Received: 14 December 2020 | Revised: 1 June 2021 | Accepted: 15 June 2021

DOI: 10.1002/aic.17358

PROCESS SYSTEMS ENGINEERING

AIChE JOURNAL

An integrated dimensionality reduction and surrogate optimization approach for plant-wide chemical process operation

Thomas R. Savage[1,2] | Fernando Almeida-Trasvina[1] | Ehecatl A. del-Rio Chanona[3] | Robin Smith[1] | Dondga Zhang[1,3]

[1]Centre for Process Integration, The Mill, University of Manchester, Manchester, UK

[2]Department of Chemical Engineering and Biotechnology, West Cambridge Site, University of Cambridge, Cambridge, UK

[3]Centre for Process Systems Engineering, Roderic Hill Building South Kensington Campus London, London, UK

Correspondence
Dondga Zhang, Centre for Process Integration, The Mill, University of Manchester, Manchester M1 3AL, UK.
Email: dongda.zhang@manchester.ac.uk

Abstract

With liquefied natural gas becoming increasingly prevalent as a flexible source of energy, the design and optimization of industrial refrigeration cycles becomes even more important. In this article, we propose an integrated surrogate modeling and optimization framework to model and optimize the complex CryoMan Cascade refrigeration cycle. Dimensionality reduction techniques are used to reduce the large number of process decision variables which are subsequently supplied to an array of Gaussian processes, modeling both the process objective as well as feasibility constraints. Through iterative resampling of the rigorous model, this data-driven surrogate is continually refined and subsequently optimized. This approach was not only able to improve on the results of directly optimizing the process flow sheet but also located the set of optimal operating conditions in only 2 h as opposed to the original 3 weeks, facilitating its use in the operational optimization and enhanced process design of large-scale industrial chemical systems.

KEYWORDS

dimensionality reduction, Gaussian process, liquefied natural gas production, operational optimization, surrogate modeling

1 | INTRODUCTION

Large scale production of liquefied natural gas (LNG) typically involves the use of complex and energy-intensive cascade refrigeration cycles. Each cascade configuration comprises of a precooling cycle and a liquefaction cycle: the former cycle cools the natural gas stream to temperatures usually below −20°C, whereas the latter cycle cools and liquefies the natural gas stream down to its target temperature, near −161°C. The current LNG industry is strongly dominated by the propane precooled mixed refrigerant (C3MR) cycle, licensed by Air Products & Chemicals Inc.,[1] with over 70% of the LNG installed capacity worldwide utilizing this process.[2] The C3MR cycle uses propane as refrigerant in the precooling cycle, and a mixed refrigerant in the liquefaction cycle. Two other cascade cycles that are commercially well-established in the LNG industry are Shell's dual mixed refrigerant (DMR) cycle,[3] which uses two mixed refrigerants, and the ConocoPhillips Cascade cycle,[4] which uses three pure component refrigerants: propane in the precooling cycle, and ethylene and methane in the liquefaction cycle.

Specifically, the shaft work energy required for refrigerant compression is by far the largest contributor to operating costs (up to 40%–50%) throughout the entire LNG plant.[2] Reducing the shaft work demand associated with a refrigeration cycle would thus likely bring significant savings in energy as well as operating costs. As a

result, many research publications have focused on applying stochastic search-based optimization methods to reduce the overall shaft work demand in these cascade cycles via first-principle rigorous modeling. For instance, Alabdulkarem et al.[5] optimized the C3MR cycle using a genetic algorithm to achieve 9% savings in shaft work compared to the base case simulation. Wang et al.[6] also minimized the shaft work demand of the C3MR cycle, but using the nonlinear optimization algorithm embedded within HYSYS commercial software. Hwang et al.[7] optimized the DMR cycle to minimize shaft work demand using a genetic algorithm and sequential quadratic programming (SQP), yielding energy savings of 1.2% against an optimized example provided in.[8] Khan et al.[9] also optimized the DMR cycle for minimum shaft work demand, using a stochastic search algorithm known as the Box method, to evaluate two different inlet conditions (temperature and pressure) of the natural gas stream. Table 1 summarizes previous works that have aimed to improve the energy efficiency of the commercial cascade cycles.

While great effort has been made to improve existing commercial cascade cycles, Almeida-Trasvina and Smith[16] proposed a novel configuration named the CryoMan Cascade cycle, developed to be competitive against these commercial benchmarks. This CryoMan Cascade cycle showed potential to achieve shaft work savings of over 13% compared to the C3MR cycle, and up to 5% compared to the DMR cycle.[16] This CryoMan Cascade cycle was developed by making structural modifications in the liquefaction cycle, and adjusting the composition of the mixed refrigerant to reduce the overall shaft work demand. Despite the efforts of previous works to provide thorough analyses and clear systematic methodologies for energy efficiency improvements, optimization of these cascade cycles is still severely restricted by several practical challenges.

The first challenge arises from the complexity of the rigorous mathematical model. For example, a single cascade has 13 degrees of freedom (e.g., refrigerant flow rates, refrigerant composition, evaporating pressures) for both the precooling and liquefaction cycles, and these operating variables interact with each other through highly nonlinear expressions (e.g., thermodynamic equations of state for gas–liquid equilibrium systems) within process simulators. This therefore implies a large number of nonlinear mathematical constraints for the optimization problem. In addition, compared to many other optimization problems in the process industry, optimizing a cascade cycle is particularly difficult as its feasible input region is disjoint (i.e., several discrete feasible subregions) due to a number of feasibility constraints such as minimum temperature differences are imposed, impeding the use of many gradient-based deterministic optimization algorithms (e.g., trapped within a local feasible subregion thus only finding a local optimum). As a result, stochastic search-based optimization algorithms are predominantly used in the previous studies.

TABLE 1 Publications on energy efficiency improvements of commercial liquefied natural gas (LNG) cascade cycles

Reference	Cascade cycle	Shaft work demand ($MW \cdot MTPA^{-1}$ LNG)
Alabdulkarem et al.[5]	C3MR	32.34
Mortazavi et al.[10]	C3MR	34.07
Wang et al.[6]	C3MR	48.17
Wang et al.[11]	C3MR	60.25
Fahmy et al.[12]	C3MR	–
Wang et al.[13]	C3MR	45.37
	DMR	39.67
Hwang et al.[7]	DMR	27.08
Khan et al.[9]	DMR	38.80
Vink and Nagelvoort[14]	C3MR	33.42
	DMR	34.25
	Phillips Cascade	38.63
Fahmy et al.[15]	Phillips Cascade	37.56

Although previous studies have shown stochastic optimization methods (e.g., genetic algorithms) to be effective to find relatively good solutions, the use of these algorithms directly introduces the second challenge, namely high computational time cost. As these algorithms require substantial samplings and iterations over the entire input space, the overall optimization procedure can become time consuming (e.g., taking from days to weeks) if the rigorous model consists of a high dimensional input space and is difficult to converge within each iteration (due to the complex mathematical structure). This becomes even more challenging when decisions of process operation must be updated over a short time period in a real plant due to the variability of natural gas (feedstock) composition. A potential solution to this challenge is the adoption of surrogate modeling and optimization.

Throughout previous decades, surrogate models (also known as meta-models or reduced order models) have been used to aid the design and optimization of chemical engineering processes. By replacing a computationally expensive rigorous model with a computationally tractable surrogate model, the optimization of chemical processes can be performed more rapidly and more efficiently. Typical surrogate models include polynomial functions,[17–19] artificial neural networks (ANNs),[20,21] Gaussian processes (GPs; Kriging models),[22–24] and radial basis functions[25–27] . For example, Palagi et al.[20] utilized neural networks to optimize an Organic Rankine cycle over seven design variables. Cabellaro and Grossman[28] utilized GP-based surrogate models to replace individual unit operations, while also recommending a maximum number of design variables as 9 or 10. Boukouvala and Ierapetritou[29] applied GPs for the optimization of a continuous pharmaceutical manufacturing case study over five input variables, highlighting the need for investigation into global optimization approaches with an increased number of decision variables (>9) and constraints (>6). Another approach is deterministic global optimization (DGO),[30–32] while ensuring high quality global solutions these approaches have been known to run into 10–100 s of hours in the case of neural network surrogate models thus are infeasible for operational optimization. DGO is known to be effective for problems with <10 design variables, however, suffers with increasingly large dimension, inhibiting their use on large-scale process flow sheets. Outside of chemical engineering Eriksson et al.[33] propose performing a series

of local Bayesian optimization routines in an attempt to enable surrogate modeling for high-dimensional problems, motivating a local approach to global optimization. With the adoption of surrogate models, it has been shown that it is possible to implement global deterministic optimization algorithms to identify a high-quality solution.

