Computers and Chemical Engineering 174 (2023) 108252

Contents lists available at ScienceDirect

Computers and Chemical Engineering

journal homepage: www.elsevier.com/locate/cace

Development of a surrogate model of an amine scrubbing digital twin using machine learning methods

Andrea Galeazzi [a], Kristiano Prifti [a], Carlo Cortellini [a], Alessandro Di Pretoro [b], Francesco Gallo [c], Flavio Manenti [a,*]

[a] Dipartimento di Chimica, Materiali e Ingegneria Chimica "Giulio Natta", Politecnico di Milano, Piazza Leonardo Da Vinci, 32, Milan, 20133, Italy
[b] Laboratoire de Génie Chimique, Université de Toulouse, CNRS/INP/UPS, Allée E. Monso, 4, Toulouse, 31431, France
[c] Itelyum Regeneration S.p.A., Via Tavernelle, 19, Pieve Fissiraga, 26854, Italy

ARTICLE INFO

Keywords:
Machine-learning
Surrogate modeling
Digital twin
Amine scrubbing
Design of experiments
Latin hypercube

ABSTRACT

Advancements in the process industry require building more complex simulations and performing computationally intensive operations like optimization. To overcome the numerical limit of conventional process simulations a surrogate model is a viable strategy. In this work, a surrogate model of an industrial amine scrubbing digital twin has been developed. The surrogate model has been built based on the process simulation created in Aspen HYSYS and validated as a digital twin against real process data collected during a steady-state operation. The surrogate relies on an accurate Design of Experiments procedure. In this case, the Latin-Hypercube method has been chosen and several nested domains have been defined in ranges around the nominal steady state operative condition. Several machine learning models have been trained using cross-validation, and the most accurate has been selected to predict each target. The resulting surrogate model showed a satisfactory performance, given the data available.

1. Introduction

The role of the digital twin will be central in revolutionizing the chemical engineering industry in the near future (Liu et al., 2021; VanDerHorn and Mahadevan, 2021). More and more industrial applications start from the development of a digital twin of the process (Kritzinger et al., 2018). The role of the digital twin is becoming critical to generate competitive solutions in a wide variety of use cases, such as process operation or maintenance (Errandonea et al., 2020), risk control and prevention (Bevilacqua et al., 2020), process design (Damiani et al., 2018), smart manufacturing (Hu et al., 2018), process optimization (Jeon and Schuesslbauer, 2020), asset lifecycle management (Macchi et al., 2018), process monitoring (Zipper et al., 2018), decision making support (Zhou et al., 2021).

VanDerHorn and Mahadevan (2021) defined the digital twin as a virtual representation of a physical system that is updated through information exchange between the physical and virtual systems. Based on this definition, the digital twin can exist only when a physical asset is present, thus a process simulation model may not be technically defined as a digital twin unless the process is actually built and the simulation is updated accordingly. However, in the design phase of engineering a new process, even though it should not be called a rigorous digital twin, a reliable process simulation can be used for the same kind of applications. Moreover, designing an optimal process configuration is of utmost importance, and the usage of *non-rigorous* digital twins helps in achieving it (Tian et al., 2018). In any case, the process simulation, also called *non-rigorous digital twin*, might be the first building block for constructing a rigorous digital twin, dynamically integrated with the physical system.

Process simulations and digital twins may be based on fundamental models which can be computationally demanding, thus impeding real-time or highly iterative applications, e.g. optimization (Zhao et al., 2022), model predictive control (Kannapinn et al., 2022), etc. The alternative to fundamental models is offered by data-driven approaches. Even though their training or fitting phase could be slow, e.g. for deep neural networks (Bishop, 2008), their application is orders of magnitude faster than fundamental models since the data-driven ones are direct and do not require iterative solutions, differently from any kind of differential systems of equations or convergence procedures, e.g. for solving the multi-component vapor–liquid (flashing) problem. However, purely data-driven models cannot make accurate predictions far from the domain within which they were trained.

The scientific challenge is to retain the accuracy of the rigorous models with the time-to-solution of data-driven models. One of many

* Corresponding author.
E-mail address: flavio.manenti@polimi.it (F. Manenti).

https://doi.org/10.1016/j.compchemeng.2023.108252
Received 25 November 2022; Received in revised form 29 March 2023; Accepted 1 April 2023
Available online 5 April 2023

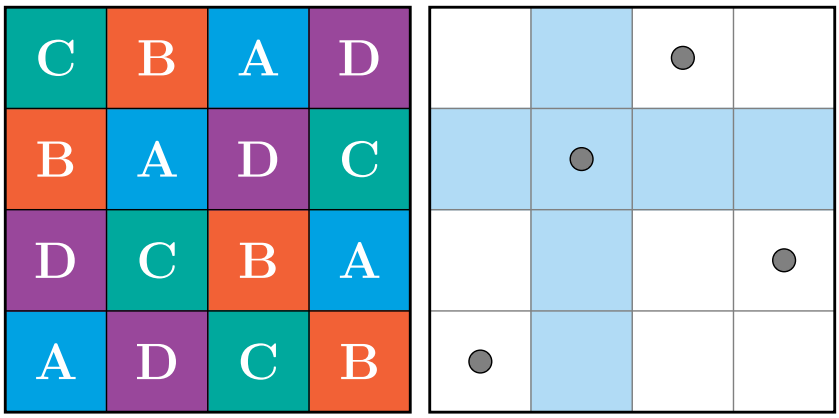


Fig. 1. Visual representations of (a) a 4 × 4 latin square with 4 objects (A, B, C, D), and (b) a two-dimensional latin hypercube design of experiment with 4 total samples (in light blue, it is highlighted that a sample does not share the row or the column with any other sample).

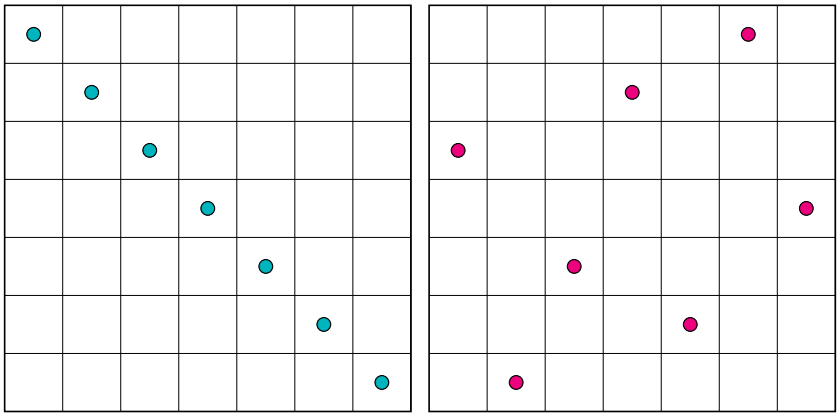


Fig. 2. Comparison of a two-dimensional uniform latin hypercube design of experiments with 7 samples between (a) a poorly space-filled configuration and (b) a better space-filling design.

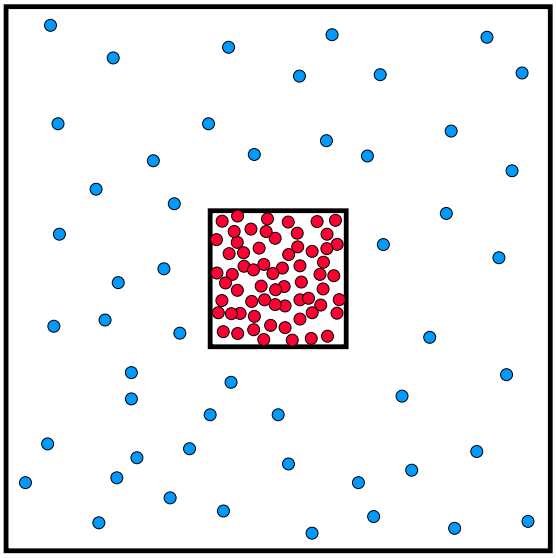


Fig. 3. Two-dimensional qualitative example for the stratified design of experiments using two domains. In red, data points contained in the internal, more concentrated, domain. In blue, data points belonging to the external, more sparse, domain.

approaches for solving this challenge might be model order reduction (Agarwal and Biegler, 2013), extensively used in computational fluid dynamics (Lassila et al., 2014). This method has the advantage of maintaining a great extent of the rigorous approach. However, one of its drawbacks is that model reduction must be performed after an accurate mathematical study of the specified system, thus precluding an automated and repetitive approach adaptable to different systems. Moreover, accuracy cannot be guaranteed if the assumptions on which the order reduction is based are too heavy. Another promising path is surrogate modeling. Surrogate models are input–output metamodels that approximate mathematical functions (Barton, 1992). These black-box models can be constructed with many different strategies, starting from simpler linear regression models all the way to more complex deep neural networks (Bhosekar and Ierapetritou, 2018; Westermann and Evins, 2019; Shokry et al., 2020). The main idea is to take data samples from the mathematical function to be approximated, e.g. digital twins, and tune the data-driven models to obtain the best fit on the data samples.

Finally, a very interesting alternative to the aforementioned approaches relies in the gray-box modeling, or hybrid modeling. This hybrid approach has been recently gaining more and more interest given the high accuracy obtainable in the prediction and the physical reliability of the generated models (Xiong and Jutan, 2002; von Stosch et al., 2014; Zendehboudi et al., 2018; Asprion et al., 2019; Sansana et al., 2021; Guo et al., 2022; Rajulapati et al., 2022). In fact, gray-box modeling tries to merge the rigorousness of first-principles models,

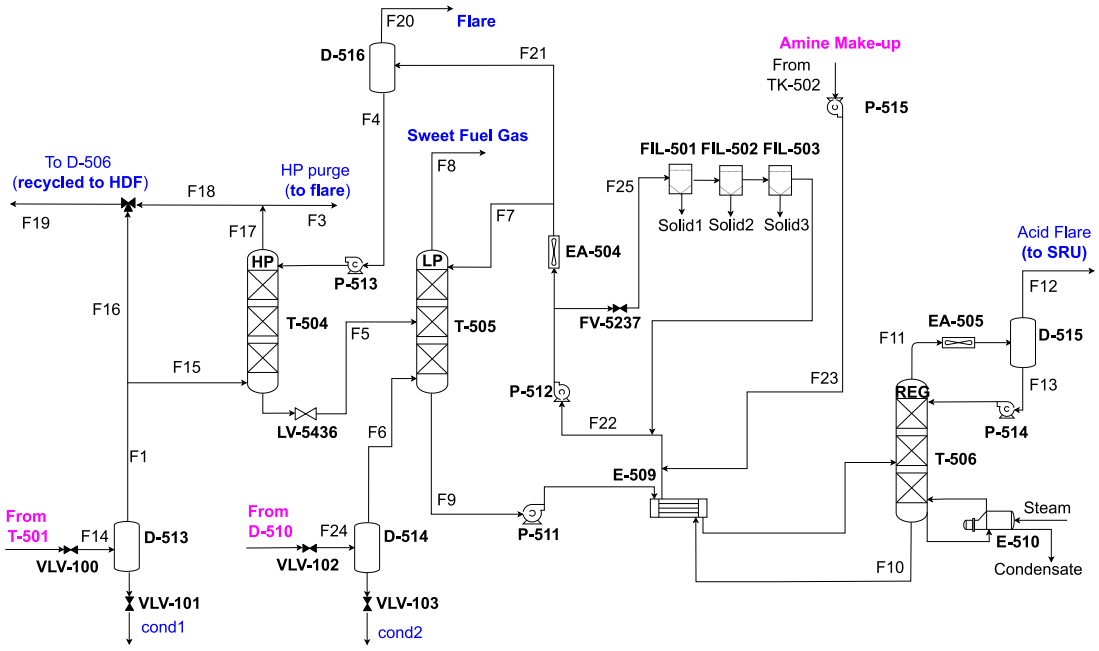


Fig. 4. Process flow diagram of the amine scrubbing process of the Itelyum exhausted oil refinery of Pieve Fissiraga.

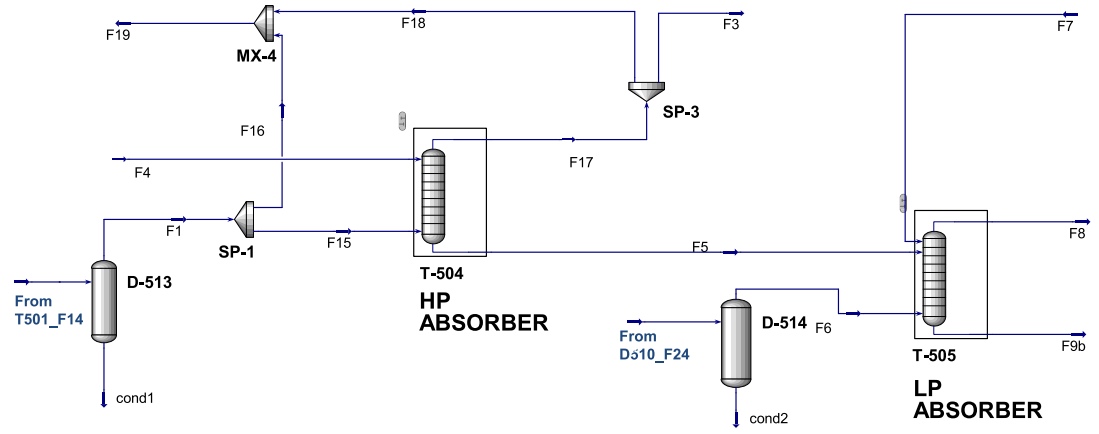


Fig. 5. Aspen HYSYS simulation for the absorber section of the amine scrubbing process.

or white-box models, with the versatility and accuracy of data-driven and machine learning models, or black-box models. This approach has been applied successfully for the metamodeling of chemical processes, especially for surrogate-based optimization (Bajaj et al., 2018; Pedrozo et al., 2021).

In the framework of black-box and gray-box modeling, the data sampling represents a very crucial procedure since the quality of the sampling greatly affects the quality of the surrogate model. Moreover, it is usually the most time-consuming step in the whole surrogate model creation. Data is generally collected through Design of Experiments (DoE) techniques that can be classified as conventional or unconventional. In the conventional category lie the traditional space-filling sampling techniques (Kleijnen, 2010; Pronzato and Müller, 2012; Kleijnen, 2015; Damblin et al., 2013; Sanchez and Wan, 2015) with the Latin Hypercube Design (LHD) being the most popular, as reported by Viana (2013). The unconventional DoE comprises mainly the adaptive sampling techniques of which an extensive review has been conducted by Liu et al. (2018).

