Optimizing the Performance of Analytical Chemistry Instrumentation

By

Srividya Sekar

Supervisors:

Dr. Victoria C.P. Chen

Dr. Jay M. Rosenberger

Committee Members:

Dr. Kevin A. Schug

Dr. Shouyi Wang

Dr. Chen Kan

Department of Industrial, Manufacturing and Systems Engineering,

The University of Texas at Arlington

May, 2022

Acknowledgments

First, I would like to thank my supervising professors Dr. Victoria Chen and Dr. Jay Rosenberger, for their encouragement, constant support and inspiration. They have been a second family to me, away from home, throughout the years I have been here and are one of the main reasons I decided to pursue my PhD at UTA.

I would also like to thank my committee member Dr. Kevin Schug, for helping me familiarize with the topics on analytical chemistry and answering all my questions patiently. I would like to extend my thanks to my other committee members Dr. Shouyi Wang and Dr. Chen Kan, for all their valuable inputs on the research. I would like to appreciate Shannon Thomas for her conscientiousness and for helping me collect data in the laboratory on time.

Additionally, I would like to thank Dr. Dwight Stoll for sharing the code for his ChromSim application with us and for providing training and support on using it.

I am very thankful to Dr. Ronald Cross, who has been a great mentor to me from my first day at UTA. He has always been there for me and has provided unending motivation and support throughout these years.

I am very grateful to my parents and my brother for all their sacrifices and encouragement to pursue my PhD degree and constantly showering me with their love and motivation every day, even from thousands of miles away. I am also very lucky to have been blessed with wonderful friends who have helped me a lot throughout this journey and for being the best support system, I could ask for.

I would like to dedicate this dissertation to my husband and best friend, Ashok. He has been extremely patient, loving and constantly pushing me to do better. I could not

imagine getting through this incredible journey successfully without his persistent help and motivation.

Finally, I would like to thank the National Science Foundation for partially supporting this research under Grant NSF-CHE-2108767 and the Trinity River Authority of Texas for funding me both as an intern and Graduate Research Assistant.

Abstract

Surrogate Optimization and Global Optimization approaches to optimize underlying functions have been studied and used extensively in the field of Operations Research. However, there are very few instances where these approaches have been applied and tested in applications with uncertainty. Additionally, extensive focus and effort has been put into developing highly complex metamodels rather than globally optimizing these metamodels.

In this study, we propose a Mixed Integer Quadratically Constrained Program (MIQCP) based approach that globally optimizes a Quintic Multivariate Adaptive Regression Splines (QMARS) metamodel. The QMARS-MIQCP based optimization is applied to a global optimization framework called QMARS-MIQCP-OPT to optimize several test functions with both lower and higher dimensions.

The QMARS-MIQCP based optimization is also applied to a surrogate optimization framework called QMARS-MIQCP-SUROPT. In any surrogate optimization algorithm, the process of exploring the space to find new points to evaluate plays a key role in the performance of the algorithm. Even though the traditional Exploration and Exploitation Pareto Approach (EEPA) works well with surrogate optimization algorithms that use EEPA to select candidate points for evaluation, because our proposed approach optimizes the space around the selected EEPA candidates, there is inherent exploration in the algorithm itself. Using traditional EEPA caused too much exploration, slowing down the algorithm's capability to find optimal solutions in fewer function evaluations. So, a modified Sorted EEPA approach was developed in the study. Additionally, two different candidate selection strategies are also studied. The QMARS-MIQCP-SUROPT algorithm is also applied to test the

performance when optimizing standard test functions that have been used as benchmarks in both surrogate optimization and global optimization applications.

Optimal parameter settings play an important role in the efficiency of analytical chemistry instrumentation. The application of the developed QMARS-MIQCP-SUROPT algorithm to analytical chemistry instrumentation will provide an abundance of knowledge about the instrument, eliminate trial-and-error runs, as well as help in reducing the sample preparation time and cost of materials used. The QMARS-MIQCP-SUROPT algorithm is applied to ChromSim, a simulation software developed by Dr. Dwight Stoll and team to optimize the parameter settings of a 2D – Liquid Chromatography (LC) system for efficient separation of different analytes. The developed QMARS-MIQCP-SUROPT algorithm is also applied to guide a series of real-world laboratory experiments to optimize the parameter settings of the Shimadzu Liquid Chromatography Mass Spectrometry (LCMS) 2020 instrument for efficient flow injection analysis of Acetaminophen.

List of Figures

List of Tables

Contents

Chapter 1

Introduction

The 2019 National Academies report, A Research Agenda for Transforming Separation Science, specifies in Chapter 3 that, "key fields of research that can provide insight and knowledge to advance separation science: materials synthesis, systems engineering, responses to external stimuli, instrumentation and characterization tools, and data science and analytics [1]." The proposed work will integrate statistics and operations research to establish an effective strategy for analysis of chemical compounds.

1.1 Overview of Proposed Research

The decision and response variables in the proposed research are grouped into three spaces: The first is the input space (X), which is usually a combination of Sample materials, target analytes and instrument properties. The next is decision space (Y), which is comprised of the variable parameters of the analytical chemistry instruments like Voltage, Pressure, Temperature, etc. Finally, we have the outcome space (Z), which is characterized by the extraction performance metrics detected by the equipment like Peak Area, Peak Width, Ionization Intensity, etc. These three spaces together represent the analytical chemistry instrument. The relationships between these three variable spaces are unknown and are likely nonlinear and non-convex.

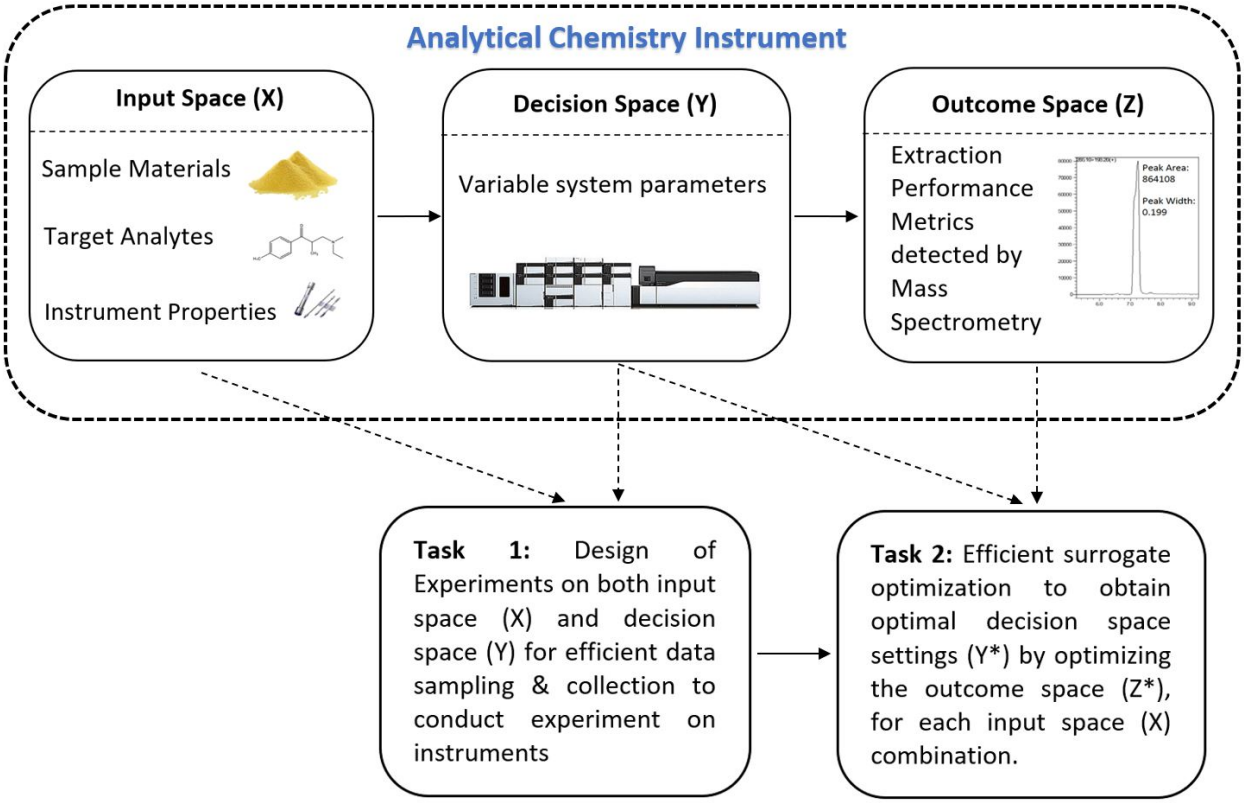

Figure 1.1: *Flowchart of proposed research*

A flowchart of the proposed research is given in Figure 1.1. The main aim of this research is to find optimal parameter settings in the decision space (Y) by maximizing one of the performance metrics in the outcome space (Z) for each combination of the input space (X). This will be achieved by performing a series of tasks.

Task 1: Design of Experiments (DoE): Using the descriptive features to represent the input space efficiently, the aim of this task is to create an efficient Design of experiments on both the input and decision space. The DoE on the input and Decision space will be incorporated within the surrogate optimization algorithm in task 2 to efficiently search the spaces.

Task 2: Surrogate Optimization: In this task, an efficient global surrogate optimization algorithm is developed to obtain optimal decision space settings by maximizing the outcome space for each combination of the input space.

1.2 Motivation

The motivation for this research can be looked at from two perspectives. One from the analytical chemistry side of it and another from the optimization side of it.

Looking at the analytical chemistry perspective, there are several new and advanced technologies that facilitate seamless analyte extraction, separation and detection that can be done while significantly reducing the sample preparation times and contamination. However, with every new technology, there is inherent lack of knowledge about the complexity of the system and its interactions with parameters. This leads to poor efficiency and undesirable results and calls for better understanding of the system and optimization of the parameters.

Looking at the optimization perspective, due to the nature of this application, collecting large amount data costs time and money. Surrogate optimization has been typically used with highly computational computer models and works well with small data [2, 3, 4, 5]. Surrogate models also do a great job in efficiently searching the space. There are several highly complex surrogate models that have been developed. A major shortcoming that needs to be addressed is that a lot of time and effort has been put into developing these complex meta models but not a lot of work has been done in globally optimizing these models. Most literature tend to use commercially available optimization toolboxes like MATLAB's FMINCON [2] and Surrogate Modeling (SUMO) toolbox [6, 7] or a pool of candidate solutions [3, 8, 9], which are unlikely to globally optimize these complex meta models.

1.3 Research Gap

In the optimization realm, there is an abundance of literature on surrogate optimization, and there is also an abundance of literature on global optimization. However, there is very little literature on global optimization of surrogate models. Additionally, there are very little literature that addresses optimization under uncertainty.

In the analytical chemistry realm, there is an abundance of literature on using DoE

and response surface methodology (RSM) in optimizing extraction parameters. There is also considerable literature on using predictive models to understand the relationship between the different extraction parameters. There are only a couple of instances that use surrogate optimization in analytical chemistry. However, all of them apply it to superstructure optimization.

Consequently, this research will focus on bridging the gaps between the two realms by creating a globally optimal surrogate optimization algorithm that can be applied to analytical chemistry instruments. A visual overview of the research gap is given in Figure 1.2.

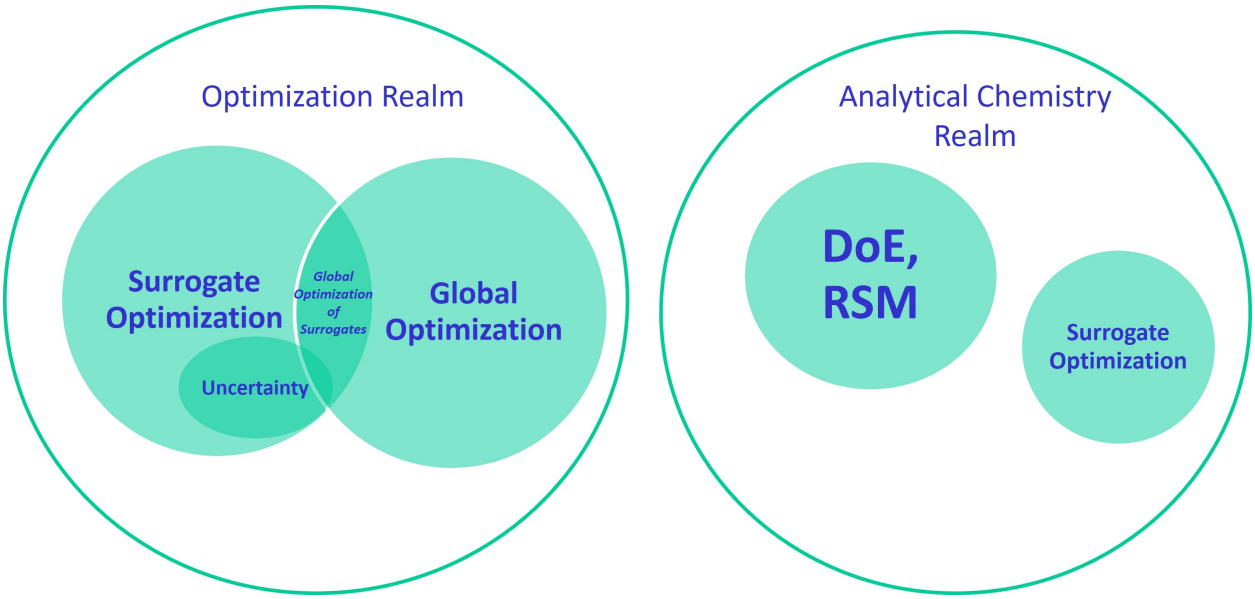

Figure 1.2: *Research gap*

1.4 Contribution

The proposed methodology aims at overcoming the above-mentioned shortcomings by using efficient DoE in two phases: (i) in the representation of the design space and (ii) a sequential DoE integrated within the surrogate optimization algorithm to effectively explore the design space. A smooth quintic MARS meta model developed by Chen [10] is used in combination with the optimization methodology discussed in Dickson [11] and Anahideh et al. [9].

Combining these, an MIQCP based surrogate optimization algorithm is developed. The proposed methodology is generic and can be implemented for a wide variety of applications that require global optimization and the use of small data.

The methodology is also tested on ChromSim, a simulation software and to guide a real-world laboratory experiment on the Shimadzu LCMS-2020 instrument. Implementing this algorithm to find optimal system parameter settings for efficient separation and flow injection analysis of analytes will provide an abundance of knowledge about the system. With the set of optimal system parameter settings for a given analyte, the sample preparation time is greatly reduced as well as the trial-and-error runs needed to achieve the optimal and efficient analysis of the analytes. A flowchart of contributions is given in Figure 1.3

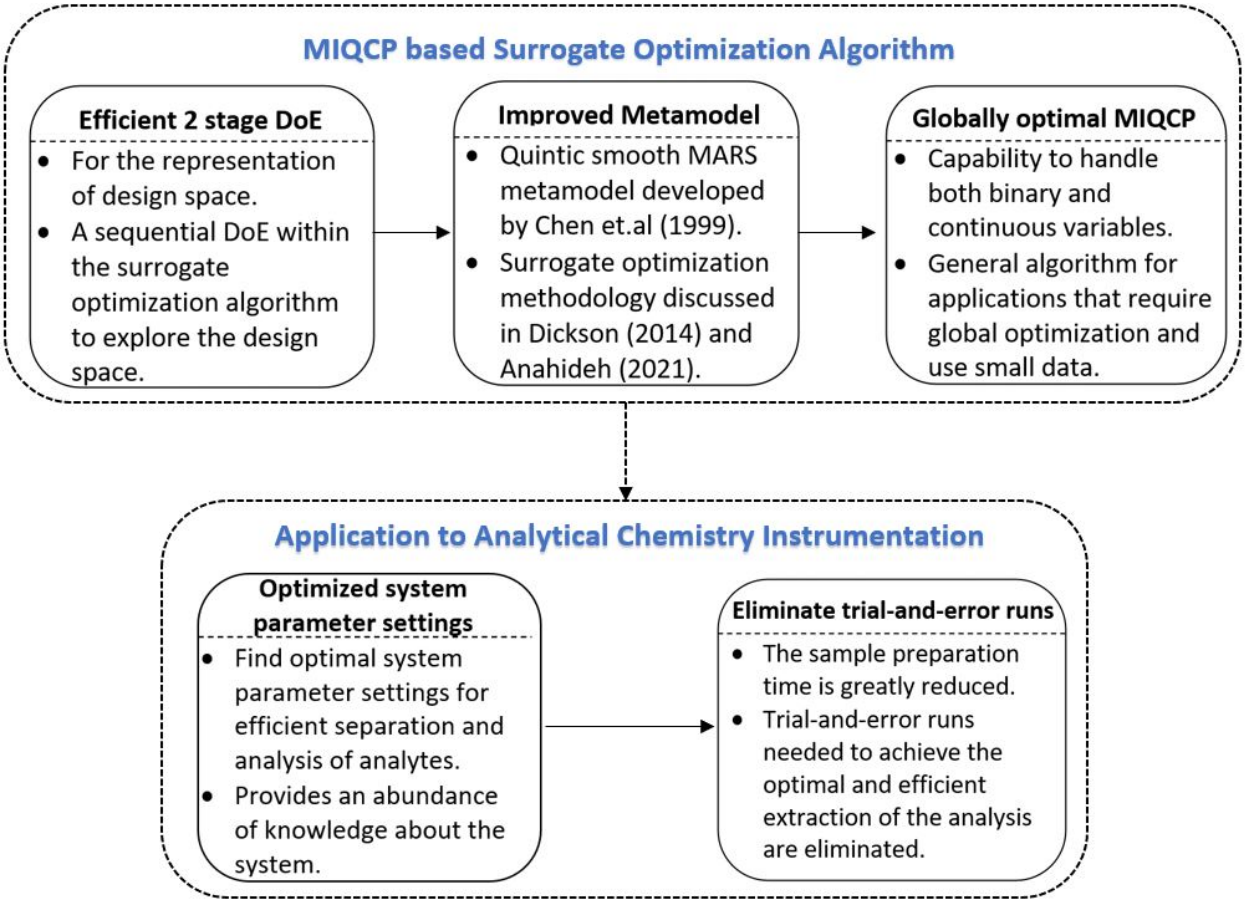

Figure 1.3: *Flowchart of contribution*

Chapter 2

Literature Review

The term optimization is the process of making a design, process, or decision fully perfect or efficient as possible. The term was first known to be used in 1857 [12]. Optimization has been used in a wide variety of applications and has drastically improved the way data are collected and decisions are made. Since its first use in 1857, the various methodologies and tools used in optimization have grown exponentially.

2.1 DoE and RSM in optimization

The concept of DoE was first developed by Ronald A. Fisher in 1926. It was first developed to arrange agricultural field experiments and to overcome the errors in measurements and environmental biases [13]. The techniques used in designing these experiments were the foundation of modern statistical principles. Over the years, this methodology gained a lot of popularity because of its ability to efficiently model the relationship between the input and the response variables and providing insights in running experiments and cutting costs. Because of its ability to clearly identify the effect of changes of one or several of the inputs simultaneously on the output, DoE has not only been used in designing physical experimental runs in laboratories or agricultural fields, but it has also been used to design computer experiments to run a wide range of computer simulations as well. There is an abundance of

literature in various applications in electrical engineering, chemical engineering, mechanical engineering, and dynamic programming [14, 15, 16, 17, 18]. It is commonly known as Design and Analysis of Computer Experiments (DACE).

