Front. Chem. Sci. Eng. 2022, 16(2): 251–273
https://doi.org/10.1007/s11705-021-2071-9

RESEARCH ARTICLE

Synergistic optimization framework for the process synthesis and design of biorefineries

Nikolaus I. Vollmer[1], Resul Al[2], Krist V. Gernaey[1], Gürkan Sin (✉)[1]

1 Process and Systems Engineering (PROSYS) Research Center, Department of Chemical and Biochemical Engineering, Technical University of Denmark, 2800 Kgs. Lyngby, Denmark
2 Novo Nordisk A/S, 2880 Bagsværd, Denmark

Abstract The conceptual process design of novel bioprocesses in biorefinery setups is an important task, which remains yet challenging due to several limitations. We propose a novel framework incorporating superstructure optimization and simulation-based optimization synergistically. In this context, several approaches for superstructure optimization based on different surrogate models can be deployed. By means of a case study, the framework is introduced and validated, and the different superstructure optimization approaches are benchmarked. The results indicate that even though surrogate-based optimization approaches alleviate the underlying computational issues, there remains a potential issue regarding their validation. The development of appropriate surrogate models, comprising the selection of surrogate type, sampling type, and size for training and cross-validation sets, are essential factors. Regarding this aspect, satisfactory validation metrics do not ensure a successful outcome from its embedded use in an optimization problem. Furthermore, the framework's synergistic effects by sequentially performing superstructure optimization to determine candidate process topologies and simulation-based optimization to consolidate the process design under uncertainty offer an alternative and promising approach. These findings invite for a critical assessment of surrogate-based optimization approaches and point out the necessity of benchmarking to ensure consistency and quality of optimized solutions.

Keywords biotechnology, surrogate modelling, superstructure optimization, simulation-based optimization, process design

Received December 17, 2020; accepted May 10, 2021

E-mail: gsi@kt.dtu.dk

1 Introduction

The global necessity of novel process solutions to meet the demands of a growing society and find adequate solutions, matching the increasing need for sustainable process solutions, is high [1]. In general, biotechnological process solutions are deemed to be conceptually more sustainable [2]. Recent developments in synthetic biology allow for producing a vast palette of biofuels, chemicals, foods, and even pharmaceutical ingredients in cell factories. The global biofoundry initiative aims at accelerating the acquisition of knowledge, the integration of data, and the development of new cell factories [3,4].

On the other side, the implementation of biorefinery concepts that incorporate these chemicals' sustainable production from sustainable feedstocks, as, e.g., lignocellulosic biomass or other residues or dedicated crop plants is dramatically low. Up to the current date, a total number of less than 100 active lignocellulosic biorefineries are operated around the world [5]. This is mainly due to these biorefinery concepts' critical economic robustness, as the operative margins for most products are narrow and chemical processes are competitive [6]. Despite several ideas of designing these biorefineries and improving their economic robustness by mass and heat integration or producing several products simultaneously in a so-called multi-product biorefinery, a conceptual design strategy that significantly promotes the implementation of these biorefineries is still an active area of research [7–10]. Current research directions span from the integration and expansion of established approaches [11,12], over the use of novel approaches and models [13,14], up to the inclusion of further economic, environmental, and sustainability factors in particular [15–19], as well as the accommodation of specificities for fermentation-based processes [20]. The vast majority of these studies primarily follow three approaches for conceptual process design that are briefly

introduced in the following paragraphs.

Classically, process design for biorefineries is performed by so-called hierarchical decomposition and involves domain knowledge from different expert areas to yield a consolidated solution [21]. However, this methodology has several shortcomings as it usually involves several iterations, which results in a long idea-to-process time, and does not necessarily yield an optimal solution for the process design, especially considering that biotechnological processes suffer from issues with the scale-up from laboratory scale to production scale [3,22].

A computational and more conceptual approach to process design is superstructure optimization (SSO) [23]. Out of all possible process options provided in a superstructure, the optimal configuration is chosen through mathematical optimization. In comparison to the previous alternative, the resulting process design is globally optimal. However, the methodology is initially limited by the number of alternatives that are included in the superstructure. Furthermore, the resulting optimization problem can become very large even though a considerable number of possible solutions can be infeasible a priori. Also, the use of high-fidelity models is limited in this approach as this can be computationally challenging to solve. Lastly, the incorporation of uncertainty increases the complexity of formulating the optimization problem through stochastic or robust optimization [24]. SSO has been successfully performed for several chemical production processes; however, the design of biotechnological processes comprises several new challenges and uncertainties which are difficult to address [25].

A second, more recent approach to process design is simulation-based optimization (SBO). The methodology builds upon the evaluation of process simulations where high-fidelity models are used [26]. Furthermore, uncertainties can be easily included in a Monte Carlo-based approach [27]. Several proposed frameworks utilize machine learning surrogate models as, e.g., stochastic kriging (SK), coupled with a Bayesian optimization approach to iteratively improve the objective function results, leading to an optimal process design [27,28]. Despite this approach being able to incorporate complex physiological models of cell factories and processes, it is heavily limited by available computational power. The number of simulations can get very high for several alternative process configurations, which constrains this approach severely to a small design space.

After elucidating the challenges within the process design of novel bioprocesses in biorefinery setups, it becomes evident that novel solutions are required, which incorporate knowledge from different fields and approach the problem in an interdisciplinary manner. Hence, we propose a novel synergistic optimization framework for the conceptual process design of biorefineries, based on a hybrid approach integrating surrogate-based SSO with SBO. It harnesses both the power of the SSO for process synthesis and the potential of SBO for detailed design optimization. The framework itself capitalizes biotechnological knowledge to guide decisions in a bottom-up strategy for both SSO and SBO, which finally yields a consolidated process design.

The remainder of the paper is structured as follows: The used surrogate models for the superstructure are introduced in section 2.1 and the methods of performing SSO in section 2.2, and the SBO in section 2.3. The detailed components of the proposed framework and the applied workflow are introduced in section 2.5. For the application of the framework, a case study is introduced in section 3.1. The corresponding high-fidelity models for the specific case study are introduced in appendix A.1 (cf. Electronic Supplementary Material, ESM), including the corresponding methodology for uncertainty and sensitivity analysis in appendix A.2 (cf. ESM). The evaluation of the framework's SSO step is presented in section 3.2 and the SBO step in section 3.3. Conclusions and an outlook on future research work are given in section 4.

2 Theoretical background

2.1 Surrogate models

Surrogate models describe various types of models, which mimic the behavior of original first-principle models. The primary reason for developing and using a surrogate model, described by McBride and Sundmacher, is the reduction of computational costs for model evaluations, which is an inherent part of the SSO, as explained in section 2.2. This comes at the cost of a certain error in the surrogate model's predictions due to imperfect resemblance. The applied technique for creating a surrogate model highly depends on the computational constraints regarding the original and the surrogate model, as well as the type of application of the surrogate model [29]. The simplest model types are linear, piecewise linear, or polynomial functions, which are fitted to the original data [30,31]. However, the majority of the used surrogate models are based on machine learning techniques, among others radial basis functions, Gaussian process regressors, and at the higher end artificial neural networks (ANNs) and even deep neural networks [29,32]. The following subsections describe first the development procedure for these model types and subsequently four different surrogate model types which are used in this framework, one based on algebraic equations, one piecewise linear model, and two machine learning models.

The development procedure for surrogate models — independent of the model type is similar and commonly starts with sampling the model input space by, e.g., Latin hypercube (LHS) or Sobol sampling. To ensure a space-filling sampling for a sufficient interpolative quality of the model, the choice of an adequate sampling technique is

crucial [33]. Subsequently, the surrogate model is fitted to a part of the data, dependent on the chosen validation strategy.

A standard method, which is also applied in this framework, is called k-fold cross-validation: the sampled dataset is divided into k equally sized subsets. In a routine, each subset is once used to validate the surrogate model, and the remaining $k-1$ subsets are used for the fitting of the model. The total validation error is calculated by the average error of the k validation folds [34].

2.1.1 Automated learning of algebraic models

The automated learning for algebraic models (ALAMO) toolbox was developed to create algebraic models for applications like SSO [35,36]. As described in Wilson's and Sahinidis' work, the ALAMO toolbox first fits simple, algebraic models, consisting of several nonlinear terms and their linear combinations to the points in the input space. Then, by employing derivative-free optimization, the best suitable combination of terms is determined through error maximization sampling. Further constraints on the model outputs can be imposed, and the possibility to perform adaptive sampling to increase the number of points in the input space at specific locations [35]. Especially for process design purposes, the ALAMO toolbox has been applied widely as a surrogate modeling technique with promising results [22,37,38].

2.1.2 Delaunay triangulation regression (DTR)

The second surrogate model is based on the concept of regression by a piecewise linear functional relationship. In the case of a one-dimensional input space, the piecewise linear functional relationship is a set of composed line segments. For a general definition of this piecewise linear functional relationship, be $P \subset \mathbb{R}^{n\times r}$ a set of points in the $n \times r$-dimensional space and a subset of real numbers. All elements $p \in P$ are considered vectors in the Euclidean vector space over $\mathbb{R}^{n\times r}$ and consist of two elements $p = \{p^x,p^z\}$. All elements p^x and p^z are respectively elements of the sets $P_x \subset P$ and $P_z \subset P$, both subsets of P. Furthermore, be $X \subset \mathbb{R}^n$ the input space and $Z \subset \mathbb{R}^r$ the output space of our functional relationship with all elements $x \in X$ and $z \in Z$ and $P \subset X \times Z$.

Defining the input space X as n-manifold with boundary, the triangulation T_X of X is a homogeneous simplicial k-complex Σ, being homeomorphic to X and $k = n$, referring to that the S consists only of n-simplices. An n-simplex σ is the special form of an n-polytope and consists of exactly $n+1$ vertices v, so $\{v_1,...,v_{n+1}\} \subseteq \sigma \in \Sigma$. For illustrative purposes, for $n = 2$ the 2-simplex is a triangle. Hence the n-simplex is the equivalent to a triangle in any dimension n. At this point, we define the set of vertices V and every vertex $v \in V$ as the element p^x of the point p. Furthermore, an important property of an n-simplex σ is it being an affine and hence convex space, which allows any point x^* inside the simplex to be described as a linear combination of the vertices with the coefficients $\alpha_i < 1$:

$$x^* = \sum_{i=1}^{n+1} \alpha_i \cdot v_i = \sum_{i=1}^{n+1} \alpha_i \cdot p_i^x \ \forall\ v_i,\ p_i^x \in \sigma. \quad (1)$$

Furthermore, the functional relationship of the surrogate model f is defined as:

$$f : \begin{cases} \mathbb{R}^n \longrightarrow \mathbb{R}^r \\ x \mapsto z \end{cases}. \quad (2)$$

The set of real numbers is denoted as $\mathbb{R}$. Given this, the conclusion for the surrogate model is that any point z^* can also be described by a linear combination of elements p^z of the point p by Eq. (1), which is equivalent to the concept of regression. In conclusion, the triangulation of the input space T_X with the functional relationship to the output space $Z = f(T_X)$ represents the surrogate model based on the described regression concept.

Reconsidering the one-dimensional case, a 1-simplex, also called edge, is nothing else than a line segment; the one 1-simplicial complex corresponds to the set of composed line segments, where each point on any line segment can be described as a linear combination of the two vertices or endpoints of the line segment it lies on. Lastly, the functional relationship describes the functional value of the point of the line segment by a linear combination of the vertices' functional values.

A prominent type of triangulation is the so-called Delaunay triangulation, which imposes the criterion that the circumcircle or its higher-dimensional equivalent, its n-hypersphere of the vertices of the simplex cannot contain another vertex [39]. There exist published algorithms to create a DTR for a given set of points [40,41].

2.1.3 Gaussian process regression (GPR)

The third surrogate model is based on the concept of regression by a stochastic process — the eponymous GPR and particular kernel functions to determine parameters of the following functional relationship:

$$f : \begin{cases} \mathbb{R}^n \longrightarrow \mathbb{R} \\ x \mapsto z \end{cases}, \quad (3)$$

with the expression, as explained by Al et al.:

$$z = \mu_{\mathrm{GP}}(x) + \sigma^2_{\mathrm{GP}} \cdot \mathcal{F}(x,\omega),\ \mu_{\mathrm{GP}}(x) = \rho \cdot \beta(x), \quad (4)$$

where, $\mu_{\mathrm{GP}}(x)$ denotes the mean value of the Gaussian process and σ^2_{GP} its variance. For the mean value $\mu_{\mathrm{GP}}(x)$, ρ are estimated parameters from the input data, as well as the variance σ^2_{GP} and $\beta(x)$ relate to a group of basis functions. Furthermore, $\mathcal{F}(x,\omega)$ describes a zero mean unit variance

stochastic process. The correlation or kernel functions ω are used to correlate any point in the input space x^* with existing points x [42]. There are various available kernel functions; hence, the reader is referred to the book by Rasmussen [43]. A couple of remarks on GPR as a surrogate model shall be made: it is very well capable of displaying highly nonfunctional relationships, and the amount of data necessary to obtain a good surrogate model is relatively low and favorably low-dimensional [29]. GPR surrogate models have seen a wide range of applications in machine learning, particularly also in different process design tasks [44–46].

