.01)	6581.0 (SD = 0.0)
8D	Solvate class	0.67 (SD = 0.01)	0.69 (SD = 0.01)	0.68 (SD = 0.01)	3291.3 (SD = 34.1)
	Non-solvate class	0.68 (SD = 0.01)	0.66 (SD = 0.01)	0.67 (SD = 0.01)	3289.8 (SD = 34.1)
	Average of all classes	0.68 (SD = 0.00)	0.68 (SD = 0.00)	0.68 (SD = 0.00)	6581.0 (SD = 0.0)

6.5.5 Machine learning for the prediction of hydrate classes

The models for predicting hydrate structures were built in the same way as the models for predicting solvate structures. Peak counts in 0.2° 2-theta step size from 5° to 10° were used as the models' variables. *Model 9* was used to predict 3 hydrate subclasses (ionic hydrate, solvate hydrate, and regular hydrate subclasses) and non-hydrate class (four outputs in total). *Model 10A* and *Model 10B* were used for predicting the hydrate class and non-hydrate class (two outputs for each model). The difference between the dataset used in *Model 10A* and *Model 10B* was the subclass included in the hydrate class. *Model 10A* used all hydrate subclasses combined, while *Model 10B* used only the structures in the regular hydrate subclass. All models have the same number of data points in each prediction class (i.e., no class imbalance). The accuracies of the models by train-test split and 4-fold cross-validation are shown in **Table 32**. The model performances as represented by precision, recall, and F1-

scores of *Model 9*, *Model 10A*, and *Model 10B*, are shown in **Table 33**, **Table 34**, and **Table 35**, respectively. **Figure 63** presents the confusion matrix of the three models.

Table 32. *The prediction accuracies of Model 9 and Model 10 as calculated by train-test split and 4-fold cross-validation*

Model accuracy	Three hydrate subclasses and non-hydrate class (Model 9)	Hydrate class VS Non- hydrate class	
		Hydrate class consists of all subclasses (Model 10A)	Hydrate class consists of only regular hydrate subclass (Model 10B)
Average from 100 iterations of train-test split (75/25)	41.15 % (SD = 1.67 %)	64.95 % (SD = 2.66 %)	65.16 % (SD = 2.96 %)
4-fold cross-validation	40.48 % (SD = 1.66 %)	65.86 % (SD = 2.88 %)	62.90 % (SD = 3.25 %)

Table 33. *Precision, recall, and F1-scores for Model 9*

Model Prediction	Precision	Recall	F1-score	Support
Regular hydrate	0.46 (SD = 0.04)	0.63 (SD = 0.05)	0.53 (SD = 0.03)	112.9 (SD = 7.5)
Ionic hydrate	0.32 (SD = 0.04)	0.28 (SD = 0.04)	0.30 (SD = 0.03)	115.9 (SD = 8.1)
Solvate hydrate	0.48 (SD = 0.04)	0.50 (SD = 0.05)	0.49 (SD = 0.03)	112.9 (SD = 7.6)
Non-hydrate	0.31 (SD = 0.04)	0.23 (SD = 0.04)	0.26 (SD = 0.04)	115.3 (SD = 7.4)
Average of all classes	0.40 (SD = 0.02)	0.41 (SD = 0.02)	0.40 (SD = 0.02)	457.0 (SD = 0.0)

Table 34. *Precision, recall, and F1-scores for Model 10A*

Model prediction	Precision	Recall	F1-score	Support
Hydrate	0.68 (SD = 0.05)	0.54 (SD = 0.05)	0.60 (SD = 0.03)	113.8 (SD = 6.0)
Non-hydrate	0.62 (SD = 0.03)	0.75 (SD = 0.05)	0.68 (SD = 0.03)	115.2 (SD = 6.0)
Average of all classes	0.65 (SD = 0.03)	0.65 (SD = 0.03)	0.64 (SD = 0.03)	229.0 (SD = 0.0)

Table 35. *Precision, recall, and F1-scores for Model 10B*

Model prediction	Precision	Recall	F1-score	Support
Hydrate	0.67 (SD = 0.04)	0.58 (SD = 0.06)	0.62 (SD = 0.04)	113.9 (SD = 7.2)
Non-hydrate	0.63 (SD = 0.04)	0.72 (SD = 0.05)	0.67 (SD = 0.03)	115.1 (SD = 7.2)
Average of all classes	0.65 (SD = 0.03)	0.64 (SD = 0.03)	0.64 (SD = 0.03)	229.0 (SD = 0.0)

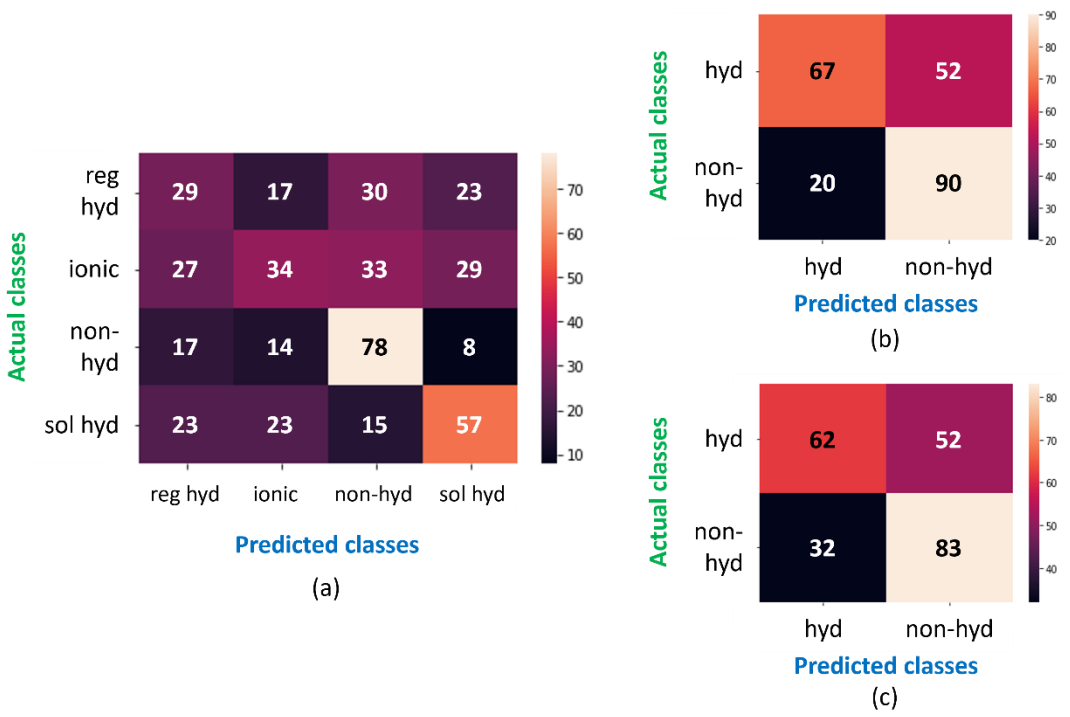


Figure 63. *The confusion matrices of the RF classification model for the prediction of hydrate structures (a) prediction of ionic hydrate subclass, solvate hydrate subclass, regular hydrate subclass, and non-hydrate class (Model 9), (b) prediction between hydrate class and non-hydrate class, in which hydrate class consists of all subclasses (Model 10A), (c) prediction between hydrate class and non-hydrate class, in which hydrate class consists of only regular hydrate subclass (Model 10B)*

As demonstrated by the prediction accuracy of 41.15% in *Model 9*, the performance of the model is superior to that of random guessing, which is 25% for four-class classification. Furthermore, the prediction accuracies of *Models 10A* and *Model 10B*, which are around 65%, are also better than random guessing for two-class classification (50%). The F1-scores presented in **Table 34** and **Table 35** suggest that both models exhibit superior performance in the prediction of the non-hydrate class as compared to the hydrate class. These findings are consistent with the performance of the similar models, namely *Models 4A* and *Model 4B*, of which the F1-scores suggest a better performance of the prediction in non-solvate class than those in solvate class, and exhibit a prediction accuracy of approximately 65%.

The important score of each variable in *Model 10A* is shown in **Figure 64**. The pattern is similar to the important scores of the prediction between the solvate class and the non-solvate class (*Model 4A* - **Figure 58**).

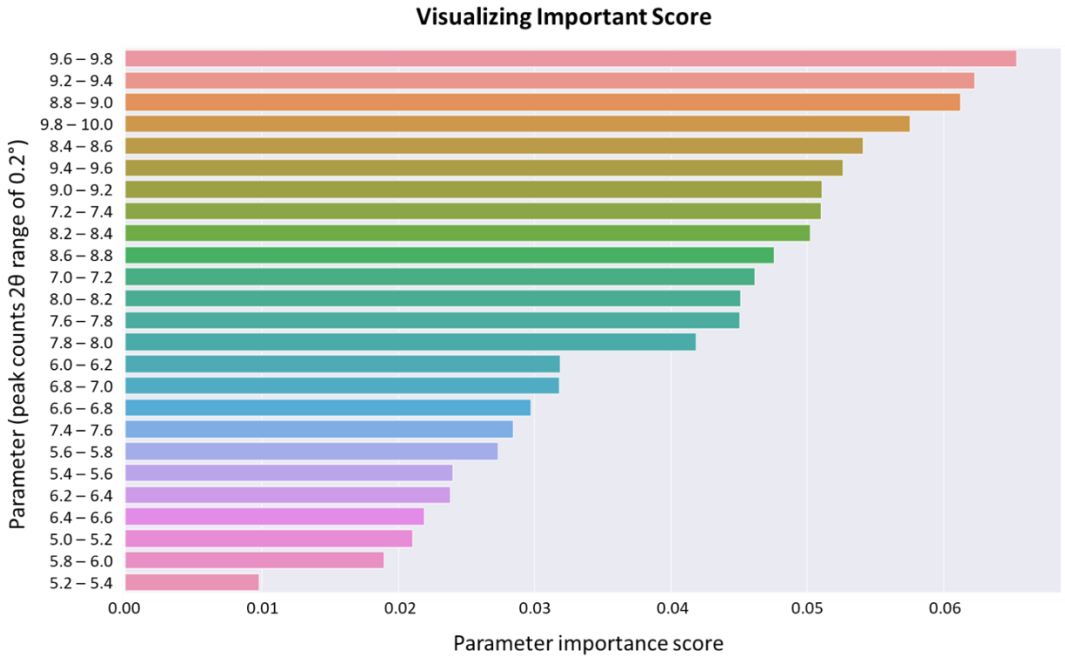


Figure 64. *Important scores of the prediction between hydrate class and non-hydrate class (Model 10A)*

6.5.6 *The effect including of weak reflection peaks in machine learning models*

To determine if including the weak reflection peaks in the powder patterns was important for the model prediction accuracy, two datasets were built:

1) no cut-off value for peak height

2) peaks included only if peak height >100 units **Table 36** shows the prediction accuracies for solvate vs. non-solvate classes and hydrate vs. non-hydrate classes for the models with and without peak height cut-offs.

