# Equation-based and data-driven modeling strategies for industrial coating processes

Paris Papavasileiou[a,b], Eleni D. Koronaki[c,b,*], Gabriele Pozzetti[d], Martin Kathrein[d], Christoph Czettl[e], Andreas G. Boudouvis[b], Stéphane P.A. Bordas[a]

[a] *Faculty of Science, Technology and Medicine, University of Luxembourg, Maison du Nombre, 6 Avenue de la Fonte, Esch-sur-Alzette, L-4364, Luxembourg*
[b] *School of Chemical Engineering, National Technical University of Athens, 9 Heroon Polytechniou str., Zographos Campus, 15780, Attiki, Greece*
[c] *Interdisciplinary Center for Security, Reliability and Trust, University of Luxembourg, 29 John F. Kennedy Avenue, Luxembourg, L-1855, Luxembourg*
[d] *CERATIZIT Luxembourg S.à r.l., Mamer, L-8201, Luxembourg*
[e] *CERATIZIT Austria GmbH, Reutte, A-6600, Austria*

## Abstract

Computational Fluid Dynamics (CFD) and Machine Learning (ML) approaches are implemented and compared in an industrial Chemical Vapor Deposition process for the production of cutting tools. In this work, the aim is to analyze the pros and cons of each method and propose a blend of the two approaches that is suitable in industrial applications, where the process is too complicated to address with first-principles models and the data do not allow the implementation of data-hungry methods. Both approaches accurately predict the coating thickness (Mean Absolute Percentage Error (MAPE) of 6.0% and 4.4% for CFD and ML respectively for the test case reactor). CFD, despite its increased computational cost, both in terms of de-

[*] Corresponding author
*Email address:* ekor@mail.ntua.gr (Eleni D. Koronaki)

veloping and also calibrating for the application at hand, provides meaningful insight and illuminates the process. On the other hand, ML can provide predictions in a time-efficient manner, and is thus appropriate for inline and concurrent predictions. However, it is limited by the available data and has low extrapolation ability. Equation-based and data-driven methods are combined by exploiting a handful of CFD results for efficient interpolation in a reduced space defined by the principal components of the dataset, by implementing Gappy POD. This allows for the accurate reconstruction of the full state-space with limited data.

*Keywords:* Chemical Vapor Deposition, Coating, supervised learning, CFD modeling, industrial application, production data

Published journal article available at:
https://doi.org/10.1016/j.compind.2023.103938

## 1. Introduction

Chemical Vapor Deposition (CVD) processes are popular in a wide range of applications, including microelectronics (Creighton and Parmeter, 1993), sensors (Ozaydin-Ince et al., 2011) and wear resistant coatings (Kathrein et al., 2003). The coating process involves the nonlinear interplay of physical mechanisms, such as diffusion and convection, with a plethora of homogeneous and heterogeneous chemical reactions. The competition between the different mechanisms determines the process outcome and the product quality. It is therefore a fine example of a process that is too complicated to study with first-principles models, such as Computational Fluid Dynamics (CFD) and where the data is often not enough to implement sophisticated data-driven strategies. Taking all of the above into consideration, the objective of this work is to investigate the potential benefit of simplified CFD process models, accompanied by purely data-driven predictions using Machine Learning (ML) approaches. Both methods are driven by the size and type of the available production data. In the CFD case, the data are used for calibration and validation and in the ML case for regression.

Computational Fluid Dynamics is a valuable tool for studying deposition processes (Kleijn and Hoogendoorn, 1991; Kleijn et al., 2007; Cheimarios et al., 2012; Cho and Mountziaris, 2013; Psarellis et al., 2018; Koronaki et al., 2019; Papavasileiou et al., 2022), since it allows the investigation of the flow field inside the reactor, as well as the main physical and chemical pathways that lead to the deposition of thin film coatings. Nevertheless, modeling industrial-scale deposition applications using CFD presents several challenges: Firstly, dealing with the complexity of the process, which often has several unknowns and secondly, the large scale of real applications.

Specifically, the actual chemical reactions that lead to deposition, including their rates, are often unknown. Therefore, it is not possible to predict the effect of the interplay between transport phenomena and chemical kinetics on the deposition rate, necessitating the development of a kinetic model (Topka et al., 2022). Even when a chemical reaction scheme is available, some of its parameters may need to be fitted for the specific application. This parameter fitting involves an increased computational cost, as it usually requires numerous simulations (Gakis et al., 2015; Koronaki et al., 2016; Gkinis et al., 2017, 2019). Nevertheless, CFD has been applied to several CVD applications, shedding light on previously “opaque” processes (Fotiadis et al., 1990; Liu and Xiao, 2015; Aviziotis et al., 2017) while also allowing to predict their outcomes (Endo et al., 2004). Although attempts have been made towards increasing the efficiency of CFD models by implementing reduced order modeling methods (Gkinis et al., 2019; Spencer et al., 2021), developing an efficient and accurate model in an industrial setting remains a challenging and time-consuming task.

In the era of Industry 4.0, digitalization has become one of the main drivers of innovation (Kagermann, 2015) and production data are becoming more and more available. The industry is trying to exploit this data, seeking improvement in several domains, including: maintenance management (Saxena and Saad, 2007; Susto et al., 2015; Wu et al., 2019; Dalzochio et al., 2020) quality management (Kim et al., 2012, 2018; Carvajal Soto et al., 2019; Iqbal et al., 2019; Wang et al., 2022), production planning and control (Priore et al., 2018; Tulsyan et al., 2018; Ma et al., 2019; Agarwal et al., 2020; Deng et al., 2022), supply chain management (Du and Jiang, 2019), process outcome predictions (Cai et al., 2020; Azadi et al., 2022; Dai et al., 2022; Malley et al., 2022) and process optimization (He et al., 2021; Galvis et al., 2022). Furthermore, digital twins (Boyes and Watson, 2022) are becoming increasingly popular in the process industry (Hürkamp et al., 2020; Rasheed et al., 2020; Perno et al., 2022), as well as in other, diverse applications (Urcun et al., 2021; Kalaboukas et al., 2023). Although the application of sophisticated methods such as Deep Neural Networks (DNNs) (Blakseth et al., 2022; Deshpande et al., 2022), Physics Informed Neural Networks (PINNs) (Raissi et al., 2019) and manifold learning (Koronaki et al., 2023) has been demonstrated on controlled small scale problems, several challenges still remain when incorporating ML in everyday industrial practice. Addressing these challenges is one of the main objectives of this work.

The industrial application in this work is the coating of cutting tools with $\alpha$-$Al_2O_3$ for increased wear resistance. Concerning CFD, the goal is to propose the best possible simplified model, based on the available data which are necessary for verification and validation. This leads to a 2D, time-dependent CFD model, presented in detail in previous work (Papavasileiou et al., 2022). The proposed model implements representative boundary conditions and employs a simple reaction scheme for the $\alpha$-$Al_2O_3$ deposition with the goal of reducing the computational cost.

Concerning ML, the first task is to pre-treat the available data, upon which the choice of method depends on. Addressing mixed types of data (categorical and numerical) is a common challenge in many applications, not restricted to deposition processes. Several regression models are trained to predict the $\alpha$-$Al_2O_3$ coating thickness using characteristics of the reactor set-up and process conditions as inputs. In this work, the focus lies more on tree-based methods (James et al., 2021b) which are the best-performing for the given data-set.

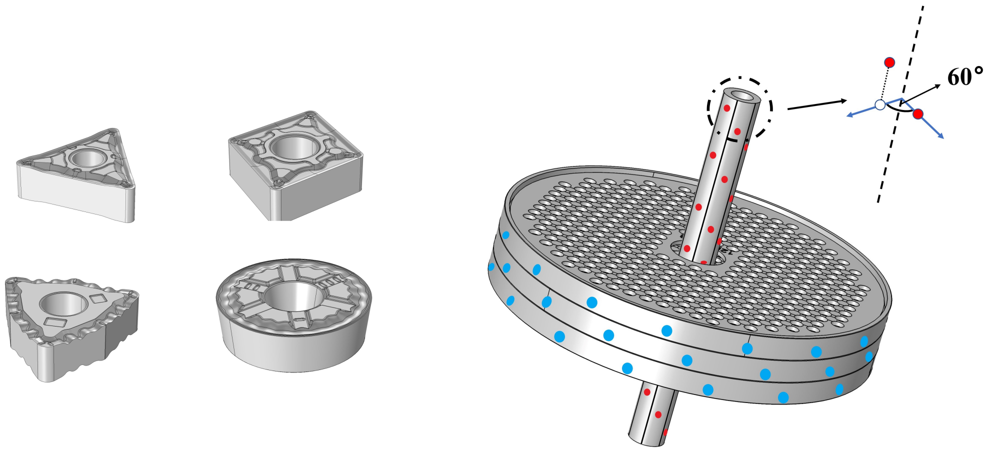
The two approaches are initially compared in their ability to accurately and efficiently predict the alumina coating thickness of the cutting tool inserts. Specifically, the advantages and disadvantages of each strategy are assessed in terms of accuracy, interpretability, extrapolation ability and computational cost. As a final step, the two approaches are merged through the implementation of the Gappy Proper Orthogonal Decomposition (Gappy POD) method (Everson and Sirovich, 1995; Willcox, 2006). The latter, is popular for optimal sensor placement, and here it adapted to propose a sufficient number of known data from which we can infer quantities that are not measurable.

The manuscript is structured as follows: A concise overview of the process and the available production data is given in Section 2. The implemented methods (CFD, ML and Gappy POD) are presented in Sections 3 and 4. The results of each method are analyzed and compared in Section 5, followed by the conclusions in Section 6.

## 2. Process description

A two-step coating process takes place inside the studied industrial-scale, commercial CVD reactor (Sucotec SCT600TH). First, a Ti(C,N) base layer of about 9 $\mu$m is grown on the cemented carbide cutting inserts, such as the ones shown in Fig. 1a. Subsequently, an alumina layer is deposited under a $AlCl_3$–$CO_2$–HCl–$H_2$–$H_2S$ chemical system. The temperature and pressure for the alumina coating step are $T$=1005°C and $p$=80 mbar, respectively (Hochauer et al., 2012). The alumina coating deposition step of the process takes approximately 3 hours.

