

RESEARCH ARTICLE

Process Systems Engineering

Data-driven decision-focused surrogate modeling

Rishabh Gupta  | Qi Zhang 

Department of Chemical Engineering and Materials Science, University of Minnesota, Minneapolis, Minnesota, USA

Correspondence

Qi Zhang, Department of Chemical Engineering and Materials Science, University of Minnesota, Minneapolis, MN 55455, USA.
Email: qizh@umn.edu

Funding information

Division of Chemical, Bioengineering, Environmental, and Transport Systems, Grant/Award Number: 2044077; University of Minnesota

Abstract

We introduce the concept of decision-focused surrogate modeling for solving computationally challenging nonlinear optimization problems in real-time settings. The proposed data-driven framework seeks to learn a simpler, for example, convex, surrogate optimization model that is trained to minimize the *decision prediction error*, which is defined as the difference between the optimal solutions of the original and the surrogate optimization models. The learning problem, formulated as a bilevel program, can be viewed as a data-driven inverse optimization problem to which we apply a decomposition-based solution algorithm from previous work. We validate our framework through numerical experiments involving the optimization of common nonlinear chemical processes such as chemical reactors, heat exchanger networks, and material blending systems. We also present a detailed comparison of decision-focused surrogate modeling with standard data-driven surrogate modeling methods and demonstrate that our approach is significantly more data-efficient while producing simple surrogate models with high decision prediction accuracy.

KEYWORDS

decision-focused learning, hybrid modeling, inverse optimization, surrogate modeling

1 | INTRODUCTION

Efficient and safe process operations require decision-making in real time; this is more important than ever as the chemical industry faces new fast-changing markets, greater feedstock variability, and increasingly time-sensitive availability of resources such as intermittent renewable energy. Many online decision-making frameworks, including model predictive control (MPC) and real-time optimization (RTO), involve the solving of mathematical optimization problems. Often, the computational complexity of the optimization problem presents a major challenge such that the long solution time renders it ineffective in online applications. A common approach to tackling this challenge is to perform the online optimization with a *surrogate model*, which is an approximation of the original model that can be solved more efficiently.

Typically in surrogate modeling, one tries to replace complicating functions that are embedded in the optimization problem with simpler

ones. This approach is widely used in process systems engineering (PSE) where the complicating functions are often associated with process units that exhibit complex nonlinear behavior. Over the years, the PSE community has developed a myriad of surrogate process models that include shortcut and lumped models derived from first principles and engineering assumptions, reduced-order dynamic models constructed using model reduction methods such as singular perturbation analysis, data-driven models based on, for example, Gaussian processes and artificial neural networks, and many more. Here, a major underlying assumption is that a surrogate model that provides a good approximation of the original functions will, once incorporated in the optimization problem, also lead to solutions that are close to the true optimal solutions. However, it is unclear whether or under what conditions this assumption holds and how accurate the surrogate model needs to be. In fact, as we show in this article, one can easily find examples in which the original optimization problem

This is an open access article under the terms of the [Creative Commons Attribution](https://creativecommons.org/licenses/by/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2024 The Authors. *AIChE Journal* published by Wiley Periodicals LLC on behalf of American Institute of Chemical Engineers.

and the surrogate optimization problem achieve very different optimal solutions despite having a highly accurate embedded surrogate model.

In this article, we propose a new surrogate modeling framework that explicitly aims to construct surrogate models that minimize the *decision prediction error* defined as the difference between the optimal solutions of the original and the surrogate optimization problems; we hence refer to it as *decision-focused surrogate modeling*. We take a data-driven approach in which the original optimization problem is solved offline with different model inputs, resulting in a dataset where each data point is an input-decision pair. We then develop an inverse optimization¹ approach that directly learns from the given data a surrogate optimization model that has a simpler (e.g., convex) form and minimizes the decision prediction error subject to the restrictions on the form of the surrogate model. Results from multiple computational case studies show that the proposed approach outperforms alternative surrogate modeling approaches in decision accuracy, data efficiency, and/or computational efficiency of the resulting surrogate optimization model.

In the remainder of this article, we first provide a systematic review of the main existing surrogate modeling frameworks and other related works in Section 2. We introduce the concept of decision-focused surrogate modeling, provide the corresponding mathematical problem formulation, and present a solution algorithm in Section 3. Several numerical case studies based on typical nonlinear chemical engineering systems are presented in Section 4. Finally, we conclude in Section 5.

2 | BACKGROUND AND RELATED WORK

In this section, we present an overview of the major surrogate modeling frameworks and other related work. We formally describe the different approaches to make clear the main conceptual differences, which will help us motivate the proposed decision-focused surrogate modeling framework and highlight its unique features in Section 3. We focus on data-driven approaches but make reference to other related methods wherever appropriate.

Without loss of generality, we assume the original optimization problem to be of the following form:

$$\begin{aligned} & \underset{x \in \mathcal{X}}{\text{minimize}} && f(x, u) \\ & \text{subject to} && g(x, u) \leq 0, \end{aligned} \quad (1)$$

where x are the decision variables and $\mathcal{X} \subseteq \mathbb{R}^n$. We assume that the input parameters u can be chosen from a given set $\mathcal{U} \subseteq \mathbb{R}^p$. The constraints describing the feasible region of (1) are defined by the functions $g: \mathbb{R}^n \times \mathbb{R}^p \rightarrow \mathbb{R}^k$, and $f: \mathbb{R}^n \times \mathbb{R}^p \rightarrow \mathbb{R}$ is the objective function of the problem.

2.1 | Optimization with embedded surrogate models

We first describe the traditional surrogate modeling framework as outlined in Section 1, which we call optimization with *embedded*

surrogate models. To explain the main idea of this approach, we rewrite (1) as follows:

$$\begin{aligned} & \underset{x \in \mathcal{X}}{\text{minimize}} && f(x, u) \\ & \text{subject to} && h(x, u) \leq 0 \\ & && d(x, u) \leq 0, \end{aligned} \quad (2)$$

where we divide the constraint functions g into two sets of functions h and d . This surrogate modeling approach is typically applied when problem (2) is such that its computational complexity mainly stems from the functions d ; hence, the goal is to replace them with a simpler set of functions \hat{d} . Such surrogate models are typically generated using simplifying assumptions based on physical and engineering insights, model order reduction techniques, or data-driven methods.² In a data-driven approach, one first generates a set of N data points where for each point $i \in \mathcal{I} = \{1, \dots, N\}$, \bar{x}_i and \bar{u}_i are sampled from \mathcal{X} and \mathcal{U} , respectively, and the corresponding function evaluations $\bar{y}_i = d(\bar{x}_i, \bar{u}_i)$ is computed. Assuming that $\hat{d}(\cdot)$ can be chosen from the set of functions \mathcal{D} , one then solves the following empirical risk minimization problem:

$$\underset{\hat{d}(\cdot) \in \mathcal{D}}{\text{minimize}} \frac{1}{N} \sum_{i \in \mathcal{I}} \ell(\bar{y}_i, \hat{d}(\bar{x}_i, \bar{u}_i)), \quad (3)$$

where $\ell(\cdot, \cdot)$ denotes a loss function, for example, the Euclidean distance, that is a measure of the difference between the observation and the estimate.

Once we have obtained \hat{d} , the surrogate model for d , we can formulate the following surrogate optimization model:

$$\begin{aligned} & \underset{x \in \mathcal{X}}{\text{minimize}} && f(x, u) \\ & \text{subject to} && h(x, u) \leq 0 \\ & && \hat{d}(x, u) \leq 0. \end{aligned} \quad (4)$$

Note that we use the term *surrogate model* for the alternative model trained using input-output data from the original model and *surrogate optimization model* for the resulting optimization problem that incorporates the surrogate model.

There exist a large number of statistical learning methods that are used to train these embedded surrogate models; popular choices in PSE include response surface methods,^{3–5} kriging (or Gaussian process regression),^{6–8} and deep learning.^{9–11} Data-driven approaches are often combined with first-principles modeling, resulting in gray-box models. For many physical systems, gray-box models have proven to perform better in terms of model accuracy and interpretability compared to purely data-driven models; hence, first-principles elements are often incorporated as long as they are not overly complex.^{12–14} A key challenge in surrogate modeling is the balance between model accuracy and computational efficiency; recent efforts aim at generating surrogate models composed of the simplest functional forms possible, given a desired model accuracy, to facilitate their use in mathematical optimization. A prominent example of such an approach is the ALAMO framework proposed by Sahinidis et al.^{15,16}

The constraints in an optimization problem define the feasible region for the variables of the problem; thus, generating surrogate constraint functions can also be interpreted as creating a surrogate representation of the feasible region. Approaches that are directly derived from this interpretation of surrogate modeling consider datasets that consist of feasible and infeasible points. With that, constructing a surrogate model essentially becomes a binary classification problem. Existing works have performed this kind of surrogate modeling using classical classification approaches such as support vector machines (SVMs).^{17,18} The intricate shapes of the feasible regions arising in many process systems applications have further motivated the development of more complex geometry-based approaches. Ierapetritou¹⁹ proposes to approximate a convex feasible region with the convex hull of a set of points sampled at the boundary. For nonconvex models, Goyal and Ierapetritou²⁰ develop an approach in which the feasible region is approximated by subtracting outer polytopes around the infeasible region from an expanded convex hull obtained from simplicial approximation.²¹ The nonconvex case has also been addressed using shape reconstruction techniques.²² Zhang et al.²³ propose an algorithm for constructing a union of multiple polytopes that approximates the nonconvex feasible region from which the data points are sampled. In a similar spirit, Schweidtmann et al.²⁴ apply persistent homology, a topological data analysis method, to identify potential holes and clusters in the data; this information is then used to obtain a representation of the nonconvex feasible region using one-class SVMs. In another approach, feasibility is captured using a so-called feasibility function,²⁵ and data-driven approaches are applied to approximate that function; see Bhosekar and Ierapetritou²⁶ for a comprehensive review of this type of methods.

