Mechanochemical Co-crystallization: Insights and Predictions

Jan Roland Grölsl[1], Bernardo Castro-Dominguez[1,*]

[1] Centre for Sustainable and Circular Technologies (CSCT), Department of Chemical Engineering, University of Bath, Claverton Down, Bath BA2 7AY, UK

Abstract

Mechanochemistry integrates mechanical and chemical phenomena to spawn new modes of reactivity, faster reaction kinetics and the discovery of novel materials, in a solvent-free and environmental manner. This simple, yet efficient technique has become a well-established screening technique to discover pharmaceutical co-crystals - components that incorporate a secondary crystalline structure into the lattice of an active pharmaceutical ingredient (API) to improve its physicochemical properties. Today, predicting the reactivity of solid-state materials under mechanochemical conditions remains a major challenge. Here, we explore various machine learning algorithms and identified XGBoost ideal to accurately predict mechanochemical co-crystallization. The model was trained using 1000 co-crystallization events and 2083 chemical descriptors, revealing fundamental insights about mechanochemistry. The model was implemented to screen secondary crystalline structures against a model API, yielding three new mechanochemically-formed co-crystals. This predictive model will accelerate the discovery of novel pharmaceuticals while its insights aid at developing a more sustainable chemical industry.

Keywords: Machine Learning, Mechanochemistry, Co-crystallization, Solvent-free manufacturing

[*]**Corresponding author:** ORCID ID: https://orcid.org/0000-0001-5913-305X Email: bcd28@bath.ac.uk, Office phone: +44 1225384946

1. Introduction

Named by the International Union of Pure and Applied Chemistry (IUPAC) as a chemical innovation with the potential of changing the world, mechanochemistry, or grinding, offers the opportunity to generate a paradigm shift in science and society by unlocking a new realm of chemical possibilities (Baláž et al., 2013; Gomollón-Bel, 2019). Chemical reactions today are not only dependent on solvents but also deeply rely on the use of heat, light or electricity to initiate reactions. However, by delivering extremely high localized heating through mechanical action, mechanochemistry enables us to target molecules and materials previously deemed impossible to synthesize and to access these in a solvent-free and therefore highly sustainable manner (Baláž et al., 2013). High reagent concentrations and nonexistence solvation during grinding facilitate new modes of reactivity, not viable in solution conditions, thus changing product selectivity. At present, our understanding is primitive and continues relying on trial-and-error approaches. The literature considers it a “magic box” with “some mystique guards the rules of Mechanochemistry”; indeed, this field is “in its infancy, akin to organic chemistry in the early 1800s” (Andersen and Mack, 2018).

With over 200 publications, neat grinding and liquid assisted grinding (LAG) have become well-established screening techniques for the discovery of novel pharmaceutical co-crystals (SciFinder; Chemical Abstracts Service). These compounds emerge from the incorporation of secondary crystalline structure into the lattice of active pharmaceutical ingredients (APIs) (see Figure 1) to form new hydrogen bonds, which direct the association of a covalently-bonded hydrogen atom with other molecules into an aggregate structure (Etter and Adsmond, 1990). These new associations spawn changes to the critical quality attributes (CQA) of drugs. In specific, co-crystals enhance the water solubility of drugs, a CQA particularly relevant as 40% of all developed drugs have solubility challenges and 80% of in-production line drugs are classified as poorly soluble (Aakeröy and Sinha, 2018; Karagianni et al., 2018). Nevertheless, to unleash

the power of mechanochemistry, regulatory agencies (e.g. Food & Drug Administration (FDA)) require predicting reactivity to enable quality-by-design (QbD). Unfortunately, our poor understanding of mechanochemistry, coupled with the 10^6 known organic compounds and 10^{60} potential secondary crystalline structures, make predicting the mechanochemical reactivity of solid-state materials a major challenge (Ghosh et al., 2019; Rupakheti et al., 2015).

In this work, we implemented a high-throughput experimental screening approach to train a machine learning (ML) model capable of probabilistically predicting mechanochemical co-crystallization events. Recent trends emphasize the importance of ML in chemical engineering, including the prediction of thermophysical properties, the design and modeling of processes as well as the design of materials (Bañares-Alcántara et al., 1985; Stephanopoulos, 1990; Stephanopoulos et al., 1987; Venkatasubramanian, 2019; Venkatasubramanian et al., 2006). ML models also have been applied to polymer informatics, molecular sciences, and crystal structure prediction, methodically realizing novel materials with desired properties and avoiding expensive trial-and-error experiments (Butler et al., 2018; Sha et al., 2021). To develop this model, various APIs and secondary structures were treated via neat grinding in an oscillatory ball mill to generate >800 in-house training experiments. Powder X-Ray diffraction (PXRD) was used to characterize the crystalline phases of the products and determine the reactivity of the reactants. The physicochemical properties of each reactant (e.g. molecular weight, functional groups) were uniquely embodied by 1825 chemical or molecular descriptors. Each co-crystallization event contained the fingerprint of 2 different molecules as inputs, generating two potential numeric outcomes: “1” co-crystal formation was successful and “0” co-crystal formation was not successful (only raw material was identified). Next to the in-house data, relevant experiments from related literature were adapted and added to the data set (all together adding up to 1000 data points) (Wicker et al., 2017).

