Computers and Chemical Engineering 182 (2024) 108563

Contents lists available at ScienceDirect

Computers and Chemical Engineering

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

Hybrid analytical surrogate-based process optimization via Bayesian symbolic regression

Sachin Jog [a], Daniel Vázquez [b], Lucas F. Santos [c,d], José A. Caballero [d], Gonzalo Guillén-Gosálbez [a,*]

[a] Department of Chemistry and Applied Biosciences, Institute for Chemical and Bioengineering, ETH Zurich, Vladimir-Prelog-Weg 1, Zurich 8093, Switzerland
[b] IQS School of Engineering, Universitat Ramon Llull, Via Augusta 390, Barcelona 08017, Spain
[c] Chemical Engineering Department, State University of Maringá, Avenida Colombo 5790, Maringá 87020900, Brazil
[d] Institute of Chemical Process Engineering, University of Alicante, Ap. Correos 99, Alicante 03080, Spain

ARTICLE INFO

Keywords:
Process optimization
Hybrid surrogate models
Black-box surrogate models
Bayesian symbolic regression

ABSTRACT

Modular chemical process simulators are widespread in chemical industries to design and optimize production processes with sufficient accuracy. However, convergence issues and entrapment in local optima during process optimization are still challenges to overcome. To circumvent them, surrogate models of first principles simulations have attracted attention as they are easier to handle, with hybrid surrogates combining data-driven surrogate models with mechanistic equations becoming particularly appealing. In this context, this work explores the use of Bayesian symbolic regression to construct and globally optimize hybrid analytical surrogate models of process flowsheets, where some units are approximated with tailored analytical expressions rather than with neural networks or Gaussian processes, which might be harder to globally optimize. Comparing with other prevalent black-box surrogate modeling & optimization approaches, such as kriging and Bayesian optimization, we find that our approach can find better solutions than those identified with pure black-box methodologies, yet model building is much more computationally demanding.

1. Introduction

Simulation software is commonly used in Process Systems Engineering to model complex chemical processes. Particularly, standard process units, such as distillation columns and reactors, can be modeled *via* first principles, which allows for their analysis and optimization. However, using standard simulators, such as the Aspen® suite, for optimization is not exempt from limitations. Specifically, while many simulation packages have built-in optimization capabilities, numerical noise makes it challenging to estimate derivatives accurately, leading to a prohibitive number of function evaluations and a substantial computation time (McBride and Sundmacher, 2019). Additionally, due to the non-convex nature of the nonlinear equations derived from first principles, multi-modality might become an issue, resulting in potential entrapment in local optima or even failure to converge to a feasible point.

If the process simulation model could be approximated with a simpler formulation using explicit input-output relationships (Caballero and Grossmann, 2008), state-of-the-art commercial solvers could be used to efficiently identify optimal solutions. Developing and optimizing such models, also called surrogate models, metamodels, or reduced order models (Papalambros and Wilde, 2000), is an active research field. A variety of techniques, such as polynomial regression (Ostertagová, 2012), support vector regression (Xiang *et al.*, 2017), artificial neural networks (ANNs) (Kahrs and Marquardt, 2007; Lim *et al.*, 2002), radial basis function networks (Kramer *et al.*, 1992), random forests (Williams and Cremaschi, 2019), and kriging (Santos *et al.*, 2022), among others, have been employed for surrogate model generation.

Surrogate models constructed through a direct mapping between some process inputs and the corresponding outputs, without any prior process knowledge, *i.e.*, a black-box model, (also referred to as surrogate model at the system level, in which a single surrogate model is used to represent the whole system (Misener and Biegler, 2023)) often lack interpretability due to the absence of mechanistic insights. This, in turn, can lead to overfitting and poor generalization (Psichogios and Ungar, 1992), which is why considerable attention has been devoted recently to developing hybrid surrogate models. A hybrid model improves upon the black-box model by adding mechanistic equations, such as mass balances, thus combining the strengths of data-driven approaches and first

* Corresponding author.
E-mail address: gonzalo.guillen.gosalbez@chem.ethz.ch (G. Guillén-Gosálbez).

https://doi.org/10.1016/j.compchemeng.2023.108563
Received 16 August 2023; Received in revised form 25 November 2023; Accepted 19 December 2023
Available online 20 December 2023

Nomenclature

Abbreviations

ANN	Artificial neural network
CSTR	Continuous stirred tank reactor
AIC	Akaike information criterion
BIC	Bayesian information criterion
BO	Bayesian optimization
NLP	Nonlinear programming
GP	Gaussian process
KP	Kaizen programming
MINLP	Mixed-integer nonlinear programming
ALAMO	Automated learning of algebraic models
BMS	Bayesian machine scientist
MCMC	Markov chain Monte Carlo
GAMS	General algebraic modeling system
OF	Objective function
LHS	Latin hypercube sampling
HY	Hybrid
BB	Black-box
TAC	Total annualized cost
ACCR	Annual capital charge ratio
CAPEX	Capital expenditures
OPEX	Operational expenditures
MSE	Mean squared error
MAPE	Mean absolute percentage error
GA	Genetic algorithm
EI	Expected improvement
DAC	Direct air capture
PFR	Plug flow reactor

Sets

I	$\{i$: Set of process units that are replaced with a surrogate$\}$
J	$\{j$: Set of surrogate process models$\}$
C	$\{c$: Set of continuous process variables$\}$
U	$\{u$: Set of continuous structural variables$\}$
Z	$\{z$: Set of integer variables$\}$
A	$\{a$: Set of all components in the flowsheet of case study 1$\}$
B	$\{b$: Set of all components in the flowsheet of case study 2$\}$
$JI_i \subset J$	Set defining each surrogate model j of each process unit i
$N^C \subset C$	$\{c$: Set of continuous process variables defined as degrees of freedom of the whole flowsheet$\}$
$N^U \subset U$	$\{u$: Set of continuous structural variables defined as degrees of freedom of the whole flowsheet$\}$
$N^Z \subset Z$	$\{z$: Set of integer variables defined as degrees of freedom of the whole flowsheet$\}$
$IM_{ij}^C \subset C$	$\{c$: Set of continuous process variables that act as inputs for process unit i and model $j\}$
$IM_{ij}^U \subset U$	$\{u$: Set of continuous structural variables that act as inputs for process unit i and model $j\}$
$IM_{ij}^Z \subset Z$	$\{z$: Set of integer variables that act as inputs for process unit i and model $j\}$
$OM_{ij} \subset C$	$\{c$: Set of continuous process variables that are obtained as outputs from process unit i and model $j\}$

Variables

x	Continuous variable
c	Continuous process variable
u	Continuous structural variable
z	Integer variable
x_c	Value of continuous process variable c
d_u	Value of continuous structural variable u
y_z	Value of integer variable z

principles knowledge. Sansana *et al.* (2021) provided a detailed review of hybrid modeling in chemical process modeling, including (and not limited to) applications in process optimization, monitoring, and control. ANNs have been mostly employed for the hybrid modeling and optimization of process units or even entire process flowsheets, *e.g.*, Psichogios and Ungar (1992) successfully modeled the cell growth rate in a bioreactor using an ANN complemented with mass balances. They concluded that the hybrid model had a greater ability for generalization and extrapolation. Henao and Maravelias (2011) and Fahmi and Cremaschi (2012) substituted individual process units with ANN-based surrogate models, which were complemented with mechanistic equations, in a superstructure-based optimization. Caballero and Grossmann (2008) used kriging (Krige, 1952) metamodels to substitute a distillation column and a plug flow reactor and iteratively optimized the process unit using a local solver. This work was later extended by Quirante and Caballero (2016) and Quirante *et al.* (2018), who optimized process flowsheets that included both surrogate kriging units and standard process units in a hybrid model. In the field of continuous pharmaceutical manufacturing, Bhalode *et al.* (2022) developed a hybrid adaptive modeling framework, with the data-driven model being a neural network. The hybrid model results were compared to plant outputs, and the model was then updated (if necessary) based on pre-defined criteria. Additionally, they integrated multi-scale information using hybrid multi-zonal compartment models. These two hybrid strategies were envisaged to aid in the creation of an integrated digital twin in continuous pharmaceutical processes. Further, Chen and Ierapetritou (2020) investigated the plant-model mismatch in first principles models (*i.e.*, cases in which historical plant data do not match well with the predictions of a model). The first principles models were then re-analyzed to generate hybrid models (with serial, parallel, and combined structures, using ANN for the data-driven part), and their implementation was discussed for a continuous stirred tank reactor (CSTR), as well as in two case studies in continuous pharmaceutical processes. In order to assess different hybrid modeling approaches, Bradley and Boukouvala (2014) first used a sequential approach (*i.e.*, where the neural network, used as the data-driven model, and the mechanistic model are generated independently). They compared such an approach with an integrated one, in which the solver optimized the neural network weights while simultaneously considering the mechanistic model-based constraints. They concluded that the integrated approach performed better in terms of constraint violation, but highlighted the computational limitations of using neural networks. In the domain of dynamic modeling of bioprocesses, Wang *et al.* (2023) demonstrated the use of different statistical methods for hybrid model selection (*e.g.*, Akaike information criterion (AIC), Bayesian information criterion (BIC), etc.). Applying the methodology to a microalgae cultivation case study, they concluded that the hybrid model selected using the BIC performed well for different noise levels.

While hybrid models provide more versatility than black-box models (Psichogios and Ungar, 1992), the non-convex nature of the optimization problem can still lead to suboptimal solutions if a global solver is not used (Schweidtmann *et al.*, 2021). While derivative-free optimization techniques, such as genetic algorithms (GAs), simulated annealing, and Bayesian optimization (BO), can provide high-quality solutions, they cannot guarantee the solution's global (nor local) optimality. In contrast, deterministic global solvers ensure convergence to the global optimum within a given tolerance, but often lead to large CPU times. The ARGONAUT framework (Boukouvala and Floudas, 2017) provides

an iterative methodology for the global optimization of constrained grey-box problems using the global solver ANTIGONE (Misener and Floudas, 2014). This approach solves subproblems to global optimality at each iteration, which increases the computational burden substantially, thus potentially limiting its practical use in large problems. Bongartz *et al.* (2018) introduced the tailored spatial branch and bound solver MAiNGO based on the propagation of McCormick relaxations (McCormick, 1976) in a reduced space. They showed the successful implementation of this approach in nonlinear programming (NLP) problems modeled using ANNs (Schweidtmann and Mitsos, 2019) as well as Gaussian processes (GPs) (Schweidtmann *et al.*, 2021).

Here we shall explore the construction and optimization of hybrid models for process optimization, where the surrogate component is built using symbolic regression. Symbolic regression, which has recently gained wide interest, represents any closed-form mathematical expression as an expression tree (Koza, 1992). This tree structure can be used to build regression models to fit given data. Specifically, given some data, suitable regression models can be found by the simultaneous optimization of the tree structure and the operators and values at each node of the tree (Cozad and Sahinidis, 2018).

There have been numerous recent contributions in symbolic regression, yet the applications in chemical engineering are still scarce. Symbolic regression approaches include genetic programming (Orzechowski *et al.*, 2018) and Kaizen programming (KP). KP utilizes genetic programming to construct basis functions, which are then linearly combined *via* linear regression. This approach has been shown to outperform standalone genetic programming-based models in some problems (Ferreira *et al.*, 2019). Specifically, focusing on chemical engineering applications, Ferreira *et al.* (2022) used real data from an oil refinery to build models based on KP to predict the composition of C4 hydrocarbons in the distillate stream of a splitter column and compared their results with a GP model, showing that the KP model outperformed the GP model. In a recent work, Cozad and Sahinidis (2018) developed an elegant mixed integer nonlinear programming (MINLP) formulation for symbolic regression using disjunctive programming, which was solved *via* deterministic global optimization (BARON (Tawarmalani and Sahinidis, 2005)).

Focusing on closed-form algebraic expressions for surrogate modeling, Ma *et al.* (2022b) recently investigated surrogate models constructed using ALAMO (Automated learning of algebraic models) (Cozad *et al.*, 2014), which generates closed-form algebraic expressions by choosing from a user-defined list of basis functions and applying various model selection criteria. They compared the optimization of various surrogate models constructed using ALAMO for an extractive distillation process. Ma *et al.* (2022a) also investigated the deterministic, discrete-time optimization of an integrated chemical plant using different abstraction levels of surrogate modeling (*i.e.*, unit-level and plant-level), where the surrogate models were constructed using ALAMO. They concluded that the highest level of abstraction, *i.e.*, plant-level abstraction, performed the best for enterprise-wide optimization.

Recently, Guimerà *et al.* (2020) introduced the Bayesian machine scientist (BMS), a symbolic regression tool that generates closed-form algebraic expressions from data. Their method employs an empirical corpus of equations to quantify prior expectations about the model, while the exact marginal posterior over models is computed using explicit approximations. The search space is navigated using Markov chain Monte Carlo (MCMC) sampling, and successive expressions are generated *via* heuristics and accepted or rejected according to the Metropolis' rule (Metropolis *et al.*, 1953). Focusing on chemical engineering applications, Negri *et al.* (2022) applied the BMS to build closed-form analytical expressions of two CO_2 capture processes simulated in Aspen HYSYS®. More recently, the BMS was used in the global optimization of process flowsheets by Forster *et al.* (2023), who constructed black-box surrogate models using the BMS and GPs. They showed that the former models could be globally optimized in shorter CPU times using state-of-the-art solvers, yet they required more time for model building.

Fig. 1 shows a simple example of expression trees of mathematical expressions generated using the BMS. Considering a CSTR with inputs F_1 and T_1, the figure shows representative equations to model output F_2. The different moves (*i.e.*, node replacement, root addition, root removal, and elementary tree replacement) that are used by the BMS to explore the space of closed-form expressions are also depicted.

In this work, we shall use Bayesian symbolic regression to generate a hybrid model of a process flowsheet and subsequently optimize it using state-of-the-art solvers. In essence, we capitalize on the ability of the BMS to generate closed-form analytical expressions which can be easier to optimize relative to other general-purpose black-box models (*e.g.*, ANNs, GPs). These analytical equations are then inserted into algebraic formulations containing mechanistic equations that can be optimized in any modeling system (*e.g.*, general algebraic modeling system (GAMS), Pyomo) using off-the-shelf solvers, including global optimization algorithms. The approach is compared against other surrogate modeling methodologies, such as kriging and GP, optimized using BO and algebraic solvers.

