ACS OMEGA

http://pubs.acs.org/journal/acsodf

Mini-Review

Progress in Research on Artificial Intelligence Applied to Polymorphism and Cocrystal Prediction

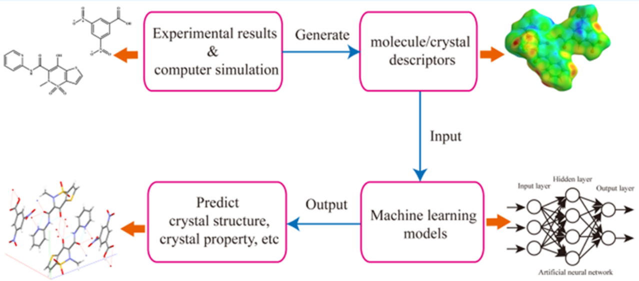
Tianyu Heng, Dezhi Yang, Ruonan Wang, Li Zhang,* Yang Lu,* and Guanhua Du

Cite This: *ACS Omega* 2021, 6, 15543−15550

Read Online

ACCESS | Metrics & More | Article Recommendations

ABSTRACT: Artificial intelligence (AI) is a technology that builds an artificial system with certain intelligence and uses computer software and hardware to simulate intelligent human behavior. When combined with drug research and development, AI can considerably shorten this cycle, improve research efficiency, and minimize costs. The use of machine learning to discover novel materials and predict material properties has become a new research direction. On the basis of the current status of worldwide research on the combination of AI and crystal form and cocrystal, this mini-review analyzes and explores the application of AI in polymorphism prediction, crystal structure analysis, crystal property prediction, cocrystal former (CCF) screening, cocrystal composition prediction, and cocrystal formation prediction. This study provides insights into the future applications of AI in related fields.



■ INTRODUCTION

Artificial intelligence (AI) is a cutting-edge comprehensive discipline that integrates various fields, such as computer science, statistics, neurology, and social science. Research on AI includes robotics, language or image recognition, natural language processing and expert systems. It investigates the laws of human intelligence activities, constructs artificial systems with certain intelligence, and explores how computer software and hardware can be used to simulate certain intelligent human behaviors.

AI includes machine learning, deep learning, data analysis, and data mining. Machine learning is an important application of AI and a powerful tool for finding relevant patterns in high-dimensional data. Computers can learn from empirical data and use algorithms to simulate linear or nonlinear relationships between material properties and related factors. Machine learning can be classified to four types, supervised machine learning, unsupervised machine learning, semisupervised machine learning and reinforcement machine learning. Data mining is a science of extracting useful information from large data sets or databases and it overlaps with some machine learning algorithms.

The studies of polymorphism and cocrystals are at the frontier and a hotspot in the field of solid drugs. Polymorphism is deemed as crystal systems of substances with different unit cells and with the same elemental composition.[1] Solvates and amorphous forms are also included in the category of polymorphism. Pharmaceutical cocrystals refer to the multicomponent crystals formed by the active pharmaceutical ingredient (API) and the cocrystal former (CCF) through hydrogen bonds or other noncovalent bonds with a fixed stoichiometric ratio between them. Polymorphism can influence many physical and chemical properties of solid drugs, such as melting point, density, solubility, dissolution rate, bioavailability, clinical efficacy, and toxicity. Pharmaceutical cocrystals can improve physicochemical properties of solid drugs as well. Meanwhile, the introduction of CCFs in pharmaceutical cocrystals, such as drug−drug cocrystals, is likely to produce the synergistic or complementary effects of pharmacological activities. Moreover, polymorphs and cocrystals of drugs can be the ways to protect intellectual property rights and extend the patent protection period. In terms of polymorphs and cocrystal screening, traditional experiments and computer modeling consume a lot of time and resources and are limited by experimental conditions and theoretical foundations. Thus, the use of AI to predict material properties and discover new materials has become a new research direction. A combination of AI and drug polymorphism or cocrystals can greatly shorten solid drug research and development cycle and costs. In addition, raw material particles of API with designed properties and functions are the research trends of solid drugs. Studies combining AI and drug

Received: March 11, 2021
Accepted: May 28, 2021
Published: June 11, 2021

https://doi.org/10.1021/acsomega.1c01330

polymorphism or cocrystals complement experimental studies, can effectively assess the risk of polymorphism, provide a deeper understanding of crystal structures, and make the control of drug solid forms more probable.

This mini-review is divided into four sections. We summarize previous work in Table 1. Section 1 introduces concepts of common algorithms. Section 2 demonstrates studies of AI applied to polymorphism, including polymorphism prediction, crystal structure analysis, and crystal property prediction. Section 3 demonstrates combinations of AI and cocrystal, including CCF screening, cocrystal composition prediction, and cocrystal formation prediction. Section 4 is the conclusion, and in this part we discuss the challenges and future development trends in this field.