2.1.4 ANN

The last type of surrogate model is the class of ANNs. ANNs are used in a plethora of applications, especially complex machine learning tasks as image and voice recognition, natural language processing, and artificial intelligence. Al et al. give the following description: in general, neural networks consist of at least three layers, one input, one hidden, and one output layer. This case is considered as a shallow neural network and, in particular, a multi-layer perceptron. Each layer contains a certain number of nodes that relate their inputs to their outputs with a so-called transfer or activation function of a specific type. When several hidden layers are added to the network, it is referred to as a deep neural network [42]. Due to their flexible architecture and possibilities in terms of composing and learning a network, neural networks are an ideal candidate for the use as a surrogate model in process design tasks and have seen a widespread application in the area of process systems engineering [29,47–49].

2.2 SSO

SSO as a method for process synthesis and design is a computational method based on mathematical optimization. Following Chen and Grossmann's elaboration, SSO involves three steps, namely (1) the definition of a set of process alternatives in a superstructure representation, (2) the formulation of the corresponding optimization problem, and (3) solving the optimization problem with an adequate solver in order to obtain the optimal process design. The primary limitations remain firstly with the initial definition of the design space, implying that only solutions can be found that are initially considered; secondly, these limitations are also fueled by the assumptions taken for the models and the capabilities of the employed solver for the optimization problem [23]. Moreover, a capital restriction is the inability to account for uncertainty in deterministic optimization approaches such as SSO [24].

The superstructure itself can be postulated in various ways. One of the most common superstructure formulations is a state-task-network, in which each unit operation model forms a task in the network, the flows are described as states, and both are connected via nodes that correspond to mixers or splitters. This composition highly resembles an actual process flowsheet, with the main difference being that the nodes represent binary decisions on whether a process path exists [50]. For the interested reader, a description of other existing superstructure formulations can be found in a recent review by Mencarelli et al. [51].

As described by Chen and Grossman, the resulting optimization problem is a mixed-integer program (MIP) of the form:

$$\mathrm{MIP}: \begin{cases} \min\ z = f(x,y) \\ s.t.\ h(x,y) = 0 \\ g(x,y) \leqslant 0 \\ x \in X,\ y \in [0,1] \end{cases}, \tag{5}$$

with a defined objective function $z = f(x,y)$, referring to a certain metric, e.g., product purity, a key performance indicator or sustainability measures, and subject to equality and inequality constraints $h(x,y)$ and $g(x,y)$, describing physical constraints, system and equipment specifications and their limits, as well as other process constraints, e.g., product purity. The continuous variables $x \in X$ denote process variables as states, mass and energy flows, and design parameters, all within a specific input space $X \subseteq \mathbb{R}$. The binary variables y denote the mentioned decisions on the existence of equipment or process paths. Depending on the underlying physical system, both the objective function f and the constraints can be nonlinear; hence the optimization problem is a mixed-integer nonlinear program (MINLP).

The objective function evaluation can be theoretically calculated based on the unit operation models and additional equations. However, the complexity of the formulation and the computational cost of evaluating high-fidelity models commonly makes the solution of the optimization problem by currently available solvers complicated or even intractable [49,51]. Consequently, there have been various research efforts in surmounting these hurdles in SSO, e.g., linearization of the objective function, surrogate model-assisted SSO, or decomposition algorithms [50,51]. Especially the capacity of surrogate model-assisted SSO alleviating the computational burden has been exploited in various studies [49,52,53]. This allows for an elegant solution to integrate complex high-fidelity models from different platforms indirectly via their surrogates into a simple superstructure formulation, compared to extensive equation-based approaches as, e.g., generalized disjunctive programming [49,51].

In the following section, four different superstructure formulations with respectively one of the introduced surrogate models in section 2.1 are introduced. The

benchmark of all four approaches is part of sections 3.2.2 and 3.2.3.

2.2.1 Surrogate-assisted MINLP

The first option under investigation are ALAMO surrogate models in an MINLP and no reformulation of the actual optimization problem. The resulting algebraic model equations from the ALAMO surrogate models can be introduced as the function for the objective function $f(x,y)$ and the equality constraints $h(x,y)$ in the MINLP formulation described in Eq. (5), the problem can be solved with an adequate solver. As the ALAMO surrogates are fitted from flowsheet simulations, a binary variable y_m for all flowsheet options $m \in I_M$ is introduced, as well as an SOS1 constraint, indicating that the optimal solution can only lie in one flowsheet:

$$\sum\nolimits_{m \in I_M} y_m = 1. \tag{6}$$

2.2.2 Surrogate-assisted mixed-integer linear program (MILP)

The second option for the SSO suggested here involves developing a superstructure based on DTR surrogate models and the reformulation of the underlying MINLP into an MILP.

For the set of possible process design configurations M with an index set $I_M = \{1,2,...,M\}$, a certain number of flowsheet simulations are performed for each configuration $m \in M$. A DTR surrogate model is then fitted for each configuration $m \in M$ from a set of sample points $P_{m,x}$ over the input space $X \subset \mathbb{R}^n$ and a set of simulation points $P_{m,z}$ over the output space $Z \subset \mathbb{R}^r$. Every sample point p_m^x and respectively p_m^z equally is a set of the type $p_m^x = \{p_{m,1}^x, p_{m,2}^x, ..., p_{m,n}^x\}$ and $p^z = \{p_{m,1}^z, p_{m,2}^z, ..., p_{m,r}^z\}$ with every $p_{m,d}^j$ being a scalar value. Besides, two index sets $I_X = \{1,2,...,n\} \subset \mathbb{N}$ and $I_Z = \{1,2,...,r\} \subset \mathbb{N}$ for the dimension of the input and output space are defined and an index set I_P for set P_x and P_z as $|P_x| = |P_z|$. The created triangulation T_X yields a set of simplices S, fulfilling the definitions stated in section 2.1.2. Be $I_S = \{1,2,...,|S|\} \subset \mathbb{N}$ the index set of S, then each simplex $\sigma_{m,s}$ again is a specific set of points $p_{m,d,i}^x$ with $i \in I_{SP} = \{i \mid p_{m,d,i}^x \in \sigma_{m,s} \forall d \in I_X\} \subseteq I_P$ that are all part of the simplex. Lastly, a binary variable $y_{m,s}$ is introduced, which relates to whether a point x is part of a simplex $\sigma \in S$, in the flowsheet configuration $m \in M$.

The MILP is postulated as follows: The objective function is the minimum value of one element of the output variable z:

$$\min z_k, \tag{7}$$

with $z_k \in z$ and $k \leqslant r$. Furthermore, for any input variable x on the boundary ∂X of X and output variable z under the given functional relationship Eq. (1) for T_X, we can rewrite Eq. (2) in the following way:

$$y_{m,s} \cdot x_d = \sum\nolimits_{i \in I_{SP}} y_{m,s} \cdot \alpha_{m,s,i} \cdot p_{m,d,i}^x$$

$$\forall\ d \in I_X, s \in I_S, m \in I_M, \tag{8}$$

$$y_{m,s} \cdot z_d = \sum\nolimits_{i \in I_{SP}} y_{m,s} \cdot \alpha_{m,s,i} \cdot p_{m,d,i}^z$$

$$\forall\ d \in I_Z, s \in I_S, m \in I_M. \tag{9}$$

Furthermore, due to the convex property of the simplex, the following equation considering the linear combination coefficients $\alpha_{m,s,i}$, which are treated as variables by the optimizer, must be fulfilled:

$$\sum\nolimits_{i \in I_{SP}} \alpha_{m,s,i} = 1\ \forall\ s \in I_S, m \in I_M. \tag{10}$$

Since an MILP only allows the formulation of linear equations, a Big-M notation for Eqs. (8) and (9) is introduced with the three variables ρ, φ and ψ replacing the three products $y_{m,s} \cdot x_d$, $y_{m,s} \cdot \alpha_{m,s,i}$ and $y_{m,s} \cdot z_d$. The following are the four equations for the product $y_{m,s,d} \cdot x_d$:

$$\left.\begin{aligned} \rho_{m,s,d} &\leqslant M_d^x \cdot y_{m,s} && (11)\\ \rho_{m,s,d} &\leqslant x_d && (12)\\ \rho_{m,s,d} &\geqslant x_d - M_d^x \cdot (1 - y_{m,s}) && (13)\\ \rho_{m,s,d} &\geqslant 0 && (14) \end{aligned}\right\} \forall s \in I_S, d \in I_X, m \in I_M.$$

In these equations, M denotes so-called Big-M parameters; their values are set to the upper bound of the respective variable they represent. Similarly, also the Big-M notation for the second product $y_{m,s} \cdot \alpha_{m,s,i}$ is written:

$$\left.\begin{aligned} \varphi_{m,s,i} &\leqslant y_{m,s} && (15)\\ \varphi_{m,s,i} &\leqslant \alpha_{m,s,i} && (16)\\ \varphi_{m,s,i} &\geqslant \alpha_{m,s,i} - (1 - y_{m,s}) && (17)\\ \varphi_{m,s,i} &\geqslant 0 && (18) \end{aligned}\right\} i \in I_{SP}, \forall s \in I_S, m \in I_M.$$

In this case, the Big-M parameter is equal to 1 and hence left out of the equations. Lastly, also the Big-M notation for the third product $y_{m,s} \cdot z_d$ is listed here:

$$\left.\begin{aligned} \psi_{m,s,\mathrm{d}} &\leqslant M_d^y \cdot y_{m,s,d} && (19)\\ \psi_{m,s,d} &\leqslant z_d && (20)\\ \psi_{m,s,d} &\geqslant z_d - M_d^y \cdot (1 - y_{m,s,d}) && (21)\\ \psi_{m,s,d} &\geqslant 0 && (22) \end{aligned}\right\} \forall s \in I_S, d \in I_Z, m \in I_M.$$

Equations (11)–(22) now allow to express Eqs. (8) and (9) without the multiplication of variables in the following way:

$$\rho_{m,s,d} = \sum\nolimits_{i \in I_{SP}} \varphi_{m,s,d,i} \cdot p^{x}_{m,d,i} \;\forall\; d \in I_X,\, s \in I_S,\, m \in I_M, \tag{23}$$

$$\psi_{m,s,d} = \sum\nolimits_{i \in I_{SP}} \varphi_{m,s,d,i} \cdot p^{z}_{m,d,i} \;\forall\; d \in I_Z,\, s \in I_S,\, m \in I_M. \tag{24}$$

To satisfy Eqs. (23) and (24), there exists exactly one simplex in one flowsheet configuration in which the point x is located, and all other simplices in this flowsheet and all other flowsheets do not contain the point. Hence, an SOS1 constraint is added to the postulation in order to express this:

$$\sum\nolimits_{m \in I_M} \sum\nolimits_{s \in I_S} y_{m,s} = 1. \tag{25}$$

With Eqs. (7) and (10), the MILP is now well defined and can be solved with a suitable optimization algorithm. Alternatively, further constraints on other elements of the output variables can be imposed, referring to other process metrics, e.g., economic indicators. For different types of triangulations, this MILP has been postulated in similar ways by Misener et al. [54,55].

2.2.3 Surrogate-assisted series of nonlinear programs (NLP)

The third option for the SSO suggested by the authors involves developing a superstructure based on GPR or ANN surrogate models and the reformulation of the underlying MINLP into a series of NLP.

To remove the integer variables y from the problem as stated in Eq. (5), which are introduced for different flowsheet options, the series of NLP has to involve an NLP for each of these flowsheet options. For the set of possible process design configurations M with an index set $I_M = \{1,2,...,M\}$, a certain number of flowsheet simulations are performed for each configuration $m \in M$. For each flowsheet configuration $m \in M$ a GPR or ANN surrogate is then fitted from a set of sample points $P_{m,x}$ over the input space $X \subset \mathbb{R}^n$ and a set of simulation points $P^{\mathrm{OBJ}}_{m,z}$ for the metric to be evaluated in the objective function over the output space $Z^{\mathrm{OBJ}} \subset \mathbb{R}$. This procedure is performed for each of the metrics j to be considered as constraint with a set of simulation points $P^{\mathrm{CON}}_{m,z,j}$ over the respective output space $Z^{\mathrm{CON}}_j \subset \mathbb{R}$. The boundaries for the input variables are set to the bounds of the input space ∂X and a set of initial points in the input space $X_0 \in X$ are declared, in order to assure global optimality by performing a multi-start optimization with an amount of $|X_0| = s$. The objective function is equally formulated as for the MILP with Eq. (7). The series of NLP can now be solved sequentially for each NLP performing being solved s times with the different initial points $x_0 \in X_0$ by using a suitable optimization algorithm.

2.3 SBO

All the presented SSO approaches in section 2.2 are deterministic approaches to optimization, relating to the found optimum, not incorporating any kind of uncertainty. In contrast to that, the process design of a novel chemical or biochemical process inherently represents a significant uncertainty by itself [56]. Due to the development of computational power and the ever-more growing use of simulation software, making the computational tractability of complex systems possible, the concept of SBO or, in particular, stochastic simulation optimization has seen increasing interest over the past years. This concept allows to incorporate stochastic considerations as uncertainty into an optimization formulation and solve them, given sufficient computational capacities [26–28]. In contrast to the prior presented approaches, the system to be optimized does not need to follow a particular mathematical structure, as the systems are commonly treated as black boxes [27].

Drawbacks of SBO are that the ability to find an optimum to a given optimization problem with SBO is heavily constrained by the computational tractability, which can be easily exceeded by excessively high computational costs for simulation evaluations, a high dimensionality of the problem, or the description of multiple objectives and constraints [26]. Furthermore, as information about derivatives in black-box systems is not readily available, the proof of global optimality for the obtained solution by SBO remains challenging [57].