The structure of the paper is as follows. We first describe the problem statement and the methodology adopted in this work. This is followed by two case studies and the associated results. The final section presents the conclusions of the work.

2. Problem statement

We aim to optimize a general process flowsheet model effectively. In what follows, we shall consider the following general mathematical formulation:

$$\begin{aligned} &min\ F(x,z) \\ &s.t.\ h(x,z)=0 \\ &\quad g(x,z)\le 0 \underline{x}\le x\le \overline{x} \\ &\quad \underline{z}\le z\le \overline{z} \\ &\quad x\in R^n,\ z\in Z \end{aligned} \tag{1}$$

where $F(x,z)$ is the objective function (OF) to be minimized, which is a function of the variables of the process flowsheet, denoted by x and z, where x represents continuous variables, and z represents integer variables. The equality and inequality constraints are denoted by $h(x,z)$ and $g(x,z)$, respectively. The lower bounds of the continuous and integer variables are denoted by $\underline{x}$ and $\underline{z}$, respectively, while the upper bounds are represented by $\overline{x}$ and $\overline{z}$, respectively. While any integer variable can be expressed as a linear combination of binary variables, here we shall keep integers in the formulation as they already represent some design decisions, *e.g.*, the number of trays in a distillation column.

Here we approximate the problem above with a hybrid analytical surrogate process model, which can then be optimized with deterministic global optimization algorithms using off-the-shelf global solvers. The main idea is illustrated in Fig. 2, which shows how the hybrid formulation consists of two complementary parts: (i) the data-driven part, built by solving a symbolic regression problem using the BMS, and (ii) the mechanistic part, which is based on first principles (*i.e.*, mass and energy balances, thermodynamic constraints, etc.).

As an example, the process flowsheet in Fig. 2 consists of a mixer, CSTR, distillation column, and splitter. Here, the mixer and splitter could be modeled with mechanistic models. Meanwhile, the CSTR and distillation column would be replaced with surrogate models. Accordingly, for each output of interest, an analytical surrogate would be built using Bayesian symbolic regression (applying the BMS). For example, F_2 and T_2 in the CSTR would be modeled with functions f_1 and f_2, using the degrees of freedom of the input stream as inputs (*i.e.*, F_1 and T_1). The mechanistic and surrogate models would be then combined into a hybrid analytical surrogate model, optimized using deterministic solvers.

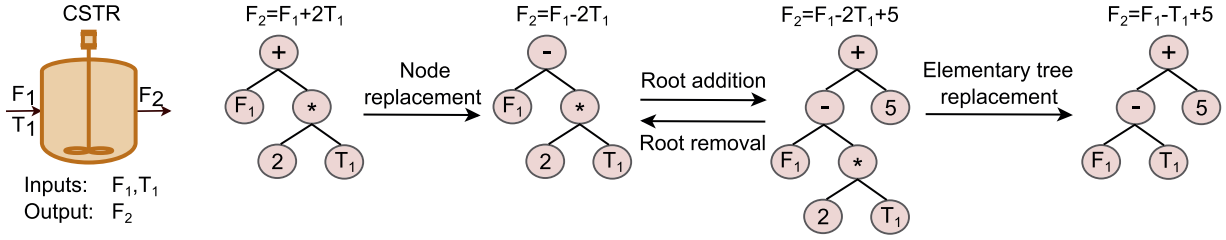

Fig. 1. Representation of mathematical equations as expression trees and different moves used by the BMS to explore the space of closed-form expressions (Guimerà et al., 2020).

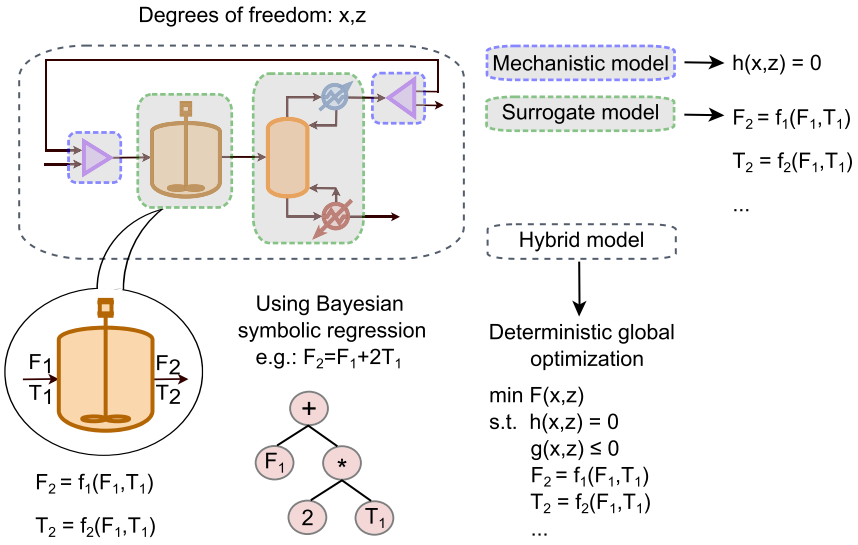

Fig. 2. Hybrid analytical surrogate modeling using Bayesian symbolic regression.

3. Methodology

This section describes the surrogate modeling and optimization framework adopted in this work. A schematic representation of the methodology is shown in Fig. 3, and each of its steps is described in the following sub-sections. In essence, we first run detailed simulations from which we build surrogate models that are combined with mechanistic equations prior to their optimization.

3.1. Data sampling and generation of outputs from the process flowsheets

The input dataset for each degree of freedom is generated *via* Latin hypercube sampling (LHS) applied within the allowable range. For each point of the dataset, the simulation is reinitialized and run by fixing the values of the degrees of freedom in the process simulator. If the simulation converges, the relevant values of the outputs from each process unit are stored, otherwise, the point is discarded and we move on to the next point.

3.2. Surrogate modeling and optimization of the process flowsheets using different methods

In this work, we create several surrogate models for a set of replaced process units, which are then combined with mechanistic equations in a hybrid (HY) model. For each of these surrogate models, the different outputs are modeled using the relevant inputs of the individual process unit being considered. The general formulation of the HY model is as follows:

$$x_c = m_{ij}\left(x_{c'\in IM_{ij}^C},\ d_{u\in IM_{ij}^U},\ y_{z\in IM_{ij}^Z}\right)\ \forall\ i\in I, j\in JI_i, c\in OM_{ij} \tag{2}$$

The sets $C, U,$ and Z contain the continuous process variables c, continuous structural variables u, and integer variables z, respectively. Note that these sets do not include all the possible variables in the process flowsheet, but only those used in the analysis. For example, the temperature of a tray in a column might be omitted if the column is approximated with a surrogate, but is needed to evaluate the performance of the column when mechanistic tray-by-tray equations are applied. The values of the continuous process variables are denoted by x_c, those of continuous structural variables by d_u, and those of integer variables by y_z. I is the set of process units i that are replaced with a surrogate, and J is the set of surrogate process models j for each process unit i (recall that more than one surrogate might be required to model one process unit depending on the outputs of interest). Thus, each surrogate process model j of process unit i is defined by the subset $JI_i \subset J$. The input-output relationship is given by the function m_{ij}, which is a function of the inputs of the specific surrogate process unit. Further, $IM_{ij}^C \subset C$ denotes the subset of continuous process variables c that are fed to model j of process unit i. Similarly, $IM_{ij}^U \subset U$ and $IM_{ij}^Z \subset Z$ are analogous to IM_{ij}^C for the continuous structural and integer variables respectively. All three subsets of input variables are defined for every process model j and its associated process unit i. $OM_{ij} \subset C$ denotes the subset of continuous process variables c that are obtained as the output of process unit i in model j. Our surrogate models have a single output variable c for each unit i in model j. Note that in this work, for simplicity, we only model output continuous process variables, while the continuous structural variables and integer variables are inputs, thus not needing to define their respective output subsets.

The HY model is compared with a fully black-box (BB) model, in which the OF to be optimized is directly modeled in terms of the degrees

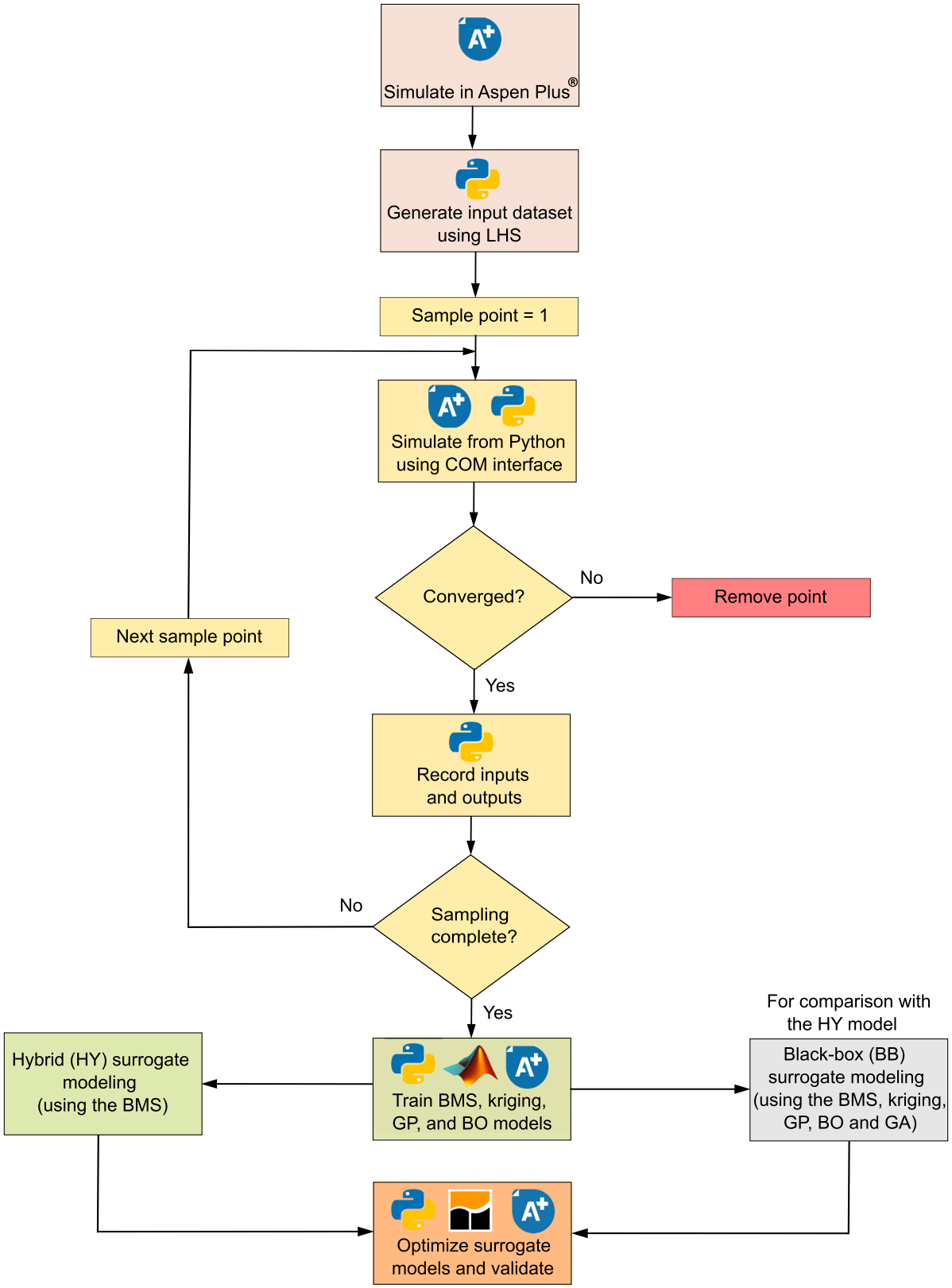

Fig. 3. Flowchart of the methodology adopted for surrogate modeling and optimization of process flowsheets.

of freedom of the whole flowsheet. The general formulation of the BB model is as follows:

$$f^{BB}\left(x_{c\in N^C},\ d_{u\in N^U}, y_{z\in N^Z}\right) \tag{3}$$

where f^{BB} directly approximates F in the set of equations defined in Equation (1) using a surrogate model. It is a function of the continuous process, continuous structural, and integer variables defined as degrees of freedom of the whole flowsheet. They are contained in the subsets $N^C \subset C$, $N^U \subset U$, and $N^Z \subset Z$, respectively.

The surrogate models defined in Equations (2) and (3) for the HY and BB models respectively, are now explained with the help of a representative process flowsheet, as shown in Fig. 4.

In this flowsheet, the pressure of all streams is assumed to be constant at 1 bar. The feed stream contains only one component A, with molar flow rate F_A^{S1} at a constant temperature T_{S1}. The pressure of stream S1 is denoted by P_{S1} (the molar flow rates, temperatures, and pressures of all other streams are defined analogously). Component A reacts in reactor R (operating isothermally at temperature T) to produce component B, as shown in reaction RE1:

$$\text{A} \rightarrow \text{B} \tag{RE1}$$

Further, a design specification (DS) ensures a 99.5% molar purity of component B in stream S4 in distillation column D, resulting in 0.5% of component A in stream S4. The rest of the components A and B (a majority of A and trace amounts of B) are recovered in stream S3. The molar flow rate of A (F_A^{S1}), reactor temperature (T), reactor volume (V), number of stages of the column (NS), and the reflux ratio of the column (RR) are the five degrees of freedom of the whole flowsheet.

Let us now consider the formulation of the HY model for this flowsheet. For the HY model, we replace the process units R and D with surrogate models. The outputs of R, estimated by surrogate models, are the conversion of component A in the reactor (X) and the heat duty of the reactor (H_R). These two outputs are denoted by R_1 and R_2, respectively. The outputs of D, estimated by surrogate models, are the molar flow rate of component B in stream S4 (F_B^{S4}), the heat duty of the condenser (H_{cond}), and the heat duty of the reboiler (H_{reb}). These outputs are denoted by D_1, D_2, and D_3, respectively. Note that the outputs approximated by surrogate models can also be obtained using mechanistic equations. For example, the heat duty of the reactor can be obtained from the conversion in the reactor and the energy balance. Therefore, the approach is flexible and the choice of using a mechanistic equation or a surrogate model for each output is up to the user.