Table 36. *Prediction accuracies for the RF classification models using datasets with and without the peaks with peak heights lower than 100*

Model accuracy	**Solvate class VS Non-solvate class**		**Hydrate class VS Non-hydrate class**	
	All peaks were included (Model 4A)	**Weak reflections with height < 100 were excluded (Model 5)**	**All peaks were included (Model 10A)**	**Weak reflections with height < 100 were excluded (Model 11)**
Average from 100 iterations of train-test split (75/25)	65.22 % (SD = 0.53 %)	65.53 % (SD = 0.48 %)	64.95 % (SD = 2.66 %)	64.76 % (SD = 2.67 %)
4-fold cross-validation	64.29 % (SD = 2.21 %)	64.97 % (SD = 2.54 %)	65.86 % (SD = 2.88 %)	63.35 % (SD = 1.76 %)

Table 37 and **Table 38** present the precision, recall, F1-score, and number of test data points of the models for solvate vs. non-solvate classes and hydrate vs. non-hydrate classes, respectively. The values in **Table 37** and **Table 38** are comparable to those in **Table 26** and **Table 34**, respectively.

Table 37. *Classification report of Model 5 for the prediction of solvate and non-solvate classes*

Model prediction	Precision	Recall	F1-score	Support
Solvate	0.69 (SD = 0.01)	0.57 (SD = 0.01)	0.62 (SD = 0.01)	3287.6 (SD = 35.2)
Non-solvate	0.63 (SD = 0.01)	0.74 (SD = 0.01)	0.68 (SD = 0.01)	3293.4 (SD = 35.2)
Average of all classes	0.66 (SD = 0.01)	0.65 (SD = 0.01)	0.65 (SD = 0.01)	6581.0 (SD = 0.0)

Table 38. *Classification report of Model 11 for the prediction of hydrate and non-hydrate classes*

Model prediction	Precision	Recall	F1-score	Support
Hydrate	0.68 (SD = 0.04)	0.58 (SD = 0.05)	0.62 (SD = 0.03)	114.5 (SD = 5.9)
Non-hydrate	0.63 (SD = 0.04)	0.72 (SD = 0.05)	0.67 (SD = 0.03)	114.5 (SD = 5.9)
Average of all classes	0.65 (SD = 0.03)	0.65 (SD = 0.03)	0.65 (SD = 0.03)	229.0 (SD = 0.0)

Similar model performance of *Model 4A* and *Model 10A*, as well as *Model 5* and *Model 11*, suggest that the inclusion of low-intensity peaks does not have a significant impact on the overall performance of the models. These findings suggest that the decision-making process of the models is not influenced by those weak reflections.

The accuracies and F1-score of all models are summarized in **Table 39**.

Table 39. *Summary of the model for the prediction of solvate and hydrate structures*

Model	Prediction class	Model's variables	Accuracy (4-fold CV)	Improved accuracy from random guess	F1-score
1	Heterosolvate – 336 Solvate hydrate – 952 Ionic solvate – 2,706 Regular solvate – 9,167 Non-solvate – 24,143	Peak counts in every 2θ range of 0.2°, from 5° to 10°.	62.86 %**	-	0.35
2	Solvate – 13,161 Non-solvate – 24,143	Peak counts in every 2θ range of 0.2°, from 5° to 10°.	67.75 %**	-	0.63

Table 39 (Cont.) *Summary of the model for the prediction of solvate and hydrate structures*

Model	Prediction class	Model's variables	Accuracy (4-fold CV)	Improved accuracy from random guess	F1-score
3	Heterosolvate – 336 Solvate hydrate – 336 Ionic solvate – 336 Regular solvate – 336 Non-solvate – 336	Peak counts in every 2θ range of 0.2°, from 5° to 10°.	35.54 %	15.54 %	0.35
4A	Solvate (combined all subclasses) – 13,161 Non-solvate – 13,161	Peak counts in every 2θ range of 0.2°, from 5° to 10°.	64.29 %	14.29 %	0.65
4B	Solvate (only regular subclass) – 9,167 Non-solvate – 9,167	Peak counts in every 2θ range of 0.2°, from 5° to 10°.	62.43 %	12.43 %	0.63
5	Solvate (combined all subclasses) – 13,161 Non-solvate – 13,161	Peak counts in every 2θ range of 0.2°, from 5° to 10°. Peaks with intensity lower than 100 were removed.	64.97 %	14.97 %	0.65
6	Solvate (combined all subclasses) – 13,161 Non-solvate – 13,161	Peak counts in every 2θ range of 0.2°, from 5° to 15°.	67.00 %	17.00 %	0.69
7	Solvate (combined all subclasses) – 13,161 Non-solvate – 13,161	Peak counts in every 2θ range of 0.2°, from 5° to 20°.	67.62 %	17.62 %	0.69
8A	Solvate (combined all subclasses) – 13,161 Non-solvate – 13,161	Peak position (1^{st}, 5^{th}, 10^{th}), peak distance (1^{st} and 5^{th} ,1^{st} and 10^{th}), peak count (5-7.5°, 5-10°, 5-15°, 5-20°)	68.74 %	18.74 %	0.69
8B		Peak position (1^{st}, 5^{th}), peak distance (1^{st} and 5^{th}), peak count (5-7.5°, 5-10°, 5-15°, 5-20°)	68.43 %	18.43 %	0.68
8C		Peak position (1^{st}, 10^{th}), peak distance (1^{st} and 10^{th}), peak count (5-7.5°, 5-10°, 5-15°, 5-20°)	67.96 %	17.96 %	0.68
8D		peak distance (1^{st} and 5^{th} ,1^{st} and 10^{th}), peak count (5-7.5°, 5-10°, 5-15°, 5-20°)	67.72 %	17.72 %	0.68
9	Solvate hydrate – 457 Ionic hydrate – 457 Regular hydrate – 457 Non-hydrate – 457	Peak counts in every 2θ range of 0.2°, from 5° to 10°.	40.48 %	15.48 %	0.40

Table 39 (Cont.) *Summary of the model for the prediction of solvate and hydrate structures*

Model	Prediction class	Model's variables	Accuracy (4-fold CV)	Improved accuracy from random guess	F1-score
10A	Hydrate (only regular hydrate subclass) – 457 Non-hydrate – 457	Peak counts in every 2θ range of 0.2°, from 5° to 10°.	65.86 %	15.86 %	0.64
10B	Hydrate (combined all subclasses) – 457 Non-hydrate – 457	Peak counts in every 2θ range of 0.2°, from 5° to 10°.	62.90 %	12.90 %	0.64
11	Hydrate (combined all subclasses) – 457 Non-hydrate – 457	Peak counts in every 2θ range of 0.2°, from 5° to 10°. Peaks with intensity lower than 100 were removed.	63.35 %	13.35 %	0.65

** CV = cross-validation, ** The model is bias to non-solvate class due to class-imbalance. The accuracies of Model 1 and Model 2 did not reflect the true model performance.*

6.6 Conclusions

This work investigated the use of previously established CMAC solvate libraries created by mining the CSD database to probe for the correlations between low 2-theta PXRD peak position and solvate formation. The recrystallisation solvents used in this work included acetone, acetonitrile, chloroform, DCM, DMF, ethanol, ethyl acetate, hexane, IPA, methanol, THF, and water. Most solvate and non-solvate structures had peaks in their powder patterns between 5° and 10° 2-theta values. Decreasing the 2-theta range from 5° - 10° to 5° - 7.5° better differentiated solvate vs. non-solvate structures as more solvate structure powder patterns had at least one peak between 5 to 7.5 2-theta, while more non-solvate structures do not have peaks in this 2-theta range.

Solvates were divided into two groups. Solvents that were more likely to form solvated than non-solvated crystalline structures were placed in the first group, while solvents that were more likely to form non-solvated than solvated structures were put into the second group. Group 1 solvent included: water, THF, chloroform, DCM and DMF, while group 2 solvent included acetonitrile, methanol, IPA, acetone, ethanol, ethyl acetate and hexane. In the first group, the highest number of solvated structures was observed for water and THF and the lowest for DMF. In the second group, the highest number of solvated forms was observed for acetonitrile and the lowest for hexane. These observations likely result from the fact that water and THF are small molecules that readily form hydrogen bonds and can be

incorporated into crystal structures more easily than hexane which has a long hydrophobic chain that is unable to form strong interactions with organic molecules.

The analysis of the preferred space group from the available dataset for solvates and non-solvates suggests that six space groups (P-1, P21/c, P21/n, P212121, P21, C2/c) were preferred over the remaining 224 space groups. Space group P-1 was observed for the most solvates (28.3%) followed by P21/c (17.0%) and P21/n (12.5%). For non-solvated structures, P21/c was the most common space group (23.9%) followed by P-1 (20.1%) and P21/n. When considering individual solvents, THF was the only solvent in which the majority of molecules did not crystallise in space group P-1 (for molecules crystallised in THF, the space group P21/c was most common).

Based on prior likelihood analysis, solvent density has the potential to be a predictor for solvate formation as structures were more likely to crystalize as solvates from high-density solvents. The other solvent properties considered, specifically molecular weight, boiling point, and dielectric constant, did not correlate with solvate formation.

As statistical analysis showed no significant trends in low 2-theta ranges for solvates vs non-solvated crystal structures, machine learning models were trained on the data. In summary, eleven models were developed and evaluated for their performance in predicting solvate and non-solvate, as well as hydrate and non-hydrate structures. Class imbalance was present in Models 1 and 2, resulting in overprediction of the non-solvate class. To mitigate this issue, some data were removed from the dataset in Models 3-11 to balance the classes. This led to improved precision, recall, and F1-score values for each class. However, the models for solvate prediction (Models 4A and 4B) still showed superior performance in predicting the solvate class compared to the non-solvate class, possibly due to the more complex variables present in the solvate class. The prediction accuracy of the same type of models for hydrate prediction (Models 9-11) was comparable to that of the solvate prediction models (Models 3-5). The models that excluded weak reflection peaks (Model 5 and Model 11) showed similar performance as the models that used all peaks (Model 4A and Model 10A), suggesting that the weak reflection peaks did not significantly contribute to the model's decision-making. Analysis of Models 4A and Model 10A indicated that the variables in the higher 2-theta region had higher importance scores, but these scores may have been inflated by the higher feature values present in this region.

The examination of the importance of peak data in a wider 2-theta range in PXRD patterns was conducted through the construction of *Models 6* and *Model 7*, which included additional peak data within the range of 10° to 20° 2-theta. The results of these models, as evidenced by the improved precision, recall, and F1-score, particularly in the solvate class, suggest that the inclusion of this additional peak data has a positive impact on the model's ability to predict solvate structures.

In terms of overall model accuracy, *Model 8A* emerged as the best performer with a prediction accuracy of 68.7%. However, when evaluating the overall F1-score of the model, *Model 8A* was found to be comparable to *Models 6* and *Model 7*, and only slightly superior to *Models 8B, 8C,* and *8D*. These findings suggest that the model performance can be enhanced by transforming the representation of the data to mitigate bias associated with trends in feature values. However, some additional data such as peak position which relevant to the available variables like peak distance and peak count did not have a significant impact on the performance of the model and can be excluded. This also decrease the size of the dataset, resulting to the reduced computational resources.

In summary, this work explores the use of statistical analysis and machine learning algorithm application in predicting solvate co-crystallisation. Statistical analysis of the peak distribution in the powder patterns showed that solvated crystal structures are likely to have more peaks in PXRD patterns at low 2-theta ranges than non-solvated structures. The machine learning models presented here suggest that PXRD peak positions can contribute to determining whether a PXRD pattern resulted from a solvated or non-solvated structure. While including molecular descriptors and unit cell densities as features in the machine learning models was beyond the scope of this project, including such features would likely further improve the model's accuracy.

7. Overall Conclusions

& Further works

7 Overall Conclusions & Further Works

This thesis probed the parameters relevant to the crystal nucleation process and investigated the use of machine learning to predict crystallisation outcomes.