Table 1. Application of AI on Polymorphism and Cocrystal Prediction

field	application	algorithm
polymorphism	polymorphism prediction	random forest
	crystal structure analysis	artificial neural network
	crystal property prediction	support vector machine, logistic regression
cocrystal	CCF screening	cluster analysis
	cocrystal composition prediction	principal component analysis
	cocrystal formation prediction	multivariate adaptive regression splines

1. COMMON ALGORITHMS

Commonly used machine learning families are seen in Figure 1. Supervised machine learning utilizes knowledge from labeled data to forecast events. It compares the obtained results with the actual or expected results to identify errors to change the model. Unsupervised machine learning analyzes how to explain the hidden patterns from the unlabeled data and does not identify the proper output.[2] Supervised learning includes classification and regression. Unsupervised learning includes clustering and dimension reduction. Regression analysis is a statistical method to find a correlation between the response and predictors. Clustering divides data into groups based on a similarity metric to uncover patterns and categories but does not directly predict new values. Regression and classification algorithms can predict material properties, and the clustering algorithm can be adopted in discovering novel materials. The steps of machine learning include first building data sets, then establishing models, and finally evaluating models. The data sources are from computer simulation and experimental results.

Some of the common AI algorithms are random forest (RF), artificial neural network (ANN), support vector machine (SVM), and logistic regression.

Random forests are a combination of tree predictors such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest. The right kind of randomness makes them accurate classifiers and regressors. They are relatively robust to outliers and noise and do not overfit easily.[3] Prediction is done by weighting votes (in classification) or averages (in regression) of the ensemble outputs. They can also evaluate the importance of input variables.

Artificial neural networks are mathematical models simulating biological nervous systems with strong self-learning and adaptive capabilities. The basic processing elements of neural networks are called artificial neurons. An artificial neural network consists of an input layer, one or two (or even three) hidden layers, and an output layer. The input layer receives input information, and the output layer generates outputs relating to the property to be predicted. ANN models work by processing input values through networks of hidden layers. Each connection in the network is called a synapse. ANN models are trained by adjusting the weights of each synapse until the output is close to the training data. Large data sets are usually needed to adequately train ANNs due to their lack of interpretability. ANNs are vulnerable to overfit, and there is danger to learn the noise from the data set as well. So it is required that the training processes are stopped close to the optimal time.[4]

The support vector machine algorithm is a supervised learning algorithm used for both data classification and regression analysis. It maps the input vectors into some high-dimensional feature space Z with a linear decision surface, ensuring the high generalization ability of the network. To construct such optimal hyperplanes, one only has to take into account a small amount of the training data, the so-called support vectors, which determine this margin. Characteristics like capacity control and ease of changing the decision surface render the support−vector network an extremely powerful and universal learning machine.[5]

Logistic regression analysis is a statistical technique to evaluate the relationship between various variables (either categorical or continuous) and a binary outcome. Fitting a logistic relation between a predictor x and a proportion of

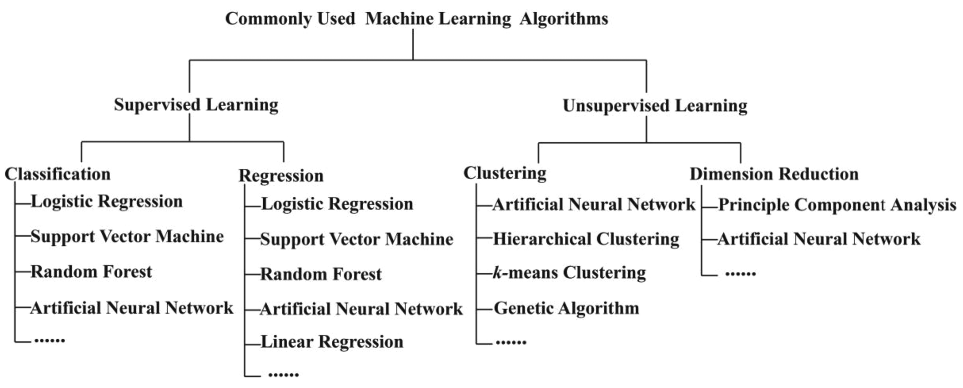


Figure 1. Commonly used machine learning algorithms.

success *y* is done by fitting a linear relation between predictor *x* and the logit of *y*.[6]

2. AI AND POLYMORPHISM

2.1. Polymorphism Prediction. The software for predicting drug crystal form is divided into two categories: one is based on molecular mechanics, and the other is based on quantum mechanics to predict molecular crystal form. The software based on molecular mechanics is mainly represented by the Polymorph module in the Material Studio software. It is a set of algorithms designed to determine low-energy forms in polymorphs. For example, if the molecular structure is known, the Polymorph module can be used to find all possible energy crystal structures and molecular arrangement rules by calculating the minimum lattice energy. The most likely crystal form is calculated by cluster analysis and energy arrangement. This method can be correlated with X-ray diffraction experimental data or achieved by examining the molecular structure of a drug. The software based on quantum mechanics is mainly represented by the universal crystal structure prediction software, that is, the evolutionary crystallography software package (universal structure predictor: evolutionary Xtallography). It can be used to predict both atomic and molecular crystal structures. The stable and metastable structures of drug molecular crystals can be predicted on the basis of molecular structures only. They can also quickly simulate and search for stable compositions and structures according to experimentally obtained unit cell parameters, fixed unit cell shapes, and unit cell volumes.[7]

Crystal structure prediction (CSP) is a part of crystal form prediction. Ten international research institutes have organized six worldwide crystal structure predictions in 1999, 2001, 2004, 2007, 2010, and 2015. The correctness of the prediction algorithm is judged by evaluating the consistency between several known but unpublished crystal structures and the structures predicted by the software. The fourth time, software with a prediction success rate of 100% appeared.[17] The following section elaborates on the application of AI to crystal form prediction.