Nonetheless, a key challenge in the surrogate modeling-based chemical process optimization is the ability to accurately portray information over an increasing number of design variables, for example, process flow sheet scale (e.g., 29 input variables in this study) as opposed to unit operation scale (e.g., less than 10 variables as previous study[28] suggested). Due to the unknown disjoint feasible input space and complexity of the rigorous model, sampling data over a high dimensional space to generate a meaningful training dataset (e.g., consisting of a large portion of feasible solutions) for surrogate construction can become extremely time consuming. This means that the primary challenge for surrogate modeling-based optimization will switch from the selection of an efficient optimization algorithm to the construction of a reliable surrogate model. Therefore, to enable a responsive plant-wide optimization (i.e., capable of updating optimization solution rapidly), how to construct an accurate surrogate model and identify a high-quality solution within a short time window is of critical importance for practical applications.

As a result, in this article, we propose a framework integrating dimensionality reduction and surrogate optimization to effectively optimize the CryoMan Cascade LNG refrigeration cycle. Specifically, partial least squares (PLS) is used to reduce the input dimension in tandem with an array of GPs in order to accurately capture the nonlinearities of the CryoMan Cascade LNG refrigeration cycle (e.g., both the constraints and the objective function). This latent space-based surrogate model, that is incrementally updated in areas of low objective value (promising solution subregions), enables access to gradient-based optimization as well as improved computational time. Through the use of this approach, we not only improve on the optimal solution generated when directly optimizing the HYSYS flow sheet but most importantly finish this within 2 h compared to a total previous time of 3 weeks. This advantage directly enables its transferable use for general industrial operational optimization whereby an optimal solution must be updated within a short time window.

The main procedures of this article are as follows: a large-scale and computationally expensive model of the CryoMan Cascade refrigeration cycle is systematically decomposed into its fundamental outputs; then partial least squares is used to reduce the input dimension to GP surrogate models providing an efficient and reduced dimension representation of each process output; the subsequent surrogate model was optimized and improved using an iterative data-resampling regime; and eventually efficient and rapid optimization of large-input, highly nonlinear refrigeration cycles is enabled with computational time reduced by multiple orders of magnitude. Therefore, Section 2 will introduce the CryoMan cascade refrigeration cycle and provide a basis for the subsequent surrogate optimization to take place. Section 3 will cover the construction of surrogate models, detailing the incorporation of dimensionality reduction techniques. Section 4 describes the specific optimization problem, and the subsequent transcription to surrogate model space. Finally, Section 5 presents the results of the surrogate optimization of the CryoMan cascade cycle both with and without the use of resampling techniques, both of which are thoroughly compared with the result of the previous rigorous optimization scheme, demonstrating the efficiency of the currently proposed methodology.

2 | CRYOMAN CASCADE LNG REFRIGERATION CYCLE

2.1 | Introduction to the CryoMan Cascade cycle

Figure 1 shows the configuration of the CryoMan Cascade cycle. This new cascade configuration is developed based on the CryoMan process developed by Kim and Zheng at the University of Manchester,[34] a single mixed refrigerant (SMR) cycle for LNG production at small scale. The promising performance of the CryoMan SMR cycle, saving 8% in shaft work demand against the PRICO® cycle,[35] motivated the development of the CryoMan Cascade cycle. In terms of its configuration, the precooling cycle in the CryoMan Cascade configuration utilizes a "heavy" mixed refrigerant that comprises of ethane, propane and n-butane, to cool down the natural gas stream and the refrigerant stream from the liquefaction cycle. This heavy refrigerant provides cooling, in a series of two multistream heat exchangers (MSHEs), at two evaporating pressures (low pressure (LP) heavy refrigerant and high pressure (HP) heavy refrigerant). The liquefaction cycle uses a lighter range of hydrocarbons as refrigerant: methane, ethane and propane, plus nitrogen. This "light" mixed refrigerant is partially condensed in the precooling cycle. A phase separator splits the refrigerant into vapor and liquid streams; a portion of vapor is mixed with a portion of the liquid stream to create LP light refrigerant, and the remainder of each phase is mixed together to create HP light refrigerant. Thus, the resulting compositions of both LP light refrigerant and HP light refrigerant are different from those obtained by phase separation alone. Each refrigerant stream provides cooling at independent evaporating pressures. There are a total of six stages for refrigerant compression: three for compression of the heavy refrigerant, and three for compression of the light refrigerant.

Almeida-Trasvina and Smith[16] assessed the performance of the CryoMan Cascade cycle, in terms of energy efficiency, against three commercial benchmarks, the C3MR, DMR and ConocoPhillips Cascade cycles. All four cascade cycles were optimized for minimum shaft work demand. The performance of the CryoMan Cascade cycle (25.97 MW·MTPA LNG^{-1}) was equivalent to shaft work savings of 13.7% and 4.9% compared to the C3MR cycle (30.09 MW·MTPA LNG^{-1}) and the DMR cycle (27.32 MW·MTPA LNG^{-1}), respectively. The potential shaft work savings offered by the CryoMan Cascade cycle would represent minor structural modifications in comparison to the DMR cycle. These structural changes imply that the first MSHE in the liquefaction cycle may require a more sophisticated internal design, as it would now require the accommodation of two cold

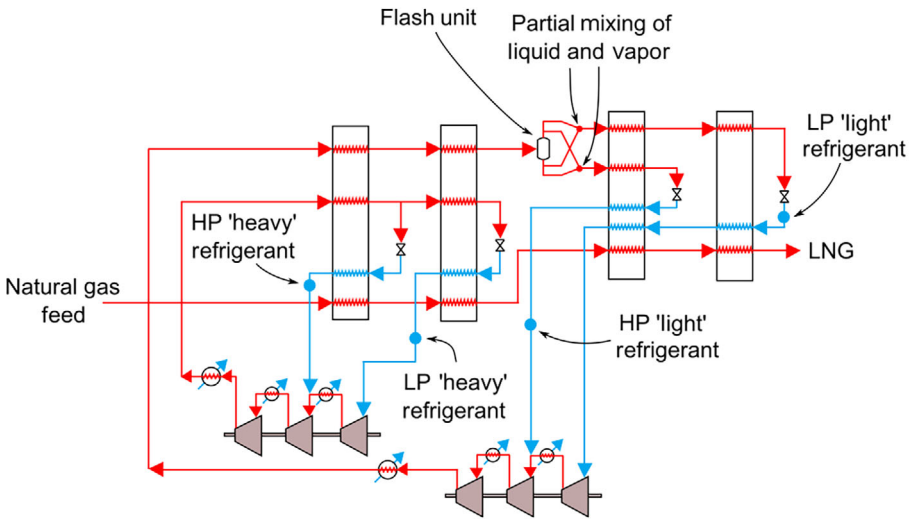

FIGURE 1 CryoMan Cascade cycle. LNG, liquefied natural gas [Color figure can be viewed at wileyonlinelibrary.com]

streams, each at a different pressure, to cool down the tube bundle of hot streams. These complications with the internal design of the MSHE mean that the CryoMan Cascade could be more attractive for grassroot designs, as opposed to retrofit projects. The CryoMan Cascade has not yet been applied industrially, but the potential economic benefits associated to its performance could be of enough industrial interest to develop this cascade configuration (saving up to billions of pounds annually estimated based on the current energy cost price).

The optimization in Almeida-Trasvina and Smith[16] was carried out using a genetic algorithm and a nonlinear optimization algorithm (SQP); the computational time required to perform the optimization was 17 h on average. As the feasible input region is disjoint (i.e., multiple discrete feasible subregions) due to the process constraints, sensitivity analysis was performed prior to optimization in order to narrow down the solution space. Sensitivity analysis took 3 weeks due to the large number of input variables (29 in this case). While the 1–2 s required for the multiple recycle loops inherent in the cascade cycle to converge can be seen as reasonably cheap, what causes the excessive optimization times mentioned here is the need for an evolutionary algorithm (EA) to locate a good initial solution for SQP optimization (as shown later in Section 4.2, 3000 samples were generated randomly, none of them was feasible). As evolutionary and other stochastic search optimization techniques require many function evaluations, this in combination with the 1–2 s evaluation time results in a large optimization time over 3 weeks.[16] In industrial practical application, however, updating optimal solution is often required to complete within hours (e.g., 4 h specified by Shell) given the frequent variation of natural gas composition.

2.2 | Introduction to the rigorous model

The full list of input and output variables in the CryoMan cascade refrigeration cycle is presented in Table 2. As shown in Table 2, the CryoMan cascade refrigeration cycle consists of 29 inputs and 13 outputs. The outputs concerned with the energy demand of the process, that is, those concerning shaft work, sum to produce the objective of the optimization problem. The remaining outputs such as the four MSHE approach temperatures, four vapor fractions, and four compression ratios, serve to enforce process constraints. The rigorous model consists of four MSHEs, six compressors, two pumps, and two flash units. Natural gas is assumed to enter at an ambient temperature of 24.85°C and atmospheric pressure. It exits the process under atmospheric pressure at −162.3°C as a liquid. The Peng–Robinson equation of state is used as the thermodynamic property model. The model makes use of two separate recycle operations within HYSYS which enable the mass and energy balance of recycle streams to converge before any data regarding the process is returned. The full list of the upper and lower bound of the input variables is presented in Table 2. A more detailed introduction to this rigorous model can be found in Section 3 in.[16] The original HYSYS code for the CryoMan cascade rigorous model can be shared upon request.