A principal component analysis (PCA) was performed to identify and remove non-relevant descriptors (Jolliffe and Cadima, 2016); this followed the implementation of different

classification algorithms with one from the XGBoost library displaying the highest performance (Chen and Guestrin, 2016). The predictive capabilities of the model were used to identify potential secondary crystalline structures with the capacity of forming a co-crystal with diclofenac. Those pairs with a high possibility/ranking of mechanochemical co-crystallization were experimentally synthesized and characterized, leading to the synthesis of a set of diclofenac-co-former never reported before via neat grinding. Finally, we identified the key chemical descriptors that drive mechanochemical co-crystallization and provided guidelines for engineering novel products in a solvent-free manner.

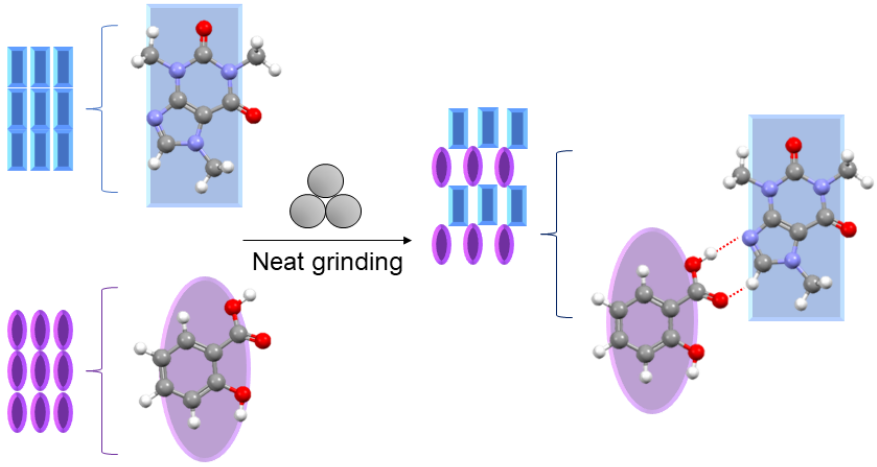

Figure 1. Schematic of a mechanochemical formed pharmaceutical co-crystal via grinding

2. Materials and Modelling Methods

Data preparation. The data set was produced by experimentally screening a range of APIs and secondary crystals (Supplementary Information, Table S1) in a combinatorial manner (Note that the dataset is publicly available at: https://doi.org/10.15125/BATH-00963). Eppendorf tubes (1.5 mL) were filled with equimolar an API/secondary crystals mixture and 30 steel balls of 2 mm in diameter. Twenty reaction tubes were loaded at once, onto a MM400 Retsch ball mill and run

for 30 min at 30 Hz. Using this systematic high throughput approach, 847 experiments were conducted and within 24 hrs after synthesis, characterized via powder X-ray diffraction (PXRD). Although, PXRD is a powerful characterization method that aids in determining the supramolecular structure of solid-state materials; to differentiate between co-crystals and salts, an acidic strength (pK_a) analysis was carried out (Supplementary Information, Section S1). Figure 2 shows how the dataset matrix was arranged for this work. All molecular descriptor inputs were estimated using the Mordred descriptor database, which computed 1825 chemical descriptors per molecule (Moriwaki et al., 2018). These descriptors were processed to avoid unassigned numerical chemical descriptors and ensure numerical variability. The total number of distinct descriptors used in this work was 1041 per molecule. Therefore, each mechanochemical event was represented by a total of 2083 descriptors, corresponding to the chemical descriptors of the API, the secondary crystal and the success classification rating. The success classification rating was determined via PXRD analysis and followed these general guidelines:

(i) "Unsuccessful" cases were denoted with a value of "0" and occurred when the ground samples displayed a PXRD spectrum containing the characteristic structure of the raw materials, only.

(ii) "Successful" events were denoted with a value of "1" and were deemed present when PXRD spectra of the samples displayed peaks other than that of the raw materials, the pK_a analysis discarded salts and the characteristic peaks of the resulting new structure were correlated with the theoretical structures found at the Cambridge Structural Database (CSD). It is worth noting that since all materials were processed via neat grinding, the presence of solvates and hydrates was not found in any of the samples produced in this work.