2.1.1. Data Mining and Polymorphism Prediction. Polymorphism prediction can help researchers understand and evaluate the reliability of the crystal structure. The combination of knowledge-based theoretical analysis and existing experimental techniques can make a better judgment of hidden threats related to solid form as soon as possible, effectively reduce risks, and help people select better commercial solid form. It can also rationalize the failed and successful experiments. Thus, polymorphism prediction is crucial from the perspective of drug production, drug processing, property, and stability. Mining information from existing solid forms can be used to evaluate the possibility of the existence of material solid forms. The information contained in the Cambridge Structural Database (CSD) has been utilized to build training data sets for statistical modeling. CSD has added a Python-based Application Programming Interface (CSD Python API) to the search method. The CSD Python API allows users to create custom scripts to perform multifunctional searches. A series of software in the CSD system can be used for molecular, intermolecular, and supramolecular analysis. Some of the commonly used CSD tools are ConQuest, Mogul, and IsoStar.[10] The data of bond length, valence angle, and torsion angle obtained from CSD can be analyzed by Mogul. Isostar is used to analyze the molecular interaction. It has been extended to generate a Full Interaction Map (FIM), which can provide a qualitative description of structural stability. The Materials module in Mercury is used to analyze the intermolecular interaction and packing order,as well as the crystal packing similarity. By combining chemical data with crystallographic data like crystal packing description, interaction frequency, interaction potential energy calculation, and hydrogen bond tendency, many problems can be solved, such as the reliability of crystal structure, the possibility of polymorphs, and the crystal morphology of compounds. The hydrogen bond propensity (HBP) tool is provided in Mercury's Solid Form module, which can quantitatively determine the possibility of hydrogen bond formation between various functional groups from two-dimensional aspects. It can be used to evaluate the possibility of a polymorph. All possible combinations of hydrogen bond donor and acceptor of target molecule can be plotted by graph and ranked according to propensity and coordination score. The disadvantage of the HBP algorithm is that it does not give the properties of the predicted crystal form.

Nauha et al.[8] applied the HBP algorithm to some pharmaceutical compounds that do not yet have reported polymorphic forms. They discovered polymorphic forms in two drugs. Data mining is performed on over 600 000 entries in the database to predict various chemical properties of molecules, and the most likely hydrogen bonds are calculated by statistical methods. In another study, Nauha et al.[9] combined the HBP algorithm with experiments and found three crystal forms of probenecid, which all showed the same hydrogen bonding pattern. Feeder et al.[10] used the Mogul software to evaluate molecular conformations in the context of CSD structures. Moreover, they used the Solid Form module in Mercury and the Full Interaction Maps method to evaluate polymorphic stability. They discovered an unusual supramolecular structure and determined the most stable polymorph in a follow-up study.

2.1.2. Machine Learning and Crystal Form Prediction. Machine learning is presently applied to predict the different crystal forms of solvates, especially to analyze the formation of solvate crystal forms. The RF algorithm is the most dominant among various multivariate data analysis tools. Johnston et al.[11] were the first to apply machine learning to predict solvate formation. By establishing an RF classification model that includes solvent properties, experimental conditions, and known crystallization results, they performed directional crystallization experiments and obtained three new solvates of carbamazepine. Takieddin et al.[12] used molecular descriptors and machine learning methods to extract over 19 000 molecular structures from CSD to predict solvates and hydrates and determined the structural features that facilitate solvate production. They compared ANN, SVM, and logistic regression models. Xin et al.[13] employed CSD Python API to screen drug molecular structures. They compared RF and SVM models and predicted solvate formation propensity for pharmaceutical molecules with a prediction success rate of 86%. They also compared the importance of different driving forces on the formation of different solvates. The challenge of applying machine learning on crystal prediction is that the influencing factors of a specific crystal form are uncertain, and the relationship between the obtained crystals and experimental conditions is unclear. The selection of descriptors needs to be a priori and must be designed according to the actual situation of crystallization and the complexity of molecular structure and crystal packing.

Predicting crystal structures is an important direction. Predicting crystal structures can help design crystals with specific properties. On the other hand, first-principles calculations for CSP can be used to screen materials before synthesis. Crystal structure prediction needs density functional theory (DFT) calculation. Calculating the accurate lattice energy of a large number of crystal structures is challenging because of the high cost of calculation. Previous techniques involved calculating and comparing the energy of crystal structures in order to predict the most thermodynamically stable crystal form. Machine learning can be used to work with experiments and first-principles methods to rapidly provide probabilistic predictions rather than calculations.

