Explaining and Integrating Machine Learning Models with Rigorous Simulation

Jan C. Schöneberger*, Burcu Aker, and Armin Fricke

DOI: 10.1002/cite.202100089

Supporting Information available online

First-principle flowsheet simulation is a reliable data resource for training machine learning (ML) models for process industry, especially when plant data is not available. Process simulators play an even more important role for evaluation and validation of ML models. In this work, we present a workflow for building and evaluating ML models based on data generated by a commercial flowsheet simulator. The resulting hybrid models, combining data-driven predictions with mass and energy balances, have much lower calculation times than the rigorous models. The implementation of such models shows great potential for solving more complex process engineering problems on the high-dimensional space in the future, while saving the process engineer's time in the present.

Keywords: Artificial neuronal network, Flowsheet simulation, Machine learning, Pre-reformer reactor, Surrogate model

Received: May 31, 2021; *revised:* July 13, 2021; *accepted:* September 06, 2021

1 Introduction

One of the twelve hypothesis form the Tutzing symposium in 2018 states: "The full potential of digitalization in the process industries can only be unleashed by using artificial intelligence." [1]. In the subsequent hypothesis it is mentioned that a symbiosis between the engineer and the artificial intelligence must be established. A newer publication points out the importance of combining process simulation and artificial intelligence, namely data-driven models [2]. In this paper we focus on machine learning models (ML) as a synonym for data-driven models [3] and their integration with rigorous models. Our hypothesis is that engineers understand rigorous models. Therefore, we use them to explain data-driven models. This is a step towards the symbiosis mentioned above.

The training of machine learning models requires large amounts of informative data. Chollet and Allaire point out that the best solution to prevent a model from learning misleading or irrelevant patterns is to get more data [4]. In the process industry, a lot of data is available [5]. This data, e.g., from historians of the distributed control systems (DCS) [6], often is not very informative. Only a few different steady states of the processes are tracked over long periods of time, because operators in general desire a continuous and disturbance free operation. Transient data, which is collected during start-up and shut-down procedures or set-point changes, is influenced by too many parameters and physical effects. This makes it hard to extract reliable information, which, in addition, is time dependent. For its evaluation, the commonly unknown and largely varying dead times and delays of the measured process variables must be considered.

First-principle flowsheet simulation provides reliable information about chemical processes, assuming that these follow the fundamental rules of mass and energy conservation. Therefore, flowsheet simulators play an important role in the generation, evaluation, and validation of ML models for the process industry. They are used to generate data when data from the plant is not available. The results of ML models are tested against flowsheet simulation results to check their consistency with the named fundamental rules. When ML models are combined with flowsheet simulation to automatically follow these rules, these become so-called hybrid models.

In this work, we present the workflow to build and evaluate a hybrid ML model for the simulation of a pre-reformer reactor based on data generated by flowsheet simulation. A parameter study with Latin hypercube sampling (LHS) [7] is conducted to fill the ML model's input space. An artificial neuronal network (ANN) [8] is trained with this data and implemented as a unit operation in the flowsheet simulator,

Dr. Jan C. Schöneberger, Burcu Aker, Prof. Dr. Armin Fricke
js@capital-gain.eu
CGC Capital-Gain Consultants GmbH, Poststraße 12, 10178 Berlin, Germany.

making it a hybrid model. Sensitivity studies are used to compare the ANN predictions with the rigorous flowsheet calculations.

2 Pre-reformer Reactor Model

A pre-reformer reactor boosts the efficiency of a steam reformer plant and protects the catalyst in the primary reformer [9]. The pre-reformer reactor is allocated before the primary reformer. A simplified process flow diagram of the plant can be found in the Supporting Information (SI). The overall plant's energy efficiency increases because more of the waste gas stream's low-temperature heat is recovered. Furthermore, the inlet temperature for the primary reformer is stabilized and coking problems are shifted to the pre-reformer reactor. In this far smaller reactor, the catalyst can be replaced much easier and without shutting down the whole plant.

For modeling purposes, we focus on the feed gas composition, the steam-to-methane ratio (*SC*), the reactors inlet temperature (T_{in}) and pressure (P_{in}), and the reactor performance parameters, namely the approach temperatures (ATR_1 and ATR_2) and the pressure drop (dP). The model includes the six components nitrogen (N_2), hydrogen (H_2), methane (CH_4), carbon monoxide (CO), carbon dioxide (CO_2), and water (H_2O) and two reactions, the steam-methane reforming (Eq. (1)) and the water gas shift (Eq. (2)).

$$CH_4 + H_2O \rightleftharpoons CO + 3\ H_2 \tag{1}$$

$$CO + H_2O \rightleftharpoons CO_2 + H_2 \tag{2}$$

The model outputs are summarized by three values, the temperature change (d$T = T_{in} - T_{out}$), the extent of reaction 1 (RZ_1) and the extent of rection 2 (RZ_2). In addition, the outlet temperature (T_{out}), the outlet pressure (P_{out}) and the outlet gas composition can be calculated. The eleven resulting MESH equations (material balances, equilibrium conditions, sum conditions, and heat balance) and the equation for the pressure are given in Eqs. (3)–(7).

$$F_{in}x_{in,c} = F_{out}x_{out,c} + \sum_{r=1}^{2}\left(\nu_{r,c}RZ_r\right) \quad c = 1\ldots6 \tag{3}$$

$$\prod_{c=1}^{6} x_{out,c}^{\nu_{r,c}} = K_r(T_{out} + ATR_r) \quad r = 1,\ 2 \tag{4}$$

$$\sum_{c=1}^{6} x_{s,c} = 1 \quad s = in, out \tag{5}$$

$$F_{in}h_{in} = F_{out}h_{out} \tag{6}$$

$$P_{out} = P_{in} - \mathrm{d}P \tag{7}$$

We use temperature-dependent equilibrium constants (Eq. (8)) in this adiabatic model of a single-phase equilibrium reactor.

$$K_r = exp\left(-\frac{\Delta_R G^o}{RT}\right) \tag{8}$$

The numerically most challenging part in this model is introduced by the property model for the enthalpy (h). Following the previous work of Schmidt et al. [10] we use the Peng-Robinson equation of state (Eq. (9)) with classic mixing rules for a_M and b_M to calculate the departure enthalpy as a function of temperature and pressure (Eq. (10)), see e.g., Gmehling et al. for details [11].

$$P = \frac{RT}{\nu - b_M} - \frac{a_M(T)}{\nu^2 + 2b_M\nu - b_M^2} \tag{9}$$

$$h - h_{ideal} = \int_{\infty}^{\nu}\left[\frac{RT}{\nu} - P\right]\mathrm{d}\nu \tag{10}$$

The integral on the right-hand side of Eq. (10) can be solved analytically using the pressure explicit Peng-Robinson equation, but Eq. (9) must be solved numerically or by using Cardano's formulas to get the specific volume ν as a function of temperature and pressure. Fortunately, efficient solution algorithms for this purpose are implemented in most flowsheet simulators. Here we take the required pure component property data (critical data, acentric factors, ideal gas heat capacities, enthalpies, and entropies of formation) from the DIPPR database [12].

We have 12 MESH+P equations and 23 variables, giving a degree of freedom of 11. With a specified reactor feed (8 variables: F_{in}, $x_{in,c=1\ldots5}$, T_{in}, P_{in}) and specified reactor parameters (3 variables: ATR_1, ATR_2, dP) the equation system is fully specified and can be solved.

The operators of a steam reformer plant commonly cannot change the dry feed gas composition nor the pre-reformer reactor feed pressure, which is determined by the feed pressure of the primary reformer. They can change the steam-to-methane ratio (*SC*) and the reactor inlet temperature though. Therefore, the inlet mole fractions are converted to dry gas mole fractions (not manipulatable by operators) and a steam-to-methane ratio, see Eq. (11).

$$x_{in,c} = \frac{x_{dry,c}}{1 + x_{dry,CH_4}SC} \tag{11}$$

In an adiabatic equilibrium reactor, the reactor feed flowrate does not affect the relative reaction extents nor the outlet temperature. We set this variable to an arbitrary value of 100 mol h^{-1}. Tab. 1 lists the ranges of the remaining input variables.