Next, to construct surrogate models of the five outputs defined previously, we use as inputs the degrees of freedom of streams S1 and S2 (*i.e.*, the continuous process variables), and the continuous structural and integer variables of the process units R and D. The degrees of freedom of streams S1 and S2 are as follows:

Stream S1: P_{S1}, T_{S1}, F_A^{S1}

Stream S2: P_{S2}, T_{S2} (same as reactor temperature T), F_A^{S2}, F_B^{S2}

The continuous structural variables and integer variables of reactor R and distillation column D are as follows:

Process unit R:

Continuous structural variables: T, V; No integer variables

Process unit D:

Continuous structural variables: RR, DS; Integer variable: NS

Further, the sets and subsets defined above in Equations (2) and (3) contain the following elements in this example:

Sets

$$I := \{R, D\}$$
$$J := \{R_1, R_2, D_1, D_2, D_3\}$$
$$C := \left\{P_{S1}, T_{S1}, F_A^{S1}, P_{S2}, F_A^{S2}, F_B^{S2}, F_A^{S3}, F_B^{S3}, F_A^{S4}, F_B^{S4}, X, H_R, F_B^{S4}, H_{cond}, H_{reb}\right\}$$
$$U := \{T, V, RR, DS\}$$
$$Z := \{NS\}$$

Subsets for degrees of freedom of the whole flowsheet

$$N^C := \left\{P_{S1}, T_{S1}, F_A^{S1}\right\}$$
$$N^U := \{T, V, RR, DS\}$$
$$N^Z := \{NS\}$$

Subsets for process unit R

$$JI_R := \{R_1, R_2\}$$
$$IM^C_{R,R_1} := \left\{P_{S1}, T_{S1}, F_A^{S1}\right\} IM^U_{R,R_1} := \{T, V\} IM^Z_{R,R_1} := \{\}$$
$$OM_{R,R_1} := \{X\}$$
$$IM^C_{R,R_2} := \left\{P_{S1}, T_{S1}, F_A^{S1}\right\} IM^U_{R,R_2} := \{T, V\} IM^Z_{R,R_2} := \{\}$$
$$OM_{R,R_2} := \{H_R\}$$

Subsets for process unit D

$$JI_D := \{D_1, D_2, D_3\}$$
$$IM^C_{D,D_1} := \left\{P_{S2}, T, F_A^{S2}, F_B^{S2}\right\} IM^U_{D,D_1} := \{RR, DS\} IM^Z_{D,D_1} := \{NS\}$$
$$OM_{D,D_1} := \left\{F_B^{S4}\right\}$$
$$IM^C_{D,D_2} := \left\{P_{S2}, T, F_A^{S2}, F_B^{S2}\right\} IM^U_{D,D_2} := \{RR, DS\} IM^Z_{D,D_2} := \{NS\}$$
$$OM_{D,D_2} := \{H_{cond}\}$$
$$IM^C_{D,D_3} := \left\{P_{S2}, T, F_A^{S2}, F_B^{S2}\right\} IM^U_{D,D_3} := \{RR, DS\} IM^Z_{D,D_3} := \{NS\}$$
$$OM_{D,D_3} := \{H_{reb}\} \tag{4}$$

Thus, the outputs R_1, R_2, D_1, D_2, and D_3, *i.e.*, the data-driven part of the HY model, are defined as follows (with the general definition shown in Equation (2)):

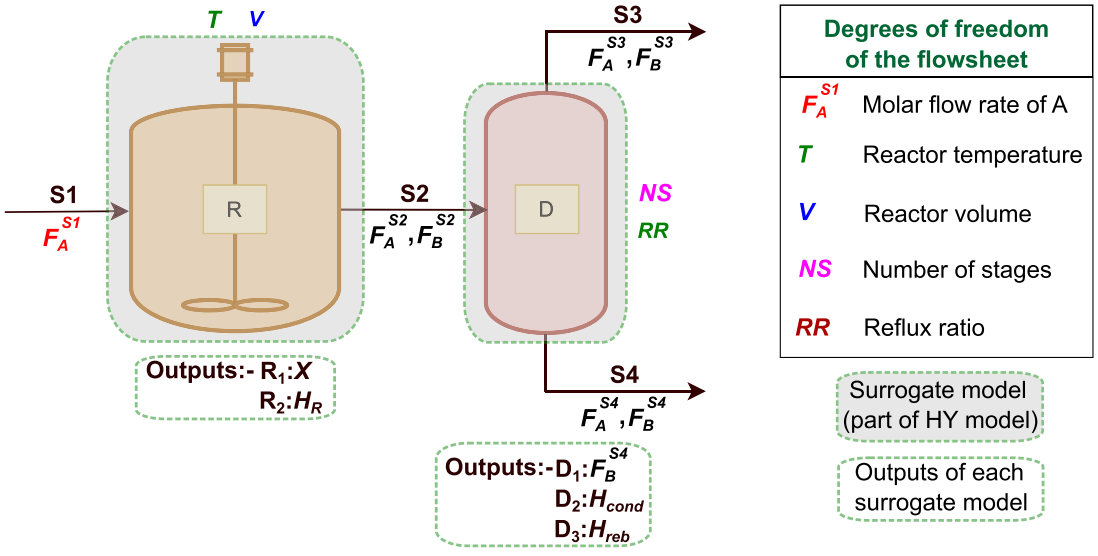

Fig. 4. Representative process flowsheet containing reactor (R) and distillation column (D).

$$\begin{aligned} x_X &= m_{R,R_1}\left(x_{P_{S1}},\ x_{T_{S1}},\ x_{F_A^{S1}},\ d_T, d_V\right) \\ x_{H_R} &= m_{R,R_2}\left(x_{P_{S1}},\ x_{T_{S1}},\ x_{F_A^{S1}},\ d_T, d_V\right) \\ x_{F_B^{S4}} &= m_{D,D_1}\left(x_{P_{S2}},\ x_T, x_{F_A^{S2}}, x_{F_B^{S2}},\ d_{RR}, d_{DS}, y_{NS}\right) \\ x_{H_{cond}} &= m_{D,D_2}\left(x_{P_{S2}},\ x_T, x_{F_A^{S2}}, x_{F_B^{S2}},\ d_{RR}, d_{DS}, y_{NS}\right) \\ x_{H_{reb}} &= m_{D,D_3}\left(x_{P_{S2}},\ x_T, x_{F_A^{S2}}, x_{F_B^{S2}},\ d_{RR}, d_{DS}, y_{NS}\right) \end{aligned} \tag{5}$$

Note that for completeness, the values of the temperature of stream S1 ($x_{T_{S1}}$), the pressures of streams S1 and S2 ($x_{P_{S1}}$ and $x_{P_{S2}}$), and the design specification (d_{DS}) have been shown as inputs in the equations in Equation (5), although they are constants.

The other part of the HY model is the mechanistic equations. In this example, we define mass balance equations based on the outputs of the surrogate models and the design specification previously mentioned. Accordingly, the following mechanistic equations defining the molar flow rates of components A and B in streams S2, S3, and S4 are part of the HY model:

$$\begin{aligned} & x_{F_A^{S2}} = x_{F_A^{S1}} \cdot (1 - x_X) \\ & x_{F_B^{S2}} = x_{F_A^{S1}} \cdot (x_X) \\ & \frac{x_{F_B^{S4}}}{x_{F_A^{S4}} + x_{F_B^{S4}}} = 0.995 \Rightarrow x_{F_A^{S4}} = \frac{0.005 \cdot x_{F_B^{S4}}}{0.995} \ \text{(as per the design specification)} \\ & x_{F_A^{S3}} = x_{F_A^{S2}} - x_{F_A^{S4}} \\ & x_{F_B^{S3}} = x_{F_B^{S2}} - x_{F_B^{S4}} \end{aligned} \tag{6}$$

The HY model thus contains the surrogate modeling equations as defined in Equation (5) and the mechanistic equations as defined in Equation (6). Further, the general mathematical formulation for the optimization of this HY model is as follows:

$$\begin{aligned} & \min f^{HY}\left(x_c,\ d_u, y_z\right) \\ & s.t.\ x_c = m_{ij}\left(x_{c' \in\ IM_{ij}^C},\ d_{u \in IM_{ij}^U},\ y_{z \in\ IM_{ij}^Z}\right) \forall i \in I, j \in JI_i, c \in OM_{ij} \\ & h\left(x_c,\ d_u, y_z\right) = 0 \\ & g\left(x_c, d_u, y_z\right) \le 0 \\ & \underline{x_c} \le x_c \le \overline{x_c} \\ & \underline{d_u} \le d_u \le \overline{d_u} \\ & \underline{y_z} \le y_z \le \overline{y_z} \\ & x_c,\ d_u \in R^n,\ y_z \in Z^+ \end{aligned} \tag{7}$$

where f^{HY} is the OF to be minimized. The first constraint describes the surrogate model equations, which are defined in Equation (5) for the representative flowsheet in Fig. 4. The explicit equality constraints $h(x_c, d_u, y_z)$ are based on first principles that connect the different surrogate and rigorous models, and denote the mass and energy balances in the HY model. These are defined in Equation (6) for the representative flowsheet. The inequality constraints $g(x_c, d_u, y_z)$ denote additional constraints imposed on the variables, such as product quality constraints, or the minimum demand of a product (there are no inequality constraints defined in the representative flowsheet). The lower bounds of the continuous process, continuous structural, and integer variables are denoted by $\underline{x_c}$, $\underline{d_u}$, and $\underline{y_z}$, respectively, while their respective upper bounds are denoted by $\overline{x_c}$, $\overline{d_u}$, and $\overline{y_z}$.

Let us now consider the formulation of the BB model for this flowsheet, which will be used for comparison with the HY model. The OF to be optimized (f^{BB}) is directly modeled with a surrogate model, and is defined as follows:

$$f^{BB}\left(x_{P_{S1}}, x_{T_{S1}}, x_{F_A^{S1}}, d_T, d_V, d_{RR},\ d_{DS}, y_{NS}\right) \tag{8}$$

Analogous to the HY model, the values of the temperature of stream S1 ($x_{T_{S1}}$), the pressure of stream S1 ($x_{P_{S1}}$), and the design specification (d_{DS}) are constants, but have been shown here for completeness.

Now, we define the specific OF quantifying the economic performance, *i.e.*, the total annualized cost (*TAC*), to be used in the optimization of the HY and BB models described above. The OF for the HY and BB models, and the *TAC* (in \$·kg^{-1} of product) are then defined as follows:

$$\begin{aligned} f^{HY} &= TAC\left(x_c, d_u, y_z\right) \\ f^{BB} &= TAC\left(x_{P_{S1}}, x_{T_{S1}}, x_{F_A^{S1}}, d_T, d_V, d_{RR},\ d_{DS}, y_{NS}\right) \\ TAC &= (ACCR \times CAPEX) + OPEX \end{aligned} \tag{9}$$

where *ACCR* denotes the annual capital charge ratio, *CAPEX* denotes the capital expenditures (such as the capital costs of the reactor and distillation column), and *OPEX* denotes the operational expenditures (such as the costs of component A, heating utilities and cooling utilities).

3.3. Performance metrics

Errors in the training and test sets are obtained for each surrogate modeling method prior to their optimization. For this purpose, we use the coefficient of determination (R^2), mean squared error (MSE), and mean absolute percentage error (MAPE) as performance metrics. Each of these metrics is defined as follows:

$$\begin{aligned} R^2 &= 1 - \frac{SSR}{SST} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \\ MSE &= \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2 \\ MAPE &= \frac{1}{n}\sum_{i=1}^{n}\frac{|(y_i - \hat{y}_i)|}{y_i} \end{aligned} \tag{10}$$

where *SSR* and *SST* refer to the sum of squares of residuals and the total sum of squares, respectively. The total number of data points is denoted by n, while each data point is denoted by i. The variables y_i, $\hat{y}_i$ and $\bar{y}$ are the actual output value, predicted output value (from the surrogate model), and the mean of the actual output values, respectively.

Further, after the optimization of the surrogate models, the optimal values of the degrees of freedom are inserted back into the rigorous simulation in Aspen Plus®, and the true value of the optimized OF is obtained from the rigorous simulation, denoted by F. This value is then compared with the OF value obtained from the optimization of the surrogate model, denoted by f, to quantify the deviation (in %) between the surrogate model and the simulation in Aspen Plus®:

$$Deviation\ [\%] = \frac{|F - f|}{F} * 100 \tag{11}$$

In the final step, the results obtained from the surrogate modeling and optimization using the various methods described above are compared.

3.4. Software implementation

We employ several software packages to construct and optimize the surrogate models. All the calculations are performed on an Intel® Core i7 10700 CPU @ 2.90 GHz computer. The process flowsheets are constructed in Aspen Plus® v11, generating the dataset *via* LHS in pyDOE v0.3.8. The final dataset consists of 500 converged data points for the training set and 125 converged data points for the test set.

For the approach developed in this work (the HY model), the surrogate models are generated using Bayesian symbolic regression through the BMS (with 6,000 MCMC steps) run in Python v3.7. The HY model is then optimized using BARON v21.1.13 in GAMS v35.2.0. As shown in the flowchart in Fig. 3, we compare our approach with five

alternative methods, all of which are based on BB models. For these alternative methods, we use different surrogate modeling techniques to construct (and subsequently optimize) the BB models, *i.e.*, Bayesian symbolic regression, kriging, GP, BO, and GA. The first alternative method consists of using Bayesian symbolic regression through the BMS (also with 6,000 MCMC steps; run in Python v3.7) to construct the BB model. The second alternative method uses kriging to generate the BB model. The DACE kriging toolbox v2.0 (Lophaven *et al.*, 2002) in MatLab® vR2021b (The MathWorks Inc., 2021a) is used for building the kriging model (with the regression model being a zero-order polynomial, and the correlation model being a Gaussian correlation). These BB models of the first two alternative methods are each optimized using BARON v21.1.13 in GAMS v35.2.0. The third alternative method consists of using a GP to generate the BB model. MeLON v0.0.8 (Schweidtmann *et al.*, 2020) is used to construct the GP (using the package GPyTorch, with the Matern kernel with smoothness parameter 5/2, constant mean function, 250 training iterations with the Adam optimizer, and a learning rate of 0.1). This model is subsequently optimized using MAiNGO v0.5.0 (Bongartz *et al.*, 2018). The fourth alternative method consists of using BO through the package GPyOpt v1.2.6 (The GPyopt authors, 2016). GPyOpt internally uses the package GPy v1.10.0 (GPy, 2012) to create a GP surrogate model to be used in the BO. Thereafter, Python v3.7 is connected to Aspen Plus® v11 using the COM interface to query the simulation directly at each iteration of the optimization. The acquisition function used is the expected improvement (EI). The fifth alternative method is to use a derivative-free optimization algorithm without a surrogate model, for which we use a GA implemented in the Global Optimization Toolbox in MatLab® vR2021b (The MathWorks Inc., 2021b). In order to be comparable to the optimization using BARON, the constraint tolerance for the GA was set to $1.00{\cdot}10^{-5}$, while the function tolerance was set to $1.00{\cdot}10^{-9}$. To avoid convergence issues in the derivative-free optimization techniques, we penalize non-converged iterations in the OF with a very high value, as discussed in further detail in Section 5.

