ACS OMEGA

Open Access

This article is licensed under CC-BY-NC-ND 4.0

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

Article

Data-Driven Modeling of the Spray Drying Process. Process Monitoring and Prediction of the Particle Size in Pharmaceutical Production

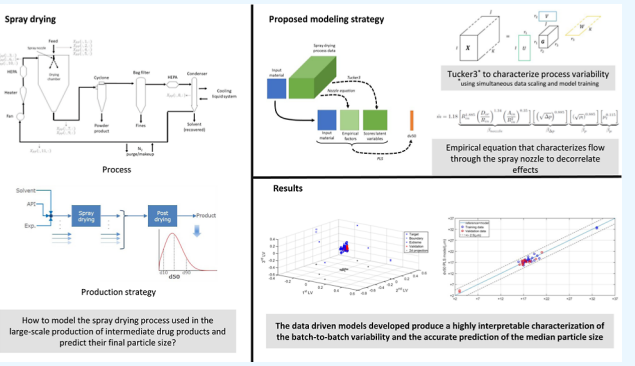
Carlos André Muñoz López, Kristin Peeters, and Jan Van Impe*

Cite This: https://doi.org/10.1021/acsomega.3c08032

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Spray drying is used in the pharmaceutical industry for particle engineering of amorphous solid dispersions (ASDs). The particle size of the spray-dried (SD) powders is one of their key attributes due to its impact on the downstream processes and the drug product's functional properties. Offline and inline laser diffraction methods can be used to estimate the product's particle size; however, the final release of these ASDs is based on offline instruments. This paper presents a novel data-driven modeling approach for predicting the particle size of SD products. The model-based characterization of the process and the product's particle size, as a critical quality attribute, follows the quality by design principles. The resulting model can be used for online process monitoring, reducing the risks of out-of-specifications products and supporting their real-time release. A Tucker3 model is trained to capture and factorize the deterministic variability of the process. Subsequently, a partial least-squares regression model is calibrated to model the impact that variability in the input material properties, the process parameters, and the spray nozzle have on the products' particle size. This strategy has been calibrated and validated using large scale production data for two intermediate drug products under high sparsity of particle size data. Despite the challenges, high accuracy was obtained in predicting the median particle size (dv50) for release. The 99% confidence interval results in an error of maximum 2.5 μm, which is less than 10% of the allowed range of variation.



Downloaded via 95.91.208.235 on June 10, 2024 at 10:47:19 (UTC).
See https://pubs.acs.org/sharingguidelines for options on how to legitimately share published articles.

INTRODUCTION

Sound scientific understanding of the pharmaceutical manufacturing processes and, specifically, the impact that the variability in the material attributes (MAs) and process parameters (PPs) has on the critical quality attributes is a core element of the quality by design (QbD) framework.[1] Achieving this goal comes with challenges as well as opportunities, some of them specific to large scale pharmaceutical manufacturing. One of these challenges is the use of mathematical models to characterize the large scale production processes including sources of variation that were not present during the product and process development.[1,2] However, this challenge also poses an opportunity; the model obtained, once it is validated, can aid the manufacturing activities. Thus, applications such as real time release (RTR) and multivariate statistical process control can be developed and deployed. As identified in the QbD framework, these models and their applications have a strong potential for impact at the later stages of the product's life cycle, i.e., at defining the control strategies, and for the continual improvement.[3] However, if the models developed are informative and interpretable, the opportunity for a feedback loop toward previous stages is possible, i.e., for the risk assessment and the design space definition.

Spray drying is a technology that is largely used in the pharmaceutical industry. It is a process used in the late stages in the production of active pharmaceutical ingredients (APIs) or during the manufacture of intermediate and final drug products (DP). The production of poorly water-soluble drugs, delivered in the form of amorphous solid dispersion (ASDs) or nanocomposites, are examples of processes where spray drying can be a key step.[4−7] Broadly speaking, spray drying is a (semi)continuous process used in the production of nano- to micro-sized particles with a reasonably narrow size distribution.[8] In any case, the spray drying process requires tight control over the product's particle properties. Among these properties, particle size distribution (PSD) is of primary interest; it can be a CQA and/or it can impact the functional

Received: November 12, 2023
Revised: February 7, 2024
Accepted: March 5, 2024

© XXXX The Authors. Published by American Chemical Society

properties, of the product. Additionally, the particle size can impact the performance during the downstream processing, e.g., powder compressibility.[9]

Despite the long-standing presence of the spray drying process in industry, research related to this process is still very active. Research efforts are focused on the development of novel formulations with advanced physical, biological, and/or chemical functional properties, as well as the advance in process understanding, modeling,[10−17] scale-up, and control. Novel and specialized spray-dried (SD) products such as microparticles containing plasmid nanocomplex,[18] microparticles of ciprofloxacin hydrochloride for pulmonary delivery,[19] and hollow spherical aggregates of silica nanoparticles[20] are a strong focus of research. These works and many others have also demonstrated the importance of understanding the impact that variations in the process conditions and input materials can have on the CQA of the products. In this regard, the QbD concept provides a framework to formulate these challenges. During the past decade, the QbD approach has serve to bridge the product-process development and the large scale manufacturing.[21,22]

The sustained increase of computational power and the scientific progress in the field of modeling of particulate solid processes in the pharmaceutical industry are paving the way toward models that can support the QbD objectives. For spray drying, models based on computational fluid dynamics, Lagrangian−Eulerian modeling, discrete element method, and population balance modeling have been investigated to develop mechanistic models of the process.[10−13] These modeling strategies provide a representation of some of the physical phenomena involved in the process, making them good for process understanding and development. However, they are not yet the best solution for process monitoring, prediction, and control. These methods are highly computationally demanding, which hinders their use in online applications and they normally require detailed data, not always available, to be calibrated for a specific process.[13,15] Other mechanistically inspired models, based on approximations and formulated as simpler one-dimensional ordinary differential equations[14−17] might be more suitable for efficient process development and for online applications. However, these models include several assumptions to approximate the real process, while many sources of variability have to be handled as sources of uncertainty. As Sloth et al.[23] have shown, the spray drying process combines two phenomena occurring simultaneously, i.e., the morphology formation and the drying. Sturm et al.[24] developed a model that describes these phenomena; however, this model is only demonstrated in the spray drying of hypromellose acetate succinate (HPMCAS), and it is valid only until the glass transition point of the material is reached. These are a few examples that show how mechanistic and semiempirical models still have limited applicability to support the large scale spray drying process.

Data-driven modeling approaches can offer a favorable trade-off between complexity and accuracy for process modeling, especially regarding online applications. Although these black box modeling approaches are specific to the product/process and more generally to the data used for training, the outcome of their application can be highly valuable regarding process understanding and model accuracy. For instance, Gil-Chávez et al.[25] use a response surface for the optimization of the spray drying process of aquasolv lignin and Milanesi et al.[13] use machine learning to extend a thermodynamic balance model of the spray drying process and to estimate the outlet temperature accurately. In recent years, multivariate statistical process monitoring (MSPM) and machine learning have become more standard approaches. These methods aid in the modeling of large scale batch processes that are subject to disturbances that were not foreseen or could not be modeled based on mechanistic or semiempirical models.[26−29] In recent reviews, Ramos et al.[30] discusses some methods commonly used in applications dealing with batch processes, while Ebadi et al.[31] focus on specific methods that target the covariance matrix of the process. Despite the diversity of existing and novel methods, the black box nature of these modeling strategies results in one common limitation, i.e., their limited interpretability. Improvement in this regard is one of the main drivers for research in this field.

A novel hierarchical strategy to model the spray drying process and predict the particle size of the product is presented. This strategy is based on MSPM methods and follows QbD precepts. It is intended to exploit data that are commonly available in the large scale, pharmaceutical manufacturing environment, with no need for additional experimental data. It is a data-driven approach that integrates a recently developed tensor decomposition training method and a linear regression model. Given the black box nature of the models obtained, the contribution of this work is the modeling strategy presented and not the specific calibrated models. These models are valid only for the products and production conditions included in the calibration data. The goal of this novel modeling strategy, applied to the spray drying process, is to (i) build a better understanding of the impact that uncontrolled process variability has on the particle size of the product, (ii) provide a strategy for interpretable and reliable process monitoring, and (iii) predict accurately the particle size of the products.

At the core, the proposed modeling strategy uses a Tucker3 model, which is calibrated using an algorithm for simultaneous data scaling and training.[32] The tensor decomposition via Tucker3 results in a multilinear rank approximation of the process variability. This method is well suited to factorize batch process data because it preserves the three-dimensional tensor structure of the data, where each mode is one direction of the process variability, i.e., batches, variables, and time. Fanaee-T and Gama[33] discuss empirical evidence that shows the advantages of using tensor methods for anomaly detection in batch process monitoring; this in comparison with matrix-based methods, such as principal component analysis, which require tensor unfolding. Better interpretability of the score plots, higher classification accuracy, lower approximation error, better identification of the variance, and a lower risk of overfitting are some of the identified advantages. Multilinear partial least-squares (PLS) for batch processes have also shown a reduction in noise propagation and a higher accuracy in the predictions, compared to traditional PLS.[27] More recently, Sun and Braatz[34] have highlighted the need for more systematic and in-depth research focused on the improvement, use, and applications of tensorial data analytics in chemical and biological manufacturing. Thus, the Tucker3 method is used in this work in conjunction with a recently developed strategy for simultaneous data scaling and training. As shown by Muñoz et al.,[32] using this calibration algorithm with a Tucker3 model results in a better factorization of the deterministic variability

https://doi.org/10.1021/acsomega.3c08032
ACS Omega XXXX, XXX, XXX−XXX

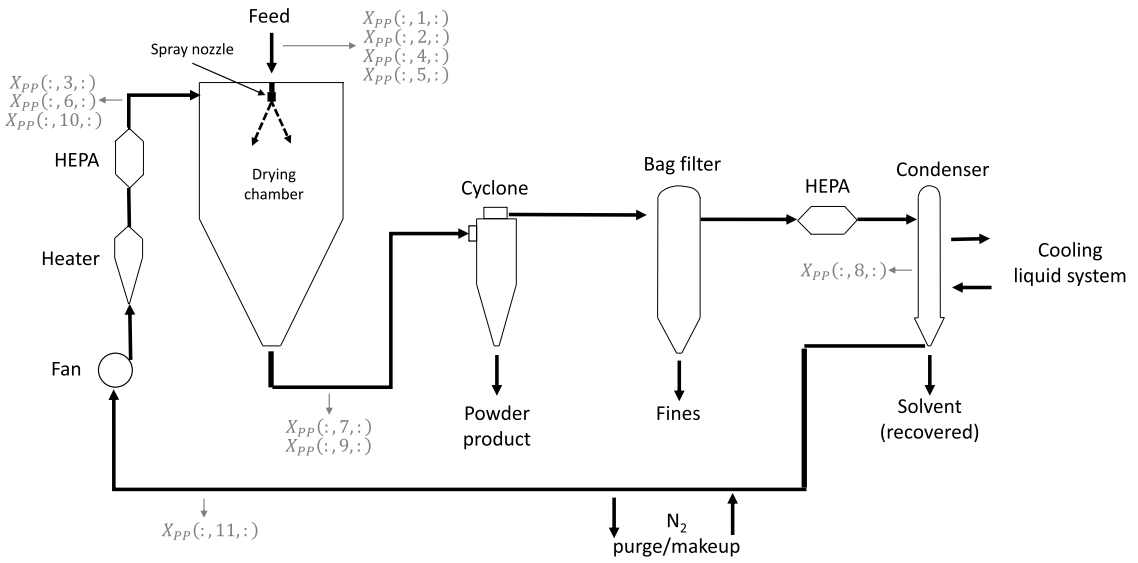


Figure 1. Spray drying process.

of the process, which is more interpretable, and provides insights into the correlations of the variables and the dynamic behavior of the process.

