ACS OMEGA

http://pubs.acs.org/journal/acsodf

Article

Bayesian Symbolic Learning to Build Analytical Correlations from Rigorous Process Simulations: Application to CO_2 Capture Technologies

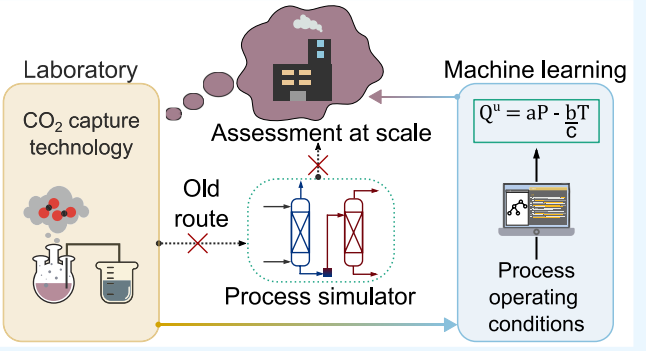
Valentina Negri, Daniel Vázquez, Marta Sales-Pardo, Roger Guimerà,* and Gonzalo Guillén-Gosálbez*

Cite This: *ACS Omega* 2022, 7, 41147−41164

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Process modeling has become a fundamental tool to guide experimental work. Unfortunately, process models based on first principles can be expensive to develop and evaluate, and hard to use, particularly when convergence issues arise. This work proves that Bayesian symbolic learning can be applied to derive simple closed-form expressions from rigorous process simulations, streamlining the process modeling task and making process models more accessible to experimental groups. Compared to conventional surrogate models, our approach provides analytical expressions that are easier to communicate and manipulate algebraically to get insights into the process. We apply this method to synthetic data obtained from two basic CO_2 capture processes simulated in Aspen HYSYS, identifying accurate simplified interpretable equations for key variables dictating the process economic and environmental performance. We then use these expressions to analyze the process variables' elasticities and benchmark an emerging CO_2 capture process against the business as usual technology.



1. INTRODUCTION

In the current emissions reduction scenario and transition toward a greener energy system, sustainable technology development has become key in every industrial sector. Nonetheless, the diffusion and application of new technologies is a lengthy process requiring multiple intermediate steps,[1,2] from the early conceptualization and planning phase to laboratory testing, pilot scale, and industrial operation. Moreover, every step calls for specific experimental and modeling skills and tools in the quest for more sustainable technologies.

Standard Process Systems Engineering (PSE) tools and, more recently, also machine learning (ML) methods are being used at different stages of such technology development process to assist in the transition from laboratory to pilot or industrial scale. Notably, a critical step for scientists at the early development stage is to compare the performance of a novel technology relative to the business as usual (BAU). Information on emerging and established technologies might not always be readily available, making it necessary to generate *in silico* data using modeling tools to ensure meaningful comparisons. In this context, experimentalists often collaborate with modeling experts to conduct technoeconomic assessments of competing technologies. These analyses might be challenging and time-consuming, particularly when process simulations need to be developed from scratch and/or lead to convergence issues. In this context, simple closed-form mathematical expressions describing the performance of technologies could simplify preliminary technoeconomic and environmental assessments during the early stages of technology development, avoiding the need for complex simulations. In addition to being easier to develop and use compared to rigorous simulations, such equations could also be employed for simplifying the optimization of the original processes, feasibility analyses, and hybrid modeling.

Among the wide range of emerging technologies under investigation, here we focus on CO_2 capture processes. This technology, which is expected to play a significant role in meeting the Paris agreement goals,[3,4] has been the focus of intense modeling efforts. Applications of CO_2 capture include flue gas treatment (e.g., pre- or post-combustion),[5] process streams purification (e.g., natural gas sweetening),[6] and CO_2 removal from the atmosphere (e.g., direct air carbon capture and storage (DACCS)).[7] Among all of the available options for CO_2 capture, post-combustion chemical absorption using amine-based solvents, historically developed to remove CO_2 and hydrogen sulfides from natural gas,[8−10] is considered the

Received: July 27, 2022
Accepted: October 11, 2022
Published: November 2, 2022

https://doi.org/10.1021/acsomega.2c04736
ACS Omega 2022, 7, 41147−41164

most mature technology. Despite its high technology readiness level, chemical absorption still leads to significant energy requirements due to the solvent regeneration step.[11] Consequently, novel solvents,[12] hybrid configurations,[13] and new strategies aiming at reducing energy consumption are under investigation.[14,15]

Peters and co-workers carried out a technoeconomic analysis to compare chemical absorption with membrane technologies for a natural gas sweetening process using Aspen HYSYS.[16] Two different inlet gases were tested, and the processes were optimized to reduce capital costs. The results showed that absorption leads to higher purity in the vented and sold gases (at the expense of higher capital costs). Other studies have also used Aspen HYSYS to optimize the CO_2 capture cost considering multiple configurations based on membranes, i.e., number of stages and recycle streams.[17,18] Hasan and co-authors[19] modeled and optimized a flue gas dehydration and CO_2 capture process based on absorption and membranes. They concluded that the CO_2 composition and gas flow rate dictate the most suitable technology. Other works modeled separations of CO_2/N_2 mixtures based on membranes to minimize the membrane area and energy consumption.[20] Hybrid configurations of membranes and cryogenic processes were also investigated to improve the energy consumption compared to monoethanolamine (MEA) absorption for flue gas mixtures with CO_2 content within 12−25%.[21] The number of publications in this area (over 7000 on hybrid CO_2 capture technologies in the last decade, 16% of them about membrane-based processes[22]) highlights the scientific community's great interest in alternative, more sustainable capture processes. Standalone cryogenic separation of CO_2 from flue gas[23] and hydrate-based gas separation[24] have also been studied, showing promising results in terms of energy consumption.

Traditionally, technoeconomic assessments of these and other chemical technologies have been based on first-principles models. Aspen Plus, Aspen HYSYS, and gPROMS are examples of process simulation packages based on mass and energy balances, transfer phenomena, and thermodynamic equations widespread in the modeling and optimization of chemical processes and energy systems. However, the advent of ML algorithms has opened new avenues for data-driven process modeling. Artificial neural networks (ANN), Gaussian processes, and random forest, among others, are increasingly being used in process modeling,[25−28] mostly to simulate complex unit operations hard to model based on first principles. For example, modeling bioreactors following complex kinetics is challenging and might be simplified using pure data-driven or hybrid models.[29−32] These approaches lead to mathematical models that often provide good approximations for time-constrained applications but are hard to interpret due to the absence of closed analytical expressions. Additionally, the ability to extrapolate is usually limited.

Analytical expressions can be explicitly obtained from data using symbolic regression, an application of genetic programming where the algorithm is trained to solve high-level problems combining simple functions.[33] Later, the expressions can be manipulated algebraically, differentiated, and more easily interpreted to generate valuable insights into the underlying principles governing the phenomena observed. As discussed in more detail later in the article, standard symbolic regression approaches rely on symbolic regression trees, i.e., superstructures of mathematical expressions, which can be coupled with optimization algorithms to find the best possible models. These representations can be optimized using either deterministic or stochastic optimizers. Deterministic methods guarantee convergence to a local solution or even to the global optimum within an epsilon tolerance.[34] In contrast, stochastic methods need to be run for an infinite time to guarantee global optimality, yet they tend to lead to lower CPU times to provide a satisfactory solution.

The pioneering ALAMO algorithm (automated learning of algebraic models for optimization) emerged in the PSE literature to address the symbolic regression problem using mixed-integer linear programming (MILP).[35] This work was enlarged in scope to include *a priori* physical knowledge[36] and applied to a range of chemical reaction problems.[37] The main limitation of ALAMO stems from the use of a finite number of basis functions. This assumption constrains the search space drastically, eventually hindering the algorithm's ability to reproduce the data precisely. Cozad and Sahinidis overcame this shortcoming by formulating an elegant mixed-integer nonlinear programming (MINLP) model for symbolic regression that can be solved with deterministic optimization methods like the nonlinear branch and bound and outer approximation algorithms.[38] Moreover, deterministic global optimizers (e.g., BARON) can also be applied to this MINLP to compute rigorous bounds on the minimum error that could be attained in the best possible regression model in the symbolic tree.[33]

To the best of our knowledge, the first studies that aimed at identifying interdependencies of process variables in CO_2 capture and storage (CCS) systems were presented by Rao et al.[39] and Zhou et al.[40], who applied the response surface and multiple-regression techniques, respectively. Zhou and co-workers later applied ANN and neurofuzzy modeling to the same set of pilot plant data.[41] The predictive accuracy of the models developed by Zhou et al. using the aforementioned techniques ranges between 70 and 99%.[42] The response surface methodology has been used later in other works to retrieve technical and technoeconomical equations from CCS process simulation data.[43] Focusing on examples that employ symbolic regression, a very recent application of ALAMO to post-combustion CO_2 capture using an MEA solvent was proposed by Danaci and co-authors,[44] who provided the capture costs for a range of input conditions. This work is based on an accurate model of the system and it explores a wide spectrum of operating conditions. However, it suffers from the limitations of the algorithm described above. The works by Pascual-González et al.[45] and Miró et al.[46] also applied mixed-integer programming (MIP) to address symbolic regression problems constrained within the limits of a reduced set of canonical expressions. In addition, the generation of tree regression models has been investigated, where the size of a tree can be controlled to balance the model's accuracy and complexity.[47] Differently, Ferreira and co-workers applied Kaizen Programming to solve symbolic regression problems, obtaining multioutput models in a single run, which were tested experimentally.[48] Recently, an MINLP for symbolic regression successfully recovered the relationship between shear stress and shear rate for both Newtonian and non-Newtonian fluids and chemical reactions kinetic laws.[49] Ansari and colleagues investigated relationships between variables in computational fluid dynamics simulations combining artificial intelligence and symbolic regression using sure-independence screening and sparsifying operators.[50,51] Lastly, linear sparse regression techniques, such as LASSO or elastic

https://doi.org/10.1021/acsomega.2c04736
ACS Omega 2022, 7, 41147−41164

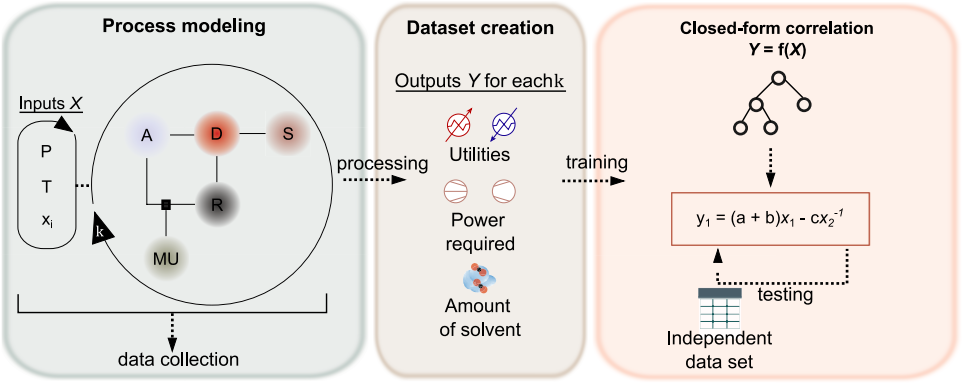


Figure 1. Sketch of the methodology adopted in this work. We first develop a process flowsheet, which is used to obtain data on key process variables Y linked to the economic and environmental performance using a set of inputs X (pressure (P), temperature (T), composition (x_i)). Then, the data is processed to obtain a dataset used for the BMS algorithm training. Later, the equations derived from the data by the BMS are validated using an independently generated dataset. The flowsheet encompasses absorption (A), desorption (D), CO_2 storage (S), recycle (R), and makeup (MU) units.

nets, can be deployed as an alternative to MI(N)LP formulations. ALVEN[52] is a recent approach part of the SPA framework[53] based on these methods, which was explicitly designed for modeling manufacturing data. An exhaustive literature review of ML models in chemical engineering, comparing strengths and weaknesses of the previous cited approaches, was given by Dobbelaere et al.[54] and Schweidtmann et al.,[55] while a methodological review of interpretability in machine learning was presented by Otte.[56]

Moving to stochastic symbolic regression, some of us introduced a Bayesian machine scientist (BMS)[57] for symbolic regression in a recent publication. Unlike genetic programming, this approach uses a Markov chain Monte Carlo (MCMC) algorithm and a principled performance metric, the description length, to find expressions representing a good balance between accuracy and model complexity. This algorithm proved to be more robust than other data-driven approaches also when data is scarce and noisy.[57]

Lately, the BMS has been employed to identify energy consumption and pollution drivers in an automated way in the work by Vázquez et al., outperforming the well-established STIRPAT empirical method.[58] Similarly, the relationship among emissions and economic parameters was previously assessed using symbolic regression, automatic identification, and search methods.[59,60]

Here, we apply this novel ML method based on symbolic regression to simplify the modeling of a CO_2 capture process, providing explicit equations that represent the interdependencies of variables in the whole system. Most of the existing models for CO_2 capture are based on first principles, and analytical expressions to streamline the calculations and enable more straightforward comparisons are missing,[43] such as in the work of Danaci et al.[44] and Subraveti et al.[61] From a survey of the literature as reported above, many applications of ML to PSE tackle very specific problems, often focusing on single process units or academic examples. Morgan and co-workers highlighted in their recent review that most of the applications of artificial intelligence or ML applied to CO_2 capture are about predicting physical properties, such as the components' miscibility and solubility,[62−64] rather than process performance,[65] e.g., CO_2 storage efficiency.[66] Among the few studies that analyze the latter aspect, the majority employs ANN and similar conventional ML tools.[67−69] Bearing this in mind, here, we apply the BMS to two CCS processes, generating closed-form expressions to estimate the economic and environmental performance considering the entire process for a range of feed conditions. In this first attempt, our results show that the BMS can be applied to identify simple analytical expressions that reproduce the process precisely and can easily be used to facilitate comparisons and carry out further in-depth analyses. Notably, these equations can be reworked or studied analytically using the concept of elasticity, borrowed from economics, to investigate the effect of the operating conditions on the process' performance, as shown at the end of this article. Our simplified equations could assist experimental scientists in benchmarking emerging CO_2 capture technologies, such as membranes, cryogenic separation, or adsorption,[70,71] in their early development stages. From a broader perspective, this work opens up new avenues to bridge the gap between modeling and experimental communities by simplifying the adoption of modeling tools by experimental groups and streamlining the modeling calculations. Moreover, our models can be applied to solve standard PSE problems, especially in the areas of surrogate-based process optimization, feasibility analysis, and hybrid modeling, by exploiting their analytical structure.