3 | METHODOLOGY

3.1 | Construction of GP-based surrogate models

The surrogate modeling procedure used takes advantage of the GP framework for function approximation. GPs specify a probabilistic model, with the method facilitating a major benefit over other surrogate modeling techniques such as ANNs through the evaluation of prediction uncertainties. Although uncertainties are not included in the current work, given that the data is generated by the rigorous model, this advantage can be extremely useful if process data is directly compiled from a genuine chemical plant in which significant measurement noise is embedded. In the absence of measurement noise, another key aspect is that GPs, with no assumed noise in measurements, provide interpolation through data points, and are therefore more sample efficient that other techniques such as ANNs. In other words, to construct an accurate GP, it requires much less data compared to that required by an ANN. As generating data from a rigorous model could be also time consuming, particularly if the feasible solution space is disjoint, this advantage allows a GP-

TABLE 2 Inputs and outputs of the CryoMan cascade refrigeration cycle

Inputs				
Heavy mixed refrigerant (HMR) precooling cycle	**HMR bounds**	**Light mixed refrigerant (LMR) liquefaction cycle**	**LMR bounds**	**Outputs**
HMR flow rate (kg/s)	[480, 600]	LMR flow rate (kg/s)	[300, 400]	HMR shaft work (1st stage)
HMR discharge pressure (bar)	[10, 20]	LMR discharge pressure (bar)	[25, 50]	HMR shaft work (2nd stage)
HMR flow rate split fraction (–)	[0.3, 0.2]	LMR HP evaporating pressure (bar)	[2.5, 8]	HMR shaft work (3rd stage)
HMR HP evaporating pressure (bar)	[3, 8.5]	LMR LP evaporating pressure (bar)	[1.2, 3]	HMR shaft work (pump)
HMR LP evaporating pressure (bar)	[1.2, 3]	NG stream precooling temp (1st MSHE) (K)	[125, 155]	LMR shaft work (1st stage)
HMR 2nd stage compression ratio (–)	[1, 3.5]	LMR HP stream precooling temp (1st MSHE) (K)	[125, 160]	LMR shaft work (2nd stage)
NG stream precooling temp (1st MSHE) (K)	[250, 270]	LMR LP stream precooling temp (1st MSHE) (K)	[125, 155]	LMR shaft work (3rd stage)
HMR stream precooling temp (1st MSHE) (K)	[250, 270]	LMR LP stream precooling temp (2nd MSHE) (K)	[105, 118]	LMR shaft work (pump)
LMR stream precooling temp (1st MSHE) (K)	[250, 270]	LMR LP stream outlet temp (1st MSHE) (K)	[195, 228]	ΔT_{min} HMR cycle (1st MSHE)
NG stream precooling temp (2nd MSHE) (K)	[215, 242]	LMR 2nd stage compression ratio (–)	[1, 3.5]	ΔT_{min} HMR cycle (2nd MSHE)
HMR stream precooling temp (2nd MSHE) (K)	[215, 242]	LMR flash vapor split fraction (–)	[0.6, 0.9]	ΔT_{min} LMR cycle (1st MSHE)
LMR stream precooling temp (2nd MSHE) (K)	[215, 242]	LMR flash liquid split fraction (–)	[0.4, 0.8]	ΔT_{min} LMR cycle (2nd MSHE)
HMR composition (Ethane) (mol fraction)	[0.15, 0.32]	LMR composition (Methane) (mole fraction)	[0.3, 0.47]	LMR vapor fraction inlet compressor (1st stage)
HMR composition (Propane) (mol fraction)	[0.38, 0.65]	LMR composition (Ethane) (mole fraction)	[0.25, 0.5]	LMR vapor fraction inlet compressor (2nd stage)
		LMR composition (Propane) (mole fraction)	[0.1, 0.27]	HMR vapor fraction inlet compressor (1st stage)
				HMR vapor fraction inlet compressor (2nd stage)
				LMR compression ratio (1st stage)
				LMR compression ratio (2nd stage)
				HMR compression ratio (1st stage)
				HMR compression ratio (2nd stage)

based surrogate model to be constructed within a comparatively short time period.[36]

A GP is defined as a collection of random variables, any finite number of which have a joint Gaussian distribution, specified by a mean function $m(\mathbf{x})$ and covariance function $k(\mathbf{x}, \mathbf{x}_*)$.[37] This mean function is often defined as a constant 0 for computational simplicity as well as functional flexibility through the specification of as little information regarding the underlying function as possible. By conditioning training data on this joint Gaussian distribution, a resulting function approximation can be derived.

$$f(\mathbf{x}_*) \sim \mathcal{GP}(\mathbf{0}, k(\mathbf{x}, \mathbf{x}_*)) \tag{1}$$

The use of a covariance function (or kernel) describes how the variance of two points in input space is effected by the distance between them. A common covariance function is the squared-exponential function as follows[37]:

$$k(\mathbf{x}, \mathbf{x}_*) = \sigma^2 \prod_{i=1}^{d} exp\left(-\theta_i \left(\mathbf{x}_i - \mathbf{x}_i^*\right)^2\right) \tag{2}$$

where $\mathbf{x} \in \mathbb{R}^d$ and $\mathbf{x}_* \in \mathbb{R}^d$ are two input vectors of design variables, $\sigma \in \mathbb{R}$ and $\theta \in \mathbb{R}^d$ are covariance function hyperparameters. The covariance function shown in Equation (2) is used to produce a covariance matrix of size $n \times m$ between two datasets $X \in \mathbb{R}^{d \times n}$ and $X_* \in \mathbb{R}^{d \times m}$ defined as $K(X, X_*)$, where n is the number of training points and m is the number of prediction points (i.e., new points of which the output needs to be predicted).

By conditioning the Gaussian prior distribution on the inputs and selected outputs of the CryoMan cascade refrigeration cycle, a

posterior distribution representing the underlying process model can be derived. The resulting surrogate model representing each process output has the form of a Gaussian distribution as follows[37]:

$$\mathbf{f}_* \mid X_*, X, \mathbf{f} \sim \mathcal{N}\left(K(X_*,X)K(X,X)^{-1}\mathbf{f}, K(X_*,X_*) - K(X_*,X)K(X,X)^{-1}K(X,X_*)\right) \tag{3}$$

where K is the covariance matrix relating input data X and data to be evaluated X_*, $\mathbf{f}$ is the corresponding output vector of input data, and $\mathbf{f}_*$ is the vector of function evaluations taken from the posterior distribution. Equation (3) is derived through conditioning the prior GP distribution (Equation 1) on a set of test data, resulting in the posterior distribution shown.[37] Observations of a GP derived from Equation (3) result not in scalar values but instead in mean and variances that describe probability distributions, in particular a Gaussian distribution, over possible function values.

Initially, data sampled from the rigorous HYSYS model is used to create and train a GP model to correlate the inputs and total shaft work (raw data can be shared upon request). As constraints also need to be enforced, GPs are used to model each individual constraint consisting of four minimum MSHE approach temperatures, a vapor fraction, and a compression ratio. A separate GP model must be created for each specific output, objective or constraint, as GPs can most commonly only predict single outputs. As a result, the current surrogate model in total consists of seven separate GPs, which use the same inputs (i.e., 29 design variables) but each returns an approximation of their respective rigorous output. This combination of GPs defines the surrogate model. The modeled outputs are listed below:

- Total shaft work (MW): *objective function*
- MSHE ΔT_{min} 1 (°C): *process constraint*
- MSHE ΔT_{min} 2 (°C): *process constraint*
- MSHE ΔT_{min} 3 (°C): *process constraint*
- MSHE ΔT_{min} 4 (°C): *process constraint*
- Refrigerant vapor fraction at inlet of single compressor: *process constraint*
- Compression ratio of single compressor: *process constraint*

It is worth noticing that for the outputs concerning process constraints, the constraint *violation* is modeled as opposed to the process output itself. A more detailed explanation of this implementation as well as its advantage is presented in Section 3.4.

3.2 | Dimensionality reduction

While the output dimension of the surrogate model has been reduced from 20 to 7 as described in Section 3.1, the input dimension remains high. Therefore, to reduce the number of inputs to each GP, take advantage of underlying relations between input variables and enable more efficient modeling, PLS, a widely used dimensionality reduction technique, is exploited in this work.