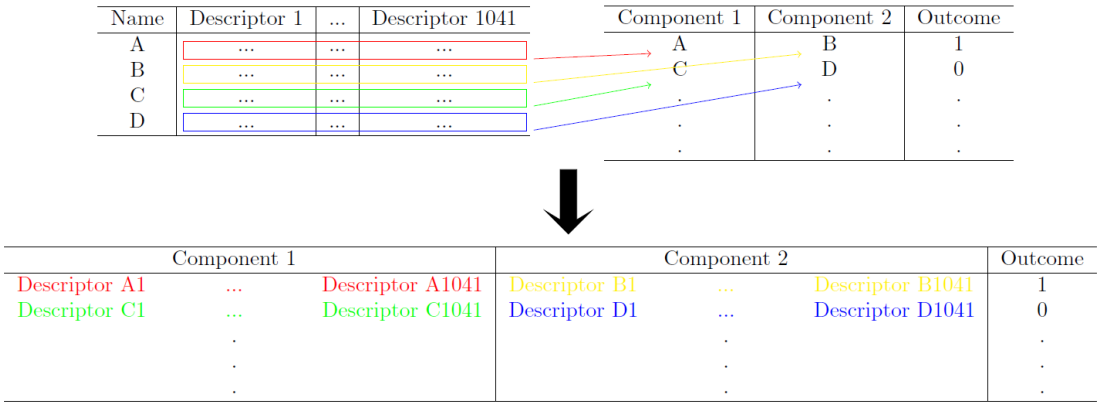

Figure 2. Schematic of dataset matrix arrangement containing 1041 molecular descriptors per component, success classification outcome and 1000 co-crystallization events.

Machine learning methods. This work assessed the performance of various ML classifiers to ensure the highest predictive capabilities. These classifiers include:

a) Random forests. This ensemble model combines predictions of multiple decision trees to enhance its performance. In this model, each decision tree is constructed differently via bagging and random feature subsets, thus generating predictions in an independent manner. The model often displays a robust performance due to the combinatorial influence of the different trees (Criminisi et al., 2012).

b) Extremely randomized trees. This model uses decision trees, each of which is trained using the whole dataset. In this model, the attributes and cut-point choices are randomized to decrease the variance (Geurts et al., 2006).

c) Adaptive boosting (AdaBoost). AdaBoost is a meta-estimator that fits a classifier on the original dataset. Copies of the classifier are further utilized, and the weights of incorrectly classified instances adjusted accordingly (Zhu et al., 2006).

d) Extreme gradient boosting (XGBoost). XGBoost is an optimized distributed gradient boosting library designed to be highly efficient and flexible. It implements machine learning algorithms under the gradient boosting framework and provides a parallel tree boosting, penalization of

trees, proportional shrinking of leaf nodes and a simple implementation for optimization techniques (Chen and Guestrin, 2016; Mahmud et al., 2019; Mitchell and Frank, 2017).

e) Multi-layer perceptron neural network or Artificial Neural Networks (ANN). ANN models use neurons (also called nodes) arranged in layers and connected in a hierarchical manner. Neurons transmit information through the different layers while adjusting a numerical weight, thus enabling the learning process (Abiodun et al., 2018; Abraham, 2005).

f) Supported vector machines (SVMs). In SVMs, a hyperplane is constructed and used to classify or regress the dataset based on distance between the data point and the hyperplane (Meyer et al., 2003; Noble, 2006).

3. Results

3.1 Dimensionality reduction and ML algorithm performance

Since the number of input features exceed the number of experiments, dimensionality reduction was implemented to avoid overfitting. Certainly, the literature suggests that to effectively train a ML model, the number of observations in the training set should be larger than the number of input features (Dietterich, 1995; Jain and Chandrasekaran, 1982; Raudys and Jain, 1991). Furthermore, the dimension of the molecular descriptors was reduced, ensuring that only those features relevant to co-crystal formation were used to build the ML classification model. For this purpose, a Principal Component Analysis (PCA) was implemented to the data set after normalization. As shown in Figure 3, PCA was implemented in 4 different manners, to understand how information is lost as the number of chemical descriptors reduces; these include:

(i) "Combined" approach, where the input matrix (e.g. the molecular descriptors for both molecules) is reduced together.

(ii) “Component” approach, where the molecular descriptors of each molecule are reduced separately and concatenated afterward.

(iii) “Original” approach, where the dimensionality was reduced using the training data in the order obtained.

(iv) “Random” approach, here the training data was ordered randomly before PCA was implemented.

The analysis showed that performing a “random” “component” approach was the best way to reduce the dimensionality while minimizing the information lost. The “component” approach showed less information lost than the “combined” information, indicating that the relevant information of the API and the secondary structure must be preserved for better performance. In this work, we utilized 25 principal components which resulted in an information loss of 1.7%.

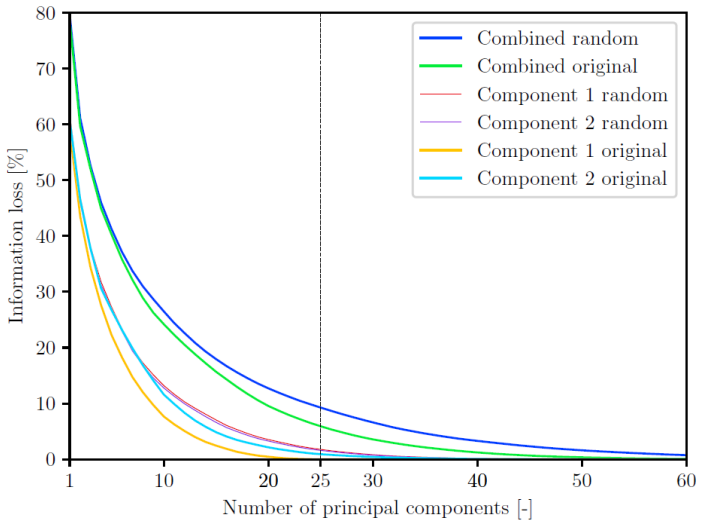

Figure 3. Dimensionality reduction and loss of information for various methods.