2.2 | Learning optimization proxies

The tremendous advances in machine learning, especially deep learning, have enabled the modeling of highly complex relationships for accurate prediction. Recently, this has also spurred a growing interest in learning directly the mapping from the input parameters of an optimization problem to its optimal solution. In this setting, the required dataset is $\{(\bar{u}_i, x_i^*)\}_{i \in \mathcal{I}}$, where x_i^* denotes an optimal solution to the original problem (1) for the input \bar{u}_i . The goal is to learn a model $m(\cdot) \in \mathcal{M}$ that returns an optimal (or near-optimal) solution for a given input by solving the following expected risk minimization problem:

$$\underset{m(\cdot) \in \mathcal{M}}{\text{minimize}} \frac{1}{N} \sum_{i \in \mathcal{I}} \ell(x_i^*, m(\bar{u}_i)). \quad (5)$$

The corresponding surrogate model is simply

$$x = m(u), \quad (6)$$

which arrives at the solution through a function evaluation rather than by solving an optimization problem. Thus, m serves as a computationally efficient proxy for the original optimization model. Importantly, as

indicated by the loss function in problem (5), this approach explicitly aims to minimize the decision prediction error. This is in contrast to the methods reviewed in Section 2.1, which minimize the prediction error for individual constraint function values but cannot provide any guarantees in terms of how it affects the decision prediction error.

Although, in theory, any type of machine learning model (specified through \mathcal{M}) can be used in problem (5), the preferred choice in the literature has been deep neural networks.^{27–31} A major challenge in deep learning is the difficulty to enforce constraints on the predictions, which in this case often leads to predicted solutions that are infeasible in the original optimization problem. Zamzam and Baker³² address this challenge, when constructing neural network proxies for the AC optimal power flow (OPF) problem, by generating a training set of strictly feasible solutions through a modified AC OPF formulation and by using the natural bounds of the sigmoid activation function to enforce generation and voltage limits. Van Hentenryck and coworkers³³ apply Lagrangian duality to consider constraints in their proposed deep learning framework, where the loss function in (5) is augmented with penalty terms derived from the Lagrangian dual of the original optimization problem and the corresponding Lagrange multipliers are updated using a subgradient method. This approach has been applied in several applications, including AC OPF,³³ security-constrained OPF,³⁴ and job shop scheduling.³⁵ Another remedy is to correct an infeasible prediction by projecting it onto a suitably chosen set such that the projection is a feasible solution. For example, in MPC, such a set can be derived from the maximal control invariant set of the system to ensure recursive feasibility.^{36,37}

While the development of the machine learning approaches described above is a more recent trend, exact methods for the construction of optimization proxies have been studied in the area of multiparametric programming for a long time. In multiparametric programming, optimal solutions are derived as explicit functions of the model parameters; these functions (or policies) can change depending on the region in which the specific parameter vector lies. The goal is to determine these so-called critical regions and the optimal policy associated with each critical region. Once these are obtained offline, the online optimization reduces to simply selecting and applying the right function from a look-up table. This approach forms the basis for explicit MPC, which has been successfully applied in various real-world settings.^{38,39} However, a major challenge in multiparametric programming is the curse of dimensionality as the number of critical regions grows exponentially with the problem size. We refer the reader to Oberdieck et al.⁴⁰ for a review of the extensive literature on multiparametric programming.

Remark 1. In addition to the methods reviewed in this section, there is an extensive body of literature on the use of surrogate-based derivative-free optimization (DFO) techniques to solve complex optimization problems.⁴¹ We have consciously omitted referencing such work here. This is because our focus is specific, namely surrogate modeling for online optimization applications, where the same optimization model must be frequently solved with different values for its input parameters;

hence, the surrogate model's parametric nature is crucial. The DFO approach lacks parametric characteristics such that one must solve the problem using the chosen DFO strategy every time the model parameters change, which means that the construction of the surrogate models within the DFO algorithm would have to happen online. As a result, DFO is typically not efficient enough to allow its use in real-time applications. On the other hand, parametric methods, like the ones reviewed here, offer a distinct advantage. With these methods, one trains a surrogate optimization model just once offline, and it can then be employed online to solve for any inputs $u \in \mathcal{U}$ without the need of retraining the surrogate model.

3 | DECISION-FOCUSED SURROGATE MODELING

The surrogate modeling frameworks reviewed in Section 2 (also summarized in Figure 1) all have their advantages and disadvantages. Optimization with embedded surrogate models is an intuitive approach that allows preservation of much of the structure of the original model. Domain knowledge can be effectively leveraged since for someone familiar with the physical system, it is usually easy to identify the part of the model that needs to be replaced as well as determine a suitable structure for the corresponding surrogate model. Often, only a small number of constraints are complicating; in that case, simply keeping the remaining constraints can help ensure

feasibility. However, surrogate models generated using these methods may lead to solutions that are quite different from the optimal solutions to the original problem. This shortcoming can be overcome by learning optimization proxies in a decision-focused fashion. Using tailored deep learning architectures, this approach can achieve highly accurate and fast surrogate models. However, it is often less data-efficient, and cannot easily incorporate safety-critical hard constraints. In the following, we introduce the concept of decision-focused surrogate modeling (Figure 1C), where we try to combine the desirable characteristics of the existing surrogate modeling approaches. We further present an inverse optimization approach to constructing such surrogate models for a certain class of problems.

3.1 | General formulation

In the proposed framework, the data generation process is the same as in optimization proxy learning, where the dataset is $\{(\bar{u}_i, x_i^*)\}_{i \in \mathcal{I}}$ with x_i^* denoting the true optimal solution to the original problem with input \bar{u}_i . Given such data, we directly train a surrogate optimization model defined by objective function \hat{f} and constraint functions \hat{g} that minimizes the decision prediction error. This learning problem can be formulated as follows:

$$\underset{\hat{f}(\cdot) \in \mathcal{F}, \hat{g}(\cdot) \in \mathcal{G}}{\text{minimize}} \quad \frac{1}{N} \sum_{i \in \mathcal{I}} \ell(x_i^*, \hat{x}_i), \tag{DFSLPa}$$

$$\text{subject to} \quad \hat{x}_i \in \arg \min_{x \in \mathcal{X}} \{ \hat{f}(x, \bar{u}_i) : \hat{g}(x, \bar{u}_i) \leq 0 \} \quad \forall i \in \mathcal{I}, \tag{DFSLPb}$$

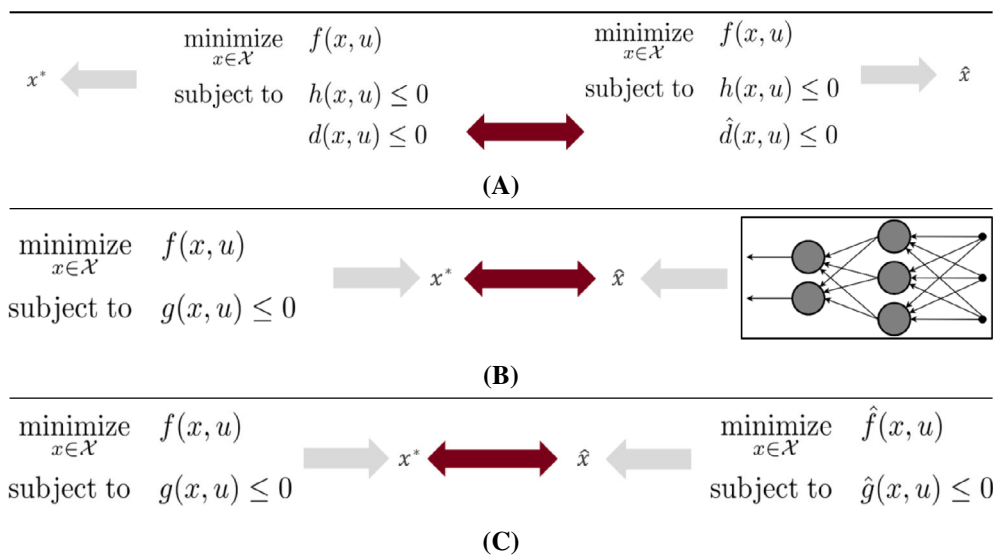


FIGURE 1 An overview of the surrogate modeling frameworks described in Section 2 (A and B) and decision-focused surrogate modeling (C). The true optimal solution of the original optimization model is denoted by x^* , whereas \hat{x} is the prediction generated by the surrogate. The gray arrows indicate the outputs of the various models while the red left-right arrows indicate what is considered in the loss function of each of the corresponding learning problems. (A) Optimization with embedded surrogate models; (B) Learning optimization proxies; (C) Decision-focused surrogate modeling.

where $\hat{f}(\cdot)$ and $\hat{g}(\cdot)$ can be chosen from some sets of functions \mathcal{F} and \mathcal{G} , respectively. The solution predicted by the surrogate optimization model for input \bar{u}_i is denoted by \hat{x}_i ; hence, the loss function is analogous to the one in the optimization proxy learning problem (5). As a result, this approach is decision-focused while allowing full flexibility in specifying the structure of the surrogate optimization model, including which original constraints to keep.

Note that since (DFSLP) replaces the original constraints g in (1) with their surrogates \hat{g} , solving the resulting surrogate optimization problem could lead to solutions that are infeasible to (1). This presents a trade-off for enhancing computational efficiency. This trade-off is also present in many other methods that we reviewed in Section 2. Typical strategies for handling infeasible solutions involve projecting the predicted solution to the closest point on the feasible set or using the prediction to warm-start a fast local solver. With our surrogate modeling approach, infeasibility can be eased by retaining most of the original constraints and replacing only those posing computational challenges (similar to methods outlined in Section 2.1). In a subsequent paper,⁴² we introduce a robust optimization method to identify regions of potential infeasibility within \mathcal{U} , which are then used to improve the surrogate model by acquiring additional training samples from these identified regions.

