Received: 28 August 2020 | Revised: 20 May 2021 | Accepted: 15 June 2021

DOI: 10.1002/aic.17357

PROCESS SYSTEMS ENGINEERING

AIChE JOURNAL

Constrained adaptive sampling for domain reduction in surrogate model generation: Applications to hydrogen production

Julian Straus | Jabir Ali Ouassou | Brage Rugstad Knudsen | Rahul Anantharaman

SINTEF Energy Research, Trondheim, Norway

Correspondence
Julian Straus, Department of Gas Technology, SINTEF Energy Research, Kolbjørn Hejes vei 1B, Trondheim 7491, Norway.
Email: julian.straus@sintef.no

Funding information
Equinor; Gassco; Gassnova; Total; European Commission under the Horizon 2020 programme, Grant/Award Number: 691712; BEIS; RVO; Bundesministerium für Wirtschaft und Energie (BMWi); DETEC

Abstract

We propose a new approach for sampling domain reduction for efficient surrogate model generation. Currently, the standard procedure is to use box constraints for the independent variables when sampling the exact simulator. However, by including additional inequality constraints to account for interdependencies between these variables, we can drastically reduce the sampling domain and ensure consistency of unit operations. Moreover, we present a methodology for constructing surrogate models based on penalized regression and error-maximization sampling. All these algorithms have been implemented as a free and open-source software package. Through a case study on the water–gas shift reaction for hydrogen production, we show that sampling domain reduction reduces the required number of sampling points significantly and improves the accuracy of the surrogate model.

KEYWORDS
adaptive sampling, hydrogen production, sampling domain reduction, surrogate modeling

1 | INTRODUCTION

Surrogate models, also known as *reduced-order models*, *response-surface models*, *metamodels*, or *proxy models*, are a class of regression models that have gained much attention in recent years. They are frequently used in a variety of engineering fields, including chemical process engineering. Surrogate models can drastically reduce computation time by replacing potentially expensive or noisy low-level models with simpler regressions. Surrogate models are most commonly used as substitutes for computationally expensive models in optimization[1] or to identify the best process structure based on a superstructure optimization approach.[2] Bhosekar and Ierapetritou[3] and McBride and Sundmacher[4] reviewed recent advances in the field of surrogate modeling.

Recent research on surrogate modeling can be grouped into two main categories: (i) Basis function selection and associated regression methodology and (ii) Improved sampling routines, also referred to as design of computer experiments. The choice of basis function often has a significant impact on the achievable accuracy of the surrogate model. It also determines for what purposes the surrogate model is suitable since the basis choice affects the model size, complexity, and differentiability. A trend in emerging software packages for surrogate-model generation is to include broad flexibility in the selection of basis functions, either embedded as a part of the regression method or overall surrogate-modeling framework or as an *a priori* choice in the surrogate fitting options taken from a roll-down list. To this end, the Automated Learning of Algebraic Models for Optimization (ALAMO) framework, as an example, selects the best subset from a large set of algebraic basis functions using a mixed-integer approach.[1] In a similar vein, the ARGONAUT framework allow the user to select the basis functions from an implemented pool, including polynomials,

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.
© 2021 The Authors. *AIChE Journal* published by Wiley Periodicals LLC on behalf of American Institute of Chemical Engineers.

Kriging models, and radial basis function as part of a derivative-free optimization (DFO) of gray-box problems.[5] The implementation of linear, polynomial, and other simple algebraic basis functions aid the application of the surrogate models for large-scale optimization purposes. Additionally, there is also a substantial effort to explore other common basis functions in surrogate modeling for process-system application including Kriging or Gaussian process models,[2,6] artificial neural networks,[7,8] and splines.[9,10]

Another important factor for the performance of surrogate model generation is the design-of-experiment (DoE) method used for the sampling. Garud et al. provided a detailed overview of DoE methods.[11] They distinguished between two main strategies: *static sampling* and *adaptive sampling*. Static sampling means that one first collects all samples via a predefined experimental design and then performs a batch fitting of the surrogate model to the obtained data. Quirante et al., for example, use static sampling for the development of surrogate models for superstructure optimization of vinyl chloride production.[2] They used 200–250 points to fit models with 4–5 independent variables and Kriging models[12] as basis functions. As another example, Ochoa-Estopier et al. sampled 3000 points describing a Latin hypercube to train an artificial neural network with 10 independent variables.[13]

One issue with static sampling approaches is that it is not known *a priori* how many sample points are required. Hence, both under- and oversampling may occur. As an alternative, adaptive sampling routines have been developed. These routines utilize a strategy of *exploration* and *exploitation*: The first refers to space-filling techniques, while the second is accomplished by iteratively selecting new sample points based on the surrogate model itself. Crombecq et al. developed a hybrid, sequential sampling strategy using a Monte Carlo-based approximation of a Voronoi tessellation for exploration and local linear approximations of the simulator for exploitation.[14] The ALAMO framework implements an adaptive sampling scheme where new sample points are placed to maximize the discrepancy between the surrogate model and the actual model.[1] Hence, it only utilizes exploitation. The smart sampling algorithm (SSA) developed by Garud et al. incorporates both exploitation and space-filling metrics.[15] Space-filling is quantified via a *crowding distance*, that is, the distance between a new point and all existing points. Exploitation is measured by a *departure function*, which quantifies the modeling error caused by excluding points. Both adaptive sampling approaches require numerical optimization. The error-maximization sampling implemented in ALAMO uses DFO as the optimization approach, which involves running the simulator or evaluating the black-box model. In contrast, SSA uses an optimization routine only involving the surrogate model. To avoid numerical optimization, Eason, and Cremaschi used jack-knifing to identify regions where the output of the surrogate model has a large variance, and combined this with a space-filling measure based on the Euclidean distance between sample points.[7] Garbo and German introduced a model-independent sequential adaptive sampling technique called nearest neighbors adaptive sampling (NNAS). A refinement metric based on local linear models was used together with a Pareto-ranking-based criterion to achieve a refinement-exploration balance of the surrogate model.[16] In general, adaptive sequential sampling routines often improve overall surrogate-model performance compared to single-shot static approaches and reduce the required number of samples.[14]

In addition to the exploration and exploitation metrics used to improve the sampling routines in surrogate-model generation, constraints on the sampling domain are frequently used to effectively reduce the search space. Different approaches have been explored for constraining the sampling domain and seeking to prevent clustering of sampling points. Li et al. proposed maximizing a distance metric between all newly sampled points and one unobserved (not sampled) point, while also including a threshold constraint to prevent clustering of new points with existing experiments.[17] They applied their clustering-constraint approach, which lends itself to leave-one-out (LOO) methods, to a Kriging model. Zhou et al. implemented a space-filling approach that minimizes a prediction-error metric of the surrogate model with a constraint added to ensure a minimum distance between the new sample points and the existing sampling points.[18] Other works have utilized governing properties of the surrogate-model, such as Gaussian process models, to impose an expected improvement metric with added constraints on the sampling.[19] Yet, with respect to surrogate-model generation for process-system applications, simple box constraints are usually used to confine the sampling domain.[2,6,20] That is, for each independent variable x_i, simple lower and upper bounds are imposed,

$$x_{i,\min} \le x_i \le x_{i,\max}, \quad \forall i \in \mathcal{I}. \tag{1}$$

However, this leads to a dilemma related to the choice of the bounds $x_{i,\min}$ and $x_{i,\max}$:

1. Using tight constraints limits the applicability of the surrogate model. Potentially interesting regions are not included during model construction which can lead to extrapolation.
2. Using loose constraints may cause sampling in regions that are never encountered in practice and potentially highly nonlinear. As a result, more sampling points are required to achieve a satisfactory fit in the region of interest.

Particularly important are constraints related to molar or mass flows in a chemical process, as these are in general defined by the required stoichiometry and extent of a chemical reaction. To address this problem of sampling with simple box constraints, Straus, and Skogestad proposed introducing molar ratios for feed flows to account for dependencies between different chemical components in the inlet of a chemical process.[21] The error of the surrogate model was reduced when the same number of sample points was used for one dependent variable, whereas the other two dependent variables remained unaffected. Straus and Skogestad further showed that sampling with standard box constraints on inlet stream and manipulated variables for an ammonia reactor may result in sampling in undesired regions, that is, sampling in regions where the reactor is extinct or shows limit-cycle behavior.[10] Introducing constraints on the manipulated variables (e.g., split ratios or compressor duties) and internal variables based on concepts from process control can avoid sampling in undesired regions.

This article addresses sampling-domain reduction for surrogate model generation through constrained adaptive sampling, particularly suited for interconnected process models. Constrained adaptive sampling can utilize dependencies between independent variables of connected submodels to construct constraints on the sampling domain. Our main contributions are:

1. We present a generalized constraint formulation for sampling domain reduction.
2. We propose a penalized adaptive sampling method that prevents oversampling of small regions during error-maximization sampling and thereby improves exploration in the sampling routine.
3. We implement and present the proposed sampling methodology in `Consumet`—a Python-based open-source package for surrogate-model generation.[22]
4. We present and demonstrate the merits of our proposed approach on an auto-thermal reformer (ATR) and the water–gas shift section of hydrogen production with two reactors.