The implementation idea is to input the results of machine learning into energy-based algorithms to create more accurate predictions. Another idea is to apply energy-based algorithms, such as the density functional method, to output data of compounds to train the machine learning algorithm. Oliynyk et al.[14] devised an SVM model to predict the crystal structures of binary and ternary inorganic compounds. They achieved 93.2% prediction accuracy on a training set of 706 compounds. David[15] combined machine learning and CSP and offered a means to expand the range of available energy models used in CSP without largely increasing computational cost, which benefited polymorphic screening and computer-guided material discovery. Their future applications are to predict larger molecules with more flexible conformation.

The success in predicting crystal structure proves the effectiveness of machine learning methods in exploring chemical white space. Due to the need for large training sets and knowledge of coding and algorithm deployment, there are problems in physical science with machine learning algorithms. Researchers should combine each technology to make up for the shortcomings of each other. Generating reliable solid structure details and properties from molecular descriptors by the calculation method is also important. The challenge is that there is still no general machine learning algorithm to predict crystal structure from molecular structure.

2.2. Crystal Structure Analysis. Crystal packing has a substantial influence on physiochemical properties and is at the core of describing and understanding polymorphism. One of the urgent needs of crystal engineering is to be able to compare crystal structures. Ideally, people want to get the similarity index between the two crystal structures. The combination with machine learning can provide a new solution.

Collins et al.[16] adopted a data mining method to combine cluster analysis with a fingerprint of the Hirshfeld surface to compare overall crystal structure similarities of a series of compounds. They provided a simple method to obtain information on the crystal packing trend. Bhardwaj et al.[17] developed an RF classification model using calculated solvent physicochemical properties and previous experimental crystal packing analysis. They were the first to predict three-dimensional crystal packing of different types of olanzapine solvates. They obtained a new solvate and identified three factors affecting the type of crystal packing.

Principal component analysis (PCA) is a statistical nonparametric method for extracting relevant information from redundant and noisy data sets. Gavezzotti et al.[18] used the TANAGRA data mining software and applied the PCA model to obtain multivariate correlations between global descriptive variables such as molecular mass, overall polarity, lattice energies, and their Coulombic, polarization-dispersion components. It provided and explained a lot of geometric and energy data on the interactions between molecules. Yang et al.[19] employed unsupervised machine learning methods to reveal the effects of organic structure and crystal symmetry on lattice energy diagrams.

In aspects of other algorithms, Phillips et al.[20] combined shape-matching and machine-learning algorithms to identify simple and complex crystal structures and to discover new crystal structures.

2.3. Crystal Property Prediction. *2.3.1. Crystallinity Prediction.* The crystallization process can be time-consuming and expensive. Therefore, prediction tools have considerable value and can be used in the early stage of development to identify molecular systems with possible crystallization problems. Molecular chemical descriptors related to crystal formation are needed to build a crystallinity prediction model, and a series of molecular descriptors are evaluated to select the important ones.

Bernard et al.[21] used the SIMCA-P software to evaluate the influence of melting point, glass transition temperature, and heat of fusion on crystallization behavior by using various thermal and molecular input parameters preprocessed by principal component analysis. Bhardwaj et al.[22] were the first to use the RF model to predict the crystallinity of organic molecules with an accuracy rate of about 70%. The training set includes two-dimensional and three-dimensional molecular descriptors and experimental crystallization results. Further development should include crystallization conditions and improvement of the ability to remove uncertainty from training data sets to enhance prediction capabilities. This method can be promoted on salt and cocrystal systems to help understand the crystallization tendency of multicomponent systems.

Wicker et al.[23] used Python and the RDKit cheminformatics toolkit to build models on the basis of chemical descriptors and unsupervised machine learning methods. They predicted whether the molecule will crystallize without considering crystal growth mechanism or conditions. This method mainly focused on the properties and interactions of individual molecules. By comparing SVM and RF models, they found that the SVM model had the highest prediction accuracy (90.3%). This prediction can be used to find synthetic modifications enhancing the crystallization tendency or to find materials with large surface area and poor crystallization. They created and optimized a new molecular descriptor which captures the conformational flexibility of a molecule based on its 2D chemical connectivity in 2016 and established an SVM prediction model. The descriptor also has the potential to solve other chemical problems where flexibility is a key factor, such as prediction of polymorphism. On the basis of their work, Pillong et al.[24] established an RF model to evaluate the solubility and crystallization tendency of 319 small molecules in 18 different solvents. This model can guide the selection of suitable crystallization solvent and effectively reduce the workload to a third of the initial plan while ensuring the crystallization success rate exceeds 92%. Their advantage is having established a unified crystallization database for machine learning.

One of the challenges in predicting crystallinity is that it requires consistent conditions in terms of concentration, evaporation rate, temperature, and pressure in crystallization experiments. Combining data from inconsistent experimental methods will introduce a lot of noise into the data and mislead machine learning methods. Researchers need to obtain enough comprehensive data sets and relevant crystallinity data to establish a durable crystallinity prediction model. It is of great value to archive relevant data in accessible electronic database format for further data processing.

