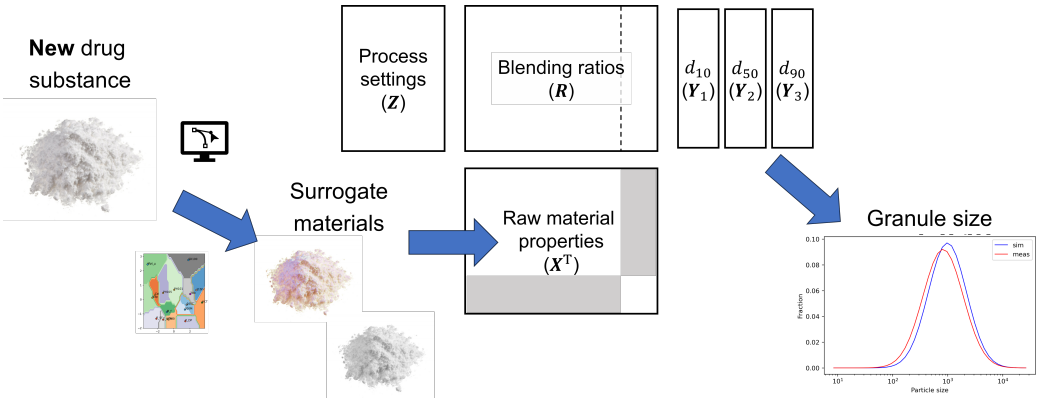
Graphical Abstract

T-shaped partial least squares for high-dosed new active pharmaceutical ingredients in continuous twin-screw wet granulation: granule size prediction with limited material information

Kensaku Matsunami, Jonathan Meyer, Martin Rowland, Neil Dawson, Thomas De Beer, Daan Van Hauwermeiren



Highlights

T-shaped partial least squares for high-dosed new active pharmaceutical ingredients in continuous twin-screw wet granulation: granule size prediction with limited material information

Kensaku Matsunami, Jonathan Meyer, Martin Rowland, Neil Dawson, Thomas De Beer, Daan Van Hauwermeiren

- Granule size was predicted for high-dosed new drug substances.
- Nonlinear classification algorithms identified surrogate materials of new active pharmaceutical ingredients.
- Surrogate material selection was possible with limited material information.
- T-shaped partial least squares successfully predicted granule size for new formulations.
- The impact of materials, blending ratios, and process settings on granule size was analyzed.

T-shaped partial least squares for high-dosed new active pharmaceutical ingredients in continuous twin-screw wet granulation: granule size prediction with limited material information

Kensaku Matsunami[a,b,*], Jonathan Meyer[c], Martin Rowland[c], Neil Dawson[c], Thomas De Beer[a], Daan Van Hauwermeiren[a,b]

[a] *Laboratory of Pharmaceutical Process Analytical Technology, Department of Pharmaceutical Analysis, Ghent University, Ottergemsesteenweg 460, Ghent, 9000, Oost-Vlaanderen, Belgium*
[b] *BIOMATH, Department of Data Analysis and Mathematical Modelling, Ghent University, Coupure links 653, Ghent, 9000, Oost-Vlaanderen, Belgium*
[c] *Worldwide Research and Development, Pfizer Inc., Sandwich, Kent, , , UK*

Abstract

This work presents a granule size prediction approach applicable to diverse formulations containing new active pharmaceutical ingredients (APIs) in continuous twin-screw wet granulation. The approach consists of a surrogate selection method to identify similar materials with new APIs and a T-shaped partial least squares (T-PLS) model for granule size prediction across varying formulations and process conditions. We devised a surrogate material selection method, employing a combination of linear pre-processing and nonlinear classification algorithms, which effectively identified suitable surrogates for new materials. Using only material properties obtained through four characterization methods, our approach demonstrated its predictive prowess. The selected surrogate methods were seamlessly integrated with our developed T-PLS model, which was meticulously validated for high-dose formulations involving three new APIs. When surrogating new APIs based on Gaussian process classification, we achieved the lowest prediction errors, signifying the method's robustness. The predicted d-values were within the range of uncertainty bounds for all cases, except for d90 of API C. Notably, the ap-

*Corresponding author

proach offers a direct and efficient solution for early-phase formulation and process development, considerably reducing the need for extensive experimental work. By relying on just four material characterization methods, it streamlines the research process while maintaining a high degree of accuracy.

Keywords: Continuous manufacturing, wet granulation, data-driven model, surrogate material selection, nonlinear classification, process and product design.

List of Abbreviations

API	active pharmaceutical ingredient
CQA	critical quality attribute
CV	cross-validation
DoE	Design of Experiments
GPC	Gaussian process classifier
L/S	liquid-to-solid
LV	latent variable
MFR	mass feed rate
PCA	principal component analysis
PLS	partial least squares
rbf	radial basis function
SVC	support vector classification
TPLS	T-shaped partial least squares
TSWG	twin-screw wet granulation

1. Introduction

The pharmaceutical industry has been facing the problem of large and expensive experimental designs in the formulation and process design phase of solid dosage forms. In this phase, active pharmaceutical ingredients (APIs) are expensive, and the available quantity is limited [1]. To reduce the necessary amount of experiments in the whole process design, continuous manufacturing has been considered as a solution because less scale-up studies are required compared to conventional batch production [2, 3, 4]. Numerous studies have been performed to understand the mechanisms of continuous manufacturing at the unit operation level. One of the most focused unit operations is twin-screw wet granulation (TSWG) [5, 6, 7].

In spite of numerous studies as well as the advancement of continuous manufacturing, challenges still exist in the design phase. Critical quality attributes (CQAs) of granules are influenced by selected materials and the composition ratio of each material as well as process settings. Due to combinatorial explosion, it is not possible to test each and every combination experimentally. Therefore, simulations play a key role in screening appropriate ranges of all factors.

Experimental studies of TSWG have been reported, which quantify the impacts of the factors on granule CQAs. The effects of process parameters, e.g., mass feed rate (MFR) and liquid-to-solid (L/S) ratio have been analyzed by a large variety of research groups [8, 9, 10, 11, 12, 13, 14]. There is a consensus that the L/S ratio is the most influential process parameter on granule CQAs [15, 16]. Effects of material properties have also been studied [17, 18, 19], where the main focus is the properties of a single material by performing multiple characterization methods. In addition, Portier et al. [20] assessed the impacts of process settings and screw design with different formulations. Whereas these research investigations provide practical knowledge to understand the impacts of process settings and material properties on CQAs, it is still challenging to define process settings for a given formulation.

As a potential approach to solve the challenge of analyzing the effect of process settings for different formulations, T-shaped partial least squares (TPLS) was applied to TSWG by Ryckaert et al. [21]. TPLS was originally proposed by Garcia-Munoz [22] to model the effects of material properties, composition ratios of materials, and process settings simultaneously. Through the proof of concept by Ryckaert et al. [21], TPLS was recognized as a tool to work for complicated design problems, which can reduce the

experimental effort enormously. However, the current method still has three main challenges toward the practical application. Firstly, a valid method is missing, which defines surrogate materials of new APIs. Due to how the TPLS algorithm works, TPLS can be used only for materials listed in training data. New API information should be replaced by surrogate materials that have similar material properties in TPLS. While Ryckaert et al. [21] proposed to use a k-nearest neighbours algorithm [23], the validity of the algorithm for TPLS has yet to be tested without any comparisons with other potential approaches. Appropriate surrogate selections become more important when formulations contain high concentrations of new APIs, which requires a comprehensive approach to identify a suitable surrogate method. Secondly, necessary material characterization methods are undefined; most existing studies have performed a large number of characterization methods. As noted above, the availability of expensive APIs is limited in the design phase, which makes it difficult to perform many characterization methods for new APIs. Lastly, the prediction accuracy of TPLS should be discussed. While TPLS can capture a larger data space of different process settings and formulations than conventional partial least squares PLS regression, the prediction accuracy tends to be much lower due to a linear regression of huge data space. Ryckaert et al. [21] mentioned that TPLS showed a lower predictive ability because it is a linear technique. In the previous studies using TPLS, the percentages of explained variance of the output parameters were 0.255 [22] and 0.257 [24]. This aspect should be further interpreted; otherwise, the prediction results cannot be reliable.

To tackle these challenges, this paper presents a method to predict the granule size of high-dosed new APIs in continuous TSWG, using a novel surrogate method and TPLS. Multiple nonlinear classification algorithms are tested based on the proposed validation method of surrogate material selection. For the TPLS construction, the information content in the training data was enriched with feature engineering. The developed model was tested for the granule size prediction of three new APIs. Further interpretation of fitting and prediction accuracy of the model was discussed subsequently.

2. Materials and Methods

2.1. Design of Experiments

The same experimental data was used with Ryckaert et al. [21] with some additional experimental data. This paper describes brief experimental

settings that are relevant to this study. For a full description of the whole experimental setup, the reader is referred to the Appendix as well as the original work of Ryckaert et al. [21].