The remainder of the article is organized as follows. First, in Section 2, we develop the theory for simple, linear inequality constraints for sampling approaches. Furthermore, we elaborate gray-box modeling related to surrogate model generation to justify the use of different dependent variables. Section 3 provides an exposition of the algorithm we propose for surrogate model generation, based on adaptive sampling, penalized regression, and information criteria. We have implemented the proposed methods as a free and open-source software solution. Section B in the Appendix summarizes its implementation details. Finally, Section 4 applies these methods to a case study where we construct surrogate models for an ATR and the water–gas shift section of hydrogen production with two reactors. A detailed description of the regression and penalty selection can be found in Section A in the Appendix.

2 | SAMPLING DOMAIN REDUCTION

2.1 | Surrogate model structure

Sampling domain reduction is vital for surrogate models representing subsections of an overall process. The surrogate models can be subsequently used for superstructure optimization. To this end, it is essential to develop a consistent structure for the surrogate models of the different subsections.

Consider surrogate model i, located between subsection k and subsection q. The inlet connection variables can then be described as $\mathbf{z}_{k,i} \in \mathbb{R}^{n_z}$ while the outlet connection variables are given as $\mathbf{z}_{i,q} \in \mathbb{R}^{n_z}$. These n_z connection variables usually include composition, flow, temperature, and pressure. Furthermore, the subsection has n_u additional input variables $\mathbf{u}_i \in \mathbb{R}^{n_u}$ corresponding to heat exchanger duties, compressor duties, or split ratios. Suppose the surrogate models are later connected. In that case, it is necessary to express the connection variables $\mathbf{z}_{k,i}$ from surrogate model k to surrogate model i in the same fashion in each subsection. One advantage of this approach is that it is unnecessary to limit the nonlinear surrogate models to the connection variables; that is, we are not limited to surrogate models that map the input connection variables $\mathbf{z}_{k,i}$ to the output connection variables $\mathbf{z}_{i,q}$. Instead, it is possible to include extents of reaction and separation coefficients to obtain mass consistency in the surrogate model. Switching from connection variables as dependent variables to auxiliary variables like the extent of reaction was described in detail by Henao et al. in developing surrogate models for superstructure optimization.[23] Consequently, we introduce a new gray-box model structure for each subsection that should be represented by a surrogate model. Using extent of reactions and separation coefficients is especially beneficial if there are mass recycle streams resulting in positive feedback as it ensures mass consistency.[10,21,24]

Figure 1 illustrates the gray-box structure and the flow of information involved in the process. This gray-box model uses as input the connection variables of the inlet streams (or calculated variables from the inlet streams) and the manipulated variables. The output of this model is then given by the connection variables of the outlet streams, that is, the variables $\mathbf{z}_{i,q}$. The gray-box model structure includes the following different sections and calculates the values sequentially:

1. Calculation of the input to the nonlinear surrogate models if ratios or other nonstream variables are used as input to the surrogate models. The calculation of the input to the nonlinear surrogate models may not be necessary if the inlet connection variable $\mathbf{z}_{k,i}$ and the additional input $\mathbf{u}_i$ are used directly in the surrogate model;
2. Nonlinear surrogate models $\mathbf{g}_i$ for the calculation of the separation coefficients, rates of the extent of reaction, temperature differences, pressure differences, and so on, summarized in a variable $\mathbf{y}_i \in \mathbb{R}^{n_y}$;
3. Exact mass balance and balances for the outlet pressure and temperature ($\mathbf{f}_i$) to calculate the connection variables of the outlet streams. These balances may also include the inlet connection variables $\mathbf{z}_{k,i}$.

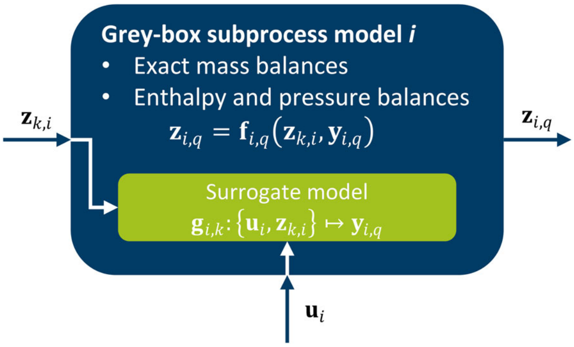
Grey-box subprocess model i
Exact mass balances
Enthalpy and pressure balances
$\mathbf{z}_{i,q} = \mathbf{f}_{i,q}(\mathbf{z}_{k,i}, \mathbf{y}_{i,q})$
$\mathbf{z}_{k,i}$
$\mathbf{z}_{i,q}$
Surrogate model
$\mathbf{g}_{i,k}: \{\mathbf{u}_i, \mathbf{z}_{k,i}\} \mapsto \mathbf{y}_{i,q}$
$\mathbf{u}_i$

FIGURE 1 Illustration of the gray-box structure for the development of consistent surrogate models [Color figure can be viewed at wileyonlinelibrary.com]

A key question is in which range we should vary the input variables to a subsection. Here, the input variables to a subsection are generally not independent of each other as they may originate from, for example, a chemical reaction. The following sections will first illustrate the concept of dependencies between the inlet variables and then propose a novel approach for avoiding sampling in unimportant regions.

2.2 | Motivating example

Straus and Skogestad already investigated flow dependencies for proportional dependencies of different inlet molar flow rates.[21] Using proportional dependencies is especially useful for surrogate models whose feed composition is not determined by a chemical reaction or a separation step. To illustrate the concept of proportional dependencies again, consider the steam–methane reforming reaction given by:

$$CH_4 + H_2O \rightarrow 3H_2 + CO \tag{2}$$

The main reactants in the feed of a steam–methane reformer (SMR) are methane and steam. Based on the chemical reaction (2), a steam–methane molar ratio of 1 would be optimal. However, operating the reactor at the stoichiometric limit could cause coking in the reactor catalysts. Steam flows greater than the stoichiometric one are required to prevent coking. Additionally, the excess steam supplied to the SMR can be utilized in the subsequent water–gas shift reactors in the reforming process. Thus, a steam–methane molar ratio of 2.5 is frequently used. Table 1 shows box constraints for the reaction for both the methane and steam molar flow rate using a steam–methane molar ratio of 2.5 to calculate the box constraints for steam at the nominal operating point. However, due to the application of box constraints we observe, extrema of:

$$\max \dot{n}_{H2O} \oslash \dot{n}_{CH4} = 7.5 \quad \min \dot{n}_{H2O} \oslash \dot{n}_{CH4} = 0.83 \tag{3}$$

where $\oslash$ corresponds to element-wise division. These values are encountered when one of the independent variables is at the upper bound and the other at the lower bound. In practice, these ratios will never be observed due to either below-stoichiometric amounts of steam, which results in coking of the catalyst bed, or too high amounts of steam, which results in unnecessary compression and steam generation costs. Figure 2 illustrates the behavior. Here, 1000 random points were created using both box constraints and proportional dependencies. We can directly see that using box constraints, most of the sampling domain is, in fact, in regions that will not be encountered in practice. Furthermore, important regions at the upper and lower methane bound are omitted as the steam–methane ratio is fixed to the nominal value if both independent variables are at the same bound.

TABLE 1 Bounds and units for the different investigated configurations

	$\dot{n}_{H_2O}$ (mol/s)	$\dot{n}_{CH_4}$ (mol/s)	$\frac{\dot{n}_{H_2O}}{\dot{n}_{CH_4}}$ (−)
Lower bound	1000	400	2.0
Nominal value	2000	800	2.5
Upper bound	3000	1200	3.0

The variation in the steam–methane ratio does not correspond to ±50% as for the box constraints. Hence, it is not surprising that the overall domain using proportional dependencies is much smaller than the one based on box constraints. In general, one should avoid sampling at too large variations around the nominal ratio, as these regions are not important in day-to-day operation due to the reasons outlined above. Hence, a variation of ±20% in the molar ratio of steam and methane can be considered sufficient for the subsequent applications of a surrogate model. If we incorporate such proportional dependencies, we can reduce the sampling domain to about 40% of the domain required with box constraints.