Remark 2. A closely related framework to our proposed decision-focused surrogate modeling approach is *decision-focused learning*.⁴³ However, it is not motivated by the need for fast online optimization, but instead, it addresses the following problem: In traditional data-driven optimization, we often follow a two-stage predict-then-optimize approach, that is, we first predict unknown input parameters u from data with external features r and then solve the optimization problem with those predicted inputs u . Here, the learning step focuses on minimizing the parameter estimation error; however, this does not necessarily lead to the best decisions (evaluated with the true parameter values) in the optimization step. In contrast, *decision-focused learning*, also known as *smart predict-then-optimize*,⁴⁴ *predict-and-optimize*,⁴⁵ and *end-to-end learning for optimization*,⁴⁶ integrates the two steps to explicitly account for the quality of the optimization solution in the learning of the model parameters.

3.2 | An inverse optimization approach

Problem (DFSLP) can be viewed as a data-driven inverse optimization problem. Given an observed decision made by some agent, the goal of inverse optimization¹ is to determine an optimization problem whose optimal solution matches and hence best explains the agent's decision; traditionally, only the objective function of the optimization model is assumed to be unknown.⁴⁷ In the (noisy) data-driven setting, multiple decisions for different inputs are observed, and the goal is to

find an optimization problem whose optimal solutions match the observations as closely as possible.⁴⁸ Applying this interpretation to the decision-focused surrogate modeling problem, the original optimization problem acts as the agent, and for each observation i , x_i^* represents the observed decision, and \bar{u}_i is the corresponding input; the surrogate optimization problem is then the optimization problem that we try to find such that its optimal solutions best resemble the observations.

Interpreting (DFSLP) as an inverse optimization problem allows us to leverage existing methods from that literature to solve the problem. Recently, we developed an efficient data-driven inverse optimization framework that can incorporate both unknown objective functions and constraints.⁴⁹ This approach can be readily applied to a certain class of decision-focused surrogate modeling problems, which we formally define in the following. Here, we consider the original optimization problem to be a generally nonconvex nonlinear program (NLP) of the following form:

$$\begin{aligned} & \underset{x \in \mathbb{R}^n}{\text{minimize}} && f(x, u) \\ & \text{subject to} && g(x, u) \leq 0 \\ & && h(x, u) = 0. \end{aligned} \quad (\text{OP})$$

We assume that (OP) can be solved to generate the dataset \mathcal{I} in (DFSLP). We use this dataset to learn a surrogate optimization problem which is a strictly convex optimization problem of the following form:

$$\begin{aligned} & \underset{x \in \mathbb{R}^n}{\text{minimize}} && \hat{f}(x, u; \theta) \\ & \text{subject to} && \hat{g}(x, u; \omega) \leq 0 \\ & && \hat{h}(x, u) = 0, \end{aligned} \quad (\text{SP})$$

where functions \hat{f} is strictly convex, \hat{g} are convex, and \hat{h} are linear in x . In (SP), we parameterize the objective function and constraints with parameters θ and ω , respectively. This parameterized surrogate problem can be learned in a decision-focused manner by solving (DFSLP) with θ and ω as its decision variables; we use the term *decision-focused surrogate optimization model* (DFSOM) to describe the resulting (SP). In this article, we use mean squared error as the loss function, that is, we solve (DFSLP) with $\ell(x_i^*, \hat{x}_i) = \|x_i^* - \hat{x}_i\|_2^2$.

3.3 | Solution method

The learning problem (DFSLP) is a bilevel program⁵⁰ with multiple lower-level optimization problems. Despite the convexity of the lower-level problems, (DFSLP) is difficult to solve due to its poor scalability.⁵¹ In our previous work,⁴⁹ we addressed this problem with an efficient decomposition-based strategy that generates high-quality solutions; the same algorithm is employed here. For the sake of brevity, we only provide a brief overview of our solution method in this subsection and refer the interested reader to Gupta and Zhang⁴⁹ for a more detailed discussion.

Our solution approach involves reformulating the bilevel problem into a single-level optimization problem by replacing the lower-level problems with their *Karush-Kuhn-Tucker* (KKT) optimality conditions. It is important to highlight that this reformulation requires that (SP) satisfies some constraint qualification conditions. This requirement can be met by designing (SP) with all linear constraints (as done in our computational experiments in Section 4), or by ensuring that the set defined by $\hat{g} < 0$ and $\hat{h} = 0$ is always feasible. The reformulation of (DFSLP) using KKT conditions results in an NLP as follows:

$$\underset{\theta \in \Theta, \omega \in \Omega, \hat{x}, \lambda, \mu}{\text{minimize}} \quad \sum_{i \in \mathcal{I}} \|x_i^* - \hat{x}_i\|_2^2 \quad (7a)$$

$$\text{subject to} \quad \nabla \hat{f}(\hat{x}_i, \bar{u}_i; \theta) + \lambda_i^\top \nabla \hat{g}(\hat{x}_i, \bar{u}_i; \omega) + \mu_i^\top \nabla \hat{h}(\hat{x}_i, \bar{u}_i; \omega) = 0 \quad \forall i \in \mathcal{I} \quad (7b)$$

$$\hat{g}(\hat{x}_i, \bar{u}_i; \omega) \leq 0 \quad \forall i \in \mathcal{I} \quad (7c)$$

$$\hat{h}(\hat{x}_i, \bar{u}_i; \omega) = 0 \quad \forall i \in \mathcal{I} \quad (7d)$$

$$\lambda_i^\top \hat{g}(\hat{x}_i, \bar{u}_i; \omega) = 0 \quad \forall i \in \mathcal{I} \quad (7e)$$

$$\lambda_i \geq 0, \hat{x}_i \in \mathbb{R}^n \quad \forall i \in \mathcal{I}, \quad (7f)$$

where the dual variables for the inequality and equality constraints of the lower-level problems in (DFSLP) are respectively denoted by λ and μ . Constraints (7b)–(7e) formulate the stationarity, primal feasibility, and complementary slackness conditions for the lower-level problems.

The reformulated problem is a nonconvex NLP with complementarity constraints that violate standard constraint qualification conditions. As these problems are known to cause convergence difficulties for NLP solvers, we further consider an *exact penalty reformulation*⁵² of (7). This reformulation yields the following problem with a regularized feasible region that satisfies the necessary constraint qualifications if the description of sets Θ and Ω also satisfy necessary regularity conditions:

$$\underset{\theta \in \Theta, \omega \in \Omega, \hat{x}, \lambda, \mu}{\text{minimize}} \quad \sum_{i \in \mathcal{I}} \|x_i - \hat{x}_i\|_2^2 + c^\top \left[\begin{array}{l} \sum_{i \in \mathcal{I}} |\nabla \hat{f}(\hat{x}_i, \bar{u}_i; \theta) + \lambda_i^\top \nabla \hat{g}(\hat{x}_i, \bar{u}_i; \omega) + \mu_i^\top \nabla \hat{h}(\hat{x}_i, \bar{u}_i; \omega)| \\ \sum_{i \in \mathcal{I}} \max\{0, \hat{g}(\hat{x}_i, \bar{u}_i; \omega)\} \\ \sum_{i \in \mathcal{I}} |\hat{h}(\hat{x}_i, \bar{u}_i; \omega)| \\ \sum_{i \in \mathcal{I}} |\lambda_i^\top \hat{g}(\hat{x}_i, \bar{u}_i; \omega)| \end{array} \right] \quad (8)$$

subject to $\lambda_i \geq 0, \hat{x}_i \in \mathbb{R}^n \forall i \in \mathcal{I},$

where c are the positive penalty parameters. The exactness of this reformulation holds only for penalty parameters greater than a certain threshold, which is hard to determine *a priori* in practice. As a result, we take an iterative approach, starting with small values for c and gradually increasing them in subsequent iterations (by a factor of ρ) until the solution of (8) is also a feasible solution for (7). We use a feasibility tolerance of ϵ as the termination criterion for the algorithm. A pseudocode that summarizes our overall approach is shown in Algorithm 1.

While (9) can be solved using the penalty reformulation, the reformulated problem (8) remains a nonconvex NLP whose large instances are generally difficult to solve. Here, we notice that for certain classes of (SP), including quadratic programs (QPs), (8) becomes a multiconvex optimization problem⁵³ (MCP) if the sets Θ and Ω are convex. This feature of (8) allows us to solve it with an efficient block-coordinate-descent (BCD) algorithm. We specifically highlight QPs as an example here because QPs are commonly used to model/approximate RTO and MPC problems. In fact, we show several examples of real-world systems in Section 4 where we use a QP to approximate their original nonconvex RTO problem. Therefore, the applicability to QPs makes our decomposition scheme especially appealing for decision-focused surrogate modeling.

Remark 3. The objective function of (8) contains non-smooth terms due to the use of ℓ_1 -norm-based penalty functions. This problem can be exactly reformulated into

Algorithm 1 A penalty block coordinate descent algorithm for solving (7)

- 1: initialize: $k \leftarrow 1, (\theta, \omega, \hat{x}, \lambda, \mu) \leftarrow (\theta_0, \omega_0, \hat{x}_0, \lambda_0, \mu_0)$ and $c \leftarrow c_1$
- 2: **while** $\|P\| > \epsilon$ **do**
- 3: solve (8) with BCD or an NLP solver (e.g., IPOPT)
- 4: $c_{k+1} \leftarrow c_k + \rho c_k$
- 5: $k \leftarrow k + 1$
- 6: **end while**
- 7: **return** $(\theta_k, \omega_k, \hat{x}_k, \lambda_k, \mu_k)$

one with a smooth objective function by introducing additional variables to linearize the penalty terms. The reformulation results in an increased size of the problem; however, in our previous work,⁴⁹ we show that (8) provides significantly better solutions than when IPOPT⁵⁴ is applied directly on (7).

