Systems & Control Transactions

Research Article - Peer Reviewed Conference Proceeding
ESCAPE 35 - European Symposium on Computer Aided Process Engineering
Ghent, Belgium. 6-9 July 2025
Jan F.M. Van Impe, Grégoire Léonard, Satyajeet S. Bhonsale,
Monika E. Polańska, Filip Logist (Eds.)

PSE PRESS

Data-driven Digital Design of Pharmaceutical Crystallization Processes

Yash Barhate[a], Yung Shun Kang[a], Neda Nazemifard[b], Ben Renner[b], Yihui Yang[b], Charles Papageorgiou[b], and Zoltan K. Nagy[a]*

[a] Purdue University, Davidson School of Chemical Engineering, West Lafayette, Indiana, United States
[b] Takeda Pharmaceuticals International Company, Cambridge, Massachusetts, United States
* Corresponding Author: zknagy@purdue.edu

ABSTRACT

Mechanistic population balance modeling (PBM) has advanced the design of pharmaceutical crystallization processes, enabling the production of active pharmaceutical ingredient (API) crystals with desired critical quality attributes (CQAs), such as purity and crystal size distribution. However, PBM development can sometimes be resource-intensive, requiring extensive design of experiments (DoE) and high-quality process data, making it impractical under fast-paced industrial development timelines. This study proposes a machine learning (ML)-based workflow for developing 'fit-for-purpose' digital twins of crystallization processes, leveraging industrially available DoE data to link operating conditions with CQAs. Validated on industrial data for a commercial API with complex crystallization challenges, the workflow efficiently identifies optimal operating conditions, demonstrating the potential of data-driven digital twins to accelerate the development of pharmaceutical processes.

Keywords: Artificial Intelligence, Machine Learning, Process Design, Modelling and Simulations, Optimization

1. INTRODUCTION

Crystallization is a critical purification and particle control operation in the pharmaceutical industry, influencing critical quality attributes (CQAs) such as crystal size distribution (CSD), yield, and polymorphic form. These attributes, substantially affect pharmaceutical drug properties, including bioavailability, tablet stability, and manufacturability. As a result, designing robust crystallization processes is crucial for the consistent production of Active Pharmaceutical Ingredients (APIs) that meet stringent quality and regulatory standards. To do so, mechanistic population balance model (PBM)-based digital design has become prominent in the industry, facilitating the optimal design and control of crystallization processes through in-silico simulations, mathematical optimization, and model-based control [1]. However, accurately determining PBM kinetic parameters for specific API-solvent systems poses challenges, requiring a well-designed design of experiments (DoE), along with high-quality real-time process data (e.g., concentration, CSD) from process analytical technology (PAT) tools or offline analyses [2]. The extensive resources—time, material, and labor—required for data acquisition, combined with rapid industrial timelines, often render PBM-based digital design resource-intensive and impractical [2,3].

Recently, there has been growing interest in data-driven digital design approaches to overcome the limitations of traditional crystallization process development methods. Xiouras et al. [3] provide a detailed review of studies on the application of data-driven modeling and control techniques in crystallization. While significant progress has been made in dynamic crystallization process modeling, such frameworks rely on the availability of real-time or historical time-series data from PAT tools. However, as previously noted, obtaining reliable PAT data is industrially challenging due to issues such as stringent calibration requirements, probe fouling, and availability constraints.

To address these limitations, data-driven frameworks must align with the realities of industrial pharmaceutical process development by leveraging available DoE data. This data typically includes only the process operating conditions and the output CQA measurements

obtained at the end of the batch. Operating in the absence of real-time state measurements places these frameworks in a low-data regime, which makes ML model development particularly challenging. Generative modeling techniques offer a promising approach to construct data-driven models under data-scare conditions by training models with synthetic experimental data generated from existing experimental inputs, thus mitigating data limitations [4, 5]. This study evaluates the effectiveness of various generative models in addressing data scarcity and enabling the development of accurate predictive models. Building on this, a generalized framework is proposed for developing data-driven digital twins and deploying them through mathematical optimization to identify operating conditions that yield the desired CQAs.

