A Performance Comparison of Two Machine Learning Models to Predict the Formation of Pharmaceutical Cocrystals

Joaquin F. Urbina[1]
Paul M. Morgan[2]
Alex Moralez[1]
Chelsea Herrera[1]

Abstract

The use of machine learning has recently attracted the pharmaceutical industry and academia because it is able to reliably predict the cocrystal formation outcomes of API-coformer combinations and thus lead to an efficient cocrystal screening approach. In this study, binary logistic regression and random forest models were developed with the intention of comparing their performance against predicting the cocrystal outcomes of a dataset of API-coformer combinations using a variety of inherent molecular features, and identifying which of these features tend to influence cocrystal formation more than others. The feature importance data of both models revealed that the most basic acceptor site on an API (basic pK_{a1}) seemed to be one of the most important features that can reliably predict the formation of cocrystals. It was also found that the random forest model showed superior performance over the binary logistic regression model in its predictive accuracy (0.901 vs 0.811 respectively) based on the ROC plots and confusion matrices. The cocrystal prediction power of these and other models will be further investigated by expanding the number and types of molecular properties and the size of the dataset.

Keywords: Pharmaceutical cocrystals, machine learning, cocrystal prediction, binary logistic regression model, random forest model

Introduction

The pharmaceutical industry and academia have seen many advances in the solid-state chemistry of oral dosage forms of active pharmaceutical ingredients (APIs), particularly in their physicochemical properties (*e.g.* aqueous solubility, dissolution rate, thermal stability, *etc.*) (Aakeröy et al., 2009; Aakeröy et al., 2014; Almeida e Sousa et al., 2016; Hickey et al., 2007; Laitinen et al., 2013; Sopyan et al., 2017). Such improvements prove vital because BCS Class II APIs are frequently hampered by poor aqueous solubility and limited bioavailability, which typically pose a significant challenge to the overall performance of the formulated drug product. Several solid forms have been formulated to overcome this barrier, such as salts,

[1] University of Belize, [2] Icahn School of Medicine at Mount Sinai, NY, USA
Corresponding Authors: Joaquin F. Urbina, Faculty of Science and Technology, University of Belize, Hummingbird Avenue, Belmopan, Belize. email: jurbina@ub.edu.bz;
Paul M. Morgan, Icahn School of Medicine at Mount Sinai, New York, NY, 10029, USA. email: paul.marcel.morgan@gmail.com
*Supplementary material published online at research.ub.edu.bz

solvates/hydrates, polymorphs, amorphous solid dispersions, and cocrystals, Figure 1 (Kavanagh et al., 2019).

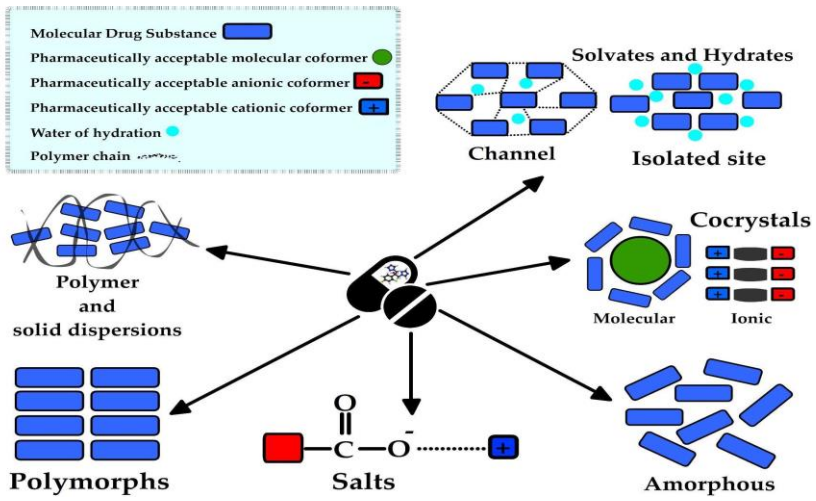

Figure 1. Possible solid forms that can be prepared from an API. (Kavanagh et al., 2019)

While the other solid forms have enhanced the physicochemical characteristics of APIs, cocrystals have attracted the attention of solid-state chemists in the field of crystal engineering because the improved properties do not compromise the structural integrity of APIs at the molecular level and no covalent modification is conducted on them. A cocrystal may exist as either a molecular cocrystal (MCC) or ionic cocrystal (ICC) (Aakeröy & Sinha, 2018; Kavanagh et al., 2019). An MCC is generally defined as a crystalline material composed of two or more molecular components (often involving a combination of APIs and coformers) in stoichiometric amounts stabilized by intermolecular interactions, notably halogen bonds and hydrogen bonds. On the other hand, an ICC is based upon a combination of neutral organic molecules and salts in a stoichiometric ratio.

It is well documented in the literature that hydrogen bonds are strong and directional non-covalent interactions that can be reliably deployed to form MCCs in a predictable manner (Aakeröy et al., 2001, 2002; Aakeröy, Desper, Elisabeth, et al., 2005; Aakeröy et al., 2004; Aakeröy, Desper, Leonard, et al., 2005; Aakeröy, Desper, & Urbina, 2005; Aakeröy et al., 2007). Based on these studies, a library of supramolecular synthons consisting of complementary hydrogen bond donor and acceptor moieties is now available to the solid-state chemist to perform non-covalent synthesis, Figure 2.