A PLS model is integrated into the proposed hierarchical model structure to predict the particle size of the SD material. The PLS model takes the time-invariant data and the scores of the Tucker3 model as inputs to predict the median particle size of the product. The scores of the Tucker3 model serve as the fingerprints for each batch produced in the spray dryer. Thus, the PLS regression is calibrated to model the impact that the variability in the input material properties, the PPs, and the spray nozzle have on the products' particle size. PLS is a common technique used in chemometrics to develop bilinear regression models exploiting advanced characterization techniques.[35] PLS models have also been used in some applications of process modeling such as the tableting[36] and granulation.[37,38] However, the direct use of PLS was discarded due to the three-dimensional nature of the process data and the need to combine these inputs with time-invariant inputs such as the critical material attributes (CMAs).

Two major challenges were addressed during the development of the proposed modeling strategy. These challenges arise from the incomplete data found in real industrial scenarios. First, only a subset of the independent input variables (e.g., feed flow rate, temperature, and density) as well as the response to the feed flow condition are measured in industrial spray dryers. However, the sensitivity analysis of the spray drying process has shown that the particle size is mostly affected by variations in the viscosity of the feed solution and the feed flow condition through the pressure nozzle.[39] To work out this limitation, the proposed regression model uses a set of empirical factors that serve to decorrelate the sources of variation in the flow through the nozzle. These empirical factors are derived from the empirical equation formulated for the flow through swirl nozzles.[40] The second challenge is the high sparsity of the particle size data. Although inline particle size analyzers are available in the market, their use in manufacturing environments is limited due to high cost/benefit ratios, low reliability, and difficulty to validate. In this work, an iterative training strategy is used to address data sparsity. This training strategy allows missing values to be inferred, based on the knowledge captured by the model about the process and the input material variability.

The paper first discusses specific aspects of the materials and methods used in this work, i.e., the products used to develop and validate the modeling strategy, the spray drying unit, the data sets available, and the modeling methods implemented. The second part of the paper presents the results obtained at every step of the modeling effort and discusses the most relevant aspects regarding the process understanding and the validity of the proposed method. Finally, the conclusions drawn from the main findings of this research are summarized.

MATERIALS AND METHODS

The data sets and the methods used in this work are described in this section. Details of the two formulations used and the spray drying unit are provided first. The second part discusses the methods and modeling strategy applied to monitor the spray drying operation and predict the product's particle size.

Products. Historical production data and process introduction data for two ASDs intermediate DP are part of this work. These products are obtained at large scale after the final steps of spray drying and post drying. They are formulations based on an API and a single excipient (different in each formulation). Generic labels, i.e., formulations A and B, are assigned to these formulations. The solution fed to the spray dryer consisted of the solid formulation dissolved in a blend of solvents.

Since the excipients are polymeric chains, the number and type of substitutions determine their functional properties. Therefore, the variations that can occur in these substitutions have a direct impact on the properties of the excipient and therefore the spray drying feed solution.[41] To a large extent, the polymerization and substitution conditions of these materials are controlled by the manufacturers to offer fixed ranges of variation in the functional properties. Thus, excipients are available at various commercial grades, which provide specific and controlled quality ranges. However, there is still a certain level of variability that cannot be neglected when the objective is to predict the variability in the particle size accurately. The uncontrolled variability in the input material can have an impact on the properties of the feed

https://doi.org/10.1021/acsomega.3c08032
ACS Omega XXXX, XXX, XXX−XXX

solution, especially its viscosity, and therefore the final particle size of the product. This is one of the primary sources of variability considered in this work.

Large Scale Spray Dryer. The products described previously are SD in a GEA PHARMA-SD type PSD-4 unit. Figure 1 depicts the flow diagram for this process, including the PPs measured, controlled, and monitored. The symbols used for these PPs are given in Table 1. The liquid feed (i.e., the intermediate DP formulation in the solvent blend) is brought to the drying chamber via the spray nozzle. Droplets are formed once the liquid, at high pressure, is released into the drying chamber. The sudden drop in pressure and the nozzle geometry are responsible for the formation of a thin layer of liquid that rapidly bursts to form droplets. The droplets flow cocurrent with the hot gas that also enters the drying chamber from the top. While the droplets are falling down, they are progressively dried to form amorphous solid particles. The solvents in the droplets are transferred to the gas phase, increasing the relative saturation. The particulate product leaves the drying chamber from the bottom, being dragged by the gas flow. First, a cyclone is used to separate the solids from the gas flow; the recovered powder is the product. Later, the gas flow passes through a bag filter to remove fines. Finally, a condenser is used to remove the vaporized solvents. Thus, the relative saturation of the gas is lowered and it can be used as drying gas again. The gas is heated, and makeup/purge streams are used. In some operations, the liquid solvent recovered in the condenser is used again in the feed solution, and the mixture of condensed solvents is denominated recovered solvent. Two high-efficiency particulate air filters are placed in the process to guarantee that the gas stream is free of particles.

Table 1. PP Spray Drying

process parameter	symbol	units
feed flow rate	$X_{PP}(:, 1,:)$	[kg/h]
feed density	$X_{PP}(:, 2,:)$	[kg/m^3]
gas flow rate	$X_{PP}(:, 3,:)$	[kg/h]
atomization pressure	$X_{PP}(:, 4,:)$	[Pa]
feed temperature	$X_{PP}(:, 5,:)$	[$°C$]
gas temperature inlet	$X_{PP}(:, 6,:)$	[$°C$]
gas temperature outlet	$X_{PP}(:, 7,:)$	[$°C$]
condenser temperature	$X_{PP}(:, 8,:)$	[$°C$]
gas pressure outlet	$X_{PP}(:, 9,:)$	[Pa]
gas pressure inlet	$X_{PP}(:, 10,:)$	[Pa]
pressure dryer	$X_{PP}(:, 11,:)$	[Pa]

The large scale production of these intermediate DPs is accomplished following the strategy presented in Figure 2. The input materials, i.e, API, excipient and solvents, are combined in the feed solution, which is SD, to obtain the wet product. The production is completed in batches. Each batch consists of the particulate product collected during operation of the spray dryer. Several batches of the wet particulate product are combined into a larger batch of material that is post dried in a dynamic dryer. The final drying step aims to further reduce the residual content of solvents in the solids to obtain the desired dried powder. The final quality attributes of the product, including the median particle size (dv50), are measured only at the end, i.e., for each batch of postdried (PD) powder.

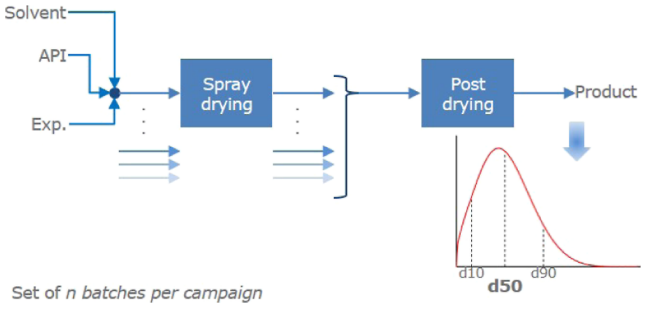


Figure 2. Production strategy.

The production strategy considers that the spray dryer is first fed with the blend of solvents with no dissolved solids. This is the flushing step that occurs for some time before every batch of production. The input flow is switched to the solution, including the dissolved formulation, once the flushing step is completed. The spray drying process applied to the two formulations differs in two operational aspects. First, formulation A uses a cycle of solvent recuperation per campaign. In this cycle, the solvent captured in the condenser for one batch is used as part of the feed solution of the next batch. Pure solvents are added to the recovered solvent mixture as makeup and to adjust the mixture's density. Formulation B does not use the recovered solvent. The second difference is the variable used to control the flow through the spray nozzle. Given the relation between the feed flow rate and the atomization pressure, only one of the two can be controlled independently. In the case of formulation A, the atomization pressure is independently controlled, while the control strategy for formulation B is based on the feed flow rate. All of the other PPs are controlled and monitored equally for the two formulations.

Data Sets. The data used to develop and validate the models of the spray drying process consist of three different data sets, two inputs, and one output. The input data include the continuously measured spray drying PPs, e.g., feed flow rate, atomization pressure, feed temperature, and properties of the excipient material. The output data set is the median particle size of the product.

PPs. Table 1 provides the list of 11 (11) PPs measured continuously during the operation of the spray dryer. Every variable is considered an input in the modeling strategy. These PPs are measured at different locations, as shown in Figure 1. The input flow rate, the atomization pressure, the temperature, and the liquid density are measured in the liquid feed stream at the input of the spray drying unit. Temperature and flow rate are also measured at the input of the gas stream after the gas heater. Pressure and temperature are measured in the drying chamber. Additional temperature and pressure sensors help to track these variables at the output of the drying chamber and downstream in the condenser.

The data set corresponding to the PPs is a third-order tensor. This structure is shown in Figure 3, where $X_{CPP[I\times J\times K]}$ is the third-order tensor that contains data for I batches of the spray drying process, with J variables measured per batch during K time points. This structure can be seen as I horizontal slices of data, one per batch, which are stacked in the vertical direction. In this work, the inherent structure of the process data is kept, and tensor methods are exploited to develop the models of the spray drying operation. The symbols used in Table 1 for the PPs correspond to their tensorial representation, as individual slices $X_{PP}(:, j,:)$ of the tensor. Each of these slices contains all measurements in time for the