Table 1. Lower and upper bounds of the input variables.

Parameter	Lower bound	Upper bound
x_{dry,CH_4}	0.44	0.98
$x_{dry,CO}$	0	0.12
x_{dry,H_2}	0	0.12
x_{dry,CO_2}	0	0.12
x_{dry,N_2}	0	0.2
T_{in} [°C]	350	600
P_{in} [bar-g]	10	50
dP [bar]	0	5
ATR_1 [K]	–50	50
ATR_2 [K]	–50	50
SC	1	3

3 Workflow

We use the Latin hypercube sampling method of the Dice Design package [13] in R-Studio [14] to generate 100 000 input data sets, each containing ten variables. Mole fractions of all six species are used to specify the range for the dry gas composition, see Tab. 1. This leads to a so-called mixture design problem. The six variables to be sampled are coupled by the sum equation. Buruk et al. [15] give a summary of common methods to tackle this problem. In the current case the six compositions must be sampled together with the five other variables. We solve this problem by sampling the five other variables with five dry gas flow rates while fixing the methane flowrate.

We use the commercial flowsheet simulator CHEMCAD 7 [16] to solve the model equations described in the previous chapter. The simulations are executed in parallel on a computer cluster using the software architecture presented by Fricke and Schöneberger [17].

Rhe 14-dimensional space (11 inputs, 3 outputs) is visualized in Microsoft PowerBI [18]. With the corrplot package [19] in R-Studio we analyze the linear variable dependencies and detect insensitive variables. The regression plots show that a linear analysis is not sufficient to describe the pre-reformer reactor's behavior. In a next step, an artificial neuronal network (ANN) is trained using the Keras library [20]. Keras only supports ANN's as ML models. Therefore, other ML models were not evaluated. The problem of selecting the best ML model for a specific problem is not addressed in this work. Furthermore, the ANN's hyper structure is not optimized. Instead, we compare different activation functions trained on the same ANN with one hidden layer and 100 neurons, as proposed by Schmidt et al. [10].

The trained ANN is used to calculate 4^{10} data sets based on a factorial design. With the factorial design the isolated impact of the individual input variables on the output variables can be visualized. By design, the analysis in PowerBI is limited to four data points per input dimension. But these four points are already helpful to analyze and understand the physical dependencies of the variables. The ANN predictions are a nonlinear interpolation of the original 100 000 data sets. We transfer the ANN back into the flowsheet simulator in order to analyze the quality of this interpolation at selected points. Then the the ANN is implemented as user added module (UAM) in CHEMCAD. Our software architecture allows implementing almost any ANN that was trained by using the Keras library if the ANN structure is not too complex. CHEMCAD's built-in analysis tool is used to compare predicted and rigorously calculated trends with sensitivity analysis and to find points of high deviation with the optimizer.

The ANN unit operation in CHEMCAD is a hybrid model that follows the material balances by predicting the reaction conversions instead of predicting the outlet composition. The energy balance is fulfilled too, because the implemented flash routines are used to calculate the reactor outlet temperature. This thermodynamically consistent temperature is compared with the temperature predicted by the ANN.

Fig. 1 shows the individual steps of the proposed workflow together with rough time estimates based on the presented case study.

4 Results

The presentation of the results follows the workflow introduced above. First, the data generation is evaluated. Then the data is analyzed with linear and nonlinear approaches. A unit operation is built with the ML models and used to compare the model predictions and to find the maximum deviation between the rigorous and one selected data-driven model.

4.1 Data Generation

Fig. 2 gives a clear visual impression of what a space filling design can do. The 11-D space is widely covered by the 100 000 sampled data sets. In comparison, the space covered by a factorial design with four points per dimension (1 048 576 data sets) looks particularly empty. The first five inputs show the interdependency of the dry gas molar fractions, which must sum up to one for each data set.

Samples generated with a space filling design rarely hit the bounds of the design variables. Using molar flow rates instead of a mixture design for the dry gas feed compositions worsens this situation slightly. Tab. 2 shows the ranges and the mean values of the sampled variables. With the

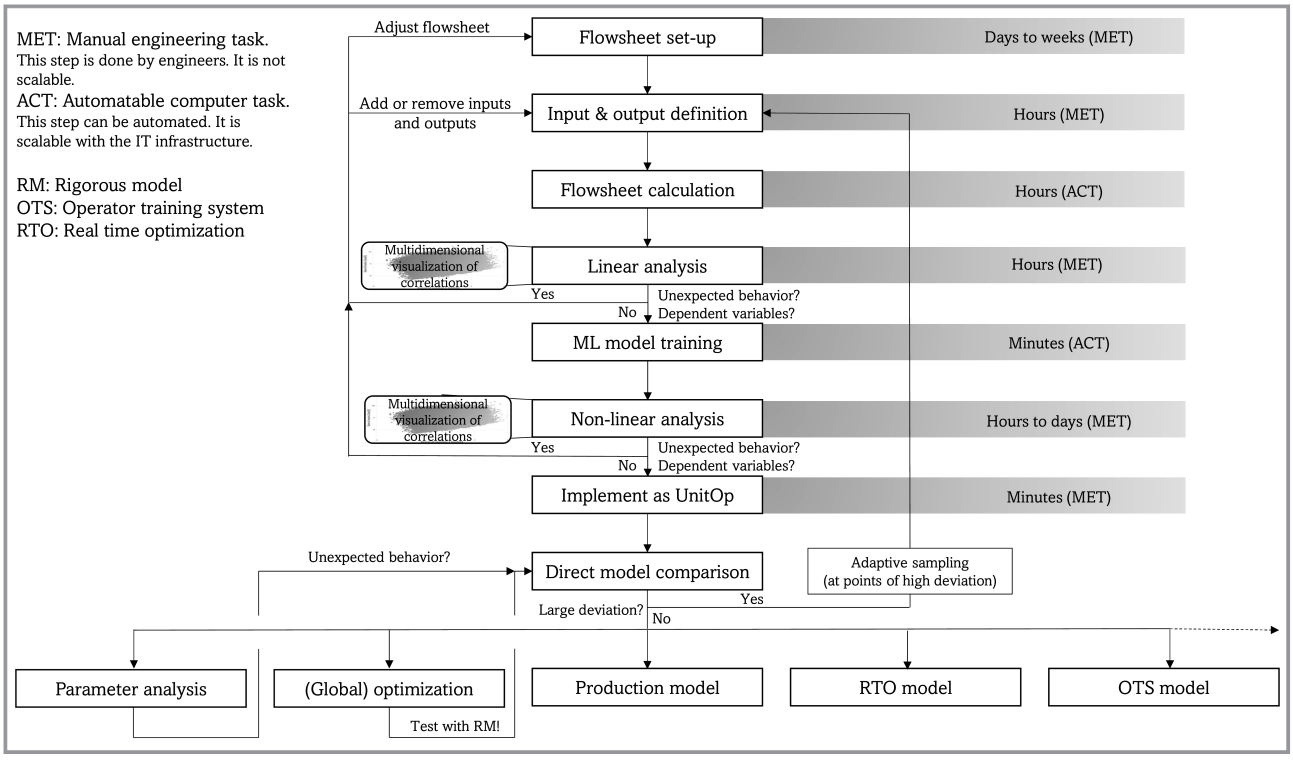


Figure 1. General workflow for building ML models from rigorous flowsheet simulations with time estimates for the individual steps.

ranges we selected for the molar flows some upper bounds for the dry gas compositions are exceeded. A more sophisticated sampling method is required if maintaining these bounds is crucial. Lueg et al. [21] present a solution for this problem. They developed and compared different sampling strategies for problems with variables that are partially constrained. With their approach maintaining the bounds is possible.

