Digital Chemical Engineering 11 (2024) 100150

Contents lists available at ScienceDirect

Digital Chemical Engineering

journal homepage: www.elsevier.com/locate/dche

Case Studies

OpenCrystalData: An open-access particle image database to facilitate learning, experimentation, and development of image analysis models for crystallization processes.

Yash Barhate [a,1], Christopher Boyle [b,c,1], Hossein Salami [d,1,*], Wei-Lee Wu [a,1], Nina Taherimakhsousi [e], Charlie Rabinowitz [e], Andreas Bommarius [d], Javier Cardona [b,c,f], Zoltan K. Nagy [a], Ronald Rousseau [d], Martha Grover [d]

[a] Davidson School of Chemical Engineering, Purdue University, West Lafayette, IN, US
[b] CMAC EPSRC Future Manufacturing Research Hub, University of Strathclyde, Glasgow, UK
[c] Department of Chemical and Process Engineering, University of Strathclyde, Glasgow, UK
[d] School of Chemical & Biomolecular Engineering, Georgia Institute of Technology, Atlanta, GA, USA
[e] Mettler Toledo AutoChem, Inc., Columbia, MD 21046, USA
[f] Department of Electronic and Electrical Engineering, University of Strathclyde, Glasgow, UK

ARTICLE INFO

Keywords:
Crystallization
Process analytical technology
Imaging
Open-access database
Machine learning

ABSTRACT

Imaging and image-based process analytical technologies (PAT) have revolutionized the design, development, and operation of crystallization processes, providing greater process understanding through the characterization of particle size, shape and crystallization mechanisms in real-time. The performance of corresponding PAT models, including machine learning/artificial intelligence (ML/AI)-based approaches, is highly reliant on the data quality used for training or validation. However, acquiring high quality data is often time consuming and a major roadblock in developing image analysis models for crystallization processes.

To address the lack of diverse, high-quality, and publicly available particle image datasets, this paper presents an initiative to create an open-access crystallization-related image database: OpenCrystalData (OCD, at www.kaggle.com/opencrystaldata/datasets). The datasets consist of images from different crystallization systems with different particle sizes and shapes captured under various conditions. The initial release consists of four different datasets, addressing the estimation of particle size distribution using *in-situ* images for different categories of particles and detection of anomalous particles for process monitoring purposes. Images are collected using various instruments, followed by case-specific processing steps, such as ground-truth labeling and particle size characterization using offline microscopy. Datasets are released on the online collaborative platform Kaggle, along with specific guidelines for each dataset. These datasets are aimed to serve as a resource for researchers to enable learning, experimentation, development, and evaluation and comparison of different analytical approaches and algorithms. Another goal of this initiative is to encourage researchers to contribute new datasets focusing on various systems and problem statements. Ultimately, OpenCrystalData is intended to facilitate and inspire new developments in imaging-based PAT for crystallization processes, encouraging a shift from time-consuming offline analysis towards comprehensive real-time process insights that drive product quality.

Introduction

Crystallization process has applications in pharmaceutical, food, and chemical industries where it is used to produce crystals with desired properties, such as purity, shape, and crystal size distribution (CSD). (Arruda et al., 2023; McGinty et al., 2020) Among different process analytical technology (PAT) tools, image-based PATs have recently gained popularity in facilitating rapid crystallization process design, and process monitoring and control. The growing interest in using these tools is mainly because of their ability to function as an

* Corresponding author.
E-mail address: HSalami3@Gatech.edu (H. Salami).
[1] Authors have equal contribution.

https://doi.org/10.1016/j.dche.2024.100150
Received 16 February 2024; Received in revised form 29 March 2024; Accepted 30 March 2024
Available online 4 April 2024
2772-5081/© 2024 The Authors. Published by Elsevier Ltd on behalf of Institution of Chemical Engineers (IChemE). This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

information-gathering tool, supporting process understanding and robustness assessment. Furthermore, using image-based PAT tools can facilitate process design and control by replacing traditional off-line measurement-based control with an on-line monitoring-based control strategy. In particular, in-situ imaging offers the potential for real-time CSD characterization and measurement, which can be directly utilized in a feedback control loop, enabling adjustment of process parameters in real-time to maintain the desired crystal quality outcomes. (Simon et al., 2014; Yu et al., 2004; Barrett et al., 2005; Nagy and Braatz, 2012; Nagy et al., 2013)

In-situ imaging probes (e.g., Particle Vision and Measurement or EasyViewer from Mettler Toledo, Blaze Micro from Blaze Metrics, etc.) can collect significant amounts of real-time image data from a crystallization process. This data can provide both qualitative and quantitative insights into the underlying processes by tracking the onset and progression of different crystallization mechanisms, enable monitoring critical quality attributes of products, characterize steady states or equilibration, and enable implementation of external feedback control loops for process control. (E. Simone et al., 2015; E. Simone et al., 2015; Wu et al., 2022; Agimelen et al., 2016; Borsos et al., 2017) Moreover, there is a growing interest in utilizing *in-situ* imaging data to assist the development of population balance models and estimation of crystallization kinetic parameters. (Szilagyi et al., 2020; B. Szilágyi et al., 2022; B. Szilágyi et al., 2022; Barhate et al., 2024)