2.3 | Generalization of dependencies

Suppose the feed streams to a surrogate model are outputs from previous unit operations with chemical reactions and/or separation steps. In that case, it is in general not possible to define simple proportional dependencies through a change in independent variables. If we consider the SMR example above, the subsequent unit would be the water–gas shift section in which CO is converted to H_2 according to:

$$CO + H_2O \rightleftharpoons H_2 + CO_2 \tag{4}$$

The feed composition to this section depends on the extent of reaction of the steam–methane reforming reaction in the previous section. The following contradicting conclusions can be drawn based on Equation (2):

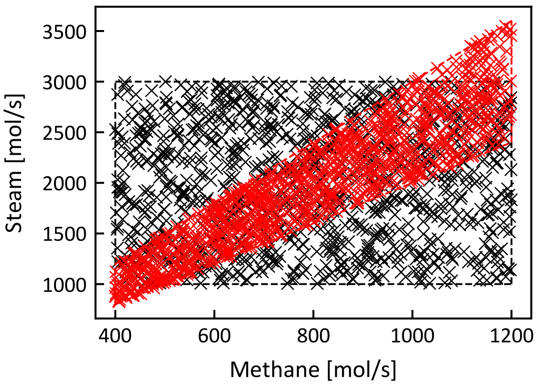

FIGURE 2 Sampling domain when using box constraints (black) and proportional dependencies (red) for the steam and methane flow rates in the feed to a steam–methane reformer [Color figure can be viewed at wileyonlinelibrary.com]

1. The more steam is in the feed to the water–gas shift section, the more hydrogen is in the feed due to a larger inlet flow rate of steam to the SMR and correspondingly a larger extent of reaction (proportional dependency);
2. The more steam is in the feed, the less hydrogen is in the feed due to a reduced extent of reaction in the SMR caused by a lower residence time in the reactor (inverse proportional dependency).

Therefore, it is impossible to predict the dependency without knowledge of the outlet composition of the previous unit operations. It can, however, have a significant influence on the sampling domain.

Consider the steam–methane reforming reaction (2) again as an example to illustrate these dependencies. Let us take the methane flow rate $\dot{n}_{CH_4}$ and the steam–methane ratio $\dot{n}_{H_2O}/\dot{n}_{CH_4}$ as our independent variables with the bounds described in Table 1, and use the reactor temperature in centigrades $T_{reac} \in [900, 1000]$ as an additional independent variable. We can analyze the outlet stream composition to identify dependencies between its chemical components.

The reactor section was modeled in Aspen HYSYS using the equilibrium reactor model. In total, 1000 random points were sampled. Reaction (4) was also included in the set of reactions in the SMR. Figure 3 illustrates the dependencies between the important outlet stream component flow rates from the steam–methane reactor. These components are the steam, hydrogen, CO, and CO_2 molar flow rates. Furthermore, calculated sampling domains using proportional constraints with hydrogen as the basis chemical component are included. The molar flow rate of hydrogen, $\dot{n}_{H_2}$, and the ratio between hydrogen and CO, $R_{H_2/CO}$, the ratio between hydrogen and CO_2, R_{H_2/CO_2}, and the ratio between hydrogen and steam, R_{H_2/H_2O} are declared as new independent variables. The ratios result in box constraints values for CO, CO_2, and steam. Note that, carbon monoxide and carbon dioxide are important chemical components in the feed stream since both are

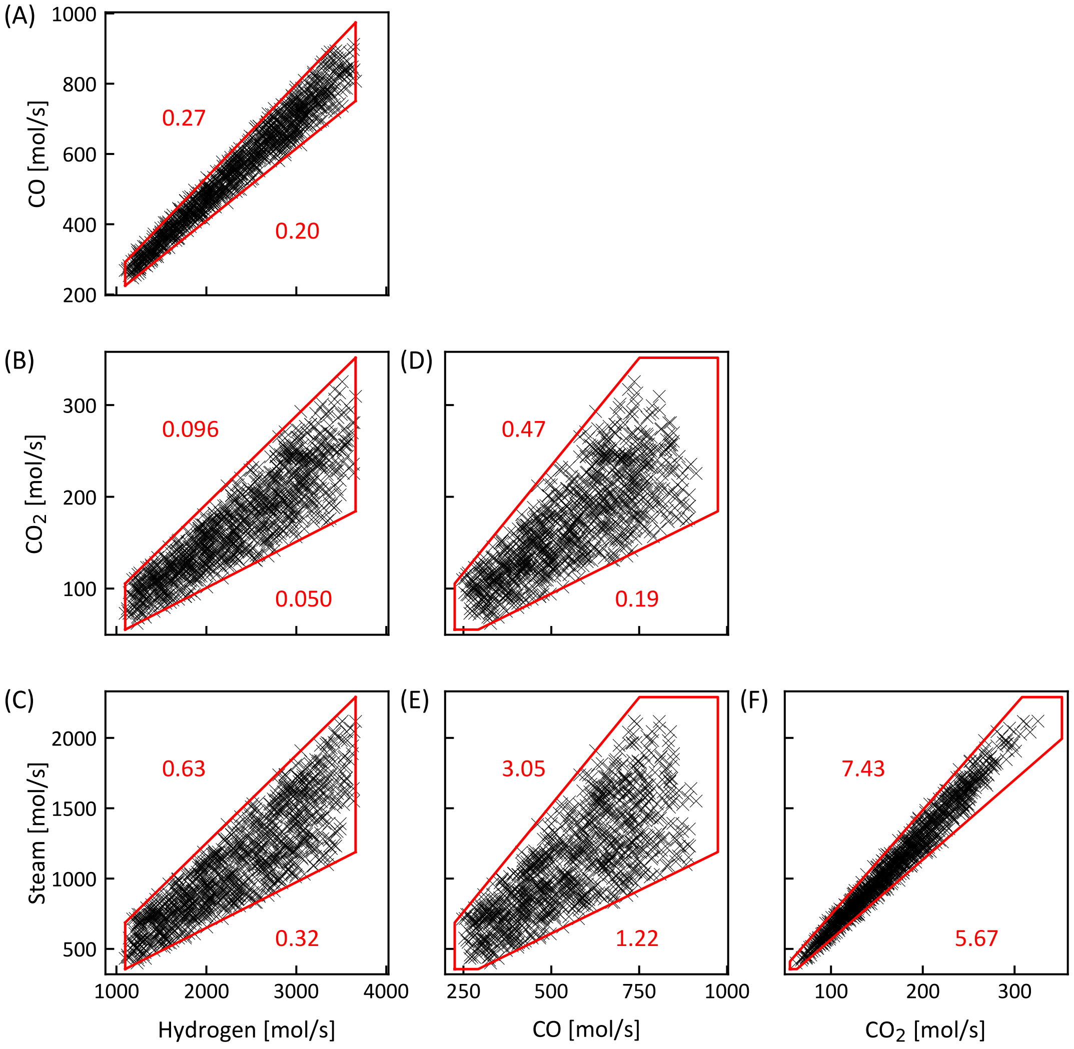

FIGURE 3 Illustration of the dependencies between the component flow rates of (A) CO and hydrogen, (B) CO_2 and hydrogen, (C) steam and hydrogen, (D) CO_2 and CO, (E) steam and CO, and (F) CO_2 and steam in the reactor outlet of a steam–methane reformer. The plotted line correspond to the sampling domain defined using proportional constraints while the number correspond to calculated upper and lower bounds for the respective ratios [Color figure can be viewed at wileyonlinelibrary.com]

present in the water–gas shift Reaction (4). From Figure 3, we can draw the following conclusions:

1. Proportional dependencies between all chemical components dominate the system.
2. Using a change in independent variables from molar flow rates to proportional constraints results in sampling in regions not encountered in practice as the individual flow rates may be higher by introducing proportional dependencies (see Figure 2 in Section 2.2). One example is Figure 3(D)), where the sampling domain includes values both at the upper and lower boundary for both CO and CO_2 that are not originating from the sampled data.

Hence, utilizing only box constraints or proportional constraints may result in sampling in regions never observed in the model due to (a) the chemical reaction (when using only box constraints) or (b) due to the total flow rate in the system (when using only proportional constraints). Furthermore, bounds on the sum of the flow rates can be advantageous. Therefore, it is necessary to restrict the sampling domain by both box constraints and proportional constraints. The following inequality constraints for the sampling domain of the water–gas shift section can be hence derived, which can be generalized to an arbitrary number of chemical components n_{chem}:

$$\min\left(\dot{\mathbf{n}}_i^{\mathrm{dat}}\right) \le \quad \dot{n}_i \quad \le \max\left(\dot{\mathbf{n}}_i^{\mathrm{dat}}\right) \tag{5}$$

$$\dot{n}_j \min\left(\dot{\mathbf{n}}_i^{\mathrm{dat}} \oslash \dot{\mathbf{n}}_j^{\mathrm{dat}}\right) \le \quad n_i \quad \le \dot{n}_j \max\left(\dot{\mathbf{n}}_i^{\mathrm{dat}} \oslash \dot{\mathbf{n}}_j^{\mathrm{dat}}\right) \tag{6}$$

$$\min\left(\dot{\mathbf{n}}_i^{\mathrm{dat}} + \dot{\mathbf{n}}_j^{\mathrm{dat}}\right) \le \dot{n}_i + \dot{n}_j \le \max\left(\dot{\mathbf{n}}_i^{\mathrm{dat}} + \dot{\mathbf{n}}_j^{\mathrm{dat}}\right) \tag{7}$$

In (5), $\dot{\mathbf{n}}_i^{\mathrm{dat}} \in \mathbb{R}^{n_{\mathrm{dat}}}$ corresponds to the inlet flow rate of chemical component i sampled from the outlet of the last subsection, n_{dat} to the number of sampled data from the previous section for creating the constraints, and $n_i \in \mathbb{R}$ to the inlet flow rate of chemical component i as used as an independent variable to the surrogate model. These inequalities hold $\forall i,j \in \{1,2,\dots,n_{\mathrm{chem}}\}$ such that $i \neq j$. The first set of inequality constraints represented in Equation (5) corresponds to box constraints, the second set (6) to proportional dependencies, and the third set (7) to inverse proportional dependencies. The set of constraints will always define a convex set and thus polytopic constraints in $\mathbb{R}^{n_{\mathrm{chem}}}$. The total number of constraints is $2\left(n_{\mathrm{chem}} + 2\sum_{i=1}^{n_{\mathrm{chem}}-1} i\right) = 2n_{\mathrm{chem}}^2$. The constraints (5)–(7) can be rearranged in the standard format $\mathbf{Ax} \le \mathbf{b}$, where $\mathbf{A}$ is a $2n_{\mathrm{chem}}^2 \times n_{\mathrm{chem}}$ matrix defined by the constraint inequalities, $\mathbf{x}$ is the vector $\dot{\mathbf{n}}$ of length n_{chem}, and $\mathbf{b}$ is a vector of length $2n_{\mathrm{chem}}^2$ corresponding to the leftmost and rightmost sides of the inequalities above.