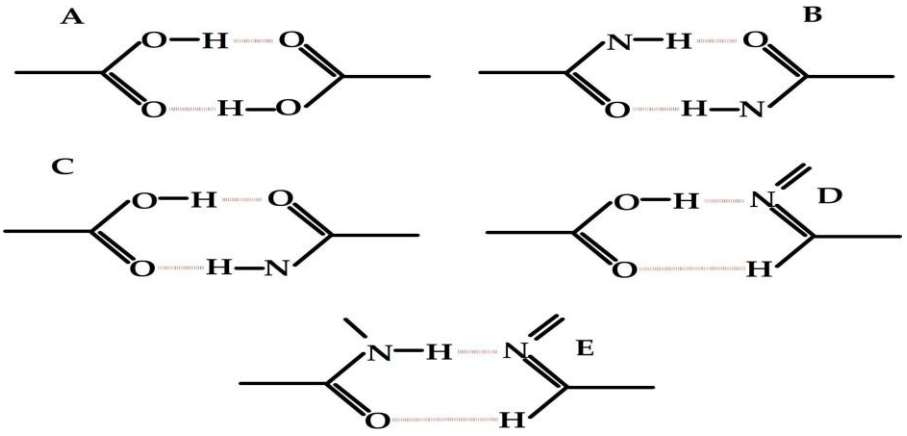

Figure 2. Examples of common supramolecular synthons used in crystal engineering (A) carboxylic acid dimer (B) amide dimer (C) carboxylic acid—amide (D) carboxylic acid—N-heterocycle, and (E) amide—N-heterocycle.

It is now possible to form MCCs through a hierarchical interplay of intermolecular interactions using the Etter rules whereby the best donor will hydrogen bond with the best hydrogen bond acceptor, the second best donor will hydrogen bond with the second best acceptor, and so forth, until all donor and acceptors are satisfied (Etter, 1991). Also, when the ΔpK_a between the two hydrogen bond moieties of the components is less than 0, proton transfer is less likely to occur and a cocrystal results (Cruz-Cabeza, 2012; Enkelmann et al., 2021; Qiao et al., 2011). These design concepts are therefore often integrated into experimental cocrystal screening, mostly involving combinatorial chemistry, mechanochemistry (Aakeröy et al., 2011; Delori et al., 2012; Tan et al., 2016), and crystallization techniques (Malamatari et al., 2017).

Even though these experimental methods have been used in the pharmaceutical industry to screen for cocrystals, they are largely based on a trial and error approach, produce unnecessary chemical waste, require more work and time, and are more expensive. This means that a more efficient and greener cocrystal screening process is needed. To this end, *in silico* screening strategies have been implemented to predict structural outcomes prior to experimentation. Examples include the use of the Cambridge Crystallographic Structural Database (CCSD) for hydrogen bond propensity, conductor-like screening model for real solvents (COSMO-RS), molecular electrostatic potential surfaces (MEPS), lattice energy calculations, Hirshfeld surface analysis, the Hansen solubility parameter (HSP), Gibbs free energy, cocrystallization propensity, and solubility advantage (Cysewski, 2017; Khalaji et al., 2021; Kumar & Nanda, 2021).

More recently, machine learning algorithms have been developed to perform virtual cocrystal screening, such as logistic regression, artificial neural networks based on CCSD data mining, a high throughput cocrystal screening model, and network science and link prediction algorithms using molecular descriptors (Devogelaer et al., 2020; Mswahili et al., 2021; Wang et al., 2020; Zheng et al., 2020), and these studies have subsequently led to increased cocrystal prediction power. In particular, a recent cocrystal prediction study revealed that a random forest model out-performed several other machine learning models, including logistic regression, AdaBoost, GradientBoosting, Multinomial Naïve Bayes, and Deep Neural Network, based on receiver operating characteristic (ROC) area under the curve (AUC) values, and was thus validated through the experimental screening of captopril cocrystals (Wang et al., 2020). The dataset used in that study was comprised of 2D structural information based on molecular descriptors extracted from the CCSD.

Herein, the novelty in our current machine learning approach is that it samples a training dataset that is founded on an array of molecular properties that cannot be extracted through the molecular descriptors afforded by RDKit. Although molecular descriptors provide a wealth of structural information about successful cocrystals (Wicker, 2017), we hypothesize that there are many critical molecular features buried within an API and a coformer that cannot be mined through molecular descriptors. Such a framework would provide a unique opportunity to explore the most inherently reliable features in an API and a coformer that would likely lead to predictable cocrystallization outcomes.

This study compares the performance of two machine learning models to predict the formation of binary cocrystals of an array of APIs and coformers possessing diverse hydrogen bond functional groups and molecular properties using a binary logistic regression model and a random forest model. The former model is inherently more simplistic, and it is therefore hypothesized that the random forest model would be more fitting to predict cocrystal outcomes considering the robustness of input variables in the training dataset. The performance of machine learning models is also dependent on the number and types of molecular properties and the size of the dataset.

Based on the hypotheses proposed above, the current study has two main objectives to achieve. First, it seeks to determine which molecular properties in APIs and coformers tend to influence cocrystal formation more than others. It also aims to determine and compare the predictive accuracies of both machine learning models and see the extent to which further optimization is needed.