2.3.2. Other Property Prediction. The physical properties of the drug can affect drug safety, stability, and effectiveness. Applying machine learning methods is another way to predict crystal property, which can help people design and control solid states.

Bryant et al.[25] used CSD Python API to develop several topology-based crystal structure descriptors to predict crystal plasticity and compressibility. Salahinejad et al.[26] used the Gaussian software and ANN method to predict sublimation enthalpy, lattice energy, and crystal melting point of small-molecule organic compounds with diverse structures. They used Bayesian regularized artificial neural networks to select the most relevant molecular descriptors. They developed a multiple linear regression model (MLR) for comparison. The average melting point prediction error of ANN was 5 K lower than that of the linear model, showing the former was more accurate than the latter. Velásco-Mejí[27] combined ANN and the genetic algorithm to model the crystallization process by considering temperature, water content, concentration, solvent addition time, pH value, and stirring speed as input parameters. They established a neural network model for predicting crystal density. By optimizing the experimental conditions, crystal density value increased from 0.61g·cm^{-3} to 0.737g·cm^{-3}, indicating that the physical and crystallographic properties substantially improved.

Perlovich et al.[28] established a database of melting temperatures and developed an algorithm for predicting the melting point of cocrystals to guide cocrystal design. However, these algorithms require high-cost calculations. Krishna et al.[29] carried out quantitative structure−activity relationship (QSAR) analysis using ANN to analyze the dependence of API's melting point on the properties of CCFs. Unlike the previous work, this model uses molecular weight, functional group type, melting temperature, etc. as input information and does not require presynthesis of cocrystals to obtain measurement data. Fathollahi et al.[30] established a QSAR model by ANN to predict the density of high-energy cocrystals. They built an MLR model using the same molecular descriptors for comparison. The results demonstrated that the ANN model can more accurately simulate the relationship between structure descriptors and cocrystal density.

3. AI AND COCRYSTAL

3.1. CCF Screening.

Traditional cocrystal screening is expensive in terms of time, energy, and laboratory resources. Using machine learning can save time and resources, especially in the early stage of CCF selection. It is valuable to reduce the list of CCFs to the most likely ones.

Researchers can adopt CSD to screen cocrystal formers (CCFs). By using the large amount of crystal structure data and selection of supramolecular synthons, CCFs with suitable conformation can be selected. Some of the methods proposed for selecting knowledge-based CCFs are as follows. (1) The HBP algorithm mentioned above, which is initially used to predict possible polymorphs, can determine cocrystal formation by evaluating the possibility of homogeneous and heterogeneous interactions. (2) The molecular complementarity method, which was proposed by Fábián, evaluates the possibility of cocrystal formation according to the analysis of calculated descriptors of molecules. The method is based on the idea that cocrystallization molecules tend to have similar molecular properties and that some properties are more strongly correlated than others. It has now been implemented as a tool in a development version of Mercury for further testing and validation. The future direction of virtual cocrystal screening is to carry out multistage and automatic CCF selection workflow.

Galek et al.[31] used CSD to predict potential lamotrigine CCFs and automatically selected the best coformers. Wicker et al.[32] utilized the SVM algorithm and the RDKit cheminformatics toolkit to calculate descriptors of coformer molecules and establish models that classify whether they can form cocrystals with specific APIs. The disadvantage is that it needs a lot of experiments to generate the initial training set with successful and unsuccessful results.

3.2. Cocrystal Composition Prediction.

The quantitative prediction of cocrystal composition can improve the online controllability of the production process and verify the quality of the final products. Barmpalexis et al.[33] quantitatively analyzed cocrystal sol-based mixtures by applying ANNs and partial least-squares (PLS) regression to spectral data modeling. They examined the influence of structure (number of hidden units) and training (number of iteration cycles) parameters, spectral range, and data preprocessing on ANN's fitting performance. They also performed PCA to reduce the dimension of input space and accelerate neural network training. Results show that ANN performs better than PLS.

3.3. Cocrystal Formation Prediction.

Some researchers use classification and regression algorithms to predict the cocrystal formation process to accelerate cocrystal screening and gain high-quality cocrystals.

Multivariate adaptive regression splines (MARSplines) is an effective nonlinear method to solve various quantitative structure−property/activity problems. It is an effective alternative to the ANN algorithm. Przybyłek et al.[34] established a dicarboxylic acid cocrystal screening model on the basis of the MARSplines algorithm using 1D and 2D molecular descriptors. The classification success rate of the cocrystal was 91%. The advantage is that the descriptors can be calculated in a few seconds using free software, and professional knowledge is not necessary. The limitation is that it cannot distinguish between typical cocrystals and salts. The method can be used as a preliminary screening tool for excluding the possible formation of immiscible solid states. Devogelaer et al.[35] introduced two neural network models that accept a pair of molecules as input and classify whether they can form a cocrystal. Two models differed in their input molecular representations and initial preprocessing steps. They used the link-prediction method to generate the invalid cocrystal set and finally discovered a new drug−drug cocrystal.