In this work, a framework for the automated surrogate model creation of digital twins using black-box machine learning models is proposed. The choice of applying this particular modeling technique resides in the fact that the whole work is aimed at setting a base ground for crafting automation procedures for creating metamodels of any generic chemical process where a digital twin or process simulation is available. Both white-box and gray-box models have been neglected given the need for a human intervention in the pipeline of the model creation, by means of development and definition of rigorous models and associated constraints. Thus rendering the automated procedure useless. The digital twins are exploited through a nested Latin-Hypercube Design of Experiments to generate sample points to be used in the training, performed through cross-validation, of several machine learning algorithms. The machine learning algorithms applied in this study (listed in Section 2.3) are among the most common ones (with an exception for shallow artificial neural networks and Gaussian process regression models which both may require a separate special treatment) and they vary in model complexity and number

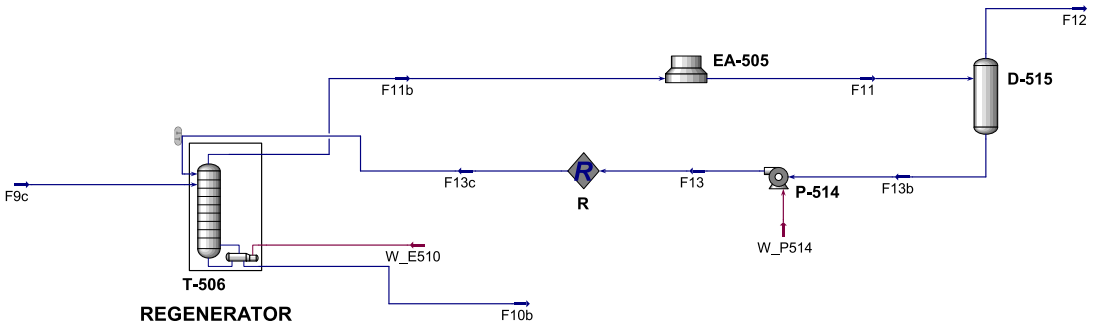
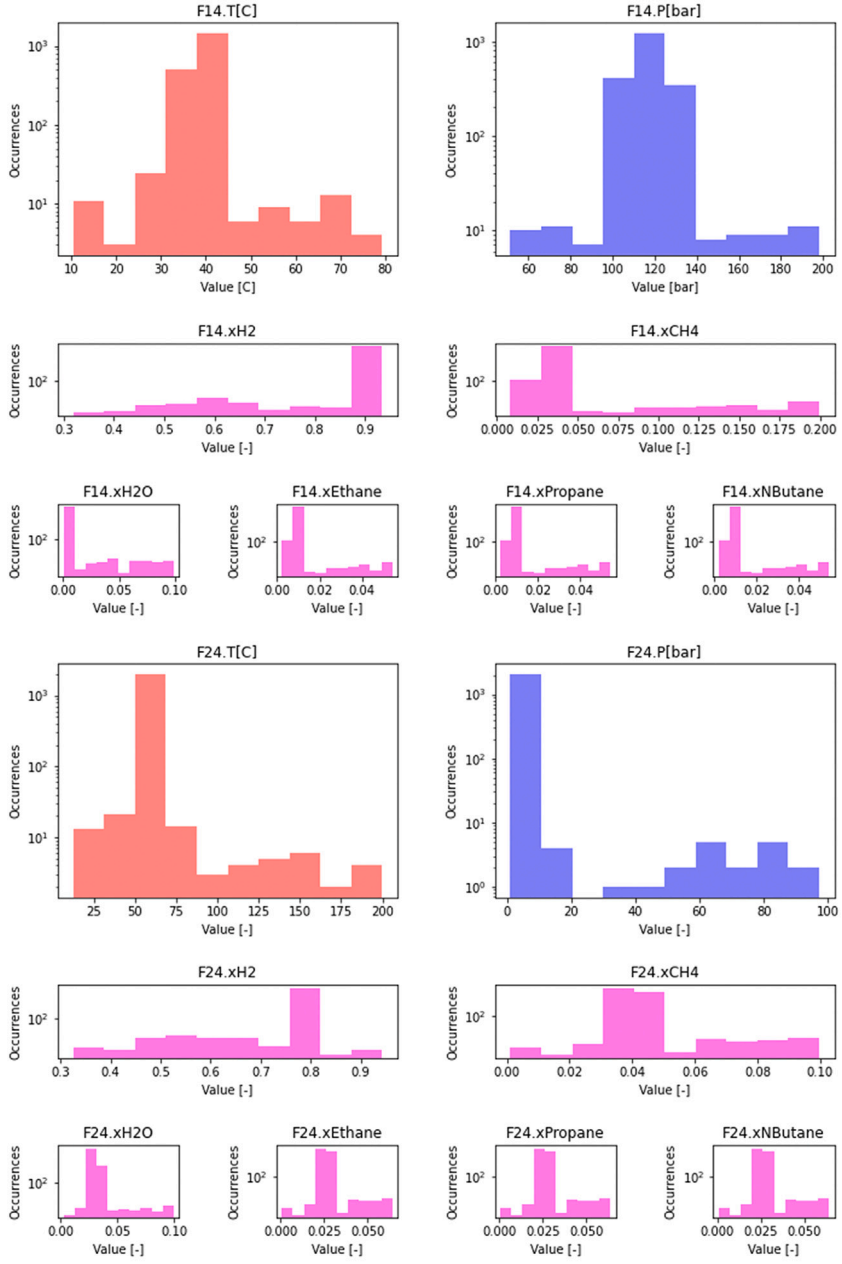
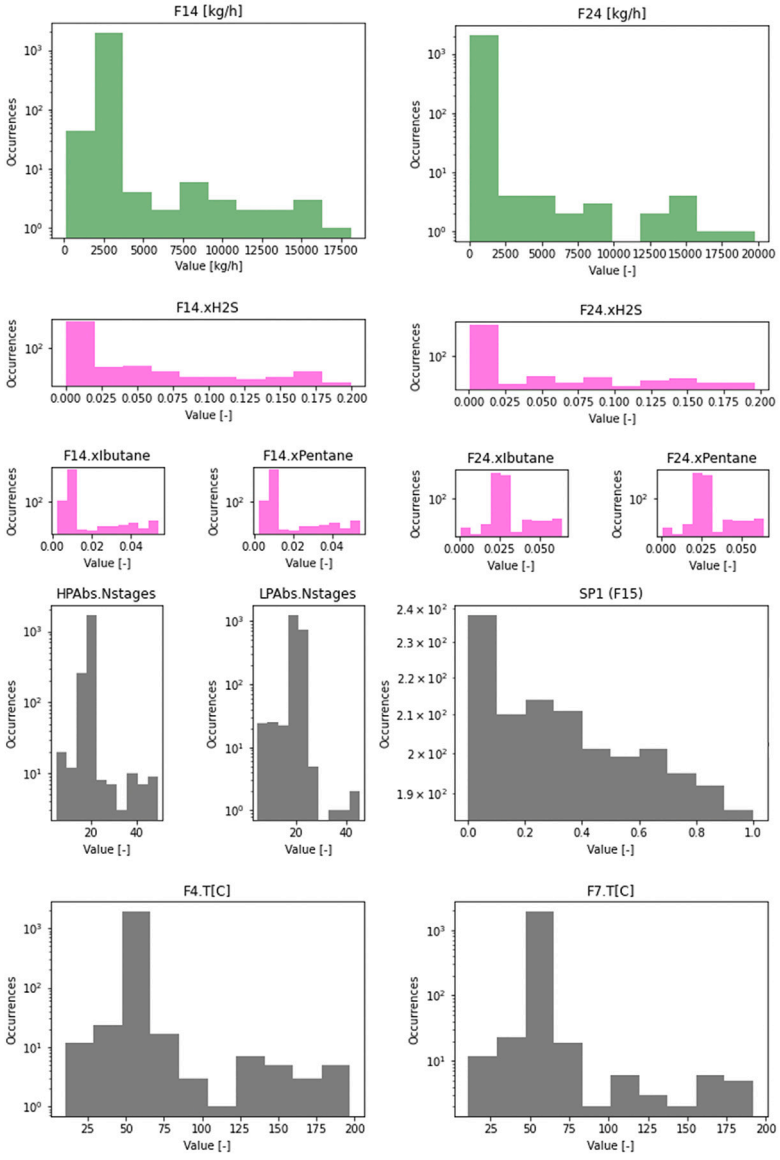


Fig. 6. Aspen HYSYS simulation for the regeneration section of the amine scrubbing process.

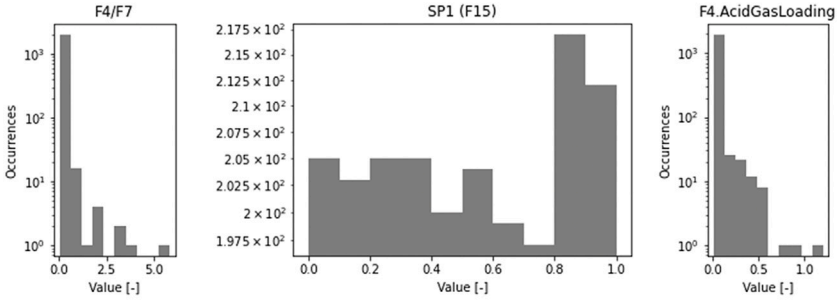


(a) Absorbers section features histograms (1/3 of Figure 7)

Fig. 7. Histograms of the distributions of features for the absorbers section. On the y-axes, the number of samples is reported and its corresponding value is indicated on the x-axes. The plot is in log scale on the y-axis.



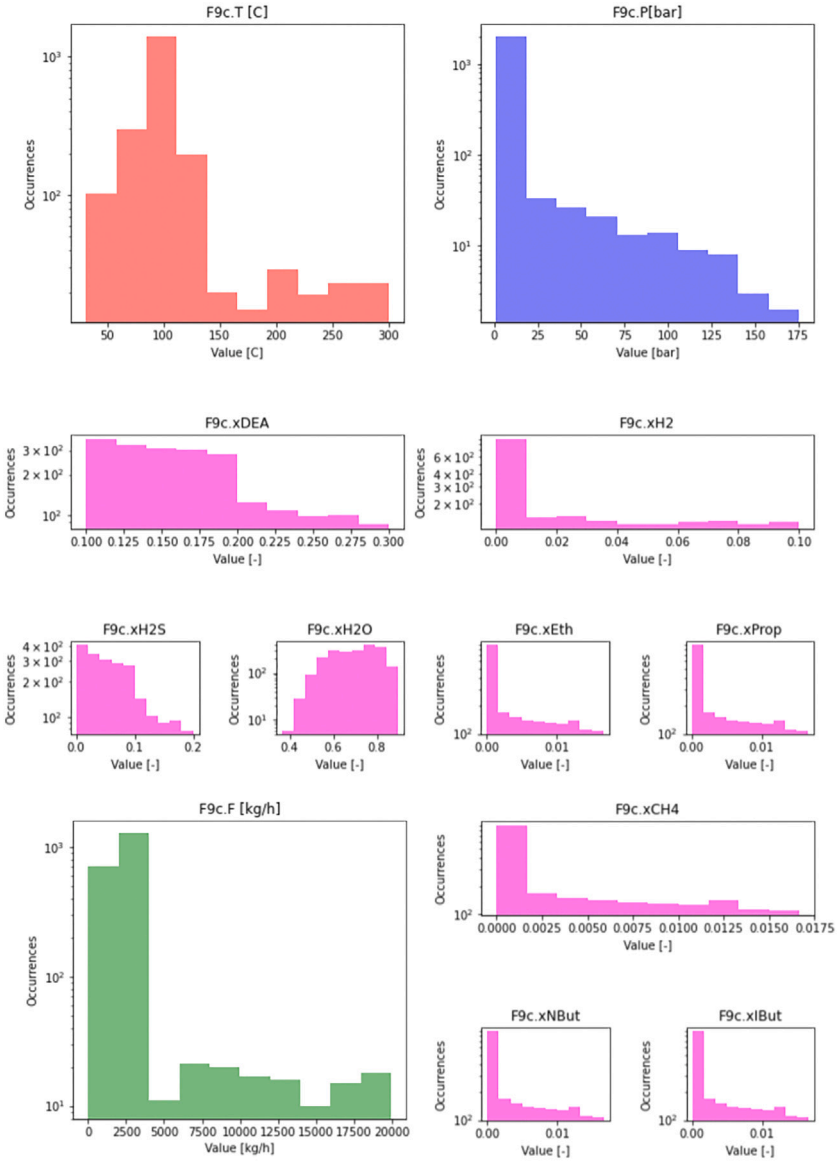
(b) Absorbers section features histograms (2/3 of Figure 7)



(c) Absorbers section features histograms (3/3 of Figure 7)

Fig. 7. (*continued*).

of parameters. Even though it is always better to favor less complex models, for increased explainability and reduced overfitting potential, in this case complexity is not accounted for and the sole parameter for model selection is accuracy. However, large scale machine learning models, such as Deep Neural Networks, are neglected entirely given the necessary conditions required to obtain useful results. In particular, the data availability and variability must be extremely high (Thebelt et al., 2022). Generally, this condition may not hold true when the



(a) Regenerator section features histograms (1/2 of Figure 8)

Fig. 8. Histograms of the distributions of features for the regenerator section. On the y-axes, the number of samples is reported and its corresponding value is indicated on the x-axes. The plot is in log scale on the y-axis.

fundamental model, to which a surrogate is needed, is already computationally expensive and cannot be exploited to maximize data volume and variance effectively for a large scale machine learning model.