To maximize the information gained from each experiment during crystallization process development and to aid the aforementioned efforts, there is a need for image analysis models that transform raw image data collected during experiments to actionable insight by extracting qualitative and quantitative information. (Xiouras et al., 2022) Various image analysis models, including a large number of machine learning (ML)-based models trained using data collected from *in-situ* PAT tools have been developed for different crystallization-related tasks including, online process monitoring and impurity detection, (Salami et al., 2023; Salami et al., 2021; Tachtatzis et al., 2015) real-time particle size and shape characterization, (Jaeggi et al., 2021; Salami and Skomski, 2023; Manee et al., 2019) and to control crystal quality attributes by manipulating process parameters. (Öner et al., 2020; Irizarry et al., 2017; Wu and Wu, 2023; Benyahia et al., 2021; de Moraes et al., 2023) However, in most cases, the models developed are specific to the system in consideration. This limited applicability is either due to selecting specific hyper parameters or processing workflows or the data-driven models being trained on experimental datasets collected for specific case studies.

To enable model development and experimentation efforts and enhance the overall capabilities *in-situ* crystallization image analysis, access to large and information-rich datasets is necessary. However, despite the abundance of experimental data being collected in various academic and industrial settings, access to high quality data, annotated and validated by human experts, or offline particle size and shape characterization tools is limited. Typically, individual labs conduct specific experimental campaigns and build isolated datasets and models. While the developed image analysis model might be introduced to the public through publications or informal communications, the experimental data underlying the models are usually not shared and remain on local hard drives or access-restricted data repositories. (Xiouras et al., 2022)

In this regard, the collaborative effort described in this manuscript envisions the creation of an open-access crystallization process image database, featuring a collection of high-quality, microscopic images captured across diverse crystallization systems and focusing on different crystallization-specific tasks such as particle size estimation or impurity detection. Depending on the task, sets of annotated images along with corresponding attributes are prepared, serving as the ground truth to enable easy and direct application of the datasets to evaluate, develop, and build different image analysis algorithms. The database, named OpenCrystalData, has been made accessible online through the well-known Kaggle data science community, and will be consistently maintained, fostering collaboration among researchers by enabling the shared use of datasets. The datasets can be searched on Kaggle via the name "OpenCrystalData" (www.kaggle.com/search?q=opencrystaldata). In general, datasets that are available on Kaggle contain key descriptors, data collection methodology, and metadata. Each dataset can be referenced with their own individual Digital Object Identifier (DOI). Users can import the desired datasets or algorithms from Kaggle and implement their own workflow on the cloud that allows for quick testing and ease of collaboration. The hope is that this initiative opens avenues to capitalize on the existing experimental data to support the advancement of algorithmic developments for image-based monitoring, design, and control of crystallization processes.

The OpenCrystalData (OCD) initiative draws inspiration from similar endeavors in different fields, including the creation of open-access databases such as the magnetic particle imaging database, (Knopp et al., 2020) the transmission electron microscopy image database for catalytic applications, (Nartova et al., 2022) and the chemical microparticle image database in multiphase flows. (Liu et al., 2023) Another notable example is the machine recognition of crystallization outcomes (MARCO) dataset (Bruno et al., 2018) that comprises of approximately half a million off line microscopic images of singular protein crystals curated together to facilitate the development of powerful ML-based image analysis algorithms for crystal recognition tasks. This diverse and large database was sourced from various organizations, including academic institutions and several pharmaceutical companies, fostering a collaborative environment, and benefitting both industrial and academic applications. (Bruno et al., 2018) While the microscopic images in the MARCO dataset were captured under ideal lighting conditions, it is essential to recognize that in-situ microscopic images obtained through PAT tools often exhibit significant variations. These variations arise due to factors such as focus, lighting, orientation, and particle density, which depend on the specific crystallization system. Consequently, the development of image analysis algorithms capable of universally extracting information from these images presents greater challenges compared to images resembling those in the MARCO dataset. The OpenCrystalData initiative aims to address this gap by hosting a collection of images, fostering the development of specialized image analysis algorithms for diverse crystallization tasks.

The current version of the database (Fig. 1) encompasses image datasets derived from two distinct crystallization systems and two standard systems including polystyrene and l-Glutamic acid particles for tackling various types of crystallization-related problems such as impurity detection (object classification), instance segmentation, and *in-situ* crystal size estimation. For each dataset, first, the necessary background and a general description of the dataset and the associated problem are provided. Subsequently, details related to the dataset generation such as the crystallization system employed, the instrumentation hardware utilized, and the methods employed for annotation and validation are discussed (Table 1). Fig. 2 shows some of the example image analysis workflows that can be employed to approach tasks related to the presented datasets. Lastly, details on how to access and get started with each dataset are presented along with a brief discussion and recommendations of some of the image analysis algorithms and modeling paradigms that could be used to approach the problem. While the volume and diversity of the datasets discussed herein, does not fully capture the vast heterogeneity observed in different crystallization systems, the problem types include the most important tasks related to crystallization process monitoring and characterization. The goal is to set task-specific standards and encourage the addition of new data by other members of the crystallization and particle engineering community.