In the first two research chapters, MFA was used in the solubility screening and cooling crystallisation studies in various organic solvents. Due to the variety of MFA crystal shapes, ranging from plate-like to needle-like crystals, MFA is a good compound for crystallisation shape screening. The research here showed that machine learning models can be implemented as a predictive tool to help guide solvent selection to achieve desirable crystal attributes and also reduce experimental time and material consumption. These models are currently restricted to MFA crystal shape prediction so future work, as discussed later in this chapter, is needed to expand model applicability to other compounds in a wider range of APIs and solvents.

In Chapter 4, the thermodynamic and kinetic parameters for the crystallisation of MFA in thirty-two organic solvents were studied. The measurement of crystallization enthalpy (ΔH^0_{cryst}) can be carried out using calorimetry, which involves determining the heat released during crystallization and scaling it with the amount of crystalline material produced at a constant temperature, T and pressure, p. However, the accuracy of this method is affected by limitations such as the cleanliness and roughness of the calorimetric cells, which can lead to crystallisation in a metastable, supersaturated solution. Therefore, an alternative method is to calculate ΔH^0_{cryst} from the solubility (C_e) of the crystals at different temperatures, using thermodynamics relation and C_e(T). This approach assumes that the heat of crystallisation is equal in magnitude but opposite in sign to the heat of dissolution. This correlation is also proved by the comparative values between the heat of crystallization and the heat of dissolution in various molecules (details are provided in the Appendix).

By determining crystallisation thermodynamics from the solubility data, it was found that the disparities in ΔH^0_{cryst}, ΔS^0_{cryst}, and ΔG^0_{cryst} among different solvents highlight the varying state of the solute, as all the crystals grown in the tested solvents belong to the same polymorphic form. Additionally, the results showed that thermodynamic parameter B derived from the plot of nucleation rate as a function of supersaturation of MFA in six studied solvents varied linearly with $\left(\Delta H^0_{cryst}\right)^3$. This observed correlation conforms with the equation of Turnbull's rule and enables us to predict the nucleation kinetics of the given

compound in solution based on their solubility data. This knowledge could help the solvent selection in the initial stage of crystallisation by determining the solution thermodynamics which reflects the interactions between the molecules of solute and solvent. In the crystallisation of MFA in six solvents, a two-step nucleation mechanism was anticipated due to the smaller value of surface free energy than the calculation based on CNT. However, other methods, such as light scattering spectroscopy or AFM, are required to support the presence of pre-nucleation clusters forming in this mechanism.

Further work can be carried out to cover more compounds in a larger variety of solvents. By increasing the number of experiments and a variety of crystallisation compounds and solvents, the correlation between thermodynamic parameter B and ΔH^0_{cryst} and ΔS^0_{cryst} may be more obvious. Additionally, future studies could explore whether certain compounds nucleate via different nucleation mechanisms (CNT and two-step nucleation) in different solvent systems. These studies would enable us to compare the differences in the thermodynamic parameters which control the energy barrier for nucleation between two different mechanisms and, thus, improve our understanding of nucleation behaviour. Furthermore, crystal growth and nucleation behaviours can also be studied by implementing analytical techniques such as light scattering spectroscopy, oblique Illumination Microscopy (OIM), and atomic force microscopy (AFM). In work done by Malkin and McPherson, light scattering techniques can be used to detect the formation of protein and viral particle aggregates in a supersaturated solution. In this study, the size of the aggregates gradually increases over time, and the increased size was assumed to correspond to the formation of metastable clusters in the two-step nucleation pathway.[259,260] Furthermore, OIM can be used to study the mesoscopic clusters by tracking the clusters' sizes and positions.[149] Additionally, AFM can be used to characterize the surface of crystals in nanoscale.[149] Implementing these additional techniques along with expanding current experiments to include more solvents and APIs would continue to improve our understanding of these systems and provide data which could be used to train future ML prediction models.

In Chapter 5, eighty-seven RF classification models were developed for the prediction of the shape of MFA crystals in a variety of crystallisation solvents. The two-class prediction model (polyhedral vs. needle) had a prediction accuracy of 93% as determined by 10-fold cross-validation. To determine the ability of the models to predict crystal shape for crystallisation solvents not present in the training set, all observations for each solvent were systematically

removed from the training data and were instead used as the test data. For these later models, different training data sets were explored. Overall, using solvent physical property descriptors in the training dataset resulted in the best model performance (as evaluated by 4-fold cross-validation), compared to using all solvent molecular descriptors or using atom count, bond count, and pharmacophore feature descriptors. When using the solvent physical properties as the model variables, the most important variables were aqueous solubility and molecular refractivity. As these variables relate to H-bonds and London dispersion forces, respectively, these results suggest that a solvent's propensity to form H-bonds and London affects the resulting crystal shape. Although the predictive models developed in this work were specific to MFA, these results suggest that with adequate data from crystallisation experiments, RF classification models could predict crystal attributes for a wider range of APIs. Thus, this study highlights the potential role of machine learning and data-driven predictive tools to support decision-making at the initial stage of pharmaceutical process development.

Further experimental crystallisation was conducted in solvents for which the models showed poor prediction accuracy when data for that solvent was used as the test set but not included in the training set. In these experiments, MFA crystals grown from trimethylamine exhibited a distinct PXRD pattern which could not be identified as polymorphic form-I, II, or III. Thus, these crystals were likely a new solvated form of MFA from triethylamine not previously present in the literature. The poor performance of the machine learning model for trimethylamine may have therefore resulted from the presence of a polymorphic form not present in the rest of the dataset.

The RF models built in Chapter 5 focused only on the prediction of the crystal shape of MFA crystallised from 30 organic solvents. While MFA is a good compound for the study of crystal shape due to the variety of crystal shapes dependent on the crystallisation solvent and supersaturation, studying the crystallisation of more organic compounds is also of interest. Moreover, studying only MFA has limited the number of solvents studied to those suitable for MFA crystallisation. By studying the crystallisation of more organic compounds, we could explore a more diverse range of solvents, which, in turn, would improve our understanding of the effect of different solvent properties on the shape of crystals of various organic compounds. Additionally, other crystallisation parameters such as temperature or stirring rate could be varied to investigate their influences on crystal shape. Other crystallisation

techniques, such as evaporation, anti-solvent crystallisation, or crystallisation from melt, could also be included in the study. These proposed studies would provide additional data for future ML models to both improve model performance and expand the types of predictions made by the models. Furthermore, the correlation between molecular descriptors as well as other crystallisation process parameters and the shape of targeted crystals could be observed. The results from these studies could facilitate solvent selection and the control of the crystallisation process to obtain the crystals with desired shapes. Undertaking experiments that explored all of these parameters would be facilitated by automated platforms for high-throughput crystallisation screening. ML imaging analysis could be implemented for the automated identification of crystal shapes in this high-throughput platform.

In Chapter 6, RF classification models were developed for predicting solvate and non-solvate crystal structures from PXRD patterns. Statistical analysis of the peak distribution in the powder patterns showed that solvated crystal structures that are likely to be denser than their non-solvated counterparts are more likely to have peaks at low 2-theta ranges. Eleven RF classification models were built and, the best model had 68.7% prediction accuracy for two-class prediction (solvates vs. non-solvates). Overall, the machine learning models presented in Chapter 6 suggest that while PXRD peak positions can contribute to determining whether or not a PXRD pattern results from a solvated or non-solvated structure, this data, in isolation, does not yield prediction accuracies above 70% and F1-score above 0.70.

The RF models built in Chapter 6 were used to predict whether a given compound forms solvates or non-solvates by considering the PXRD pattern with emphasis on the peaks in the low 2-theta region. The training dataset for these models used only the data of PXRD peak location to predict the solvated and non-solvated forms of the compounds as the inclusion of additional features was beyond the scope of this work. However, to develop the model performance, molecular descriptors of crystallisation compounds and solvents and unit cell densities could be added to the training dataset for the machine learning models. This additional data could add more value to this work by correlating the molecular descriptors to the formation of solvated structures.

This thesis expands current knowledge of thermodynamic and kinetic parameters of MFA crystallisation while also exploring further applications of machine learning models in the field of crystallisation. The work shows the potential to be expanded to the prediction of

crystallisation of other APIs in other solvents using other experimental parameters, such as crystallisation temperature, stirring rate, or even different crystallisation techniques. Improving our understanding of crystallisation and being able to predict crystal shape will facilitate us to better design the experiment prior to any wet-lab work. This work could contribute to better control of crystal attributes in pharmaceutical manufacturing by enabling us to predict and therefore avoid undesirable crystal attributes like needle-shaped crystals which cause issues in downstream pharmaceutical manufacturing processes. This informed design of crystallisation experiments for manufacturing processes could in turn cut down on the early-stage screening experiments, thereby reducing associated time, material costs and environmental impact.

8 References

1. Silva, M., Vieira, B. & Ottens, M. Preferential crystallization for the purification of similar hydrophobic polyphenols. *Journal of Chemical Technology and Biotechnology* **93**, 1997–2010 (2018).

2. Chen, W. *et al.* Biopurification of monoclonal antibody (mAb) through crystallisation. *Sep Purif Technol* **263**, 118358 (2021).

3. Verdoes, D. & Bassett, J.-M. High purity products by crystallisation. *Speciality Chemicals Magazine* 32–35 (2009).

4. Dikshit, R. C. & Chivate, M. R. Separation of ortho and para nitrochlorobenzenes by extractive crystallisation. *Chem Eng Sci* **25**, 311–317 (1970).

5. Verlag, S. New concepts for enantioselective crystallisation. *Deutsche Nationalbibliothek* (2012).

6. Li, X., Chen, W., Yang, H., Yang, Z. & Heng, J. Y. Y. Protein crystal occurrence domains in selective protein crystallisation for bio-separation. *CrystEngComm* **22**, 4566–4572 (2020).

7. Gao, Z., Rohani, S., Gong, J. & Wang, J. Recent Developments in the Crystallization Process: Toward the Pharmaceutical Industry. *Engineering* **3**, 343–353 (2017).

8. Wong, S. Y., Chen, J., Forte, L. E. & Myerson, A. S. Compact crystallization, filtration, and drying for the production of active pharmaceutical ingredients. *Org Process Res Dev* **17**, 684–692 (2013).

9. Brown, C. J. *et al.* Enabling precision manufacturing of active pharmaceutical ingredients: workflow for seeded cooling continuous crystallisations. *Mol Syst Des Eng* **3**, 518–549 (2018).

10. Siegrist, T. *et al.* Enhanced Physical Properties in a Pentacene Polymorph. *Communications* **40**, 1732–1736 (2001).

11. Antonio, M. & Maggio, R. M. Assessment of mefenamic acid polymorphs in commercial tablets using chemometric coupled to MIR and NIR spectroscopies. Prediction of dissolution performance. *J Pharm Biomed Anal* **149**, 603–611 (2018).

12. Grzesiak, A. L., Lang, M., Kim, K. & Matzger, A. J. Comparison of the Four Anhydrous Polymorphs of Carbamazepine and the Crystal Structure of Form I. *J Pharm Sci* **92**, 2260–2271 (2003).

13. Parks, C. *et al.* Molecular Dynamics Electric Field Crystallization Simulations of Paracetamol Produce a New Polymorph. *Cryst Growth Des* **17**, 3751–3765 (2017).

14. Censi, R. & Martino, P. Di. Polymorph Impact on the Bioavailability and Stability of Poorly Soluble Drugs. *molecules* **20**, 18759–18776 (2015).