Note that in the worst-case scenario, we have:

$$\min\left(\dot{\mathbf{n}}_i^{\mathrm{dat}} \oslash \dot{\mathbf{n}}_j^{\mathrm{dat}}\right) = 0 \quad \max\left(\dot{\mathbf{n}}_i^{\mathrm{dat}} \oslash \dot{\mathbf{n}}_j^{\mathrm{dat}}\right) = \infty$$

which corresponds to the trivial statements $0 \le \dot{n}_i \le \infty$ for dependency constraints (6). Hence, the dependency constraints will not be active in the sampling in this limit. Similarly, the worst case scenarios for the inverse proportional dependencies are:

$$\begin{aligned} \min\left(\dot{\mathbf{n}}_i^{\mathrm{dat}} + \dot{\mathbf{n}}_j^{\mathrm{dat}}\right) &= \min\left(\dot{\mathbf{n}}_i^{\mathrm{dat}}\right) + \min\left(\dot{\mathbf{n}}_j^{\mathrm{dat}}\right) \\ \max\left(\dot{\mathbf{n}}_i^{\mathrm{dat}} + \dot{\mathbf{n}}_j^{\mathrm{dat}}\right) &= \max\left(\dot{\mathbf{n}}_i^{\mathrm{dat}}\right) + \max\left(\dot{\mathbf{n}}_j^{\mathrm{dat}}\right) \end{aligned}$$

In the worst case scenario, Equation (7) reduces to simple box constraints. If desired, it is possible to include a slack on the constraints. The slack would correspond to utilizing a slightly bigger value for the maximum or minimum values in Equations (5–7) to account for limits in the number of sampled data points. The slack can then represent unsampled extreme values in the flow rates.

Similar conclusions can be drawn for several inlet streams to a submodel. In this case, proportional and inverse proportional dependencies are frequently encountered between streams. As an example, the steam flow to a SMR is often considered a separate stream.

3 | IMPLEMENTATION OF SAMPLING-DOMAIN REDUCTION

3.1 | Adaptive sampling

The best approach for incorporating the aforementioned type of constraints for sampling-domain reduction in surrogate-model generation is using an optimization-based algorithm as frequently used in *adaptive sampling routines*. In this type of algorithms, regions of the input space are iteratively sampled based on the current fit of the surrogate model (i.e., prediction error), the set of already sampled points (i.e., degree of exploration of the input space), and the required fidelity of the model.

In the following sections, we will use the following notation:

- $\mathbf{x} \in \mathbb{R}^d$ corresponds to the input variables to the true and surrogate models, where d is the input dimension. In the notation of Figure 1, $\mathbf{x}$ is constructed from the components of $\mathbf{z}_{k,i}$ and $\mathbf{u}_i$.
- $f : \mathbb{R}^d \to \mathbb{R}$ corresponds to a true model or simulator, and $\mathbf{f} : \mathbb{R}^d \to \mathbb{R}^r$ to a collection of such models, where r is the number of submodels. For instance, if we have one detailed model for calculating a temperature $T(\mathbf{x})$ and one for calculating a pressure $P(\mathbf{x})$, then $\mathbf{f} = (P,T)$ and $r = 2$.
- $g : \mathbb{R}^d \to \mathbb{R}$ or $\mathbf{g} : \mathbb{R}^d \to \mathbb{R}^r$ corresponds to a surrogate model used to approximate the above.
- $\boldsymbol{\theta}$ refers to the fitting parameters of a generic surrogate model g or $\mathbf{g}$. The size and dimension of this object depend on the chosen surrogate model class.

In this article, we explore *error-maximization sampling* that seeks to the sample regions of the input space with the poorest fit of the surrogate model.[1] In its simplest form, error-maximization sampling selects new samples by solving the unconstrained optimization problem

$$\hat{\mathbf{x}} = \operatorname{argmax}[f(\mathbf{x}) - g(\mathbf{x}|\theta)]^2 \tag{8}$$

Note that we only consider scalar functions $f(\mathbf{x})$ and $g(\mathbf{x})$ here; the higher-dimensional generalization will be treated in Section 3.2. Solving (8) requires for most cases using a DFO solver since evaluating $f(\mathbf{x})$ is obtained by running a simulator with no derivative information available.

A challenge with the error-maximization sampling in its form (8) is that it tends to sample the same regions when applied iteratively in the surrogate-model generation. While this does locate local maxima in the model error, which can be valuable for constructing more accurate surrogate models, it will also consume substantial computational resources without extending the exploration of the input space. Exploration–exploitation wise, sequential sampling by solving (8) to select new samples points iteratively is heavily biased toward exploitation. To remedy this clustering of sampling points, we propose adding a penalty to the objective function, which causes the algorithm to avoid previous sampling points $\mathbf{x}_k$. Additionally, we add, as introduced in the previous sections, linear constraints on the input variables. This gives the optimization problem

$$\hat{\mathbf{x}} = \operatorname{argmax}\left\{ [f(\mathbf{x}) - g(\mathbf{x}|\theta)]^2 - \mu \sum_k h(\mathbf{x} - \mathbf{x}_k) \;\middle|\; \mathbf{A}\mathbf{x} \le \mathbf{b} \right\} \tag{9}$$

Here, the linear inequality constraint $\mathbf{A}\mathbf{x} \le \mathbf{b}$ corresponds to the sampling domain constraints described in Equations (5–7). $\mu \ge 0$ is a nonnegative penalty parameter. The penalty function h, summed over previously sampled points $\mathbf{x}_k$, should satisfy three criteria. First, we should require that $h(\mathbf{u}) \to 0$ as $\|\mathbf{u}\| \to \infty$, to prevent penalizing sampling-point selection away from all existing points. Second, we should require that $\forall \mathbf{u} : h(\mathbf{u}) > 0$, so that it does indeed punish instead of rewarding clustering of sampling points. Third, $h(\mathbf{u})$ should preferably be a monotonically decreasing function of $\|\mathbf{u}\|$. There are many ways to select such a penalty function, including exact and nonexact, smooth penalty functions.[25] In our implementation, we use the latter type in the form of a Gaussian (exponential) penalty function,

$$h(\mathbf{u}) = \exp\left(-\mathbf{u}^2/\rho^2\right) \tag{10}$$

where $\rho > 0$ is a scalar parameter that limits how far away from previous sampling points such a penalty should occur, while μ parametrizes the exploration–exploitation trade-off in the sampling.

If $\mathbf{x}$ is normalized, then the squared differences $\|\mathbf{u}\| \le 1$. In the most extreme edge case, where one point is at the origin $(0,0,\ldots)$ and the other at the distant corner $(1,1,\ldots)$, then $h(\mathbf{u}) < 0.02$ when $\rho \lesssim \sqrt{d}/2$, where d is the dimension of $\mathbf{x}$. Since the penalty serves no purpose for points at opposite ends of the sampling area, this value is a reasonable upper limit on ρ. Below this value, any value of ρ can be justified, depending on what balance one desires between exploration and exploitation. For the penalty μ, a zero value reproduces nonpenalized error-maximization sampling (i.e., pure exploitation), while setting it to infinity causes sampling of points that are as far as possible away from each other (i.e., pure exploration). Since both are valid approaches, any non-negative value for the penalty μ is a valid choice, but the most useful is normally a balance between the exploration with exploitation terms.

3.2 | Generalization to higher dimensions

The method presented in Section 3.1 for sampling data for generating surrogate models of black-box functions is limited to scalar functions $f : \mathbb{R}^d \to \mathbb{R}$. In this subsection, we generalize our approach to functions $\mathbf{f} : \mathbb{R}^d \to \mathbb{R}^r$, d and r are the numbers of input and output dimensions, respectively. The algorithm presented here covers general class of processes: for instance, the components of matrices, tensors, complex numbers, and quaternions can all be trivially mapped onto real vectors in $\mathbb{R}^d$ or $\mathbb{R}^r$. The main processes that cannot be represented in this way are functions involving logic, integers, categories, or other discrete data, where different algorithms are required.