4 | COMPUTATIONAL CASE STUDIES

We present numerical results from three case studies based on typical nonlinear chemical engineering systems where we apply the proposed decision-focused surrogate modeling strategy to simplify the optimization task. The first two case studies are based on single-input systems for which we present detailed analyses demonstrating how our approach can provide excellent decision prediction accuracy with a much simpler model than is typically required by traditional methods that construct surrogate optimization problems with embedded surrogate models. The third case study is based on a larger multi-input multi-output blending network system where we show the data efficiency of our approach in comparison to black-box optimization proxies. We also demonstrate the effectiveness of the resulting DFSOM in reducing the computational burden of optimizing the system in real time. All computer code along with the datasets used for these case studies is available at <https://github.com/ddolab/DecFocSurrMod>.

4.1 | RTO of a continuous stirred tank reactor system

We consider the problem of RTO of a continuous stirred tank reactor (CSTR) operating in a stochastic environment. Specifically, we optimize the operation of an ideal adiabatic CSTR⁵⁵ that is carrying out an exothermic reversible reaction between reactant A and product R. The concentration of the inlet stream, which does not contain any R, is subject to observable disturbances. Here, the primary goal of RTO is to maximize the concentration of the product R in the outlet stream by manipulating the temperature of the inlet stream. This can be formulated as the following optimization problem:

$$\begin{aligned} & \underset{T_i, T_o, A_o, R_o}{\text{maximize}} && 2.009R_o - 1.657 \times 10^{-3}(T_i - 410)^2 \\ & \text{subject to} && 0 = 1/\tau(A_i - A_o) - k_1(T_o)A_o + k_{-1}(T_o)R_o \\ & && 0 = -R_o/\tau + k_1(T_o)A_o - k_{-1}(T_o)R_o \\ & && 0 = 1/\tau(T_i - T_o) + \frac{-\Delta H_R}{\rho C_p}(k_1(T_o)A_o - k_{-1}(T_o)R_o), \end{aligned} \quad (9)$$

where A_i is the concentration of A in the inlet stream, and T_i is the inlet temperature. Similarly, variables with the subscript o characterize the properties of the outlet stream. In (9), the first two constraints are the mass balances whereas the third constraint specifies the energy balance for the reactor. The paper by Economou et al.⁵⁵ provides details of all model parameters used in this case study.

Problem (9) is a nonconvex NLP due to the dependence of the forward and backward reaction rate constants (k_1 and k_{-1}) on the reactor temperature (T_o). We use decision-focused surrogate modeling to learn a convex surrogate optimization model for the RTO problem. We achieve this by replacing the original rate law expression by the following approximation that is linear in the decision variables of (9):

$$k_1(T_o)A_o - k_{-1}(T_o)R_o \rightarrow a(A_i)T_o + b(A_i)A_o + c(A_i)R_o, \quad (10)$$

where a , b , and c are functions of the input parameter A_i . In addition, we also learn scalar coefficient values for a new quadratic objective function for the DFSOM.

4.1.1 | Surrogate model complexity and data efficiency

We begin by examining how using different functional forms for the coefficients $a(A_i)$, $b(A_i)$, and $c(A_i)$ in (10) affects the accuracy of the learned DFSOM. To simplify our analysis, we focus on these coefficients being polynomial functions of the input disturbance value. We test our model's accuracy by varying the complexity of the functions from constants up to cubic polynomials of A_i . In each case, we train the DFSOM using a dataset \mathcal{I} of 1000 samples, with each sample containing the input parameter A_i and the corresponding optimal (T_i, A_o, R_o) values obtained by solving the original NLP (9). After training, we evaluate the accuracy of the model by testing it on a separate dataset of 100 samples.

The results of the model complexity analysis are shown in Figure 2A. In the plot, the ‘‘Sparse 3’’ column represents the case in which we specify the coefficient models to be cubic but add a term to the objective function of (DFSOM) that penalizes the sum of absolute values of the polynomial coefficients. This technique is commonly used in machine learning to control model complexity. We find that this approach results in simple yet highly accurate polynomials, so we choose ‘‘Sparse 3’’ as the model for our further analysis.

Figure 2A shows that increasing the model complexity generally leads to a more accurate optimization problem. However, it is important to note that this added complexity lies in the input parameter space, and the optimization problem in the decision variable space stays convex. This is one of the unique features of our approach, allowing us to transfer the complexity from the decision variable space to the input space while still learning the relevant characteristics of the original optimization model. None of the traditional methods that employ embedded surrogate models can learn a similar model because their model learning step is entirely separate from the optimization step, so there is no differentiation between input parameter space and decision variable space when constructing the surrogate model.

We then investigate how the size of the training dataset affects the accuracy of our model. This is important because constructing the dataset requires solving a complex optimization problem. Figure 2B shows that our method can produce an accurate and robust model

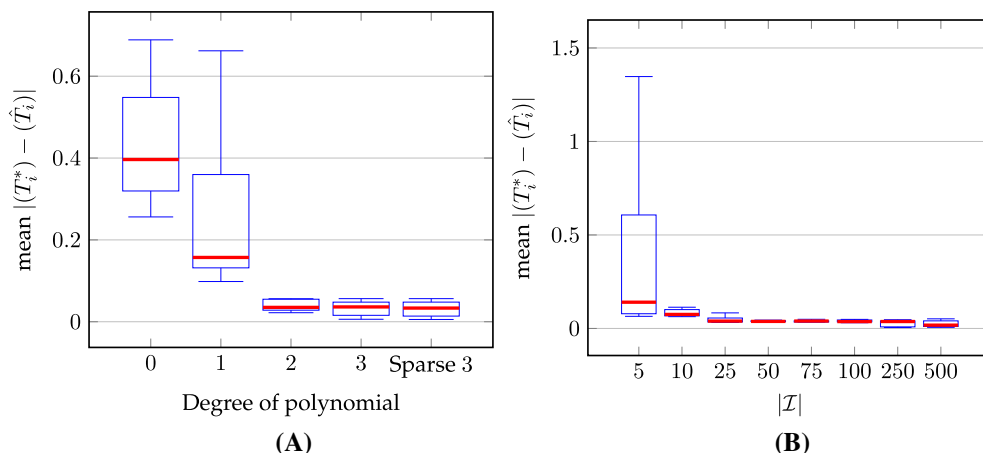


FIGURE 2 The true optimal solution obtained by solving (9) is denoted by T_i^* , and \hat{T}_i represents the solution of the decision-focused surrogate optimization model. The box plots show the interquartile ranges of prediction error for the surrogate models obtained with 10 different instances of randomly generated training data. We use a separate test dataset of 100 samples to compute the prediction error. (A) Effect of model complexity; (B) Effect of training dataset size.

with as few as 25 samples, indicating that a large dataset may not be necessary. In the next subsection, we compare the performance and data efficiency of our method with traditional methods using embedded surrogate models.

4.1.2 | Comparison with traditional embedded surrogate model approach

To compare our proposed strategy to other surrogate modeling methods popular in PSE, we construct surrogate models for the rate law expression. We use a large training dataset of 3000 points, where each point consists of (A_o, R_o, T_o) as inputs and the true rate value as the output. These surrogate models learn a function $\hat{r}(A_o, R_o, T_o)$ that produces approximately the same reaction rate as the original model (on the left-hand side) in (10). We then substitute the original rate law in (9) with \hat{r} to obtain a surrogate optimization model, which we solve using the local NLP solver IPOPT.⁵⁴ Note that although the resulting surrogate optimization models are generally nonconvex, we use a local solver because it is typically preferred in RTO settings due to the high computational effort required by global solvers.

We consider the following two surrogate modeling methods for the rate law expression:

- ES-ALAMO - We use ALAMO¹⁵ (version 2022.10.7) to estimate \hat{r} . We allow all available basis functions except for the sine and cosine functions. We manually substituted the algebraic \hat{r} expressions obtained using ALAMO into an RTO problem implementation in the JuMP⁵⁶ modeling environment available in the Julia programming language.
- ES-NN - We train a neural network (NN) to estimate the reaction rate value. The NN model comprises three inputs, four hidden layers with 10 nodes, and a single output node. We use smooth sigmoid activation functions for the hidden layers to ensure that the resulting optimization model remains solvable with IPOPT. We used the Python library TensorFlow⁵⁷ v2.10.0 to train the NN models. These models were incorporated into the optimization problems using the OMLT⁵⁸ package v1.0.

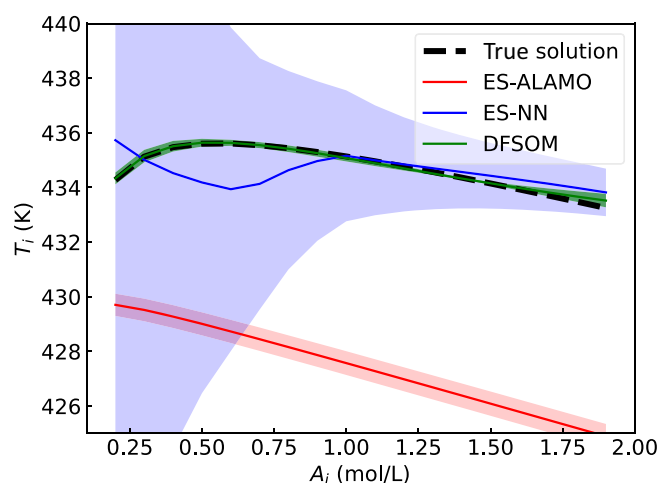


FIGURE 3 Line plots show the mean optimal T_i values obtained using different surrogate modeling methods, with shaded areas indicating one standard deviation. Five surrogate optimization models were developed for each method using different datasets.