2.1.1. Material selection

Materials included in the training data are summarized in Table A1 in the Appendix. Only materials shown with asterisks in ID were newly added from Ryckaert et al. [21]. To account for batch variability, the same materials with different batches were counted as different IDs in this study, e.g., HCT2. In total, 12 APIs (15 IDs), 4 binders, and 7 fillers (9 IDs) were used to develop a model.

2.1.2. Material characterization

Material properties used in this study are listed in Table A3 in the Appendix. The raw material property database consists of two parts. For the APIs and fillers, 19 physicochemical and solid state properties were used, whereas 4 properties were used for the binders used in this study [21]. Detailed descriptions of the material characterization methods used for surrogate API selection are provided in the Appendix.

2.1.3. Experimental set-up

37 different formulations were used in the training model of TPLS, as summarized in Table A2 in the Appendix. Seven new formulations were added to the original 30 formulations in Ryckaert et al. [21]. Each formulation consists of one API, one or two fillers, and one binder. As the ranges of content are summarized in Table 1, formulations cover from low-dose to high-dose.

After pre-blending of each formulation, granulation experiments were performed using the granulation module of the ConsiGma^{TM}-25 unit (GEA Pharma Systems, Collette, Wommelgem, Belgium) [21]. In the granulator, preblends were wetted by granulation liquid (i.e., demineralized water) and sheared by two 25 mm diameter co-rotating screws. The kneading zones are made up of six kneading elements, each with a length of L=D/4, arranged at a 60° angle. When granulation reached the steady-state, granules were sampled and oven dried (24 h, 40 °C, 25% RH) for granule characterization.

For each formulation, MFR and L/S ratio were changed based on a full factorial Design of Experiments (DoE). The ranges of process settings are shown in Table 1. Whereas the range of MFR was almost the same for

all formulations, L/S ratio ranges were experimentally determined for each formulation. In total, 283 unique experimental settings were performed.

Table 1: Ranges of composition ratio and process settings used in the TPLS training data.

	Minimum	Maximum
Formulation		
API content (%)	5	70
Filler content (%)	25	90
Binder content (%)	2	5
Process settings		
Mass feed rate (kg/h)	5	25
Liquid-to-solid ratio (% w/w)	5.65	84.6

2.1.4. Granule characterization

The granule size distribution was measured by an image analysis system (QICPIC particle size analyzer, Sympatec, Clausthal-Zellerfeld, Germany). Samples of 15 g were measured in triplicate. 10% (d10), 50% (d50), and 90% (d90) cumulative undersize fractions of volumetric granule size distribution were used as indicators of granule CQAs.

2.2. Surrogate API selection

Since in the first stages of process development, information on material characteristics of new APIs is limited, the proposed method only focuses on eight material properties: 50% cumulative undersize fraction of volumetric particle size distribution (d50p), bulk, tapped, and true densities (ρ_b, ρ_t, and ρ_{true}), porosity (ε), flow function coefficient (ffc), powder cohesivity (τc), and specific surface area (SSA). These material properties are typically measured for new materials, according to the guidance from the industry partner in this study. Surrogate selection is thus performed with limited material characterization.

2.2.1. Concept of selection methods

Surrogate API selection consists of two parts: data pre-processing and classification. To enable visual interpretation of surrogate API selection, the dimension of the data was reduced from eight to two in data pre-processing. Proper dimensional reduction could improve prediction accuracy as it reduces collinearity in the data. Two-dimensional reduction methods were considered in this study: principal component analysis (PCA) and kernel PCA. PCA is a technique for the dimensional reduction of a data-set while retaining as much of its variability as possible. PCA finds the directions of the greatest variance in the data and projects the data onto those directions. The data can be represented in a lower-dimensional space but preserve its important features. Kernel PCA works well if interesting data distributions are non-linear by performing PCA in some high-dimensional feature space [25]. Both linear and non-linear dimensional reduction methods were performed and compared.

Pre-processed two-dimensional information was used for classification. Classification is a supervised machine-learning technique for categorical variables, which are discrete and can only take a limited number of values. In this study, API names were used as categorical variables. Multiple classification algorithms were tested using `Python` package `scikit-learn` to identify the suitable classifiers for surrogate selections comprehensively. Most classification algorithms can simulate the probability that a test material belongs to each training material as well as the classification result. Initial screening was performed by classifying training materials using classification methods. If a method could not classify training materials, it was already excluded from the candidates of the methods. In this study, 25 classifiers, e.g., random forest, support vector, and Gaussian process, were used and compared. The full list of the 25 classifiers is described in the result section.

2.2.2. Test of selection methods

We introduced a test method, where the robustness of classification could be confirmed rather than appropriateness of selected surrogate materials. Since material properties of the same material can be slightly changed by different upstream processes, e.g., crystallization, the combination of material properties that lies within the vicinity of the original material should be classified as the same material. This robustness was tested by adding some noise to original data.

Classification tests were performed based on the following steps. These

steps can verify the robustness of each classification method.

1. One material was selected from materials in the training data.
2. A random, Gaussian noise was added to each material property of the selected material. (Mean and standard deviation of the Gaussian distributions were set as zero and 0.1 times of the original value of the material property.)
3. Surrogate API selection methods were applied to the noise-added data set.
4. If a classification result was the same with the original material, the classification was judged as *correct*.
5. These steps were performed 100 times for each material.
6. Percentage of *correct* was calculated.

Afterward, percentages of *correct* were compared among different classifiers to identify appropriate classifiers. Only the materials used in Ryckaert et al. [21] were used in the selection of surrogate API methods.

2.3. TPLS development

The schematic overview of matrices used in this study is illustrated in Figure 1. The TPLS model considers four matrices: process settings ($\boldsymbol{Z} \in \mathbb{R}^{283\times2}$), raw material properties ($\boldsymbol{X} \in \mathbb{R}^{28\times23}$), blending ratios ($\boldsymbol{R} \in \mathbb{R}^{283\times28}$), and granule quality attributes ($\boldsymbol{Y} \in \mathbb{R}^{283\times1}$) for each attribute). Note that "$\boldsymbol{\alpha} \in \mathbb{R}^{\beta\times\gamma}$" means that the matrix $\boldsymbol{\alpha}$ consists of real numbers with β rows and γ columns. In this study, MFR and L/S ratio were used as the elements of process settings ($\boldsymbol{Z}$); d-values (d10, d50, and d90) of granule size distribution were chosen as granule quality attributes ($\boldsymbol{Y}$). The elements of blending ratios ($\boldsymbol{R}$) and raw material properties ($\boldsymbol{X}$) were all items listed in Tables A1 and A3, respectively. The TPLS models were developed based on the algorithm described by Garcia-Munoz [22].

The following changes were made from the methodology described by Ryckaert et al. [21] to improve the accuracy:

- A TPLS model was developed for each quality indicator ($\boldsymbol{Y}_1$, $\boldsymbol{Y}_2$, and $\boldsymbol{Y}_3$ in Figure 1) instead of for all quality descriptors simultaneously.
- Quality attributes (d-values) were transformed using the common logarithm to introduce non-linearity and correct for d-values spanning multiple orders of magnitude.

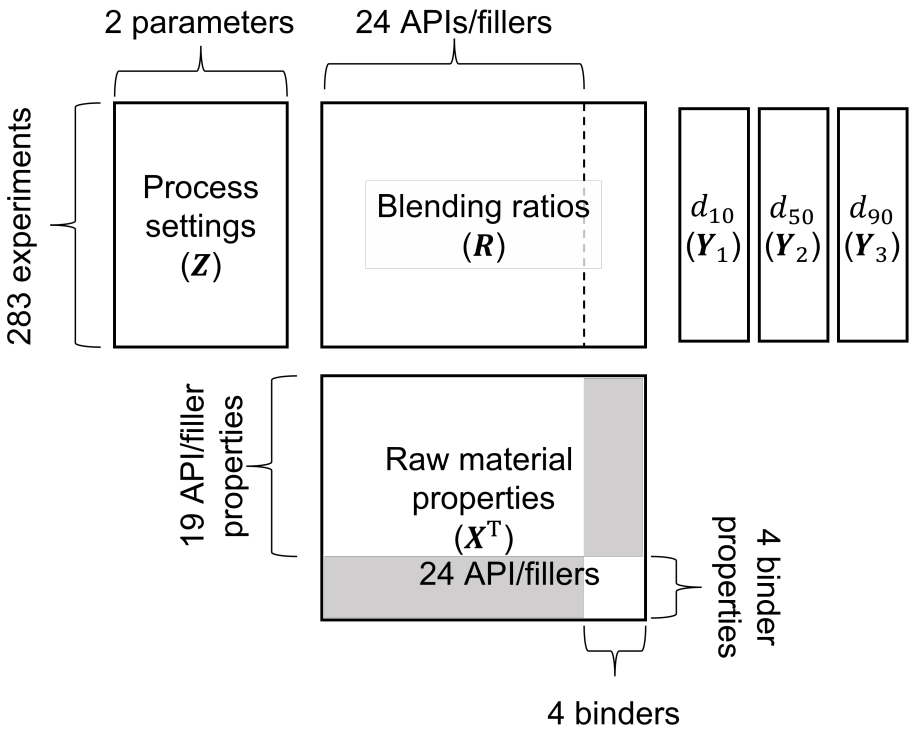


Figure 1: Overview of matrices used for the TPLS model.