This article is structured as follows. In the next section, we state the problem of interest and introduce the two CCS case studies. Later, the methods employed to solve it are presented. Then, the results are analyzed and discussed for both cases. Lastly, we show two possible applications of the obtained surrogate models and discuss their use in different PSE areas before the conclusions and outlook for future works.

2. PROBLEM STATEMENT

Figure 1 outlines the overall methodology adopted here. In essence, we are interested in generating simple analytical equations from process simulation data, which experimentalists could use to evaluate their technologies. We consider a process simulation model implemented in Aspen HYSYS, which we run iteratively to generate $|K|$ scenarios for different inlet conditions. This data shall then be used to build an analytical expression reproducing the model precisely, as explained later.

Let us consider a set of data points K, corresponding to experimental observations or generated *in silico* with a process model. These points are the basis for constructing the data-

https://doi.org/10.1021/acsomega.2c04736
ACS Omega 2022, 7, 41147−41164

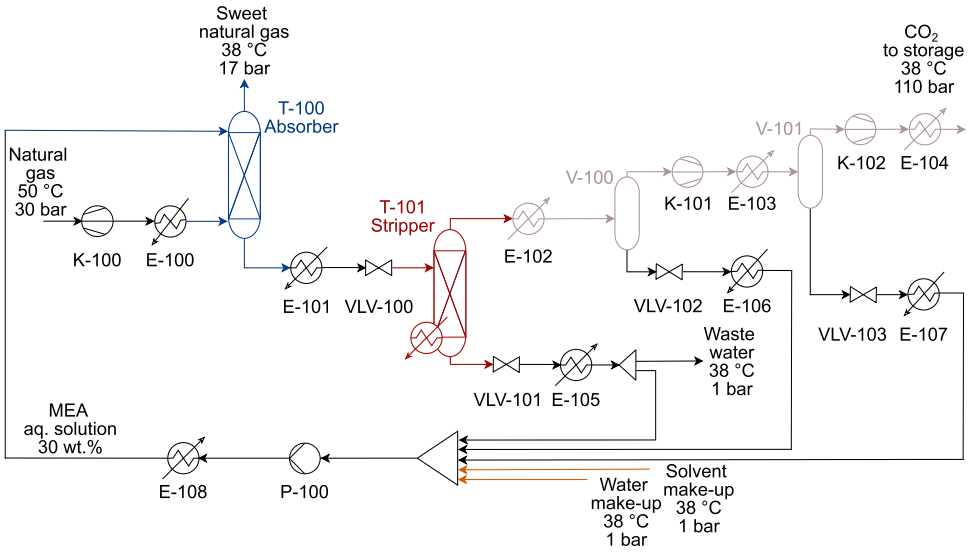


Figure 2. Natural gas sweetening process flow diagram. The process consists of an absorption stage (blue), desorption (red), CO_2 compression (sand), and recycle with solvent makeup (orange).

driven model. We classify the variables in the dataset as independent or dependent. The former refer to the degrees of freedom in the experimental setting (or process model), while the latter are obtained once the former are fixed, by either solving the process model or running the associated experiment. Let I denote the set of independent variables and J the set of dependent ones. The following notation is adopted: x_{ki} is the value of independent variable i in the observed point k, while y_{kj} is the value of the dependent variable j in the same point. Therefore, the independent data takes the form of a matrix with dimension $|K| \times |I|$, while the dependent data is represented by a matrix with dimension $|K| \times |J|$.

The analysis aims to find analytical expressions of the form given in eq 1 that predict the output data (values of the dependent variables, $\tilde{y}_{kj}$) from the input data, while minimizing the approximation error (e_{kj}) and the risk of overfitting.

$$y_{kj} = f(x_{k1}, \cdots, x_{ki}, \cdots, x_{k|I|}) + e_{kj} \quad \forall\, j \in J,\ k \in K \tag{1}$$

$$\tilde{y}_{kj} = f(x_{k1}, \cdots, x_{ki}, \cdots, x_{k|I|})$$

In eq 1, $f(x_{ki})$ is unknown, meaning that both the structure of the model and its parameters are to be learned from the data. Hereafter we refer to $f(x_{kj})$ as $\tilde{y}_{kj}$. Hence, three problems need to be solved simultaneously to find the best expressions. The first one is the features selection problem, i.e., identifying which independent variables are statistically relevant from the viewpoint of the dependent variables. The second problem is finding the model structure, i.e., identifying the best mathematical formulation to explain the data, which requires solving the previous task. Lastly, the third task is solving the parameter estimation problem, i.e., finding the best model parameters for a given mathematical structure. We next explain how to tackle these three problems simultaneously using the BMS.

3. METHODS

3.1. Process Models Used for Data Generation. We consider the natural gas sweetening and flue gas treatment processes simulated in Aspen HYSYS V11 with the *Acid gas—chemical solvents* fluid package. The synthetic data generated is based on simulation results at steady state. The Supporting Information provides more details about the assumptions and limitations of the model flowsheets.

The first case study represented in Figure 2 refers to the sweetening of natural gas with CO_2 sequestration and storage. The process model considers a feed of sour natural gas, the absorption and desorption columns operating with an MEA aqueous solution, and the CO_2 compression stage. The natural gas (4986 kmol h^{-1}) is assumed to be a binary mixture of CH_4 and CO_2, 80 and 20% molar fraction (mol), respectively, at 30 bar and 50 °C. For simplicity, the presence of H_2S in the feed stream is neglected. We note that the flowsheet is based on published studies.[16,72] Moreover, sensitivity analyses were carried out to adjust the operating conditions for our case study, as explained in the Supporting Information.

The process operation is as follows. First, the natural gas pressure is decreased from 30 to 17 bar in an expander. Then, the stream is heated up to the absorber operating condition (50 °C) and fed to the last stage, where CO_2 is recovered. The sweet natural gas is obtained at the top at 99.6% mol CH_4, meeting the standard required for pipeline injection and distribution. The CO_2-rich liquid stream is sent to the top of the stripper, where it is regenerated by CO_2 desorption with steam and subsequently recycled. The absorption and stripping columns operate at 17 and 11 bar, respectively, taking advantage of the inlet condition of the natural gas at high pressure. Indeed, in the first tower, a higher than atmospheric pressure favors the absorption of CO_2, while in the second one, higher pressure is meant to lower the reboiler duty, decreasing water evaporation, based on Schach et al.[73] Both columns are designed with 12 stages and are packed with plastic material to avoid corrosion due to the CO_2−MEA mixture.[74] The stripper operates without a condenser, while the reboiler energy consumption is 5.4 MJ/kg CO_2 removed for 90% mol CO_2 recovery. The lean load in the recycle is 0.053 mole CO_2/mole MEA. The CO_2-rich stream extracted at the top of the stripper is compressed to a supercritical state (110 bar and 38 °C) for pipeline transportation and injection at a selected storage site (not considered in this work) with a purity of 99% mol. The MEA and water makeup maintain the solvent solution in the recycle at 30% wt MEA at 38 °C.

https://doi.org/10.1021/acsomega.2c04736
ACS Omega 2022, 7, 41147−41164

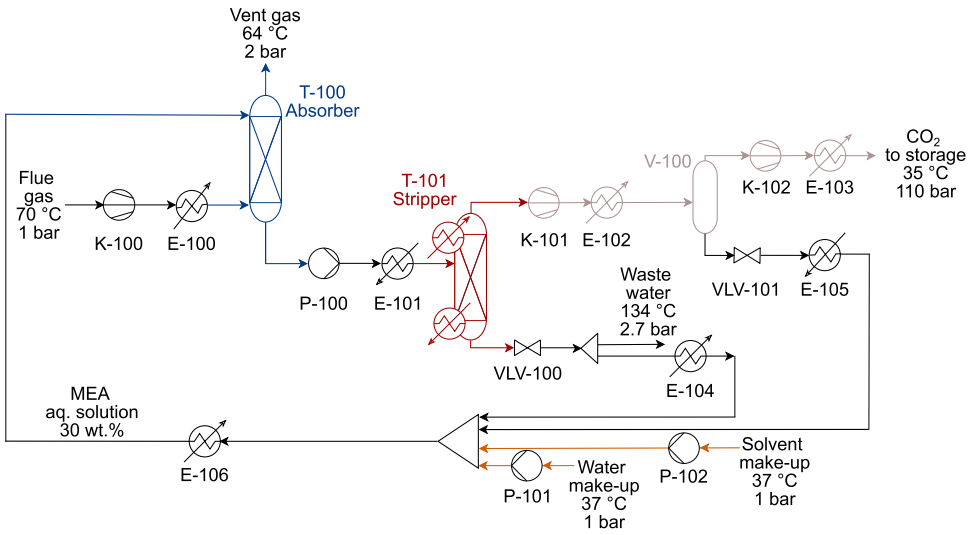


Figure 3. Flue gas treatment process flow diagram. The process consists of an absorption stage (blue), desorption (red), CO_2 compression (sand), and recycle with solvent makeup (yellow).

The second case study investigates post-combustion CO_2 removal from a typical power plant flue gas. The flowsheet is based on similar studies[70,75−77] and adjusted with sensitivity analyses, described in the Supporting Information. The flue gas composition at the inlet can vary significantly depending on the power plant. This study focuses on flue gas in coal-fired power plants after the SO_2 scrubbing pretreatment.[78] The process flowsheet can be divided into three main stages: absorption, desorption, and CO_2 compression.

The mixture of N_2, CO_2, O_2, and H_2O (1000 kmol h^{-1}) enters the post-combustion plant in Figure 3 at 1 bar and 70 °C. The feed is compressed to 2 bar to overcome the column pressure drop and cooled to 50 °C. The CO_2 lean gas at the top contains 3% mol CO_2. We highlight that an even lower CO_2 concentration can be achieved by increasing the amount of MEA and consequently the reboiler duty, as discussed in the Supporting Information. The CO_2-rich solution leaving the bottom of the absorber is sent to the first stage of the stripper to separate CO_2 using steam. The absorber and stripper columns operate under slight pressure at 2 and 5 bar to favor the absorption process and lower the reboiler duty, respectively, as done in the previous case study. The first tower has 17 stages and the second has 14. Both columns are packed with plastic material due to the corrosivity of the CO_2−MEA mixture. The energy consumption of the stripper is 2.7 MJ/kg CO_2 removed for a CO_2 loading of 0.052, in accordance with the literature.[79] The stripper operates under two design specifications: 90% mol CO_2 recovery and 90% mol CO_2 purity in the distillate. The CO_2-rich stream is compressed to supercritical conditions, at 110 bar and 38 °C, prior to being transported and stored underground (not included in our analysis) with a purity of 99.6% mol CO_2. The recycle stream is a 30% wt MEA aqueous solution at 37 °C whose concentration is maintained constant with fresh water and MEA makeup.

In both case studies, we focus on predicting the cooling and heating utilities [kW], net power required [kW], and amount of MEA [kg/h] as output variables from the following input variables: feed pressure [bar], temperature [°C] and composition, and product composition. We hereafter refer to the product as the stream leaving at the top of the absorber in both examples. We note that the product composition in the absorber is a variable that depends on the amount of MEA in the recycle stream for a given inlet gas composition.

The values of the input variables to the process are obtained using Latin hypercube sampling (LHS), which returns the desired number of randomly distributed points for each independent variable in given intervals. We carry out the calculations of the absorber top product purity and amount of MEA in MATLAB. This approach allows us to simplify the solution of the flowsheet by reducing the number of loops to one (the recycle stream) and prevent dependencies between the variables MEA and product composition, while maintaining the number of degrees of freedom. More precisely, we define the variable MEA within an interval of interest using LHS and we run the process models to obtain a range of compositions of the absorber top product. Later, the composition is considered as an independent variable for the surrogate model.