Methods

BCS Class II APIs were chosen for this study due to their low solubility and high permeability. The coformers were selected based on their potential to hydrogen-bond with the APIs through the common supramolecular synthons illustrated in Figure 2. From the CCSD, there were several experimental crystal structures that contained these API-coformer combinations which resulted in either a 'cocrystal' or 'no cocrystal', which were subsequently entered into a 'Master Sheet' and used to train the dataset. The molecular properties of each API and coformer molecule were obtained through the use of several software tools, compiled, and organized into a 'Master Sheet', which served as the training dataset with an initial cohort of 542 observations, Figure 3 (also see Supplementary Information).

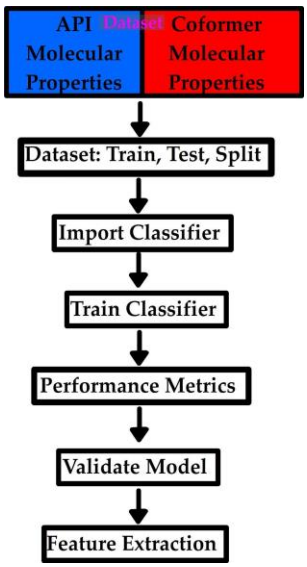

Figure 3. Workflow of methodology used to construct the machine learning models.

We employed Python's machine learning library - scikit learn (version 0.24.1) to develop a binary logistic regression classifier, and a random forest classifier (https://github.com/paulmorganjr/urbina_morgan_cocrystals). To construct our machine learning models, our dataset contained a total of 31 variables corresponding to several molecular properties of both the API and coformer. The binary logistic regression classification is a statistical model that uses one or more predictor variables that may be categorical or continuous to predict a target variable class, *i.e.* cocrystal formation or no cocrystal formation (Estiri et al., 2021). Alternatively, the random forest model is comprised of a host of decision trees that are based on several true or false conditions using the input data (Heo et al., 2019). The final classification is based on the sum of decisions made by the decision tree. Both models were fine-tuned with the most optimal parameters to ensure that the validation dataset would not be over fitted (Pfaff et al., 2022).

Python version 3.10.6 was used for statistical analyses (Patel et al., 2020). The receiver operating characteristic curve (ROC) analysis and the area under the curve (AUC) were calculated using an in-house python script to compare the efficacy of each model. All P values were 2-sided with a significance threshold value of <0.05 (Heo et al., 2019).

Results

Figure 4 compares the feature importance for the (A) binary logistic regression and (B) random forest models whereby the scores of all the molecular properties for both APIs and coformers were factored into each model. These scores are useful because they play a critical role in deciphering how useful a variable is in predicting an outcome. While most properties tend to influence cocrystal formation at varying degrees in each model, it seems that the most basic pK_a of an API is one of the most important properties, given its highest score, regardless of the type of model used.