With the input space from Fig. 2 the flowsheet simulator calculates the output space shown in Fig. 3. The resulting

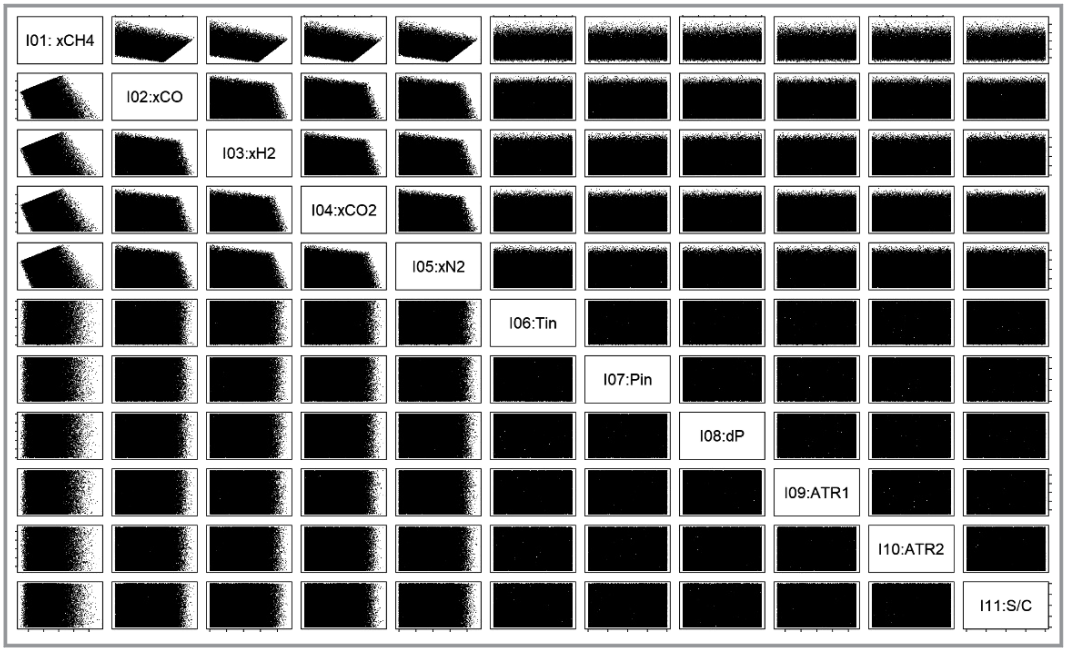


Figure 2. Input space for the rigorous reactor model created with latin hypercube sampling.

Table 2. Variable ranges of the sampled input data sets.

Parameter	Lowest value	Mean value	Highest value
x_{dry,CH_4}	0.448	0.618	0.970
$x_{dry,CO}$	0.000	0.096	0.233
x_{dry,H_2}	0.000	0.096	0.233
x_{dry,CO_2}	0.000	0.096	0.235
x_{dry,N_2}	0.000	0.096	0.232
T_{in} [°C]	350.002	475.000	600.000
P_{in} [bar-g]	10.000	30.000	50.000
dP [bar]	0.000	2.500	5.000
ATR_1 [K]	–49.999	0.000	49.999
ATR_2 [K]	–49.999	0.000	49.999
SC	1.000	2.000	3.000

temperature increase ranges from –150 K to 210 K, the extent of reaction 1 from $-3\,mol\,h^{-1}$ to $4\,mol\,h^{-1}$, and the extent of reaction 2 from $-2\,mol\,h^{-1}$ to $12\,mol\,h^{-1}$. One flowsheet calculation takes less than a second. The 100 000 flowsheets can be calculated on a computer cluster within hours. This value is scalable with the size of the cluster.

4.2 Data Analysis

We want to see the impact of individual input variables on the outputs, called isolated impact. This helps to understand the physical behavior of the observed complex and nonlinear system. In flowsheet simulators, this is done with so-called sensitivity studies, where one or two variables are changed while all other design variables are held constant. In an 11D space this strategy is very time consuming, and we see a lack of systematic approaches for this problem, e.g., due to limitations in the functionality of commercial flowsheet simulators. Schöneberger et al. [22] present a stepwise sensitivity study based procedure for the derivation of a batch distillation recipe. In this work we use a different approach because the design variables cannot be decoupled into independent groups, and a stepwise approach cannot be applied.

The outputs calculated from the inputs generated by a space filling design are useless when we want to see the isolated impacts, because too many influences are mixed. This is demonstrated with Fig. 4 where the calculated values of the extent of reaction 1 are plotted over this input variable. One can hardly detect a trend in this figure.

When a factorial design is used, we can see the isolated impacts. But the calculation of a factorial design with a sufficient number of discrete points is computationally too expensive, as discussed above. To find the isolated impacts we interpolate on the points calculated with the space filling design and try to detect dependencies.

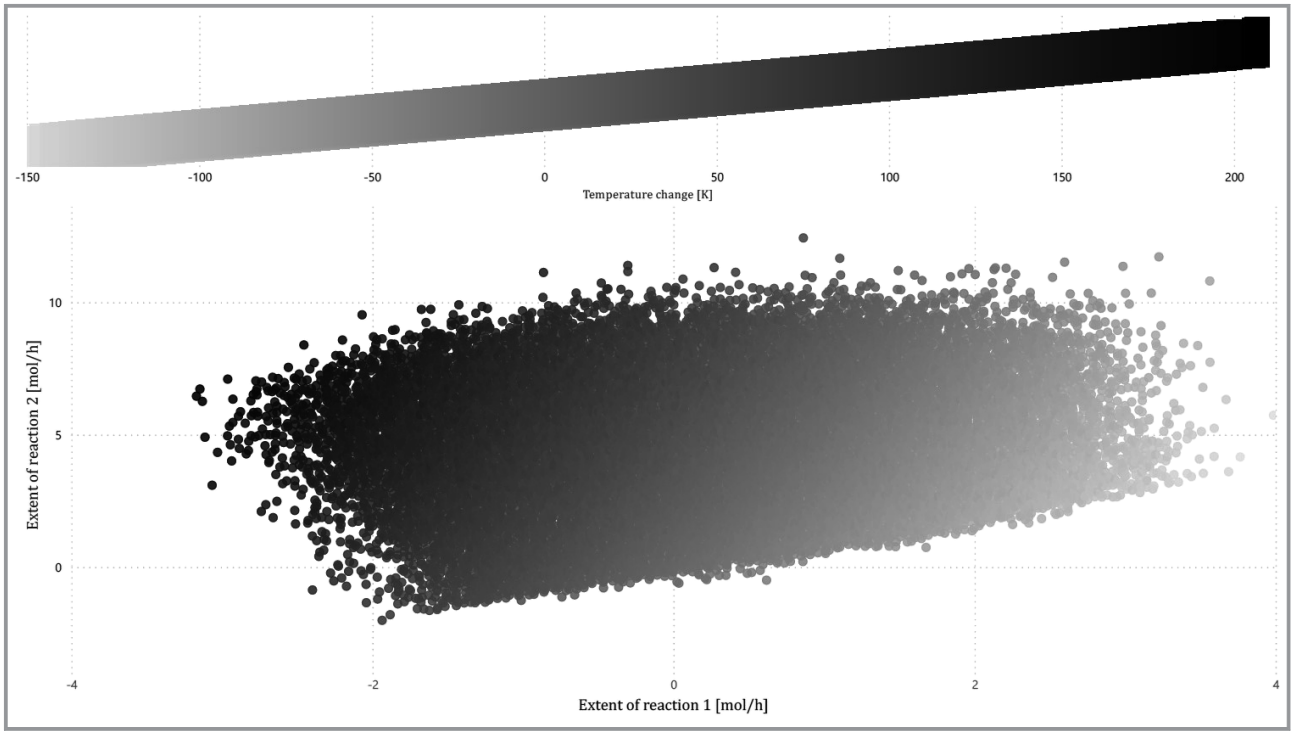


Figure 3. Calculated output variables temperature change (shading), the extent of reaction 1 and the extent of reaction 2 based on the input space from Fig. 2.