Chabalenge et al.[36] adopted a decision tree algorithm to study the influencing factors of cocrystallization processes. By using the open-sourced machine learning software WEKA, they examined the effects of different experimental conditions on the cocrystal conversion rate. This model can obtain high-quality cocrystals by selecting the correct operating conditions and shorten the development time of cocrystals. Wang et al.[37] developed a machine learning model trained on the Cambridge Structural Database. Taking 2D structures as input, the probability of cocrystal formation is returned for two given molecules. All the cocrystal records in the CSD were used as positive samples, while negative samples were constructed by randomly combining different molecules into chemical pairs. They also studied the impact of training set size on model performance, which improved by increasing data size. The ROC-AUC of the consensus model improved from 0.754 to 0.852 by increasing the size of the training set. They predicted two cocrystals successfully by their model.

4. CONCLUSION

With the rapid development of AI, it can be applied to many aspects of polymorphism and cocrystal research, such as polymorphism prediction, crystal structure analysis, crystal property prediction, CCF screening, cocrystal composition prediction, and cocrystal formation prediction. It is a new research direction that can save experimental costs and provide a theoretical guidance for future studies. Nevertheless, there is still plenty of room for improvement.

First, developing robust descriptors for crystalline solids is challenging. The selection of descriptors should be meaningful and universal, and the relationship with output should be simpler. New descriptors should be developed to encode more complex material data. Algorithms for generating descriptors that can be used by experts in nonrelated fields are still lacking.

Second, there is a need to improve the quantity and quality of machine learning data. A large database including both positive and negative results is crucial for balanced model training. It is also important that the data come from uniform and comparable experiments. At present, databases are independent and not unified in data format, which limits the usage of machine learning. In addition, when using the CSD and machine learning techniques in big-data analysis, it is likely that the database errors bury so deeply to become impossible to detect.

Third, new techniques have appeared to improve algorithm efficiency such as parallel computing and cloud computing. No single algorithm can fit for all applications. Comparison among different algorithms is necessary to choose the best one.

In short, researchers should integrate the big data generated in computational chemistry and crystallization experiments and explore the hidden rules in complex data to promote the development and application of AI in pharmaceutical crystals.

■ AUTHOR INFORMATION

Corresponding Authors

Li Zhang − *Beijing City Key Laboratory of Polymorphic Drugs, Center of Pharmaceutical Polymorphs, Institute of Materia Medica, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing 100050, P.R. China;* orcid.org/0000-0003-3115-8196; Phone: +86 10 63165310; Email: zhangl@imm.ac.cn

Yang Lu − *Beijing City Key Laboratory of Polymorphic Drugs, Center of Pharmaceutical Polymorphs, Institute of Materia Medica, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing 100050, P.R. China;* orcid.org/0000-0002-2274-5703; Phone: +86 10 63165212; Email: luy@imm.ac.cn

Authors

Tianyu Heng − *Beijing City Key Laboratory of Polymorphic Drugs, Center of Pharmaceutical Polymorphs, Institute of Materia Medica, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing 100050, P.R. China;* orcid.org/0000-0002-0417-3932

Dezhi Yang − *Beijing City Key Laboratory of Polymorphic Drugs, Center of Pharmaceutical Polymorphs, Institute of Materia Medica, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing 100050, P.R. China;* orcid.org/0000-0002-3159-4126

Ruonan Wang − *Beijing City Key Laboratory of Polymorphic Drugs, Center of Pharmaceutical Polymorphs, Institute of Materia Medica, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing 100050, P.R. China;* orcid.org/0000-0003-3885-221X

Guanhua Du − *Beijing City Key Laboratory of Drug Target and Screening Research, National Center for Pharmaceutical Screening, Institute of Materia Medica, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing 100050, P.R. China*

Complete contact information is available at: https://pubs.acs.org/10.1021/acsomega.1c01330

Notes

The authors declare no competing financial interest.

Biographies

Dr. Li Zhang received her Ph.D. in 2008 from the Institute of Materia Medica, Chinese Academy of Medical Sciences and Peking Union Medical College, China. Presently, she is working as an associate professor in the Institute of Materia Medica, Chinese Academy of Medical Sciences and Peking Union Medical College, China. Her research interests focus on drug polymorphism research, new technology, and method research of pharmaceutical analysis and development of certified reference materials.

Prof. Yang Lu is head of Beijing City Key Laboratory of Polymorphic Drugs in Institute of Materia Medica, Chinese Academy of Medical Sciences and Peking Union Medical College. Her research topics include drug quality analysis, drug polymorphism research, pharmaceutical reference materials research, new technology and method research of pharmaceutical analysis, and three-dimensional structure determination of drug and biological receptor target molecules. She is a member of the National Pharmacopoeia Committee. She also serves as vice president of the Chinese Crystallography Society and director of Pharmaceutical Crystallography Professional Committee.

Tianyu Heng received her B.Sc in 2019 from China Pharmaceutical University, China. Presently, she is working as a postgraduate in the Institute of Materia Medica, Chinese Academy of Medical Sciences and Peking Union Medical College, China. Her research interest is based on drug polymorphism and cocrystal research as well as new method research of pharmaceutical analysis.