The nomenclature used henceforth for the different surrogate modeling and optimization strategies with the software described above is summarized in Table 1. The model names for the deterministic optimization approaches have been chosen based on the software used to construct the surrogate models, while the Bayesian optimization-based model and the optimization using genetic algorithm are directly named BO and GA, respectively.

An important point to note here is that the same initial point, which corresponds to one of the points in the training set, is used in all the optimization approaches (except for the BO and GA, which do not require an initial point to be provided). Using this same initial point, we also attempted to compare the results of the surrogate modeling and optimization approaches with those obtained using the built-in optimizer in Aspen Plus®. However, the optimization algorithm in Aspen Plus® was unable to converge the flowsheets to a feasible point for the MINLP problems considered in this work, and thus these values have not been reported.

4. Case studies

We consider two case studies, *i.e.*, propylene glycol and green methanol production. The first case study is a simplified version of the industrial production process of propylene glycol from propylene oxide and water, used as a proof-of-concept. The second case study focuses on green methanol production from electrolytic hydrogen and carbon dioxide obtained from direct air capture (DAC). For simplicity, the unitary production cost (in $\$\cdot kg^{-1}$ of product), used as the OF for the optimization in the case studies, accounts only for the *OPEX* computed for 2018, as it often dominates the *CAPEX*. A further analysis of the *CAPEX* for the optimum points found for each of the models is provided in section S3 of the supplementary material. Thus, the OF (to be minimized) is defined as follows:

$$OF = C_{Feed} + C_{CU} + C_{HU} + C_{Electricity} + C_{Wastewater} \tag{12}$$

where C_{Feed} refers to the costs associated with the raw materials, C_{CU} and C_{HU} refer to the cooling and heating utilities' costs respectively, $C_{Electricity}$ refers to the cost of electricity, and $C_{Wastewater}$ refers to the cost of wastewater treatment, all these by kilogram of product. The cost parameters used for calculating the OF in both case studies are described in Table S1 and Table S2 in the supplementary material (Ghanta *et al.*, 2013; Keith *et al.*, 2018; Parkinson *et al.*, 2019; Turton *et al.*, 2018).

In both case studies, the optimization is formulated as an MINLP problem, as described in Section 3. The termination criteria for the optimization of all models are a maximum CPU time of four hours (*i.e.*, 14,400.00 seconds), or a relative optimality gap of 10^{-9} (this second criterion does not apply to the BO and GA), whichever is attained first. Reaching an optimality gap of 10^{-9} guarantees that the surrogate model (either BB or HY) is solved to global optimality within this tolerance, which is not accomplished in some cases. Hence, additionally, we select exceeding a maximum CPU time of four hours as a secondary termination criterion. As will be discussed later, the BMS approaches spend more time on model building, while the surrogate optimization is usually faster. Specifically, in these case studies, considering parallelization of the BMS calculations to generate analytical equations for each output of the BMS HY model, the training time was up to about six hours for the whole BMS HY model. The training time to generate an analytical equation for the OF in the BMS BB model was between two to three hours. On the other hand, the DACE BB, MeLON BB, and BO models completed their training in a few seconds. The question we want to address is whether the BMS models, owing to the more tractable, closed-form equations generated, lead to savings in the optimization that can offset their larger training times.

Additionally, an important point to note here is that the analytical equation provided by the BMS might not include all the degrees of freedom of the whole flowsheet (elements of the subsets N^C, N^U, and N^Z defined in Section 3.2). This can be the case in the BMS HY model, where some of the degrees of freedom of the whole flowsheet may not be present in the analytical equations of the outputs of the relevant process units. For example, in the flowsheet shown in Fig. 4, and the BMS equations in Equation (5), the reactor volume V might not be present in the final analytical equation of the conversion X. It could also be the case in the BMS BB model, where the OF modeled using the BMS might not include some of the degrees of freedom of the whole flowsheet. Taking the same example as before, the BMS equation (to get output f^{BB}) of the BB model in Equation (8) might not include the reactor volume V. In such cases, the values of such degrees of freedom not present in the BMS equation (*i.e.*, the degrees of freedom of the whole flowsheet not present in the final analytical equation/s in the BMS HY model and/or BMS BB model, and thus not optimized subsequently) are taken from the initial point provided for the optimization, which as mentioned previously in Section 3.4, is one of the points in the training set. A sensitivity analysis of these degrees of freedom omitted in the BMS models has been added in section S8 of the supplementary material.

Table 1
Nomenclature used for different surrogate modeling and optimization strategies.

Model type	Model name	Surrogate model	Optimization
HY	BMS HY	BMS	BARON
BB	BMS BB	BMS	BARON
BB	DACE BB	DACE kriging toolbox	BARON
BB	MeLON BB	MeLON	MAiNGO
BB	BO	GPy	GPyOpt
BB	GA	-	MatLab

4.1. Case study 1 – propylene glycol production

Propylene glycol ($C_3H_8O_2$) is commonly used as an intermediate for various chemicals and also in food, cosmetics, and pharmaceuticals. Commercially, it is produced by the hydrolysis of propylene oxide (C_3H_6O) using an excess of water (H_2O), and without a catalyst (Trent, 2001). This reaction also results in the production of dipropylene glycol, tripropylene glycol, and higher polyglycols, and an excess of H_2O is used to maximize the commercial production of $C_3H_8O_2$. This work considers the simplifying assumption that only $C_3H_8O_2$ is produced in reaction RE2 shown below.

$$C_3H_6O + H_2O \rightarrow C_3H_8O_2 \quad \text{(RE2)}$$

As the flowsheet depicted in Fig. 5a shows, a saturated liquid fresh feed of C_3H_6O and H_2O at 25 °C and 1 bar pressure enters the CSTR, operating isothermally at 1 bar pressure, in which the hydrolysis reaction takes place. As mentioned earlier, an excess of H_2O is used in the commercial process to ensure higher conversion to $C_3H_8O_2$. We thus enforce that the molar ratio of H_2O to C_3H_6O in the fresh feed ($F^{S1}_{H_2O}/F^{S1}_{C_3H_6O}$) falls in the range 2 to 5 to ensure an excess of H_2O in the inlet to the CSTR. Subsequently, after the reaction in the CSTR, the stream S3 enters the distillation column D, in which a design specification ensures a 99.5% molar purity of $C_3H_8O_2$ (by varying the distillate-to-feed ratio) in S5, while the split fraction of the purge from the splitter (S7) is fixed to 0.1. The molar flow rates of C_3H_6O and H_2O ($F^{S1}_{C_3H_6O}$ and $F^{S1}_{H_2O}$, respectively), the CSTR temperature and volume (T and V, respectively), and the number of stages and reflux ratio of the distillation column (N and

Table 2
Degrees of freedom and their associated ranges for case study 1.

Degrees of freedom	Symbol	Unit	Lower bound	Upper bound
Flow rate of C_3H_6O	$F^{S1}_{C_3H_6O}$	kmol·hr^{-1}	100.00	500.00
Flow rate of H_2O	$F^{S1}_{H_2O}$	kmol·hr^{-1}	200.00	2,500.00
CSTR temperature	T	°C	40.00	80.00
CSTR volume	V	m^3	6.00	10.00
Reflux ratio	RR	-	0.50	10.00
No. of stages	N	-	5.00	15.00

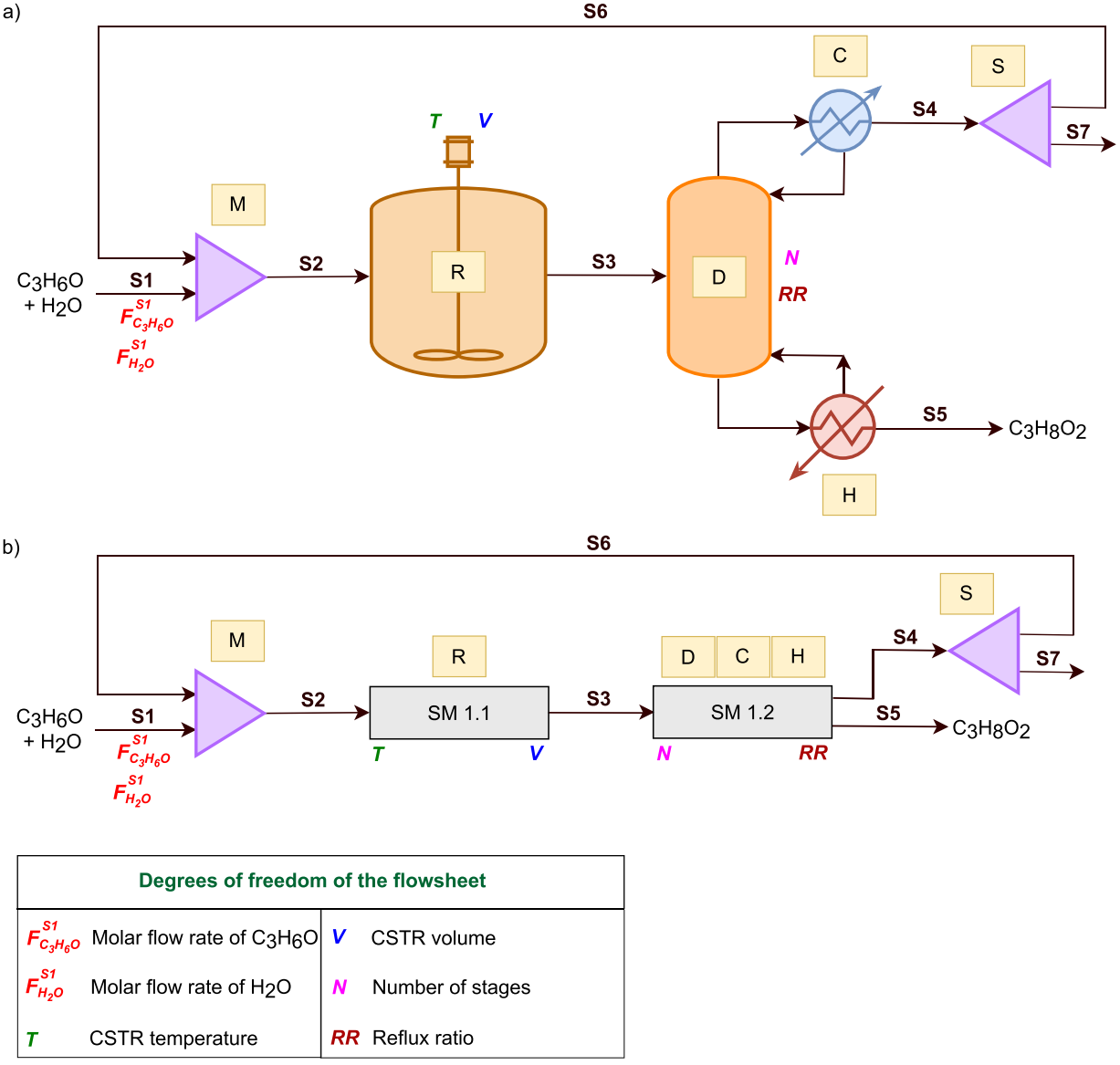

Fig. 5. a) Process flowsheet for case study 1, in which the mixer (M), CSTR (R), distillation column (D), condenser (C), reboiler (H), and splitter (S) are shown. b) Process flowsheet for case study 1 denoting the process units replaced by surrogate models SM 1.1 and SM 1.2.

RR, respectively) are the six degrees of freedom. The degrees of freedom, along with their associated ranges, are shown in Table 2. The thermodynamic package NRTL is used in the flowsheet. The utilities consumed are cooling water (20 °C to 30 °C) and high-pressure steam (250 °C, 40 bar). The OF quantifies the total cost expressed as $\$\cdot kg^{-1}$ of $C_3H_8O_2$ produced and considers C_{Feed} (C_3H_6O and H_2O), C_{CU} (cooling water), and C_{HU} (high-pressure steam).

For the BMS HY model, the process units replaced by surrogate models are labeled as SM 1.1 and SM 1.2 in Fig. 5b, while the mixer and splitter are modeled using mass balances for each of them alongside a simplified energy balance described later. The inputs and outputs of each BMS equation of SM 1.1 and SM 1.2 of the BMS HY model are shown in Table 3. The set *A* in Table 3 contains all three components of the flowsheet, *i.e.*, C_3H_6O, H_2O, and $C_3H_8O_2$. F^{S2} denotes the total molar flow rate of stream S2, while the mole fraction of H_2O in stream S2 is denoted by $x^{S2}_{H_2O}$.

As SM 1.1 requires the temperature of stream S2 (T_{S2}) to be known, it has been calculated in the BMS HY model using a simplified energy balance, in which the specific heat capacities at constant pressure (C_p) of streams S1, S2, and S6 have been assumed to be equal (for the training set, this assumption results in a MAPE of $1.14{\cdot}10^{-2}$ between the actual and predicted values of the temperature of stream S2). As the mole fraction of $C_3H_8O_2$ in stream S2 is less than 10^{-4}, it is assumed to be negligible. Therefore, the inputs to model SM 1.1 are only the total molar flow rate and the mole fraction of H_2O.

4.2. Case study 2 – green methanol production

As methanol (CH_3OH) production currently relies on natural gas (for syngas production), it is necessary to look for alternative routes to reduce its dependence on fossil carbon. Accordingly, pathways based on carbon dioxide (CO_2) hydrogenation have been proposed. The flowsheet shown in Fig. 6a is based on the works by Van-Dal and Bouallou (2013) and Vázquez and Guillén-Gosálbez (2021). It considers CO_2 obtained from DAC powered by natural gas and electricity from the current mix, while hydrogen (H_2) is obtained from wind-powered water splitting.

