

Artificial Intelligence Assisted Pharmaceutical Crystallization

Zuoxuan Zhu,^{||} Yuan Zhang,^{||} Zhixuan Wang, Weiwei Tang,^{*} Jingkang Wang, and Junbo Gong^{*}Cite This: *Cryst. Growth Des.* 2024, 24, 4245–4270

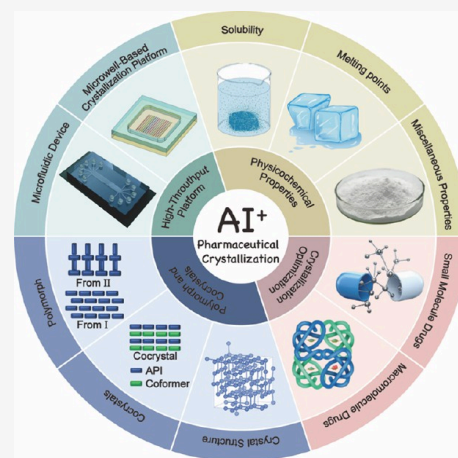
Read Online

ACCESS |

Metrics & More

Article Recommendations

ABSTRACT: The ever-increasing demand for novel drug development has spurred the adaptation of conventional research methods in the era of artificial intelligence. Pharmaceutical crystallization, as an essential part of drug development, has become a thrilling research frontier that reduces screening labor via integrating automated high-throughput platforms with *in situ* monitoring and data-driven algorithms (e.g., machine learning) to predict physicochemical properties and various solid-state forms. In this review, we started with a primer to introduce the machine learning algorithms that are widely used in pharmaceutical crystallization. Then, we systematically summarized recent advancements on high-throughput platforms to acquire huge amounts of data sets, prediction of physicochemical properties based on abundant experimental data, optimization and monitoring of pharmaceutical crystallization process to screen crystallization conditions, and prediction of polymorphs and cocrystals. Finally, we discussed the challenges and opportunities in an endeavor to develop a fully automated pharmaceutical crystallization screening paradigm for ultimately realizing a self-driving screening laboratory. This review highlights the frontier of artificial intelligence in pharmaceutical crystallization and offers a guideline for beginners to not only understand the basic principles of machine learning algorithms but also learn how to utilize machine learning to accelerate pharmaceutical crystallization development.



■ INTRODUCTION

The demand for pharmaceutical development is continually increasing,¹ as it prioritizes the well-being of all humankind. Crystallization of active pharmaceutical ingredients (API) plays a central role in the development of various pharmaceutical solid forms such as polymorph, cocrystals/salt, and solvate,² which significantly impacts product solubility, bioavailability, and stability.^{3,4} Approximately 80% of APIs are manufactured through the unit operation of crystallization.⁵ Obtaining the desired pharmaceutical crystals (i.e., specific crystal structure and morphology) requires significant investments in screening solid-state forms, identifying the most suitable crystalline form for further development, and exploring crystallization conditions for reproducible production of exact polymorph with superior crystal morphology.⁶ Moreover, prior to human trials, adequate information on drug candidates is required, such as potential crystal forms, physical properties, and manufacturing capabilities. This step is crucial in preventing issues, for example, failure to scale up in later stages of development. However, the current approach to fulfill these tasks still relies on trial and error, the success of which depends upon the well-trained skills of human scientists, with the guidance of crystallization mechanisms and physicochemical rules.⁷ This paradigm rarely evolved in past decades, creating a bottleneck that hinders the advancement of crystallization screening to meet the demands of drug discovery

companies and pharmaceutical manufacturers. The remarkable progress of pharmaceutical crystallization is to transform the current paradigm into a new one via the assistance of artificial intelligence (AI).

The trial-and-error approach has been widely employed for screening and optimizing crystallization conditions, as well as for developing cocrystals and selecting conformers.⁶ High-throughput experiments excel in simultaneously performing multiple tasks, thereby greatly enhancing the efficiency of a single batch of experiments.⁸ Nonetheless, a challenge arises in analyzing the abundant characteristic data, which often lags behind the rate at which data is acquired. On the contrary, AI encompasses a range of technologies that enable computational systems to perform functions conventionally necessitating human cognitive capabilities, including the acquisition of knowledge, logical deduction, sensory interpretation, and the act of making informed judgments.⁹ Its importance is growing in the realm of big data analytics, as it enables the rapid and efficient

Received: November 27, 2023

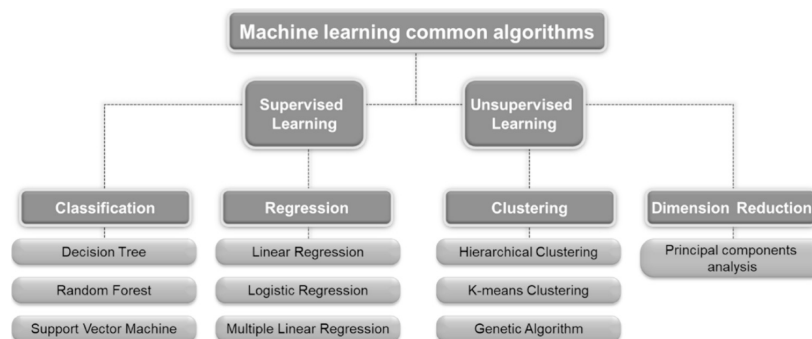
Revised: April 26, 2024

Accepted: April 26, 2024

Published: May 3, 2024



Scheme 1. Common Algorithms of Machine Learning



processing and analysis of vast amounts of data, surpassing the capabilities of human analysts. Machine learning (ML) stands as an important subdivision of AI, exerting a substantial influence on nearly every aspect of the world in recent decades. Its application spans various research fields, such as computer vision,¹⁰ natural language processing,⁹ health care,¹¹ and beyond, driving paradigm shifts in these fields. Deep learning (DL) algorithms, in particular, demonstrate exceptional capabilities by learning from data sets and extracting insights and patterns from big data that might otherwise be arduous or unattainable compared to traditional algorithmic models.

In recent times, ML has emerged as a valuable instrument within the domain of crystallization, offering fresh perspectives and showcasing its capacity to confront enduring challenges encountered in the advancement of the crystallization process. These encompass data-centric monitoring, modeling, and control, as well as the prediction of physicochemical properties and crystallization behavior.¹² By integrating ML into pharmaceutical crystallization, a promising new paradigm in drug development has emerged, showing tremendous potential in accelerating drug discovery. These applications primarily revolve around result prediction and process monitoring. ML algorithms enable the prediction of various aspects of crystalline materials, such as physicochemical properties,¹³ crystal structures,¹⁴ and cocrystal conformers.¹⁵ This prediction capability reduces the burdensome workload involved in labor-intensive solid-form screening experiments. Besides, ML algorithms can be utilized in process analytical techniques for *in situ* monitoring of crystallization process,¹⁶ identification and classification of crystals,¹⁷ collection of multidimensional crystal morphology data,¹⁸ and optimization and screening of crystallization conditions.¹⁹ Apart from process monitoring and result prediction, a substantial amount of data from high-throughput experiments is necessary to train artificial neural networks. This necessitates the use of microfluidic or microwell platforms with *in situ* monitoring technologies.

Recently, Xiouras et al. reviewed applications of AI and ML algorithms to industrial crystallization.²⁰ A systematic review on data-driven algorithms and high-throughput platforms which stand for the combination of software (algorithms) and hardware (platforms), nonetheless, is still lacking for screening and optimization conditions of AI-assisted pharmaceutical crystallization. Therefore, this progress report studies the recent advancements in this field and primarily centers around the early stage of drug development. In particular, our focus lies on the following areas, where significant progress has been achieved in recent years: (a) the construction of high-throughput crystallization platforms to acquire abundant data for ML

(Section 2); (b) the prediction of physicochemical properties of pharmaceutical crystals (Section 3); (c) the screening and optimization of drug crystallization conditions with the aid of convolutional neural network based on the image segmentation algorithm (Section 4); (d) the prediction of polymorph and cocrystals with the aid of AI algorithms (Section 5).

1.1. Overview of Machine Learning Algorithms. ML is one of the branches of AI and is at its core. ML research aims to allow computers to learn, simulate human learning behavior, build learning capabilities, and finally achieve recognition and judgment. Data is an essential component of ML. Data labels are the key information to characterize the sample. According to the presence or absence of labels, ML can be divided into two categories: supervised learning and unsupervised learning. Common algorithms of ML are presented in Scheme 1.

1.1.1. Supervised Learning. Supervised learning can be divided into regression models and classification models depending on whether the labels are continuous or discrete. Supervised learning is the more mature and powerful of the ML algorithms and is used for most ML studies in drug development for example mapping chemical components to properties of interest.²¹

Regression Model. The regression model learns the mapping function between features and continuous labels, and the characteristics of the function determine the type of regression model. If there is a linear relationship between features and labels, which satisfies a multiple primary function, the model is a linear regression. For polynomial linear regression, the higher complexity often introduces overfitting problems, which can be avoided by adding regular terms, and linear regression that leads to regularization operations is called regular linear regression.

Classification Model. Classification models are designed to classify data into discrete categories. Different from the regression model, the classification task can be accomplished by learning the mapping relationship between samples and discrete labels, as well as from the perspective of data generation. If the mapping relationship can be directly learned by the mapping function, the model is called a nonprobabilistic discriminant model, which is widely used in pharmaceutical crystallization.

K-Nearest Neighbor Algorithm. The K-nearest Neighbor Algorithm²² classifies a sample into a certain category if the majority of its k-most similar samples (i.e., nearest in feature space) in the feature space also belong to the same category. This method relies exclusively on the classification decisions made by its nearest one or a few samples to determine the class to which a sample should be assigned.

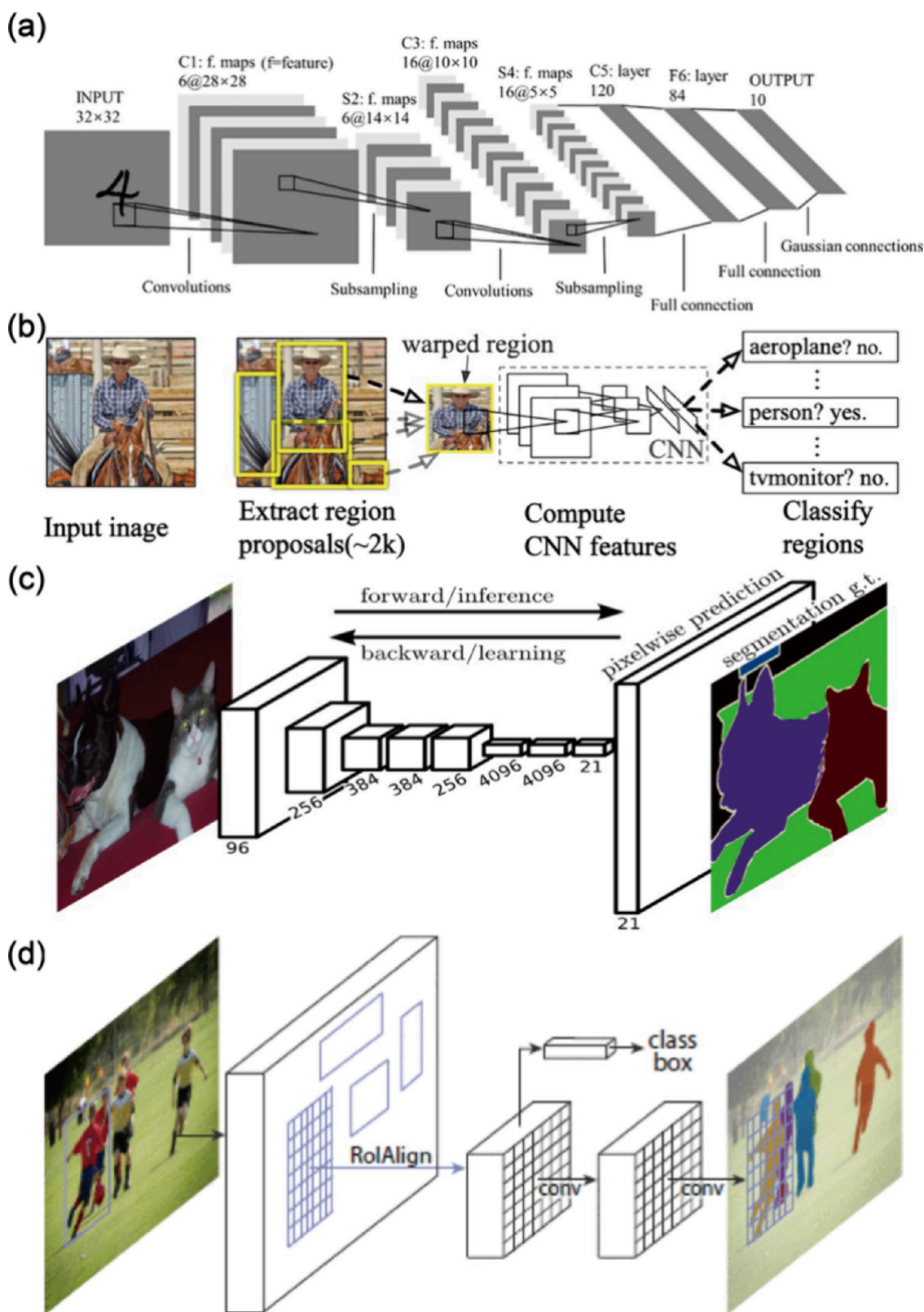


Figure 1. Structure of deep learning algorithms based on convolutional neural networks. Reproduced with permission from ref 42. Copyright 2022 Elsevier. (a) LeNet-5 model for image classification;⁴³ (b) R-CNN model for object detection;⁴⁴ (c) FCN model for semantic segmentation;⁴⁵ (d) Mask R-CNN model for instance segmentation.⁴⁶

Support Vector Machine Algorithm. The Support Vector Machine (SVM) algorithm²³ operates by transforming input

vectors into a high-dimensional feature space, establishing a linear decision boundary that enhances the network's general-

ization capacity. Only a small subset of training data, known as support vectors, is crucial to create these optimal hyperplanes, as they define the margin between classes. For this reason, SVM is featured as its ability to control model capacity and effortlessly adjust the decision boundary. In addition, the SVM algorithm has significant advantages in the prediction of physicochemical properties of pharmaceutical crystals.²⁴

Random Forests Algorithm. Random forests²⁵ is an ensemble learning algorithm comprised of multiple tree predictors, where each tree relies on the values of a random vector sampled independently and from the same distribution for all trees within the forest. This strategic use of randomness enhances their accuracy as classifiers and regressors. Random Forests exhibit robustness against outliers and noise, and they demonstrate resilience against overfitting. Additionally, they possess the capability to assess the importance of input variables, making them a valuable tool in various ML applications. It has been applied to predict the crystallization propensity of API²⁶ and protein.²⁷

1.1.2. Unsupervised Learning. Unsupervised learning²⁸ does not rely on the labels of the sample, and it can learn the internal structure or deep features of the data. Unsupervised learning is commonly applied for broad data analysis and classification purposes, or to discover hidden patterns within extensive data sets. Clustering and data dimensionality reduction are two typical unsupervised learning methods. Particularly, the application of dimensionality reduction techniques is quite necessary for feature extraction of raw data.²⁹

Clustering Method. Segmented clustering is often used when the number of categories to be divided is known or can be preset, otherwise hierarchical clustering can be considered to divide the sample into structural categories with hierarchical relationships.

Dimension Reduction Method. For data dimensionality reduction, a more common approach is to utilize Principal Component Analysis (PCA),³⁰ which can significantly reduce the dimensions of a data set while retaining as much of the data's variation as achievable. By capturing the essential patterns in the data, PCA simplifies complex data sets, making them more manageable for further analysis and interpretation.

Different prediction tasks correspond to different models. The prediction of physicochemical properties may be a regression problem (e.g., predicting solubility³¹) or a classification problem (e.g., crystal structure prediction³²). For example, regression models such as linear regression, and logistic regression can be used to predict continuous physicochemical properties, whereas classification models such as random forests, decision trees, or convolutional neural networks can be used to predict classification results. In addition, the interpretability of models is important in drug discovery. Some models such as decision trees and linear regression have good interpretability, whereas deep learning models may be more difficult to interpret.²⁴ In conclusion, the choice of algorithm is the result of multiple factors.

1.1.3. Artificial Neural Networks. The concept of artificial neural networks (ANN) originated from the simulation and study of the working principles of biological neural systems. The development of ANN started with supervised learning and gradually expanded to unsupervised learning. The fundamental structure of ANN consists of neurons organized into layers, along with a learning algorithm (e.g., backpropagation) that adjusts the weights during training. Neurons receive multiple inputs and generate a single output. DL algorithms leverage neural networks with numerous layers to learn and model

intricate patterns in data. Here we introduce three common DL algorithms.

Feedforward Neural Networks. Feedforward neural networks (FFNN)³³ are the predominant type of ANN. These networks are composed of multiple interconnected layers, where each neuron in a particular layer receives input from the neurons in the preceding layer. The training of FFNNs commonly involves the backpropagation algorithm. In the field of crystallization, FFNNs are frequently employed for various applications, including the prediction of optimal crystallization conditions and the optimization of parameters in industrial crystallization processes.³⁴

Recurrent Neural Networks. Unlike FFNNs, RNNs possess feedback connections, Recurrent neural networks (RNN)³⁵ enable them to effectively process sequential data, including time series data and crystal growth data. These capabilities make RNNs suitable for a multitude of applications for example prediction of crystal growth rate.³⁶

Convolutional Neural Networks. Convolutional neural networks (CNN)³⁷ are widely recognized for their effectiveness in image processing, making them a valuable tool for accelerating crystal identification in the optimization of crystallization conditions. Furthermore, CNNs have proven useful in other areas of crystallization research such as the analysis of crystal structures³⁸ and the classification of crystallographic images.³⁹

1.2. Image Segmentation Models. Traditional image processing methods typically convert images to an 8-bit format and segment crystals based on grayscale features,⁴⁰ such as the open-access software ImageJ.⁴¹ However, it is sensitive to background noise and has poor robustness. Other traditional image processing methods their characteristics and applications have been reviewed in another review.⁴² In brief, traditional methods have limitations when it comes to the automated handling of images and distinguishing crystals with varying morphologies. Moreover, these methods necessitate the use of high-quality images, requiring complex preprocessing, including adjusting contrast, manually separating polycrystals, etc.⁴¹ Based on DL image segmentation methods, crystals can be classified or identified using the pre-trained model from simple preprocessed images (e.g., resizing images). By leveraging computer vision models, thousands of outputs can be analyzed within minutes from high-throughput screen experiments.

Deep Learning Based Image Segmentation Models. The classification of objects in images and the measurement of their morphology can be divided into four stages of image classification, ranging from simple to complex. We illustrated each stage with a representative model.