In this regard, it is worth mentioning that, like other ML tools, the BMS has no physical knowledge about any of the two processes that are regarded as a black box of which only inputs and outputs are relevant for building the simplified equations. The lack of information about physical and chemical laws that leads to poor interpretability of many black-box models is a well-known drawback extensively discussed in the literature,[54,55] where the hybrid modeling approach is preferred instead.[80] However, here we claim that our approach, although it cannot be directly interpreted in terms of chemistry and physics first principles, offers a mathematical form that links input and output variables more intuitively than other ML algorithms previously published. Specifically, we argue that interpretability is a continuum rather than a binary feature, with simple, first-principles models at one end and very complex models with many parameters (such as state-of-the-art deep-learning models) at the other end. Certainly, the models derived by the BMS are not directly relatable to first principles, but they are interpretable in many important regards. For example, they can be used directly to answer questions: How does property Y scale when property X tends to 0, or in the limit of large X? Or why are predictions of Y very large at some values of X? Or what is the derivative of Y with respect to X? In

https://doi.org/10.1021/acsomega.2c04736
ACS Omega 2022, 7, 41147−41164

Table 1. Independent and Dependent Variables with Their Respective Ranges Explored for the Natural Gas Sweetening and Flue Gas Treatment Case Studies[a]

		independent variables	lower bound	upper bound	reference	dependent variables	lower bound	upper bound	design specification
natural gas	x_1	pressure [bar]	18.0	32.0	[18,72]	y_{MinCU} minimum cooling utilities [kW]	64 559.8	99 100.6	CO_2 mol recovery > 90%
	x_2	temperature [°C]	35	50	[17,18,72]	y_{MinHU} minimum heating utilities [kW]	63 465.7	93 289.2	
	x_3	CO_2 mol feed	0.2000	0.3000	[81]	$y_{NetPower}$ net power required [kW]	1048.0	3267.4	
	x_4	CH_4 mol product	0.8786	0.9999	for pipeline injection CH_4 > 0.98% mol[16,82]	$y_{AmountMEA}$ amount of MEA [kg/h]	380 000.0	550 000.0	
flue gas	x_1	pressure [bar]	1.0	2.8	[78,83]	y_{MinCU} minimum cooling utilities [kW]	1737.1	4746.4	CO_2 mol recovery > 90%
	x_2	temperature [°C]	35	70	[75,77]	y_{MinHU} minimum heating utilities [kW]	1168.6	4527.2	CO_2 mol > 90% in distillate
	x_3	CO_2 mol feed	0.0700	0.1500	[78]	$y_{NetPower}$ net power required [kW]	174.9	1484.1	
	x_4	H_2O mol feed	0.0500	0.1500	[78]	$y_{AmountMEA}$ amount of MEA [kg/h]	12 000.0	17 000.0	
	x_5	O_2 mol feed	0.0200	0.1200	[78]				
	x_6	CO_2 mol product	0.0228	0.1858	calculated				

[a]The product stream always refers to the stream leaving at the top of the absorber.

these important ways, the models proposed in this work are much closer to first-principles ones than to deep-learning models. Moreover, the prior used in the calculations could be modified to consider specific problem-related equations, e.g., from the chemical engineering literature. Lastly, we build the simplified equations from rigorous process simulations, so the model should ultimately capture the main trends dictated by the first principles.

The dependent variables we are interested in refer to the energy consumption of each process, which contributes most significantly to its economic and environmental performance. The utilities are calculated by computing the grand composite curves, assuming full heat integration (minimum utilities consumption), and that the utility requirements of the optimal heat exchanger network would approach the thermodynamic targets. The electricity consumption corresponds to the net power required to operate pumps and compressors, discounting the energy gained from the expander in the case of natural gas. The range of the sampling variables is defined based on the literature, as reported in Table 1. The feed composition in Table 1 is given for all of the components but one, which is adjusted such that the sum of the component molar fractions is equal to one. Note that here we work under the strong assumption that the inlet conditions vary without any change in the design of the plant, assumed to be fixed.

In this work, we use a dataset for training and an independently generated set for validation. First, we generated 1200 and 2500 scenarios in the natural gas and flue gas cases, respectively, out of which 1174 and 1245 converged in the simulation. More initial points were run for the flue gas simulation to account for the increased complexity of the flowsheet, e.g., two more independent variables (two more components in the feed) and two design specifications for the stripper (CO_2 mol purity and recovery). Out of the total points, those that did not satisfy the conditions of 30% wt MEA solution were discarded. This data was used to generate the models reported in Section 4. Then, the expressions were validated with additional points generated for the same ranges of input variables using LHS. The validation set includes 199 converged points for the natural gas sweetening and 195 for the flue gas treatment process, as reported in Section 4. We refer to the Supporting Information for the results of the training set.

3.2. Mathematical Approach for Symbolic Regression. Typical symbolic regression methods combine three main ingredients: (1) a suitable representation of the problem based on symbolic trees; (2) an appropriate objective function to drive the search; and (3) an optimization engine to identify the best expressions. Although the BMS operates in slightly different terms (it samples models from the Bayesian posterior distribution and is not based on any MINLP formulation), it can also be cast into this scheme. The three ingredients are described in detail next.

Closed-form mathematical expressions can be represented as trees: the internal nodes are simple mathematical operations (e.g., sum or exponential), while the leaves are variables and parameters, as represented in Figure 4.

Concerning the objective function, the BMS uses the description length (approximated as in eq 2) to select the best model. The description length is calculated from the Bayesian Information Criterion (BIC) reported in eq 3, which

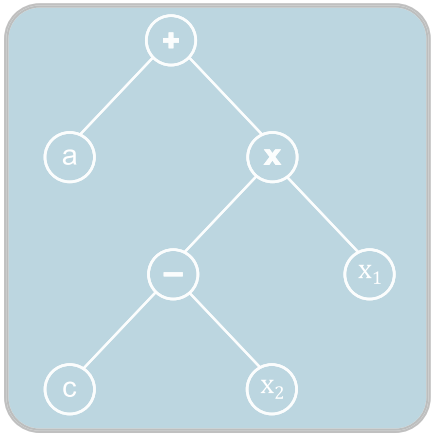


Figure 4. Example of a symbolic tree representing the function $f(x) = a + (c - x_2)\, x_1$.

https://doi.org/10.1021/acsomega.2c04736
ACS Omega 2022, 7, 41147−41164

considers the number of model parameters, the sample size and the mean square error of the model (MSE, see eq 4), and the prior over expressions (POE). Note that the sample size is only considered for estimating the BIC of each model explored.

$$L \approx \frac{\mathrm{BIC}}{2} - \log(\mathrm{POE}) \tag{2}$$

$$\mathrm{BIC} = p\cdot\log(|K|) + |K|[\log(2\pi) + \log(\mathrm{MSE}) + 1] \tag{3}$$

$$\mathrm{MSE} = \frac{\sum_k (y_k - \tilde{y}_k)^2}{|K|} \tag{4}$$

In eq 3, p is the number of parameters of the learned model plus one. In eq 4, y_k is the real value observed of a generic variable i, while $\tilde{y}_k$ is the value predicted by the BMS for each point k.

Based on the symbolic tree representation above, an MCMC algorithm explores the space of all of the possible mathematical expressions implementing three moves on the trees: (i) node replacement, (ii) root addition or removal, and (iii) elementary tree replacement. Each of these moves affects the mathematical expressions differently, by introducing minor variations or significant changes in the structure, or causing the trees to shrink/grow. Alternatively, deterministic optimization methods could be used to explore the tree, e.g., by capitalizing on the MINLP formulation of Cozad and Sahinidis[33] coupled with a standard MINLP solution algorithm.[38] Finally, the BMS selects the most plausible model in an MCMC run, namely, the one with the minimum description length.

The prior used in the description length calculations is the maximum entropy distribution consistent with a corpus of 4080 closed-form mathematical expressions retrieved from Wikipedia.

We apply the algorithm available in the online repository provided by the authors in a similar fashion as Žegklitz and Pošík[84] previously did to compare different ML tools, and we adjust only the number of MCMC steps. We refer to the original article[57] for further details regarding the BMS algorithm.

3.3. Mathematical Implementation. The inputs to the BMS include the training dataset (as small as 100 points[57]), the hyperparameters of the priors (given in ref 57 for a fixed number of independent variables and model parameters), and the number of MCMC steps.

The maximum number of parameters in an expression is fixed to twice the number of independent variables. This choice controls the size of the regression tree to avoid too large search spaces, which would lead to large CPU times. The number of steps has been chosen based on the coefficient of determination (R^2, described later) and description length obtained throughout the MCMC steps, reported in the Supporting Information. Notably, the description length tends to improve as iterations proceed, often reaching a plateau after a sufficiently large number of steps, which is case-dependent. Here, we generated the training and validation data by sampling on process models. Therefore, the amount of data that can be obtained is in principle infinite, provided that the flowsheet converges. When dealing with an experimental setting, fewer points might be available as experiments are costly and time-consuming, and additional design of experiments tools might be coupled with the BMS.[65,85] To highlight the power of the BMS, we here report the results obtained at a relatively low number of MCMC steps to keep the computational time low.

The sampling was performed using MATLAB R2020a interfacing with Aspen HYSYS v11. Then, the outputs were processed in Python 3.8 with Numpy and Pandas and used to determine the values of the four dependent variables that dictate the operating costs. The BMS was trained using the Jupyter notebook code available online.[57] The algorithm returns one closed-form mathematical expression for each dependent variable of interest as a function of the independent variables and some parameters (multiple regression). Lastly, additional points were generated in the same interval of the training variables using LHS to validate the expressions. The methodology applied is summarized in Figure 5.

We computed some statistical metrics for each output model to assess the goodness of the model fit in both the training and validation steps. For regression models, the R^2 in eq 5 represents a measure of how well the regression predictions

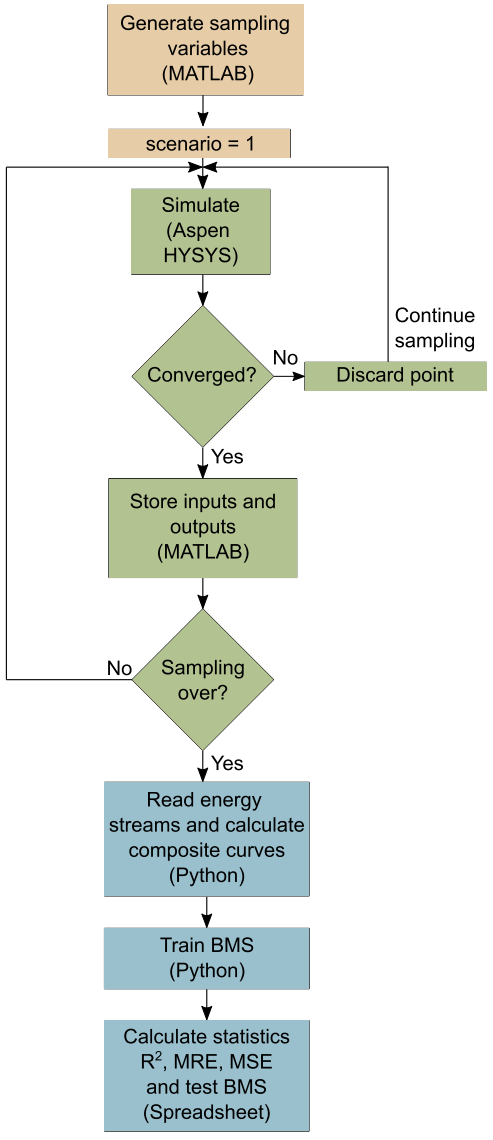


Figure 5. Outline of the procedure to obtain closed-form mathematical expressions: from data sampling, through simulation and data processing, to the application of the BMS and model validation.

https://doi.org/10.1021/acsomega.2c04736
ACS Omega 2022, 7, 41147−41164

approximate the real data points on a convenient scale from 0 to 100%. Therefore, an R^2 of one indicates that the regression predictions fit the data perfectly.

$$R^2 = 1 - \frac{\sum_k (y_k - \tilde{y}_k)^2}{\sum_k (y_k - \overline{y})^2} \tag{5}$$

where y_k is the real value, $\tilde{y}_k$ is the value predicted for each point k, and $\overline{y}$ is the average value.

Additionally, the mean relative error (MRE) in eq 6 measures the precision of the model. The MRE is calculated as the absolute value of the relative error between real and predicted data, normalized by the number of data points.

$$\mathrm{MRE} = \frac{\mathrm{abs}\left(\frac{\sum_k (y_k - \tilde{y}_k)}{y_k}\right)}{|K|} \tag{6}$$

The elasticities can be calculated once $f(x)$ is obtained. They provide insight into the extent to which changes in the various inputs affect the process performance. Elasticities quantify the proportionate change in a dependent variable y relative to a change in an independent variable x_i, keeping the other independent variables ($x_{j\neq i}$) and parameters constant. In eq 7, we report the generic formula employed to calculate the elasticity (E).

$$E = \frac{\delta y}{\delta x}\frac{x}{y} \tag{7}$$

4. RESULTS AND DISCUSSION

4.1. Natural Gas. We run the BMS for the data collected as described above, obtaining the closed-form mathematical expressions for the cooling (MinCU) and heating (MinHU) in eqs 8 and 9, respectively. The parameters (a) are available in the Supporting Information.

$$\mathrm{MinCU} = \left(a_0 \cdot \left(a_3 + \exp\left(a_2 \cdot \exp\left(\frac{x_4}{a_1}\right) + a_5^{x_4^{a_4}} + x_3\right)\right)\right)^{a1} \tag{8}$$

$$\mathrm{MinHU} = \left(\sinh\left(a_4^{x_4^{a_3}} \cdot x_4^{a_5} \cdot a_2^2\right) + a_0 \cdot \frac{x_3}{a_2 \cdot x_4}\right) \cdot a_1 + \left(\frac{a_0}{a_6 \cdot x_4}\right) + a_2 \tag{9}$$

As seen, the cooling and heating utilities equations only select two out of the four independent variables reported in Table 1: the concentration of CO_2 in the feed (x_3) and the CH_4 product purity (x_4). Notably, for a fixed input flow, these variables are strongly connected to the cooling needs in the CO_2 compression stage and heating requirements in the stripper reboiler, which represent a large percentage of the overall cooling and heating, respectively. In contrast, the feed pressure and temperature weakly influence the utilities consumption in the design reported in Figure 2, which is fixed in all of the scenarios.

The net electricity consumption (Net power) is calculated as the duty required by compressors, pumps, and expanders.

$$\text{Net power} = a_6 \cdot \left(a_0^{a_2 + x_3} + a_4 \cdot x_4 + x_1^{a_3} + \frac{a_6 + a_1}{a_5 + x_2}\right) \tag{10}$$

As expected, the expression reported in eq 10 relates Net power to all of the independent variables: feed pressure (x_1) and temperature (x_2), the concentration of CO_2 (x_3) in the feed, and the CH_4 product purity (x_4). The concentrations influence the duty of the CO_2 compressors and pumps for a fixed input flow, while the feed pressure and temperature determine how much power can be gained from the expansion.