- Among raw material properties, S_{true} and $S40$ were transformed using the common logarithm due to spanning multiple orders of magnitude.
- Three latent variables (LVs) were used to avoid over-fitting based on R^2 values.

To assess the fitting quality of the TPLS models, the percentage of explained variance of the Y explained by the TPLS model (R^2Y) was computed, as shown in Equation 1:

$$R^2Y = \sum_{i=1}^{n} \left(1 - \frac{\sum (y_j - y'_{j,i})^2}{\sum y_j^2}\right), \tag{1}$$

where y_j and $y'_{j,i}$ represent a quality attribute (e.g., d10) in experiment j and the simulated quality attribute value at the i-rd LV in experiment j, respectively. Note that the predicted quality attribute is the sum of $y'_{j,i}$ from the first to the n-rd LV. In addition to R^2Y, robust R2 value, R^2rY was also used, as shown in Equation 2:

$$R^2rY = \sum_{i=1}^{n} \left(1 - \left(\frac{med\left(y_j - y'_{j,i}\right)}{med\left(y_j\right)}\right)^2\right), \tag{2}$$

where $med\left(\boldsymbol{A}\right)$ represents the median of the set $\boldsymbol{A}$. For model fitting, R^2 that measures the proportion of variances was chosen as an indicator because TPLS was developed by capturing variances in the data.

Cross-validation (CV) was also performed, where $RMSE$ was used as the objective function. 5% of experimental data was excluded from training data at random and used as test data. This step was repeated 30 times, and the average of $RMSE$ was computed as the output to get an estimation on the robustness of the model. For CV, $RMSE$ was used instead of R^2 because understanding the expected prediction errors was more important.

Key material properties and process parameters were interpreted based on the results of the Variable Importance for the Projection (VIP) scores. VIP scores for predictor variable j, $VIP(j)$, can be calculated as shown in Equation 3:

$$VIP(j) = \sqrt{\frac{p \sum_{k=1}^{p} w(j,k)^2 R^2(k)}{p}}, \tag{3}$$

where $w(j,k)$, $R^2(k)$, and p represent the weight of predictor variable j in component k of the PLS model, the R-squared value for LV k, and the number of LVs in the PLS model.

The developed TPLS could be applicable for the same equipment and screw configurations described in the section of the experimental set-up. To use it for different settings, it is necessary to either consider the effect of screw configurations (e.g., [26]) or develop another TPLS model.

2.4. Model validation

A combination of surrogate API selection and TPLS was validated by applying it to the prediction of granule size containing new APIs. Three new APIs (API A, B, and C) were used, and the test formulations are summarized in Table 2. Since multiple batches of APIs were used except for API C, batch differences were also reflected in API name, e.g., A_1. All formulations contained a new API, i.e., untrained data in TPLS, and the loading of new API was in excess of 5%. The ranges of MFR and L/S ratio were 5–10 kg/h and 30–48%. In total, 23 experimental data points were used for model validation.

Table 2: Overview of the test formulations. The content is denoted in mass percentage of the total blend content.

Formulation ID	API name	Content (m%)	Filler 1 name	Content (m%)	Filler 2 name	Content (m%)	Binder name	Content (m%)
1	A_1	53.8	200M	22.6	PH101	21.5	KEF	2.2
2	A_2	75.3	200M	11.3	PH101	11.3	KEF	2.2
3	A_2	64.5	200M	16.7	PH101	16.7	KEF	2.2
4	A_3	64.5	200M	16.7	PH101	16.7	KEF	2.2
5	B_1	53.8	200M	21.0	PH101	21.0	K30	4.3
6	B_2	53.8	200M	21.0	PH101	21.0	K30	4.3
7	B_3	53.8	200M	21.0	PH101	21.0	K30	4.3
8	B_3	67.2	200M	14.2	PH101	14.2	K30	4.3
9	C	64.5	200M	16.1	PH101	16.1	KEF	3.2

Multiple surrogate API selection methods were tested to find the best method in terms of the prediction accuracy of granule size. While the work of Ryckaert et al. [21] chose the weights of five materials as a surrogate of a new API, the number of surrogate materials n_s for a new API was also changed among 1, 2, 4, or 24 (all) in this study. Concentration of surrogate material h, C_h, was regarded as shown in Equation 4:

$$C_h = \frac{p_h}{\sum_{k=1}^{n_s} p_k} C_{\mathrm{API}}, \tag{4}$$

where p_k and C_{API} represent a probability that a new API belongs to material h and API content, respectively. If all 24 materials are chosen as the number of surrogate materials, $\sum_{k=1}^{n_s} p_k$ becomes 1.

3. Results and Discussion

3.1. Surrogate API selection

The initial screening result of classification algorithms is presented in Table 3; 14 out of 25 methods were excluded from the candidates. More than a half of methods were unable to classify materials properly even though the same data sets were included in training data.

A test of selection methods was performed for the remaining eleven classification algorithms. For the support vector classification (SVC), the linear and radial basis function (rbf) kernels were also compared. The test result is summarized in Table 4. For most of the cases, PCA as a dimension reduction pre-processing step showed higher prediction accuracy than kernel PCA; this was seen in the algorithm test scores. Since non-linearity of data was considered in classification algorithms, simple and linear pre-processing worked better. Among different classification algorithms, SVC-related algorithms, i.e., `NuSVC`, `SVC(kernel="linear")`, `SVC(kernel="rbf")`, showed good scores, and Gaussian process classifier (GPC) was next to them. The accuracy was used as the scoring metric and is defined as:

$$\mathrm{accuracy}(y, \hat{y}) = \frac{1}{n_{\mathrm{samples}}} \sum_{i=1}^{n_{\mathrm{samples}}} \mathbf{1}(\hat{y}_i = y_i), \tag{5}$$

with $\mathbf{1}(x)$ is the indicator function.

To understand the performances of classification algorithms, mappings of classified materials were created, as shown in Figure 2. In the graphs, classified materials are illustrated for different scores of principal component (PC) 1 and 2. In addition to SVC and GPC, two different classification algorithms were compared: random forest classification and extra tree classification (score was close to GPC). The mapping result of random forest had several areas that were difficult to interpret. For instance, there are multiple zones where M_f is selected as the classified material. While the area

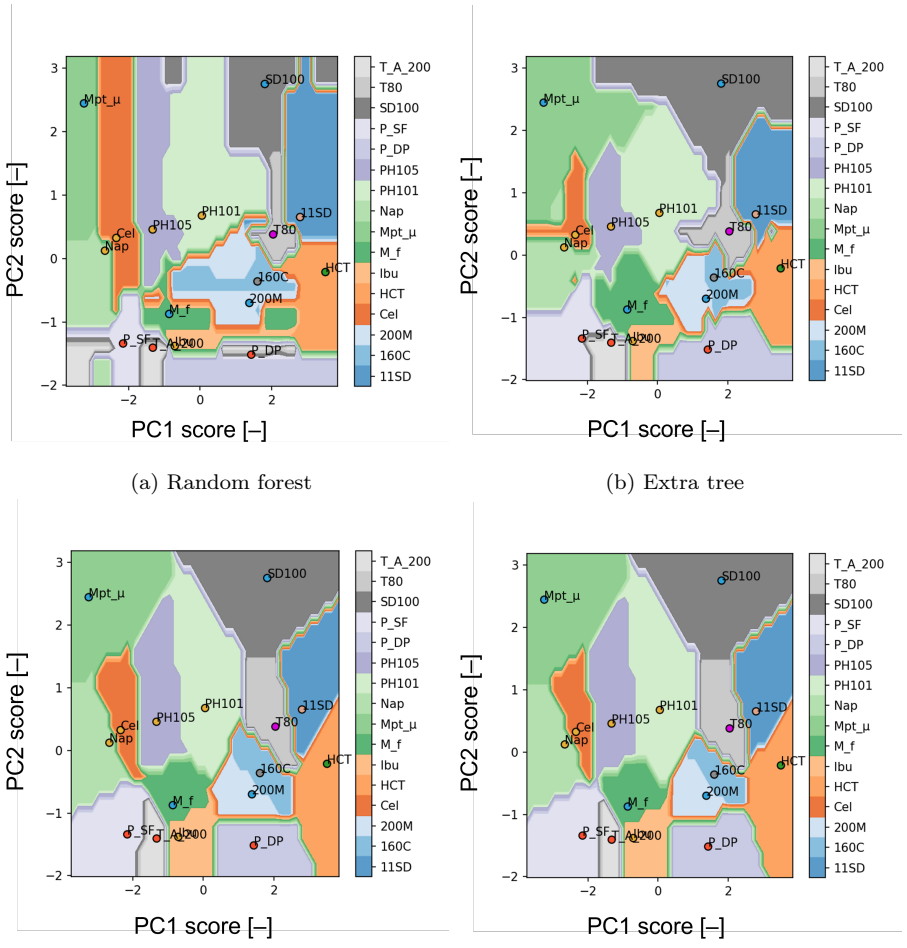
Table 3: Initial screening of classification algorithms.