PLS is a machine learning method to reduce a large number of input variables to a smaller dimensional latent representation, specifically via a linear combination. By incorporating information about the respective outputs, the original input space can be reduced into a latent space with minimal loss of information. In this study, PLS performs dimensionality reduction through the projection of the input data matrix $\mathbf{X} \in \mathbb{R}^{n\times 29}$, where n is the number of data points, and output matrix $\mathbf{Y} \in \mathbb{R}^{n\times 1}$ (either the objective or a constraint) onto lower dimensional latent structures. The latent structures are chosen to maximize the covariance between each of the latent structures themselves, thus capturing the most information possible from this reduced dimension representation.[38] Mathematically, the general PLS transformation is as follows:

$$\mathbf{X} = \mathbf{T}\mathbf{P}^T + \mathbf{E} \tag{4}$$

$$\mathbf{Y} = \mathbf{U}\mathbf{Q}^T + \mathbf{F} \tag{5}$$

where $\mathbf{T} \in \mathbb{R}^{n\times l}$ and $\mathbf{U} \in \mathbb{R}^{n\times l}$ representing the latent projection of inputs $\mathbf{X}$ and outputs $\mathbf{Y}$ respectively to l dimensions as opposed to this original 29 in the case of the inputs. The output dimension in this case is not reduced as it is already at a minimum value of 1 pertaining to the output of each GP. $\mathbf{P} \in \mathbb{R}^{d\times l}$ and $\mathbf{Q} \in \mathbb{R}^{p\times l}$ represent loading matrices that perform the projection onto the latent input and output space respectively. $\mathbf{E}$ and $\mathbf{F}$ are the error terms.

To effectively reduce the input dimension meanwhile minimizing the loss information, in this study, the latent dimension l is chosen such that 95% of the variance between the inputs and the outputs is explained by the latent variables. This allows for the majority of information from the input space to be maintained while minimizing the amount of latent variables projected to. In this work, PLS only reduces the dimension of input variables $\mathbf{X}$, as the output variable $\mathbf{Y}$ of each GP is only one dimensional, thus cannot be further reduced. An advantage of the PLS method over other methods of unsupervised dimensionality reduction such as principal component analysis (PCA) is that PLS allows for the incorporation and consideration of the output variables. Whereas other unsupervised methods such as PCA project the input variables onto a latent space with no consideration onto their effect of the resulting outputs.

3.3 | Kriging partial least squares

Now attention is turned to combining the two previously mentioned modeling techniques of GPs and PLS. Kriging partial least squares (KPLS) is a GP architecture, introduced by Bouhlel et al.[39] Motivation for the method stems from the selection of hyperparameters $\boldsymbol{\theta}$ in Equation (2). As the length of the vector $\boldsymbol{\theta}$ increases with the number of input variables, the Gaussian process hyperparameter selection problem becomes increasingly more complex.

Through the integration of PLS within the covariance function (Equation 2), the two respective inputs $\mathbf{x}$ and $\mathbf{x}_*$ are projected onto a lower dimension latent space as shown in Equation (6). Subsequently,

there exists fewer hyperparameters and this reduced problem becomes considerably easier. This not only facilitates more rapid and efficient training of high dimensional GPs, but also enables a better choice of hyperparameters through a simplified optimization landscape.

$$k(\mathbf{x},\mathbf{x}^*)=\sigma^2\prod_{l=1}^{h}\prod_{i=1}^{d}exp\left[-\theta_l\left(\mathbf{w}_{*i}^{(l)}\mathbf{x}_i-\mathbf{w}_{*i}^{(l)}\mathbf{x}_i^*\right)^2\right] \tag{6}$$

Equation (6) shows the PLS extension of the squared exponential covariance function previously shown in Equation (2), where l is the dimension of the latent space and $\mathbf{w}^{(l)}$ are vectors derived from PLS describing the lth principal directions in X space that maximize the covariance between $\mathbf{X}$ and $\mathbf{y}$. h is the total dimension of the latent space (i.e., total number of principal components). For an extended discussion into KPLS interested readers can see.[39] In the context of updating each GP in the presence of new data, PLS weights are first calculated and subsequently optimal hyperparameters are found.

The final surrogate model structure is specified by 7, separate, reduced dimension GPs. Each representing either the total shaft work objective function, or a process constraint violation (Figure 2). This completes the construction of the current surrogate model. Specifically, implementation of the KPLS was carried out in Python 3.7.4 using the SMT Toolbox package[40] on an Intel i7 8th Gen processor.

3.4 | Optimization of the surrogate model

Once constructed, attention is now turned to the subsequent optimization of the CryoMan cascade cycle by exploiting the surrogate model. The objective function to be minimized is the total shaft work of the compressors within the process. A number of important constraints must be enforced to ensure the resulting set of operating conditions are not only physically realizable but also safe within an industrial context. The optimization problem is described as follows:

$$\min_{\phi}\left(\sum_{i=1}^{N}W_i(\phi)\right)/m_{\mathrm{LNG}} \tag{7}$$

$$\text{s.t.}\quad \Delta T_{\min}^{i}(\phi)\geq 2^oC\quad \Delta T_{\min}^{i}(\phi)\in \mathrm{MSHE}_{\Delta T_{\min}} \tag{8}$$

$$P_{\mathrm{rat}}(\phi)\leq 3.5 \tag{9}$$

$$\sum_{j=1}^{m}x_j(\phi)=1\quad x_j(\phi)\in X^{MR} \tag{10}$$

$$\mathrm{VF}^{\mathrm{ref}}(\phi)=1.00 \tag{11}$$

$$\mathrm{lb}\leq\phi\leq\mathrm{ub} \tag{12}$$

The specific shaft work shown in Equation (7) is defined as the sum of the individual shaft works $W_i(\phi)$ from each of process compressors divided by the mass flow rate of LNG m_{LNG}. This is the objective function to be minimized. The 29 inputs to the rigorous model, shown previously in Table 2, are represented by ϕ with the set of inputs having corresponding upper and lower bounds ub and lb, respectively (Equation 12). Constraints include: minimum approach temperatures for each of the MSHEs $\Delta T_{\min}{}^{i}(\phi)$ respectively where $\mathrm{MSHE}_{\Delta T_{\min}}$ is the set of all approach temperatures within the process (Equation 8) ensuring feasible heat transfer between refrigerant and natural gas streams, maximum compression ratios $P_{\mathrm{rat}}(\phi)$ (Equation 9) to discourage mechanical damage to compressors, valid molar compositions represented by $x_j(\phi)$ (Equation 10), and no wetness within

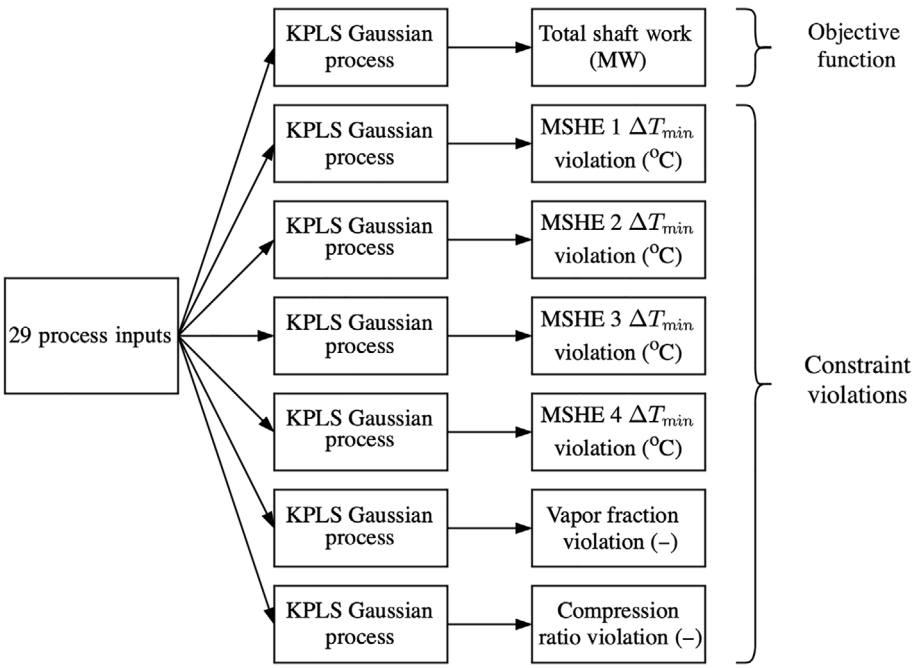

FIGURE 2 Surrogate model structure

compressors to reduce mechanical damage to turbine blades which is enforced by constraining vapor fractions at the inlet of compressors $VF^{ref}(\phi)$ to one (Equation 11).

The problem is transcribed to surrogate model space using a weighted penalty approach as follows:

$$\min_{\phi} \; GP_{SW}(\phi) + w_{con} \sum_{i=1}^{6} \max(GP_i(\phi), 0) \tag{13}$$

$$\text{s.t.} \;\; lb \le \phi \le ub \tag{14}$$

where GP_{SW} is the GP modeling the total specific shaft work. The use of a weighted penalty approach enables stochastic search optimization to be used to optimize the highly nonlinear process objective and constraints as opposed to traditional deterministic methods that solve the KKT matrix and may result in suboptimal local solutions. Each constraint GP returns a positive value if the respective constraint is violated, increasing linearly as the violation increases. Taking an MSHE approach temperature, for example, if the return temperature itself is 0.5°C, the GP will instead return a value of 1.5 representing the constraint violation. However, negative values representing feasibility can still be returned from the GP, that is, with an approach temperature >2°C. The magnitude of this negative value has no practical meaning as no feasible solution can be more feasible than another feasible solution. Thus, the max operator is used to limit values to either positive real numbers or 0.