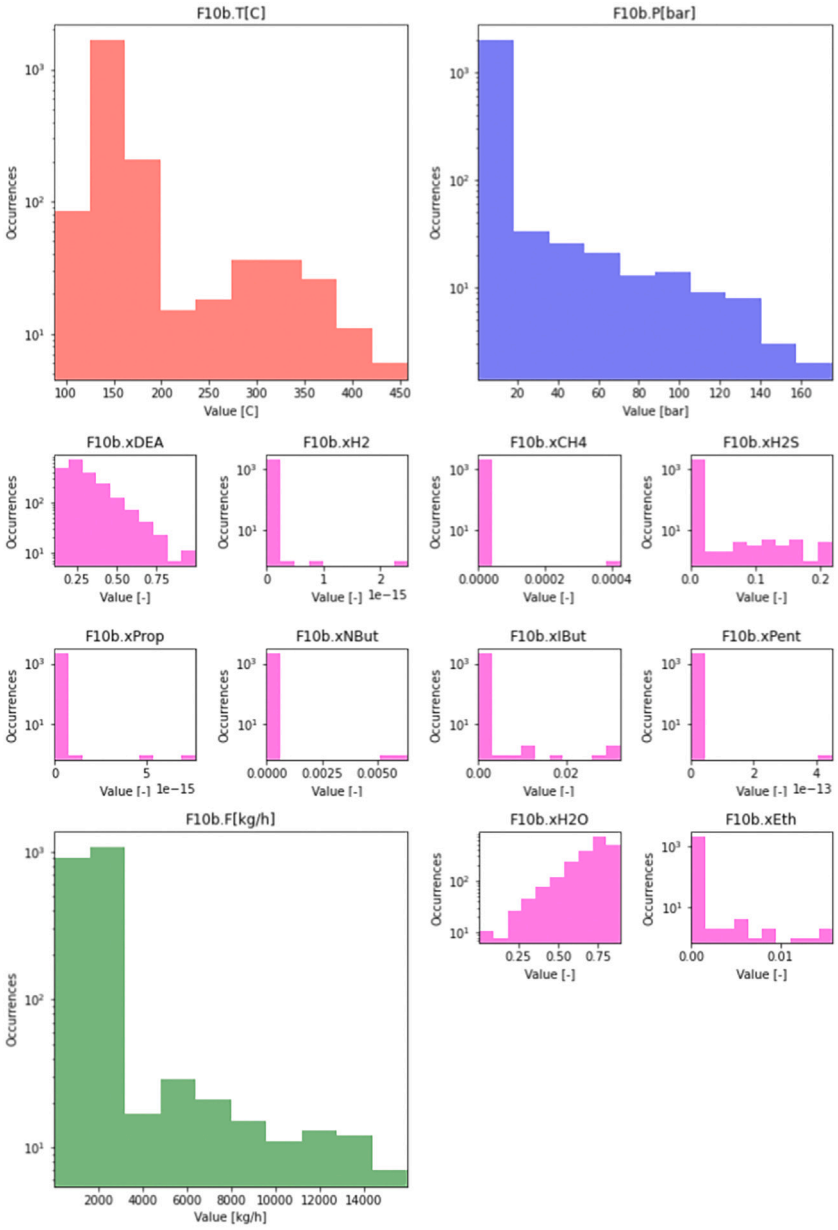
The remaining of the article is structured as follows: in Section 2 the general methods for the simulation, the design of experiments, and the machine learning modeling are introduced; in Section 3, the results of the application of the methods proposed are presented and a broad description of the process is given; in Section 4, the conclusions are drawn.

2. Methods

2.1. Industrial process modeling

The process selected for the digital twin development and consequent surrogate model creation is an amine scrubbing process, from the exhausted oil refinery of Itelyum Regeneration S.p.A. in Pieve Fissiraga (Lodi), Italy. Itelyum currently owns two oil refineries located in Pieve Fissiraga (LO, Italy) and Ceccano (FR, Italy), which have a combined treatment capacity of approximately 200 kt per year. Waste oil is produced by regenerating lube oil through combined thermal de-asphalting and hydrofinishing, according to the patented Revivoil technology (Minana et al., 1994). The off-gases coming from the hydrofinishing process are treated in the amine scrubbing process, using Diethanolamine (DEA). A more extensive process description is given in Section 3.1.

The digital twin has been created by modeling the amine scrubbing process with Aspen HYSYS V11. The simulation is developed in steady-state, due to the process being subjected to small time-dependent operating condition variations during normal operations. The validation of the simulation model, required to define it as a digital twin, has been done with plant data taken from the Yokogawa Exaquantum data historian complemented by previous Itelyum laboratory measurements, especially for stream compositions. The thermodynamic modeling of the chemical components has been carried out using the Aspen Technology "Acid Gas - Chemical Solvents" property package (Dyment



(b) Regenerator section features histograms (2/2 of Figure 8)

Fig. 8. (*continued*).

and Watanasiri, 2015) which employs the Peng and Robinson (1976) equation of state for the vapor phase and the electrolyte non-random two-liquid (eNRTL) activity coefficient model for the liquid phase (Song and Chen, 2009).

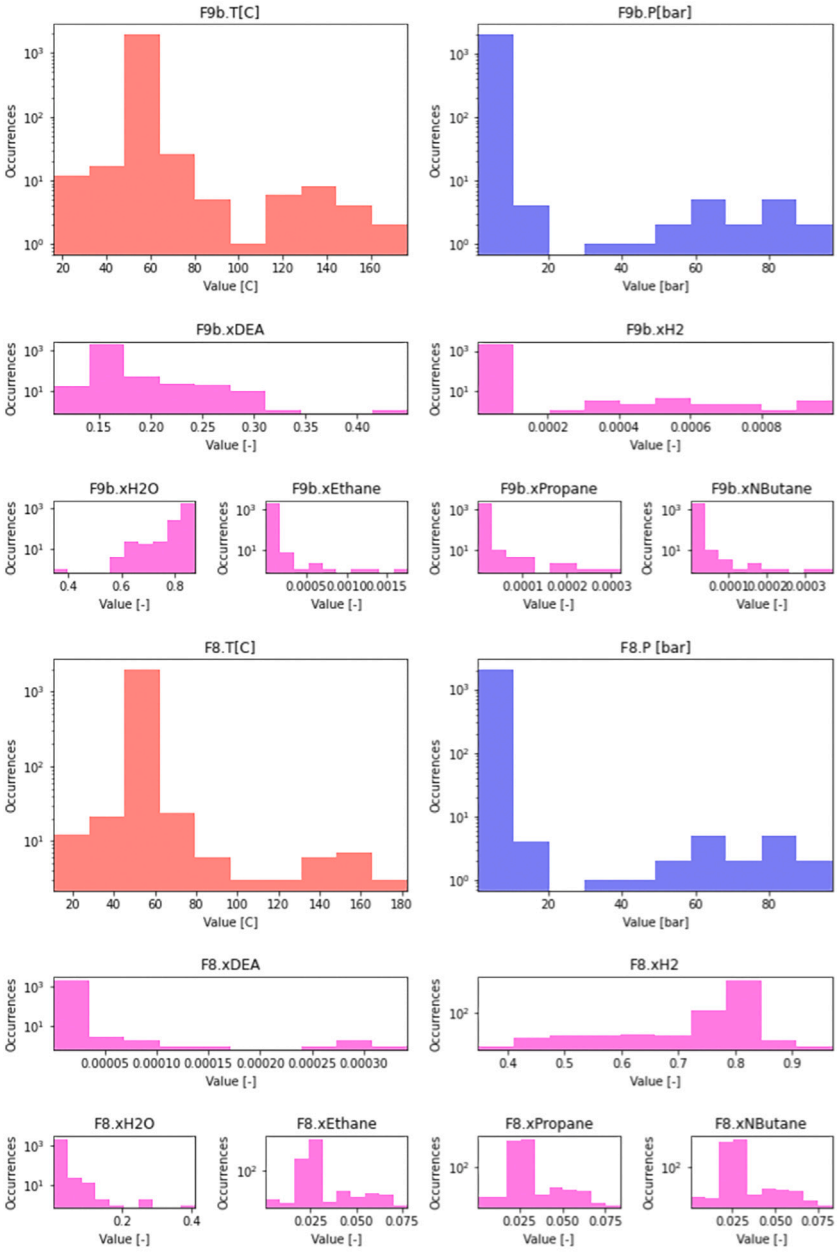
2.2. Data sampling

Before starting the data generation procedure it is very important to carefully define the input and output variables. In surrogate models, the former are referred as features, while the latter become the predicted variables, or targets (Jiang et al., 2020).

The quality of the surrogate model is directly related to the data quality. The data exploited for the surrogate model has been generated by automatically solving multiple times, at different input conditions, the HYSYS simulation. These input conditions have been selected through the DoE. Precisely, the conventional Latin Hypercube Design (LHD) method has been applied (McKay et al., 1979).

The LHD method has been selected among other strategies after an extensive literature review that demonstrated its capabilities regarding flexibility and adaptability to more complex DoE strategies (McKay et al., 1979; Morris and Mitchell, 1995; Loh, 1996; Helton and Davis, 2003; Cioppa and Lucas, 2007; Viana, 2013, 2016; Li et al., 2017; Sheikholeslami and Razavi, 2017; Panwar and Michael, 2018; Donovan et al., 2018). In particular, when tackling high dimensional problems, with exponentially high sampling volumes, most of the one-shot DoE strategies that show a remarkable efficiency in low dimensional problems (Navid et al., 2018), e.g. the Sobol' (1967) quasi-random low discrepancy sequences, do not produce a sensible improvement over LHD (Manteufel, 2000; Bhattacharyya, 2018) and, moreover, do not provide the same amount of flexibility that LHD does (Helton and Davis, 2003).

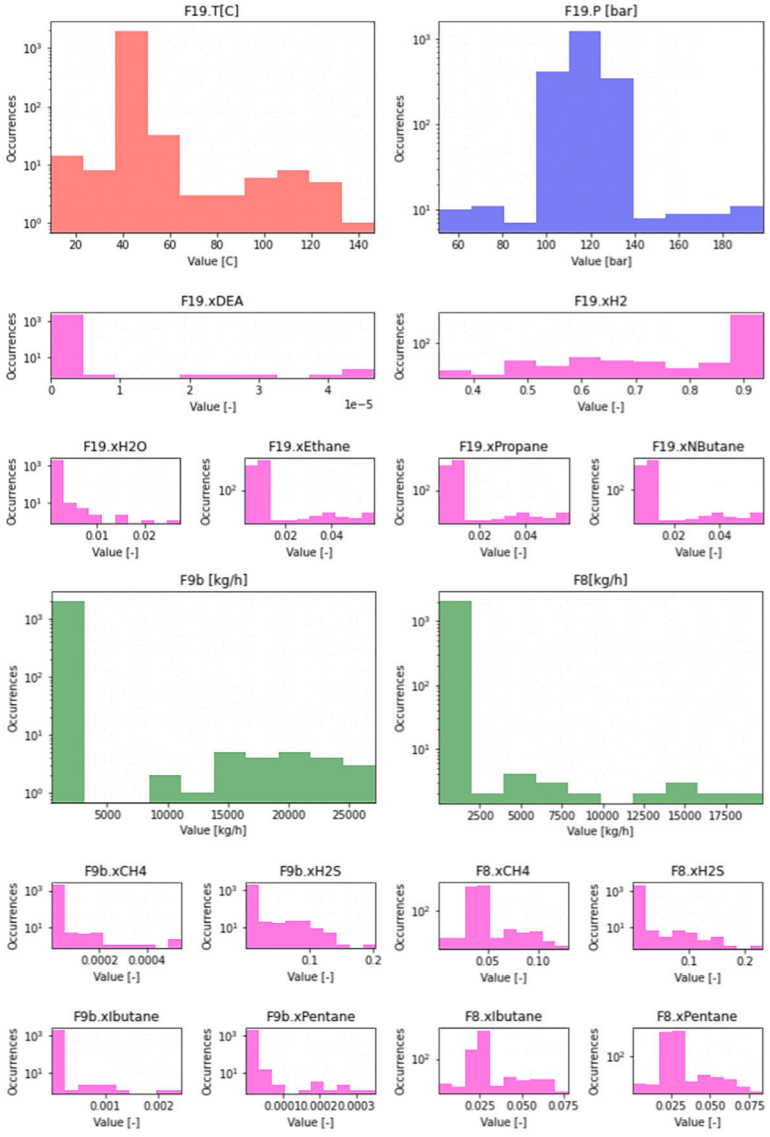
Latin Hypercube Sampling (LHS) is a sampling strategy derived from the "Latin square" concept of combinatorial mathematics in which an $m \times n$ matrix is filled with objects that only appear once in each



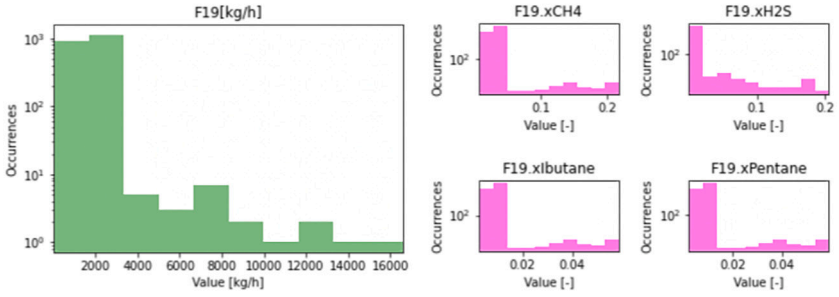
(a) Absorbers section targets histograms (1/3 of Figure 9)

Fig. 9. Histograms of the distributions of targets for the absorbers section. On the y-axes, the number of samples is reported and its corresponding value is indicated on the x-axes. The plot is in log scale on the y-axis.

row and each column as shown in Fig. 1(a). The latin square can be generated with many different permutations. The number of permutations increases exponentially with the increase of the array dimensions. Since LHS can have many different configurations for the same design of experiments, as shown in Fig. 2, it can be upgraded with optimization techniques to maximize particular features (Helton and Davis, 2003; Sheikholeslami and Razavi, 2017; Li et al., 2017; Vořechovský and Mašek, 2020) and find the optimal one that satisfies the specified constraints. One example is space-filling maximization (see Fig. 2), where an optimization routine is iteratively applied to a conventional LHS in order to find a DoE configuration that is able to better homogeneously fill the whole experimental domain. However, these optimization procedures may be computationally costly, especially when the dimensions of the problem are high and the LHS could be shaped in a huge amount of different configurations. Moreover, when applying the simpler and faster LHD in high dimensional problems, the probability of obtaining a particular unwanted shape, like the one of Fig. 2(a), is exponentially lower as the amount of input variables increases and the slower optimization methods may fail to provide a sensible improvement. For higher dimensional problems, sequential strategies are better suited to define a more efficient DoE (Li et al., 2017; Sheikholeslami and Razavi, 2017; Bhattacharyya, 2018; Xu et al., 2018). In this case, the experimental domain may be sliced in several portions, e.g. nested