Description	Name of `scikit-learn` library
Classification was not possible.	`CategoricalNB`
	`ComplementNB`
	`MultinomialNB`
	`RadiusNeighborsClassifier`
Percentage of "correct" was $<50\%$ even for training materials without noise.	`LinearSVC`
	`KNeighborsClassifier`
	`BernoulliNB`
	`DecisionTreeClassifier`
	`AdaBoostClassifier`
	`SGDClassifier`
	`HistGradientBoostingClassifier`
	`PassiveAggressiveClassifier`
	`RidgeClassifier`
	`RidgeClassifierCV`
Classification was generally possible for training materials without noise.	`NuSVC`
	`SVC`
	`BaggingClassifier`
	`ExtraTreesClassifie`
	`GaussianProcessClassifier`
	`LogisticRegression`
	`RandomForestClassifier`
	`DecisionTreeClassifier`
	`ExtraTreeClassifier`
	`MLPClassifier`
	`GradientBoostingClassifier`

Table 4: Test score accuracy of classification algorithms.

Name of `scikit-learn` library	Pre-processing method	
	PCA	Kernel PCA
`NuSVC`	0.830	0.681
`SVC(kernel="linear")`	0.831	0.711
`SVC(kernel="rbf")`	0.848	0.701
`BaggingClassifier`	0.606	0.536
`ExtraTreesClassifier`	0.811	0.656
`GaussianProcessClassifier`	0.826	0.686
`LogisticRegression`	0.801	0.564
`RandomForestClassifier`	0.725	0.637
`DecisionTreeClassifier`	0.506	0.522
`ExtraTreeClassifier`	0.643	0.585
`MLPClassifier`	0.765	0.610
`GradientBoostingClassifier`	0.452	0.532

around $PC1 = -1$ and $PC2 = -1$ includes the point of the original values of M_f, the area around $PC1 = 1$ and $PC2 = -1$ was not close to the original location of M_f. Extra tree classification had right-angled regions, e.g., Cel ($PC1 = -2$ and $PC2 = 0$), which reduced the robustness of the classification. On the other hand, the mapping results of SVC and GPC, which are almost the same, show clear classifications of materials, which is much easier to interpret. The result of SVC shown in Figure 2 is based on a linear kernel, and decision boundaries were almost the same for the other SVCs.



(a) Random forest

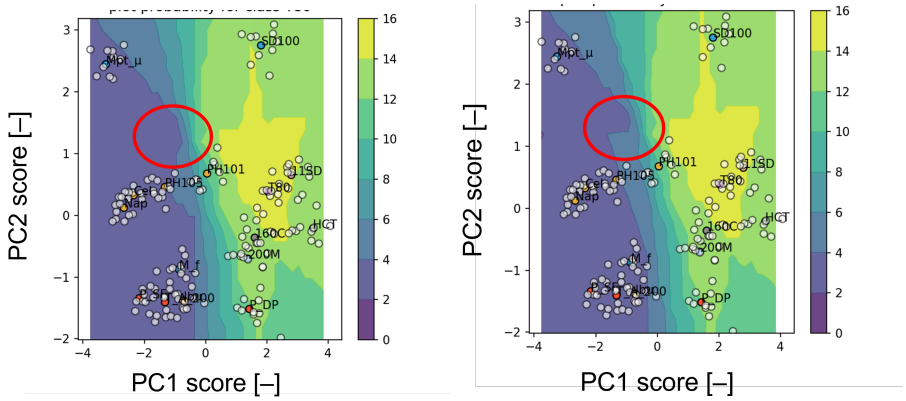
(b) Extra tree

(c) SVC (linear)

(d) GPC

Figure 2: Mapping of classified materials in a two-dimensional score space.

Since differences among multiple SVCs and GPC could not be detected by mapping, decision function plots of SVCs were compared with probability plots of GPC for specific materials. Decision function represents the distance to the separating hyperplane in SVCs, where larger absolute values mean higher confidence in classifications. The probability plots of material XX show the probabilities that materials belong to XX with different values of PC1 and PC2. Firstly, different kernels of SVCs were compared by plotting classification probabilities, as an example of T80 can be seen in Figure 3. Probability plots seem quite similar, where only visible difference is the area circled in red. Considering the fact that both kernels behaved similarly and showed high accuracy, we concluded that there was no need of introducing complexity nor further optimization. Therefore, the following parts of the paper only focused on SVC using a linear kernel.



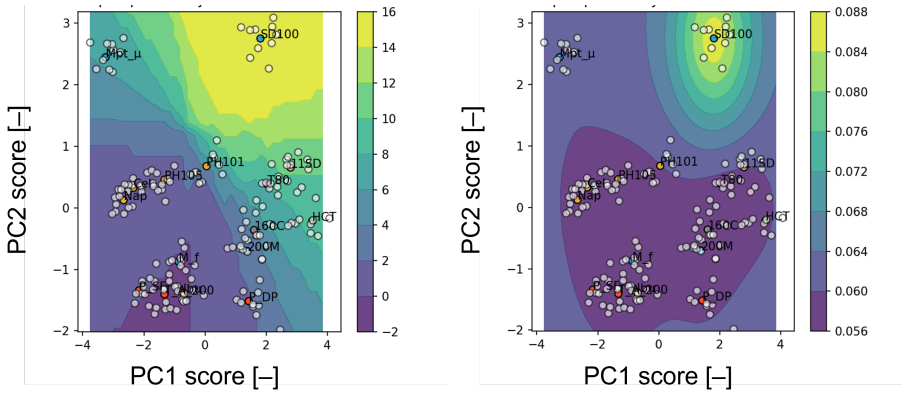
(a) Rbf kernel

(b) Linear kernel

Figure 3: Decision function plot of T80 using SVC with rbf and linear kernels. The grey dots are examples of data points that were used to assess the robustness of the classification: the values of the original material properties with some added Gaussian noise.

The differences between SVC and GPC were also analyzed in terms of decision function and probability plots. Figure 4 shows decision function plots of SVC and probability plots of GPC in case of SD100. While the slope of probability or decision function depends on directions in SVC, the slope seems similar for any directions in GPC. GPC introduces Gaussian distribution for the classification, which brought the smooth probability changes by score shifts. Since it is difficult to judge the best classification algorithm

281 from these differences, both algorithms were used for the model validation.



(a) SVC (linear) (b) GPC

Figure 4: Decision function plots of SVC and probability plot of GPC, in case of SD100.

282 *3.2. TPLS development*

283 The quality of the developed TPLS is summarized in Table 5. For all
284 indicators, d10 showed the best model quality, which means the prediction
285 of large granules was more difficult. The model quality of all attributes
286 was huge improved by introducing log-transformation of d-values as well
287 as some material properties. The R^2Y of d10, d50, and d90 without log-
288 transformation were 0.1365, 0.1797, and 0.2073, respectively.

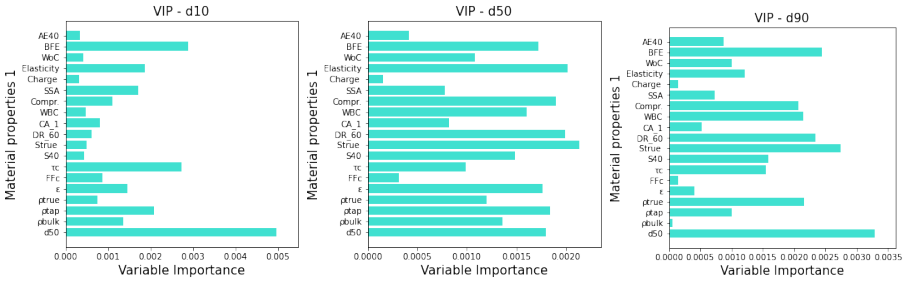
Table 5: Model quality descriptors of the TPLS model (robust) R-squared values and the root mean squared error on the prediction, averaged over all CV folds.

Attribute	R^2Y	R^2rY	$RMSE$ [μm] (CV)
d10	0.3132	0.4965	368.4
d50	0.2528	0.2919	813.7
d90	0.2988	0.1900	830.8

289 On the other hand, the presented R^2 values in Table 5 are still low com-
290 pared to other PLS models of TSWG, e.g., Van Snick et al. [27]. As discussed
291 in Ryckaert et al. [21], TPLS has a limitation of model quality due to a linear
292 regression technique. The total search space of TPLS is a lot larger because
293 of the multiple matrices, i.e., material properties, blending ratios, and process
294 settings, and loadings. Linear regressions of such a large space could cause

lower R^2 values than PLS models. In this study, the training data includes a large ranges of explanatory variables, e.g., from hydrophilic to hydrophobic materials and from low to high L/S ratio. For instance, when only 71 data points that are all of available data of L/S ratio $> 30\%$ were used as the training data, the R^2Y of d10, d50, and d90 were 0.7769, 0.8080, and 0.8345, respectively. The TPLS developed by the 71 data points was applied to the model validation for the comparison as presented in the next section.