https://doi.org/10.1021/acsomega.3c08032
ACS Omega XXXX, XXX, XXX−XXX

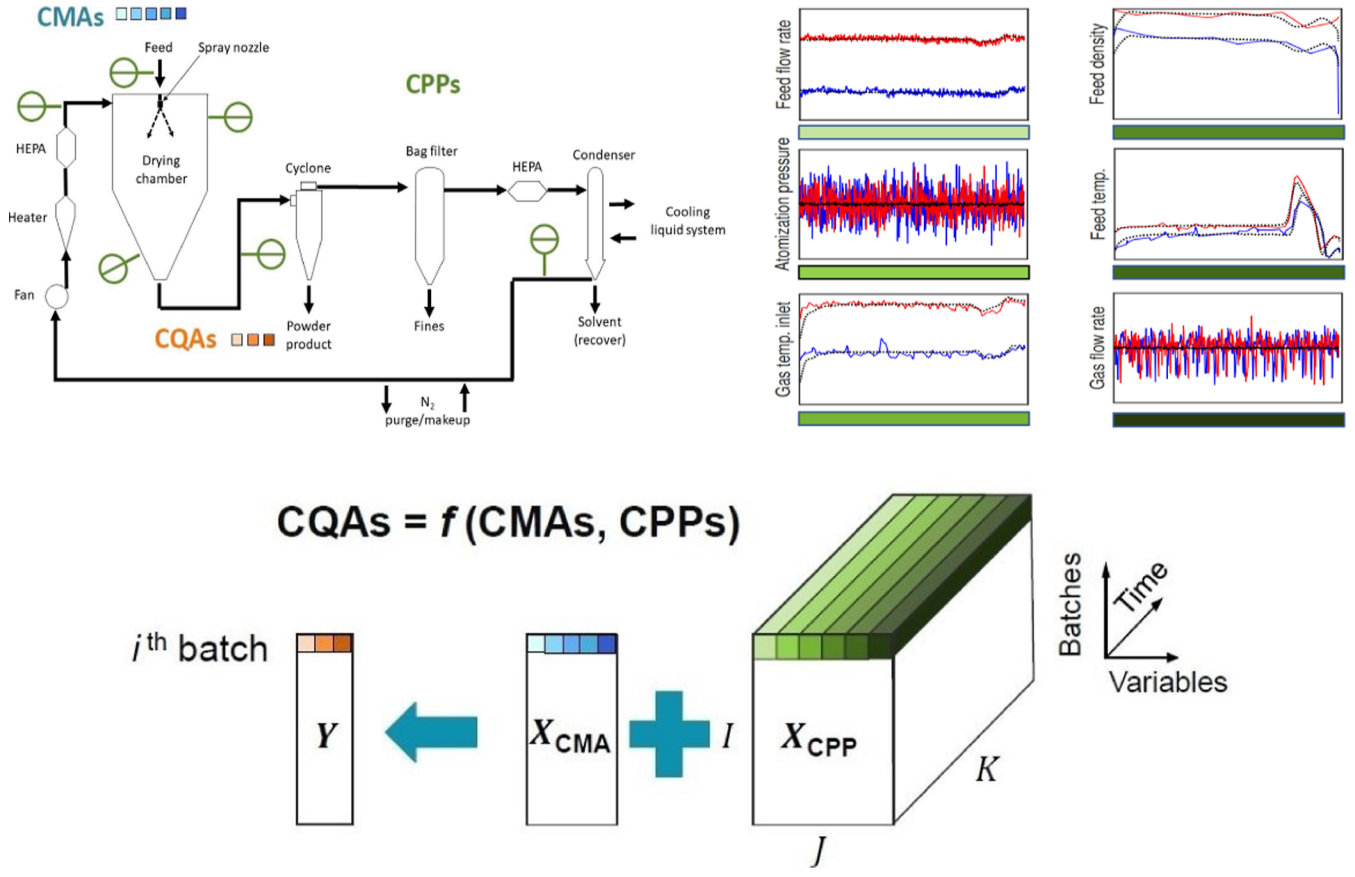


Figure 3. Structure of the data from batch processes.

given variable and for all batches in the data set. The subindices critical process parameter (CPP) and PP are used to indicate whether the data set consists only of CPPs, or if it also includes noncritical PPs.

MAs. Table 2 shows the list of MAs known for the lots of excipients used in each batch of each formulation. These data

Table 2. MAs Available for the Excipient

material attribute	symbol	units
loss on drying	$X_{MA}(:, 1)$	[mg/g]
residue on ignition/sulfated ash	$X_{MA}(:, 2)$	[mg/g]
viscosity in aqueous solution	$X_{MA}(:, 3)$	[cP]
content of substituent groups	$X_{MA}(:, 4: L)$	[mg/g]

are time-invariant because the formulation for every batch of SD product is prepared in advance using a single lot of excipient or a known blend. In Figure 3 the data set of the MAs is represented by $X_{CMA[I\times L]}$. This is a matrix with a set of I batches for which the L attributes of the excipient are reported. When more than one lot of excipient is used in the formulation of a single batch of SD product, the attributes in X_{CMA} correspond to the weighted sum of the attributes for the individual lots of excipient. The subindices CMA and MA are used to indicate whether the data set consists only of critical MAs (CMAs), or if it also includes noncritical MAs.

Particle Size. The median particle size of the product is the goal CQA to be estimated using the modeling strategy proposed in this work. The release value of this CQA is derived from the PSD of the product, which is determined via offline analytical testing using a Malvern Mastersizer 2000 laser diffraction equipment (Malvern panalytical, The Netherlands). Although these data are available for all batches of product included in the model development, it is a sparse data set with respect to the input data $\boldsymbol{X}_{PP}$ and X_{MA}. Data sparsity is a common challenge in large scale industrial data sets. In this particular case, the reason for sparsity in the output data is the combination of multiple batches of wet product into a single batch going to the post dryer. Since the particle size is measured only for the PD powder, a single value of dv50 is reported for several batches of SD product. Thus, the output data set is a sparse column vector $Y_{[I\times 1]}$. Given the production scales of each formulation, the sparsity of Y varies. The approach followed to deal with the sparsity in the output data is discussed later in the Methods section.

Data from a process analytical technology (PAT) tool for the inline measurement of the PSD at the outlet of the spray dryer chamber are available for a limited set of batches. This PAT produces a discrete signal of the inline estimate of the dv50 at high frequency (i.e., 1/s). The instrument used is the particle size analyzer Malvern Insitec (Malvern panalytical, The Netherlands). The main limitation of the resulting measurement is the large variability in the output signal due to measurement noise and deviations introduced due to vibrations in the drying chamber produced due to hammering. Wavelet analysis is used for the data processing of the signal. Wavelet analysis was applied to decompose the raw signal into the corresponding time series at different frequency bands.[42] This allows to filter out the undesired variability of the measurement and to obtain a much more reliable signal for the continuous inline measurement of the particle size. The denoised signal is used to estimate the median particle size of

https://doi.org/10.1021/acsomega.3c08032
ACS Omega XXXX, XXX, XXX−XXX

the product from individual batches of SD material. Since these data are available only for a limited set of batches, they were used for validation purposes only.

Data Structure. Figure 3 represents the QbD approach applied to model the spray drying process. The QbD framework establishes that the control strategy must be defined in terms of the CQAs, which in turn are function of the CMAs and CPPs. Thus, in terms of the regression model, the output space (Y) is regressed into the input space which consists of the variables contained in X_{MA} and $\boldsymbol{X}_{PP}$.

The data used for the development of the monitoring and predictive models were split to have independent sets for model training and validation. Table 3 provides details on the

Table 3. Data Distribution for Model Training and Validation

formulation	data set purpose	excipient lots	number of spray-dried batches	post-dried batches
A	training	7	216	40
	validation	4	139	24
B	training	12	68	26
	validation	8	51	19

number of batches used in each data set. The validation data set of each formulation accounts for around 40% of the total number of batches available for model development, i.e., training and validation. The training data set for formulation A consisted of 216 batches of SD product, which in turn correspond to 40 batches of the final PD product. A total of 7 different lots of excipient were used in the production of those batches. Equivalently, the validation data set for this formulation consisted of 139 SD batches that resulted in 24 batches of the final PD product; 4 independent lots of excipient were used in the production of the validation batches. The ratio between SD batches and PD batches is not constant, because the number of SD batches that are combined into a single PD batch depends on the scale ratio between the two. Since two different dryers are used in the PD step, the number of SD batches that go into a PD batch varies.

Figure 4 depicts the distribution of the data in the training and validation data sets. This figure shows the reported dv50 for the batches of the PD product. The data available for formulation A consist only of production data at target conditions. The resulting variability observed in the particle size remains within the control limits defined for this product. The data available for formulation B are a combination of production data at target conditions and process introduction data. As it is shown in Figure 4b, the introduction data for formulation B consists of four boundary batches and four extreme batches. On the one hand, the boundary batches were completed under process conditions that are at the limits of the proven acceptable ranges using the same spray nozzle. On the other hand, the extreme batches were produced using nozzles of different dimensions with variations in the feed flow rate and atomization pressure to obtain products with extremely low and high particle dimensions. All other PPs remained under target conditions. In the case of formulation B, the validation set was intentionally selected to not coincide with the calibration set. The goal is to assess the linearity of the model beyond the calibration range. This strategy would not be applied to the calibration and validation of a model that is intended to be deployed and used in practice. However,

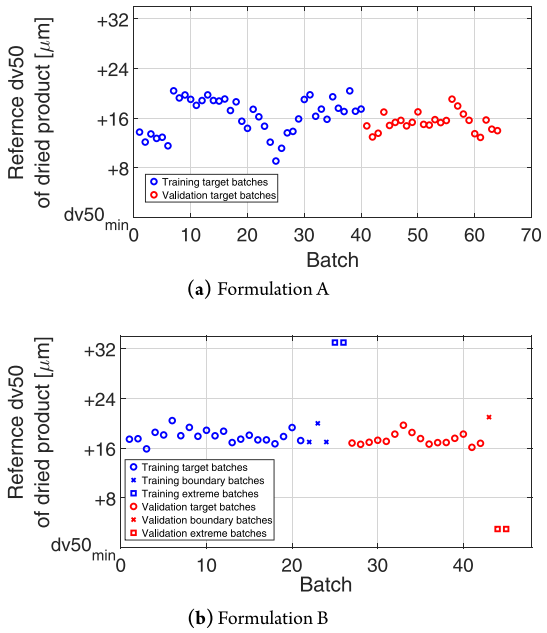


Figure 4. dv50 reported by the reference method for the training and validation data sets.

considering the research purpose of this work, this strategy provides insights about the validity of the linear correlations captured by the model beyond its calibration range.

The scale of the y-axis in Figure 4 is relative. A minimum particle size was selected as a reference, i.e., $dv50_{min}$, and all other particle sizes are reported relative to it in μm. This scale was selected to provide the magnitude of the variability in the median particle size. The absolute values of particle size are not needed to assess the accuracy of the model. Other figures that report dv50 values use the same scale.

Modeling Methods. Tucker3 and PLS are the two methods used in the proposed modeling strategy to monitor spray dryer operation and predict the median particle size of the particulate product.

Tucker3. Tucker3 is a tensor decomposition method used for the factorization of batch process data. A third-order tensor is factorized into three factor matrices (i.e., U, V, and W) and a core tensor ($\boldsymbol{G}$), as shown in Figure 5. The variability in the

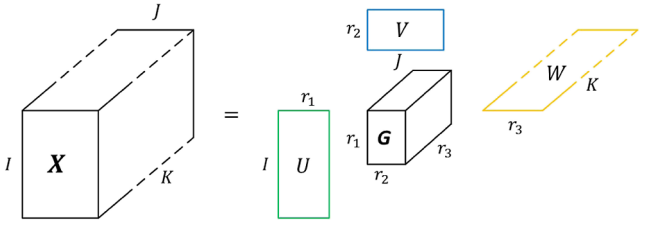


Figure 5. Graphical representation of the Tucker3 factorization.

input data is expressed as the multilinear combination of the columns in the factor matrices. The elements of the core tensor are used as the weight coefficients of the linear combinations.[43] The decomposition obtained via the Tucker3 method can be interpreted as the factorization of the overall process variability into batch-to-batch variability, correlations among variables, and the linear time-series basis of the dynamic behavior. This means that U contains information regarding batch-to-batch variability, V contains information about the variables

https://doi.org/10.1021/acsomega.3c08032
ACS Omega XXXX, XXX, XXX−XXX

responsible for these variations and how they are correlated, and W represents the dynamic behavior of the system. The information obtained in W can be seen as a set of time series that work as the linear basis for the dynamic behavior of the PPs. The Tucker3 factorization is given in eq 1 where $\otimes$ is the Kronecker product defined in eq 2.

$$\mathbf{X}(i, j, k) \approx \sum_{l=1}^{r_1} \sum_{m=1}^{r_2} \sum_{n=1}^{r_3} \boldsymbol{G}(l, m, n) U(i, l) V(j, m) W(k, n)$$
$$X_{(1)} = U G_{(1)} (W \otimes V)^{\mathrm{T}} + E_{(1)} \tag{1}$$

$$W_{[K\times r_3]} \otimes V_{[J\times r_2]} = \begin{bmatrix} w_{11}V & w_{12}V & \cdots \\ w_{21}V & w_{22}V & \cdots \\ \vdots & \vdots & \end{bmatrix}_{[KJ\times r_3 r_2]} \tag{2}$$