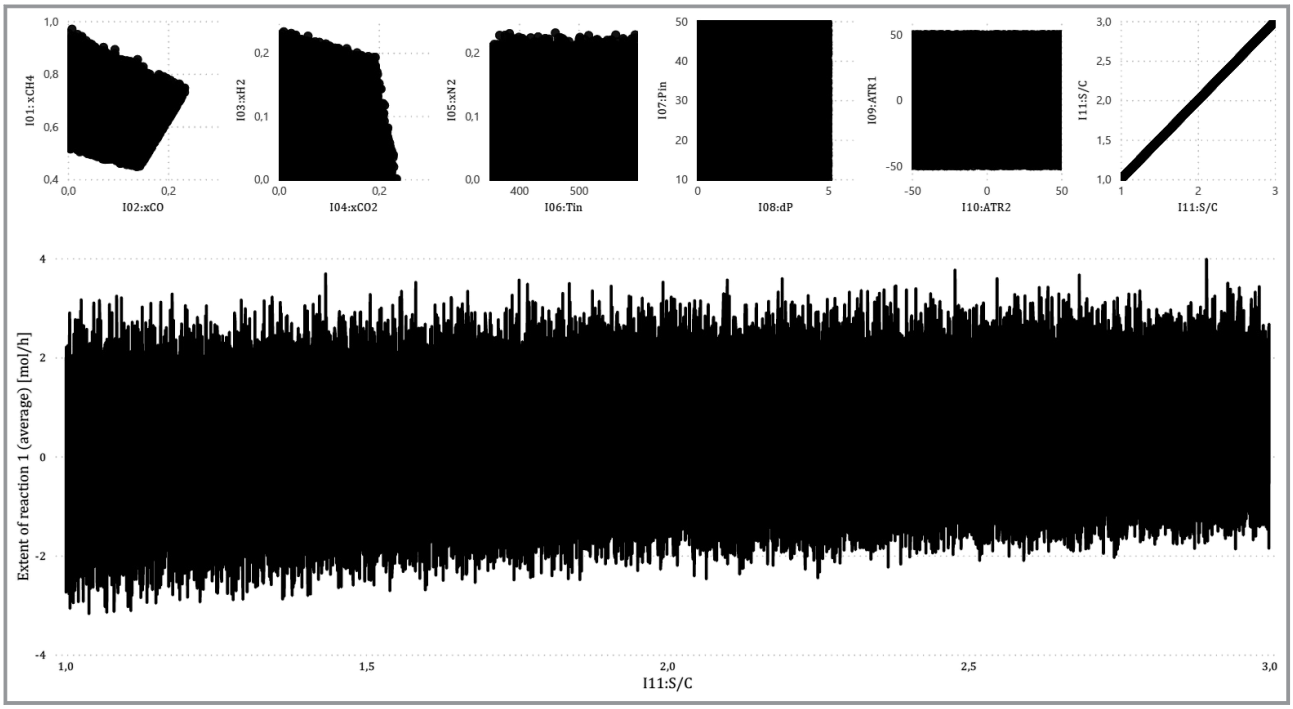


Figure 4. Influence of the steam-to-methane ratio (SC) on the extent of reaction 1 using the data from the latin hypercube sampling.

4.2.1 Linear Interpretation and Interpolation

A darker gray in Fig. 3 relates to a higher temperature increase. With this indication the correlation between the reaction extents and the temperature increase is visualized. Reaction 1 is endothermal and reaction 2 is exothermal. Fig. 3 also reveals a correlation between the two reaction extents. A higher extent of reaction 1 leads to a higher extent of reaction 2. The physical reason for this is that reaction 1 produces the educt CO for reaction 2 and due to its endothermic behavior, it reduces the temperature. The latter increases the extent of the exothermal reaction 2 because of Le Chatelier's principle.

We confirm the almost linear dependency between the temperature increase and the two reaction extents with a linear regression. Fig. 5 shows the deviation of the linear prediction and the rigorously calculated values for the temperature increase. Physically this deviation shows the impact of the non-ideal enthalpy model (Eqs. (9) and (10)). With a pressure-independent enthalpy and constant heat capacities we expect this correlation to be fully linear.

The dependency of the reaction extents regarding the input variables is not linear. Linear regressions fail, as shown in Fig. 6. The residual standard error of the regression is 0.26 for RZ_1 and 0.47 for RZ_2. Beside the nitrogen dry gas inlet mole fraction all the input variables are statistically significant regarding the estimated outputs. We conclude that nonlinear regression techniques are required to get a better fit.

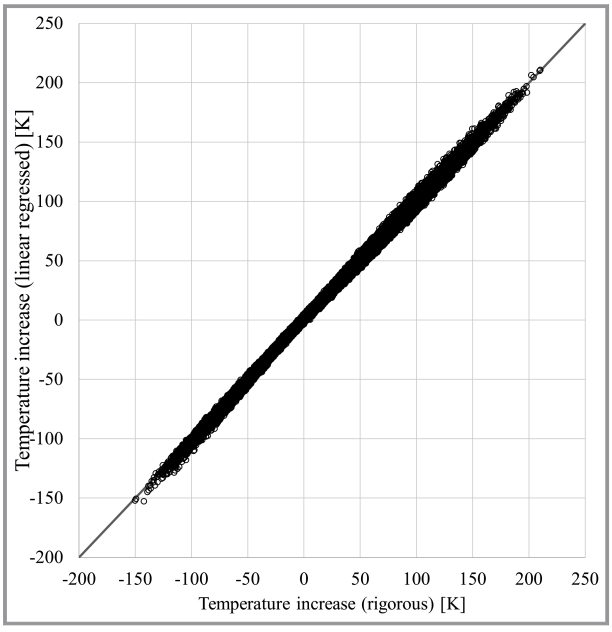


Figure 5. Deviation plot for the linear regression of the temperature change as a function of the extent of reaction 1 and the extent of rection 2.

4.2.2 Interpolation with an ANN

Supervised machine learning groups several nonlinear regression techniques. Currently, popular methods in the process industry are support vector machines, kriging, and

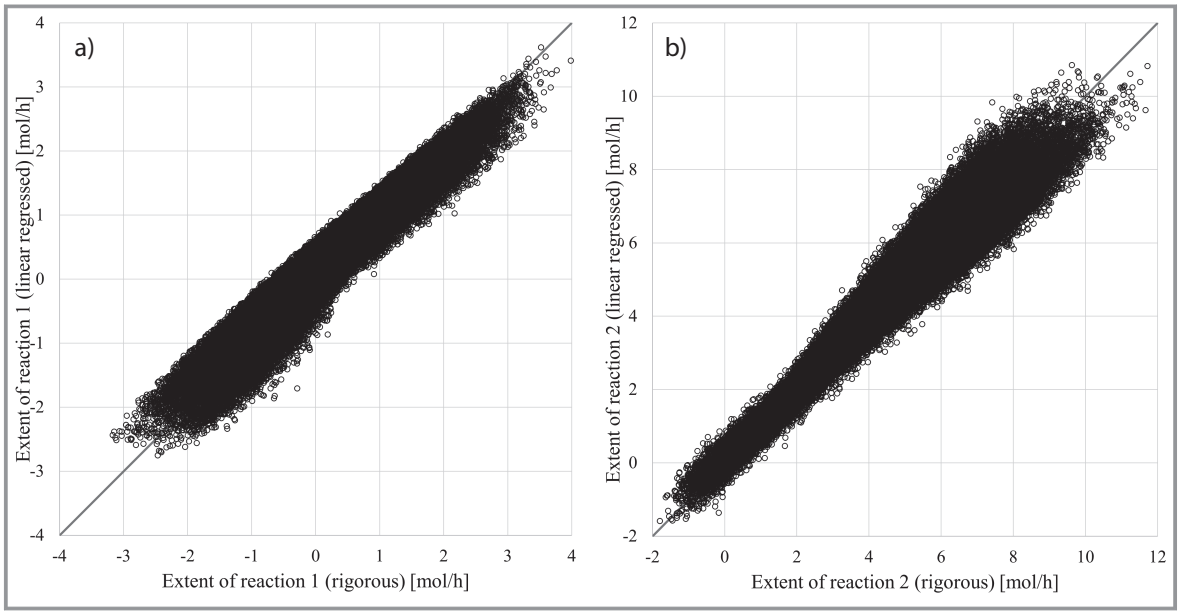


Figure 6. Deviation plots for the linear regression of the extent of reaction 1 (a) and the extent of reaction 2 (b) as a function of the eleven input variables.