By modeling the constraints in this way, the sharp transition in the GP landscape that results from modeling flat area of 0 violation to a linear penalty is avoided by instead modeling the complete positive and negative constraint landscape, and constricting penalties to positive values outside of the GP using the max operator. w_{con} specifies the magnitude of the constraint penalty which is able to be changed throughout the optimization scheme as described below in Section 3.5.

To optimize the surrogate model, an evolutionary algorithm is taken advantage of, enabling search throughout the high dimensional disjoint search space. Through the creation of a population of solutions over which operations such as mutations, crossover and selection occur, the bounded optimization problem described in Equations (13) and (14) can be solved. While evolutionary algorithms have been proven very successful on highly nonlinear problems, their convergence to a final solution can also be relatively passive. Therefore, following the evolutionary algorithm, the best solution within the population was used as an initial guess for a traditional Newton-based BFGS solver in order to gain a final set of decision variables. An advantage of GPs over neural networks is that when applying deterministic optimization solvers, it is computationally straightforward to calculate the gradient of a GP, particularly when the kernel function is squared exponential which is itself infinitely differentiable; whereas calculating the gradient of a neural network requires more computation. This was implemented in Python 3.7.4 using a custom evolutionary algorithm solver* and the package SciPy for the BFGS solver.

As previously mentioned, the variance of a prediction was not used within each GP prediction. This is due to the fact that while PLS aids the training of GPs, the number of input variables to each GP remains the same. With a relatively large number of optimization variables the variance term in a GP prediction (i.e., output) can dominate making its incorporation into the problem increasingly challenging. This motivates the use of a considered data resampling regime following the optimization of the surrogate model in order to maximize useful information obtained after each optimization.

3.5 | Iterative data resampling

In order to improve the accuracy of the surrogate model, while also converging to an optimal solution for Equation (13), an iterative resampling regime is taken advantage of in order to guide the surrogate model optimization toward promising regions.

The main aspect of a surrogate model resampling regime is to iteratively solve the optimization problem described by Equations (13) and (14), then evaluate this solution or some nearby solutions within the rigorous Aspen HYSYS model. This new dataset is then used to retrain the GP surrogate models based on the new information. The optimization is then performed again on the resulting surrogate model, and the entire procedure is iteratively repeated. Through this approach, information is gained regarding the underlying rigorous model in promising regions around plausibly optimal solutions (rather than sampling over 5.3×10^8 points to explore the entire high dimensional solution space). The resampling regime used in this study consists of three main features:

- **Annealing the penalty parameter w_{con} from a low to high value**. Initially the GP surrogate models are not completely accurate over the entire optimization landscape as both the objective function and the constraints have minimal information, resulting in solutions that are deemed feasible with respect to the surrogate constraints, however, when evaluated are in fact infeasible. By initializing w_{con} small, we allow the resampling regime to explore promising areas of low objective value and gain information about constraint violations. Subsequently, as this weight is increased throughout the overall resampling regime, the optimization scheme transitions from an exploratory nature to the more rigid enforcement of constraint violations in order to gain a valid final solution.
- **Local resampling**. When a promising solution is located, additional samples are generated from the Aspen HYSYS model around this solution. This local region is defined as ±5% of the overall bounds centered around this promising solution. Only 50 data points are sampled using an efficient, space-filling, sampling regime within this local region. To further investigate this region, GPs are constructed using only this resampled data, and optimized within these constricted bounds. The secondary solution generated from this subproblem is used to update the original dataset along with a proportion of the 50 resampled data points selected at random. This allows for this promising region to be modeled more accurately in

subsequent iterations. If a false positive situation occurs and there is no feasible solution or rather no feasible solution of worth in this region then the scheme simply returns to optimizing over the complete original bounds.

- **Refining solution**. If the subsequent solution after local resampling and optimization is feasible and still has a comparatively low objective value, a third phase begins. Similar to the local resampling, another 50 data points are generated from the Aspen HYSYS model, this time within more constricted bounds of ±2.5% of the overall bounds centered around the most recent data point. Constructing a further refined GP and performing optimization within further constricted bounds allows for a final solution to be obtained, as an accurate representation of the nonlinearities of the rigorous model are able to be captured.

It is vital to emphasize that the introduction of this three-phase resampling approach is attributed to the high dimensional, nonlinear and disjoint solution space caused by a series of hard feasibility constraints, thus traditional resampling frameworks mainly applied to low dimensional continuous optimization problems are not applicable to the current study. A flowchart detailing the overall surrogate optimization scheme is detailed in Figure 3.

In this work, data resampling was implemented in Python 3.7.4 which was linked to Aspen HYSYS using the COM interface allowing for interaction between the rigorous model and the optimization implementation within Python.

4 | RESULTS AND DISCUSSION

To demonstrate the efficiency of the proposed dimensionality reduction and surrogate modeling framework, two case studies are provided to optimize the CryoMan Cascade refrigeration cycle. The first case study will test the accuracy of the surrogate model with the use of prior domain knowledge in the form of constricted bounds on the input variables to the process. These constricted bounds are generated at a significant time cost through extensive sensitivity analysis.

Due to this time-consuming sensitivity analysis step, the second case study will test the surrogate model without the incorporation of prior domain knowledge, that is, over the complete range of possible

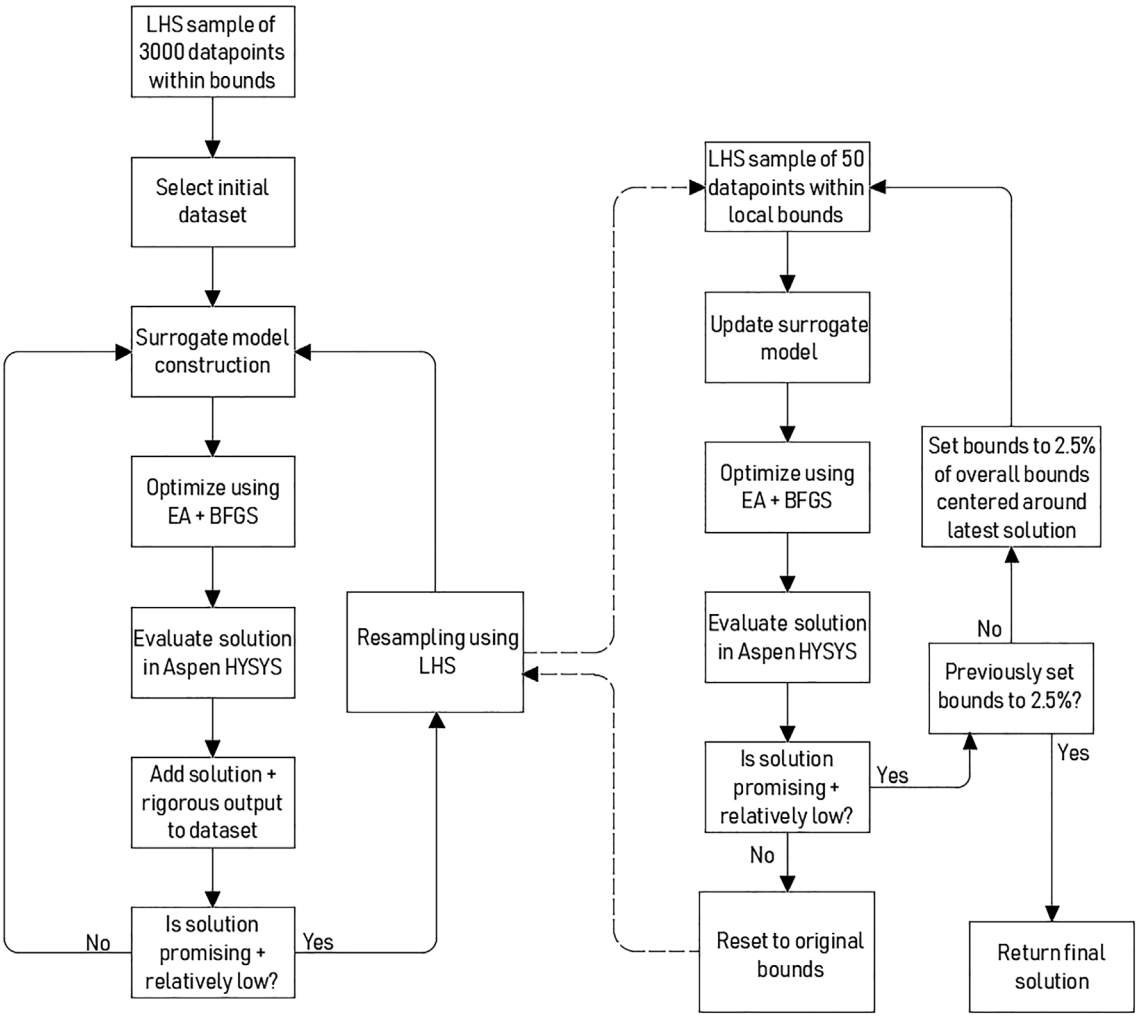

FIGURE 3 Flowchart describing the overall surrogate model resampling optimization scheme. Abbreviations: LHS, Latin hypercube sample; BFGS, Broyden–Fletcher–Goldfarb–Shanno

operating conditions. Subsequently, the resampling regime described in Section 3.5 is incorporated into the optimization scheme in order to gain an accurate optimal solution.