Image Classification. Image classification refers to the process of labeling images with their corresponding categories. As illustrated in Figure 1 (a), the algorithms for image classification typically follow a fundamental structure: Convolutional layer → Batch normalization layer → Activation function → pooling layer.⁴³ The convolutional layer is responsible for extracting features from the image. After the data is normalized, the model approximates the nonlinear function by applying the ReLu function to the input. The pooling layer is then employed to reduce the spatial dimensions. An example of a classical convolutional network is LeNet-5, which is often used for handwriting character recognition.⁴³

Object Detection. Image classification involves the assignment of a single label to the entire image, whereas object detection goes a step further by localizing and classifying

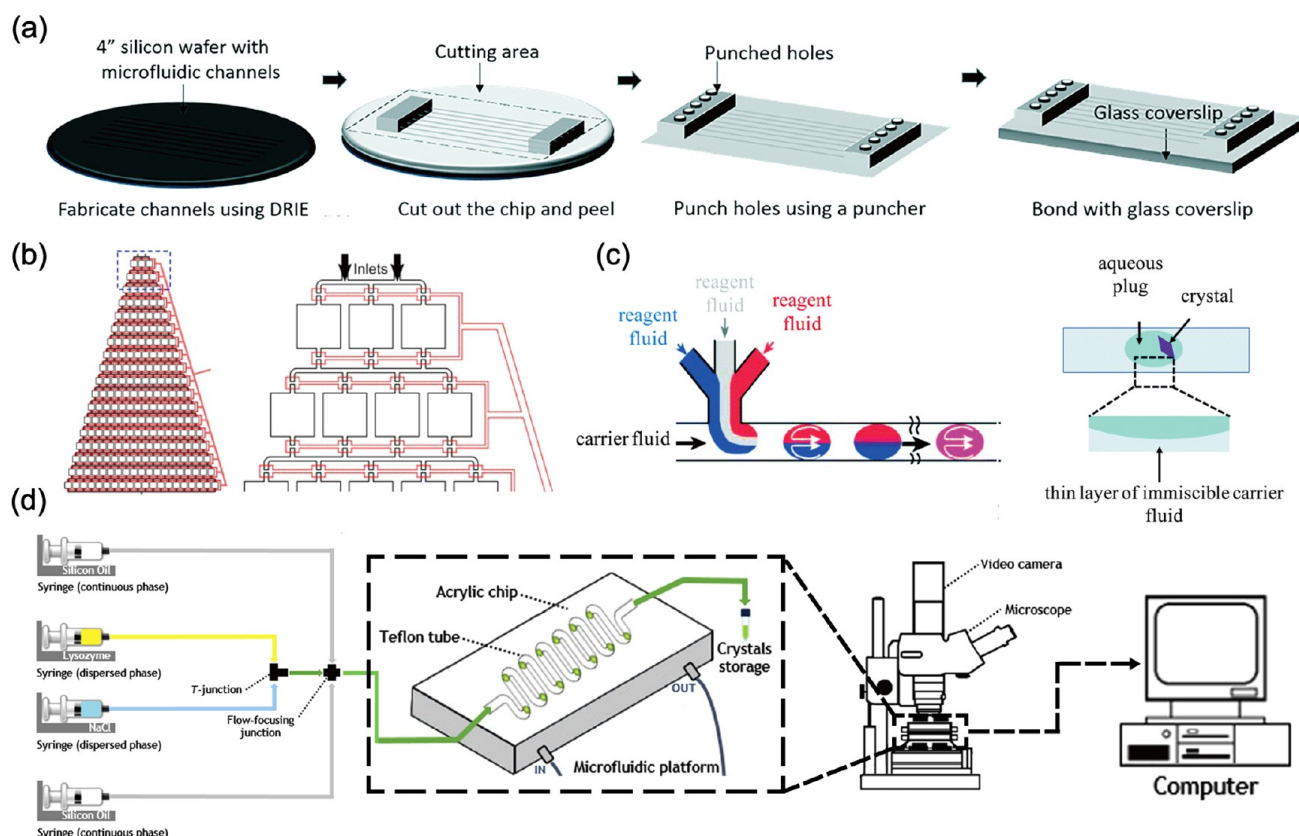


Figure 2. Fabrication of a microfluidic chip, channel structure, and experiment procedures. (a) Fabrication procedure of a microfluidic device. Reproduced with permission from ref 60. Copyright 2017 Royal Society of Chemistry; (b) channel structure of concentration gradient generator chip. Reproduced with permission from ref 62. Copyright 2017 American Chemical Society; (c) droplet generation in microfluidic devices. Reproduced with permission from ref 63. Copyright 2014 Royal Society of Chemistry; (d) microfluidic-based crystallization screening platform for protein crystallization. Reproduced with permission from ref 64. Copyright 2018 Elsevier.

multiple objects within the image using bounding boxes. This capability enables object detection to handle more complex images. The region-based convolutional neural network (R-CNN) algorithm is the representative algorithm, as shown in Figure 1 (b).⁴⁴ It begins by extracting region proposals from the input image. These proposals are then processed using a large CNN to calculate their features. Finally, the outcome provides information about the likelihood of an object being classified as a specific class within a given region.

Object classification is of utmost importance for the recognition of crystals, which enables the differentiation of crystals with distinct morphologies into separate categories. An example of such an algorithm is CrystalNet, which is a protein crystal image labeling algorithm proposed by Yann and Tang in 2016.⁴⁷ It achieves a classification accuracy of 90.8%, a significant improvement when compared to the traditional models based solely on image grayscale values.

Semantic Segmentation. Semantic segmentation involves the segmentation of an image at the pixel level, where each pixel is classified into a specific class. Generally, it outputs a bounding box that depicts the object edge, enabling to identify and localize objects. The fully convolutional network (FCN), Figure 1 (c), generates predictions for each pixel, thereby preserving the spatial information on the original input image.⁴⁵ Ultimately, the model assigns a respective label to each pixel in the feature map. Object detection enables the identification of crystals with various types or morphologies in an image. However, it is unable to separate single objects belonging to the same classes, having

difficulty in acquiring the morphology feature of individual objects.⁴⁸

Instance Segmentation. Instance segmentation involves the separation of objects into individual instances within the same class. For example, if there are three planes in an image, instance segmentation will split each plane into a separate object, all labeled as “Plane”. Semantic segmentation, on the other hand, labels the entire group of planes as a single “Plane” entity. Mask R-CNN,⁴⁶ which is built upon the foundation of Fast R-CNN, is a classical instance segmentation model that predicts the category of each object by creating a mask that covers each instance. Figure 1 (d) illustrates the process of object detection and mask generation by instance segmentation. Mask R-CNN incorporates an RPN layer, which is responsible for identifying potential object locations. These branches perform tasks such as classification, coordinate regression, and segmentation, enabling the creation of masks for each instance.

The introduction of instance segmentation algorithms has significantly enabled the acquisition of multidimensional information and advanced the capabilities of particle analytical technology tools. Crystal morphology-related information could be obtained quickly and in large quantities, including parameters such as count, surface area, length, and width of crystals. This information facilitates the studies of crystallization kinetics, the crystallization process itself, and the phase transformation process, which will be reviewed in detail in Chapter 4.

2. THE CONSTRUCTION OF A HIGH-THROUGHPUT CRYSTALLIZATION PLATFORM

In the pharmaceutical industry, API crystals often require high-quality attributes such as high chemical purity, unique crystal structure, good crystal shape, appropriate size, and size distribution, which are influenced by various factors, including the supersaturation,⁴⁹ temperature,⁵⁰ solvent composition,⁵¹ diffusion rate,⁵² etc. To obtain the optimum crystallization technology, high-throughput screening experiments are essential to optimize crystallization conditions.^{8,53} Moreover, the effectiveness of ML algorithms heavily depends on having access to large quantities of high-quality data for modeling training. Without these data sets, AI systems might struggle to recognize patterns and make accurate predictions in new situations. Over the past decade, two high-throughput crystallization platforms were developed based on microfluidic devices and well-based experiments. These platforms could provide highly efficient approaches to collecting high-quality data, facilitating the modeling and optimization of crystallization processes. By integrating the advantages of high-throughput platforms with AI, the efficiency of process development of crystallization has been significantly improved with better control of product quality.

2.1. Microfluidic Devices. Microfluidic devices provide a unique advantage over traditional methods in terms of creating a free-convection environment that promotes the formation of high-quality crystals.⁵⁴ This technique has been widely adopted in high-throughput screening.⁵⁵ By virtue of micrograms drug, a diverse array of operational parameters can be systematically investigated, affording an expedient means of assessing various crystallization conditions. Within the pharmaceutical development context, microfluidic platforms were meticulously tailored to facilitate the screening of distinct solid-state forms of APIs,⁵⁶ facilitating the early-stage screening of solid forms in drug development.

In this subsection, we first reviewed the fabrication of microfluidic devices and the associated materials, followed by three types of microfluidic devices that are extensively used in high-throughput crystallization screening. Then, the progress of on-chip *in situ* monitoring technology was summarized to improve the efficiency of the screening of crystallization conditions

Fabrication of Microfluidic Devices. Selecting the optimum material is the fundamental step in microfluidic device fabrication. Since the effect of surface properties at a microscale level on crystallization may be amplified, the materials of the platform are likely to affect the product quality.⁵⁷ To address the issue of solvent incompatibility in the device,⁵⁸ hybrid materials like polydimethylsiloxane (PDMS) and glass were chosen as the replacement. PDMS is commonly used as a substrate to create microchannels and chambers, while glass is used for sealing and providing optical windows. This combination offers great advantages such as reduced swelling, improved solvent compatibility, and enhanced optical transparency.⁵⁹

The common technique used for microfluidic device fabrication is soft lithography, which involves several steps illustrated in Figure 2 (a).⁶⁰ The process begins with the creation of a hard master, typically using silicon as the substrate. PDMS is subsequently poured into the mold, subjected to a heat-curing process, removed, and ultimately affixed to a glass substrate through the application of plasma treatment. Soft lithography offers several benefits, including the ability to produce high-resolution replicas at a lower cost and faster pace compared to

traditional photolithography methods. It also allows for the creation of intricate fluidic paths and pneumatic control lines by stacking multiple layers, enabling the generation of a larger number of channels or chambers to accommodate diverse crystallization conditions. Furthermore, the flexibility and exceptional optical transparency of the chip structure render it a convenient choice for the *in situ* monitoring of the crystallization process.⁶¹ Nevertheless, it is imperative to acknowledge that soft lithography is a relatively cost-intensive procedure, frequently necessitating a cleanroom environment.

Versatile Microfluidic Device. One notable advantage of microfluidic devices in screening crystallization conditions is their ability to effortlessly generate a series of solute concentration combinations. Concentration gradient generator microfluidic devices, in particular, are often used in this context.

In concentration gradient generators, an array of channels is employed to create a sequence of concentration combinations by dividing and subsequently recombining input solutions.⁶⁵ Each channel containing a specific combination of concentrations represents an independent crystallization condition, as shown in Figure 2 (b). This method relies on controlling the diffusive mixing of the flows within the network of microchannels. Through the controlled flow of multiple fluid streams, each containing distinct solute concentrations, in a laminar and parallel configuration, stepwise concentration gradients are established perpendicular to the direction of fluid flow.⁶⁶

This technique for gradient generation facilitates the convenient exploration of an extensive spectrum of solvent compositions and concentrations within a single microfluidic device. By varying the input solutions and their flow rates, different gradients can be generated, enabling efficient screening of crystallization parameters under various conditions. The capacity to generate multiple autonomous conditions within a compact device enhances the throughput and efficiency of the screening process, thereby expediting the determination of optimal crystallization operation conditions.

In recent years, there has been a notable surge of interest in droplet-based microfluidic crystallization, as it permits the swift acquisition of data for studying nucleation kinetics and conducting screening of crystallization conditions.⁶⁴ In this methodology, reactions are encapsulated within monodisperse droplets, with each droplet serving as a microchamber. This encapsulation is achieved by introducing reagent phases into an immiscible carrier fluid. The size and uniformity of the generated droplets can be finely regulated through manipulation of the relative velocity and viscosity within four distinct channel configurations: capillaries,⁶⁷ flow-focusing,⁶⁸ Y and T-junctions.⁶⁹ Researchers have developed cost-effective and easy-to-use droplet-based microfluidic platforms for screening protein crystallization conditions.⁶⁴ By controlling the droplet volume through the velocity of the injection pump, the impact of droplet size on crystal quality was studied. As illustrated in Figure 2 (c), the protein and precipitant are initially blended at a T-junction and subsequently amalgamated with oil via a flow-focusing technique. This sequence of actions culminates in the formation of a droplet array, encompassing a broad spectrum of concentration combinations of both protein and precipitant.⁶³ This allows for the simultaneous monitoring of hundreds of crystallization processes through an optical microscope, greatly enhancing the efficiency of screening, Figure 2(d).⁶⁴

These microfluidic devices can be combined with droplet-based platforms with gradient generators to further improve screening efficiency. For example, Abdallah et al. developed a

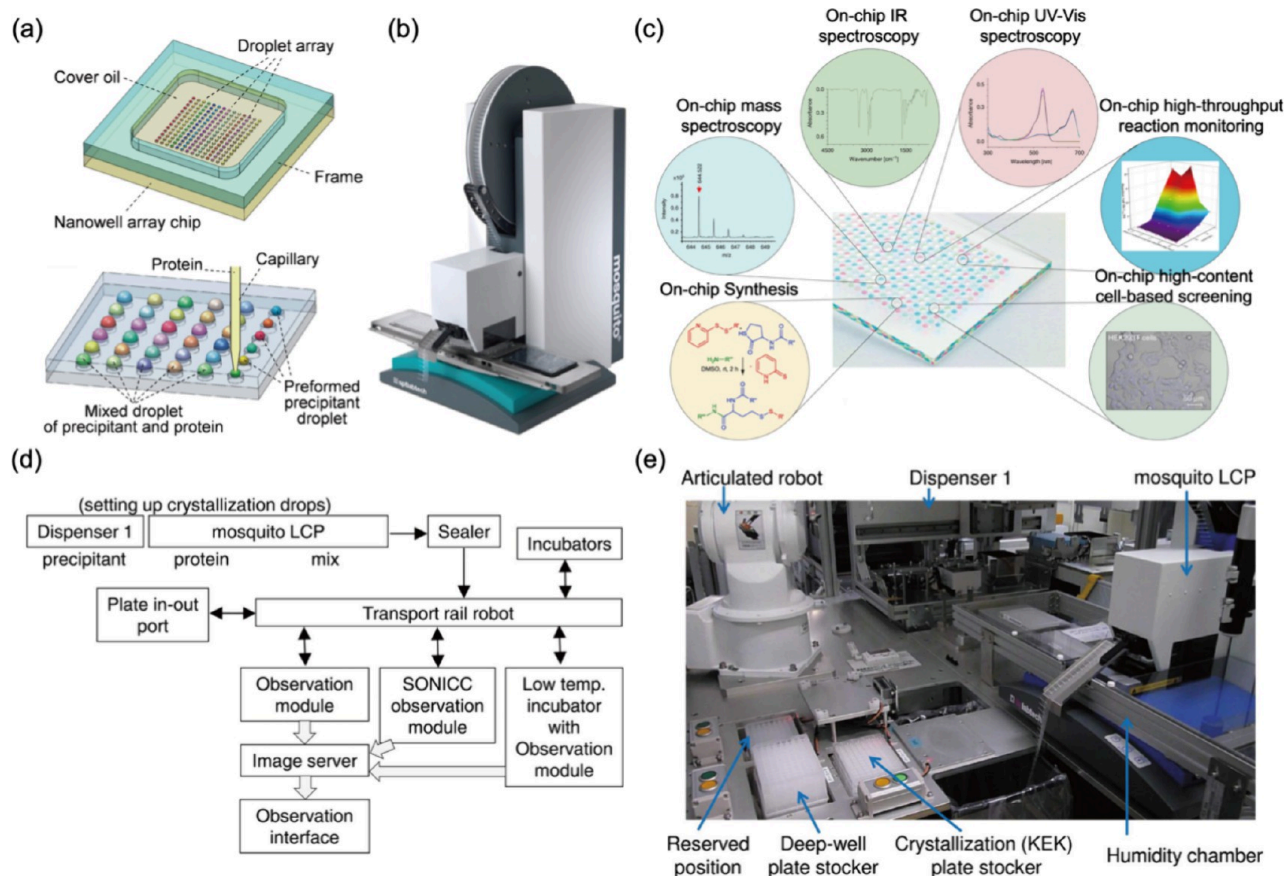


Figure 3. Droplet-based microwell screening strategy. (a) Droplet-based microwell plate: platform structure (up), semicontact liquid handling (down). Reproduced with permission from ref 78. Copyright 2014 Springer Nature; (b) liquid handling robot, Mosquito (TTP Labtech, Ltd.); (c) microwell plate integrating *in situ* characterizing techniques. Reproduced with permission from ref 79. Copyright 2020 Springer Nature; (d) the scheme of automatic protein crystallization screen platform, named XPS. Reproduced with permission from ref 80. Copyright 2006 Blackwell Munksgaard; (e) the upgraded version of XPS, XPS2. Reproduced with permission from ref 81. Copyright 2021 Blackwell Munksgaard.

droplet-based microfluidic platform through a gradient generation chip in which approximately 25 nL of protein sample is used for each condition, allowing for 170 different conditions to be tested simultaneously.⁶² Nonetheless, it is imperative to acknowledge that achieving precise control over the concentration of the protein and the precipitant within the microfluidic device can pose a notable challenge. Typically, the concentration information can be estimated by preliminary experiments using colorimetric assays.⁶²

On-Chip *in Situ* Monitoring Techniques. Furthermore, recent advancements in on-chip *in situ* monitoring technology have dramatically improved the efficiency of crystallization screening, particularly in the field of pharmaceutical crystallization. In addition to optical microscopy, on-chip X-ray diffraction, Raman spectroscopy, and confocal Raman microscopy have been employed to characterize the crystallization process within microfluidic devices, which has been systematically reviewed by Shi et al.⁷⁰ These advanced techniques allow for real-time observation and analysis and provide further insights into the structural information and crystallization kinetics for better understanding and optimization of crystallization operation conditions.

Overall, microfluidic techniques offer significant advantages in terms of high-throughput screening, rapid data collection, and the ability to characterize crystallization processes. Integrating various process analytical techniques further enhances the

capabilities of these devices for screening and optimizing crystallization conditions.

2.2. Microwell-Based Crystallization Platform. Droplet-Based Microwell Plate. Compared to conventional well plates, the droplet-based microwell plate platform accelerates screening efficiency by reducing API consumption while enabling more parallel experiments. The foundational element of this platform is the generation of microscale solutions. Currently, there are two main strategies: using liquid handling robots with integrated microdispensing heads⁷¹ and generating water-in-oil microdroplets on a microfluidic chip.⁷² Liquid handling robots can dispense and mix protein/precipitant solutions at volumes as low as 25–100 nL. These robotic systems find extensive utility in the domains of synthesis, crystallization, and biology, primarily attributed to their advantages, which encompass minimal sample utilization, stability, high throughput, and complete automation. For droplet-based well-plates, there is a need to perform high throughput crystallization experiments in small-volume volume well-plates with accurate manipulation of microdroplets.

The droplet-based microwell plate is widely utilized in protein crystallization and can be categorized into three screening modes: microbatch,⁷³ free interface diffusion,⁷⁴ and vapor diffusion.⁷⁵ Researchers currently face two main challenges when automating high-throughput screening under these methods, namely, (i) fabricating chips with high-density

microwells and (ii) achieving controlled mixing of precipitant and protein solutions at nanoliter or lower volumes.

Zhu et al. proposed one solution addressing the first challenge by fabricating a micron-precision microwell chip, as illustrated in Figure 3 (a). Standard photolithography and wet chemical etching methods were applied to fabricate chips featuring a high-density nanoarray well pattern (672 droplets/cm²) on a glass substrate coated with chromium and photoresist. Each droplet in the nanoarray well had a volume of only 4–8 nL, leading to a dramatic reduction in protein consumption by a factor of 50–500. However, this fabrication technique using lithography methods such as SU-8 lithography⁷⁶ and laser ablation⁷⁷ requires sophisticated and expensive instruments, making it less accessible.

For the second challenge, Liang et al. utilized 3D printing technology to fabricate a massively parallel protein screening chip based on the vapor diffusion mode.⁸² However, the chips produced by 3D printing are porous and permeable, which does not meet the requirements for airtight chambers in vapor diffusion crystallization. They further addressed airtightness by immersing the chip in dissolved paraffin wax to fill the micropores and applying a parylene coating. To provide hermetic sealing for the entire device, a glass sheet coated with PDMS film was used. Notably, the paraffin-filled chips are translucent, enabling direct imaging of crystals through optical microscopy. In comparison to standard 96-well plates, this chip demonstrated a 100-fold reduction in well size and a 7-fold increase in number density. Moreover, each independent experiment consumed just 10 μ L of protein solution, resulting in a significant 20–100-fold reduction when compared to standard plates.

Single- and multichannel liquid dispersion equipment enables precise mixing of precipitants with protein solutions at nanoliter or even lower volumes. Typically, liquid handling using a syringe pump is employed in single-channel liquid dispersion equipment. In Figure 3 (a), a tapered capillary tube is connected to a picoliter precision syringe pump for pumping and mixing liquids in nanoliter quantities or below, using a semicontact mode.⁷⁸ In this mode, the droplet at the tip contacts with a predroplet in the well, facilitating mixing. Semicontact mode offers advantages over contact or noncontact modes, where the capillary tip either directly contacts the substrate surface or does not make contact at all. The movement of the capillary tip is adjusted based on the droplet size, preventing the tip from breaking due to excessive movement of the carrier table or small droplets from not falling due to surface tension. The microplates are affixed to a three-axis (x-y-z) moving carrier table with nanometer-level precision. By programming simple commands, the carrier table can be synchronized with the syringe pump, allowing for the automated formation of droplets with customizable compositions by controlling the flow rate.