artificial neuronal networks. McBride and Sundmacher [23] presented an exhaustive overview about the usage of surrogate and machine learning models in chemical engineering. From the large toolbox of ML models, we choose an ANN with eleven inputs, one hidden layer of 100 neurons, and three outputs to keep the findings comparable with the work of Schmidt et al. [10].

Two ANNs were trained, one with a reluctant linear unit activation function (ANN-ReLU) and one with a hyperbolic tangent activation function (ANN-TANH). Both ANNs use a linear activation function for the three outputs. We scaled and centered the input data and the first output, i.e., the outlet temperature. 80 % of the 100 000 data sets were used as training data and 20 % for validation. We use the RMSProp optimizer to minimize the mean squared error (MSE). The training of one ANN (1503 parameters) takes some minutes on a common desktop PC. With the first iterations the MSE decreases very fast. After 10 iterations the objective function flattens. The MSE of the training data and the validation data show the same trend and remain in the same order of magnitude. This indicates that overfitting is not an issue here. A typical training plot can be found in the SI.

With the regression for the ANN-TANH a lower value of the objective function is achieved. This activation function has the advantage to be continuously differentiable. That is an important feature when the ANN is to be used as a surrogate model for optimization. Fig. 7 shows the deviation between the rigorously calculated and the predicted reaction extents. The prediction accuracy is much higher than with the linear regression, see Fig. 6. There are still some data points that are not predicted well, but we decided to not tune the ANN further and to proceed with this model.

The ANN-TANH predicts the results for the 4^{10} data sets of the factorial design within seconds. With this data the isolated impact of the input variables can be analyzed. Fig. 8 shows how the *SC* affects the average of the output variables. It increases the extent of reaction 1 and decreases the extent of reaction 2 and the outlet temperature. In an interactive PowerBI report the ranges of input and output variables can be adjusted. This enables the analysis of isolated impacts in specific regions of the multidimensional space. We give access to a demo report at the Simulate365 platform [24].

4.3 ANN Unit Operation

The flowsheet simulator CHEMCAD can run custom unit operations coded in C++. Via this interface, we create a PreReformer ANN unit operation that reads the structure and the weights of an ANN from a data file using the hierarchical data format HDF5 [25]. ANNs that were created and regressed with Keras/Tensorflow can directly be stored in this format. The data mapping and scaling must be defined once, together with the input menu (GUI). The GUI we designed for this case is shown in the SI. With this implementation, we can easily switch between ANNs, load new ANNs or update the model weights. For this work, we included the ANN-ReLU, the ANN-TANH, and the ANN trained by our KEEN project partner Air Liquide (ANN-AL) [10].

The PreReformer ANN unit operation reads temperature, pressure, and composition from the inlet stream and converts these values to the ANN input parameters dry gas composition, and steam-to-methane ratio. The user of the flowsheet simulator specifies the reactor performance

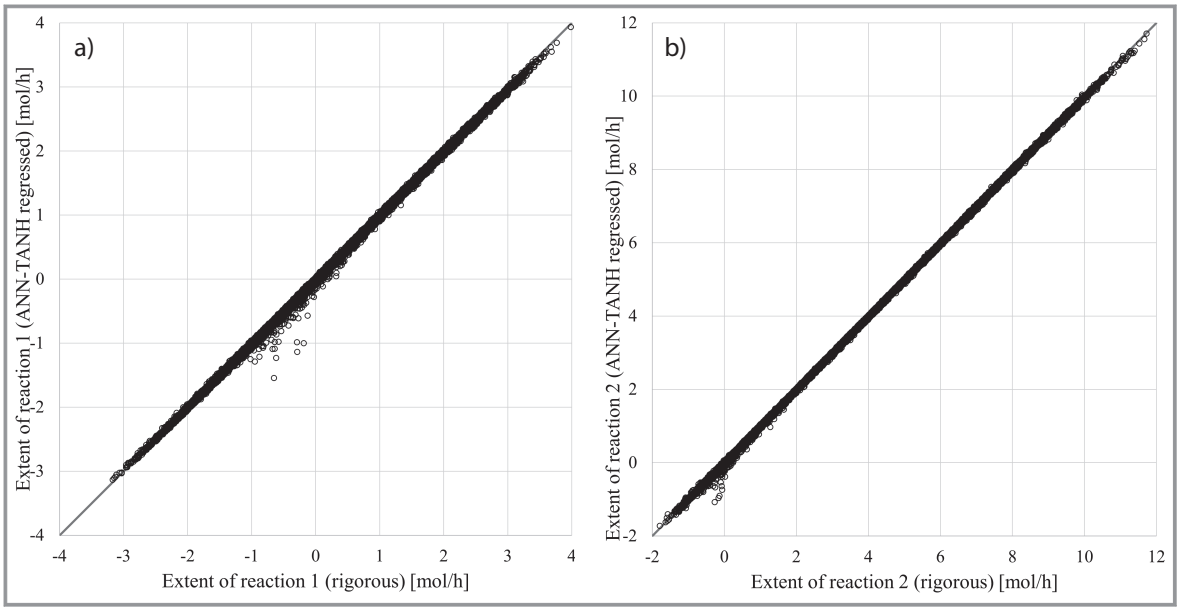


Figure 7. Deviation plots for the nonlinear regression of the extent of reaction 1 (a) and the extent of reaction 2 (b) as a function of the eleven input variables with the ANN-TANH.

parameters via the GUI. With these inputs, the PreReformer ANN unit operation predicts the outlet temperature, and the reaction extents. It uses the reaction extents to calculate the outlet composition. This assures that the mass balance is fulfilled. We fulfill the energy balance for this adiabatic reactor surrogate model by performing a flash calculation with the inlet enthalpy, the outlet pressure, and the outlet composition to get the outlet temperature. The outlet temperature predicted by the ANN is not used anymore. We still show it, as it is an indicator for the model accuracy. Furthermore, we use it to compare the more rigorous hybrid surrogate model (outlet temperature by flash calculation) with the less rigorous hybrid surrogate model (outlet temperature by ANN prediction). Note that the latter model

