Efficient Data-based Methodology for Model enhancement and Flowsheet analyses for Continuous Pharmaceutical Manufacturing

Pooja Bhalode, Nirupaplava Metta, Yingjie Chen, Marianthi Ierapetritou

*Department of Chemical and Biochemical Engineering, Rutgers University, 98 Brett Road, Piscataway, New Jersey, 08854, USA
bDepartment of Chemical and Biomolecular Engineering, University of Delaware, 150 Academy Street, Newark, Delaware, 19716, USA
mgi@udel.edu

Abstract

Accurate process predictability and efficient flowsheet analysis has been the focus of research in pharmaceutical manufacturing. Currently available mechanistic models, being unable to account for all complex powder variabilities, need to be improved for accurate digital twin framework of the manufacturing plant. In this work, we propose a novel methodology of integrated plant data for model enhancement. These enhanced models are applied for flowsheet analyses focusing on design space identification. Surrogate-based feasibility analysis is implemented using artificial neural network with adaptive sampling techniques. This strategy is demonstrated for a direct compaction line capturing realistic scenarios observed in the manufacturing plant.

Keywords: Pharmaceutical, Surrogate, Feasibility, Hybrid modeling, Industry 4.0

1. Introduction

Research efforts driven by FDA’s Quality-by-design (QbD) initiative has led to development of detailed predictive models of continuous pharmaceutical manufacturing lines. These studies mainly focus on process flowsheet simulation using a combination of mechanistic and empirical models to demonstrate dynamic effects of process and material parameters on critical quality attributes, with tracking of powder material properties (Boukouvala et al. 2011; Wang et al. 2017). However, these models are not able to capture all observed process variabilities due to the inherent hard-to-predict nature of pharmaceutical powder materials. In this study, we aim to address these limitations by employing a data-based methodology for model enhancement focusing on inclusion of process data collected under Industry 4.0 framework. Developed process flowsheets models are applied for data-based system analyses. This study focuses on feasibility analysis to address the goal of identification of the design space of a process. In this analysis, a feasibility function which characterizes the maximum constraint violation, is evaluated. For a process with design variables $d$ and uncertain parameters $\theta$, the $J$ constraints that represent feasible operation can be expressed as $g_j(d, \theta) \leq 0$. Following this, the feasibility function is defined as shown in Eq (1).

$$\psi(d, \theta) = \max_{j \in J} g_j(d, \theta)$$  \hspace{1cm} (1)

Since the constraints in the process models are not always available in closed form or are computationally expensive to evaluate, surrogate-based feasibility analysis methodology
is implemented wherein a computationally efficient surrogate model is used to represent the feasibility function. The following sections focus on describing Industry 4.0 dataflow and methodology for model enhancement, followed by feasibility studies using the enhanced models. Concluding remarks are provided in Section 6.

2. Industry 4.0 Dataflow

Recent interest in application of Internet of Things (IoT) (Liao et al. 2017) aimed at developing an integrated data collection, storage, and knowledge extraction framework for efficient and effective manufacturing. This framework for pharmaceutical industry is illustrated in Figure 1, wherein the overall manufacturing plant data is stored to be used for predictive modelling. Data acquisition starts from the bottom up where all data are being generated. Data from Process Analytical Tools (PAT) and sensors are collected and sent to different prediction tools and OPC servers, whereas the raw data from the continuous pilot plant are sent to control software. To facilitate overall data storage and integration, data from these sources are sent to a local historian and then to a cloud storage infrastructure for ease of accessibility for predictive digital twin models. Material calibration and experimental data is stored using standardized data structures within cloud-based electronic laboratory notebooks (ELNs) for efficient data integration with the overall plant and extraction to modelling and simulation platforms.

![Figure 1: Data flow within Industry 4.0 framework for continuous pharmaceutical manufacturing](image)

3. Methodology for Model Enhancement

In this section, the methodology for model enhancement based on the Industry 4.0 framework is presented. With this framework, process model outputs and actual plant data can be compared and analyzed to identify if a significant residual exists, indicating a potential departure between the model and the plant. The following sub-sections entail different methods for model enhancement with process data.