(b) Absorbers section targets histograms (2/3 of Figure 9)



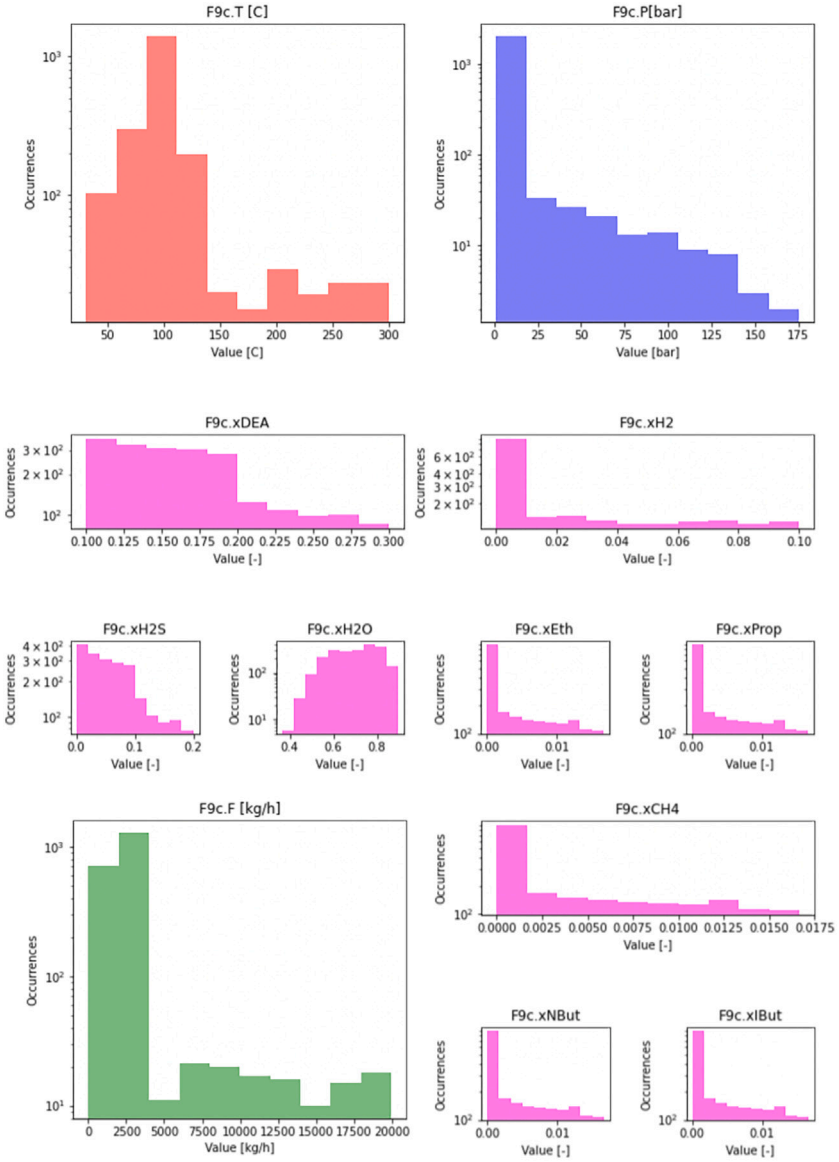
(c) Absorbers section targets histograms (3/3 of Figure 9)

Fig. 9. (*continued*).

hypercubes, and each one of them studied with a specific design of experiments. The sequential strategy is applied in order to add new sample points where required based on a particular metric, e.g. to increase prediction accuracy of specific areas.

An important aspect in the DoE creation is the definition of the boundaries within which each variable can vary. In this study, several nested domains have been defined for the DoE, in a similar fashion to what already proposed by Qian (2012), Liu et al. (2016) and Xu et al. (2018). The inner domains, closer to the nominal condition of steady state operation, have a higher concentration of points, while the outer ones, further away from the actual operating conditions, are more and more sparse. A graphical representation is shown in Fig. 3.

The internal domains are supposed to be formulated in a way to guarantee feasibility everywhere throughout the entire sub-domain.



(a) Regenerator section features histograms (1/2 of Figure 10)

Fig. 10. Histograms of the distributions of targets for the regenerator section. On the y-axes, the number of samples is reported and its corresponding value is indicated on the x-axes. The plot is in log scale on the y-axis.

External domains, however, are more explorative in nature and can be constructed in a way to gather knowledge in really uncommon areas of operation. Sparsity helps in reducing the total amount of computation expense required. One of the benefits of this exploration is to corroborate the actual feasibility domain, by possibly studying how far it can extend.

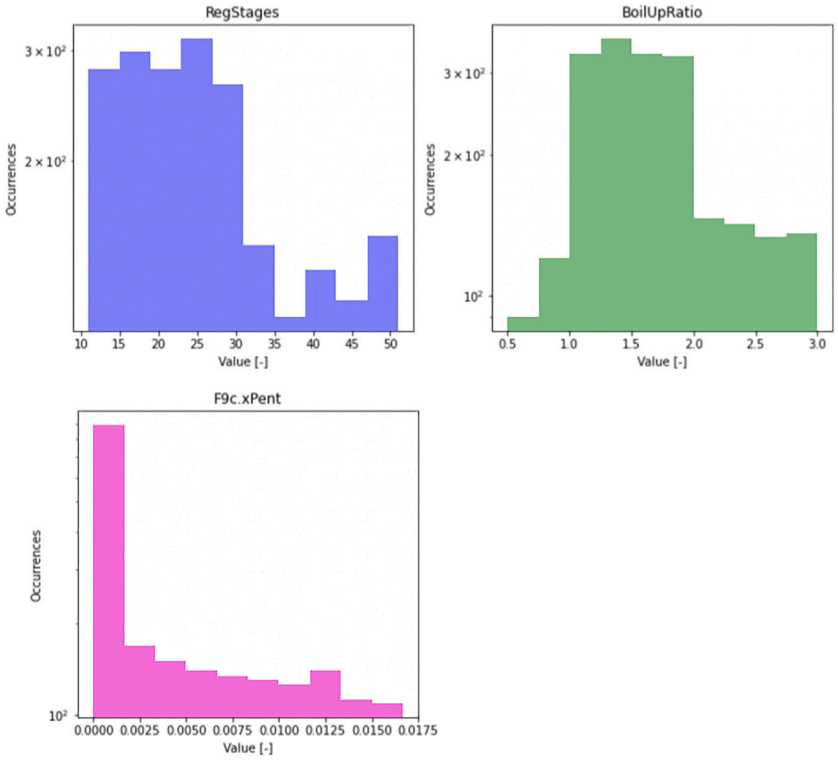
The volume of a single hypercube generated by a uniformly distributed and discretized sampling domain can be calculated as follows:

$$V = \prod_{i=1}^{m} \frac{m_i^{UB} - m_i^{LB}}{s_i} \tag{1}$$

where V is the hypercube volume, m is the number of feature variables, m_i^{UB} and m_i^{LB} are, respectively, the upper and lower bounds of feature m_i, and s_i is the discretization step of each variable m_i. Thus, the sampling frequency grows exponentially with $\mathcal{O}(n^m)$ as the number of dimensions added to the problem increases. For example, with an hypercube of order 10, i.e. with 10 input features, and a grid of 10 nodes the total amount of points that can be sampled is 1e+10. If we let a rigorous model to take just 1 second to obtain a solution we would need approximately 317 years of computational time to sample the whole discretized hypercube.

The trade-off that must be sought after is the one between the amount of samples, the computational expense and the explained variance (Loh, 1996; Manteufel, 2000; Helton and Davis, 2003; Donovan et al., 2018; Ledolter and Kardon, 2020). A higher amount of samples means a higher computational expense but does not necessarily mean an improvement in explaining the variance. Thus, the optimum should be to maximize the explained variance by keeping both number of sample and computational expense to a minimum.

The features selected for the black box model are, first of all, those stream related variables that are necessary to fix all the degrees of freedom of the system, in addition to several other design variables. In particular, the former are Temperature (T), Pressure (P), array of compositions ($\bar{x}$), and Flowrate (F). However, this is just one of the many ways in which a flow-stream can be thermodynamically and



(b) Regenerator section targets histograms (2/2 of Figure 10)

Fig. 10. (*continued*).

physically defined. For example, instead of pressure, the density could be specified, and, instead of compositions and flowrate the individual components flowrate could be fixed. Nevertheless, the surrogate model and the framework developed should be agnostic with respect to the specific variable selected, since no fundamental relationships or rules are implemented for each kind of variable. The approach can be eventually defined as purely data-driven.

However, one important factor affecting the number of sample points usable in generating the metamodel, during data generation, is the feasibility of the process simulation. Especially when complex non-linear processes and advanced thermodynamics are taking place, it is not rare to encounter unfeasible working conditions if moving too far from the nominal operating point. This problem consequently affects the quality of the surrogate model itself by reducing the data available during the training of the machine learning models.

2.3. Modeling approach

The models investigated in this study for the creation of the surrogate metamodel fall under the umbrella of machine learning algorithms. Precisely, the ones considered are, in order of complexity: *linear regression*, *polynomial regression*, *support vector regression (SVR)* (Awad and Khanna, 2015), *decision tree regression*, *random forest* (Ho, 1995, 1998), *AdaBoost* (Freund and Schapire, 1995), *gradient boosting* (Friedman, 2001, 2002). For an extensive explanation of the aforementioned techniques the reader can refer to Bishop (2008). These algorithms were chosen among the most widely utilized regarding the surrogate modeling topic, and applied in their conventional form using the Scikit-Learn library (Pedregosa et al., 2011; Hao and Ho, 2019) for the Python programming language (Van Rossum and Drake, 2009).

Each model is fit to the sampled data using all the feature variables, i.e. the input to the black box as described in Section 2.2, while the target is a single output of the applied black box at a time. The data set available is initially split with a 80/20 rule in training and test, respectively, then the models are trained with a k-fold Cross-Validation (CV) (with $k = 5$) on the training set. After the training phase, only the best-fit model is selected as the definitive predictor of the target variable and, after retraining it on the entire training set, its performance is evaluated on the test set. The procedure is then repeated for every target, generating a set of optimal models (a single one for each target). The selection of the best models goes through the minimization of the cross-validation MAE of the target.

The indicators used to evaluate the performance are defined as follows,

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

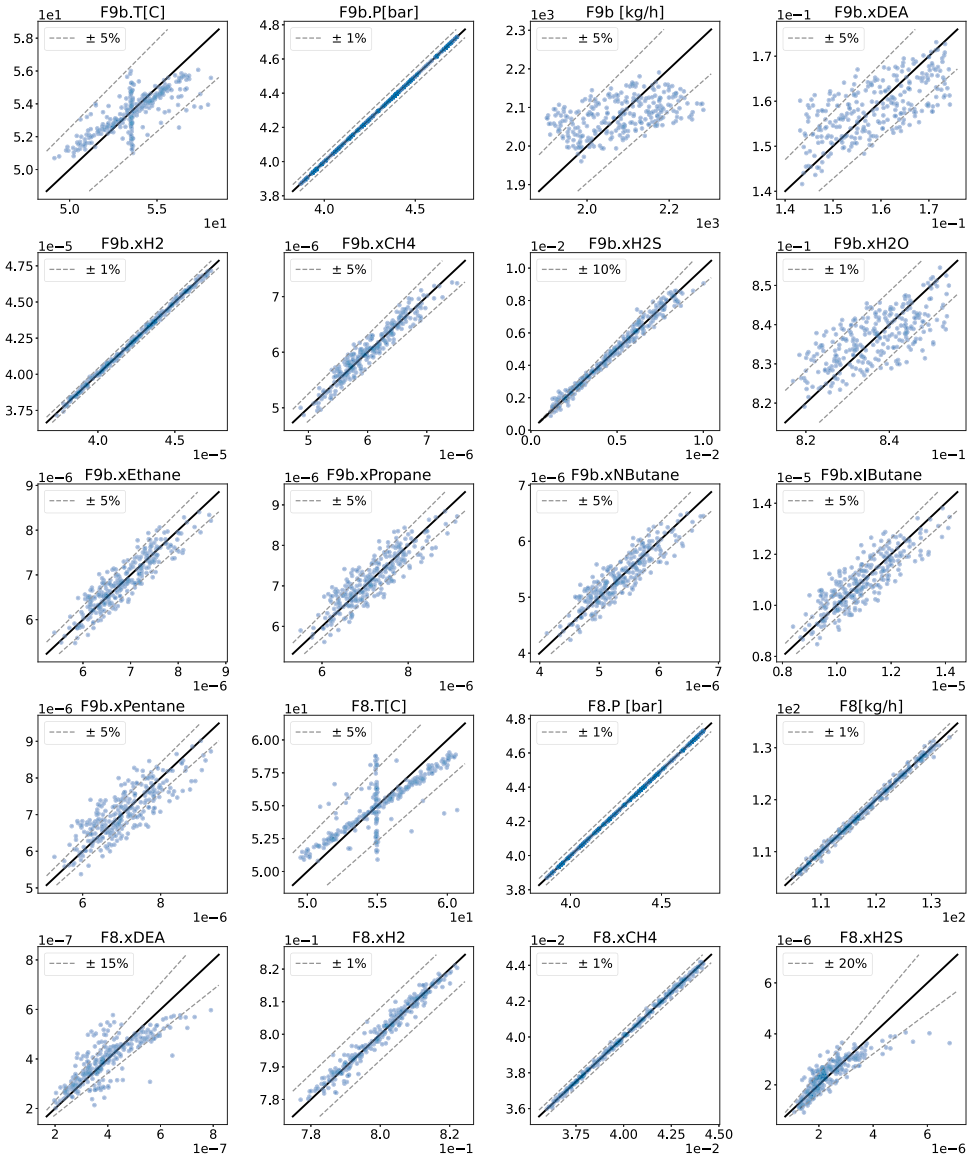
$$\mathrm{RMSE}(y,\hat{y}) = \sqrt{\frac{1}{n}\sum_{i=0}^{n-1}(y_i - \hat{y}_i)^2} \tag{3}$$

$$\mathrm{MAE}(y,\hat{y}) = \frac{1}{n}\sum_{i=0}^{n-1}|y_i - \hat{y}_i| \tag{4}$$

$$\mathrm{MAPE}(y,\hat{y}) = \frac{1}{n}\sum_{i=0}^{n-1}\frac{|y_i - \hat{y}_i|}{|y_i|} \tag{5}$$

where i is the ith data sample, y is the true value, $\hat{y}$ is the value predicted by the model, and n is the total number or samples.