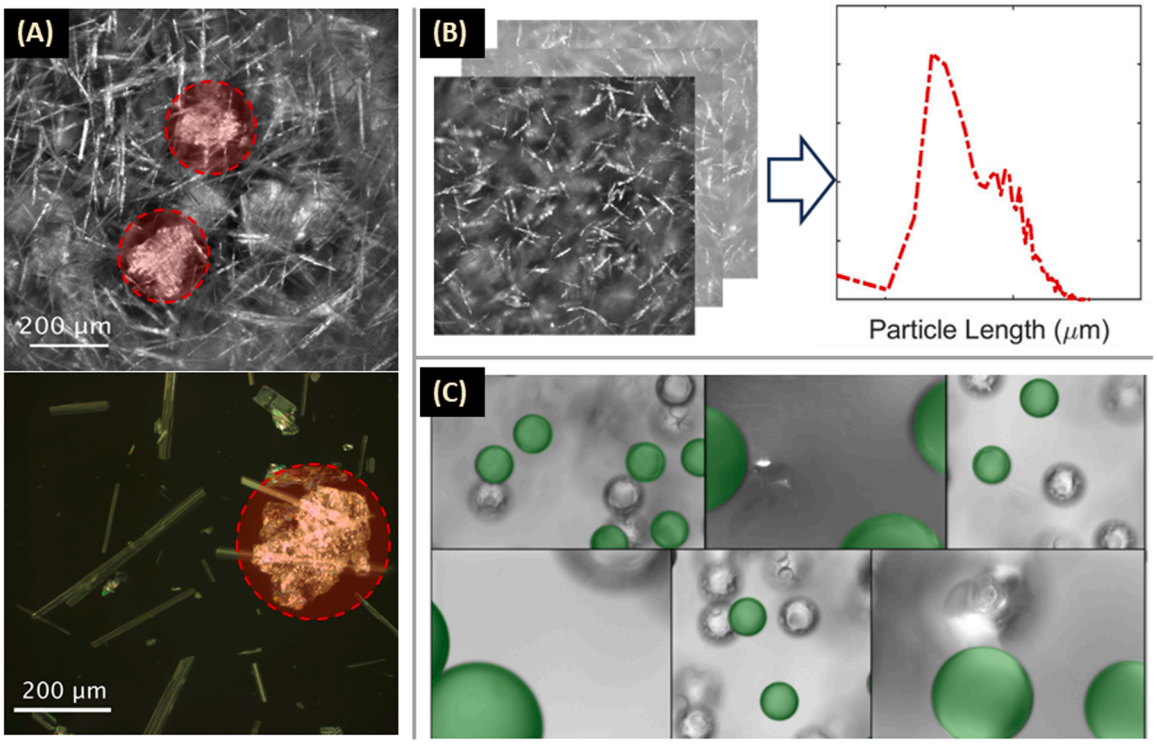

Fig. 1. Example images from three datasets in OpenCrystalData. (A) cephalexin reactive crystallization with phenylglycine impurities highlighted and circled with dashed lines, (B) needle-like particle images from an industrial agrochemical crystallization process used to build models that generate PSD data comparable with offline PSD, (C) images of polystyrene standard sphere particles with particle-level annotations shown by the green circles. Object-level annotations can be used to train deep-learning-based image analysis models.

Table 1
Information about the datasets in the first release of OCD on the Kaggle platform.

Name	Impurity in cephalexin reactive crystallization	AgCrystal images	Standard polystyrene microspheres	VANSIL® Wollastonite and l-Glutamic acid
Problem type	Morphology/anomaly detection	Particle Characterization	Particle Characterization	Particle Characterization
Number of images	400 Raw images, 6000 Cropped images	3888 images	2300 images	120 images
Microscopy type	*In-situ*	*In-situ*	*In-situ*	*In-situ*
PAT hardware	EasyViewer-100	EasyViewer-100	PVM v819	EasyViewer-100 EasyViewer-400 ParticleTrack-G400
Other metadata	–	Solids concentration loading and offline particle size distributions	Particle size distribution	Particle size and chord length distribution
DOI	10.34740/kaggle/dsv/6581298	10.34740/kaggle/dsv/6743111	10.34740/kaggle/dsv/6376944	10.34740/kaggle/dsv/7414048

Datasets description

Dataset-1: impurity in cephalexin reactive crystallization

This dataset highlights the use case of detecting the presence of impurity or undesired particles in a crystallization process. The dataset can be used to inform the development or evaluate the performance of image anomaly detection algorithms. It also presents an example to encourage further research into the application of image-based process monitoring.

Background: One of the most common applications of crystallization in chemical and pharmaceutical industries is for purification. Examples include crystallizing a specific, therapeutically active enantiomer, (Lorenz et al., 2006; Bredikhin and Bredikhina, 2017) or crystallizing a target product from a reaction solution at the end of a multi-step synthesis process involving multiple reactants and products. Ideally, the molecule of interest is the only crystallizing species in such a system resulting in a pure product upon filtering the slurry, and with an acceptable yield. However, there might be cases in which deviations in process parameters might lead to conditions under which other molecules in the system form a solid phase. Detecting the formation of this second solid phase is critical for process monitoring and quality purposes. Enzymatic reactive crystallization of cephalexin monohydrate is such a case, where a byproduct of the cephalexin synthesis reaction (phenylglycine, PG) has a low solubility in the reaction medium (water) and will precipitate if generated in excess amounts. (Salami et al., 2021)