The CVD reactor consists of 40-50 perforated disks, stacked one on top of the other, whereon the inserts are placed. In Fig. 1b, a schematic of three such disks is shown for clarity. The mixture of gas reactants, enters the reactor via perforations on a rotating cylindrical tube, placed in the center of the structure of the stacked disks. There are two antipodal perforations for each disk level. There is a 60° angle difference between the axis connecting the inlet holes for each disk level. The rotational motion of the inlet tube (rotating with a rotational speed of 2 RPM) causes the process to have an inherent periodic nature. The interior geometry of the reactor changes from production run to production run, since the geometry of the inserts (and the disks on which they are placed), changes based on production requirements.



(a)

(b)

Figure 1: (a) Indicative geometries of the coated cutting tools. (b) A 3D representation of a 3-disk part of the reactor. The inlet perforations on the rotating inlet tube are shown in red. The outlet perforations for each disk are shown in blue.

The main goal of the process is to achieve uniform coating thickness, since this uniformity also leads to uniform product longevity (Bar-Hen and Etsion, 2017). Ideally, coating thickness uniformity would be achieved across all production runs, reactors, and production sites. However, this is not always the case. For this reason, a way of predicting the coating thickness of the inserts given the reactor set-up is needed. Furthermore, coming up with a systematic way of assessing the factors that influence the coating thickness uniformity is also highly important.

### *2.1. Available data*

For the Ti(C,N)/$\alpha$-$Al_2O_3$ multi-layer coating, the thickness measurements are performed via the Calotest method. A small spherical cavity is ground on the coated inserts using a rotating ball of known geometry, providing a tapered cross-section of the film when viewed under an optical microscope (Łępicka and Grądzka-Dahlke, 2019). This way, the thickness of both the Ti(C,N) and $\alpha$-$Al_2O_3$ coating layers can be calculated. Measurements are usually taken for 3 positions on 5 disks of interest. Therefore, 15 thickness measurements are available for each production run. A 2D representation of the reactor indicating the points where thickness is typically measured is shown in Fig. 2. These measurements allow for not only for the calibration and validation of the CFD model, but also for several ML approaches.

Apart from coating thickness measurements, the dataset also contains several features concerning the process and the reactor setup, which will serve as inputs to the machine-learning model. The production "recipe" used for the coating is the available feature providing information regarding the process. Setup-wise, there is a plethora of available features for each disk of the reactor, including:

1. The position of each disk inside the reactor.
2. The number of inserts placed on each disk.
3. The type of insert placed on each disk. Each type of insert has different geometrical characteristics.
4. The type of disk used. The type of disk used is always relative to the type of insert placed on top of it.

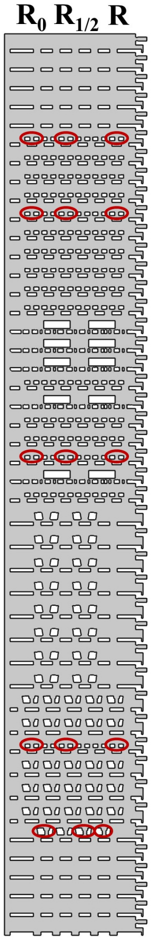


Figure 2: Positions with available $\alpha$-$Al_2O_3$ thickness values from the production data for our test case. In general, across different production runs, the R position (the one closest to the reactor outlet) is the one with the highest amount of data. For this reason, the ML models are trained to make predictions for inserts placed in this position.

5. The surface area of the inserts placed on the disk.

These features allow for the creation of more features, such as the total surface area and the standard deviation of the surface area of the inserts that are coated inside the reactor. Another feature that can be created is the difference between the nominal surface area of the production "recipe" and the actual insert surface area inside the reactor. Furthermore, for each disk, we can exploit the information available for its neighboring disks.

This way, we end up with several features, of which thirteen are used as inputs after being pre-processed. These features are summarized in Table 1. Considering the coating thickness measurements as outputs, we can train several supervised learning models to make coating thickness predictions per disk. In this context, during training, a labeled set of inputs is provided and specifically here, the inputs are the aforementioned features and the labels are the $\alpha$-$Al_2O_3$ coating thickness measurements.

## 3. Computational ingredients

### *3.1. ML methods*

For the data-driven approach to the problem, the implementation of an assortment of machine learning methods for the prediction of coating thickness inside the reactor is investigated. All methods implemented fall into supervised learning methods.

In supervised learning, each one of the input variables $x_i$ is associated with a response (or output) $y_i$ (James et al., 2021a). The goal of the ML strategy is to train a model able to relate the input variables $x_i$ to the output $y_i$. This way, future observations can be predicted and the relationship between the

Table 1: Summary of the features included in the training of the regression models.

| **Feature** | **Type** | **Pre-processing** |
|---|---|---|
| Number of inserts on disk | Numerical (integer) | standardization |
| Surface area of inserts on disk | Numerical (float) | standardization |
| Disk position | Numerical (integer) | standardization |
| Total surface area of inserts inside the reactor | Numerical (float) | standardization |
| Surface area standard deviation | Numerical (float) | standardization |
| \|Nominal “recipe” surface area - actual surface area\| | Numerical (float) | standardization |
| Production “recipe” | Categorical | binary encoding |
| Insert geometry | Categorical | binary encoding |
| Disk geometry | Categorical | binary encoding |
| Insert geometry – disk above | Categorical | binary encoding |
| Insert geometry – disk below | Categorical | binary encoding |
| Disk geometry – disk above | Categorical | binary encoding |
| Disk geometry – disk below | Categorical | binary encoding |

inputs and the output can be interpreted. Here, the goal is to predict the $\alpha$-$Al_2O_3$ coating thickness (a continuous target variable) from several inputs, using a regression method. The specific methods include but are not limited to:

- Linear methods, such as linear, lasso or ridge regression.

- Non-linear methods, such as polynomial regression.

- Tree-based methods, such as regression trees and their ensemble versions: random forests, gradient boosted regression trees and extreme gradient boosted regression trees.

- Artificial neural networks.

During the early phases of this research, several techniques were utilized, including linear, lasso, and ridge regression, as well as support vector machines and Gaussian process regression. Preliminary findings indicated that tree-based methods outperformed the other techniques, and as a result, the focus of this study is on tree-based methods.

The models' accuracy will be evaluated via two different metrics, namely the mean absolute error (MAE) and the mean absolute percentage error (MAPE). When the model is trained or tested on $N$ observations and for each observation $i$ the prediction is $\hat{y}_i$ while the actual value is $y_i$, MAE and MAPE can be written as follows:

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N} |\hat{y}_i - y_i| \tag{1}$$

$$\text{MAPE} = \frac{1}{N}\sum_{i=1}^{N} \frac{|\hat{y}_i - y_i|}{y_i} \tag{2}$$

Two different computational costs pertain to each ML model, the training time ($t_{\text{train}}$) and the prediction time ($t_{\text{pred}}$) of the model. Both of these costs are expressed in CPU time.

### *3.1.1. Tree-based methods*

Tree-based methods work by partitioning the space of the inputs $X$ into a set of rectangles. Afterwards, a simple model (e.g. a constant) is fit in each partition. The process starts by splitting the entire input space in two based on a variable of the input space and its value. The optimal variable and split point are chosen in order to achieve an accurate fit. Then, either or both of the resulting regions are split again in two, once again using the optimal input and split point. This procedure continues until a stopping criterion has been met. The occurring binary splits allow for model interpretability since the entire sample space can be described by a single tree. Tree-based methods can be used for both regression and classification purposes (Hastie et al., 2009a).

The prediction accuracy of a single tree is often not as high as that of other methods. Furthermore, a small change in the data can lead to an entirely different tree layout. These two issues and especially the predictive performance of the trees can be rectified by combining multiple trees through the implementation of ensemble methods such as bagging and boosting (James et al., 2021b).

The concept behind ensemble methods is to build a prediction model by combining a number of simpler base methods, in two steps: First, a number of base learners must be created from the available data. The second step involves the combination of these learners into one ensemble predictor. The most common ensemble tree-based methods are random forests, bagged trees and gradient boosted trees. These methods, however, have some key differences between them.

Random forests and bagged trees, discussed here, operate similarly. They both build $B$ regression trees and each tree is trained using bootstrap-sampled (i.e. sample a particular data-point and then reintroduce it to the dataset), versions of the original dataset. Bagging regression methods provide a prediction by averaging the outputs of the $B$ trees that they consist of. If $\hat{y}_{i,b}$ is the prediction of each grown tree, then the final prediction of the bagging method $\hat{y}_{i,bag}$ is given by:

$$\hat{y}_{i,bag} = \frac{1}{B} \sum_{b=1}^{B} \hat{y}_{i,b} \tag{3}$$

Random forests and bagged trees differ only in the amount of input features $N_{\text{input}}$ that are considered when building each tree. In bagged trees, all available features are considered. On the contrary, in random forests, a random subset of $p$ input features is considered. This serves the purpose of de-correlating the individual trees, since the trees are not always built by selecting the global optimal features, but by selecting the optimal feature from a randomly sampled subset of the input features (James et al., 2021b).

Gradient boosting and extreme gradient boosting are boosting methods. In the case of boosting methods, contrary to bagging methods, the $B$ base trees are created sequentially. First, the first tree of the ensemble is created. Afterwards, each created tree is fitted to the difference between the value predicted by the previous tree and the real output. This way, each tree improves the shortcomings of the previous one. There is no averaging of the result of the $B$ trees in this case (Hastie et al., 2009b).

Therefore, after building the $b^{th}$ tree which outputs $\gamma_{jb}$ and is trained on the residual of the output of the ensemble after the previous tree has been built, the output of the ensemble $f_b(x)$ can be written as:

$$f_b(x) = f_{b-1}(x) + \lambda \cdot \sum_{j=1}^{J} \gamma_{jb} I(x \in R_{jm}) \tag{4}$$

where $I$ is the indicator function, and $\lambda$ is the learning rate of the boosting procedure. $\lambda$ serves the purpose of scaling the contribution of the output of each tree to the final prediction of the ensemble.