Various ML algorithms were implemented to explore the 1000 co-crystallization events, from which 350 resulted in co-crystal formation. Scikit-learn v.23 was used in PyCharm to explore the performance of seven different ML algorithms, which were comparatively assessed in terms of their accuracy (area under the curve (AUC)) and processing time. As shown in Table 1, a XGBoost algorithm displayed high AUC values and rapid processing time; therefore, this

method was chosen for further tuning. A hyperparameter optimization process was executed via multidimensional optimization to identify the parameters required to maximize the accuracy and processing time of this ML algorithm (Supplementary Information, Section S2). As shown in Figure 4, after hyperparameter optimization, a mean AUC value of 80% with a 1.7% standard deviation was achieved, outperforming previous classifiers. Note that although a hyperparameter optimization process could have been implemented to all the classifier models shown in Table 1; for the purposes of this work, only the parameters of the best performing model, at the same baseline, were optimized.

Table 1. Performance of various classification models, where the AUC value is calculated in a 5-fold cross-validation process with a mean of $\bar{x}$, and standard deviation of σ.

	$AUC_{k=5}$		
Classifier model selection	$\bar{x}$ [%]	σ [%]	Processing time [s]
Random forest[a]	72.97	1.95	1.12
Extremely randomized trees[a]	71.19	3.07	0.78
Gradient Boosting[a]	71.31	3.67	1.59
AdaBoost[a]	68.95	2.54	0.61
XGBoost[a]	74.50	2.77	0.40
Multi-layer perceptron neural network[a]	72.14	2.45	2.46
SVM[a]	69.18	3.00	0.19

[a]Default values

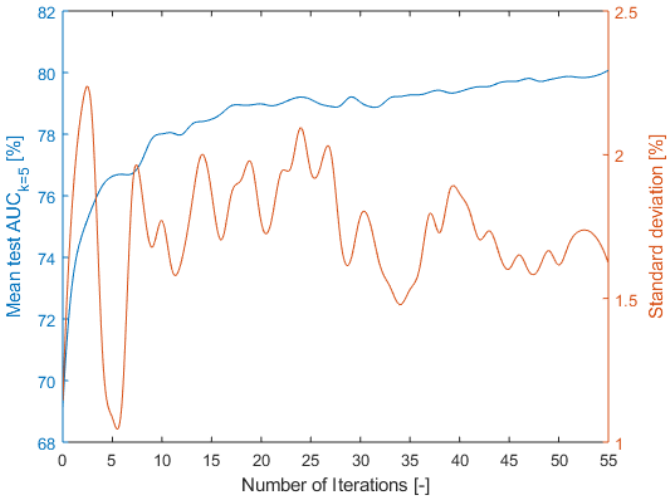

Figure 4. Prediction progress with an increasing number of iterations leading to the optimal set of parameters.

3.2. ML model validation and synthesis of novel solventless co-crystals

The optimized XGBoost algorithm was validated by testing its predictive performance using molecules unknown to the model. To illustrate this, we used paracetamol, a molecule not used in our training data set, but which has been screened thoroughly for co-crystallization across the literature (Srirambhatla et al., 2012; Wicker et al., 2017). We used the work presented by Wood et al., which reported 35 different co-crystallization events, from which 16 of these produced a verified co-crystal (Wood et al., 2014). The same classification criterium has been used for this dataset (0 or 1). This comparative validation test shows an unbiased evaluation, displaying equally distributed false-negative and false-positive cases. Based on the achieved AUC of 80% and the confusion matrix shown in (Figure 5), we can conclude that the performance of the model has reasonable predictive capabilities and can be used to extract insights behind mechanochemical co-crystallization.

	Predicted Yes	Predicted No
Actual Yes	16	5
Actual No	5	9

Figure 5. Confusion matrix for paracetamol co-crystals.

The ML model achieved a high level of accuracy for the prediction of co-crystal formation via mechanochemistry, using known and unknown training data. Next, we assessed the discovery potential of the model by screening 14 secondary structures (Table 2) against the molecule *diclofenac*, which is considered to have low solubility, according to the Biopharmaceutics Classification System (B.Chuasuwan et al., 2009). Diclofenac was not part of the training data repertoire, allowing to further assess the predictive accuracy of the model. As shown in Table 2, our ML model predicted that through mechanochemistry, 5 co-crystals could be formed. This outcome was experimentally assessed, revealing that the model produced two false-positive

predictions. This screening exercise generated the formation of three new co-crystals under neat grinding, so far only reported under solvent-depended methods (Aakeröy et al., 2011; Báthori et al., 2011): (i) Diclofenac-Isonicotinamide, (ii) Diclofenac-2-pyrrolidinone, and (iii) Diclofenac-4,4'-Bipyridine.