Estimating the parameters of the Tucker3 decomposition requires determining the set of ranks $[r_1, r_2,$ and $r_3]$ that produce the best approximation of the tensor data. This is possible using a cross-validation strategy. The approach proposed by Louwerse et al.[44] is followed in this work to perform a structured search that reduces the redundancy in the multirank search and avoids the evaluation of unfeasible solutions. The error in the approximation obtained via the Tucker3 decomposition is given by $\boldsymbol{E}$ which is a tensor of the same dimensions as the input data. The Tucker3 method and the tensor handling is implemented in this work with the help of Tensorlab which is a Matlab based tool design for this purpose.[45]

PLS. PLS is a standard modeling tool used in chemometrics that has shown potential in application of MSPM. PLS is used to develop regression models where the covariance between the set of input variables and out variable(s) is captured in a latent space through a bilinear transformation. Equation 3 is the formulation of the PLS method, where X and Y are the input output matrices. The factor matrices P and Q are the input and output loadings, which are the linear transformations from the original multidimensional spaces into a reduced latent space, where the input−output covariance is expressed. T and U are the scores that represent the original data in the model space. The scores in T, which are the projections of the input data in the model space, work as the set of predictors of the output variable(s). Finally, W, the weight matrix, provides a way to compute the inverse transformation of the method given by W^*, which in turn allows us to compute the scores from new input data.

$$\begin{aligned} X &= TP^T + E \qquad T = XW^* \\ Y &= UQ^T + H \qquad W^* = W(P^TW)^{-1} \\ Y &= TQ^T + F \end{aligned} \tag{3}$$

The estimation of the parameters in the PLS regression model requires the determination of the best rank approximation of the covariance of X and Y. Cross-validation is used for this purpose; details on the specific conditions used to calibrate the model for each formulation are provided in the Results section.

MODELING STRATEGY

The modeling strategy proposed for the statistical process monitoring of the spray dryer operation and the prediction of the product's dv50 is depicted in Figure 6. The aim of this

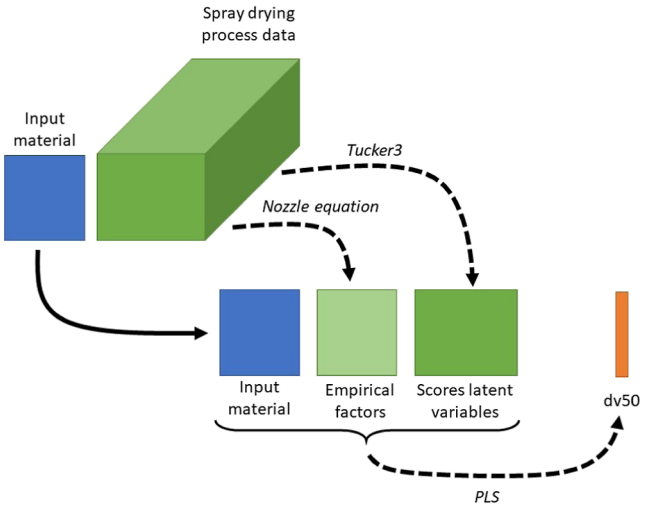


Figure 6. Modeling strategy for process monitoring and dv50 prediction.

modeling strategy is to first obtain a reliable description in a reduced space of the overall spray drying operation and then use this description to predict the median particle size of the product. The proposed strategy consists of two models acting in series. A process monitoring model was followed by a predictive model. The Tucker3 factorization is used as the model that describes the variability observed in the operation of the spray dryer. This model takes the time series of the spray drying PPs and evaluates the condition of every batch completed. The scores of the Tucker3 model are used as the fingerprints of the deterministic variability of the spray drying process for each batch. These scores are used as inputs in the second model of this strategy. The predictive model aims to relate variations in the input conditions, i.e., MAs and process conditions, with the variations in the particle size of the obtained product. A PLS model is trained for this purpose, and 3 different data sets are used as inputs. First, the material attributes, secondly, the scores of the Tucker3 model, and finally, a set of empirical factors. These empirical factors are included to decorrelate the variability observed in the flow conditions of the liquid feed through the spray nozzle.

Monitoring Model. The Tucker3 model was calibrated using a constrained version of the least-squares optimization problem to achieve simultaneously the best fit of the model and the optimal scale of the data. This strategy for simultaneous scaling and training was proposed by Muñoz et al.[32] and comprises a way to approximate evenly the variability exhibited by each PP. Equation 4 is the optimization formulation of this training strategy. The extra term in the optimization problem is added as a cost function of the variance of the residuals. The minimization of this cost function pushes the residuals in the approximation of each variable to have a unit variance distribution. S is the vector of scaling parameters that is optimized to obtain a solution to the problem. Each term in S is the scaling factor applied to each variable in the data set. This method has been shown to be effective at reducing the noise propagation and improving the

https://doi.org/10.1021/acsomega.3c08032
ACS Omega XXXX, XXX, XXX−XXX

rank approximation of the data; it also favors the capture of the deterministic variability by the model.

$$\min_{S,U,V,W} \|(X_{(2)}^{T} \odot S)T - VG_{(2)}(W \otimes U)T\|_{F}^{2} + \left\| \frac{\mathrm{diag}(E_{(2)}E_{(2)}^{T})}{IK} - 1 \right\|_{F}^{2} \quad (4)$$

The strategy for simultaneous scaling and training of the Tucker3 model exploits the structural decomposition characteristics of this tensor factorization method and improves the information captured in the model. Thus, this strategy addresses some of the limitations identified in previous benchmarking studies of the Tucker3 method.[46,47]

Predictive Model. The PLS model takes the scores of the Tucker3 model, the excipient attributes (Table 2), and a set of empirical factors as inputs. The empirical factors are included to account for the independent variability in the spray nozzle and the viscosity of the liquid feed. Since the variations in these conditions impact the flow of liquid through the nozzle, their effects are observed in variations of the flow rate or the pressure drop. However, the resulting effects on the flow condition are not sufficient to estimate the effect in the particle size of the product because each source of variability influences differently the particle size and the flow condition.

Empirical Factors. The empirical factors proposed in this work are derived from the empirical relation that describes flow through swirl nozzles given in eq 5.[40] This equation relates the volumetric flow through the nozzle with the nozzle's dimensions, the fluid's properties (i.e., viscosity and density), and the pressure drop. In large scale production setups, the data available regarding these parameters are limited. In this case, only the feed flow rate, the atomization pressure drop, and the liquid density are available. Thus, applying eq 5 directly to calibrate a model for this application is unfeasible. In fact, even in the scenario where all of the process conditions are known, this empirical expression could still fail to produce an accurate result because it assumes that there is no variability in the nozzle. As it will be shown in the Results section, the variability associated with the nozzle results in variability in the particle size even when geometrically identical nozzles are used. Similarly, small variations in the polymeric structure of the excipient could induce variability, mainly in the viscosity of the liquid formulations, which is not measured in industrial setups.

$$\frac{\dot{m}}{\mu_1 R_{in}} = 1.18\left(\frac{D_{or}}{R_{in}}\right)^{1.34}\left(\frac{A_{in}}{R_{in}^{2}}\right)^{0.35}\left(\frac{\sqrt{\Delta p \rho_1}\, R_{in}}{\mu_1}\right)^{0.885} \quad (5)$$

$$\dot{m} = 1.18\underbrace{\left[R_{in}^{1.885}\left(\frac{D_{or}}{R_{in}}\right)^{1.34}\left(\frac{A_{in}}{R_{in}^{2}}\right)^{0.35}\right]}_{\beta_{nozzle}}\underbrace{[(\sqrt{\Delta p})^{0.885}]}_{\beta_{\Delta p}}\underbrace{[(\sqrt{\rho_1})^{0.885}]}_{\beta_{\rho}}\underbrace{[\mu_1^{0.115}]}_{\beta_{\mu}} \quad (6)$$

β_{nozzle} and β_{μ} are the factors that, if computed, can quantify the variability of the system for which there is no direct data available. The variation in the nozzle originates from small differences in the adjustment of the nozzle when it is manipulated for cleaning and maintenance or due to the continual use resulting in fouling and wear out. This means that the effect of this variability can be expected to be significant only when looking at the overall difference between batches, and not within a single batch. Thus, β_{nozzle} can be computed as shown in eq 7. The variability from batch-to-batch can be captured, independently of other variations impacting the flow condition, by computing β_{nozzle}. Given that during the flushing step the atomized fluid consists only of solvents, the variability in the flow condition will respond mainly to variations introduced due to changes in the nozzle. Since the impact of viscosity and density can be neglected, β_{nozzle} can be estimated based on the solvent flow rate and the pressure drop.

$$\beta_{nozzle} = \frac{\dot{m}}{\beta_{\Delta p}\beta_{\rho}\beta_{\mu}} \quad (7)$$

The second factor, i.e., β_{μ}, is intended to account for variability in the viscosity of the liquid feed due to variability in the properties of the excipient. This empirical factor is related to the other terms, as shown in eq 8.

$$\beta_{\mu} = \frac{\dot{m}}{\beta_{nozzle}\beta_{\Delta p}\beta_{\rho}} \quad (8)$$

The relation between conditions during the flushing and drying steps can be used to eliminate some of the terms in eq 8. This is possible because the viscosity of the solvent mixture is invariant, no changes in the nozzle are expected to occur between the flushing and drying steps of the same batch, and the variations in density are expected to occur equally in the solvent and the feed mixture.

Model Training. The training algorithm of the PLS model was adapted to infer the missing values of dv50. First, the known median particle size of every PD batch is used as the starting guessed value of dv50 for every corresponding batch of SD material. Then, the PLS model is fitted, and a new estimate of the dv50 value for each batch of SD powder is obtained. These estimations are adjusted to guarantee that the dv50 values for the PD batches remain equal to the known values while keeping the inferred batch-to-batch variability. The adjusted dv50 values are then used to train the PLS model again. Thus, the PLS model is fitted multiple times, following this iterative strategy. After each iteration, the training dv50 values are updated based on the results of the previous iteration. The training algorithm stops when the change in the estimated values is negligible.