Multichannel liquid dispersion equipment is commercially available in liquid handling robots.⁸³ The liquid handling robot depicted in Figure 3 (b) is equipped with 4/8/16 channel modules and is compatible with standardized well plates, allowing for precise dosing of picoliter liquids. Compatible liquid dispensers include Phoenix (Art Robbins Instruments, Inc.), Mosquito (TTP Labtech, Ltd.), and OryxNano (Douglas Instruments, Ltd.). The high-throughput crystallization screening mode facilitates the swift exploration of crystallization operational parameters, ultimately leading to the acquisition of high-quality single crystals of pharmaceutical drugs, well-suited for single-crystal X-ray diffraction analysis. Tyler et al.

discovered the crystallization pathway of a pesticide, dithianon, which was previously thought to be noncrystallizable.⁸⁴ Besides, they also found 13th polymorphs of the precursor of the drug olanzapine.

The incompatibility among complex chemical environments, specialized infrastructure, existing analytical methods, and high-throughput screening poses a significant challenge in the screening of nanoscale crystallization conditions on microwell plates. Typically, the inspection of single crystals requires offline analysis using specialized devices. Standard well plates, commonly made of polystyrene, have physical properties that hinder their suitability for specialized analytical methods for example the excessive scattering of X-rays. Moreover, commercially available well plates designed specifically for *in situ* XRD are unable to effectively detect smaller crystals due to their relatively thick composition.⁸⁵ To satisfy the demand for modular equipment, Soliman et al. proposed a novel approach. Using a standard 96-well plate as a template, they etched patterns onto a transparent material, providing a more suitable platform for the screening of nanoscale crystallization conditions.⁸⁶ Through the use of contact line pins, the drops can be accurately positioned and confined, allowing for low-background *in situ* X-ray detection, regardless of plate orientation.

The chemBIOS well plate,⁷⁹ Figure 3 (c), is a nanodroplet based drug screening platform that integrates Nuclear Magnetic Resonance (NMR), Ultraviolet (UV), and Infrared (IR) spectroscopies. These spectroscopic techniques are commonly used to characterize pharmaceutical solid forms. The chemBIOS plate enables simultaneous synthesis, *in situ* analysis, and biological applications, with a screening capability of more than 5,000 results per experiment. Apart from the capability of chemical composition analysis, the device exhibits optical properties that lend themselves to on-chip characterization and real-time monitoring of reactions across the visible (utilizing optical microscopy), UV, and IR spectroscopy regions. The platform facilitates the simultaneous initiation of all synthetic reactions by clamping two slides together, resembling the process of crystallization where two droplets are mixed to form supersaturated solutions. Although primarily used for drug synthesis and analytical assays, the chemBIOS platform also proves valuable for high-throughput screening of polymorphs, cocrystals, and salts.

Continuous Automated Drug Crystallization Platform. With the advancement of automation techniques, there has been a development of fully automated crystallization screening infrastructure by both academic and industrial groups.⁸⁷ Typically, an automatic crystallization screening system consists of two main components: an automatic dispenser for setting droplets and a characterizing device for the automatic observation of these droplets. The integration of multiple computer programs to control these various modules is complex, ranging from the automated dispensing of droplets to the photography of crystals in each well.

In 2003, Kato and colleagues introduced a comprehensively automated system for protein crystallization, denoted as XPS, illustrated in Figure 3 (d).⁸⁰ This system integrates several components, including (i) protein and precipitant dispensers for crystallization, (ii) a plate sealer, (iii) a plate-transport rail robot, (iv) plate incubators, (v) an observation module, and (vi) an image-storage server. Setting up a crystallization condition can be accomplished within as little as 36 s in a 96-well plate.

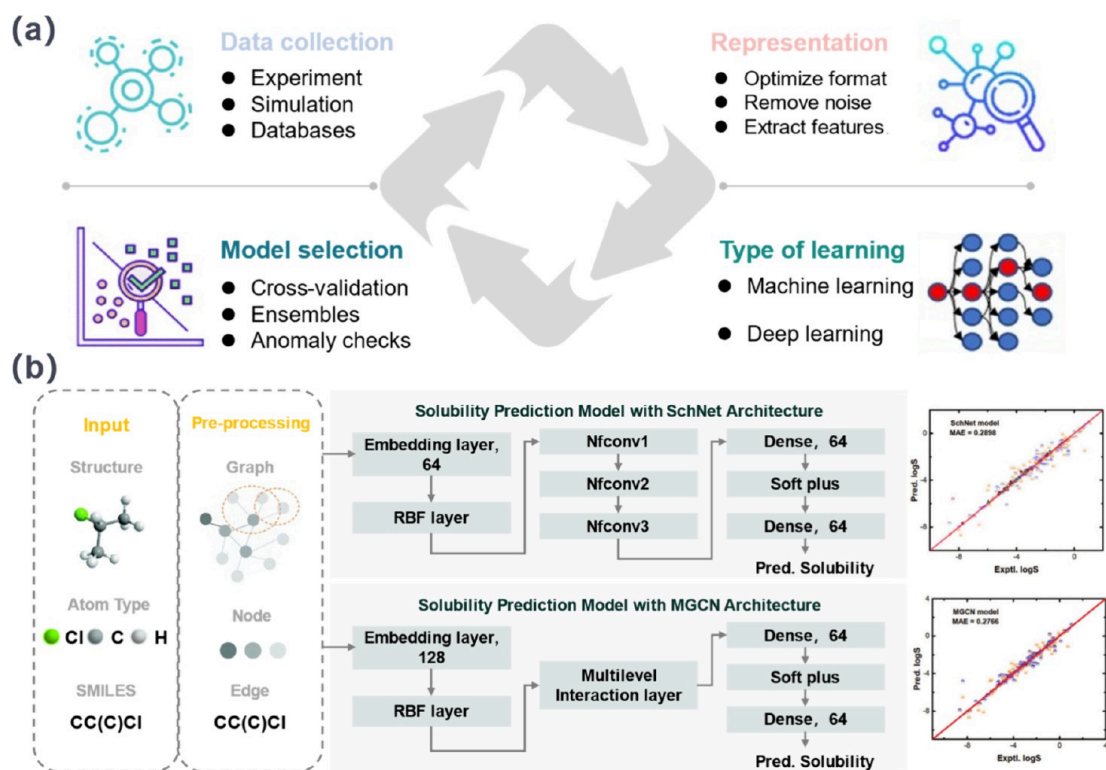


Figure 4. (a) General flowchart for predicting physicochemical properties from molecular and crystal structure data. (b) The overall architecture and prediction results of multilayer graph convolutional network (MGCN) and SchNet structure. Reproduced with permission from ref 90. Copyright 2020 Royal Society of Chemistry.

The XPS2 system,⁸¹ which represents an enhanced iteration of the original XPS, Figure 3(e), incorporates second harmonic generation and ultraviolet two-photon excited fluorescence within the crystal characterization module. Moreover, it integrates a cryogenic incubator accompanied by an observation module. Operating at a temperature of 4 °C, this cryogenic incubator augments the success rate of crystallization and is compatible with imaging instrumentation. It permits scheduled observations to transpire without necessitating a temperature adjustment of the crystallization plate. In addition, the XPS2 system consolidates all constituent modules under a unified software framework. This streamlined integration empowers users to command the XPS2 system through a singular control software, thereby automating all procedures ranging from the configuration of crystallization conditions to the observation of crystals.

In summary, microfluidic and automatic devices offer numerous advantages over traditional methods, allowing for free-convection environments or labor-free platforms that optimize crystal quality. These systems have evolved into readily accessible tools for high-throughput screening of crystallization parameters and the discernment of preferred solid forms of APIs.

3. PREDICTION OF PHYSICOCHEMICAL PROPERTIES

Physicochemical properties are intrinsic characteristics of drugs. The solubility, lipophilicity, and partition coefficient of the drug will indirectly affect the pharmacokinetic properties and drug target binding efficiency. The flowability and filterability of drug powders have important influences on the processes of production, mixing, transfer, and storage, and are of great significance for controlling product quality. Therefore, the determination of physicochemical properties is an essential step

in drug design and development.^{20,88} However, laborious and expensive experiments hinder rapid property prediction. For this reason, different artificial intelligence-based tools have been developed to predict these properties. Figure 4 (a) illustrates the general flowchart of physical and chemical property prediction.⁸⁹

Step I. Data Collection. The machine learning models trained in the prediction effort are constructed by learning patterns and regularities from large amounts of data. The data related to drug crystallization include molecular structure, solvent type, crystallization conditions (e.g., temperature, pressure), etc.

Step II. Data Representation. Effective representation can optimize the data format and make it easier for the model to recognize and process patterns and structures in the data, thereby improving the performance of subsequent learning tasks such as classification, regression, or clustering. In cases where the raw data may contain a lot of complex and redundant information, and where high-dimensional data may lead to a “dimensionality catastrophe”, the problem can be solved by data denoising, feature extraction, and dimensionality reduction techniques. In crystallography, this includes the extraction of chemical descriptors from molecular structures, such as intermolecular forces, stereochemical parameters, and process parameters from crystallization experiments. These features will be used as input to the machine learning model.

Step III. Type of Learning Algorithms. The type of learning algorithm requires a combination of many aspects. Data characteristics (data volume, data dimensions, data balance, data complexity, etc.), type of prediction task (regression, classification), training performance, interpretability, computational

resources (ML, DL), and a priori knowledge are all key factors in algorithm selection.

Step IV. Selection of Model. Combining the above three steps, a model with relatively optimal performance is constructed for the specific problem of physical and chemical property prediction.⁸⁸ During the model selection process, cross-validation can help evaluate the performance of different models and select the one that performs best on the validation set. This helps to prevent overfitting. In addition, when faced with complex or high-dimensional data, integration methods can be used to improve the performance of a single model and reduce model bias and variance. Also, anomaly checking is performed after model training to identify anomalous data results and feed them back into the data set for secondary cleaning. In practice, model selection is usually an iterative process that requires combining multiple techniques to obtain the most suitable model for solving a particular problem, while taking into account the robustness and generalization capabilities of the model.

3.1. Solubility. Solubility is a crucial physical property in the design and development of pharmaceutical crystallization. Solubility prediction has been the focus of intensive research as an experimental alternative that can reduce waste and improve crystallization efficiency. Pioneers have made some progress using AI algorithms.⁸⁸ However, the attainment of a reliable methodology for the direct and precise prediction of solubility remains a challenging pursuit due to factors including polymorphism, solid-solution interactions, data sample imbalances, appropriate descriptor selection, and the intricacy of computational models.²⁰

In ML-based studies on solubility prediction, differences arise in the choice of training and test data sets, the selection of descriptors, the types and quantity of descriptors, as well as the mathematical models employed for both classification and regression tasks. The data source is typically an open-source database (e.g., CSD) or undisclosed internal experimental data, and molecular descriptors are used in prediction models.⁹¹

The Determination of Descriptors. For physicochemical property prediction, molecular descriptors are used to refer to a measure of a molecule's properties in some aspects, either physicochemical properties of the molecule or numerical indices derived from the molecular structure by various algorithms. Certain quantitative relationships are established from the information carried by the molecular descriptors on the physicochemical properties of the molecule.

These descriptors can be categorized into four classes based on their dimensions:⁹² zero-dimensional (0D) descriptors are the most basic molecular representations, including essential properties like molecular weight, atom count, and atom type frequencies; one-dimensional (1D) descriptors represent drugs based on their substructures, such as ring counts and functional groups; two-dimensional (2D) descriptors provide further information by taking into account atomic adjacency, connectivity, and other topological features; three-dimensional (3D) descriptors depict molecules in the 3D spatial domain, with each atom defined by x, y, and z coordinates, capturing spatial and geometric configuration details.⁹³ Besides, the Simplified Molecular Input Line Entry System (SMILES) is another 1D linear notation for describing the structure of molecules and is used for the input and representation of molecular reactions with well-defined canonical formulas.⁹⁴ Based on this string, it can be converted into a 2D graph or 3D model of the molecule via the open-source cheminformatics

library RDKit.⁹⁵ In the ANN algorithm, the SMILES string of each molecule can serve as character-level input.⁹⁶ These descriptors encapsulate intricate physicochemical properties of drugs and find extensive utility in drug discovery and screening endeavors.

Regarding drug crystal solubility prediction, the principle of descriptor selection is based on the understanding and quantification of the relationship between molecular properties and solubility. Theories related to solubility encompass a range of approaches, including fragment-based semiempirical methods like general solubility equations, thermodynamic cycles, and thermodynamic parameters. Additionally, concepts such as Hansen solubility parameters and Hildebrandt solubility parameters are utilized. Various molecular dynamics methods, along with first-principles *ab initio* calculations, also contribute to the understanding of solubility.^{24,26,97,98}

Also, the relevance and number of descriptors directly affect the complexity of model training. Too many descriptors may lead to model overfitting. Therefore, finding a balance between the choice of descriptors and the generalization ability of the model has been a concern for researchers. As it has long been demonstrated, simple linear regression methods outperform more complex modeling methods when choosing three descriptors in terms of chemical structure to predict the aqueous solubility of certain available drugs.⁹⁹ The overfitting of data by artificial neural networks leads to a significant reduction in predictive performance when faced with solubility predictions for compounds outside the database.

To prevent an undue focus on the role of chemical structure in solubility prediction models, several advances have focused on quantifying structure–activity/property relationships and recognizing the importance of solute–solvent and solvent–solvent interactions. By meticulously considering the distinct contributions of the three energy-quantized interactions emerging from API dissolution behavior, Ma et al. found that solute–solvent interactions and the API properties are equally important in solubility prediction.²⁶ Guided by the purely perturbed chain statistical associative fluid theory (PC-SAFT, an important theory in molecular thermodynamics),⁹⁷ Ge et al. created 16 molecular descriptors to take into account intermolecular inert interactions, laying the foundation for solubility prediction. There are of course drawbacks, as the above work only used a small data set. Continuous efforts to improve predictive models primarily focus on enriching descriptors for solute–solvent interactions in aqueous/organic systems and interactions among solute molecules, as well as expanding the data set to comprehensively capture various solute molecule characteristics.

The Development of Models. The selection of learning algorithms and the optimization of prediction models are important parts of the development of property prediction. Among the comparison of ANN, SVM, RF, and Extra Trees models, the SVM model performed the best for its capability to solve high-dimensional problems and generalize the data set.²⁴ However, in a broader data set, ANN ultimately performed as the optimized model (among MLR, ANN, RF, ET, and SVM models) for predicting 103 drug molecules in water, acetonitrile, ethanol, acetone, and ethyl acetate systems.⁹⁷ In addition, prediction models with optimized data structures and high accuracy could be achieved by integrating two different algorithms (ANN and RDF algorithm), which was validated by our research group where excellent prediction results in estimating the solubility of certain APIs in ethanol.²⁶

DL algorithms represented by ANNs have exhibited great potential for solubility prediction.⁹⁶ Instead of manually extracting features, DL automatically filters the data and extracts high-dimensional features from the data. Recent work has explored molecular representations and DL modeling architectures. For example, the molecular graph can be constructed using SMILES, and thus the graph neural network (GNN) model can be built from scratch for molecular property prediction.^{100,101} The GNN model utilizes a molecular graph representation, where atoms are transformed into nodes and bonds are represented as edges of the graph.¹⁰¹ During the training process, the convolution operator is introduced into the graph structure to derive the graph convolution network (GCN).⁹⁰ In addition, compared to GCN, a new SchNet model is also applied for the prediction of the physicochemical properties of drug molecules. Gao et al. predicted the aqueous solubility of drugs well based on multilayer GCN and SchNet structures on two original data sets (Delaney¹⁰² and Huuskonen¹⁰³). The overall architecture of the two structures is shown in Figure 4 (b).⁹⁰ Both systems are designed to accurately capture and extract molecular structure information, making them well-suited for predicting various chemical properties. Further, later advances by Gao's team continue to confirm that GCN can directly utilize 3D structures to efficiently extract useful molecular information and is suitable for complex molecular property prediction.¹⁰⁴ In comparison, the DNN model is relatively less accurate but more transferable between independent data sets. It may be because the overly complex network structure of GCN is affected by data noise leading to overfitting. Lee et al. also highlighted the advantage of GCN in predicting solubility in several solute systems.¹⁰⁵ In addition to GCN, the traditional GNN has gradually evolved to derive other rich graph-based neural network models in recent years. Ahmad et al. used SMILES strings as molecular representations to develop models using simple graph convolution, graph homomorphism networks, graph attention networks, and AttentiveFP networks.¹⁰⁶ On the anticancer test data set, the AttentiveFP-based network model showed optimal performance.

As aforementioned above, prediction of aqueous solubility using multilevel GCNs, SchNet, graph attention networks, graph homomorphic networks, and others have been developed, however, they can lead to discrepant estimates because of different databases used.^{96,104,106,107} Current DL in property prediction is also striving to improve prediction performance with a broader sample of data.¹⁰⁴ Overall, striking a balance between model accuracy and generalization capability is a perpetual concern for researchers in both ML and DL. The quality and size of the data set, the principles, and dimensions of descriptor selection, as well as the complexity of the model and optimization strategies, are all critical factors. Insufficient data volume or imbalanced distribution, excessive or inappropriate feature selection, overtraining, or overly complex models can lead to overfitting in the final model. To mitigate this, appropriately increasing the number of training samples, employing regularization methods, reducing feature dimensionality and algorithm integration are all effective measures to enhance generalization capabilities.

3.2. Melting Point. Melting point is one of the basic physical and chemical properties of crystalline materials, which is determined not only by the properties of single molecules but also by the intermolecular interactions within the crystal structure.^{20,108} Currently, the prediction works that have been reported include melting points of pharmaceutical organic

compounds, thermal stability of macromolecules, melting temperature of polymers, and melting points of cocrystal crystals.

Some of the published work on the prediction of drug molecule melting points (models, database size, performance statistics) is summarized in Table 1. All of these works achieved

Table 1. Prediction of Melting Point through ML Method

reference	method	training set	test set	Performance (r^2/q^2)	
				training set	test set
Bergström ¹⁰⁹	PLS ^a	185	92	0.57	0.54*
Karthikeyan ¹¹⁰	ANN ^b	2089	1042	0.66	0.66*
Johnson ¹¹¹	MLR	106	—	0.90	—
Modarresi ¹¹²	MLR ^c	278	45	0.66	0.79*
Nigsch ¹⁰⁸	KNN ^d	3119	1000	0.33	0.44
Hughes ¹¹³	SVM ^e	150	50	0.84	0.50
Lazzús ¹¹⁴	ANN	260	73	0.96	0.98
Bhatarai ¹¹⁵	PLS	94	15	0.81	0.85
Guendouzia ¹¹⁶	MLR	42	20	0.95	0.96
Salahinejad ¹¹⁷	BRANNLP ^f	6593	1648	0.79	0.78

^aPartial least-squares. ^bArtificial neural network. ^cMultiple linear regression. ^dk-nearest neighbor. ^eSupport vector machine. ^fBayesian network employing the Laplacian prior. *Value of " q^2 " in performance.

good prediction performance at the time, but the results of these works are not comparable due to the different classes of drug molecules targeted, nonuniform database sizes and sources, various strategies for descriptor selection, and different model optimization processes.

Conclusively, the principles used to predict the melting points of organic compounds consist of the following categories: mechanics,¹⁰⁹ quantitative structure–property/activity relationship,¹¹² quantum mechanics calculation,¹¹⁶ and method based on fragment contribution.¹¹⁸ Most of these models have been developed using a limited set of specific organic compounds, restricting their applicability to a broader range of diverse compounds. Alternatively, methods based on group contribution can be employed to predict the thermodynamic properties of various chemical groups within compounds.¹¹⁶ However, many parameters and complex calculation programs were required to accurately estimate the melting point.

The enthalpy of sublimation can be used as a descriptor in a quantitative structure–property relationship (QSPR) model along with other molecular descriptors for modeling melting point predictions. The sublimation enthalpy serves as a thermochemical measure signifying the intensity of intermolecular interactions within a crystal lattice, reflecting the stability of the crystal structure. Additionally, it offers valuable insights into other thermodynamic properties, including enthalpy of formation and lattice energy. Salahinejad et al. developed a QSPR model based on sublimation heat, lattice energy, and melting point of organic compounds, which applies to compounds with diverse and complex structures.¹¹⁷ By combining sparse Bayesian feature selection with ML techniques to identify the most pertinent molecular descriptors for the model, it is possible to create a concise quantitative model that achieves highly accurate predictions of melting points.