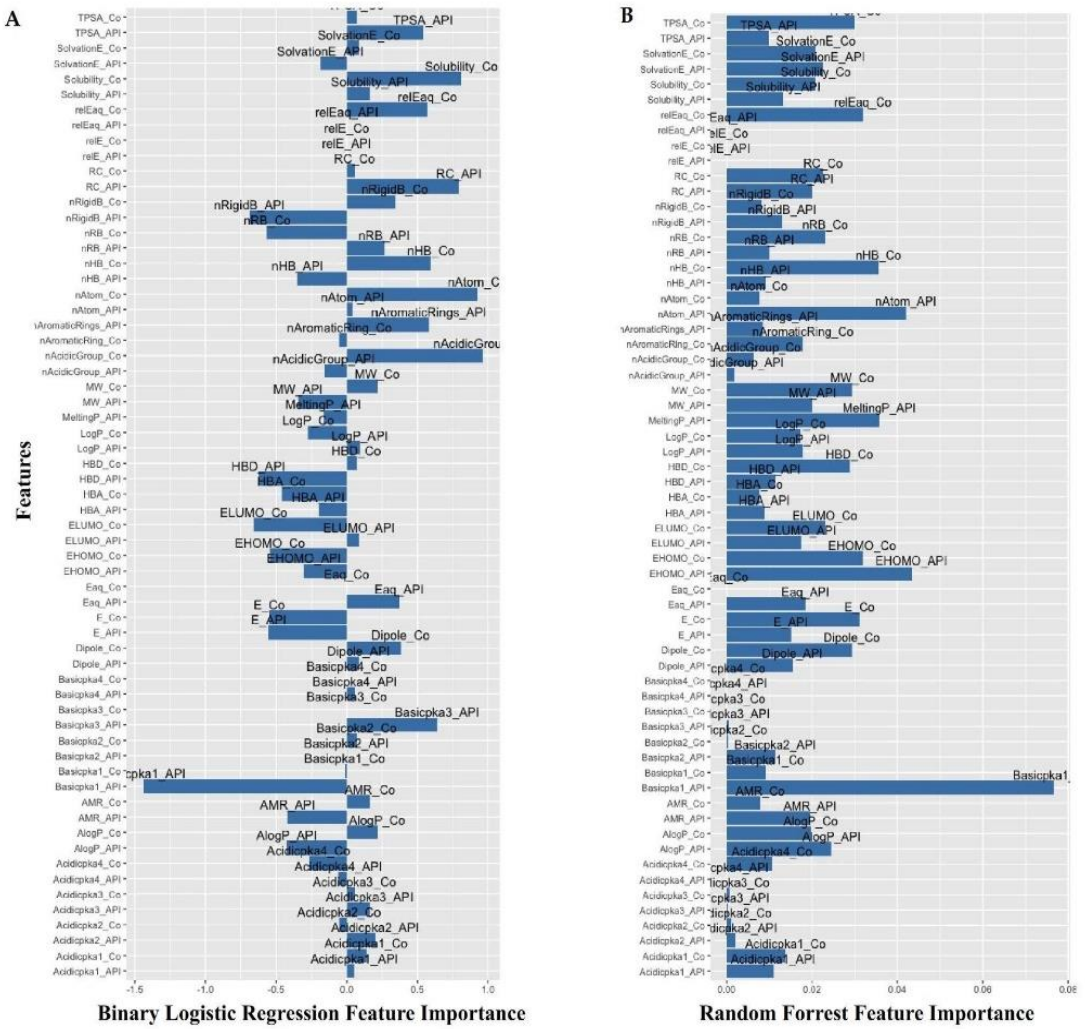

Figure 4. A. Binary logistic regression feature importance. B. Random Forest feature importance. Feature importance refers to scores that are assigned to molecular properties of both the API and coformer from the input dataset.

The receiver operating characteristic (ROC) curve and the confusion matrix for each model were also generated, Figure 5(A)-(D). While the area under the ROC (AUROC) curve was greater than 0.500 in each model, the AUROC value was greater (0.901) and the curve was steeper for the ROC plot in the random forest model compared to that in the binary logistic regression model (0.811), meaning that the rate of true positive cocrystal formation outcomes is greater than the rate of false positive outcomes. This observation agrees with the confusion matrix results because there is a greater number of predicted true positive (54) and true negative (79) cocrystal outcomes in the random forest model that coincided with the number of experimental cocrystal outcomes obtained from the CCSD in the training set compared to those in the binary logistic regression model (34 true positive and 45 true negative outcomes). At the same time, the number of predicted false positive (19 versus 17) and false negative (11 versus 13) cocrystal outcomes was comparable between the binary logistic regression and random forest models respectively.

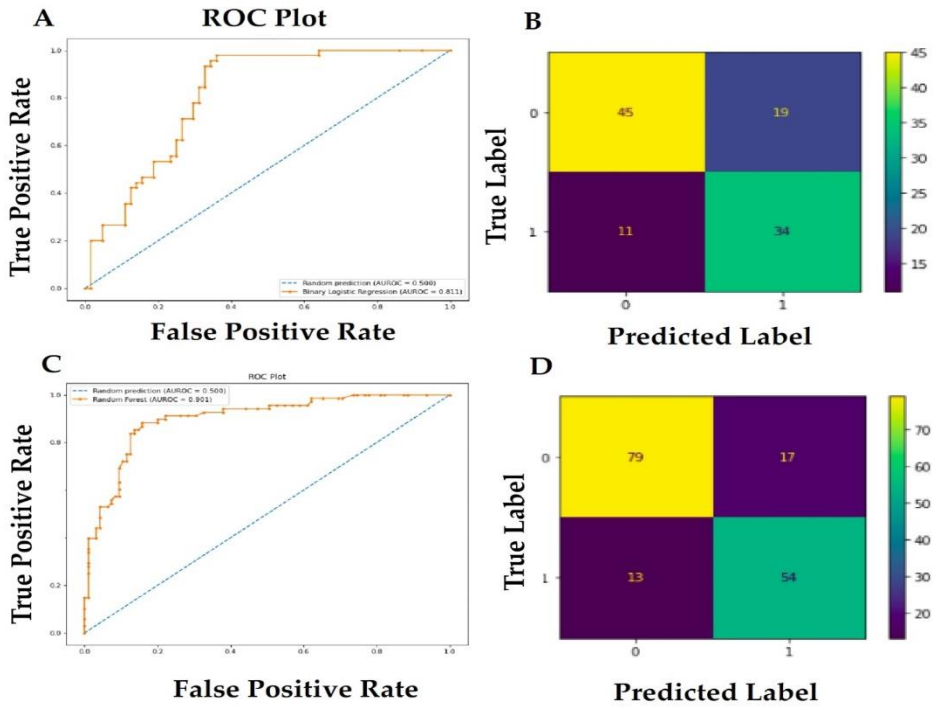

Figure 5. A. Receiver operating characteristic curve (ROC)

Figure 5. A. Receiver operating characteristic curve (ROC) for the binary logistic regression model with an area under the curve of 0.811. B. The confusion matrix performance metrics calculated from predictions of the binary logistic regression model on the test dataset. C. ROC for the random forest model with an area under the curve of 0.901. D. The confusion matrix performance metrics calculated from predictions of the random forest model on the test dataset.