Figure 3 shows a comparison of optimal T_i values obtained using different methods. The DFSOM, trained with only 50 data points, significantly outperforms the two traditional methods. While ES-ALAMO yields a robust model that does not vary much with changing data, the predicted optimal T_i values differ significantly from their true values. In contrast, while ES-NN yields mean T_i values close to the true optimum, its output is highly sensitive to the training dataset used. The decision-focused model provides both high accuracy and robustness, independent of the quality of the dataset.

We further evaluate the quality of different models by comparing the feasible regions of the surrogate optimization models with the true feasible region. This analysis is shown in Figure 4. Methods that embed surrogate models in the optimization problem, as compared in Figure 4A, aim to exactly replicate the feasible region, often requiring complex models that need larger training datasets. However, even with a good surrogate model that closely replicates the true feasible region, the decision prediction error can still be high due to small discrepancies, as we observe for ES-NN. In contrast, decision-

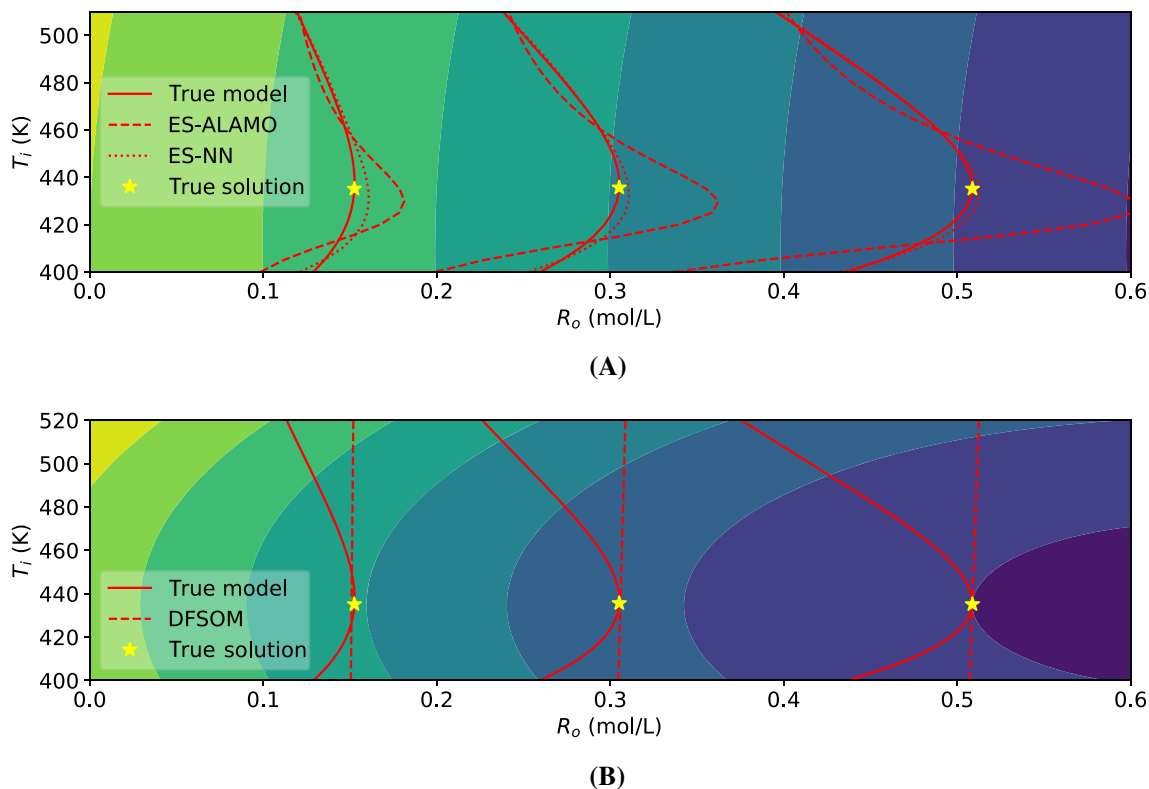


FIGURE 4 The feasible regions of the original and surrogate real-time optimization problems projected onto the two-dimensional variable space between R_o and T_i . From left to right, the three line plots depict the feasible regions for $A_i = 0.3, 0.6$, and 1.0 mol/L, respectively. The contours in the background represent the objective function for the true model in (A) and for the decision-focused surrogate optimization model in (B), with darker colors indicating larger values.

focused learning achieves high accuracy with simple models by focusing only on predicting the optimal solutions. As shown in Figure 4B, although the feasible region of the DFSOM is different from the true feasible region, they coincide exactly at the optimum for the original model. The learning problem (DFSOP) adjusts the surrogate optimization model's objective function to ensure this point is the optimal solution. Overall, we find that decision-focused learning allows us to significantly reduce the complexity of the surrogate optimization model while retaining its accuracy in predicting the optimal solutions.

4.2 | RTO of a heat exchanger network

Our second case study considers a heat exchanger network adapted from Biegler et al.,⁵⁹ as shown in Figure 5. The inlet temperature of stream H2, denoted by T_5 , has a nominal value of 583 K but is subject to random disturbances. Upon a change in T_5 , we optimize the operation of the heat exchanger network by solving the following nonconvex NLP:

$$\underset{Q_c, F_{H2}}{\text{minimize}} \quad 10^{-2} Q_c + 4(F_{H2} - 1.7)^2, \quad (11a)$$

$$\text{subject to} \quad 0.5Q_c + 165 \geq 0 \quad (11b)$$

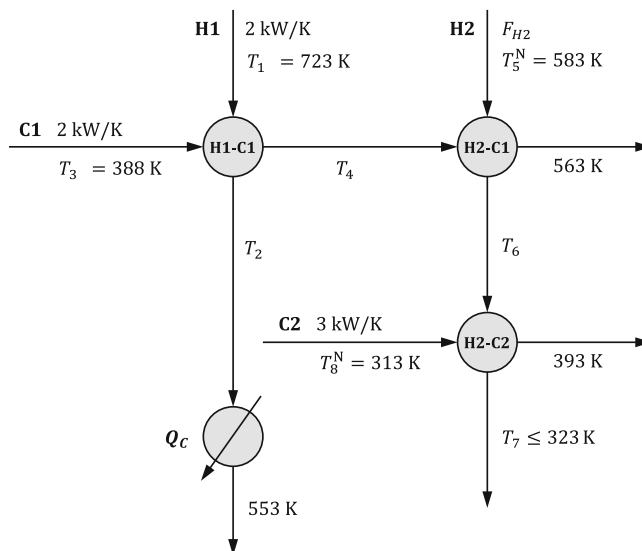


FIGURE 5 Given heat exchanger network.

$$-10 - Q_c + (T_5 - 558 + 0.5Q_c)F_{H2} \geq 0 \quad (11c)$$

$$-10 - Q_c + (T_5 - 393)F_{H2} \geq 0 \quad (11d)$$

$$-250 - Q_c + (T_5 - 313)F_{H2} \geq 0 \tag{11e}$$

$$-250 - Q_c + (T_5 - 323)F_{H2} \leq 0 \tag{11f}$$

$$Q_c \geq 0, F_{H2} \geq 0, \tag{11g}$$

where the cooling duty Q_c and the heat capacity flow rate F_{H2} are adjustable variables.

The nonconvexity in (11) arises from the bilinear term in constraint (11c). Therefore, we apply our approach to replace the nonconvex term with the following approximation that is linear in the decision variables of (11):

$$Q_c F_{H2} \rightarrow a(T_5) Q_c + b(T_5) F_{H2}, \tag{12}$$

where a and b are functions of the input parameter T_5 . Along with this, we also learn new coefficients for the quadratic objective function. Unlike in the previous case study, here these coefficients are also functions of the input parameter T_5 . We specify the model for all unknown parameters as ‘‘Sparse 3’’ polynomials in T_5 (which we defined in Section 4.1.1).

To train the models for a , b , and the objective coefficients, we solve (DFSOP) with datasets of varying sizes. Each data point in the dataset consists of an input T_5 and the corresponding optimal Q_c^* and F_{H2}^* values that we obtain by solving (11). We evaluate the performance of the resulting surrogate optimization models using a test dataset of 100 data points. The analysis of these results is available in Figure 6A, where we show the evolution of prediction quality as a function of the training dataset size. We find that the prediction error of the models converges to its minimum value with as few as 50 data points. Therefore, we set $|I|$ to 50 for our further analysis.

Next, we compare the performance of our decision-focused approach with the traditional approach with embedded surrogate models. To do this, we train an NN with two hidden layers, each consisting of ten nodes. The objective is to output the value of the bilinear term, given Q_c and F_{H2} as inputs. The dataset used for this purpose consists of 1000 data points. Similar to the first case study,

we maintain the sigmoid activation function for the hidden layers to guarantee the solvability of the surrogate optimization model with IPOPT. We then embed this NN in (11) by replacing the bilinear term, resulting in a surrogate optimization model that we call ES-NN.

Due to the simplicity of the function that the NN intends to replace, the NN learns to approximate the bilinear term almost perfectly. However, as we show in Figure 6B, ES-NN does not always yield the true optimal Q_c values as its solution. This happens because the NN-based problem is nonconvex and IPOPT recovers one of the two local solutions that are possible for a given T_5 value. Here, a global solver can be used to remedy this situation, but this may not be practical for RTO due to the associated computational expense. In contrast, the DFSOM, which is convex by design, correctly identifies the existence of the discontinuity in the global optimal solution space with IPOPT. Additionally, we find that the DFSOM produces solutions that are nearly identical to those produced by the true problem for all values of T_5 considered in this case study.

In conclusion, this case study showcases a notable benefit of decision-focused surrogate modeling, whereby the surrogate optimization models resulting from this approach exhibit convexity, rendering them amenable to efficient local solvers. This obviates concerns regarding the convergence to suboptimal solutions, which is a common issue when using nonconvex models in RTO applications.