One surrogate model per output dimension must be created for processes with higher-dimensional output, following the same approach as outlined in Appendix A for Equation (A3). Thus, we can summarize our approximation as $\mathbf{f}(\mathbf{x}) \approx \mathbf{g}(\mathbf{x})$, where each submodel in the vector $\mathbf{g} = (g_0, g_1, \ldots)$ is defined as:

$$\forall m : \quad g_m(\mathbf{x}) := \sum_{\mathbf{n} \in \mathbb{N}^d} \theta_{m,\mathbf{n}} b_{\mathbf{n}}(\mathbf{x}) \tag{11}$$

Here, m is an index corresponding to the output dimension, and $\mathbf{n}$ is an index array corresponding to the basis function order in each input dimension. Each submodel g_m is then fitted using the procedures from Sections A.2 and A.3 in the Appendix. Note that, the number of nonzero model parameters $\theta_{m,\mathbf{n}}$ produced by these fits can be very different depending on the output m. This is because different observables such as pressure, temperature, and chemical composition may differ significantly as functions of the input variables. Each submodel g_m is also fitted using a different penalty λ_m, obtained via a separate optimization of, for example, the corrected Akaike information criterion (AIC).[26] Section A.3 in the Appendix discusses various information criteria and the and the implementation of the optimization procedure.

There are two ways to adjust the adaptive sampling procedure described in Section 3.1 to higher dimensions. One approach is *sequential error-maximization sampling*. Here, we iterate through output variables, and perform a separate error-maximization of each variable:

$$\forall m : \quad \hat{\mathbf{x}}_m = \operatorname{argmax}\left\{ [f_m(\mathbf{x}) - g_m(\mathbf{x}|\theta)]^2 - \mu \sum_k h(\mathbf{x} - \mathbf{x}_k) \;\middle|\; \mathbf{A}\mathbf{x} \le \mathbf{b} \right\} \tag{12}$$

where $\{\hat{\mathbf{x}}_m\}$ refers to new sample points obtained via error-maximization, while $\{\mathbf{x}_k\}$ refers to old sample points. After performing these error-maximization samplings, the submodels g_m are refitted using the new data. Data from all sampled points $\{\mathbf{f}(\hat{\mathbf{x}}_m)\}$ are used to improve all submodels g_m. Once a submodel g_m converges within the desired error tolerance, we stop performing new error-maximization samplings for that output variable.

As an alternative, *simultaneous error-maximization sampling* can be used, maximizing a weighted average error:[1]

$$\hat{\mathbf{x}} = \text{argmax}\left\{\sum_m w_m[f_m(\mathbf{x}) - g_m(\mathbf{x}|\theta)]^2 - \mu\sum_k h(\mathbf{x}-\mathbf{x}_k) \;\middle|\; \mathbf{A}\mathbf{x} \le \mathbf{b}\right\} \tag{13}$$

It is essential to introduce weights w_m in this approach. For instance, pressures typically vary by $\sim 10^6\,\text{Pa}$ and temperatures by $\sim 10^{2\circ}\text{C}$. Thus, if these output variables are not scaled accordingly, the error-maximization procedure would effectively *only* minimize errors in the pressure submodel, as its average error would likely be 3–4 orders of magnitude larger than the error in the temperature model. One simple and efficient solution is to set $w_m := \left(f_m^{\max} - f_m^{\min}\right)^{-1}$, where $f_m^{\max} := \max_k f_m(\mathbf{x}_k)$ and $f_m^{\min} := \min_k f_m(\mathbf{x}_k)$ are the maximum and minimum previously sampled values for that output.

We have implemented both approaches in `Consumet`.[22] In general, we expect simultaneous sampling to be more efficient if each submodel has (i) similar complexity and (ii) large variations in the same regions of input space. Conversely, if the high-error regions of each submodel are disjoint, it may be more efficient to perform sequential sampling. Moreover, if we have one trivial and one strongly nonlinear output, sequential sampling would quickly mark the simple one as "converged," while simultaneous sampling would waste computation time re-evaluating both.

3.3 | Implementation of Consumet

We have implemented the algorithm for automated surrogate model construction described in the previous subsections and in Section A in the Appendix in Python, using only free and open-source libraries, and have released the resulting tool `Consumet`[22] under an MIT open-source license. The source code is available at GitHub. A description of `Consumet` can be found in Section B in the Appendix.

`Consumet` uses the LASSO method for surrogate model fitting.[27] LASSO regression is a well-explored method for combined model fitting and model selection through the ℓ_1 penalty on coefficients for the basis functions. It has the advantage of only requiring the solution of a continuous optimization problem, is easy to implement, and does not require the use of licensed software. For a comparison and alternative regression and model selection methods, see, for example, Cozad et al.[1] or Bhosekar et al.[3] In the Consumet implementation, the penalty λ is autoselected via information criterion optimization as elaborated in Section A.3 in the Appendix.[28] Numerically, the regression problem is implemented in `Pyomo`,[29,30] and the resulting equations are solved using `IPOPT`.[31] The error-maximization problem (12) or (13) is solved using `NOMAD`,[32] which is based on the mesh-adaptive direct search algorithm.

4 | RESULTS

4.1 | Introduction

The investigated case studies are part of the methane reforming process for hydrogen production from natural gas. Due to the large selection of separation technologies for both CO_2 and H_2,[33] surrogate models for common process sections may allow a fast evaluation of a large number of combinations of separation technologies. In this respect, we created surrogate models for both a reforming section and the subsequent water–gas shift section as illustrated in the process flow diagram 4. The water–gas shift section, being downstream of the reforming section, allows a comparison of the proposed sampling domain reduction procedure developed in Section 2 with standard box constraints. Extents of reactions as described in Section 2.1 were used to achieve mass consistency in the surrogate model. The overall process was modeled in Aspen HYSYS.

We now discuss the procedure we used to construct and evaluate our surrogate models. Referring to Figure B1 in the Appendix, we first sampled 100 initial points via the *Batch sampling* routine. The *Adaptive sampling* box was allowed to sample the simulation maximum 10 times before proceeding to the *Convergence* check. At this point, we used the result of the error-maximization procedure as an estimate for the prediction error of our surrogate model. The convergence criterion, the relative prediction error, was set to 1%. If we executed the *Adaptive sampling* routine a total of 15 times without achieving convergence, the surrogate construction was terminated. Hence, the maximum number of sampling points was given by 250 sampling points. To determine the regression penalties, we used the corrected AIC.[26,34] All parameters of the surrogate model with a value smaller than 10^{-4} were discarded. This entire procedure was first applied to the reforming section to highlight the impact of adaptive sampling with LASSO regression and demonstrate the benefit of the clustering penalty. Afterward, it was repeated using both unconstrained sampling (box constraints) and constrained sampling (inequality constraints) for the water–gas shift section to compare the efficacy of these approaches. Simultaneous sampling was used for both the reforming and the water–gas shift section.

We created a validation set $\{\mathbf{x}_k\}$ by sampling $N = 2{,}000$ new points within the box constraint sampling domain (reforming section) and the constrained region (water–gas shift section) to evaluate the performance of the generated surrogate models. We then calculated the *relative absolute error* of each surrogate model:

$$\text{RAE}_k := \frac{|f(\mathbf{x}_k) - g(\mathbf{x}_k|\theta)|}{f_{\max} - f_{\min}} \tag{14}$$

Here, $f_{\max}$ and $f_{\min}$ were the maximum and minimum values for the true function $f(\mathbf{x})$ estimated from the full sample set. These results were used to determine the *maximum relative absolute error* and *relative root-mean-square error*:

$$\mathrm{MRAE} := \max_{k=1}^{N} \mathrm{RAE}_k \qquad \mathrm{RRMSE} := \sqrt{\frac{1}{N}\sum_{k=1}^{N}\mathrm{RAE}_k^2} \tag{15}$$

The performance of our surrogate generation method was then evaluated based on MRAE and RRMSE values, which highlight the accuracy of the models, and based on the number of iterations before termination and the number of parameters selected by Lasso, which highlight the complexity of the response surface.

4.2 | Reforming section

4.2.1 | Process description

The reforming section in a hydrogen production facility is responsible for converting all hydrocarbons in the natural gas feed. The main reforming reactor can be either an externally fired SMR or an ATR. Surrogate modeling enables changing between these two reactor types without developing new flowsheets.

The natural gas feed is preheated, mixed with intermediate-pressure (IP) steam, and fed to a prereformer to convert higher alkanes. Further IP steam is mixed with the prereformer outlet and then heated. Oxygen is provided from an air separation unit (ASU) and compressed via a compressor train with intermediate cooling. The combined stream is then fed to an ATR. The ATR outlet is cooled to the inlet temperature of the high-temperature water–gas shift reactor (HT-WGS). The boundaries of the corresponding surrogate models are outlined in Figure 4.