2. PROPOSED METHODOLOGY

2.1 Generalized data-driven digital twin development workflow

A systematic workflow is essential for building robust data-driven digital twins that generalize across diverse problem statements (Figure 1). The workflow begins with 'System definition' (Step 1), where the CQAs of interest are identified as model outputs, and the operational parameters to be varied are defined. Step 2 focuses on experimental data collection from historical or newly executed DoEs, designed using heuristics, statistical, or model-driven approaches. Step 3 involves data pre-processing, including missing value removal, reserving 10-15 % of data as a hold-out test set, and standardizing inputs for robust model training.

The 'ML model architecture selection and hyperparameter optimization', evaluates and trains multiple regression models including Random Forest (RF), gradient-boosted trees (GBT), Support Vector Regressors (SVR), and Distributed gradient-boosted trees (e.g., XGBoost). Given the limited size of the training dataset, simpler models are prioritized over deep neural networks to reduce the risk of overfitting and maintain model interpretability. Hyperparmater tuning is conduced using an exhaustive search over specified parameters for each model architecture (GridSearchCV) with a K-fold cross validation strategy (K = 3 or 5), optimizing for mean squared error (MSE) as the loss function. The best-performing models for each CQA were selected based on lowest validation MSE across folds and highest test R^2.

When constructing data-driven digital twins using operating parameter-CQA structures, the absence of dynamic interaction with the physical system necessitates the development of static digital twins. These models provide predictions alongside quantified uncertainty measures [2], which can be achieved using ensemble-based uncertainty quantification (UQ) methods such as Monte Carlo cross-validation, bootstrapping, and out-of-bag estimation.

2.1.1 Synthetic data generation from DoE data and augmentation

The central concept in Step 6, is the use of generative AI models to generate new datasets (referred to as 'synthetic data' (z_i^s)) that capture the relationship between operational parameters and CQAs, using existing experimental DoE data (referred to as 'real data' (z_i)). These synthetic datasets serve as proxies for new laboratory experimental data and can be combined with real data to improve the predictive capabilities of ML models [4]. In this study, three widely studied generative models were explored: the Gaussian Copula Synthesizer, Generative Adversarial Networks (GANs), and Variational Autoencoders (VAEs). Each of these models is based on distinct underlying principles and hence, generates synthetic data of varying quality, where *quality* refers to the fidelity of the synthetic data in replicating the statistical properties of the real data. The ability of generative models to accurately capture the real data distribution depends on factors such as the size and distribution of the input real data and the effectiveness of model training [4, 5].

An essential consideration in synthetic data generation is the enforcement of domain-specific constraints to ensure the generated data is meaningful. It was observed that incorporating constraints directly into the training phase of generative models yielded superior performance compared to post-processing methods that discard non-compliant synthetic samples.

Evaluating the quality of synthetic data prior to experimental testing or validation presents significant challenges. To address this, we implemented a recently proposed *weighted expected risk minimization (ERM)* framework [6]. This approach systematically integrates synthetic data with real data to enhance the predictive performance of the resulting model on unseen data. This is done using a weighted loss function (Equation 1),

$$\mathcal{L}_{n,m}(\theta;\ \alpha) = \frac{1-\alpha}{n}\sum_{i=1}^{n}\mathcal{L}\left(\theta, z_i\right) + \frac{\alpha}{m}\sum_{i=1}^{m}\mathcal{L}\left(\theta, z_i^s\right) \quad (1)$$

where the weighting parameter (α) adjusts the contribution of synthetic data based on its quality and its impact on model performance on unseen real test data. Here, n and m refer to the number of real and synthetic samples respectively, $\mathcal{L}$ is the training loss function such as MSE and θ are model hyperparameters. The value of α can be determined systematically for different (n, m) either theoretically, using scaling laws, or experimentally, by minimizing the test error on a validation split of the original dataset [6]. Thus, by shifting the emphasis from solely generating high-quality synthetic data to effectively leveraging it in combination with real data, this approach provides a practical framework for improving ML model performance, particularly in limited data applications.