15. Morissette, S. L., Soukasene, S., Levinson, D., Cima, M. J. & Almarsson, Ö. Elucidation of crystal form diversity of the HIV protease inhibitor ritonavir by high-throughput crystallization. *Proc Natl Acad Sci U S A* **100**, 2180–2184 (2003).

16. Bakar, M. R. A., Nagy, Z. K., Saleemi, A. N. & Rielly, C. D. The Impact of Direct Nucleation Control on Crystal Size Distribution in Pharmaceutical Crystallization Processes.pdf. *Cryst Growth Des* **9**, 1378–1384 (2009).

17. Liu, W. *et al.* Solvent – Solvent Cooling Crystallization : An Effective Method to Control the Morphology and Size of Ammonium Perchlorate Crystals. *Crystal Research and Technology* **54**, 1–8 (2019).

18. Mcginty, J. *et al.* Effect of Process Conditions on Particle Size and Shape in Continuous Antisolvent Crystallisation of Lovastatin. *Crystals (Basel)* **10**, 925–941 (2020).

19. Beck, R., Häkkinen, A., Malthe-Sørenssen, D. & Andreassen, J. P. The effect of crystallization conditions, crystal morphology and size on pressure filtration of l-glutamic acid and an aromatic amine. *Sep Purif Technol* **66**, 549–558 (2009).

20. Yu, Z. Q., Chew, J. W., Chow, P. S. & Tan, R. B. H. Recent advances in crystallization control: An industrial perspective. *Chemical Engineering Research and Design* **85**, 893–905 (2007).

21. Nagy, Z. K. & Braatz, R. D. Advances and new directions in crystallization control. *Annu Rev Chem Biomol Eng* **3**, 55–75 (2012).

22. Kee, N. C. S., Tan, R. B. H. & Braatz, R. D. Selective crystallization of metastable α-Form of L- glutamic acid through feedback concentration control. *Cryst Growth Des* **9**, 3044–3051 (2009).

23. Dunuwila, D. D., Carroll, L. B. & Berglund, K. A. An investigation of the applicability of attenuated total reflection infrared spectroscopy for measurement of solubility and supersaturation of aqueous citric acid solutions. *J Cryst Growth* **137**, 561–568 (1994).

24. Nagy, Z. K., Fujiwara, M. & Braatz, R. D. Monitoring and advanced control of crystallization processes. in *Handbook of Industrial Crystallization* (ed. Allan S. Myerson, E. Deniz, L. A. Y.) 313–345 (Cambridge University Press, 2019). doi:10.1017/9781139026949.011.

25. Luft, J. R., Newman, J. & Snell, E. H. Crystallization screening: The influence of history on current practice. *Acta Crystallographica Section F:Structural Biology Communications* **70**, 835–853 (2014).

26. Carlson, R. & Carlson, J. E. *Data Handling in Science and Technology*. *Elsevier Science* vol. 24 (Elsevier Science, 2005).

27. Gurung, R. Application of machine learning methods for the design of crystallisation processes. (University of Strathclyde, 2018).

28. Wengert, S., Csányi, G., Reuter, K. & Margraf, J. T. Data-efficient machine learning for molecular crystal structure prediction. *Chem Sci* **12**, 4536–4546 (2021).

29. 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).

30. Xin, D., Gonnella, N. C., He, X. & Horspool, K. Solvate Prediction for Pharmaceutical Organic Molecules with Machine Learning. *Cryst Growth Des* **19**, 1903–1911 (2019).

31. 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* **10**, 23–25 (2008).

32. Wicker, J. G. P. & Cooper, R. I. Will it crystallise? Predicting crystallinity of molecular materials. *CrystEngComm* **17**, 1927–1934 (2015).

33. Valavi, M. Crystallisation Thermodynamics. (University of Limerick, 2016). doi:10.13140/RG.2.2.22700.46726.

34. McGinty, J., Yazdanpanah, N., Price, C., Joop, H. & Sefcik, J. *The Handbook of Continuous Crystallization*. *The Handbook of Continuous Crystallization* (The Royal Society of Chemistry, 2020). doi:10.1039/9781788013581.

35. Xu, S., Wang, Y., Hou, Z. & Chuai, X. Overview of secondary nucleation: From fundamentals to application. *Ind Eng Chem Res* **59**, 18335–18356 (2020).

36. RICHARDSON, J. F. & HARKER, J. H. Crystallisation. in *Chemical Engineering* (ed. BACKHURST, J. R.) 827–900 (Elsevier, 2002). doi:10.1016/B978-0-08-049064-9.50026-6.

37. Linse, S. Monomer-dependent secondary nucleation in amyloid formation. *Biophys Rev* **9**, 329–338 (2017).

38. K.F.Kelton & A.L.Greer. Heterogeneous nucleation. in *Pergamon Materials Series* vol. 15 165–226 (2010).

39. Richardson, J.F., H. J. H. *Chemical Engineering*. vol. 2 (Butterworth-Heinemann, 2002).

40. Dirksen, J. A. & Ring, T. A. Fundamentals of crystallization: Kinetic effects on particle size distributions and morphology. *Chem Eng Sci* **46**, 2389–2427 (1991).

41. Wu, Y., Wang, D. & Li, Y. Understanding of the major reactions in solution synthesis of functional nanomaterials. *Sci China Mater* **59**, 938–996 (2016).

42. Davey, R. J., Schroeder, S. L. M. & Horst, J. H. ter. Nucleation of Organic Crystals — A Molecular Perspective Angewandte. *Angewandte Chemie - International Edition* **52**, 2166–2179 (2013).

43. Jiang, S. & Ter Horst, J. H. Crystal nucleation rates from probability distributions of induction times. *Cryst Growth Des* **11**, 256–261 (2011).

44. Vekilov, P. G. Nucleation.pdf. *Cryst Growth Des* **10**, 5007–5019 (2010).

45. Jawor-baczynska, A., Moore, D. & Sefcik, J. Effect of mixing, concentration and temperature on the formation of mesostructured solutions and their role in the nucleation of DL -valine crystals. *The Royal Society of Chemistry* **179**, 141–154 (2015).

46. Jawor-baczynska, A., Moore, B. D., Lee, S., Mccormick, A. V & Sefcik, J. Population and size distribution of solute-rich mesospecies within mesostructured aqueous amino acid solutions. *The Royal Society of Chemistry* **167**, 425–440 (2013).

47. Galkin, O. *et al.* Two-Step Mechanism of Homogeneous Nucleation of Sickle Cell Hemoglobin Polymers. *Biophys J* **93**, 902–913 (2007).

48. Tang, X. *et al.* Local structure order assisted two-step crystal nucleation in polyethylene. *Phys Rev Mater* **1**, 1–7 (2017).

49. Peng, Y. *et al.* Two-step nucleation mechanism in solid–solid phase transitions. *Nat Mater* **14**, 101–108 (2015).

50. Krautwurst, N. *et al.* Two-Step Nucleation Process of Calcium Silicate Hydrate, the Nanobrick of Cement. *Chemistry of Materials* **30**, 2895–2904 (2018).

51. Bera, M. K. & Antonio, M. R. Crystallization of Keggin Heteropolyanions via a Two-Step Process in Aqueous Solutions. *J Am Chem Soc* **138**, 7282–7288 (2016).

52. Erdemir, D., Lee, A. Y. & Myerson, A. S. Nucleation of crystals from solution: Classical and two-step models. *Acc Chem Res* **42**, 621–629 (2009).

53. Kashchiev, D. Classical nucleation theory approach to two-step nucleation of crystals. *J Cryst Growth* **530**, 1–13 (2020).

54. Burley, J. C., Duer, M. J., Stein, R. S. & Vrcelj, R. M. Enforcing Ostwald ' s rule of stages : Isolation of paracetamol forms III and II. *European Journal of Pharmaceutical Sciences* **31**, 271–276 (2007).

55. Schmelzer, J. W. P. & Abyzov, A. S. How Do Crystals Nucleate and Grow : Ostwald ' s Rule of Stages and Beyond Chapter 9 How Do Crystals Nucleate and Grow : Ostwald ' s Rule of Stages and Beyond. *Thermal Analysis and Calorimetry* **11**, 195–211 (2017).

56. Gavezzotti, A. Molecular aggregation of acetic acid in a carbon tetrachloride solution: a molecular dynamics study with a view to crystal nucleation. *Chemistry - A European Journal* **5**, 567–576 (1999).

57. Shore, Joel D.; Perchak, Dennis; Shnidman, Y. Simulations of the nucleation of AgBr from solution. *Journal of Chemical Physics* **113**, 6276–6284 (2000).

58. Georgalis, Y.; Umbach, P.; Raptis, J.; Saenger, W. Lysozyme aggregation studied by light scattering. I. Influence of concentration and nature of electrolytes. *Acta Crystallogr D Biol Crystallogr* **53**, 691–702 (1997).

59. Haas, C. & Drenth, J. The Interface between a Protein Crystal and an Aqueous Solution and Its Effects on Nucleation and Crystal Growth. *Journal of Physical Chemistry B* **104**, 368–377 (2000).

60. Cubillas, P. & Anderson, M. W. *Synthesis Mechanism: Crystal Growth and Nucleation*. *Zeolites and Catalysis: Synthesis, Reactions and Applications* vol. 1 (2010).

61. Kashchiev, D. On the critical supersaturation for nucleation. *Journal of Chemical Physics* **134**, 10–12 (2011).

62. Ichiro, S. Growth and morphology of quasicrystals. *Phase Transitions* **14**, 69–79 (1999).

63. Muller, F. L., Fielding, M. & Black, S. A Practical Approach for Using Solubility to Design Cooling Crystallisations. *Org Process Res Dev* **13**, 1315–1321 (2009).

64. Singh, B. Yield of Crystallization Process. https://www.chemicalslearning.com/2022/07/yield-of-crystallization-process.html (2022).

65. Prasad, M. R., Deb, P. K., Chandrasekaran, B., Maheshwari, R. & Tekade, R. K. Basics of Crystallization Process Applied in Drug Exploration. in *Dosage Form Design Parameters* vol. 2 67–103 (Elsevier, 2018).

66. Peter W. Cains. *Polymorphism in Pharmaceutical Solids - Google Books*. (Informa Healthcare, 2009).

67. Schall, J. M. & Myerson, A. S. Solutions and Solution Properties. in *Handbook of Industrail Crystallization* (eds. Myerson, A. S., Erdemir, D. & Lee, A. Y.) (Butterworth-Heinemann, 2002).

68. Threlfall, T. L. & Coles, S. J. A perspective on the growth-only zone, the secondary nucleation threshold and crystal size distribution in solution crystallisation. *The Royal Society of Chemistry* **18**, 369–378 (2016).

69. Jones, A. G. Crystallization principles and techniques. in *Crystallization Process Systems* 58–79 (Butterworth-Heinemann, 2002). doi:10.1016/b978-075065520-0/50004-3.

70. Kubota, N. A new interpretation of metastable zone widths measured for unseeded solutions. *J Cryst Growth* **310**, 629–634 (2008).

71. Bukovec, P., Meden, A., Smrkolj, M. & Vre˜er, F. Influence of Crystal Habit on the Dissolution of Simvastatin Single Crystals. *Acta Chim. Slov* **62**, 958–966 (2015).

72. Bukovec, P., Benkič, P., Smrkolj, M. & Vrečer, F. Effect of crystal habit on the dissolution behaviour of simvastatin crystals and its relationship to crystallization solvent properties. *Pharmazie* **71**, 263–268 (2016).

73. Ostendorf, M. Particle Engineering of an API for Improved Powder-Flow Properties Particle Engineering of an API for Improved Powder-Flow Properties. (2016).