The result of the model is the output of the ensemble after the final tree has been built. Boosting methods are more prone to overfitting for large values of $B$ than bagging methods. For this reason, $B$ needs to be carefully selected through cross-validation.

### *3.1.2. Challenges*

Applying data-driven methods to a real-world dataset presents several challenges. First and foremost, the dataset needs to be “cleaned”: Given that the production dataset is derived from different production sites, different reactors, and different people, it is bound to contain some errors. These errors must be identified and corrected before any type of analysis. Then, there is the question of the format of the data. Even when the data is neatly organized in an SQL database, it still needs to be extracted and formatted (using the pandas python library (McKinney, 2010), for example) so that it can be used to train models in a python framework. Afterwards comes the question of data type. In this particular application, there are both numeric and alphanumeric features (features that contain names instead of values). Since several of the implemented methods are not compatible with alphanumeric (categorical) features, those features need to be encoded in a way (i.e. binary encoding, one-hot encoding (Potdar et al., 2017)) that allows them to be used in our models. Finally, once the data is ready, the task is to find the best performing model and to determine the hyperparameters that influence performance. Therefore, a hyperparameter optimization step must also be included. By following this step-by-step approach, we can establish a data pipeline specific to our data that allows us to overcome all the aforementioned challenges. This however requires experience, input from the process experts, along with a clear understanding of the data.

### *3.2. CFD modeling: Implementation & challenges*

For this specific application, a digital “replica” of the process would have to be a 3D, time-dependent full reactor (40-50 disks) model which would include a complex reaction scheme. A complex reaction scheme, would lead to more degrees of freedom and an increased number of kinetic parameters that would need to be fitted. Apart from this, given the rotation of the inlet tube (and therefore the fact that the problem is not axisymmetric) a moving mesh would also need to be implemented. This would translate into a computationally intractable task. If we consider that the reactor interior geometry changes on a day-to-day basis, since the geometries of inserts and the disks on which they are placed change based on production quotas, a computationally expensive model is not a suitable method to study this industrial application. For this reason, aiming to drive the computational cost down, the problem was approached as follows:

- The problem is modeled in 2D.

- The boundary conditions for both the inlet and the outlet are selected in a way that is representative of their 3D characteristics.

- The model takes into account only 7-disk parts of the reactor in a divide and conquer approach.

- A simpler reaction scheme that still leads to accurate results is used.

To efficiently tackle the challenges of the process, a 2D, time-dependent model that accounts for the transport of mass, momentum, and species inside the reactor is proposed. The COMSOL Multiphysics® software was used for the CFD modeling. The interested reader can seek detailed information in the recent work of Papavasileiou et al. (2022); here the key points are summarized for completeness.

A reaction scheme consisting of a homogeneous reaction in the gas phase and a heterogeneous reaction for the deposition of $\alpha$-$Al_2O_3$ is part of the model. The following assumptions are made: a) laminar and incompressible flow, b) constant temperature of in the entire reactor domain, c) ideal gas phase. The CFD model accounts for 7-disk "building blocks" of the reactor, in order to keep the computational cost low. To account for the rotation of the inlet tube, pulse velocity boundary conditions are applied at the inlets. To represent the placement of the holes on the inlet tube in the 2D computational geometry, a phase difference is included between the boundary conditions of each disk. A similar approach is taken for the outlet perforations. Since they are not aligned, pressure boundary conditions are applied at every other disk (1st open, 2nd closed and so forth). In order to model the deposition of $\alpha$-$Al_2O_3$ under the $AlCl_3$–$CO_2$–HCl–$H_2$–$H_2S$ chemical system, we implement a simple reaction scheme based on the work of Schierling et al. (1999). Implementing this simpler scheme results in a lower computational cost. The simulations account for two full rotations (or periods) of the feeding tube.

## 4. Combining equation-based and data-driven approaches using Gappy POD

In this work, the Gappy POD method is used for the reconstruction of several 7-disk reactor snapshots acquired using the aforementioned CFD model using limited - or “gappy” data. Gappy POD was first introduced by Everson and Sirovich (1995) and then implemented, among others, to a CFD airfoil application by Willcox (2006) and for non-linear fracture mechanics modeling (Kerfriden et al., 2013). Optimal sensor placement is another problem that can be solved using the Gappy POD method, as indicated in the works of Willcox (2006) and Jo et al. (2019). This is achieved by finding the optimal way of filling the “gaps” in the data, or in other words, selecting the sensor positions that give the most information possible.

A concise overview of the method, along with the procedure followed for the acquisition of data and the metrics used for the evaluation of the method, are presented in the following paragraphs.

### *4.1. Overview*

In this section, the Gappy POD method is summarized for completeness. Let’s consider a dataset $\mathbf{X}$ of M vectors (represented as $d$-dimensional real vectors $x_1, \dots, x_M$). A POD basis, $\mathbf{\Phi} \in \mathbb{R}^{N \times M}$, of $\mathbf{X}$ is computed, such that $\mathbf{X}$ can be approximated as a linear combination of p vectors:

$$\widetilde{\mathbf{X}} = \sum_{j=1}^{p} c^j \mathbf{\Phi}^j \tag{5}$$

or in matrix-vector format:

$$\widetilde{\mathbf{X}} = \mathbf{\Phi} \cdot c \tag{6}$$

The size of the truncated POD basis $\mathbf{\Phi}$ is selected based on the error between the actual vector $\mathbf{X}$ and the reconstructed approximation $\tilde{\mathbf{X}}$ :

$$\text{reconstruction error} = \|\mathbf{X} - \tilde{\mathbf{X}}\| \tag{7}$$

Another factor that can be taken into account when selecting the size of the truncated basis is the total energy retained by the selected number of modes. For each basis vector $j$, the relative importance $(E_j)$ is given by:

$$E_j = \frac{\lambda_j}{\sum_{i=1}^{p} \lambda_i} \tag{8}$$

and therefore, the total energy retained for the $k$ retained modes is given by:

$$E_{\text{total}} = \sum_{j=1}^{k} E_j \tag{9}$$

Let us consider a vector $X'$ that is spanned by the same basis $\Phi$ and that only $m$ values of this vector are known, such that the partial vector $X'_{partial}$ can be defined:

$$X'_{\text{partial}} = m \cdot X', m \in \mathbb{R}^{m \times N} \tag{10}$$

The goal is to find coefficients $c'$, such that an approximation $\tilde{X}'$ of the vector $X'$ can be defined as :

$$\tilde{X'} = X' \cdot c' \tag{11}$$

then:

$$X'_{\text{partial}} \approx m \cdot X' \cdot c' \tag{12}$$

Finding the values of $c'$ that satisfy the above leads to an optimization problem, which results in the solution of the linear system:

$$M \cdot c' = (m \cdot \mathbf{\Phi})' \cdot X'_{\text{partial}} \tag{13}$$

with $M = (m \cdot \mathbf{\Phi})' \cdot (m \cdot \mathbf{\Phi})$

### *4.2. CFD data sampling*

Snapshots, i.e. vectors containing information regarding the system's state at a specific time, of 12 different 7-disk reactor parts will be used for the implementation of the Gappy POD method. For each reactor part, there 31 available time-instances (each one with 1 second time difference from the previous). This way, the full dataset consists of 372 vectors.

At each time-instance, 4 quantities of interest are sampled along the lines connecting inlet-outlet at each disk level. The points of these lines are then interpolated at 250 specific query points using linear interpolation. In this manner, 250 evenly spaced points along each line are obtained. An example of the lines along which the quantities of interest are sampled is demonstrated in Fig. 3.

The quantities of interest at each point are:

1. The velocity magnitude ($U$).

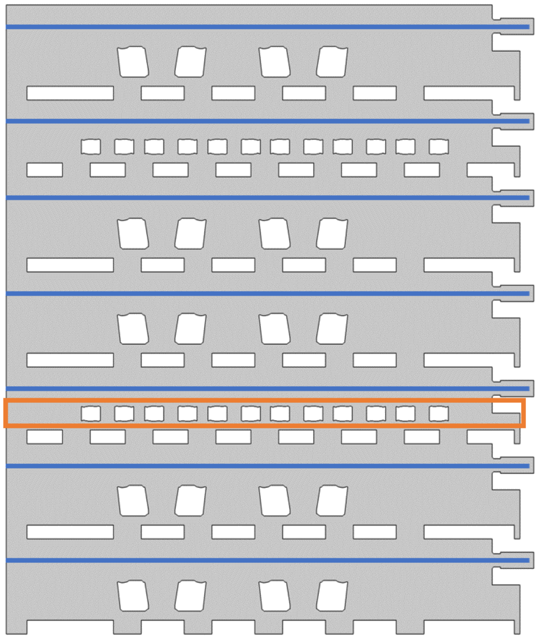


Figure 3: In blue: The seven lines along which the 4 quantities of interest $(U, p, C_{AlCl_3}, C_{H_2O})$ are sampled. In orange: The disk with available thickness measurements. The thickness measurements, as well as the $\alpha$-$Al_2O_3$ deposition rates at the inserts of this disk, are also included for our implementation of Gappy POD.

2. The pressure $(p)$.
3. The concentration of the precursor $AlCl_3$ $(C_{AlCl_3})$.
4. The concentration of water $(C_{H_2O})$.

Furthermore, the deposition rates as predicted by the CFD model along with the available thickness data for 3 positions ($R_0$, $R_{1/2}$, $R$) for each 7-disk reactor part, are included in each snapshot. An overview of the resulting dataset after sampling and organizing the vectors is presented in Fig. 4.

It is worth noting that a plethora of input parameters influences the final product, the most important of which include the configuration of the reactor's interior geometry and the production "recipe". The latter includes all the steps and chemical species involved in the production of a single coating layer. In this work, to make the simulations tractable, the focus lies on a single "recipe" for a single product and various geometries, without loss of generality.