We assume that there is a DAC facility integrated with the methanol production plant. The CO_2 feed, which enters at 25 °C and 1 bar at a constant flow rate of 2000 $kmol{\cdot}hr^{-1}$, is compressed with the help of a compression train with intermediate cooling to reach the desired pressure at the reactor inlet. This reactor pressure (*P*) is one of the degrees of freedom. The compression train is required to keep the compression ratio below three in each compressor, while intermediate cooling is required because of the high temperatures resulting from compression. The H_2 feed is assumed to be available at 30 bar, with its molar flow rate ($F^{S9}_{H_2}$) being a degree of freedom. The H_2 feed needs to be compressed to reach the desired pressure. The CO_2 and H_2 are then fed into a process heater along with the recycle stream S19, where the outlet temperature (*T*) is modeled as a degree of freedom. This heated stream (S12) then enters the plug flow reactor (PFR) operating adiabatically and containing a Cu-ZnO-Al_2O_3 catalyst, in which the CO hydrogenation reaction RE3 and water-gas shift reaction RE4 take place. The volume of the PFR (*V*) is another degree of freedom.

$$CO + 2H_2 \leftrightarrows CH_3OH \quad \text{(RE3)}$$

$$CO_2 + H_2 \leftrightarrows CO + H_2O \quad \text{(RE4)}$$

Reactions RE3 and RE4 lead to the global reaction RE5:

$$CO_2 + 3H_2 \leftrightarrows CH_3OH + H_2O \quad \text{(RE5)}$$

The kinetic model developed by Van-Dal and Bouallou (2013) has been used in this work. The outlet stream from the reactor (S13) is cooled down in two successive coolers, the first one is enforced to achieve a vapor fraction of one, and the second one to cool the stream to 35 °C. This cooled stream (S15) is then sent to the first flash unit, where most of the unreacted CO and H_2 are recycled back to the reactor (with intermediate compression). A small part of the recycle stream is purged, with the split fraction (*SF*) of the purge from the splitter (S17) being another degree of freedom. The stream coming out at the bottom of the first flash unit (S20) is expanded to 2 bar, before being sent to the second flash unit, and then to a distillation column. The bottom of the distillation column yields a wastewater stream S26 (containing about 0.1 mole% CH_3OH) that is treated in a separate facility. The distillate stream S27, containing mainly CO_2 and CH_3OH (with a 99.9% molar recovery of CH_3OH, obtained by varying the distillate-to-feed ratio of the column) is then sent to the third flash unit to obtain a 99.5% (by mole) pure stream of CH_3OH (S30) at the bottom. This purity is attained by varying the operating temperature of flash F3. The reflux ratio and the number of stages of the distillation column (*RR* and *N*, respectively) are the remaining two degrees of freedom, thus resulting in a total of seven degrees of freedom for the flowsheet. These degrees of freedom, along with their associated ranges, are shown in Table 4. There are no explicit constraints associated with the degrees of freedom in this case study.

The thermodynamic packages Peng-Robinson (for streams at pressures greater than 2 bar) and NRTL (for streams at pressures less than 2 bar) have been used in the simulation. A pressure drop of 0.1 bar is considered in the distillation column, 0.2 bar in the process heaters and coolers (except H1), and 5 bar in the PFR. Although an explicit pressure drop is not considered in H1, it is assumed to be included in the pressure drop defined for the PFR. The utilities used are electricity from the grid, cooling water (20 °C to 30 °C), and high-pressure steam (250 °C, 40 bar). Heat integration considers the following streams: the process heaters (H1 and H3), process coolers (C1 to C5), the condenser (C6) and reboiler (H4) of the distillation column D, and the flash units F2 and F3 (for which we define a heater (H2) and cooler (C7) before flash units F2 and F3, respectively). The minimum temperature difference considered is 10 °C. We use pinch analysis to calculate the minimum heating and cooling utilities' targets for each point in the training set in the BB models. In the HY model, we carry out the optimization of process variables and heat integration in GAMS simultaneously, using the model developed by Duran and Grossmann (1986).

The OF quantifies the total cost expressed in $\$\cdot kg^{-1}$ of CH_3OH produced and considers the terms C_{Feed} (CO_2 and H_2), C_{CU} (cooling water), C_{HU} (high-pressure steam), $C_{Electricity}$ (electricity from the grid), and $C_{Wastewater}$.

For the BMS HY model, the process units replaced by surrogate models are labeled from SM 2.1 to SM 2.12 in Fig. 6b, while the splitter S is modeled using its mass balance. Note that the first three compressors (K1 to K3) and the first three coolers (C1 to C3) are not modeled, as the molar flow rate of CO_2 is not a degree of freedom and is instead kept constant at 2000 $kmol{\cdot}hr^{-1}$. Thus, the variables of these process units (such as the electricity requirement of compressors K1 to K3, heat duties of the coolers C1 to C3, etc.) are constants in this case study and are defined as such in the calculations.

The inputs and outputs of each BMS equation of SM 2.1 to SM 2.12 of the BMS HY model are shown in Table 5. The set *B* in Table 5 contains all

Table 3
Inputs and outputs associated with each BMS equation of SM 1.1 and SM 1.2 in the BMS HY model for case study 1.

Surrogate Model	Inputs	BMS equation	Output	Unit
SM 1.1	$\boldsymbol{V}$, $\boldsymbol{T}$, $\boldsymbol{F^{S2}}$, $\boldsymbol{T_{S2}}$, $\boldsymbol{x^{S2}_{H_2O}}$	SM 1.1.1	Fractional conversion of C_3H_6O in R	-
		SM 1.1.2	Heat duty of R	$GJ{\cdot}hr^{-1}$
SM 1.2	$\boldsymbol{T}$, $\boldsymbol{RR}$, $\boldsymbol{N}$, $\boldsymbol{F^{S3}_a}\ \forall\ \boldsymbol{a} \in \boldsymbol{A}$	SM 1.2.1	Temperature of S4	°C
		SM 1.2.2	Heat duty of C	$GJ{\cdot}hr^{-1}$
		SM 1.2.3	Heat duty of H	$GJ{\cdot}hr^{-1}$
		SM 1.2.4	Molar flow rate of $C_3H_8O_2$ in S5	$kmol{\cdot}hr^{-1}$

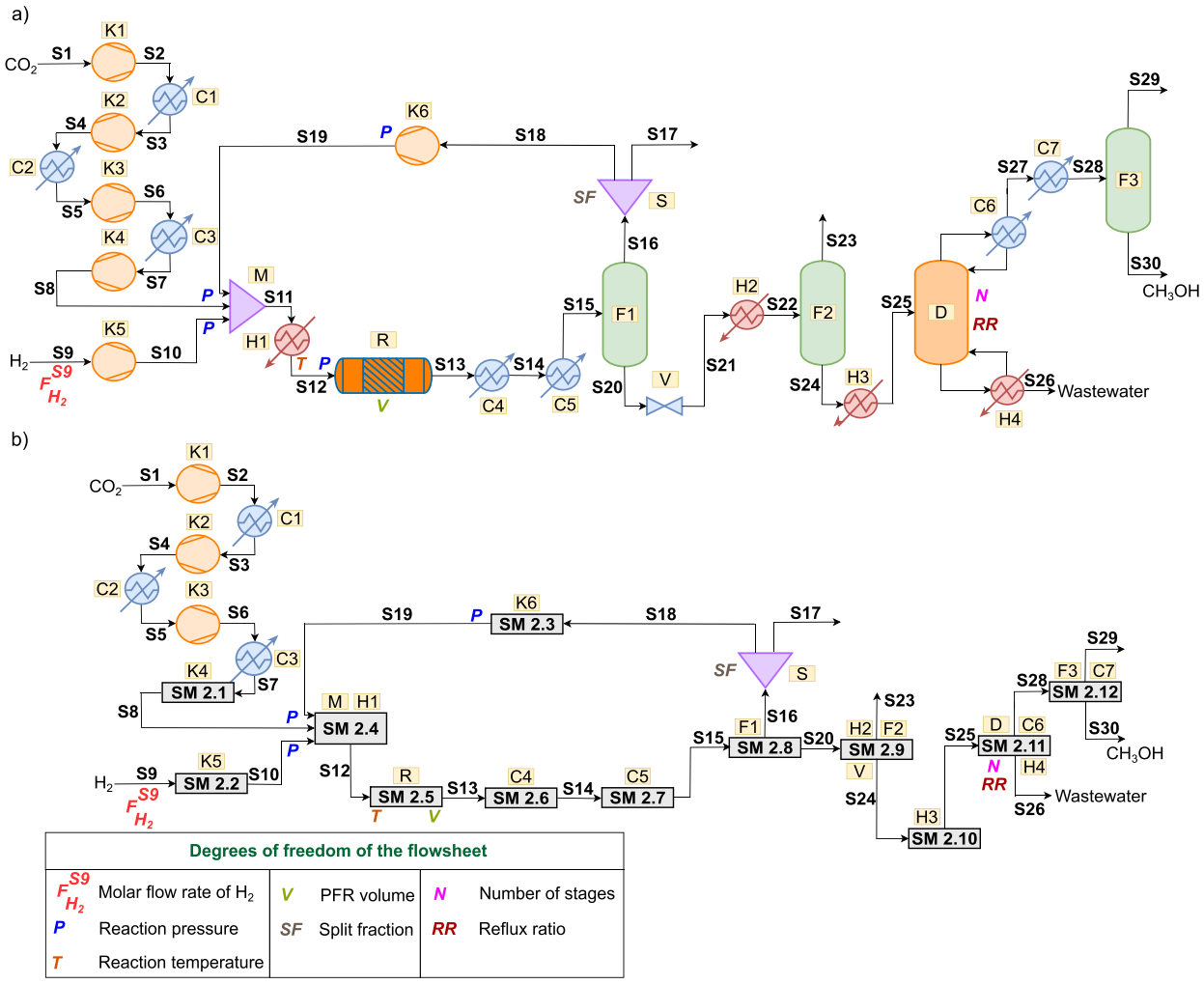

Fig. 6. a) Process flowsheet for case study 2, in which the compressors (K), mixer (M), coolers (C), heaters (H), PFR (R), splitter (S), flash separators (F), distillation column (D), and valve (V) are shown. b) Process flowsheet for case study 2 denoting the process units replaced by surrogate models SM 2.1 to SM 2.12.

Table 4
Degrees of freedom and their associated ranges for case study 2.

Degrees of freedom	Symbol	Unit	Lower bound	Upper bound
Molar flow rate of H_2	$F_{H_2}^{S9}$	kmol·hr^{-1}	4,500.00	6,000.00
PFR temperature	T	°C	180.00	240.00
PFR pressure	P	bar	45.00	55.00
PFR volume	V	m^3	35.00	54.00
Split fraction	SF	-	$1.00 \cdot 10^{-3}$	$5.00 \cdot 10^{-2}$
Reflux ratio	RR	-	1.25	1.80
No. of stages	N	-	45.00	55.00

five components of the flowsheet, *i.e.*, CO, CO_2, H_2, CH_3OH, and H_2O.

As heat integration requires the temperature of stream S11 to be known, it has been predicted in model SM 2.4. We did not use correlations to calculate the enthalpies of the individual process streams to predict this temperature, as we want to avoid introducing an additional source of error due to the correlations. Additionally, the molar flow rates of only some of the components have been used as inputs in models SM 2.9 to SM 2.12. This is because components that have a mole fraction of less than 10^{-4} in a stream are neglected, to simplify the calculations in the BMS HY model.

5. Results and discussion

5.1. Case study 1 – propylene glycol production

5.1.1. Results of the surrogate modeling

The sampling time taken for generating the required outputs from the flowsheet for the training dataset is 842.00 seconds.

In the BB models, the training dataset consisting of 500 × 6 inputs is sent to the BMS, DACE kriging toolbox, MeLON, and GPyOpt to generate the BMS BB, DACE BB, MeLON BB, and BO models, respectively. Fig. 7 shows the quality of the BMS BB, DACE BB, and MeLON BB models for the test set. All three surrogate models approximate the OF accurately, with the minimum R^2 value being 99.63% for the BMS BB model. Similarly, the highest MSE and MAPE values are observed for the BMS BB model. Even so, the model obtained shows an acceptable level of accuracy in the data fitting (defining the acceptable level of accuracy as $R^2 > 99\%$).

Concerning the HY model, the quality of the fitting for the test set is shown in Figures S1 and S2 in the supplementary material. The R^2 values range from 99.15% to 100.00%, indicating that the BMS HY model performs very well. The individual training times for each BMS equation of the BMS HY model are reported in Table 6. The process units mentioned in Table 6 refer to the process flowsheet shown in Fig. 5.

Table 5
Inputs and outputs associated with each BMS equation of SM 2.1 to SM 2.12 in the BMS HY model for case study 2.