4.1 | Case study 1—Surrogate modeling with prior domain knowledge

Within the first case study the structure of the surrogate model was constructed using data points sampled within the narrow search space identified from the prior domain knowledge (i.e., our previous sensitivity analysis[16]). Extensive sensitivity analyses were carried out on the degrees of freedom to identify a region of the solution space in which the CryoMan Cascade cycle performs well in terms of energy efficiency. Both the precooling and liquefaction cycle were analyzed. Shaft work demand and feasibility in the MSHE ($\Delta T \geq \Delta T_{min}$) were used as indicators to help define the solution space to be searched with the genetic algorithm. These sensitivity analyses include parametric studies in which a single degree of freedom is varied while the rest remained the same. Studies on refrigerant composition were also performed: the mole fraction of each component was allowed to changed, while the relative contribution of the remaining components is kept fixed. For a given initial composition, Equation (15) calculates the relative contribution ϕ_j of each component x_j; Equation (16) calculates the mole fraction of components x_j as component x_i is varied. Interactions between degrees of freedom were also accounted for by simultaneously varying two degrees of freedom, for example, refrigerant composition and refrigerant condensing pressure.

$$\phi_j = \frac{x_j^{\text{initial}}}{\left[1 - x_i^{\text{initial}}\right]} \quad x_i, x_j \in X^{\text{MR}} \tag{15}$$

$$x_j^{\text{calc}} = \phi_j \cdot \left[1 - x_i^{\text{varied}}\right] \tag{16}$$

As a result, once the narrow solution space is found, only 350 feasible data points (generated at considerable computational cost approximately 40 min as most points sampled are still infeasible within this region) and 500 infeasible data points distributed throughout this 29-dimensional subspace are used to construct the surrogate model. No resampling regime is used as prior knowledge has been incorporated for model construction. The model is directly used to optimize the process. The evolutionary algorithm employed for surrogate model optimization used tournament selection with a tournament size of 2 to combat elitism within the population and maintain global representation. A single crossover was used to generate new generations. The mutation rate was set as 0.05 and the probability of an individual surviving to the next generation was set as 0.95. The number of generations was set as 600 after a number of test runs assessing convergence. The optimal solution generated through the surrogate model was then evaluated within the rigorous Aspen HYSYS simulation and the accuracy of the outputs compared. The results are as follows with Table 3, demonstrating the optimal output as predicted from the surrogate model when compared to the verification from the rigorous Aspen HYSYS model.

With the rigorous optimization conducted in our previous work, directly optimizing the Aspen HYSYS model returned a valid solution with a respective shaft work of 144.7 MW.[16] In spite of the slightly lower (i.e., 1.8%) total energy cost compared to that identified by the surrogate model, this rigorous solution was obtained in a time of approximately 17 h as opposed to the surrogate model time of 1 h (40 min for data sampling and 10 min for surrogate model construction and optimization) conducted in this study. Due to the frequent feedstock variation, updating the process operating conditions must be completed within a short time frame (e.g., 4 h in industrial plants). As a result, the computational cost (i.e., 17 h) required for rigorous model optimization is not practical. Using the surrogate model-based approach as presented in this work, on the other hand, is much more promising.

From Table 3, it can be seen that the surrogate solution obtains a solution with shaft work similar to that of the rigorous optimization, albeit slightly higher. The optimal solution identified in this model successfully satisfied all the constraints with the exception of mispredicting the vapor fraction on entry to the third compressor (VF_3). This constraint can be considered important with respect to the mechanical integrity of compressors. Any liquid on entry will cause mechanical damage contributing to possibly catastrophic and unsafe scenarios as well as increased long-term maintenance costs. All other constraints after rigorous validation remain feasible. Most importantly of these are the highly nonlinear temperature differences. Any positive real value can represent feasible heat transfer; however, a minimum of 2°C is enforced to ensure economic viability with respect to the size of the MSHEs. It is possible for HYSYS to return a negative value for approach temperature which would represent a physical impossibility.

From Table 3, it is concluded that not only is the solution gained from the surrogate model physically realizable but also feasible with respect to the minimum approach temperature. The results show that

TABLE 3 Results of the Kriging partial least squares surrogate model superstructure without resampling

	SW_{total}	VF_3	P_3^{rat}	$\Delta T_{min\ 1}$	$\Delta T_{min\ 2}$	$\Delta T_{min\ 3}$	$\Delta T_{min\ 4}$
Unit	**(MW)**	**(−)**	**(−)**	**(K)**			
Constraint	**NA**	**=1**	**≤3.5**	**≥2**			
Surrogate model prediction	146.3	0.994	3.01	2.60	3.99	2.52	2.82
Aspen HYSYS output	147.4	0.983	3.05	2.73	2.92	2.10	2.23

Notes: The optimal solution predicted by this surrogate model is validated against the rigorous model. VF stands for vapor fraction, and P_{rat} stands for compression ratio.

GPs can be used to accurately represent the total shaft work as well as accurately represent highly nonlinear constraints for the purposes of optimization. To ensure reliability and safety within an industrial setting, small errors in the prediction such as that which occur with the vapor fraction in this case study, must be avoided. It can be hypothesized that the introduction of resampling may reduce these small errors, specifically around promising regions.

However, despite the promising result, the current surrogate model is built upon data carefully sampled from a narrow range determined using the prior domain knowledge. As explained before, obtaining this domain knowledge is extremely time consuming (i.e., 3 weeks for the comprehensive sensitivity analysis). As a result, when considering the prior domain knowledge integrated within this solution in the form of the refined data sampling and optimization bounds, this approach can be seen as infeasible when considering the overall timeframe to gain the solution (i.e., 3 weeks for sensitivity analysis and 1 h for surrogate modeling and optimization) if such prior knowledge does not exist. Therefore, to extend the applicability of the current surrogate modeling strategy and to gain a solution not only comparable with the original rigorous optimization but also without the necessity of prior domain knowledge (i.e., over the complete initial sampling and optimization bounds), a resampling regime is essential to enable accurate operational optimization of large-scale LNG refrigeration cycles. Nevertheless, this initial case study has confirmed that the CryoMan cascade cycle can be accurately modeled by a set of reduced dimension GPs.

4.2 | Case study 2—Surrogate modeling and data resampling

The second case study will now test the model accuracy and optimization scheme *without* the incorporation of prior domain knowledge, that is, without reduced optimization bounds as well as without the initially selected dataset. Firstly, only 3000 data points were sparsely collected from the broad 29-dimensional solution space (approximately 1.3 points sampled per dimension) using a Latin hypercube sample, taking approximately 1 h. The initial number of points was chosen due to the aforementioned time constraint of 4 h relating to operational optimization, allowing for an optimal solution to be located within the remaining hours. This is then reduced to 100 data points based on minimal constraint violation. From this initial sampling, we noticed that none of the original 3000 data points were feasible in the context of the optimization problem, illustrating the need for a resampling regime without prior information and that the underlying feasible solution space is not continuous. The surrogate model was constructed using the 100 infeasible data points and was subsequently exploited to predict the initial optimal operation condition.

The resampling optimization scheme shown in Figure 3 was then used to generate optimal candidates. If a solution gained is deemed promising with respect to the constraint violation, that is, does not violate any constraints or lies near the boundary of a constraint, then it is further investigated through additional sampling of the rigorous model in this region. Initially, new bounds of 5% of the original bounds are used for resampling, centered around the previously generated promising solution. However, if after another iteration of optimization, this resampling the solution remains promising, the bounds are again lowered to 2.5% and sampling is performed. This layered approach allows for promising regions, particularly around the boundary of constraints, to be modeled and explored to sufficient accuracy. In total, 43 resampling and optimization iterations were performed consisting of either just the addition of a single data point, a sampled set of refined data points, or a sampled set of further refined data points (decided by the resampling framework as shown in Figure 3). The total resampling time for the scheme was 37 min until the surrogate model identified a better solution compared to the previous rigorous optimal result.[16] When considering the time to generate the initial sample of an hour, the total time for the entire surrogate optimization without prior domain knowledge was 97 min.

As previously mentioned, an advantage of GPs over other methods of surrogate modeling is that uncertainty can be obtained from a prediction. However, in this situation due to the large number of design variables, the uncertainty term was found to dominate the objective function as the optimization is forced to explore the entire search space containing large areas of uncertainty, therefore slowing down convergence as significantly more resampling iterations would be needed to converge on an optimal solution. Therefore, only the mean value of each of the GP predictions was used. The initial average prediction error of the surrogate model across the complete search space was calculated as 4.3%; however, it should be noted that this technique only needs to improve the accuracy of the objective in areas of interest, identified via optimization and thus a global prediction accuracy is of less importance than the result of the final solution. Finally, the results for this optimization scheme are as follows.