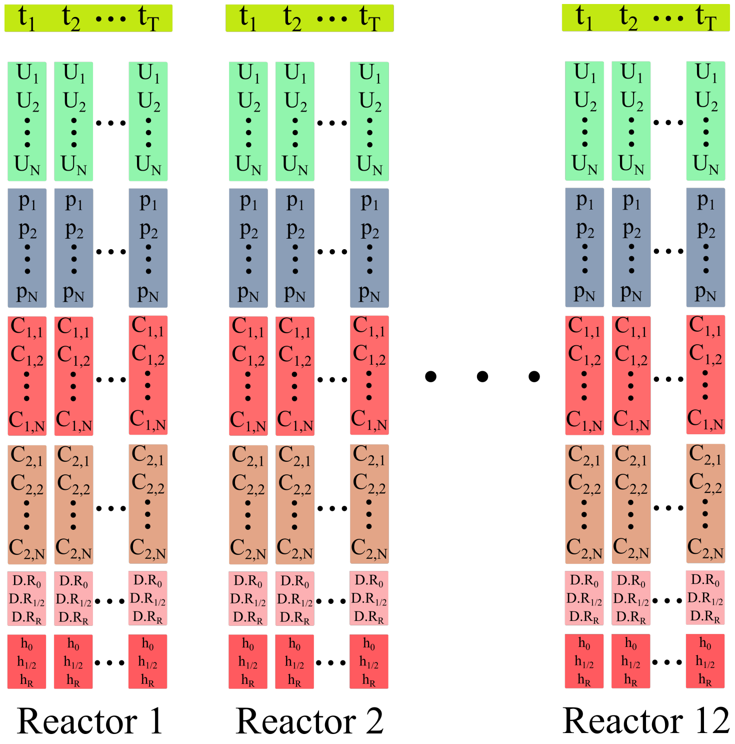


Figure 4: The final matrix considered for the Gappy POD method. A total of 31 time-instances for 12 different reactor geometries have been sampled. These contain all 4 quantities of interest (velocity magnitude, pressure, precursor concentration ($C_1$), water concentration ($C_2$) along with the calculated deposition rates (D.R) and the coating thickness measurements (h) taken from the production data. In our case, T = 31 (number of time-instances per reactor) and N = 1750 (total number of points: 7 lines containing 250 points each).

### 4.3. Performance metrics

The performance of the Gappy POD approach will be evaluated using the Root Mean Squared Error (RMSE) between: a) the Gappy POD reconstruction and the POD reconstruction, b) the Gappy POD reconstruction and the snapshots of the reactor given by the CFD model. The RMSE between two values ($\hat{y}_i$ and $y_i$) for $N$ observations can be written as follows:

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2} \tag{14}$$

### 4.4. Mask selection

The effectiveness of Gappy POD depends on the condition number of matrix $M$, which is defined in Eq. (13). The matrix $M$ is created from the inner products of the "gappy" POD vectors, which are the elements of the original POD vectors corresponding to the known elements of $X'$. Since these vectors are no longer orthogonal, the matrix $M$ is fully populated. For orthogonality to be preserved, the known element positions and non-zero elements of $M$ must be appropriately arranged. Additionally, the diagonal entries of $M$ must not be too small, indicating that the POD basis element at that point should not be small. The condition number of the matrix $M$ reflects these requirements, with a smaller condition number indicating greater satisfaction of these conditions. This analysis is detailed in (Willcox, 2006), in the context of optimal sensor placement, and in (Alonso et al., 2004a,b), which consider the angle between the measurement subspace and the low dimensional space that spans the data.

To determine the known values of the vector $X'$ in a more systematic manner, a greedy algorithm similar to the one proposed by Willcox (2006) is implemented. However, in our case, the mask elements are selected in a way that reduces the reconstruction error. Considering $m$ known values of each snapshot $X'$, then the greedy algorithm implemented works as follows:

1. Initialize by randomly selecting $m$ known values.
2. Starting with the first mask element, loop through all the possible positions for the known values and calculate the reconstruction error for each resulting mask.
3. Find the position of the element that minimizes the reconstruction error and place the first element there.
4. Repeat steps 2-3 for all remaining mask elements.

This way, we can efficiently find positions for the mask elements that yield an acceptable reconstruction error. It should be noted, however, that this does not always lead to the globally optimal positions.

## 5. Results

### *5.1. CFD model*

#### *5.1.1. CFD model parameters*

To elaborate on the model summary made in Section 3.2, further information regarding the CFD model parameters is given in this section.

The prescribed inlet boundary conditions are inlet velocity conditions. For each disk, the gas feed velocity is a time-dependent pulse function that mirrors the inlet tube rotation, varying between 0 and $V_{\max}$. There is a phase difference between the pulses of each disk. $V_{\max}$ and the aforementioned phase difference are determined based on the experimental conditions and geometry, taking into account: a) the 2 RPM rotational speed of the inlet tube, b) the total inlet gas flow rate, c) the number of disks per run, d) the two antipodal perforations per disk, e) the diameter of the perforations (0.002 m), and f) the 60° angle difference between the perforations of each disk.

Outlet pressure boundary conditions are applied at every other disk level. This way, we account for the real geometry where the outlet perforations are not aligned. This results in a model where only the first, the third, the fifth, and the seventh outlet from the top are considered open.

Seven different chemical species are considered, along with a simplified reaction scheme for the deposition of $\alpha$-$Al_2O_3$. The molar fractions at the inlet are the following: $CO_2$ (0.0385), $AlCl_3$ (0.0169), HCl (0.0210), $H_2O$ ($10^{-6}$), CO ($10^{-6}$), $H_2$ (0.9203), and $H_2S$ (0.0033).

The process conditions for the alumina coating step are $T$=1005°C and $p$=80 mbar, as indicated in (Hochauer et al., 2012). Further information can be found in the recent work of Papavasileiou et al. (2022).

### *5.1.2. CFD model predictions*

The CFD model has been tested for 4 different 7-disk reactor geometries. All four 7-disk geometries are building blocks of the test case reactor, whose 2D representation is shown in Fig. 2. It is possible to predict the $\alpha$-$Al_2O_3$ coating thickness with a maximum relative error of 8% and within 5% mean absolute percentage error for each 7-disk geometry, when compared to the available production data. The maximum observed mean absolute percentage error for the $\alpha$-$Al_2O_3$ coating thickness is 4.33%. Simulations for each geometry consist of about $10^6$ degrees of freedom. The solution time for each geometry is approximately 3 core hours on an $11^{\text{th}}$ Gen Intel(R) Core(TM) i7-1185G7 processor. The results of the CFD simulations are summarized in Fig. 5.

### *5.2. Data-driven predictions*

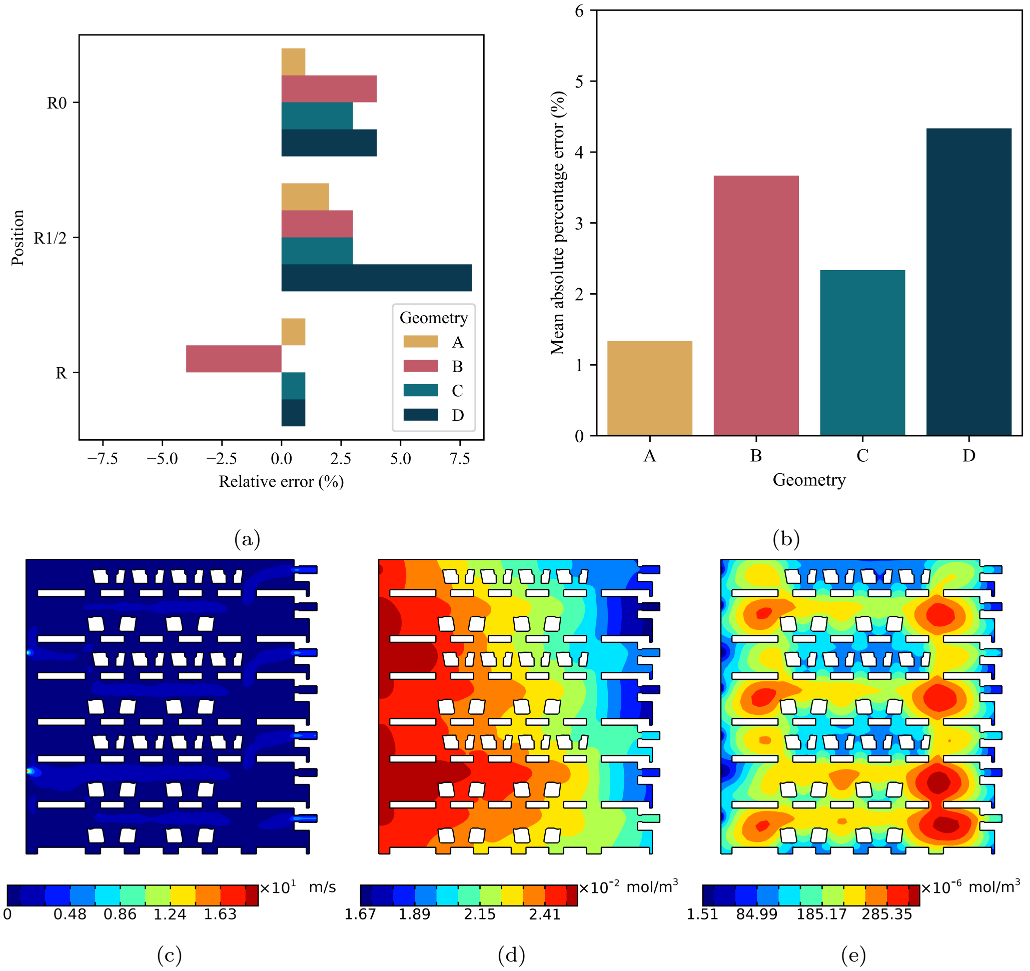
We implement the following tree-based methods: a) Regression Trees, b) Random Forests, c) Gradient Boosting Regression Trees (GBRT) and eXtreme Gradient Boosting Regression Trees (XGBoost). All the methods have comparable performance. Among them, the best performing is XGBoost and the results below focus on its predictions.

The dataset contains a total of 6114 observations and is split into a training set and a test set, using a ratio of 75/25. Each one of these observations contain thickness measurements at the $R$ position for a particular disk (cf. Fig. 2), corresponding to a number of inputs, detailed in Section 2.1. The numerical features were standardized, and the categorical features were encoded using binary encoding.