The similar model can also predict the related properties of macromolecular substances, such as the thermal stability of proteins and the melting point of polymers. Thermal stability is a

fundamental biophysical property of protein. It is closely related to protein expression, folding, activity, and function. Gorania et al. proved that it is feasible and effective to predict the thermal stability of protein using ANN on the premise that its amino acid sequence is known.¹¹⁹ Moreover, protein engineering frequently involves point mutations. Addressing the matter of thermal stability concerning these protein point mutations is a prevalent challenge in protein engineering. Jia et al. established a classification and regression analysis model to predict the thermal stability of single-point mutations, which had the potential to guide the decision-making processes during protein design via mutagenesis.²⁷ This model was specifically designed for predicting the thermal stability of single-point mutations, while also emphasizing the potential for predicting the thermal stability of multipoint mutations.

Other approaches to estimating polymer melt points primarily revolve around the creation of group contribution methods and quantitative structure–property relationships, which rely on physicochemical, structural, and topological parameters.¹²⁰ For instance, a polymer can be represented numerically through descriptors derived from the topology, shape, and charge/polarity of a specific motif constituting the polymer backbone.¹²¹ An RF model, developed using existing data, was employed to swiftly predict the properties of multicomponent polyhydroxyalkanoate (PHA) copolymers.¹²² The results emphasized the physical significance of the relationship between the composition, structural units, and thermal properties of PHA.

On another point, in the drug development and design process, where cocrystals are developed to improve the physicochemical properties of drug products, AI techniques can also be used for the property prediction of unsynthesized cocrystals. Gamidi et al. established a model for predicting the melting point of unsynthesized cocrystals using ANN.¹²³ By selecting the ideal cofomers of cocrystals, the melting point of the new solid form of drug molecules can be set within a target range for more accurate and efficient melting point prediction.

3.3. Miscellaneous Properties. Lipophilicity. Drug companies face challenges in penetrating cell membranes with low lipophilicity while encountering issues during transportation in body fluids or plasma with high lipophilicity.¹²⁴ To predict this property, Zhang et al. employed SVM and genetic algorithms for feature selection and the conjugate gradient method for parameter optimization to establish a three-class classification model of lipophilicity.¹²⁴ The proposed three-class classification model can correctly distinguish the types of compounds according to their water solubility and lipophilicity.

Flowability. The flowability of powders of APIs is a key determinant impacting the manufacturing and production of solid oral dosage forms. Powder flowability can affect the mixing efficiency, production rate, and uniformity of APIs with excipients. Previous studies have focused on relatively free-flowing systems, with particle size distribution and particle shape as key factors for describing the flowability, allowing for reasonable predictions based on particle characteristics.¹²⁵ However, the powder flowability behavior is also associated with many other factors, such as the shape and roughness of the particles, specific surface, surface energy, and static electricity. Barjat et al. attempted to predict flowability across diverse materials and various flow conditions. They utilized analytical data from hundreds of materials to train a radial basis SVM.¹²⁶ They found the measured physical properties are reasonably

related to the flow behavior, leading to a good flowability classification model.

Filtration. Filter cake filtration stands to separate crystals from the mother liquor, a process greatly impacted by the size and shape of the crystals. Regrettably, determining the particle size and shape distribution (PSSD) is a challenging task, and its precise quantitative relationship with filter cake permeability remains unknown. Existing literature offers insights into predicting filterability through modeling, often relying on theories founded on strong assumptions, such as the Carman-Kozeny equation.¹²⁷ Certainly, the constraints in identifying nonspherical particles using current equipment have posed challenges in conducting prompt and efficient PSSD measurements.

Nevertheless, progress in imaging and computational technologies is steadily surmounting these hurdles, ensuring the accuracy and reliability of PSSD measurements. Perini et al. correlated the relative filter cake resistance with the particle size and shape distribution properties of two compounds that had needle crystal morphology.¹²⁸ In subsequent research, Monte Carlo sampling techniques were employed to create a stack/filter cake composed of needle-like crystals, characterizing the cake porosity, particle orientation, and size bias. Moreover, it successfully predicted the porosity trends for the experimental system involving L-glutamic acid.¹²⁹

4. SCREENING AND OPTIMIZATION OF CRYSTALLIZATION CONDITIONS

As introduced in Section 2, automated tools and laboratory-built platforms are now available for performing high-throughput experiments. However, the analysis of experimental results still heavily relies on human experiences. This manual analysis may also introduce biases and errors that lead to the exploration of incorrect parameter spaces in subsequent screening, making it challenging to identify the optimal conditions. In this regard, AI excels at rapidly analyzing vast amounts of data and performing repetitive tasks objectively and thus can offer an unbiased solution to overcome these limitations. ML algorithms, in particular, can aid in decision-making for exploring chemical spaces, thereby accelerating the screening process.

Crystallization is a process that combines scientific principles and artistic intuition, and the factors influencing the outcome of crystallization are not yet fully understood. Therefore, the screening or optimization of crystallization conditions may be considered as a “black-box” problem. AI can provide predictions based on existing experimental data, offering relatively rational guidance for subsequent experiments. In the following, we summarized the applications of AI-assisted crystallization condition screening and optimization for both small-molecule and macromolecule drugs.

4.1. Small Molecule. Controlling and engineering the solid form of APIs is critical in the development of small molecule drugs.^{2,3} However, the traditional approach to acquiring desired solid forms mainly depends on screening and optimization crystallization conditions of API, which involves exhaustively exploring all combinations in a high-dimensional chemical/operational space. This process generates abundant data, but conventional analysis methods struggle to keep up with the rate of data collection.

To address this challenge, image classification technology has emerged as a useful tool for accurately categorizing crystals. Its application in identifying crystals with desired characteristics, such as polymorph, crystal habit, and particle size, has

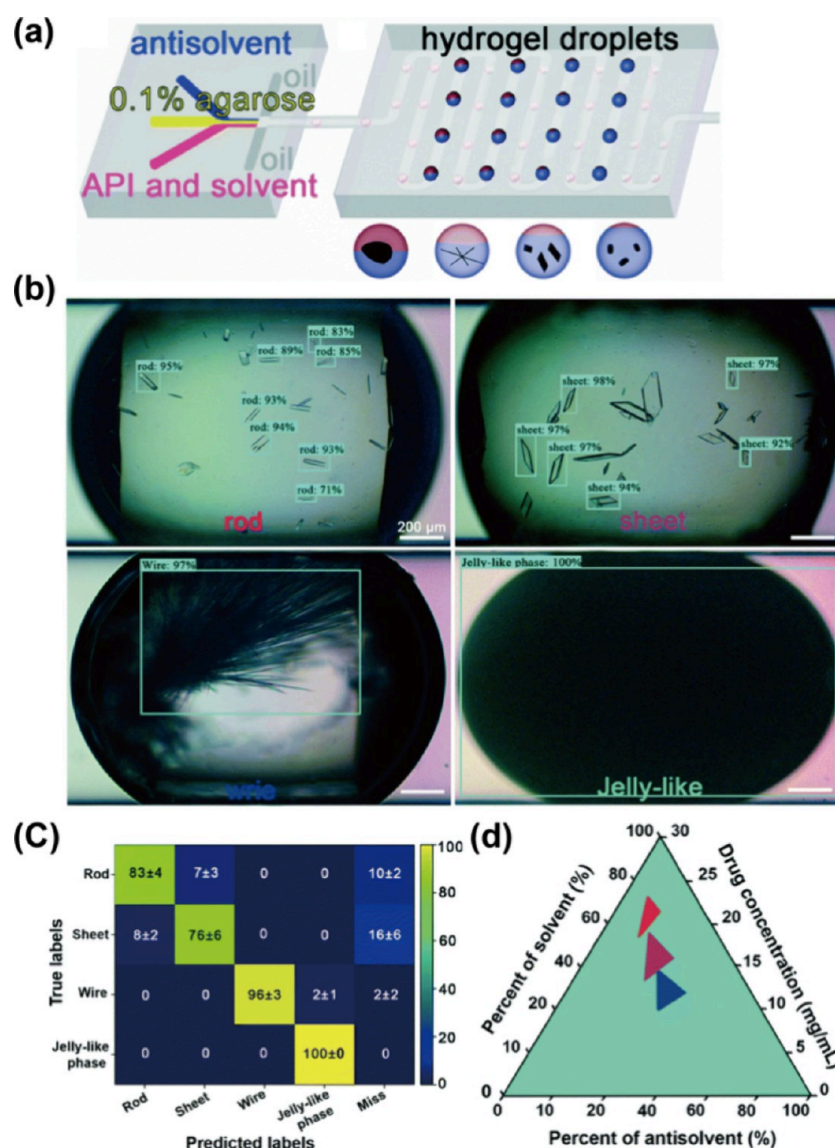


Figure 5. Integration of the hydrogel droplet platform with a deep learning approach. Reproduced with permission from ref 130. Copyright 2020 Royal Society of Chemistry. (a) Scheme of the platform. (b) The recognition and categorization of four distinct crystal morphologies within the image. (c) The confusion matrix was employed as an assessment tool to measure the efficiency of the Faster R-CNN model in classifying crystal morphologies within test data sets. (d) A ternary phase diagram was utilized to visualize the areas of operation parameters.

broadened considerably. For instance, Su et al.¹³⁰ applied DL algorithms to classify API crystal morphology in high-throughput hydrogel droplets, as illustrated in Figure 5 (a). In their study, 144 stable hydrogel droplets containing the API, solvent, and antisolvent were stored on an S-shaped chip using microfluidics technology. An object detection model called faster R-CNN was employed to identify crystals within the captured images from an optical microscope. The model outputted the possibility of a crystal being present in a given region of interest (ROI) within the image, allowing for accurate classification into four categories: jelly like, rod, sheet, and wire (Figure 5 (b)). The performance of crystal morphology identification was evaluated using a confusion matrix, presented in Figure 5 (c), which revealed the deviations between the predicted and actual results. The results demonstrated an accuracy of above 80% in identifying crystal morphology. Subsequently, the ternary phase diagram, as illustrated in Figure 5(d), was generated to graphically represent the crystallization

landscape for the four distinct crystal morphologies, contingent upon the proportion of antisolvent to solvent and the drug concentration. Overall, this work combined high-throughput crystallization experiments in hydrogel droplets with rapid processing of large amounts of experimental data using DL algorithms, providing valuable guidance for the precise preparation of pharmaceutical crystals.

While object detection algorithms are useful for classifying crystal images, real-time monitoring of API crystallization processes is also essential for capturing key information such as polymorph transformation, particle size changes, and morphology variations.¹⁸ These data can provide valuable insights into crystallization kinetics and the regulation of particle morphology/polymorph, thus guiding the optimization of crystallization conditions.¹³¹ The introduction of instance segmentation algorithms has enabled not only crystal classification but also the identification of crystal edges, making it possible to quantify multidimensional crystal morphology.¹³²

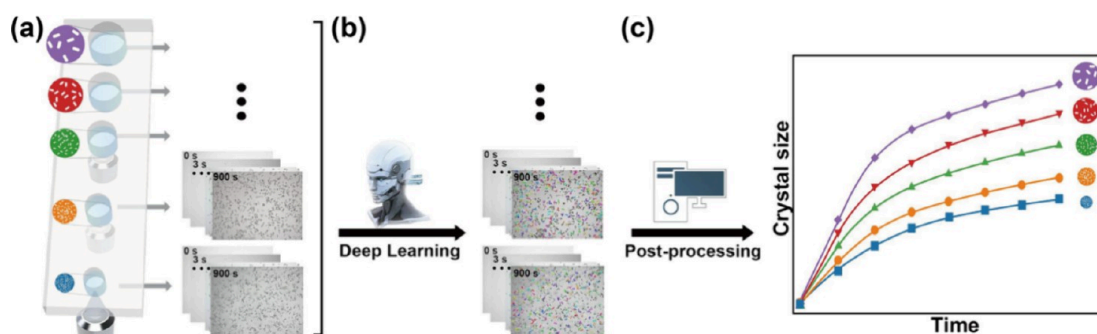


Figure 6. Application of instance segmentation in detecting crystal morphology data. Reproduced with permission from ref 133. Copyright 2022 Elsevier. (a) Crystallization experiments were conducted using a microwell plate. (b) Analysis crystals with DL. (c) Multiple facets of crystal data were obtained through postprocessing and subsequently plotted as functions of time over the course of the crystallization process.

He et al. used Mask R-CNN, an instance segmentation model, to analyze crystal size during API crystallization, as depicted in Figure 6.¹³³ The DL model efficiently and accurately analyzed thousands of crystal images of indomethacin and ibuprofen, providing information on crystal morphology, crystal size, and crystal size distribution (CSD) over time. Based on this time-resolved crystal morphology data set, growth kinetic profiles can be obtained in different microwell volumes using an established empirical crystal growth model. By examining crystallization kinetics in various sizes of microwells, they uncovered the surprising impact of microreactor volume on crystal growth rates. These insights can be used to guide the optimization of crystallization conditions for obtaining API crystals of specific sizes, which is crucial for subsequent scaling up and industrial production.

Instance segmentation algorithms can be applied to the study of polymorph control in API crystallization. Previously, we utilized the Mask R-CNN model to investigate solvent-mediated polymorphic transitions of APIs, enabling the tracking of crystal size, shape, and polymorphic changes.¹³⁴ In the phase transition from α to β crystal forms, the Mask R-CNN model accurately distinguished between the two forms. Additionally, the ML model allowed for the measurement of multidimensional data of crystal size and shape. Overall, the application of DL-based image processing techniques offers a noninvasive and high-throughput approach to studying crystallization processes. It improves the efficiency of data processing during the screening of crystallization conditions, making it a valuable tool for understanding polymorphic transformation and optimizing crystallization processes in pharmaceutical development.

Sometimes, one may improve the product quality attributes of small molecule drugs by forming cocrystals, salts, spherical particles, or inclusion complexes. The understanding of the complex interactions between APIs and solvents in these processes makes rational screening challenging. However, with the assistance of AI, key factors can be predicted at an early stage, reducing the workload of screening experiments significantly. AI models can be trained to predict the formation of multi-component solid forms using inputs that typically include the main body (API), guest molecules, and solvent environment represented by molecular descriptors. Together with existing experimental results, this data set is used to train and validate ML algorithms. Once candidate molecules or operation conditions are provided, the model outputs the probability of forming the desired multicomponent solid forms with the API.

Our group has developed prediction models for screening the components of spherical particles¹³⁵ and inclusion com-

plexes,¹³⁶ using ANN, SVM, and logistic regression (LR) algorithms in combination with molecular descriptors and solvation-free energy descriptors. Notably, models based on a combination of multiple ML algorithms (i.e., ANN, SVM, and LR) provided more accurate outcomes compared to those with a single algorithm. These prediction models enable the screening of potential components for various solid forms of small-molecule drugs. By incorporating molecular descriptors and solvation-free energy descriptors, the models can provide valuable insights and guide the selection of candidates efficiently and rationally.

Furthermore, the prediction of crystallization tendencies in API development is also a crucial aspect. Several approaches have been explored in this topic. Wicker and Copper employed one-dimensional computational chemical descriptors to forecast the crystallization propensity of small organic molecules by examining atom connectivity.¹³⁷ However, their data set for the noncrystallized outcomes was relatively small, limiting the robustness of their predictions. Pereira et al. introduced 3D chemical structure descriptors and ranked them using an RF algorithm to predict crystallization tendencies.¹³⁸ The consensus model, which combined RF, SVM, and multilayer perceptron (MLP) models, exhibited improved prediction results. Using untrained data sets, the consensus model achieved an accuracy of 80% in correctly predicting the crystallization propensity of molecules.

These approaches primarily focused on predicting the possibility of molecular crystallization based on physical and chemical properties, without considering the solvent. However, the solvent plays a significant role in crystallization processes. Hosokawa et al. established a classification model using decision trees and Bayesian probability methods to screen solvents for API crystallization, considering both chemical structures and solvents.¹³⁹ Ghosh et al. classified Pfizer's crystallization data based on experimental type and solvent system, but found that including crystallization methods (evaporation, melting, cooling, and antisolvent crystallization) introduced noise and reduced the performance of ML methods.¹⁴⁰

Considering the crystallization conditions and methods is crucial for accurate prediction. Data sets obtained from the published work multiply prediction difficulty, where crystallization conditions/methods were diverse. Pillong et al. reconstructed a data set with consistent conditions, including concentration, evaporation rate, temperature, and pressure, and used a random forest model to predict API crystallization propensity in multiple solvents.¹⁴¹ Their approach led to a one-

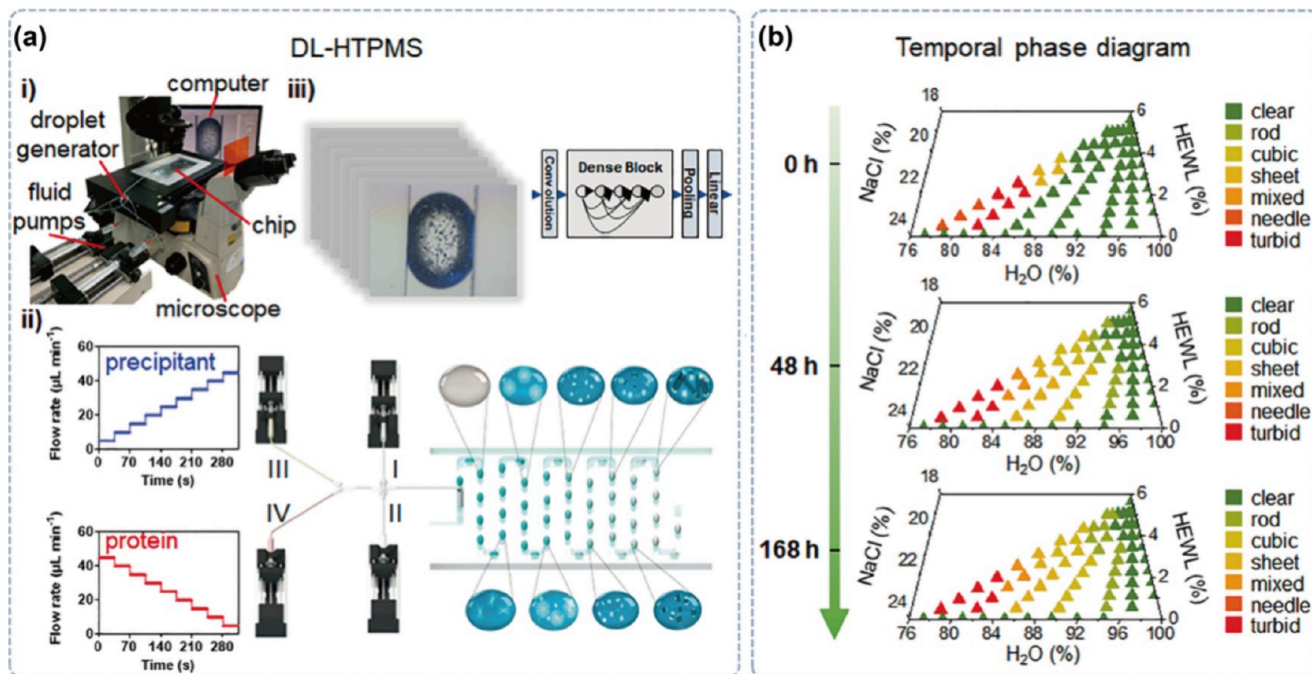


Figure 7. Screening for protein crystallization conditions. Reproduced with permission from ref 151. Copyright 2022 Elsevier. (a) High-throughput programmable microliter-droplet systems assisted by DL. (b) The evolution of protein crystalline or morphologies overtime during the crystallization process.

third reduction in the number of experiments needed for solvent screening.

Nevertheless, despite these endeavors, the task of forecasting the crystallization tendencies of APIs remains a challenging and ongoing pursuit. The constraints of limited data sets, the intricate nature of the crystallization process, and the wide array of conditions and methodologies documented in existing literature all present substantial obstacles for ML models. Consequently, additional research and development efforts are imperative to enhance the precision and dependability of prediction models within this domain.

4.2. Macromolecule. Therapeutic proteins, including monoclonal antibodies, have played a crucial role in treating various diseases, leading to significant revenue generation in the healthcare industry, which yielded a cumulative revenue of \$188 billion in the single year of 2017.¹⁴² However, the downstream purification process for these proteins, primarily relying on multistep chromatography techniques, accounts for a substantial portion (around 80%) of the manufacturing cost. This purification process has become a major bottleneck in achieving cost-effective production of protein products.