Discussion

A comparison of the feature importance between the two models revealed that the basic pK_{a1} of an API seems to be one of the most important properties towards cocrystal formation, which relates to the first hypothesis. This is in line with previous work conducted by our research group on cocrystal predictions of azole APIs and mono- and dicarboxylic acids (Herrera, 2021), as well as with one of the Etter rules which states that the strongest hydrogen bond donor will interact with the strongest hydrogen bond acceptor, the second strongest donor with the second strongest acceptor, etc. in a hierarchical fashion until all sites have been satisfied (Etter, 1991). Ever since this concept has been postulated by Etter, a significant body of publications have supported it for the purposes of academic and pharmaceutical cocrystal research. At this time, it may be speculated that the choice of API needs to involve a careful consideration of the strongest acceptor site on the heterocyclic API such that a cocrystal is formed and not a salt. However, further computational and experimental validation research is required in order to better understand the degree of importance of the basic pK_{a1} of an API in cocrystal prediction when the dataset is expanded with additional features such as hydrogen bond propensity, lattice energy, Gibbs free energy, *inter alia*, as well as with a greater diversity of BCS Class II APIs and coformers. It will also be important to note whether there will be any significant inherent molecular properties in a coformer that need to be considered in cocrystal design after these further studies are conducted.

It was expected that the random forest model would out-perform the binary logistic regression model (AUROC 0.811 vs 0.901 respectively) as per our second hypothesis, since the former involved more robust input variables in the training dataset compared to the latter, suggesting that it was better fit as a predictor of cocrystal outcomes. Our work also implies that the use of machine learning models can accurately predict the cocrystallization outcome of an API. The ability to both be a predictor and identify relevant features makes the binary logistic regression and random forest models ideal in elucidating key molecular descriptors for cocrystal formation (Pfaff et al., 2022).

There are several limitations in this study. First, although a robust array of 31 molecular properties were considered in our machine learning models, there are still other features like hydrogen bond propensity, that are experimentally influential, which we did not consider. Integrating more experimental parameters from the CCSD is likely to improve the outcomes of predictions and validation but is also likely to change the weightings of feature importance. As to the magnitude of influence on feature importance, this is currently unknown, and is a priority in our future work. Second, the size of our dataset is relatively small with a total of 542 observations. We hope to robustly expand this dataset in an effort to increase the diversity. This expansion should improve our models and increase the accuracy of our experimentally predicted outcomes. Finally, we have only explored two machine learning models to classify cocrystallization prediction outcomes. A neural network based deep learning model is likely to out-perform the conventional machine learning models we have employed in this study. Within this scope of work, it is also important to systematically test the dataset using models of varying complexity in order to more accurately establish which intrinsic molecular features consistently play a salient role in predicting cocrystal formation. From these future studies, we intend to generate a list of predicted cocrystal outcomes involving new API-coformer combinations that are not part of the training dataset and validate these predictions against the experimental cocrystal screening results of these combinations.

Conclusion

We developed two machine learning algorithms that can accurately predict API cocrystal formation based on the molecular properties of an API and a coformer. The random forest model out-performed the binary logistic model with AUROC values of 0.901 and 0.811 respectively. Both models were also mined for molecular features that can provide further insight into molecular properties that seem critical to API cocrystallization. Of the 31 molecular properties that were explored, the basic pK_{a1} of an API appears to be a heavily weighted feature. We plan to explore the feature importance of key molecular properties that are critical to cocrystallization in a sequel study, as well as to validate our *in silico* predictions with experimental data through cocrystal screening studies.

Acknowledgements

The authors thank faculty members Dr. Apolonio Aguilar, Ms. Juliane Pasos, and Ms. Karen Link from the Chemistry program under the Faculty of Science and Technology at the University of Belize for their feedback on Chelsea and Alex's undergraduate research projects.

References

Aakeröy, C. B., Beatty, A. M., & Helfrich, B. A. (2001). "Total synthesis" supramolecular style: Design and hydrogen-bond-directed assembly of ternary supermolecules. *Angewandte Chemie International Edition*, *40*(17), 3240-3242.

Aakeröy, C. B., Beatty, A. M., & Helfrich, B. A. (2002). A high-yielding supramolecular reaction. *Journal of the American Chemical Society*, *124*(48), 14425-14432.

Aakeröy, C. B., Desper, J., Elisabeth, E., Helfrich, B. A., Levin, B., & Urbina, J. F. (2005). Making reversible synthesis stick: competition and cooperation between intermolecular interactions. *Zeitschrift für Kristallographie-Crystalline Materials*, *220*(4), 325-332.

Aakeröy, C. B., Desper, J., Helfrich, B. A., Desper, J., & Helfrich, B. A. (2004). Heteromeric intermolecular interactions as synthetic tools for the formation of binary co-crystals.

Aakeröy, C. B., Desper, J., Leonard, B., & Urbina, J. F. (2005). Toward High-Yielding Supramolecular Synthesis: Directed Assembly of Ditopic Imidazoles/Benzimidazoles and Dicarboxylic Acids into Cocrystals via Selective O– H⊙⊙⊙ N Hydrogen Bonds. *Crystal Growth & Design*, *5*(3), 865-873.

Aakeröy, C. B., Desper, J., & Urbina, J. F. (2005). Supramolecular reagents: versatile tools for non-covalent synthesis. *Chemical Communications*(22), 2820-2822.

Aakeröy, C. B., Fasulo, M. E., & Desper, J. (2007). Cocrystal or salt: does it really matter? *Molecular Pharmaceutics*, *4*(3), 317-322.