Several authors have argued that the MAE is a better estimator with respect to the Root Mean Squared Error (RMSE) (Willmott and Matsuura, 2005; Willmott et al., 2009), shown in Eq. (3), and while this could be true the RMSE is still a good indicator of model performance (Chai and Draxler, 2014) and it is a good strategy to look at both the MSE and the RMSE in assessing the capabilities of a fit model (Chai and Draxler, 2014; Karunasingha, 2022). For a better understanding of the quality of the fit, the other statistical metrics described above are evaluated on the test set, i.e. the R^2 coefficient (Eq. (2)), the RMSE (Eq. (3)), the MAE (Eq. (4)), and the Mean Average Percent Error (MAPE, in Eq. (3)). R^2 is a traditional indicator of the



(a) Absorbers section parity plots (1/2 of Figure 11)

Fig. 11. Parity plot of the target variables for the absorbers section. Reported on the y-axes are the model predictions while on the x-axes are the digital twin ground truth values. The solid black lines represent the parity line while the dashed lines are selected error intervals.

goodness of fit of linear regressions that scales from $-\infty$, when there is no correlation between the model and the data, to 1, when they are perfectly correlated. However, it must be noted that the R^2 estimator may not be used in the formulation of Eq. (2) for comparing non-linear regression models. In that case, the mathematical formulation must be adjusted accordingly (Kvalseth, 1985; Miles, 2014).

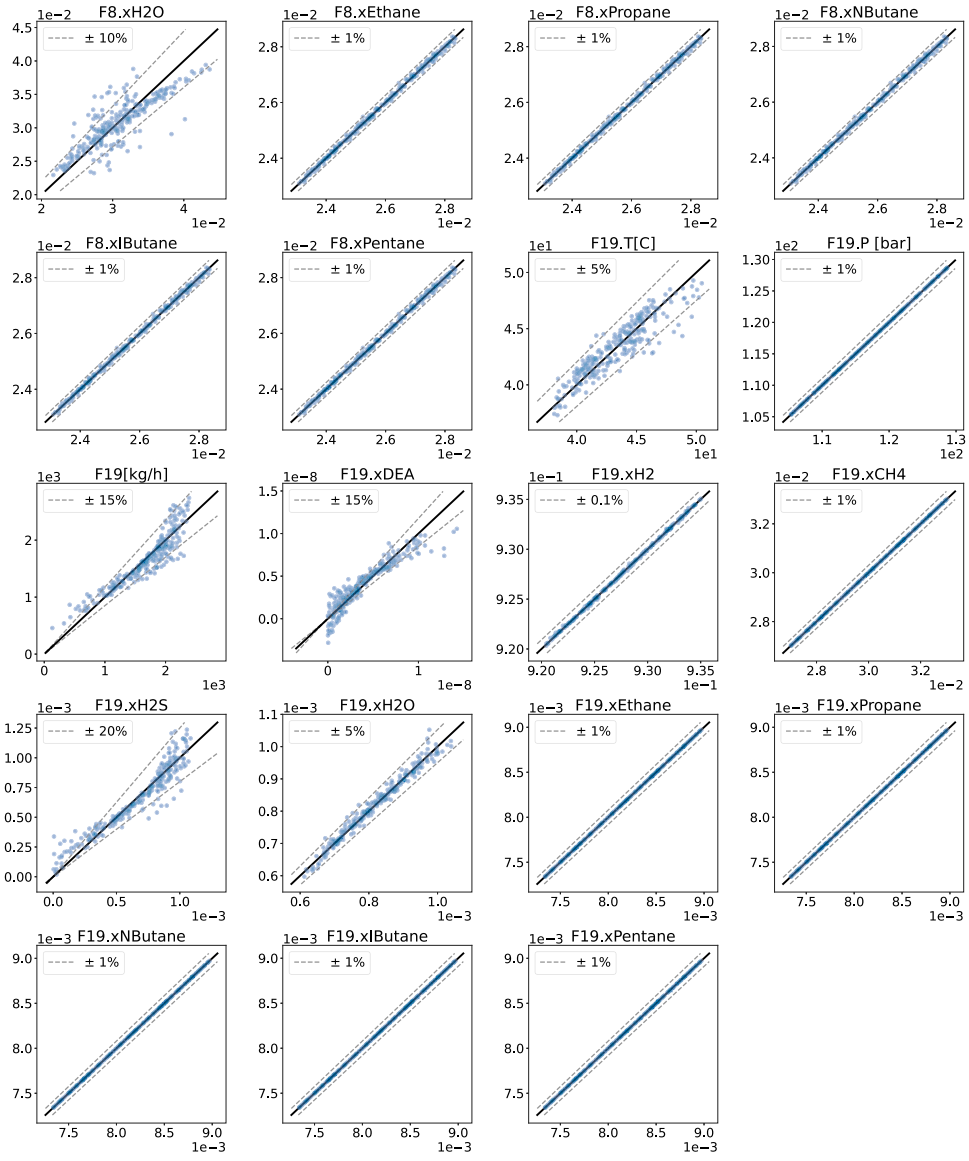
The MAPE, on the other hand, gives better insights on errors of values with small absolute magnitude since the MAPE tends to infinity when y_i, at the denominator of Eq. (5), tends to zero.

3. Results & discussion

3.1. Amine scrubbing process

The chemical process chosen as the case study for the creation of the surrogate model is the amine scrubbing section of the exhausted oil refinery of Itelyum located in Pieve Fissiraga (LO, Italy), as mentioned in Section 2.1. This process is operated to treat the off-gases coming from the hydrofinishing section of the plant. The treatment of such gases is necessary to remove dangerous pollutants, i.e. H_2S, and recover valuable components, i.e. H_2 and light hydrocarbons (C_1–C_5). The solvent used for the chemical absorption is Diethanolamine (DEA) in a water mixture of a molar ratio of 25/75, respectively. The layout of the process is reported in the Process Flow Diagram (PFD) of Fig. 4.

Two main sections compose the process, i.e. the absorbing section and the regeneration section. The absorbing section includes a high-pressure absorber (HP or T-504, in Fig. 4) and a low-pressure absorber (LP or T-505, in Fig. 4). The regeneration section only has a single column (REG or T-506, in Fig. 4) used to regenerate the spent amine coming from the upstream HP and LP absorbers. The actual number of stages of the columns was not at disposal for this study since only the theoretical stages were reported and due to the aging of the plant those values were deemed inaccurate. Thus, a correction of the theoretical number of stages has been assumed from literature studies (Kohl and Nielsen, 1997) and set equal to 20 for all the columns. The nominal operating conditions of the columns are shown in Table 1 while the design specifications for the absorbers and the regeneration are shown in Table 2.



(b) Absorbers section parity plots (2/2 of Figure 11)

Fig. 11. (*continued*).

Table 1
Operating nominal conditions for the absorbers and regenerator column. The temperature value is referred to in the middle tray while pressure is referenced at the bottom of the unit.

Unit	Tag	Temperature (°C)	Pressure (bar)	Pressure drop (mbar)
HP absorber	T-504	30	107.6	40
LP absorber	T-505	55	4.3	0
Regenerator	T-506	135	2.8	0

Table 2
Design specification used to simulate the amine scrubbing system.

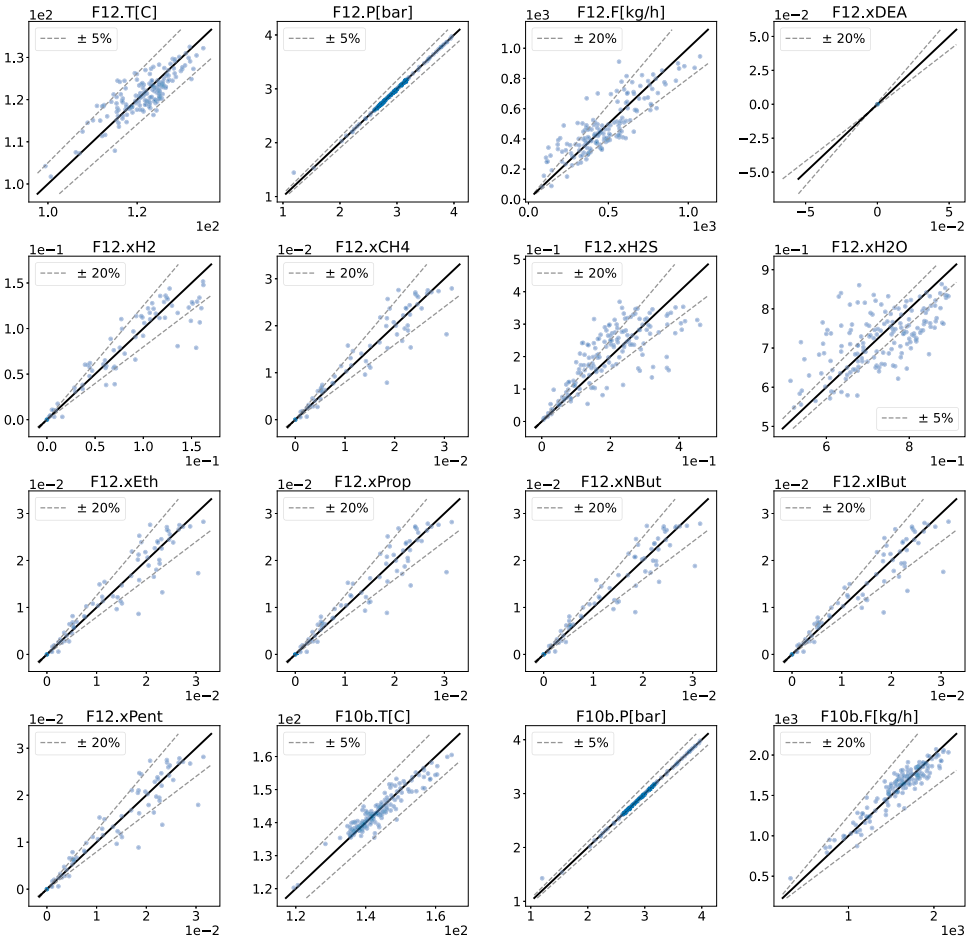
Unit	Tag	Specification	Value
HP absorber	T-504	Number of stages	20
LP absorber	T-505	Number of stages	20
Regenerator	T-506	Number of stages	20
Regenerator	T-506	H_2S recovery	0.98

Between the two sections, the E-509 heat exchanger is an economizer used to heat up the spent amine exiting the absorbers and at the same time cool down the regenerated amine flowing from the bottom of the regenerator T-506. The air cooler EA-504 is used to further cool down the regenerated DEA entering the LP absorber. The top product of column T-506 is partially condensed using the air cooler EA-505 and the uncondensed gases are transferred to a Sulfur Recovery Unit (SRU), as shown in Fig. 4.

The composition and flowrate values of all the input and output streams are reported in Table 3. Except for the utilities, namely, the steam sent to the reboiler of the T-506 column and the amine make-up, the only input streams entering the process are F14 and F24. These two streams originate from different sections of the plant and differ slightly in composition but heavily in flowrate value.

3.2. Process simulation and validation

As mentioned in Section 2.1 the process simulation has been performed using Aspen HYSYS software with the "Acid Gas - Chemical Solvents" thermodynamic property package. Temperatures, pressures, and flowrate values come from the DCS during a steady-state operation of the process. On the other hand, due to the lack of on-line



(a) Regeneration section parity plots (1/2 of Figure 12)

Fig. 12. Parity plot of the target variables for the regenerator section. Reported on the y-axes are the model predictions while on the x-axes are the digital twin ground truth values. The solid black lines represent the parity line while the dashed lines are selected error intervals.

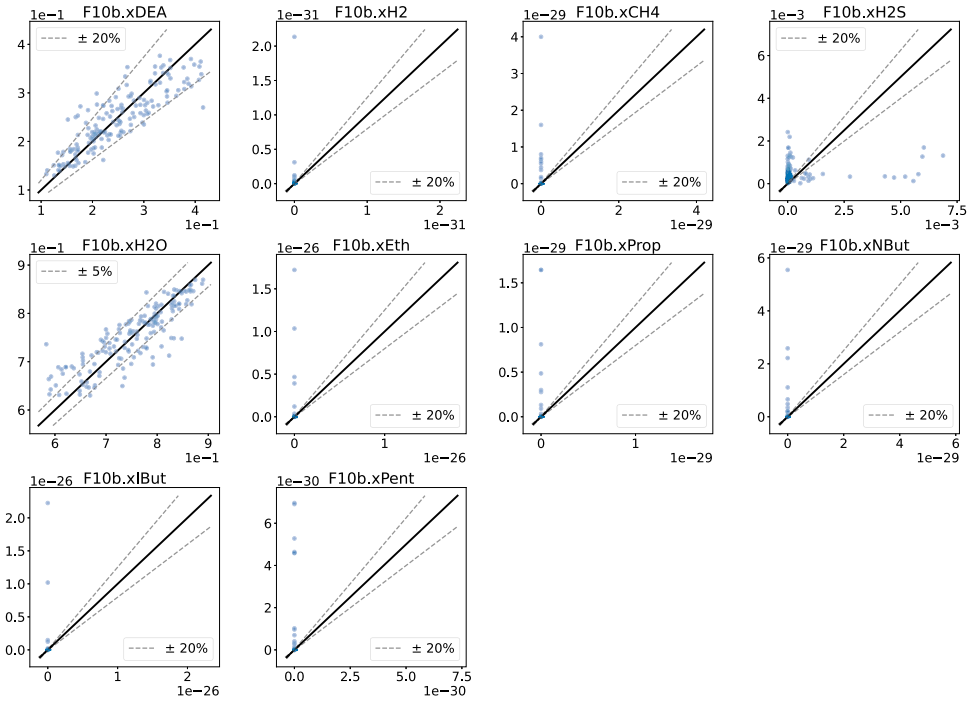
Table 3
Flowrates and normalized compositions averaged between the start of the run and the end of the run for the input–output streams of the amine scrubbing process.