The ML model was further assessed by evaluating its performance against Mercury – a virtual co-crystal screening model from the CSD (Macrae et al., 2020). This model uses quantitative structure-activity relationships to determine co-crystallization; nonetheless, it does not consider the peculiarities behind mechanochemistry. Table 2 also shows that the CSD co-crystal prediction tool was able to predict only 48.5% of the events with 4 false-positive and 2 false-negative predictions. These results indicate that our ML model effectively captured the phenomenon that drives mechanochemical events. Additionally, it is important to mention that adding additional chemical descriptors did not show noteworthy statistical fluctuations on the predictive performance of the model (Figure 5 and Table 2), as the information loss due to dimensionality reduction is no greater than 1.7%.

Table 2. Predictive comparison between the in-house algorithm and the CSD Mercury prediction tool against the experimental results of several secondary crystalline structures with diclofenac.

Secondary structures	ML Prediction	CSD Mercury Prediction	Experimental result
Isonicotinamide	Positive	Negative	Positive
2-Pyrrolidinone	Positive	Positive	Positive
4,4'-Bipyridine	Positive	Negative	Positive
3,3'-Thiodipropionic acid	Positive	Negative	Negative
Folic acid	Positive	Positive	Negative
Phenazine	Negative	Negative	Negative
Salicylic acid	Negative	Negative	Negative
Nicotinamide	Negative	Negative	Negative
Caffeine	Negative	Positive	Negative
Theophylline	Negative	Negative	Negative
L-Glutamic acid	Negative	Negative	Negative
Ibuprofen	Negative	Positive	Negative
L-Ascorbic acid	Negative	Negative	Negative
Acetylsalicylic acid	Negative	Positive	Negative

4. Discussion

4.1 Mechanochemical reactivity and ML insights

The literature has reported important theoretical frameworks that aid in designing a set of co-crystal pairs (Wood et al., 2014). Even so, the incorporation of the phenomenological components caused by mechanochemical conditions has not been reported, due to its complexity caused by the inhomogeneous reactants, the presence of localized mechanical stresses, severe and transient conditions, and relaxation processes (e.g., heating, aggregation, recombination, adsorption, etc.) and mostly due to the lack of systematic studies and difficulties of directly observing materials under mechanical stresses (Baláž et al., 2013). In this work, we analyzed the molecular descriptors used for the prediction of mechanochemical co-crystallization events to provide insightful guidelines for the design of advanced materials. The molecular descriptors were ranked based on their cumulative importance for the prediction of mechanochemical co-crystal formation (Supplementary Information, Table S3).

The molecular charge or electrostatic features of the molecules had the highest impact on mechanochemical co-crystallization. We found that the relative negative charge (a descriptor that relates the partial charge of the most negative atom to the total negative charge of the molecule) was a critical feature for the generation of co-crystals (ranked 1). Having a negative charge present in the acceptor molecule's surface area was found to be quite important (ranked 3 and 4). Certainly, having a high negative charge on the proton-acceptor increases the probability of co-crystal formation. In contrast, the relative positive charge of a proton donor appeared to be of less importance (ranked 31), suggesting that electronegative moieties of the proton acceptor drive bond formation and supramolecular arrangement. Etter's rules for hydrogen-bond formation ranks "*all good proton donors and acceptors are used in hydrogen bonding"* as the most critical condition, further supporting our results (Etter and Adsmond, 1990); nonetheless, we have found that electronegativity has a higher degree of importance if the hydrogen bond is formed under solid-state conditions.

The structure of molecules had a contribution towards predicting neat grinding co-crystallization. The distance and bonding between intermolecular atoms (ranked 5 and 6) were found to be important features as they can be perceived as the probability for hydrogen bond formation, which is a strong intermolecular force created by a hydrogen atom and an electronegative atom. Etter's rules point out that "*the best proton donors and acceptors remaining after intramolecular hydrogen-bond formation form intermolecular hydrogen bonds to one another*" (Etter and Adsmond, 1990)), thus supporting the results found in this work. In line with this rule, we found that intramolecular characteristics, such as charge imbalance, are an important co-crystallization contributor.

The contributions of topochemical descriptors (e.g. polarizability, volume, electronegativity, ionization potential, atomic number, and mass) suggest that to enable the reactivity of mechanochemical co-crystallization, the API and secondary crystal need to be similar in shape and size. This isostructural requirement coupled with the solventless environment of the process supports the notion that solid solutions are formed during neat grinding (Cherukuvada et al., 2016; Lusi et al., 2011). It is known that if two solid-state materials are isostructural, then a solid solution will exist. The Hume-Rothery rules suggest that optimal formation of substitutional solid solutions occurs when both molecules have similar atomic radii, crystal structures, electronegativities and valances (Callister, William D.; Rethwisch, 2018; Hume-Rothery and Powell, 1935). These characteristics parallel the information contained in highly ranked topochemical descriptors. Although solid solutions in organic molecules have not been well documented across the literature, it is well known that mechanochemical processes allow their formation during metal alloying and oxidic systems (Baláž et al., 2013). We hypothesize that when an API and a secondary crystal are isostructural, at their point of collision, lattice dislocation and intramolecular mixing occurs, yielding a compositional gradient dictated by the native molecules. The formed solid solution rearranges according to the molecular charge or electrostatic features of the molecules present, yielding the co-crystal product.