Among the several approaches for performing SBO, a surrogate model-based method with SK surrogate models will be elucidated in this section, as their use is favorable for computationally expensive simulations [44]. The interested reader is referred to a comprehensive summary of the benefits and drawbacks of several other approaches by Amaran et al. [57]. Besides the mentioned benefits of SBO, a surrogate-based approach gives into the structure of the search space and location of a possible global optimum [58]. SK as surrogate model type is an extended variant to the presented GPR in section 2.1.3. For the individual differences, the reader is referred to the original contributions [28,59].

Following the framework described by Al et al., the SK surrogate is described with the following functional relationship for any input point $x \in \mathbb{R}^n$, closely related to Eq. (4):

$$z = \mu + \varepsilon(x) + \mathcal{L}(x), \tag{26}$$

where μ is now a constant term referring to the mean value of the prediction and $\varepsilon(x)$ and $\mathcal{L}(x)$ representing extrinsic

and intrinsic uncertainty in the prediction. Extrinsic uncertainty describes the uncertainty of the surrogate model with regards to the high-fidelity model due to an imperfect representation of the input space $X \subset \mathbb{R}^n$ by the set of sampling points P_X. Intrinsic uncertainty represents uncertainty in the original model with regards to the original physical system it represents. The model is fitted to the initial set of sampling points of the input space P_X and simulation points P_Z in the output space $Z \in \mathbb{R}$.

The performed optimization works as an evolutionary program, where an adaptive search with infill optimization under a given infill criterion is performed. With an initially small set of sample points, the infill optimization directs the search towards more promising areas in the search space by adding new sample points to the set P_X. In this iterative procedure, the SK model is updated and hence improved. The iterations are terminated upon reaching a specific criterion, and the simulation results from each step can be investigated, and the optimum can be determined from the set of sample points of the last iteration i^* [27]:

$$\min z = \min P_{i^*,X}. \tag{27}$$

Different infill criteria can be used for this task, and the reader is referred to the literature for an overview [27].

2.4 Optimization under uncertainty

Process design inherently involves various sources of uncertainty. These uncertainties can be accommodated in the optimization; however, this always implies a certain imparity in the calculated objective function, a so-called "price of robustness" [60]. Several factors influence this tradeoff: Grossmann et al. mention the availability of information on the type of uncertainty, as well as general data on the uncertainty, the way of hedging against the uncertainty, a tremendous computational burden, as well as difficult tractability of the results for this [24].

Established practices in performing optimization under uncertainty in connection with mathematical programming are robust optimization, chance-constraint programming, or stochastic programming, which are investigated for several decades and recently received attention by the use of data-driven modeling techniques as, e.g., deep and reinforcement learning in order to alleviate certain shortcomings of the classical mathematical programming approaches [61]. However, for increased tractability of the optimization problem, simulation-based approaches that use high-fidelity simulation models and a metamodeling approach seem to be a more viable solution [24,62].

Here, Monte Carlo methods, as introduced in section A.2.1 (cf. ESM), represent a straightforward approach to include the optimization under uncertainty in SBO and allow a simple statistical quantification of the objective and the constraint values. This has been performed successfully for similar process design tasks [27,63]. This procedure is integrated with the presented solver in section 2.3 and will thus be used as such in the scope of this study.

2.5 Framework S3O

As mentioned, the conceptual design of novel bioprocesses in a biorefinery setup is a highly complex challenge. They arise from the difficulty of utilizing sustainable feedstocks as lignocellulosic biomass, which requires additional unit operations in the upstream process, the problematic nature of microorganisms of showing scale-dependent dynamics due to the intricate interactions in the regulation within the transcriptional and metabolic network, and the arduous conception of the downstream process, which has to account for the specifications in the upstream process while being constrained by economic limitations. While all three introduced process design strategies can be utilized, the resulting process design can face several hurdles deriving from a lack of conceptuality, intrinsic limitations in their feasibility, or disproportionate computational burdens. To surmount these hurdles, the proposed framework aims to leverage synergies in applying all proposed strategies in a hybrid manner, where the benefits of each methodology are harnessed to expedite the global task of designing a process conceptually.

It comprises three sequential steps, namely (1) the selection of product sets, substrates, and operations, (2) SSO for determining candidate process topologies, and (3) simulation optimization for consolidating an optimal process design. It is illustrated in Fig. 1.

2.5.1 Selection of product sets, substrates, and operations

The overarching idea in the framework is to "begin with the end in mind" [64,65]. Applying this principle in a biorefinery context, the first thing to be defined is the set of products. Due to the critical economic viability of biorefinery concepts, it is of utmost importance to choose an appropriate portfolio of products, which exploits the available substrate to the maximally possible extent and potentially maximizes the economic key performance indicators of the biorefinery. Once the set of products is defined, the feedstock for the biorefinery has to be chosen accordingly, where a feedstock candidate should contain reasonable amounts of the respective substrate that is needed to produce the desired set of products.

Based on both a defined set of products and a feedstock, all potential process candidates' necessary unit operations can be defined. As an integral part of this framework, this step is heavily influenced by domain knowledge, which allows for a bottom-up assessment of the possible alternatives reducing the workload in this step immensely. With the defined number of alternatives for unit operations and process routes, a mechanistic model is developed for

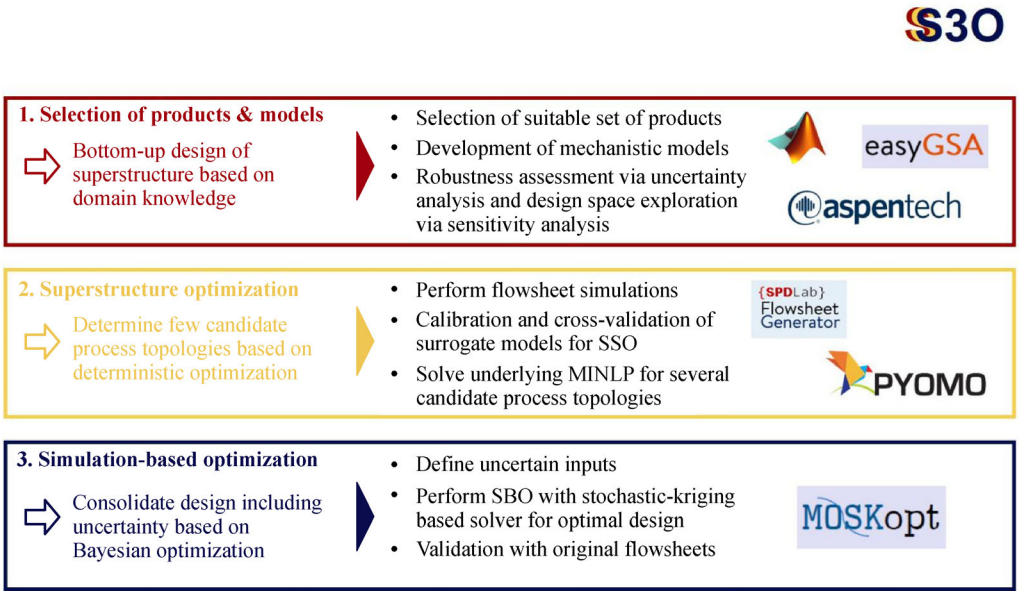

Fig. 1 Illustration of the proposed framework S3O with its three stages: (1) selection of products and models, (2) SSO, and (3) SBO, as well as the employed software and toolboxes.

each unit operation, incorporating the domain knowledge in the form of mathematical equations to describe the underlying physical, chemical, and biological phenomena.

2.5.2 SSO

With all developed models and the set of all process alternatives, a superstructure is composed. By the prior selection of possible process alternatives through expert knowledge and ultimately subjected to economic considerations, the size of the superstructure is a priori heavily reduced and hence faster to solve. Classically, by performing SSO, the result is a deterministically optimal process design. A second paramount benefit of SSO is the possibility to account for nontrivial design decisions, deriving from the nature of binary decisions in the process design. However, this does not allow directly to incorporate possible uncertainties, which is why the SSO in this framework serves as a selection tool for not only one deterministic optimal process design but rather several candidate process topologies. Furthermore, additional constraints regarding the operability of the process, product quality constraints, or others can be included in this stage.

2.5.3 Simulation optimization

By selecting several candidate process topologies through SSO, the number of candidates subjected to simulation optimization is again reduced and hence computationally facilitated. The candidate topologies, which are prone to be the best process design, are subjected to simulation optimization under uncertainty to ultimately yield a consolidated process design. Possible uncertainties that can be included span from uncertainties in the technical domain, e.g., scale-up issues for fermentation reactors or fluctuations in impurities that are separated in the downstream processing, over the operational domain, e.g., varying product and feedstock prices and supply, up to the computational domain, e.g., uncertainties in model parameters, error propagation properties of models and uncertainties in design parameters, as already mentioned in section A.2 (cf. ESM). Constraints from the prior stage can be added, as well as additional constraints. The entire workflow of the implemented framework is shown in Fig. 2.

For the SSO, it is crucial to assess which option delivers the best results under the given objective of determining a small set of candidate process topologies. The constraints here are that (1) the methodology should be able to pick candidates that turn out to be optimal, (2) delivers results consistently with respect to differently shaped design spaces and flowsheet alternatives, (3) the SSO can be performed within a reasonable amount of time with reasonable computational resources, (4) the results are consistent with each other, e.g., considering different sample sizes for the input space and (5) should yield solutions which are close to the theoretical underlying global optimum of the original flowsheet model. In the following section 3.2, all proposed options are evaluated and benchmarked regarding these criteria to define which methodology to choose ultimately.

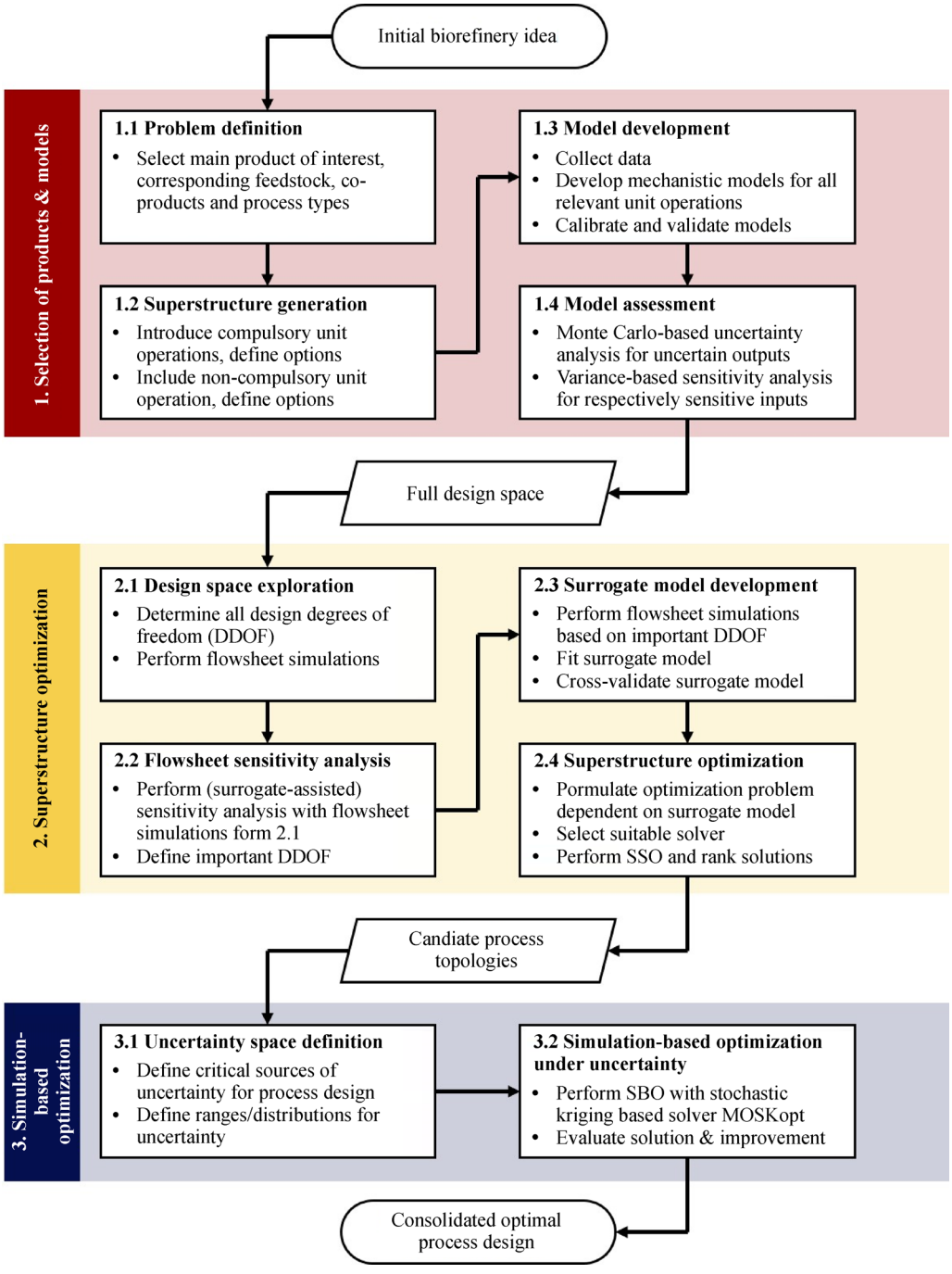

Fig. 2 Workflow of the proposed framework S3O indicating the tasks in the three stages and its intermediate and the final results.

3 Case study and results

To demonstrate the feasibility and capabilities of the introduced framework, a relevant industrial case study is selected. After a concise description of the case study and its relevance to industry, the application of the three steps of the framework and the results are described and analyzed.