#### *5.2.1. Hyperparameter selection*

Optimal model performance, is influenced by the choice of hyperparameters for each method. The most important hyperparameters of the implemented tree-based ensemble methods are:

1. The maximum depth of the trees ($d_{\max}$), i.e. the number of bifurcations of the main "branch" of the tree. Selecting too large a tree depth can lead to overfitting, which in essence means that the model fails to generalize accurately.



(a)

(b)

(c)

(d)

(e)

Figure 5: (a) Relative error for the CFD predictions for 3 different positions with available production data inside the reactor. Simulations are performed for four different 7-disk geometries in total. (b) Mean absolute percentage error (averaged over the 3 positions for which data are available) for the CFD simulations for the 4 different reactor geometries. (c) Velocity magnitude, (d) Precursor Concentration and (e) Water Concentration inside the reactor at a certain time during the deposition.

2. The number of trees ($B$). A large number of trees reduces the variance of bagging methods, however it can lead to overfitting in the case of

boosting methods.

3. For boosting methods specifically, another important hyperparameter is the learning rate ($\lambda$). The choice of $\lambda$ usually affects the optimal $B$. For example, a very small $\lambda$ usually requires a large $B$ to achieve satisfactory performance.

Searching for the optimal model hyperparameters in an exhaustive manner is a computationally expensive task. The time required for all 5 tree-based methods using an exhaustive grid search approach performing 10-fold cross-validation was 43 core hours on an 11$^{\text{th}}$ Gen Intel(R) Core(TM) i7-1185G.

To demonstrate here the effect of $d_{\max}$, results are shown for fixed values of $B$ and $\lambda$ (cf. Fig. 6). For a constant number of trees ($B = 10000$), boosting methods show better performance for low values of $d_{\max}$. On the contrary, bagging methods indicate better performance for higher values of $d_{\max}$.

Overall, for all the hyperparameters tested, boosting methods appear to outperform their bagging counterparts. Out of the two boosting methods, the XGBoost method displays higher training and predicting speed. Specifically, for the same training set and the same hyperparameters ($B = 10000$, $d_{\max} = 5$ and $\lambda = 0.01$), the average training time over 10 cross-validation splits is 16.5s for the XGBoost model and 99.5s for the GBRT model. Moreover, the average prediction time is 20ms for the XGBoost model and 333ms for the GBRT model. Therefore, due to its lower computational cost, further hyperparameter tuning will take place for the XGBoost algorithm, in order to find the optimal hyperparameter combination.

After selecting the optimal value of maximum depth, we further inves-

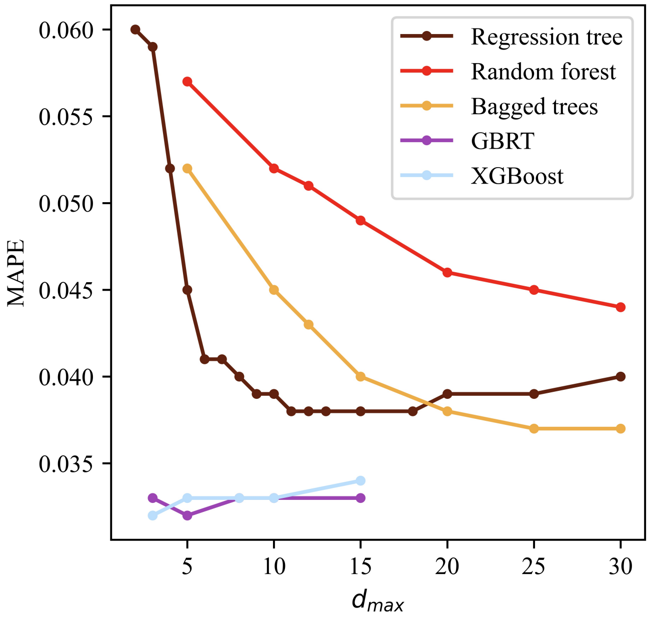


Figure 6: MAPE vs $d_{\max}$ for all methods after 10-fold cross-validation. $B = 10000$ for all ensemble methods. $\lambda = 0.01$ for the boosting methods. For the base method (regression tree) and bagging methods (Bagged Trees and Random Forests) increasing the maximum depth of the trees leads to a reduced MAPE. For the boosting methods (GBRT and XGBoost), the MAPE increases when increasing the maximum depth of the trees. Random forest regression performing worse than the simple regression tree can be attributed to the fact that it only considers a subset of available features when building each tree of the ensemble.

tigate the effect of the number of trees $B$ on the accuracy of the XGBoost model. As indicated in Table 2, the accuracy of the model drastically improves when $B \geq 500$, nevertheless, the trade-off is in the form of increased computational cost.

Following hyperparameter optimization and tuning, the final values selected for the XGBoost model are the following: $d_{\max} = 5$, $B = 10000$,

Table 2: XGBoost model results after cross-validation for various values of $B$, where $d_{\text{max}} = 5$ and $\lambda = 0.01$. As expected, an increased number of base predictors improves the performance of the ensemble boosting method. However, it also increases the training time and prediction time of the model. All metrics are averaged over 10 cross-validation splits.

| Number of trees ($B$) | MAPE | $\bar{t}_{\text{train}}$ (s) | $\bar{t}_{\text{pred}}$ (ms) |
|---|---|---|---|
| 10000 | 3.1% | 16.3 | 20 |
| 5000 | 3.3% | 8.0 | 14 |
| 2000 | 3.4% | 3.3 | 9 |
| 1000 | 3.6% | 1.7 | 9 |
| 500 | 3.9% | 0.9 | 8 |
| 200 | 12.6% | 0.4 | 8 |
| 100 | 33.8% | 0.2 | 10 |

$\lambda = 0.01$.

### 5.2.2. Machine learning outcomes

Two more accuracy metrics are introduced here, the mean square error (MSE) and the coefficient of determination (R$^2$). When the model is trained or tested on $N$ observations and for each observation $i$ the predicted value is $\hat{y}_i$ while the actual value is $y_i$ and the average of the actual values is $\bar{y}$, MSE and R$^2$ can be written as follows:

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2 \tag{15}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(\hat{y}_i - y_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2} \tag{16}$$

The prediction error of XGBoost regression model for the training set, reaches a MAPE of 0.9%, versus 3.1% for the test set. The prediction accuracy of the XGBoost model on the training set and on the test set can be summarized in Figs. 7a and 7b respectively. Due to the confidentiality of the production data, absolute $\alpha$-$Al_2O_3$ thickness values cannot be presented. Therefore, only relative error values and normalized thickness values are presented.

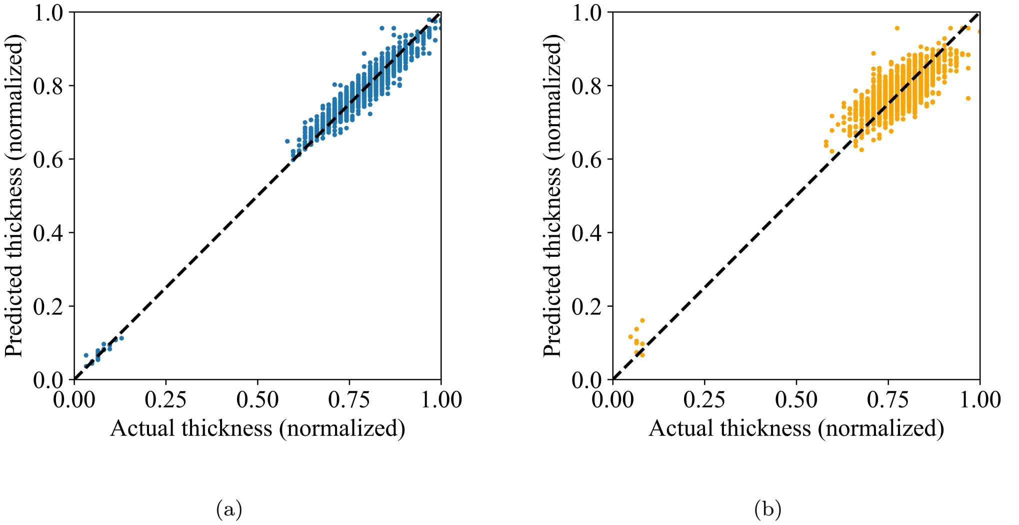


Figure 7: (a) Training set performance: MSE:0.005 | MAE:0.051 | MAPE:0.9% | $R^2$:0.980. (b) Test set performance: MSE:0.059 | MAE:0.187 | MAPE:3.1% | $R^2$:0.753.

### *5.3. CFD vs ML*

#### *5.3.1. Predictive accuracy*

For the test-case reactor set-up presented in Fig. 2, the prediction results for the position closest to the outlet for both methods are given in Table 3. Disk position is counted from the bottom to the top of the reactor.

Table 3: XGBoost prediction accuracy vs CFD prediction accuracy for the coating thickness of inserts closest to the reactor outlet ($R$ position). Errors relative to the available production data are presented. The high error in the prediction of the CFD model for the 6th reactor disk can be attributed to the fact that it is the bottom-most disk of the simulated 7-disk geometry, and therefore the effect of the inlets and outlets that are below it is not taken into account.

| Disk position | CFD prediction | XGBoost prediction |
|---|---|---|
| 39 | 3.2% | 3.5% |
| 35 | 1.0% | -3.1% |
| 23 | -4.0% | -7.0% |
| 10 | 1.0% | -5.5% |
| 6 | 20.6% | -2.8% |
| **MAPE** | 6.0% | 4.4% |
| Total prediction time (s) | 43200 | 0.1 |

Despite the significant difference in the computational effort involved in the CFD model in comparison to the ML regression model, both methods have comparable accuracy on the test-case. CFD predictions for the test reactor have a mean absolute percentage error of 6%, while XGBoost makes predictions with a mean absolute percentage error of 4.4%. The high error in the prediction of the CFD model for the 6th reactor disk (20.6%) can be attributed to the fact that it is the bottom-most disk of the simulated 7-disk geometry and therefore the effect of the inlets and outlets that are below it is not taken into account. This can be solved by an extra 7-disk simulation, where the disk of interest won't be in the bottom-most position. This would of course further increase the computational cost of the CFD approach. The maximum observed absolute relative error for the predictions of the XGBoost model on the test-case reactor is 7%.