Description: Images in this dataset are captured from a cephalexin monohydrate crystallization process (Fig. 1, Panel A). Briefly, a cephalexin monohydrate crystal slurry was prepared with a seeded batch crystallization process and using pH swings to generate supersaturation. Upon the completion of crystallization, a set of *in-situ* images were captured from the slurry using a Mettler-Toledo EasyViewer-100 probe (pure cephalexin monohydrate slurry, no byproduct crystals present). In the next step, phenylglycine crystals, pre-made using a pH swing

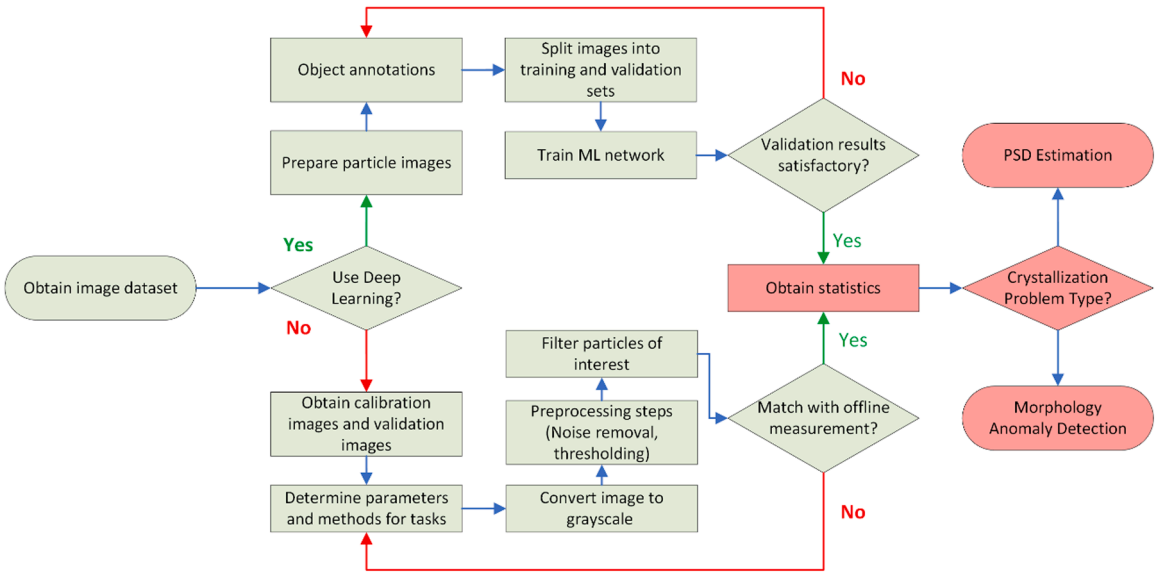

Fig. 2. Examples of the image analysis workflows that were used to tackle different crystallization problem types related to the presented datasets. Access to particle-level annotations facilitates training of deep learning-based segmentation models (top path). More conventional methods such as thresholding or watershed, can be applied in combination with a feedback loop from offline measurements to optimize model hyperparameters such as threshold level (bottom path). Other alternatives include use of pretrained general-purpose vision models such as SAM (Kirillov et al., 2023).

method, were added to the slurry in incremental amounts. *In-situ* images were captured from the slurry (impure slurry, phenylglycine crystals present). A complete description of the experiments and a discussion on potential impact of process conditions on the shape of impurity crystals can be found elsewhere. (Salami et al., 2021) This image dataset is divided into two parts of Raw and Cropped images corresponding to the treatment of the original image. Raw images are original images captured by the probe, and Cropped images are instances of cephalexin or phenylglycine crystals. Images in both raw and cropped folders are then divided into two sub-folders, cephalexin crystals and phenylglycine crystals corresponding to the target product and the byproduct. All phenylglycine crystal containing images were selected by a manual review process of images from the corresponding system and drawing a bounding box around the identified impurity particle.

Potential approaches: Depending on the application, the problem of monitoring a crystallization process for detecting impurities can be approached from two angles. If the impurity to be controlled for is known and has a fixed morphology one might treat the problem as a binary classification with two classes of target and impurity crystal images. This approach might require taking necessary steps to address the possibly imbalanced dataset as in most applications it is likely that most image datapoints are from the target class. More generally however, the problem can be formulated as an anomaly detection task in which impurity crystals (ideally, particles with morphologies significantly different than that of target crystals' distribution) images are to be identified.

Dataset-2: agcrystal images

This dataset was used to calibrate an image segmentation algorithm for the analysis of *in-situ* microscopy images, and the results were compared with the measurements obtained from offline microscopy.

Background: Monitoring and control of crystal size and shape during crystallization is critical in various industrial applications (agrochemicals, pharmaceuticals, or energetics). Lack of control in particle size can cause poor final product performance, manifesting as failures in product formulation or dissolution testing, and negatively impact downstream processing such as product filtration, washing, and drying. (MacLeod and Muller, 2012) Traditionally, measurements are obtained offline after slurry filtration, washing, and drying. While sampling during the crystallization process can improve data scarcity, oversampling of the crystallization slurry can impact the crystallization kinetics and process. Unprocessed *in-situ* images can provide qualitative information throughout the process. An adaptive image analysis algorithm can be developed, and its hyperparameters fine-tuned as a function of solids loading to extract quantitative information (size distribution, shape, aspect ratio) from these images.