3.1 Selection of product sets, substrate and operations

As the first and main produced chemical in the biorefinery concept in the case study, xylitol is selected. Xylitol is a sugar substitute with a similar taste to sucrose and manifold beneficial health properties, as around 40% fewer calories than sucrose, a very low glycemic index, which makes it very suitable for diabetic nutrition, and

anti-cariogenic properties [66]. It has been determined as one of the top 12 chemicals to be produced in a biorefinery concept by the US Department of Energy already in 2004, maintained that status in 2010, and is still attracting researcher's interest in order to facilitate a biotechnological production in cell factories [67–71]. Nevertheless, and despite a market volume estimate of 1 Bio USD and more in 2022, there are only a few companies worldwide, e.g., DuPont Nutrition Biosciences, producing xylitol in a chemical process from wood biomass or corn in a chemical conversion process [70,72].

Proposed process designs in literature are scarce; however, those who approach the idea of a process design for the biotechnological production of xylitol either conclude that the chemical production process is economically more safe and promising or that xylitol can increase the added value of a biorefinery with other principal products, as for example ethanol [73–75]. Hence, the process design of a biotechnological production process for xylitol is a perfectly suitable case study for demonstrating the proposed framework.

Beginning with the end in mind, according to the framework, a suitable set of products is supposed to be chosen in the first step of the workflow. As xylitol is favorably produced from the hemicellulosic fraction of lignocellulosic biomass, the product set can involve a product for the cellulosic fraction of the biomass and one for the lignin fraction. For the cellulosic fraction, succinic acid is considered as a product. It is also one of the top 12 chemicals to be produced in a biorefinery setup and has potential as a platform chemical, which makes it attractive for several industrial branches, which benefits the economic resilience of the biorefinery [67,76]. The lignin fraction can be either used in a combustion process to provide heat for steam generation in order to integrate heat over the different process trains in the biorefinery or can be further converted in a pyrolysis process into sustainable aviation fuels, for which there is a high demand with higher economic margins than for common biofuels as, e.g., bioethanol [77–79]. The downstream process for both the cellulosic and the hemicellulosic process train can involve a classic setup with an evaporation unit and following crystallization units or involve alternative technologies as membrane separation; both approaches are applied commercially and have been investigated for their use in biorefinery downstream processes for both xylitol and succinic acid [53,68,76]. As possible feedstock, wheat straw with a high hemicellulosic content is chosen. A potential superstructure formulation for this base-case process design is illustrated in Fig. 3.

It becomes evident that by following the concept of beginning with the end in mind and rigorously applying expert knowledge in a bottom-up composition approach for the superstructure, the initial search space is kept comparatively small, which expedites all the following steps in the workflow.

However, in order to analyze the results of the application of this framework more tractable and thus accessible to the reader, we reduce the superstructure of this base case design to a smaller subset of only the xylitol production train with a limited amount of unit operations. Once the framework is validated, the whole superstructure, as in Fig. 3, can be processed nonetheless. The reduced superstructure involves six unit operations, namely a biomass pretreatment unit operated as dilute acid pretreatment (PT), an upconcentration unit (UCH), a fermentation

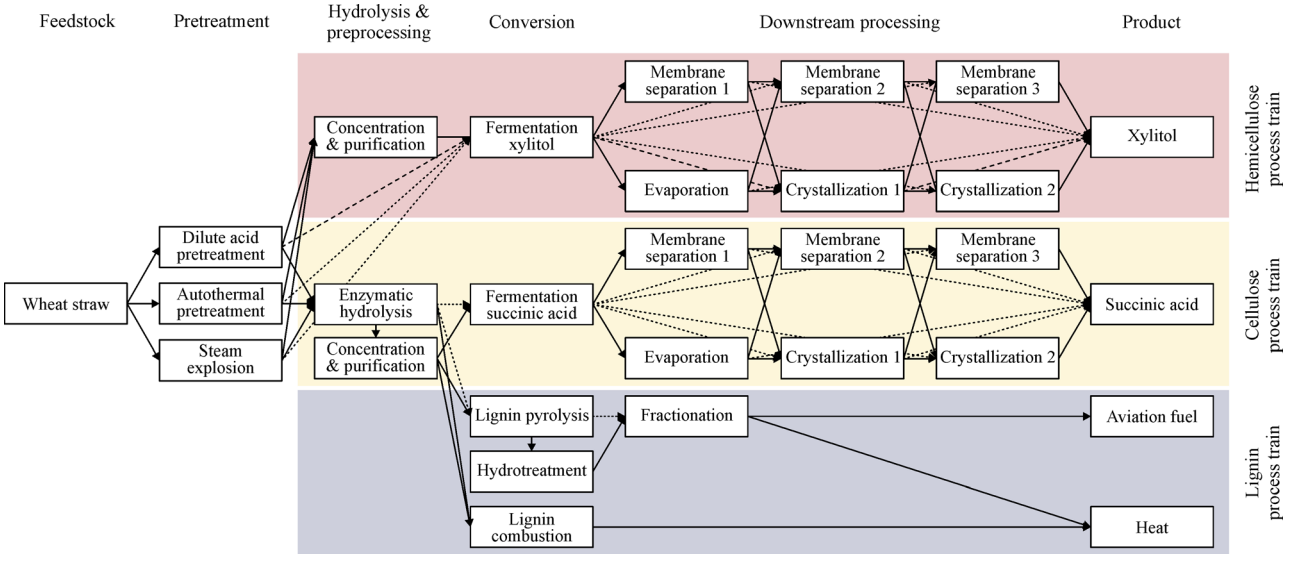

Fig. 3 Illustration of the entire initial bottom-up composed superstructure for the base-case process design of the introduced case study with a hemicellulose, a cellulose, and a lignin process train; the reduced superstructure which will serve as the base case in this study is marked in bold.

unit operated as batch fermentation (FX), an evaporation unit (EX), a first crystallization unit operated as cooling crystallization (CX1) and a second crystallization unit operated as antisolvent crystallization (CX2). Out of these six unit operations, three are compulsory (PT, FX, CX1), and three are optional (UCH, EX, CX2), which results in eight binary decisions or eight different flowsheet alternatives. These are listed in Table 1 together with their configuration ID (cID), which will be used as an identifier throughout the section.

Table 1 Overview of all flowsheet options with their respective cID and the units composing the flowsheet.

cID	Flowsheet
1	PT-UCH-FX-EX-CX1-CX2
2	PT-UCH-FX-EX-CX1
3	PT-UCH-FX-CX1-CX2
4	PT-UCH-FX-CX1
5	PT-FX-EX-CX1-CX2
6	PT-FX-EX-CX1
7	PT-FX-CX1-CX2
8	PT-FX-CX1

With eight flowsheet options as listed in Table 1, the following step 2 of the framework could also be solved by a purely enumeration-based approach instead of formulating an SSO problem. However, to rigorously search through the entire design space for globally optimal solutions and to better account for nontrivial design decisions while accelerating calculation times, the SSO approach is the favorable option. The small problem size is chosen as mentioned to enhance the tractability in the scope of this manuscript.

For all the unit operations, mechanistic models are developed: the PT model is set up as described in section A.1.1 (cf. ESM) while considering the same components as Prunescu et al. [80]. The data for the calibration of the model derives from proprietary experiments. The operational variables considered in the pretreatment model are the pretreatment temperature T_{PT} in °C, the pretreatment time t_{PT} in min and the acid concentration *acid* in wt-% to the original biomass. The fermentation model (FX) is set up as described in section A.1.2 (cf. ESM), while using data from batch experiments from Tochampa et al. [81]. The considered operational variables for the fermentation model are the fermentation time t_{FX} in h as well as the inoculum concentration *inoc* in $g \cdot g^{-1}$ *broth*. The evaporation model (UCH, EX) is set up as described in section A.1.3 (cf. ESM) in the ASPEN Plus process simulation software while using the DIPPR property database and the NRTL equation of state. The considered operational variable is the vapor fraction for either evaporation v_{EX} or upconcentration v_{UCH}. The crystallization model (CX1, CX2) is set up as described in section A.1.4 (cf. ESM) and subsequently validated and calibrated with proprietary experimental data. The considered operational variables are the crystallization time $t_{CX1,2}$ in h, the flowrate of coolant for CX1 $F_{C,CX1}$ in $kg \cdot s^{-1}$ and the flowrate of antisolvent for CX2 $F_{AS,CX2}$ in $kg \cdot s^{-1}$ as well as the cooling temperature $T_{C,CX1}$ in °C for CX1. The considered model outputs for this study are the mass of produced xylitol M_{Xyo} in kg, the concentration of the inhibitory compounds 5-hydroxymethylfurfural and acetic acid in the final process stage C_{5HMF},C_{Aac} in $g \cdot L^{-1}$ as well as a CO_2 ratio φ in $kg \cdot kg^{-1}$ indicating how much CO_2 is produced per kilogram of xylitol by the generation of steam to provide heat in the pretreatment and the evaporation units. The considered uncertainty in the third step of the framework is the composition of the feedstock. All models are implemented in MATLAB. The evaporation model is interfaced with a COM interface to MATLAB. All models are assessed regarding their robustness by means of Monte Carlo-based uncertainty and sensitivity analysis as described in sections A.2.1 and A.2.2 (cf. ESM). Hence, the full design space for the case study is set up by factorial selection in SPDlab, while excluding infeasible options a priori. All model implementations are available through the S3O GitHub repository [82].

3.2 SSO

3.2.1 Flowsheet sensitivity analysis

In order to perform the design space exploration, all operational variables for each flowsheet are considered as input variables. With the functionalities of SPDlab, $N = 2000$ samples with LHS sampling for each flowsheet (cIDs 1–8) are simulated [83]. The flowsheet samples are used as input for the flowsheet sensitivity analysis by using the easyGSA toolbox in MATLAB [42] to determine which of the operational variables are most important regarding the model output and should hence be considered as variables in the optimization problem. The design space exploration results are illustrated in Fig. 4 in violin plots for each model output for each relevant flowsheet option. The results of the ANN-assisted flowsheet sensitivity analysis are illustrated in a heatmap in Fig. 5. All scripts and implementations regarding the flowsheet sensitivity analysis are available through the S3O GitHub repository [82].

As a first major result, the flowsheets with the cIDs 3, 4, 7, and 8 turn out to be infeasible with the given design space, as no set of input variables results in produced xylitol. Furthermore, it becomes evident that there are major differences between the flowsheets' design spaces with cIDs 1, 2, 5, and 6. Where cID 6 seems to have an evenly distributed number of points regarding the mass of xylitol produced, for cID 6, most of the sets of input variables turn out to be infeasible with only a small feasible fraction. Regarding the inhibitory compounds, especially

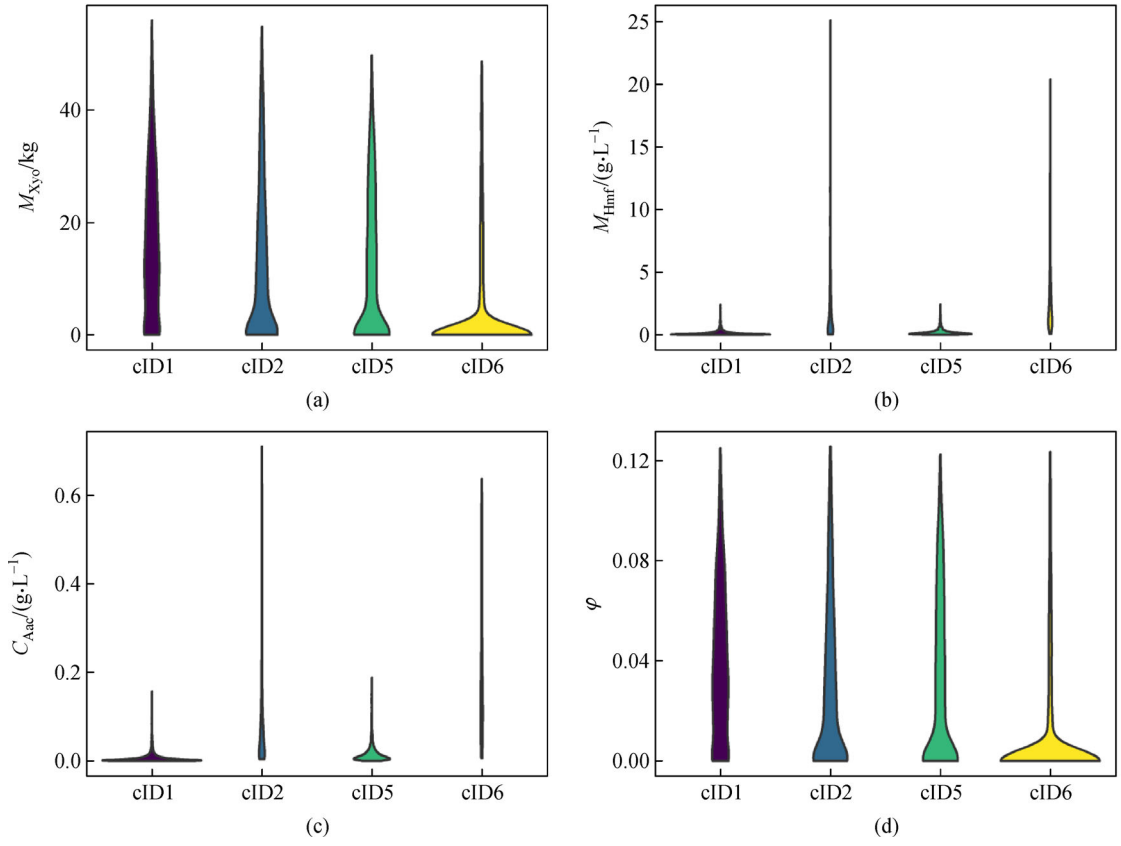

Fig. 4 Violin plots of the results from the design space exploration of flowsheets (a) cID 1, (b) cID 2, (c) cID 5, and (d) cID 6 with the outputs: the mass of produced xylitol (upper left), the concentration of 5-hydroymethylfurfural in the final stage (upper right), the concentration of acetic acid in the final stage (lower left) and the CO_2 ratio (lower right).