An extension of the classic design of experiments is sequential design of experiments, which is a popularly used technique for selecting new points in the design space to evaluate. The main difference of sequential designs from the classic design of experiments is that the sample size and composition of the samples need not be fixed in advance, thereby reducing the number of function evaluations. Dodge and Romig [19] were one of the pioneers who recognized the drawback of the fixed sample size in the classic DoE and designed an experiment that used a double sampling method in industry quality control to identify defective parts in 1929. In 1940, another important contributor to this methodology was Mahalanobis [20], who also focused on the importance of designing and revising experiments as data are gathered and more knowledge about the system is gained. Following this in 1943, Wald [21] developed a more logical approach during World War II in which observations were made one by one, and the decision to stop sampling could be made at any point time. However, an important issue with the sequential design approach that needed addressing was the issue of deciding which population to select in each iteration in applications with multiple populations. Herbert [22] addresses the issue using a two-population application and provides a solution using Markov Chains. Similarly, Chernoff [23, 24] addresses the same issue for a two-population problem using the maximum likelihood at each iteration. Albert [25] in 1961 extends Chernoff's methodology from two states of nature to infinite states of nature and also defines an optimal strategy using game theory and a payoff matrix to determine which population to choose in the next iteration. There is not a lot of literature with contributions succeeding these developments on the methodology, but it has been popularly used in combination with the surrogate optimization algorithm. The common take away from this literature is that all of them provide a three-alternative hypothesis: to reject the null hypothesis, to fail to reject the null hypothesis or to continue the experiment at any stage of the experiment, unlike testing at the end of the experiment like in a classic DoE [26].

Even though, DoE was very efficient and useful in designing experiments, it was not created with the capability to optimize the response. Consequently, Response Surface Methodology (RSM) was developed by Box and Wilson in 1951 [27]. RSM uses a series of sequential DoE to identify the points and the regions to explore in the response surface. Box and Draper [28] emphasized the importance of the initial DoE in the performance of RSM and also suggested some DoE that can be used in the literature. RSM starts with a first-degree model of the response surface and as it gets closer to the maximum, the peak is modeled using a second-degree polynomial function to approximate the response surface. This concept was developed by Myers and Montgomery in 2002 [29]. An efficient metric to decide when to switch from the first-degree design to the second-degree design was then developed by Nuran in 2007 [30]. There are several second-order designs, but the most efficient second-order design is the CCD. One of the main advantages of RSM is its ability to optimize the response. As mentioned above, this is done by moving around the design space. Gradient descent has been found to be an efficient method to make a decision as to which direction to move iteratively to find an optimum response quickly [31]. There is an abundance of literature that review the methodology and its applications in various fields such as agriculture, chemical sciences, food sciences and engineering [32, 33, 34, 35, 36, 37, 38]. A major assumption in the literature when using RSM is that the form of the underlying function is known, which is not the case for most of the applications. Consequently, the solution obtained using this method may suboptimal. Another disadvantage in using RSM is that it can optimize the objectives/responses only one-at-at a time [36].

2.2 Surrogate optimization

Surrogate optimization is a technique that uses meta models or surrogate models to optimize the response the surface. Surrogate optimization is a more advanced version of RSM that has capability to learn and model the underlying response surface. In applications where the underlying model of the response surface is unknown, also known as black box systems,

surrogate optimization has been widely known to decipher the structure of the underlying surface by making use of iterative DoE, fitting and validation of surrogate models [39, 9] This methodology has been widely known to be used in applications involving simulations with computation times in hours, days or even weeks. The major advantage of using the surrogate optimization technique is its capability to converge to a solution with comparatively fewer simulation or experimental runs and computation time [40].

It is also noted that surrogate optimization has been widely used in hyper parameter optimization of machine learning models and has been commonly referred to as Bayesian Optimization in most machine learning journals [41, 42, 43].

Based on literature that uses surrogate optimization [44, 45, 40, 46, 39], a generalized structure of the surrogate optimization procedure and the key components in the surrogate optimization technique are identified. They are initialization, running computer simulations/laboratory experiments, building surrogate models, finding new points to run the simulation/experiments, and evaluation of stopping criteria. A basic process flow chart of the steps in surrogate optimization is denoted in figure 2.1.

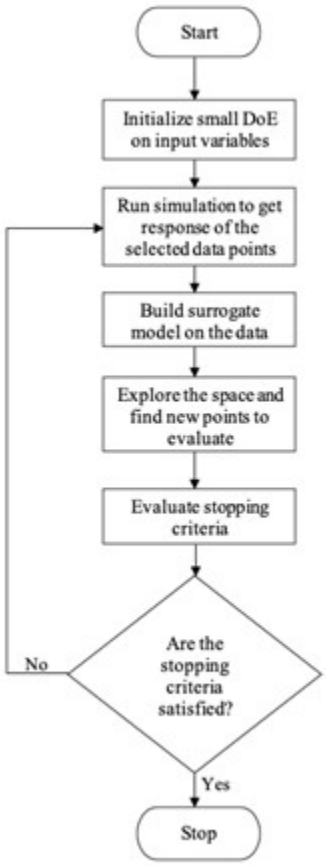

Figure 2.1: *Flowchart of surrogate optimization based on literature*

2.2.1 Initialization

The performance of the surrogate model depends on how well the input space is represented. The most common initialization techniques found in the literature are studying the input space and using space filling designs such as Orthogonal Arrays (OA) [47, 48], Latin Hypercube Designs (LHD) [49, 50], Sobol sequence [51, 52], etc. The main aim of using the surrogate optimization technique is to reduce the number of experimental runs. Keeping that in mind, in this research, Sobol sequence and LHD are used are used to test the performance of the methodology.

2.2.2 Surrogate Models

A surrogate model or metamodel is a mathematical approximation of the relationship between the input parameters and response variable. Selecting or developing a model that can closely mimic the response surface yields the best performance in the optimization.

Response surface models have been used as surrogate models in several applications because of their ability to closely approximate computationally expensive simulations with low order polynomials and their relatively easy comprehensibility. It has been predominantly used in applications with low dimensions of about 10-15 variables. It is seen that as the dimension of the problem increases, the computational cost increases and the accuracy of the model in approximating the underlying function decreases [53]. To overcome this disadvantage, Kaufman et al. [54] suggests applying constraints, typically using an upper and lower bound on the design space, enabling a more efficient construction of the response surface. Response surface models have also been used in multi-objective optimization applications because of their various advantages. However, they are incapable of efficiently optimizing multi-objective problems. Consequently, response surface models have been used in combination with evolutionary algorithms. The literature proposes using response surface models to approximate the pareto frontier, because they are computationally inexpensive, and then using evolutionary algorithms to optimize the pareto frontier [55, 56]. It has also been used in combination with simulated annealing by creating a quadratic approximation and globally optimizing the objective function, by gradually reducing the design space [57].

Kriging is an interpolating model that has been widely used as a surrogate model because of its ability to perform better than polynomial regression models. It was first developed by Danie Krige to find the concentration of ores in gold mines [14]. A major advantage of Kriging models stated in the literature is the possibility of controlling the variance of the basis function [44]. There are several variations of the Kriging model: Ordinary Kriging, Simple Kriging, universal Kriging, and Bayesian Kriging [44, 58, 59]. Even though the basis of Kriging is assumed to be continuous between 0 and 1, it has been seen to approximate

the underlying function efficiently in design optimization and structural reliability analysis applications [58, 60, 61, 62] with discrete variables. A major disadvantage of the Kriging model seen in the literature is the decrease in performance for high-dimensional applications.

Radial Basis Functions (RBF) are another modeling technique popularly used for a surrogate model. In contrast to Kriging, RBF models have the capability to handle high-dimensional problems efficiently [63, 5, 64]. A major advantage of RBF models is the ability to tune their parameters easily to suit the application, and the model form is similar to a neural network with a single hidden layer. Consequently, it has been widely used in the literature in multi-objective surrogate black-box optimization applications, global optimization of highly non-convex functions and constrained surrogate assisted optimization [65, 66, 67, 68, 69, 70].

Support Vector Regression (SVR) is one of the commonly used surrogate models. SVR emerged as a modification of Support Vector Machines (SVM), which were mainly used for classification. It was first developed by AT&T Bell Laboratories [71]. A major advantage of SVR models is the ability to specify an acceptable error margin for the prediction results without affecting the prediction accuracy of the model. This is very useful in data with inherent noise and when collecting data using physical experiments [44]. Since SVR models are specifically designed for dealing with high-dimensional noisy data and because of the structure of the model, it is rendered useless if the cost of collecting large amounts of data to fill the design space is inefficient. Consequently, SVR is rather seen in applications where simulation of the data are possible [72, 73, 74, 75].

MARS was first developed by Friedman [76] in 1991. It has been used as a very popular surrogate model because it is a non-parametric regression model and is considered an extension of the linear regression model. Tsai and Chen [10] propose a robust version of MARS within an orthogonal array (OA)/MARS continuous-state stochastic dynamic programming (SDP) method. The method reduces the computational complexity of a wastewater treatment SDP by introducing two automatic stopping rules that determine the maximum number of basis functions and a robust version of MARS that prefers lower

order interaction terms to higher order interaction terms. Crino and Brown [77] propose the use of MARS in combination with RSM in obtaining a single objective global optimum for a structural design. The literature makes use of the advantages of both the MARS model and RSM. MARS has an excellent capability to reduce the dimension of a problem, and RSM can quickly converge to an optimum for low dimensional problems. Costas et al. [78] use MARS in the multi objective global optimization of a crashworthiness problem. In the literature, MARS was compared against a quadratic polynomial model, cubic polynomial model and gaussian kriging model. The application had highly noisy data adding to the complexity of finding a pareto optimal solution. It was seen that MARS outperformed the other models. Anahideh et al. [9] proposes Tree Knot MARS (TK-MARS) as a surrogate model in black box optimization. It uses a classification and regression tree partitioning (CART) method to identify the locations of knots. A smart replication method to identify promising candidates to replicate is also introduced. In addition, there is significant literature that discusses the use of MARS for optimal fleet assignment [79] and also optimizing decisions based upon a MARS system model [80, 81, 82, 83, 84].

Ensemble of surrogates is another popular technique where different surrogate models are combined using a weighted approach or some other metric and used as metamodels. While most literature focuses on selecting one model from a set of choices, the newly developed ensemble of surrogates focuses on choosing a combination of surrogates, which work best for the application [85, 86]. Goel et al. [87] propose a weighted approach in combining Polynomial Response Surface (PRS), Kriging and RBF models. The literature uses ensemble of surrogates to predict regions of high uncertainty and to provide robust approximations of response surfaces. It was seen that the DoE used in the initialization phase had a huge impact on the performance of the ensemble of surrogates, and some of DoE made the surrogate models perform worse than they would have if used separately. The literature suggests the use of ensemble of surrogates if the best surrogate model for the application is unknown beforehand. Ouyang et al. [88] used a combination of five different surrogate models – PRS, RBF, Kriging, SVR and Gaussian Process (GP). The literature studied the performance

of five models separately as well as all the combinations of the models. It was seen that the model with all the surrogates performed consistently better for the application. The literature suggests that the DoE used for sampling plays a major role in the performance of the model. It also mentions that adding different surrogate models, like Artificial Neural Networks (ANN), to the ensemble or adding more than five surrogate models may not have the same accuracy as the ensemble of surrogates proposed in the literature.

2.2.3 Decision Space Search

Effective searching of the decision space to find new points to evaluate is an important part of the process because it helps find an optimal solution with few function evaluations. In many cases, it is very expensive and time consuming to setup and run the experiments to evaluate a point; using an efficient decision space search algorithm helps save cost and time. The literature discusses several algorithms to select the next point(s) for evaluation, which are discussed in this section.

2.2.3.1 Sequential Search Algorithms

Sequential search algorithms search through the decision space to iteratively select new points and sample the entire decision space by using several evaluation criteria to identify the next location to sample new points. Jones et al. [89] introduce a criteria based sequential design approach to minimize response in computer experiments. It starts with an initial space filling DoE and chooses points in regions where the predicted response is minimized, or prediction uncertainty is relatively large. After, each point is selected based on the criteria, the true response is calculated, and the response is updated. The process is repeated until the change in the criteria is negligible. Whereas Williams et al. [90] use a maximin approach to minimize the weighted response based on the inputs. It uses a maximin LHS design to initialize the sample points. The literature also splits the input variables into control and environmental variables. It uses a posterior expected improvement metric to select new points and as the stopping criteria. New points in the control variable space and

the environmental variable space are selected based on the posterior expected improvement metric. The literature also keeps track of the best know minimum solution as the global minimizer and is updated in each iteration with a new minimizer if one is found. Crombecq et al. [6] present a comparison and analysis of different exploration based space-filling designs in a surrogate modelling setting for time-consuming simulated computer experiments. The literature compares random sampling, Voronoi-based sequential design and Delaunay-based sequential design against traditional one-shot LHD. It is found that random sampling does not consider the points that were evaluated previously and randomly searches through the space to suggest new points to evaluate. It works well in applications were the number of allowed sample points is large, which is not the case in most applications. With Voronoi-based sequential design, the Voronoi volume estimation must be performed, and to get a good estimate of the estimation, at least 100 random samples per data point is required, which is not a lot in computer-based simulations. However, it is not a viable approach for applications that require expensive laboratory runs. Delaunay-based sequential design requires the calculation of Delaunay's triangulation, which is rendered infeasible for problems with more than 6 dimensions. A major disadvantage mentioned is that it does not have good space filling properties in the corners. In other literature, Crombecq et al. [91] propose four new space filling designs that achieve maximum intersite distance as well the projected distance for simulation based surrogate modelling. The first method is called Sequential nested Latin hypercubes in which the iteration starts with an initial grid of candidate points. Where, a new point is chosen sequentially from the grid. Once all points from the grid have been chosen, midpoints between the samples are considered, and the process is repeated, thus (asymptotically) doubling the grid size at each iteration and the point that is farthest away from the grid is chosen as the new point in each iteration. The second method developed is called Global Monte Carlo methods, which uses Monte Carlo methods to generate (100 $\times$ the number of sample points evaluated in an iteration) random points for each sample a score to calculate the distance of the new points from the previously evaluated points and whichever point has the maximum score is picked as the new candidate point. The third and

fourth methods use the Direct Search Toolbox of MATLAB as a local optimizer to optimize the projected distance and the intersite distance. These methods use Monte Carlo method to generate a surface for the projected distance and intersite distance and locally optimize the surfaces. Following that, Xiao et al. [92] introduce an adaptive sequential sampling method to build surrogate models with applications in structural reliability analysis. The literature uses Monte Carlo Simulation to initialize the design and uses a kriging surrogate model. Three learning models are used to sequentially sample new points in the designs: the first learning function focuses on the distances to the limit-state functions and to the existing training sample points. The second learning function focuses on the distances to the limit-state functions and the points with higher uncertainty. The third learning function focuses on the properties and advantages of the first two learning functions. The performance of these learning functions was evaluated on both small and medium-high dimensional problems. It was found that there was no significant difference in the different learning functions for small problems. For medium-high dimensional problems, learning functions two and three performed better.

2.2.3.2 Pareto Frontier Approach

The Pareto frontier approach to selecting new points is the most popular due to their ability to find a balance between exploration and exploitation. Exploration is used to explore the unexplored regions in the response surface to identify new points to run the experiments. Exploration does not use surrogate models to identify these points, instead it uses distance metrics like maximin criteria. Whereas exploitation makes use of surrogate models to explore the space to find points that optimize the response. A pareto frontier with non-dominating solutions combining the points obtained from exploration and exploitation is constructed. New points to run the experiments are then identified from this pareto frontier. Jones et al. [93] introduce an Efficient Global Optimization (EGO) algorithm that selects new points based on the Expected Improvement (EI) of the objective function. The algorithm selects new points with highest EI. The algorithm searches through the entire design space by max-

imizing the EI. The search is stopped when the EI is less than 1%. Knowles [94] extends the EGO algorithm introduced by Jones to optimize Multi-objective problems and calls it ParEGO. It uses Evolutionary Algorithms (EA) to maximize the Expected Improvement of the objective and select new points for evaluation. The approach also uses a weighted cost function to select non-dominated pareto optimal solutions. Unlike the other literature, Simpson et al. [58] propose the use of global surrogate models to explore the design space and predict a pareto frontier all at once without having to specify weights for objective functions in each iteration or use optimization algorithms to select new points. The predicted Pareto frontier can then be explored graphically to determine suitable design solutions that yield the best compromise between the multiple competing objectives. Dickson et al. [11] present the idea of finding a non-dominated pareto optimal solution by finding the balance between exploration and exploitation called Exploration Exploitation Pareto Approach (EEPA). The method uses a random pool (weak points) to build the pareto frontier from which new points are selected for evaluation. Nezami and Anahideh [95] propose a improved version of EEPA called EEPA$^+$ utilizing a dynamic coordinate discretization scheme similar to DYCORS developed by Regis and Shoemaker [96]. Krityakierne et al. [97] use the concept of multi-objective optimization in single objective optimization. The literature introduces Surrogate Optimization with Pareto Selection (SOP) where simultaneous surrogate-assisted candidate searches are performed around selected evaluated points, and a bi-objective optimization is performed to find the balance between exploration and exploitation, which are set to be the two conflicting objectives. Erikkson et al. [98] use an acquisition function and a set of weighted functions defined by the distance of the candidate points from the previously evaluated points. A promising candidate for evaluation is chosen when the acquisition function is minimized. The weights are used to balance exploration and exploitation. Akthar and Shoemaker [99] use a list of sampling radii to select a new candidate point for evaluation. The method keeps track of the improvements in the solution and uses a larger sample radius when there is no improvement and a smaller radius when there is improvement, to balance between exploration and exploitation. This research uses a methodology similar to

this approach.

2.2.4 Evaluation of Stopping Criteria

The stopping criteria in a surrogate optimization algorithm plays an important role in minimizing the number of function evaluations and in making sure the algorithm does not get stuck at a local optimum. Expected Improvement (EI) has been used as popular stopping criteria for expensive optimization problems. Several modifications and adaptions of the method have been used in a wide variety of optimization problems [100]. Anahideh et al. [9] uses a maximum of 1000 function evaluations in the optimization of a single objective surrogate based optimization of a complex black box system. Caballero and Grossmann [101] present the optimization of a modular chemical process simulator. The surrogate optimization algorithm in the literature uses a Kriging metamodel to mimic the simulation model. The algorithm is stopped when the size of the sampling region is considerably small enough and if the same set of independent variables are returned in two consecutive iterations. Chaudhuri and Haftka [102] introduce a stopping criterion for the Efficient Global Optimization (EGO) algorithm called EGO-AT, which uses Probability of targeted Improvement (PI) with an adaptive target. The literature introduces a reasonable target for improvement in the next cycle, and the probability of achieving that target are used as stopping criteria. It assumes that a threshold for improvement is already defined by the user based on the problem, and if the threshold is less than the absolute tolerance and relative tolerance, calculated using formula defined in the literature, the algorithm does not proceed with the next iteration. Another stopping criteria defined is if the probability of achieving the target is less than 20% then algorithm is stopped. Wan et al. [103] use absolute expected improvement and relative expected improvement defined by the ratio of the current expected improvement versus the initial expected improvement as the threshold and the algorithm is stopped when the improvement is less than the threshold. In addition, the literature also uses total simulation time and total number of sampled points as stopping criteria in a simulation-based supply chain optimization problem. Voutchkov and Keane [104] suggest several possible

stopping criteria for a multi-objective pareto frontier based optimization algorithm, such as a fixed number of iterations, when all updated points are non-dominated, when a percentage of new updated points belonging to the pareto frontier falls below a pre-defined value, when a percentage of new updated points belonging to the pareto frontier falls above a pre-defined value, and when there is no more improvement of the pareto frontier quality.

2.2.5 Applications of Surrogate Modeling and Optimization in Chemistry

Analytical chemistry is a branch of chemistry that uses methods and instruments to efficiently separate, identify and quantify matters. The applications of analytical chemistry play a key role in our day to day lives in identifying any contaminants in the water we drink and the air we breathe. On a broader level, it has applications in forensic science, bio analysis and clinical analysis and has led to a lot of important discoveries [105]. A key to proper and efficient extraction and detection of compounds is the selection of the settings of the experimental parameters like temperature, pressure, etc. There are several new technologies that enable the efficient extraction of compounds [106, 107, 108, 109, 110, 111]. Due to the high level of complexity involved in the process, there is not a lot of knowledge on the effect of these experimental parameters on the efficiency of the process. In addition, the preparation of samples and each run of the equipment is time and resource intensive. This leads to working with limited data and the need for creative methodologies that can make efficient use of these limited data points and optimize the experimental parameters.