The Variable Importance for the Projection (VIP) plots of the TPLS models were generated to find key material properties and process settings. Figures 5, 6, and 7 present the VIP plots of API/filler properties, binder properties, and process settings, respectively. As for material properties of APIs and fillers (Figure 5), particle size (i.e., d50p) had high impacts on d10 and d90, whereas moisture-related properties (e.g., S_{true} and DR60) were influential on d50 and d90. Figure 6 shows that CA1 was the most influential factor for all d-values among binder properties. In addition, L/S ratio generally had higher impacts on d-values than MFR (Figure 7).

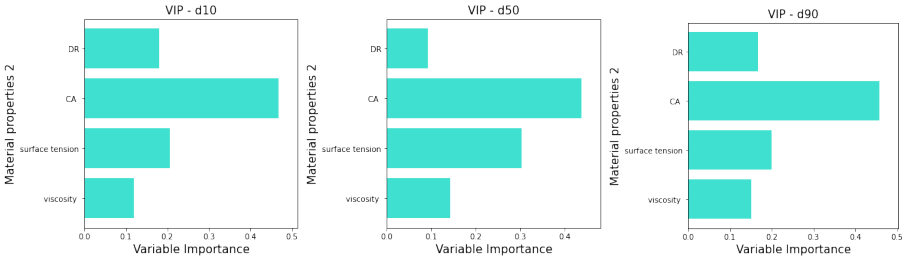


(a) d10 (b) d50 (c) d90

Figure 5: VIP plots of material properties of APIs and fillers.

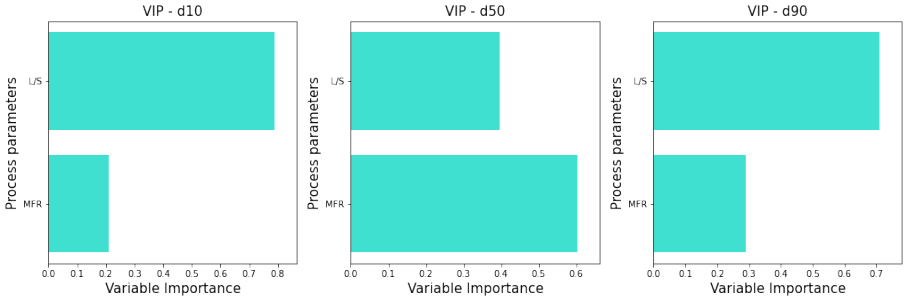
3.3. Model validation

The developed method was validated using test data using new APIs that are not included in any training data. Validation results with different classification algorithms and the number of surrogate APIs are summarized in Table 6. As a general trend, $RMSE$ became lower, i.e., higher prediction accuracy, if the number of surrogate materials was increased. A smaller number of surrogate materials could cause a higher risk of inappropriate selection, which can be reduced by using a larger number of materials. For example, the selected single material was different between different batches



(a) d10 (b) d50 (c) d90

Figure 6: VIP plots of material properties of binders.



(a) d10 (b) d50 (c) d90

Figure 7: VIP plots of process settings.

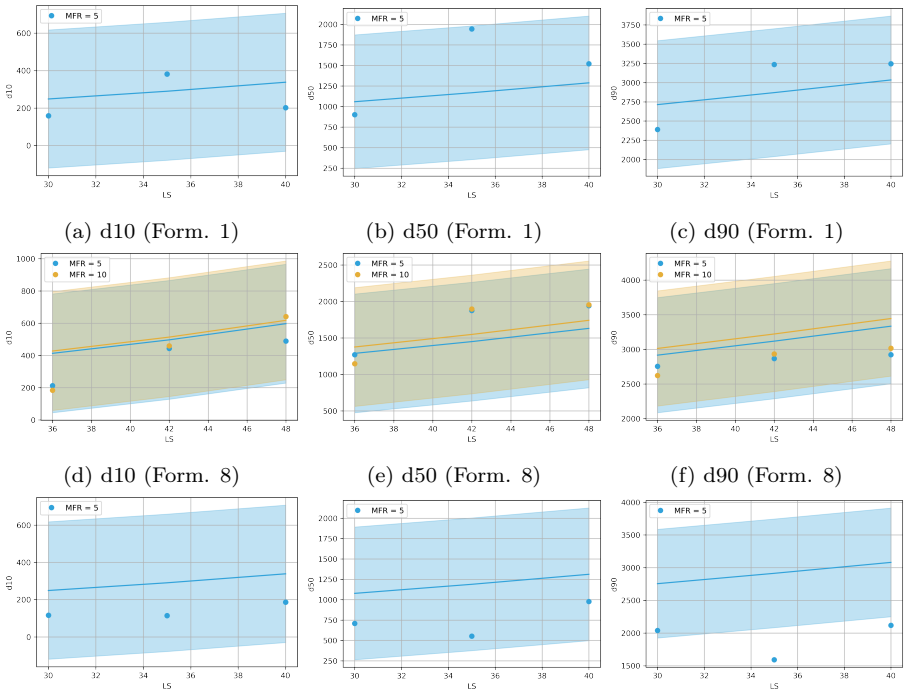
of the same APIs (e.g., A_1 and A_2). Overall, GPC using all APIs and fillers as surrogate materials showed the best scores for d10 and d50, and the third best score for d90. This combination was used for the further analysis of the validation results.

Table 6: Validation results with different classification algorithms and number of surrogate APIs.

Classification	Number of surrogate materials	$RMSE$ value [μm] d10	d50	d90
SVC	1	191.04	577.10	651.31
	2	176.18	561.93	534.03
	4	193.30	500.59	487.15
	all	168.86	415.50	501.24
GPC	1	343.20	678.88	985.31
	2	283.95	605.48	893.03
	4	168.42	521.95	591.09
	all	137.34	413.65	509.22

Figure 8 shows the comparisons between the measured and simulated values for formulations 1, 8, and 9. Dots, lines, and filled areas in Figure 8 represent measured values, predicted values, and uncertainty bounds based on $RMSE$ of CV, respectively. The widths of uncertainty bounds were fixed for each of d10, d50, and d90, which caused the prediction of negative values if predicted values were lower than $RMSE$. While experimental data of one formulation are presented for each API (API A, B, and C), other results also showed the similar prediction accuracy. The measured values were close to predicted lines and within the range of the uncertainty bounds for API A and B. For API C, the measured d90 values were out of the uncertainty bounds. Material properties of API C were exceptional and far from all of the training materials, which made surrogate material selection difficult.

To visually understand the prediction accuracy of d-values, the entire granule size distribution was reconstructed assuming a log-normal distribution. A high L/S ratio usually makes granule size distribution mono-modal [28], and a log-normal distribution has been used to fit the granule size distribution (e.g., [29]). A comparison of granule size distribution is summarized in Figure 9, where the results of formulations 1, 8, and 9 are shown with



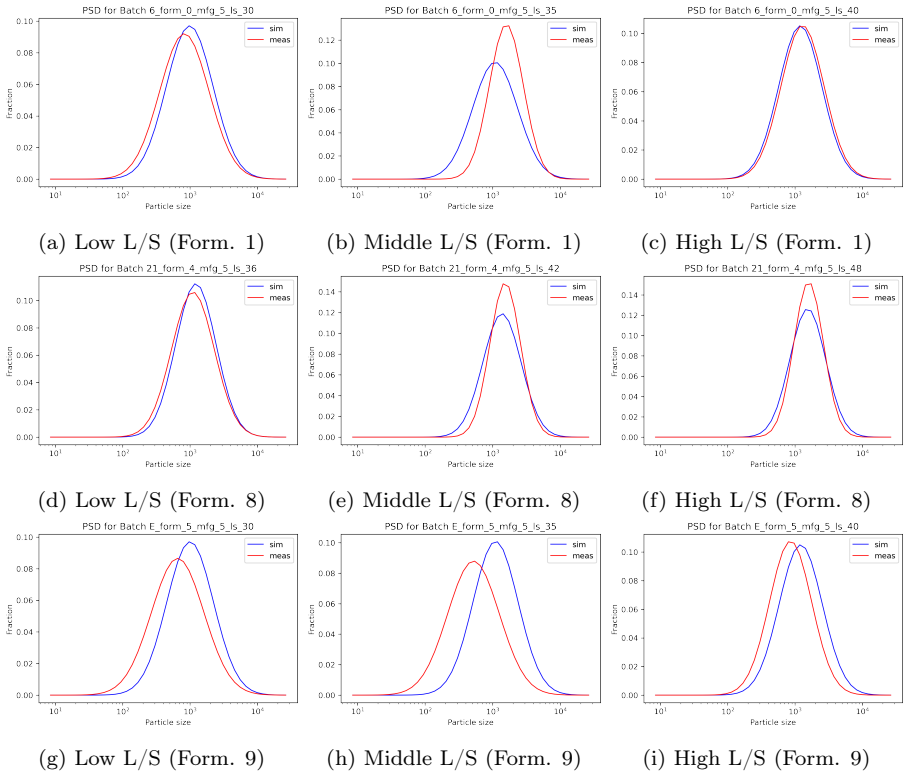
(a) d10 (Form. 1) (b) d50 (Form. 1) (c) d90 (Form. 1)

(d) d10 (Form. 8) (e) d50 (Form. 8) (f) d90 (Form. 8)

(g) d10 (Form. 9) (h) d50 (Form. 9) (i) d90 (Form. 9)

Figure 8: Comparison between the measured and predicted values with uncertainty bounds for validation.

three different L/S ratios. The predictions of granule size distributions were accurate, especially for low and high L/S of API A (formulation 1) as well as low L/S of API B (formulation 8). For middle and high L/S of API B, prediction captured the peak of the distribution. The prediction was slightly difficult for middle L/S of API A, in addition to API C. The measured d-values of middle L/S were higher than those of low and high L/S in API A, whereas TPLS can predict monotonic increase or decrease by changes in L/S due to a linear regression model. In addition, powder cohesivity (τc) of API C was much higher than any materials used in training data. Adding a large variety of materials would improve the prediction accuracy of API C.