### *5.3.2. Computational performance*

Although the predictive accuracy of the two approaches is similar, they demonstrate a very noticeable contrast when it comes to their computational performance. Specifically, in the case of CFD, making predictions for an entire production run would require 4 or 5 7-disk simulations. This corresponds to a computational cost of 12 to 15 core hours. On the other hand, using the XGBoost model to make predictions for an entire production run comes with a computational cost of less than 1 core second. This translates to a reduction of more than 99.99% in required resources.

## *5.4. Gappy POD*

Results of our Gappy POD implementation will be presented for two different cases:

1. The case of the full dataset.
2. The case of a single reactor.

In each case, the dataset consists of time-instances of the state vector, over a period of 30 secs. Therefore, the full dataset eventually consists of 372 snapshots, whereas in the single reactor dataset, it consists of 31 vectors.

In both cases, 87.5% of the available snapshots are used to derive the POD basis of the training set. The rest of the snapshots (12.5%) are kept and used for the validation of the method. For both cases, the data are scaled in the range of $[0, 1]$ using min-max normalization.

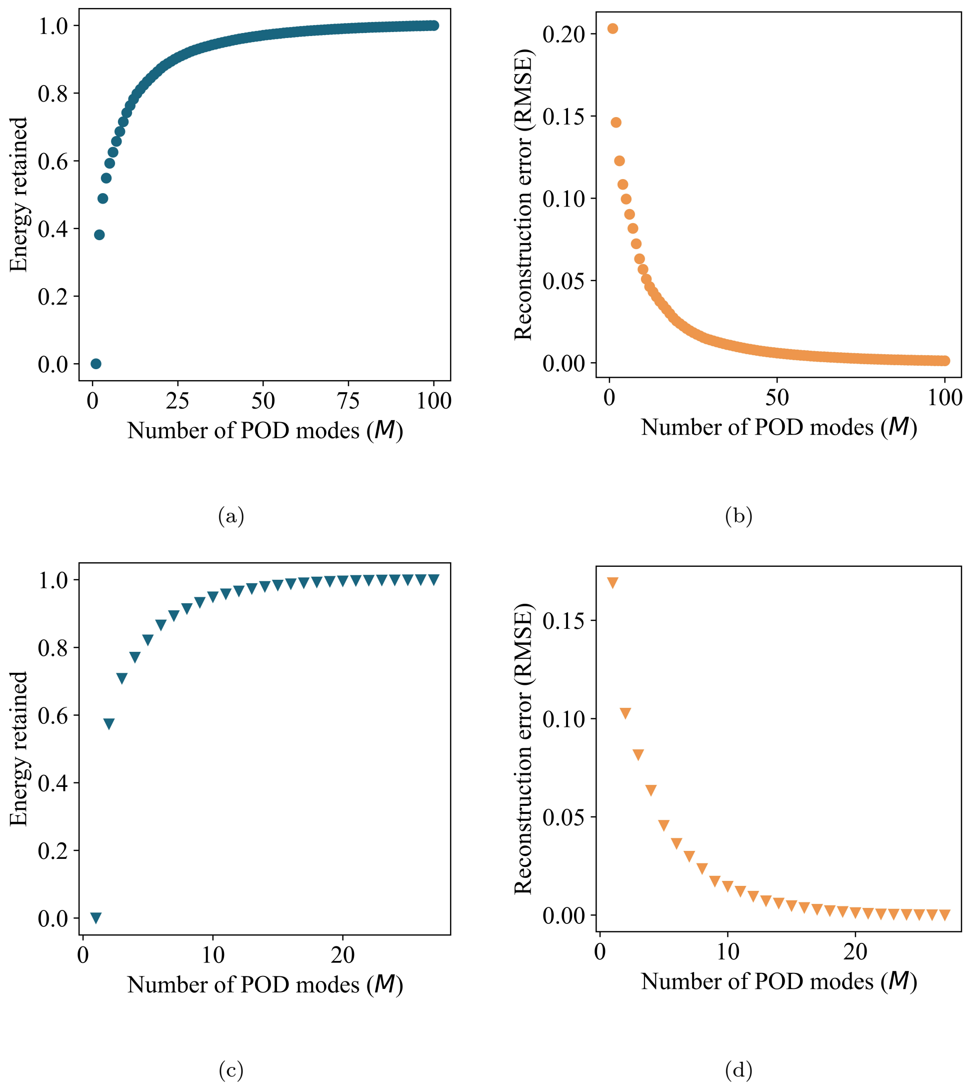
The number of modes used for the POD basis are selected after checking the energy retained by the modes and the resulting reconstruction error. The total retained energy for the full dataset and the single reactor dataset, is shown in Fig. 8a and Fig. 8c respectively, whereas the reconstruction error as a function of the basis size is shown in Fig. 8b and Fig. 8d respectively.

The full reactor dataset requires at least 50 POD modes to capture more than 95% of the energy of the system, with a corresponding reconstruction error (RMSE) of 0.0059. The single reactor dataset, is accurately represented by 15 POD modes that reflect more than 98 % of the energy with a reconstruction error (RMSE) of 0.004. Eventually, for the immediate comparison of the results, the same basis size is considered, equal to 15 POD modes. The corresponding retained energy and error are shown in Table 4.

Table 4: Number of POD modes selected for each case, along with the corresponding retained energy and reconstruction error.

| Case | # POD modes | Energy retained | Recon. error (RMSE) |
|---|---|---|---|
| Full dataset | 15 | 81.69% | 0.0373 |
| Single reactor | 15 | 98.70% | 0.0040 |
| Single reactor | 5 | 82.74% | 0.0456 |

After selecting the size of the POD basis for each case, the mask elements for Gappy POD are obtained using the greedy algorithm described in



(a)

(b)

(c)

(d)

Figure 8: The energy retained (in blue) and the reconstruction error (in orange) of the POD approximation using $M$ modes. (a), (b): Energy and reconstruction error for the full dataset. Only the first 100 modes are shown. (c), (d): Energy and reconstruction error for the single reactor case.

Section 4.4. It should be noted that the mask length should be greater or equal to the size of the POD basis. For all three cases, we allow one mask element more than the size of the POD basis. It should be noted that in all cases the mask elements acquired consist of all the quantities of interest (velocity magnitude, pressure, precursor concentration, water concentration) discussed in Section 4.2.

After acquiring the mask elements, the RMSE between the Gappy POD approximation and the test set, along with the RMSE between the Gappy POD approximation and the POD reconstruction, can be calculated. Specifically, for the case of the full dataset, the RMSE between the Gappy POD approximation and the test set is 0.0648 while the RMSE between the Gappy POD approximation and the POD reconstruction is 0.0512 (cf. Fig. 9). For the case of the single reactor, the RMSE between the Gappy POD approximation and the test set is 0.0099 while the RMSE between the Gappy POD approximation and the POD reconstruction is 0.0064. If we choose to make a comparison using the number of POD modes with the same retained energy and reconstruction error, we choose 5 POD modes (82.74% retained energy and 0.046 reconstruction error) and 6 mask elements for the single reactor case. Then, the RMSE between the Gappy POD approximation and the test set is 0.0474 while the RMSE between the Gappy POD approximation and the POD reconstruction is 0.0143.

The performance of the method, is linked to how well the dataset is spanned by the selected POD vectors, generally implying that a larger POD basis is beneficial for the results. Nevertheless, since the ambition of this approach is to select only a few measurements as mask elements, it is more beneficial to work with the smallest possible number of POD vectors.

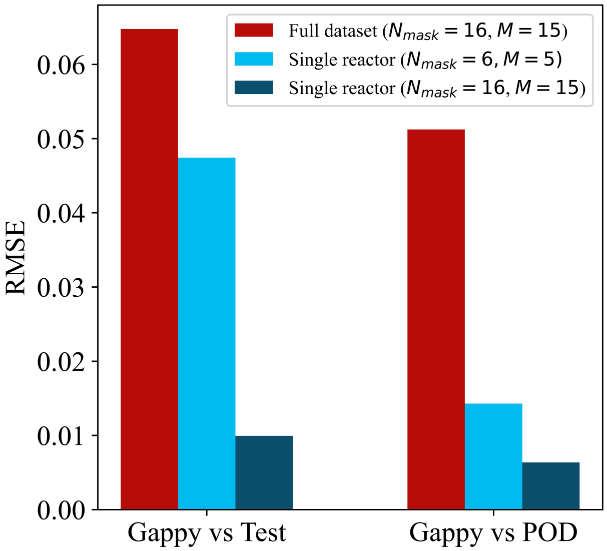


Figure 9: On the left: Error between the Gappy POD approximation and the snapshots of the test set for all cases. On the right: Error between the Gappy POD approximation and the POD approximation for all cases. It is evident that the single reactor case shows the lowest errors. This is probably due to the lower variance observed in the dataset of the single reactor when compared with the full dataset. For the case of the single reactor, using a smaller POD basis (5 modes instead of 15) leads to an increase in both errors.

## 6. Conclusions

This work presents an overview of the implementation of equation-based and machine-learning methods in industrial-scale deposition applications. The challenges associated with the complexity of the process and the characteristics of real production data are discussed and the methods to overcome them are presented.

In the equation-based approach, a reduced model is presented and validated with production measurements of the coating thickness. The simplifications introduced and the pertinent assumptions upon which they are based are discussed, along with the results. The trade-off between the computational cost associated with the CFD model and the physical insight obtained, is discussed and compared to the ML approach. Coating thickness predictions are possible with an average error of 6%. In addition, the CFD model, predicts the distributions of velocity, and reactive species, illuminating thus, the mechanisms that contribute to the final product. Furthermore, it can be used to predict the thickness achieved in parts of the reactor where there are no measurements. Moreover, the CFD approach also allows extrapolating for different process conditions and different inlet reactant concentrations. For the 7-disk CFD approach, the results of Table 3, show that appropriate selection of the 7-disk "building blocks" for the simulations is of high importance for the accuracy of the prediction.