74. Beck, R., Nyster, T. O., Enstad, G. G., Malthe-Sørenssen, D. & Andreassen, J. P. Influence of crystal properties on powder flow behavior of an aromatic amine and l-glutamic acid. *Particulate Science and Technology* **28**, 146–160 (2010).

75. Perini, G., Salvatori, F., Ochsenbein, D. R., Mazzotti, M. & Vetter, T. Filterability prediction of needle-like crystals based on particle size and shape distribution data. *Sep Purif Technol* **211**, 768–781 (2019).

76. SUN, C. & GRANT, D. J. W. Influence of Crystal Shape on the Tableting Performance of L-Lysine Monohydrochloride Dihydrate. *J Pharm Sci* **90**, 569–579 (2001).

77. Lovette, M. A. *et al. Crystal Shape Engineering*. (2008).

78. Yang, G., Kubota, N., Sha, Z., Louhi-Kultanen, M. & Wang, J. Crystal shape control by manipulating supersaturation in batch cooling crystallization. *Cryst Growth Des* **6**, 2799–2803 (2006).

79. Camacho, D. M., Roberts, K. J., Lewtas, K. & More, I. The crystal morphology and growth rates of triclinic N-docosane crystallising from N-dodecane solutions. *J Cryst Growth* **416**, 47–56 (2015).

80. Camacho, D. M. *et al.* Morphology & growth of methyl stearate as a function of crystallization environment. *Cryst Growth Des* **17**, 563–575 (2017).

81. Liang, Z. *et al.* Supersaturation controlled morphology and aspect ratio changes of benzoic acid crystals. *Comput Chem Eng* **99**, 296–303 (2017).

82. S. Sarig et al. The Effect of Supersaturation on the Crystal Characteristics of Potassium Chloride. *J. appl. Chem. Biotechnol* **28**, 663–667 (1978).

83. Finnie, S. D., Ristic, R. I., Sherwood, J. N. & Zikic, A. M. Morphological and growth rate distributions of small self-nucleated paracetamol crystals grown from pure aqueous solutions. *J Cryst Growth* **207**, 308–318 (1999).

84. Stoica, C. *et al.* Understanding the effect of a solvent on the crystal habit. *Cryst Growth Des* **4**, 765–768 (2004).

85. Wang, Y. & Liang, Z. Solvent effects on the crystal growth structure and morphology of the pharmaceutical dirithromycin. *J Cryst Growth* **480**, 18–27 (2017).

86. Lynch, A., Verma, V., Zeglinski, J., Bannigan, P. & Rasmuson, Å. Face indexing and shape analysis of salicylamide crystals grown in different solvents. *CrystEngComm* **21**, 2648–2659 (2019).

87. Wang, Y. & Liang, Z. Solvent effects and its role in quantitatively manipulating the crystal growth: Benzoic acid as case study. *CrystEngComm* **19**, 3198–3205 (2017).

88. Wang, H., Lin, Q., Dou, X., Yang, T. & Han, Y. A Different View of Solvent Effects in Crystallization. *Crystals (Basel)* **7**, 357 (2017).

89. Toro-Vazquez, J. F. & Gallegos-Infante, A. Viscosity and its relationship to crystallization in a binary system of saturated triacylglycerides and sesame seed oil. *JAOCS, Journal of the American Oil Chemists' Society* **73**, 1237–1246 (1996).

90. Jones, A. G. Particulate crystal characteristics. *Crystallization Process Systems* 1–25 (2007) doi:10.1016/b978-075065520-0/50002-x.

91. Van Rosmalen, G. M. & Bennema, P. Characterization of additive performance on crystallization: Habit modification. *J Cryst Growth* **99**, 1053–1060 (1990).

92. Lin, C. H., Gabas, N., Canselier, J. P. & Hiquily, N. Influence of additives on the growth morphology of γ-aminobutyric acid. *J Cryst Growth* **166**, 104–108 (1996).

93. Lin, C. H., Gabas, N., Canselier, J. P. & Pèpe, G. Prediction of the growth morphology of aminoacid crystals in solution I. α-Glycine. *J Cryst Growth* **191**, 791–802 (1998).

94. Dowling, R. *et al.* Acceleration of crystal growth rates: an unexpected effect of tailor-made additives. *Chemical Communications* **46**, 5924–5926 (2010).

95. Morinaga, K., Oikawa, N. & Kurita, R. Emergence of different crystal morphologies using the coffee ring effect. *Sci Rep* **8**, (2018).

96. Zeng, X. M., Martin, G. P., Marriott, C. & Pritchard, J. The Influence of Crystallization Conditions on the Morphology of Lactose Intended for Use as a Carrier for Dry Powder Aerosols. *Journal of Pharmacy and Pharmacology* **52**, 633–643 (2010).

97. Department of Chemistry (MIT). Growing Quality Crystals. https://web.mit.edu/x-ray/cystallize.html.

98. Girolami, M. Introduction. in *Introduction to Machine Learning* 1–20 (2006).

99. Dada, E. G. *et al.* Machine learning for email spam filtering: review, approaches and open research problems. *Heliyon* **5**, e01802 (2019).

100. Guzella, T. S. & Caminhas, W. M. A review of machine learning approaches to Spam filtering. *Expert Systems with Applications* vol. 36 10206–10222 Preprint at https://doi.org/10.1016/j.eswa.2009.02.037 (2009).

101. Awad, W. A. & Elseuofi, S. M. *Machine Learning methods for E-mail Classification*. *International Journal of Computer Applications* vol. 16 (2011).

102. Joseph, F. *A study on Deep Machine Learning Algorithms for diagnosis of diseases*. *International Journal of Applied Engineering Research* vol. 12 (2017).

103. Dhahri, H., Al Maghayreh, E., Mahmood, A., Elkilani, W. & Nagi, M. F. Automated Breast Cancer Diagnosis Based on Machine Learning Algorithms. (2019) doi:10.1155/2019/4253641.

104. Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V. & Fotiadis, D. I. Machine learning applications in cancer prognosis and prediction. *Computational and Structural Biotechnology Journal* vol. 13 8–17 Preprint at https://doi.org/10.1016/j.csbj.2014.11.005 (2015).

105. EBioMedicine. Machine learning in cancer diagnostics. *EBioMedicine* vol. 45 1–2 Preprint at https://doi.org/10.1016/j.ebiom.2019.07.029 (2019).

106. Sajda, P. MACHINE LEARNING FOR DETECTION AND DIAGNOSIS OF DISEASE. *Annu Rev Biomed Eng* **8**, 537–565 (2006).

107. Butler, K. T., Davies, D. W., Cartwright, H., Isayev, O. & Walsh, A. Machine learning for molecular and materials science. *Nature* **559**, 547–555 (2018).

108. Girolami, M. Decision Trees. in *Introduction to Machine Learning* 185–208 (2019).

109. Brodley, C. E. & Utgoff, P. E. Multivariate Decision Trees. *Mach Learn* **19**, 45–77 (1995).

110. Ho, T. K. Random decision forests. in *Proceedings of the International Conference on Document Analysis and Recognition, ICDAR* vol. 1 278–282 (1995).

111. Zhang, W. F. K. Bagging. *Encyclopedia of Database Systems* 206–209 (2009) doi:10.1007/978-0-387-39940-9_979.

112. Breiman, L. Random Forests. *Mach Learn* **45**, 5–32 (2001).

113. Liaw, A. & Wiener, M. Classification and Regression by randomForest. *R News* **2**, 18–22 (2002).

114. Bhardwaj, R. M., Johnston, A., Johnston, B. F. & Florence, A. J. A random forest model for predicting the crystallisability of organic molecules. *The Royal Society of Chemistry* **17**, 4272–4275 (2012).

115. Xin, D., Gonnella, N. C., He, X. & Horspool, K. Solvate Prediction for Pharmaceutical Organic Molecules with Machine Learning. *Cryst Growth Des* **19**, 1903–1911 (2019).

116. 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* **10**, 23–25 (2008).

117. Pillong, M. *et al.* A publicly available crystallisation data set and its application in machine learning. *CrystEngComm* **19**, 3737–3745 (2017).

118. Yao, T. S. *et al.* Machine Learning to Instruct Single Crystal Growth by Flux Method. *Chinese Physics Letters* **36**, 1–5 (2019).

119. Ghosh, A. *et al.* Assessment of machine learning approaches for predicting the crystallization propensity of active pharmaceutical ingredients. *The Royal Society of Chemistry* **21**, 1215–1223 (2019).

120. Pereira, F. Machine learning methods to predict the crystallization propensity of small organic molecules. *CrystEngComm* **22**, 2817–2826 (2020).

121. LaValley, M. P. Logistic regression. *Circulation* **117**, 2395–2399 (2008).

122. Park, H. A. An introduction to logistic regression: From basic concepts to interpretation with particular attention to nursing domain. *J Korean Acad Nurs* **43**, 154–164 (2013).

123. Hilbe, J. M. *Practical Guide to Logistic Regression*. *CRC Press* (Taylor & Francis Group, 2015). doi:10.18637/jss.v071.b03.

124. Sun, H., Deng, X., Wang, K. & Jin, R. Logistic regression for crystal growth process modeling through hierarchical nonnegative garrote-based variable selection. *IIE Transactions (Institute of Industrial Engineers)* **48**, 787–796 (2016).

125. Yan, S. & Wu, G. Predicting Crystallization Propensity of Proteins from Arabidopsis Thaliana. *Biol Proced Online* **17**, 1–12 (2015).

126. Shimono, E., Inoue, K., Kurita, T. & Ichiraku, Y. Logistic regression analysis for the material design of chiral crystals. *Chem Lett* **47**, 611–612 (2018).

127. L'Heureux, A., Grolinger, K., Elyamany, H. F. & Capretz, M. A. M. Machine Learning with Big Data: Challenges and Approaches. *IEEE Access* **5**, 7776–7797 (2017).

128. Zhou, L., Pan, S., Wang, J. & Vasilakos, A. V. Machine learning on big data: Opportunities and challenges. *Neurocomputing* **237**, 350–361 (2017).

129. Wuest, T., Weimer, D., Irgens, C. & Thoben, K. D. Machine learning in manufacturing: Advantages, challenges, and applications. *Prod Manuf Res* **4**, 23–45 (2016).

130. Sze, V., Chen, Y. H., Einer, J., Suleiman, A. & Zhang, Z. Hardware for machine learning: Challenges and opportunities. in *Proceedings of the Custom Integrated Circuits Conference* (Institute of Electrical and Electronics Engineers Inc., 2017). doi:10.1109/CICC.2017.7993626.

131. Jabbar, H. K. Methods to Avoid Over-fitting and Under-fitting in Supervised Machine Learning (Comparative Study). in *Computer Science, Communication & Instrumentation Devices* (eds. Stephen, A., Rohil, H. & Vasavi, S.) 163–171 (2014).

132. Goodfellow, I., Bengio, Y. & Courville, A. Machine Learning Basics. in *Deep Learning* 97–165 (MIT Press, 2016).

133. Kulkarni, S. A., Kadam, S. S., Meekes, H., Stankiewicz, A. I. & ter Horst, J. H. Crystal nucleation kinetics from induction times and metastable zone widths. *Cryst Growth Des* **13**, 2435–2440 (2013).

134. James R.Connolly. Diffraction Basics , Part 2 Overview Diffraction Basics , Part 2. *Introduction to XRay Powder Diffraction* **2**, 1–12 (2012).