The weighted sum according to eq 9 is used to relate the PLS results for the median particle size of n SD batches to the particle size of the corresponding PD batch. The relative contribution in mass from every SD batch to the final PD batch (w_i^{SD}/w^{PD}) is the weighting parameter. The weight fraction is used under the assumption that the variation in the density of the product is negligible. Evidence collected on the boundary batches showed that this is a valid assumption. The ratio observed between the variation in the median particle size (dv50) and the variation in density is 10 [um] to 0.05 [g/mL]. This means that in the worst case, the variation in density could impact the final estimation of the particle size by ±0.3 [um]. This deviation is equivalent to around 15% of the error in the model. Thus, although the assumption made has an impact on the accuracy of the proposed model, this is a minor contribution to the overall error of the model.

https://doi.org/10.1021/acsomega.3c08032
ACS Omega XXXX, XXX, XXX−XXX

Table 4. Training, Cross-Validation, and Validation of the Tucker3 Models

formulation	best multilinear rank	data set	SSE	relative error (%)
A	[4,4,2]	training	60.6	73.9
		cross-validation	60.7	74.1
		validation	66.0	73.4
B	[6,6,2]	training	29.8	61.8
		cross-validation	32.0	65.8
		validation	28.7	76.1

$$\mathrm{dv50}^{\mathrm{PD}} = \frac{1}{w^{\mathrm{PD}}} \sum_{i=1}^{n} w_i^{\mathrm{SD}} \mathrm{dv50}_i^{\mathrm{SD}}$$

$$w^{\mathrm{PD}} = \sum_{i=1}^{n} w_i^{\mathrm{SD}} \tag{9}$$

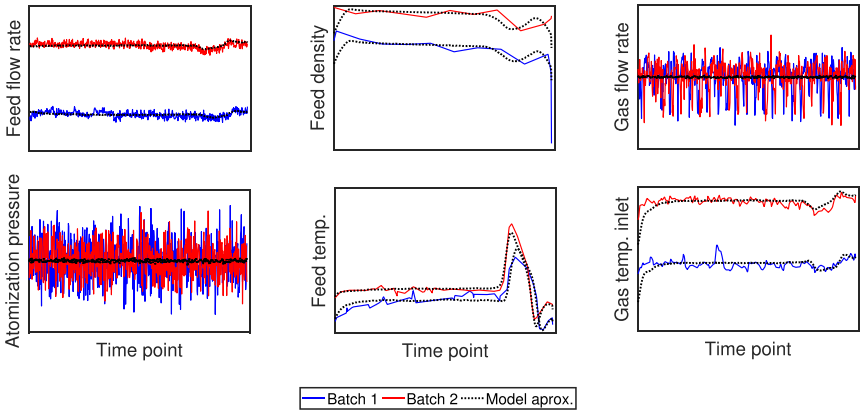
The update rule used at iteration k of the PLS training algorithm is given by eq 10. The dv50 values used to train the model at each iteration are the values estimated in the previous iteration, after being centered around the reference value of the corresponding PD batch. This approach guarantees that the dv50 value of each PD batch used in the model calibration remains equal to the value reported by the reference method ($\mathrm{dv50}_{\mathrm{ref}}^{\mathrm{PD}}$) across all iterations. Moreover, the variability present in the input space is used to infer the variability in the dv50 values of individual SD batches. The training algorithm stops when convergence is reached. The difference in the values computed for $\mathrm{dv50}_{i,\mathrm{new}}^{\mathrm{SD}}$ between two consecutive iterations is used to define the convergence criterion according to eq 11.

$$\mathrm{dv50}_{i,\mathrm{new}}^{\mathrm{SD}} = \mathrm{dv50}_{i,k}^{\mathrm{SD}} \frac{\mathrm{dv50}_{\mathrm{ref}}^{\mathrm{PD}}}{\mathrm{dv50}_{k}^{\mathrm{PD}}} \tag{10}$$

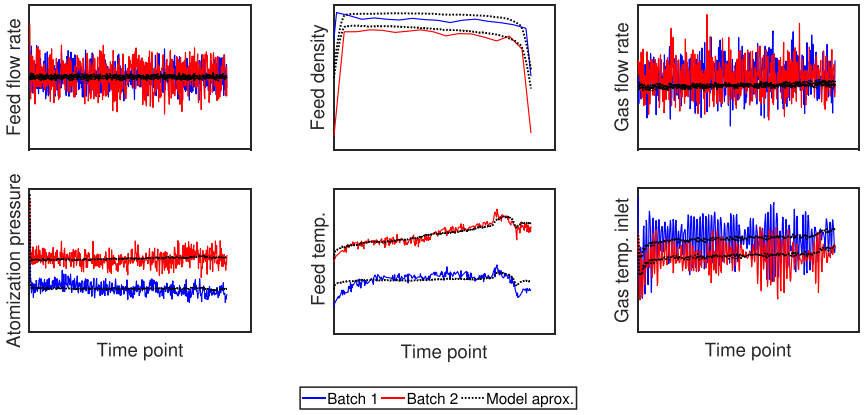
$$\max_i |\mathrm{dv50}_{i,\mathrm{new},k+1}^{\mathrm{SD}} - \mathrm{dv50}_{i,\mathrm{new},k}^{\mathrm{SD}}| < \alpha \tag{11}$$

RESULTS AND DISCUSSION

Tucker3 Process Monitoring Model. The Tucker3 models have been trained using 10-fold cross-validation to determine the best multilinear rank approximations of the spray drying process data. The method discussed by Louwerse et al.[44] was used to determine the set of feasible and nonredundant solutions. The strategy described in the section for simultaneous scaling and training was implemented to fit the Tucker3 models. Detailed results of the model training are presented in the Supporting Information. The best multilinear rank approximations are [4,4,2] for formulation A and [6,6,2] for formulation B. These results highlight some of the differences between the two processes and their data sets.



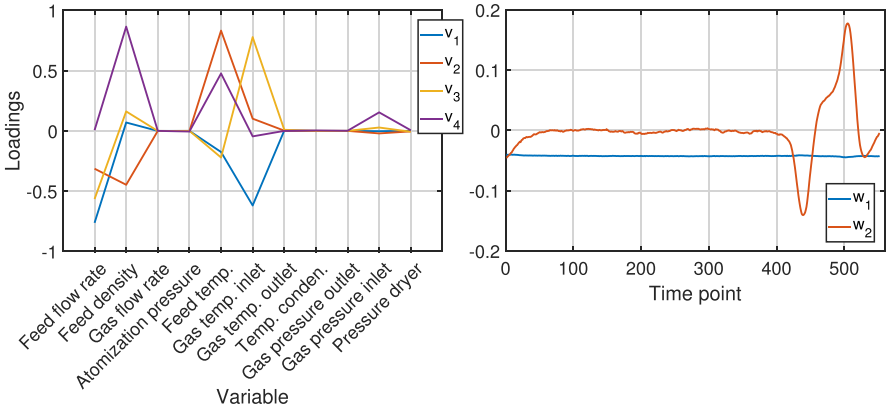
(a) Formulation A



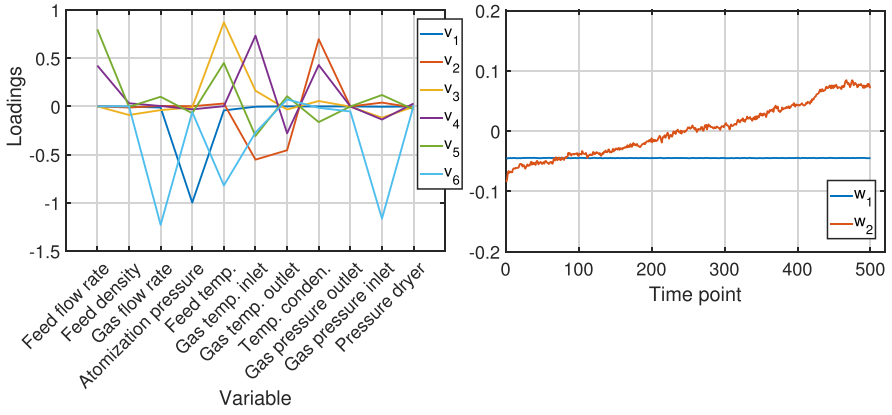
(b) Formulation B

Figure 7. Tucker3 model approximation of the PPs for two batches.

https://doi.org/10.1021/acsomega.3c08032
ACS Omega XXXX, XXX, XXX−XXX



(a) Formulation A



(b) Formulation B

Figure 8. Loadings of the Tucker3 models for the spray drying process of the two formulations. v_i and w_i are vectors of the factor matrices V and W respectively.

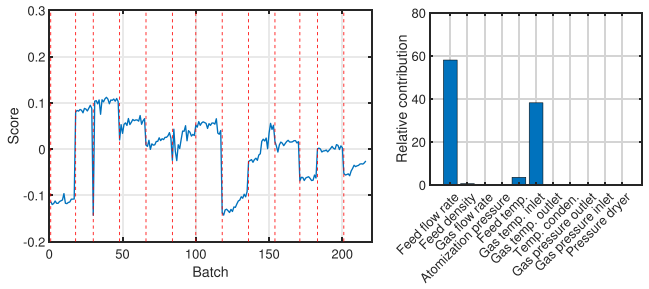
The higher ranks obtained for formulation B, in the first and second modes, are mostly related to the presence of introduction batches in the training set. The presence of independent sources of variation, introduced to test the boundary conditions of the process, demands additional factors to approximate the data set.

Table 4 summarizes the results obtained for the training and validation of the Tucker3 models for the two formulations. The relative error represents the fraction of variability in the data that is not captured by the models. Since the simultaneous scaling and training of the Tucker3 model favors capturing the deterministic variability of the process,[32] the unexplained variability is mostly associated with noise in the measurements and stochastic variability of the system. Thus, the relatively large errors obtained for the two formulations suggest that a significant part of the variability present in the data is not related to deterministic phenomena in the processes. The sum squared error (SSE) reported in Table 4 provides an estimate of the models' reconstruction error. The SSE is reported as the mean value per batch in each data set. The similarity in the magnitudes of the errors obtained from each model when applied to each data set demonstrates the robustness of the models. Figure 7a,b shows how the Tucker3 model approximates some of the PPs for two batches of each formulation. The results obtained in the case of the liquid feed and gas flow rates show how the model approximates the mean trajectory of the variables and leaves out the signals' noise and stochastic variability.

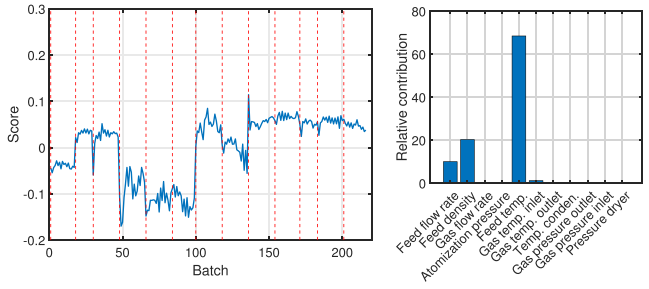
Figure 8 depicts two of the factor matrices in the Tucker3 models of the spray drying operation. Figure 8a corresponds to formulation A, while Figure 8b corresponds to formulation B. The loadings of the PPs are depicted at the left of each figure. These loadings are given by matrix V of the tensor factorization, with $r_2 = 4$ and $r_2 = 6$ for formulations A and B, respectively. Each of these loadings contains information on the numerical correlations among PPs. In the case of formulation A, the process variability occurs mainly in the feed stream, its flow rate, its density, and temperature, but also in the gas temperature at the inlet. In contrast, the model of formulation B shows that variability has been introduced or is present in almost every PP. The main reason for this condition is the inclusion of the introduction batches. However, despite the diversity in the variability of the process for formulation B, the variance in the feed density for this formulation is less pronounced when compared to formulation A. The interpretability of these results demonstrates the added value of the information captured by these models.