The ML approach is discussed in detail, as far as the possible specific methods are concerned. The suitability of each is assessed, based on the data available. Eventually the best performing ML method, XGBoost, is able to deliver accurate and time-efficient coating thickness predictions, but cannot provide insight into the transport of species that determines the coating thickness.

The implementation of Gappy POD for this specific application, shows how data-driven methods and CFD results can be intertwined to provide further insight on the important quantities of interest inside the reactor. By further analysis of the resulting mask elements, we can explore the hypothetical scenario of sensor placement inside such reactors. Furthermore, we can reconstruct entire snapshots from a few measurements inside the reactor, reducing in this way the computational cost of the problem.

It should be noted that the strategy employed here is not exclusive to CFD modeling. The same workflow could still be implemented in other applications, regardless of the equation-based modeling approach used. The only limiting factor would be the amount and type of available data for the application.

Another important observation is that specific *combinations* of inputs can lead to the same outputs. This merits further investigation, due to its importance in the actual production process, which is the topic of future work.

To conclude, it is clear that each individual approach is a valuable tool in studying a complex process offering different advantages: physical insight and extrapolation abilities in CFD and time-efficient, accurate predictions in ML. It is therefore worth investing the effort in each one of them, and ultimately, in merging them in a hybrid approach with additional benefits. Ideally, the resulting model could combine high accuracy, time-efficient predictions, and excellent extrapolation ability, moving in this way toward a digital twin of the process.

## Acknowledgements

P.P. and E.D.K. acknowledge financial support by the Fonds National de la Recherche (FNR) Luxembourg (BRIDGE grant HybridSimCVD). E.D.K. is supported by the EU under a MSCA Individual Fellowship. S.P.A.B received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 811099 TWINNING Project DRIVEN for the University of Luxembourg.

## References

Agarwal, P., Tamer, M., Sahraei, M.H., Budman, H., 2020. Deep Learning for Classification of Profit-Based Operating Regions in Industrial Processes. Ind. Eng. Chem. Res. 59, 2378–2395. doi:10.1021/acs.iecr.9b04737.

Alonso, A.A., Frouzakis, C.E., Kevrekidis, I.G., 2004a. Optimal sensor placement for state reconstruction of distributed process systems. AIChE Journal 50, 1438–1452. doi:10.1002/aic.10121.

Alonso, A.A., Kevrekidis, I.G., Banga, J.R., Frouzakis, C.E., 2004b. Optimal sensor location and reduced order observer design for distributed process systems. Computers & Chemical Engineering 28, 27–35. doi:10.1016/S0098-1354(03)00175-3.

Aviziotis, I.G., Duguet, T., Vahlas, C., Boudouvis, A.G., 2017. Combined Macro/Nanoscale Investigation of the Chemical Vapor Deposition of Fe from Fe(CO)5. Advanced Materials Interfaces 4, 1601185. doi:10.1002/admi.201601185.

Azadi, P., Winz, J., Leo, E., Klock, R., Engell, S., 2022. A hybrid dynamic model for the prediction of molten iron and slag quality indices of a large-scale blast furnace. Computers & Chemical Engineering 156, 107573. doi:10.1016/j.compchemeng.2021.107573.

Bar-Hen, M., Etsion, I., 2017. Experimental study of the effect of coating thickness and substrate roughness on tool wear during turning. Tribology International 110, 341–347. doi:10.1016/j.triboint.2016.11.011.

Blakseth, S.S., Rasheed, A., Kvamsdal, T., San, O., 2022. Deep neural network enabled corrective source term approach to hybrid analysis and modeling. Neural Networks 146, 181–199. doi:10.1016/j.neunet.2021.11.021.

Boyes, H., Watson, T., 2022. Digital twins: An analysis framework and open issues. Computers in Industry 143, 103763. doi:10.1016/j.compind.2022.103763.

Cai, H., Feng, J., Yang, Q., Li, W., Li, X., Lee, J., 2020. A virtual metrology method with prediction uncertainty based on Gaussian process for chemical mechanical planarization. Computers in Industry 119, 103228. doi:10.1016/j.compind.2020.103228.

Carvajal Soto, J.A., Tavakolizadeh, F., Gyulai, D., 2019. An online machine learning framework for early detection of product failures in an Industry 4.0 context. International Journal of Computer Integrated Manufacturing 32, 452–465. doi:10.1080/0951192x.2019.1571238.

Cheimarios, N., Koronaki, E.D., Boudouvis, A.G., 2012. Illuminating nonlinear dependence of film deposition rate in a CVD reactor on operating conditions. Chemical Engineering Journal 181–182, 516–523. doi:10.1016/j.cej.2011.11.008.

Cho, J., Mountziaris, T.J., 2013. Onset of flow recirculation in vertical rotating-disc chemical vapor deposition reactors. AIChE Journal 59, 3530–3538. doi:10.1002/aic.14179.

Creighton, J.R., Parmeter, J.E., 1993. Metal CVD for microelectronic applications: An examination of surface chemistry and kinetics. Critical Reviews in Solid State and Materials Sciences 18, 175–237. doi:10.1080/10408439308242560.

Dai, W., Mohammadi, S., Cremaschi, S., 2022. A hybrid modeling framework using dimensional analysis for erosion predictions. Computers & Chemical Engineering 156, 107577. doi:10.1016/j.compchemeng.2021.107577.

Dalzochio, J., Kunst, R., Pignaton, E., Binotto, A., Sanyal, S., Favilla, J., Barbosa, J., 2020. Machine learning and reasoning for predictive maintenance in Industry 4.0: Current status and challenges. Computers in Industry 123, 103298. doi:10.1016/j.compind.2020.103298.

Deng, J., Sierla, S., Sun, J., Vyatkin, V., 2022. Reinforcement learning for industrial process control: A case study in flatness control in steel industry. Computers in Industry 143, 103748. doi:10.1016/j.compind.2022.103748.

Deshpande, S., Lengiewicz, J., Bordas, S.P.A., 2022. Probabilistic deep learning for real-time large deformation simulations. Computer Methods in Applied Mechanics and Engineering 398, 115307. doi:10.1016/j.cma.2022.115307.

Du, H., Jiang, Y., 2019. Backup or Reliability Improvement Strategy for a Manufacturer Facing Heterogeneous Consumers in a Dynamic Supply Chain. IEEE Access 7, 50419–50430. doi:10.1109/access.2019.2911620.

Endo, H., Kuwana, K., Saito, K., Qian, D., Andrews, R., Grulke, E.A., 2004. CFD prediction of carbon nanotube production rate in a CVD reactor. Chemical Physics Letters 387, 307–311. doi:10.1016/j.cplett.2004.01.124.

Everson, R., Sirovich, L., 1995. Karhunen–Loève procedure for gappy data. J. Opt. Soc. Am. A 12, 1657. doi:10.1364/JOSAA.12.001657.

Fotiadis, D.I., Boekholt, M., Jensen, K.F., Richter, W., 1990. Flow and heat transfer in CVD reactors: Comparison of Raman temperature measurements and finite element model predictions. Journal of Crystal Growth 100, 577–599. doi:10.1016/0022-0248(90)90257-L.

Gakis, G., Koronaki, E., Boudouvis, A., 2015. Numerical investigation of multiple stationary and time-periodic flow regimes in vertical rotating disc CVD reactors. Journal of Crystal Growth 432, 152–159. doi:10.1016/j.jcrysgro.2015.09.026.

Galvis, L., Offermans, T., Bertinetto, C.G., Carnoli, A., Karamujić, E., Li, W., Szymańska, E., Buydens, L.M.C., Jansen, J.J., 2022. Retrospective quality by design r(QbD) for lactose production using historical process data and design of experiments. Computers in Industry 141, 103696. doi:10.1016/j.compind.2022.103696.

Gkinis, P., Aviziotis, I., Koronaki, E., Gakis, G., Boudouvis, A., 2017. The effects of flow multiplicity on GaN deposition in a rotating disk CVD reactor. Journal of Crystal Growth 458, 140–148. doi:10.1016/j.jcrysgro.2016.10.065.

Gkinis, P., Koronaki, E., Skouteris, A., Aviziotis, I., Boudouvis, A., 2019. Building a data-driven reduced order model of a chemical vapor deposition process from low-fidelity CFD simulations. Chemical Engineering Science 199, 371–380. doi:10.1016/j.ces.2019.01.009.

Hastie, T., Tibshirani, R., Friedman, J., 2009a. Additive Models, Trees, and Related Methods, in: Hastie, T., Tibshirani, R., Friedman, J. (Eds.), The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer, New York, NY. Springer Series in Statistics, pp. 295–336. doi:10.1007/978-0-387-84858-7_9.

Hastie, T., Tibshirani, R., Friedman, J., 2009b. Boosting and Additive Trees, in: Hastie, T., Tibshirani, R., Friedman, J. (Eds.), The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer, New York, NY. Springer Series in Statistics, pp. 337–387. doi:10.1007/978-0-387-84858-7_10.

He, Z., Tran, K.P., Thomassey, S., Zeng, X., Xu, J., Yi, C., 2021. A deep reinforcement learning based multi-criteria decision support system for optimizing textile chemical process. Computers in Industry 125, 103373. doi:10.1016/j.compind.2020.103373.

Hochauer, D., Mitterer, C., Penoy, M., Puchner, S., Michotte, C., Martinz, H., Hutter, H., Kathrein, M., 2012. Carbon doped $\alpha$-Al2O3 coatings grown by chemical vapor deposition. Surface and Coatings Technology 206, 4771–4777. doi:10.1016/j.surfcoat.2012.03.059.

Hürkamp, A., Gellrich, S., Ossowski, T., Beuscher, J., Thiede, S., Herrmann, C., Dröder, K., 2020. Combining Simulation and Machine Learning as Digital Twin for the Manufacturing of Overmolded Thermoplastic Composites. JMMP 4, 92. doi:10.3390/jmmp4030092.

Iqbal, R., Maniak, T., Doctor, F., Karyotis, C., 2019. Fault Detection and Isolation in Industrial Processes Using Deep Learning Approaches. IEEE Trans. Ind. Inf. 15, 3077–3084. doi:10.1109/tii.2019.2902274.

James, G., Witten, D., Hastie, T., Tibshirani, R., 2021a. Statistical Learning. Springer US, New York, NY. pp. 15–57. doi:10.1007/978-1-0716-1418-1_2.