Description: Images in this dataset originated from an industrial agrochemical crystallization process with needle-like crystal particles (Fig. 1, Panel B). Three batches of the active ingredient with different crystal size distributions were used for the calibration and validation of an image analysis algorithm. Pre-weighed samples were suspended in a known volume of mother liquor to create slurries with known solids loading. Particles were imaged with an EasyViewer-100 in situ probe immersed in the vessel. Images were acquired starting at the lowest solids concentration and increased to higher concentrations by adding more solid material. A Morphologi G3 microscope was used to measure the particles off-line via image analysis and establish the ground truth values. The dataset contains a folder with images for model calibration and a folder with images for model validation. Each folder is divided into subfolders labeled with particle size and solids concentration associated with the images contained inside. The off-line measured particle size distributions, the ground truth, for all batches are compiled into a spreadsheet for reference.

Potential approaches: Extracting quantitative descriptors (length, width) of particles from an image is the main goal for this dataset. An image can be processed by an image analysis model with the goal of segmenting in-focus particles from the background noise such as light artifact, overlapping particles, and particles that are out-of-focus. Different background separation and elimination algorithms can be explored by the user to attain adequate segmentation, with image analysis software such as ImageJ, Matlab, Python packages (OpenCV, Scikit-Image, PIL). Basic methods from ImageJ, such as rolling ball, thresholding, and average background subtraction, which are computationally fast, can be tried at first. However, these methods fall short if the image background or illumination are varying. More complex algorithms such as MOG (Gaussian) and GMG (Bayesian) from OpenCV's background subtractor methods can address the variability in the image

background at the expense of more computational time and power needed. Afterwards, object detection algorithms can be implemented to detect particles of certain morphology, needle-like particles for this dataset, and the characteristics of particles can be obtained. Results obtained from analyzing the higher solids concentration images can be compared with lower solids concentration data as image segmentation is typically simpler to perform at lower concentrations. User can also choose to use particle size distribution obtained from an off-line image analysis instrument as the ground truth to evaluate or calibrate the image analysis algorithm as done in a previous work. (Wu et al., 2023)

Dataset-3: standard polystyrene microspheres

This dataset is a collection of microscope images taken of polystyrene standard sphere particles (Fig. 1, Panel C). These particles are of National Institute of Standards and Technology (NIST)-traceable size distribution, making them an excellent case for training and/or evaluating image analysis methods. Along with the images, particle location annotations are provided (*i.e.*, object masks). Together this forms a dataset suitable for training a deep learning model for object detection or segmentation. Furthermore, this dataset can be used to evaluate the performance of pre-trained foundational object detection models such as SAM (Kirillov et al., 2023) on example domain specific tasks.

Background: Particle characterization can be used for gathering data for process modelling. Off-line analysis is typically used to obtain samples of particle size, to which models such as population balance models are fit. *In-situ* imaging offers a higher sampling rate and has the potential of providing richer data for model fitting. However, image analysis of *in-situ* images is typically more challenging. (Lins et al., 2022) Deep learning-based models such as YOLO (ultralytics/yolov5), Mask R-CNN (He et al., 2020), and more recently SAM (Kirillov et al., 2023) are becoming popular for object detection and segmentation tasks, where an image is analyzed to give a list of objects found, including each object's location and border. In the case of object detection, a training set would be a set of images annotated with object locations (e.g., a bounding contour or pixel-mask). Such dataset can be used to train a deep learning model from scratch or fine-tune a pre-trained model for use with microscopic images.

Description: Standard Polystyrene particles were obtained from Fisher Scientific. Particles of different sizes (150, 300, 400, and 500 μm) at different solids concentrations (1, 2, 2.5, 3.3, and 5 wt% for all sizes; and for 400 μm size up to 12.5 wt%) were suspended in de-ionized water and were imaged using a Mettler-Toledo Particle Vision and Measurement (PVM v819) probe. Further details on the data collection can be referred elsewhere (Cardona et al., 2018). While many images were collected, a random subset was selected to make the size of the resulting dataset manageable. In each image the location of particles was annotated and stored in the provided JSON file, following the COCO Dataset format. (Lin et al., 2014) Annotations are in the form of a bounding contour. Particle locations were annotated using a combined approach of manual and automatic annotation. The first pass of machine-assisted annotation using a pre-trained model was subsequently checked by a human such that those annotations deemed inaccurate were refined manually. Manual annotation was achieved using the Computer Vision Annotation Tool (CVAT) (openvino/cvat). Particles were annotated by drawing a polygon around their border. If the border was unsharp, a best guess approach was employed to select the appropriate border for the particle. Particles which were out of focus were not annotated. Particles falling off the edge of the frame were annotated despite not being able to be sized. This is to provide object detection models with consistent information on what a particle edge looks like. Particles on the edge can be trivially filtered later before forming a PSD. Along with the annotated images, the ground truth particle size distribution given by the manufacturer is included in the dataset.

Potential approaches: A typical desired output of image analysis is the size distribution (or quantiles) of a population of particles. To answer this, the location of each particle in each image needs to be found and the particle dimensions need to be measured. The sizes can be histogrammed to form a PSD, if enough particles are counted for statistical robustness. Otherwise, quantiles or mean size can be obtained. A deep learning-based model such as Mask R-CNN (He et al., 2020) can be trained using the provided dataset to resolve particle size from the *in-situ* microscopy data. The obtained PSD can then be compared to the NIST-traceable PSD provided by the manufacturer. The results can be validated at two levels: segmentation-level evaluation is performed by comparing the results of image analysis against the ground-truth annotations (*i.e.*, checking whether the model detects the particles in an image accurately); PSD-level evaluation is made with reference to the manufacturer specification (*i.e.*, checking whether the obtained PSD matches the expected one).