From Table 4, it is concluded that with the use of the surrogate optimization regime a slightly better optimal solution was able to be found compared the rigorous optimization approach (i.e., saving 0.2% energy cost). When comparing Table 3 and Table 4, within the first

TABLE 4 Results of the Kriging partial least squares surrogate model superstructure with resampling

Unit	SW_{total} (MW)	VF_3 (–)	P_3^{rat} (–)	$\Delta T_{min\ 1}$ (K)	$\Delta T_{min\ 2}$	$\Delta T_{min\ 3}$	$\Delta T_{min\ 4}$
Constraint	(–)	=1	≤3.5	≥2			
Surrogate model prediction	144.81	0.995	3.52	2.06	2.00	2.05	2.11
Aspen HYSYS output	144.37	0.995	3.50	2.01	2.00	2.02	2.04

Notes: The optimal solution predicted by this surrogate model is validated against the rigorous model. VF stands for vapor fraction, and P_{rat} stands for compression ratio.

TABLE 5 Optimized design variables of case study 1 and case study 2

Design variable	Unit	Case study 1	Case study 2
HMR flowrate	(kg/s)	550	600
HMR discharge pressure	(bar)	14.5	12.9
HMR flow rate split fraction	(–)	0.659	0.645
HMR HP evaporating pressure	(bar)	3.77	4.36
HMR LP evaporating pressure	(bar)	1.21	1.48
HMR 2nd stage compression ratio	(–)	1.79	2.18
NG stream precooling temp (1st MSHE)	(K)	261	262
HMR stream precooling temp (1st MSHE)	(K)	260	263
LMR stream precooling temp (1st MSHE)	(K)	262	262
NG stream precooling temp (2nd MSHE)	(K)	230	233
HMR stream precooling temp (2nd MSHE)	(K)	233	232
LMR stream precooling temp (2nd MSHE)	(K)	234	231
HMR composition (ethane)	(mole fraction)	0.152	0.171
HMR composition (Propane)	(mole fraction)	0.510	0.536
LMR flow rate	(kg/s)	388	379
LMR discharge pressure	(bar)	27.9	26.1
LMR HP evaporating pressure	(bar)	2.71	2.96
LMR LP evaporating pressure	(bar)	1.59	1.52
NG stream precooling temp (1st MSHE)	(K)	150	148
LMR HP stream precooling temp (1st MSHE)	(K)	154	156
LMR LP stream precooling temp (1st MSHE)	(K)	147	151
LMR LP stream precooling temp (2nd MSHE)	(K)	110	109
LMR LP stream outlet temperature (1st MSHE)	(K)	226	221
LMR 2nd stage compression ratio	(–)	3.47	3.50
LMR flash vapor split fraction	(–)	0.833	0.813
LMR flash liquid split fraction	(–)	0.482	0.514
LMR composition (nitrogen)	(mole fraction)	0.0620	0.0590
LMR composition (methane)	(mole fraction)	0.367	0.356
LMR composition (ethane)	(mole fraction)	0.346	0.344

Abbreviations: HMR, heavy mixed refrigerant; HP, high pressure; LMR, light mixed refrigerant; LP, low pressure; MSHE, multistream heat exchanger; NG, natural gas.

case-study the model was unable to find a solution wherein all constraints are active. This can be attributed to the lack of resampling, resulting in slightly inaccurate predictions on the boundary of the process constraints inhibiting the optimization algorithm from finding a solution with active constraints. However, in the second case-study with the inclusion of resampling it can be seen that the solution does involve a number of active constraints. Through the annealing of the constraint penalty parameter ω_{con}, the algorithm located a number of solutions on the boundary of, or near a constraint. As the representation of these constraint functions iteratively improved with more datapoints, the constraints were able to be modeled more accurately and a solution on the boundary was located.

Most importantly, the resampling framework identified the final optimal solution within only 36 min. When considering the time taken to generate the training data points, the overall surrogate modeling and optimization time from start to finish was 97 min, significantly shorter than that spent for rigorous optimization (i.e., over 3 weeks including sensitivity analysis to reduce search space and 17 h for stochastic optimization). This directly suggests that the current surrogate modeling and resampling framework can enable an efficient operational optimization and online planning for the CryoMan cascade refrigeration cycle when its operating conditions are required to update within a short time frame.

As shown in Table 4, it can be seen that the prediction error of the outputs is much less than the first case study, owing to an increased amount of data in promising regions, generated by the resampling regime. The objective, and all constraints are able to be modeled accurately, allowing for a solution to be found on the boundary of each constraint where an optimal solution would traditionally be expected to be found. When comparing the first and the second case study, it can be seen that the introduction of resampling enables more accurate predictions with respect to both the shaft work as well as the constraints allowing for improved optimization results. The resulting shaft work value improves on the shaft work value without

resampling by 2.09%, saving a significant amount of energy. The inclusion of this framework can be attributed to this reduced shaft-work value, as most commonly optimal solutions are observed on the boundary of constraints (i.e., constraints are active). In producing a solution on the boundary of constraints we can be reasonably confident that this framework has located an improved optimal solution through a better representation of these constraint functions as opposed to this happening per-chance. However, with the introduction of resampling comes associated drawbacks, namely the increased computational cost of generating additional samples. There is clearly a tradeoff that must be balanced, too many additional samples will make the resampling regime inefficient and only undermines the process of surrogate modeling. While the lack of resampling completely may result in infeasible operating conditions or a suboptimal solution.

In an industrial setting, if prior knowledge is available, for example, in an existing well-established LNG refrigeration process, then optimizing a surrogate model itself will be the better approach due to increased time savings. However, if the process is still under investigation or many process alternatives are being considered at the design stage, then the integration of resampling and surrogate optimization can significantly decrease the time for process optimization without the need of prior knowledge. This will accelerate the procedure to evaluate and develop novel configurations which must be optimized to ensure a fair comparison. Finally, the optimal operating conditions identified in case study 1 and case study 2 are summarized in Table 5.

5 | CONCLUSIONS

The optimization of the energy demand for large-scale refrigeration cycles is an important problem, particularly when considering the design, online planning, and operational optimization of promising configurations such as that of the CryoMan cascade refrigeration cycle. Current rigorous model-based optimization schemes have runtimes on the order of weeks for a single refrigeration cycle, slowing down development times and preventing the use of operational optimization which must take place every few hours. This article combines the following techniques within a surrogate modeling framework in order to reduce the optimization time from 3 weeks to 2 h while also improving on the solution gained when optimizing the rigorous model:

- GPs to model process objectives and constraints;
- PLS to reduce the input dimension to each GP;
- evolutionary algorithm to optimize nonlinear, computationally economic surrogate model; and
- an iterative data resampling regime to improve the accuracy of surrogate model and guide subsequent optimization.

Based on significant improvements in process optimization time, operational optimization and improved rapid assessment of process alternatives is enabled for large-scale LNG production processes. Therefore, significant energy savings can be made not only when compared to existing LNG systems but also when compared to the current optimization techniques used to evaluate novel refrigeration cycles. Assessing the accuracy of surrogate models, built upon datapoints sampled in specific regions is an aspect of the work we wish to note. Due to this behavior induced by our specific methodology, the surrogate model is indeed accurate in optimal regions of search space. However, achieving global surrogate model accuracy is not the goal in this methodology and creating a perfect global approximation may bring about original issues that we wish to address such as computational expense and extreme nonlinearity. The optimization of processes implemented within process simulators is a considered task, to ensure success many moving parts were combined in this work. By starting with a difficult case-study, we encountered many issues that were incrementally addressed while performing this work; however, there is no standard procedure for large-scale surrogate optimization as of yet. We hope that by presenting the aspects we deemed most useful in this work, such as the use of dimensionality reduction, more attention will be paid to the use of surrogate optimization for larger, more complex systems and concepts which find success across a broad number of case studies will emerge. While success in this domain does rely on the overall iterative approach taken, it also relies on the choice of approximating model itself. With the rapid growth of machine learning particularly in the context of chemical engineering, there is significant scope for new structures such as graph neural networks, Bayesian neural networks and sparse GPs to name a few. Of particular relevance will be the performance of these new surrogate models over a large number of decision variables. The inclusion of hybrid, or physics informed data-driven models is also an interesting direction to further improve the accuracy of the surrogate model. This also raises a number of interesting questions regarding how to maintain the computational benefits of a surrogate model while also including a computationally cheap and tractable source of prior physical knowledge. A possible route for future work may also involve the investigation of real-time optimization of large-scale LNG plants, as well as how specific approaches into the optimization of the subproblem effects convergence. The methodology presented here is also transferable to generic large-scale (e.g., high dimensional) physically constrained processes. Investigating the problem under the framework of DGO may also be interesting, particularly as the large number of input variables provides a barrier to existing approaches.