DoE enables direct control of the experiments and the number of runs [112]. One of the major applications of DoE and RSM is in the optimization of parameter settings to maximize extraction of chemical compounds. The focus of this research is on this application as well. There is an abundance of literature that use RSM to optimize parameter settings in analytical chemistry in general [113, 114, 115] and also specifically in mass spectrometry [116, 117, 110], which is the application of interest in this research.

Surrogate optimization was originally intended for use with highly computational

computer models [8, 3, 118, 4], such as finite-element simulation models, and uses surrogates (a.k.a., metamodels) to infer information from very few evaluations of the computer models. In this research, the laboratory experiments take the role of the computer model. This section of the literature review focuses on identifying literature that has applied optimization to chemical processes. Put et al. [119] use a MARS model to predict the retention times of a Reversed-phase high performance liquid chromatography (RPLC), which is one of the most frequently used techniques to separate pharmaceutical mixtures. The prediction from the MARS model helps in the selection of an optimal set of starting chromatographic system parameters, which are usually determined by trial and error. The literature compares the performance of the MARS model developed by Friedman with MLR models and a neural model, and it is found that the MARS model performs better the other two model types. Cozad et al. [120] introduce Automated learning of algebraic models for optimization (ALAMO), which uses surrogate optimization for superstructures in simulation based applications and builds the model iteratively by selecting new points in each iteration using adaptive sampling. The methodology generates algebraic metamodels based on user input and basic engineering and statistical relationships and uses Kriging and ANN to model some more complex basis functions. The literature tests the proposed methodology on the simulation of a Carbon Capture Adsorber. Quirante et al. [121] utilize a Kriging based surrogate optimization methodology in the design of distillation columns to optimize the sequence of columns in the superstructure and the number of trays in the column with an objective to minimize the Total Annual Cost (TAC) of the system. The methodology in the literature only guarantees a local optimum, and to guarantee global optimum, it is suggested to increase the degrees of freedom and increase the number of sample points. However, doing so results in slower performance of the Kriging model and more CPU time. Nentwich and Engell [122] employ Artificial Neural Networks (ANN) in the superstructure optimization of the hydroformylation process. The literature compares the performance of the simulation using the traditional thermodynamic models used in the simulators versus replacing the thermodynamic model with the multi-layer perceptron ANN models. It is seen that ANNs

perform better with significantly lower computation times. Schweidtmann et al. [123] use the recently developed Thompson Sampling Efficient Multi-Objective (TS-EMO) algorithm in combination with an automated continuous reaction system for super structure optimization. The literature uses Latin Hypercube design (LHD) to initialize the algorithm and uses Gaussian Process surrogate models. The algorithm also uses a pareto frontier to optimize and select new points in consecutive iterations. The experiment was set up, so the entire system ran automatically without any user intervention overnight by using the TS-EMO algorithm to generate parameters to control the reactor. The experiment was stopped when a dense front of at least 20 experimental Pareto data points were collected.

A major takeaway from the literature using surrogate optimization in the analytical chemistry realm is that it is mostly used in superstructure optimization. There is very little literature that uses surrogate optimization for obtaining optimal parameter settings and even less literature that has achieved global optimization of these surrogate models.

Chapter 3

MIQCP Formulation

3.1 Introduction to MARS

A General MARS model with a response surface, $f(\mathbf{x})$, fitted on independent variables x_i can be represented as:

$$\hat{f}(\mathbf{x}) = a_0 + \sum_{m=1}^{M} \{a_m B_m(\mathbf{x})\} \tag{3.1}$$

where, the first term, a_0, is the intercept. The second term is a weighted sum of the basis functions, where a_m is a constant coefficient for each basis function, $B_m(\mathbf{x})$. The basis function depends on $\mathbf{x}$, which is a n-dimensional vector of input variables. In the weighted sum, M represents the total number of linearly independent basis functions.

Usually, each basis function is univariate and the product of two or more basis functions is used to model the interactions between the input variables. The basis function for interaction terms is usually modeled as:

$$B_m(\mathbf{x}) = \prod_{k=1}^{K_m} \eta_{km} \tag{3.2}$$

where η_{km} is the k^{th} univariate term of the m^{th} basis function and K_m is the number of interaction terms in the m^{th} basis function. In this research $K_m \leq 2$ is considered. In addition, the formulation is set up such that it can be expanded easily to higher order

interactions.

Figure 3.1 obtained from [81] shows a graphical representation of the two way interaction term of a MARS function.

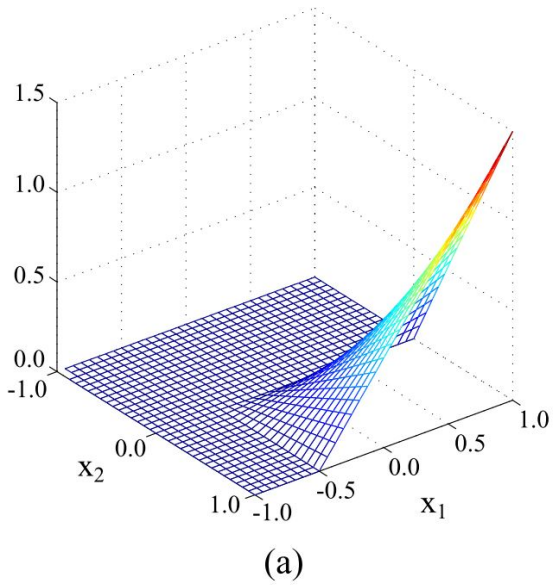

Figure 3.1: *Two-way interaction term of MARS function*

There is significant literature that have worked on developing a global optimization algorithm for various types of MARS models [80, 82, 83, 84]. Specifically, Ju et al. [83] optimized a two way interaction piece-wise linear function of MARS of the form in equation 3.3:

$$\eta_{km} = [s_{km}\left(x_{j(k,m)} - t_{km}\right)]_+ = \max\left\{s_{km}\left(x_{j(k,m)} - t_{km}\right),\ 0\right\} \tag{3.3}$$

where $j(k,m)$ is the index of the variable within the vector $\mathbf{x}$ for the k^{th} univariate term of the m^{th} basis function and t_{km} is the knot location corresponding to $x_{j(k,m)}$. The value s_{km} is the direction of the hinge function and can be either +1 or -1.

3.2 Quintic MARS (QMARS)

The aspect in which this research is different from Ju et al. [83] is that this research focuses on optimizing the QMARS model developed by Chen [10], where the MARS basis function contains cubic, quadratic and quintic terms. Figure 3.2 obtained from [81] shows the graphical representation of the difference between optimizing two-way interaction term with a piecewise-linear fit and quintic fit.

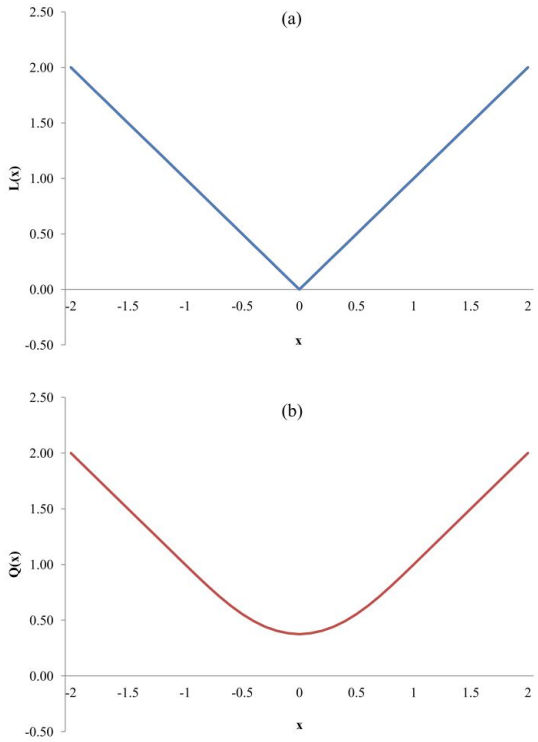

Figure 3.2: *Two-way interaction term with (a) piecewise-linear fit vs. (b) quintic fit*

The QMARS function has three knots: a lower knot denoted by t_{km}^{-}, a center knot denoted by t_{km}, and an upper knot denoted by t_{km}^{+}, corresponding to the variable $x_{j(k,m)}$ of the k^{th} univariate term of the m^{th} basis function. Then the MARS basis function η_{km}, for the negative direction denoted by s_{km}= -1 can be defined as:

$$\eta_{km}$$

$$= \begin{cases} t_{km} - x_{j(k,m)}, & x_{j(k,m)} \leq t^-_{km} \\ \alpha^-_{km}\left(x_{j(k,m)} - t^+_{km}\right)^3 + \beta^-_{km}\left(x_{j(k,m)} - t^+_{km}\right)^4 + \gamma^-_{km}\left(x_{j(k,m)} - t^+_{km}\right)^5, & t^-_{km} < x_{j(k,m)} < t^+_{km} \\ 0, & x_{j(k,m)} \geq t^+_{km} \end{cases} \tag{3.4}$$

where,

$$\alpha^-_{km} = \frac{(-1)\left(4t^+_{km} - 10t_{km} + 6t^-_{km}\right)}{\left(t^-_{km} - t^+_{km}\right)^3}$$

$$\beta^-_{km} = \frac{(-1)\left(-7t^+_{km} + 15t_{km} + 8t^-_{km}\right)}{\left(t^-_{km} - t^+_{km}\right)^4},$$

$$\gamma^-_{km} = \frac{(-1)\left(3t^+_{km} - 6t_{km} + 3t^-_{km}\right)}{\left(t^-_{km} - t^+_{km}\right)^5}$$

Similarly, the MARS basis function η_{km}, for the positive direction denoted by s_{km}= 1 can be defined as:

$$\eta_{km}$$

$$= \begin{cases} 0, & x_{j(k,m)} \leq t^-_{j(k,m)} \\ \alpha^+_{km}\left(x_{j(k,m)} - t^-_{km}\right)^3 + \beta^+_{km}\left(x_{j(k,m)} - t^-_{km}\right)^4 + \gamma^+_{km}\left(x_{j(k,m)} - t^-_{km}\right)^5, & t^-_{km} < x_{j(k,m)} < t^+_{km} \\ x_{j(k,m)} - t_{km}, & x_{j(k,m)} \geq t^+_{km} \end{cases} \tag{3.5}$$

where,

$$\alpha^+_{km} = \frac{4t^+_{km} - 10t_{km} + 6t^-_{km}}{\left(t^+_{km} - t^-_{km}\right)^3}$$

$$\beta^+_{km} = \frac{-7t^+_{km} + 15t_{km} + 8t^-_{km}}{\left(t^+_{km} - t^-_{km}\right)^4}$$

$$\gamma^+_{km} = \frac{3t^+_{km} - 6t_{km} + 3t^-_{km}}{\left(t^+_{km} - t^-_{km}\right)^5},$$

3.3 Formulation of QMARS using MIQCP

The research focuses on QMARS by formulating the model using Mixed Integer Quadratically Constrained Program (MIQCP) defined by the above equations. Table 3.1 shows the variable notations and its definitions used in the formulation.

Notation	Definition
M	Number of unique univariate basis functions
a_0	Constant Intercept coefficient of the QMARS model
a_m	Coefficient of the m^{th} basis function,$\forall m \in M$
K_m	Number of interactions terms in the m^{th} basis function,$\forall k \in K_m, \forall m \in M$
s_{km}	Direction of the k^{th} univariate term of the m^{th} basis function, $\forall k = 1, ..., K_m, m = 1, ..., M$
η_{km}	Value of the k^{th} univariate term of the m^{th} basis function, $\forall k = 1, ..., K_m, m = 1, ..., M$
t^-_{km}	Lower knot of the k^{th} univariate term of the m^{th} basis function, $\forall k = 1, ..., K_m, m = 1, ..., M$
t_{km}	Center knot of the k^{th} univariate term of the m^{th} basis function, $\forall k = 1, ..., K_m, m = 1, ..., M$
t^+_{km}	Upper knot of the k^{th} univariate term of the m^{th} basis function, $\forall k = 1, ..., K_m, m = 1, ..., M$
T_j	Set of knots for variable j. where $T_j = \{l_j, u_j\} \bigcup_{k=1,...,K_m, m=1,...,M} \{t^-_{km} \cup t_{km} \cup t^+_{km}\}$ $\forall j = 1, ..., n, k = 1, ..., K_m, m = 1, ..., M$
$t_{j(r)}$	r^{th} smallest knot within the set T_j, $\forall j = 1, ..., n, r = 0, \ldots, \lvert T_j \rvert$ Note: $T_j = (l_j, t_{j(1)}, t_{j(2)}, ...t_{j(\lvert T_j \rvert - 2)}, u_j)$
r^-_{km}	Index of the piece in the ordered knots set immediately to the left of t^-_{km}, $\forall k = 1, ..., K_m, m = 1, ..., M$
r^+_{km}	Index of the piece in the ordered knots set immediately to the right of t^+_{km},

	$\forall k = 1, ..., K_m, m = 1, ..., M$
x_{jr}	Linear piecewise variable representing variable j in each of the pieces, st. $x_{jr} = x_j$ if $t_{j(r)} \leq x_j < t_{j(r+1)}$, 0 otherwise, $\forall j = 1, ..., n, r = 0, ..., \lvert T_j \rvert - 1$
y_{jr}	Binary piecewise variable representing variable j in each of the pieces, st. $y_{jr} = 1$ if $t_{j(r)} \leq x_j < t_{j(r+1)}$, 0 otherwise, $\forall j = 1, ..., n, r = 0, ..., \lvert T_j \rvert - 1$
h^i_{km}	Difference between $x_{j(k,m)}$ and the associated lower or upper knot for variable j taken to the power of i when $t^-_{km} < x_j < t^+_{km}$, $\forall i = 1, ..., 5, k = 1, ..., K_m, j = 1, ..., n, m = 1, ..., M$,
n	Number of explanatory variables
x_j	Value of variable j, $\forall j = 1, ..., n$
l_j	Lower bound on each of the explanatory variables,$\forall j = 1, ..., n$
u_j	Upper bound on each of the explanatory variables,$\forall j = 1, ..., n$
E	Set of evaluated points
x	Value of variables x_j in the evaluated points set, st. $x \in E$
C	Set of center points available for evaluation
c	Index of the center point selected for evaluation
x_c	Value of center point selected for evaluation, st. $x_c \in C$
ρ	Radius of the circle within which the QMARS model is optimized

Table 3.1: Table of variables

The formulation is given as follows:

$$\max \hat{f}(x) = a_0 + \sum_{m=1}^{M} \left\{ a_m \prod_{k=1}^{K_m} \eta_{km} \right\} \tag{3.6}$$

$$\eta_{km} = \begin{cases} \alpha^- h^3_{km} + \beta^- h^4_{km} + \gamma^- h^5_{km} \\ + \sum_{r=r^-_{km}}^{|T_j|-1} t_{j(r)} y_{j(k,m)r} - x_{j(k,m)r}, & s_{km} = -1, \forall k = 1, ..., K_m, m = 1, ..., M \\ \alpha^+ h^3_{km} + \beta^+ h^4_{km} + \gamma^+ h^5_{km} \\ + \sum_{r=r^+_{(km)}}^{|T_j|-1} x_{j(k,m)r} - t_{j(r)} y_{j(k,m)r}, & s_{km} = 1, \forall k = 1, ..., K_m, m = 1, ..., M \end{cases} \tag{3.7}$$

$$t_{j(r)} y_{jr} \leq x_{jr} \leq t_{j(r+1)} y_{jr}; \qquad \forall j = 1, ..., n, r = 0, ..., |T_j| - 1 \tag{3.8}$$

$$x_j = \sum_{r=0}^{|T_j|-1} x_{jr}; \qquad \forall j = 1, ..., n \tag{3.9}$$

$$\sum_{r=0}^{|T_j|-1} y_{jr} = 1; \qquad \forall j = 1, ..., n \tag{3.10}$$

$$h^1_{km} = \begin{cases} \sum_{r=r^-_{km}}^{r^+_{km}} x_{jr} - t_{km} y_{jr}, & s_{km} = -1, \forall k = 1, ..., K_m, m = 1, ..., M \\ \sum_{r=r^-_{km}}^{r^+_{km}} x_{jr} - t_{km} y_{jr}, & s_{km} = 1, \forall k = 1, ..., K_m, m = 1, ..., M \end{cases} \tag{3.11}$$

$$h^2_{km} = h^1_{km}.h^1_{km}; \qquad \forall k = 1, ..., K_m, m = 1, ..., M \tag{3.12}$$

$$h^3_{km} = h^1_{km}.h^2_{km}; \qquad \forall k = 1, ..., K_m, m = 1, ..., M \tag{3.13}$$

$$h^4_{km} = h^2_{km}.h^2_{km}; \qquad \forall k = 1, ..., K_m, m = 1, ..., M \tag{3.14}$$

$$h^5_{km} = h^2_{km}.h^3_{km}; \qquad \forall k = 1, ..., K_m, m = 1, ..., M \tag{3.15}$$

$$\|x - x_c\| \leq \rho \tag{3.16}$$

$$l_j \leq x_j \leq u_j; \qquad \forall j = 1, ..., n \tag{3.17}$$

$$\eta_{km} \in \mathbb{R}; \qquad \forall k = 1, ..., K_m, m = 1, ..., M \tag{3.18}$$

$$x_{jr} \in \mathbb{R}; \qquad \forall j = 1, ..., n, r = 1, ..., |T_j| - 1 \tag{3.19}$$

$$h^i_{km} \in \mathbb{R}; \qquad \forall i = 1, ..., 5, k = 1, ..., K_m, m = 1, ..., M \tag{3.20}$$

$$x_j \in \mathbb{R}; \qquad \forall j = 1, ..., n \tag{3.21}$$

$$y_{jr} \in \mathbb{B}; \qquad \forall j = 1, ..., n, r = 1, ..., |T_j| - 1 \tag{3.22}$$

The objective (3.6) is the QMARS model. Equations (3.7) - (3.10) represent the basis functions and knots formulated as linear constraints. Equations (3.11) - (3.15) represent the use of linear formulation and consequently the use of recursive substitution to handle the cubic, quadratic and quintic terms in the basis function. Equation (3.16) is the euclidean distance constraint which represents the ball constraint within which the QMARS model is optimized. If the radius of the ball is large enough, it is similar to globally optimizing the entire space. Equations 3.17 represents the bounds on the decision variables. Finally, equations (3.18) - (3.22) represent the variable types of the variables used in the formulation.

3.4 Global Optimization of QMARS using MIQCP

The process of optimizing the QMARS model using the MIQCP will be to referred to as QMARS-MIQCP-OPT henceforth. The algorithm for QMARS-MIQCP-OPT is given by Algorithm 1. The first step in the algorithm is generating a large data set of decision variables and corresponding response variables. The QMARS model is then fit on this large data set. Finally, the QMARS-MIQCP-OPT algorithm optimizes the QMARS model using the MIQCP. The objectives and constraints were formulated in MALTAB and combined with the GUROBI optimizer for optimization.

Algorithm 1:
General Algorithm for QMARS-MIQCP-OPT

Result: Optimal Solution: x^* and $f(x^*)$

1 Generate a sample data set of decision variables and corresponding response variables.
2 Fit QMARS model on the data.
3 Optimize the QMARS model using MIQCP to obtain x^* and $f(x^*)$.