The field of protein crystallization, historically oriented toward the acquisition of sizable, top-quality single crystals for structural elucidation, has more recently garnered attention for its prospective utility in downstream bioprocessing and bioseparation applications. Crystallized proteins offer advantages such as high purity and increased stability, which are advantageous in formulation, storage, and drug delivery stages.¹⁴³ This has captured attention as a practical and scalable method for purifying and isolating proteins in downstream processes.¹⁴⁴

The concept of a “crystallization cocktail” refers to a mixture of different solutions or reagents that are combined to create the optimal conditions for protein crystals to form.¹⁴⁵ However, the complexity of protein structures and the environmental factors

involved make it challenging to predict the ideal crystallization cocktails based solely on the crystallization mechanism. As a result, the exploration of various factors including the concentrations of protein and precipitant, solution pH, temperature, etc., has been carried out blindly. Consequently, the investigation of conditions conducive to protein crystallization has constituted a central research focus for numerous years, leading to the development of automated/semiautomated high-throughput screening platforms and data processing methods.

AI has been applied to the field of protein crystallization, primarily in crystal identification and classification. The combination of automated high-throughput screening and AI-assisted classification and screening protocols has significantly improved the efficiency of optimizing crystallization conditions. In the early 21st century, protein crystals were evaluated using SVM algorithms for screening crystallization cocktails, achieving an 84.4% consistency rate compared to assessments made by human experts.¹⁴⁶ Buchala and Wilson¹⁴⁷ combined different feature extraction methods with multiple classifiers to achieve more accurate classification results. With the advent of convolutional neural networks (CNN), the MARCO model was trained and tested using a data set of 500,000 macromolecular crystal images, achieving an accurate identification rate of 94%.¹⁴⁸ The accuracy of crystal recognition was further improved to 97.23% using a scalable deep CNN, named EfficientNet.¹⁴⁹ Another CNN-based model called CrystalNet was developed, enabling the classification of protein crystals at different stages in time resolution during high-throughput screening, thereby facilitating the monitoring of crystal evolution.⁴⁷ The automated analysis system known as CrytPro enables stable and reliable monitoring of changes in crystal size during crystallization.¹⁵⁰

Huang et al.¹⁵¹ combined an object detection algorithm with a high-throughput screening platform to optimize protein

crystallization at the milliliter scale, as illustrated in Figure 7(a). Different protein and precipitant concentrations were programmed to generate microliter droplets (2160 droplets, 72 concentration combinations, and 30 replicate experiments per group). A CNN algorithm-based model (DenseNet) classified images into seven categories. Crystal morphology at different stages was then correlated with the crystallization conditions to obtain a temporal phase diagram, Figure 7(b), enabling accurate scale-up experiments with a 93% accuracy rate. The temporal phase diagram provided guidance for producing rod-shaped crystals at the milliliter scale, addressing the challenge of scaling up after high-throughput screening of crystallization conditions.

Over the past two decades, there have been various approaches to predicting protein crystallization propensity using ML techniques. These methods aim to forecast or evaluate the likelihood of a protein sequence crystallizing by leveraging sequence properties as key factors. The general workflow of these prediction methods can be summarized as follows: (1) the transformation of the protein sequence into a feature vector of fixed dimensions, and (2) using these feature vectors as inputs to a classification model to predict the crystallizable or noncrystallizable type of proteins.

Early prediction methods used statistical approaches to identify features associated with protein crystallization propensity.¹⁵² In 2006, the first ML-based predictor, known as SECRET, was introduced.¹⁵³ OB-Score, developed by Overton and Barton in the same year, relied solely on two features of protein sequences: isoelectric point and hydrophobicity.¹⁵⁴ Subsequent models such as CRYSTALP¹⁵⁵ and CRYSTALP2¹⁵⁶ incorporated additional features for example amino acid composition. Zucker et al. integrated experimental data to improve prediction performance for proteins with more than 200 residues.¹⁵⁷ Improving with the analysis of grand average hydropathy, the model performed well for both full-length proteins and protein domain constructs.¹⁵²

Other models, e.g., CRYSpred, conducted comprehensive searches of amino acid features within the amino acid index database to identify relevant indices for predicting crystallization propensity.¹⁵⁸ The results highlighted that residue characteristics located on the protein surface were more informative, compared with that encompassing all constituent residues, and the SCMCrys model further quantified the impact of amino acids using optimization techniques.¹⁵⁹ CrysaliS concentrated on analyzing the attributes of noncrystallizable proteins, and it pinpointed suboptimal sites within these proteins for their crystallization. This analysis led to the identification of potential single-point mutations that could potentially improve the propensity of crystallization.¹⁶⁰ Some prediction methods go beyond physicochemical properties and consider the stages of protein crystallization. For example, Wang et al. predicted overall protein crystallization propensity by quantifying the outcomes of five key stages in the crystallization process, leading to improved predictive performance.¹⁶¹ In addition to stand-alone algorithms such as SVM,¹⁵² SVR,¹⁶⁰ and Naïve Bayes¹⁵⁵ algorithms, researchers have explored fusion approaches and DL algorithms to improve the accuracy of predictions.

TargetCrys tackles the challenge of fusing protein sequences featured by adopting a decision-level fusion approach. It utilizes the intermediate decisions made by the first-layer SVM, based on sequence features, as input to the second-layer SVM.¹⁶² This decision-level fusion strategy helps enhance predictions. The XtalPred-RF model, an enhancement of XtalPred, incorporates the random forest (RF) algorithm.¹⁶³ This improved model

yields significant improvements in predicting successful protein crystallization, resulting in a notable reduction in screening workload. With the advent of DL algorithms, predicting protein crystallization propensity has become more accessible. DeepCrystal, for example, employs CNNs to identify proteins that have the potential to form high-quality crystals.¹⁶⁴ It eliminates the need to search for or design biochemical and structural features from protein sequences. XANNpred, another neural network-based model, can guide the construction design for high-quality crystals by focusing on residue sequence position.¹⁶⁵ Much like the prediction of crystallization tendencies for small molecules, forecasting the propensity for protein crystallization relies exclusively on the protein's sequence. This reliance on sequence-based prediction arises from the restricted alignment of technical and sampling parameters across high-throughput protein crystallization screening efforts.

5. POLYMORPH AND COCRYSTAL PREDICTION

During drug development, scientists pay much attention to the polymorphism and cocrystals which exert great influence on the drug's bioavailability, efficacy, toxicity, and shelf life.¹⁶⁶ Investigators undertake the examination of diverse polymorphs and cocrystals to select or engineer the most suitable drug forms for specific applications. This process encompasses the assessment of various parameters such as melting points and solubility mentioned in Section 3 to ensure both the efficacy and safety of the pharmaceutical compounds. Consequently, the investigation of polymorphism and eutectics constitutes a pivotal component within the realms of pharmaceutical crystallization.

However, experimental screening of polymorph and cocrystals is often tedious and expensive. Recently, AI has been used to improve the efficiency of polymorph and cocrystal screening by rebuilding the methodology of high-throughput experiments and computation predictions.¹⁶⁷ In this section, we reviewed fundamental principles and recent advancements in crystal structure prediction of polymorphs and coformer prediction of cocrystals using AI algorithms.

5.1. Data Mining. Polymorphism refers to the fact that a compound can exist in different crystal structures. The understanding of crystal structures is fundamental to the effective screening of pharmaceutical polymorphs. Crystal structure prediction (CSP) is one of the most vibrant research topics in pharmaceutical crystallization, which aims to predict the most reasonable crystal structure of a molecule without any a priori experimental knowledge. In drug development, CSP is frequently conducted within a search space confined by experimental data. In this manner, data mining from structure databases is particularly necessary for the use of AI techniques for polymorph prediction and screening.

Crystal Structure Database. The Cambridge Structure Database (CSD) contains crystal structures of all recognized organic and metal–organic compounds worldwide.^{168,169} In addition, the Inorganic Crystal Structure Database (ICSD) collects and provides information on the crystal structure of inorganic matter without C–H bonds, except for metals and alloys, as determined by experiments so far, including chemical names and formulas, unit cell parameters, space groups, atomic coordinates, and so on.¹⁷⁰ The Protein Data Bank (PDB) is currently the most important database for collecting the structure of biological macromolecules in 2.5 dimensions (representing three-dimensional data in two-dimensional form) and is a three-dimensional structure database of proteins,

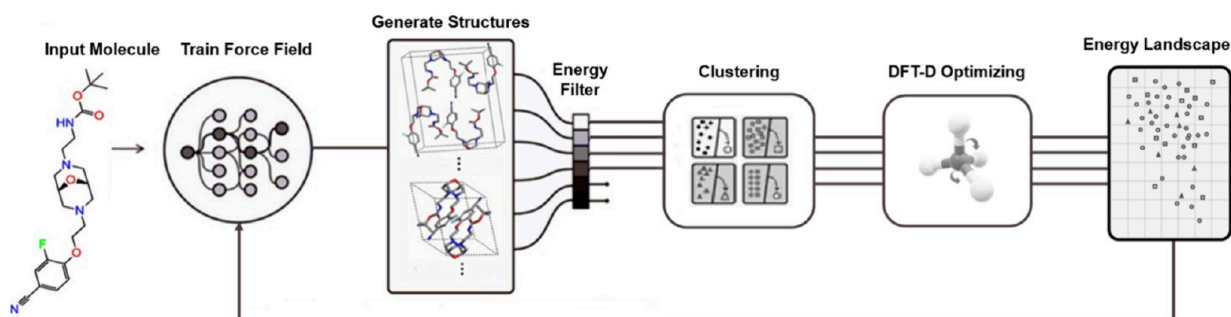


Figure 8. Workflow of cloud-computing-based CSP platform.

polysaccharides, nucleic acids, viruses, and other biological macromolecules identified by X-ray single crystal diffraction, nuclear magnetic resonance, electron diffraction, and other experimental methods.¹⁷¹ The Joint Committee on Powder Diffraction Standards (JCPDS) collects powder diffraction data for the identification of crystalline materials, which can serve as a data source in AI-assisted crystal phase analysis. Crystallography Open Database (COD), a computer-controlled database, stores crystallographic data, atomic coordinate parameters, and detailed chemical content and references.¹⁷² It has comprehensively and extensively collated, checked, and evaluated the quality of a large amount of molecular structure data, providing more accurate data than the original literature. These databases have been extensively used to construct data sets for model training purposes.

Data Extraction. The CSD has introduced a Python-based Application Programming Interface (CSD Python API) for its search method, enabling users to develop custom scripts for versatile searches. Within the CSD system, a suite of software tools is available for diverse molecular, intermolecular, and supramolecular analyses. These tools facilitate activities such as exploring crystal structures, examining molecular parameters, investigating conformational orientations, searching for pharmacophore models, and predicting intermolecular interactions.¹⁶⁹ For example, some software that is also affiliated with CCDC can help with conformational analysis and prediction. Mogul is a knowledge database based on the distillation of massive small molecule crystal structure data. Using Mogul, any unusual conformational features in small molecule conformations, such as unreasonable dihedral angles, bond lengths, and bond angles, can be examined.¹⁷³ A successful example has indicated that the Mogul software can be applied to assess the conformations of small organic molecules.¹⁷⁴

The Hydrogen Bonding Propensity (HBP) tool is a feature within the Solid Morphology module of Mercury software. It quantifies the potential for hydrogen bond formation between different functional groups in two dimensions.¹⁷⁵ This tool can be employed to assess the potential occurrence of a polymorphic form, which successfully identified some cocrystal conformer (CCF) of acetaminophen drug.¹⁷⁶ Galek et al. employed CSD to predict the potential CCFs of lamotrigine, ultimately identifying the most favorable cofomer.¹⁷⁷

5.2. AI-Assisted Polymorph Prediction. Predicting methods of pharmaceutical polymorphs are mainly based on molecular mechanics and quantum chemistry calculations.¹⁷⁵ In the former method, all possible energetic crystal structures are generated from the given molecular structure and ranked according to the values of the calculated lattice energy. The most likely crystal form usually has the lowest lattice energy. In the

latter method, energy evaluation based on quantum mechanics, for example, density functional theory (DFT) calculation, can be directly used or trained in the force field of molecular mechanics to optimize lattice energy.

At present, the more advanced CSP method mainly combines DFT calculation with structure search algorithms such as (quasi) random search, genetic algorithm, particle swarm optimization, and differential evolution process.¹⁷⁸ Due to the high cost of calculation, it is challenging to accurately calculate the lattice energy of a large number of crystal structures. The main challenge lies in the large number of possible crystal structures and their small energy differences.

Crystal Structure Prediction Method. The current mainstream CSP method is based on (i) *ab initio* global optimization with the first-principle calculations, (ii) ML technique that accelerates and supports decision and clustering results in the middle steps of the workflow, and (iii) template-based element replacement.¹⁷⁹ Figure 8 (a) outlines the steps of the CSP workflow. Among them, the lattice energy diagram has high requirements on the calculation methods, usually involving the advanced quantum mechanical calculations of molecules and crystals. However, these calculations are for a completely static lattice, ignoring the influence of temperature.

In addition, polymorphic prediction is challenging due to the existence of thermodynamically possible polymorphs, which usually contain more structure than experimentally obtained ones.³² Sun et al. conducted crystal structure prediction through a procedure akin to the one delineated earlier.¹⁸⁰ They employed Monte Carlo and particle swarm global optimization algorithms to generate crystal structures. This was underpinned by a combination of SVM classifiers and neural network models. Furthermore, they executed several iterations for relative free energy calculations, following high-precision DFT-D calculations. This prediction method was successfully applied to the stable polymorph prediction of oxabicyclidine AZD1305.

Furthermore, given the imperative to screen potential distinct structures within a constrained energy range, a comprehensive assessment of a substantial number of experimental structures becomes requisite. In such context, the adoption of an efficient maximum likelihood method for data analysis emerges as a highly favorable and suitable choice. The maximum likelihood method aims to optimize model parameters to maximize the probability of observing the data. In this regard, Wengert et al. developed a model enabling the exploration of an extensive array of candidate crystal structures. This model accurately captures the subtle interactions between molecules, including hydrogen bonding and many-body dispersion interactions.¹⁶⁷ This achievement was made possible by refining the physics-based description of remote interactions using the density generalized

tight binding (DFTB) level. The proposed workflow demonstrates broad applicability across various molecular materials.

Optimization of the Algorithm—Reduction on the Cost of Calculation. The exploration of structural candidates can be significantly expanded through search algorithms and by utilizing DFT-calculated energy as a metric for relative stability. However, the structure search process necessitates DFT calculations for numerous candidate structures, making it a time-consuming endeavor.

To reduce the computational cost, Cheng et al. introduced an optimization algorithm and constructed an innovative model combining the database, graph network (GN), and optimization algorithm.¹⁷⁸ Remarkably, its computational efficiency was approximately 1000 times higher than that of traditional methods for structure searching. The combination of the parallelism of the optimization algorithm and the adaptive capability of the neural network results in fast convergence compared to traditional search algorithms, which is suitable for such nonlinear and high-dimensional problems as polymorph prediction. In particular, BO has a significant advantage in tuning the hyperparameters of the neural network, which further reduces the computational cost of model training. This flexible framework, integrating material databases, graphical networks, and optimization algorithms, paves the way for innovative data-driven crystal structure prediction methods.

In contrast to the methods mentioned earlier, which heavily depend on the iterative use of computationally intensive *ab initio* calculations, Kusaba et al. highlighted a CSP approach that incorporates metric learning of crystal structure similarity.¹⁸¹ In this method, crystal structures are selected from an existing database, and predicted by a constructed binary classifier to be similar to stable structures corresponding to the given query compositions. Notably, initial calculations from first principles are necessary solely during the last stage, involving the local optimization of the proposed structure, leading to a substantial decrease in computational expenses. However, this technique depends on the substitution of elements and is applicable only if an appropriate template for substitution is accessible. As crystal structure databases continue to expand, the utilization of substitution-based CSP methods is expected to broaden in the future.

5.3. Coformer Prediction of Cocrystals. A large number of trial-and-error attempts are required to identify suitable cocrystal formers, highlighting the urgent need for prediction that can enhance the efficiency of cocrystal screening. Previous studies on predicting cocrystal formation have encompassed various computational methodologies, including molecular complementarity (MC), electrostatic potential, COSMO-RS,¹⁸² ML, molecular dynamics,¹⁸³ and hydrogen bonding propensity.¹⁷⁵ The principles of the aforementioned approaches primarily converge within the calculations of intermolecular interactions, requiring extensive calculations for each new set of potential cofomer components and generating test structures.

AI algorithms can be used to generate predictive empirical models from suitable data and to implement cocrystal screening. Table 2 lists several successful cases of cofomer predictions. Przybyłek et al. developed models based on simple categorical regression and multiple adaptive regression splines (MARS-plines) algorithms for cocrystal prediction of some amide API molecules such as meloxicam, deferiprone, and chlorzoxazone) with phenolic acid cofomers.¹⁸⁸ Wicker et al. reported the use of SVM to generate prediction models that can classify cofomers as successful or unsuccessful cocrystals using simple descriptors

Table 2. Some Successful Prediction Methods of Cocrystal Cofomers

reference	method	API	CCF
Przybyłek ¹⁸⁴	regression	amide	phenolic acid
Wicker ¹⁷⁶	SVM	paracetamol	Phenazine
Galek ¹⁷⁷	HBP	lamotrigine	methylparaben
Kim ¹⁸⁵	ANN	Pelubiprofen	Isonicotinamide/ nicotinamide
Hao ¹⁸⁶	DCN	Norfloxacin	phenazine
Wang ¹⁶⁶	DNN	Captopril	L-proline/sarcosine
Jiang ¹⁸⁷	CCGNet	pyrene	1,3,5-Trifluoro-2,4,6-triiodobenzene

of cofomer molecules, successfully predicting several cofomers that can bind to paracetamol drug.¹⁷⁶ Wang et al. developed a ML model trained on all cocrystal records within CSD and negative samples created by randomly pairing different molecules into chemical pairs.¹⁶⁶ Through experimental and virtual cocrystal screenings of the drug captopril, two cocrystals of captopril with L-proline and sarcosine were successfully discovered.

Some of the above studies have made meaningful attempts at cocrystal screening. However, the generalization ability of current ML methods in independent test sets remains unsatisfactory, only leading to the moderate prediction accuracy of the test set, which should be related to their deficiencies in data sets, descriptors, and models. In this regard, some of the following works give more importance to features (i.e., descriptors) in their prediction studies.

Mswahili et al. extracted descriptor values from the SMILE of compound simulations and found, in comparison with traditional ML approaches (RF, SVM), that ANNs exhibited optimal accuracy, thus showcasing significant potential.¹⁸⁹ Hao et al. incorporated two space-charge descriptors into ML modeling. They introduced two innovative ML algorithms: COSMO-SVM and 3D CNN.¹⁹⁰ Utilizing these prediction methods, researchers successfully identified two previously unknown cofomers for the drug norfloxacin. In a separate study, the introduction of the descriptors of pK_a rule successfully predicted the most likely cofomers capable of forming cocrystals with Pelubiprofen (PF), which exhibited significantly improved water solubility and dissolution behaviors compared to PF alone.¹⁸⁵

It is well-known that the independent testing ability or generalization ability of a model is particularly important for ML in practice, which represents the prediction ability of a model for new samples in an unknown space. Indeed, enhancing the independent testing capability of an ML model poses a challenge.

Data-driven ML heavily depends on substantial volumes of high-quality data. While the CSD provides a comprehensive collection of cocrystal structures suitable for supporting DL, it primarily comprises positive samples and lacks records of invalid cofomers (negative samples). Furthermore, characterizing the feature representation of samples is essential for machine learning models, especially when dealing with unbalanced data. Limited and unbalanced sample data is a challenge for DL applications and has been considered a difficult task.