Rules for mechanochemical co-crystallization. There are various empirically derived guidelines for the design of co-crystals and other supramolecular structures (Corpinot and Bučar, 2019); the most important include: (i) Etter's general rules for Hydrogen bonding and (ii) the proton transfer and pK_a rule. These rules have been applied to solvent-based co-crystallization techniques such as solvent evaporation and cooling co-crystallization. In these techniques, a liquid/vapor solvent is used to disrupt the crystal lattice of the API and secondary crystal. Once dissolved, the solvent is removed to yield solid-state materials with a thermodynamically driven structure.

On the other hand, during mechanochemical co-crystallization, the incorporation of a solute molecule into the lattice of a solvent crystal occurs without the aid of a liquid/vapor solvent and does not require the complete dissolution of the raw materials. Hence, the rules that drive solvent-based co-crystallization events are not necessarily the same as the ones that drive mechanochemical events. The ML algorithm developed in this work suggests that although Etter's and the pK_a rule are essential attributes; mechanochemical conditions need to consider the rules for solid solutions. Solid-state molecular miscibility is key and therefore when designing co-crystals, Hume-Rothery rules should be incorporated onto Etter's and pK_a.

5. Conclusions

In this work, we developed a ML model to predict the mechanochemical co-crystallization of organic molecules. The model was trained by a dataset of 1000 experimental co-crystallization events and a comprehensive set of 2083 molecular descriptors. The features of the model were subject to a reduction in dimensionality using a PCA. The accuracy and processing speed of different algorithms were assessed to determine the best classification algorithm (XGBoost) which was subject to hyperparameter optimization. The model showed to be an efficient screening tool; therefore, it has been used to assess a data set found in the literature and experimentally discover diclofenac co-crystals - structures never synthesized in a solvent-free

manner. The ML algorithm was used to generate insights related to phenomena that drive mechanochemical co-crystallization. In specific, we found that the formation of solid solutions during neat grinding is key for the incorporation of a solute molecule into the lattice of a solvent crystal, and therefore must be considered when designing novel co-crystals.

Acknowledgments

This research was supported by Royal Society-Research Grant RSG\R1\180090. Jan Gröls (JC) would like to thank the University of Bath for his Ph.D. studentship and the Centre for Doctoral Training in Sustainable Chemical Technologies. The authors gratefully acknowledge the Material and Chemical Characterization Facility (MC2) at the University of Bath for technical support and assistance in this work.

Data and Code Availability Statement

The following items are deposited at the University of Bath's Research Data Archive (Gröls and Castro Dominguez, 2021):

1. The generated in-house data used to train the models (including general datasets, raw PXRD data, description, and outcomes)
2. The machine learning training and prediction codes

Author contributions

B.C.D conceived of the presented idea and designed/directed the project. J.G. developed the ML model, designed, and performed the experiments, and analyzed the data. All authors discussed the results and contributed to the final manuscript.

References

Aakeröy, C.B., Grommet, A.B., Desper, J., 2011. Co-Crystal Screening of Diclofenac. Pharmaceutics 3, 601–614. https://doi.org/10.3390/pharmaceutics3030601

Aakeröy, C.B., Sinha, A.S. (Eds.), 2018. Co-crystals, Monographs in Supramolecular Chemistry. The Royal Society of Chemistry. https://doi.org/10.1039/9781788012874

Abiodun, O.I., Jantan, A., Omolara, A.E., Dada, K.V., Mohamed, N.A., Arshad, H., 2018. State-of-the-art in artificial neural network applications: A survey. Heliyon 4, e00938. https://doi.org/https://doi.org/10.1016/j.heliyon.2018.e00938

Abraham, A., 2005. Artificial Neural Networks. Handb. Meas. Syst. Des., Major Reference Works. https://doi.org/https://doi.org/10.1002/0471497398.mm421

Andersen, J., Mack, J., 2018. Mechanochemistry and organic synthesis: from mystical to practical. Green Chem. 20, 1435–1443. https://doi.org/10.1039/C7GC03797J

B.Chuasuwan, Binjesoh, V., Polli, J.E., Zhang, H., Amidon, G.L., Junginger, H.E., Midha, K.K., Shah, V.P., Stavchansky, S., Dressman, J.B., Barends, D.M., 2009. Biowaiver Monographs for Immediate Release Solid Oral Dosage Forms: Diclofenac Sodium and Diclofenac Potassium. J. Pharm. Sci. 98, 1206–1219. https://doi.org/https://doi.org/10.1002/jps.21525