3.1. Identification and correction of process-model mismatch

A typical cause of departure can be attributed to the parameters used in the model (namely parametric mismatch). It is important to identify a suitable set of parameters that can appropriately describe the system. Using Industry 4.0 data framework, different parameter estimation methods including least squares fitting and Kalman filtering (Zhang 1997) can be applied to obtain a set of physically suitable parameters. Kravaris et al. (2013) identifies global optimization-based algorithms for parameter estimations under measurement uncertainty, and Wang et al. (2018) introduces a Bayesian approach for data from different scales. In certain situations, if the best set of parameters still does not yield
good model predictability, structural mismatches exist within the model, meaning that the functional forms within the model are not suitable to describe the actual system. For identification of structural mismatches, principal component analysis and correlation methods like partial correlation and mutual information are implemented (Badwe et al. 2009; Chen et al. 2013; Meneghetti et al. 2014). These methods can qualitatively determine the sources of structural mismatch. After identifying the source of mismatch, model selection and discriminations methods like posterior probability study and Akaike Information Criterion can be used to identify the most probable model (Wu et al. 2011; Wang et al. 2018).

3.2. Non-Parametric or Surrogate modeling

The non-parametric or surrogate model development method is implemented if the desired model predictability is not achieved after addressing process-model mismatch or when the cost of mismatch identification and correction is high. Non-parametric modeling includes developing an efficient surrogate based on input-output datasets. Kriging model, Artificial Neural Networks, and related methods are often used in developing the surrogate (Rogers et al. 2013). Datasets availed from the Industry 4.0 framework are applied to develop the computationally efficient surrogate models which captures the overall process without inclusion of mechanistic information.

3.3. Hybrid Modeling

Hybrid modeling approach focuses on capturing the mechanistic information along with data-driven surrogate models. The essence is to combine a priori knowledge like conservation and kinetic laws with nonparametric models built using process data (Stosch et al. 2014). Proposed hybrid structures can be broadly categorized as parallel and serial structure. For parallel structure, inputs are fed into both the mechanistic part and data-driven part of the model, and the model outputs are combined by superposition, multiplication, or weighting. This structure is often used to correct for errors of the original model. For serial structure, inputs are fed into the mechanistic and data-driven models in a sequential way, meaning that the output of the first part of the model serves as an input to the next, and the final outputs can be combined with appropriate weighting. The serial structure is prevalent when mechanism of the original model is not well-understood (Stosch et al. 2014). By integrating the mechanistic and data-driven models, the overall predictability of the model can be increased, and the enhanced model can then be integrated into flowsheet to conduct further flowsheet analyses.

4. Surrogate-based Feasibility Analysis

This section focuses on surrogate based feasibility analysis methodology where, an inexpensive feed-forward ANN model combined with Bayesian regularization (Burden and Winkler 2008) for network training is explained. The number of hidden neurons is determined in an initial model selection, minimizing the sum of squared errors between target and prediction data. In the model improvement stage, adaptive sampling is used to choose additional samples for improving the ANN model. In this work, a modified expected improvement function ($EI_{feas}$) proposed by Boukouvala and Ierapetritou (2012) as given in Eq. (2) is used to implement the adaptive sampling strategy.

$$EI_{feas}(x) = \Phi \left( \frac{-\mu}{s} \right) = s \frac{1}{\sqrt{2\pi}} e^{-0.5 \frac{y^2}{s^2}} \tag{2}$$
where, \( y \) and \( s \) are the surrogate model predictor and standard error of the prediction at \( x \) respectively. Maximization of \( EI_{feas} \) identifies sample points close to the feasible region boundary or in the region of high prediction uncertainty. Thus, as samples are added, the surrogate model accuracy is improved focusing on identification of feasible region. In this work, the variance \( s^2 \) of ANN model predictor \( y \) at a sample point \( x \) is estimated using a statistical technique known as jack-knifing (Eason and Cremaschi 2014) wherein, sample set is divided into \( K \) disjoint sample sets. \( K \) neural networks are built for the sets and variance of prediction \( s^2 \) is estimated using Eq (3).

\[
s^2(x) = \frac{1}{K} \sum_{k=1}^{K} \left( U_k(x) - \frac{1}{K} \sum_{k=1}^{K} U_k(x) \right)^2
\]

Figure 2 illustrates the algorithm for implementation of the ANN based feasibility analysis methodology. This methodology is tested on Continuous Direct Compaction (CDC) process explained by (Wang et al. 2017), where the uncertain parameters used for the problem and their bounds are tabulated in Table 1. The constraints for the problem that define the feasible region is given in Table 2.