4.3 | RTO of a blending network

We consider a network of blending nodes that mix material streams of various specifications to produce products with desired qualities. An optimization problem to minimize the cost of operation of this network is as follows⁶⁰:

$$\underset{\bar{f}, x, q}{\text{minimize}} \quad \sum_{j \in J} \sum_{i \in N_j} c_{ij} \left(f_{ij} + \frac{1}{1000} (f_{ij} - \bar{f}_{ij})^2 \right) - \sum_{k \in K} d_k \sum_{j \in J} x_{jk}, \tag{13a}$$

$$\text{subject to} \quad \sum_{i \in N_j} f_{ij} = \sum_{k \in K} x_{jk} \quad \forall j \in J \tag{13b}$$

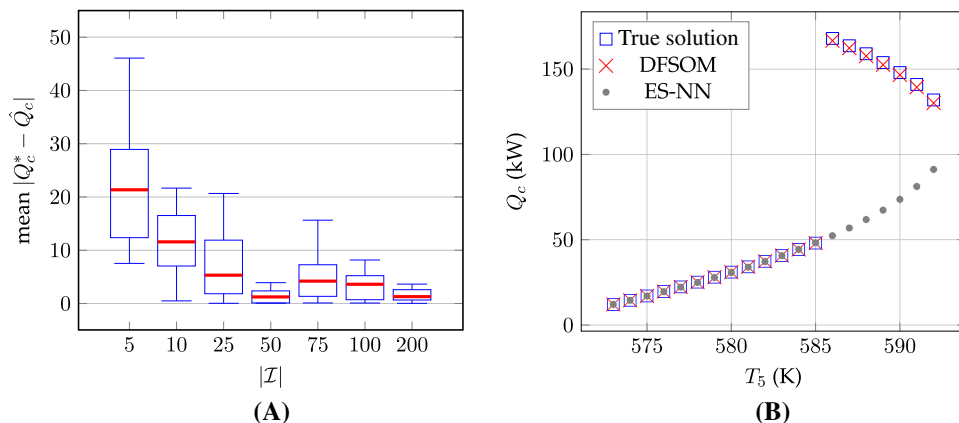


FIGURE 6 The true optimal solution obtained by solving (11) is denoted by Q_c^* , and \hat{Q}_c represents the solution of the surrogate optimization model. The box plots show the interquartile ranges of prediction error for the surrogate optimization models obtained with 10 different instances of randomly generated training data. (A) Effect of training dataset size on model accuracy. We use a separate test dataset of 100 samples to compute the prediction error. (B) Comparison of DFSOM with embedded surrogate model.

$$q_{jw} \sum_{k \in \mathcal{K}} x_{jk} = \sum_{i \in \mathcal{N}_j} \lambda_{ijw} f_{ij} \quad \forall j \in \mathcal{J}, w \in \mathcal{W} \quad (13c)$$

$$\sum_{j \in \mathcal{J}} x_{jk} \leq S_k \quad \forall k \in \mathcal{K} \quad (13d)$$

$$\sum_{j \in \mathcal{J}} q_{jw} x_{jk} \leq Z_{kw} \sum_{j \in \mathcal{J}} x_{jk} \quad \forall k \in \mathcal{K}, w \in \mathcal{W} \quad (13e)$$

$$f_{ij} \geq 0 \quad \forall j \in \mathcal{J}, i \in \mathcal{N}_j \quad (13f)$$

$$q_{jw} \geq 0 \quad \forall j \in \mathcal{J}, w \in \mathcal{W} \quad (13g)$$

$$x_{jk} \geq 0 \quad \forall j \in \mathcal{J}, k \in \mathcal{K}, \quad (13h)$$

where \mathcal{J} is the set of blending nodes, and \mathcal{N}_j are the sets of material streams entering node j ; we use variable f_{ij} to denote the flow rate of stream i entering node j . The product streams are generated by combining the outflows from different blending nodes and are represented by the set \mathcal{K} . The outflow rate from node j to product stream k is denoted by the variable x_{jk} . Additionally, the set \mathcal{W} contains the various components present in the material streams whose specifications need to be maintained in the output streams. We determine the quality of node j for a specific component w using the variable q_{jw} .

In problem (13), we enforce mass balance for each node through constraints (13b). The quality of a blending node j is determined as a function of the specifications, λ_{ijw} , of the streams entering that node, through constraints (13c). Additionally, we ensure that the outflow rate of a product stream k does not exceed its market demand, S_k , using constraints (13d). Finally, constraints (13e) set the specification of component w in product stream k below its permissible level, Z_{kw} . The objective function of (13) comprises of three terms. The first represents the linear cost of acquiring a material stream, while the second penalizes the deviation of inlet flow rates from their nominal values, \bar{f}_{ij} . The third term represents the revenue generated through the product streams.

We consider a scenario in which the RTO of the blending network operation is desired in the face of changing product demands, $S = \{S_k\}_{k \in \mathcal{K}}$. However, solving (13) is highly challenging due to the presence of bilinearities in constraints (13c) and (13e). In fact, efficient solution of the blending problem has been the subject of much research due to its importance in the operation of petroleum refineries and waste-water treatment plants, among others.^{61–63} To address this issue, we propose a solution that involves training a convex DFSOM of (13) offline for online use. In what follows, we show that our approach can significantly speed up the online solution of the blending problem while preserving solution accuracy, as measured against the *global solutions* of (13).

4.3.1 | Design and training of the DFSOM

Our approach involves linearizing the nonconvex terms in (13). For each $j \in \mathcal{J}$, $k \in \mathcal{K}$, and $w \in \mathcal{W}$, there is a bilinear term $q_{jw} x_{jk}$ in (13), which we replace with the following approximation:

$$\sum_{k \in \mathcal{K}} \sum_{\ell=0}^2 \left(p'_{jkwk\ell} S_k^\ell q_{jw} + q'_{jkwk\ell} S_k^\ell x_{jk} \right), \quad (14)$$

where $p'_{jkwk\ell}$ and $q'_{jkwk\ell}$ are scalar parameters to be determined by solving (DFSPLP). Similar to the previous two problems, we penalize the sum of absolute values of $p'_{jkwk\ell}$ and $q'_{jkwk\ell}$ in the objective function to induce sparsity. In this problem, we keep the objective function in the decision-focused surrogate problem the same as in the original problem.

We test the proposed decision-focused surrogate modeling approach on the blending network presented in Example 2 of Foulds et al.⁶⁴ This network has four pooling nodes blending a total of six inlet streams that result in four final products; for the exact parameter values used in this case study, we refer to Foulds et al.⁶⁴ To generate training data for (DFSPLP), we solve (13) with S_k values sampled from the uniform distribution $\mathcal{U}(100,200)$ for all k in \mathcal{K} . A single data point here consists of the input vector S and the corresponding optimal solution vector (f, x, q) .

4.3.2 | Data efficiency of DFSOM in comparison to black-box optimization proxies

We examine the minimum amount of training data needed to create a high-quality surrogate. To do this, we solve (DFSPLP) using different sizes of \mathcal{I} , and then assess the prediction error on a separate test dataset of 100 points. Figure 7A displays the results of this analysis. Each box plot in the figure represents the interquartile range of prediction error for ten decision-focused surrogate models, each constructed using a different random training dataset of size $|\mathcal{I}|$. Remarkably, we find that, even for this high-dimensional problem, a low prediction error can be achieved with just 20 samples. Additionally, the prediction error achieves its minimum value with as few as 75 samples. These findings demonstrate the effectiveness of the proposed approach in developing surrogates with a small dataset, which is crucial since solving (13) to construct a large training dataset can be computationally burdensome.

The results in Figure 7A are in direct contrast to the ones presented in Figure 7B where we assess the accuracy of NN models that are trained to be optimization proxies for (13). We train three different types of networks to produce optimal f and x values given inputs S . The main difference between these networks is their size, which allows them to have different numbers of learnable parameters (our DFSOMs have 288 learnable parameters); the key characteristics of these networks are summarized in Table 1. We train these NNs with different numbers of training data points ranging from 10 to 4000. For each NN and $|\mathcal{I}|$ combination, we train ten different models, each with a different random training dataset. Similar to the DFSOM case, we evaluate the prediction accuracy of the NNs on a separate test dataset of 100 samples.

As shown in Figure 7B, even with 4000 data points, the NNs exhibit significantly worse performance compared to the DFSOMs. We believe this is a direct consequence of the black-box nature of

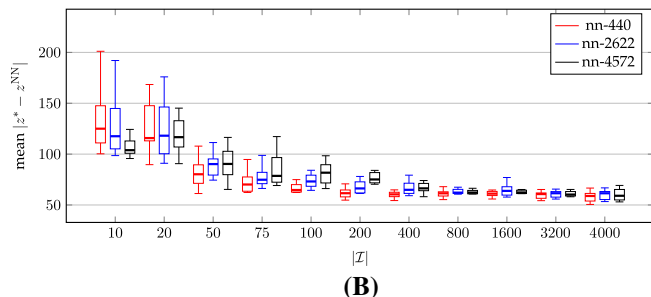
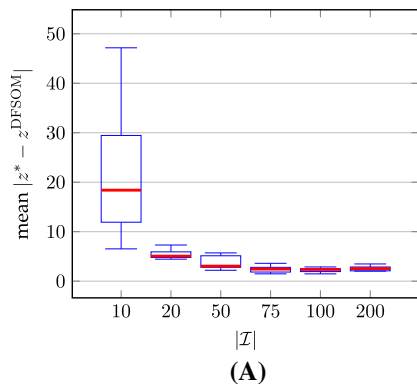


FIGURE 7 The true optimal objective function value of (13) is denoted by z^* , whereas z^{DFSOM} and z^{NN} denote the objective values obtained by using the solutions generated by decision-focused surrogate optimization model (DFSOM) and neural network (NN) proxies, respectively. The box plots show the interquartile ranges of prediction error for the surrogate models obtained with ten different instances of randomly generated training data. In both figures, we use a separate test dataset of 100 samples to compute the prediction error. (A) Effect of training dataset size on the accuracy of DFSOMs. (B) Effect of training dataset size on the accuracy of NNs.