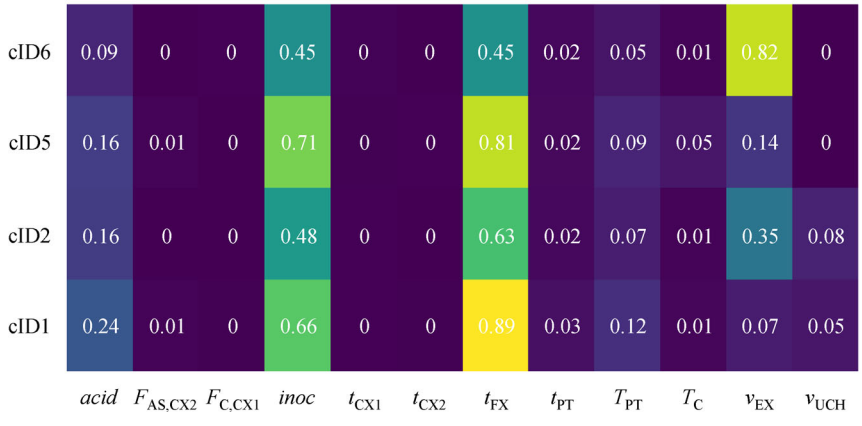

Fig. 5 Heatmap of the total sensitivity indices (S_{Ti}) calculated with the easyGSA toolbox by using ANN surrogates for all flowsheet options and all operational variables.

cIDs 2 and 6, as well as cID 5 to minor extents, seem to produce significant amounts of inhibitors. This allows for the conclusion that both inhibitory compounds should be constrained to a certain level in the optimization problem formulation to ensure product quality.

Regarding the flowsheet sensitivity analysis, a more uniform picture emerges. For all four flowsheets, similar operational variables seem vital despite a different order, and also similar variables are insensitive concerning the output. Hence, for cIDs 1, 5, and 6, we define T_{PT}, *acid*, *inoc*, t_{FX} and v_{EX} and for cID 2, we define *acid*, *inoc*, t_{FX}, v_{EX} and v_{UCH} as the variables to be considered in the optimization problem. All simulations consider an initial amount of lignocellulosic biomass of $M = 1000$ kg.

In conclusion, the design space exploration helped reduce the initial search problem by 50%, and the flowsheet sensitivity analysis can be used as a tool for identifying crucial variables for SSO.

3.2.2 Surrogate model performance assessment

For the surrogate model development and validation, N = 500 and N = 1000 flowsheet simulations with LHS sampling are performed to compare their performance with a different number of sampling points. The flowsheets simulations are again performed with the SPDlab functionalities. All four surrogate model types are fitted and cross-validated with these sampling points, as explained in section 2.1. The ALAMO surrogates are fitted within the ALAMO software. For cross-validation, 20% of the input samples are split off, and ALAMO is given this fraction as a validation dataset to calculate validation metrics internally. The DTR surrogates are all created with the Delaunay functionality in Python's scipy package, utilizing the QHull algorithm [41]. For their cross-validation, due to the model's inexistent extrapolative capabilities, the boundary points, which form the convex hull of the design space, are added to the sampling points. By random selection, a fraction of 20% of the sample points is split off to calculate the validation metrics for the training and testing dataset. This procedure is repeated ten times to calculate the range of the cross-validation metrics.

The GPR surrogates are fitted in MATLAB while utilizing the statistics and machine learning toolbox. For each dataset and each output, an internal routine to optimize hyperparameters, e.g., the kernel functions in MATLAB, is used with default settings to fit the GPR model to the data. Subsequently, a 5-fold cross-validation, as described in section 2.1, is performed to obtain the validation metrics. The ANN surrogate is equally fitted in MATLAB while using the deep learning toolbox. For each dataset and each output, a grid-search algorithm as implemented in the easyGSA toolbox is employed to optimize several parameters, e.g., the number of nodes in the hidden layer to fit the ANN model to the data. Afterward, also a 5-fold cross-validation is performed in order to obtain the validation metrics. All scripts and implementations regarding the surrogate performance analysis are available through the S3O GitHub repository [82].

The calculated validation metrics are the coefficient of determination R^2 and the root mean square error (RMSE) for the whole dataset, as well as the testing and the training dataset, dependent on the applied validation method. Figure 6 shows the parity plots for both N = 500 and N = 1000 flowsheet samples for each surrogate model predicting the amount of produced xylitol. For the ALAMO surrogate models, the parity plots show a relatively high variance in the model and the tendency to underpredict higher values for the amount of xylitol. Secondly, the GPR surrogates show almost a perfect fit, apart from cID 6, where the accuracy is slightly lower than for the other three flowsheets. Lastly, the ANN surrogates show equally good fits to the data, however, with higher variances than the GPR models and a decrease in the prediction quality for cID 6. This can be attributed to the low number of feasible sampling points in the design space of cID 6, which is diametric to both surrogate model methodologies as they rely heavily on the amount of provided input data. The DTR surrogates are not shown as parity plots as the simulation points become an inherent part of the model, which makes the coefficient of determination $R^2 = 1$ by definition and a parity plot thus dispensable.

For all illustrated models in Fig. 6, the cross-validation metrics for the full, the testing, and the training data set for flowsheet cID 1 for the output variable being the amount of produced xylitol are listed in Table 2.

The cross-validation metrics for all other flowsheet cIDs and the other output variables are listed in the supporting material. For the ALAMO surrogate models, the impression from the parity plots is confirmed by average values for R^2 between 0.6 and 0.8 and RMSE values, which are significantly high. As described earlier, the R^2 and RMSE values for the DTR surrogate model are immanently 0 or 1 for the full and the training data set, but for the testing dataset, it becomes obvious that the quality of fit for unseen data is insufficient, expressed by R^2 values around 0.6 and

Table 2 Cross-validation metrics of all surrogate models for flowsheet option cID 1 for both N = 500 and N = 1000 samples for the output variable being the amount of produced xylitol

Model	ALAMO		DTR		GPR		ANN	
	N = 500	N = 1000	N = 500	N = 1000	N = 500	N = 1000	N = 500	N = 1000
R2	0.822	0.765	1	1	1	1	0.997	0.994
RMSE	5.27	6.29	0	0	0.007	0.017	0.597	0.922
R2train	0.817	0.762	1	1	0.997	1	0.997	0.994
R2test	0.722	0.724	0.487	0.642	0.933	0.952	0.895	0.956
RMSEtrain	5.35	6.31	0	0	0.423	0.121	0.674	0.944
RMSEtest	6.54	6.99	8.802	7.677	2.945	2.66	4.002	2.535

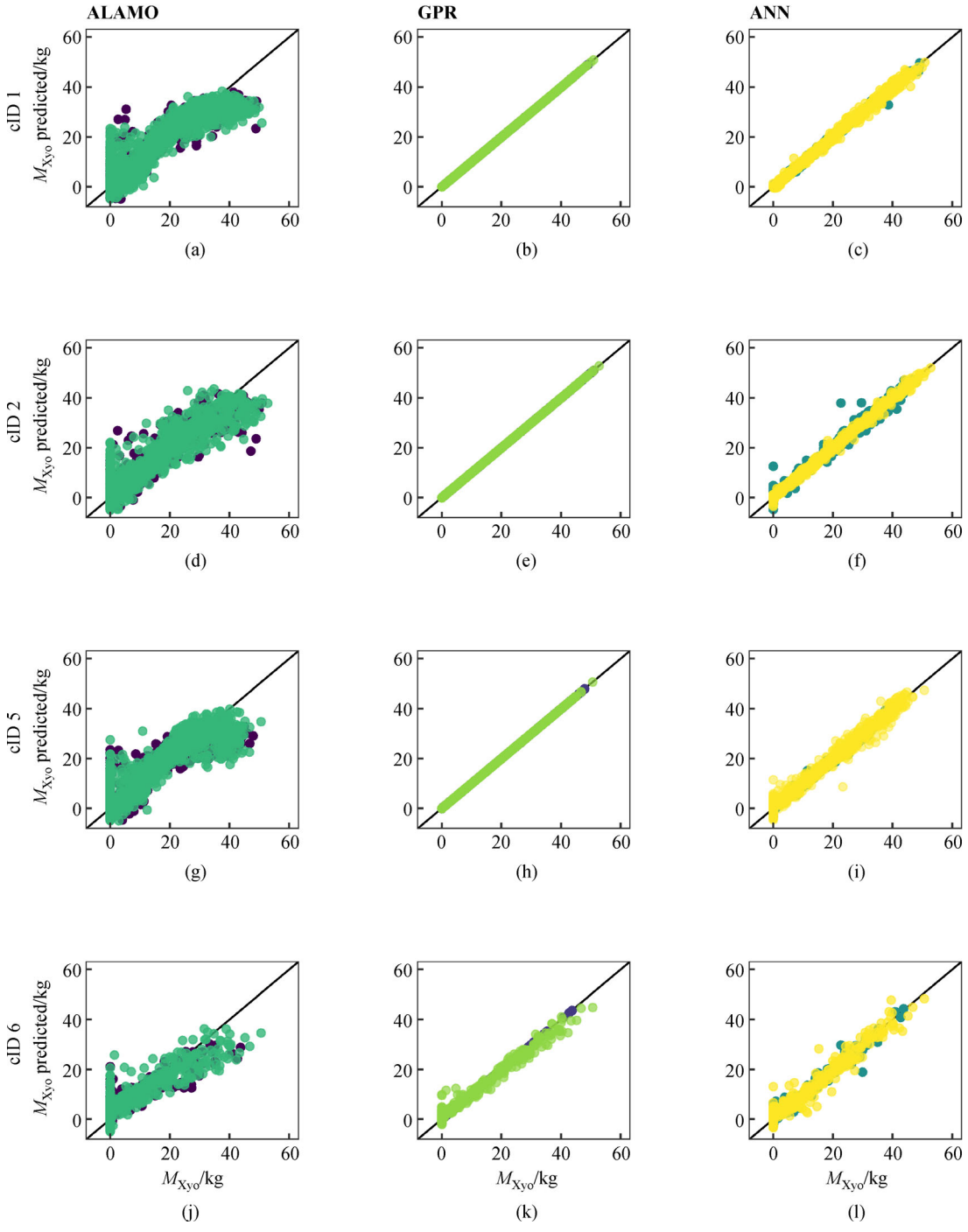

Fig. 6 Parity plots of the ALAMO, the GPR, and the ANN surrogate models for all flowsheets (ALAMO: (a) cID 1, (d) cID 2, (g) cID 5, and (j) cID 6; GPR: (b) cID 1, (e) cID 2, (h) cID 5, and (k) cID 6; ANN: (c) cID 1, (f) cID 2, (i) cID 5, and (l) cID 6) indicating the predicted outputs over the simulated outputs for $N = 500$ (dark blue, blue, turquoise) and $N = 1000$ samples (green, bright green, yellow).

RMSE values up to almost 10. Equally, for the GPR and the ANN surrogate, the validation metrics confirm the parity plots' results; both model types show excellent validation scores even for the testing data sets. As a general trend, it is also to denote that for the DTR, the GPR, and the ANN surrogates, the model quality overall increases with $N = 1000$ instead of $N = 500$ samples, which is in agreement with the described properties in section 2.1. The ALAMO surrogates do not show a consistent improvement of the validation metrics with increasing sample size, which is relatable to the difficulty of fitting a given number of algebraic terms to datasets of increasing size.

Overall, it is to state that both machine learning surrogate models show the best validation metrics and both ALAMO and especially the DTR surrogates reveal insufficient predictive abilities for unseen data points in the test dataset.

3.2.3 SSO results

In order to define the underlying optimization problem to the SSO properly, the following objective function and constraints are introduced:

$$\text{MINLP}: \begin{cases} \max\ M_{\text{Xyo}} = f(x,y) \\ s.t.\ C_{\text{Hmf}} = g_1(x,y) \leqslant 0.5\ \text{g}\cdot\text{L}^{-1} \\ C_{\text{Aac}} = g_2(x,y) \leqslant 0.5\ \text{g}\cdot\text{L}^{-1} \\ \varphi = g_3(x,y) \geqslant 0{,}1 \\ x \in X,\ y \in [0,1] \end{cases} . \quad (28)$$

The operational variables for the solver to choose are *acid*, *inoc*, t_{FX},v_{EX} for all flowsheet cIDs, T_{PT} for flowsheet cIDs 1, 2, and 6 and v_{UCH} for flowsheet cID 2. In order to solve the underlying optimization to each SSO, suitable solvers have to be chosen. For the ALAMO surrogates and the resulting MINLP, the BARON solver is chosen [84]. All algebraic equations for each flowsheet option are instantiated in PYOMO. However, due to the reduced design space of four flowsheet options, the binary variable is removed, and the four flowsheets are solved separately to accelerate the calculation process. This converts the MINLP given as in Eq. (28) into four NLPs; however, the results do not differ as the underlying optimization problem does not change. Hence this simplification is valid. All optimization problems are solved with the default solver settings. If a solution is found, the problem is solved to global optimality; otherwise, the optimization problem is found infeasible.