Jiang et al. developed a highly accurate and comprehensive cocrystal screening model, Cocrystal Graph Network (CCGNet), based on the framework of graph neural networks, encompassing three key aspects: data set, sample representation, and model construction,¹⁹¹ which integrates feature complementation strategies to precisely forecast cocrystal formation

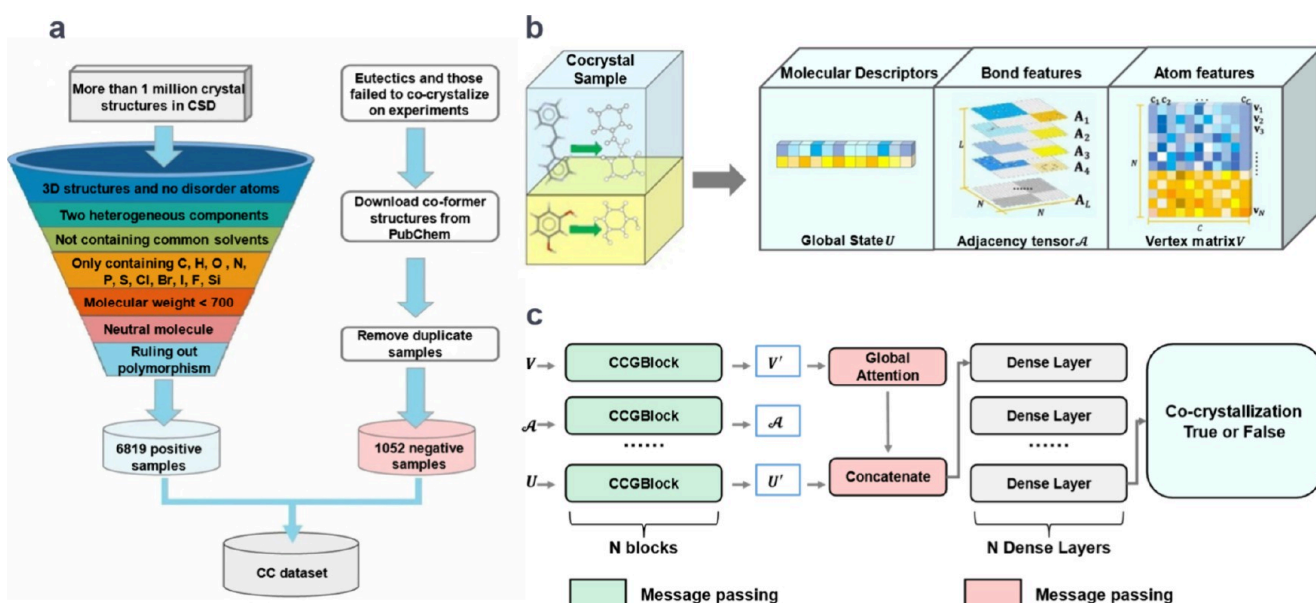


Figure 9. Overview of CCGNet cocrystal-screening framework and model performances on the independent test sets. Reproduced with permission from ref 191. Copyright 2021 Springer Nature. (a) The flowchart of sample collection. Left: the collection of cocrystal positive samples. Right: the collection of cocrystal negative samples. (b) Sample Representation. (c) The architecture of the CCGNet model. Green and pink denote the calculation block at the message passing phase and the readout phase, respectively.

(Figure 9). The prediction accuracy of the CCGNet model was much better than seven commonly used ML models in the prediction of four API cocrystal coformers, showing strong robustness and generalization ability. For limited and unbalanced data, this model provides a possibility to explore the application of DL in many fields lacking big data.

SUMMARY AND OUTLOOK

In recent years, the emergence of AI has provided a powerful tool for rebuilding the paradigm of pharmaceutical crystallization development.

AI algorithms are the foundation and the key for this paradigm revolution. ML algorithms have been used to predict the formation of crystalline materials; more profoundly, the use of deep learning algorithms, particularly in computer vision, has allowed for the accurate identification of crystals. The combination of DL algorithms like Graph Neural Networks with molecular mechanics or quantum chemistry calculations has shown superiority in crystal structure prediction for pharmaceutical polymorph development. Moreover, neural network-based approaches can capture intricate feature relationships and apply them to the analysis of both structural and attribute data in pharmaceutical cocrystal coformer prediction. Additionally, various ML algorithms were developed for predicting physicochemical properties.

Besides, huge amounts of data produced by high throughput platforms and automation provide the data basis for speeding up AI applications. On the one hand, with the advent of computer science, the availability of big data and various databases has greatly promoted the development of AI. On the other hand, the development of high throughput crystallization platforms including microfluidic devices and microwell-based platforms, catalyzed the generation of substantial data sets capable of training various predictive models in pharmaceutical crystallization. Moreover, the integration of automated crystallization robots and *in situ* characterization techniques further accelerated

the process efficiency and data collection of pharmaceutical crystallization development.

Overall, AI has revolutionized pharmaceutical crystallization by offering efficient (prediction) algorithms for predicting and screening pharmaceutical solid forms, as well as optimizing crystallization conditions during drug discovery and development processes. Albeit its prosperous future, artificial intelligence in the applications of pharmaceutical crystallization still faces several challenges:

- (1) Instruments are fundamental to the application of AI in pharmaceutical crystallization. Constructing and maintaining commercial automated screen systems are still challenges in many laboratories, due to their high-cost equipment investment. Besides, the incompatibility of high throughput platforms with *in situ* characterizing techniques hinders the efficient screening of pharmaceutical solid forms.
- (2) Image analysis technology in pharmaceutical crystallization needs to be extensively developed. Apart from crystal identification and multidimensional measurement, deeper applications such as 3D crystal reconstruction need to be developed through the computer version. Meanwhile, the analysis of large-scale characterization data obtained from spectroscopies and microscopies lags behind the rate at which data is collected.
- (3) The prediction of properties and polymorphs in drug crystallization will be a long-standing challenge for crystallography and computational chemistry. The computational principles include, but are not limited to, molecular dynamics, atomic mechanics, chemical bonding structure parameters, quantitative structure–property relationships, first principles based on density functional theory, etc. The choice of empirical methods, selection, and definition of parameters will fundamentally affect the underlying logic and comprehensive performance of the subsequent calculations. Correspondingly, the above

empirical methods will use application-specific generated descriptors as features, and the choice of descriptors is mainly reflected in the number and dimensionality of descriptors, the correlation between descriptors, which is directly related to the prediction results, generalization ability, and representation ability of the final model. Currently, the development of enhanced and efficient low-dimensional descriptors is still a direction of endeavor.

- (4) Machine learning algorithms have limitations on the size and dimensionality of the data, and they are key factors in determining the upper limit of machine learning performance. The quality and quantity of data directly affect the machine learning algorithm's ability to understand the data and learn patterns. For the process of specific drug crystal properties and crystal shape prediction, although some open-source databases exist, data collection and preprocessing are still a major problem, both to ensure that a certain amount of random noise is added to increase the data diversity and to achieve a certain degree of regularity, and to prevent high-dimensional data from causing a rise in the complexity of the model, leading to model overfitting.

Addressing these challenges, much attention should be put into these areas in the future. First, for exploring wide crystallization parameter space, the development of the automated high throughput screen is an important research hotspot. Meanwhile, the high-throughput platform combined with on-chip analysis technology needs to meet the demand for *in situ* characterization of crystals. Second, the empirical principles, data limitations, feature dimensions, model overfitting, and computational costs faced in the process of drug crystal property prediction and crystal shape prediction should be considered comprehensively. Data and features determine the upper limit of machine learning performance, and the ability to approach the upper limit can be improved through models and algorithms, while facing the need to constantly avoid/mitigate the occurrence of model overfitting through regularization, cross-validation, feature dimensionality reduction, etc., and comprehensively optimize to improve the model representation and generalization ability. Lastly, ML models should rely on exploring comprehensive sets of collected data for decision-making, rather than relying solely on the intuition of skilled experts. Based on the data sets acquired from an automated high-throughput platform, ML models iteratively determine their exploration of the crystallization parameter space, thus developing a fully automated pharmaceutical crystallization screening paradigm for ultimately realizing a self-driving screening laboratory.

AUTHOR INFORMATION

Corresponding Authors

Weiwei Tang — School of Chemical Engineering and Technology, State Key Laboratory of Chemical Engineering, Tianjin University, The Co-Innovation Center of Chemistry and Chemical Engineering of Tianjin, Tianjin 300072, P. R. China; orcid.org/0000-0002-7998-4350; Phone: 86-22-27405754; Email: wwtang@tju.edu.cn; Fax: +86-22-27314971

Junbo Gong — School of Chemical Engineering and Technology, State Key Laboratory of Chemical Engineering, Tianjin University, The Co-Innovation Center of Chemistry and

Chemical Engineering of Tianjin, Tianjin 300072, P. R. China; orcid.org/0000-0002-3376-3296; Phone: 86-22-27405754; Email: junbo_gong@tju.edu.cn; Fax: +86-22-27314971

Authors

Zuoxuan Zhu — School of Chemical Engineering and Technology, State Key Laboratory of Chemical Engineering, Tianjin University, The Co-Innovation Center of Chemistry and Chemical Engineering of Tianjin, Tianjin 300072, P. R. China

Yuan Zhang — School of Chemical Engineering and Technology, State Key Laboratory of Chemical Engineering, Tianjin University, The Co-Innovation Center of Chemistry and Chemical Engineering of Tianjin, Tianjin 300072, P. R. China

Zhixuan Wang — Shandong NHU Amino Acid Co., Ltd., Shandong 261108, P. R. China

Jingkang Wang — School of Chemical Engineering and Technology, State Key Laboratory of Chemical Engineering, Tianjin University, The Co-Innovation Center of Chemistry and Chemical Engineering of Tianjin, Tianjin 300072, P. R. China

Complete contact information is available at:
<https://pubs.acs.org/10.1021/acs.cgd.3c01408>

Author Contributions

^{||}Zuoxuan Zhu and Yuan Zhang contributed equally to this work.

Notes

The authors declare no competing financial interest.

Biographies



Zuoxuan Zhu graduated from Jiangsu Ocean University in 2021 with a bachelor's degree. He is currently a master student at the School of Chemical Engineering, Tianjin University. His research focuses on biomineralization and pathological crystallization.



Yuan Zhang graduated from Tianjin University in 2022 with a bachelor's degree. She is currently a master student at the School of Chemical Engineering, Tianjin University. Her research focuses on biomineralization and pathological crystallization.



Zhixuan Wang graduated from Chongqing University in 2007 with a bachelor's degree and completed his master's degree while working in 2015. He is currently a Ph.D. candidate at Tianjin University. His research focuses on industrial crystallization and crystallization kinetics.



Weiwei Tang studied Chemical Engineering at Tianjin University and completed his Ph.D. in 2018, where he investigated the mechanism of crystal nucleation at the molecular level. He was a visiting scholar (joint Ph.D. program) at the Department of Industrial and Physical Pharmacy (IPPH), Purdue University from 2014 to 2017. Subsequently, his research is focused on biomineralization and pathological crystallization at the University of Houston as a postdoctoral fellow. He is currently an associate professor in the School of Chemical Engineering and Technology at Tianjin University. His research centers on crystal engineering, biological (pathological) crystallization, industrial crystallization, and chiral separation via crystallization.



Jingfang Wang completed her bachelor's degree in Engineering from the Department of Chemical Engineering and her Master's Degree in Chemical Engineering at Tianjin University. She is a full professor at Tianjin University and focuses on the research of industrial crystallization fundamentals and applications. She was elected a member of the Chinese Academy of Engineering in 1999 and founded the National Engineering Research Center of Industrial Crystallization Technology (NERCICT) in China. She has received a number of awards and titles in the field of chemical engineering in both teaching and research aspects, including the member of the Academic Degrees Committee of the State Council, the Authorized Expert of Tianjin Municipal City, the National 8th Five-year Plan Advanced Individual, National March 8th Red Banner Holder, Outstanding Teacher in Tianjin, and many more.



Junbo Gong studied Chemical Engineering at Tianjin University and completed his Ph.D. in 2001. He was a visiting scholar at the University of Michigan from 2006 to 2008. He is currently a full professor at the School of Chemical Engineering and Technology at Tianjin University. He serves as the head of the Department of Chemical Engineering and the director of the National Engineering Research Center of Industrial Crystallization Technology (NERCICT). His current research mainly focuses on industrial crystallization, crystal engineering, crystallization process intensification, biological (pathological) crystallization, and functional crystalline materials.

■ ACKNOWLEDGMENTS

The authors are grateful for the financial support of National Natural Science Foundation of China (NNSFC 22278300) and the Natural Science Foundation of Tianjin (22JCZDJC00040).

■ REFERENCES

(1) Guo, M.; Sun, X.; Chen, J.; Cai, T. Pharmaceutical cocrystals: A review of preparations, physicochemical properties and applications.

- Acta Pharmaceutica Sinica B* **2021**, *11* (8), 2537–2564. Bolla, G.; Sarma, B.; Nangia, A. K. Crystal Engineering of Pharmaceutical Cocrystals in the Discovery and Development of Improved Drugs. *Chem. Rev.* **2022**, *122*, 11514. Healy, A. M.; Worku, Z. A.; Kumar, D.; Madi, A. M. Pharmaceutical solvates, hydrates and amorphous forms: A special emphasis on cocrystals. *Adv. Drug Delivery Rev.* **2017**, *117*, 25–46.
- (2) Chen, J.; Sarma, B.; Evans, J. M. B.; Myerson, A. S. Pharmaceutical Crystallization. *Cryst. Growth Des.* **2011**, *11* (4), 887–895.
- (3) Varshosaz, J.; Ghassami, E.; Ahmadipour, S. Crystal Engineering for Enhanced Solubility and Bioavailability of Poorly Soluble Drugs. *Curr. Pharm. Des.* **2018**, *24* (21), 2473–2496.
- (4) Detoisien, T.; Forite, M.; Taulelle, P.; Teston, J.; Colson, D.; Klein, J. P.; Veesler, S. A Rapid Method for Screening Crystallization Conditions and Phases of an Active Pharmaceutical Ingredient. *Org. Process Res. Dev.* **2009**, *13* (6), 1338–1342.
- (5) Lu, J.; Rohani, S. Polymorphism and Crystallization of Active Pharmaceutical Ingredients (APIs). *Curr. Med. Chem.* **2009**, *16* (7), 884–905.
- (6) Rodrigues, M.; Baptista, B.; Lopes, J. A.; Sarraguca, M. C. Pharmaceutical cocrystallization techniques. Advances and challenges. *Int. J. Pharm.* **2018**, *547* (1–2), 404–420.
- (7) Lutsko, J. F. How crystals form: A theory of nucleation pathways. *Sci. Adv.* **2019**, DOI: [10.1126/sciadv.aav7399](https://doi.org/10.1126/sciadv.aav7399). Cerreia Vioglio, P.; Chierotti, M. R.; Gobetto, R. Pharmaceutical aspects of salt and cocrystal forms of APIs and characterization challenges. *Adv. Drug Delivery Rev.* **2017**, *117*, 86–110.
- (8) Morissette, S. L.; Almarsson, Ö.; Peterson, M. L.; Remenar, J. F.; Read, M. J.; Lemmo, A. V.; Ellis, S.; Cima, M. J.; Gardner, C. R. High-throughput crystallization: polymorphs, salts, co-crystals and solvates of pharmaceutical solids. *Adv. Drug Delivery Rev.* **2004**, *56* (3), 275–300.
- (9) LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *nature* **2015**, *521* (7553), 436–444.
- (10) Szegedy, C.; Toshev, A.; Erhan, D. Deep neural networks for object detection. In *Advances in Neural Information Processing Systems 26 (NIPS 2013)*; NeurIPS, 2013; Vol. 26.
- (11) Aggarwal, R.; Sounderajah, V.; Martin, G.; Ting, D. S. W.; Karthikesalingam, A.; King, D.; Ashrafian, H.; Darzi, A. Diagnostic accuracy of deep learning in medical imaging: a systematic review and meta-analysis. *Npj Digital Med.* **2021**, DOI: [10.1038/s41746-021-00438-z](https://doi.org/10.1038/s41746-021-00438-z).
- (12) Price, S. L. Predicting crystal structures of organic compounds. *Chem. Soc. Rev.* **2014**, *43* (7), 2098–2111.
- (13) Nagy, Z. K.; Fevotte, G.; Kramer, H.; Simon, L. L. Recent advances in the monitoring, modelling and control of crystallization systems. *Chemical Engineering Research & Design* **2013**, *91* (10), 1903–1922.
- (14) Wengert, S.; Csanyi, G.; Reuter, K.; Margraf, J. T. Data-efficient machine learning for molecular crystal structure prediction. *Chemical Science* **2021**, *12* (12), 4536–4546.
- (15) Wang, D. Y.; Yang, Z.; Zhu, B. Q.; Mei, X. F.; Luo, X. M. Machine-Learning-Guided Cocrystal Prediction Based on Large Data Base. *Cryst. Growth Des.* **2020**, *20* (10), 6610–6621.
- (16) Yu, L. X.; Lionberger, R. A.; Raw, A. S.; D'Costa, R.; Wu, H.; Hussain, A. S. Applications of process analytical technology to crystallization processes. *Adv. Drug Delivery Rev.* **2004**, *56* (3), 349–369.
- (17) Huo, Y.; Liu, T.; Liu, H.; Ma, C. Y.; Wang, X. Z. In-situ crystal morphology identification using imaging analysis with application to the L-glutamic acid crystallization. *Chem. Eng. Sci.* **2016**, *148*, 126–139.
- (18) Wu, Y. Y.; Gao, Z. G.; Rohani, S. Deep learning-based oriented object detection for in situ image monitoring and analysis: A process analytical technology (PAT) application for taurine crystallization. *Chemical Engineering Research & Design* **2021**, *170*, 444–455.
- (19) Larsen, P. A.; Rawlings, J. B.; Ferrier, N. J. Model-based object recognition to measure crystal size and shape distributions from in situ video images. *Chem. Eng. Sci.* **2007**, *62* (5), 1430–1441.
- (20) Xiouras, C.; Cameli, F.; Quillo, G. L.; Kavousanakis, M. E.; Vlachos, D. G.; Stefanidis, G. D. Applications of Artificial Intelligence and Machine Learning Algorithms to Crystallization. *Chem. Rev.* **2022**, *122* (15), 13006–13042.
- (21) Gupta, R.; Srivastava, D.; Sahu, M.; Tiwari, S.; Ambasta, R. K.; Kumar, P. Artificial intelligence to deep learning: machine intelligence approach for drug discovery. *Molecular Diversity* **2021**, *25* (3), 1315–1360.
- (22) Wu, X.; Wang, S.; Zhang, Y. Survey on theory and application of k-Nearest-Neighbors algorithm. *Computer Engineering and Applications. Comput. Eng. Appl.* **2017**, *53* (21), 1–7. Keller, J. M.; Gray, M. R.; Givens, J. A. J. I. A fuzzy k-nearest neighbor algorithm. *IEEE Trans. Syst., Man, Cybern.* **1985**, *SMC-15* (4), 580–585.
- (23) Evgeniou, T.; Pontil, M.; Poggio, T. Regularization networks and support vector machines. *Adv. Comput. Math.* **2000**, *13* (1), 1–50.
- (24) Boobier, S.; Hose, D. R. J.; Blacker, A. J.; Nguyen, B. N. Machine learning with physicochemical relationships: solubility prediction in organic solvents and water. *Nat. Commun.* **2020**, *11* (1), 5753.
- (25) Breiman, L. Random forests. *Mach. Learn.* **2001**, *45* (1), 5–32.
- (26) Ma, Y.; Gao, Z.; Shi, P.; Chen, M.; Wu, S.; Yang, C.; Wang, J.; Cheng, J.; Gong, J. Machine learning-based solubility prediction and methodology evaluation of active pharmaceutical ingredients in industrial crystallization. *Frontiers of Chemical Science and Engineering* **2022**, *16* (4), 523–535.
- (27) Jia, L.; Yarlagaadda, R.; Reed, C. C. Structure Based Thermostability Prediction Models for Protein Single Point Mutations with Machine Learning Tools. *PLoS One* **2015**, *10* (9), No. e0138022.
- (28) Ghahramani, Z. Unsupervised learning. In *Advanced Lectures on Machine Learning*; Bousquet, O., VonLuxburg, U., Ratsch, G. Eds.; Lecture Notes in Artificial Intelligence, Vol. 3176; 2004; pp 72–112.
- (29) Modarresi, H.; Dearden, J. C.; Modarress, H. J. J. QSPR correlation of melting point for drug compounds based on different sources of molecular descriptors. *J. Chem. Inf. Model.* **2006**, *46* (2), 930–936.
- (30) Shlens, J. A Tutorial on Principal Component Analysis. *arXiv*, April 3, 2014. DOI: [10.48550/arXiv.1404.1100](https://doi.org/10.48550/arXiv.1404.1100).
- (31) Hewitt, M.; Cronin, M. T.; Enoch, S. J.; Madden, J. C.; Roberts, D. W.; Dearden, J. C. In silico prediction of aqueous solubility: the solubility challenge. *J. Chem. Inf. Model.* **2009**, *49* (11), 2572–2587.
- (32) Price, S. L. Predicting crystal structures of organic compounds. *Chem. Soc. Rev.* **2014**, *43* (7), 2098–2111.
- (33) Bebis, G.; Georgiopoulos, M. Feed-forward neural networks. *Ieee Potentials* **1994**, *13* (4), 27–31.
- (34) Rohani, S.; Haeri, M.; Wood, H. C. Modeling and control of a continuous crystallization process Part I. Linear and non-linear modeling. *Comput. Chem. Eng.* **1999**, *23* (3), 263–277.
- (35) Rumelhart, D. E.; Hinton, G. E.; Williams, R. J. Learning representations by back-propagating errors. *Nature* **1986**, *323* (6088), 533–536.
- (36) Suarez, L. A. P.; Georgieva, P.; de Azevedo, S. F. Nonlinear MPC for fed-batch multiple stages sugar crystallization. *Chem. Eng. Res. Des.* **2011**, *89* (6A), 753–767.
- (37) Forsyth, D. A.; Mundy, J. L.; di Gesù, V.; Cipolla, R.; LeCun, Y.; Haffner, P.; Bottou, L.; Bengio, Y. Object recognition with gradient-based learning. *Shape, Contour Grouping Comput. Vis.* **1999**, *1681*, 319–345.
- (38) Ke, T.-W.; Brewster, A. S.; Yu, S. X.; Ushizima, D.; Yang, C.; Sauter, N. K. A convolutional neural network-based screening tool for X-ray serial crystallography. *Journal of Synchrotron Radiation* **2018**, *25* (3), 655–670.
- (39) Souza, A.; Oliveira, L. B.; Hollatz, S.; Feldman, M.; Olukotun, K.; Holton, J. M.; Cohen, A. E.; Nardi, L. DeepFreak: Learning crystallography diffraction patterns with automated machine learning. *arXiv*, April 26, 2019. DOI: [10.48550/arXiv.1904.11834](https://doi.org/10.48550/arXiv.1904.11834).
- (40) Otsu, N. A threshold selection method from gray-level histograms. *IEEE transactions on systems, man, and cybernetics* **1979**, *9* (1), 62–66.
- (41) Schumacher, A. B.; Englert, A. H.; Susin, J. B.; Marczak, L. D. F.; Cardozo, N. S. M. An Automated Measuring Methodology for Crystal Size in Sweetened Condensed Milk Using Digital Image Processing and Analysis. *Food Analytical Methods* **2015**, *8* (7), 1858–1867.