The loadings of the PPs are the mixture patterns applied to the trajectories obtained with the time loadings. The product of the two sets of loadings results in the mean trajectory of every variable in the process. Therefore, the trajectories contained in the time loadings can be seen as the linear basis for the dynamic behavior of the process. The plots at the right

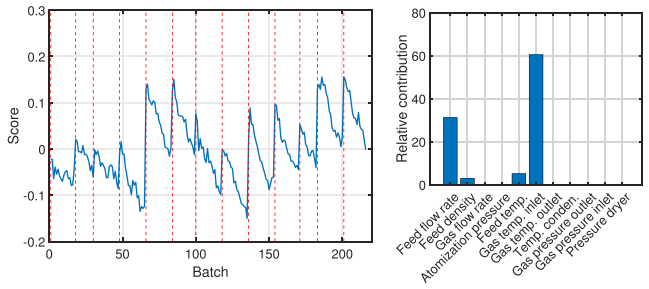
https://doi.org/10.1021/acsomega.3c08032
ACS Omega XXXX, XXX, XXX−XXX



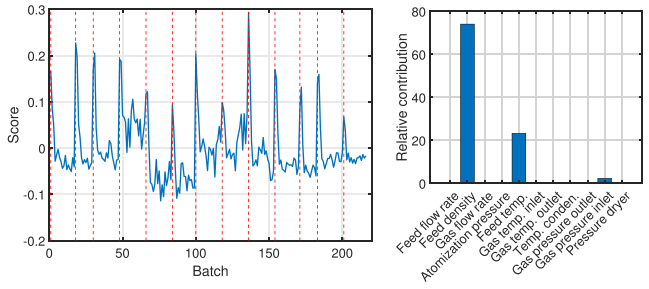
(a) First Latent Variable



(b) Second Latent Variable



(c) Third Latent Variable



(d) Fourth Latent Variable

Figure 9. Scores and contributions for each LV of the tucker3 model for formulation A.

in Figure 8 correspond to the time loadings of each model, i.e., matrix W of the tensor factorization, with $r_3 = 2$ for both formulations. Two main behaviors are observed in the linear basis obtained for formulation A. The first loading reflects the steady state condition of this process. The second loading also reflects, for the most part, a steady-state condition. However, this loading also shows some dynamic variability at the start and end of the process. The results obtained in the case of formulation B are partially similar. The first loading is almost identical in both cases. However, the second loading of formulation B reflects another dynamic behavior. The second loading can be seen as the linear basis for the monotonic change in one or more variables of the process. The difference between the second time loadings of each model means that these processes experience differences in the dynamic behaviors.

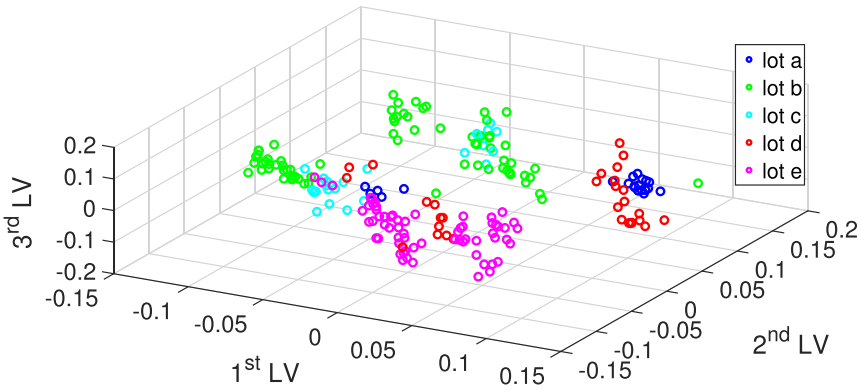
Figure 9 depicts the scores obtained for the training batches of formulation A. These are obtained as matrix U of the Tucker3 model and contain information about the batch-to-batch variability. In Figure 9, the scores for every latent variable (LV) of the model are plotted separately, and these are accompanied by the corresponding contribution plot. The vertical red dashed lines are used to mark the changes in the production campaign. Thus, these plots allow provide a visualization of trajectories and differences among batches and production campaigns. The trends, trajectories, and events that are identified in these plots reflect the deterministic variability occurring in the process. Although this variability remains inside the normal operating range (NOR) for formulation A, it is important to understand the sources of uncontrolled variation and characterize their impact on the product's particle size.

The variation captured in the scores of every LV is analyzed with the help of contributions from the PPs to each of the latent variables. Thus, the periodic variability described by the third and fourth LVs (Figure 9c,d) can be associated with the recurrent changes that occur in each campaign in the gas temperature and the density of the liquid feed. The scores of the third LV (Figure 9c) show that variations in the gas temperature occur from the start to the end of every campaign, following an almost monotonic trend. This means that the gas temperature varies in the same way in every campaign, i.e., increasing with every new batch of the campaign. The scores of the fourth LV (Figure 9d) show the effect of solvent recuperation on the process. The manufacturing strategy for formulation A includes the use of recovered solvent. The first two batches of every campaign use a mixture of pure solvents, while the remaining batches use the solvent recovered from previous batches. Although the process includes a step to adjust the density of the recovered solvent mixture, the trend displayed by the scores shows that this mode of operation induces some variability in the feed density. The peak observed in the scores of this LV, at the beginning of every campaign, corresponds with the two first batches, which are the ones running with a mixture of pure solvents.

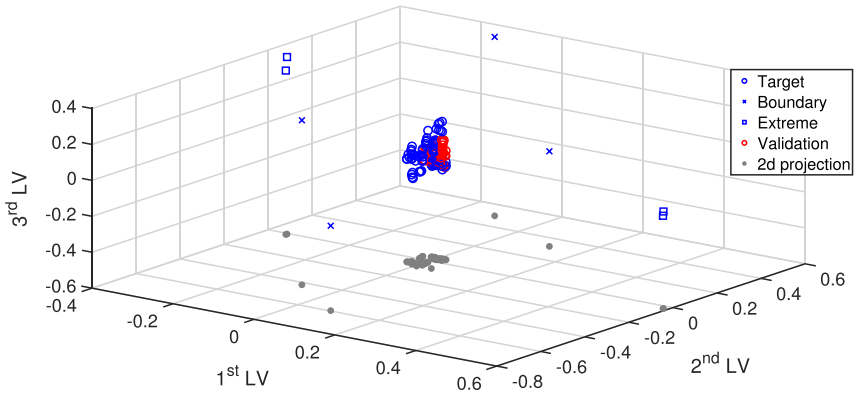
Regarding the scores of the first and second LVs (Figure 9a,b), the contribution plots show that the variability corresponds to changes in the feed flow rate and the feed temperature. These variations are not structured as those seen in the other scores. However, apart from a few extreme values, the variations in these LVs seem to be significant when looking across different production campaigns and not between batches of the same campaign. This suggests that the sources of variation have a low frequency of change. In the case of the feed flow rate, the variation was found to be associated with the change in the excipient's lot and the manipulation of the spray nozzle for cleaning or maintenance tasks.

The score plot shown in Figure 10a is used to visualize the relative location of each batch in the latent space of the Tucker3 model. This figure depicts the scores for the first 3 LVs of the Tucker3 model. These 3 LVs account for 99.3% of the variability captured by the model, which in turn is 26.1% of the variability in the training data set. The colors used in this score plot indicate the excipient's lot used in each case. The presence of groups of batches (i.e., clusters) in the model's space corroborates once more that the spray drying process

https://doi.org/10.1021/acsomega.3c08032
ACS Omega XXXX, XXX, XXX−XXX



(a) Formulation A. Colors given by the lot of excipient used in each batch.



(b) Formulation B. 2d projections of the scores added to improve visualization.

Figure 10. Score plot for the first 3 LVs of the Tucker3 models.

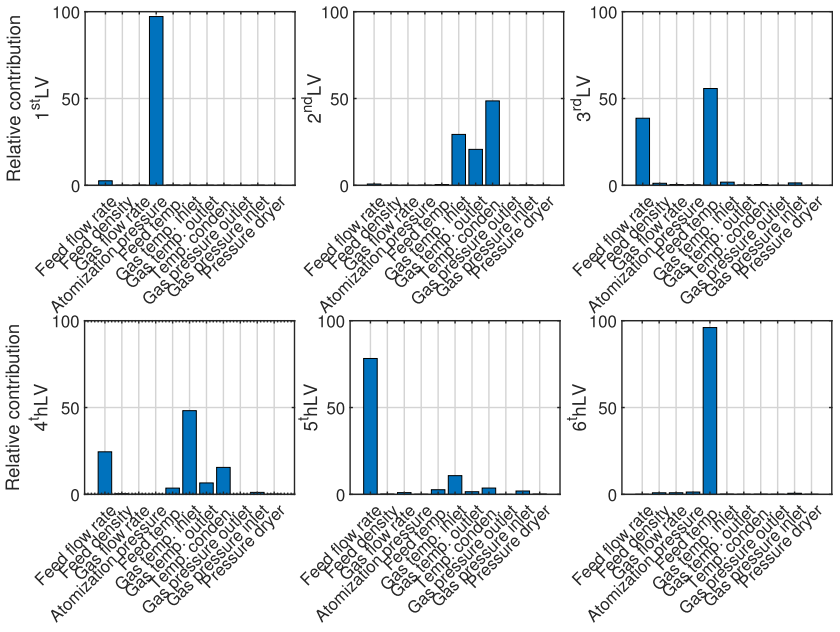


Figure 11. Contributions to the LVs of the Tucker3 model of formulation B.

operates under uncontrolled deterministic variability, but in all cases inside the NOR of the process. The fact that some of the clusters can be discriminated by the excipient's lot used suggests that a correlation exists between the variation in the process and the excipient's lot. However, there are also some batches produced using the same material lot that appear in separated clusters; this suggests that another source of variation might be impacting the process.

https://doi.org/10.1021/acsomega.3c08032
ACS Omega XXXX, XXX, XXX−XXX

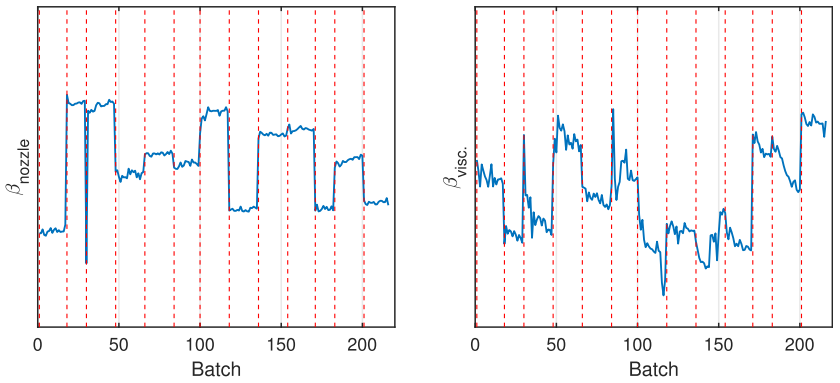


Figure 12. Empirical factors for the training data set of formulation A.

Figure 10b shows the score plot obtained for formulation B. In this case, the results highlight the differences between the batches produced around target conditions inside the NOR, and the batches completed during the process introduction. The results demonstrate that the operation at target conditions occurs well inside the design space defined by the batches performed at the borders of the proven acceptable ranges. Additionally, the 4 extreme batches, which were produced using different spray nozzles, are outside the design space of the process.

The variability present on the introduction data of formulation B can be seen in Figure 10b along the first and second LVs. The study of the contribution plots in Figure 11 shows that the model has captured the variation in the flow condition through the nozzle, i.e., atomization pressure and feed flow rate, in the first and the third to fifth LVs, respectively. Variations in the gas temperature are seen in the second LV. On the one hand, this explains the relative location of the extreme batches in the score plot, which are located at the extremes of the first LV but with no relative change in the second and third LV. On the other hand, the boundary batches show variation in all 3 LVs. This is the expected behavior because the conditions of these batches were designed to explore the variation in multiple PPs.

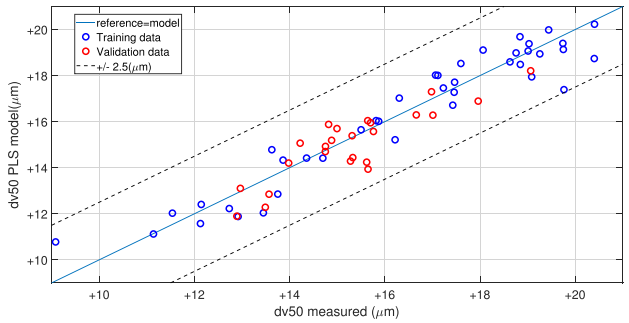
The assessment of the contribution plots for formulations A and B (Figures 9 and 11) with respect to the PPs depicted in Figure 7a,b helps to illustrate the validity of the contribution plots. The batches depicted in Figure 7a,b are batches produced around target conditions for both formulations; these illustrate the differences and similarities between the two processes. In formulation A, the variability in the flow condition through the nozzle is observed in the feed flow rate and not the atomization pressure, which is why the first LV for this formulation carries the contribution from this PP (Figures 7a and 9). The opposite is observed for formulation B in Figures 7b and 11. Variability on the feed density is more significant in formulation A, which explains why a LV of its model characterizes this variability. The feed temperature and the gas inlet temperature vary in both processes, while the gas flow rate does not in either of them. The contribution plots show how in both formulations there are contributions from the feed and gas inlet temperature, while there are no contributions from the gas flow rate.