3.5 Experiments and Results

The performance of the QMARS-MIQCP-OPT methodology was tested by optimizing functions f_1, f_2, f_3 and f_4 found in Ju et al. [83]. The true function values of the optimal solution obtained by minimizing and maximizing the QMARS model using the QMARS-MIQCP-OPT algorithm are compared with the results of the TITL-MARS-OPT in Ju et al. [83]. Functions f_1 and f_2 are 2-dimensional. The number of dimensions is denoted by the parameter, d. Whereas, functions f_3 and f_4 are 10-dimensional. All the functions have significant interactions. The results are summarized in table 3.2.

Function	**d**	**Measurement**	**QMARS-MIQCP-OPT**	**TITL-MARS-OPT**
f_1	2	Minimum	**-4.78**	-4.56
		Maximum	5.88	**7.67**
f_2	2	Minimum	**-0.95**	-0.61
		Maximum	0.83	**0.86**
f_3	10	Minimum	**-891.64**	-826.08
		Maximum	**6415.68**	6029.13
f_4	10	Minimum	**-4,696,047.51**	-4,581.942.36
		Maximum	**-206.18**	-1991.28

Table 3.2: Comparison of QMARS-MIQCP-OPT and TITL-MARS-OPT

The results show that QMARS-MIQCP-OPT achieves better solutions consistently when minimizing functions f_1 to f_4 in comparison to TITL-MARS-OPT. As for maximizing

the functions, QMARS-MIQCP-OPT performs better for functions f_3, f_4. For function f_2 the optimal solution from QMARS-MIQCP-OPT is really competitive with TITL-MARS-OPT. Overall, QMARS-MIQCP-OPT outperforms TITL-MARS-OPT for all the four functions when minimizing and outperforms TITL-MARS-OPT for two out of the four when maximizing.

Chapter 4

Surrogate Optimization of QMARS-MIQCP

4.1 Existing Surrogate Optimization Methodology

A general framework for existing surrogate optimization methodology is presented in Algorithm 2 and illustrated in Figure 4.1. In line 1 of Algorithm 2, a finite set of points is selected from the input space. Because existing surrogate optimization algorithms are limited to the points in this point set, a very large DoE is generated to ensure good representation of the space. In Figure 4.1, the input space is represented by the gray rectangle, and the DoE points are represented by a combination of the blue hexagons, orange circles and yellow stars. In line 2 of the algorithm, a small initial set of points is selected from the large DOE, shown as orange circles in the figure. In line 3, the objective function is evaluated for this initial set of points. In the figure, the evaluated points are represented by the blue hexagons. In line 4 of the algorithm, the surrogate model is fit to the set of evaluated points. Exploitation of the surrogate model is then used in line 5 to identify promising new points. In addition, exploration using a distance metric or similar approach identifies new points in unexplored areas of the input space. These new points are limited to the unevaluated points in the large DOE from line 1 and are represented by the orange circles in the figure. A key component of

surrogate optimization is to balance exploration and exploitation of the space. The number of points selected in line 5 can also vary based on the specific methodology. In the figure, these new points are represented by the yellow stars. The new points are evaluated in line 6, and the evaluated and unevaluated point sets are updated in line 7. Correspondingly, in the figure, the yellow circles become blue hexagons. The while loop conducted iterations until specified stopping criteria are met.

Existing algorithms rely on a maximum number of function evaluations to stop the algorithm. However, this is a major shortcoming because the solution found is the best known solution at that point in the algorithm and may not necessarily be close to the optimal solution. A second important drawback is how the search in line 5 of the algorithm is limited to the points in the large DOE from line 1. Failure to explore the space in between the DOE points guarantees a suboptimal search.

Algorithm 2:
Existing Surrogate Optimization Methodology

Result: Optimal Parameter Settings: x^* and $f(x^*)$
1 Generate DoE with maximum number of points considered.
2 Select a small set of initial points from DoE.
3 Evaluate the objective for the initial set.
while *Stopping criteria not met* **do**
4 Fit surrogate model on set of evaluated points.
5 Search space to find new points to evaluate from the set of candidate points.
6 Evaluate the objective for the selected new points.
7 Remove selected points from set of candidate points and add to set of evaluated points.
end

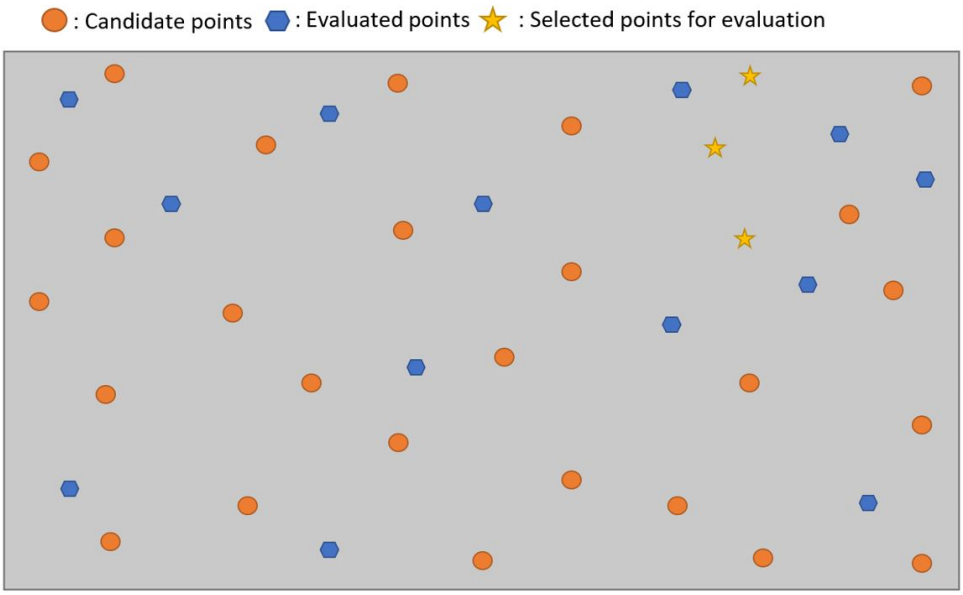

Figure 4.1: *Existing surrogate optimization methodology*

4.2 Proposed Surrogate Optimization Methodology

Our proposed methodology aims to overcome the shortcomings in the existing methodology by formally optimizing the surrogate model and enabling the selection of new points that are no longer limited to the large DOE point set. Algorithm 3 presents the proposed surrogate optimization methodology, which we refer to as QMARS-MIQCP-SUROPT. Lines 1–4 are the same as Algorithm 2, but the algorithms differ in the while loop. In line 5, the proposed approach specifically employs the Exploration-Exploitation Pareto Approach (EEPA), developed by Dickson [11], although an alternate search could be employed. The key difference comes in line 6, in which the proposed QMARS-MIQCP-SUROPT approach explicitly conducts the MIQCP optimization discussed in Section 3.3.

As mentioned, the proposed methodology also starts by initializing a large DoE and selecting a small set of points, at which the objective is evaluated. These are also similarly represented by the gray rectangle and a combination of the blue hexagons, orange circles and yellow stars in Figure 4.2. Our contribution to the methodology is in the process of searching the space and selecting new points for evaluation in subsequent iterations. Our approach uses

a modified EEPA algorithm to efficiently balance both exploration and exploitation while searching the space. The modified EEPA algorithm is discussed in Section 4.2.2. The points generated by the EEPA algorithm are represented by the purple triangles in the figure. Our proposed methodology creates a ball, with the EEPA points as centers and uses the MIQCP formulation discussed in Section 3.3 to optimize the surrogate model and generate new points for evaluation. In the formulation in Section 3.3, equation (3.16) represents the formulation of the ball. In the figure, the ball is represented by the green circle surrounding the purple triangle. The MIQCP optimization considers the space within the bounds of the circle, and the optimal solutions are selected as new points for evaluation. These new points are represented by the yellow stars in the figure and are not limited to the DOE in line 1. Similar to Algorithm 2, the last two lines of Algorithm 3 evaluate the objective at the new points and then update the evaluated and unevaluated point sets. However, because the new points are no longer limited to the DOE from line 1, only those new points that coincide with candidate points and the EEPA points chosen as centers in the DOE are removed.

Algorithm 3:
Proposed Surrogate Optimization Methodology

Result: Optimal Parameter Settings: x^* and $f(x^*)$
1 Generate DoE with maximum number of points considered.
2 Select a small set of initial points from DoE.
3 Evaluate the initial set and add points to the set of evaluated points.
while *Stopping criteria not met* **do**
4 Fit surrogate model on set of evaluated points.
5 Use EEPA to search the space to find points from the set of candidate points.
6 Optimize the surrogate within a ball with EEPA points as centers to find new points for evaluation.
7 Evaluate new points selected.
8 Remove EEPA points and coinciding new points from set of candidate points and add to set of evaluated points.
end

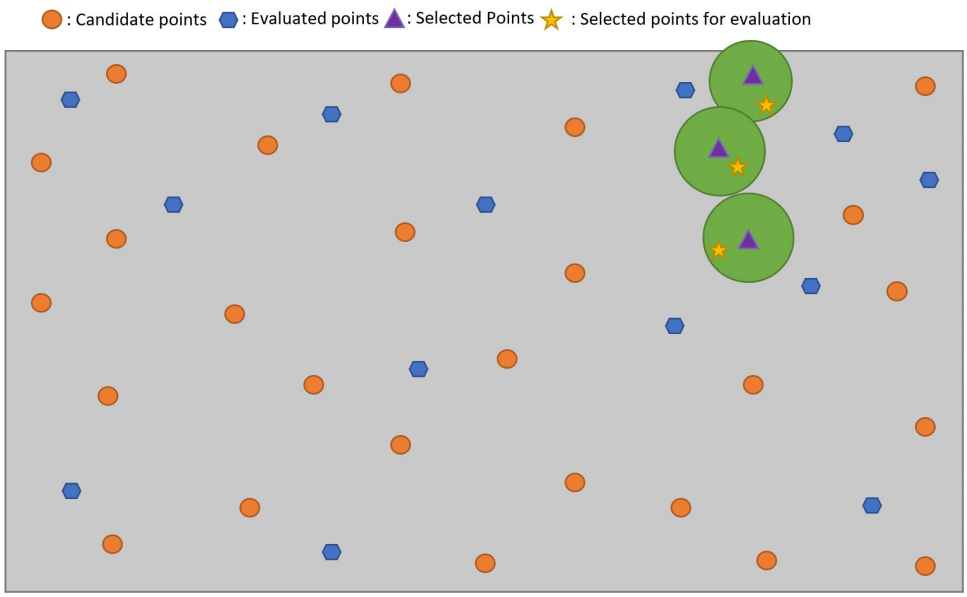

Figure 4.2: *Proposed surrogate optimization methodology*

The rest of the section discusses the details of the proposed surrogate optimization methodology QMARS-MIQCP-SUROPT, specifically in regards to the surrogate model used, the candidate search algorithm used and the candidate selection strategy.

4.2.1 Surrogate Model

Our proposed methodology, QMARS-MIQCP-SUROPT uses the Tree-Knot MARS (TK-MARS) surrogate model developed by Anahideh et al. [9]. The TK-MARS surrogate model is a modification of the QMARS model developed by Chen [10], upon which the MIQCP formulation discussed in Section 3.3 is based. Like all regression based statistical models, MARS is intended for modeling data observed under uncertainty. Consequently, unlike common choices for surrogate optimization, the MARS approximation is non-interpolating. Anahideh et al. [9] demonstrated the benefit of a MARS surrogate when the objective is evaluated under uncertainty and when the input space includes unimportant variables. Analytical chemistry surrogate optimization applications potentially involve both issues.

A drawback the original MARS algorithm concerns knot selection. The linear model form of MARS defines basis functions that depend on knot locations based on values in the

data. The knot locations selected by the MARS model are usually distinct data values and as the number of data points increase, the number of potential knot locations increases. To enable a more efficient knot selection for the purposes of surrogate optimization, Anahideh et al. [9] utilized a Classification And Regression Tree (CART) model to identify promising knot locations. They compared TK-MARS with an interpolating radial basis function (RBF) surrogate, a non-interpolating RBF surrogate, and a non-interpolating Gaussian Process (GP) surrogate. TK-MARS outperformed all others in a deterministic environment. In a noisy uncertain environment, TK-MARS outperformed non-interpolating GP and interpolating RBF surrogate optimization, while non-interpolating RBF was competitive in some instances. Given these results and the demonstrated ability of TK-MARS in intelligently identifying knot locations, this model was chosen as the surrogate in our study.

4.2.2 Candidate Search Algorithm

Our proposed methodology, QMARS-MIQCP-SUROPT uses a modified version of the EEPA algorithm developed by Dickson [11]. Algorithm 4 presents the original EEPA algorithm. The algorithm starts with the set of candidate points (orange circles in Figure 4.2) and their predicted response values. In line 1 of the algorithm, the minimum distance to the set of evaluated points (blue hexagons in Figure 4.2) is calculated for each point in the candidate set (orange circles in Figure 4.2). In line 2, a Pareto-frontier approach is used to determine a set of non-dominated points based on the minimum distance and the predicted response values of the set of candidate points. The while loop is used to iterate through the set of non-dominated Pareto points to limit its size to a maximum number of points depending on the cost of each function evaluation. In line 3 of the algorithm, a Maximin distance metric is used to select a set of diverse candidates and to eliminate the selection of points that are too close to each other. Finally, line 4 of the algorithm returns the selected limited set of EEPA candidates (purple triangles in Figure 4.2).

The reason that this methodology was selected as the desired candidate search algorithm was that it was demonstrated to work well in the literature [9] with the chosen

Algorithm 4:
EEPA

Result: Set of EEPA candidates
Input: Set of evaluated points, Set of candidate Points, Predicted response values for candidate points.
1 Calculate minimum distance of each candidate point to the set of evaluated points.
2 Determine non-dominated set of points using Pareto approach.
while *EEPA Candidates < Maximum number of EEPA Candidates* **do**
3 | Apply maximin criteria to non-dominated set of Pareto points.
4 | Return set of EEPA candidates.
end

surrogate, TK-MARS. However, some modifications were needed for use with the proposed QMARS-MIQCP-SUROPT approach. The main difference between TK-MARS based surrogate optimization and QMARS-MIQCP-SUROPT is that TK-MARS directly evaluates the points suggested by EEPA. Whereas QMARS-MIQCP-SUROPT optimizes the space around the suggested EEPA points and evaluates the optimal points found. We found that the additional optimization in our algorithm inherently included exploration of the space, and the Maximin distance metric applied in the EEPA algorithm led our optimization algorithm to take longer to converge to the optimum. To overcome this, we proposed a modified version of the EEPA in Algorithm 5 that that uses a sorting approach instead of the Maximin distance metric. This is referred to as sorted EEPA.

Lines 1 and 2 of Algorithm 5 are the same as Algorithm 4. The way in which sorted EEPA varies from the traditional EEPA algorithm is in lines 3–5. Once the set of non-dominated points based on the minimum distance and the predicted response values of the candidate points are identified, in line 3 of the algorithm, the set is then sorted based on the predicted response values from minimum to maximum. Then in line 4 of the algorithm, the maximum number of points, depending on the cost of each function evaluation, is initialized. In line 5 of the algorithm, the first N points are selected from the set of non-dominated Pareto points. Finally, line 6 of the algorithm returns the selected limited set of EEPA candidates (purple triangles in Figure 4.2).

Algorithm 5:
Sorted EEPA

Result: Set of EEPA candidates
Input: Set of evaluated points, Set of candidate points, Predicted response values for candidate points.

1 Calculate minimum distance of each candidate point to the set of evaluated points.
2 Determine non-dominated set of points using Pareto approach.
3 Sort non-dominated set of Pareto points based on predicted response values from minimum to maximum.
4 Set N = Maximum number of EEPA candidates.
5 Save top N sorted non-dominated set of Pareto points as EEPA candidates.
6 Return set of EEPA candidates.

Figure 4.3 shows the average response values of each function evaluation when optimizing the Rastrigin test function over 30 different iterations. The same set of initial points and candidate points were used for all three comparisons. In the figure, the yellow, blue and orange lines represent the average response values of each function evaluation for TK-MARS, QMARS-MIQCP-SUROPT using EEPA and QMARS-MIQCP-SUROPT using sorted EEPA respectively. Function evaluations 1 to 20 are the average response values for the initial set of points. Function evaluations 20 to 120 are the average response values obtained in consecutive iterations as a result of the surrogate optimization algorithm using EEPA and Sorted EEPA. For this experiment, the number of EEPA candidates were limited to 2. So, as an example, the average response values for function evaluations 20 and 21 were the result of one run of the corresponding surrogate optimization algorithms. It can be seen that even though, there is large variation in the average response values, TK-MARS consistently manages to identify points with lower response values.

For QMARS-MIQCP-SUROPT with EEPA it can be seen that the variation causes instability in our algorithm, preventing it from identifying points with lower response values. On the contrary, when using QMARS-MIQCP-SUROPT with sorted EEPA, it can be seen that it identifies points with lower response values and continues to trend downwards with each run of the surrogate optimization algorithm. QMARS-MIQCP-SUROPT with sorted EEPA even manages to beat TK-MARS on the average response value in a few instances,

resulting in a lower overall best known solution (BKS) when the algorithm is terminated.

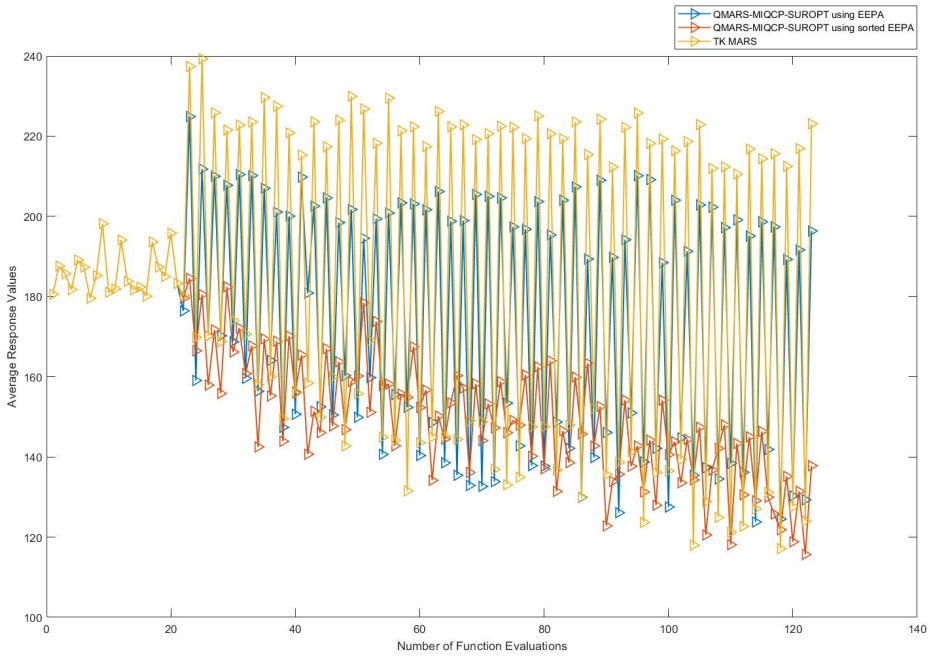

Figure 4.3: *Comparison of average response values*

4.2.3 Candidate Selection Strategy

As mentioned in section 4.2.2, QMARS-MIQCP-SUROPT optimizes the space around the selected EEPA candidates and evaluates the optimal solutions rather than directly evaluating the EEPA candidates. The optimization of the space is done by creating a ball in the space with the EEPA candidates as centers. This is formulated as an Euclidean distance constraint in equation 3.16 of the MIQCP formulation. An important factor to be considered in the formulation of the constraint, is the radius of the ball, ρ. The radius ρ has to be selected such that the previously evaluated points do not lie within or on the circumference of the ball to avoid selecting previously evaluated points as optimal solutions in consecutive iterations. To overcome this, QMARS-MIQCP-SUROPT uses the minimum distance of the EEPA candidate to the closest evaluated point $\delta(x)$ as the radius but reduces it by a factor γ to ensure that the closest previously evaluated point does not lie on the circumference of

the ball. Essentially, the radius of the ball can be calculated as:

$$\rho = \delta(x) - (\delta(x) * \gamma), \text{ where } 0 \leq \gamma \leq 1. \tag{4.1}$$

Evidently, the factor γ plays an important role in the radius ball over which the optimization is performed. We studied two different strategies for selecting the factor γ and it performance on the MIQCP:

- **Cyclical γ :** Cycling through different values of γ.
- **Fixed γ :** Fixing γ at one value for the entire optimization algorithm.