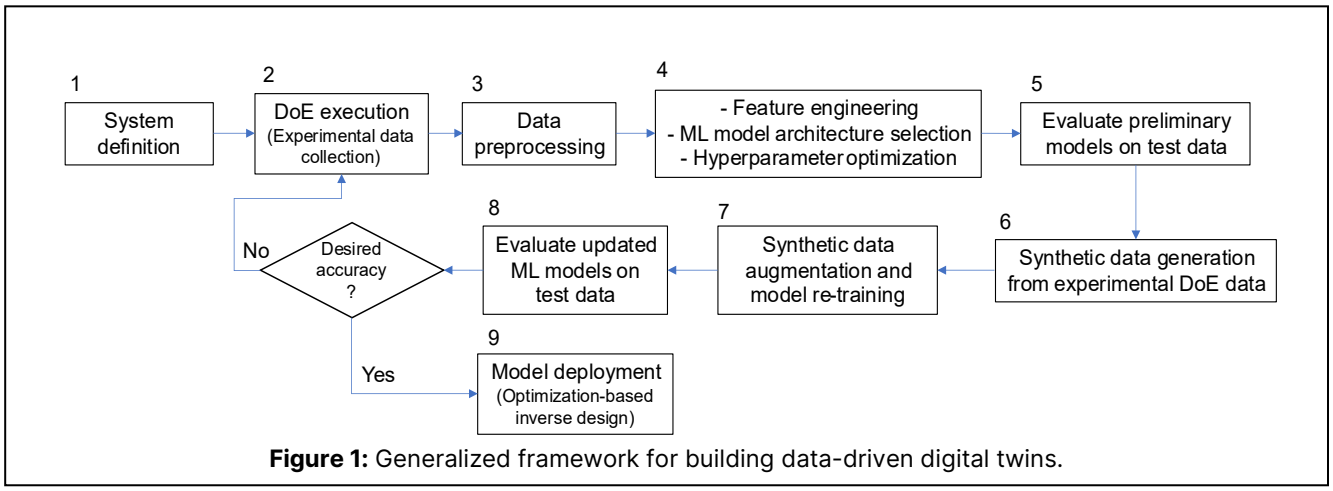

Figure 1: Generalized framework for building data-driven digital twins.

In Step 8, the refined models were re-evaluated using the test dataset. If the satisfactory models are obtained for all CQAs—with prediction uncertainties comparable to measurement uncertainties—the workflow proceeds to the model deployment phase. Otherwise, the process iterates back to step 2, where additional experiments were performed to augment the training data.

Step 9 involves deploying trained models through various approaches to enhance process development and understanding, including in-silico simulations, global system analysis, and optimization. This study demonstrates their use in a mathematical optimization framework to determine operating conditions that address crystallization challenges.

2.2 Key design objectives for the model compound

An industrial compound (Compound A) was selected as the model compound to demonstrate the proposed data-driven modeling workflow. Preliminary experimental investigations into the batch cooling crystallization of A in a fixed, proprietary solvent revealed the formation of a product mixture comprising both amorphous and crystalline particles at the end of the batch. This observation was corroborated through offline characterization techniques, including powder X-ray diffraction (PXRD), and thermogravimetric analysis (TGA). Additionally, significant agglomeration tendencies were observed in the API crystals, as indicated by inline EasyViewer, and offline particle size analysis using the Morphologi 4. Furthermore, slow crystal growth kinetics presented challenges in achieving larger particle sizes. These observations informed the formulation of three key crystallization design objectives for this compound:

1. **Increase Crystallinity**: Increase the percentage of crystalline particles in the product batch.
2. **Minimize Agglomeration**: Reduce the degree of agglomeration (DoA) in the crystals.
3. **Increase Particle size**: Increase the mean volumetric particle size (D-50).

Each of these CQAs was quantified through offline analysis of the crystals obtained at the end of batch crystallization. Crystallinity was quantified via PXRD and TGA, DoA was quantified using the Morphologi 4, and D-50 was measured using the Malvern Mastersizer. Detailed experimental protocols for crystallization experiments and offline product characterization are detailed elsewhere [7].