- (42) Liu, J.; Kuang, W. J.; Liu, J. Q.; Gao, Z. G.; Rohani, S.; Gong, J. B. In-situ multi-phase flow imaging for particle dynamic tracking and characterization: Advances and applications. *Chem. Eng. J.* **2022**, 438, 135554.
- (43) Lecun, Y.; Bottou, L.; Bengio, Y.; Haffner, P. Gradient-based learning applied to document recognition. *Proceedings of the Ieee* **1998**, 86 (11), 2278–2324.
- (44) Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*; IEEE, 2014; pp 580–587.
- (45) Shelhamer, E.; Long, J.; Darrell, T. Fully Convolutional Networks for Semantic Segmentation. *Ieee Transactions on Pattern Analysis and Machine Intelligence* **2017**, 39 (4), 640–651.
- (46) He, K. M.; Gkioxari, G.; Dollar, P.; Girshick, R. Mask R-CNN. *Ieee Transactions on Pattern Analysis and Machine Intelligence* **2020**, 42 (2), 386–397.
- (47) Yann, M.; Tang, Y. Learning Deep Convolutional Neural Networks for X-Ray Protein Crystallization Image Analysis. *AAAI* **2016**, DOI: 10.1609/aaai.v30i1.10150.
- (48) Chen, Y. T.; Chen, S. F. Localizing plucking points of tea leaves using deep convolutional neural networks. *Comput. Electron. Agric.* **2020**, 171, 105298.
- (49) Patel, D. D.; Joguparthi, V.; Wang, Z.; Anderson, B. D. Maintenance of Supersaturation I: Indomethacin Crystal Growth Kinetic Modeling Using an Online Second-derivative Ultraviolet Spectroscopic Method. *J. Pharm. Sci.* **2011**, 100 (7), 2623–2641.
- (50) Malwade, C. R.; Qu, H. Y. Cooling Crystallization of Indomethacin: Effect of Supersaturation, Temperature, and Seeding on Polymorphism and Crystal Size Distribution. *Org. Process Res. Dev.* **2018**, 22 (6), 697–706.
- (51) Tona, R. M.; McDonald, T. A. O.; Akhavein, N.; Larkin, J. D.; Lai, D. Microfluidic droplet liquid reactors for active pharmaceutical ingredient crystallization by diffusion controlled solvent extraction. *Lab Chip* **2019**, 19 (12), 2127–2137.
- (52) Liotta, S. Kinetics of Precipitation. *J. Am. Chem. Soc.* **1965**, 87 (6), 1414–1414.
- (53) Mayr, L. M.; Bojanic, D. Novel trends in high-throughput screening. *Current Opinion in Pharmacology* **2009**, 9 (5), 580–588.
- (54) Sauter, C.; Dhoub, K.; Lorber, B. From macrofluidics to microfluidics for the crystallization of biological macromolecules. *Cryst. Growth Des.* **2007**, 7 (11), 2247–2250.
- (55) Wang, L.; Sun, K. L.; Hu, X. J.; Li, G.; Jin, Q. H.; Zhao, J. L. A centrifugal microfluidic device for screening protein crystallization conditions by vapor diffusion. *Sensors and Actuators B-Chemical* **2015**, 219, 105–111.
- (56) Candoni, N.; Grossier, R.; Lagaize, M.; Veisler, S. Advances in the Use of Microfluidics to Study Crystallization Fundamentals. *Annual Review of Chemical and Biomolecular Engineering* **2019**, 10, 59–83.
- (57) Besenhard, M. O.; Neugebauer, P.; Ho, C. D.; Khinast, J. G. Crystal Size Control in a Continuous Tubular Crystallizer. *Cryst. Growth Des.* **2015**, 15 (4), 1683–1691.
- (58) Lee, J. N.; Park, C.; Whitesides, G. M. Solvent compatibility of poly(dimethylsiloxane)-based microfluidic devices. *Anal. Chem.* **2003**, 75 (23), 6544–6554.
- (59) Ren, K. N.; Zhou, J. H.; Wu, H. K. Materials for Microfluidic Chip Fabrication. *Acc. Chem. Res.* **2013**, 46 (11), 2396–2406.
- (60) Kotnala, A.; Zheng, Y.; Fu, J. P.; Cheng, W. Microfluidic-based high-throughput optical trapping of nanoparticles. *Lab Chip* **2017**, 17 (12), 2125–2134.
- (61) Jasch, K.; Barth, N.; Fehr, S.; Bunjes, H.; Augustin, W.; Scholl, S. A Microfluidic Approach for a Continuous Crystallization of Drug Carrier Nanoparticles. *Chem. Eng. Technol.* **2009**, 32 (11), 1806–1814.
- (62) Abdallah, B. G.; Roy-Chowdhury, S.; Fromme, R.; Fromme, P.; Ros, A. Protein Crystallization in an Actuated Microfluidic Nanowell Device. *Cryst. Growth Des.* **2016**, 16 (4), 2074–2082.
- (63) Puigmartí-Luis, J. Microfluidic platforms: a mainstream technology for the preparation of crystals. *Chem. Soc. Rev.* **2014**, 43 (7), 2253–2271.
- (64) Ferreira, J.; Castro, F.; Rocha, F.; Kuhn, S. Protein crystallization in a droplet-based microfluidic device: Hydrodynamic analysis and study of the phase behaviour. *Chem. Eng. Sci.* **2018**, 191, 232–244.
- (65) Dertinger, S. K. W.; Chiu, D. T.; Jeon, N. L.; Whitesides, G. M. Generation of gradients having complex shapes using microfluidic networks. *Anal. Chem.* **2001**, 73 (6), 1240–1246.
- (66) Jeon, N. L.; Dertinger, S. K. W.; Chiu, D. T.; Choi, I. S.; Stroock, A. D.; Whitesides, G. M. Generation of solution and surface gradients using microfluidic systems. *Langmuir* **2000**, 16 (22), 8311–8316.
- (67) Tice, J. D.; Song, H.; Lyon, A. D.; Ismagilov, R. F. Formation of droplets and mixing in multiphase microfluidics at low values of the Reynolds and the capillary numbers. *Langmuir* **2003**, 19 (22), 9127–9133.
- (68) Anna, S. L.; Bontoux, N.; Stone, H. A. Formation of dispersions using “flow focusing” in microchannels. *Appl. Phys. Lett.* **2003**, 82 (3), 364–366.
- (69) Dreyfus, R.; Tabeling, P.; Willaime, H. Ordered and disordered patterns in two-phase flows in microchannels. *Phys. Rev. Lett.* **2003**, DOI: 10.1103/PhysRevLett.90.144505.
- (70) Shi, H. H.; Xiao, Y.; Ferguson, S.; Huang, X.; Wang, N.; Hao, H. X. Progress of crystallization in microfluidic devices. *Lab Chip* **2017**, 17 (13), 2167–2185.
- (71) Stevens, R. C. High-throughput protein crystallization. *Curr. Opin. Struct. Biol.* **2000**, 10 (5), 558–563.
- (72) Li, L.; Ismagilov, R. F. Protein Crystallization Using Microfluidic Technologies Based on Valves, Droplets and SlipChip. *Annual Review of Biophysics* **2010**, 39, 139–158.
- (73) Li, L.; Du, W. B.; Ismagilov, R. F. User-Loaded SlipChip for Equipment-Free Multiplexed Nanoliter-Scale Experiments. *J. Am. Chem. Soc.* **2010**, 132 (1), 106–111. Zheng, B.; Roach, L. S.; Ismagilov, R. F. Screening of protein crystallization conditions on a microfluidic chip using nanoliter-size droplets. *J. Am. Chem. Soc.* **2003**, 125 (37), 11170–11171.
- (74) Hansen, C. L.; Skordalakes, E.; Berger, J. M.; Quake, S. R. A robust and scalable microfluidic metering method that allows protein crystal growth by free interface diffusion. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, 99 (26), 16531–16536.
- (75) Zheng, B.; Tice, J. D.; Ismagilov, R. F. Formation of Arrayed Droplets by Soft Lithography and Two-Phase Fluid Flow, and Application in Protein Crystallization. *Adv. Mater.* **2004**, 16 (15), 1365.
- (76) del Campo, A.; Greiner, C. SU-8: a photoresist for high-aspect-ratio and 3D submicron lithography. *J. Micromech. Microeng* **2007**, 17 (6), R81.
- (77) Selimovic, S.; Piraino, F.; Bae, H.; Rasponi, M.; Redaelli, A.; Khademhosseini, A. Microfabricated polyester conical microwells for cell culture applications. *Lab Chip* **2011**, 11 (14), 2325–2332.
- (78) Zhu, Y.; Zhu, L. N.; Guo, R.; Cui, H. J.; Ye, S.; Fang, Q. Nanoliter-Scale Protein Crystallization and Screening with a Microfluidic Droplet Robot. *Sci. Rep.* **2014**, DOI: 10.1038/srep05046.
- (79) Benz, M.; Asperger, A.; Hamster, M.; Welle, A.; Heissler, S.; Levkin, P. A. A combined high-throughput and high-content platform for unified on-chip synthesis, characterization and biological screening. *Nat. Commun.* **2020**, DOI: 10.1038/s41467-020-19040-0.
- (80) Hiraki, M.; Kato, R.; Nagai, M.; Satoh, T.; Hirano, S.; Ihara, K.; Kudo, N.; Nagae, M.; Kobayashi, M.; Inoue, M.; et al. Development of an automated large-scale protein-crystallization and monitoring system for high-throughput protein-structure analyses. *Acta Crystallogr. D Biol. Crystallogr.* **2006**, 62 (9), 1058–1065.
- (81) Kato, R.; Hiraki, M.; Yamada, Y.; Tanabe, M.; Senda, T. A fully automated crystallization apparatus for small protein quantities. *Acta Crystallographica Section F-Structural Biology Communications* **2021**, 77, 29–36.
- (82) Liang, Y. R.; Zhu, L. N.; Gao, J.; Zhao, H. X.; Zhu, Y.; Ye, S.; Fang, Q. 3D-Printed High-Density Droplet Array Chip for Miniaturized Protein Crystallization Screening under Vapor Diffusion Mode. *ACS Appl. Mater. Interfaces* **2017**, 9 (13), 11837–11845.
- (83) Zhu, Y.; Zhang, Y. X.; Cai, L. F.; Fang, Q. Sequential Operation Droplet Array: An Automated Microfluidic Platform for Picoliter-Scale

- Liquid Handling, Analysis, and Screening. *Anal. Chem.* **2013**, *85* (14), 6723–6731.
- (84) Tyler, A. R.; Ragbirsingh, R.; McMonagle, C. J.; Waddell, P. G.; Heaps, S. E.; Steed, J. W.; Thaw, P.; Hall, M. J.; Probert, M. R. Encapsulated Nanodroplet Crystallization of Organic-Soluble Small Molecules. *Chem.* **2020**, *6* (7), 1755–1765.
- (85) Eisenstein, M. The shape of things. *Nat. Methods* **2007**, *4* (1), 95–102.
- (86) Soliman, A. S. M.; Warkentin, M.; Apker, B.; Thorne, R. E. Development of high-performance X-ray transparent crystallization plates for in situ protein crystal screening and analysis. *Acta Crystallographica Section D-Biological Crystallography* **2011**, *67*, 646–656.
- (87) Gorrec, F.; Löwe, J. Automated protocols for macromolecular crystallization at the MRC laboratory of molecular biology. *J. Vis. Exp.* **2018**, No. 131, No. e55790.
- (88) Ma, Y.; Gao, Z.; Shi, P.; Chen, M.; Wu, S.; Yang, C.; Wang, J.; Cheng, J.; Gong, J. Machine learning-based solubility prediction and methodology evaluation of active pharmaceutical ingredients in industrial crystallization. *Front. Chem. Sci. Eng.* **2022**, *16* (4), 523–535.
- (89) Butler, K. T.; Davies, D. W.; Cartwright, H.; Isayev, O.; Walsh, A. Machine learning for molecular and materials science. *Nature* **2018**, *559* (7715), 547–555.
- (90) Gao, P.; Zhang, J.; Sun, Y.; Yu, J. Accurate predictions of aqueous solubility of drug molecules via the multilevel graph convolutional network (MGCN) and SchNet architectures. *Phys. Chem. Chem. Phys.* **2020**, *22* (41), 23766–23772.
- (91) Sun, H. M. A universal molecular descriptor system for prediction of LogP, LogS, LogBB, and absorption. *J. Chem. Inf. Comput. Sci.* **2004**, *44* (2), 748–757.
- (92) Chen, W.; Liu, X.; Zhang, S.; Chen, S. Artificial intelligence for drug discovery: Resources, methods, and applications. *Mol. Ther. Nucleic Acids* **2023**, *31*, 691–702.
- (93) Matter, H.; Potter, T. Comparing 3D pharmacophore triplets and 2D fingerprints for selecting diverse compound subsets. *J. Chem. Inf. Comput. Sci.* **1999**, *39* (6), 1211–1225.
- (94) Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28* (1), 31–36.
- (95) Riniker, S.; Landrum, G. A. Better Informed Distance Geometry: Using What We Know To Improve Conformation Generation. *J. Chem. Inf. Model.* **2015**, *55* (12), 2562–2574.
- (96) Panapitiya, G.; Girard, M.; Hollas, A.; Sepulveda, J.; Murugesan, V.; Wang, W.; Saldanha, E. Evaluation of Deep Learning Architectures for Aqueous Solubility Prediction. *ACS Omega* **2022**, *7* (18), 15695–15710.
- (97) Ge, K.; Ji, Y. Novel Computational Approach by Combining Machine Learning with Molecular Thermodynamics for Predicting Drug Solubility in Solvents. *Ind. Eng. Chem. Res.* **2021**, *60* (25), 9259–9268.
- (98) Wyttenbach, N.; Niederquell, A.; Kuentz, M. Machine Estimation of Drug Melting Properties and Influence on Solubility Prediction. *Mol. Pharmaceutics* **2020**, *17* (7), 2660–2671.
- (99) Hewitt, M.; Cronin, M. T. D.; Enoch, S. J.; Madden, J. C.; Roberts, D. W.; Dearden, J. C. In Silico Prediction of Aqueous Solubility: The Solubility Challenge. *J. Chem. Inf. Model.* **2009**, *49* (11), 2572–2587.
- (100) Zhou, Z.; Kearnes, S.; Li, L.; Zare, R. N.; Riley, P. Optimization of Molecules via Deep Reinforcement Learning (vol 39, pg 532, 2019). *Sci. Rep.* **2020**, DOI: 10.1038/s41598-020-66840-x.
- (101) Wieder, O.; Kohlbacher, S.; Kuenemann, M.; Garon, A.; Ducrot, P.; Seidel, T.; Langer, T. A compact review of molecular property prediction with graph neural networks. *Drug Discovery Today Technol.* **2020**, *37*, 1–12.
- (102) Delaney, J. S. ESOL: Estimating aqueous solubility directly from molecular structure. *J. Chem. Inf. Comput. Sci.* **2004**, *44* (3), 1000–1005.
- (103) Huuskonen, J. Estimation of aqueous solubility for a diverse set of organic compounds based on molecular topology. *J. Chem. Inf. Comput. Sci.* **2000**, *40* (3), 773–777.
- (104) Gao, P.; Liu, Z.; Tan, Y.; Zhang, J.; Xu, L.; Wang, Y.; Jeong, S. Y. Accurate predictions of drugs aqueous solubility via deep learning tools. *J. Mol. Struct.* **2022**, *1249*, 131562.
- (105) Lee, S.; Lee, M.; Gyak, K. W.; Kim, S. D.; Kim, M. J.; Min, K. Novel Solubility Prediction Models: Molecular Fingerprints and Physicochemical Features vs Graph Convolutional Neural Networks. *ACS Omega* **2022**, *7* (14), 12268–12277.
- (106) Ahmad, W.; Tayara, H.; Chong, K. T. Attention-Based Graph Neural Network for Molecular Solubility Prediction. *ACS Omega* **2023**, *8* (3), 3236–3244.
- (107) Gao, P.; Zhang, J.; Sun, Y.; Yu, J. Accurate predictions of aqueous solubility of drug molecules via the multilevel graph convolutional network (MGCN) and SchNet architectures. *Phys. Chem. Chem. Phys.* **2020**, *22* (41), 23766–23772.
- (108) Nigsch, F.; Bender, A.; van Buuren, B.; Tissen, J.; Nigsch, E.; Mitchell, J. B. O. Melting point prediction employing k-nearest neighbor algorithms and genetic parameter optimization. *J. Chem. Inf. Model.* **2006**, *46* (6), 2412–2422.
- (109) Bergstrom, C. A. S.; Norinder, U.; Luthman, K.; Artursson, P. Molecular Descriptors influencing melting point and their role in classification of solid drugs. *J. Chem. Inf. Comput. Sci.* **2003**, *43* (4), 1177–1185.
- (110) Karthikeyan, M.; Glen, R. C.; Bender, A. General melting point prediction based on a diverse compound data set and artificial neural networks. *J. Chem. Inf. Model.* **2005**, *45* (3), 581–590.
- (111) Johnson, J. L. H.; Yalkowsky, S. H. Two new parameters for predicting the entropy of melting: Eccentricity (epsilon) and spirality (mu). *Ind. Eng. Chem. Res.* **2005**, *44* (19), 7559–7566.
- (112) Modarresi, H.; Dearden, J. C.; Modarress, H. QSPR correlation of melting point for drug compounds based on different sources of molecular descriptors. *J. Chem. Inf. Model.* **2006**, *46* (2), 930–936.
- (113) Hughes, L. D.; Palmer, D. S.; Nigsch, F.; Mitchell, J. B. O. Why are some properties more difficult to predict than others? A study of QSPR models of solubility, melting point, and Log P. *J. Chem. Inf. Model.* **2008**, *48* (1), 220–232.
- (114) Lazzus, J. A. Neural Network Based on Quantum Chemistry for Predicting Melting Point of Organic Compounds. *Chin. J. Chem. Phys.* **2009**, *22* (1), 19–26.
- (115) Bhatarai, B.; Teetz, W.; Liu, T.; Oberg, T.; Jeliakova, N.; Kochev, N.; Pukalov, O.; Tetko, I. V.; Kovarich, S.; Papa, E. CADASTER QSPR Models for Predictions of Melting and Boiling Points of Perfluorinated Chemicals. *Mol. Inform.* **2011**, *30* (2–3), 189–204.
- (116) Guendouzi, A.; Mekelleche, S. M. Prediction of the melting points of fatty acids from computed molecular descriptors: A quantitative structure-property relationship study. *Chem. Phys. Lipids* **2012**, *165* (1), 1–6.
- (117) 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.
- (118) Bejagam, K. K.; Lalonde, J.; Iverson, C. N.; Marrone, B. L.; Pilania, G. Machine Learning for Melting Temperature Predictions and Design in Polyhydroxyalkanoate-Based Biopolymers. *J. Phys. Chem. B* **2022**, *126* (4), 934–945.
- (119) Gorania, M.; Seker, H.; Haris, P. I. Predicting a protein's melting temperature from its amino acid sequence. In *32nd Annual International Conference of the IEEE Engineering-in-Medicine-and-Biology-Society (EMBC 10)*, Buenos Aires, Argentina, Aug 30–Sep 04, 2010; pp 1820–1823. DOI: 10.1109/iembs.2010.5626421.
- (120) Venkatram, S.; Batra, R.; Chen, L.; Kim, C.; Shelton, M.; Ramprasad, R. Predicting Crystallization Tendency of Polymers Using Multifidelity Information Fusion and Machine Learning. *J. Phys. Chem. B* **2020**, *124* (28), 6046–6054.
- (121) Pilania, G.; Iverson, C. N.; Lookman, T.; Marrone, B. L. Machine-Learning-Based Predictive Modeling of Glass Transition

Temperatures: A Case of Polyhydroxyalkanoate Homopolymers and Copolymers. *J. Chem. Inf. Model.* **2019**, *59* (12), 5013–5025.