4.2.3.1 Cyclical γ

For the cyclical γ strategy, the γ value was chosen from a arbitrary list of values: [0.8,0.5,0.3,0.1], where the value 0.8 creates the smallest ball and the value 0.1 creates the largest ball. In addition to choosing γ from the list values, γ was cycled based on the number of function evaluations completed. Essentially, for the first 25% of evaluations QMARS-MIQCP-SUROPT uses a γ of 0.8, then for the next 25% of the evaluations uses a γ of 0.5 and so on.

The reason for not cycling through the list for γ at each function evaluation is because the radius is based on the distance to the closest previously evaluated point, and this distance decreases as more points are evaluated. Also, since QMARS-MIQCP-SUROPT selects optimal solutions for evaluation in consecutive iterations by optimizing the surrogate, using large radii to optimize the space in the initial iterations can result in in a misleading selection of points that subsequently reduces the efficiency of the algorithm.

Figure 4.4 shows the realized values of ρ for each function evaluation over an execution of the QMARS-MIQCP-SUROPT algorithm with cyclical γ. It can be seen that, realized ρ values follow a step-wise pattern starting with the smallest radius and then increasing in size as the algorithm progresses. While within each of the steps, a gradual decrease in the radius can be observed, starting with exploration and inherently shifting towards exploitation, there is an overall increasing pattern in the radius encouraging more exploration

in each step using the cyclical γ.

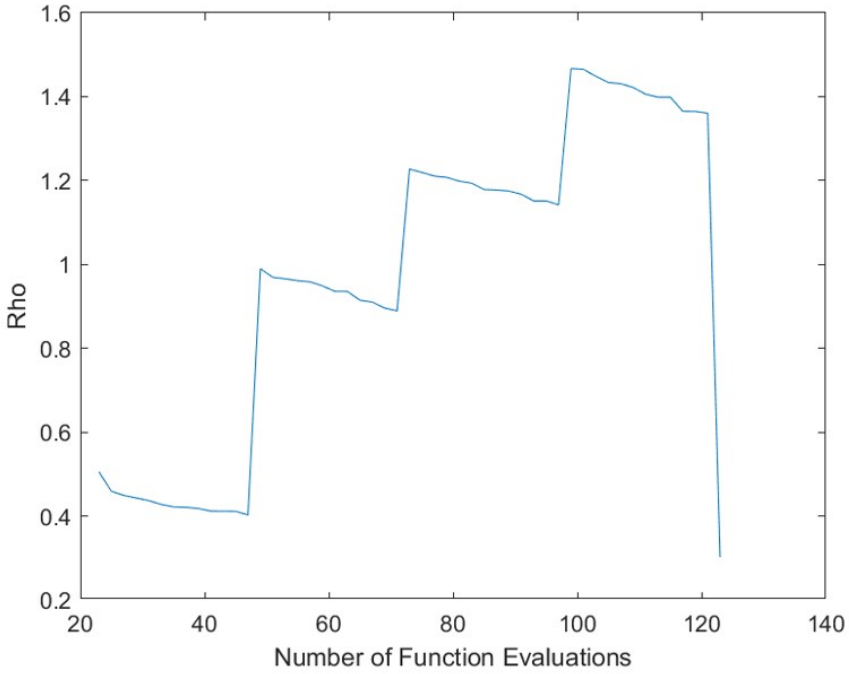

Figure 4.4: *Realized ρ values using Cyclical γ*

4.2.3.2 Fixed γ

For the fixed γ strategy, the γ value was set at 0.8. An ideal strategy to balance exploration and exploitation in any surrogate optimization algorithm is to start by doing exploration in the initial iterations, predominantly to gather information quickly about the space, and then switch to exploitation in the later iterations to make use of the gathered information to effectively reach the optimum. Because the radius ρ in equation (4.1) is based on the distance to the closest previously evaluated point, as the number of evaluated points increases, the radius ρ will decrease. Correspondingly, even if the γ value in equation (4.1) is fixed at 0.8, the associated radius ρ will gradually decrease as the algorithm iterates, facilitating exploration early in the algorithm and exploitation towards the end.

Figure 4.5 shows the realized values of ρ for each function evaluation over an execution of the QMARS-MIQCP-SUROPT algorithm with fixed γ. With this strategy, it can be seen that, the realized ρ values start with the largest radius circle and gradually decrease

in size as the algorithm progresses. This demonstrates the inherent exploration-exploitation nature of the algorithm, even with a fixed γ.

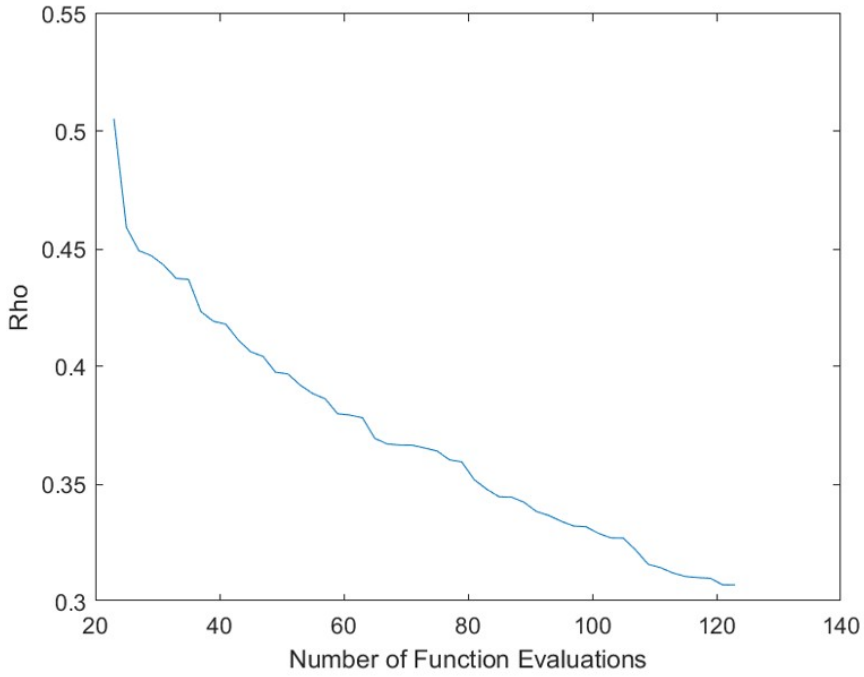

Figure 4.5: *Realized ρ values using Fixed γ*

4.2.4 Types of QMARS-MIQCP-SUROPT Algorithms

Based on the different strategies for candidate search and candidate selection, there are 4 types of QMARS-MIQCP-SUROPT algorithms that were studied:

- **QMS Algorithm 1:** QMARS-MIQCP-SUROPT with Sorted EEPA and Fixed γ
- **QMS Algorithm 2:** QMARS-MIQCP-SUROPT with EEPA and Fixed γ
- **QMS Algorithm 3:** QMARS-MIQCP-SUROPT with Sorted EEPA and Cyclical γ
- **QMS Algorithm 4:** QMARS-MIQCP-SUROPT with EEPA and Cyclical γ

Algorithm 6 is a more detailed version of Algorithm 3 that includes the steps of the QMARS-MIQCP-SUROPT algorithm with Sorted EEPA and Fixed γ.

In line 1 of Algorithm 6, as before, a very large DoE is generated to ensure good representation of the space. In line 2, the large DoE is split into a small initial set of points, which are also the first set of evaluated points represented as $X_{Evaluated}$, and a comparatively larger set with the remaining points called the candidate set. In the algorithm, since points chosen from the candidate set are used as centers for the optimization, the larger set is referred to as $X_{Centers}$. In line 3, the objective function is evaluated for the $X_{Evaluated}$ set. In line 4, the number of function evaluations completed $N_{Evaluations}$ is updated to the number of points evaluated so far. Line 5 initializes the maximum number of function evaluations, which is the stopping criteria for the algorithm and the list of γ values. Since this is the algorithm with fixed γ, the γ_{idx} value is set to 1, forcing the γ value to be fixed at 0.8. In line 6 of the algorithm, the surrogate model is fit to the set of evaluated points. The sorted EEPA algorithm is designed such that multiple centers can be chosen from the candidate set using exploration and exploitation and optimized simultaneously. Line 7 of the algorithm initializes the number of centers chosen in each iteration. In line 8, the sorted EEPA algorithm (algorithm 5) is run to identify promising center points X_{EEPA} for optimization. The distance of the center points to closest previous evaluated point $\delta(X_{EEPA})$ is retained for the formulation of the ball (refer to equation (4.1)). Lines 9–13 run the actual optimization of the fitted surrogate model in line 8 for the selected center points with the corresponding radius formulated using the obtained distances and γ. The optimization returns the set of promising new points X_{New}. The new points are evaluated in line 17, and the evaluated point set is updated in line 18. In line 19, the number of function evaluations completed is updated, correspondingly. The while loop is iterated until the number of function evaluations is equal to the maximum number of function evaluations set in line 7. The result at the end of the algorithm is the optimal setting and its corresponding objective value.

Algorithm 6:
Detailed QMARS-MIQCP-SUROPT Algorithm with Sorted EEPA and Fixed γ

Result: Optimal Parameter Settings: $X^*_{Evaluated}$ and $Y^*_{Evaluated}$
Generate Initial DoE with maximum number of points considered.
Split Initial DoE into $X_{Evaluated}$ and $X_{Centers}$ $s.t$ $|X_{Evaluated}| << |X_{Centers}|$.
Run function evaluation on $X_{Evaluated}$ to get $Y_{Evaluated}$;
Set $N_{Evaluations} = |X_{Evaluated}|$;
Initialize $Max_{Evaluations}$, $\gamma \in [0.8, 0.5, 0.3, 0.1]$, $\gamma_{idx} = 1$;
while $N_{Evaluations} <= Max_{Evaluations}$ **do**
    Fit QMARS model on $X_{Evaluated}$ and $Y_{Evaluated}$;
    Initialize N_{Points};
    Run sorted EEPA to obtain X_{EEPA} and $\delta(X_{EEPA})$;
    for *i = 1 to* $min(N_{Points}, |X_{EEPA}|)$ **do**
        $x_j = X_{EEPA}(i)$;
        Set $\rho = \delta(X_{EEPA}(i)) - (\delta(X_{EEPA}(i)) * \gamma(\gamma_{idx}))$;
        Formulate QMARS-MIQCP using fitted QMARS model, x_j and ρ;
        Optimize QMARS-MIQCP to obtain X_{New};
        Discard x_j and coinciding X_{New} from $X_{Centers}$;
    end
    Run function evaluation on X_{New} to get Y_{New};
    Add X_{New} to $X_{Evaluated}$ and Y_{New} to $Y_{Evaluated}$;
    $N_{Evaluations} = N_{Evaluations} + N_{Points}$;
end

4.3 Performance Metric for Comparing Surrogate Optimization Algorithms

For comparing the performance of the different QMARS-MIQCP-SUROPT algorithms, as well as comparing QMARS-MIQCP-SUROPT algorithms with other surrogate optimization algorithms, we use the Maximal True Function Area Under the Curve (MTFAUC) performance metric developed by Anahideh et al. [9]. The MTFAUC metric was chosen to assess performance because we were interested in testing QMARS-MIQCP-SUROPT algorithms in highly uncertain environments, and the MTFAUC metric was illustrated to be robust in evaluating the algorithms when the objective is observed under uncertainty. In uncertain environments, the best known solution is subject to random error and can oscillate up and down. Consequently, the traditionally used Area Under the Curve (AUC) metric will not show a monotonically non-decreasing shape (for minimization) as the number of iterations increases, and the AUC value could be misleading. A robust surrogate should have fewer and shorter oscillations in the final iterations, which results in a lower overall MTFAUC value.

4.4 Experiments and Results

To evaluate the performance of the algorithms, we use the different test functions listed in Table 4.1 that have commonly been used to test both global and surrogate optimization algorithms and have been deemed challenging [124]. We tested the performance of the algorithm under a highly uncertain setting, for which it is assumed that the simulated random noise follow a Gaussian distribution. For fitting the surrogate model, the observed response is the true function value plus the simulated random noise due to uncertainty. We follow the procedure similar to the discussion in [9]. The Gaussian noise is formulated and added as follows: $\tilde{f}(x) = f(x) + \epsilon$, where $\epsilon \sim N(0, p_\epsilon * \sigma_0)$, where $p_\epsilon \in (0, 1)$ is the percentage of noise level, and σ_0 is given by: $\sigma_0 = \max\{f(x^1, ..., f(x^N)\} - \min\{f(x^1), ..., f(x^N))\}$. In our study p_ϵ was set to 0.25.

Function	Formulation	Range	Global Min
Rosenbrock	$f(x) = \sum_{i=1}^{d-1} \left[100 \left(x_{i+1} - x_i^2\right)^2 + \left(x_i - 1\right)^2\right]$	$[-5, 10]$	$f(x^*) = 0;$ $x^* = (1, ..., 1)$
Rastrigin	$f(x) = 10d + \sum_{i=1}^{d} \left[x_i^2 - 10\cos\left(2\pi x_i\right)\right]$	$[-5.12, 5.12]$	$f(x^*) = 0;$ $x^* = (0, ..., 0)$
Ackley	$f\left(x_0 \cdots x_n\right) = -20\exp\left(-0.2\sqrt{\frac{1}{d}\sum_{i=1}^{d} x_i^2}\right)$ $-\exp\left(\frac{1}{d}\sum_{i=1}^{d}\cos\left(2\pi x_i\right)\right) + 20 + e$	$[-32.768, 32.768]$	$f(x^*) = 0;$ $x^* = (1, ..., 1)$

Table 4.1: Global optimization test functions

Anahideh et al. [9] also mention that the TK-MARS surrogate model is highly capable of identifying unimportant factors. We also study the performance of our algorithm and its effect on TK-MARS in handling unimportant factors, and we set the fraction of important variables to 0.5. The results compare the data from 30 runs of the algorithms. The 30 runs use the same initial pool and 30 different candidate pools. The same data sets were used for both the QMARS-MIQCP-SUROPT and TK-MARS algorithms. Table 4.2 shows the test parameters and the parameters specific to the algorithms that have been used in the study and shown in the results. In addition to the parameters specific to the algorithms, there are certain parameters specific to the surrogate QMARS model, which are listed in Table 4.3. Section 4.4.1 compares the different QMARS-MIQCP-SUROPT algorithms to identify the best algorithm. Section 4.4.2 compares the best algorithm identified in Section 4.4.1 with its competitor TK-MARS.

Parameters	**Settings**
Test Functions	Rosenbrock, Rastrigin, Ackley
Dimensions (d)	30
Fraction of Important Variables (fiv)	0.50
Noise Level (np)	0.25
Initial Pool Size (N)	$d + 1$
DoE	LHD
Surrogate Model	TK-MARS
Candidate Search Algorithm	EEPA, Sorted EEPA
Number of EEPA Candidates	3
Candidate Selection	Fixed γ, Cyclical γ
Number of Function Evaluations	125
Number of Runs	30

Table 4.2: Experimental parameters

Parameters	Settings
$M_{\max}$	$\left\lfloor \frac{2n+3}{2+3} \right\rfloor$
circle	0
levels	0
knots	0
alg3	0
ASR	1
difference	0.02
interactions	1
robust	0
tolerance	0.5
convex	0
gap	0
smooth	0

Table 4.3: QMARS parameters

4.4.1 Comparison of QMARS-MIQCP-SUROPT Algorithms

Figure 4.6 shows the box plots of MTFAUC over 30 runs for each QMS algorithm type mentioned in Section 4.2.4 for the Rosenbrock test function. It can be seen that the medians for the boxes for QMS Algorithms 1 and 2, which use fixed γ, are comparatively lower, and the medians for QMS Algorithms 3 and 4, which use cyclical γ, are comparatively higher. We can infer that using fixed γ has an effect on providing lower MTFAUC. Overall, it can be seen that both the median and mean for QMS Algorithm 1 - QMARS-MIQCP-SUROPT with sorted EEPA and fixed γ are the lowest.

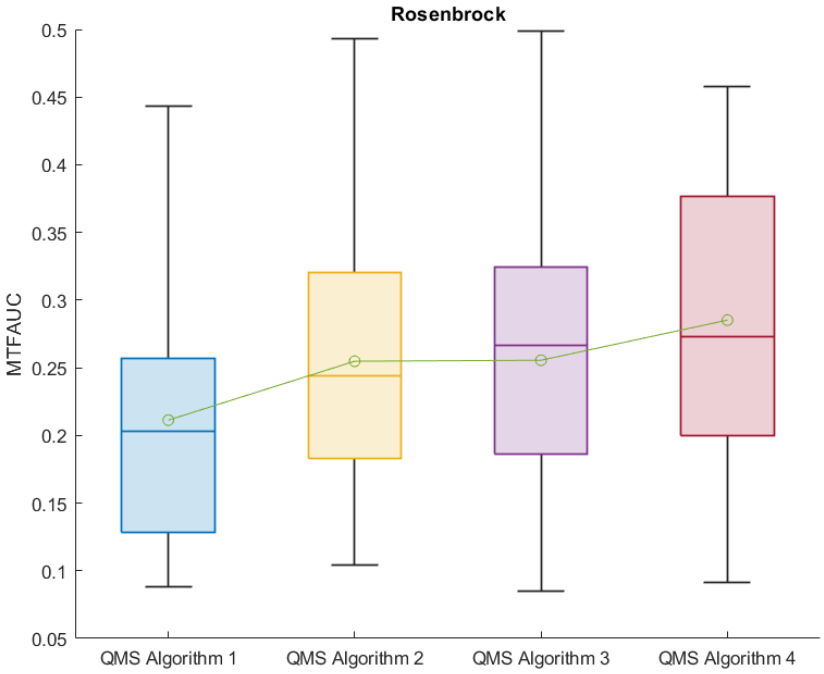

Figure 4.6: *Comparison of different QMARS-MIQCP-SUROPT algorithms - Rosenbrock*

Figure 4.7 shows the box plots of MTFAUC over 30 runs for each QMS algorithm type mentioned in Section 4.2.4 for the Rastrigin test function. It can be seen that the means for the boxes for QMS Algorithms 1 and 3, which use sorted EEPA, are comparatively lower. Even though the median for QMS Algorithm 3 is slightly higher than QMS Algorithm 4, which uses EEPA, Algorithm 3 has an outlier with the lowest MTFAUC. We can infer that using fixed γ has an effect on providing lower MTFAUC. Overall, it can be seen that both the median and mean for QMS Algorithm 1 - QMARS-MIQCP-SUROPT with sorted EEPA and fixed γ are the lowest.