For the DTR surrogates and the resulting MILP, the Gurobi solver is chosen. The formulation of the problem is equally given with Eq. (28), with f, g_1, g_2 and g_3 being linear functional relationships as given with Eq. (2). Equations (6) through (25) are implemented in PYOMO. Equally for the MILP, the four flowsheets are solved separately, which reduces the dimensionality of the binary variable. However, the resulting simplified MILP yields the same solution; hence the simplification is valid. For N = 500 flowsheet samples the resulting optimization problem results in around 1000000 continuous and 50000 integer variables and double the number for N = 1000 flowsheet samples. All problems are solved to global optimality.

For both the GPR and the ANN surrogates, the MATLAB solver fmincon with the sequential quadratic programming algorithm is chosen. The formulation of the problem is equally given with Eq. (28), with f, g_1, g_2 and g_3 being functional relationships as given with Eq. (3) while excluding the binary variable y. The surrogates are passed to the solver together with a multi-start option, indicating that the optimization problem should be solved for $q = 1000$ times in order to ensure global optimality. For flowsheets where the optimization problem yields a high fraction of all the multi-start solutions converging, the results are within the same magnitude as the other flowsheets' results. With cID 2 and N = 1000 samples and also for cID 6 and the ANN, however, the observed solutions diverge, which is also reflected in the small number of multi-starts converging. All scripts and implementations regarding the SSO are available through the S3O GitHub repository [82].

All results from the SSO for all four flowsheets for N = 500 samples are listed in Table 3 until Table 6. Results from the SSO of flowsheet cID 6 with N = 500 samples with all surrogate models and their respective solvers. All tables indicate the predicted objective function values and constraints (opt) and the results from the corresponding validation simulation with the original flowsheet for the same conditions (val) — according to optimal operational conditions and their lower (lb) and upper (ub) bounds.

For the MINLP problems utilizing ALAMO surrogates and the BARON solver, it becomes clear that the low surrogate quality, expressed by the poor performance metrics, heavily affects the feasibility and the quality of the

Table 3 Results from the SSO of flowsheet cID 1 with N = 500 samples with all surrogate models and their respective solvers.

cID 1-500	ub	lb	ALAMO/BARON		DTR/Gurobi		GPR/fmincon		ANN/fmincon	
			opt	val	opt	val	opt	val	opt	val
T_{PT}	173	195	179.581		184.31		187.74		177.24	
Acid	0.5	2	0.672		1.456		1.337		2.000	
Inoc	0.5	3	3.000		1.523		1.497		1.191	
t_{FX}	8	16	47.938		43.207		42.656		47.727	
v_{EX}	0.99	0.998	0.995		0.996		0.998		0.998	
M_{Xyo}			59.852	0.000	49.094	48.964	54.083	43.410	85.240	45.682
C_{Hmf}		0.5	0.000	0.028	0.058	0.060	0.034	0.006	0.020	0.007
C_{Aac}		0.5	0.002	0.004	0.002	0.002	0.001	0.000	0.001	0.000
φ	0.1		0.140	0.000	0.118	0.117	0.116	0.100	0.100	0.114

Table 4 Results from the SSO of flowsheet cID 2 with $N = 500$ samples with all surrogate models and their respective solvers.

cID 2-500	ub	lb	ALAMO/BARON		DTR/Gurobi		GPR/fmincon		ANN/fmincon	
			opt	val	opt	val	opt	val	opt	val
Acid	0.5	2	0.685		0.762		0.874		2.000	
Inoc	0.5	3	2.998		1.493		0.500		2.736	
t_{FX}	12	48	47.999		46.427		40.493		12.490	
v_{EX}	0.99	0.998	0.996		0.997		0.990		0.998	
v_{UCH}	0.4	0.6	0.512		0.520		0.585		0.423	
M_{Xyo}			53.004	0.000	50.017	51.123	13.315	0.078	4.410	11.938
C_{Hmf}		0.5	0.500	3.935	0.500	0.451	1.830	0.779	1.467	0.136
C_{Aac}		0.5	0.000	0.289	0.022	0.021	0.082	0.048	0.057	0.008
φ	0.1		0.123	0.000	0.117	0.120	0.028	0.000	0.037	0.028

Table 5 Results from the SSO of flowsheet cID 5 with $N = 500$ samples with all surrogate models and their respective solvers.

cID 5-500	ub	lb	ALAMO/BARON		DTR/Gurobi		GPR/fmincon		ANN/fmincon	
			opt	val	opt	val	opt	val	opt	val
T_{PT}	173	195			193.69		185.50		186.6	
Acid	0.5	2			0.776		1.188		1.127	
inoc	0.5	3			2.315		0.963		0.782	
t_{FX}	8	16			28.415		45.079		48.000	
v_{EX}	0.99	0.998			0.993		0.998		0.998	
M_{Xyo}			Infeasible		47.915	47.86	54.829	48.12	67.400	46.86
C_{Hmf}		0.5			0.152	0.152	0.057	0.038	0.044	0.022
C_{Aac}		0.5			0.012	0.126	0.003	0.002	0.002	0.001
φ	0.1				0.118	0.118	0.132	0.123	0.168	0.117

Table 6 Results from the SSO of flowsheet cID 6 with $N = 500$ samples with all surrogate models and their respective solvers.

cID 6-500	ub	lb	ALAMO/BARON		DTR/Gurobi		GPR/fmincon		ANN/fmincon	
			opt	val	opt	val	opt	val	opt	val
T_{PT}	173	195			184.00		184.00		191.04	
Acid	0.5	2			0.960		1.265		1.609	
inoc	0.5	3			2.536		2.507		0.976	
t_{FX}	8	16			23.221		22.465		45.041	
v_{EX}	0.99	0.998			0.998		0.998		0.997	
M_{Xyo}			Infeasible		43.688	43.95	47.727	43.58	0.016	26.35
C_{Hmf}		0.5			0.373	0.347	0.500	0.322	11.321	4.627
C_{Aac}		0.5			0.022	0.021	0.014	0.018	0.243	0.134
φ	0.1				0.112	0.112	0.101	0.111	0.000	0.066

optimization results. For this option, none of the obtained solutions were satisfactory. However, for the MILP problems utilizing the DTR surrogates, all solutions show a very close agreement between the predicted solutions and their validations, despite the insufficient validation metrics. Furthermore, the problem is solved for all flowsheets for all sample sizes. However, it becomes clear that for all flowsheets, the theoretical underlying global optimum is not reached. This is attributable to the number of sampling points provided as the DTR surrogates are strictly interpolative and can only predict according to the provided input samples.

For the GPR and the ANN surrogates, for $N = 500$ samples, the results are mostly not in agreement with the validation simulations despite the good validation metrics. For $N = 1000$ samples, however, the results from the optimization improve and converge with the validation results. This is explained by the general trend of machine learning models to show an improved prediction when provided with larger amounts of training data. Overall it shows that the GPR-assisted NLP predicts more consistency for different flowsheets and sampling sizes. In contrast to that, the ANN assisted NLP predicts more inconsistently, depending on the sampling size and flowsheet, but with successful predictions, the predicted values for the objective function are higher than the ones predicted by the GPR and thus closer to what would correspond to the global optimum for the rigorous flowsheet.

The illustration Fig. 7 indicates both the prediction of each surrogate model for each flowsheet and sample size as the center of the circle and the root mean square error of the testing data set of the surrogate model as the radius of the circle. Furthermore, the results from the validation simulations are added (cross).

Again, it becomes clear that the most consistent combination of the surrogate model, optimization problem, and solver is the choice of DTR surrogates despite its validation metrics, as overall, the consistency is highest. For both the GPR and the ANN, it is visible that the models predict higher objective function values, but the validation simulations are less in agreement than with the DTR surrogates. For the ALAMO surrogates, it becomes apparent that their performance in the given optimization problem is impaired. This can potentially be attributed to the underprediction surrogate models, as seen in Fig. 6, which do not allow for an optimal solution under the given constraints.

Overall, it is to point out that after analyzing the quality of the surrogate models and the results from the SSO, the indication regarding the quality between the different surrogate models is ambiguous. However, regarding the underlying case study, it becomes apparent that flowsheet cID 1 shows the best objective function values for both sample sizes; hence it should be subjected to investigation in the third step of the framework. Both cID 2 and cID 5

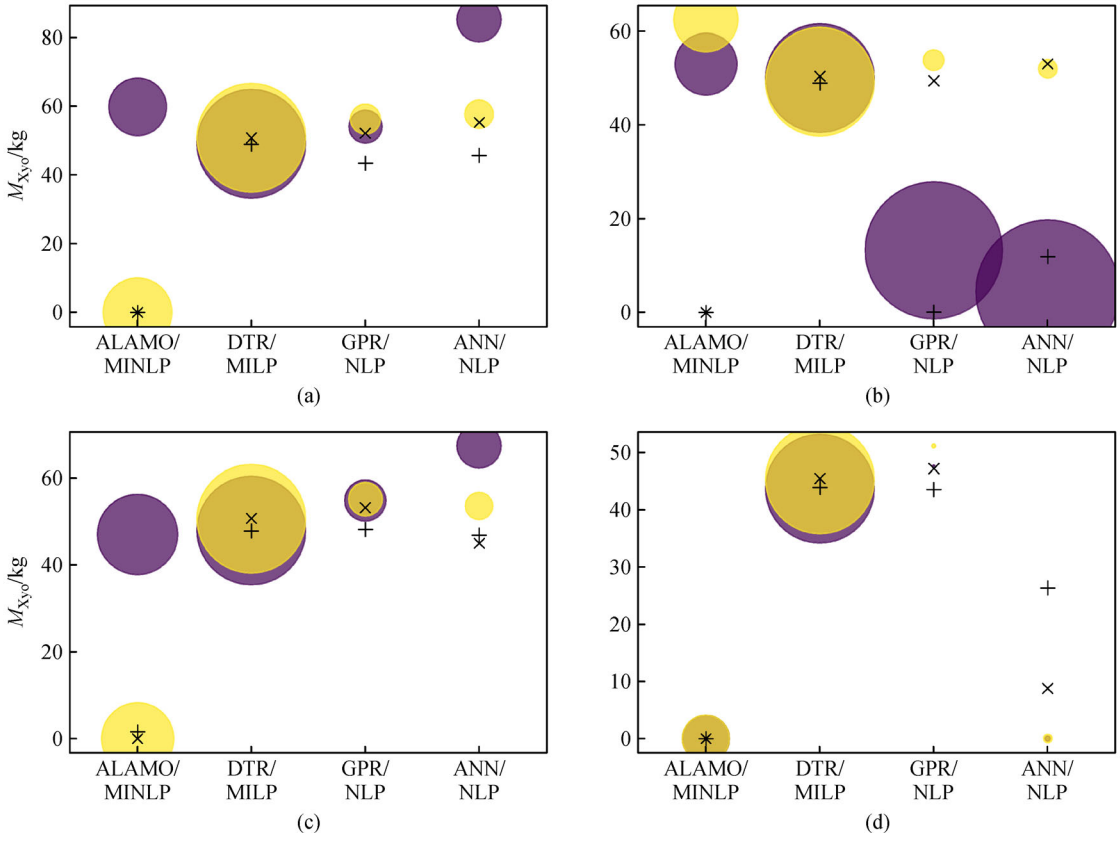

Fig. 7 Bubble plot for the visualization of the consistency metrics of the different superstructure modeling approaches, the center of each sphere indicating the predicted value in the optimization problem, the radius of the sphere being the RMSE of the testing dataset in the cross-validation, and the cross/saltire indicating the respective validation simulation for (a) cID 1, (b) cID 2, (c) cID 5 and (d) cID 6 for respectively $N = 500$ samples (blue, cross) and $N = 1000$ samples (yellow, saltire).

show very similar objective function values and constraint values, which is why both should be equally subjected to investigation in the third step of the framework.

3.3 SBO results

All flowsheets considered candidate process topologies from step two of the framework are now subjected to SBO using the MOSKopt solver [27], utilizing SK surrogate models. As uncertain input, the wheat straw composition is chosen to vary around 5% by the nominal value. The composition of the feedstock is highly dependent on climate effects as well as geological conditions of the fields, amongst others, which lead to varying compositions [85]. Logically, a varying feedstock composition influences the product yields, explaining the importance of critically assessing this uncertainty. The initially assumed composition is 31.3% hemicellulose, 42.7% cellulose, and the residual as lignin; the uncertainty is supposed to be uniformly distributed. For the SBO, twenty-five initial samples and each sample with 100 Monte Carlo samples are chosen. The optimization criterion for the underlying Bayesian Optimization is the multi-constraint FEI, as explained in Al et al.'s work [27]. The solver hedges against uncertainty with the simulation's mean values and performs 75 iterations, which results in 100 total calculation steps. The starting points are chosen to be the optimal results for each flowsheet from step two of the framework. All scripts and implementations regarding the SBO are available through the S3O GitHub repository [82]. The results from the three SBO runs are listed in the following Table 7 (opt) together with the corresponding validation simulations (val).

Firstly, the results from the SSO are confirmed, referring to that cID 1 appears to be the optimal flowsheet in the given design space, including a pretreatment unit, an upconcentration unit, a fermentation unit, an evaporation unit, and two crystallization units. This is also the flowsheet option with the highest number of possible unit operations to maximize the product yield. Both cID 2 and cID 5 seem to perform equally well, despite their difference being once performing an upconcentration before the fermentation and once performing two crystallization steps in the downstream processing. Overall, the yield for all flowsheet options is comparatively low. This indicates the possible use of alternative unit operations and operation modes in the initial superstructure to increase the amount of final product.