Baláž, P., Achimovičová, M., Baláž, M., Billik, P., Cherkezova-Zheleva, Z., Criado, J.M., Delogu, F., Dutková, E., Gaffet, E., Gotor, F.J., Kumar, R., Mitov, I., Rojac, T., Senna, M., Streletskii, A., Wieczorek-Ciurowa, K., 2013. Hallmarks of mechanochemistry: from nanoparticles to technology. Chem. Soc. Rev. 42, 7571–7637. https://doi.org/10.1039/c3cs35468g

Bañares-Alcántara, R., Westerberg, A.W., Rychener, M.D., 1985. Development of an expert system for physical property predictions. Comput. Chem. Eng. 9, 127–142. https://doi.org/https://doi.org/10.1016/0098-1354(85)85003-1

Báthori, N.B., Lemmerer, A., Venter, G.A., Bourne, S.A., Caira, M.R., 2011. Pharmaceutical Co-crystals with Isonicotinamide—Vitamin B3, Clofibric Acid, and Diclofenac—and Two

Isonicotinamide Hydrates. Cryst. Growth Des. 11, 75–87. https://doi.org/10.1021/cg100670k

Butler, K.T., Davies, D.W., Cartwright, H., Isayev, O., Walsh, A., 2018. Machine learning for molecular and materials science. Nature 559, 547–555. https://doi.org/10.1038/s41586-018-0337-2

Callister, William D.; Rethwisch, D.G., 2018. Materials Science and Engineering: An Introduction, 10th ed.

Chen, T., Guestrin, C., 2016. XGBoost: A Scalable Tree Boosting System, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '16. Association for Computing Machinery, New York, NY, USA, pp. 785–794. https://doi.org/10.1145/2939672.2939785

Cherukuvada, S., Kaur, R., Guru Row, T.N., 2016. Co-crystallization and small molecule crystal form diversity: from pharmaceutical to materials applications. CrystEngComm 18, 8528–8555. https://doi.org/10.1039/C6CE01835A

Corpinot, M.K., Bučar, D.-K., 2019. A Practical Guide to the Design of Molecular Crystals. Cryst. Growth Des. 19, 1426–1453. https://doi.org/10.1021/acs.cgd.8b00972

Criminisi, A., Shotton, J., Konukoglu, E., 2012. Decision Forests: A Unified Framework for Classification, Regression, Density Estimation, Manifold Learning and Semi-Supervised Learning.

Dietterich, T., 1995. Overfitting and Undercomputing in Machine Learning. ACM Comput. Surv. 27, 326–327. https://doi.org/10.1145/212094.212114

Etter, M.C., Adsmond, D.A., 1990. The use of cocrystallization as a method of studying hydrogen bond preferences of 2-aminopyrimidine. J. Chem. Soc.{,} Chem. Commun. 589–591. https://doi.org/10.1039/C39900000589

Geurts, P., Ernst, D., Wehenkel, L., 2006. Extremely randomized trees. Mach. Learn. 63, 3–42. https://doi.org/10.1007/s10994-006-6226-1

Ghosh, A., Louis, L., Arora, K.K., Hancock, B.C., Krzyzaniak, J.F., Meenan, P., Nakhmanson, S., Wood, G.P.F., 2019. Assessment of machine learning approaches for predicting the crystallization propensity of active pharmaceutical ingredients. CrystEngComm 21, 1215–1223. https://doi.org/10.1039/C8CE01589A

Gomollón-Bel, F., 2019. Ten Chemical Innovations That Will Change Our World: IUPAC identifies emerging technologies in Chemistry with potential to make our planet more sustainable. Chem. Int. 41, 12–17. https://doi.org/https://doi.org/10.1515/ci-2019-0203

Gröls, J., Castro Dominguez, B., 2021. Dataset supporting the paper: Predicting Mechanochemical Co-crystallization via Machine Learning [WWW Document]. Bath Univ. Bath Res. Data Arch. https://doi.org/https://doi.org/10.15125/BATH-00963

Hume-Rothery, W., Powell, H.M., 1935. On the Theory of Super-Lattice Structures in Alloys. Zeitschrift für Krist. - Cryst. Mater. 91, 23–47. https://doi.org/https://doi.org/10.1524/zkri.1935.91.1.23

Jain, A.K., Chandrasekaran, B.B.T.-H. of S., 1982. 39 Dimensionality and sample size considerations in pattern recognition practice, in: Classification Pattern Recognition and Reduction of Dimensionality. Elsevier, pp. 835–855. https://doi.org/https://doi.org/10.1016/S0169-7161(82)02042-2

Jolliffe, I.T., Cadima, J., 2016. Principal component analysis: a review and recent developments. Philos. Trans. R. Soc. A Math. Phys. Eng. Sci. 374, 20150202. https://doi.org/10.1098/rsta.2015.0202

Karagianni, A., Malamatari, M., Kachrimanis, K., 2018. Pharmaceutical Cocrystals: New Solid Phase Modification Approaches for the Formulation of APIs. Pharmaceutics 10. https://doi.org/10.3390/pharmaceutics10010018