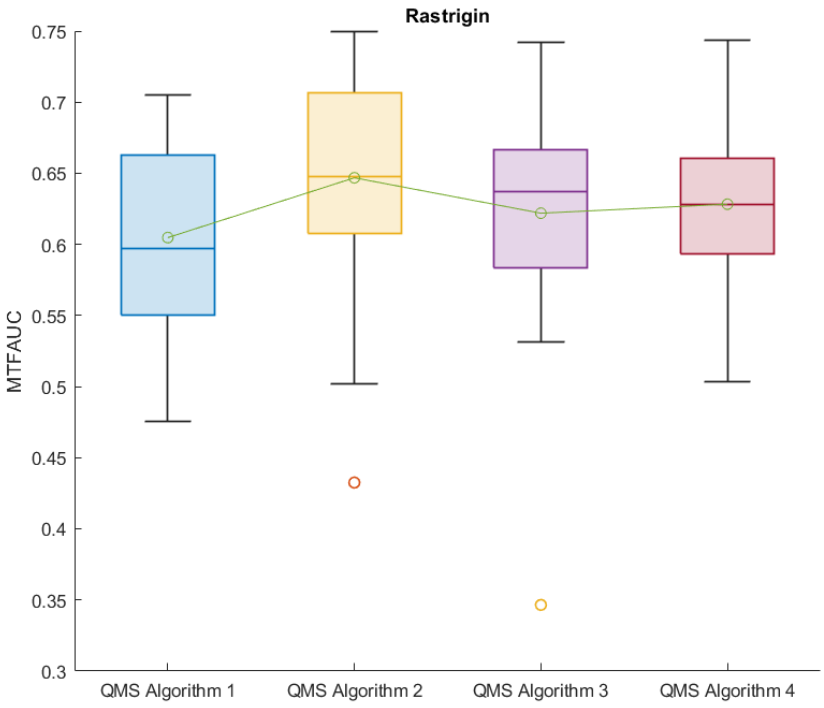

Figure 4.7: *Comparison of different QMARS-MIQCP-SUROPT algorithms - Rastrigin*

Figure 4.8 shows the box plots of MTFAUC over 30 runs for each QMS algorithm type mentioned in Section 4.2.4 for the Ackley test function. It can be seen that the median for the boxes for QMS Algorithms 1 and 3, which use sorted EEPA, are comparatively lower. Whereas, the median for the boxes for QMS Algorithms 2 and 4, which use EEPA, are comparatively higher. It is also notable that the medians for QMS Algorithms 1 and 3 are similar to each other, but the variation for the box for QMS algorithm 1 is lower. We can infer that using fixed γ has an effect on providing lower MTFAUC. Overall, it can be seen that both the median and mean for QMS Algorithm 1 - QMARS-MIQCP-SUROPT with sorted EEPA and fixed γ are the lowest.

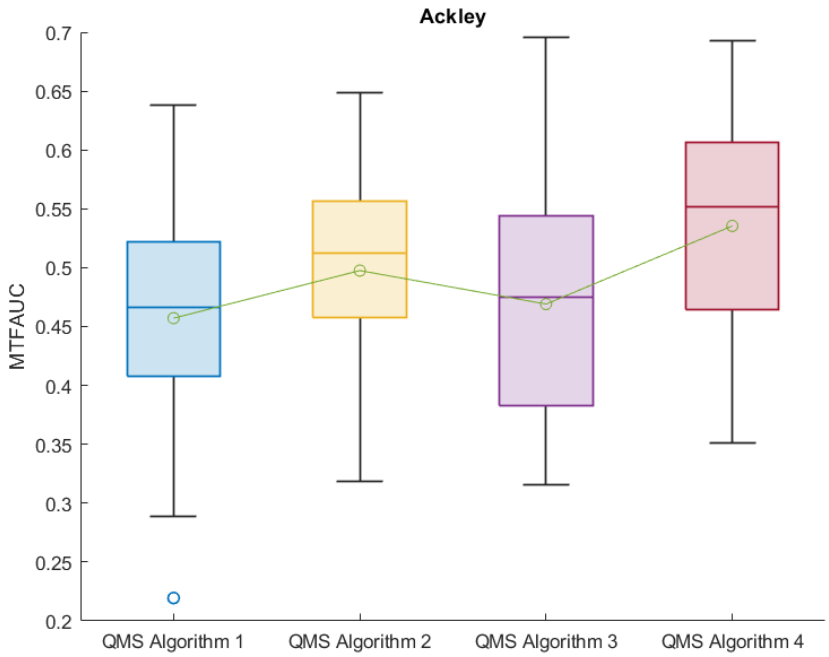

Figure 4.8: *Comparison of different QMARS-MIQCP-SUROPT algorithms - Ackley*

In conclusion, QMS Algorithm 1 - QMARS-MIQCP-SUROPT with sorted EEPA and fixed γ had the lowest mean and median for all three test functions, and boxes that were overall lower. QMS Algorithm 1 - QMARS-MIQCP-SUROPT with sorted EEPA and fixed γ is considered to perform better than the other algorithm types, and subsequent computational results will focus on this algorithm. The next section presents a comparison to TK-MARS.

4.4.2 Comparison of QMARS-MIQCP-SUROPT vs TK-MARS

In this section we use MTFAUC to compare TK-MARS from Anahideh et al. [9] against QMARS-MIQCP-SUROPT with sorted EEPA and fixed γ, which was found to perform best among the QMS algorithms. Figures 4.9 - 4.11 show the box plots of MTFAUC over 30 runs for each algorithm tested on the Rosenbrock, Rastrigin and Ackley test functions. It can be seen that the medians for the boxes for our proposed methodology, QMARS-MIQCP-SUROPT with sorted EEPA and fixed γ, is considerably lower than TK-MARS in all three

cases. We can infer that our proposed methodology yields better performance.

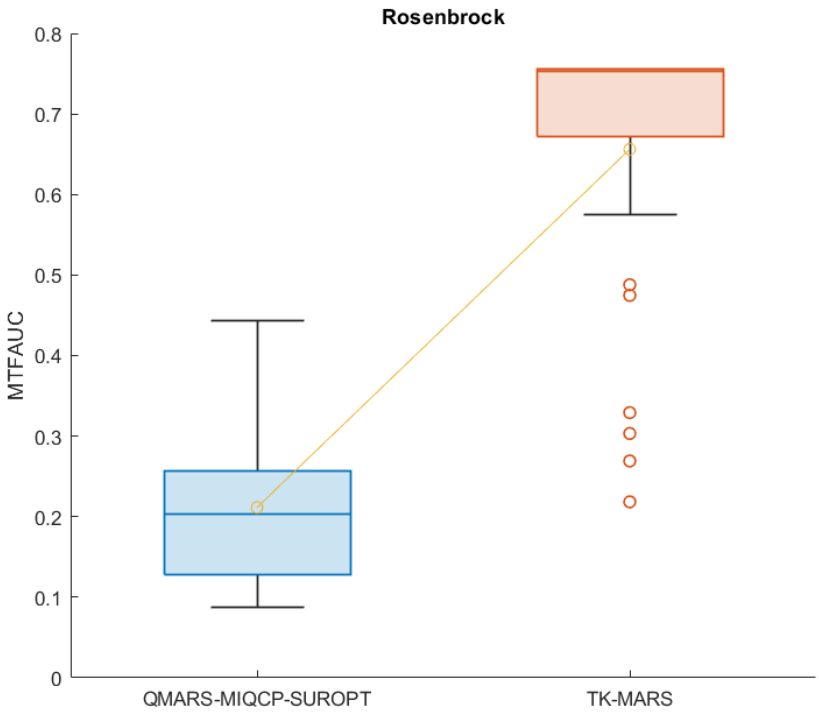

Figure 4.9: *Comparison of QMARS-MIQCP-SUROPT vs TK-MARS - Rosenbrock*

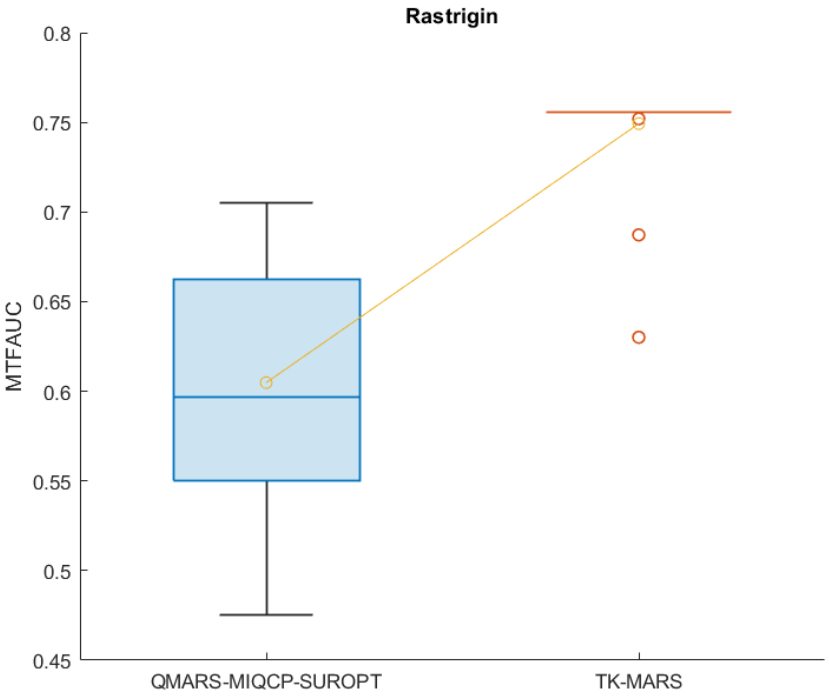

Figure 4.10: *Comparison of QMARS-MIQCP-SUROPT vs TK-MARS - Rastrigin*

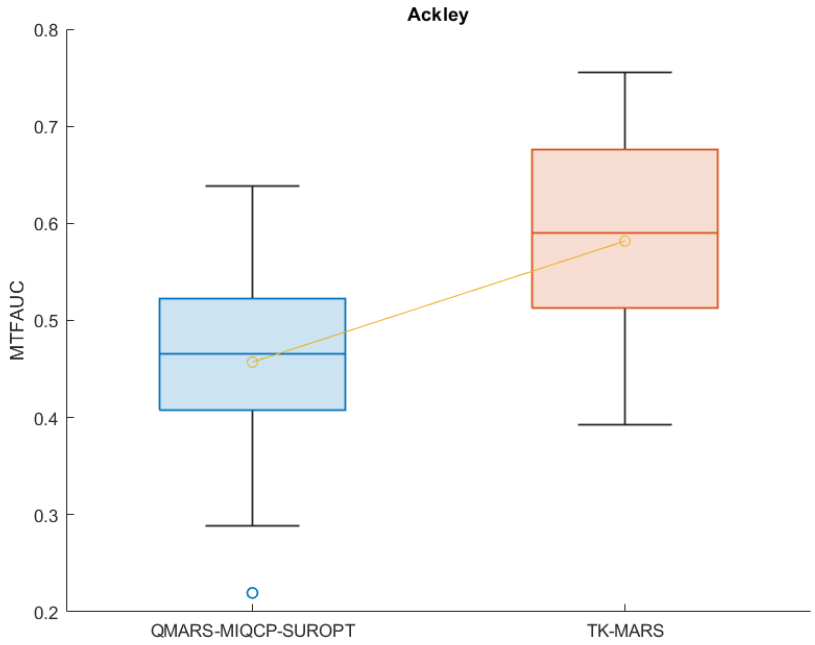

Figure 4.11: *Comparison of QMARS-MIQCP-SUROPT vs TK-MARS - Ackley*

Chapter 5

Application of QMARS-MIQCP-SUROPT in Analytical Chemistry

As mentioned in earlier sections, one of the motivations for this research is addressing the lack of optimization algorithms applied specifically in optimizing ionization efficiency using mass spectrometry. It was seen in the literature that most closely related applications in analytical chemistry use statistical analysis methods such as DOE, RSM or even MARS models to predict the relationships between the different parameter settings of the instrument, but no actual optimization [119, 113, 116, 115, 117, 114, 108, 110, 111]. There were two instances in the literature that used surrogate optimization with a Kriging or Gaussian Process (GP) as the surrogate model [122, 123]. There is research that use mathematical modeling and formal optimization, but in all of the instances, the application was superstructure optimization. Consequently, there is no evidence of using surrogate optimization to guide a series of laboratory runs for efficient analyte extraction in mass spectrometry.

In this chapter, we demonstrate our QMARS-MIQCP-SUROPT methodology for a real-world application that requires a series of laboratory runs. The goal is to identify optimal parameter settings for the flow injection analysis of an analyte in the Shimadzu LCMS-2020

instrument. The performance of the methodology was also tested for five different analytes on ChromSim, a simulation software developed by Dr. Dwight Stoll and team [125, 126].

5.1 Experiments and Results for ChromSim

First, we tested our methodology for five different analytes on the simulation software, ChromSim. ChromSim was developed to simulate a 2D-LC machine with pumps, filters and columns and the separated components being detected by a Diode-Array UV Detector. The schematic diagram of the system being simulated in ChromSim is shown in figure 5.1. The figure was obtained from the supplemental materials for [126]. The literature compares chromatograms, retention times and peak widths for various analytes from the simulation software ChromSim, to the actual laboratory experiments and were able to achieve similar results to the laboratory experiments.

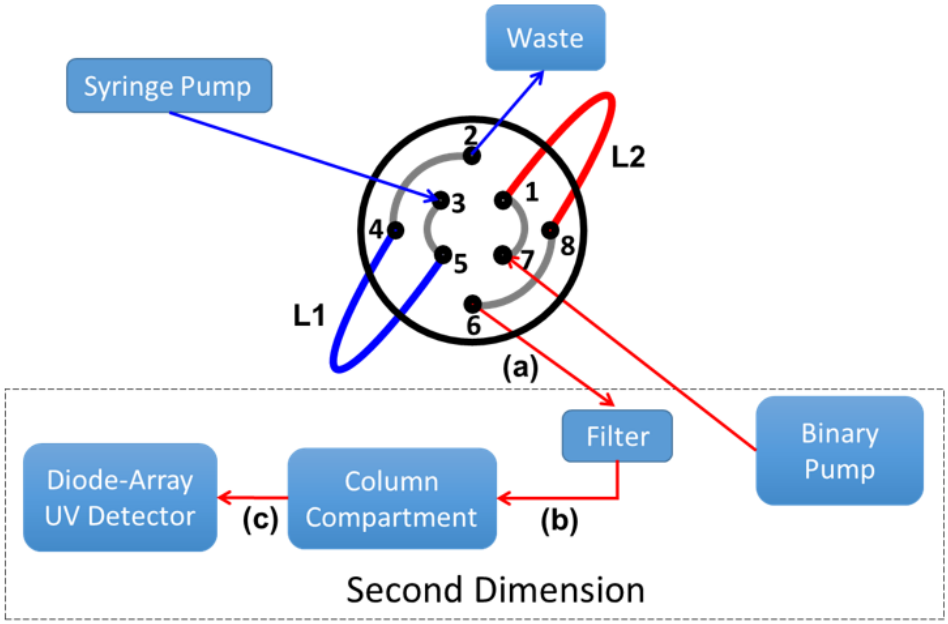

Figure 5.1: *Schematic diagram of the system being simulated in ChromSim*

Even though, the setup of the simulation software is entirely different in comparison to the real-world Shimadzu LCMS-2020 instrument, testing our methodology using simulation software provides guidance for how our methodology works for an analytical chemistry

instrument and the setup needed to do the actual laboratory runs. This will also help demonstrate that our methodology is versatile and can be applied to any analytical chemistry instrument in which there is a need to optimize parameter settings of the instrument with minimal expensive laboratory experiments.

We tested the performance of the QMARS-MIQCP-SUROPT with sorted EEPA and fixed γ methodology described in Section 4.2 on ChromSim in extracting five different sample analytes defined based on their descriptors in Table 5.1. Our methodology was run five times for each of the different analytes. Recalling Figure 1.1, in this application, the input space (X) is represented by the five target analytes in Table 5.1. The decision space (Y) is represented by six different system parameters that were varied to test their effects on the performance of the separation the analytes:

- Number of Plates - plates
- Void Volume - $V_m(ml)$
- Mobile phase composition before gradient - Φ_{dip}
- Initial mobile phase composition - Φ_{initial}
- Total dwell volume of injection loop - V_{loop}
- Percentage of the loops filled - %Fill

Finally, the outcome space is represented by the width of the chromatograms at 50% of the height ($Width_{50}$), which is minimized. Figure 5.2 shows how the value of $Width_{50}$ is calculated and returned as an output in ChromSim.

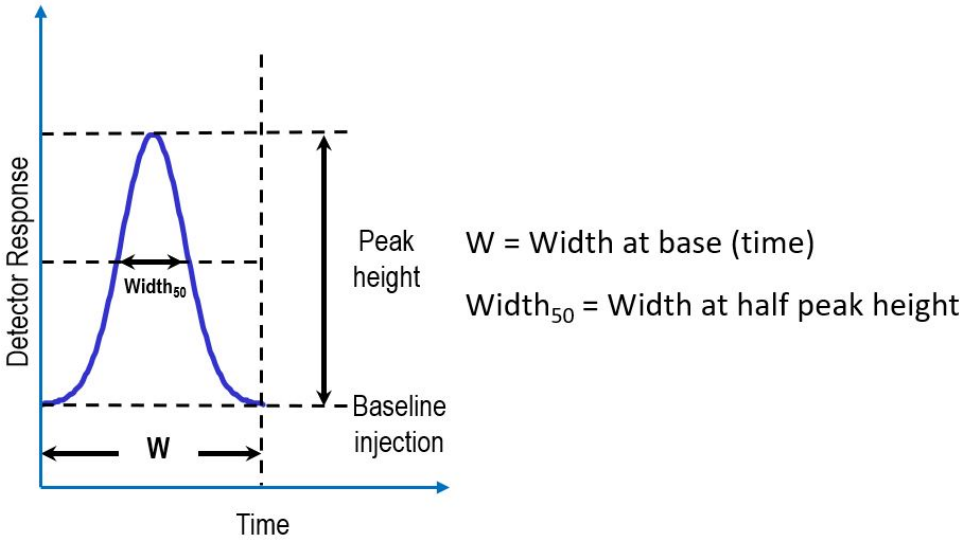

Figure 5.2: *Sample diagram representing calculation of* $Width_{50}$

Analytes	$\mathbf{K_w}$	**B**	**a**
Analyte 1	37.337	20.96	1.903
Analyte 2	53.1964	24.8364	0
Analyte 3	39.5963	18.1927	0
Analyte 4	41.8065	19.3678	0
Analyte 5	451.8365	28.9407	0

Table 5.1: Sample analytes tested on ChromSim

For the QMARS-MIQCP-SUROPT algorithm, it is necessary to set the parameter ranges on the system parameters to obtain meaningful optimal solutions. These ranges are shown in Table 5.2. For this experiment, a large DOE based on an LHD with 1000 points was generated. Out of the 1000 points, 7 points were randomly selected for the initial evaluation, and the remaining points were chosen as the set of candidate points. The number of EEPA candidates was set to 3, and a maximum of 35 function evaluations was set as the stopping criteria for the algorithm. The same set of points were chosen for the initial evaluation and as the set of candidate points for all 5 analytes. A process similar to what was conducted in the laboratory was followed for the simulation. To start the algorithm, the 7 initial points

were run using ChromSim and $Width_{50}$ was recorded as the response. After this, QMARS-MIQCP-SUROPT with sorted EEPA and fixed γ was run to generate 3 potential parameter settings for evaluation. These settings were then run using ChromSim, and $Width_{50}$ was recorded until the stopping criteria was met. Figures 5.3–5.7 show the path of the best known minimum $Width_{50}$ for each of the QMARS-MIQCP-SUROPT iterations and analyte. It can be seen that path is different for each of the analyte.

Parameter	**Minimum**	**Maximum**
plates	1000	3000
V_m (ml)	0.055	0.5
Φ_{dip}	0.1	0.9
Φ_{initial}	0.1	0.5
V_{loop}	5	500
%Fill	0.1	0.9

Table 5.2: System parameter ranges for ChromSim

In addition to running the QMARS-MIQCP-SUROPT algorithm with the parameters mentioned in Tables 4.2 and 4.3, we also wanted to leverage the ability of the QMARS surrogate model to automatically detect and model interactions. The methodology was tested with setting the parameter *interactions* as follows:

- *interactions* = 1, meaning no interactions (represented by the blue line),
- *interactions* = 2 with *robust* = 1 (represented by the orange line) throughout the experiment, and
- starting with *interactions* = 1 and then switching the setting to *interactions* = 2 with *robust* = 1 (represented by the yellow line) after $2d$ parameter settings have been evaluated.

For example for the last setting, if the dimension (d) is 7, then after 14 runs have been evaluated, the settings are switched automatically.