Stream	H_2O	DEA	H_2	H_2S	Methane	Ethane	Propane	i-Butane	n-Butane	n-Pentane	Flowrate (kg/h)
F14	1.14e−3	0.00e+0	9.43e−1	1.08e−3	3.08e−2	8.36e−3	3.96e−3	1.25e−3	1.59e−3	8.92e−3	2221.96
F24	3.41e−2	0.00e+0	7.77e−1	4.85e−3	3.81e−2	2.45e−2	1.69e−2	6.51e−3	9.37e−3	8.89e−2	118.29
cond1	9.98e−1	0.00e+0	1.29e−3	1.53e−4	6.49e−5	2.15e−5	7.21e−6	2.41e−6	2.36e−6	8.87e−6	7.02
cond2	–	–	–	–	–	–	–	–	–	–	0
F19	4.99e−4	3.29e−9	9.45e−1	3.02e−8	3.08e−2	8.37e−3	3.97e−3	1.25e−3	1.60e−3	8.93e−3	2192.65
F3	–	–	–	–	–	–	–	–	–	–	0
F8	3.03e−2	3.83e−7	7.84e−1	2.07e−6	3.85e−2	2.47e−2	1.70e−2	6.55e−3	9.43e−3	8.94e−2	111.92
F20	–	–	–	–	–	–	–	–	–	–	0
F12	5.80e−2	0.00e+0	3.66e−3	9.34e−1	5.14e−4	5.83e−4	4.24e−4	2.45e−4	1.77e−4	2.20e−3	24.71

measurements of chemical compositions, the available data for such quantities are infrequent laboratory measurements. With such measurements, only light hydrocarbons up to C_4 are sampled, thus leaving the remaining heavier hydrocarbons unmeasured. The assumption made in this regard is to consider the trailing amount of undistinguished hydrocarbons as pure n-Pentane. The normalized average composition between the start and the end of the run, i.e. from a clean start-up to a manual shut-down for maintenance and cleaning, have been taken and used for the purpose of simulating the amine scrubbing process. Such values are reported in Table 3.

The convergence of the simulation is rather slow given the presence of two recycle loops, one inside the regeneration section and the other between them, and, most importantly, the complex thermodynamic system. The time required to obtain such a solution is in the order of one to ten seconds and for the data generation procedure described in Section 2.2 this is hindering, since a large number of runs is needed. To overcome this obstacle, the simulation has been split into two independent, individually sampled, sections, shown in Figs. 5 and 6.

In order to validate the process simulation and be able to define it as a digital twin, it is necessary to compare the simulation results with process data. To do so, a single variable has been chosen as the key performance indicator for the accuracy of prediction and it is the stream flowrate since it is the most abundant measurement throughout the process. Unfortunately, it is not possible to measure every variable or stream inside a real plant due to the costs associated with the instrumentation. For this particular case, the flowrate variables measured are the ones of stream F3, F4, F6, F7, F8, F9, F12, F13, F19, and F23. The residual absolute relative error calculated on the data available is approximately 4 %. Considering the assumptions made and the quality of data this residual amount can be considered low enough



(b) Regeneration section parity plots (2/2 of Figure 12)

Fig. 12. (*continued*).

Table 4
Mass flow, in kg/h, for selected streams in the amine scrubbing process.

	F18	F9	F8	F5	F4	F7
Digital twin	2196.52	4019.19	111.92	2428.13	2307.40	1371.34
Process data	2202.56	4029.00	113.00	2466.00	2448.00	1568.00

to define this process simulation as a digital twin. An example is shown in Table 4.

3.3. Data generation

Given the process model described in Section 3.2, the data generation can start only after the definition of the feature variables, as mentioned in Section 2.2, and their variability domains. The variables chosen as features are reported in Table 5 for the absorbers section while in Table 6 for the regeneration section.

To reduce the dimensionality, thus the complexity, of the problems only the methane concentration is made to vary, and the other hydrocarbons are bounded in a constant proportion with respect to the methane content during the design of experiments procedure. This ratio is calculated using the averaged experimental measurements for both streams F14 and F24. The assumption made is that the distribution of hydrocarbon molecules found in the flue gases exiting the hydrofinishing process is constant. In Table 7 the molar ratios with respect to methane are reported.

For each of the two sections in Figs. 5 and 6, 6000 simulations have been run with inputs generated through the Latin-Hypercube design of experiments described in Section 2.2. However, the simulation convergence is not guaranteed everywhere inside the domains reported in Tables 5 and 6. For the absorbers section, 2000 simulations have been designed each inside the nominal domain, the relaxed domain, and the extended domain. The resulting convergence rates are respectively, 97.35 %, 3.50 %, and 7.50 %. The converging simulations of the absorbing split are therefore only 2086 out of 6000. For the regeneration section, the convergence rate reported was lower than the absorbers thus two additional DoE runs have been added, as reported in Table 6 with the names *Extra range 1* and *Extra range 2*. In this case, 2000 simulations were designed for the nominal domain, 2000 more for the extended domain, 1000 for the first extra domain, and another 1000 for the second extra domain. The convergence rate reported is respectively 40.35 %, 12.50 %, 40.40 %, and 70.20 %. Globally, only 2203 simulations out of 6000 found a feasible solution. In Figs. 7 and 8, the resulting distributions of feature variables are presented while in Figs. 9 and 10, the distributions of target variables generated through the DoE are shown.

3.4. Surrogate model

The surrogate model creation starts with the training of the selected machine learning models described in Section 2.3. The training of such algorithms is performed using a 5-fold cross-validation method in order to select the best performing model which is then retrained on the entire CV data set without folding and the resulting error is evaluated on the test set, previously separated from the original whole data set, as described in Section 2.3.

3.4.1. Absorbers section results

The results for the absorbing section are reported in Table 8 and shown graphically in Fig. 11. For this case, the majority of variables have been modeled using the random forest method, since this algorithm outperformed all the others in the CV phase. A few notable exceptions are the pressure variables which are modeled using linear regression. This is reasonable since in the process simulation all the pressure drops of the units are linear, thus this method is able to perfectly interpolate that behavior. By looking at streams F19 and F8 in Table 8, it can be noted that the error found in the composition values of the hydrocarbon components is very similar. This is explained by the fact that during the data generation procedure discussed in Section 3.3 these variables are strongly correlated due to the ratio imposed as a constraint, as reported in Table 7.

To better understand why some particular models are under-performing the expectations it is necessary to look at the distribution of data reported in Figs. 7 and 9. For example, both temperature

Table 5
Nominal values and design of experiments domains for the absorbers section. Variable SR is defined as the split ratio of each splitter, SP1 and SP3. N_{Stages} is the number of stages of each absorbing unit.

Stream	Variable	UoM	Nominal value	Nominal range	Relaxed range	Extended range
F14	T	(°C)	40	36–44	24–56	10–80
F14	P	(bar)	117	105.3–128.7	70.2–163.8	50–200
F14	F	(kg/h)	2200	1980–2420	1320–3080	1–20 000
F14	x_{H_2O}	(–)	0.001	0.0009–0.0011	0–0.1	0–0.1
F14	x_{CH_4}	(–)	0.03	0.027–0.033	0.001–0.2	0.001–0.2
F14	x_{H_2S}	(–)	0.001	0.0009–0.0011	0–0.2	0–0.2
F24	T	(°C)	60	54–66	36–84	10–200
F24	P	(bar)	4.3	3.87–4.73	2.58–6.02	1–100
F24	F	(kg/h)	120	108–132	72–168	1–20 000
F24	x_{H_2O}	(–)	0.03	0.027–0.033	0–0.1	0–0.1
F24	x_{CH_4}	(–)	0.04	0.036–0.044	0.001–0.1	0.001–0.1
F24	x_{H_2S}	(–)	0.005	0.0045–0.0055	0–0.2	0–0.2
F4	T	(°C)	55	49.5–60.5	33–77	1–15 000
F4	F	(kg/h)	504.3	453.87–554.73	302.58–706.02	0–1
F7	T	(°C)	55	49.5–60.5	33–77	1–15 000
F7	F	(kg/h)	1568	1411.2–1724.8	940.8–2195.2	5–50
F4, F7	x_{DEA}	(–)	0.16	0.144–0.176	0.1–0.3	10–200
F4, F7	x_{H_2S}	(–)	0.0002	0.00018–0.00022	0–0.2	10–200
SP1, SP3	SR	(–)	0	0–1	0–1	0.1–0.3
T-504, T-505	N_{Stages}	(–)	20	18–22	5–50	0–0.2

Table 6
Nominal values and design of experiments domains for the regeneration section. Variable α represents the vapor fraction, $x_{C_1-C_5,\text{tot}}$ is the total content of light hydrocarbons, and R_{BoilUp} is the boil up ratio of the column.

Tag	Variable	UoM	Nominal value	Nominal range	Extended range	Extra range 1	Extra range 2
F9c	T	(°C)	95	57–133	30–300	1890–2310	1890–2310
F9c	P	(bar)	2.9	1.74–4.06	1–300	85.5–104.5	85.5–104.5
F9c	F	(kg/h)	2100	1260–2940	1–20 000	2.61–3.19	2.61–3.19
F9c	$x_{C_1-C_5,\text{tot}}$	(–)	0	–	0–0.1	0–0.1	–
F9c	x_{H_2S}	(–)	0.01	0.001–0.1	0.001–0.2	0.001–0.2	0.001–0.1
F9c	x_{H_2}	(–)	0	–	0–0.1	0–0.1	–
F9c	x_{DEA}	(–)	0.16	0.1–0.2	0.1–0.3	0.1–0.3	0.1–0.2
F11	α	(–)	0.073	0.005–0.5	0.005–1	0.005–1	0.005–0.5
T-506	N_{Stages}	(–)	20	10–30	10–50	10–50	10–30
T-506	R_{BoilUp}	(–)	1.4	1–2	0.5–3	0.5–3	1–2

Table 7
Molar ratio of the hydrocarbon concentration with respect to methane in streams F14, entering the HP absorber, and F24, entering the LP absorber. The ratio is calculated as x_i/x_{CH_4}.

Component	F14	F24
Methane	1	1
Ethane	0.2717	0.6430
Propane	0.1288	0.4441
i-Butane	0.0405	0.1708
n-Butane	0.0518	0.2460
n-Pentane	0.2898	2.3332

variables of streams F9b and F8 show (see Fig. 11) an unusually straight vertical cloud of points in the middle of the parity plot. These vertical conglomerates mirror exactly the distribution of data found in Fig. 9. In fact, around that value it is possible to find a dense cluster of numerous points. It is plausible that such skewed distribution of data is introducing an unwanted bias during the training of the models.

The worst-performing model of Table 8, which is the random forest of the flowrate variable of stream F9b, is not performing well in the test set, and is also near to the nominal value, i.e., the range in which the sample points are more numerous. This unsatisfactory performance is possibly originating from the data samples of higher magnitude, found on the far right of the distribution (see Fig. 9), which could have a bias effect in shifting the error reduction during training at higher absolute values.

3.4.2. Regenerator section results

The results concerning the regeneration sub-process are shown in Table 9 and in Fig. 12. In this case, it is possible to note that several target variables have an R^2 coefficient which is highly negative, indicating that there is no correlation between the data and the models. This is true for most of the compositions of stream F10b (see Table 9), especially the hydrocarbon molecules. The reason for this behavior can be better understood by looking at stream's F10b compositions in Fig. 12. Indeed the ground truth values (on the x-axes) are actually zero while the models predict values which are non-zero but yet very small, in the order of 10^{-26} to 10^{-30}. Moreover, from a physical point of view, stream F10b is the outlet at the bottom of the regenerator and it should contain only water and DEA, thus the negligible amount of other components is explaining the extremely low values of the MAE and RMSE. On the other hand, it is possible to see huge MAPE values for H_2 and H_2O compositions of stream F12 even though the other metrics are indicating a fairly good fit. This condition is signaling the presence of errors where the truth value is very low, or maybe even zero (Pedregosa et al., 2011), and the prediction has a higher order of magnitude (for example the same order of magnitude of the MAE or RMSE), see Eq. (5). In fact, by inspecting the parity plot of variable x_{H_2} of stream F12 it is possible to find a small cluster of points in the bottom left of the graph, where the true values are closer to zero. Again, as in the case of the absorbing section described above, the metrics of the hydrocarbon compositions are very close to each other for stream F12, since they are highly correlated, see Table 7.

4. Conclusions

Complex process simulations or digital twins have inherent numerical difficulties in terms of computational time required to obtain a solution plus the feasibility of the solution, or convergence, is often not guaranteed in an extended domain around the design specification

Table 8
Performance results on the test set of the best models selected after the cross-validation step and retrained on the entire cross-validation set.