Surrogate Model	Inputs	BMS equation	Output	Unit
SM 2.1	P	SM 2.1.1	Power required for K4	$GJ{\cdot}hr^{-1}$
SM 2.2	$F_{H_2}^{S9}$, P	SM 2.2.1	Power required for K5	$GJ{\cdot}hr^{-1}$
SM 2.3	P, $F_b^{S18}\ \forall\ b \in B$	SM 2.3.1	Power required for K6	$GJ{\cdot}hr^{-1}$
		SM 2.3.2	Temperature of S19	°C
SM 2.4	$F_{H_2}^{S9}$, P, T_{S19}, $F_b^{S19}\ \forall\ b \in B$	SM 2.4.1	Temperature of S11	°C
		SM 2.4.2	Heat duty of H1	$GJ{\cdot}hr^{-1}$
SM 2.5	P, T, V, $F_b^{S12}\ \forall\ b \in B$	SM 2.5.1	CH_3OH produced in R	$kmol{\cdot}hr^{-1}$
		SM 2.5.2	CO produced in R	$kmol{\cdot}hr^{-1}$
		SM 2.5.3	Temperature of S13	°C
SM 2.6	T_{S13}, P_{C4}, $F_b^{S13}\ \forall\ b \in B$	SM 2.6.1	Heat duty of C4	$GJ{\cdot}hr^{-1}$
		SM 2.6.2	Temperature of S14	°C
SM 2.7	T_{S14}, P_{C5}, $F_b^{S14}\ \forall\ b \in B$	SM 2.7.1	Heat duty of C5	$GJ{\cdot}hr^{-1}$
SM 2.8	P_{F1}, $F_b^{S15}\ \forall\ b \in B$	SM 2.8.1	Split fraction of CH_3OH in F1	-
		SM 2.8.2	Split fraction of CO_2 in F1	-
		SM 2.8.3	Split fraction of H_2O in F1	-
		SM 2.8.4	Heat duty of F1	$GJ{\cdot}hr^{-1}$
SM 2.9	$F_b^{S20}\ \forall\ b \in \{CO_2, CH_3OH, H_2O\}$	SM 2.9.1	Split fraction of CO_2 in F2	-
		SM 2.9.2	Heat duty of F2	$GJ{\cdot}hr^{-1}$
		SM 2.9.3	Temperature of S21	°C
SM 2.10	$F_b^{S24}\ \forall\ b \in \{CO_2, CH_3OH, H_2O\}$	SM 2.10.1	Heat duty of H3	$GJ{\cdot}hr^{-1}$
SM 2.11	N, RR, $F_b^{S25}\ \forall\ b \in \{CO_2, CH_3OH, H_2O\}$	SM 2.11.1	Heat duty of C6	$GJ{\cdot}hr^{-1}$
		SM 2.11.2	Heat duty of H4	$GJ{\cdot}hr^{-1}$
		SM 2.11.3	Temperature at inlet of C6	°C
		SM 2.11.4	Temperature of S27	°C
		SM 2.11.5	Temperature at inlet of H4	°C
		SM 2.11.6	Temperature of S26	°C
SM 2.12	T_{S28}, $F_b^{S28}\ \forall\ b \in \{CO_2, CH_3OH\}$	SM 2.12.1	Split fraction of CH_3OH in F3	-
		SM 2.12.2	Heat duty of F3	$GJ{\cdot}hr^{-1}$
		SM 2.12.3	Temperature of S28	°C

5.1.2. Results of the optimization

The sizes of the deterministic optimization models (*i.e.*, BMS HY, BMS BB, DACE BB, and MeLON BB models) are summarized in Table 7.

As seen in Table 7, the BMS HY model has more equations and variables than any other method. This is due to the fact that the surrogate models are defined for several process units and also because we combine them with mass and energy balance equations. All the BB models have only three equations, one for the objective function directly modeled with a surrogate model, and the other two for the constraints defined previously. Further, the BMS BB model does not contain any integer variable, as it is not present in the final analytical equation of the surrogate model.

The results of the optimization (*i.e.*, minimization of the unitary production cost in $\${\cdot}kg^{-1}$) are summarized in Table 8, *i.e.*, training time, optimization time to find the best solution (*i.e.*, the solution showing the minimum OF value among all the solutions explored within the maximum CPU time defined as the termination criterion), total optimization time (all in seconds), and OF values obtained from the surrogate model (f) and rigorous simulation in Aspen Plus® (F), both in $\${\cdot}kg^{-1}$, along with their associated deviation (in percentage).

Note that the training time reported for the BMS HY model in Table 8 corresponds to the maximum of the individual training times of the six surrogate models it contains, as we assume that the BMS calculations could be parallelized.

As seen in Table 8, the training of the BMS BB model is faster than in the BMS HY model, yet both are much slower in the training phase than the other approaches. Specifically, the training of the DACE BB, MeLON BB, and BO models is very fast (*i.e.*, 1.12 seconds, 11.14 seconds, and 1.00 second, respectively). As there is no surrogate model used by the GA, no training time has been reported. Considering the optimization performance, the BMS HY model can be globally optimized with a relative optimality gap of 10^{-9} in just 2.14 seconds, and it finds the best solution in only 0.25 seconds. However, the BMS BB, DACE BB, and MeLON BB models are not able to close the same optimality gap within the maximum CPU time specified as the termination criterion (*i.e.*, 14,400.00 seconds). Moreover, all of them find their respective best solutions in much lesser time, being found during the pre-processing for the BMS BB model, and only in 14.84 seconds for the MeLON BB model (also during the pre-processing). For the DACE BB model, the best solution is found in 5,409.98 seconds, and for the BO model in 8,556.40 seconds, although the maximum time is set to 14,400.00 seconds in all cases. For the GA model, the optimization time to find the best solution is 12,878.59 seconds, even though the maximum time is also set to 14,400.00 seconds.

Further, comparing the values of F for each model, it is observed that the BMS HY model is the best in terms of the minimum value found (*i.e.*, 2.42 $\${\cdot}kg^{-1}$). It is followed by the GA approach, which also finds a value of 2.42 $\${\cdot}kg^{-1}$ (after rounding off to three digits, the values are the same, but for five significant digits, the OF values from the BMS HY and GA models are 2.4205 $\${\cdot}kg^{-1}$ and 2.4211 $\${\cdot}kg^{-1}$, respectively). Further, GA was able to complete 3,963 evaluations of the simulation in 14,400.00 seconds. Thus, if the training time of the BMS HY model could be improved, it could potentially find a better solution than GA in much lesser optimization time. The BMS HY and GA models are followed by the BMS BB and DACE BB models, which find the same value of 2.45 $\${\cdot}kg^{-1}$ (rounding off the result to three significant digits results in the same value, but the DACE BB model actually finds a 0.12% higher value than the BMS BB model). The MeLON BB model returns a slightly higher value of 2.46 $\${\cdot}kg^{-1}$, while the BO model performs the worst with a value of 2.49 $\${\cdot}kg^{-1}$ using a penalty parameter value of 75. The results for other penalty parameter values have been reported in section S4 of the supplementary material.

In terms of the percentage deviation between the f and F values, the BMS HY model presents the minimum deviation, 0.23%. The BMS BB and MeLON BB models show deviations of 4.42% and 4.68% respectively, while the DACE BB model has a higher deviation of 8.16%. There is no deviation reported for the BO and GA models (NA, *i.e.*, not applicable) as they sample directly from the flowsheet.

The optimized values of the six degrees of freedom for each model are shown in Table 9. As we can see, for the best-performing model (*i.e.*, BMS HY), the optimized values of T, V, RR, and N are at the bounds of their respective ranges, while for the second best-performing approach,

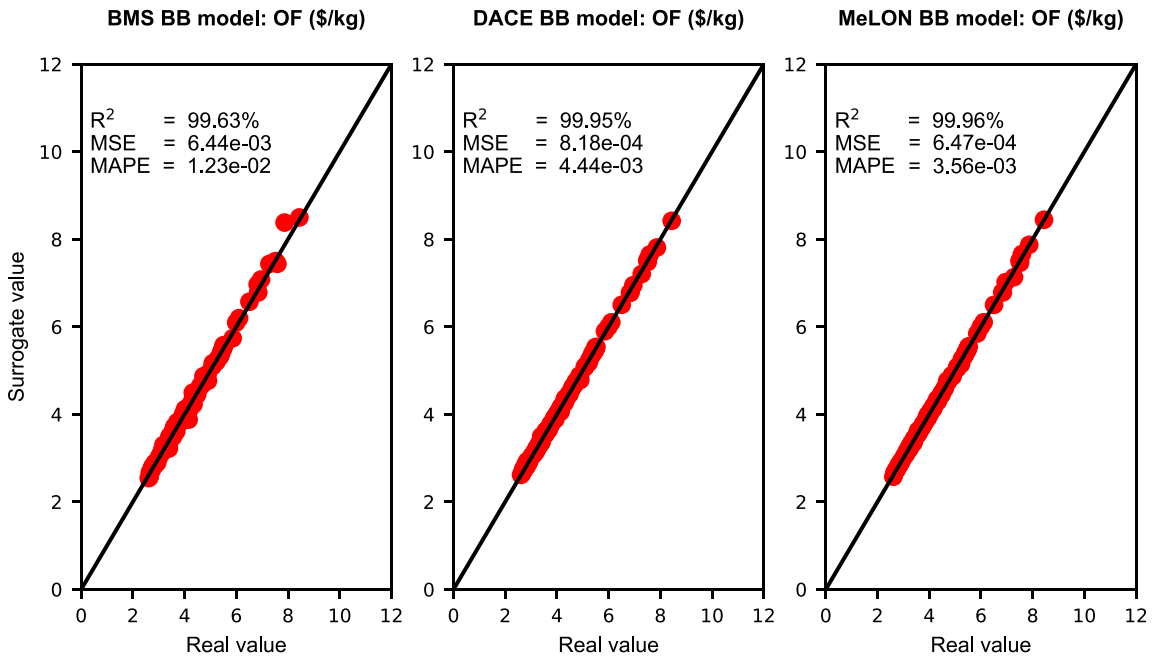

Fig. 7. Comparison between the real (rigorous simulation) and surrogate values of the OF for the test data of case study 1.

Table 6
Training time [s] for each BMS equation of the BMS HY model for case study 1.

Surrogate Model	BMS equation	Output	Unit	Training time [s]
SM 1.1	SM 1.1.1	Fractional conversion of C_3H_6O in R	-	12,030.00
	SM 1.1.2	Heat duty of R	$GJ{\cdot}hr^{-1}$	10,590.00
SM 1.2	SM 1.2.1	Temperature of S4	°C	1,760.00
	SM 1.2.2	Heat duty of C	$GJ{\cdot}hr^{-1}$	9,777.00
	SM 1.2.3	Heat duty of H	$GJ{\cdot}hr^{-1}$	11,407.00
	SM 1.2.4	Molar flow rate of $C_3H_8O_2$ in S5	$kmol{\cdot}hr^{-1}$	5,327.00

Table 7
Size of the mathematical programming models for case study 1.

Model	Equations	Continuous variables	Integer variables
BMS HY	27	31	1
BMS BB	3	6	0
DACE BB	3	7	1
MeLON BB	3	7	1

i.e., GA model, the feed flow rates are also at their lower bounds. Further, the imposed constraint on the ratio ($F^{S1}_{H_2O}/F^{S1}_{C_3H_6O}$) is at its lower bound of 2 for all six models, being only marginally higher for the BO model. This is in line with expectations, as the optimization tries to minimize $F^{S1}_{C_3H_6O}$, which shows the largest contribution to the OF. Further, the DACE BB model reports optimal values of the feed flow rates which are much higher compared to those from the other approaches. Although this would result in a larger production of $C_3H_8O_2$, it also results in larger heat duties in the CSTR, condenser, and reboiler, thus increasing the utility costs as a trade-off. This results in the OF value for this approach being higher than that for the BMS HY and GA models (the two best-performing approaches). Considering the CSTR temperature and volume, higher values for both these degrees of freedom enable greater conversion of C_3H_6O to $C_3H_8O_2$, while a lower value of the reflux ratio in the column reduces the heat duties of the condenser and reboiler. Similarly, a higher value of the number of stages enables a more effective separation in the column. Thus, the optimal values of the reflux ratio and the number of stages found for the best-performing approach (BMS HY model) are at their lower bound and upper bound, respectively. Additionally, the number of stages is not present in the final BMS equation obtained for the BMS BB model. Therefore, as mentioned in section 4, the number of stages is fixed at its value in the initial point used in the optimization of the BMS BB model. A sensitivity analysis of the influence of the number of stages on the BMS BB model has been shown in section S8 of the supplementary material.

5.2. Case study 2 – green methanol production

5.2.1. Results of the surrogate modeling

The sampling time taken for generating the required outputs from the flowsheet for the training dataset is 2,440.00 seconds.

In the BB models, the training dataset consisting of 500 × 7 inputs is sent to the BMS, DACE kriging toolbox, MeLON, and GPyOpt to generate the BMS BB, DACE BB, MeLON BB, and BO models, respectively. Fig. 8 shows the quality of the fitting generated by the BMS BB, DACE BB, and MeLON BB models for the test set. It is seen that all three surrogate

Table 8
Results of the minimization problem for case study 1.

Model	Training time [s]	Optimization time for best solution [s]	Total optimization time [s]	f [$\cdot kg^{-1}$]	F [$\cdot kg^{-1}$]	Deviation [%]
BMS HY	12,030.00	0.25	2.14	2.43	2.42	0.23%
BMS BB	7,527.00	Pre-processing	14,400.00	2.34	2.45	4.42%
DACE BB	1.12	5,409.98	14,400.00	2.25	2.45	8.16%
MeLON BB	11.14	14.84	14,400.00	2.35	2.46	4.68%
BO	1.00	8,556.40	14,400.00	NA	2.49	-
GA	NA	12,878.59	14,400.00	NA	2.42	-

Table 9
Optimized values of degrees of freedom for case study 1.

Degrees of freedom	Symbol	Unit	Model					
			BMS HY	BMS BB	DACE BB	MeLON BB	BO	GA
Flow rate of C_3H_6O	$F^{S1}_{C_3H_6O}$	kmol·hr^{-1}	157.59	100.00	409.75	100.00	174.89	100.00
Flow rate of H_2O	$F^{S1}_{H_2O}$	kmol·hr^{-1}	315.17	200.00	819.90	200.00	370.86	200.00
CSTR temperature	T	°C	80.00	57.07	80.00	59.64	61.64	80.00
CSTR volume	V	m^3	10.00	10.00	6.00	6.00	8.36	6.00
Reflux ratio	RR	-	0.50	0.50	0.50	0.50	0.50	0.50
No. of stages	N	-	15.00	9.00	8.00	13.00	15.00	15.00

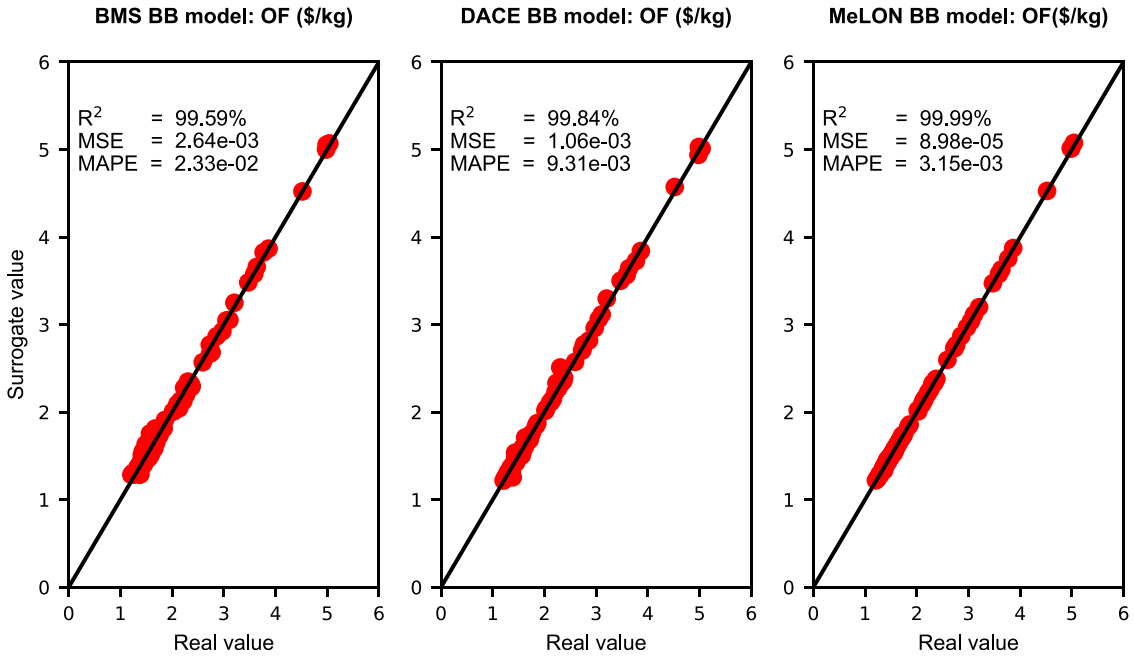

Fig. 8. Comparison between the real (rigorous simulation) and surrogate values of the OF for the test data of case study 2.

models perform very well in modeling the OF, with the minimum R^2 value being 99.59% for the BMS BB model. Similarly, the highest MSE and MAPE values are found in the BMS BB model. Even so, the model obtained shows an acceptable level of accuracy in the data fitting (again defining the acceptable level of accuracy as $R^2 > 99\%$).