135. Pecharsky, V. K. & Zavalij, P. Y. *Fundamentals of powder diffraction and structural characterization of materials*. *Fundamentals of Powder Diffraction and Structural Characterization of Materials* (Springer Science+Business Media, 2005). doi:10.1007/978-0-387-09579-0.

136. Glusker, J. P. *Crystal Structure Analysis for Chemists and Biologists*. (VCH Publishers, Inc., 1994).

137. Laue, M. Von, Bragg, W. L. & Ewald, P. P. X-ray diffraction : the contributions. (2015).

138. Raschka, S. Model Evaluation, Model Selection, and Algorithm Selection in Machine Learning. (University ofWisconsin–Madison, 2018).

139. Berrar, D. Cross-validation. in *Encyclopedia of Bioinformatics and Computational Biology: ABC of Bioinformatics* vols 1–3 542–545 (Elsevier, 2018).

140. Dalianis, H. Evaluation Metrics and Evaluation. in *Clinical Text Mining* 45–53 (Springer International Publishing, 2018). doi:10.1007/978-3-319-78503-5_6.

141. Goos, G. *et al.* Advances in Information Retrieval. in *27th European Conference on IR Research (ECIR)* (ed. Hutchison, D.) (Springer, 2005).

142. David M W Powers. *Evaluation: From Precision, Recall and F-Factor to ROC, Informedness, Markedness & Correlation*.

143. Japkowicz, N. Why question machine learning evaluation methods? (An illustrative review of the shortcomings of current methods). in *AAAI Workshop - Technical Report* vol. WS-06-06 6–11 (2006).

144. Moriwaki, H., Tian, Y. S., Kawashita, N. & Takagi, T. Mordred: A molecular descriptor calculator. *J Cheminform* **10**, 1–14 (2018).

145. Guha, R. & Willighagen, E. A Survey of Quantitative Descriptions of Molecular Structure. *Curr Top Med Chem* **12**, 1946–1956 (2013).

146. Chemical Computing Group Inc. QuaSAR-Descriptor. *1997* http://www.cadaster.eu/sites/cadaster.eu/files/challenge/descr.htm.

147. Chemical Computing Group Inc. Chemical Computing Group (CCG) | Computer-Aided Molecular Design. https://www.chemcomp.com/.

148. Chakrabarti, R. & Vekilov, P. G. Attraction between Permanent Dipoles and London Dispersion Forces Dominate the Thermodynamics of Organic Crystallization. *Cryst Growth Des* **20**, 7429–7438 (2020).

149. Warzecha, M., Safari, M. S., Florence, A. J. & Vekilov, P. G. Mesoscopic solute-rich clusters in olanzapine solutions. *Cryst Growth Des* **17**, 6668–6676 (2017).

150. Chung, S. Y., Kim, Y. M., Kim, J. G. & Kim, Y. J. Multiphase transformation and Ostwalds rule of stages during crystallization ofametal phosphate. *Nat Phys* **5**, 68–73 (2009).

151. Ostwald, W. Z. Studien über die Bildung und Umwandlung fester Körper. *Phys. Chem.* **22**, (1897).

152. Cardew, P. T., Davey, R. J. & Ruddick, A. J. Kinetics of polymorphic solid-state transformations. *Journal of the Chemical Society, Faraday Transactions 2: Molecular and Chemical Physics* **80**, 659–668 (1984).

153. Hedges, L. O. & Whitelam, S. Limit of validity of Ostwalds rule of stages in a statistical mechanical model of crystallization. *Journal of Chemical Physics* **135**, (2011).

154. Abu Bakar, M. R., Nagy, Z. K., Rielly, C. D. & Dann, S. E. Investigation of the riddle of sulfathiazole polymorphism. *Int J Pharm* **414**, 86–103 (2011).

155. Atkins, P., Paula, J. de & Keeler, J. *Physical Chemistry*. *Physical Chemistry* (Oxford University Press, 2018). doi:10.1201/9781315156910.

156. Navrotsky, A. Progress and New Directions in Calorimetry: A 2014 Perspective. *Journal of the American Ceramic Society* **97**, 3349–3359 (2014).

157. Warzecha, M., Safari, M. S., Florence, A. J. & Vekilov, P. G. Mesoscopic solute-rich clusters in olanzapine solutions. *Cryst Growth Des* **17**, 6668–6676 (2017).

158. Adam, A., Schrimpl, L. & Schmidt, P. C. Factors influencing capping and cracking of mefenamic acid tablets. *Drug Dev Ind Pharm* **26**, 489–497 (2000).

159. Panchagnula, R., Sundaramurthy, P., Pillai, O., Agrawal, S. & Raj, Y. A. Solid-State Characterization of Mefenamic Acid. *Journal of Pharmaceutical Sciences* vol. 93 1019–1029 Preprint at https://doi.org/10.1002/jps.20008 (2004).

160. Abbas, N., Oswald, I. D. H. & Pulham, C. R. Accessing mefenamic acid form II through high-pressure recrystallisation. *Pharmaceutics* **9**, 1–11 (2017).

161. Aguiar, A. J. & Zelmer, J. E. Dissolution Behavior of Polymorphs of Chloramphenicol Palmitate and Mefenamic Acid. *J Pharm Sci* **58**, 983–987 (1969).

162. Yang, X., Sarma, B. & Myerson, A. S. Polymorph control of micro/nano-sized mefenamic acid crystals on patterned self-assembled monolayer islands. *Cryst Growth Des* **12**, 5521–5528 (2012).

163. Abdul Mudalip, S. K., Abu Bakar, M. R., Jamal, P. & Adam, F. Solubility and dissolution thermodynamic data of mefenamic acid crystals in different classes of organic solvents. *J Chem Eng Data* **58**, 3447–3452 (2013).

164. Kato, F., Otsuka, M. & Matsuda, Y. Kinetic study of the transformation of mefenamic acid polymorphs in various solvents and under high humidity conditions. *Int J Pharm* **321**, 18–26 (2006).

165. SeethaLekshmi, S. & Guru Row, T. N. Conformational Polymorphism in a Non-steroidal Anti-inflammatory Drug, Mefenamic Acid. (2012) doi:10.1021/cg300812v.

166. Svartaas, T. M., Ke, W., Tantciura, S. & Bratland, A. U. Maximum Likelihood Estimation-A Reliable Statistical Method for Hydrate Nucleation Data Analysis. *Energy and Fuels* **29**, 8195–8207 (2015).

167. Svartaas, T. M., Ke, W., Tantciura, S. & Bratland, A. U. Maximum Likelihood Estimation-A Reliable Statistical Method for Hydrate Nucleation Data Analysis. *Energy and Fuels* **29**, 8195–8207 (2015).

168. General notices: Description and Solubility. in *USP-NF* (ed. Rockville, M.) (2020).

169. Vekilov, P. G., Galkin, O., Pettitt, B. M., Choudhury, N. & Nagel, R. L. Determination of the Transition-State Entropy for Aggregation Suggests How the Growth of Sickle Cell Hemoglobin Polymers can be Slowed. *Journal of Molecular Biology* vol. 377 882–888 Preprint at https://doi.org/10.1016/j.jmb.2008.01.025 (2008).

170. Choudhury, N. & Pettitt, B. M. On the mechanism of hydrophobic association of nanoscopic solutes. *J Am Chem Soc* **127**, 3556–3567 (2005).

171. Brandel, C. & Ter Horst, J. H. Measuring induction times and crystal nucleation rates. *Faraday Discuss* **179**, 199–214 (2015).

172. Archibald, E. H. & URe, W. The Density and Viscosity of Acetone at Low Tennperatures. 726–731 (1923).

173. Metz, D. J. & Glines, A. Density , Viscosity , and Dielectric Constant of Tetrahydrofuran between -78 and 30. *The Jpurnal of Physical Chemistry* 1158 (1965).

174. Song, K., Koo, J. Y. & Choi, H. C. Viscosity effect on the strategic kinetic overgrowth of molecular crystals in various morphologies: concave and octapod fullerene crystals. *RSC Adv* **11**, 20992–20996 (2021).

175. Radu, M. & Schilling, T. Solvent hydrodynamics speed up crystal nucleation in suspensions of hard spheres. *Epl* **105**, 1–7 (2014).

176. Kashchiev, D. *Nucleation: Basic Theory with Applications*. (Butterworth-Heinemann, 2000).

177. Vekilov, P. G. Dense liquid precursor for the nucleation of ordered solid phases from solution. *Cryst Growth Des* **4**, 671–685 (2004).

178. Clouet, E. Modeling of Nucleation Processes. *Fundamentals of Modeling for Metals Processing* **22**, 203–219 (2018).

179. Han, J. *et al.* Understanding Nucleation Mechanism of Mefenamic Acid: An Examination of Relation between Pre-assembly Structure in Solution and Nucleation Kinetics. *Cryst Growth Des* **21**, 6473–6484 (2021).

180. Vekilov, P. G. Non-classical Nucleation. in *Crystallization via Nonclassical Pathways Volume 1: Nucleation, Assembly, Observation & Application* 19–46 (American Chemical Society, 2020).

181. Kaissaratos, M., Filobelo, L. & Vekilov, P. G. Two-Step Crystal Nucleation Is Selected because of a Lower Surface Free Energy Barrier. *Cryst Growth Des* **21**, 5394–5402 (2021).

182. Karthika, S., Radhakrishnan, T. K. & Kalaichelvi, P. A Review of Classical and Nonclassical Nucleation Theories. *Cryst Growth Des* **16**, 6663–6681 (2016).

183. Pan, W., Kolomeisky, A. B. & Vekilov, P. G. Nucleation of ordered solid phases of proteins via a disordered high-density state: Phenomenological approach. *Journal of Chemical Physics* vol. 122 Preprint at https://doi.org/10.1063/1.1887168 (2005).

184. Kashchiev, D., Vikilov, P. G. & Kolomeisky, A. B. Kinetics of two-step nucleation of crystals.pdf. *J Chem Phys* **122**, (2005).

185. Durán-Olivencia, M. A., Yatsyshin, P., Kalliadasis, S. & Lutsko, J. F. General framework for nonclassical nucleation. *New J Phys* **20**, (2018).

186. Peterson, J. J., Snee, R. D., McAllister, P. R., Schofield, T. L. & Carella, A. J. Statistics in pharmaceutical development and manufacturing. *Journal of Quality Technology* **41**, 111–134 (2009).

187. Ding, B. Pharma Industry 4 . 0 : Literature review and research opportunities in sustainable pharmaceutical supply chains. *Process Safety and Environmental Protection* **119**, 115–130 (2018).

188. Reinhardt, I. C., Oliveira, J. C. & Ring, D. T. Journal of Industrial Information Integration Current Perspectives on the Development of Industry 4 . 0 in the Pharmaceutical Sector. *J Ind Inf Integr* **18**, 100131 (2020).

189. Marosi, G. *et al.* Pharmaceutical and Macromolecular Technologies in the Spirit of Industry 4 . 0. **62**, 457–466 (2018).

190. Zawbaa, H. M. *et al.* Computational intelligence modelling of pharmaceutical tabletting processes using bio-inspired optimization algorithms. *Advanced Powder Technology* **29**, 2966–2977 (2018).

191. Ündey, C., Ertunc, S., Mistretta, T. & Looze, B. Applied advanced process analytics in biopharmaceutical manufacturing : Challenges and prospects in real-time monitoring and control. **20**, 1009–1018 (2010).

192. Chen, Y. *et al.* Digital twins in pharmaceutical and biopharmaceutical manufacturing: A literature review. *Processes* **8**, 1088 (2020).