Lastly, we find that the amount of solvent (Amount of MEA) required to achieve a specific product purity is proportional to the amount of CO_2 in the feed.[86]

$$\text{Amount of MEA} = a_0 \cdot a_3^{x_4} \cdot a_7 \cdot x_3 \cdot \left(\left(\frac{x_4}{a_2}\right)^{a_6} + \left(a_4 \cdot x_4 + \frac{a_5 \cdot \tan(a_5 + x_4^{a_0}) \cdot (a_1 + 2 \cdot a_6 + x_4)}{x_3}\right)^2\right) \tag{11}$$

Therefore, the amount of MEA in eq 11 is a function of the CO_2 molar fraction in the feed (x_3) and the CH_4 purity in the product (x_4). Consequently, the expression found by the BMS does not select the feed pressure and temperature.

The variable selection problem (or features selection problem) is summarized in Table 2. Notably, the BMS identifies that the feed pressure and temperature do not influence three out of the four dependent variables selected, while the CO_2 composition in the feed and the CH_4 product purity are included in all of the expressions generated. All of the closed-form mathematical expressions referring to the dependent variables in the natural gas flowsheet include fewer parameters than the maximum allowed, with the exception of the simplified equation predicting the MEA consumption. Equations 8−11 are quite compact and include elementary operations, such as additions, multiplications, and exponentials. Trigonometric functions also appear in the case of MinHU and Amount of MEA.

Table 2. Summary of the Features Selection Problem for the Natural Gas Sweetening Process[a]

independent/ dependent variables	feed pressure [bar]	feed temperature [°C]	CO_2 mol feed	CH_4 mol product
MinCU	0	0	1	1
MinHU	0	0	1	1
Net power	1	1	1	1
Amount of MEA	0	0	1	1

[a]The dependent variables are listed per row, while the independent ones are reported in the columns with a one if selected and zero otherwise.

The data scatter around the regression line is shown in Figure 6 for each dependent variable (eqs 8−11). The corresponding values of R^2, as well as MRE and MSE, are reported in Table 3.

The statistics indicate that the model for Net power shows the best match between the observed and predicted data. This is shown in Figure 6c, where the data of the validation set lies very close to the diagonal. Contrarily, MinCU leads to the highest MRE and a slightly lower R^2 value than the other variables, as shown by the broader distribution of the points on the diagonal in Figure 6a. The scatter plots of MinHU and

https://doi.org/10.1021/acsomega.2c04736
ACS Omega 2022, 7, 41147−41164

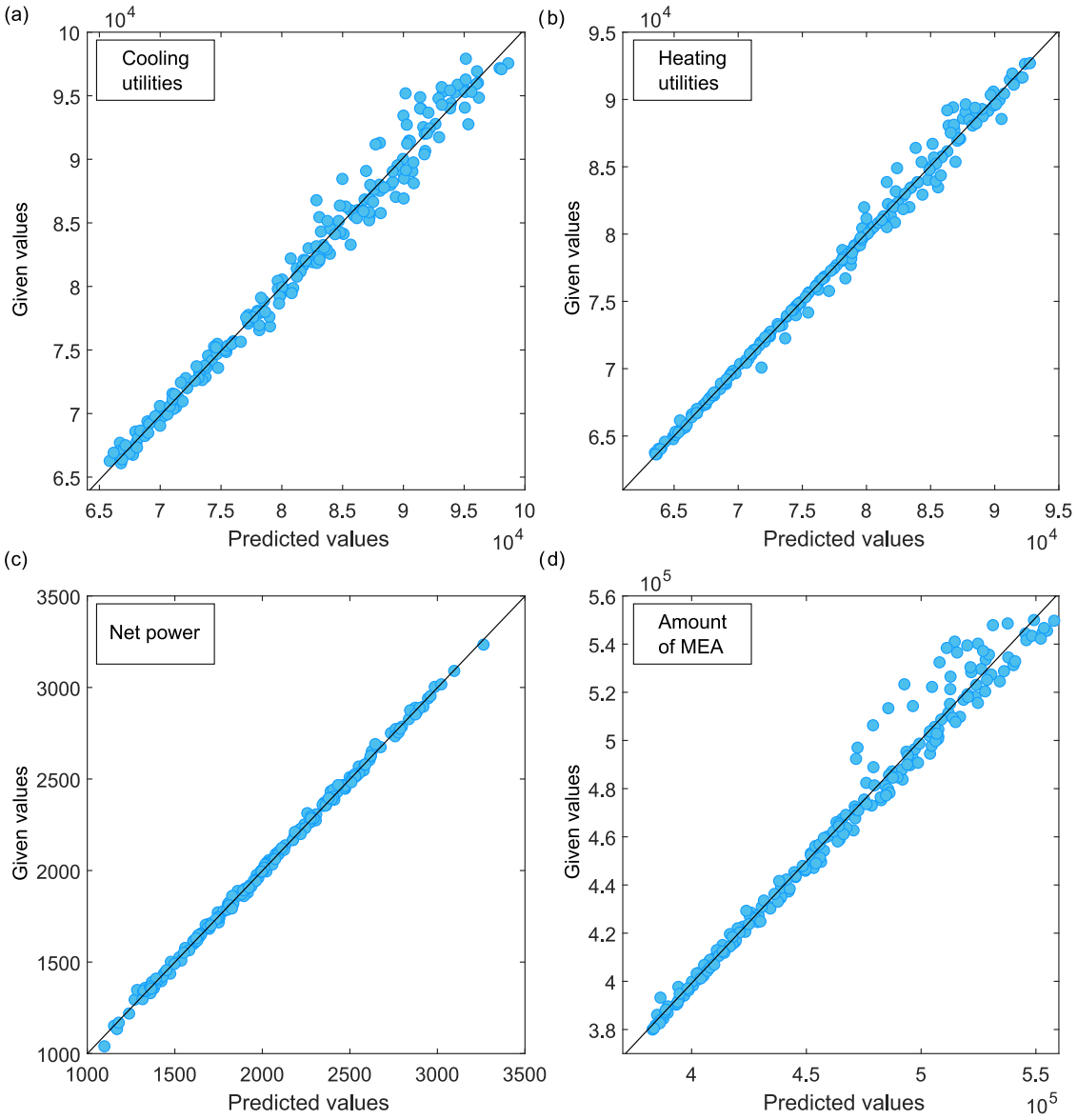


Figure 6. Given vs predicted values correlation for the four output variables in the validation dataset: (a) cooling and (b) heating utilities, (c) net power, and (d) amount of MEA for the natural gas sweetening process.

Table 3. Coefficient of Determination (R^2), Mean Relative Error (MRE), and Mean Square Error (MSE) Statistics for Each Output Variable in the Validation Dataset of the Natural Gas Sweetening Process

case study	variable	R^2	MRE	MSE
natural gas	MinCU	0.9818	0.0103	1.50E+06
	MinHU	0.9921	0.0051	5.21E+05
	Net power	0.9986	0.0072	3.15E+02
	Amount of MEA	0.9922	0.0050	1.94E+07

Amount of MEA show a predicted vs observed data pattern lying in between the previous two. The variable Amount of MEA is depicted in Figure 6d. As seen, the data is distributed close to the regression line in the lower interval of values explored. However, it tends to scatter close to the upper bound of the variable.

It is also worth recalling that the utilities and power variables are processed data, i.e., they are not direct outputs of the simulation. Nevertheless, overall, the BMS is able to recover highly accurate models and identify the independent variables that physically influence the process the most, even when these have been processed.

4.2. Flue Gas. We repeat the analysis for the same outputs in the flue gas process, modifying the prior to consider six independent variables.

The minimum cooling (MinCU) utilities of the flue gas treatment process are found by the BMS to be a function of five independent variables reported in Table 1: feed pressure (x_1) and temperature (x_2), CO_2 (x_3) and O_2 (x_5) molar concentrations in the feed, and CO_2 molar concentration in the product (x_6).

$$\text{MinCU} = \left(\frac{x_3 + a_0 \cdot x_1}{x_2} + \left(a_2 + \frac{\left(x_3 + \frac{a_3 \cdot x_5 \cdot x_6}{x_5^{a_1}} \right) \cdot (a_5 + x_1)}{a_4 \cdot x_2} \right)^{a_2} \right)^4 \quad (12)$$

https://doi.org/10.1021/acsomega.2c04736
ACS Omega 2022, 7, 41147−41164

In particular, feed pressure and temperature determine the amount of cooling necessary to reach the absorber operating conditions after the compression, while the CO_2 concentration in the feed and the remaining CO_2 in the product affect the cooling in the compression stage. The coolers in the CO_2 compression stage consume most of the total utilities reported in eq 12. For a fixed inlet flue gas stream, a change in the water concentration mainly affects the heating and the recycle streams, while the utility requirements of the coolers are negligible compared to the compression stage. Therefore, eq 12 omits water concentration (x_4). The expression obtained for MinCU is rather simple, as it only considers additions, multiplications, and exponentials.

On the contrary, the total heating utility (MinHU) is a function of all six independent variables (eq 13).

$$\begin{aligned}\text{MinHU} = \Bigg(& a_7 + x_2 + a_3 \cdot (a_2 + x_4) \cdot \\ & \left(a_5 \cdot x_6 \cdot \exp(-x_3) + \frac{a_2 \cdot x_5 \cdot (a_1 \cdot (a_0 + x_2))^{x_1}}{x_2}\right) \\ & + \frac{x_6 \cdot (a_4 + a_3 \cdot x_6 + a_6 \cdot x_5)}{x_3}\Bigg) / \cos(a_6 + x_5)\end{aligned} \tag{13}$$

The water content (x_4) in the feed stream is linked to the steam consumed by the reboiler of the stripper that regenerates the solvent. The amount of steam also depends on the CO_2 (x_3) and O_2 (x_5) feed concentrations and on the CO_2 concentration in the clean flue gas (x_6). We note that this expression is less compact than the previous ones and contains a trigonometric function. However, it can be simplified as follows. Parameter a_6 is quite large (5.6e+3), so the cos() function can be considered a constant that becomes zero for any value of x_5 (molar fraction between 0 and 1).

The net electricity consumption (Net power) in the flue gas treatment process accounts for the energy consumed by pumps and compressors.

$$\begin{aligned}\text{Net power} = & a_0 \cdot x_5 \\ & + (a_6 + (x_5 + \sinh(a_7 - x_1 \cdot \cos(a_2) \cdot (x_6 + x_3^{\cos(a_1 + x_6)})) \\ & + \frac{a_3}{a_2 + x_2 + x_5 + a_4 \cdot x_2} + a_5 \cdot \exp(a_7))^{a_5})^4 / x_2^2\end{aligned} \tag{14}$$

In eq 14, Net power is expressed as a function of the feed pressure (x_1) and temperature (x_2), the CO_2 (x_3) and O_2 molar concentrations in the feed (x_5), and the CO_2 molar concentration in the product (x_6). As in eq 12, the concentration of water in the feed is omitted. The compressors contribute much more to the total energy consumed than the pumps, and the flow rate of CO_2 mainly determines the compression duty.

Lastly, the closed-form expression of the amount of MEA (Amount of MEA) includes all of the concentrations in the feed and the product (eq 15) because the stream composition dictates the amount of solvent needed.

$$\begin{aligned}\text{Amount of MEA} = & \left(\frac{a_8 + \frac{a_8 + x_3}{a_9^{(x_6/2)}} + a_0 \cdot \left(x_5 + \sin\left(a_7 \cdot x_6 + \frac{\text{abs}(a_1 \cdot x_3 \cdot (a_5 + x_5))}{\text{abs}(a_5 + a_3 \cdot x_3^{a_4})}\right)\right)}{a_2}\right)^{a_1} \\ & + a_6 \cdot x_4\end{aligned} \tag{15}$$

Once again, feed pressure and temperature are not linked to the amount of MEA required. The abs() function can be simplified because all of the independent variables and parameters a_1 and a_5 are positive (see the Supporting Information). Conversely, a_4 is negative but $x_3^{a_4}$ is positive. Thus, the ratio of abs() functions can be calculated as the ratio of the arguments.

As observed, the simplified equation for the amount of MEA includes 10 parameters (out of the 12 allowed). A summary of the selection of the independent variable for each dependent one is reported in Table 4. At first glance, the expressions reported in eqs 12−15 seem more complex than in the previous case. However, the building blocks in the formulas are still very simple additions, multiplications, and some trigonometric and exponential functions. Although the second case considers two more independent variables and one more design specification, we can still obtain equations that fit the data with an R^2 value greater than 0.94 for three of the four variables, two above 0.99, as reported in Table 5. Even in the case of minimum cooling utilities, the MRE remains below 8%. We note that our ultimate goal is to predict the economic and environmental performance, so estimating the cooling utilities with less accuracy is not an issue as their contribution to the overall performance is low.

Table 4. Summary of the Features Selection Problem for the Flue Gas Treatment Process[a]

independent/ dependent variables	feed pressure [bar]	feed temperature	CO_2 mol feed	H_2O mol feed	O_2 mol feed	CO_2 mol product
MinCU	1	1	1	0	1	1
MinHU	1	1	1	1	1	1
Net power	1	1	1	0	1	1
Amount of MEA	0	0	1	1	1	1

[a]The dependent variables are listed per row, while the independent ones are reported in the columns with a one if selected and zero otherwise.