Regression Model. Figure 12 depicts the empirical factors obtained for the training batches of formulation A. These, as explained previously, aim to approximate the variability in the viscosity of the liquid feed and in the state of the spray nozzle. The vertical dashed lines in Figure 12 mark the changes in the production campaign. These results show that the variation in the nozzle factor occurs when one production campaign ends and another starts. These changes correspond to times when the spray nozzle was replaced or handled for cleaning and maintenance. On the contrary, the viscosity factor shows variation both between campaigns and within certain campaigns. These variations are expected since they reflect the changes in the viscosity of the feed mixture due to differences in the excipient properties when different lots are used.

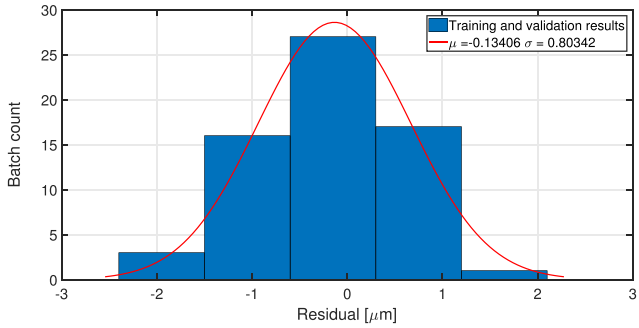
Since the dv50 of the individual SD batches is unknown, the iterative training strategy described in the Model training section was implemented to infer these values from the variability captured by the model and the know dv50 values of the PD batches. The PLS model was trained by using cross-validation to determine the number of LVs. Detailed results of the PLS model training are presented in the Supporting Information. A PLS model with 3 LVs was chosen in the case of formulation A. Backward feature elimination was used to identify the subset of input variables that resulted in the most accurate regression model. The elimination of 3 properties of the excipient, i.e., the viscosity in aqueous solution, the content of one of the substituent groups, and the pH, resulted in the improved accuracy of the PLS model. It is important to notice that the viscosity in aqueous solution, which is the standard measure of viscosity for the excipient material, is only mildly correlated with β_μ and it was found to be uninformative for the prediction of the variability in the particle size of the product.

Figure 13 shows the results of the PLS model prediction. The predicted dv50 values for the particulate product, i.e., the PD powder resulting from multiple combined batches of SD powder, are plotted against the reference value determined by laser diffraction. These figures show the accuracy of the model and the good agreement between the results obtained for the training and the validation data sets. Reference lines that mark deviations of ± 2.5 μm are plotted to highlight the accuracy of the model predictions. The accuracy and linearity of the dv50 PLS regression models can be assessed from these results. High linearity is evidenced in the goodness of fit given by the R^2 values of the linear relation between the predicted and measured dv50 values. The R^2 value obtained in the case of formulation A is 0.898, while it is 0.969 for formulation B. The model accuracy is assessed based on the distribution of the residuals. These residuals are shown in Figure 13b,d. In both cases, the models yield estimations with very low bias and

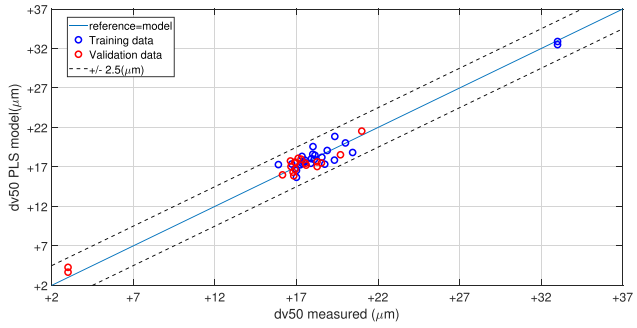
https://doi.org/10.1021/acsomega.3c08032
ACS Omega XXXX, XXX, XXX−XXX



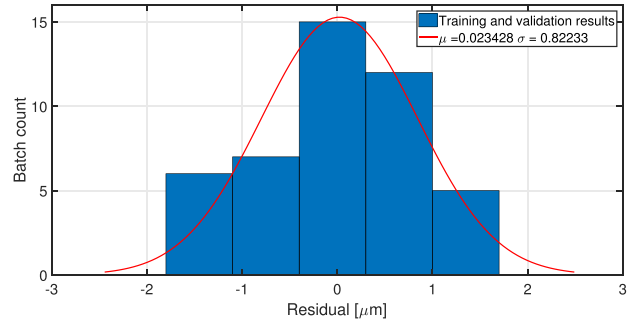
(a) PLS model prediction Formulation A



(b) PLS model residuals Formulation A



(c) PLS model prediction Formulation B



(d) PLS model residuals Formulation B

Figure 13. dv50 PLS model predictions and residuals for the SD particulate products.

variance, and the error distribution is normally distributed with $|\mu| < 0.15$ μm and $\sigma < 0.9$ μm.

In the case of formulation B, the extreme introduction batches that resulted in the lowest particle sizes ($\approx dv50_{min} + 2$ μm) were intentionally added to the validation set. The aim was to evaluate whether the correlations learned by the model are representative enough. The accuracy of the prediction for these extreme cases proved that the empirical factor used to capture the variability in the nozzle together with the correlations captured by the PLS model are consistent with the actual response of the system.

Finally, the available particle size data from the inline PAT was used to validate the estimation obtained by the model on individual batches of SD material (formulation A). The consistency between the inline measurement obtained from the PAT and the offline release testing was verified first. As reported by Medendorp et al.,[48] the two methods, i.e., offline and inline laser diffraction, result in different dv50 estimations, but the two are linearly correlated. The inline measurement normally underestimates the large particles, resulting in a lower dv50. The linear equation that characterizes the relation between the two methods was used as a reference to evaluate the estimations obtained by using the predictive model. Figure 14, shows the results in terms of the dv50 model estimation for the SD batches as well as the PAT measurements. These results demonstrate the high accuracy of the inference done via a data-driven modeling strategy. The variability present in the process and the input excipient material was correctly captured by the modeling strategy. The effects from variations in feed density, nozzle, and excipient material, as discussed before, are observed in the results from the PAT estimation, and they are well reproduced by the developed model.

CONCLUSIONS

A novel data-driven modeling strategy for the spray drying process in the large scale production of pharmaceuticals has been presented. This strategy follows a hierarchical approach that consists of two models, a Tucker3 process monitoring model and a PLS model for the prediction of the median particle size. The strategy used for simultaneous data scaling and training of the Tucker3 model resulted in highly deterministic and interpretable monitoring models for the two formulations. The loadings serve to generate a basic understanding of the process for each formulation, and the scores of each LV provided a detailed view on the factorized batch-to-batch variability. Additionally, the LVs served as good predictors for variability in the particle size of the product. The regression model takes the LVs of the Tucker3 model, the CMAs of the excipient, and two empirical factors as inputs to predict the variability in the median particle size of the product. The two empirical factors were formulated to help decorrelate the variations in the flow through the spray nozzle. These factors discriminate between changes in the spray nozzle and variations in the viscosity of the feed flow, which could originate from variations in the excipient properties. The results on the training and validation data sets for the two formulations have shown that the proposed modeling strategy is successful at describing the variation observed in the process and also at predicting the variation in the output. The results showed that the dv50 prediction is very accurate, with 99% of the training and validation errors below 2.5 μm, which is less than 10% of the allowed range of variation. The strategy designed to infer the missing output values resulted in a highly accurate model with respect to the reference release method and the inline PAT. Regarding the spray drying process and

https://doi.org/10.1021/acsomega.3c08032
ACS Omega XXXX, XXX, XXX−XXX

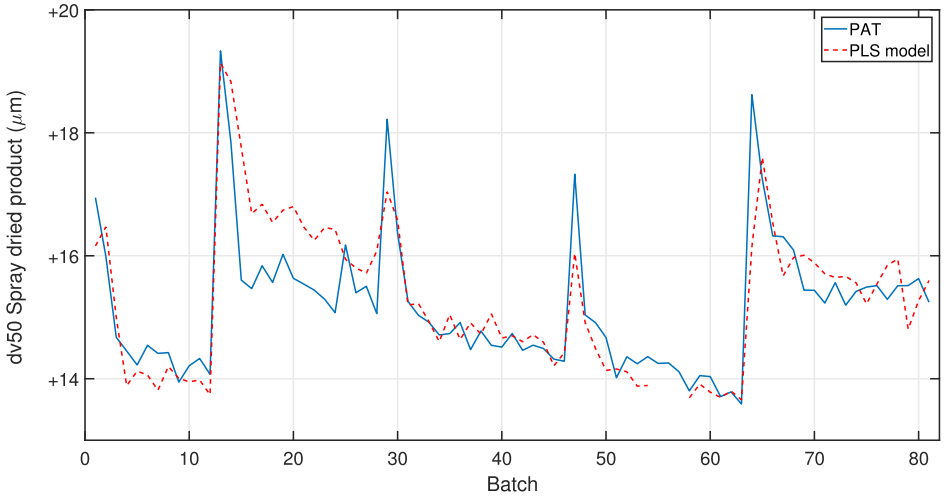


Figure 14. dv50 for individual batches of SD material, model inference vs PAT estimation.

the response of the particle size, this work has demonstrated that the variability in the spray nozzle, viscosity, and density of the liquid feed can have a significant impact on the uncontrolled variability of the particle size and that the method proposed here is able to capture these relations from the production data. This modeling strategy and the corresponding model structure were validated, proving to be accurate and reliable enough to replace traditional laboratory testing. Strategies such as right first-time and RTR can be supported in large-scale production based on these methods.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acsomega.3c08032.

Model training results (PDF)

■ AUTHOR INFORMATION

Corresponding Author

Jan Van Impe − *BioTeC+ Chemical & Biochemical Process Technology & Control, Campus Gent, KU Leuven, 9000 Ghent, Belgium*; orcid.org/0000-0002-5904-1638; Email: jan.vanimpe@kuleuven.be

Authors

Carlos André Muñoz López − *BioTeC+ Chemical & Biochemical Process Technology & Control, Campus Gent, KU Leuven, 9000 Ghent, Belgium*

Kristin Peeters − *Technical Operations, Geel Chemical Production Site, Janssen Pharmaceutica, J&J, 2440 Geel, Belgium*

Complete contact information is available at: https://pubs.acs.org/10.1021/acsomega.3c08032

Notes

The authors declare the following competing financial interest(s): Carlos Andre Munoz Lopez and Kristin Peeters are currently employees of Johnson and Johnson.

■ ACKNOWLEDGMENTS

This work was supported by the Fund for Scientific Research Flanders (FWO), project G0B4121N. CAML holds a VLAIO-Baekeland [HBC.2017.0239] grant.

■ REFERENCES

(1) Kim, J. Y.; Chun, M. H.; Choi, D. H. Control strategy for process development of high-shear wet granulation and roller compaction to prepare a combination drug using integrated quality by design. *Pharmaceutics* **2021**, *13*, 80.