![Algorithm for ANN based feasibility analysis methodology](image)

**Table 1: Variable bounds for CDC process feasibility case study**

<table>
<thead>
<tr>
<th>Factors</th>
<th>Units</th>
<th>Nominal</th>
<th>LB</th>
<th>UB</th>
</tr>
</thead>
<tbody>
<tr>
<td>API flowrate</td>
<td>kg/h</td>
<td>3</td>
<td>2.85</td>
<td>3.15</td>
</tr>
<tr>
<td>Excipient flowrate</td>
<td>kg/h</td>
<td>26.7</td>
<td>25.36</td>
<td>28.03</td>
</tr>
<tr>
<td>Mill impeller speed</td>
<td>rpm</td>
<td>1120</td>
<td>1064</td>
<td>1176</td>
</tr>
<tr>
<td>Blender blade speed</td>
<td>rpm</td>
<td>250</td>
<td>237.5</td>
<td>262.5</td>
</tr>
<tr>
<td>Fill depth</td>
<td>m</td>
<td>0.01</td>
<td>0.0095</td>
<td>0.0105</td>
</tr>
<tr>
<td>Thickness</td>
<td>m</td>
<td>0.0025</td>
<td>0.002375</td>
<td>0.002625</td>
</tr>
</tbody>
</table>

Overall, the ANN based feasibility analysis is applied to the six-dimensional problem with 20 inequality constraints. \( 2^6 \) Latin Hypercube samples are used in the initial model selection phase. Using the initial samples, number of hidden neurons are varied. 8 hidden neurons yield the least sum of squared errors between predicted and target feasibility function values. 500 samples are iteratively added using the adaptive sampling strategy. Since this is a high dimensional problem, the feasible region is represented using a matrix of contour plots of feasibility function values as shown in Figure 3. For each contour plot, only two factors are varied, and the remaining two factors are set at nominal values. The feasible region boundary predicted by the final ANN model is represented using a red
Efficient Data-based Methodology for Model Enhancement and Flowsheet analyses for Continuous Pharmaceutical Manufacturing

line. Feasible region boundary from the original process model is also plotted using red dashed line. Good agreement between the original and predicted feasible regions is observed.

Table 2: Constraints for the CDC process feasibility case study

<table>
<thead>
<tr>
<th>Unit operation</th>
<th>Variable</th>
<th>Units</th>
<th>Limits based on nominal value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blender</td>
<td>Mean residence time</td>
<td>s</td>
<td>+/- 20%</td>
</tr>
<tr>
<td></td>
<td>Delay time</td>
<td>s</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Mass hold SS</td>
<td>kg</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Mean bulk, true density</td>
<td>kg/m³</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Mean d10, d50, d90</td>
<td>µm</td>
<td></td>
</tr>
<tr>
<td>Comill</td>
<td>Mean d10, d50, d90</td>
<td>µm</td>
<td>+/-20%</td>
</tr>
<tr>
<td></td>
<td>Mean bulk, true density</td>
<td>kg/m³</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Mean residence time</td>
<td>s</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Mass holdup SS</td>
<td>kg</td>
<td></td>
</tr>
<tr>
<td>Tablet Press</td>
<td>Concentration</td>
<td>%</td>
<td>+/-5%</td>
</tr>
<tr>
<td></td>
<td>Weight</td>
<td>kg</td>
<td>+/-10%</td>
</tr>
<tr>
<td></td>
<td>Hardness</td>
<td>kp</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Main compression pressure</td>
<td>MPa</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Pre compression pressure</td>
<td>MPa</td>
<td></td>
</tr>
</tbody>
</table>

Figure 3: Matrix of feasibility function contour plots representing the feasible region of the continuous direct compaction process (Feasible region boundary from the original model and predicted by the ANN model are represented as red dashed lines and red solid lines respectively)
5. Conclusions

We propose an integrated methodology for data-based model enhancement using Industry 4.0 framework for surrogate-based feasibility analysis. Pilot plant and laboratory data collected from Industry 4.0 framework is applied for enhancement of mechanistic models. The enhanced flowsheets are proposed to be applied for feasibility analysis on a continuous direct compaction line. A good agreement between the actual and the predicted feasible regions is observed. For future work, we aim to extend the methodology to other continuous manufacturing lines with the inclusion of process dynamics.

Acknowledgement

The authors would like to acknowledge funding from U.S. Food and Drug Administration (FDA) through grant 1U01FD006487-01.

References