TABLE 1 Key characteristics of neural networks (NNs) trained as optimization proxies for (13).

NN name	Number of hidden layers	Number of nodes/layer	Number of learnable parameters
nn-440	2	11	440
nn-2622	4	25	2622
nn-4572	7	25	4572

Note: We use ReLU activation functions for the hidden layers.

NNs, which does not allow for the incorporation of known correlations about the data, in contrast to the DFSOM. It is also important to note that the DFSOM output will always satisfy the crucial mass balances in (13b) and demand satisfaction constraints in (13d), whereas there is no trivial way of ensuring the same for the NN outputs. These findings further highlight the fact that the proposed decision-focused surrogate modeling approach can be a superior alternative to NNs when it comes to generating high-quality surrogates for computationally challenging optimization problems.

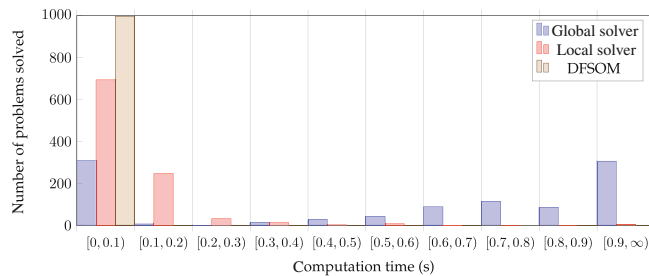


FIGURE 8 Computational performance of decision-focused surrogate optimization model (DFSOM) in comparison to when (13) is solved with local and global optimization solvers.

TABLE 2 Optimality gap for solutions produced by decision-focused surrogate optimization model (DFSOM) compared with when (13) is solved by IPOPT.

	Mean	Maximum	Minimum
$\frac{ z^{\text{local}} - z^* }{ z^* }$	0.61	2.30	0
$\frac{ z^{\text{DFSOM}} - z^* }{ z^* }$	0.002	0.009	0

4.3.3 | Computational performance of DFSOMs for RTO

We analyze the computational performance of the DFSOM in contrast to (13). We solve 1000 instances of a DFSOM and (13), with each instance corresponding to different S values following the same distribution as the training data. The optimization solver Gurobi is used to solve (DFSOP), while we solve each instance of (13) twice, once with the global solver Gurobi and again with the local solver IPOPT. A histogram in Figure 8 shows the distribution of computation times for the three cases. We find that, depending on S , the global solution of (13) can take more than 100 s to find, possibly making it intractable for online application. While the local solver provides a fast solution in most cases, the generated solutions generally suffer from large optimality gaps as shown in Table 2. IPOPT-generated solutions had a mean optimality gap of 61%, which can be as high as 230% in the worst case.

In contrast, DFSOM is highly effective in alleviating computational burden associated with the global solution of this challenging optimization problem. We find that all 1000 instances of DFSOM solve in less than 0.1 s, and even the worst-case optimality gap is only 0.9%. Therefore, in addition to being a superior surrogate modeling strategy, using decision-focused surrogate modeling can also be a better alternative to suboptimal local solvers.

Remark 4. In all three numerical case studies, we employ polynomial functions of inputs to represent the nonlinear coefficient models. For instance, in (10), $a(A_i)$, $b(A_i)$, and $c(A_i)$ are modeled as cubic polynomials in A_i . The choice of polynomials stems from their linearity in

learnable parameters, thus preserving the multiconvex nature of (8). While the proposed framework can accommodate more complex functions like neural networks with nonlinear activation functions, this would preclude the use of BCD for solving the ensuing learning problem, potentially complicating the solution of larger instances. Nonetheless, our computational findings, particularly depicted in Figure 4, demonstrate that the proposed framework can achieve highly accurate decision predictions from simpler models. This can be attributed to the framework's focus on learning only the optimal solution space of an optimization problem, rather than attempting to model the entire feasible solution space.

5 | CONCLUSIONS

In this article, we introduced a novel decision-focused surrogate modeling approach for the online solution of computationally challenging nonlinear optimization problems. Our data-driven framework produces a surrogate optimization problem that minimizes the decision prediction error on a training dataset containing optimal solutions of the original optimization problem corresponding to different model inputs. We showed that the learning problem can be formulated and solved as a data-driven inverse optimization problem. Through three computational case studies, we demonstrated that:

1. Our approach produces high-quality surrogates with much simpler surrogate representations of the feasible regions of the original problem compared to traditional methods that involve optimization with embedded surrogate models. This key benefit arises from the decision-focused nature of our approach as it does not seek to learn the entire feasible solution space.
2. Simpler models of the feasible region lead to convex surrogate optimization problems, which obviates the need for expensive global solvers while still generating solutions that are close to globally optimal solutions.
3. Compared to black-box models used as optimization proxies, our approach is significantly more data-efficient, allowing the user to retain a large part of the original optimization problem that does not contribute to the problem's nonconvexity.
4. Our framework produces an optimization problem as the resulting surrogate model, making it easier to incorporate essential system constraints as hard constraints, which is typically not straightforward with black-box optimization proxies based on, for example, neural networks.

In summary, our decision-focused surrogate modeling paradigm presents a promising new avenue for the solution of time-critical optimization problems, offering both higher quality and more efficient surrogate optimization problems in comparison to traditional surrogate modeling methods. In a follow-up work,⁴² we extended

the proposed framework to include a mechanism that minimizes potential infeasibility (with respect to the original problem) of the DFSOM's solutions. Several important directions for future work still remain, including the nonparametric construction of convex surrogate models to replace the nonconvex parts of the original problem, and the development of adaptive sampling algorithms to further improve data efficiency.

AUTHOR CONTRIBUTIONS

Rishabh Gupta: conceptualization (equal); data curation (lead); formal analysis (equal); investigation (lead); methodology (lead); software (lead); validation (lead); visualization (lead); writing – original draft (lead); writing – review and editing (equal). **Qi Zhang:** conceptualization (equal); formal analysis (equal); funding acquisition (lead); methodology (supporting); supervision (lead); writing – original draft (supporting); writing – review and editing (equal).

ACKNOWLEDGMENTS

The authors thank Sayandeep Biswas and Joshua Larson, who were at the time undergraduate students at the University of Minnesota, for conducting preliminary computational analysis for this article. The authors gratefully acknowledge the financial support from the National Science Foundation under Grant #2044077 as well as the Minnesota Supercomputing Institute (MSI) at the University of Minnesota for providing resources that contributed to the research results reported in this article. Rishabh Gupta acknowledges financial support from a departmental fellowship sponsored by 3M and a Doctoral Dissertation Fellowship from the University of Minnesota.

DATA AVAILABILITY STATEMENT

The computer code and datasets that support the findings of this study are available under the MIT license at <https://github.com/ddolab/DecFocSurrMod>. Specifically, plots in Figure 2 can be regenerated by running the Julia code in the “cstr” folder. Steps to reproduce Figures 3 and 4 are shown in the `plots.ipynb` Jupyter notebook in the same folder. The Julia and Python code to reproduce results in Figure 6 is available in the “heat_exchanger” folder. Finally, the input files to regenerate the results in Figures 7 and 8 as well as Table 2 are available in the “blending_network” folder. The computer code in the GitHub repository also makes available the exact algorithmic hyperparameter settings used for the results shown in this article.

ORCID

Rishabh Gupta  <https://orcid.org/0000-0002-7940-2701>

Qi Zhang  <https://orcid.org/0000-0001-8862-4675>

REFERENCES

1. Ahuja RK, Orlin JB. Inverse optimization. *Oper Res.* 2001;49(5): 771-783.
2. Biegler LT, Lang Y, Lin W. Multi-scale optimization for process systems engineering. *Comput Chem Eng.* 2014;60:17-30.
3. Boukouvala F, Muzzio FJ, Ierapetritou MG. Predictive modeling of pharmaceutical processes with missing and noisy data. *AIChE J.* 2010; 56(11):2860-2872.