Dr. Dezhi Yang received his Ph.D. in 2015 from the Institute of Materia Medica, Chinese Academy of Medical Sciences and Peking Union Medical College, China. Presently, he is working as an associate professor in the Institute of Materia Medica, Chinese Academy of Medical Sciences and Peking Union Medical College, China. He has expertise in drug polymorphism research, cocrystal design, and synthesis of drugs and lead compounds.

Ruonan Wang received her B.Sc in 2018 from Shenyang Pharmaceutical University, China. Presently, she is working as a postgraduate in the Institute of Materia Medica, Chinese Academy of Medical Sciences and Peking Union Medical College, China. Her research interests include drug polymorphism and cocrystal research of drugs and lead compounds.

Dr. Guanhua Du received his Ph.D. in 1995 from the Institute of Materia Medica, Chinese Academy of Medical Sciences and Peking Union Medical College, China. Presently, he is head of Beijing City Key Laboratory of Drug Target and Screening Research, National Center for Pharmaceutical Screening, Institute of Materia Medica, Chinese Academy of Medical Sciences and Peking Union Medical College, China. His research interests are based on drug discovery, high-throughput drug screening, neuropharmacology, and cardio cerebrovascular pharmacology research.

■ ACKNOWLEDGMENTS

This work was financially supported by the National Key R&D Program of China (grant number 2016YFC1000900), National Science and Technology Major Project of China (grant numbers 2017ZX09101001003, 2018ZX09711001-003-022, and 2018ZX09711001-010), and CAMS Innovation Fund for Medical Sciences (grant number 2020-I2M-1-003).

■ REFERENCES

(1) Brittain, H. G. Polymorphism and solvatomorphism 2010. *J. Pharm. Sci.* **2012**, *101*, 464−484.

(2) Saravanan, R.; Sujatha, P. A state of art techniques on machine learning algorithms: a perspective of supervised learning approaches in data classification. *In 2018 Second International Conference on Intelligent Computing and Control Systems (ICICCS).* **2018**, 945−949.

(3) Breiman, L. Random forests. *Machine learning.* **2001**, *45* (1), 5−32.

(4) Mitchell, J. B. Machine learning methods in chemoinformatics. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2014**, *4* (5), 468−481.

(5) Cortes, C.; Vapnik, V. Support-vector networks. *Machine learning* **1995**, *20* (3), 273−297.

(6) Speelman, D. Logistic regression. *Corpus methods for semantics: Quantitative studies in polysemy and synonymy.* **2014**, *43*, 487−533.

(7) Li, D. X.; Luan, H. S.; Guo, B. S.; Xi, Q.; Wang, J. X.; Wang, H. Applied Computational Pharmaceutics: an Efficient Pharmaceutical Tool. *Chin. J. Pharm.* **2017**, *48* (12), 1673−1684.

(8) Nauha, E.; Bernstein, J. Predicting" crystal forms of pharmaceuticals using hydrogen bond propensities: Two test cases. *Cryst. Growth Des.* **2014**, *14* (9), 4364−4370.

(9) Nauha, E.; Bernstein, J. Predicting" Polymorphs of Pharmaceuticals Using Hydrogen Bond Propensities: Probenecid and Its Two Single-Crystal-to-Single-Crystal Phase Transitions. *J. Pharm. Sci.* **2015**, *104* (6), 2056−2061.

(10) Feeder, N.; Pidcock, E.; Reilly, A. M.; Sadiq, G.; Doherty, C. L.; Back, K. R.; Meenan, P.; Docherty, R. The integration of solid-form informatics into solid-form selection. *J. Pharm. Pharmacol.* **2015**, *67* (6), 857−868.

(11) Johnston, A.; Johnston, B. F.; Kennedy, A. R.; Florence, A. J. Targeted crystallisation of novel carbamazepine solvates based on a retrospective Random Forest classification. *CrystEngComm* **2008**, *10* (1), 23−25.

(12) Takieddin, K.; Khimyak, Y. Z.; Fábián, L. Prediction of Hydrate and Solvate Formation Using Statistical Models. *Cryst. Growth Des.* **2016**, *16*, 70−81.

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

(14) Oliynyk, A. O.; Adutwum, L. A.; Harynuk, J. J.; Mar, A. Classifying Crystal Structures of Binary Compounds AB through Cluster Resolution Feature Selection and Support Vector Machine Analysis. *Chem. Mater.* **2016**, *28*, 6672−6681.

(15) Mcdonagh, D.; Skylaris, C. K.; Day, G. M. Machine-Learned Fragment-Based Energies for Crystal Structure Prediction. *J. Chem. Theory Comput.* **2019**, *15*, 2743−2758.

(16) Collins, A.; Wilson, C. C.; Gilmore, C. J. Comparing entire crystal structures using cluster analysis and fingerprint plots. *CrystEngComm* **2010**, *12*, 801−809.

(17) 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* **2018**, *20* (28), 3947−3950.