(a) Low L/S (Form. 1) (b) Middle L/S (Form. 1) (c) High L/S (Form. 1)

(d) Low L/S (Form. 8) (e) Middle L/S (Form. 8) (f) High L/S (Form. 8)

(g) Low L/S (Form. 9) (h) Middle L/S (Form. 9) (i) High L/S (Form. 9)

Figure 9: Comparison of granule size distribution assuming a log-normal distribution.

In short, the developed model enabled to prediction of d-values of different formulations containing new API and different process settings. While R^2 values of the TPLS models were low, the entire model could be used for a wide range of predictions by using different types of training data. Note that the TPLS models developed by L/S $>$ 30% resulted in lower prediction accuracy

($RMSE$ of d10, d50, and d90 were 180.49, 563.15, and 541.94, respectively). The model can be used only by performing five characterization techniques of a new API, so it is beneficial to use it for the screening of formulations and process settings in the early development phase.

On the other hand, the model has two limitations toward the further applications. Firstly, the presented model predicted only d-values, not the entire granule size distributions, which makes it impossible to predict bimodal distributions. In addition, further improvement of prediction accuracy might be difficult considering low R^2 values of the TPLS models. Within the framework of TPLS, only reducing the number of training data points was a way to increase R^2, which decreased the prediction accuracy for new materials. This makes the training set very specific to particular use cases and decreases the ability to generalise beyond the training data set. Introducing non-linearity would be inevitable to increase the prediction accuracy with the same applicability.

4. Conclusions

We developed the method to predict granule size for new formulations by integrating a novel surrogate API selection method and TPLS. The proposed surrogate API selection method, which consists of PCA and SVC/GPC, worked by using material properties based on the five basic characterization techniques (particle size, densities, ring shear test, solubility, and specific surface area). TPLS was developed by training data consisting of a large variety of material properties, blending ratios, and process settings. The proposed surrogate material method can be used for other purposes in the earlier design phase, e.g., preliminary experiments without using new APIs, since the use of surrogate materials is often valuable [30].

The proposed method was validated by applying it to formulations containing high-dosed APIs. After the selection of the classification algorithm and the number of surrogate materials, the model could predict d-values of granulation size with acceptable errors. We identified that the best surrogate method in this study was the combination of PCA-based pre-processing of data and GPC to compute surrogates of new APIs from all known materials. The developed model can guide formulation and process design in the earlier phase with much less experimental work.

At the same time, the limitation of TPLS was observed throughout the work in terms of low R^2Y values. To increase accuracy and expand appli-

cability further, switching from TPLS to other non-linear models would be required.

5. Outlook

TPLS can be used for the initial screening of formulations and process settings for new drugs. The applicability of TPLS with surrogate API methods is the largest among available granule CQA prediction methods (e.g., [28, 12]). This benefit works especially for the early phase of process design, where design freedom and uncertainty is high. Prediction accuracy of TPLS may be improved by further enriching training data and increasing the number of material properties. In addition, fitting accuracy of TPLS can be improved by limiting the applicability of the model, e.g., hydrophilic or hydrophobic materials. In this case, a valid method to clarify appropriate TPLS models for new products is necessary; otherwise, prediction can be worse as presented in this study.

It is possible that the layer of complexity in the current model, i.e. from material properties to surrogate material to prediction, has a negative impact on the predictive power of this methodology. All information of materials and process settings should be included in the training data for the usage of TPLS, which requires an additional surrogate material method to fill in the gaps. In an ideal case, the input to the model would be material properties instead of specifying a material to circumvent this issue.

In place of TPLS, other data-driven methods could be explored. In this work, the information that the data is a distribution, i.e. the particle size distribution or the mixing rules of the formulation, was not included in the model. The research group of the authors published on the application of kernel mean embedding to predict particle size distributions from process settings [31]. An extension of the former proposed model to distribution-to-distribution regression would be an interesting approach to apply to the data used in this study. Further, neural networks might also be applicable: for instance the special type mixture density networks can be used to not only predict a target value, but also the underlying probability distribution.

Appendix A.

Appendix A.1. Material selection

Table A1: Overview of raw materials used in the experiments.

Type	Material	Brand name	ID	Manufacturer
API	Theophylline Anh. Powder 200M		*T_A_200M	BASF
	Theophylline Anh. Powder 200		T_A_200	Siegfried
	Hydrochlorothiazide		HCT *HCT2 *HCT3	UTAG
	Metoprolol Tartrate micronized		MPT_μ	Polydrug Laboratories
	Metoprolol Tartrate powder		*MPT_p	UTAG
	Celecoxib		Cel	UTAG
	Metformin fine		M_f	FARMHISPANIA
	Ibuprofen		Ibu *Ibu2	BASF
	Naproxen		Nap	UTAG
	Paracetamol semi fine		P_sf	Mallinckrodt
	Paracetamol micronized		*P_μ	Mallinckrodt
	Paracetamol dense powder		P_dp	Mallinckrodt
Binder	Hydroxypropyl methylcellulose (HPMC)	Methocel E5	E5	Dow Chemical Company
		Methocel E15	*E15	Dow Chemical Company
	Hydroxypropyl cellulose (HPC)	Klucel EF	KEF	Ashland
	Polyvinylpyrrolidone (PVP)	Kollidon 30	K30	BASF
Filler	Lactose	Tablettose T80	T80	Meggle
		Pharmatose 200 M	200M *200M2	DEF Pharma
		Subertab 11SD	11SD	DEF Pharma
	Microcrystalline cellulose	Avicel PH105	PH105	FMC BioPolymer
		Avicel PH101	PH101 *PH101_2	FMC BioPolymer
	Mannitol	Pearlitol 160C	160C	Roquette
		Pearlitol 100SD	100SD	Roquette

*Materials added to training data from Ryckaert et al. [21]

Table A2: Overview of the formulation blend compositions.

Code	API	Content (%)	Filler	Content (%)	Binder	Content (%)
F1B1	Ibu	10	200M	85	E5	5
F1B2	Ibu	40	200M	55	E5	5
F1B3	Ibu	70	200M	25	E5	5
F2B1	P_sf	10	160C	85	K30	5
F2B2	P_sf	40	160C	55	K30	5
F2B3	P_sf	70	160C	25	K30	5
F3B1	T_A_200	10	PH101	85	KEF	5
F3B2	T_A_200	40	PH101	55	KEF	5
F3B3	T_A_200	70	PH101	25	KEF	5
F4B1	M_f	10	T80	85	E5	5
F4B2	M_f	40	T80	55	E5	5
F4B3	M_f	70	T80	25	E5	5
F5B1	Nap	10	PH105	85	K30	5
F5B2	Nap	40	PH105	55	K30	5
F5B3	Nap	70	PH105	25	K30	5
F6B1	HCT	10	200M/PH105	44/44	KEF	2
F6B2	HCT	40	200M/PH105	29/29	KEF	2
F6B3	HCT	70	200M/PH105	14/14	KEF	2
F7B1	MPT_μ	10	PH101	88	E5	2
F7B2	MPT_μ	40	PH101	58	E5	2
F7B3	MPT_μ	70	PH101	28	E5	2
F8B1	P_dp	10	11SD	88	KEF	2
F8B2	P_dp	40	11SD	58	KEF	2
F8B3	P_dp	70	11SD	28	KEF	2
F9B1	M_f	10	T80	88	E5	2
F9B2	M_f	40	T80	58	E5	2
F9B3	M_f	70	T80	28	E5	2
F10B1	Cel	10	SD100	88	KEF	2
F10B2	Cel	40	SD100	58	KEF	2
F10B3	Cel	70	SD100	28	KEF	2
*F11B1	HCT2	50	200M2/PH101_2	30/15	E15	5
*F11B2	HCT2	5	200M2/PH101_2	75/15	E15	5
*F12B1	Ibu2	50	200M2/PH101_2	30/15	E15	5
*F13B1	T_A_200M	50	200M2/PH101_2	30/15	E15	5
*F14B1	MPT_p	50	200M2/PH101_2	30/15	E15	5
*F15B1	HCT2	63.2	200M2/PH101_2	16.8/16.8	KEF	3.2
*F16B1	P_μ	63.2	200M2/PH101_2	16.8/16.8	KEF	3.2

*Formulations added to training data from Ryckaert et al. [21]

Appendix A.2. Material characterization

Four characterization techniques used for surrogate API selection were performed based on the following procedures.