Table 5. Coefficient of Determination (R^2), Mean Relative Error (MRE), and Mean Square Error (MSE) Statistics for Each Output Variable in the Validation Dataset of the Flue Gas Treatment Process

case study	variable	R^2	MRE	MSE
flue gas	MinCU	0.4506	0.0744	1.22E+05
	MinHU	0.9405	0.0381	1.72E+04
	Net power	0.9995	0.0090	5.74E+01
	Amount of MEA	0.9965	0.0046	7.22+03

The scatter plots in Figure 7 represent the goodness of fit for the four output variables in eqs 12−15 considering the data in the validation set. Once again, Net power is the variable that shows the best model performance (R^2 of 99.95%) and for which the data lies precisely on the regression line in Figure 7c,

https://doi.org/10.1021/acsomega.2c04736
ACS Omega 2022, 7, 41147−41164

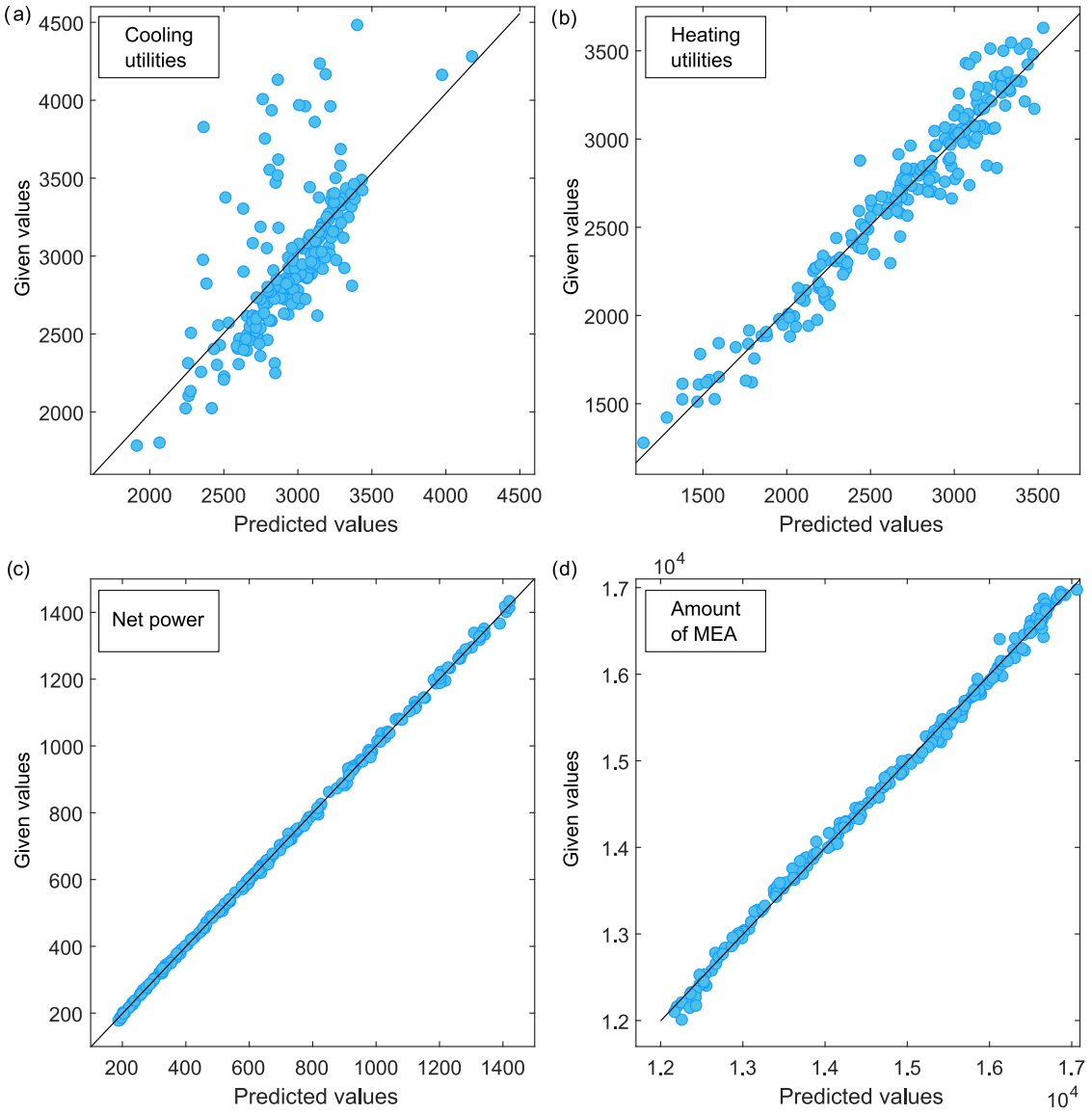


Figure 7. Given vs predicted values correlation for the four output variables in the validation dataset: (a) cooling and (b) heating utilities, (c) net power and (d) amount of MEA for the flue gas treatment process.

while MinCU leads to the worst fit (R^2 of 45.06%) for the MCMC steps selected. The relationship between predicted and real values for MinCU shown in Figure 7a indicates that the model reproduces very well the data in the range 2500−3500, where it accumulates. However, some points are far from the regression line. We note that a low R^2 value doen not always imply that the model is unacceptable. If, for example, the variability of the data is low, the MRE will likely be low (indeed, here is around 7%), and the model will still provide reliable predictions.

On the contrary, the model for MinHU improves compared to MinCU, while it still shows some data variability across the regression line. Finally, the data fit of Amount of MEA shows that the data is more aggregated along the diagonal, approaching the goodness of fit of Net power. The R^2, MRE, and MSE values of the dependent variables are reported in Table 5.

The residual plots for each output variable of the two case studies can be found in the Supporting Information, where the training results and the corresponding R^2, MRE, and MSE values are also given.

Additionally, the same dataset used to train the BMS was used to train an ANN with Bayesian regularization for both cases. The results are reported in the Supporting Information. ANN models lead to an R^2 above 99%, both in the training and validation dataset. However, the obtained models are not easily interpretable and hard to employ in further analyses such as those presented in the following sections.

5. ANALYTICAL APPLICATION OF THE EXPRESSIONS

5.1. Analysis of the Elasticities. The BMS has the advantage over other ML algorithms of providing closed-form mathematical expressions. In turn, these can be manipulated analytically, differentiated, or used in optimization frameworks to investigate the performance of the processes further or compare different alternatives. In this regard, we claim that our models are more interpretable than conventional black-box tools and can be used to answer questions about the influence of each variable on the process as explained in Section 1.

https://doi.org/10.1021/acsomega.2c04736
ACS Omega 2022, 7, 41147−41164

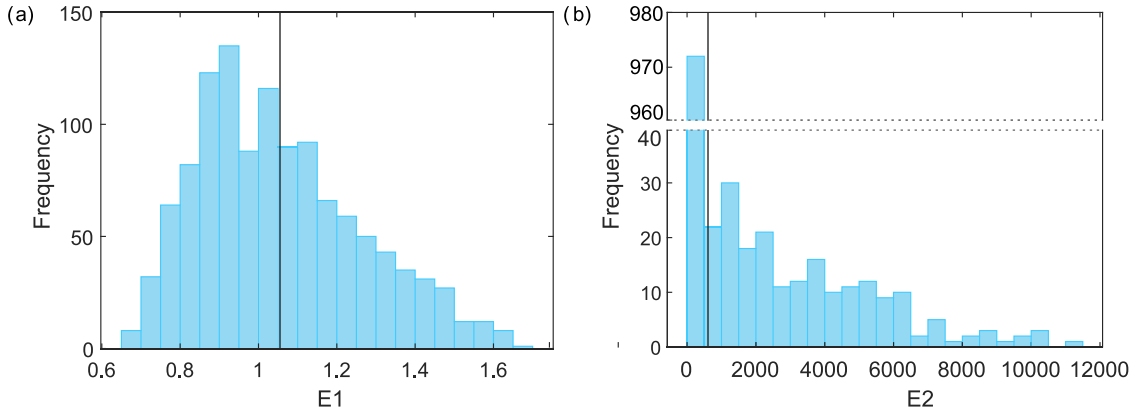


Figure 8. Elasticities of the natural gas dependent variables for MinHU in the training dataset. (a) $E1$ is the elasticity corresponding to the independent variable x_3 CO_2 concentration in the feed (mean = 1.0544) and (b) $E2$ to x_4 CH_4 product (mean = 620.6721).

Hence, in this section, we take a step forward to study the strength of the link between the independent variables selected to predict the minimum heating utilities (MinHU) and the value of this output variable. The choice of this variable is motivated by the high energy requirements of the absorption process, which can be ascribed almost entirely to the regeneration of the solvent (reboiler duty), ultimately dictating the economic and environmental performance. To carry out the calculations, we use eq 7 for each point of the training set, where y is the dependent variable chosen (MinHU) and x each independent variable y. We calculate the elasticities as described above for each x in all of the points k and then plot the distribution of these values. From a practical viewpoint, the elasticities provide insight into the relationships between independent and dependent variables. Although the analysis of the elasticities is still possible using other ML tools, the BMS allows for a more in-depth study of the change in a dependent variable as a result of an increment of an independent one. For example, it is possible to calculate the elasticity using ANNs, but the result would be a numerical value and not an expression that can be manipulated further.

In Figure 8, we provide the histogram of the elasticities for the dependent variable MinHU in the natural gas sweetening process. Recall that the expression of this independent variable only includes two independent variables out of four: the CO_2 concentration in the feed (x_3) and the CH_4 product purity (x_4) (see eq 9). The mean of the CO_2 feed concentration elasticity is above 1 (subplot a), representing a positive elastic relationship: for a x% increase in the independent variable, the dependent variable increases by y%, where $y > x$, denoting a strong response in the output to changes in the input. The CH_4 product purity (subplot b) is also positive elastic. However, the high value of the mean (620) is not representative of the majority of the points (median = 4.23). This behavior is due to the instability of the derivative whose value skyrockets for x_4 above 0.999, which, however, does not influence the accuracy of the model itself. On the contrary, this provides an interesting insight into the physical model by implying that for purities above 0.999, unattainable from a practical standpoint, the energy consumption required for an increment of the purity would be prohibitive.

Next, we analyze the elasticities of the minimum heating utilities of the flue gas process (eq 13). We recall that the dependent variable chosen is a function of all six independent variables whose histograms of elasticities are shown in Figure 9. The feed pressure (subplot a) shows a mean elasticity between 0 and 1, implying that the relationship is positive inelastic. An increase in the feed pressure (x_1) leads to higher heating utilities due to the lower compression ratio in K-100. On the contrary, the elasticity of the feed temperature (subplot b) lies from −1 to 0, denoting a negative inelastic response. As expected, for an increase in the feed temperature (x_2), the minimum heating utilities decrease. The mean elasticity of the CO_2 feed concentration (x_3) is positive elastic (subplot c), as its mean value is greater than 1. As said before, the reboiler duty, which is the most significant contribution to the total heating utilities, depends on the initial amount of CO_2. On the contrary, the concentration of water (subplot d) in the feed shows a negative inelastic relationship with a mean value between 0 and −1. It is worth noting that the mean elasticity of water is −0.02, which shows that the influence of this variable is almost negligible for the dependent variable considered. We recall that water is not chosen in the simplified equation of the minimum cooling (eq 12). Increasing O_2 (subplot e) increases, in turn, the energy consumption since the mean elasticity of x_4 is between 0 and 1—positive inelastic. Finally, the elasticity of the CO_2 concentration (x_6) in the product stream at the top of the absorber (subplot f) is smaller than −1, therefore, indicating a negative elastic response; i.e., an increase in the CO_2 concentration of the product stream lowers the heating requirements.

Hence, the analysis of elasticities clearly shows that the CO_2 concentration in the feed is the variable influencing the heating needs the most, as expected. Moreover, the effect of the water concentration is negligible, despite appearing in the simplified equation.

5.2. Emerging Technologies Assessment. We next illustrate how the simplified equations could be used to benchmark emerging technologies. To this end, let us consider an alternative CO_2 capture technology still under development, namely, the cryogenic CO_2 separation from flue gas based on the Stirling cooler system developed by Song and co-authors.[23] We shall compare the performance of the latter against that of the BAU using the expressions in eqs 8−15.