(2) von Stosch, M.; Schenkendorf, R.; Geldhof, G.; Varsakelis, C.; Mariti, M.; Dessoy, S.; Vandercammen, A.; Pysik, A.; Sanders, M. Working within the design space: Do our static process characterization methods suffice? *Pharmaceutics* **2020**, *12*, 562−615.

(3) Djuris, J.; Djuric, Z. Modeling in the quality by design environment: Regulatory requirements and recommendations for design space and control strategy appointment. *Int. J. Pharm.* **2017**, *533*, 346−356.

(4) Paudel, A.; Worku, Z. A.; Meeus, J.; Guns, S.; Van Den Mooter, G. Manufacturing of solid dispersions of poorly water soluble drugs by spray drying: Formulation and process considerations. *Int. J. Pharm.* **2013**, *453*, 253−284.

(5) Singh, A.; Van den Mooter, G. Spray drying formulation of amorphous solid dispersions. *Adv. Drug Delivery Rev.* **2016**, *100*, 27−50.

(6) Malamatari, M.; Charisi, A.; Malamataris, S.; Kachrimanis, K.; Nikolakakis, I. Spray Drying for the Preparation of Nanoparticle-Based Drug Formulations as Dry Powders for Inhalation. *Processes* **2020**, *8*, 788.

(7) Tambe, S.; Jain, D.; Meruva, S. K.; Rongala, G.; Juluri, A.; Nihalani, G.; Mamidi, H. K.; Nukala, P. K.; Bolla, P. K. Recent Advances in Amorphous Solid Dispersions: Preformulation, Formulation Strategies, Technological Advancements and Characterization. *Pharmaceutics* **2022**, *14*, 2203.

(8) Salama, A. H. Spray drying as an advantageous strategy for enhancing pharmaceuticals bioavailability. *Drug Delivery Transl. Res.* **2020**, *10*, 1−12.

(9) Vicente, J.; Pinto, J.; Menezes, J.; Gaspar, F. Fundamental analysis of particle formation in spray drying. *Powder Technol.* **2013**, *247*, 1−7.

(10) Mezhericher, M.; Levy, A.; Borde, I. Spray drying modelling based on advanced droplet drying kinetics. *Chem. Eng. Process.* **2010**, *49*, 1205−1213.

(11) Breinlinger, T.; Hashibon, A.; Kraft, T. Simulation of the in fl uence of surface tension on granule morphology during spray drying using a simple capillary force model. *Powder Technol.* **2015**, *283*, 1−8.

(12) Seydel, P.; Blömer, J.; Bertling, J. Modeling Particle Formation at Spray Drying Using Population Balances. *Dry. Technol.* **2007**, *24*, 137.

(13) Milanesi, A.; Rizzuto, F.; Rinaldi, M.; Foglio Bonda, A.; Segale, L.; Giovannelli, L. Thermodynamic Balance vs. Computational Fluid

Dynamics Approach for the Outlet Temperature Estimation of a Benchtop Spray Dryer. *Pharmaceutics* **2022**, *14*, 296.
(14) Cotabarren, I. M.; Bertín, D.; Razuc, M.; Ramírez-Rigo, M. V.; Piña, J. Modelling of the spray drying process for particle design. *Chem. Eng. Res. Des.* **2018**, *132*, 1091−1104.
(15) Lisboa, H. M.; Duarte, M. E.; Cavalcanti-Mata, M. E. Modeling of food drying processes in industrial spray dryers. *Food Bioprod. Process.* **2018**, *107*, 49−60.
(16) Oakley, D. E. Spray dryer modeling in theory and practice. *Dry. Technol.* **2004**, *22*, 1371−1402.
(17) Negiz, A.; Lagergren, E. S.; Cinar, A. Mathematical Models of Cocurrent Spray Drying. *Ind. Eng. Chem. Res.* **1995**, *34*, 3289−3302.
(18) Mohajel, N.; Najafabadi, A. R.; Azadmanesh, K.; Vatanara, A.; Moazeni, E.; Rahimi, A.; Gilani, K. Optimization of a spray drying process to prepare dry powder microparticles containing plasmid nanocomplex. *Int. J. Pharm.* **2012**, *423*, 577−585.
(19) Razuc, M.; Piña, J.; Ramírez-Rigo, M. V. Optimization of Ciprofloxacin Hydrochloride Spray-Dried Microparticles for Pulmonary Delivery Using Design of Experiments. *AAPS PharmSciTech* **2018**, *19*, 3085−3096.
(20) Cheow, W. S.; Li, S.; Hadinoto, K. Spray drying formulation of hollow spherical aggregates of silica nanoparticles by experimental design. *Chem. Eng. Res. Des.* **2010**, *88*, 673−685.
(21) Huang, J.; Kaul, G.; Cai, C.; Chatlapalli, R.; Hernandez-Abad, P.; Ghosh, K.; Nagi, A. Quality by design case study: An integrated multivariate approach to drug product and process development. *Int. J. Pharm.* **2009**, *382*, 23−32.
(22) Kumar, S.; Gokhale, R.; Burgess, D. J. Quality by Design approach to spray drying processing of crystalline nanosuspensions. *Int. J. Pharm.* **2014**, *464*, 234−242.
(23) Sloth, J.; Jørgensen, K.; Bach, P.; Jensen, A. D.; Kiil, S.; Dam-Johansen, K. Spray drying of suspensions for pharma and bio products: Drying kinetics and morphology. *Ind. Eng. Chem. Res.* **2009**, *48*, 3657−3664.
(24) Sturm, D. R.; Moser, J. D.; Sundararajan, P.; Danner, R. P. Spray drying of hypromellose acetate succinate. *Ind. Eng. Chem. Res.* **2019**, *58*, 12291−12300.
(25) Gil-Chávez, J.; Padhi, S. S. P.; Hartge, U.; Heinrich, S.; Smirnova, I. Optimization of the spray-drying process for developing aquasolv lignin particles using response surface methodology. *Adv. Powder Technol.* **2020**, *31*, 2348−2356.
(26) Meneghetti, N.; Facco, P.; Bezzo, F.; Himawan, C.; Zomer, S.; Barolo, M. Knowledge management in secondary pharmaceutical manufacturing by mining of data historians - A proof-of-concept study. *Int. J. Pharm.* **2016**, *505*, 394−408.
(27) Luo, L.; Bao, S.; Mao, J.; Tang, D. Phase Partition and Phase-Based Process Monitoring Methods for Multiphase Batch Processes with Uneven Durations. *Ind. Eng. Chem. Res.* **2016**, *55*, 2035−2048.
(28) Muñoz López, C. A.; Bhonsale, S.; Peeters, K.; Van Impe, J. F. M. Manifold Learning and Clustering for Automated Phase Identification and Alignment in Data Driven Modeling of Batch Processes. *Front. Chem. Eng.* **2020**, *2*, 1−18.
(29) Davidopoulou, C.; Ouranidis, A. Pharma 4.0-Artificially Intelligent Digital Twins for Solidified Nanosuspensions. *Pharmaceutics* **2022**, *14*, 2113.
(30) Ramos, M.; Ascencio, J.; Hinojosa, M. V.; Vera, F.; Ruiz, O.; Jimenez-Feijoó, M. I.; Galindo, P. Multivariate statistical process control methods for batch production: a review focused on applications. *Prod. Manuf. Res.* **2021**, *9*, 33−55.
(31) Ebadi, M.; Chenouri, S.; Lin, D. K.; Steiner, S. Statistical monitoring of the covariance matrix in multivariate processes: A literature review. *J. Qual. Technol.* **2022**, *54*, 269−289.
(32) Muñoz, C. A.; Telen, D.; Nimmegeers, P.; Impe, J. V. Feature extraction for batch process monitoring and fault detection via simultaneous data scaling and training of tensor based models. *IFAC-PapersOnLine* **2018**, *51*, 433−440.
(33) Fanaee-T, H.; Gama, J. Tensor-based anomaly detection: An interdisciplinary survey. *Knowl. Base Syst.* **2016**, *98*, 130−147.
(34) Sun, W.; Braatz, R. D. Opportunities in tensorial data analytics for chemical and biological manufacturing processes. *Comput. Chem. Eng.* **2020**, *143*, 107099.
(35) Rajalahti, T.; Kvalheim, O. M. Multivariate data analysis in pharmaceutics: A tutorial review. *Int. J. Pharm.* **2011**, *417*, 280−290.
(36) Van Snick, B.; Grymonpré, W.; Dhondt, J.; Pandelaere, K.; Di Pretoro, G.; Remon, J. P.; De Beer, T.; Vervaet, C.; Vanhoorne, V. Impact of blend properties on die filling during tableting. *Int. J. Pharm.* **2018**, *549*, 476−488.
(37) Mathe, R.; Casian, T.; Tomuţă, I. Multivariate feed forward process control and optimization of an industrial, granulation based tablet manufacturing line using historical data. *Int. J. Pharm.* **2020**, *591*, 119988.
(38) Poutiainen, S.; Matero, S.; Hämäläinen, T.; Leskinen, J.; Ketolainen, J.; Järvinen, K. Predicting granule size distribution of a fluidized bed spray granulation process by regime based PLS modeling of acoustic emission data. *Powder Technol.* **2012**, *228*, 149−157.
(39) Bhonsale, S.; Muñoz López, C. A.; Van Impe, J. Global sensitivity analysis of a spray drying process. *Processes* **2019**, *7*, 562.
(40) Wimmer, E.; Brenn, G. Viscous effects on flows through pressure-swirl atomizers. *ICLASS 2012, 12th Triennial International Conference on Liquid Atomization and Spray Systems*: Heidelberg, Germany, 2012.
(41) Mašková, E.; Kubová, K.; Raimi-Abraham, B. T.; Vllasaliu, D.; Vohlídalová, E.; Turánek, J.; Mašek, J. Hypromellose − A traditional pharmaceutical excipient with modern applications in oral and oromucosal drug delivery. *J. Controlled Release* **2020**, *324*, 695−727.
(42) Mallat, S. G. A. A theory for multiresolution signal decomposition: the wavelet representation. *IEEE Trans. Pattern Anal. Mach. Intell.* **1989**, *11*, 674−693.
(43) Sidiropoulos, N. D.; De Lathauwer, L.; Fu, X.; Huang, K.; Papalexakis, E. E.; Faloutsos, C. Tensor Decomposition for Signal Processing and Machine Learning. *IEEE Trans. Signal Process.* **2017**, *65*, 3551−3582.
(44) Louwerse, D. J.; Smilde, A. K.; Kiers, H. A. L. Cross-validation of multiway component models. *J. Chemom.* **1999**, *13*, 491−510.
(45) Vervliet, N.; Debals, O.; Sorber, L.; Van Barel, M.; De Lathauwer, L. *Tensorlab 3.0*, 2016. https://www.tensorlab.net.
(46) Liu, J.; Wang, D.; Chen, J. Monitoring Framework Based on Generalized Tensor PCA for Three-Dimensional Batch Process Data. *Ind. Eng. Chem. Res.* **2020**, *59*, 10493−10508.
(47) Westerhuis, J. a.; Kourti, T.; Macgregor, J. F. Comparing alternative approaches for multivariate statistical analysis of batch process data. *J. Chemom.* **1999**, *13*, 397−413.
(48) Medendorp, J.; Bric, J.; Connelly, G.; Tolton, K.; Warman, M. Development and beyond: Strategy for long-term maintenance of an online laser diffraction particle size method in a spray drying manufacturing process. *J. Pharm. Biomed. Anal.* **2015**, *112*, 79−84.

https://doi.org/10.1021/acsomega.3c08032
ACS Omega XXXX, XXX, XXX−XXX