Appendix A.2.1. Laser diffraction

Laser diffraction (Mastersizer 2000, Malvern Instruments, Worcestershire, UK) was employed for volume-based particle size distribution analysis using dry dispersion. Powders were fed to a 550 RF lens at 3.0 G with a standard jet pressure of 2.0 bar. Measurements were conducted in triplicate and analyzed with Mastersizer 2000 software, yielding d50p as 50% cumulative undersize fraction.

Appendix A.2.2. Bulk and tapped density

A measured quantity of each powder was introduced into a 250 mL graduated cylinder to determine the bulk volume. Subsequently, the graduated cylinder underwent 1250 taps via a tapping machine (J. Engelsman, Ludwigshafen, Germany) to establish the tapped volume. Bulk density (ρ_b [g/mL]) and tapped density (ρ_t [g/mL]) were computed based on the mass of powder per unit bulk and tapped volumes, respectively.

An AccuPyc 1330 helium pycnometer (Micromeritics, Norcross, GA, USA) determined the true density (ρ_{true} [g/mL]) of each powder. This was accomplished with an equilibration rate set at 0.0050 psig/min, involving 10 purges per test. Each powder underwent duplicate measurements. Powder bed porosity (ε, dimensionless) was computed by ρ_b and ρ_{true}.

Appendix A.2.3. Ring shear test

Each powder underwent triplicate analysis in a ring shear tester (RST-XS, Dietmar Schulze Schüttgutmesstechnik, Wolfenbüttel, Germany) to measure powder flowability and cohesivity. Initially, a pre-shear step applied a normal load of 1000 Pa to the sample. Subsequently, shear stress was introduced to the powder through three consolidation stresses: 400, 600, and 800 Pa. Powder flowability was assessed by the unconfined yield strength to the major principal stress ratio, denoted as the dimensionless flow function coefficient (ffc). Moreover, powder cohesivity (τc) was defined as the yield locus intercept with the shear stress axis. For a comprehensive analysis and interpretation of ring shear test data, Schulze [32] provides further information.

Appendix A.2.4. Specific surface area

Samples underwent a degassing process in a vacuum using VacPrep Micrometrics, Norcross, USA for 24 hours at 60 °C, followed by a 1-hour nitrogen purge to eliminate impurities. Subsequently, the nitrogen adsorption isotherm was assessed through nitrogen adsorption (TriStar 3000 V6.08A, Micrometrics, Norcross, USA) Specific surface area (SSA [m^2/g]) was determined using the Brunauer-Emmett-Teller theory.

Table A3: Raw material properties used in the study.

Characterization technique	Descriptor	Abbreviation
API and filler		
Laser diffraction	50% cumulative undersize fraction of volumetric particle size distribution	d50p
Powder density and porosity	Bulk, tapped and true density	ρ_b, ρ_t, ρ_{true}
	Porosity	ε
Ring shear test (400, 600, 800, and 1000 Pa)	Flow function coefficient	ffc
	Powder cohesivity	τc
Powder elasticity and plasticity	Elastic recovery	Elas
	Work of compaction	WoC
Powder rheometer	Basic flow energy	BFE
	Compressibility at 15 kPa	Com
	Ratio of flow energy at 40 and 0 mm/s	AR40
Charge density	Charge-to-mass ratio	CD
Specific surface area	Specific surface area	SSA
Water binding capacity	Water binding capacity	WBC
Solubility	Maximum powder solubility	S_{true}
Dissolution rate	Fraction powder dissolved after 60 min	DR60
Dynamic vapour sorption	Moisture content in sorption mode at 40% RH	S40
Contact angle	Hydrophilic contact angle after 1 s	CA1
Binder		
Dissolution rate	Fraction binder dissolved after 90 s	DR90
Contact angle	Hydrophilic contact angle after 1 s	CAb1
Viscosity	Dynamic viscosity	Vis
Surface tension	Surface tension	ST

Author contributions

Kensaku Matsunami: Conceptualization; Methodology; Software; Formal analysis; Writing - original draft; Writing - review & editing; Visualization. **Jonathan Meyer**: Conceptualization; Methodology; Validation; Investigation; Data Curation; Writing - review & editing; Project administration; Funding acquisition. **Martin Rowland**: Writing - review & editing; Project administration; Funding acquisition. **Neil Dawson**: Writing - review & editing; Project administration; Funding acquisition. **Thomas De Beer**: Conceptualization; Resources; Writing - review & editing; Supervision; Project administration. **Daan Van Hauwermeiren**: Conceptualization; Methodology; Software; Formal analysis; Writing - review & editing; Visualization; Supervision.

Acknowledgement

The authors would like to acknowledge, Pfizer Inc. for their financial support and fruitful collaboration in this project.

References

[1] B. V. Snick, J. Dhondt, K. Pandelaere, J. Bertels, R. Mertens, D. Klingeleers, G. D. Pretoro, J. P. Remon, C. Vervaet, T. D. Beer, V. Vanhoorne, A multivariate raw material property database to facilitate drug product development and enable in-silico design of pharmaceutical dry powder processes, International Journal of Pharmaceutics 549 (2018) 415–435. doi:10.1016/j.ijpharm.2018.08.014.
URL https://linkinghub.elsevier.com/retrieve/pii/S0378517318305763

[2] G. Allison, Y. T. Cain, C. Cooney, T. Garcia, T. G. Bizjak, O. Holte, N. Jagota, B. Komas, E. Korakianiti, D. Kourti, R. Madurawe, E. Morefield, F. Montgomery, M. Nasr, W. Randolph, J.-L. Robert, D. Rudd, D. Zezza, Regulatory and quality considerations for continuous manufacturing may 20–21, 2014 continuous manufacturing symposium, Journal of Pharmaceutical Sciences 104 (2015) 803–812. doi:10.1002/jps.24324.
URL http://linkinghub.elsevier.com/retrieve/pii/S0022354916300077

[3] S. L. Lee, T. F. O'Connor, X. Yang, C. N. Cruz, S. Chatterjee, R. D. Madurawe, C. M. V. Moore, L. X. Yu, J. Woodcock, Modernizing pharmaceutical manufacturing: from batch to continuous production, Journal of Pharmaceutical Innovation 10 (2015) 191–199. doi:10.1007/s12247-015-9215-8.
URL http://link.springer.com/10.1007/s12247-015-9215-8

[4] K. Matsunami, F. Sternal, K. Yaginuma, S. Tanabe, H. Nakagawa, H. Sugiyama, Superstructure-based process synthesis and economic assessment under uncertainty for solid drug product manufacturing, BMC Chemical Engineering 2 (2020) 6. doi:10.1186/s42480-020-0028-2.
URL https://bmcchemeng.biomedcentral.com/articles/10.1186/s42480-020-0028-2

[5] P. Suresh, I. Sreedhar, R. Vaidhiswaran, A. Venugopal, A comprehensive review on process and engineering aspects of pharmaceutical wet granulation, Chemical Engineering Journal 328 (2017) 785–815. doi:10.1016/J.CEJ.2017.07.091.

[6] P. Thapa, J. Tripathi, S. H. Jeong, Recent trends and future perspective of pharmaceutical wet granulation for better process understanding and product development, Powder Technology (2019). doi:10.1016/j.powtec.2018.12.080.

[7] C. Vervaet, J. P. Remon, Continuous granulation in the pharmaceutical industry, Chemical Engineering Science 60 (2005) 3949–3957. doi:10.1016/j.ces.2005.02.028.

[8] A. Ito, P. Kleinebudde, Influence of granulation temperature on particle size distribution of granules in twin-screw granulation (tsg), Pharmaceutical Development and Technology 24 (2019) 874–882. doi:10.1080/10837450.2019.1615089.
URL https://www.tandfonline.com/doi/full/10.1080/10837450.2019.1615089

[9] S. H. Kim, K. M. Hwang, C. H. Cho, T. T. Nguyen, S. H. Seok, K. M. Hwang, J. Y. Kim, C. W. Park, Y. S. Rhee, E. S. Park, Application of continuous twin screw granulation for the metformin hydrochloride extended release formulation, International Journal of Pharmaceutics 529 (2017) 410–422. doi:10.1016/j.ijpharm.2017.07.019.

[10] A. Kumar, J. Dhondt, J. Vercruysse, F. D. Leersnyder, V. Vanhoorne, C. Vervaet, J. P. Remon, K. V. Gernaey, T. De Beer, I. Nopens, Development of a process map: A step towards a regime map for steady-state high shear wet twin screw granulation, Powder Technology 300 (2016) 73–82. doi:10.1016/j.powtec.2015.11.067.