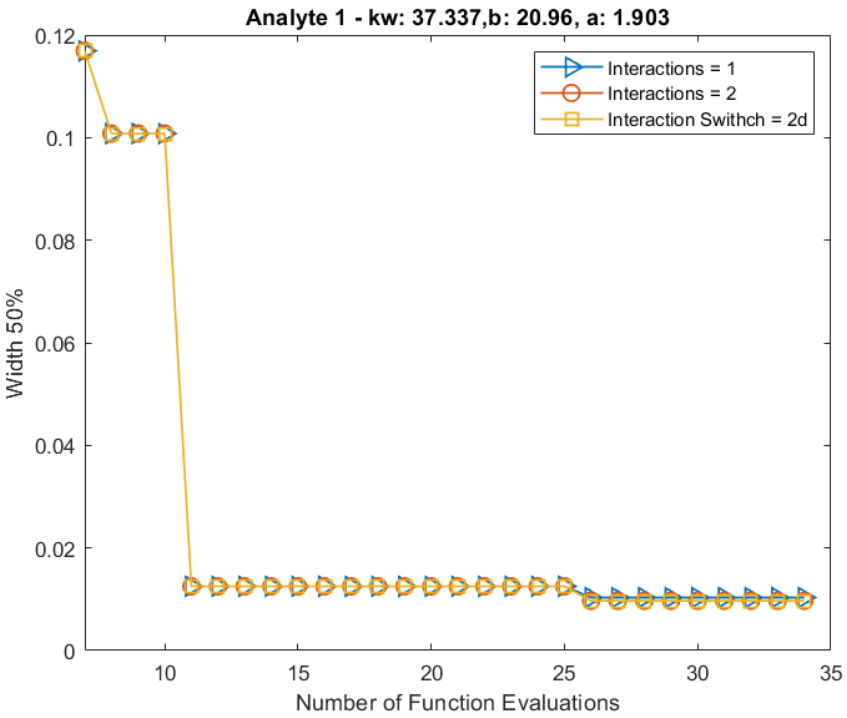

Figure 5.3: *Best Known Width 50% - Analyte 1*

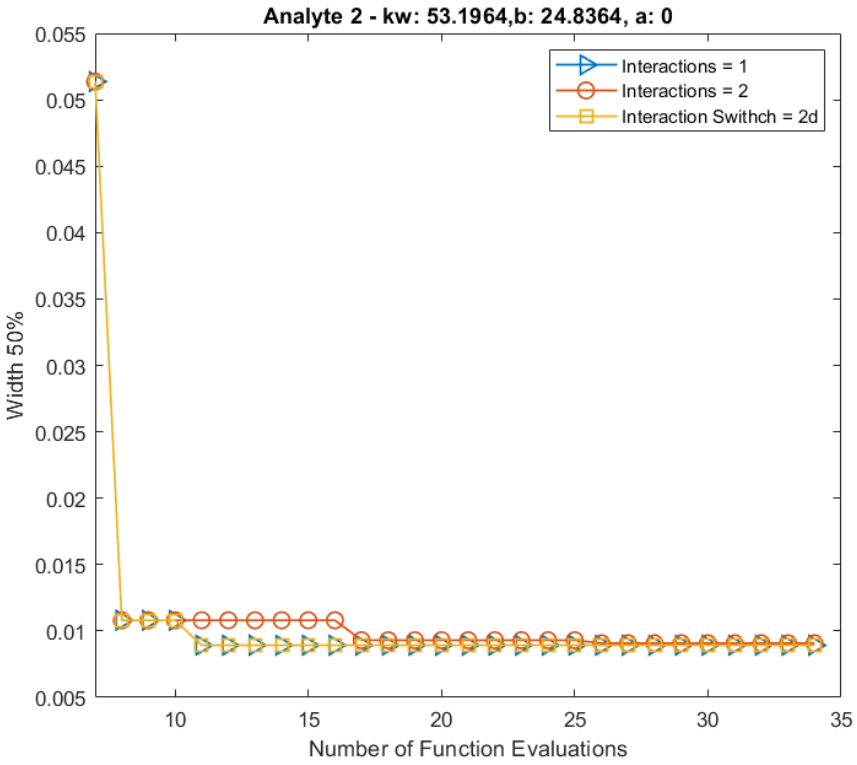

Figure 5.4: *Best Known Width 50% - Analyte 2*

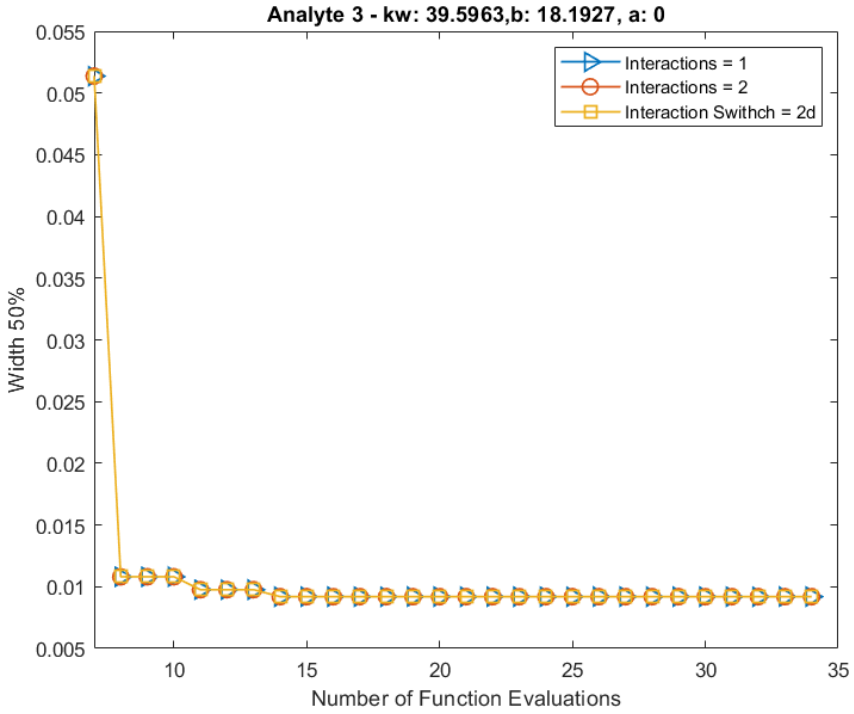

Figure 5.5: *Best Known Width 50% - Analyte 3*

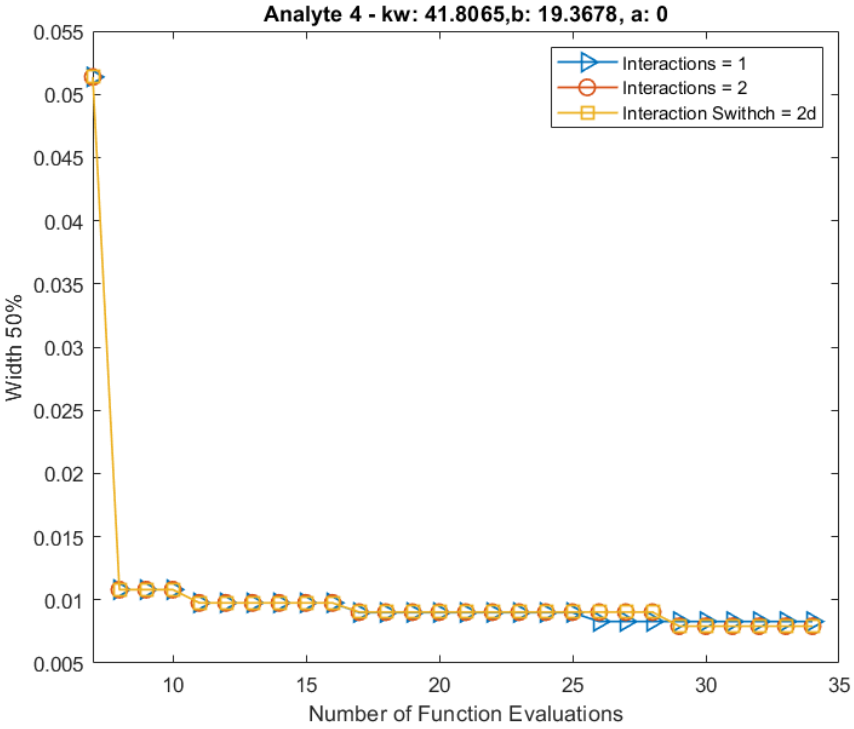

Figure 5.6: *Best Known Width 50% - Analyte 4*

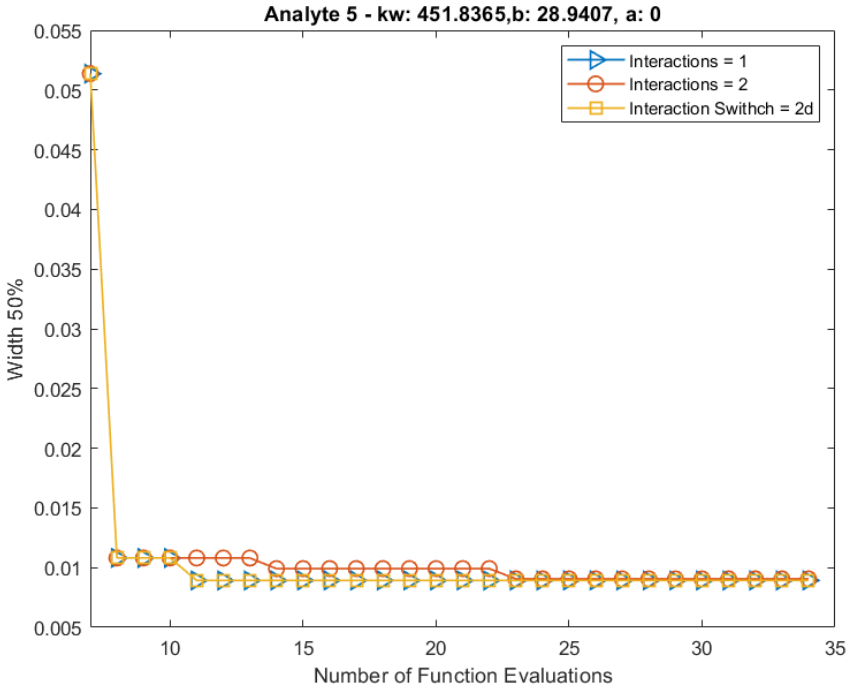

Figure 5.7: *Best Known Width 50% - Analyte 5*

An optimal setting for each analyte and QMARS setting combination is also shown in table 5.3.

Analytes	**Interaction**	**plates**	$\mathbf{V_m}$**(ml)**	$\mathbf{\Phi_{dip}}$	$\mathbf{\Phi_{initial}}$	$\mathbf{V_{loop}}$	**%Fill**	$\mathbf{Width_{50}}$
Analyte 1	1	2062	0.07	0.9	0.5	5	0.2	0.0103
	2	2237	0.06	0.9	0.5	5	0.2	0.0097
	2d	2237	0.06	0.9	0.5	5	0.2	0.0097
Analyte 2	1	2998	0.06	0.9	0.2	5	0.5	0.0089
	2	2465	0.06	0.4	0.3	5	0.1	0.0091
	2d	2998	0.06	0.9	0.2	5	0.5	0.0089
Analyte 3	1	2703	0.07	0.4	0.2	5	0.1	0.0092
	2	2703	0.07	0.4	0.2	5	0.1	0.0092
	2d	2703	0.07	0.4	0.2	5	0.1	0.0092
Analyte 4	1	2009	0.09	0.2	0.3	5	0.4	0.0083
	2	2990	0.09	0.2	0.1	5	0.9	0.0079
	2d	2990	0.09	0.2	0.1	5	0.9	0.0079
Analyte 5	1	2998	0.06	0.9	0.2	5	0.5	0.0089
	2	2704	0.06	0.3	0.3	5	0.1	0.0091
	2d	2998	0.06	0.9	0.2	5	0.5	0.0089

Table 5.3: Optimal parameter settings - ChromSim

It can also be seen from Figures 5.3–5.7 that switching the setting to *interactions* = 2 with *robust* = 1 after $2d$ parameter settings (represented by the yellow line) performs similarly to setting *interactions* = 1 for most of the cases, which is why the lines for the other settings are not visible. In a couple of instances, it has a better minimum $Width_{50}$. We interpret this to imply that the ChromSim application does not involve interaction effects between the studied input parameters.

5.2 Experiments and Results for Shimadzu LCMS-2020

The term LC-MS stands for Liquid Chromatography Mass Spectrometry. Liquid chromatography is a separation technique commonly used in analytical chemistry to separate the components of a sample based on their retention strength in the stationary phase and mobile phase inside a LC column. Over the years, there have been several advances in LC resulting in techniques such as high-performance liquid chromatography (HPLC) that are capable of separating components with higher speeds and improved sensitivity. Upon separation, the different components in the samples can be detected/analyzed using different techniques. One such technique is electrospray ionization, where the separated components are ionized at atmospheric pressure and the generated ions are then extracted into a highly vacuum based environment of the mass spectrometer after much a majority of the solvent is evaporated where the ions are sorted based on their mass-to-charge ratios (m/z), resulting in a time-series plot of the ion intensities for each component in the sample. A key information to note in our application is that, we simply performed flow injection analysis to create a temporal slug of the analyte, to flow into the mass spectrometer. So, the analyte did not pass through the LC system of the instrument. Instead, the analyte was injected through the auto sampler directly into the MS system.

Figure 5.8 obtained from [127] shows the basic components of a MS system. A typical MS system consists of an ion source which ionizes the components and uses an ion guide to precisely create and deliver a fine spray of the ionized components into the mass analyzer, where the ions are sorted based on their mass-to-charge ratios (m/z) and finally, the detector detects and quantifies the ion intensities of the separated components.

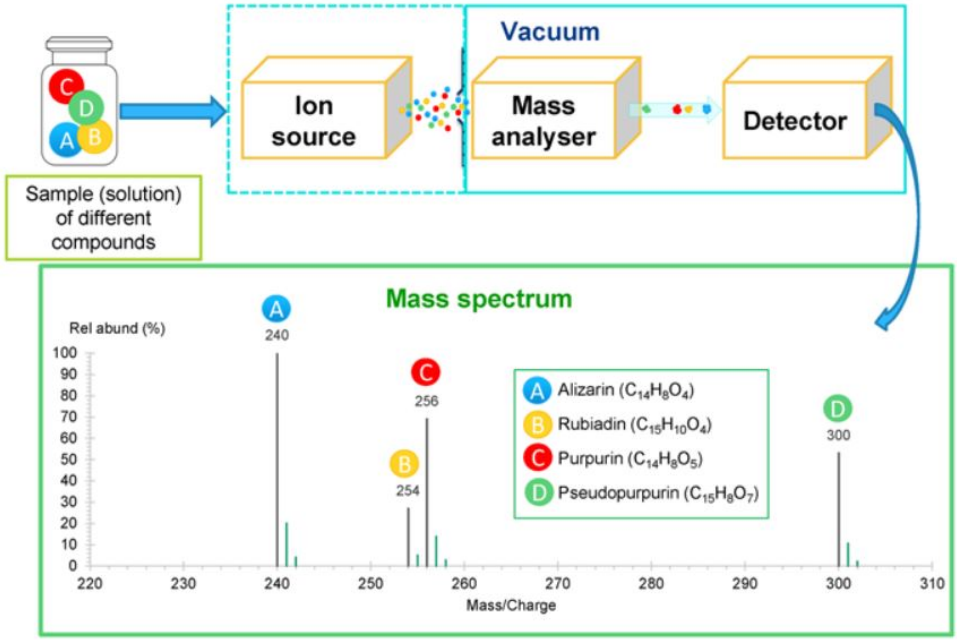

Figure 5.8: *Basic components of a MS system*

Figure 5.9 obtained from [128] shows a closer look at the components of the Electrospray Ionization Source in the MS system. This component is specifically of interest in this application because the system parameter settings that we optimize to test their effects on the ionization efficiency pertain specifically to this component of the system. The figure was modified to show the system parameter settings that were varied specific to the application in the boxes with the dotted lines.

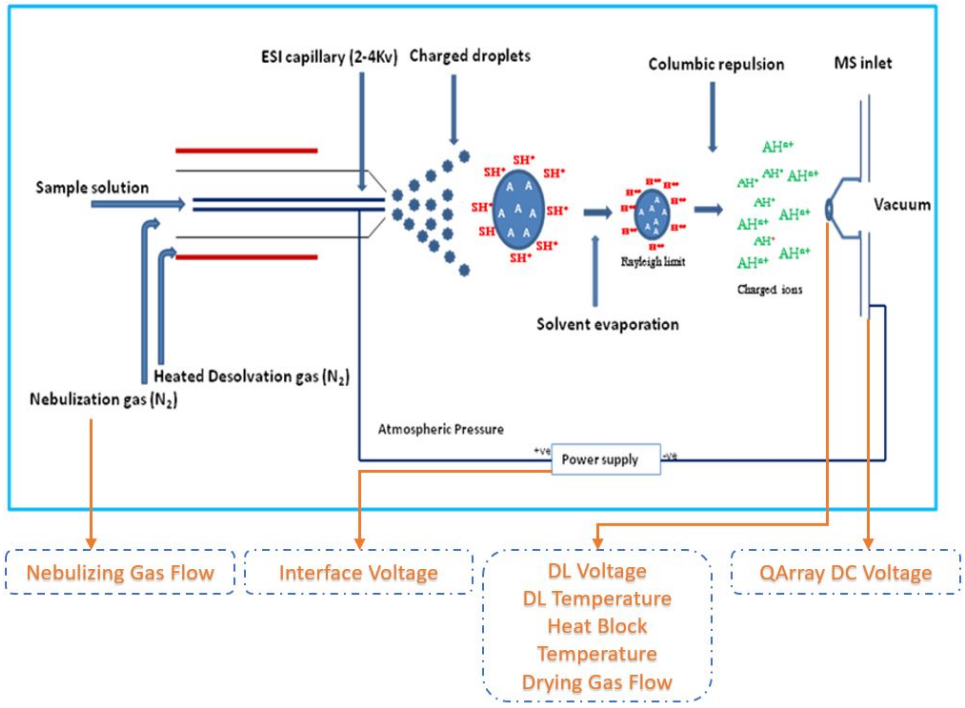

Figure 5.9: *Basic components of an Electrospray Ionization source*

The performance of the QMARS-MIQCP-SUROPT with sorted EEPA and fixed γ methodology described in Section 4.2 was tested in the LCMS-2020 instrument in extracting a sample analyte, Acetaminophen. Recalling Figure 1.1, in this application, the input space (X) is represented by the target analyte Acetaminophen. The decision space (Y) is represented by seven different system parameters that were varied to test their effects on the ionization efficiency:

- Interface Voltage (V)
- DL Voltage (V)
- Qarray DC Voltage (V)
- DL Temperature $°C$
- Heat Block Temperature $°C$
- Nebulizing Gas Flow (mL/min) and Drying Gas Flow (mL/min)

Specifically these parameter settings pertain to the different components of the ion source shown in figure 5.9. Finally, the outcome space is represented by the detected Ion Intensity, and is goal is to maximize this. For the QMARS-MIQCP-SUROPT algorithm, it is necessary to set the parameter ranges on the system parameters to obtain meaningful optimal solutions. These ranges are shown in Table 5.4.