Secondly, it is noteworthy that for all the three flowsheets, the predicted optimum for each respectively, despite being subjected to uncertainty in the feedstock, is globally higher than by any of the introduced SSO alternatives. Figure 8 indicates for each cID how much the objective function value improved quantitatively over the 100 iterations when providing the operational conditions found to be optimal in step 2.

In conclusion, the performed SBO based on the SSO results from step 2 appears to be an excellent combinatorial solution leveraging synergistic effects between screening a multitude of alternatives and thoroughly designing few alternatives.

4 Conclusions and future research

4.1 Conclusions

In this paper, we presented a framework to expedite the conceptual process design of novel bioprocesses in biorefinery setups by leveraging synergistic effects from applying expert knowledge and combining SSO and SBO in a hybrid manner. In this process, four different SSO alternatives are investigated and benchmarked. The proposed bottom-up approach to compose the initial superstructure by beginning with the end in mind and first selecting a product set, a feedstock, and subsequently,

Table 7 Results from the SBO for all candidate process topologies with the MOSKopt solver, using 25 initial sampling points, 75 iterations, the mean value as hedge against uncertainty, and the multi-constraint FEI criterion

Item	ub	lb	cID 1		cID 2		cID 5	
			opt	val	opt	val	opt	val
T_{PT}	173	195	195.000				195	
Acid	0.5	2	0.715		0.984		0.879	
Inoc	0.5	3	1.611		3.000		3	
t_{FX}	8	16	44.994		30.367		24.271	
v_{EX}	0.99	0.998	0.997		0.998		0.998	
v_{UCH}	0.4	0.6			0.400			
M_{Xyo}			56.310	56.736	53.760	54.224	53.96	54.17
C_{Hmf}		0.5	0.045	0.032	0.492	0.471	0.062	0.061
C_{Aac}		0.5	0.001	0.001	0.019	0.020	0.002	0.002
φ	0.1		0.119	0.125	0.127	0.129	0.128	0.128

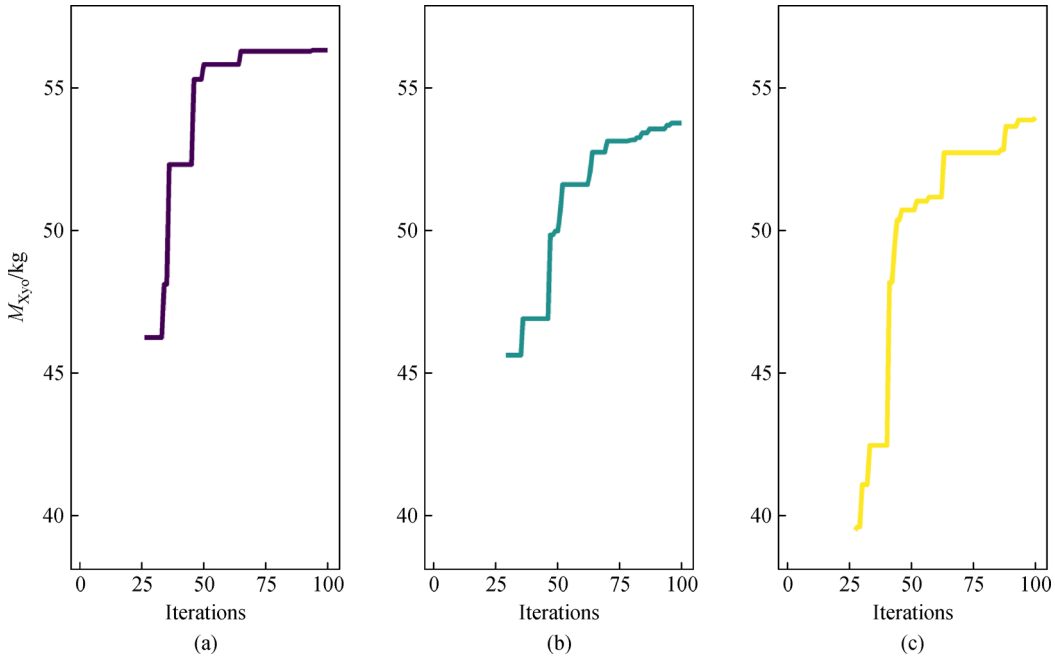

Fig. 8 Visualization of the improvement of the objective function value over the iterations in the SBO with the MOSKopt solver for flowsheet (a) cID 1, (b) cID 2, and (c) cID 5.

the processing units reduce the initial maximum size of the superstructure heavily.

The second step of the proposed framework is based on using SSO to solve the process synthesis problem underlying the superstructure and determine several candidate process topologies. Four different alternatives of solving the optimization problem by surrogate model-assisted SSO are proposed: the use of ALAMO surrogate models in an MINLP, the use of DTR surrogate models in an MILP, and the use of either GPR or ANN surrogate models in a series of NLP with a multi-start solution strategy. The latter three options serve the purpose of predicting several candidate process topologies. A thorough benchmark of all four alternatives reveals that all surrogate models have different shortcomings when applied in the framework. A general property of surrogate models is the shortcoming in terms of adhering to mass balance restrictions. Furthermore, they do not rely on physical constraints but rather on an alternative mathematical formulation and the quantity of provided data to fit the behavior of the underlying high-fidelity model accurately.

In consequence, the predictions of surrogate models inherently deviate to a certain extent. Hence, a combined validation of both surrogate performance and optimization results is essential to benchmark the proposed alternatives. In other words, the mere selection of a surrogate model based on convincing cross-validation metrics is not a guarantee to obtain a consistent and qualitatively good result from the resulting optimization problem in which the surrogate model is used. We have empirically presented in this study that cross-validation metrics alone, i.e., the coefficient of determination and the root mean square error, are solitarily not a good indicator of the surrogate model's quality in the subsequent optimization application. For example, the GPR and ANN surrogates suffer from several shortcomings in the optimization problem, whereas their validation metrics are superior. Opposed to this, the DTR surrogates' validation metrics are inferior compared to the GPR and the ANN, but the optimization results are consistent and qualitatively better. Hence, we recommend a critical assessment of the quality of a chosen surrogate model and its validation in the respective optimization problem. We also point out that the combination of DTR surrogates in an MILP has shown promising results in using the proposed framework in this particular case study.

Regarding the size of the resulting superstructure formulations for larger design cases, it is to state that also the computational burden will consequently increase but can be alleviated in several ways. For example, the necessary flowsheet simulations can be expedited by using both a parallelization approach to evaluate bigger numbers of flowsheet options and utilizing cloud or cluster computing solutions to accelerate the speed of a single simulation. Furthermore, by decreasing the initial size of the superstructure as in step 1 of the framework and solely focusing on several candidate process topologies in step 3, the overall computational burden is reduced optimally. It can also facilitate the solution of larger problems than the presented one.

Lastly, the third step of the framework involves the SBO of all candidate process topologies from step 2. The used solver MOSKopt can improve the objective function value

consistently for all subjected candidates under a given uncertainty scenario and yields a consolidated optimal process design as the final result of the framework.

Overall, it remains to conclude that our proposed framework was validated successfully. It considerably facilitates and accelerates the conceptual process design of novel bioprocesses in biorefinery setups, as shown with an example of a xylitol production process in a biorefinery.

4.2 Future research

Despite the impaired validation metrics, the proposed framework with the DTR surrogate models performed best in the optimization task and is hence the suggested alternative in this case. However, to improve the validation metrics and improve the predictions in terms of global optimality, a first option would be to base the framework on adaptive sampling strategies instead of a static sampling strategy as Latin hypercube sampling [86–88]. Alternatively, algorithms that reduce the mesh's original size by iteratively removing edges of the triangulation can be employed [89]. The improvement for the triangulation and subsequently the MILP is a reduced number of simplices by having fewer sampling points in regions, where a linear interpolation with a lower error is possible, and increasing the number of sampling points in regions, where the linear interpolation causes a greater error and improving the prediction in the vicinity of the global optimum. This has been proposed in the literature by Chen et al. for the case of a triangulation model but also as assistance for improving the fit of other machine learning models as GPRs, which can be equally used for this optimization task, as shown in this work [90,91].

Overall, we acknowledge that the presented results here are specific to the selected case study. Therefore, the application in further case studies is necessary in order to consolidate these conclusions further. Indeed, we believe there are concrete empirical observations, which indicate that the validation issue of emerging surrogate-based/machine learning-based optimization approaches needs to be more critically analyzed and assessed to ensure their appropriate use in further process systems engineering applications.

From a more holistic perspective, the use of this framework is not only limited to process design but can equally be applied for the optimization of whole value chains, involving the choice of different feedstocks, plant locations, and logistic constraints. Lastly, especially for the used fermentation models, the black-box kinetics can be replaced by genome-scale metabolic models in order to perform also cell factory optimization, which can contribute to the further expedition of designing bioprocesses and promote the transition towards a bio-based and circular economy as instigated by the 2030 Sustainability Agenda of the United Nations.

Acknowledgements The authors would like to express their gratitude to the Novo Nordisk Foundation (Grant No. NNF17SA0031362) for funding the Fermentation-Based Biomanufacturing Initiative of which this project is a part.

Electronic Supplementary Material Supplementary material is available in the online version of this article at https://dx.doi.org/10.1007/s11705-021-2071-9 and is accessible for authorized users.

References

1. United Nations. Transforming our world: the 2030 agenda for sustainable development, 2015
2. Ubando A T, Felix C B, Chen W H. Biorefineries in circular bioeconomy: a comprehensive review. Bioresource Technology, 2020, 299: 122585
3. Straathof A J J, Wahl S A, Benjamin K R, Takors R, Wierckx N, Noorman H J. Grand research challenges for sustainable industrial biotechnology. Trends in Biotechnology, 2019, 37(10): 1042–1050
4. Hillson N, Caddick M, Cai Y, Carrasco J A, Chang M W, Curach N C, Bell D J, Feuvre R L, Friedman D C, Fu X, et al. Building a global alliance of biofoundries. Nature Communications, 2019, 10 (1): 1038–1041
5. Hassan S S, Williams G A, Jaiswal A K. Lignocellulosic biorefineries in Europe: current state and prospects. Trends in Biotechnology, 2019, 37(3): 231–234
6. Hassan S S, Williams G A, Jaiswal A K. Moving towards the second generation of lignocellulosic biorefineries in the EU: drivers, challenges, and opportunities. Renewable & Sustainable Energy Reviews, 2019, 101: 590–599
7. Moncada B J, Aristizábal M V, Cardona A C A. Design strategies for sustainable biorefineries. Biochemical Engineering Journal, 2016, 116: 122–134
8. Chaturvedi T, Torres A I, Stephanopoulos G, Thomsen M H, Schmidt J E. Developing process designs for biorefineries-definitions, categories, and unit operations. Energies, 2020, 13(6): 1493
9. Kokossis A C, Yang A. On the use of systems technologies and a systematic approach for the synthesis and the design of future biorefineries. Computers & Chemical Engineering, 2010, 34(9): 1397–1405
10. Chemmangattuvalappil N G, Ng D K S, Ng L Y, Ooi J, Chong J W, Eden M R. A review of process systems engineering (PSE) tools for the design of ionic liquids and integrated biorefineries. Processes (Basel, Switzerland), 2020, 8(12): 1–29
11. Tey S Y, Wong S S, Lam J A, Ong N Q X, Foo D C Y, Ng D K S. Extended hierarchical decomposition approach for the synthesis of biorefinery processes. Chemical Engineering Research & Design, 2021, 166: 40–54
12. Clauser N M, Felissia F E, Area M C, Vallejos M E. A framework for the design and analysis of integrated multi-product biorefineries from agricultural and forestry wastes. Renewable & Sustainable Energy Reviews, 2021, 139: 110687
13. Mountraki A D, Benjelloun-Mlayah B, Kokossis A C. A surrogate modeling approach for the development of biorefineries. Frontiers in Chemical Engineering, 2020, 2: 12