The authors provide the feed specifications and the cooling and electricity needs of the compressors, with and without heat integration. We analyze the case without heat integration for simplicity and determine the CO_2 concentration (mol) in the clean gas from the mass balance provided. The CO_2 concentration in the clean gas of the process developed by

https://doi.org/10.1021/acsomega.2c04736
ACS Omega 2022, 7, 41147−41164

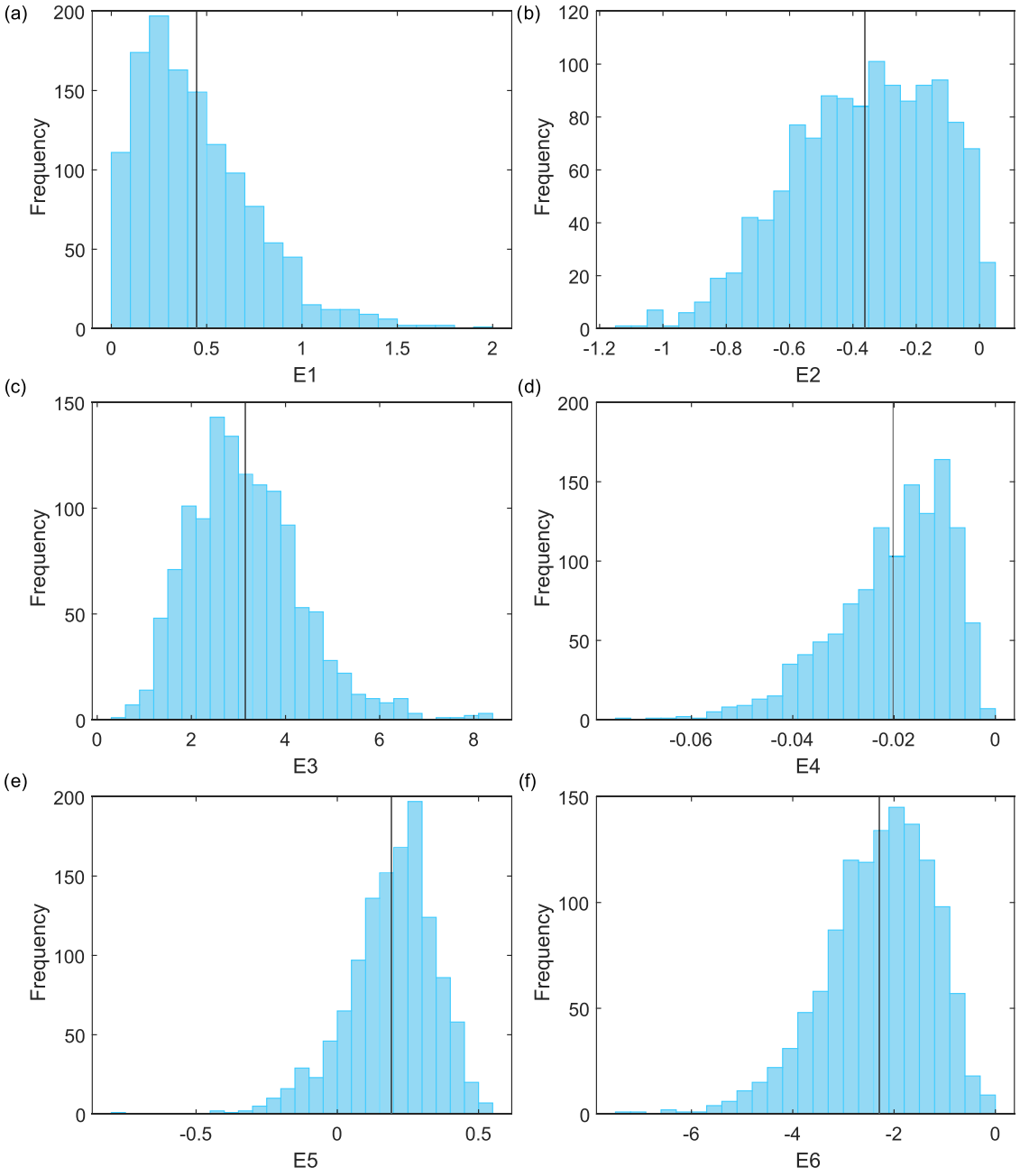


Figure 9. Elasticities of the flue gas dependent variables for MinHU in the training dataset. (a) *E*1 refers to the independent variable x_1 pressure (mean = 0.4475), (b) *E*2 to x_2 temperature (mean = −0.3632), (c) *E*3 to x_3 CO_2 mol concentration in the feed (mean = 3.1495), (d) *E*4 to x_4 water concentration in the feed (mean = −0.0202), (e) *E*5 to x_5 O_2 concentration in the feed (mean = 0.1921), and (f) *E*6 to x_6 CO_2 concentration in the product (mean = −2.3001).

Song is 0.005 mol, which is below our lower bound 0.023 (Table 1) and, thus, falls outside the limits of our training set. We then calculate the minimum cooling [kW] required and the net electricity consumption [kW] using eqs 12 and 15 for both concentrations.

Moreover, since the input flow used by Song et al. is noticeably higher than in the case we explored here, we take the energy consumption per mass flow rate of the clean gas [kJ/kg]. In our process (Figure 3), the product is almost constant in all of the scenarios analyzed. Nonetheless, we make predictions taking the maximum and minimum flow rate obtained from the sampling (which only differ by 13%, see Table S2 in the Supporting Information). The results are reported in Table 6 as the ratio between our dependent variables (subscript BAU) and the values of the Stirling process (subscript cry).

Using the BMS models reported above, we conclude that the new process reduces energy consumption for the specific conditions analyzed, mainly owing to the lack of heating required. We note that our analysis here is just a simple example of an additional application of the BMS, and the conclusions we draw are based on the available data and assumptions made. We also point out that the extrapolation performed using the equations for the CO_2 molar concentration outside of the trained bounds leads to a relative error of 8.5 and 5% for the cooling and electricity ratios given in Table

https://doi.org/10.1021/acsomega.2c04736
ACS Omega 2022, 7, 41147−41164

Table 6. Comparison of Cooling and Electricity Requirements for the Process by Song et al.[23] (cry) and Our BAU (BAU)[a]

	case I: without heat integration			
	$MinCU^{BAU}/MinCU^{cry}$		$Net\ power^{BAU}/Net\ power^{cry}$	
	prod low	prod high	prod low	prod high
CO_2 mol in ref work[23] (0.5%)	2.19	1.91	1.17	1.02

[a]The values are calculated as the ratio of the energy requirement of the processes: BAU/cryogenic per absorber top product mass flow rate.

6, respectively. Lastly, we note that the simplified equation retrieved by the BMS for the cooling duty is the one affected by the poorest performance in predicting the data (R^2 of 45%).

5.3. Potential Applications of the Bayesian Machine Scientist to Process Systems Engineering Problems. The models developed in this work aim to support experimentalists and guide their research in the quest for more sustainable technologies. For example, experimental groups could quickly benchmark their CO_2 separation technologies against standard MEA-based capture processes using simplified analytical methods without the need to carry out detailed simulations. This would allow them to identify critical hotspots concerning energy consumption or purity specifications. Moreover, the streamlined equations obtained with the BMS could find multiple applications in PSE, mostly in the areas of surrogate-based process optimization, flexibility analysis, and hybrid model building, as discussed next.

Surrogate-based optimization has recently emerged to overcome the challenges of simulation-based optimization, which attempts to optimize detailed process simulations. In the latter, functions with an algebraic form or derivative information might be absent or too costly and noisy to evaluate.[35] Stochastic algorithms, such as genetic algorithms, can be employed in these cases, requiring numerous samplings and iterations;[87] alternatively, derivative-free algorithms can also be used.[35] Here, process flowsheet optimization is treated as a black-box problem because process simulators commonly present intractable gradients.[88] In this context, the BMS could provide analytical surrogates that could be solved with state-of-the-art solvers using standard modeling systems. This would also enable the application of standard deterministic global optimization algorithms, which cannot be easily applied when dealing with ANNs and Gaussian processes (despite some recent work on tailored deterministic global optimization algorithms for the said surrogates[89,90]).

This approach could find applications in refrigeration cycles,[87] natural gas liquefaction,[91] supply chain inventory control,[92] carbon capture,[93] process synthesis,[94] pharmaceutical processes,[95] semibatch bioprocesses,[27] and biorefineries,[96] to mention a few in chemical engineering and beyond. On the other hand, the applications are not limited to technology benchmarking, as discussed below.

Our approach could also be used in the context of surrogate-based feasibility and flexibility analyses. The former addresses the question of whether a system can remain feasible within a given region of parameter values. In contrast, the latter computes the maximum deviation from the nominal conditions such that the system would still remain feasible. Seminal works by Grossmann and co-workers proposed solution strategies[97] and a two-level optimization framework[98] to tackle these problems, which cannot be directly applied to black-box problems that are not explicitly differentiable. Hence, analytical surrogates could enable the use of such algorithms based on bilevel optimization in a range of problems. Examples of applications include, but are not limited to, models with black-box constraints, computationally expensive models, and nonconvex feasible regions, particularly in pharmaceutical applications,[99] planning, scheduling and control,[100,101] or chromatographic systems.[102]

Surrogate models can be further combined with an algebraic objective, and material and energy balances to formulate algebraic optimization problems under the framework of hybrid modeling.[35] Hybrid models fill the gap linked to the lack of exact knowledge on the process, allowing the user to specify part of the model through a data-driven component, thus requiring less data than pure black-box models, which is particularly relevant in bioprocesses applications.[103] In this context, our approach could be used to build analytical hybrid models, where mechanistic equations would be combined with an analytical surrogate, leading to fully analytical formulations easier to handle. Moreover, it could also be used in tandem with deterministic global optimization algorithms for gray box model optimizations, as presented by Boukouvala and Floudas.[104] Specifically, the BMS could help to approximate black-box constraints, enabling the straightforward application of deterministic global optimization methods to hybrid models.

In particular, applications of ANNs coupled with black-box optimization, which could benefit from analytical surrogates as those developed here, include process synthesis, flexibility analysis, and dynamic optimization, as reviewed by Tsay.[105]

Overall, our approach has the advantage of providing an explicit mathematical form, which can be manipulated algebraically, differentiated, and integrated, alone or together with mechanistic equations, e.g., mass and energy balances in gray box models. Moreover, while interpretability is not a binary value, the results obtained from the BMS are more interpretable than those obtained from ANNs or Gaussian processes.

6. CONCLUSIONS

In this article, we explored the application of machine learning to simplify the benchmarking of emerging technologies with a focus on carbon capture. We applied a Bayesian machine scientist algorithm to streamline the modeling of two basic processes for carbon dioxide removal, generating simple closed-form mathematical expressions of key variables dictating the economic and environmental performance of the whole system considered.

We found that it is possible to build highly accurate simplified process model equations in an automatic manner in relatively low computational time, which can then be used to compare alternative technologies and perform further numerical analyses. The statistics of the goodness of fit, namely, the *R*-squared, mean relative error, and mean square error, indicate that the best predictions correspond to the net power requirement, while the minimum cooling utilities are harder to predict. Nonetheless, the Bayesian machine scientist is able to find precise expressions even for those variables that are not

https://doi.org/10.1021/acsomega.2c04736
ACS Omega 2022, 7, 41147−41164

directly an output of the simulations, such as process utilities and the net power consumption. It can also identify critical process variables that influence the dependent variables the most. Moreover, the number of steps for the Markov chain Monte Carlo algorithm can be increased further to identify even better expressions.

An analysis of the elasticities was carried out to provide insights into how the process variables affect the technology performance. The streamlined process equations were then used to benchmark an emerging technology using literature data with the standard amine capture process, finding that it could outperform the latter under the conditions and assumptions considered.

Overall, this study proved that advanced machine learning methods could be applied to automatically derive simplified process equations that can accurately predict the behavior of technologies in carbon capture applications and beyond. These simplified equations, in turn, can be used to analyze the influence of the independent variables on the overall performance and enable a direct comparison of emerging technologies without the need to run a process simulation in each comparative assessment sought.

This study represents a first proof of concept based on simple case studies, and future work should further explore how to control the shape and complexity of these expressions and include more specific *a priori* knowledge. Moreover, these simplified equations could also be applied to experimental and plant data and used for optimization purposes, i.e., in process design, which could open new opportunities for developing machine learning-based optimization algorithms based on explicit symbolic equations.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acsomega.2c04736.

Models MCMC steps and parameters, process flowsheet assumptions and limitations, residual plots, models training, and neural network analyses (PDF)

■ AUTHOR INFORMATION

Corresponding Authors

Roger Guimerà − *Department of Chemical Engineering, Universitat Rovira i Virgili, Tarragona 43007 Catalonia, Spain; ICREA, Barcelona 08010 Catalonia, Spain*; Email: roger.guimera@urv.cat

Gonzalo Guillén-Gosálbez − *Institute for Chemical and Bioengineering, Department of Chemistry and Applied Biosciences, ETH Zürich, 8093 Zürich, Switzerland*; orcid.org/0000-0001-6074-8473; Email: gonzalo.guillen.gosalbez@chem.ethz.ch

Authors

Valentina Negri − *Institute for Chemical and Bioengineering, Department of Chemistry and Applied Biosciences, ETH Zürich, 8093 Zürich, Switzerland*; orcid.org/0000-0003-4292-0924

Daniel Vázquez − *Institute for Chemical and Bioengineering, Department of Chemistry and Applied Biosciences, ETH Zürich, 8093 Zürich, Switzerland*; orcid.org/0000-0001-9380-3918

Marta Sales-Pardo − *Department of Chemical Engineering, Universitat Rovira i Virgili, Tarragona 43007 Catalonia, Spain*

Complete contact information is available at: https://pubs.acs.org/10.1021/acsomega.2c04736

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

Marta Sales-Pardo and Roger Guimerà acknowledge that this research was funded by Project No. PID2019−106811GB-C31 from MCIN/AEI/10.13039/501100011033 and by the Government of Catalonia (2017SGR-896). Gonzalo Guillén-Gosálbez acknowledges the support from the NCCR Catalysis (grant 180544), a National Centre of Competence in Research funded by the Swiss National Science Foundation.

■ SETS, VARIABLES, AND PARAMETERS

K	{k: set of training points}
I	{i: set of independent variables}
J	{j: set of dependent variables}
x_{ki}	value of independent variable
y_{ki}	real value of dependent variable
$\tilde{y}_{ki}$	predicted value by the BMS
R^2	coefficient of determination
Net power	net power [kW]
Amount of MEA	amount of MEA [kg/h]
BIC	Bayesian Information Criterion
MSE	mean square error
POE	prior over expressions
L	description length
p	number of parameters plus one
MRE	mean relative error
E	elasticity
MinCU	minimum cooling utilities [kW]
MinHU	minimum heating utilities [kW]

■ ACRONYMS

ALAMO	automated learning of algebraic models for optimization
ANN	artificial neural networks
BAU	business as usual
BMS	Bayesian machine scientist
CCS	CO_2 capture and storage
DACCS	direct air carbon capture and storage
LHS	Latin hypercube sampling
MCMC	Markov chain Monte Carlo
MEA	monoethanolamine
MILP	mixed-integer linear programming
MINLP	mixed-integer nonlinear programming
MIP	mixed-integer programming
ML	machine learning
mol	molar fraction
PSE	Process Systems Engineering

■ REFERENCES

(1) Cooper, R. G. Managing Technology Development Projects. *Res.-Technol. Manage.* **2006**, *49*, 23−31.