[11] H. Liu, S. C. Galbraith, B. Ricart, C. Stanton, B. Smith-Goettler, L. Verdi, T. O'Connor, S. Lee, S. Yoon, Optimization of critical quality attributes in continuous twin-screw wet granulation via design space validated with pilot scale experimental data, International Journal of Pharmaceutics 525 (2017) 249–263. doi:10.1016/j.ijpharm.2017.04.055.

[12] K. Matsunami, T. Nagato, K. Hasegawa, H. Sugiyama, Determining key parameters of continuous wet granulation for tablet quality and productivity: A case in ethenzamide, International Journal of Pharmaceutics 579 (2020) 119160. doi:10.1016/j.ijpharm.2020.119160.
URL https://linkinghub.elsevier.com/retrieve/pii/S0378517320301447

[13] W. Meng, L. Kotamarthy, S. Panikar, M. Sen, S. Pradhan, M. Marc, J. D. Litster, F. J. Muzzio, R. Ramachandran, Statistical analysis and comparison of a continuous high shear granulator with a twin screw granulator: Effect of process parameters on critical granule attributes and granulation mechanisms, International Journal of Pharmaceutics 513 (2016) 357–375. doi:10.1016/j.ijpharm.2016.09.041.
URL https://linkinghub.elsevier.com/retrieve/pii/S0378517316308717

[14] A. Ryckaert, F. Stauffer, A. Funke, D. Djuric, V. Vanhoorne, C. Vervaet, T. D. Beer, Evaluation of torque as an in-process control for granule size during twin-screw wet granulation, International Journal of Pharmaceutics 602 (2021) 120642. doi:10.1016/j.ijpharm.2021.120642.
URL https://linkinghub.elsevier.com/retrieve/pii/S0378517321004476

[15] C. Portier, C. Vervaet, V. Vanhoorne, Continuous twin screw granulation: A review of recent progress and opportunities in formulation and equipment design, Pharmaceutics 13 (2021) 668. doi:10.3390/pharmaceutics13050668.
URL https://www.mdpi.com/1999-4923/13/5/668

[16] M. Verstraeten, D. Van Hauwermeiren, K. Lee, N. Turnbull, D. Wilsdon, M. am Ende, P. Doshi, C. Vervaet, D. Brouckaert, S. T.

Mortier, I. Nopens, T. De Beer, In-depth experimental analysis of pharmaceutical twin-screw wet granulation in view of detailed process understanding, International Journal of Pharmaceutics 529 (2017) 678–693. doi:10.1016/j.ijpharm.2017.07.045.
URL https://linkinghub.elsevier.com/retrieve/pii/S0378517317306440

[17] M. Fonteyne, H. Wickström, E. Peeters, J. Vercruysse, H. Ehlers, B.-H. Peters, J. P. Remon, C. Vervaet, J. Ketolainen, N. Sandler, J. Rantanen, K. Naelapää, T. De Beer, Influence of raw material properties upon critical quality attributes of continuously produced granules and tablets, European Journal of Pharmaceutics and Biopharmaceutics 87 (2014) 252–263. doi:10.1016/j.ejpb.2014.02.011.
URL https://linkinghub.elsevier.com/retrieve/pii/S093964111400071X

[18] K. M. Hwang, C. H. Cho, S. D. Yoo, K. I. Cha, E. S. Park, Continuous twin screw granulation: Impact of the starting material properties and various process parameters, Powder Technology 356 (2019) 847–857. doi:10.1016/j.powtec.2019.08.062.

[19] S. V. Lute, R. M. Dhenge, A. D. Salman, Twin screw granulation: Effects of properties of primary powders, Pharmaceutics 2018, Vol. 10, Page 68 10 (2018) 68. doi:10.3390/PHARMACEUTICS10020068.
URL https://www.mdpi.com/1999-4923/10/2/68/htm https://www.mdpi.com/1999-4923/10/2/68

[20] C. Portier, K. Pandelaere, U. Delaet, T. Vigh, G. Di Pretoro, T. De Beer, C. Vervaet, V. Vanhoorne, Continuous twin screw granulation: A complex interplay between formulation properties, process settings and screw design, International Journal of Pharmaceutics 576 (2020) 119004. doi:10.1016/j.ijpharm.2019.119004.
URL https://linkinghub.elsevier.com/retrieve/pii/S0378517319310658

[21] A. Ryckaert, D. Van Hauwermeiren, J. Dhondt, A. D. Man, A. Funke, D. Djuric, C. Vervaet, I. Nopens, T. De Beer, Tpls as predictive platform for twin-screw wet granulation process and formulation development, International Journal of Pharmaceutics 605 (2021) 120785. doi:10.1016/J.IJPHARM.2021.120785.

[22] S. Garcia-Munoz, Two novel methods to analyze the combined effect of multiple raw-materials and processing conditions on the product's final

attributes: Jrpls and tpls, Chemometrics and Intelligent Laboratory Systems (2014). doi:10.1016/j.chemolab.2014.02.006.

[23] I. Triguero, J. Maillo, J. Luengo, S. Garcia, F. Herrera, From big data to smart data with the k-nearest neighbours algorithm, Proceedings - 2016 IEEE International Conference on Internet of Things; IEEE Green Computing and Communications; IEEE Cyber, Physical, and Social Computing; IEEE Smart Data, iThings-GreenCom-CPSCom-Smart Data 2016 (2017) 859–864doi:10.1109/ITHINGS-GREENCOM-CPSCOM-SMARTDATA.2016.177.

[24] J. Dhondt, J. Bertels, A. Kumar, D. Van Hauwermeiren, A. Ryckaert, B. Van Snick, D. Klingeleers, C. Vervaet, T. De Beer, A multivariate formulation and process development platform for direct compression, International Journal of Pharmaceutics 623 (2022) 121962. doi:https://doi.org/10.1016/j.ijpharm.2022.121962.
URL https://www.sciencedirect.com/science/article/pii/S0378517322005178

[25] B. Schölkopf, A. Smola, K.-R. Müller, Nonlinear component analysis as a kernel eigenvalue problem, Neural Computation 10 (1998) 1299–1319. doi:10.1162/089976698300017467.
URL https://direct.mit.edu/neco/article/10/5/1299-1319/6193

[26] L. Vandevivere, E. Van Wijmeersch, O. Häusler, T. De Beer, C. Vervaet, V. Vanhoorne, The effect of screw configuration and formulation variables on liquid requirements and granule quality in a continuous twin screw wet granulation process, Journal of Drug Delivery Science and Technology 68 (2022) 103042. doi:https://doi.org/10.1016/j.jddst.2021.103042.
URL https://www.sciencedirect.com/science/article/pii/S177322472100722X

[27] B. Van Snick, A. Kumar, M. Verstraeten, K. Pandelaere, J. Dhondt, G. D. Pretoro, T. De Beer, C. Vervaet, V. Vanhoorne, Impact of material properties and process variables on the residence time distribution in twin screw feeding equipment, International Journal of Pharmaceutics 556 (2019) 200–216. doi:10.1016/j.ijpharm.2018.11.076.
URL https://linkinghub.elsevier.com/retrieve/pii/S0378517318309141

[28] D. Van Hauwermeiren, M. Verstraeten, P. Doshi, M. T. am Ende, N. Turnbull, K. Lee, T. De Beer, I. Nopens, On the modelling of

granule size distributions in twin-screw wet granulation: Calibration of a novel compartmental population balance model, Powder Technology 341 (2019) 116–125. doi:10.1016/j.powtec.2018.05.025.
URL https://linkinghub.elsevier.com/retrieve/pii/S0032591018303954

[29] C. Mendez Torrecillas, G. W. Halbert, D. A. Lamprou, A novel methodology to study polymodal particle size distributions produced during continuous wet granulation, International Journal of Pharmaceutics 519 (2017) 230–239. doi:10.1016/j.ijpharm.2017.01.023.
URL https://linkinghub.elsevier.com/retrieve/pii/S0378517317300248

[30] M. S. Escotet-Espinoza, S. Moghtadernejad, J. Scicolone, Y. Wang, G. Pereira, E. Schäfer, T. Vigh, D. Klingeleers, M. Ierapetritou, F. J. Muzzio, Using a material property library to find surrogate materials for pharmaceutical process development, Powder Technology 339 (2018) 659–676. doi:https://doi.org/10.1016/j.powtec.2018.08.042.
URL https://www.sciencedirect.com/science/article/pii/S0032591018306727

[31] Van Hauwermeiren, Daan and Stock, Michiel and De Beer, Thomas and Nopens, Ingmar, Predicting pharmaceutical particle size distributions using kernel mean embedding, PHARMACEUTICS 12 (3) (2020) 26.
URL http://dx.doi.org/10.3390/pharmaceutics12030271

[32] D. Schulze, Practical determination of flow properties, Springer International Publishing, Cham, 2021, pp. 101–139. doi:10.1007/978-3-030-76720-4_4.
URL https://doi.org/10.1007/978-3-030-76720-4_4