Aakeröy, C. B., Forbes, S., & Desper, J. (2009). Using Cocrystals To Systematically Modulate Aqueous Solubility and Melting Behavior of an Anticancer Drug. *Journal of the American Chemical Society*, *131*(47), 17048-17049. https://doi.org/10.1021/ja907674c

Aakeröy, C. B., Forbes, S., & Desper, J. (2014). Altering physical properties of pharmaceutical co-crystals in a systematic manner. *CrystEngComm*, *16*(26), 5870-5877.

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

Aakeröy, C. B., & Sinha, A. S. (2018). *Co-crystals: preparation, characterization and applications* (Vol. 24). Royal Society of Chemistry.

Almeida e Sousa, L., Reutzel-Edens, S. M., Stephenson, G. A., & Taylor, L. S. (2016). Supersaturation Potential of Salt, Co-Crystal, and Amorphous Forms of a Model Weak Base. *Crystal Growth & Design*, *16*(2), 737-748. https://doi.org/10.1021/acs.cgd.5b01341

Cruz-Cabeza, A. J. (2012). Acid–base crystalline complexes and the p K a rule. *CrystEngComm*, *14*(20), 6362-6365.

Cysewski, P. (2017). In silico screening of dicarboxylic acids for cocrystallization with phenylpiperazine derivatives based on both cocrystallization propensity and solubility advantage. *Journal of molecular modeling*, *23*(4), 1-11.

Delori, A., Friščić, T., & Jones, W. (2012). The role of mechanochemistry and supramolecular design in the development of pharmaceutical materials. *CrystEngComm*, *14*(7), 2350-2362.

Devogelaer, J. J., Meekes, H., Tinnemans, P., Vlieg, E., & De Gelder, R. (2020). Co-crystal Prediction by Artificial Neural Networks. *Angewandte Chemie International Edition*, *59*(48), 21711-21718.

Enkelmann, D., Lipinski, G., & Merz, K. (2021). Cyanopyridines–Suitable Heterocycles for Cocrystal Syntheses. *European Journal of Inorganic Chemistry*, *2021*(33), 3367-3372.

Estiri, H., Strasser, Z. H., Brat, G. A., Semenov, Y. R., Patel, C. J., & Murphy, S. N. (2021). Evolving Phenotypes of non-hospitalized Patients that Indicate Long Covid. *medRxiv*, 2021.2004.2025.21255923. https://doi.org/10.1101/2021.04.25.21255923

Etter, M. C. (1991). Hydrogen bonds as design elements in organic chemistry. *The Journal of Physical Chemistry*, *95*(12), 4601-4610. https://doi.org/10.1021/j100165a007

Heo, J., Yoon, J. G., Park, H., Kim, Y. D., Nam, H. S., & Heo, J. H. (2019). Machine Learning-Based Model for Prediction of Outcomes in Acute Stroke. *Stroke*, *50*(5), 1263-1265. https://doi.org/10.1161/strokeaha.118.024293

Herrera, C. C. (2021). *Preliminary Structural Prediction of Co-crystal Formation with Antifungal Azole-Based Active Pharmaceutical Ingredients*. University of Belize.

Hickey, M. B., Peterson, M. L., Scoppettuolo, L. A., Morrisette, S. L., Vetter, A., Guzmán, H., Remenar, J. F., Zhang, Z., Tawa, M. D., Haley, S., Zaworotko, M. J., & Almarsson, Ö. (2007). Performance comparison of a co-crystal of carbamazepine with marketed product. *European Journal of*

Pharmaceutics and Biopharmaceutics, *67*(1), 112-119. https://doi.org/https://doi.org/10.1016/j.ejpb.2006.12.016

Kavanagh, O. N., Croker, D. M., Walker, G. M., & Zaworotko, M. J. (2019). Pharmaceutical cocrystals: from serendipity to design to application. *Drug Discovery Today*, *24*(3), 796-804. https://doi.org/https://doi.org/10.1016/j.drudis.2018.11.023

Khalaji, M., Potrzebowski, M. J., & Dudek, M. K. (2021). Virtual cocrystal screening methods as tools to understand the formation of pharmaceutical cocrystals—a case study of linezolid, a wide-range antibacterial drug. *Crystal Growth & Design*, *21*(4), 2301-2314.

Kumar, A., & Nanda, A. (2021). In-silico methods of cocrystal screening: A review on tools for rational design of pharmaceutical cocrystals. *Journal of Drug Delivery Science and Technology*, *63*, 102527.

Laitinen, R., Löbmann, K., Strachan, C. J., Grohganz, H., & Rades, T. (2013). Emerging trends in the stabilization of amorphous drugs. *International Journal of Pharmaceutics*, *453*(1), 65-79. https://doi.org/https://doi.org/10.1016/j.ijpharm.2012.04.066

Malamatari, M., Ross, S. A., Douroumis, D., & Velaga, S. P. (2017). Experimental cocrystal screening and solution based scale-up cocrystallization methods. *Advanced Drug Delivery Reviews*, *117*, 162-177. https://doi.org/https://doi.org/10.1016/j.addr.2017.08.006

Mswahili, M. E., Lee, M.-J., Martin, G. L., Kim, J., Kim, P., Choi, G. J., & Jeong, Y.-S. (2021). Cocrystal prediction using machine learning models and descriptors. *Applied Sciences*, *11*(3), 1323.