(2) Mulvihill, M. J.; Beach, E. S.; Zimmerman, J. B.; Anastas, P. T. Green Chemistry and Green Engineering: A Framework for

Sustainable Technology Development. *Annu. Rev. Environ. Resour.* **2011**, *36*, 271−293.
(3) Haszeldine, R. S.; Flude, S.; Johnson, G.; Scott, V. Negative Emissions Technologies and Carbon Capture and Storage to Achieve the Paris Agreement Commitments. *Philos. Trans. R. Soc., A* **2018**, *376*, No. 20160447.
(4) Bruhn, T.; Naims, H.; Olfe-Kräutlein, B. Separating the Debate on CO_2 Utilisation from Carbon Capture and Storage. *Environ. Sci. Policy* **2016**, *60*, 38−43.
(5) Oreggioni, G. D.; Brandani, S.; Luberti, M.; Baykan, Y.; Friedrich, D.; Ahn, H. CO_2 Capture from Syngas by an Adsorption Process at a Biomass Gasification CHP Plant: Its Comparison with Amine-Based CO2 Capture. *Int. J. Greenhouse Gas Control* **2015**, *35*, 71−81.
(6) Muhammad, A.; GadelHak, Y. Simulation Based Improvement Techniques for Acid Gases Sweetening by Chemical Absorption: A Review. *Int. J. Greenhouse Gas Control* **2015**, *37*, 481−491.
(7) Deutz, S.; Bardow, A. Life-Cycle Assessment of an Industrial Direct Air Capture Process Based on Temperature−Vacuum Swing Adsorption. *Nat. Energy* **2021**, *6*, 203−213.
(8) Page, B.; Turan, G.; Zapantis, A.*Global Status of CCS 2020*; 2020.
(9) Kidnay, A. J.; Parrish, W. R.*Fundamentals of Natural Gas Processing*; CRC Press, Taylor & Francis Group, 2006. DOI: 10.1201/9781420014044.
(10) Adib, H.; Sharifi, F.; Mehranbod, N.; Kazerooni, N. M.; Koolivand, M. Support Vector Machine Based Modeling of an Industrial Natural Gas Sweetening Plant. *J. Nat. Gas Sci. Eng.* **2013**, *14*, 121−131.
(11) Vega, F.; Baena-Moreno, F. M.; Gallego Fernández, L. M.; Portillo, E.; Navarrete, B.; Zhang, Z. Current Status of CO_2 Chemical Absorption Research Applied to CCS: Towards Full Deployment at Industrial Scale. *Appl. Energy* **2020**, *260*, No. 114313.
(12) Raksajati, A.; Ho, M. T.; Wiley, D. E. Comparison of Solvent Development Options for Capture of CO_2 from Flue Gases. *Ind. Eng. Chem. Res.* **2018**, *57*, 6746−6758.
(13) Asif, M.; Suleman, M.; Haq, I.; Jamal, S. A.Post-Combustion CO_2 Capture with Chemical Absorption and Hybrid System: Current Status and Challenges. In *Greenhouse Gases: Science and Technology*; Blackwell Publishing Ltd, December 1, 2018; pp 998−1031. DOI: 10.1002/ghg.1823.
(14) Lund, H.; Mathiesen, B. V. The Role of Carbon Capture and Storage in a Future Sustainable Energy System. *Energy* **2012**, *44*, 469−476.
(15) Qian, Q.; Asinger, P. A.; Lee, M. J.; Han, G.; Mizrahi Rodriguez, K.; Lin, S.; Benedetti, F. M.; Wu, A. X.; Chi, W. S.; Smith, Z. P. MOF-Based Membranes for Gas Separations. *Chem. Rev.* **2020**, *120*, 8161−8266.
(16) Peters, L.; Hussain, A.; Follmann, M.; Melin, T.; Hägg, M. B. CO2 Removal from Natural Gas by Employing Amine Absorption and Membrane Technology-A Technical and Economical Analysis. *Chem. Eng. J.* **2011**, *172*, 952−960.
(17) Hoorfar, M.; Alcheikhhamdon, Y.; Chen, B. A Novel Tool for the Modeling, Simulation and Costing of Membrane Based Gas Separation Processes Using Aspen HYSYS: Optimization of the CO_2/CH_4 Separation Process. *Comput. Chem. Eng.* **2018**, *117*, 11−24.
(18) Ahmad, F.; Lau, K. K.; Shariff, A. M.; Murshid, G. Process Simulation and Optimal Design of Membrane Separation System for CO2 Capture from Natural Gas. *Comput. Chem. Eng.* **2012**, *36*, 119−128.
(19) Hasan, M. M. F.; Baliban, R. C.; Elia, J. A.; Floudas, C. A. Modeling, Simulation, and Optimization of Postcombustion CO_2 Capture for Variable Feed Concentration and Flow Rate. 1. Chemical Absorption and Membrane Processes. *Ind. Eng. Chem. Res.* **2012**, *51*, 15642−15664.
(20) Gabrielli, P.; Gazzani, M.; Mazzotti, M. On the Optimal Design of Membrane-Based Gas Separation Processes. *J. Membr. Sci.* **2017**, *526*, 118−130.
(21) Belaissaoui, B.; Le Moullec, Y.; Willson, D.; Favre, E. Hybrid Membrane Cryogenic Process for Post-Combustion CO2 Capture. *J. Membr. Sci.* **2012**, *415*−*416*, 424−434.
(22) Song, C.; Liu, Q.; Ji, N.; Deng, S.; Zhao, J.; Li, Y.; Song, Y.; Li, H. Alternative Pathways for Efficient CO_2 Capture by Hybrid Processes—A Review. *Renewable Sustainable Energy Rev.* **2018**, *82*, 215−231.
(23) Song, C.; Liu, Q.; Ji, N.; Deng, S.; Zhao, J.; Kitamura, Y. Advanced Cryogenic CO_2 Capture Process Based on Stirling Coolers by Heat Integration. *Appl. Therm. Eng.* **2017**, *114*, 887−895.
(24) Xie, N.; Chen, B.; Tan, C.; Liu, Z. Energy Consumption and Exergy Analysis of MEA-Based and Hydrate-Based CO_2 Separation. *Ind. Eng. Chem. Res.* **2017**, *56*, 15094−15101.
(25) Thompson, M. L.; Kramer, M. A. Modeling Chemical Processes Using Prior Knowledge and Neural Networks. *AIChE J.* **1994**, *40*, 1328−1340.
(26) Lee, J. H.; Shin, J.; Realff, M. J. Machine Learning: Overview of the Recent Progresses and Implications for the Process Systems Engineering Field. *Comput. Chem. Eng.* **2018**, *114*, 111−121.
(27) Bradford, E.; Imsland, L.; Zhang, D.; del Rio Chanona, E. A. Stochastic Data-Driven Model Predictive Control Using Gaussian Processes. *Comput. Chem. Eng.* **2020**, *139*, No. 106844.
(28) Shrivastava, R.; Mahalingam, H.; Dutta, N. N. Application and Evaluation of Random Forest Classifier Technique for Fault Detection in Bioreactor Operation. *Chem. Eng. Commun.* **2017**, *204*, 591−598.
(29) Saxén, B.; Saxén, H. A Neural-Network Based Model of Bioreaction Kinetics. *Can. J. Chem. Eng.* **1996**, *74*, 124−131.
(30) Patnaik, P. Hybrid Neural Simulation of a Fed-Batch Bioreactor for a Nonideal Recombinant Fermentation. *Bioprocess Biosyst. Eng.* **2001**, *24*, 151−161.
(31) Tholudur, A.; Ramirez, W. F.Optimization of Fed-Batch Bipreactors Using Neural Network Parameter Function Models. In *Biotechnology Progress*; American Chemical Society, 1996; pp 302−309. DOI: 10.1021/bp960012h.
(32) von Stosch, M.; Oliveira, R.; Peres, J.; Feyo de Azevedo, S. Hybrid Semi-Parametric Modeling in Process Systems Engineering: Past, Present and Future. *Comput. Chem. Eng.* **2014**, *60*, 86−101.
(33) Cozad, A.; Sahinidis, N. V. A Global MINLP Approach to Symbolic Regression. *Math. Program.* **2018**, *170*, 97−119.
(34) Sahinidis, N. V. BARON: A General Purpose Global Optimization Software Package. *J. Global Optim.* **1996**, *8*, 201−205.
(35) Cozad, A.; Sahinidis, N. V.; Miller, D. C. Learning Surrogate Models for Simulation-Based Optimization. *AIChE J.* **2014**, *60*, 2211−2227.
(36) Cozad, A.; Sahinidis, N. V.; Miller, D. C. A Combined First-Principles and Data-Driven Approach to Model Building. *Comput. Chem. Eng.* **2015**, *73*, 116−127.
(37) Wilson, Z. T.; Sahinidis, N. V. The ALAMO Approach to Machine Learning. *Comput. Chem. Eng.* **2017**, *106*, 785−795.
(38) Grossmann, I. E. Review of Nonlinear Mixed-Integer and Disjunctive Programming Techniques. *Optim. Eng.* **2002**, *3*, 227−252.
(39) Rao, A. B.; Rubin, E. S. Identifying Cost-Effective CO_2 Control Levels for Amine-Based CO_2 Capture Systems. *Ind. Eng. Chem. Res.* **2006**, *45*, 2421−2429.
(40) Zhou, Q.; Chan, C. W.; Tontiwachiwuthikul, P. Regression Analysis Study on the Carbon Dioxide Capture Process. *Ind. Eng. Chem. Res.* **2008**, *47*, 4937−4943.
(41) Zhou, Q.; Wu, Y.; Chan, C. W.; Tontiwachwuthikul, P. Modeling of the Carbon Dioxide Capture Process System Using Machine Intelligence Approaches. *Eng. Appl. Artif. Intell.* **2011**, *24*, 673−685.
(42) Liang, Z. H.; Rongwong, W.; Liu, H.; Fu, K.; Gao, H.; Cao, F.; Zhang, R.; Sema, T.; Henni, A.; Sumon, K.; Nath, D.; Gelowitz, D.; Srisang, W.; Saiwan, C.; Benamor, A.; Al-Marri, M.; Shi, H.; Supap, T.; Chan, C.; Zhou, Q.; Abu-Zahra, M.; Wilson, M.; Olson, W.; Idem, R.; Tontiwachwuthikul, P. PT Recent Progress and New Developments in Post-Combustion Carbon-Capture Technology with Amine Based Solvents. *Int. J. Greenhouse Gas Control* **2015**, *40*, 26−54.