As surrogates are created for the overall section, black-box models have to be created for the rates of extent of reaction $\dot{\xi}$. The following elementary reactions take place in the reactors:

$$CH_4 + H_2O \rightleftharpoons CO + 3H_2 \tag{16}$$

$$C_2H_6 + 2H_2O \rightleftharpoons 2CO + 5H_2 \tag{17}$$

$$C_3H_8 + 3H_2O \rightleftharpoons 3CO + 7H_2 \tag{18}$$

$$C_4H_{10} + 4H_2O \rightleftharpoons 4CO + 9H_2 \tag{19}$$

$$CO + H_2O \rightleftharpoons CO_2 + H_2 \tag{20}$$

$$CH_4 + 2O_2 \rightleftharpoons CO_2 + 2H_2O \tag{21}$$

$$C_2H_6 + 3.5O_2 \rightleftharpoons 2CO_2 + 3H_2O \tag{22}$$

Note that, reactions (21) and (22) can only take place in the ATR as they require oxygen. The corresponding linear mass balances for the gray-box model are:

$$\dot{\mathbf{n}}_{\mathrm{out}} = \dot{\mathbf{n}}_{\mathrm{in}} + \boldsymbol{\nu}\dot{\boldsymbol{\xi}} \tag{23}$$

in which $\boldsymbol{\nu}$ corresponds to the matrix of stoichiometric coefficients of the chemical reactions as shown in the reaction Equations (16) and (22). Note that, the molar flows $\dot{\mathbf{n}}_{\mathrm{in}}$ are a summation of all inlet flows at point 1, that is, *Natural gas*, *IP stream*, and *Oxygen from ASU*. In total, seven surrogate models, had to be fitted to obtain an adequate representation of the reaction network.

The independent variables to the system are the inlet conditions of the natural gas, the IP steam, and the oxygen. We used proportional dependencies as outlined in Section 2.2 for the IP steam; that is, we introduce the steam to carbon ratio as a new independent variable instead of the flow rate of the IP steam. Similarly, we introduced a *control structure* for the oxygen flow rate from the ASU. The flow rate of oxygen was used as a *manipulated variable* for controlling the outlet temperature of the ATR unit operation; that is, the new independent variable was the reactor outlet temperature, and the oxygen

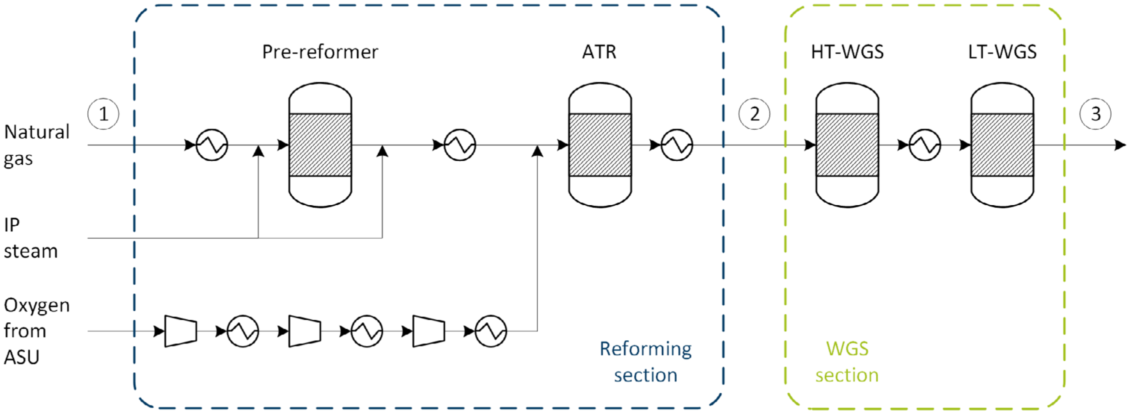

FIGURE 4 Process flow diagram for both the reforming section of an oxygen blown ATR and the water–gas shift (WGS) section with a tail gas recycle [Color figure can be viewed at wileyonlinelibrary.com]

TABLE 2 Bounds and units for the independent variables, reforming section

	$\dot{n}_{CH_4}$ (mol/s)	$\dot{n}_{C_2H_6}$ (mol/s)	$\dot{n}_{C_3H_8}$ (mol/s)	$\dot{n}_{CO_2}$ (mol/s)	$R_{H_2O/C}$ (mol/s)	$T_{ATR,out}$ (°C)
Lower bound	700	50	5	15	1.0	950
Upper bound	1200	100	15	30	2.0	1100

flow rate was a dependent variable. In total, six independent variables were identified. These are the composition of the natural gas feed (four variables: $\dot{n}_{CH_4}$, $\dot{n}_{C_2H_6}$, $\dot{n}_{C_3H_8}$, and $\dot{n}_{CO_2}$), the ratio between steam and carbon ($R_{H_2O/C}$), and the outlet temperature of the ATR ($T_{ATR,out}$).

4.2.2 | Surrogate model development

As our first case study, we construct surrogates for the rates of extent of reaction $\dot{\xi}$ for reactions (16) and (20) using simultaneous sampling. The former is the key reaction for hydrogen production in the reforming section, while the latter can be positive or negative depending on the process conditions. To construct our surrogates, we used Legendre polynomials of degree 3 as basis functions (see Section A.1 in the Appendix). Thus, the total number of model parameters can be up to 84. The rest of our simulation parameters, for example, convergence criteria, are described in Section 4.1. The utilized bounds for the box constraints for the independent variables are given in Table 2. The impact of the Gaussian clustering penalty described in Equation (10) can be best analyzed through error-maximization sampling without penalty and comparison of the results. Table 3 shows the key results for both dependent variables with and without clustering penalty. The program terminated after just two iterations of error–maximization sampling resulting in a total number of 120 sampling points when using the clustering penalty. This implies that the response surface is in general simple.

A potential reason for the simpler response surface can be the proportional dependency on the steam–carbon ratio. Furthermore, the change of variables from the oxygen flow rate to the ATR outlet temperature can simplify the response surface. The number of non-zero parameters is in the range of 50% of the total number of available parameters parameters for both surrogate models. Only 43 of the 84 parameters are used for the rate of extent of reaction of reaction (16) and 33 of the 84 for reaction (20). Hence, we can see that LASSO is efficient in reducing the number of parameters to avoid overfitting when the corrected AIC is used.

Without the clustering penalty, the error-maximization sampling stopped after five iterations corresponding to 150 sampled points, an increase of 30 sampling points compared to the case with clustering penalty. The resulting surrogate model has an error in the same range although the MRAE is reduced for the surrogate model of reaction (16) while the RRMSE is increased for both reactions. One potential reasoning for this behavior can be seen in the simultaneous sampling approach. The error-maximization sampling will focus on the surrogate model for reaction (16), as the MRAE is worse. Correspondingly, fewer points are sampled in the region in which the surrogate model of reaction (20) is worse, which resulted in an increase in the MRAE.

Analyzing the location of the sampling points reveals that we have indeed an accumulation of sampling points in certain regions of the sampling domain resulting in an improved fit in these regions at the cost of a reduced fit in other regions due to the focus on exploitation instead of exploration.

4.3 | Water–gas shift section

4.3.1 | Process description

The water–gas shift section is responsible for the conversion of CO to H_2 according to:

$$CO + H_2O \rightleftharpoons CO_2 + H_2 \tag{24}$$

The reaction is an exothermic equilibrium reaction. Hence, the highest CO conversions are achieved at lower temperatures. In current practice, two reactors are utilized, one at higher temperatures for an increased reaction rate (HT-WGS) and one at lower temperatures for an increased conversion (LT-WGS). However, depending on the process configuration, it may be beneficial to use one medium temperature water–gas shift reactor to reduce capital cost and increase the temperature of the outlet of the water–gas shift section. One medium temperature water–gas shift reactor is preferred, for example, when using high-temperature hydrogen selective membranes for hydrogen separation.[33]

The introduced exact mass balances are given as:

$$\dot{\mathbf{n}}_{out} = \dot{\mathbf{n}}_{in} + \boldsymbol{\nu}\dot{\xi} \tag{25}$$

For all nonreacting chemical components, Equation (25) can be reduced to:

$$\dot{\mathbf{n}}_{out} = \dot{\mathbf{n}}_{in} \tag{26}$$

The dependent process variables are therefore given by the single rate of extent of reaction $\dot{\xi}$, the outlet temperature of both reactors, and the duty of the heat exchanger between both water–gas shift reactors. The independent process variables are the inlet variables to the HT–WGS reactor, both composition through molar flow rates and feed temperature, and the inlet temperature of the LT-WGS reactor.