Patel, L., Shukla, T., Huang, X., Ussery, D. W., & Wang, S. (2020). Machine Learning Methods in Drug Discovery. *Molecules*, *25*(22). https://doi.org/10.3390/molecules25225277

Pfaff, E. R., Girvin, A. T., Bennett, T. D., Bhatia, A., Brooks, I. M., Deer, R. R., Dekermanjian, J. P., Jolley, S. E., Kahn, M. G., Kostka, K., McMurry, J. A., Moffitt, R., Walden, A., Chute, C. G., & Haendel, M. A. (2022). Identifying who has long COVID in the USA: a machine learning approach using N3C data. *Lancet Digit Health*, *4*(7), e532-e541. https://doi.org/10.1016/s2589-7500(22)00048-6

Qiao, N., Li, M., Schlindwein, W., Malek, N., Davies, A., & Trappitt, G. (2011). Pharmaceutical cocrystals: an overview. *International Journal of Pharmaceutics*, *419*(1-2), 1-11.

Sopyan, I., Fudholi, A., Muchtaridi, M., & Sari, I. P. (2017). Co-crystallization: a tool to enhance solubility and dissolution rate of simvastatin. *Journal of Young Pharmacists*, *9*(2), 183.

Tan, D., Loots, L., & Friščić, T. (2016). Towards medicinal mechanochemistry: evolution of milling from pharmaceutical solid form screening to the synthesis of active pharmaceutical ingredients (APIs). *Chemical Communications*, *52*(50), 7760-7781.

Wang, D., Yang, Z., Zhu, B., Mei, X., & Luo, X. (2020). Machine-Learning-Guided Cocrystal Prediction Based on Large Data Base. *Crystal Growth & Design*, *20*(10), 6610-6621. https://doi.org/10.1021/acs.cgd.0c00767

Wicker, J. G. P. C., Lorraine M; Robshaw, Oliver; Little, Edmund J; Stokes, Stephen P; Cooper, Richard I; Lawrence, Simon E. (2017). Will they co-crystallize? *CrystEngComm*, *19*(36), 5336-5340.

Zheng, L., Zhu, B., Wu, Z., Fang, X., Hong, M., Liu, G., Li, W., Ren, G., & Tang, Y. (2020). Strategy for Efficient Discovery of Cocrystals via a Network-Based Recommendation Model. *Crystal Growth & Design*, *20*(10), 6820-6830. https://doi.org/10.1021/acs.cgd.0c00911

Supplementary Information

1. Methodology for API and Coformer Molecular Property Determination and Preparation of Master Sheet for Binary Logistic Regression and Random Forest Analyses

PubChem and Drugbank

The freely accessible chemical information online database PubChem was used to download the 3D '.sdf' file of APIs and coformers for the determination of computational values for each molecule. The online Drug database Drugbank was used to obtain a general definition of each API and a brief context on its general usage. Drugbank also provided the melting point for all APIs while PubChem provided the coformer melting points. APIs and coformers that rendered no experimental melting points were given a value of zero.

Determination of pK_a Values

The pK_a values of hydrogen-bond donor and acceptor groups in APIs and coformers were determined using the database available in the Chemicalize pK_a calculator from ChemAxon. In instances where a specific coformer structure was not accessible, the software's tools were utilized to build a 2D representation of the structure and searched in that format against its database. Solubility values (mg/mL) of each component were also determined using the Chemicalize pK_a calculator and tabulated. The differences in pK_a (ΔpK_a) between the pK_a values of the APIs and coformers were calculated using Equation 1. This same procedure was repeated for the API's second most acid pK_a and the second most basic pK_a of coformers.

$$\Delta pK_a = [\text{API } pK_a]_{\text{most acidic}} - [\text{Coformer } pK_a]_{\text{most basic}} \qquad \text{Equation 1}$$

Determination of Computational Values

Wavefunction's Spartan 14.V1.1.4 molecular modeling and computational chemistry software contains codes for molecular mechanics, semi-empirical methods, *ab initio,* models, density functional models, and post-Hartree-Fock models. Based on past *in silico* studies, Density Functional Theory was used as a primary code to calculate Energy. The structures generated from PubChem were used as structural input for the calculations to be made. Due to the lack of specific mention in literature, Energy was chosen as the main parameter to render the calculations. A basis set of EDF2/6-31G* at a ground state specie in a vacuum setting was chosen to calculate the Highest Occupied Molecular Orbital (HOMO) energy values, Lowest Unoccupied Molecular Orbital (LUMO) energy values, dipole energy, and energy associated with water (E_{aq}). In cases where the coformer contained ionic charges, the total charge was determined by summing up all the individual charges present. In such cases, the total charge was changed from neutral and

replaced by the appropriate overall charge value along with the unpaired electrons if any were present. In cases where the cocrystal component already existed in the Spartan Molecular Database (SMD) and had available properties, no computational calculations were performed. All the properties collected for each API were appended using the "append molecule(s)" option within Spartan and then exported as a spreadsheet. The same step was followed for the properties of each coformer. For components that had no available 3D structure, the Spartan automatic conversion feature was used to translate the 2D structure into their 3D renditions. Other structures that had no 3D or 2D file had to be manually drawn using the Spartan 2D drawing kit and the software then allowed to transform them into their 3D representations. The arrangement of the columns in the spreadsheet all depended on the researcher's preference.