193. Salvalaglio, M., Vetter, T., Mazzotti, M. & Parrinello, M. Controlling and predicting crystal shapes: The case of urea. *Angewandte Chemie - International Edition* **52**, 13369–13372 (2013).

194. Chattoraj, S. & Sun, C. C. Crystal and Particle Engineering Strategies for Improving Powder Compression and Flow Properties to Enable Continuous Tablet Manufacturing by Direct Compression. *J Pharm Sci* **107**, 968–974 (2018).

195. MacLeod, C. S. & Muller, F. L. On the Fracture of Pharmaceutical Needle-Shaped Crystals during Pressure Filtration: Case Studies and Mechanistic Understanding. *Org Process Res Dev* **16**, 425–434 (2012).

196. Feng, Y., Grant, D. J. W. & Sun, C. C. Influence of crystal structure on the tableting properties of n-alkyl 4-hydroxybenzoate esters (parabens). *J Pharm Sci* **96**, 3324–3333 (2007).

197. Docherty, R., Cldesdale, G., Roberts, K. J. & Bennema, P. Application of BFDH, attachment energy and ising models to predicting and understanding the morphology of molecular crystals. *J. Phys. D: Appl. Phys.* **24**, 89–99 (1991).

198. Hartman, P. & Perdok, W. G. On the relations between structure and morphology of crystals. II. *Acta Crystallogr* **8**, 521–524 (1955).

199. Li, J., Tilbury, C. J., Kim, S. H. & Doherty, M. F. A design aid for crystal growth engineering. *Progress in Materials Science* vol. 82 1–38 Preprint at https://doi.org/10.1016/j.pmatsci.2016.03.003 (2016).

200. ter, H. J. H., Geertman, R. M., van, der H. A. E. & van, R. G. M. Solvent influence on the crystal morphology of RDX. *J. Cryst. Growth* **198/199**, 773–779 (1999).

201. Borsos, A., Majumder, A. & Nagy, Z. K. Multi-Impurity Adsorption Model for Modeling Crystal Purity and Shape Evolution during Crystallization Processes in Impure Media. *Cryst Growth Des* **16**, 555–568 (2016).

202. Sun, Y. *et al.* Modeling Olanzapine Solution Growth Morphologies. *Cryst Growth Des* **18**, 905–911 (2018).

203. Tilbury, C. J., Green, D. A., Marshall, W. J. & Doherty, M. F. Predicting the effect of solvent on the crystal habit of small organic molecules. *Cryst Growth Des* **16**, 2590–2604 (2016).

204. Bhardwaj, R. M., Johnston, A., Johnston, B. F. & Florence, A. J. A random forest model for predicting the crystallisability of organic molecules. *CrystEngComm* **17**, 4272–4275 (2015).

205. Heng, T. *et al.* Progress in Research on Arti fi cial Intelligence Applied to Polymorphism and Cocrystal Prediction. (2021) doi:10.1021/acsomega.1c01330.

206. van Eijkeren, M. A., Christiaens, G. C. M. L., Geuze, H. J., Haspels, A. A. & Sixma, J. J. Effects of mefenamic acid on menstrual hemostasis in essential menorrhagia. *Am J Obstet Gynecol* **166**, 1419–1428 (1992).

207. Ruoff, G. & Lema, M. Strategies in pain management: New and potential indications for COX-2 specific inhibitors. *Journal of Pain and Symptom Management* vol. 25 21–31 Preprint at https://doi.org/10.1016/S0885-3924(02)00628-0 (2003).

208. Heavner, J. E. & Cooper, D. M. Pharmacology of Analgesics. in *Anesthesia and Analgesia in Laboratory Animals* 97–123 (Elsevier Inc., 2008). doi:10.1016/B978-012373898-1.50008-5.

209. Modi, S. V. & Patel, Dr. J. Development and Evaluation of Self-emulsifying Drug Delivery of a Poorly Water Soluble NSAID. Preprint at (2015).

210. Sriamornsak, P., Limmatvapirat, S., Piriyaprasarth, S., Mansukmanee, P. & Huang, Z. A new self-emulsifying formulation of mefenamic acid with enhanced drug dissolution. *Asian J Pharm Sci* **10**, 121–127 (2015).

211. Macrae, C. F. *et al.* Mercury: Visualization and analysis of crystal structures. *Journal of Applied Crystallography* vol. 39 453–457 Preprint at https://doi.org/10.1107/S002188980600731X (2006).

212. Cesur, S. & Gokbel, S. Crystallization of mefenamic acid and polymorphs. *Crystal Research and Technology* **43**, 720–728 (2008).

213. Adam, A., Schrimpl, L. & Schmidt, P. C. Some physicochemical properties of mefenamic acid. *Drug Dev Ind Pharm* **26**, 477–487 (2000).

214. Su, C. S., Tang, M. & Chen, Y. P. Recrystallization of pharmaceuticals using the batch supercritical anti-solvent process. *Chemical Engineering and Processing: Process Intensification* **48**, 92–100 (2009).

215. Abdul Mudalip, S. K. *et al.* Effects of Solvents on Polymorphism and Shape of Mefenamic Acid Crystals. *MATEC Web of Conferences* **150**, 0–5 (2018).

216. Assafa, S. M., Khanfar, M. S., Obeidat, R., Salem, M. S. & Arida, A. I. Effect of different organic solvents on crystal habit of mefenamic acid. *Jordan Journal of Pharmaceutical Sciences* **2**, 150–158 (2009).

217. Cesur, S. & Gokbel, S. Crystallization of mefenamic acid and polymorphs. *Crystal Research and Technology* **43**, 720–728 (2008).

218. Assafa, S. M., Khanfar, M. S., Obeidat, R., Salem, M. S. & Arida, A. I. Effect of different organic solvents on crystal habit of mefenamic acid. *Jordan Journal of Pharmaceutical Sciences* **2**, 150–158 (2009).

219. Abdul Mudalip, S. K. *et al.* Effects of Solvents on Polymorphism and Shape of Mefenamic Acid Crystals. *MATEC Web of Conferences* **150**, 0–5 (2018).

220. Louppe, G. Understanding Random Forests: From Theory to Practice. (University of Liège, 2014).

221. Biau, G. & Scornet, E. A random forest guided tour. *Test* **25**, 197–227 (2016).

222. Pedregosa, Et & Al. Scikit-learn: Machine Learning in Python. *JMIR 12* 2825–2830 (2011).

223. Ying, X. An Overview of Overfitting and its Solutions. *J Phys Conf Ser* **1168**, (2019).

224. MOE (Molecular Operating Environment). Preprint at (2008).

225. PubChem, U.S. National Library of Medicine. https://pubchem.ncbi.nlm.nih.gov/.

226. Kohavi, R. A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. in *International Joint Conference of Artificial Intelligence* (1995).

227. Padrón, J. A. & Carrasco, R. Molecular descriptor based on a molar refractivity partition using Randic- type graph-theoretical invariant . **5**, 258–265 (2002).

228. Majumder, A. & Nagy, Z. K. Prediction and control of crystal shape distribution in the presence of crystal growth modifiers. *Chem Eng Sci* **101**, 593–602 (2013).

229. Li, J. W., Zhang, S. H., Gou, R. J., Han, G. & Chen, M. H. The effect of crystal-solvent interaction on crystal growth and morphology. *J Cryst Growth* **507**, 260–269 (2019).

230. Ter Horst, J. H., Geertman, R. M. & Van Rosmalen, G. M. The effect of solvent on crystal morphology. *J Cryst Growth* **230**, 277–284 (2001).

231. Chakrabarti, R. G. & Vekilov, P. G. Attraction between Permanent Dipoles and London Dispersion Forces Dominate the Thermodynamics of Organic Crystallization. *Cryst Growth Des* (2020) doi:10.1021/acs.cgd.0c01102.

232. Shekunov, B. Y. & York, P. Crystallization processes in pharmaceutical technology and drug delivery design. *J Cryst Growth* **211**, 122–136 (2000).

233. Sanii, R. *et al.* Toward an Understanding of the Propensity for Crystalline Hydrate Formation by Molecular Compounds. Part 2. *Cryst Growth Des* **21**, 4927–4939 (2021).

234. King, K. & England, J. F. Chloral hydrate (Noctec) overdose. *Medical Journal of Australia* **2**, 260–260 (1983).

235. Contardi, M. *et al.* Transparent ciprofloxacin-povidone antibiotic films and nanofiber mats as potential skin and wound care dressings. *European Journal of Pharmaceutical Sciences* **104**, 133–144 (2017).

236. Gopinath, H. *et al.* Formulation and evalution of Levofloxacin hemihydrate immediate release tablets Charecterization of Ayurvedic nanomedicine Abhrak Bhasma View project Formulation and evalution of Levofloxacin hemihydrate immediate release tablets. *Journal of Chemical and Pharmaceutical Sciences* **7**, 233–235 (2012).

237. Bhardwaj, R. M. *et al.* Exploring the experimental and computed crystal energy landscape of olanzapine. *Cryst Growth Des* **13**, 1602–1617 (2013).

238. 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).

239. Braun, D. E. *et al.* Solid-state forms of β-resorcylic acid: How exhaustive should a polymorph screen be? *Cryst Growth Des* **11**, 210–220 (2011).

240. Chavda, V. P. & Shah, D. Self-emulsifying delivery systems: One step ahead in improving solubility of poorly soluble drugs. in *Nanostructures for Cancer Therapy* 653–718 (Elsevier, 2017). doi:10.1016/B978-0-323-46144-3.00025-8.

241. Bhatia, A. *et al.* Polymorphism and its Implications in Pharmaceutical Product Development. *Dosage Form Design Parameters* **2**, 31–65 (2018).

242. Bē Rziņš, A., Kons, A., Saršū Ns, K., Belyakov, S. & Actiņš, A. On the rationalization of formation of solvates: Experimental and computational study of solid forms of several nitrobenzoic acid derivatives. *Cryst Growth Des* **20**, 5767–5784 (2020).

243. Zhang, G. G. Z. & Zhou, D. Crystalline and amorphous solids. in *Developing Solid Oral Dosage Forms: Pharmaceutical Theory and Practice: Second Edition* 23–57 (Academic Press, 2017). doi:10.1016/B978-0-12-802447-8.00002-9.

244. Gildenhuys, J., Roex, T. Le, Egan, T. J. & De Villiers, K. A. The single crystal X-ray structure of β-hematin DMSO solvate grown in the presence of chloroquine, a β-hematin growth-rate inhibitor. *J Am Chem Soc* **135**, 1037–1047 (2013).

245. Boryczka, S. *et al.* X-ray crystal structure of betulin-DMSO solvate. *J Chem Crystallogr* **42**, 345–351 (2012).

246. Cavallari, C., Santos, B. P. A. & Fini, A. Olanzapine Solvates. *J Pharm Sci* **102**, 4046–4056 (2013).

247. Yuan, L. & Lorenz, H. Solvate Formation of Bis(demethoxy)curcumin: Screening and Characterization. *Crystals 2018, Vol. 8, Page 407* **8**, 407 (2018).

248. Hamdi, N., Feutelais, Y., Yagoubi, N., De Girolamo, D. & Legendre, B. Solvates of indomethacin. *J Therm Anal Calorim* **76**, 985–1001 (2004).

249. Neville, G. A., Beckstead, H. D. & Cooney, J. D. Thermal analyses (TGA and DSC) of some spironolactone solvates. *Fresenius' Journal of Analytical Chemistry 1994 349:10* **349**, 746–750 (1994).