4. Jia Z, Davis E, Muzzio FJ, Ierapetritou MG. Predictive modeling for pharmaceutical processes using kriging and response surface. *J Pharma Innov.* 2009;4(4):174-186.
5. Jones D. A taxonomy of global optimization methods based on response surfaces. *J Global Optim.* 2001;21(4):345-383.
6. Boukouvala F, Muzzio FJ, Ierapetritou MG. Dynamic data-driven modeling of pharmaceutical processes. *Ind Eng Chem Res.* 2011;50(11):6743-6754.
7. Caballero JA, Grossmann IE. An algorithm for the use of surrogate models in modular flowsheet optimization. *AIChE J.* 2008;54(10):2633-2650.
8. Heo Y, Zavala VM. Gaussian process modeling for measurement and verification of building energy savings. *Energy Build.* 2012;53:7-18.
9. Henaou CA, Maravelias CT. Surrogate-based superstructure optimization framework. *AIChE J.* 2011;57(5):1216-1232.
10. Lee JH, Shin J, Realf MJ. Machine learning: overview of the recent progresses and implications for the process systems engineering field. *Comput Chem Eng.* 2018;114:111-121.
11. Schweidtmann AM, Mitsos A. Deterministic global optimization with artificial neural networks embedded. *J Optim Theory Appl.* 2019;180(3):925-948.
12. Asprion N, Böttcher R, Pack R, et al. *Graybox Models - New Opportunities for the Optimization of Entire Processes.* Vol 40. Elsevier Masson SAS; 2017.
13. Boukouvala F, Hasan MM, Floudas CA. Global optimization of general constrained grey-box models: new method and its application to constrained PDEs for pressure swing adsorption. *J Global Optim.* 2017;67(1-2):3-42.
14. Cozad A, Sahinidis NV, Miller DC. A combined first-principles and data-driven approach to model building. *Comput Chem Eng.* 2015;73:116-127.
15. Cozad A, Sahinidis NV, Miller DC. Learning surrogate models for simulation-based optimization. *AIChE J.* 2014;60(6):2211-2227.
16. Wilson ZT, Sahinidis NV. The ALAMO approach to machine learning. *Comput Chem Eng.* 2017;106:785-795.
17. Basudhar A, Dribusch C, Lacaze S, Missoum S. Constrained efficient global optimization with support vector machines. *Struct. Multidiscipl. Optim.* 2012;46(2):201-221.
18. Ibrahim D, Jobson M, Li J, Guillén-Gosálbez G. Optimal design of flexible heat-integrated crude oil distillation units using surrogate models. *Chem Eng Res Des.* 2021;165:280-297.
19. Ierapetritou MG. New approach for quantifying process feasibility: convex and 1-D quasi-convex regions. *AIChE J.* 2001;47(6):1407-1417.
20. Goyal V, Ierapetritou MG. Framework for evaluating the feasibility/operability of nonconvex processes. *AIChE J.* 2003;49(5):1233-1240.
21. Goyal V, Ierapetritou MG. Determination of operability limits using simplicial approximation. *AIChE J.* 2002;48(12):2902-2909.
22. Banerjee I, Ierapetritou MG. Feasibility evaluation of nonconvex systems using shape reconstruction techniques. *Ind Eng Chem Res.* 2005;44(10):3638-3647.
23. Zhang Q, Grossmann IE, Sundaramoorthy A, Pinto JM. Data-driven construction of convex region surrogate models. *Optim Eng.* 2016;17(2):289-332.
24. Schweidtmann AM, Weber JM, Wende C, Netze L, Mitsos A. Obey validity limits of data-driven models through topological data analysis and one-class classification. *Optim Eng.* 2022;23:855-876.
25. Halemane KP, Grossmann IE. Optimal process design under uncertainty. *AIChE J.* 1983;29(3):425-433.
26. Bhoosekar A, Ierapetritou M. Advances in surrogate based modeling, feasibility analysis, and optimization: a review. *Comput. Chem. Eng.* 2018;108:250-267.
27. Karg B, Lucia S. Efficient representation and approximation of model predictive control Laws via deep learning. *IEEE Trans Cybern.* 2020;50(9):3866-3878.
28. Krishnamoorthy D, Skogestad S. Real-time optimization strategies using surrogate optimizers. Paper presented at: Proceedings of the 2019 Foundations in Process Analytics and Machine Learning. 2019; Raleigh, North Carolina.
29. Kumar P, Rawlings JB, Wright SJ. Industrial, large-scale model predictive control with structured neural networks. *Comput Chem Eng.* 2021;150:107291.
30. Pan X, Zhao T, Chen M. DeepOPF: deep neural network for DC optimal power flow. Paper presented at: Proceedings of the 2019 IEEE International Conference on Communications, Control, and Computing Technologies for Smart Grids (SmartGridComm). 2019; Beijing, China.
31. Sun H, Chen X, Shi Q, Hong M, Fu X, Sidiropoulos ND. Learning to optimize: training deep neural networks for interference management. *IEEE Trans Signal Process.* 2018;66(20):5438-5453.
32. Zamzam AS, Baker K. Learning optimal solutions for extremely fast ac optimal power flow. Paper presented at: Proceedings of the 2020 IEEE International Conference on Communications, Control, and Computing Technologies for Smart Grids (SmartGridComm). 2020; Tempe, AZ.
33. Fioretto F, Mak TWK, van Hentenryck P. Predicting AC optimal power flows: combining deep learning and Lagrangian dual methods. Paper presented at: Proceedings of the AAAI Conference on Artificial Intelligence. New York; 2020:630-637.
34. Velloso A, Van Hentenryck P. Combining deep learning and optimization for preventive security-constrained DC optimal power flow. *IEEE Trans Power Syst.* 2021;36(4):3618-3628.
35. Kotary J, Fioretto F, Van Hentenryck P. Fast approximations for job shop scheduling: a Lagrangian dual deep learning method. Paper presented at: Proceedings of the AAAI Conference on Artificial Intelligence. 2022;36(7):7239-7246.
36. Chen S, Saulnier K, Atanasov N, et al. Approximating explicit model predictive control using constrained neural networks. In: Lin Z, (ed.) *Proceedings of the American Control Conference.* AACC; 2018:1520-1527.
37. Paulson JA, Mesbah A. Approximate closed-loop robust model predictive control with guaranteed stability and constraint satisfaction. *IEEE Control Syst Lett.* 2020;4(3):719-724.
38. Alessio A, Bemporad A. A survey on explicit model predictive control. *Lect. Notes Control Inform. Sci.* 2009;384:345-369.
39. Pistikopoulos EN, Diangelakis NA, Oberdieck R, Papanthanasios MM, Nascu I, Sun M. PAROC - an integrated framework and software platform for the optimisation and advanced model-based control of process systems. *Chem Eng Sci.* 2015;136:115-138.
40. Oberdieck R, Diangelakis NA, Nascu I, et al. On multi-parametric programming and its applications in process systems engineering. *Chem Eng Res Des.* 2016;116:61-82.
41. Rios LM, Sahinidis NV. Derivative-free optimization: a review of algorithms and comparison of software implementations. *J Global Optim.* 2013;56(3):1247-1293.
42. Gupta R, Zhang Q. Decision-focused surrogate modeling with feasibility guarantee. In: Yamashita Y, Kano M, (eds.) *Computer Aided Chemical Engineering.* Vol 49. Elsevier; 2022:1717-1722.
43. Wilder B, Dilkina B, Tambe M. Melding the data-decisions pipeline: decision-focused learning for combinatorial optimization. In: Van Hentenryck P, Zhou Z-H, (eds.) *33rd AAAI Conference on Artificial Intelligence, AAAI 2019, 31st Innovative Applications of Artificial Intelligence Conference, IAAI 2019 and the 9th AAAI Symposium on Educational Advances in Artificial Intelligence.* Vol 2019. EAAI; 2019:1658-1666.
44. Elmachtoub AN, Grigas P. Smart "predict, then optimize". *Manage Sci.* 2022;68(1):9-26.
45. Mandi J, Demirović E, Stuckey PJ, Guns T. Smart predict-and-optimize for hard combinatorial optimization problems. Paper presented at: AAAI 2020 - 34th AAAI Conference on Artificial Intelligence. New York; 2020:1603-1610.

46. Donti PL, Amos B, Kolter JZ. Task-based end-to-end model learning in stochastic optimization. *Advances in Neural Information Processing Systems* 30. 2017.
47. Chan TC, Mahmood R, Zhu IY. Inverse optimization: theory and applications. *arXiv preprint arXiv:2109.03920*. 2021.
48. Gupta R, Zhang Q. Decomposition and adaptive sampling for data-driven inverse linear optimization. *INFORMS J Comput*. 2022;34(5):2720-2735.
49. Gupta R, Zhang Q. Efficient learning of decision-making models: a penalty block coordinate descent algorithm for data-driven inverse optimization. *Comput Chem Eng*. 2023;170:108123.
50. Dempe S, Zemkoho A. Bilevel optimization. In: Zemkoho A, Dempe S, (eds.) *Springer Optimization and Its Applications*. Vol 161. Springer; 2020.
51. Aswani A, Shen Z-JM, Siddiq A. Inverse optimization with noisy data. *Oper Res*. 2018;66(3):870-892.
52. Nocedal J, Wright S. *Numerical Optimization*. Springer Science & Business Media; 2006.
53. Shen X, Diamond S, Udell M, Gu Y, Boyd S. Disciplined multi-convex programming. In: Wen C, Jiang Z-P, (eds.) *29th Chinese Control and Decision Conference (CCDC)*. IEEE; 2017:895-900.
54. Wächter A, Biegler LT. On the implementation of an interior-point filter line-search algorithm for large-scale nonlinear programming. *Math Program*. 2006;106:25-57.
55. Economou CG, Morari M, Palsson BO. Internal model control: extension to nonlinear system. *Ind Eng Chem Process Des Dev*. 1986;25(2):403-411.
56. Dunning I, Huchette J, Lubin M. Jump: a modeling language for mathematical optimization. *SIAM Rev*. 2017;59(2):295-320.
57. Abadi M, Agarwal A, Barham P, et al. *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems (Software)*. tensorflow.org; 2015.
58. Cecon F, Jalving J, Haddad J, et al. Omlt: optimization & machine learning toolkit. *J Mach Learn Res*. 2022;23(349):1-8.
59. Biegler LT, Grossmann IE, Westerberg AW. *Systematic Methods for Chemical Process Design*. Prentice Hall; 1997.
60. Adhya N, Tawarmalani M, Sahinidis NV. A lagrangian approach to the pooling problem. *Ind Eng Chem Res*. 1999;38(5):1956-1972.
61. Gupte A, Ahmed S, Dey S, Cheon M. Pooling problems: an overview. *Optimization and Analytics in the Oil and Gas Industry*. International Series in Operations Research and Management Science. Springer; 2015.
62. Misener R, Floudas CA. Advances for the pooling problem: Modeling, global optimization, and computational studies. *Appl Comput Math*. 2009;8(1):3-22.
63. Tawarmalani M, Sahinidis NV, Tawarmalani M, Sahinidis NV. The pooling problem. *Convexification and Global Optimization in Continuous and Mixed-Integer Nonlinear Programming: Theory, Algorithms, Software, and Applications*. Springer; 2002:253-283.
64. Foulds LR, Haugland D, Jørnsten K. A bilinear approach to the pooling problem. *Optimization*. 1992;24(1-2):165-180.

How to cite this article: Gupta R, Zhang Q. Data-driven decision-focused surrogate modeling. *AIChE J*. 2024;70(4): e18338. doi:[10.1002/aic.18338](https://doi.org/10.1002/aic.18338)