In the HY model, the quality of the fitting for the test set is shown graphically in Figures S3 to S12 in the supplementary material. With a minimum R^2 value of 80.98%, a maximum value of 100.00%, and a mean value of 98.74%, it can be concluded that most of the surrogate models of the BMS HY model perform very well in modeling each output. Moreover, only three of the 29 outputs show an R^2 value of less than 98.00%.

The individual training times for each BMS equation of the BMS HY model are reported in Table 10. The process units mentioned in Table 10 refer to the process flowsheet shown in Fig. 6.

5.2.2. Results of the optimization

The sizes of the deterministic optimization models (*i.e.*, BMS HY, BMS BB, DACE BB, and MeLON BB models) are summarized in Table 11.

As seen in Table 11, analogous to case study 1, the BMS HY model has more equations and variables than any other method. This is because the surrogate models are defined for various process outputs, and also due to the presence of mass and energy balance equations in the model. All the BB models have only one equation for the objective function directly modeled with a surrogate model. Further, the BMS BB model does not contain any integer variable, as it is not present in the final analytical equation of the surrogate model.

The results of the optimization (*i.e.*, minimization of the unitary production cost in \$·kg^{-1}) are displayed in Table 12, which shows the training time, optimization time to find the best solution, total optimization time (all in seconds), and the OF values obtained from the surrogate model (f) and rigorous simulation in Aspen Plus® (F), both in \$·kg^{-1}, along with their associated deviation (in percentage).

Note that the training time reported for the BMS HY model in Table 12 is the maximum of the individual training times of the 29 surrogate models, as we again assume the parallelization of the BMS training.

As seen in Table 12, analogous to case study 1, the training of the BMS BB model is completed in less time than with the BMS HY model. Moreover, the training of the DACE BB, MeLON BB, and BO models is again very fast (*i.e.*, 3.62 seconds, 10.96 seconds, and 1.00 second, respectively). Moreover, the BMS HY model completes the optimization with a relative optimality gap of 10^{-9} in just 0.89 seconds. However, while the BMS BB and DACE BB models find the best solution during pre-processing itself, both are unable to close the relative optimality gap of 10^{-9} within the CPU time specified as the termination criterion (*i.e.*, 14,400.00 seconds). The MeLON BB model completes the optimization in 15.66 seconds only and finds the best solution during pre-processing, but the globally optimal values reported by the model (with the default settings of MeLON, as mentioned in Section 3.4) are the same as the initial point provided to the model. For the BO and GA models, the optimization time required to find the best solution is 11,931.27 and 13,821.55 seconds, respectively, but they are allowed to continue the optimization for 14,400.00 seconds (like in all other methods). The

Table 10
Training time [s] for each BMS equation of the BMS HY model for case study 2.

Surrogate Model	BMS equation	Output	Unit	Training time [s]
SM 2.1	SM 2.1.1	Power required for K4	GJ·hr^{-1}	8,184.00
SM 2.2	SM 2.2.1	Power required for K5	GJ·hr^{-1}	8,550.00
SM 2.3	SM 2.3.1	Power required for K6	GJ·hr^{-1}	10,223.00
	SM 2.3.2	Temperature of S19	°C	2,350.00
SM 2.4	SM 2.4.1	Temperature of S11	°C	6,359.00
	SM 2.4.2	Heat duty of H1	GJ·hr^{-1}	7,401.00
SM 2.5	SM 2.5.1	CH_3OH produced in R	kmol·hr^{-1}	5,061.00
	SM 2.5.2	CO produced in R	kmol·hr^{-1}	1,817.00
	SM 2.5.3	Temperature of S13	°C	23,402.00
SM 2.6	SM 2.6.1	Heat duty of C4	GJ·hr^{-1}	12,991.00
SM 2.7	SM 2.6.2	Temperature of S14	°C	11,490.00
	SM 2.7.1	Heat duty of C5	GJ·hr^{-1}	10,434.00
SM 2.8	SM 2.8.1	Split fraction of CH_3OH in F1	-	9,738.00
	SM 2.8.2	Split fraction of CO_2 in F1	-	10,535.00
	SM 2.8.3	Split fraction of H_2O in F1	-	9,390.00
	SM 2.8.4	Heat duty of F1	GJ·hr^{-1}	10,161.00
SM 2.9	SM 2.9.1	Split fraction of CO_2 in F2	-	8,637.00
	SM 2.9.2	Heat duty of F2	GJ·hr^{-1}	21,397.00
	SM 2.9.3	Temperature of S21	°C	17,283.00
SM 2.10	SM 2.10.1	Heat duty of H3	GJ·hr^{-1}	14,814.00
SM 2.11	SM 2.11.1	Heat duty of C6	GJ·hr^{-1}	5,863.00
	SM 2.11.2	Heat duty of H4	GJ·hr^{-1}	14,625.00
	SM 2.11.3	Temperature at inlet of C6	°C	4,383.00
	SM 2.11.4	Temperature of S27	°C	2,447.00
	SM 2.11.5	Temperature at inlet of H4	°C	3,965.00
	SM 2.11.6	Temperature of S26	°C	1,835.00
SM 2.12	SM 2.12.1	Split fraction of CH_3OH in F3	-	3,276.00
	SM 2.12.2	Heat duty of F3	GJ·hr^{-1}	3,541.00
	SM 2.12.3	Temperature of S28	°C	2,988.00

Table 11
Size of the mathematical programming models for case study 2.

Model	Equations	Continuous variables	Integer variables
BMS HY	171	184	1
BMS BB	1	5	0
DACE BB	1	8	1
MeLON BB	1	8	1

value reported for the BO model is found with a penalty parameter value of 20. The results for other penalty parameter values have been reported in section S4 of the supplementary material.

Further, comparing the values of F for each model, it is observed that the GA model is the best in terms of the minimum value found (*i.e.*, 1.20 \$·kg^{-1}). It is closely followed by the BO model and BMS HY model, which find values of 1.21 \$·kg^{-1} and 1.22 \$·kg^{-1}, respectively, *i.e.*, 0.83% and 1.67% higher than that found by the GA model, respectively. Further, the GA was able to complete 2,517 evaluations of the simulation in 14,400.00 seconds. Therefore, analogous to case study 1, if the training time of the BMS HY model was improved, it would have the potential to find a better solution than GA in much lesser optimization time. The first three models in terms of best optimization performance are followed by the DACE BB model (1.26 \$·kg^{-1}), closely followed by the MeLON BB model (*i.e.*, 1.27 \$·kg^{-1}). The BMS BB model returns the worst value of 1.48 \$·kg^{-1}. It must be noted that the flowsheet in Aspen Plus® failed to converge with the default options for convergence for the solution provided by the DACE BB model. Therefore, the tolerance for tear stream convergence was adjusted from 10^{-4} to 10^{-5} for the DACE BB model. Additionally, the method for tear stream convergence in Aspen Plus® was changed from 'Newton' to 'Wegstein' for the DACE BB model. This ensures that the flowsheet converges.

In terms of the percentage deviation between the f and F values, the BMS HY model presents the minimum deviation, 0.53%. The BMS BB model and DACE BB model have higher deviations of 13.54% and 17.70%, respectively, and the maximum deviation of 49.85% is found for the MeLON BB model. Again, there is no deviation reported for the GA and BO models as they sample directly from the flowsheet.

The optimized values of the seven degrees of freedom for each model are shown in Table 13. For each of the models, we can calculate the stoichiometric number $\boldsymbol{M}$ (Medrano-García *et al.*, 2017), usually defined to measure the quality of the syngas (which is a mixture of CO, CO_2, and H_2):

$$\boldsymbol{M} = \frac{\boldsymbol{Flow\ rate}_{H_2} - \boldsymbol{Flow\ rate}_{CO_2}}{\boldsymbol{Flow\ rate}_{CO} + \boldsymbol{Flow\ rate}_{CO_2}} \tag{13}$$

Here, as there is no CO in the feed, we only consider the flow rate of H_2 obtained from the optimization ($\boldsymbol{F}^{S9}_{H_2}$), and the constant CO_2 molar flow rate of 2000 kmol·hr^{-1}. Usually, values of $\boldsymbol{M}$ close to 2 are preferred for methanol synthesis. From the values obtained from the optimization, we observe that except for the BMS BB model (with a value of 1.25 for $\boldsymbol{M}$), all other approaches result in values close to 2. While there is no clear

Table 12
Results of the minimization problem for case study 2.

Model	Training time [s]	Optimization time for best solution [s]	Total optimization time [s]	f [\$·kg^{-1}]	F [\$·kg^{-1}]	Deviation [%]
BMS HY	23,402.00	0.89	0.89	1.21	1.22	0.53%
BMS BB	10,237.00	Pre-processing	14,400.00	1.28	1.48	13.54%
DACE BB	3.62	Pre-processing	14,400.00	1.03	1.26*	17.70%
MeLON BB	10.96	15.38	15.66	1.90	1.27	49.85%
BO	1.00	11,931.27	14,400.00	NA	1.21	-
GA	NA	13,821.55	14,400.00	NA	1.20	-

*The default convergence options had to be changed as the flowsheet failed to converge

Table 13
Optimized values of degrees of freedom for case study 2.

Degrees of freedom	Symbol	Unit	Model					
			BMS HY	BMS BB	DACE BB	MeLON BB	BO	GA
H_2 flow rate	$F^{S9}_{H_2}$	kmol·hr^{-1}	5,931.04	4,500.00	6,000.00	5,792.03	5,965.86	5982.59
PFR temperature	T	°C	237.58	218.61	203.01	238.83	209.96	215.03
PFR pressure	P	bar	46.18	49.16	47.61	49.16	49.61	54.93
PFR volume	V	m^3	38.08	54.00	54.00	38.08	45.67	53.63
Split fraction	SF	-	$1.60{\cdot}10^{-3}$	$1.00{\cdot}10^{-3}$	$1.00{\cdot}10^{-3}$	$5.00{\cdot}10^{-3}$	$1.00{\cdot}10^{-3}$	$1.00{\cdot}10^{-3}$
Reflux ratio	RR	-	1.25	1.45	1.80	1.45	1.25	1.26
No. of stages	N	-	55.00	51.00	55.00	51.00	55.00	49.00

trend observed for the reactor pressure, temperature, and volume (as they influence multiple aspects of the overall process), their values are at the higher end of their bounds. For the split fraction, a value very close to the lower bound is reported by all the approaches, as a larger recycle would ensure a higher conversion of the feed to methanol. Further, the three best-performing approaches (*i.e.*, GA model, BO model, and BMS HY model, in that order) report the reflux ratio at the lower bound, which would result in lower heat duties for the condenser and reboiler of the distillation column.

For the BMS HY model, the PFR volume is not present in the final BMS equations obtained for the surrogate model of the PFR (SM 2.5). Therefore, as mentioned in Section 4, the value of the PFR volume from the initial point is used in the BMS HY model. Similarly, the PFR pressure, reflux ratio, and number of stages are not present in the final BMS equation obtained for the BMS BB model. Therefore, these values are taken from the initial point and are used in the BMS BB model. A sensitivity analysis of the influence of these degrees of freedom on the BMS BB model is given in section S8 of the supplementary material.

Overall, the BMS HY model is the best-performing optimization strategy in case study 1. It performs the third-best in case study 2, providing a value that is 1.67% higher than the best-performing optimization strategy (*i.e.*, the GA model), and where one of the other approaches (*i.e.*, DACE BB model) fails to provide a solution that converges in the simulation package.

6. Conclusions

Here we explored the use of hybrid models built using Bayesian symbolic learning to simplify the global optimization of process flowsheets. In essence, some of the process units are approximated with analytical surrogates constructed using symbolic regression, which are combined with mechanistic equations. Through comparison with other process optimization approaches, we found that the optimization of hybrid models built using Bayesian symbolic regression provides high-quality solutions in much less CPU time than their fully black-box counterparts. However, building such hybrid models requires much larger CPU times than constructing other standard surrogates (*i.e.*, kriging, GPs), and also more time than generating fully black-box analytical equations of entire flowsheets.

The hybrid approach based on symbolic regression (BMS HY model) performs the best in case study 1, and the third-best in case study 2 (with the derivative-free optimization approaches performing better), where there are no explicit constraints on the degrees of freedom of the process. Thus, while the specific outcome can differ on a case-by-case basis, the BMS HY model has the potential to outperform other alternative approaches, one of which even failed to provide a feasible point after the optimization, *i.e.*, the flowsheet did not converge for the optimized values obtained in the DACE BB model.