250. Zhoujin, Y. *et al.* A new solvate of clonixin and a comparison of the two clonixin solvates. *RSC Adv* **11**, 24836–24842 (2021).

251. Henderson, W. A., Seo, D. M., Han, S.-D. & Borodin, O. Electrolyte Solvation and Ionic Association. VII. Correlating Raman Spectroscopic Data with Solvate Species. *J Electrochem Soc* **167**, 110551 (2020).

252. Bolton, B. A. & Prasad, P. N. Laser raman investigation of pharmaceutical solids: Griseofulvin and its solvates. *J Pharm Sci* **70**, 789–793 (1981).

253. Byrn, S. R., Zografi, G. & Chen, X. S. Solvates and Hydrates. in *Solid State Properties of Pharmaceutical Materials* 38–47 (John Wiley & Sons, Ltd, 2017). doi:10.1002/9781119264408.CH3.

254. Lau, E. Preformulation Studies. in *Handbook of Modern Pharmaceutical Analysis* (ed. Satinder Ahuja) 173–233 (2001).

255. CULLITY, B. D. *Elements of X-ray Diffraction*. *Physics Bulletin* vol. 29 (ADDISON-WESLEY PUBLISHING COMPANY INC., 1978).

256. Liu, L. *et al.* Effects of Solvent Molecules on the Interlayer Spacing of Graphene Oxide. *Trans. Tianjin University* **24**, 555–562 (2018).

257. Cruz Cabeza, A. J., Pidcock, E., Day, G. M., Motherwell, W. D. S. & Jones, W. Space group selection for crystal structure prediction of solvates. *CrystEngComm* **9**, 556–560 (2007).

258. Cole, J. C., Raithby, P. R. & Taylor, R. Prior likelihoods and space-group preferences of solvates. *Cryst Growth Des* **21**, 1178–1189 (2021).

259. Malkin, A. J. & McPherson, A. Light-scattering investigations of nucleation processes and kinetics of crystallization in macromolecular systems. *Acta Crystallogr D Biol Crystallogr* **50**, 385–395 (1994).

260. Malkin, A. J. & McPherson, A. Light scattering investigations of protein and virus crystal growth: ferritin, apoferritin and satellite tobacco mosaic virus. *J Cryst Growth* **128**, 1232–1235 (1993).

261. Rychly, R. & Npvlt, J. Measuring and Calculating Heat of Crystallisation. *CRYSTAL Research&Technology* **9**, 799–810 (1974).

262. Jia, Y. & Liu, X. Y. From surface self-assembly to crystallization: Prediction of protein crystallization conditions. *Journal of Physical Chemistry B* **110**, 6949–6955 (2006).

263. Moser, M., Georg, A. G., Steinemann, F. L., Rütti, D. P. & Meier, D. M. Continuous milli-scale reaction calorimeter for direct scale-up of flow chemistry. *J Flow Chem* **11**, 691–699 (2021).

264. Francis, W. & Peters, M. C. *Data Sheet No. 121 - Measurement of the Calorific Value of Fuel Gas*. *Fuels and Fuel Technology (Second Edition)* (Pergamon, 1980). doi:https://doi.org/10.1016/B978-0-08-025249-0.50077-X.

265. Wunderlich, B. *Thermal Analysis*. *Encyclopedia of Materials: Science and Technology* (Elsevier, 2001). doi:https://doi.org/10.1016/B0-08-043152-6/01648-X.

266. Derewenda, Z. S. & Vekilov, P. G. Entropy and surface engineering in protein crystallization. *Acta Crystallogr D Biol Crystallogr* **62**, 116–124 (2006).

267. Perreu, J. : *C.r. Acad. Sci Paris* **199**, (1934).

268. Rychly, R. Thesis. (Techn. University (VSCHT), Prague, 1969).

269. A. McPherson, Y. Kuznetsov. Mechanisms, kinetics, impurities and defects: consequences in macromolecular crystallization. Acta Crystallogr F Struct Biol Commun. 2014 Apr;70(Pt 4):384-403.

Appendix

Thermodynamics of crystallisation and dissolution

Crystallisation and dissolution are both thermodynamic processes that involve the formation and breaking of solid-liquid interfaces. The fundamental difference between these two processes is the direction of heat flow.[261] In crystallisation, a solid is formed from a liquid as molecules or ions come together and organize into a repeating three-dimensional pattern. This process releases heat, known as the heat of crystallization. On the other hand, dissolution is the process by which a solid dissolves in a liquid, forming a solution. This process absorbs heat, known as the heat of dissolution.[261] In the process of crystallisation, growth units are transported from the liquid phase to the surface of the crystal nuclei and incorporated at specific locations called kink sites. This process is known as surface kinetics (**Figure S1**).[148] The nucleation kinetics, which determines the formation of crystals, is largely influenced by the nucleation barrier, the speed of transport of growth units, and the surface kinetics.[171] When the transport of growth units is faster than their integration at the surface, the formation of amorphous structures is more likely to occur. Conversely, when the integration of growth units at the surface is rapid, they are able to arrange into an orderly, compact structure that eventually forms a crystal.[262]

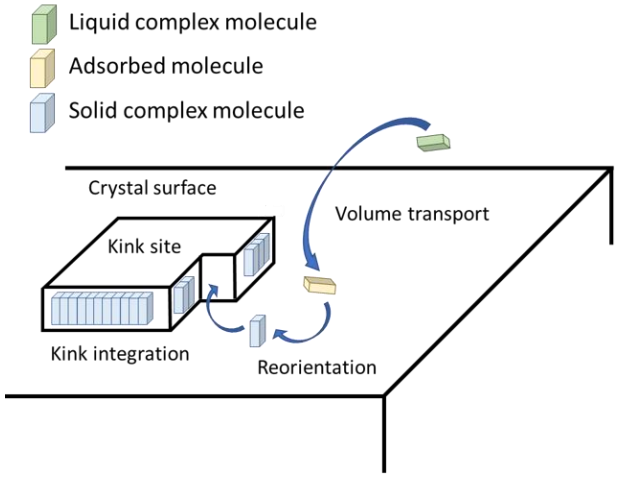


Figure S1. *The process of growing a crystal involves the transfer of growth units from the liquid phase (liquid complex molecule - Green) to specific sites on the surface of crystal nuclei, known as kink sites. The formation of crystal requires the reorientation of an adsorbed molecule (Yellow) before integrating into an orderly, compact structure (as solid complex molecule - Blue) at kink site.*

The heat of crystallisation and heat of dissolution is often assumed to be equivalent, with the amount of heat released during crystallisation equal to the amount of heat absorbed during dissolution. However, this assumption holds true only under certain conditions, such as when the solid and liquid phases are in equilibrium and the process is carried out at a constant temperature. This assumption does not take into account the effects of surface integration during crystallisation on the heat of crystallisation.[155] In reality, the crystallisation process typically occurs at the surface of the liquid and the heat is released over a range of surface areas. This can have an impact on the overall heat of crystallisation. Additionally, as the crystallisation process proceeds, the surface area may change, which can also affect the heat of crystallisation.[155]

The direct measurement of the values of heat of crystallisation is limited by the accuracy of the methods employed.[261] Calorimetry is one of the techniques used to measure the heat flow associated with a chemical or physical process. It is commonly used in thermodynamics to determine the enthalpy change (ΔH) associated with a reaction or phase change. The main advantage of calorimetry is its ability to directly measure the heat flow associated with a process, which allows for the determination of the enthalpy change.[263] However, calorimetry is not a very sensitive technique and may not be able to detect small changes in heat flow.[264,265] The major challenges include ensuring the cleanliness and smoothness of the calorimetric cells, as even small imperfections can lead to crystallisation in a metastable, supersaturated solution.[261] Another challenge is accurately determining the total mass of the formed crystals, which can be distorted by solution trapped between crystals and mislabeled as a crystalline mass. In homogeneous nucleation, where crystals are allowed to form spontaneously, the measurement can be time-consuming and difficult due to the extended duration of the process.[261]

An alternative approach for determining the heat of crystallization is to measure the solubility, C_e, of the crystals at different temperatures and use standard thermodynamics relations to calculate ΔH^0_{cryst}, ΔS^0_{cryst}, and ΔG^0_{cryst}. The equilibrium constant for the reaction of a molecule in solution to a molecule in crystal form is $K = C_e^{-1}$, assuming that the activity coefficients are close to one due to low solubility and that the solution behaves as an ideal one.[262,266]

It is assumed that at solubility, the heat of crystallization is equal in magnitude but opposite in sign to the heat of dissolution, i.e. $\Delta H^0_{cryst} = -\Delta H^0_{disso}$.[261] This leads to the equation for the equilibrium constant for crystallization as $K_{cryst} = C_e^{-1}$, at given temperature T (**Equation S1**).[148,149]

$$\Delta G^0_{cryst} = -RT ln K = RT ln C_e \qquad \textbf{Equation S1}$$

To determine ΔH^0_{cryst}, the van'tHoff relation (**Equation S2**) can be employed.[148,149]

$$\frac{\partial ln C_e}{\partial (1/T)} = \frac{\Delta H^0_{cryst}}{R} \qquad \textbf{Equation S2}$$

The comparison between the ΔH^0_{cryst} obtained through directly measuring the heat of crystallization and the measured heat of dissolution ΔH^0_{disso} is presented in **Table S1**.

***Table S1**. Comparative values between heat of crystallisation (ΔH^0_{cryst}) and heat of dissolution (ΔH^0_{disso})*

Substance	$-\Delta H^0_{cryst}$ from direct measurement [kJ/mol]	Measured ΔH^0_{disso} [kJ/mol]	% difference	Reference
$Na_2HPO_4 \cdot 12H_2O$	90.79	91.30	0.56 %	Perreu 1934[267]
$Na_2SO_4 \cdot 10H_2O$	70.08	70.33	0.36 %	
$Na_2CO_3 \cdot 10H_2O$	56.90	56.61	0.51 %	
$Na_2S_2O_3 \cdot 5H_2O$	31.30	31.46	0.51 %	
$ZnSO_4 \cdot 7H_2O$	22.72	23.01	1.28 %	
$MnCl_2 \cdot 4H_2O$	19.87	20.12	1.26 %	
$BaCl_2 \cdot 2H_2O$	19.25	19.79	2.81 %	
$MgSO_4 \cdot 7H_2O$	17.62	17.36	1.48 %	
$(NH_2)_2CO$	11.12	11.37	2.25 %	Rychly 1969[268]
$CuSO_4 \cdot 5H_2O$	10.46	10.21	2.39 %	Perreu 1934[267]

From the data presented in **Table S1**, it is evident that the values of ΔH^0_{cryst} and ΔH^0_{disso} are comparable. The results demonstrate a correlation between the heat of crystallization and heat of dissolution, and the discrepancies identified are minimal. This supports the use of solubility data for the determination of crystallization thermodynamics. Additionally, it is important to note that limitations of the calorimetry method, such as improper cleaning of the vessel wall or inaccuracies in the measurement of the actual amount of the crystal

substance at the end of the measurement, may account for some of the discrepancies between ΔH^0_{cryst} and ΔH^0_{disso} presented above.

Despite the advantages of the approach for the determination of the crystallization heat from the solubility of crystals at different temperatures, including lower cost and accessible data, it is crucial to take into account an important limitation. This approach assumes that the crystal dissolution process follows ideal thermodynamic behavior, which may not always be the case in practice. Various factors, such as crystal defects, impurities, and crystal growth, can cause crystals to display non-ideal behavior and thereby affect the thermodynamics of the process, potentially leading to errors in the determination of the crystallization heat.[148,269]