2.3 Experimental data collection and preprocessing

As per step 2 of the workflow, lab-scale experiments were conducted to collect the initial DoE for the ML model development. The DoE included 31 experiments designed using a combination of heuristic and experimental insights to systematically explore the design space. Initial experiments employed linear temperature profiles with variations in key process parameters, such as initial concentration, seed loading, cooling rates, and final batch temperature. To capture the effects of thermocycles on product CQAs, a subset of experiments incorporated thermal cycling, varying the number of cycles, cooling and heating rates, and intermediate temperature setpoints. Table 1 summarizes the distribution of experiments across four categories: linear, single-cycle, two-cycle, and multi-cycle profiles.

Table 1: Variability in the training dataset across the number of thermocycles in temperature profiles.

Profile type	No. of experiments
Linear (cooling)	18
Linear (1 cooling,1 heating)	2
1 cycle	1
2 cycle	4
3 cycle	3
4 cycle	3

The dataset exhibits inherent imbalance due to practical constraints, as not all CQAs were measured in every experiment. While D50 was measured across all 31

experiments, DoA and crystallinity were measured in only 21 and 15 experiments, respectively. To validate the data-driven framework, four additional experiments were conducted under optimal conditions identified in a prior mechanistic modeling study [7], which aimed to maximize D50 and minimize DoA. These experiments formed the hold-out dataset for evaluating the performance of the ML models. For ML model development, the experimental data from all thermocyclic temperature profiles were represented using 13 input and 3 output features, corresponding to the measured CQAs. The input features included initial concentration, seed loading, and parameters characterizing thermal profiles (e.g., intermediate temperatures, cooling rates, and heating rates).

Figure 2 illustrates the input feature representation across different thermal profiles.

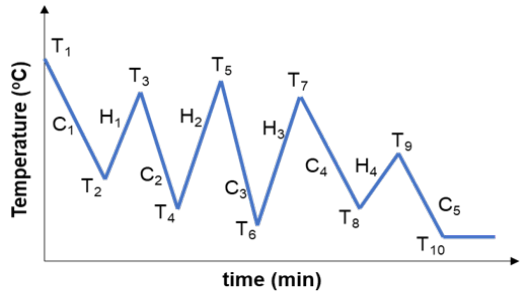

Figure 2: Input features that were used to describe the thermocyclic temperature profiles.

2.4 Computational implementation of the generalized workflow

The ML models described in section 2.1, were implemented using the scikit-learn library (version 1.4.0) and the XGBoost framework (version 2.1.1). Input features were standardized using the StandardScalar class, and hyperparameter tuning was conducted using GridSearchCV with K = 3 cross-validation, optimizing the following hyperparameters:

- RF: Number of estimators (10-200), max depth (0-30), min samples split (2, 5, 10) and min samples leaf (1, 2, 4).
- GBT: Number of estimators (10-200), learning rate (0.01-0.2), max depth (3, 5, 7), subsample (0.5, 0.75, 1.0), alpha (0.1, 0.5, 0.9).
- XGBoost: Number of estimators (10-200), learning rate (0.01-0.2), max depth (3, 5, 7), subsample (0.5, 0.75, 1.0), samples by tree (0.3, 0.7, 1.0).
- SVR: Kernels (Linear, polynomial, radial basis function, sigmoid), C parameter (0.1-20).

The best hyperparameter configuration for each model and the best model architecture were selected based on the lowest validation MSE and highest test R^2. For synthetic data generation, generative models were implemented using the Synthetic Data Vault (SDV) API (version 1.10.0). To ensure the feasibility of the thermocyclic profiles, the generated synthetic data were constrained according to Equations 2,3 and 4.

$$T_i \geq T_{i-1} \;\; \forall\, i = \{3,5,7,9\} \tag{2}$$

$$T_i \leq T_{i-1} \;\; \forall\, i = \{2,4,6,8,10\} \tag{3}$$

$$T_i = 0 \;\Rightarrow (T_j = 0) \wedge (H_j = 0) \wedge (C_j = 0)\; \forall\, \mathrm{j} > \mathrm{i} \tag{4}$$