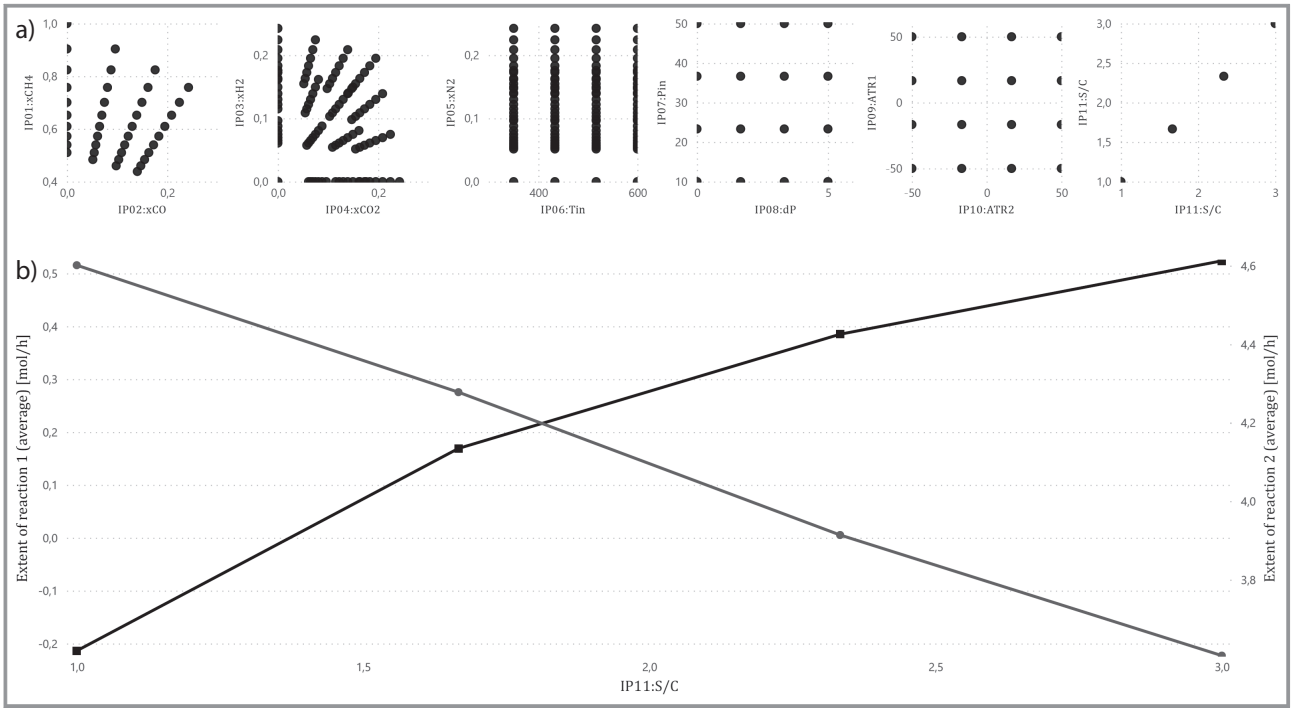


Figure 8. Influence of the steam-to-methane ratio (SC) on the extents of reaction 1 (squares) and rection 2 (circles) using the predicted data from the factorial design.

cannot be used in a flowsheet simulator, because such a tool forces the compliance with mass and energy balances.

4.4 Comparison of the Models

We have four models to describe the physical behavior of the pre-reformer reactor. The first one is the rigorous model (CC-RM) defined by the equation system Eqs. (3)–(10). The other three models are artificial neuronal networks combined with mass and energy balances (ANN-TANH, ANN-ReLU, and ANN-AL). Our work can be compared with Schmidt et al. [10], who trained the ANN-AL on a different data set and generated the data with a different flowsheet simulator.

We compare the models by running sensitivity studies in CHEMCAD. Tab. 3 shows the point in the 11D space from which we analyze the isolated impacts. The hyperbolic tangent based ANN predictions of the rection extent are in good accordance with the rigorously calculated curve, see Fig. 9a.

The ANN with the rectified linear unit activation function shows a larger offset but a similar trend. Fig. 9b shows that the hybrid model (Tout by flash) is much closer to the rigorously calculated outlet temperature than the outlet temperature predicted by the ANN.

Fig. 10 shows the predicted isolated impact of the steam-to-methane ratio on both reaction extents. All models predict an increasing extent of rection 1 and a decreasing extent of rection 2 with an increasing steam to carbon ratio. The trend predicted with the ANN-ReLU shows an off-set (reaction 1) and changing slopes (rection 2). Therefore, we consider the predictions of ANN-ReLU as not to be physically reasonable in the observed region.

Note that ANN-TANH and ANN-AL give almost the same result, even though they were trained on different data sets, which were generated with different flowsheet simulators. We conclude that rigorous flowsheet simulations can be transferred from one tool to another with machine learning models.

Table 3. Test point for the comparison of the models.

Parameter	Value
x_{dry,CH_4}	0.60
$x_{dry,CO}$	0.17
x_{dry,H_2}	0.17
x_{dry,CO_2}	0.04
x_{dry,N_2}	0.02
T_{in} [°C]	400[a)]
P_{in} [bar-g]	35
dP [bar]	1
ATR_1 [K]	–40
ATR_2 [K]	30
SC	1.5[b)]

a) Altered from 350 °C to 600 °C (Fig. 9); b) altered from 1 to 3 (Fig. 10).

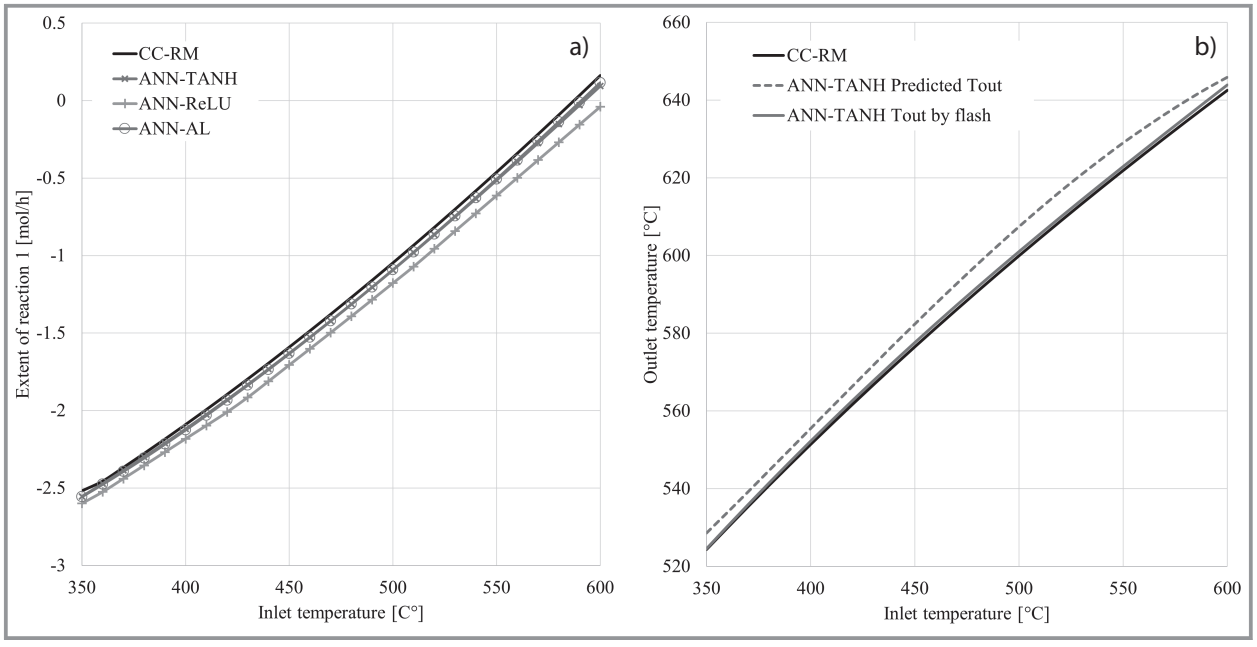


Figure 9. Comparison of the model predictions based on the isolated impact of the inlet temperature. Predicted reaction extent (a) and predicted outlet temperature (b).