TABLE 3 Results of the adaptive sampling, reforming section

Clustering penalty	Sampling points	Dependent variable	Model parameters	MRAE (%)	RRMSE (%)
Yes	120	$\dot{\xi}$ of (16)	43	1.07	0.12
		$\dot{\xi}$ of (20)	33	0.41	0.09
No	150	$\dot{\xi}$ of (16)	32	0.67	0.16
		$\dot{\xi}$ of (20)	48	0.47	0.10

TABLE 4 Bounds and units for the independent variables, water–gas shift section

	$\dot{n}_{H_2O}$ (mol/s)	$\dot{n}_{H_2}$ (mol/s)	$\dot{n}_{CO}$ (mol/s)	$\dot{n}_{CO_2}$ (mol/s)	$T_{LT-WGS,in}$ (mol/s)	$T_{HT-WGS,in}$ ($^\circ$C)
Lower bound	842	1552	540	187	300	190
Upper bound	2645	2775	1078	510	400	300

TABLE 5 Results of the constrained and unconstrained adaptive sampling of the rate of extent of reaction $\dot{\xi}$ and water–gas shift section

Sampling type	Points total	Points inside (%)	MRAE (%)	RRMSE (%)
Unconstrained	250	3	2.70	0.54
Constrained	102	100	0.91	0.17

4.3.2 | Surrogate model development

In order to evaluate the impact of constrained sampling, both constrained and unconstrained adaptive sampling were conducted. We chose Legendre polynomials of degree 3 as our surrogate models and fitted these models using the procedure described in Section A in the Appendix with the parameters described in Section 4.1. The data for calculating constraints was obtained from the ATR model with 2000 sampling points. The number of sampling points is large and chosen to ensure that the constraints describe the complete polytope. Both the inequality constraints and box constraints were calculated directly from this data set; this makes it easier to compare the surrogate models, as we avoid extrapolation when comparing them. The bounds for the box constraints can be found in Table 4. We note that the polytope from the inequality constraints has a size of only 8% compared to box constraints. Hence, it was expected that constrained adaptive sampling should let us either (i) produce a better fit using a similar number of sampling points, and/or (ii) produce a similar fit using fewer sampling points. After fitting these surrogate models, validation was performed using 2000 points sampled within the constrained region. For simplicity, only results corresponding to the rate of extent of reaction $\dot{\xi}$ are shown.

Constrained sampling required 1 iteration of adaptive sampling, while unconstrained sampling did not converge within our chosen iteration limit. NOMAD may choose to sample points that violate the imposed inequality constraints. This is a consequence of using a progressive barrier function for constraint handling in the `NOMAD` solver, as recommended by Audet and Dennis,[35] thus allowing the solver to choose infeasible trial points during execution to speed up the convergence of the direct-search algorithm. As these points are outside the regions of interest and may be located in strongly nonlinear regions, we did not perform simulations at these points, nor were they included in the regression. Hence, after iteration 1, NOMAD had attempted to sample 110 points, but we only performed simulations at 102 of them.

Table 5 summarizes the key results from both the constrained and the unconstrained adaptive sampling. We see that, only $\sim 3\%$ of the sampled points were in the region of interest for the unconstrained sampling. Moreover, despite using nearly twice as many sample points to fit the surrogate, the RRMSE is three times larger. A second important observation corresponds to the number of parameters chosen by the LASSO regression. The constrained sampling rendered 41 out of 84 parameters nonzero, while for unconstrained sampling, the corresponding number was 80 parameters. The model shrinkage obtained with the constrained sampling indicates the reduced complexity of the corresponding surrogate model. The box plots of the relative absolute error in Figure 5 illustrate the distribution of the residuals. They further substantiate the advantage of constrained sampling: it significantly reduces both the median error and the maximum error, resulting in a significantly more accurate model. Note that, outliers (error larger than the whiskers) were removed.

The impact of using a different basis function than Legendre polynomials on the performance of constrained sampling was also tested. Using Taylor polynomials of order 3, we observed that constrained sampling converged after a single iteration of the error-maximization sampling, while the unconstrained sampling did not converge. The error was in the same range as the error of the surrogate models with Legendre polynomial as basis functions.

4.3.3 | Constrained sampling area

The required number of points for constrained sampling in surrogate model development depends on the process. Figure 6(A) shows the size of the constraints for hydrogen and CO with a varying number of data points, n_{dat}, to illustrate the changes.

As we can see, increasing the number of sampling points increased the total area in which points are sampled. Furthermore, we observed that certain constraints may be the same when increasing the number of sampling points. This conclusion can be explained by the fact that the constraints were calculated from individual points as outlined in Equations (5) to (7): if these points were included in the reduced data set, the constraints would be the same.

Figure 6(B) shows the relative size of the polytope (left axis) and the MRAE (right axis) as a function of the number of data points. The relative size is high even for a small number of data points (200), while the maximum relative error is essentially flat from 200 data points onward. Hence, it may not be necessary to sample a large number of data points for calculating the constraints, even if the surrogate models are then extrapolated when used. However, this cannot be generalized due to differences in the response surface: if the response surface is steep at the border of the constraints, it may lead to problems when the constraints do not represent the overall system sufficiently well.

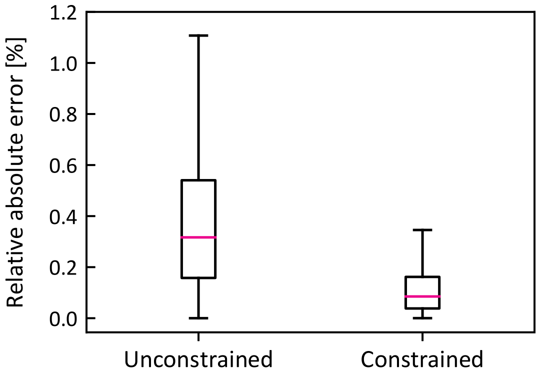

FIGURE 5 Box plots of the relative absolute error of the rate of extent of reaction $\dot{\xi}$ for unconstrained and constrained adaptive sampling. Pink lines indicate medians, boxes indicate interquartile range (IQR), and whiskers indicate the rest of the distribution within 1.5 IQR of the box endpoints. Outliers (error larger than the whiskers) are excluded [Color figure can be viewed at wileyonlinelibrary.com]

5 | DISCUSSION

Combining LASSO regression with error-maximization sampling is not new. Cozad et al. compared LASSO with the ALAMO framework and concluded that ALAMO is superior to LASSO, regardless of whether one uses error-maximization sampling or a single Latin-hypercube sampling.[1] However, our implementation fundamentally differs from theirs, as we determine an optimal regression penalty λ using an information criterion. Furthermore, we introduced a Gaussian penalty function to prevent oversampling small regions of the parameter space.

The combination of error-maximization sampling and sampling-domain reduction is our main contribution. Both reactions and separations always result in dependent component flow rates. Hence, using box constraints is generally not advisable for subsections, as they frequently require sampling of unimportant and potentially highly nonlinear regions. Particularly in the case of error-maximization sampling, this limitation of standard box constraints may result in the majority of the sampling points being outside the desired sampling region. In Section 4.3, we showed that our methodology could simultaneously reduce the number of sampling points by roughly a factor 2, and increase the accuracy of the final model by nearly one order of magnitude.

One explanation for the major difference in accuracy of the fitted surrogate models can be the chosen model order for the surrogate models. This would imply that 3rd-order polynomials cannot accurately describe the box-constrained sampling domain due to an increased nonlinearity of the response surface. In particular, when the order of the surrogate model is low, obtaining a better fit *outside* the region of interest necessarily implies a worse model *inside* the region of interest. Hence, a different set of basis functions or a higher model order would be required if the sampling domain is not constrained.

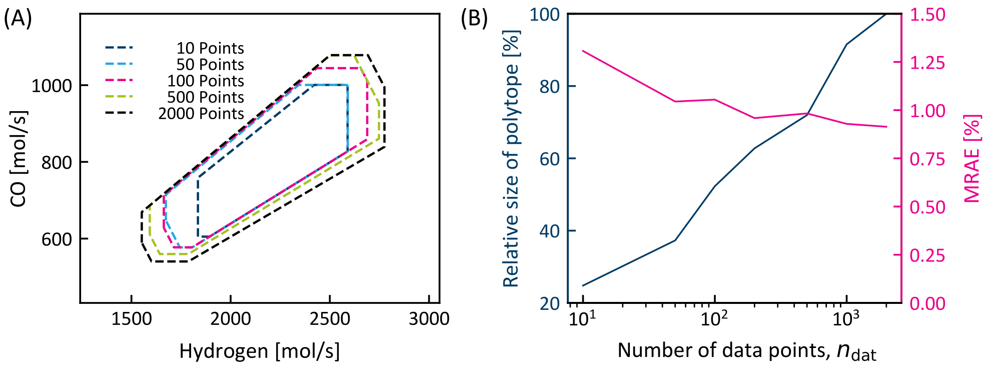

FIGURE 6 (A) Illustration of the constraints with varying number of data points used for the calculation of the constraints for CO and hydrogen and (B) the corresponding relative size of the sampling domain (left axis) and MRAE (of the validation set, right axis) as function of the number of data points. As comparison, the unconstrained relative size of the polytope corresponds to around 1300% with a MRAE of 2.7% [Color figure can be viewed at wileyonlinelibrary.com]

This is similar to the toy function utilized in Straus and Skogestad,[10] in which the change of independent variables reduced the required model order. A higher order of the surrogate model would furthermore require more sampling points due to the larger number of model parameters. Correspondingly, despite a higher proportion of the sampling points being inside the constrained sampling domain, the compromise for fitting a surrogate model within the entire box may be a worse fit in the region of interest.