Hyperparameter tuning of generative models was guided by the SDV's built-in statistical similarity metric, used as the evaluation criterion. The models' effectiveness in reducing test error was evaluated using the weighted ERM approach. For each CQA, three generative models were tested with varying quantities of synthetic samples (5, 10, 15, 20, 25) added to the existing real dataset. The optimal weighting factor for each scenario was then determined, and the corresponding hyperparameter configuration that minimized the cross-validation test error was selected. The models were then retrained with these optimized settings and saved as the final models.

3. RESULTS AND DISCUSSION

In Step 4 of the workflow, various model architectures were assessed for their ability to explain the variance in the collected experimental data. Random forest regressors excelled in modeling D-50 and DoA, whereas gradient boosted trees were most effective for crystallinity predictions. Figure 3 presents the parity plot for predictions across training and testing datasets using these models. Additionally, employing Monte Carlo cross-validation for UQ allowed for both the nominal predictions and their prediction uncertainties, as shown in Figure 4.

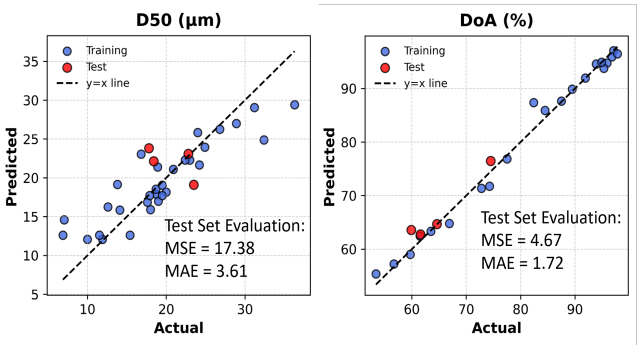

Figure 3: Parity plot of predictions for D-50 and DoA models before synthetic data augmentation.

To further enhance the model accuracy, Synthetic data generation and augmentation were explored. As elaborated in Section 2, synthetic data of varying sizes were

generated from the trained generative AI models, and the weighted ERM approach was used to identify the best α for each scenario. The optimal hyperparameter configuration for D-50 was using a VAE model, with m = 25 and α = 0.35. As shown in Figure 5, this configuration gave the highest reduction in the validation loss on the real dataset. The convex nature of the validation loss curve indicates a trade-off; beyond a certain α value, the inclusion of synthetic data ceased to reduce loss, likely due to deteriorating data quality acting as noise. Re-training the ML model for predicting D-50 with this optimized dataset and hyperparameter setting improved performance on the hold-out test dataset, reducing MSE and mean absolute error (MAE) by 44.6% and 26.9% respectively (Figure 6).

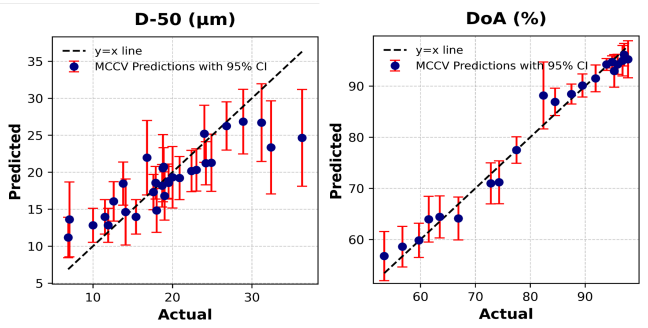

Figure 4: Nominal predictions with prediction uncertainties from the ML models of D-50 and DoA.