(18) Gavezzotti, A.; Presti, L. L. Building Blocks of Crystal Engineering: A Large-Database Study of the Intermolecular Approach between C-H Donor Groups and O, N, Cl, or F Acceptors in Organic Crystals. *Cryst. Growth Des.* **2016**, *16*, 2952−2962.

(19) Yang, J.; Li, N.; Li, S. The interplay among molecular structures, crystal symmetries and lattice energy landscapes revealed using unsupervised machine learning: a closer look at pyrrole azaphenacene. *CrystEngComm* **2019**, *21*, 6173−6185.

(20) Phillips, C. L.; Voth, G. A. Discovering crystals using shape matching and machine learning. *Soft Matter* **2013**, *9*, 8552−8568.

(21) Van Eerdenbrugh, B.; Baird, J. A.; Taylor, L. S. Crystallization tendency of active pharmaceutical ingredients following rapid solvent evaporation−classification and comparison with crystallization tendency from undercooled melts. *J. Pharm. Sci.* **2010**, *99* (9), 3826−3838.

(22) Bhardwaj, R. M.; Johnston, A.; Johnston, B. F.; Florence, A. J. A random forest model for predicting the crystallisability of organic molecules. *CrystEngComm* **2015**, *17* (23), 4272−4275.

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

(24) Pillong, M.; Marx, C.; Piechon, P.; Wicker, J. G.; Cooper, R. I.; Wagner, T. A publicly available crystallisation data set and its application in machine learning. *CrystEngComm* **2017**, *19*, 3737−3745.

(25) Bryant, M. J.; Maloney, A. G. P.; Sykes, R. A. Predicting mechanical properties of crystalline materials through topological analysis. *CrystEngComm* **2018**, *20*, 2698−2704.

(26) Salahinejad, M.; Le, T. C.; Winkler, D. A. Capturing the crystal: prediction of enthalpy of sublimation, crystal lattice energy, and melting points of organic compounds. *J. Chem. Inf. Model.* **2013**, *53* (1), 223−229.

(27) Velásco-Mejía, A.; Vallejo-Becerra, V.; Chávez-Ramírez, A. U.; Torres-González, J.; Reyes-Vidal, Y.; Castañeda-Zaldivar, F. Modeling and optimization of a pharmaceutical crystallization process by using neural networks and genetic algorithms. *Powder Technol.* **2016**, *292*, 122−128.

(28) Perlovich, G. L. Thermodynamic characteristics of cocrystal formation and melting points for rational design of pharmaceutical two-component systems. *CrystEngComm* **2015**, *17*, 7019−7028.

(29) Rama Krishna, G.; Ukrainczyk, M.; Zeglinski, J.; Rasmuson, Å. C. Prediction of Solid State Properties of co-crystals Using Artificial Neural Network Modeling. *Cryst. Growth Des.* **2018**, *18*, 133−144.

(30) Fathollahi, M.; Sajady, H. Prediction of density of energetic cocrystals based on QSPR modeling using artificial neural network. *Struct. Chem.* **2018**, *29* (4), 1119−1128.

(31) Galek, P. T.; Pidcock, E.; Wood, P. A.; Bruno, I. J.; Groom, C. R. One in half a million: a solid form informatics study of a pharmaceutical crystal structure. *CrystEngComm* **2012**, *14*, 2391−2403.

(32) Wicker, J. G.; Crowley, L. M.; Robshaw, O.; Little, E. J.; Stokes, S. P.; Cooper, R. I.; Lawrence, S. E. Will they co-crystallize? *CrystEngComm* **2017**, *19*, 5336−5340.

(33) Barmpalexis, P.; Karagianni, A.; Nikolakakis, I.; Kachrimanis, K. Artificial neural networks (ANNs) and partial least squares (PLS) regression in the quantitative analysis of cocrystal formulations by Raman and ATR-FTIR spectroscopy. *J. Pharm. Biomed. Anal.* **2018**, *158*, 214−224.

(34) Przybyłek, M.; Jeliński, T.; Słabuszewska, J.; Ziółkowska, D.; Mroczyńska, K.; Cysewski, P. Application of Multivariate Adaptive Regression Splines (MARSplines) Methodology for Screening of Dicarboxylic Acid Cocrystal Using 1D and 2D Molecular Descriptors. *Cryst. Growth Des.* **2019**, *19*, 3876−3887.

(35) Devogelaer, J.-J.; Meekes, H.; Tinnemans, P.; Vlieg, E.; Gelder, R. Co-crystal Prediction by Artificial Neural Networks. *Angew. Chem.* **2020**, *132* (48), 21895−21902.

(36) Chabalenge, B.; Korde, S.; Kelly, A. L.; Neagu, D.; Paradkar, A. Understanding Matrix Assisted Continuous Cocrystallisation using Data Mining approach in Quality by Design (QbD). *Cryst. Growth Des.* **2020**, *20* (7), 4540−4549.

(37) Wang, D.; Yang, Z.; Zhu, B.; Mei, X.; Luo, X. Machine-Learning-Guided Cocrystal Prediction Based on Large Data Base. *Cryst. Growth Des.* **2020**, *20* (10), 6610−6621.