ENDNOTE

* Found at https://github.com/tomrsavage/evolutionaryalgorithm

DATA AVAILABILITY STATEMENT

Data available on request from the authors: The data that support the findings of this study are available from the corresponding author upon reasonable request.

ORCID

Thomas R. Savage https://orcid.org/0000-0001-8715-8369
Ehecatl A. del-Rio Chanona https://orcid.org/0000-0003-0274-

2852

Dondga Zhang https://orcid.org/0000-0001-5956-4618

REFERENCES

1. Gaumer S, Newton C. Combined cascade and multicomponent refrigeration system and method, 1970. Air Products and Chemicals Inc., US patent 3 763 658.
2. Mokhatab S, Mak JY, Valappil JV, Wood DA, eds. *Handbook of Liquefied Natural Gas*. Boston: Gulf Professional Publishing; 2014:147-183.
3. Grootjans HF, Nagelvoort RK, Vink KJ. Liquefying a stream enriched in methane, 2002. Shell Oil Company, US Patent 6 370 910 B1.
4. Houser CG, Yao J, Andress DL, Low WR. Efficiency improvement of open-cycle cascaded refrigeration process, 1997. Phillips Petroleum Company, US Patent 5 669 234.
5. Alabdulkarem A, Mortazavi A, Hwang Y, Radermacher R, Rogers P. Optimization of propane pre-cooled mixed refrigerant LNG plant. *Appl Therm Eng*. 2011;31(6–7):1091-1098.
6. Wang M, Khalilpour R, Abbas A. Operation optimization of propane precooled mixed refrigerant processes. *J Nat Gas Sci Eng*. 2013;15:19-105.
7. Hwang J-H, Roh M-I, Lee K-Y. Determination of the optimal operating condition of dual mixed refrigerant cycle of LNG FPSO topside liquefaction process. *Comput Chem Eng*. 2012;49:25-36.
8. Venkatarathnam G, Timmerhaus K, Rizzuto C. *Cryogenic Mixed Refrigerant Processes*. Vol 1; Berlin: Springer Science & Business Media; 2008.
9. Khan MS, Karimi IA, Lee M. Evolution and optimization of the dual mixed refrigerant process of natural gas liquefaction. *Appl Therm Eng*. 2016;96:320-329.
10. Mortazavi A, Somers C, Hwang Y, Radermacher R, Rodgers P, Al-Hashimi S. Performance enhancement of propane pre-cooled mixed refrigerant LNG plant. *Appl Energy*. 2012;93:125-131. (1) Green Energy; (2) Special Section from papers presented at the 2nd International Enery 2030 Conference.
11. Wang M, Zhang J, Xu Q, Li K. Thermodynamic-analysis-based energy consumption minimization for natural gas liquefaction. *Ind Eng Chem Res*. 2011;50(22):12630-12640.
12. Fahmy MFM, Nabih HI, El-Aziz MRA. Investigation and performance improvement of the propane precooling cycle in the propane precooled mixed refrigerant cycle liquefaction process. *Ind Eng Chem Res*. 2016;55(10):2769-2783.
13. Wang M, Khalilpour R, Abbas A. Thermodynamic and economic optimization of LNG mixed refrigerant processes. *Energy Convers Manage*. 2014;88:947-961.
14. Vink KJ, Nagelvoort RK. Comparison of baseload liquefaction processes. *International Conference & Exhibition on Liquefied Natural Gas*; Paris: International Institute of Refrigeration; 1998:308-318.
15. Fahmy M, Nabih H, El-Nigeily M. Enhancement of the efficiency of the open cycle Phillips optimized cascade LNG process. *Energy Convers Manage*. 2016;112:308-318.
16. Almeida-Trasviña F, Smith R. *Design and Optimisation of Novel Cascade Refrigeration Cycles for LNG Production*. Vol 43. Netherlands: Elsevier BV; 2018:621-626 Paper presented at the 28th European Symposium on Computer Aided Process Engineering, ESCAPE 28; Conference date: 10-06-2018 Through 13-06-2018.
17. Chávez-Hurtado JL, Rayas-Sánchez JE. Polynomial-based surrogate modeling of RF and microwave circuits in frequency domain exploiting the multinomial theorem. *IEEE Trans Microw Theory Tech*. 2016;64(12):4371-4381.
18. Içten E, Nagy ZK, Reklaitis GV. Process control of a dropwise additive manufacturing system for pharmaceuticals using polynomial chaos expansion based surrogate model. *Comput Chem Eng*. 2015;83:221-231.
19. Zhang Y, Sahinidis NV. Uncertainty quantification in CO_2 sequestration using surrogate models from polynomial chaos expansion. *Ind Eng Chem Res*. 2013;52(9):3121-3132.
20. Palagi L, Sciubba E, Tocci L. A neural network approach to the combined multi-objective optimization of the thermodynamic cycle and the radial inflow turbine for organic Rankine cycle applications. *Appl Energy*. 2019;237:210-226.
21. Perera A, Wickramasinghe P, Nik VM, Scartezzini J-L. Machine learning methods to assist energy system optimization. *Appl Energy*. 2019;243:191-205.
22. Liu B, Zhang Q, Gielen GGE. A Gaussian process surrogate model assisted evolutionary algorithm for medium scale expensive optimization problems. *IEEE Trans Evol Comput*. 2014;18(2):180-192.
23. Su G, Peng L, Hu L. A Gaussian process-based dynamic surrogate model for complex engineering structural reliability analysis. *Struct Saf*. 2017;68:97-109.
24. Xia W, Luo B, Liao X-p. An enhanced optimization approach based on Gaussian process surrogate model for process control in injection molding. *Int J Adv Manuf Technol*. 2011;56(9):929-942.
25. Regis RG. Stochastic radial basis function algorithms for large-scale optimization involving expensive black-box objective and constraint functions. *Comput Oper Res*. 2011;38(5):837-853.
26. Regis RG. Constrained optimization by radial basis function interpolation for high-dimensional expensive black-box problems with infeasible initial points. *Eng Optim*. 2014;46(2):218-243.
27. Regis RG, Shoemaker CA. Combining radial basis function surrogates and dynamic coordinate search in high-dimensional expensive black-box optimization. *Eng Optim*. 2013;45(5):529-555.
28. Caballero JA, Grossmann IE. An algorithm for the use of surrogate models in modular flowsheet optimization. *AIChE J*. 2008;54(10):2633-2650.
29. Boukouvala F, Ierapetritou MG. Derivative-free optimization for expensive constrained problems using a novel expected improvement objective function. *AIChE J*. 2014;60(7):2462-2474.
30. Demirhan CD, Tso WW, Powell JB, Pistikopoulos EN. Sustainable ammonia production through process synthesis and global optimization. *AIChE J*. 2018;65(7):e16498.
31. Huster WR, Schweidtmann AM, Lüthje JT, Mitsos A. Deterministic global superstructure-based optimization of an organic rankine cycle. *Comput Chem Eng*. 2020;141:106996.
32. Schweidtmann AM, Huster WR, Lüthje JT, Mitsos A. Deterministic global process optimization: accurate (single-species) properties via artificial neural networks. *Comput Chem Eng*. 2019;121:67-74.
33. Eriksson D, Pearce M, Gardner J, Turner RD, Poloczek M. Scalable global optimization via local Bayesian optimization. *Advances in Neural Information Processing Systems*; 2019:5496-5507.
34. Kim J-K, Zheng X. Refrigeration process, 2011. The University of Manchester, US Patent 9 562 717 B2.
35. Almeida-Trasvina F. *Development of Novel Refrigeration Cycles for Small Scale LNG Processes*. Manchester, UK: The University of Manchester; 2016.
36. Savage T, Almeida-Trasvina F, Chanona AD-R, Smith R, Zhang D. Surrogate modelling and optimization for complex liquefied natural gas refrigeration cycles. *IFAC-PapersOnLine*. 2020;53(2):11193-11198.
37. Williams CK, Rasmussen CE. *Gaussian Processes for Machine Learning*. Vol 2. Cambridge, MA: MIT Press; 2006.
38. Wold S, Sjöström M, Eriksson L. PLS-regression: a basic tool of chemometrics. *Chemom Intel Lab Syst*. 2001;58(2):109-130. PLS Methods.
39. Bouhlel M, Bartoli N, Morlier J, Otsmane A. Improving kriging surrogates of high-dimensional design models by partial least squares dimension reduction. *Struct Multidiscip Optim*. 2016;53:935.
40. Bouhlel MA, Hwang JT, Bartoli N, Lafage R, Morlier J, Martins JRRA. A python surrogate modeling framework with derivatives. *Adv Eng Softw*. 2019;135:102662.

How to cite this article: Savage TR, Almeida-Trasvina F, del-Rio Chanona EA, Smith R, Zhang D. An integrated dimensionality reduction and surrogate optimization approach for plant-wide chemical process operation. *AIChE J*. 2021; e17358. https://doi.org/10.1002/aic.17358