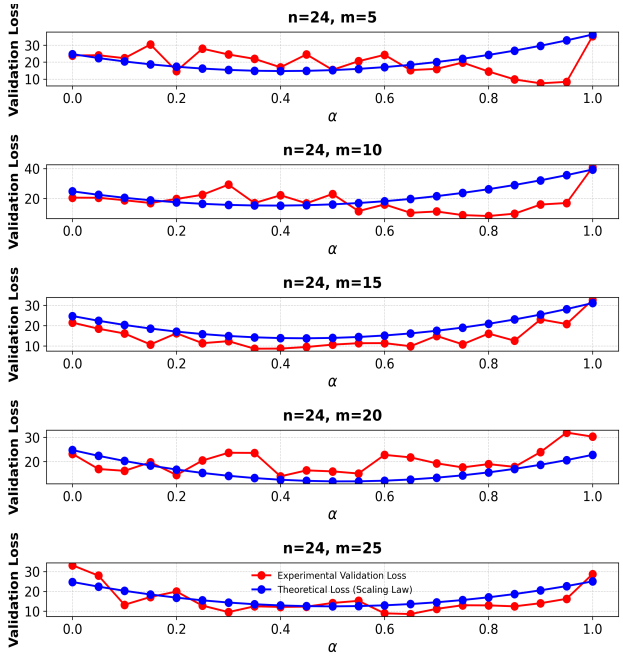

Figure 5: Weighted ERM approach to identify optimal α for synthetic data augmentation for D-50 model.

For DoA predictions, augmentation with synthetic data from generative models did not significantly improve outcomes, potentially due to the inferior quality of the synthetic data. While further model tuning could enhance this, it was deferred as the existing DoA model already demonstrated high predictive accuracy with the real dataset. Since the model predictions for both these CQAs were comparable with measurement errors (uncertainties), these were deployed using optimization frameworks to determine operational conditions that satisfy the crystallization design objectives (Step 9).

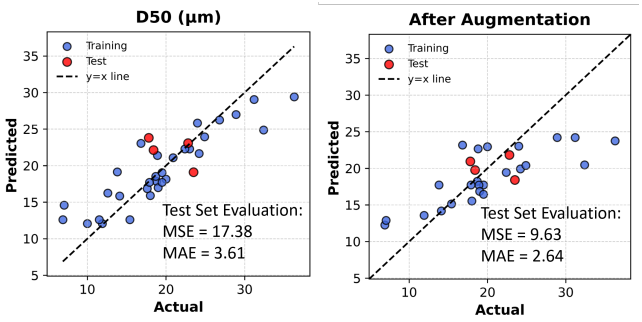

Figure 6: Parity plot for D-50 predictions before (left) and after (right) augmentation of Synthetic Data

The problem was formulated as a dual-objective optimization to maximize D-50 and minimize DoA. The decision variables ($\bar{x}$) represent the ML models' input features corresponding to various thermocyclic profiles. Inequality constraints (Equations 2, 3, and 4) ensured the feasibility of these profiles, with the decision variables' bounds detailed in . This problem was modeled using the Pymoo framework and solved using a derivative-free genetic algorithm. Figure 7 displays the Pareto fronts achieved with various thermocyclic profiles, confirming that increasing the number of temperature cycles from 1/2 to 3/4 enhances mean size and reduces DoA.

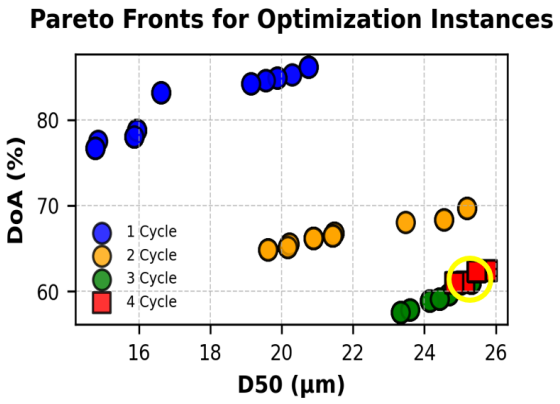

Figure 7: Pareto front obtained after solving the optimization problem.

Furthermore, mapping these optimized results with the crystallinity predictions (Figure 8) from the crystallinity model revealed increased crystallinity values with more thermocycles—an observation experimentally confirmed as reported in [7], where increased crystallinity

and reduced DoA were noted in samples from the end of each cycle in a 4-cycle temperature profile experiment.