Dataset-4: VANSIL® wollastonite and l-Glutamic acid particles

Background: In many areas such as the pharmaceutical industry product specifications are rapidly changing to become more precise. To enable this requires improving the manufacturing processes and techniques. (Bolla et al., 2022; Urwin et al., 2023) The improvement of manufacturing processes requires developing PATs to capture information concerning a process and enable process automation to reduce batch-to-batch variations. (Orehek et al., 2021) Robust process automation is highly dependent on *in-situ* monitoring systems. Therefore, developing robust monitoring systems is imperative for two main reasons. Acquiring real-time data from a process, providing real-time feedback required for process automation, and collecting data using PATs to obtain the required information for understanding different crystallization mechanisms. (Gao et al., 2021; Metherall et al., 2023) This dataset includes images from two different imaging instruments along with particle size and chord length data, enabling comparisons between different in situ particle characterization instruments.

Description: VANSIL® Wollastonite and l-Glutamic acid powders were purchased and used as received from Vanderbilt Minerals LLC and Sigma Aldrich, respectively. These two materials were used to prepare analogous suspensions of pharmaceutical particles, allowing us to create images of the particles in suspension. VANSIL® Wollastonite particles had a needle-like shape with high aspect ratios while l-Glutamic acid particles had low aspect ratios and were rod shaped. Water-based suspensions of VANSIL® Wollastonite with different weight percentages, and acetone-based suspension of l-Glutamic acid with 5 wt% was prepared. *In-situ* images of particles were captured using Mettler-Toledo EasyViewer-100 and Mettler-Toledo EasyViewer-400 probes. Concurrently, a ParticleTrack G400 instrument was used to measure the particle chord length distribution for all suspensions. Instrument default collection parameters were used throughout the data collection.

Potential approaches: The image data provided in this dataset can be used mostly in the same manner as outlined for the Dataset-2 above. Additionally, thanks to the availability of chord length distribution in this case, potential mappings between the image-based and reported particle size data and chord length distribution estimations by ParticleTrack instrument can be explored.

Discussion and outlook

Four datasets discussed above touch upon general crystallization image analysis problem types. While different image processing methods and tasks are encapsulated in this release, these datasets do not fully cover all problem types or crystallization systems. New potential additions to enrich the current database can be divided into a few categories:

Dataset-1 focuses on impurity detection and particle classification. Another related and equally important crystallization problem type is the detection of different crystal polymorphs in a vessel. Provided there are morphological differences, different forms can be detected as

different particle populations with distinct visual characteristics. The ratio of the polymorphs observed in a set of images can be used to estimate a polymorphic form ratio that can potentially be used for crystallization process control and modeling. Other relevant cases include image data from systems prone to agglomeration. Such dataset can enable building models to detect and inform on the extent of agglomeration in a system.

Dataset-2 and 4 include images of needle-like or rod-like particles along with off-line measurements of particle size as ground-truth. Similar data for new systems and more complicated shapes and morphologies, such as rod-like particles, can enable building more generalizable image analysis models, ultimately enabling real-time process control or improve calibration data for multi-dimensional crystallization population balance models.

Dataset-3 with images of standard spherical particles and similar cases enable the development of deep learning-based models for accurate image and instance segmentation. New additions can include image data to aid in training models to detect spherical droplets such as emulsion systems in active ingredient formulations or detection of oiling out phenomenon in crystallization systems. These models can be useful, for example, for measuring liquid-liquid phase separation points on the phase diagram.

Overall, the addition of new and diverse datasets in various categories of crystallization problem types and systems is necessary to enable experimentation and new developments. More importantly, building truly generalizable (unbiased, not overfit) machine learning-based image analysis models requires rich data of various particle shapes and morphologies that, most likely, cannot be supplied by a single experimental campaign.

The open-access nature of OpenCrystalData guarantees full transparency regarding the datasets utilized for training image analysis models. This transparency enables potential users to thoroughly assess the quality of datasets on the platform according to their specific requirements. OpenCrystalData also implements an internal review process focusing on both the integrity of primary data and the consistency and accuracy of the metadata, corresponding to each dataset. During the future expansions of the database, this quality review will be systematically applied to each new dataset to maintain the platform's standard for high-quality data.

Conclusion

In recent years, image-based process analytical technologies (PATs) have attracted significant attention in the crystallization community. Advancements in hardware, in the form of new imaging probes or specially designed processes to facilitate capture of high-quality images along with powerful image analysis models highlight the potential for extracting real-time and actionable insights. However, it is essential to recognize the inherent complexity of the problem, particularly when working with images under high solids loading conditions. Additionally, a significant obstacle in achieving this goal is the scarcity of high-quality image datasets with the necessary pre-processing, annotations, and ground-truth benchmarking, that is necessary for the training and enhancement of existing image analysis models. The availability of secondary data and metadata along with the primary image dataset is also crucial for a high-quality image dataset.