Lusi, M., Atwood, J.L., MacGillivray, L.R., Barbour, L.J., 2011. Isostructural coordination polymers: epitaxis vs. solid solution. CrystEngComm 13, 4311–4313. https://doi.org/10.1039/C1CE05164D

Macrae, C.F., Sovago, I., Cottrell, S.J., Galek, P.T.A., McCabe, P., Pidcock, E., Platings, M., Shields, G.P., Stevens, J.S., Towler, M., Wood, P.A., 2020. Mercury 4.0: from visualization to analysis, design and prediction. J. Appl. Crystallogr. 53, 226–235. https://doi.org/10.1107/S1600576719014092

Mahmud, S.M.H., Chen, W., Jahan, H., Liu, Y., Sujan, N.I., Ahmed, S., 2019. iDTi-CSsmoteB: Identification of Drug–Target Interaction Based on Drug Chemical Structure and Protein Sequence Using XGBoost With Over-Sampling Technique SMOTE. IEEE Access 7, 48699–48714. https://doi.org/10.1109/ACCESS.2019.2910277

Meyer, D., Leisch, F., Hornik, K., 2003. The support vector machine under test. Neurocomputing 55, 169–186. https://doi.org/https://doi.org/10.1016/S0925-2312(03)00431-4

Mitchell, R., Frank, E., 2017. Accelerating the XGBoost algorithm using GPU computing. PeerJ Comput. Sci. 3, e127. https://doi.org/10.7717/peerj-cs.127

Moriwaki, H., Tian, Y.-S., Kawashita, N., Takagi, T., 2018. Mordred: a molecular descriptor calculator. J. Cheminform. 10, 4. https://doi.org/10.1186/s13321-018-0258-y

Noble, W.S., 2006. What is a support vector machine? Nat. Biotechnol. 24, 1565–1567. https://doi.org/10.1038/nbt1206-1565

Raudys, S.J., Jain, A.K., 1991. Small sample size effects in statistical pattern recognition: recommendations for practitioners. IEEE Trans. Pattern Anal. Mach. Intell. 13, 252–264. https://doi.org/10.1109/34.75512

Rupakheti, C., Virshup, A., Yang, W., Beratan, D.N., 2015. Strategy To Discover Diverse Optimal Molecules in the Small Molecule Universe. J. Chem. Inf. Model. 55, 529–537. https://doi.org/10.1021/ci500749q

SciFinder; Chemical Abstracts Service: Columbus, OH [WWW Document], n.d. URL https://scifinder.cas.org (accessed 1.14.21).

Sha, W., Li, Y., Tang, S., Tian, J., Zhao, Y., Guo, Y., Zhang, W., Zhang, X., Lu, S., Cao, Y.-C., Cheng, S., 2021. Machine learning in polymer informatics. InfoMat 3, 353–361.

https://doi.org/https://doi.org/10.1002/inf2.12167

Srirambhatla, V.K., Kraft, A., Watt, S., Powell, A. V, 2012. Crystal Design Approaches for the Synthesis of Paracetamol Co-Crystals. Cryst. Growth Des. 12, 4870–4879. https://doi.org/10.1021/cg300689m

Stephanopoulos, G., 1990. Artificial intelligence in process engineering—current state and future trends. Comput. Chem. Eng. 14, 1259–1270. https://doi.org/https://doi.org/10.1016/0098-1354(90)80006-W

Stephanopoulos, G., Johnston, J., Kriticos, T., Lakshmanan, R., Mavrovouniotis, M., Siletti, C., 1987. Design-kit: An object-oriented environment for process engineering. Comput. Chem. Eng. 11, 655–674. https://doi.org/https://doi.org/10.1016/0098-1354(87)87010-2

Venkatasubramanian, V., 2019. The promise of artificial intelligence in chemical engineering: Is it here, finally? AIChE J. 65, 466–478. https://doi.org/https://doi.org/10.1002/aic.16489

Venkatasubramanian, V., Zhao, C., Joglekar, G., Jain, A., Hailemariam, L., Suresh, P., Akkisetty, P., Morris, K., Reklaitis, G. V, 2006. Ontological informatics infrastructure for pharmaceutical product development and manufacturing. Comput. Chem. Eng. 30, 1482–1496. https://doi.org/https://doi.org/10.1016/j.compchemeng.2006.05.036

Wicker, J.G.P., Crowley, L.M., Robshaw, O., Little, E.J., Stokes, S.P., Cooper, R.I., Lawrence, S.E., 2017. Will they co-crystallize? CrystEngComm 19, 5336–5340. https://doi.org/10.1039/C7CE00587C

Wood, P.A., Feeder, N., Furlow, M., Galek, P.T.A., Groom, C.R., Pidcock, E., 2014. Knowledge-based approaches to co-crystal design. CrystEngComm 16, 5839–5848. https://doi.org/10.1039/C4CE00316K

Zhu, J., Rosset, S., Zou, H., Hastie, T., 2006. Multi-class AdaBoost. Stat. Interface 2. https://doi.org/10.4310/SII.2009.v2.n3.a8