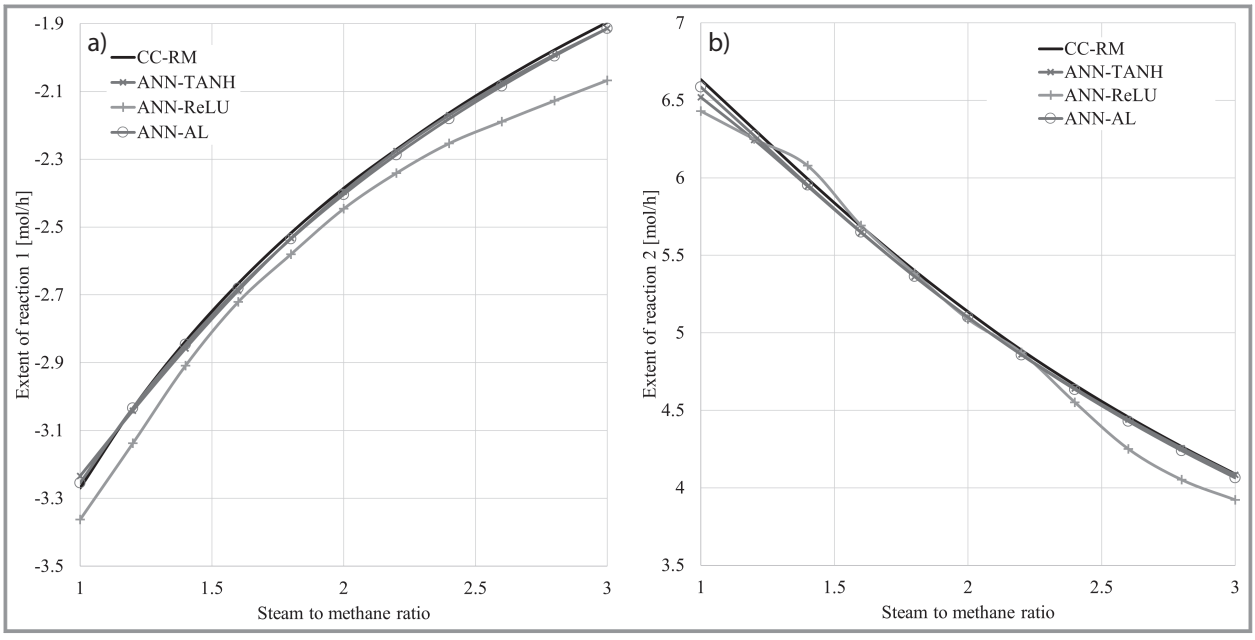


Figure 10. Comparison of the model predictions based on the isolated impact of the steam-to-methane ratio. Predicted reaction extent 1 (a) and reaction extent 2 (b).

4.5 Maximum Model Deviation (ANN-TANH)

The sequential quadratic programming algorithm (SQP) implemented in the flowsheet simulator is used to find the set of input variables at which the deviation between extent of reaction 1 calculated with the ANN-TANH und with CC-RM reaches a maximum. Due to the nature of SQP algorithms this can be a local optimum in case of a non-convex objective function [26]. We start from the point of maximum deviation that is visible in Fig. 7, i.e., the worst regressed data set from the 100 000 data points generated with LHS. The optimizer converges to a corner of the 11D hypercube, where it finds the maximum squared error between ANN-TANH and CC-RM. Tab. 4 shows this data set. These input parameters result in a very small extent of reaction 1. The extent of rection 2 is also practically zero, as no CO is present in the feed. The ANN-TANH predicts a quite large negative extent for reaction 1. We assume that the reason for this bad prediction is that all input parameters beside the CO concentration indicate a high negative reaction extent: low inlet and approach temperatures (reaction is endothermal), low steam-to-methane ratio (water is a educt), high hydrogen content in feed (hydrogen is a product), and high pressure (number of moles is increasing due to the reaction). The fact that no CO is present and hence the reverse reaction cannot take place is not taken into account by the ANN-TANH. The flowsheet simulator ignores negative component flowrates and sets them to zero. This leads to a mass balance error in the flowsheet simulation and a warning message.

Table 4. Data set with maximum deviation between the predictions of CC-RM and ANN-TANH.

Parameter	Value
x_{dry,CH_4}	0.767
$x_{dry,CO}$	0.00
x_{dry,H_2}	0.233
x_{dry,CO_2}	0.00
x_{dry,N_2}	0.00
T_{in} [°C]	350
P_{in} [bar-g]	50
dP [bar]	0
ATR_1 [K]	–50
ATR_2 [K]	50
SC	1.0
RZ_1 CC-RM [mol h^{-1}]	0.00001
RZ_1 ANN-TANH [mol h^{-1}]	–2.048

Fig. 11 shows the isolated impact of the inlet temperature on the extent of reaction 1 from this point. We expect the equilibrium constant of reaction 1 to change with the reactor temperature. This effect gets visible at an inlet temperature of around 500 °C. Even though the ANN-AL shows a

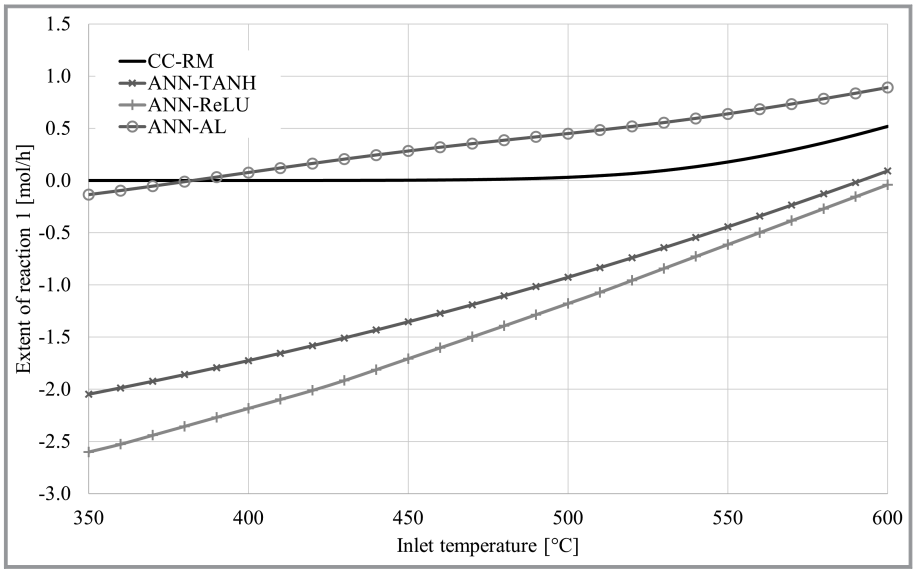


Figure 11. Isolated impact of the inlet temperature starting from the point of maximum model deviation (Tab. 4).

smaller deviation from the rigorous model, none of the ML models is able to describe the flat line predicted by CC-RM.

If this region is of major interest, we must generate more training data points there and retrain the models. If this region is not of interest, e.g., because such a feed gas composition is unrealistic, we can exclude it from the maximum error search optimization by adding bounds and constraints. No matter what the case is, we see the maximum prediction error of an ML model and the location where it occurs as important information for anyone who is going to use this model.

5 Summary and Conclusions

We demonstrated the workflow of

- generating data for an adiabatic pre-reformer reactor in a steam reformer plant by using a first-principle flowsheet simulator,
- using this data for the training of linear and nonlinear surrogate models,
- using the machine learning (ML) models to calculate data for visualizing the isolated impacts of the input variables,
- implementing the ML models as unit operation in a flowsheet simulator,
- interpreting the prediction quality of the ML models based on isolated impacts and process engineering knowledge, and
- finding the maximum deviation between the ML models and the rigorous model in the observed region with a SQP optimizer.

We conclude that ML models save a process engineer's time for analyzing flowsheet simulations. They also significantly reduce the computational effort for exploring the commonly high-dimensional solution space. We demonstrated how to systematically build ML models, how to use them in an exemplary analysis of process variables, and how to check their predictions for physical consistency. The accuracy of hybrid models is higher than the accuracy of sole data-driven models. Only hybrid ML models give consistent results when implemented in a flowsheet simulator. We see hybrid ML models as a promising technique to transfer simulation results from one simulation tool to another. The predicted results must be tested for physical reasonability though. For this step domain knowledge is required, but it can be supported by ML unit operations in flowsheet simulators.