OpenCrystalData aims to tackle these challenges by promoting learning, experimentation, development, and evaluation of image analysis models through initiating a series of publicly available image datasets. The overarching goal of OpenCrystalData is to encourage efforts for building a comprehensive database that encompass a diverse array of tasks associated with the crystallization process monitoring and characterization including but not limited to particle size estimation, particle classification, and anomaly detection. By providing these open-access resources, it is envisaged that this initiative stands as a catalyst for advancing the field of image-based crystallization process monitoring and development.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Nina Tahermakhsousi reports a relationship with Mettler-Toledo Autochem that includes: employment. Charlie Rabinowitz reports a relationship with Mettler-Toledo Autochem that includes: employment. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

CB and JC would like to thank the EPSRC Future Continuous Manufacturing and Advanced Crystallization Research Hub (Grant ref: EP/P006965/1) for funding part of the study associated with dataset-3. CB and JC would like to acknowledge that part of this work was carried out in the CMAC National Facility supported by UKRPIF (UK Research Partnership Fund) award from the Higher Education Funding Council for England (HEFCE) (Grant ref: HH13054). They also like to thank Dr. Yi-Chieh Chen and Dr. Carla Ferreira for their efforts in collecting images related to dataset-3.

References

Agimelen, O.S., et al., 2016. Integration of in situ imaging and chord length distribution measurements for estimation of particle size and shape. Chem. Eng. Sci. 144, 87–100.

Arruda, R.J., et al., 2023. Automated and material-sparing workflow for the measurement of crystal nucleation and growth kinetics. Cryst. Grow. Des. 23, 3845–3861.

Barhate, Y., Kilari, H., Wu, W.-L., Nagy, Z.K., 2024. Population balance model enabled digital design and uncertainty analysis framework for continuous crystallization of pharmaceuticals using an automated platform with full recycle and minimal material use. Chem. Eng. Sci. 287, 119688.

Barrett, P., et al., 2005. A review of the use of process analytical technology for the understanding and optimization of production batch crystallization processes. Org. Process. Res. Dev. 9, 348–355.

Benyahia, B., Anandan, P.D., Rielly, C., 2021. Control of batch and continuous crystallization processes using reinforcement learning. Comput. Aid. Chem. Eng. 50. Elsevier Masson SAS.

Bolla, G., Sarma, B., Nangia, A.K., 2022. Crystal engineering of pharmaceutical cocrystals in the discovery and development of improved drugs. Chem. Rev. 122, 11514–11603.

Borsos, Á., Szilágyi, B., Agachi, P.Ş., Nagy, Z.K., 2017. Real-time image processing based online feedback control system for cooling batch crystallization. Org. Process. Res. Dev. 21, 511–519.

Bredikhin, A.A., Bredikhina, Z.A., 2017. Stereoselective crystallization as a basis for single-enantiomer drug production. Chem. Eng. Technol. 40, 1211–1220.

Bruno, A.E., et al., 2018. Classification of crystallization outcomes using deep convolutional neural networks. PLoS ONE 13, e0198883.

Cardona, J., et al., 2018. Image analysis framework with focus evaluation for in situ characterisation of particle size and shape attributes. Chem. Eng. Sci. 191, 208–231.

openvino/cvat: v2.11.3 - Computer Vision Annotation Tool (CVAT). doi:10.5281/ZENODO.10908511.

de Moraes, M.G.F., et al., 2023. Modeling and predictive control of cooling crystallization of potassium sulfate by dynamic image analysis: exploring phenomenological and machine learning approaches. Ind. Eng. Chem. Res. 62, 9515–9532.

Gao, Y., et al., 2021. Application of PAT-based feedback control approaches in pharmaceutical crystallization. Crystals 11, 221. *2021, Vol. 11, Page 221*.

He, K., Gkioxari, G., Dollár, P., Girshick, R., 2020. Mask R-CNN. IEEE Trans Pattern Anal Mach Intell 42, 386–397.

Irizarry, R., Chen, A., Crawford, R., Codan, L., Schoell, J., 2017. Data-driven model and model paradigm to predict 1D and 2D particle size distribution from measured chord-length distribution. Chem. Eng. Sci. 164, 202–218.

Jaeggi, A., Rajagopalan, A.K., Morari, M., Mazzotti, M., 2021. Characterizing ensembles of platelike particles via machine learning. Ind. Eng. Chem. Res. 60, 473–483.

Kirillov, A., et al., 2023. Segment Anything. https://doi.org/10.48550/arXiv.2304.02643.

Knopp, T., Szwargulski, P., Griese, F., Gräser, M., 2020. OpenMPIData: an initiative for freely accessible magnetic particle imaging data. Data Brief 28, 104971.

Lin, T.Y., et al., 2014. Microsoft COCO: common objects in context. Lect. Note. Comput. Sci. (including subseries Lect. Note. Artifi. Intell. Lect. Note. Bioinform.) 740–755. **8693 LNCS**.

Lins, J., Harweg, T., Weichert, F., Wohlgemuth, K., 2022. Potential of deep learning methods for deep level particle characterization in crystallization. Appl. Sci. 12, 2465. *2022, Vol. 12, Page 2465*.

Liu, J., et al., 2023. A verified open-access AI-based chemical microparticle image database for in-situ particle visualization and quantification in multi-phase flow. Chem. Eng. J. 451, 138940.

Lorenz, H., Perlberg, A., Sapoundjiev, D., Elsner, M.P., Seidel-Morgenstern, A., 2006. Crystallization of enantiomers. Chem. Eng. Process.: Process Intensifi. 45, 863–873.