Tag	Variable	UoM	Best model	Nominal value	R^2	RMSE (UoM)	MAE (UoM)	MAPE (%)
F9b	T	(°C)	Linear Regression	5.37e+01	0.59	1.12e+0	8.51e−1	1.584
F9b	P	(bar)	Linear Regression	4.30e+00	1.00	8.49e−16	7.20e−16	0.000
F9b	F	(kg/h)	Random Forest	2.10e+03	0.23	8.29e+1	6.86e+1	3.306
F9b	x_{DEA}	(−)	Random Forest	1.58e−01	0.45	6.48e−3	5.27e−3	3.333
F9b	x_{H_2}	(−)	Random Forest	4.14e−05	1.00	1.31e−7	1.04e−7	0.248
F9b	x_{H_2S}	(−)	Random Forest	1.07e−02	0.98	3.12e−4	2.39e−4	6.239
F9b	x_{H_2O}	(−)	Random Forest	8.31e−01	0.47	6.56e−3	5.33e−3	0.637
F9b	x_{CH_4}	(−)	Random Forest	5.80e−06	0.92	1.37e−7	1.11e−7	1.838
F9b	x_{Ethane}	(−)	Random Forest	6.57e−06	0.82	2.63e−7	2.13e−7	3.118
F9b	$x_{Propane}$	(−)	Random Forest	4.73e−06	0.76	3.31e−7	2.67e−7	3.753
F9b	$x_{i-Butane}$	(−)	Random Forest	1.95e−06	0.71	6.12e−7	4.95e−7	4.562
F9b	$x_{n-Butane}$	(−)	Random Forest	2.77e−06	0.80	2.16e−7	1.74e−7	3.268
F9b	$x_{Pentane}$	(−)	Random Forest	2.44e−05	0.66	4.56e−7	3.68e−7	5.254
F8	T	(°C)	Linear Regression	5.50e+01	0.67	1.55e+0	1.18e+0	2.142
F8	P	(bar)	Linear Regression	4.30e+00	1.00	8.49e−16	7.20e−16	0.000
F8	F	(kg/h)	Random Forest	1.18e+02	0.99	5.49e−1	4.03e−1	0.340
F8	x_{DEA}	(−)	Random Forest	1.01e+01	0.65	6.58e−8	4.81e−8	12.183
F8	x_{H_2}	(−)	Polynomial Order 2	7.84e−01	0.97	1.97e−3	1.48e−3	0.185
F8	x_{H_2S}	(−)	Random Forest	3.75e−06	0.70	4.52e−7	3.19e−7	13.423
F8	x_{H_2O}	(−)	Random Forest	3.03e−02	0.70	2.39e−3	1.79e−3	5.810
F8	x_{CH_4}	(−)	Random Forest	3.85e−02	1.00	1.02e−4	7.58e−5	0.188
F8	x_{Ethane}	(−)	Random Forest	2.47e−02	1.00	6.53e−5	4.88e−5	0.188
F8	$x_{Propane}$	(−)	Random Forest	1.70e−02	1.00	6.52e−5	4.87e−5	0.188
F8	$x_{i-Butane}$	(−)	Random Forest	9.43e−03	1.00	6.53e−5	4.87e−5	0.188
F8	$x_{n-Butane}$	(−)	Random Forest	6.55e−03	1.00	6.51e−5	4.86e−5	0.188
F8	$x_{Pentane}$	(−)	Random Forest	8.94e−02	1.00	6.52e−5	4.87e−5	0.188
F19	T	(°C)	Linear Regression	4.31e+01	0.84	1.09e+0	8.59e−1	1.967
F19	P	(bar)	Linear Regression	1.17e+02	1.00	9.66e−15	6.16e−15	0.000
F19	F	(kg/h)	Random Forest	2.17e+03	0.84	1.82e+2	1.33e+2	9.719
F19	x_{DEA}	(−)	Random Forest	8.30e−09	0.87	1.08e−9	7.63e−10	158.836
F19	x_{H_2}	(−)	Polynomial Order 2	9.44e−01	1.00	8.75e−5	6.23e−5	0.007
F19	x_{H_2S}	(−)	Random Forest	1.39e−07	0.91	8.63e−5	6.08e−5	50.986
F19	x_{H_2O}	(−)	Random Forest	7.50e−04	0.98	1.50e−5	1.09e−5	1.323
F19	x_{CH_4}	(−)	Polynomial Order 2	3.08e−02	1.00	2.74e−6	2.01e−6	0.007
F19	x_{Ethane}	(−)	Polynomial Order 2	8.37e−03	1.00	7.33e−7	5.44e−7	0.007
F19	$x_{Propane}$	(−)	Polynomial Order 2	3.97e−03	1.00	7.35e−7	5.45e−7	0.007
F19	$x_{i-Butane}$	(−)	Polynomial Order 2	1.60e−03	1.00	7.28e−7	5.48e−7	0.007
F19	$x_{n-Butane}$	(−)	Polynomial Order 2	1.25e−03	1.00	7.45e−7	5.47e−7	0.007
F19	$x_{Pentane}$	(−)	Polynomial Order 2	8.93e−03	1.00	7.39e−7	5.49e−7	0.007

Table 9
Performance results on the test set of the best models selected after the cross-validation step and retrained on the entire cross-validation set for the regeneration section.

Tag	Variable	UoM	Best model	Nominal value	R^2 (1)	RMSE (UoM)	MAE (UoM)	MAPE (%)
F12	T	(°C)	Random Forest	1.23E+02	0.68	3.37e+0	2.70e+0	2.235
F12	P	(bar)	Linear Regression	2.49E+00	1.00	2.24e−15	1.64e−15	0.000
F12	F	(kg/h)	SVR	1.10E+02	0.76	9.69e+1	7.90e+1	21.434
F12	x_{DEA}	(−)	Random Forest	0.00E+00	1.00	0.00e+0	0.00e+0	0.000
F12	x_{H_2}	(−)	Random Forest	0.00E+00	0.94	1.31e−2	6.07e−3	3.25e+10
F12	x_{H_2S}	(−)	SVR	1.19E−01	0.43	7.05e−2	5.81e−2	8.168
F12	x_{H_2O}	(−)	SVR	8.81E−01	0.91	2.74e−3	1.67e−3	2.11e+14
F12	x_{CH_4}	(−)	Random Forest	0.00E+00	0.95	2.09e−3	9.21e−4	8.333
F12	x_{Ethane}	(−)	Random Forest	0.00E+00	0.95	2.06e−3	9.26e−4	11.406
F12	$x_{Propane}$	(−)	Random Forest	0.00E+00	0.95	2.05e−3	9.06e−4	8.851
F12	$x_{i-Butane}$	(−)	Random Forest	0.00E+00	0.95	2.03e−3	9.12e−4	14.519
F12	$x_{n-Butane}$	(−)	Random Forest	0.00E+00	0.95	2.02e−3	8.82e−4	9.509
F12	$x_{Pentane}$	(−)	Random Forest	0.00E+00	0.95	2.06e−3	9.12e−4	11.965
F10b	T	(°C)	Polynomial Order 2	1.35E+02	0.90	2.26e+0	1.75e+0	1.214
F10b	P	(bar)	Linear Regression	2.50E+00	1.00	1.96e−15	1.51e−15	0.000
F10b	F	(kg/h)	Polynomial Order 2	1.98E+03	0.90	1.04e+2	8.54e+1	5.715
F10b	x_{DEA}	(−)	Polynomial Order 2	1.76E−01	0.76	3.70e−2	2.86e−2	11.536
F10b	x_{H_2}	(−)	Random Forest	0.00E+00	−2.38e+7	1.51e−32	1.70e−33	0.000
F10b	x_{H_2S}	(−)	SVR	0.00E+00	−0.05	1.25e−3	4.12e−4	1.32e+6
F10b	x_{H_2O}	(−)	Polynomial Order 2	8.24E−01	0.90	2.26e+0	1.75e+0	1.214
F10b	x_{CH_4}	(−)	Decision Tree	0.00E+00	−1.10e+7	3.35e−30	5.47e−31	0.000
F10b	x_{Ethane}	(−)	SVR	0.00E+00	−7.17e+5	1.44e−27	1.22e−27	0.000
F10b	$x_{Propane}$	(−)	Decision Tree	0.00E+00	−1.50e+8	1.17e−29	2.05e−30	0.000
F10b	$x_{i-Butane}$	(−)	Decision Tree	0.00E+00	−4.21e+6	2.79e−27	2.76e−27	0.000
F10b	$x_{n-Butane}$	(−)	Decision Tree	0.00E+00	−5.90e+8	3.76e−29	1.91e−29	0.000
F10b	$x_{Pentane}$	(−)	Gradient Boosting	0.00E+00	−3.72e+7	3.12e−30	4.27e−31	0.000

of the process. Black box surrogate models can overcome this problem by using metamodels which are inherently continuous and defined in a much broader domain and, except for the training phase of the model, do not require an iterative complicated numerical solution. The models that respect these claims are the typical ones found in machine learning applications. In particular, linear regression, higher

order polynomial regression, support vector regression (SVR), decision tree regression, random forest, AdaBoost, and gradient boosting. By developing a surrogate model using these models it is possible to more easily conduct computationally heavier operations, like, e.g., process optimization.

In this work, the amine scrubbing plant of Itelyum Regeneration S.p.A. in Pieve Fissiraga has been simulated using the Aspen HYSYS process simulator and validated as a digital twin against real process data of a steady-state operation, taken from the DCS. Then, a surrogate model has been developed on top of the digital twin. In several cases, the results obtained, in terms of residual error between the surrogate model and the digital twin data of the target variables, showed a great performance. In a few cases, the performance was poorer. These results show that extreme caution must be put into the design of experiments phase and the definition of the black box boundaries because it is important to sample the operational domain with enough points in order to give the models the possibility to interpret the behavior of the target variables more accurately. However, the intrinsic dimensionality of the problem is a burden since a black box that contains too many features grows exponentially in the dimension of the sample set required to obtain enough information on the system. In other words, the more a process grows in dimension and complexity (number of units, streams, recycles, complex thermodynamics, nonlinear units, etc.) the more difficult it becomes to create an accurate surrogate model with fewer data. In any case, the methods and the framework proposed in this work can be adapted to surrogate process simulation black boxes of any dimension and of any kind of digital twin. However, it must be noted that with higher problem complexity, and especially, dimensionality, the data sampling procedure needs to scale exponentially. Thus, a compromise should be found between the problem complexity and dimensionality and the capacity of generating enough sample data.

CRediT authorship contribution statement

Andrea Galeazzi: Conceptualization, Methodology, Software, Investigation, Writing – original draft, Writing – review & editing. **Kristiano Prifti:** Methodology, Writing – original draft, Writing – review & editing. **Carlo Cortellini:** Software, Investigation, Writing – original draft, Writing – review & editing. **Alessandro Di Pretoro:** Conceptualization, Methodology, Investigation, Writing – original draft, Writing – review & editing. **Francesco Gallo:** Conceptualization, Data acquisition, Writing – original draft, Writing – review & editing. **Flavio Manenti:** Conceptualization, Methodology, Supervision, Writing – original draft, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The data that has been used is confidential

References

Agarwal, A., Biegler, L.T., 2013. A trust-region framework for constrained optimization using reduced order modeling. Opt. Eng. 14 (1), 3–35. http://dx.doi.org/10.1007/s11081-011-9164-0.

Asprion, N., Boettcher, R., Pack, R., Stavrou, M.-E., Hoeller, J., Schwientek, J., Bortz, M., 2019. Gray-box modeling for the optimization of chemical processes. Chem. Ing. Tech. 91 (3), 305–313. http://dx.doi.org/10.1002/cite.201800086.

Awad, M., Khanna, R., 2015. Support vector regression. In: Awad, M., Khanna, R. (Eds.), Efficient Learning Machines: Theories, Concepts, and Applications for Engineers and System Designers. A Press, Berkeley, CA, pp. 67–80. http://dx.doi.org/10.1007/978-1-4302-5990-9_4.

Bajaj, I., Iyer, S.S., Faruque Hasan, M.M., 2018. A trust region-based two phase algorithm for constrained black-box and grey-box optimization with infeasible initial point. Comput. Chem. Eng. 116, 306–321. http://dx.doi.org/10.1016/j.compchemeng.2017.12.011.

Barton, R.R., 1992. Metamodels for simulation input-output relations. In: Proceedings of the 24th Conference on Winter Simulation - WSC '92. ACM Press, Arlington, Virginia, United States, pp. 289–299. http://dx.doi.org/10.1145/167293.167352.

Bevilacqua, M., Bottani, E., Ciarapica, F.E., Costantino, F., Di Donato, L., Ferraro, A., Mazzuto, G., Monteriù, A., Nardini, G., Ortenzi, M., Paroncini, M., Pirozzi, M., Prist, M., Quatrini, E., Tronci, M., Vignali, G., 2020. Digital twin reference model development to prevent operators' risk in process plants. Sustainability 12 (3), 1088. http://dx.doi.org/10.3390/su12031088.

Bhattacharyya, B., 2018. A critical appraisal of design of experiments for uncertainty quantification. Arch. Comput. Methods Eng. 25 (3), 727–751. http://dx.doi.org/10.1007/s11831-017-9211-x.

Bhosekar, A., Ierapetritou, M., 2018. Advances in surrogate based modeling, feasibility analysis, and optimization: A review. Comput. Chem. Eng. 108, 250–267. http://dx.doi.org/10.1016/j.compchemeng.2017.09.017.

Bishop, C.M., 2008. Pattern Recognition and Machine Learning (Information Science and Statistics). Springer.

Chai, T., Draxler, R.R., 2014. Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature. Geosci. Model Dev. 7 (3), 1247–1250. http://dx.doi.org/10.5194/gmd-7-1247-2014.

Cioppa, T.M., Lucas, T.W., 2007. Efficient nearly orthogonal and space-filling latin hypercubes. Technometrics 49 (1), 45–55. http://dx.doi.org/10.1198/004017006000000453.

Damblin, G., Couplet, M., Iooss, B., 2013. Numerical studies of space-filling designs: Optimization of Latin hypercube samples and subprojection properties. J. Simul. 7 (4), 276–289. http://dx.doi.org/10.1057/jos.2013.16.

Damiani, L., Demartini, M., Giribone, P., Maggiani, M., Revetria, R., 2018. Simulation and Digital Twin Based Design of a Production Line: A Case Study. Hong Kong, p. 5.

Donovan, D., Burrage, K., Burrage, P., McCourt, T.A., Thompson, B., Yazici, E.Ş., 2018. Estimates of the coverage of parameter space by Latin hypercube and orthogonal array-based sampling. Appl. Math. Model. 57, 553–564. http://dx.doi.org/10.1016/j.apm.2017.11.036.

Dyment, J., Watanasiri, S., 2015. Acid Gas Cleaning Using DEPG Physical Solvents: Validation with Experimental and Plant Data. Aspen Technology Inc., Bedford, MA, USA.

Errandonea, I., Beltrán, S., Arrizabalaga, S., 2020. Digital twin for maintenance: A literature review. Comput. Ind. 123, 103316. http://dx.doi.org/10.1016/j.compind.2020.103316.

Freund, Y., Schapire, R.E., 1995. A desicion-theoretic generalization of on-line learning and an application to boosting. In: Vitányi, P. (Ed.), Computational Learning Theory. In: Lecture Notes in Computer Science, Springer, Berlin, Heidelberg, pp. 23–37. http://dx.doi.org/10.1007/3-540-59119-2_166.

Friedman, J.H., 2001. Greedy function approximation: A gradient boosting machine. Ann. Statist. 29 (5), 1189–1232.

Friedman, J.H., 2002. Stochastic gradient boosting. Comput. Statist. Data Anal. 38 (4), 367–378. http://dx.doi.org/10.1016/S0167-9473(01)00065-2.

Guo, W., Sun, Z., Vilsen, S.B., Meng, J., Stroe, D.I., 2022. Review of "grey box" lifetime modeling for lithium-ion battery: Combining physics and data-driven methods. J. Energy Storage 56 (A), 105992. http://dx.doi.org/10.1016/j.est.2022.105992.