Overall, our work showcases the advantages of the hybrid analytical surrogate modeling approach in the global optimization of process flowsheets. Specifically, we showed how hybrid modeling based on analytical equations can address the simultaneous process optimization and heat integration problem successfully. We expect that our approach becomes more competitive as more efficient algorithms for symbolic regression emerge in the literature. Moreover, our method might be particularly appealing in applications where multiple optimizations need to be executed, like in multi-objective optimization *via* single-objective reformulations, as once the model is built it can be optimized effectively iteratively. Additionally, in case the constrained conditions of the process change slightly, our method is expected to remain valid within the bounds of the training points, while it would be necessary to repeat all simulations using the derivative-free approaches. Future extensions could include the incorporation of uncertainties in the modeling framework and the simultaneous optimization of several flowsheets within an integrated cluster.

CRediT authorship contribution statement

Sachin Jog: Conceptualization, Data curation, Formal analysis, Methodology, Software, Validation, Visualization, Writing – original draft, Writing – review & editing. **Daniel Vázquez:** Conceptualization, Formal analysis, Methodology, Supervision, Writing – review & editing. **Lucas F. Santos:** Formal analysis, Methodology, Writing – review & editing. **José A. Caballero:** Formal analysis, Methodology, Supervision, Writing – review & editing. **Gonzalo Guillén-Gosálbez:** Conceptualization, Formal analysis, Funding acquisition, Methodology, Project administration, Supervision, Writing – review & editing.

Declaration of competing interest

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

Data availability

Data will be made available on request.

Acknowledgments

The authors would like to acknowledge the financial support from the Swiss National Science Foundation (Project LEARN-D, MINT 200021_214877). JAC also acknowledges the Spanish Ministerio de Ciencia y Innovación for the financial support under project PID2021-124139NB-C21.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.compchemeng.2023.108563.

References

Bhalode, P., Chen, Y., Ierapetritou, M., 2022. Hybrid modelling strategies for continuous pharmaceutical manufacturing within digital twin framework. Computer Aided Chemical Engineering. Elsevier Masson SAS, pp. 2125–2130. https://doi.org/10.1016/B978-0-323-85159-6.50354-7.

Bongartz, D., Najman, J., Sass, S., Mitsos, A., 2018. MAiNGO – McCormick-Based Algorithm for Mixed-Integer Nonlinear Global Optimization. In: Tech. rep. Process Systems Engineering. (AVT.SVT). RWTH Aachen University.
Boukouvala, F., Floudas, C.A., 2017. ARGONAUT: AlgoRithms for global optimization of coNstrAined grey-box compUTtational problems. Optim. Lett. 11, 895–913. https://doi.org/10.1007/s11590-016-1028-2.
Bradley, W.T., Boukouvala, F., 2014. Merging machine learning with mechanistic models via sequential and integrated hybrid process modeling.
Caballero, J.A., Grossmann, I.E., 2008. An algorithm for the use of surrogate models in modular flowsheet optimization. AIChE J. 54, 2633–2650. https://doi.org/10.1002/aic.11579.
Chen, Y., Ierapetritou, M., 2020. A framework of hybrid model development with identification of plant-model mismatch. AIChE J. 66, 1–16. https://doi.org/10.1002/aic.16996.
Cozad, A., Sahinidis, N.V., 2018. A global MINLP approach to symbolic regression. Math. Program. 170, 97–119. https://doi.org/10.1007/s10107-018-1289-x.
Cozad, A., Sahinidis, N.V., Miller, D.C., 2014. Learning surrogate models for simulation-based optimization. AIChE J. 60, 2211–2227. https://doi.org/10.1002/aic.14418.
Duran, M.A., Grossmann, I.E., 1986. Simultaneous optimization and heat integration of chemical processes. AIChE J. 32, 123–138. https://doi.org/10.1002/aic.690320114.
Fahmi, I., Cremaschi, S., 2012. Process synthesis of biodiesel production plant using artificial neural networks as the surrogate models. Comput. Chem. Eng. 46, 105–123. https://doi.org/10.1016/j.compchemeng.2012.06.006.
Ferreira, J., Pedemonte, M., Torres, A.I., 2022. Development of a machine learning-based soft sensor for an oil refinery's distillation column. Comput. Chem. Eng. 161, 107756 https://doi.org/10.1016/j.compchemeng.2022.107756.
Ferreira, J., Torres, A.I., Pedemonte, M., 2019. A comparative study on the numerical performance of kaizen programming and genetic programming for symbolic regression problems. In: 2019 IEEE Latin American Conference on Computational Intelligence (LA-CCI). IEEE, pp. 1–6. https://doi.org/10.1109/LA-CCI47412.2019.9036755.
Forster, T., Vázquez, D., Guillén-Gosálbez, G., 2023. Algebraic surrogate-based process optimization using Bayesian symbolic learning. AIChE J. https://doi.org/10.1002/aic.18110.
Ghanta, M., Fahey, D.R., Busch, D.H., Subramaniam, B., 2013. Comparative economic and environmental assessments of H_2O_2-based and tertiary butyl hydroperoxide-based propylene oxide technologies. ACS Sustain. Chem. Eng. 1, 268–277. https://doi.org/10.1021/sc300121j.
GPy, 2012. Gpy: A Gaussian process framework in python.
Guimerà, R., Reichardt, I., Aguilar-Mogas, A., Massucci, F.A., Miranda, M., Pallarès, J., Sales-Pardo, M., 2020. A Bayesian machine scientist to aid in the solution of challenging scientific problems. Sci. Adv. 6 https://doi.org/10.1126/sciadv.aav6971.
Henao, C.A., Maravelias, C.T., 2011. Surrogate-based superstructure optimization framework. AIChE J. 57, 1216–1232. https://doi.org/10.1002/aic.12341.
Kahrs, O., Marquardt, W., 2007. The validity domain of hybrid models and its application in process optimization. Chem. Eng. Process. Process Intensif. 46, 1054–1066. https://doi.org/10.1016/j.cep.2007.02.031.
Keith, D.W., Holmes, G., Angelo, St., Heidel, D., 2018. A process for capturing CO_2 from the atmosphere. Joule 2, 1573–1594. https://doi.org/10.1016/j.joule.2018.05.006.
Koza, J.R., 1992. Genetic Programming: On the programming of computers by means of natural selection. MIT Press, Cambridge.
Kramer, M.A., Thompson, M.L., Bhagat, P.M., 1992. Embedding Theoretical Models in Neural Networks. In: 1992 American Control Conference. IEEE, pp. 475–479. https://doi.org/10.23919/ACC.1992.4792111.
Krige, D.G., 1952. A statistical approach to some basic mine valuation problems on the witwatersrand. J. Chem. Met. Min. Soc. South Africa 201–215.
Lim, H., Kang, M., Chang, M., Lee, J., Park, S., 2002. Simulation and optimization of a styrene monomer reactor using a neural network hybrid model. IFAC Proc. Vol. 35, 175–180. https://doi.org/10.3182/20020721-6-ES-1901.01015.
Lophaven, S.N., Nielsen, H.B., Sondergaard, J., 2002. DACE - A Matlab Kriging Toolbox.
Ma, K., Sahinidis, N.V., Amaran, S., Bindlish, R., Bury, S.J., Griffith, D., Rajagopalan, S., 2022a. Data-driven strategies for optimization of integrated chemical plants. Comput. Chem. Eng. 166, 107961 https://doi.org/10.1016/j.compchemeng.2022.107961.
Ma, K., Sahinidis, N.V., Bindlish, R., Bury, S.J., Haghpanah, R., Rajagopalan, S., 2022b. Data-driven strategies for extractive distillation unit optimization. Comput. Chem. Eng. 167, 107970 https://doi.org/10.1016/j.compchemeng.2022.107970.
McBride, K., Sundmacher, K., 2019. Overview of surrogate modeling in chemical process engineering. Chemie Ing. Tech. 91, 228–239. https://doi.org/10.1002/cite.201800091.
McCormick, G.P., 1976. Computability of global solutions to factorable nonconvex programs: Part I — Convex underestimating problems. Math. Program. 10, 147–175. https://doi.org/10.1007/BF01580665.
Medrano-García, J.D., Ruiz-Femenia, R., Caballero, J.A., 2017. Multi-objective optimization of combined synthesis gas reforming technologies. J. CO2 Util. 22, 355–373. https://doi.org/10.1016/j.jcou.2017.09.019.
Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H., Teller, E., 1953. Equation of state calculations by fast computing machines. J. Chem. Phys. 21, 1087–1092. https://doi.org/10.1063/1.1699114.
Misener, R., Biegler, L., 2023. Formulating data-driven surrogate models for process optimization. Comput. Chem. Eng. 179, 108411 https://doi.org/10.1016/j.compchemeng.2023.108411.
Misener, R., Floudas, C.A., 2014. ANTIGONE: algorithms for coNTinuous / integer global optimization of nonlinear equations. J. Glob. Optim. 59, 503–526. https://doi.org/10.1007/s10898-014-0166-2.
Negri, V., Vázquez, D., Sales-Pardo, M., Guimerà, R., Guillén-Gosálbez, G., 2022. Bayesian symbolic learning to build analytical correlations from rigorous process simulations: application to CO_2 capture technologies. ACS Omega 7, 41147–41164. https://doi.org/10.1021/acsomega.2c04736.
Orzechowski, P., Cava, W.La, Moore, J.H., 2018. Where are we now? A large benchmark study of recent symbolic regression methods. In: GECCO 2018 - Proceedings of the 2018 Genetic and Evolutionary Computation Conference. Association for Computing Machinery, Inc, pp. 1183–1190. https://doi.org/10.1145/3205455.3205539.
Ostertagová, E., 2012. Modelling using polynomial regression. Procedia Eng. 48, 500–506. https://doi.org/10.1016/j.proeng.2012.09.545.
Papalambros, P.P., Wilde, D.J., 2000. Principles of Optimal Design: Modeling and Computation, 2nd ed. Cambridge University Press, Cambridge.
Parkinson, B., Balcombe, P., Speirs, J.F., Hawkes, A.D., Hellgardt, K., 2019. Levelized cost of CO_2 mitigation from hydrogen production routes. Energy Environ. Sci. 12, 19–40. https://doi.org/10.1039/C8EE02079E.
Psichogios, D.C., Ungar, L.H., 1992. A hybrid neural network-first principles approach to process modeling. AIChE J. 38, 1499–1511. https://doi.org/10.1002/aic.690381003.
Quirante, N., Caballero, J.A., 2016. Large scale optimization of a sour water stripping plant using surrogate models. Comput. Chem. Eng. 92, 143–162. https://doi.org/10.1016/j.compchemeng.2016.04.039.
Quirante, N., Javaloyes-Antón, J., Caballero, J.A., 2018. Hybrid simulation-equation based synthesis of chemical processes. Chem. Eng. Res. Des. 132, 766–784. https://doi.org/10.1016/j.cherd.2018.02.032.
Sansana, J., Joswiak, M.N., Castillo, I., Wang, Z., Rendall, R., Chiang, L.H., Reis, M.S., 2021. Recent trends on hybrid modeling for Industry 4.0. Comput. Chem. Eng. 151, 107365 https://doi.org/10.1016/j.compchemeng.2021.107365.
Santos, L.F., Costa, C.B.B., Caballero, J.A., Ravagnani, M.A.S.S., 2022. Framework for embedding black-box simulation into mathematical programming via kriging surrogate model applied to natural gas liquefaction process optimization. Appl. Energy 310, 118537. https://doi.org/10.1016/j.apenergy.2022.118537.
Schweidtmann, A.M., Bongartz, D., Grothe, D., Kerkenhoff, T., Lin, X., Najman, J., Mitsos, A., 2021. Deterministic global optimization with Gaussian processes embedded. Math. Program. Comput. 13, 553–581. https://doi.org/10.1007/s12532-021-00204-y.
Schweidtmann, A.M., Mitsos, A., 2019. Deterministic global optimization with artificial neural networks embedded. J. Optim. Theory Appl. 180, 925–948. https://doi.org/10.1007/s10957-018-1396-0.
Schweidtmann, A.M., Netze, L., Mitsos, A., 2020. The Melon Toolbox: Machine Learning Models for Optimization. In: 2020 Virtual AIChE Annual Meeting. AIChE.
The GPyopt authors, 2016. GPyOpt: A Bayesian optimization framework in python.
Tawarmalani, M, Sahinidis, N.V., 2005. A polyhedral branch-and-cut approach to global optimization. Math. Program. 103 (2), 225–249. https://doi.org/10.1007/s10107-005-0581-8.
The MathWorks Inc, 2021b. Global Optimization Toolbox version: 4.6 (R2021b). Natick, Massachusetts: The MathWorks Inc. https://www.mathworks.com.
The MathWorks Inc, 2021a. MATLAB version: 9.11.0 (R2021b). Natick, Massachusetts: The MathWorks Inc. https://www.mathworks.com.
Trent, D.L., 2001. Propylene Oxide. Kirk-Othmer Encyclopedia of Chemical Technology. John Wiley & Sons, Inc., Hoboken, NJ, USA https://doi.org/10.1002/0471238961.1618151620180514.a01.pub2.
Turton, R., Shaeiwitz, J.A., Bhattacharya, D., Whiting, W.B., 2018. Analysis, Synthesis and Design of Chemical Processes. Pearson Education.
Van-Dal, É.S., Bouallou, C., 2013. Design and simulation of a methanol production plant from CO_2 hydrogenation. J. Clean. Prod. 57, 38–45. https://doi.org/10.1016/j.jclepro.2013.06.008.
Vázquez, D., Guillén-Gosálbez, G., 2021. Process design within planetary boundaries: Application to CO_2 based methanol production. Chem. Eng. Sci. 246, 116891 https://doi.org/10.1016/j.ces.2021.116891.
Wang, H., Kontoravdi, C., del Rio Chanona, E.A., 2023. A hybrid modelling framework for dynamic modelling of bioprocesses. Computer Aided Chemical Engineering. Elsevier Masson SAS, pp. 469–474. https://doi.org/10.1016/B978-0-443-15274-0.50075-5.
Williams, B.A., Cremaschi, S., 2019. Surrogate model selection for design space approximation and surrogatebased optimization. Computer Aided Chemical Engineering. Elsevier Masson SAS, pp. 353–358. https://doi.org/10.1016/B978-0-12-818597-1.50056-4.
Xiang, H., Li, Y., Liao, H., Li, C., 2017. An adaptive surrogate model based on support vector regression and its application to the optimization of railway wind barriers. Struct. Multidiscip. Optim. 55, 701–713. https://doi.org/10.1007/s00158-016-1528-9.