Identifying constraints on the independent variables requires analysis of data from previous unit operations. Hence, it is necessary to sample data. This can be computationally expensive and complicated depending on the flowsheet topology. The previous section illustrated the impact of the number of data points on the performance of the surrogate model and showed that excessive sampling of data points may be unnecessary. However, if the overall process is converted into several surrogate models in superstructure optimization, all surrogate models will be located after a first surrogate model. As a result, it is possible to obtain these required data from sampling of preceding surrogate models, which reduces the computational cost for obtaining the data. Furthermore, even if box constraints are applied, it is necessary to carefully select their values. Hence, to prevent extrapolation, it may be always necessary to obtain data from previous unit operations. Depending on the process, it may be sufficient to sample only corner points of the previous unit operations, that is, when the constraints are at the upper or lower bounds.

6 | CONCLUSION

We have in this article developed a methodology for constructing surrogate models based on penalized regression, error-maximization sampling, and sampling-domain reduction. The methodology has been implemented as a free and open-source software solution in Python.[22] This package uses only the free and open-source components Pyomo, IPOPT, and NOMAD, and thus does not require any commercial software. The implementation is completely modular with the benefit that new or extended functionality, for example, new basis functions or model selection criteria, can easily be added. A core feature of the implementation is the automated construction of linear constraints for the sampling region resulting in improved fits of the surrogate models.

As a case study, we applied our software package to hydrogen production simulations. Specifically, we constructed surrogate models for the reforming and water–gas shift sections. The results showed that sampling domain reduction increased surrogate model accuracy by up to one order of magnitude and reduced the number of simulation sampling points by roughly a factor 2. Both these effects can be attributed to a more targeted sampling of important regions of the input space.

ACKNOWLEDGMENTS

The authors gratefully acknowledge funding from the ACT project ELEGANCY. ACT ELEGANCY, project No. 271498, has received funding from DETEC (CH), BMWi (DE), RVO (NL), Gassnova (NO), BEIS (UK), Gassco, Equinor, and Total, and is cofunded by the European Commission under the Horizon 2020 programme, ACT Grant Agreement No. 691712.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

DATA AVAILABILITY STATEMENT

Data available on request from the authors.

ORCID

Julian Straus https://orcid.org/0000-0001-8622-1936
Jabir Ali Ouassou https://orcid.org/0000-0002-3725-0885
Brage Rugstad Knudsen https://orcid.org/0000-0001-5981-9879
Rahul Anantharaman https://orcid.org/0000-0001-9228-6197

REFERENCES

1. Cozad A, Sahinidis NV, Miller DC. Learning surrogate models for simulation-based optimization. *AIChE J*. 2014;60:2211-2227.
2. Quirante N, Javaloyes-Antón J, Caballero JA. Hybrid simulation-equation based synthesis of chemical processes. *Chem Eng Res and Des*. 2018;132:766-784.
3. Bhosekar A, Ierapetritou M. Advances in surrogate based modeling, feasibility analysis, and optimization: a review. *Comp Chem Eng*. 2018; 108:250-267.
4. McBride K, Sundmacher K. Overview of surrogate modeling in chemical process engineering. *Chem Ing Tech*. 2019;91:228-239.
5. Boukouvala F, Floudas CA. ARGONAUT: AlgoRithms for global optimization of coNstrAined grey-box compUTational problems. *Opt Lett*. 2017;11:895-913.
6. Caballero JA, Grossmann IE. An algorithm for the use of surrogate models in modular howsheet optimization. *AIChE J*. 2008;54:2633-2650.
7. Eason J, Cremaschi S. Adaptive sequential sampling for surrogate model generation with artificial neural networks. *Comp Chem Eng*. 2014;68:220-232.
8. Straus J, Skogestad S. A new termination criterion for sampling for surrogate model generation using partial least squares regression. *Comp Chem Eng*. 2019;121:75-85.
9. Grimstad B, Foss B, Heddle R, Woodman M. Global optimization of multiphase flow networks using spline surrogate models. *Comp Chem Eng*. 2016;84:237-254.
10. Straus J, Skogestad S. Surrogate model generation using self-optimizing variables. *Comp Chem Eng*. 2018;119:143-151.
11. Garud SS, Karimi IA, Kraft M. Design of computer experiments: a review. *Comp Chem Eng*. 2017;106:71-95.
12. Krige DG. *A Statistical Approach to Some Mine Valuations and Allied Problems at the Witwatersrand* [MA thesis]. South Africa: University of Witwatersrand; 1951.
13. Ochoa-Estopier LM, Jobson M, Smith R. The use of reduced models for design and optimisation of heat-integrated crude oil distillation systems. *Energy*. 2014;75:5-13.
14. Crombecq K, Gorissen D, Deschrijver D, Dhaene T. A novel hybrid sequential design strategy for global surrogate modeling of computer experiments. *SIAM J Sci Comp*. 2011;33:1948-1974.
15. Garud SS, Karimi I, Kraft M. Smart sampling algorithm for surrogate model development. *Comp Chem Eng*. 2017;96:103-114.
16. Garbo A, German BJ. A model-independent adaptive sequential sampling technique based on response nonlinearity estimation. *Struct Multidiscip Optim*. 2020;61:1051-1069.

17. Li G, Aute V, Azarm S. An accumulative error based adaptive design of experiments for offline metamodeling. *Struct Multidiscip Optim*. 2010;40:137-155.
18. Zhou Q, Shao X, Jiang P, Gao Z, Zhou H, Shu L. An active learning variable-fidelity metamodelling approach based on ensemble of metamodels and objective-oriented sequential sampling. *J Eng Design*. 2016;27:205-231.
19. Parr JM, Forrester AI, Keane AJ, Holden CM. Enhancing infill sampling criteria for surrogate-based constrained optimization. *J Comp Methods Sci Eng*. 2012;12:25-45.
20. Cozad A, Sahinidis NV, Miller DC. A combined first-principles and data-driven approach to model building. *Comp Chem Eng*. 2015;73:116-127.
21. Straus J, Skogestad S. Use of latent variables to reduce the dimension of surrogate models. In: Espuña A, Graells M, Puigjaner L., eds. *Computer Aided Chemical Engineering*. Vol. 40. 27th European Symposium on Computer Aided Process Engineering; 2017; Elsevier; 445-450.
22. Ouassou JA, Straus J, Knudsen BR, Anantharaman R. Consumet: constructor of surrogates and metamodels. https://github.com/act-elegancy/consumet. Accessed: May 19, 2021.
23. Henao CA, Maravelias CT. Surrogate-based superstructure optimization framework. *AIChE J*. 2011;57:1216-1232.
24. Straus J, Skogestad S. Variable reduction for surrogate modelling. Proceedings of Foundations of Computer-Aided Process Operations 2017; January 8-12, 2017; Tucson, AZ.
25. Conn AR, Gould NIM, Toint PL. *Trust-Region Methods*. Philadelphia, PA: SIAM; 2000.
26. Hurvich CM, Tsai CL. Regression and time series model selection in small samples. *Biometrika*. 1989;76:297-307.
27. Tibshirani R. Regression shrinkage and selection via the Lasso. *J R Stat Soc B (Methodological)*. 1996;58:59-76.
28. Ding J, Tarokh V, Yang Y. Model Selection Techniques: An Overview. *IEEE Signal Processing Magazine*. 2018;35(6):16-34. http://doi.org/10.1109/msp.2018.2867638.
29. Hart WE, Watson JP, Woodruff DL. Pyomo: modeling and solving mathematical programs in python. *Mathemat Prog Comp*. 2011;3:219-260.
30. Hart WE, Laird CD, Watson JP, et al. *Pyomo–Optimization Modeling in Python*. Vol 67. 2nd ed. Springer; 2017.
31. Wächter A, Biegler LT. On the implementation of an interior-point filter line-search algorithm for large-scale nonlinear programming. *Mathemat Prog*. 2006;106:25-57.
32. Le Digabel S. Algorithm 909: NOMAD: nonlinear optimization with the MADS algorithm. *ACM Trans Math Softw*. 2011;37:44:1-44:15.
33. Voldsund M, Jordal K, Anantharaman R. Hydrogen production with CO_2 capture. *Int J Hydrogen Energy*. 2016;41:4969-4992.
34. Akaike H. Information theory and an extension of the maximum likelihood principle. In: Petrov BN, Csaki F., eds. Proceedings of the 2nd International Symposium on Information Theory; 1973; Budapest: Akademiai Kiado; 267-281.
35. Audet C, Dennis J Jr. A progressive barrier for derivative-free nonlinear programming. *SIAM J Opt*. 2009;20:445-472.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

How to cite this article: Straus J, Ouassou JA, Knudsen BR, Anantharaman R. Constrained adaptive sampling for domain reduction in surrogate model generation: Applications to hydrogen production. *AIChE J*. 2021;e17357. https://doi.org/10.1002/aic.17357