Determination of API and Coformer Drug-Like Properties

Drug-likeness rules are a set of guidelines for the structural properties of compounds and used for fast calculations of drug-like properties of a molecule. The guidelines are not absolute, nor are they intended to form strict cutoff values for which property values are drug-like and which are not drug-like. Nevertheless, they are quite effective and efficient. DrugLiTo is an open-source virtual screening tool, and its calculations are based on various drug-likeness rules like Lipinski's rule. The drug-like properties of each API and coformer were calculated using this tool, and copied and pasted on the same spreadsheet used to export the properties calculated from Spartan. Each worksheet was labeled according to the property they hold, i.e. API properties and coformer properties.

Formatting of Molecular Properties and Preparing the Master Sheet

The molecular properties of APIs and coformers were compiled into a 'Master Sheet' (see Supplementary Information) which was then used to train the binary logistic regression and random forest models. Properties that repeated themselves were eliminated to avoid repetition. Once all the properties of each component matched with each other (i.e the same properties were considered for both API and coformer), the master sheet was compiled. This was accomplished by making reference to an 'Experimental Formulation' sheet, which contained API-coformer combinations found in the CCSD software using the free online version.[1] For example, if acyclovir formed a cocrystal with aspirin, both components were located inside each of their respective worksheets and placed next to each other on the new worksheet labeled as 'Master Sheet'. In the end, two columns were created, one labeled as 'Formation' and the other as 'Binary Formation'. Depending on the structural outcome, each combination was labeled as either CC (cocrystal), S (salt), So (solvate), Ca (Clathrate), or H (Hydrate) and a 1 if CC was assigned or a 0 if it was any designation other than CC. Final Column titles were formatted so that the entire sheet can be in an acceptable format for developing the models.

2. Results

Table 1. Shows 31 molecular features and their importance weighting for the binary logistic regression model and the random forest model that comprise our dataset.

	Feature Importance	
Features (*Molecular Properties*)	**Binary Logistic Regression**	**Random Forest**
MW_API	-0.34496969	0.02011641
LogP_API	0.09257833	0.01782247
AlogP_API	-0.42462263	0.0245608
HBA_API	-0.20132495	0.00886859
HBD_API	-0.62442268	0.01134749
TPSA_API	0.54034361	0.00991255
AMR_API	-0.41851244	0.01955798
nRB_API	0.26624186	0.0100314
nAtom_API	0.04168245	0.04197165

nAcidicGroup_API	-0.16019191	0.00189427
RC_API	0.79381706	0.02001658
nRigidB_API	-0.68929588	0.0130757
nAromaticRings_API	0.58107963	0.00847413
nHB_API	-0.34935948	0.00908816
E_API	-0.55223561	0.01499251
Eaq_API	0.37131394	0.01846656
relE_API	0	0
relEaq_API	0	0
EHOMO_API	-0.30262597	0.04325314
ELUMO_API	0.08710924	0.01757253
Dipole_API	0.08449122	0.0153998
SolvationE_API	-0.18724896	0.02257039
Solubility_API	0.16372218	0.01328888

Acidicpka1_API	0.05357868	0.01089692
Acidicpka2_API	0.20130881	0.00195928
Acidicpka3_API	0.16220816	0.00014866
Acidicpka4_API	-0.05847491	0
Basicpka1_API	-1.44097491	0.07664851
Basicpka2_API	-0.00179647	0.01138154
Basicpka3_API	0.64020799	0.00059525
Basicpka4_API	0.05847491	0
MeltingP_API	-0.1661764	0.03576395
MW_Co	0.21619311	0.02943753
LogP_Co	-0.2720467	0.01732715
AlogP_Co	0.21643399	0.01971686
HBA_Co	-0.46028155	0.00769478
HBD_Co	0.06810168	0.02878199

TPSA_Co	0.06767133	0.02998355
AMR_Co	0.16483552	0.00787716
nRB_Co	-0.56388985	0.02309109
nAtom_Co	0.9255284	0.00763587
nAcidicGroup_Co	0.96255367	0.00635825
RC_Co	0.05686386	0.02258121
nRigidB_Co	0.34150863	0.00813558
nAromaticRing_Co	-0.05113562	0.01791981
nHB_Co	0.59733783	0.0355671
E_Co	-0.54851207	0.03114957
Eaq_Co	0	0
relE_Co	0	0
relEaq_Co	0.57216983	0.0319403
EHOMO_Co	-0.54257769	0.03173921

ELUMO_Co	-0.65825084	0.02301752
Dipole_Co	0.38246829	0.02932621
SolvationE_Co	0.08760538	0.02091158
Solubility_Co	0.81233268	0.02091158
Acidicpka1_Co	0.13818505	0.01383913
Acidicpka2_Co	-0.05359446	0.00117207
Acidicpka3_Co	0.05564166	0.00065581
Acidicpka4_Co	-0.26959601	0.01062214
Basicpka1_Co	-0.01439614	0.00906533
Basicpka2_Co	0.06959204	0.00041044
Basicpka3_Co	0	0
Basicpka4_Co	0	0

References

(1) Groom, C. R.; Bruno, I. J.; Lightfoot, M. P.; Ward, S. C. The Cambridge Structural Database. *Acta Crystallographica Section B* **2016**, *72* (2), 171-179. DOI: doi:10.1107/S2052520616003954.