14. Pyrgakis K A, Kokossis A C. A total site synthesis approach for the selection, integration and planning of multiple-feedstock biorefineries. Computers & Chemical Engineering, 2019, 122: 326–355
15. Meramo-Hurtado S I, González-Delgado Á D. Biorefinery synthesis and design using sustainability parameters and hierarchical/3D multi-objective optimization. Journal of Cleaner Production, 2019, 240: 118134
16. Galanopoulos C, Giuliano A, Barletta D, Zondervan E. An integrated methodology for the economic and environmental assessment of a biorefinery supply chain. Chemical Engineering Research & Design, 2020, 160: 199–215
17. Ulonska K, König A, Klatt M, Mitsos A, Viell J. Optimization of multiproduct biorefinery processes under consideration of biomass supply chain management and market developments. Industrial & Engineering Chemistry Research, 2018, 57(20): 6980–6991
18. Aristizábal-Marulanda V, Cardona Alzate C A. Methods for designing and assessing biorefineries. Biofuels, Bioproducts & Biorefining, 2019, 13(3): 789–808
19. Meramo-Hurtado S I, González-Delgado Á D. Process synthesis, analysis, and optimization methodologies toward chemical process sustainability. Industrial & Engineering Chemistry Research, 2021, 60(11): 4193–4217
20. Darkwah K, Knutson B L, Seay J R. A Perspective on challenges and prospects for applying process systems engineering tools to fermentation-based biorefineries. ACS Sustainable Chemistry & Engineering, 2018, 6(3): 2829–2844
21. Biegler L T, Grossmann I E, Westerberg A W. Systematic Methods for Chemical Process design. 1st ed. London: Pearson, 1997
22. Yuan Z, Eden M R. Superstructure optimization of integrated fast pyrolysis-gasification for production of liquid fuels and propylene. AIChE Journal. American Institute of Chemical Engineers, 2016, 62 (9): 3155–3176
23. Chen Q, Grossmann I E. Recent developments and challenges in optimization-based process synthesis. Annual Review of Chemical and Biomolecular Engineering, 2017, 8(1): 249–283
24. Grossmann I E, Apap R M, Calfa B A, García-Herreros P, Zhang Q. Recent advances in mathematical programming techniques for the optimization of process systems under uncertainty. Computers & Chemical Engineering, 2016, 91: 3–14
25. Koutinas M, Kiparissides A, Pistikopoulos E N, Mantalaris A. Bioprocess systems engineering: transferring traditional process engineering principles to industrial biotechnology. Computational and Structural Biotechnology Journal, 2012, 3(4): e201210022
26. Bhosekar A, Ierapetritou M. Advances in surrogate based modeling, feasibility analysis, and optimization: a review. Computers & Chemical Engineering, 2018, 108: 250–267
27. Al R, Behera C R, Gernaey K V, Sin G. Stochastic simulation-based superstructure optimization framework for process synthesis and design under uncertainty. Computers & Chemical Engineering, 2020, 143: 107118
28. Wang Z, Ierapetritou M. Constrained optimization of black-box stochastic systems using a novel feasibility enhanced kriging-based method. Computers & Chemical Engineering, 2018, 118: 210–223
29. McBride K, Sundmacher K. Overview of surrogate modeling in chemical process engineering. Chemieingenieurtechnik (Weinheim), 2019, 91(3): 228–239
30. Friedman M. Multivariate adaptive regression splines. Annals of Statistics, 1991, 19(1): 1–67
31. Sudret B. Global sensitivity analysis using polynomial chaos expansions. Reliability Engineering & System Safety, 2008, 93(7): 964–979
32. Williams B A, Cremaschi S. Surrogate model selection for design space approximation and surrogatebased optimization. Computer-Aided Chemical Engineering, 2019, 47: 353–358
33. Janssen H. Monte-Carlo based uncertainty analysis: sampling efficiency and sampling convergence. Reliability Engineering & System Safety, 2013, 109: 123–132
34. Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2nd ed. New York: Springer, 2009
35. Wilson Z T, Sahinidis N V. The ALAMO approach to machine learning. Computers & Chemical Engineering, 2017, 106: 785–795
36. Cozad A, Sahinidis N V, Miller D C. Learning surrogate models for simulation-based optimization. AIChE Journal. American Institute of Chemical Engineers, 2014, 60(6): 2211–2227
37. Eslick J C, Ng B, Gao Q, Tong C H, Sahinidis N V, Miller D C. A framework for optimization and quantification of uncertainty and sensitivity for developing carbon capture systems. Energy Procedia, 2014, 63: 1055–1063
38. Miller D C, Siirola J D, Agarwal D, Burgard A P, Lee A, Eslick J C, Nicholson B, Laird C, Biegler L T, Bhattacharyya D, Sahinidis N V, Grossmann I E, Gounaris C E, Gunter D. Next generation multi-scale process systems engineering framework. Computer-Aided Chemical Engineering, 2018, 44: 2209–2214
39. Delaunay B. On the empty sphere. Journal of Physics and Radium. 1934, 12(7): 793–800 (in French)
40. Žalik B. An efficient sweep-line Delaunay triangulation algorithm. CAD Computer Aided Design, 2005, 37(10): 1027–1038
41. Barber C B, Dobkin D P, Huhdanpaa H. The quickhull algorithm for convex hulls. ACM Transactions on Mathematical Software, 1996, 22(4): 469–483
42. Al R, Behera C R, Zubov A, Gernaey K V, Sin G. Meta-modeling based efficient global sensitivity analysis for wastewater treatment plants—an application to the BSM2 model. Computers & Chemical Engineering, 2019, 127: 233–246
43. Rasmussen C E. Gaussian processes in machine learning. In: Bousquet O, von Luxburg U, Rätsch G, eds. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). Berlin, Heidelberg: Springer Verlag, 2004, 63–71
44. Boukouvala F, Ierapetritou M G. Feasibility analysis of black-box processes using an adaptive sampling kriging-based method. Computers & Chemical Engineering, 2012, 36(1): 358–368
45. Caballero J A, Grossmann I E. An algorithm for the use of surrogate models in modular flowsheet optimization. AIChE Journal. American Institute of Chemical Engineers, 2008, 54(10): 2633–2650
46. Davis E, Ierapetritou M. A kriging based method for the solution of mixed-integer nonlinear programs containing black-box functions. Journal of Global Optimization, 2009, 43(2-3): 191–205
47. Hwangbo S, Al R, Sin G. An integrated framework for plant data-driven process modeling using deep-learning with Monte-Carlo

simulations. Computers & Chemical Engineering, 2020, 143: 107071
48. Schweidtmann A M, Mitsos A. Deterministic global optimization with artificial neural networks embedded. Journal of Optimization Theory and Applications, 2019, 180(3): 925–948
49. Henao C A, Maravelias C T. Surrogate-based superstructure optimization framework. AIChE Journal. American Institute of Chemical Engineers, 2011, 57(5): 1216–1232
50. Yeomans H, Grossmann I E. A systematic modeling framework of superstructure optimization in process synthesis. Computers & Chemical Engineering, 1999, 23(6): 709–731
51. Mencarelli L, Chen Q, Pagot A, Grossmann I E. A review on superstructure optimization approaches in process system engineering. Computers & Chemical Engineering, 2020, 136: 106808
52. Huster W R, Schweidtmann A M, Lüthje J T, Mitsos A. Deterministic global superstructure-based optimization of an organic Rankine cycle. Computers & Chemical Engineering, 2020, 141: 106996
53. Jones M, Forero-Hernandez H, Zubov A, Sarup B, Sin G. Superstructure optimization of oleochemical processes with surrogate models. Computer-Aided Chemical Engineering, 2018, 44: 277–282
54. Misener R, Floudas C A. Piecewise-linear approximations of multidimensional functions. Journal of Optimization Theory and Applications, 2010, 145(1): 120–147
55. Misener R, Gounaris C E, Floudas C A. Global optimization of gas lifting operations: a comparative study of piecewise linear formulations. Industrial & Engineering Chemistry Research, 2009, 48(13): 6098–6104
56. Pistikopoulos E N. Uncertainty in process design and operations. Computers & Chemical Engineering, 1995, 19(Suppl 1): 553–563
57. Amaran S, Sahinidis N V, Sharda B, Bury S J. Simulation optimization: a review of algorithms and applications. 4OR, 2014, 12(4): 301–333
58. Fu M C, Price C C, Zhu J, Hillier F S. Handbook of Simulation Optimization Associate Series Editor. New York: Springer, 2015
59. Ankenman B, Nelson B L, Staum J. Stochastic kriging for simulation metamodeling. Operations Research, 2010, 58(2): 371–382
60. Bertsimas D, Sim M. The price of robustness. Operations Research, 2004, 52(1): 35–53
61. Ning C, You F. Optimization under uncertainty in the era of big data and deep learning: when machine learning meets mathematical programming. Computers & Chemical Engineering, 2019, 125: 434–448
62. Hüllen G, Zhai J, Kim S H, Sinha A, Realff M J, Boukouvala F. Managing uncertainty in data-driven simulation-based optimization. Computers & Chemical Engineering, 2020, 136: 106519
63. Marques C M, Moniz S, de Sousa J P, Barbosa-Póvoa A P. A simulation-optimization approach to integrate process design and planning decisions under technical and market uncertainties: a case from the chemical-pharmaceutical industry. Computers & Chemical Engineering, 2017, 106: 796–813
64. Crater J S, Lievense J C. Scale-up of industrial microbial processes. FEMS Microbiology Letters, 2018, 365(13): 138
65. Noorman H J, Heijnen J J. Biochemical engineering's grand adventure. Chemical Engineering Science, 2017, 170: 677–693
66. Da Silva S S, Chandel A K. D-Xylitol: Fermentative Production, Application and Commercialization. Berlin Heidelberg: Springer-Verlag, 2012
67. Choi S, Song C W, Shin J H, Lee S Y. Biorefineries for the production of top building block chemicals and their derivatives. Metabolic Engineering, 2015, 28: 223–239
68. de Albuquerque T L, da Silva I J, de MacEdo G R, Rocha M V P. Biotechnological production of xylitol from lignocellulosic wastes: a review. Process Biochemistry, 2014, 49(11): 1779–1789
69. Venkateswar Rao L, Goli J K, Gentela J, Koti S. Bioconversion of lignocellulosic biomass to xylitol: an overview. Bioresource Technology, 2016, 213: 299–310
70. Dasgupta D, Bandhu S, Adhikari D K, Ghosh D. Challenges and prospects of xylitol production with whole cell bio-catalysis: a review. Microbiological Research, 2017, 197: 9–21
71. Felipe Hernández-Pérez A, de Arruda P V, Sene L, da Silva S S, Kumar Chandel A, de Almeida Felipe M G. Xylitol bioproduction: state-of-the-art, industrial paradigm shift, and opportunities for integrated biorefineries. Critical Reviews in Biotechnology, 2019, 39(7): 924–943
72. Delgado Arcaño Y, Valmaña García O D, Mandelli D, Carvalho W A, Magalhães Pontes L A. Xylitol: a review on the progress and challenges of its production by chemical route. Catalysis Today, 2020, 344: 2–14
73. Mountraki A D, Koutsospyros K R, Mlayah B B, Kokossis A C. Selection of biorefinery routes: the case of xylitol and its integration with an organosolv process. Waste and Biomass Valorization, 2017, 8(7): 2283–2300
74. Franceschin G, Sudiro M, Ingram T, Smirnova I, Brunner G, Bertucco A. Conversion of rye straw into fuel and xylitol: a technical and economical assessment based on experimental data. Chemical Engineering Research & Design, 2011, 89(6): 631–640
75. Giuliano A, Barletta D, De Bari I, Poletto M. Techno-economic assessment of a lignocellulosic biorefinery co-producing ethanol and xylitol or furfural. Computer-Aided Chemical Engineering, 2018, 43: 585–590
76. Mancini E, Mansouri S S, Gernaey K V, Luo J, Pinelo M. From second generation feed-stocks to innovative fermentation and downstream techniques for succinic acid production. Critical Reviews in Environmental Science and Technology, 2020, 50(18): 1829–1873
77. Ragauskas A J, Beckham G T, Biddy M J, Chandra R, Chen F, Davis M F, Davison B H, Dixon R A, Gilna P, Keller M, Langan P, Naskar A K, Saddler J N, Tschaplinski T J, Tuskan G A, Wyman C E. Lignin valorization: improving lignin processing in the biorefinery. Science, 2014, 344(6185): 1246843
78. Ponnusamy V K, Nguyen D D, Dharmaraja J, Shobana S, Banu J R, Saratale R G, Chang S W, Kumar G. A review on lignin structure, pretreatments, fermentation reactions and biorefinery potential. Bioresource Technology, 2019, 271: 462–472
79. Wang W C, Tao L. Bio-jet fuel conversion technologies. Renewable & Sustainable Energy Reviews, 2016, 53: 801–822
80. Prunescu R M, Blanke M, Jakobsen J G, Sin G. Dynamic modeling and validation of a biomass hydrothermal pretreatment process—a demonstration scale study. AIChE Journal. American Institute of

Chemical Engineers, 2015, 61(12): 4235–4250
81. Tochampa W, Sirisansaneeyakul S, Vanichsriratana W, Srinophakun P, Bakker H H C, Chisti Y. A model of xylitol production by the yeast *Candida mogii*. Bioprocess and Biosystems Engineering, 2005, 28(3): 175–183
82. S3O GitHub Repository. 2021, 10.5281/zenodo.5017353
83. Al R, Behera C R, Gernaey K V, Sin G. Towards development of a decision support tool for conceptual design of wastewater treatment plants using stochastic simulation optimization. Computer-Aided Chemical Engineering, 2019, 46: 325–330
84. Kılınç M R, Sahinidis N V. Exploiting integrality in the global optimization of mixed-integer nonlinear programming problems with BARON. Optimization Methods & Software, 2018, 33(3): 540–562
85. Vassilev S V, Baxter D, Andersen L K, Vassileva C G, Morgan T J. An overview of the organic and inorganic phase composition of biomass. Fuel, 2012, 94: 1–33
86. Eason J, Cremaschi S. Adaptive sequential sampling for surrogate model generation with artificial neural networks. Computers & Chemical Engineering, 2014, 68: 220–232
87. Garud S S, Karimi I A, Kraft M. Smart sampling algorithm for surrogate model development. Computers & Chemical Engineering, 2017, 96: 103–114
88. Garud S S, Karimi I A, Brownbridge G P E, Kraft M. Evaluating smart sampling for constructing multidimensional surrogate models. Computers & Chemical Engineering, 2018, 108: 276–288
89. Obermeier A, Vollmer N, Windmeier C, Esche E, Repke J U. Generation of linear-based surrogate models from non-linear functional relationships for use in scheduling formulation. Computers & Chemical Engineering, 2021, 146: 107203
90. Chen Y, Goetsch P, Hoque M A, Lu J, Tarkoma S. d-Simplexed: adaptive delaunay triangulation for performance modeling and prediction on big data analytics. IEEE Transactions on Big Data, 2019, in press
91. Jiang P, Zhang Y, Zhou Q, Shao X, Hu J, Shu L. An adaptive sampling strategy for kriging metamodel based on Delaunay triangulation and TOPSIS. Applied Intelligence, 2018, 48(6): 1644–1645