James, G., Witten, D., Hastie, T., Tibshirani, R., 2021b. Tree-Based Methods. Springer US, New York, NY. pp. 327–365. doi:10.1007/978-1-0716-1418-1_8.

Jo, T., Koo, B., Kim, H., Lee, D., Yoon, J.Y., 2019. Effective sensor placement in a steam reformer using gappy proper orthogonal decomposition. Applied Thermal Engineering 154, 419–432. doi:10.1016/j.applthermaleng.2019.03.089.

Kagermann, H., 2015. Change Through Digitization—Value Creation in the Age of Industry 4.0, in: Albach, H., Meffert, H., Pinkwart, A., Reichwald,

R. (Eds.), Management of Permanent Change. Springer Fachmedien, Wiesbaden, pp. 23–45. doi:10.1007/978-3-658-05014-6_2.

Kalaboukas, K., Kiritsis, D., Arampatzis, G., 2023. Governance framework for autonomous and cognitive digital twins in agile supply chains. Computers in Industry 146, 103857. doi:10.1016/j.compind.2023.103857.

Kathrein, M., Schintlmeister, W., Wallgram, W., Schleinkofer, U., 2003. Doped CVD Al2O3 coatings for high performance cutting tools. Surface and Coatings Technology 163–164, 181–188. doi:10.1016/s0257-8972(02)00483-8.

Kerfriden, P., Goury, O., Rabczuk, T., Bordas, S.P.A., 2013. A partitioned model order reduction approach to rationalise computational expenses in nonlinear fracture mechanics. Computer Methods in Applied Mechanics and Engineering 256, 169–188. doi:10.1016/j.cma.2012.12.004.

Kim, A., Oh, K., Jung, J.Y., Kim, B., 2018. Imbalanced classification of manufacturing quality conditions using cost-sensitive decision tree ensembles. International Journal of Computer Integrated Manufacturing 31, 701–717. doi:10.1080/0951192x.2017.1407447.

Kim, D., Kang, P., Cho, S., Lee, H.j., Doh, S., 2012. Machine learning-based novelty detection for faulty wafer detection in semiconductor manufacturing. Expert Systems with Applications 39, 4075–4083. doi:10.1016/j.eswa.2011.09.088.

Kleijn, C.R., Dorsman, R., Kuijlaars, K.J., Okkerse, M., van Santen, H., 2007. Multi-scale modeling of chemical vapor deposition processes for thin film technology. Journal of Crystal Growth 303, 362–380. doi:10.1016/j.jcrysgro.2006.12.062.

Kleijn, C.R., Hoogendoorn, C.J., 1991. A study of 2- and 3-D transport phenomena in horizontal chemical vapor deposition reactors. Chemical Engineering Science 46, 321–334. doi:10.1016/0009-2509(91)80141-K.

Koronaki, E., Gkinis, P., Beex, L., Bordas, S., Theodoropoulos, C., Boudouvis, A., 2019. Classification of states and model order reduction of large scale Chemical Vapor Deposition processes with solution multiplicity. Computers & Chemical Engineering 121, 148–157. doi:10.1016/j.compchemeng.2018.08.023.

Koronaki, E.D., Evangelou, N., Psarellis, Y.M., Boudouvis, A.G., Kevrekidis, I.G., 2023. From partial data to out-of-sample parameter and observation estimation with Diffusion Maps and Geometric Harmonics. doi:10.48550/arXiv.2301.11728, arXiv:arXiv:2301.11728.

Koronaki, E.D., Gakis, G.P., Cheimarios, N., Boudouvis, A.G., 2016. Efficient tracing and stability analysis of multiple stationary and periodic states with exploitation of commercial CFD software. Chemical Engineering Science 150, 26–34. doi:10.1016/j.ces.2016.04.043.

Łępicka, M., Grądzka-Dahlke, M., 2019. The initial evaluation of performance of hard anti-wear coatings deposited on metallic substrates: Thickness, mechanical properties and adhesion measurements – a brief review. REVIEWS ON ADVANCED MATERIALS SCIENCE 58, 50–65. doi:10.1515/rams-2019-0003.

Liu, S.S., Xiao, W.D., 2015. CFD–PBM coupled simulation of silicon CVD growth in a fluidized bed reactor: Effect of silane pyrolysis kinetic models. Chemical Engineering Science 127, 84–94. doi:10.1016/j.ces.2015.01.026.

Ma, Y., Zhu, W., Benton, M.G., Romagnoli, J., 2019. Continuous control of a polymerization system with deep reinforcement learning. Journal of Process Control 75, 40–47. doi:10.1016/j.jprocont.2018.11.004.

Malley, S., Reina, C., Nacy, S., Gilles, J., Koohbor, B., Youssef, G., 2022. Predictability of mechanical behavior of additively manufactured particulate composites using machine learning and data-driven approaches. Computers in Industry 142, 103739. doi:10.1016/j.compind.2022.103739.

McKinney, W., 2010. Data Structures for Statistical Computing in Python, in: Python in Science Conference, Austin, Texas. pp. 56–61. doi:10.25080/Majora-92bf1922-00a.

Ozaydin-Ince, G., Coclite, A.M., Gleason, K.K., 2011. CVD of polymeric thin films: Applications in sensors, biotechnology, microelectronics/organic electronics, microfluidics, MEMS, composites and membranes. Rep. Prog. Phys. 75, 016501. doi:10.1088/0034-4885/75/1/016501.

Papavasileiou, P., Koronaki, E.D., Pozzetti, G., Kathrein, M., Czettl, C., Boudouvis, A.G., Mountziaris, T.J., Bordas, S.P.A., 2022. An efficient chemistry-enhanced CFD model for the investigation of the rate-limiting mechanisms in industrial Chemical Vapor Deposition reactors. Chemical

Engineering Research and Design 186, 314–325. doi:10.1016/j.cherd.2022.08.005.

Perno, M., Hvam, L., Haug, A., 2022. Implementation of digital twins in the process industry: A systematic literature review of enablers and barriers. Computers in Industry 134, 103558. doi:10.1016/j.compind.2021.103558.

Potdar, K., S., T., D., C., 2017. A Comparative Study of Categorical Variable Encoding Techniques for Neural Network Classifiers. IJCA 175, 7–9. doi:10.5120/ijca2017915495.

Priore, P., Ponte, B., Puente, J., Gómez, A., 2018. Learning-based scheduling of flexible manufacturing systems using ensemble methods. Computers & Industrial Engineering 126, 282–291. doi:10.1016/j.cie.2018.09.034.

Psarellis, G.M., Aviziotis, I.G., Duguet, T., Vahlas, C., Koronaki, E.D., Boudouvis, A.G., 2018. Investigation of reaction mechanisms in the chemical vapor deposition of al from DMEAA. Chemical Engineering Science 177, 464–470. doi:10.1016/j.ces.2017.12.006.

Raissi, M., Perdikaris, P., Karniadakis, G., 2019. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. Journal of Computational Physics 378, 686–707. doi:10.1016/j.jcp.2018.10.045.

Rasheed, A., San, O., Kvamsdal, T., 2020. Digital Twin: Values, Challenges and Enablers From a Modeling Perspective. IEEE Access 8, 21980–22012. doi:10.1109/access.2020.2970143.

Saxena, A., Saad, A., 2007. Evolving an artificial neural network classifier for condition monitoring of rotating mechanical systems. Applied Soft Computing 7, 441–454. doi:10.1016/j.asoc.2005.10.001.

Schierling, M., Zimmermann, E., Neuschütz, D., 1999. Deposition kinetics of $Al_2O_3$ from $AlCl_3$-$CO_2$-$H_2$-HCl gas mixtures by thermal CVD in a hot-wall reactor. J. Phys. IV France 09, Pr8–85–Pr8–91. doi:10.1051/jp4:1999811.

Spencer, R., Gkinis, P., Koronaki, E., Gerogiorgis, D., Bordas, S., Boudouvis, A., 2021. Investigation of the chemical vapor deposition of Cu from copper amidinate through data driven efficient CFD modelling. Computers & Chemical Engineering 149, 107289. doi:10.1016/j.compchemeng.2021.107289.

Susto, G.A., Schirru, A., Pampuri, S., McLoone, S., Beghi, A., 2015. Machine Learning for Predictive Maintenance: A Multiple Classifier Approach. IEEE Trans. Ind. Inf. 11, 812–820. doi:10.1109/tii.2014.2349359.

Topka, K.C., Vergnes, H., Tsiros, T., Papavasileiou, P., Decosterd, L., Diallo, B., Senocq, F., Samelor, D., Pellerin, N., Menu, M.J., Vahlas, C., Caussat, B., 2022. An innovative kinetic model allowing insight in the moderate temperature chemical vapor deposition of silicon oxynitride films from tris(dimethylsilyl)amine. Chemical Engineering Journal 431, 133350. doi:10.1016/j.cej.2021.133350.

Tulsyan, A., Garvin, C., Ündey, C., 2018. Advances in industrial biopharmaceutical batch process monitoring: Machine-learning methods for small data problems. Biotechnology and Bioengineering 115, 1915–1924. doi:10.1002/bit.26605.

Urcun, S., Rohan, P.Y., Skalli, W., Nassoy, P., Bordas, S.P.A., Sciumè, G., 2021. Digital twinning of Cellular Capsule Technology: Emerging outcomes from the perspective of porous media mechanics. PLOS ONE 16, e0254512. doi:10.1371/journal.pone.0254512.

Wang, R., Cheung, C.F., Wang, C., Cheng, M.N., 2022. Deep learning characterization of surface defects in the selective laser melting process. Computers in Industry 140, 103662. doi:10.1016/j.compind.2022.103662.

Willcox, K., 2006. Unsteady flow sensing and estimation via the gappy proper orthogonal decomposition. Computers & Fluids 35, 208–226. doi:10.1016/j.compfluid.2004.11.006.

Wu, H., Yu, Z., Wang, Y., 2019. Experimental study of the process failure diagnosis in additive manufacturing based on acoustic emission. Measurement 136, 445–453. doi:10.1016/j.measurement.2018.12.067.