https://doi.org/10.1021/acsomega.2c04736
ACS Omega 2022, 7, 41147−41164

(43) Li, Z.; Sharma, M.; Khalilpour, R.; Abbas, A.Optimal Operation of Solvent-Based Post-Combustion Carbon Capture Processes with Reduced Models. In *Energy Procedia*; Elsevier, 2013; Vol. *37*, pp 1500−1508. DOI: 10.1016/J.EGYPRO.2013.06.025.
(44) Danaci, D.; Bui, M.; Petit, C.; Mac Dowell, N. En Route to Zero Emissions for Power and Industry with Amine-Based Post-Combustion Capture. *Environ. Sci. Technol.* **2021**, *55*, 10619−10632.
(45) Pascual-González, J.; Pozo, C.; Guillén-Gosálbez, G.; Jiménez-Esteller, L. Combined Use of MILP and Multi-Linear Regression to Simplify LCA Studies. *Comput. Chem. Eng.* **2015**, *82*, 34−43.
(46) Miró, A.; Pozo, C.; Guillén-Gosálbez, G.; Egea, J. A.; Jiménez, L. Deterministic Global Optimization Algorithm Based on Outer Approximation for the Parameter Estimation of Nonlinear Dynamic Biological Systems. *BMC Bioinf.* **2012**, *13*, No. 90.
(47) Gkioulekas, I.; Papageorgiou, L. G. Tree Regression Models Using Statistical Testing and Mixed Integer Programming. *Comput. Ind. Eng.* **2021**, *153*, No. 107059.
(48) Ferreira, J.; Torres, A. I.; Pedemonte, M.Towards a Multi-Output Kaizen Programming Algorithm. In *2021 IEEE Latin American Conference on Computational Intelligence*; Institute of Electrical and Electronics Engineers Inc., 2021. DOI: 10.1109/LA-CCI48322.2021.9769841.
(49) Neumann, P.; Cao, L.; Russo, D.; Vassiliadis, V. S.; Lapkin, A. A. A New Formulation for Symbolic Regression to Identify Physico-Chemical Laws from Experimental Data. *Chem. Eng. J.* **2020**, *387*, No. 123412.
(50) Ouyang, R.; Curtarolo, S.; Ahmetcik, E.; Scheffler, M.; Ghiringhelli, L. M. SISSO: A Compressed-Sensing Method for Identifying the Best Low-Dimensional Descriptor in an Immensity of Offered Candidates. *Phys. Rev. Mater.* **2018**, *2*, 083802.
(51) Ansari, M.; Gandhi, H. A.; Foster, D. G.; White, A. D. Iterative Symbolic Regression for Learning Transport Equations. *AIChE J.* **2022**, *68*, No. e17695.
(52) Sun, W.; Braatz, R. D. ALVEN: Algebraic Learning via Elastic Net for Static and Dynamic Nonlinear Model Identification. *Comput. Chem. Eng.* **2020**, *143*, No. 107103.
(53) Sun, W.; Braatz, R. D. Smart Process Analytics for Predictive Modeling. *Comput. Chem. Eng.* **2021**, *144*, No. 107134.
(54) Dobbelaere, M. R.; Plehiers, P. P.; Van de Vijver, R.; Stevens, C. V.; Van Geem, K. M. Machine Learning in Chemical Engineering: Strengths, Weaknesses, Opportunities, and Threats. *Engineering* **2021**, *7*, 1201−1211.
(55) Schweidtmann, A. M.; Esche, E.; Fischer, A.; Kloft, M.; Repke, J.-U.; Sager, S.; Mitsos, A. Machine Learning in Chemical Engineering: A Perspective. *Chem. Ing. Tech.* **2021**, *93*, 2029−2039.
(56) Otte, C.*Safe and Interpretable Machine Learning: A Methodological Review*; Springer: Berlin, Heidelberg, 2013; Vol. *445*. DOI: 10.1007/978-3-642-32378-2_8/COVER.
(57) Guimerà, R.; Reichardt, I.; Aguilar-Mogas, A.; Massucci, F. A.; Miranda, M.; Pallarès, J.; Sales-Pardo, M. A Bayesian Machine Scientist to Aid in the Solution of Challenging Scientific Problems. *Sci. Adv.* **2020**, *6*, No. eaav6971.
(58) Vázquez, D.; Guimerà, R.; Sales-Pardo, M.; Guillén-Gosálbez, G. Automatic Modeling of Socioeconomic Drivers of Energy Consumption and Pollution Using Bayesian Symbolic Regression. *Sustainable Prod. Consumption* **2022**, *30*, 596−607.
(59) Liu, H.; Zhang, Z. Probing the Carbon Emissions in 30 Regions of China Based on Symbolic Regression and Tapio Decoupling. *Environ. Sci. Pollut. Res.* **2022**, *29*, 2650−2663.
(60) Li, W.; Yang, G.; Li, X. Modeling the Evolutionary Nexus between Carbon Dioxide Emissions and Economic Growth. *J. Cleaner Prod.* **2019**, *215*, 1191−1202.
(61) Subraveti, S. G.; Li, Z.; Prasad, V.; Rajendran, A. Machine Learning-Based Multiobjective Optimization of Pressure Swing Adsorption. *Ind. Eng. Chem. Res.* **2019**, *58*, 20412−20422.
(62) Liu, H.; Chan, C.; Tontiwachwuthikul, P.; Idem, R. Analysis of CO_2 Equilibrium Solubility of Seven Tertiary Amine Solvents Using Thermodynamic and ANN Models. *Fuel* **2019**, *249*, 61−72.
(63) Venkatraman, V.; Alsberg, B. K. Predicting CO_2 Capture of Ionic Liquids Using Machine Learning. *J. CO_2 Util.* **2017**, *21*, 162−168.
(64) Mesbah, M.; Shahsavari, S.; Soroush, E.; Rahaei, N.; Rezakazemi, M. Accurate Prediction of Miscibility of CO_2 and Supercritical CO_2 in Ionic Liquids Using Machine Learning. *J. CO2 Util.* **2018**, *25*, 99−107.
(65) Morgan, J. C.; Chinen, A. S.; Anderson-Cook, C.; Tong, C.; Carroll, J.; Saha, C.; Omell, B.; Bhattacharyya, D.; Matuszewski, M.; Bhat, K. S.; Miller, D. C. Development of a Framework for Sequential Bayesian Design of Experiments: Application to a Pilot-Scale Solvent-Based CO_2 Capture Process. *Appl. Energy* **2020**, *262*, No. 114533.
(66) Kim, Y.; Jang, H.; Kim, J.; Lee, J. Prediction of Storage Efficiency on CO_2 Sequestration in Deep Saline Aquifers Using Artificial Neural Network. *Appl. Energy* **2017**, *185*, 916−928.
(67) Helei, L.; Tantikhajorngosol, P.; Chan, C.; Tontiwachwuthikul, P. Technology Development and Applications of Artificial Intelligence for Post-Combustion Carbon Dioxide Capture: Critical Literature Review and Perspectives. *Int. J. Greenhouse Gas Control* **2021**, *108*, No. 103307.
(68) Shalaby, A.; Elkamel, A.; Douglas, P. L.; Zhu, Q.; Zheng, Q. P. A Machine Learning Approach for Modeling and Optimization of a CO_2 Post-Combustion Capture Unit. *Energy* **2021**, *215*, No. 119113.
(69) Sipöcz, N.; Tobiesen, F. A.; Assadi, M. The Use of Artificial Neural Network Models for CO_2 Capture Plants. *Appl. Energy* **2011**, *88*, 2368−2376.
(70) Wilberforce, T.; Baroutaji, A.; Soudan, B.; Al-Alami, A. H.; Olabi, A. G. Outlook of Carbon Capture Technology and Challenges. *Sci. Total Environ.* **2019**, *657*, 56−72.
(71) Bui, M.; Adjiman, C. S.; Bardow, A.; Anthony, E. J.; Boston, A.; Brown, S.; Fennell, P. S.; Fuss, S.; Galindo, A.; Hackett, L. A.; Hallett, J. P.; Herzog, H. J.; Jackson, G.; Kemper, J.; Krevor, S.; Maitland, G. C.; Matuszewski, M.; Metcalfe, I. S.; Petit, C.; Puxty, G.; Reimer, J.; Reiner, D. M.; Rubin, E. S.; Scott, S. A.; Shah, N.; Smit, B.; Trusler, J. P. M.; Webley, P.; Wilcox, J.; Mac Dowell, N. Carbon Capture and Storage (CCS): The Way Forward. *Energy Environ. Sci.* **2018**, *11*, 1062−1176.
(72) Song, C.; Liu, Q.; Ji, N.; Deng, S.; Zhao, J.; Kitamura, Y. Natural Gas Purification by Heat Pump Assisted MEA Absorption Process. *Appl. Energy* **2017**, *204*, 353−361.
(73) Schach, M.-O.; Schneider, R.; Schramm, H.; Repke, J.-U. Techno-Economic Analysis of Postcombustion Processes for the Capture of Carbon Dioxide from Power Plant Flue Gas. *Ind. Eng. Chem. Res.* **2010**, *49*, 2363−2370.
(74) Tay, W. H.; Lau, K. K.; Lai, L. S.; Shariff, A. M.; Wang, T. Current Development and Challenges in the Intensified Absorption Technology for Natural Gas Purification at Offshore Condition. *J. Nat. Gas Sci. Eng.* **2019**, *71*, No. 102977.
(75) Adams, T. A., II; Salkuyeh, Y. K.; Nease, J.Processes and Simulations for Solvent-Based CO_2 Capture and Syngas Cleanup. In *Reactor and Process Design in Sustainable Energy Technology*; Elsevier Inc., 2014; pp 163−231. DOI: 10.1016/B978-0-444-59566-9.00006-5.
(76) Adams, D.*Flue Gas Treatment for CO_2 Capture*; **2010**.
(77) Li, K.; Leigh, W.; Feron, P.; Yu, H.; Tade, M. Systematic Study of Aqueous Monoethanolamine (MEA)-Based CO_2 Capture Process: Techno-Economic Assessment of the MEA Process and Its Improvements. *Appl. Energy* **2016**, *165*, 648−659.
(78) Wattanaphan, P.; Sema, T.; Idem, R.; Liang, Z.; Tontiwachwuthikul, P. Effects of Flue Gas Composition on Carbon Steel (1020) Corrosion in MEA-Based CO_2 Capture Process. *Int. J. Greenhouse Gas Control* **2013**, *19*, 340−349.
(79) Petrakopoulou, F.*Comparative Evaluation of Power Plants with CO2 Capture: Thermodynamic, Economic and Environmental Performance*; Technischen Universität: Berlin, 2011.
(80) Wu, X.; Wang, M.; Liao, P.; Shen, J.; Li, Y. Solvent-Based Post-Combustion CO_2 Capture for Power Plants: A Critical Review and Perspective on Dynamic Modelling, System Identification, Process Control and Flexible Operation. *Appl. Energy* **2020**, *257*, No. 113941.

https://doi.org/10.1021/acsomega.2c04736
ACS Omega 2022, 7, 41147−41164

(81) Scholes, C. A.; Stevens, G. W.; Kentish, S. E. Membrane Gas Separation Applications in Natural Gas Processing. *Fuel* **2012**, *96*, 15−28.
(82) Baker, R. W.; Lokhandwala, K.Natural Gas Processing with Membranes: An Overview. In *Industrial and Engineering Chemistry Research*; American Chemical Society, April 2, 2008; pp 2109−2121. DOI: 10.1021/ie071083w.
(83) Agbonghae, E. O.; Hughes, K. J.; Ingham, D. B.; Ma, L.; Pourkashanian, M. Optimal Process Design of Commercial-Scale Amine-Based CO_2 Capture Plants. *Ind. Eng. Chem. Res.* **2014**, *53*, 14815−14829.
(84) Žegklitz, J.; Pošík, P. Benchmarking State-of-the-Art Symbolic Regression Algorithms. *Genet. Program. Evolvable Mach.* **2021**, *22*, 5−33.
(85) Franceschini, G.; Macchietto, S. Model-Based Design of Experiments for Parameter Precision: State of the Art. *Chem. Eng. Sci.* **2008**, *63*, 4846−4872.
(86) *Amine Scrubbing with Aspen HYSYS V8. 0*, AspenTech, 2012.
(87) Savage, T. R.; Almeida-Trasvina, F.; del-Rio Chanona, E. A.; Smith, R.; Zhang, D. An Integrated Dimensionality Reduction and Surrogate Optimization Approach for Plant-Wide Chemical Process Operation. *AIChE J.* **2021**, *67*, No. e17358.
(88) van de Berg, D.; Savage, T.; Petsagkourakis, P.; Zhang, D.; Shah, N.; del Rio-Chanona, E. A. Data-Driven Optimization for Process Systems Engineering Applications. *Chem. Eng. Sci.* **2022**, *248*, No. 117135.
(89) Bongartz, D.; Najman, J.; Sass, S.; Mitsos, A.*MAiNGO − McCormick-Based Algorithm for Mixed-Integer Nonlinear Global Optimization*; 2018.
(90) Schweidtmann, A. M.; Mitsos, A. Deterministic Global Optimization with Artificial Neural Networks Embedded. *J. Optim. Theory Appl.* **2019**, *180*, 925−948.
(91) Santos, L. F.; Costa, C. B. B.; Caballero, J. A.; Ravagnani, M. A. S. S. Framework for Embedding Black-Box Simulation into Mathematical Programming via Kriging Surrogate Model Applied to Natural Gas Liquefaction Process Optimization. *Appl. Energy* **2022**, *310*, No. 118537.
(92) Ye, W.; You, F. A Computationally Efficient Simulation-Based Optimization Method with Region-Wise Surrogate Modeling for Stochastic Inventory Management of Supply Chains with General Network Structures. *Comput. Chem. Eng.* **2016**, *87*, 164−179.
(93) Hao, Z.; Barecka, M.; A Lapkin, A. Accelerating Net Zero from the Perspective of Optimizing a Carbon Capture and Utilization System. *Energy Environ. Sci.* **2022**, *15*, 2139−2153.
(94) Kim, S. H.; Boukouvala, F. Surrogate-Based Optimization for Mixed-Integer Nonlinear Problems. *Comput. Chem. Eng.* **2020**, *140*, No. 106847.
(95) Boukouvala, F.; Ierapetritou, M. G. Surrogate-Based Optimization of Expensive Flowsheet Modeling for Continuous Pharmaceutical Manufacturing. *J. Pharm. Innovation* **2013**, *8*, 131−145.
(96) Vollmer, N. I.; Al, R.; Gernaey, K. V.; Sin, G. Synergistic Optimization Framework for the Process Synthesis and Design of Biorefineries. *Front. Chem. Sci. Eng.* **2022**, *16*, 251−273.
(97) Halemane, K. P.; Grossmann, I. E. Optimal Process Design under Uncertainty. *AIChE J.* **1983**, *29*, 425−433.
(98) Grossmann, I. E.; Floudas, C. A. Active Constraint Strategy for Flexibility Analysis in Chemical Processes. *Comput. Chem. Eng.* **1987**, *11*, 675−693.
(99) Rogers, A.; Ierapetritou, M. Feasibility and Flexibility Analysis of Black-Box Processes Part 1: Surrogate-Based Feasibility Analysis. *Chem. Eng. Sci.* **2015**, *137*, 986−1004.
(100) Dias, L. S.; Ierapetritou, M. G. Integration of Planning, Scheduling and Control Problems Using Data-Driven Feasibility Analysis and Surrogate Models. *Comput. Chem. Eng.* **2020**, *134*, No. 106714.
(101) Badejo, O.; Ierapetritou, M. Integrating Tactical Planning, Operational Planning and Scheduling Using Data-Driven Feasibility Analysis. *Comput. Chem. Eng.* **2022**, *161*, No. 107759.
(102) Ding, C.; Ierapetritou, M. A Novel Framework of Surrogate-Based Feasibility Analysis for Establishing Design Space of Twin-Column Continuous Chromatography. *Int. J. Pharm.* **2021**, *609*, No. 121161.
(103) Zhang, D.; Del Rio-Chanona, E. A.; Petsagkourakis, P.; Wagner, J. Hybrid Physics-Based and Data-Driven Modeling for Bioprocess Online Simulation and Optimization. *Biotechnol. Bioeng.* **2019**, *116*, 2919−2930.
(104) Boukouvala, F.; Floudas, C. A. ARGONAUT: AlgoRithms for Global Optimization of CoNstrAined Grey-Box CompUTational Problems. *Optim. Lett.* **2017**, *11*, 895−913.
(105) Tsay, C. Sobolev Trained Neural Network Surrogate Models for Optimization. *Comput. Chem. Eng.* **2021**, *153*, No. 107419.

https://doi.org/10.1021/acsomega.2c04736
ACS Omega 2022, 7, 41147−41164