Hao, J., Ho, T.K., 2019. Machine learning made easy: A review of scikit-learn package in python programming language. J. Educ. Behav. Stat. 44 (3), 348–361. http://dx.doi.org/10.3102/1076998619832248.

Helton, J.C., Davis, F.J., 2003. Latin hypercube sampling and the propagation of uncertainty in analyses of complex systems. Reliab. Eng. Syst. Saf. 81 (1), 23–69. http://dx.doi.org/10.1016/S0951-8320(03)00058-9.

Ho, T.K., 1995. Random decision forests. In: Proceedings of 3rd International Conference on Document Analysis and Recognition, Vol. 1. IEEE, pp. 278–282.

Ho, T.K., 1998. The random subspace method for constructing decision forests. IEEE Trans. Pattern Anal. Mach. Intell. 20 (8), 832–844. http://dx.doi.org/10.1109/34.709601.

Hu, L., Nguyen, N.-T., Tao, W., Leu, M.C., Liu, X.F., Shahriar, M.R., Al Sunny, S.M.N., 2018. Modeling of cloud-based digital twins for smart manufacturing with MT connect. Procedia Manuf. 26, 1193–1203. http://dx.doi.org/10.1016/j.promfg.2018.07.155.

Jeon, S.M., Schuesslbauer, S., 2020. Digital twin application for production optimization. In: 2020 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM). pp. 542–545. http://dx.doi.org/10.1109/IEEM45057.2020.9309874.

Jiang, T., Gradus, J.L., Rosellini, A.J., 2020. Supervised machine learning: A brief primer. Behav. Ther. 51 (5), 675–687. http://dx.doi.org/10.1016/j.beth.2020.05.002.

Kannapinn, M., Pham, M.K., Schäfer, M., 2022. Physics-based digital twins for autonomous thermal food processing: Efficient, non-intrusive reduced-order modeling. Innov. Food Sci. Emerg. Technol. 81, 103143. http://dx.doi.org/10.1016/j.ifset.2022.103143.

Karunasingha, D.S.K., 2022. Root mean square error or mean absolute error? Use their ratio as well. Inform. Sci. 585, 609–629. http://dx.doi.org/10.1016/j.ins.2021.11.036.

Kleijnen, J.P., 2010. Design and analysis of computational experiments: Overview. In: Bartz-Beielstein, T., Chiarandini, M., Paquete, L., Preuss, M. (Eds.), Experimental Methods for the Analysis of Optimization Algorithms. Springer, Berlin, Heidelberg, pp. 51–72. http://dx.doi.org/10.1007/978-3-642-02538-9_3.

Kleijnen, J.P.C., 2015. Kriging metamodels and their designs. In: Kleijnen, J.P. (Ed.), Design and Analysis of Simulation Experiments. In: International Series in Operations Research & Management Science, Springer International Publishing, Cham, pp. 179–239. http://dx.doi.org/10.1007/978-3-319-18087-8_5.

Kohl, A.L., Nielsen, R.B., 1997. Chapter 14 - physical solvents for acid gas removal. In: Kohl, A.L., Nielsen, R.B. (Eds.), Gas Purification, fifth ed. Gulf Professional Publishing, Houston, pp. 1187–1237. http://dx.doi.org/10.1016/B978-088415220-0/50014-8.

Kritzinger, W., Karner, M., Traar, G., Henjes, J., Sihn, W., 2018. Digital twin in manufacturing: A categorical literature review and classification. IFAC-PapersOnLine 51 (11), 1016–1022. http://dx.doi.org/10.1016/j.ifacol.2018.08.474.

Kvalseth, T.O., 1985. Cautionary note about R2. Amer. Statist. 39 (4), 279–285. http://dx.doi.org/10.2307/2683704.

Lassila, T., Manzoni, A., Quarteroni, A., Rozza, G., 2014. Model order reduction in fluid dynamics: Challenges and perspectives. In: Quarteroni, A., Rozza, G. (Eds.), Reduced Order Methods for Modeling and Computational Reduction. In: MS&A - Modeling, Simulation and Applications, Springer International Publishing, Cham, pp. 235–273.

Ledolter, J., Kardon, R.H., 2020. Focus on data: Statistical design of experiments and sample size selection using power analysis. Invest. Ophthalmol. Vis. Sci. 61 (8), 11. http://dx.doi.org/10.1167/iovs.61.8.11.

Li, W., Lu, L., Xie, X., Yang, M., 2017. A novel extension algorithm for optimized Latin hypercube sampling. J. Stat. Comput. Simul. 87 (13), 2549–2559. http://dx.doi.org/10.1080/00949655.2017.1340475.

Liu, M., Fang, S., Dong, H., Xu, C., 2021. Review of digital twin about concepts, technologies, and industrial applications. J. Manuf. Syst. 58, 346–361. http://dx.doi.org/10.1016/j.jmsy.2020.06.017.

Liu, H., Ong, Y.-S., Cai, J., 2018. A survey of adaptive sampling for global metamodeling in support of simulation-based complex engineering design. Struct. Multidiscip. Optim. 57 (1), 393–416. http://dx.doi.org/10.1007/s00158-017-1739-8.

Liu, Z., Yang, M., Li, W., 2016. A sequential latin hypercube sampling method for metamodeling. In: Zhang, L., Song, X., Wu, Y. (Eds.), Theory, Methodology, Tools and Applications for Modeling and Simulation of Complex Systems. Springer, Singapore, pp. 176–185. http://dx.doi.org/10.1007/978-981-10-2663-8_19.

Loh, W.-L., 1996. On Latin hypercube sampling. Ann. Statist. 24 (5), 2058–2080. http://dx.doi.org/10.1214/aos/1069362310.

Macchi, M., Roda, I., Negri, E., Fumagalli, L., 2018. Exploring the role of digital twin for asset lifecycle management. IFAC-PapersOnLine 51 (11), 790–795. http://dx.doi.org/10.1016/j.ifacol.2018.08.415.

Manteufel, R., 2000. Evaluating the convergence of Latin hypercube sampling. In: 41st Structures, Structural Dynamics, and Materials Conference and Exhibit. American Institute of Aeronautics and Astronautics, Atlanta, GA, U.S.A., http://dx.doi.org/10.2514/6.2000-1636.

McKay, M.D., Beckman, R.J., Conover, W.J., 1979. A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. Technometrics 21 (2), 239–245. http://dx.doi.org/10.2307/1268522.

Miles, J., 2014. R Squared, Adjusted R Squared. In: Wiley StatsRef: Statistics Reference Online, John Wiley & Sons, Ltd, http://dx.doi.org/10.1002/9781118445112.stat06627.

Minana, J.A.G., Schieppati, R., Dalla Giovanna, F., 1994. Process to re-refine used oils.

Morris, M.D., Mitchell, T.J., 1995. Exploratory designs for computational experiments. J. Statist. Plann. Inference 43 (3), 381–402. http://dx.doi.org/10.1016/0378-3758(94)00035-T.

Navid, A., Khalilarya, S., Abbasi, M., 2018. Diesel engine optimization with multi-objective performance characteristics by non-evolutionary Nelder-Mead algorithm: Sobol sequence and Latin hypercube sampling methods comparison in DoE process. Fuel 228, 349–367. http://dx.doi.org/10.1016/j.fuel.2018.04.142.

Panwar, P., Michael, P., 2018. Empirical modelling of hydraulic pumps and motors based upon the Latin hypercube sampling method. Int. J. Hydromechatron..

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, É., 2011. Scikit-learn: Machine learning in Python. J. Mach. Learn. Res. 12 (85), 2825–2830.

Pedrozo, H.A., Rodriguez Reartes, S.B., Bernal, D.E., Vecchietti, A.R., Diaz, M.S., Grossmann, I.E., 2021. Hybrid model generation for superstructure optimization with generalized disjunctive programming. Comput. Chem. Eng. 154, 107473. http://dx.doi.org/10.1016/j.compchemeng.2021.107473.

Peng, D.-Y., Robinson, D.B., 1976. A new two-constant equation of state. Ind. Eng. Chem. Fundam. 15 (1), 59–64. http://dx.doi.org/10.1021/i160057a011.

Pronzato, L., Müller, W.G., 2012. Design of computer experiments: Space filling and beyond. Stat. Comput. 22 (3), 681–701. http://dx.doi.org/10.1007/s11222-011-9242-3.

Qian, P.Z.G., 2012. Sliced Latin hypercube designs. J. Amer. Statist. Assoc. 107 (497), 393–399. http://dx.doi.org/10.1080/01621459.2011.644132.

Rajulapati, L., Chinta, S., Shyamala, B., Rengaswamy, R., 2022. Integration of machine learning and first principles models. AIChE J. 68 (6), e17715. http://dx.doi.org/10.1002/aic.17715.

Sanchez, S.M., Wan, H., 2015. Work smarter, not harder: A tutorial on designing and conducting simulation experiments. In: 2015 Winter Simulation Conference (WSC). pp. 1795–1809. http://dx.doi.org/10.1109/WSC.2015.7408296.

Sansana, J., Joswiak, M.N., Castillo, I., Wang, Z., Rendall, R., Chiang, L.H., Reis, M.S., 2021. Recent trends on hybrid modeling for industry 4.0. Comput. Chem. Eng. 151, 107365. http://dx.doi.org/10.1016/j.compchemeng.2021.107365.

Sheikholeslami, R., Razavi, S., 2017. Progressive Latin hypercube sampling: An efficient approach for Robust sampling-based analysis of environmental models. Environ. Model. Softw. 93, 109–126. http://dx.doi.org/10.1016/j.envsoft.2017.03.010.

Shokry, A., Baraldi, P., Zio, E., Espuña, A., 2020. Dynamic surrogate modeling for multistep-ahead prediction of multivariate nonlinear chemical processes. Ind. Eng. Chem. Res. 59 (35), 15634–15655. http://dx.doi.org/10.1021/acs.iecr.0c00729.

Sobol', I.M., 1967. On the distribution of points in a cube and the approximate evaluation of integrals. USSR Comput. Math. Math. Phys. 7 (4), 86–112. http://dx.doi.org/10.1016/0041-5553(67)90144-9.

Song, Y., Chen, C.-C., 2009. Symmetric nonrandom two-liquid segment activity coefficient model for electrolytes. Ind. Eng. Chem. Res. 48 (11), 5522–5529. http://dx.doi.org/10.1021/ie900006g.

Thebelt, A., Wiebe, J., Kronqvist, J., Tsay, C., Misener, R., 2022. Maximizing information from chemical engineering data sets: Applications to machine learning. Chem. Eng. Sci. 252, http://dx.doi.org/10.1016/j.ces.2022.117469.

Tian, Y., Demirel, S.E., Hasan, M.M.F., Pistikopoulos, E.N., 2018. An overview of process systems engineering approaches for process intensification: State of the art. Chem. Eng. Process. - Process Intensif. 133, 160–210. http://dx.doi.org/10.1016/j.cep.2018.07.014.

Van Rossum, G., Drake, F.L., 2009. Python 3 Reference Manual. CreateSpace, Scotts Valley, CA.

VanDerHorn, E., Mahadevan, S., 2021. Digital twin: Generalization, characterization and implementation. Decis. Support Syst. 145, 113524. http://dx.doi.org/10.1016/j.dss.2021.113524.

Viana, F.A., 2013. Things you wanted to know about the Latin hypercube design and were afraid to ask. In: 10th World Congress on Structural and Multidisciplinary Optimization, Vol. 19. sn.

Viana, F.A.C., 2016. A tutorial on Latin hypercube design of experiments. Qual. Reliab. Eng. Int. 32 (5), 1975–1985. http://dx.doi.org/10.1002/qre.1924.

von Stosch, M., Oliveira, R., Peres, J., de Azevedo, S.F., 2014. Hybrid semi-parametric modeling in process systems engineering: Past, present and future. Comput. Chem. Eng. 60, 86–101. http://dx.doi.org/10.1016/j.compchemeng.2013.08.008.

Vořechovský, M., Mašek, J., 2020. Distance-based optimal sampling in a hypercube: Energy potentials for high-dimensional and low-saturation designs. Adv. Eng. Softw. 149, 102880. http://dx.doi.org/10.1016/j.advengsoft.2020.102880.

Westermann, P., Evins, R., 2019. Surrogate modelling for sustainable building design – A review. Energy Build. 198, 170–186. http://dx.doi.org/10.1016/j.enbuild.2019.05.057.

Willmott, C.J., Matsuura, K., 2005. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. Clim. Res. 30 (1), 79–82. http://dx.doi.org/10.3354/cr030079.

Willmott, C.J., Matsuura, K., Robeson, S.M., 2009. Ambiguities inherent in sums-of-squares-based error statistics. Atmos. Environ. 43 (3), 749–752. http://dx.doi.org/10.1016/j.atmosenv.2008.10.005.

Xiong, Q., Jutan, A., 2002. Grey-box modelling and control of chemical processes. Chem. Eng. Sci. 57 (6), 1027–1039. http://dx.doi.org/10.1016/S0009-2509(01)00439-0.

Xu, J., Duan, X., Wang, Z., Yan, L., 2018. A general construction for nested Latin hypercube designs. Statist. Probab. Lett. 134, 134–140. http://dx.doi.org/10.1016/j.spl.2017.10.022.

Zendehboudi, S., Rezaei, N., Lohi, A., 2018. Applications of hybrid models in chemical, petroleum, and energy systems: A systematic review. Appl. Energy 228, 2539–2566. http://dx.doi.org/10.1016/j.apenergy.2018.06.051.

Zhao, Y., Jiang, C., Vega, M.A., Todd, M.D., Hu, Z., 2022. Surrogate modeling of nonlinear dynamic systems: A comparative study. J. Comput. Inf. Sci. Eng. 23 (1), http://dx.doi.org/10.1115/1.4054039.

Zhou, C., Xu, J., Miller-Hooks, E., Zhou, W., Chen, C.-H., Lee, L.H., Chew, E.P., Li, H., 2021. Analytics with digital-twinning: A decision support system for maintaining a resilient port. Decis. Support Syst. 143, 113496. http://dx.doi.org/10.1016/j.dss.2021.113496.

Zipper, H., Auris, F., Strahilov, A., Paul, M., 2018. Keeping the digital twin up-to-date – Process monitoring to identify changes in a plant. In: 2018 IEEE International Conference on Industrial Technology (ICIT). pp. 1592–1597. http://dx.doi.org/10.1109/ICIT.2018.8352419.