Table 2: Lower (LB) and upper bounds (UB) of decision variables

$\bar{x}$	LB	UB
Initial Conc. (mg/g solvent)	24.8	55.2
Seed loading	0.05	0.15
T_i(oC)	0	70
C_i(oC/min)	0.05	0.3
H_i(°C/min)	0.5	1.0

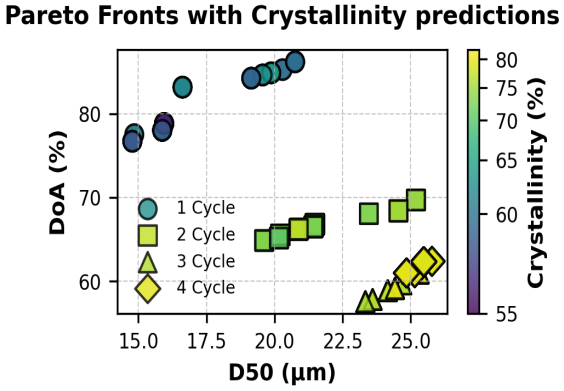

Figure 8: Crystallinity predictions from the Crystallinity ML model mapped onto Pareto front from Figure 7.

A 4-cycle optimal profile from the obtained Pareto front (circled in Figure 7) was experimentally validated, affirming the efficacy and impact of the data-driven digital design workflow in developing crystallization processes. Table 3 compares predictions across all CQAs at the validation point, showing consistency and reliability. Moreover, the optimized profile obtained using a data-driven framework not only matched the CQAs achieved with the mechanistic population-balance approach but also aimed to minimize batch time, thereby enhancing productivity [7].

Table 3: Comparison of experimental and model-predicted CQAs for the validation experiment.

CQA	Experimental	Model-predicted
D-50	17.5±0.3	22.85±6.58
DoA	64.5±12.7	64.63±5.26
Crystallinity	79.4±6.7	77.22±4.16

4. CONCLUSION

This study introduces a systematic workflow for constructing 'fit-for-purpose', data-driven digital twins for crystallization process development, linking process operating conditions with CQAs obtained during product characterization. A key feature of this workflow is the integration of synthetic data derived from experimental data, which enhances the predictive capabilities of ML models, particularly in applications with limited data. The workflow's effectiveness was validated by its application to an industrial API exhibiting agglomeration and low product crystallinity. Using this workflow, reliable data-driven models were developed and integrated into an optimization framework, successfully identifying operating conditions that addressed complex industrial challenges. This demonstrates the potential of these tools in developing complex crystallization processes.

ACKNOWLEDGEMENTS

Funding from Takeda Pharmaceuticals International Co. is gratefully acknowledged.

REFERENCES

1. Nagy Z.K., Braatz, R.D. Advances and new directions in crystallization control. *Annu. Rev. Chem. Biomol. Eng.* **3**, 55–75 (2012)
2. Barhate Y., Kilari H., Wu W.L., Nagy Z.K. 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 (2024)
3. Xiouras, C., Cameli, F., Quilló, G. L., Kavousanakis, M. E., Vlachos, D. G., Stefanidis, G. D. Applications of Artificial Intelligence and Machine Learning Algorithms to Crystallization. *Chem. Rev.* 122, 13006–13042 (2022)
4. Ma Y., Li W., Yang H., Gong J., Nagy Z.K. Digital design of cooling crystallization processes using a machine learning-based strategy. *Ind. Eng. Chem. Res.* 63, 46, 20236-20251 (2024)
5. Lu Y., Shen M., Wang H., Wang X., Rechem C., Fu T., Wei W. Machine learning for synthetic data generation: A review. *arXiv: 2302.04062*
6. Jain A., Montanari A., Sasoglu E. Scaling laws for learning with real and surrogate data. arXiv:2402.04376
7. Kang Y.S., Kilari H., Nazemifard N., Renner C.B., Yang Y., Papageorgiou C., Nagy Z.K. Optimization based digital design for agglomeration control of a pharmaceutical crystallization process. *2024 AIChE Annual Meeting* (2024)