Supporting Information

Supporting Information for this article can be found under DOI: https://doi.org/10.1002/cite.202100089.

The authors acknowledge the financial support by the Federal Ministry for Economic Affairs and Energy of Germany (BMWi) within the innovation platform "KEEN – Artificial Intelligence Incubator Laboratory in the Process Industry" (grant number 01MK20014E).

Symbols used

a_M	[m^6bar mol^{-2}]	Mixture parameter in the PR equation of state
ATR	[K]	Approach temperature
b_M	[m^3mol^{-1}]	Mixture parameter in the PR equation of state
dP	[bar]	Pressure drop
dT	[K]	Temperature change

F	[mol h^{-1}]	Molar flowrate
$\Delta_R G°$	[J mol^{-1}]	Gibbs free energy of reaction at reference state
h	[J mol^{-1}]	Specific molar enthalpy
K	[–]	Reaction equilibrium constant based on mole fractions
P	[bar-g]	Pressure
R	[m^{3}bar mol^{-1}K^{-1}]	Universal gas constant
RZ	[mol h^{-1}]	Reaction extent
SC	[–]	Steam-to-methane ratio
T	[°C]	Temperature
v	[m^{3}mol^{-1}]	Specific molar volume
x	[mol mol^{-1}]	Mole fraction
ν	[–]	Stochiometric coefficient

Sub- and Superscripts

dry	Dry basis
c	Component (1–6)
ideal	Ideal gas state
r	Reaction (1–2)
s	Stream (in/out)

Abbreviations

ACT	Automatable computer task
ANN	Artificial neuronal network
ANN-AL	Artificial neuronal network trained by Schmidt et al. [10]
ANN-ReLU	Artificial neuronal network with a rectified linear unit activation function
ANN-TANH	Artificial neuronal network with a hyperbolic tangent activation function
CC-RM	Rigorous model evaluated with the flowsheet simulator CHEMCAD 7
DCS	Distributed control system
Ixx	Input xx
KEEN	Artificial Intelligence Incubator Laboratory in the Process Industry
LHS	Latin hypercube sampling
MET	Manual engineering task
MESH	Material balances, equilibrium and sum equations, and heat balance
ML	Machine learning
Oxx	Output xx
OTS	Operator training system
PR	Peng-Robinson equation of state
RM	Rigorous model
RTO	Real time optimization

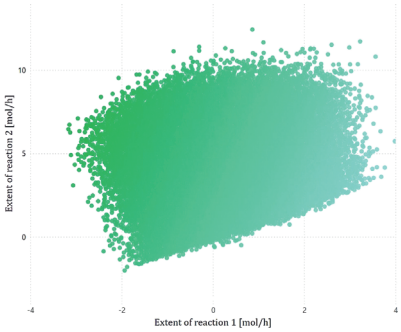
References

[1] N. Kockmann, *Chem. Ing. Tech.* **2018**, *90 (11)*, 1621–1627. DOI: https://doi.org/10.1002/cite.201800135
[2] S. Bröcker, R. Benfer, M. Bortz, S. Engell, C. Knösche, A. Kröner, *Process Simulation – Fit for the future?*, DECHEMA Position Paper, Dechema, Frankfurt **2021**. https://dechema.de/en/Positionspapier_Prozesssimulation/_/PP_Process_Simulation_2021_engl.pdf (Accessed on May 28, 2021)
[3] D. P. Solomatine, A. Ostfeld, *J. Hydroinf.* **2008**, *10 (1)*, 3–22. DOI: https://doi.org/10.2166/hydro.2008.015
[4] F. Chollet, J. Allaire, *Deep learning with R*, Manning, New York **2018**.
[5] Z. Ge, Z. Song, S. X. Ding, B. Huang, *IEEE Access* **2017**, *5*, 20590–20616. DOI: https://doi.org/10.1109/ACCESS.2017.2756872
[6] S. A. Boyer, *SCADA: Supervisory control and data acquisition*, 4th ed., International Society of Automation, Research Triangle Park, NC **2010**.
[7] M. D. McKay, R. J. Beckman, W. J. Conover, *Technometrics* **1979**, *21 (2)*, 239. DOI: https://doi.org/10.2307/1268522
[8] S. S. Haykin, *Neural networks: A comprehensive foundation*, 2nd ed., Prentice Hall, Upper Saddle River, NJ **1999**.
[9] B. J. Cromarty, C. W. Hooper, *Int. J. Hydrogen Energy* **1997**, *22 (1)*, 17–22. DOI: https://doi.org/10.1016/S0360-3199(96)00034-1
[10] R. Schmidt, A. Chattot, A. Bouchrit, M. Mighani, E. Örs, *Comput.-Aided Chem. Eng.* **2020**, *48*, 523–528.
[11] J. Gmehling, B. Kolbe, M. Kleiber, J. R. Rarey, *Chemical thermodynamics for process simulation*, Wiley-VCH Verlag, Weinheim **2012**.
[12] *DIPPR 801 Database*, Design Institute for Physical Property Research/AIChE, Houghton, MI **2011**.
[13] D. Dupuy, C. Helbert, J. Franco, *J. Stat. Software* **2015**, *65 (11)*, 1–38.
[14] *R: A Language and Environment for Statistical Computing*, R Core Team, Vienna **2020**. www.r-project.org/, (Accessed on May 28, 2021).
[15] Y. Buruk Sahin, E. Aktar Demirtaş, N. Burnak, *Pamukkale J. Eng. Sci.* **2016**, *22 (4)*, 297–304. DOI: https://doi.org/10.5505/pajes.2015.98598
[16] *CHEMCAD 7*, Chemstations Inc. **2020**. www.chemstations.com/Support/Software_Updates_and_Notes/CHEMCAD_7_Support (Accessed on May 28, 2021)
[17] A. Fricke, J. C. Schöneberger, *Chem. Ing. Tech.* **2017**, *89 (5)*, 515–526. DOI: https://doi.org/10.1002/cite.201600130
[18] https://powerbi.microsoft.com/en-us (Accessed on May 28, 2021)
[19] T. Wei, V. Simko, *R package "corrplot": Visualization of a Correlation Matrix* **2017**. https://github.com/taiyun/corrplot (Accessed on May 28, 2021)
[20] J. J. Allaire, F. Chollet, *keras: R Interface to 'Keras'* **2020**. https://cran.r-project.org/package=keras (Accessed on May 28, 2021)
[21] L. Lueg, D. Schack, R. Schmidt, M. von Kurnatowski, P. O. Ludl, M. Bortz, *Chem. Ing. Tech.*, submitted. DOI: https://doi.org/10.1002/cite.202100087
[22] J. C. Schöneberger, D. Staak, J. Rarey, in *Simulation and Optimization in Process Engineering: The Benefit of Mathematical Methods in Applications of the Chemical Industry*, Elsevier, submitted.
[23] K. McBride, K. Sundmacher, *Chem. Ing. Tech.* **2019**, *91 (3)*, 228–239. DOI: https://doi.org/10.1002/cite.201800091
[24] https://simulate365.com/ (Accessed on May 28, 2021)
[25] *HDF5*, HDF Group, Champaign, IL **2020**. www.hdfgroup.org/solutions/hdf5/ (Accessed on May 28, 2021)
[26] T. F. Edgar, D. M. Himmelblau, L. S. Lasdon, *Optimization of chemical processes*, 2nd ed., McGraw-Hill, New York **2008**.

DOI: 10.1002/cite.202100089

Explaining and Integrating Machine Learning Models with Rigorous Simulation

Jan C. Schöneberger, Burcu Aker, Armin Fricke*



Research Article: We present a workflow for building and evaluating machine learning (ML) models based on data generated by rigorous simulation. For this purpose, the ML models are implemented in a flowsheet simulator as unit operations. These unit operations are hybrid models, combining data-driven predictions with mass and energy balances.

Supporting Information available online