(122) Bejagam, K. K.; Lalonde, J.; Iverson, C. N.; Marrone, B. L.; Pilania, G. Machine Learning for Melting Temperature Predictions and Design in Polyhydroxyalkanoate-Based Biopolymers. *J. Phys. Chem. B* **2022**, *126* (4), 934–945.

(123) Gamidi, R. K. R.; Ukrainczyk, M.; Zeglinski, J.; Rasmuson, A. C. Prediction of melting point of the co-crystals: ANN models. *Acta Crystallogr. A Found. Adv.* **2017**, *73*, C780–C780.

(124) Zhang, H.; Xiang, M. L.; Ma, C. Y.; Huang, Q.; Li, W.; Xie, Y.; Wei, Y. Q.; Yang, S. Y. Three-class classification models of logS and logP derived by using GA-CG-SVM approach. *Mol. Divers* **2009**, *13* (2), 261–268.

(125) Hildebrandt, C.; Gopireddy, S. R.; Fritsch, A. K.; Profitlich, T.; Scherliess, R.; Urbanetz, N. A. Evaluation and prediction of powder flowability in pharmaceutical tableting. *Pharm. Dev. Technol.* **2019**, *24* (1), 35–47.

(126) Barjat, H.; Checkley, S.; Chitu, T.; Dawson, N.; Farshchi, A.; Ferreira, A.; Gamble, J.; Leane, M.; Mitchell, A.; Morris, C.; et al. Demonstration of the Feasibility of Predicting the Flow of Pharmaceutically Relevant Powders from Particle and Bulk Physical Properties. *Journal of Pharmaceutical Innovation* **2021**, *16* (1), 181–196.

(127) Carman, P. C. Fluid flow through granular beds. *Chem. Eng. Res. Des.* **1997**, *75*, S32–S48, DOI: 10.1016/S0263-8762(97)80003-2.

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

(129) Perini, G.; Avendaño, C.; Hicks, W.; Parsons, A. R.; Vetter, T. Predicting filtration of needle-like crystals: A Monte Carlo simulation study of polydisperse packings of spherocylinders. *Chem. Eng. Sci.* **2021**, *230*, 116151.

(130) Su, Z. N.; He, J. X.; Zhou, P. P.; Huang, L.; Zhou, J. H. A high-throughput system combining microfluidic hydrogel droplets with deep learning for screening the antisolvent-crystallization conditions of active pharmaceutical ingredients. *Lab Chip* **2020**, *20* (11), 1907–1916.

(131) Simon, L. L.; Nagy, Z. K.; Hungerbühler, K. Comparison of external bulk video imaging with focused beam reflectance measurement and ultra-violet visible spectroscopy for metastable zone identification in food and pharmaceutical crystallization processes. *Chem. Eng. Sci.* **2009**, *64* (14), 3344–3351.

(132) Qin, J. P.; Zhang, Y.; Zhou, H.; Yu, F.; Sun, B.; Wang, Q. S. Protein Crystal Instance Segmentation Based on Mask R-CNN. *Crystals* **2021**, *11* (2), 157.

(133) He, J. X.; Zhou, J. H.; Dong, J. P.; Su, Z. N.; Huang, L. Revealing the effects of microwell sizes on the crystal growth kinetics of active pharmaceutical ingredients by deep learning. *Chem. Eng. J.* **2022**, *428*, 131986.

(134) Gao, Z. G.; Wu, Y. Y.; Bao, Y.; Gong, J. B.; Wang, J. K.; Rohani, S. Image Analysis for In-line Measurement of Multidimensional Size, Shape, and Polymorphic Transformation of L-Glutamic Acid Using Deep Learning-Based Image Segmentation and Classification. *Cryst. Growth Des.* **2018**, *18* (8), 4275–4281.

(135) Ma, Y. M.; Sun, M. M.; Liu, Y. B.; Chen, M. Y.; Wu, S. G.; Wang, M. W.; Wang, L. Y.; Gao, Z. G.; Han, D. D.; Liu, L. D.; et al. Design of Spherical Crystallization of Active Pharmaceutical Ingredients via a Highly Efficient Strategy: From Screening to Preparation. *Acs Sustainable Chemistry & Engineering* **2021**, *9* (27), 9018–9032.

(136) Ma, Y. M.; Niu, Y.; Yang, H. Y.; Dai, J. Y.; Lin, J. W.; Wang, H. Q.; Wu, S. G.; Yin, Q. X.; Zhou, L.; Gong, J. B. Prediction and design of cyclodextrin inclusion complexes formation via machine learning-based strategies. *Chem. Eng. Sci.* **2022**, *261*, 117946.

(137) Wicker, J. G. P.; Cooper, R. I. Will it crystallise? Predicting crystallinity of molecular materials. *CrystEngComm* **2015**, *17* (9), 1927–1934.

(138) Pereira, F. Machine learning methods to predict the crystallization propensity of small organic molecules. *CrystEngComm* **2020**, *22* (16), 2817–2826.

(139) Hosokawa, K.; Goto, J.; Hirayama, N. Prediction of solvents suitable for crystallization of small organic molecules. *Chem. Pharm. Bull.* **2005**, *53* (10), 1296–1299.

(140) Ghosh, A.; Louis, L.; Arora, K. K.; Hancock, B. C.; Krzyzaniak, J. F.; Meenan, P.; Nakhmanson, S.; Wood, G. P. F. Assessment of machine learning approaches for predicting the crystallization propensity of active pharmaceutical ingredients. *CrystEngComm* **2019**, *21* (8), 1215–1223.

(141) Pillong, M.; Marx, C.; Piechon, P.; Wicker, J. G. P.; Cooper, R. I.; Wagner, T. A publicly available crystallisation data set and its application in machine learning. *CrystEngComm* **2017**, *19* (27), 3737–3745.

(142) Walsh, G. Biopharmaceutical benchmarks 2018. *Nat. Biotechnol.* **2018**, *36* (12), 1136–1145.

(143) Basu, S. K.; Govardhan, C. P.; Jung, C. W.; Margolin, A. L. Protein crystals for the delivery of biopharmaceuticals. *Expert Opinion on Biological Therapy* **2004**, *4* (3), 301–317. Yang, M. X.; Shenoy, B.; Distler, M.; Patel, R.; McGrath, M.; Pechenov, S.; Margolin, A. L. Crystalline monoclonal antibodies for subcutaneous delivery. *Proc. Natl. Acad. Sci. U. S. A.* **2003**, *100* (12), 6934–6939. Stolnik, S.; Shakesheff, K. Formulations for delivery of therapeutic proteins. *Biotechnol. Lett.* **2009**, *31* (1), 1–11.

(144) Judge, R. A.; Johns, M. R.; White, E. T. Protein purification by bulk crystallization: The recovery of ovalbumin. *Biotechnol. Bioeng.* **1995**, *48* (4), 316–323.

(145) Howard, E. I.; Fernandez, J. M.; Garcia-Ruiz, J. M. On the Mixing of Protein Crystallization Cocktails. *Cryst. Growth Des.* **2009**, *9* (6), 2707–2712.

(146) Kawabata, K.; Saitoh, K.; Takahashi, M.; Sugahara, M.; Asama, H.; Mishima, T.; Miyano, M. Integrated state evaluation for the images of crystallization droplets utilizing linear and nonlinear classifiers. *Acta Crystallographica Section D-Biological Crystallography* **2006**, *62*, 1066–1072.

(147) Buchala, S.; Wilson, J. C. Improved classification of crystallization images using data fusion and multiple classifiers. *Acta Crystallographica Section D-Structural Biology* **2008**, *64*, 823–833.

(148) Bruno, A. E.; Charbonneau, P.; Newman, J.; Snell, E. H.; So, D. R.; Vanhoucke, V.; Watkins, C. J.; Williams, S.; Wilson, J. Classification of crystallization outcomes using deep convolutional neural networks. *PLoS One* **2018**, *13* (6), e0198883.

(149) Edwards, D. W.; Dinc, I. Classification of Protein Crystallization Images using EfficientNet with Data Augmentation. In *CSBio '20: Proceedings of the Eleventh International Conference on Computational Systems-Biology and Bioinformatics*, Bangkok, Thailand, 2020.

(150) Sigdel, M.; Pusey, M. L.; Aygun, R. S. CrystPro: Spatiotemporal Analysis of Protein Crystallization Images. *Cryst. Growth Des.* **2015**, *15* (11), S254–S262.

(151) Huang, L.; Yang, D. Y.; Yu, Z. M.; He, J. X.; Chen, Y.; Zhou, J. H. Deep Learning-Aided High-Throughput Screening of Time-Resolved Protein Crystallization on Programmable Microliter-Droplet Systems. *Chem. Eng. J.* **2022**, *450*, 138267.

(152) Babnigg, G.; Joachimiak, A. Predicting protein crystallization propensity from protein sequence. *Journal of Structural and Functional Genomics* **2010**, *11* (1), 71–80.

(153) Smialowski, P.; Schmidt, T.; Cox, J.; Kirschner, A.; Frishman, D. Will my protein crystallize? A sequence-based predictor. *Proteins: Struct., Funct., Bioinf.* **2006**, *62* (2), 343–355.

(154) Overton, I. M.; Barton, G. J. A normalised scale for structural genomics target ranking: The OB-Score. *FEBS Lett.* **2006**, *580* (16), 4005–4009.

(155) Chen, K.; Kurgan, L.; Rahbari, M. Prediction of protein crystallization using collocation of amino acid pairs. *Biochem. Biophys. Res. Commun.* **2007**, *355* (3), 764–769.

(156) Kurgan, L.; Razib, A. A.; Aghakhani, S.; Dick, S.; Mizianty, M.; Jahandideh, S. CRYSTALP2: sequence-based protein crystallization propensity prediction. *BMC Struct. Biol.* **2009**, *9*, 50.

(157) Zucker, F. H.; Stewart, C.; dela Rosa, J.; Kim, J.; Zhang, L.; Xiao, L. R.; Ross, J.; Napuli, A. J.; Mueller, N.; Castaneda, L. J.; et al.

- Prediction of protein crystallization outcome using a hybrid method. *J. Struct. Biol.* **2010**, *171* (1), 64–73.
- (158) Mizianty, M. J.; Kurgan, L. A. CRYSpred: Accurate Sequence-Based Protein Crystallization Propensity Prediction Using Sequence-Derived Structural Characteristics. *Protein Peptide Lett.* **2012**, *19* (1), 40–49.
- (159) Charoenkwan, P.; Shoombuatong, W.; Lee, H. C.; Chaijaruwanich, J.; Huang, H. L.; Ho, S. Y. SCMCrys: Predicting Protein Crystallization Using an Ensemble Scoring Card Method with Estimating Propensity Scores of P-Collocated Amino Acid Pairs. *PLoS One* **2013**, *8* (9), e72368.
- (160) Wang, H. L.; Feng, L. B.; Zhang, Z. D.; Webb, G. I.; Lin, D. H.; Song, J. N. CrysAlis: an integrated server for computational analysis and design of protein crystallization. *Sci. Rep.* **2016**, DOI: [10.1038/srep21383](https://doi.org/10.1038/srep21383).
- (161) Wang, H.; Wang, M.; Tan, H.; Li, Y.; Zhang, Z.; Song, J. PredPPCrys: Accurate Prediction of Sequence Cloning, Protein Production, Purification and Crystallization Propensity from Protein Sequences Using Multi-Step Heterogeneous Feature Fusion and Selection. *PLoS One* **2014**, *9* (8), No. e105902.
- (162) Hu, J.; Han, K.; Li, Y.; Yang, J. Y.; Shen, H. B.; Yu, D. J. TargetCrys: protein crystallization prediction by fusing multi-view features with two-layered SVM. *Amino Acids* **2016**, *48* (11), 2533–2547.
- (163) Jahandideh, S.; Jaroszewski, L.; Godzik, A. Improving the chances of successful protein structure determination with a random forest classifier. *Acta Crystallographica Section D-Structural Biology* **2014**, *70*, 627–635.
- (164) Elbasir, A.; Moovarkumudalvan, B.; Kunji, K.; Kolatkar, P. R.; Mall, R.; Bensmail, H. DeepCrystal: a deep learning framework for sequence-based protein crystallization prediction. *Bioinformatics* **2019**, *35* (13), 2216–2225.
- (165) Overton, I. M.; van Niekerk, C. A. J.; Barton, G. J. XANNpred: Neural nets that predict the propensity of a protein to yield diffraction-quality crystals. *Proteins-Structure Function and Bioinformatics* **2011**, *79* (4), 1027–1033.
- (166) Wang, D.; Yang, Z.; Zhu, B.; Mei, X.; Luo, X. Machine-Learning-Guided Cocystal Prediction Based on Large Data Base. *Cryst. Growth Des.* **2020**, *20* (10), 6610–6621.
- (167) Wengert, S.; Csanyi, G.; Reuter, K.; Margraf, J. T. Data-efficient machine learning for molecular crystal structure prediction. *Chem. Sci.* **2021**, *12* (12), 4536–4546.
- (168) Gorbitz, C. H. The development and use of a crystallographic database. *Acta Crystallogr. Sect. B Struct. Sci. Cryst. Eng. Mater.* **2016**, *72*, 167–168.
- (169) van de Streek, J.; Motherwell, S. Searching the Cambridge Structural Database for polymorphs. *Acta Crystallogr. Sect. B Struct. Sci. Cryst. Eng. Mater.* **2005**, *61*, 504–510.
- (170) Zagorac, D.; Muller, H.; Ruehl, S.; Zagorac, J.; Rehme, S. Recent developments in the Inorganic Crystal Structure Database: theoretical crystal structure data and related features. *J. Appl. Crystallogr.* **2019**, *52* (5), 918–925.
- (171) Shinada, N. K.; Schmidtke, P.; de Brevern, A. G. Accurate Representation of Protein-Ligand Structural Diversity in the Protein Data Bank (PDB). *Int. J. Mol. Sci.* **2020**, *21* (6), 2243.
- (172) Grazulis, S.; Chateigner, D.; Downs, R. T.; Yokochi, A. F. T.; Quiros, M.; Lutterotti, L.; Manakova, E.; Butkus, J.; Moeck, P.; Le Bail, A. Crystallography Open Database - an open-access collection of crystal structures. *J. Appl. Crystallogr.* **2009**, *42*, 726–729.
- (173) Robertson, S. Mogul: Rapid Retrieval of Molecular Geometry Information From a Crystallographic Database. *Acta Crystallogr. A Found. Crystallogr.* **2004**, *60*, S96–S96.
- (174) 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.
- (175) Heng, T.; Yang, D.; Wang, R.; Zhang, L.; Lu, Y.; Du, G. Progress in Research on Artificial Intelligence Applied to Polymorphism and Cocystal Prediction. *ACS Omega* **2021**, *6* (24), 15543–15550.
- (176) Wicker, J. G. P.; Crowley, L. M.; Robshaw, O.; Little, E. J.; Stokes, S. P.; Cooper, R. I.; Lawrence, S. E. Will they co-crystallize? *CrystEngComm* **2017**, *19* (36), 5336–5340.
- (177) Galek, P. T. A.; 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* (7), 2391–2403.
- (178) Cheng, G.; Gong, X. G.; Yin, W. J. Crystal structure prediction by combining graph network and optimization algorithm. *Nat. Commun.* **2022**, *13* (1), 1492.
- (179) Dudek, M. K.; Druzbicki, K. Along the road to crystal structure prediction (CSP) of pharmaceutical-like molecules. *CrystEngComm* **2022**, *24* (9), 1665–1678.
- Ryan, K.; Lengyel, J.; Shatruk, M. Crystal Structure Prediction via Deep Learning. *J. Am. Chem. Soc.* **2018**, *140* (32), 10158–10168.
- (180) Sun, G.; Liu, X.; Abramov, Y. A.; Nilsson Lill, S. O.; Chang, C.; Burger, V.; Broo, A. Current State-of-the-art In-house and Cloud-Based Applications of Virtual Polymorph Screening of Pharmaceutical Compounds: A Challenging Case of AZD1305. *Cryst. Growth Des.* **2021**, *21* (4), 1972–1983.
- (181) Kusaba, M.; Liu, C.; Yoshida, R. Crystal structure prediction with machine learning-based element substitution. *Comput. Mater. Sci.* **2022**, *211*, 111496.
- (182) Loschen, C.; Klamt, A. Solubility prediction, solvate and cocystal screening as tools for rational crystal engineering. *J. Pharm. Pharmacol.* **2015**, *67* (6), 803–811.
- (183) Barua, H.; Gunnam, A.; Yadav, B.; Nangia, A.; Shastri, N. R. An ab initio molecular dynamics method for cocystal prediction: validation of the approach. *CrystEngComm* **2019**, *21* (47), 7233–7248.
- (184) Przybylek, M.; Cysewski, P. Distinguishing Cocystals from Simple Eutectic Mixtures: Phenolic Acids as Potential Pharmaceutical Coformers. *Cryst. Growth Des.* **2018**, *18* (6), 3524–3534.
- (185) Kim, P.; Lee, I.-S.; Kim, J.-Y.; Mswahili, M. E.; Jeong, Y.-S.; Yoon, W.-J.; Yun, H.-S.; Lee, M.-J.; Choi, G. J. A study to discover novel pharmaceutical cocystals of pelubipirofen with a machine learning approach compared. *CrystEngComm* **2022**, *24* (21), 3938–3952.
- (186) Hao, Y.; Hung, Y. C.; Shimoyama, Y. Investigating Spatial Charge Descriptors for Prediction of Cocystal Formation Using Machine Learning Algorithms. *Cryst. Growth Des.* **2022**, *22*, 6608.
- (187) Jiang, Y.; Yang, Z.; Guo, J.; Li, H.; Liu, Y.; Guo, Y.; Li, M.; Pu, X. Coupling complementary strategy to flexible graph neural network for quick discovery of coformer in diverse co-crystal materials. *Nat. Commun.* **2021**, DOI: [10.1038/s41467-021-26226-7](https://doi.org/10.1038/s41467-021-26226-7).
- (188) Przybylek, M.; Cysewski, P. Distinguishing Cocystals from Simple Eutectic Mixtures: Phenolic Acids as Potential Pharmaceutical Coformers. *Cryst. Growth Des.* **2018**, *18* (6), 3524–3534.
- (189) Mswahili, M. E.; Lee, M.-J.; Martin, G. L.; Kim, J.; Kim, P.; Choi, G. J.; Jeong, Y.-S. Cocystal Prediction Using Machine Learning Models and Descriptors. *Appl. Sci.* **2021**, *11* (3), 1323.
- (190) Hao, Y.; Hung, Y. C.; Shimoyama, Y. Investigating Spatial Charge Descriptors for Prediction of Cocystal Formation Using Machine Learning Algorithms. *Cryst. Growth Des.* **2022**, *22* (11), 6608–6615.
- (191) Jiang, Y.; Yang, Z.; Guo, J.; Li, H.; Liu, Y.; Guo, Y.; Li, M.; Pu, X. Coupling complementary strategy to flexible graph neural network for quick discovery of coformer in diverse co-crystal materials. *Nat. Commun.* **2021**, *12* (1), 5950.