MacLeod, C.S., Muller, F.L., 2012. On the fracture of pharmaceutical needle-shaped crystals during pressure filtration: case studies and mechanistic understanding. Org. Process. Res. Dev. 16, 425–434.

Manee, V., Zhu, W., Romagnoli, J.A., 2019. A deep learning image-based sensor for real-time crystal size distribution characterization. Ind. Eng. Chem. Res. 58, 23175–23186.

McGinty, J., et al., 2020. Effect of process conditions on particle size and shape in continuous antisolvent crystallisation of lovastatin. Crystals 10, 925. *2020, Vol. 10, Page 925*.

Metherall, J.P., Carroll, R.C., Coles, S.J., Hall, M.J., Probert, M.R., 2023. Advanced crystallisation methods for small organic molecules. Chem. Soc. Rev. 52, 1995–2010.

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

Nagy, Z.K., Fevotte, G., Kramer, H., Simon, L.L., 2013. Recent advances in the monitoring, modelling and control of crystallization systems. Chem. Eng. Res. Des. 91, 1903–1922.

Nartova, A.V., et al., 2022. Particle recognition on transmission electron microscopy images using computer vision and deep learning for catalytic applications. Catalysts 12, 135. *2022, Vol. 12, Page 135*.

Öner, M., et al., 2020. Comprehensive evaluation of a data driven control strategy: experimental application to a pharmaceutical crystallization process. Chem. Eng. Res. Des. 163, 248–261.

Orehek, J., Teslić, D., Likozar, B., 2021. Continuous crystallization processes in pharmaceutical manufacturing: a review. Org. Process. Res. Dev. 25, 16–42.

Salami, H., Skomski, D., 2023. Building confidence in deep Learning-based image analytics for characterization of pharmaceutical samples. Chem. Eng. Sci. 278, 118904.

Salami, H., McDonald, M.A., Bommarius, A.S., Rousseau, R.W., Grover, M.A., 2021. In Situ Imaging Combined with Deep Learning for Crystallization Process Monitoring: application to Cephalexin Production. Org. Process. Res. Dev. 25, 1670–1679.

Salami, H., Wang, S., Skomski, D., 2023. Evaluation of a self-supervised machine learning method for screening of particulate samples: a case study in liquid formulations. J. Pharm. Sci. 112, 771–778.

Simon, L.L., et al., 2014. Assessment of recent process analytical technology (PAT) trends: a multiauthor review. Org. Process. Res. Dev. 19, 203–214.

Simone, E., Zhang, W., Nagy, Z.K., 2015a. Application of process analytical technology-based feedback control strategies to improve purity and size distribution in biopharmaceutical crystallization. Cryst. Grow. Des. 15, 2908–2919.

Simone, E., Saleemi, A.N., Nagy, Z.K., 2015b. In situ monitoring of polymorphic transformations using a composite sensor array of Raman, NIR, and ATR-UV/vis spectroscopy, FBRM, and PVM for an intelligent decision support system. Org. Process. Res. Dev. 19, 167–177.

Szilágyi, B., Eren, A., Quon, J.L., Papageorgiou, C.D., Nagy, Z.K., 2022a. Digital design of the crystallization of an active pharmaceutical ingredient using a population balance model with a novel size dependent growth rate expression. From development of a digital twin to in silico optimization and experimental validation. Cryst. Grow. Des. 22, 497–512.

Szilágyi, B., Eren, A., Quon, J.L., Papageorgiou, C.D., Nagy, Z.K., 2022b. Monitoring and digital design of the cooling crystallization of a high-aspect ratio anticancer drug using a two-dimensional population balance model. Chem. Eng. Sci. 117700 https://doi.org/10.1016/J.CES.2022.117700.

Szilagyi, B., Eren, A., Quon, J.L., Papageorgiou, C.D., Nagy, Z.K., 2020. Application of model-free and model-based quality-by-control (QbC) for the Efficient design of pharmaceutical crystallization processes. Cryst. Grow. Des. 20, 3979–3996.

Tachtatzis, C., et al., 2015. Image-based monitoring for early detection of fouling in crystallisation processes. Chem. Eng. Sci. 133, 82–90.

ultralytics/yolov5: v7.0 - YOLOv5 SOTA Realtime Instance Segmentation. doi:10.5281/ZENODO.7347926.

Urwin, S.J., et al., 2023. Digital process design to define and deliver pharmaceutical particle attributes. Chem. Eng. Res. Des. 196, 726–749.

Wu, G., Wu, Z., 2023. Machine learning-based MPC of batch crystallization process using physics-informed RNNs. IFAC-PapersOnLine 56, 2846–2851.

Wu, W.L., et al., 2022. Implementation and application of image analysis-based turbidity direct nucleation control for rapid agrochemical crystallization process design and scale-up. Ind. Eng. Chem. Res. 61, 14561–14572.

Wu, W.-L., et al., 2023. Sensor fusion and calibration-based adaptive image analysis procedure for in situ crystal size measurement. Cryst. Grow. Des. 23, 7076–7089.

Xiouras, C., et al., 2022. Applications of artificial intelligence and machine learning algorithms to crystallization. Chem. Rev. 122, 13006–13042.

Yu, L.X., et al., 2004. Applications of process analytical technology to crystallization processes. Adv. Drug. Deliv. Rev. 56, 349–369.