Parameter	**Minimum**	**Maximum**
Interface Voltage (V)	1	5
DL Voltage (V)	10	150
Qarray DC Voltage (V)	10	150
DL Temperature $°C$	100	300
Heat Block Temperature $°C$	100	500
Nebulizing Gas Flow (mL/min)	0.5	1.5
Drying Gas Flow (mL/min)	3	20

Table 5.4: System parameter ranges for the LCMS-2020 instrument

In addition to these parameter ranges, there was also an inherent system constraint that the absolute difference between the DL Voltage and the Qarray DC Voltage should not exceed 100 Volts. These types of constraints are very common in real-world applications and our methodology was able adapt and include the restrictions by adding the constraint:

$$|DL\ Voltage - Qarray\ DC\ Voltage| \leq 100. \tag{5.1}$$

Additionally, any points generated in the DOE that do not satisfy this constraint were also removed as infeasible, so that they would not be considered by the surrogate model. It should also be noted that the instrument inherently rounds the parameters DL Voltage, Qarray DC Voltage, DL Temperature, Heat Block Temperature to the nearest integer and Interface Voltage, Nebulizing Gas Flow and Drying Gas Flow to one decimal point. This is also included in the algorithm when generating the DOE, as well as the new points for

evaluation in each consecutive iteration.

For this experiment, a large DOE based on an LHD with 1000 points, including all of the constraints, was generated. Out of the 1000 points, 8 points were randomly selected for the initial evaluation, and the remaining points were chosen as the set of candidate points. The number of EEPA candidates was set to 3, and a maximum of 30 function evaluations was set as the stopping criteria for the algorithm. To start the algorithm, the 8 initial points were run on the machine in triplicate, and the average ion intensities were recorded as the response. Because this is an actual laboratory experiment, the responses are subject to uncertainty. The QMARS-MIQCP-SUROPT algorithm with sorted EEPA and fixed γ was executed to generate 3 potential parameter settings for evaluation. These settings were then run on the instrument in triplicate, and the average ion intensities were recorded. This process was repeated until a maximum of 30 parameter settings were obtained.

Also, in the previous section, it was found that switching the setting to *interactions* = 2 with *robust* = 1 after $2d$ parameter settings performed similarly to setting *interactions* = 1. For the laboratory experiments, the algorithm was run with both settings, and no difference was seen in the candidate points chosen in consecutive iterations. Consequently, this study also does not appear to require modeling of interactions.

Figure 5.10 shows the path of the best known maximum ion intensities for each of the QMARS-MIQCP-SUROPT iterations. It can be seen that a maximum ion intensity of around 5.9 million was obtained in the 28^{th} function evaluation. It was also noted that the maximum ion intensity obtained is very close to the ion intensity of around 6.5 million observed in the tuning file. The difference is due to different parameter setting ranges used in the tuning file than the ones in the algorithm.

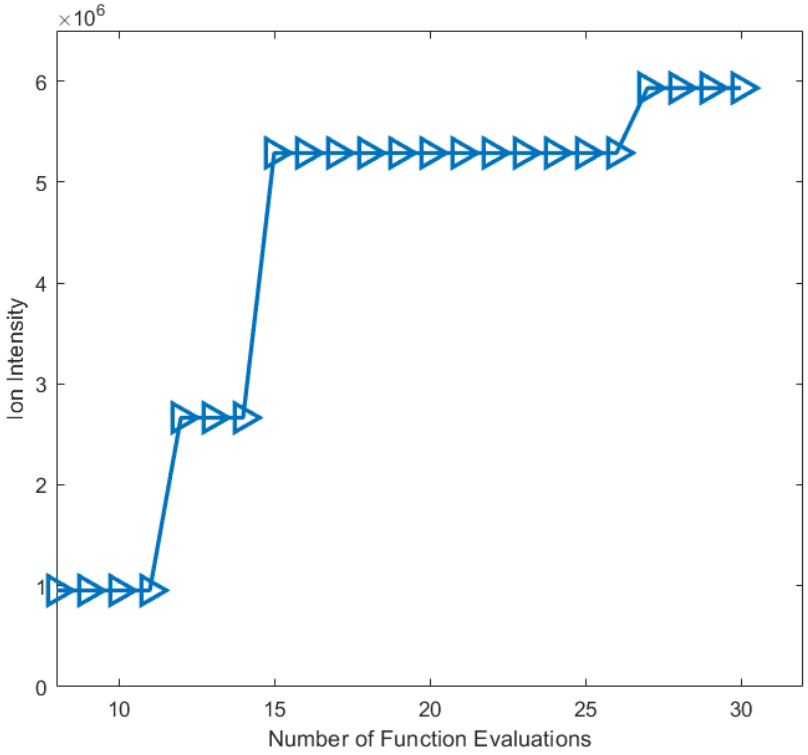

Figure 5.10: *Best Known Maximum Ion Intensity - Acetaminophen*

Figure 5.11 shows the comparison of the chromatograms for the best known solution out of the 8 initial points that were evaluated at the start of the algorithm (top) and the optimal setting (bottom). It can be seen that the chromatograms are very different from each other. The chromatogram on the top is very noisy since it is a sub optimal solution that did not yield efficient ionization. By contrast, the chromatogram for the optimal setting is more smooth and has a defined peak in the area where the ionization intensity was measured. It is also should be noted that the y-axis scales on the chromatograms are different. The identified optimal setting for achieving a maximum ion intensity of around 5.9 million is shown in Table 5.5.

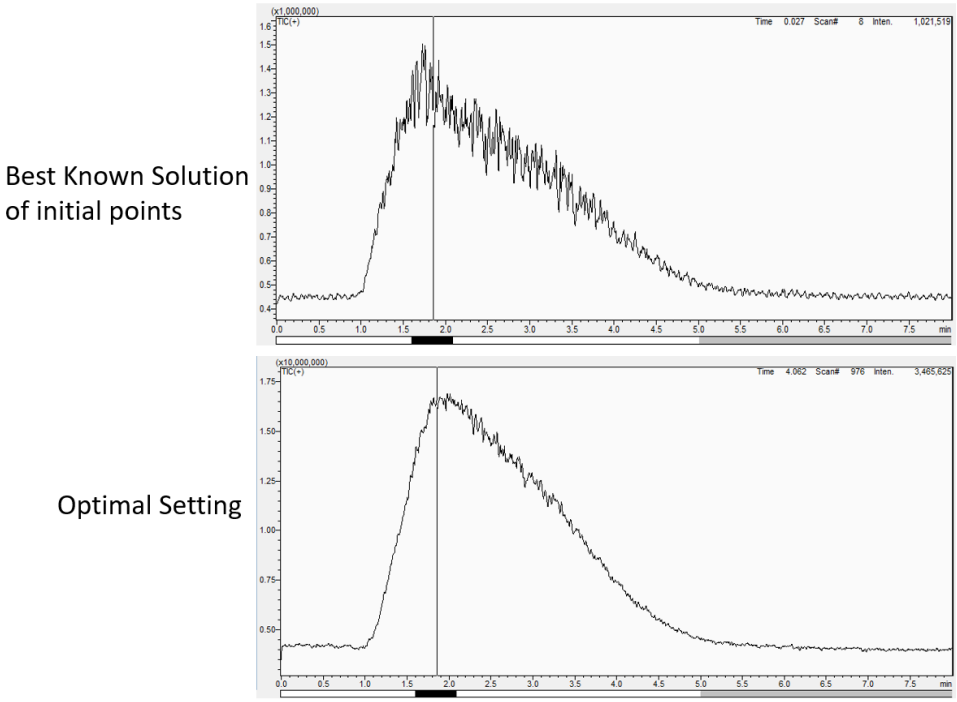

Figure 5.11: *Comparison of tuning file and optimal setting chromatograms*

Parameter	**Optimal Setting**
Interface Voltage (V)	4.8
DL Voltage (V)	83
Qarray DC Voltage (V)	16
DL Temperature $°C$	187
Heat Block Temperature $°C$	497
Nebulizing Gas Flow (mL/min)	1
Drying Gas Flow (mL/min)	17
Ion Intensity	5,932,508

Table 5.5: Optimal parameter settings for Acetaminophen

It is also worth mentioning the study by Raji and Schug [129] that examined the effect of instrumental parameters on ESI-MS analyte response using a full factorial design. The study conducted experiments on two different instruments, LCMS-2010 and Thermo

LCQ Deca XP, to identify the effect of four major parameters - spray voltage, ion transfer capillary temperature, ion transfer capillary voltage, and tube lens voltage on the analyte response - ion intensity. The LCMS-2010 is an older but very similar version of the LCMS-2020 instrument on which our algorithm was tested. It is important to note that spray voltage, ion transfer capillary temperature, ion transfer capillary voltage, and tube lens voltage in LCMS-2010 are the same as the Interface Voltage, DL Temperature, DL Voltage and Qarray DC Voltage in the LCMS-2020 instrument respectively. The study uses a full factorial design with just 2 levels - the minimum and maximum ranges of the parameters leading to a total of 16 experimental settings to identify any significant interactions.

If the study were to apply the full factorial design even with just the 2 levels, it would have increased the number of experimental settings to 128. It would still not have been possible to identify the optimal solution because of the categorical levels. However, with our methodology, we were able handle using continuous settings, as well as identify an optimal solution close to the tuning file in just 28 experimental runs, This is 78% less than a traditional RSM method using factorial designs. This illustrates that our methodology is capable of reducing the sample preparation times, the machine run times and the overall cost involved in running each experiment.

Chapter 6

Conclusion

In this research, in Chapter 3 we developed a MIQCP formulation for the QMARS model with the capability to handle the linear, cubic, quadratic and quintic terms in the QMARS basis functions. The QMARS-MIQCP formulation was incorporated within a global optimization algorithm called QMARS-MIQCP-OPT to globally optimize several test functions. The performance of the QMARS-MIQCP-OPT algorithm was compared with TITL-MARS-OPT. It was found that QMARS-MIQCP-OPT outperforms TITL-MARS-OPT for all the 4 test functions when minimizing and outperforms TITL-MARS-OPT for 2 out of the 4 test functions when maximizing.

In Chapter 4, the QMARS-MIQCP formulation was incorporated within a surrogate optimization algorithm called QMARS-MIQCP-SUROPT. A modified EEPA algorithm for candidate search and new scheme for candidate selection were also developed. First, the different QMARS-MIQCP-SUROPT algorithms - QMARS-MIQCP-SUROPT with sorted EEPA and fixed γ, QMARS-MIQCP-SUROPT with EEPA and fixed γ, QMARS-MIQCP-SUROPT with sorted EEPA and cyclical γ, and QMARS-MIQCP-SUROPT with EEPA and cyclical γ - for three different standard global optimization test functions - Rosenbrock, Rastrigin and Ackley - in uncertain environment. It was found that for all three test functions, QMARS-MIQCP-SUROPT with sorted EEPA and fixed γ outperformed the other algorithms in its ability to find better optimal solutions and converge faster. An additional

study was conducted to compare QMARS-MIQCP-SUROPT with sorted EEPA and fixed γ to its competitor TK-MARS for the same three global optimization test functions. It was seen that the QMARS-MIQCP-SUROPT approach clearly outperformed TK-MARS for all three test functions.

In Chapter 5, QMARS-MIQCP-SUROPT with sorted EEPA and fixed γ algorithm was tested on both a real-world laboratory experiment as well as a simulation. First, the algorithm was tested on a 2D-LC simulation software, ChromSim to find the optimal system parameter settings for *plates*, $V_m(ml)$, Φ_{dip}, $\Phi_{initial}$, V_{loop} and $\%Fill$ to efficiently extract five different analytes by minimizing $Width_{50}$. This helped gauge the performance of the algorithm in analytical chemistry applications and the setup that would be needed for an actual laboratory run. Additionally, the ability of the QMARS surrogate model to automatically detect and model interactions was also tested in the simulation. However, it was found that switching the interaction and robust setting automatically after $2d$ parameter settings evaluations, performs similarly to setting the setting with no interactions.

For the second real world study, the methodology was tested on the Shimadzu LCMS-2020 instrument to find the optimal system parameter settings for Interface Voltage (V), DL Voltage (V), Qarray DC Voltage (V), DL Temperature $°C$, Heat Block Temperature $°C$, Nebulizing Gas Flow (mL/min) and Drying Gas Flow (mL/min) to efficiently extract Acetaminophen from the sample by maximizing the ion intensity of the analyte. The optimal parameter settings were found in just 28 laboratory experiments achieving a maximum ion intensity of around 5.9 million. It was also seen the chromatograms for the optimal setting and the tuning file were very similar to each other indicating that the optimal solution found was reasonable. Setting the parameter $interactions = 1$ and switching the setting to $interactions = 2$ with $robust = 1$ after $2d$ parameter settings automatically was also tested. It was found that there was no difference in the candidate points chosen in consecutive iterations or in the performance of the algorithm.

In a nutshell, QMARS-MIQCP-SUROPT with Sorted EEPA and Fixed γ performed better than the other algorithms that were studied. It was also seen that it performed

well when tested on ChromSim and in actual laboratory experiments in the Shimadzu LCMS-2020 instrument.

Chapter 7

Future Work

One interesting future direction of this research is exploring the automatic switching of the interaction setting from no interactions to two-way interactions after $2d$ function evaluations. While the chemistry case studies in this dissertation did not demonstrate the need for modeling interactions, it is anticipated that interactions will be important in other applications, such as supercritical fluid extraction. In addition, it is observed that the popular surrogate/global optimization test functions (e.g., Rastrigin, Rosenbrock, Ackley) do not involve strong interactions, so they are not appropriate for testing the ability of QMARS-MIQCP-SUROPT to handle interactions. Future work will need to identify appropriate test functions with strong interaction term and further study the interactions setting in QMARS.

It was seen that QMARS-MIQCP-SUROPT with sorted EEPA and fixed γ performed better in comparison to all other algorithms. In the research the value for γ was arbitrarily fixed at 0.8. Testing the performance of the algorithm at different fixed γ values would be insightful.

Another development in the research is creating an ensemble approach of the four different QMARS-MIQCP-SUROPT algorithms that were identified and the interactions settings for QMARS. Specifically for real-world experiments, since each function evaluation is expensive and time-consuming, an ensemble approach that can take advantage of the batch runs in the laboratory or other experiments and suggest optimal parameter settings would

be useful.

It was also noted that when the number of basis functions in the QMARS model increases or when the number of dimensions increases, the MIQCP algorithm can be slow. This is a disadvantage against typical surrogate optimization algorithms that do not conduct a formal optimization. Because surrogate optimization is intended for time-consuming experiments (computer or laboratory), the computational speed of the algorithm is not considered a limitation. However, exploring an alternate MIQCP approach to speed up the algorithm would still be worthwhile for improving the proposed methodology.

Finally, the algorithm was applied only to single objective optimization problems. Many real world experiments have multiple objectives. For example, when optimizing parameter settings in applications like Supercritical Fluid Extraction-Supercritical Fluid Chromatography(SFE-SFC) systems, there is need for optimizing multiple objectives such as peak area, peak width and chromatographic efficiency simultaneously. Modifying the proposed approach to optimize multi-objective optimization problems is worth for future study.

Bibliography

[1] National Academies Press, *A Research Agenda for Transforming Separation Science*. National Academies Press, 2019.

[2] R. G. Regis and C. A. Shoemaker., "A Stochastic Radial Basis Function Method for the Global Optimization of Expensive Functions," *INFORMS Journal on Computing*, vol. 19, no. 4, pp. 497–509, 2007.

[3] J. Muller and R. Piche., "Mixture Surrogate Models Based on Dempster-Shafer Theory for Global Optimization Problems," *Journal of Global Optimization*, vol. 51, pp. 79–104, 2011.

[4] J. Muller and C. A. Shoemaker., "Influence of Ensemble Surrogate Models and Sampling Strategy on the Solution Quality of Algorithms for Computationally Expensive Black-box Global Optimization Problems," *Journal of Global Optimization*, vol. 60, pp. 123–144, 2014.

[5] R. G. Regis, "An Initialization Strategy for High-Dimensional Surrogate-Based Expensive Black-Box Optimization BT - Modeling and Optimization: Theory and Applications," in *Springer Proceedings in Mathematics & Statistics* (L. F. Zuluaga and T. Terlaky, eds.), (New York, NY), pp. 51–85, Springer New York, 2013.

[6] K. Crombecq, I. Couckuyt, D. Gorissen, and T. Dhaene, "Space-Filling Sequential Design Strategies for Adaptive Surrogate Modelling," *Proceedings of the First Inter-*

national Conference on Soft Computing Technology in Civil, Structural and Environmental Engineering, vol. 92, pp. 1–20, 2009.

[7] K. Crombecq, D. Gorrissen, D. Deschrivjver, and T. Dhaene, "A Novel Hybrid Sequential Design Strategy for Global Surrogate Modeling of Computer Experiments," *SIAM Journal on Scientific Computing*, vol. 33, pp. 1948–1974, 2011.

[8] R. G. Regis and A. Shoemaker., "Constrained Global Optimization of Expensive Black Box Functions using Radial Basis Functions," *Journal of Global Optimization*, vol. 31, pp. 153–171, 2005.

[9] H. Anahideh, J. Rosenberger, and V. Chen, "High-dimensional Black-box Optimization Under," *Computers and Operations Research*, 2021.

[10] V. C. Chen, "Application of orthogonal arrays and MARS to inventory forecasting stochastic dynamic programs," *Computational Statistics and Data Analysis*, vol. 30, no. 3, pp. 317–341, 1999.

[11] J. F. Dickson, *An Exploration and Exploitation Pareto Approach to Surrogate Optimization*. PhD thesis, The University of Texas at Arlington, Arlington, TX, 2014.

[12] Merriam-Webster. (n.d.), "Optimization," 2020.

[13] R. A. Fisher, "The Arrangement of Field Experiments," *Journal of Ministry of Agriculture of Great britain*, pp. 503–515, 1926.

[14] J. Sacks, W. J. Welch, T. J. Mitchell, and H. P. Wynn, "Design and Analysis of Computer Experiments," *Statistical Science*, vol. 4, pp. 409–423, 1989.

[15] V. C. P. Chen, D. Gunther, and E. L. Johnson, "Solving for an Optimal Airline Yield Management Policy via Statistical Learning," *Journal of the Royal Statistical Society, Series C*, vol. 52 Part 1, pp. 1–12, 2003.

[16] V. C. P. Chen, K. L. Tsui, R. R. Barton, and M. Meckesheimer, "A review on design, modeling and applications of computer experiments," *IIE Transactions (Institute of Industrial Engineers)*, vol. 38, no. 4, pp. 273–291, 2006.

[17] Z. Yang, V. C. P. Chen, M. E. Chang, M. L. Sattler, and A. Wen, "A decision-making framework for ozone pollution control," *Operations Research*, vol. 57, no. 2, pp. 484–498, 2009.

[18] M. Sayadi., *A Design and Analysis of Computer Experiments Approach for Green Building*. PhD thesis, The University of Texas at Arlington, 2016.

[19] H. F. Dodge and H. G. Romig, "A Method of Sampling Inspection," *Bell System Technical Journal*, vol. 8, pp. 613–631, oct 1929.

[20] P. C. Mahalanobis, "A Sample Survey of the Acreage under Jute in Bengal," *Sankhyā: The Indian Journal of Statistics (1933-1960)*, vol. 4, pp. 511–530, nov 1940.

[21] W. Abraham, "Sequential Analysis," *Social Forces*, vol. 27, pp. 170–171, dec 1943.

[22] R. Herbert, "Some Aspects of the Sequential Design of Experiments," *Bulletin of the American Mathematical Society*, vol. 58, no. 5, pp. 527–535, 1952.

[23] H. Chernoff, "Sequential Design of Experiments," *The Annals of Mathematical Statistics*, vol. 30, no. 3, pp. 755–770, 1959.

[24] H. Chernoff, "Approaches in Sequential Design of Experiments," Tech. Rep. January, Defense Technical Information Center, 1973.

[25] A. E. Albert, "The Sequential Design of Experiments for Infinitely Many States of Nature," *The Annals of Mathematical Statistics*, vol. 32, no. 3, pp. 774–799, 1961.

[26] N. Salkind, "Encyclopedia of Research Design," 2010.

[27] G. E. P. Box and K. B. Wilson, "On the Experimental Attainment of Optimum Conditions," *Journal of the Royal Statistical Society*, vol. 13, no. 1, pp. 1–45, 1951.

[28] G. Box and N. Draper, *Empirical Model-Building and Response Surfaces*. Scientific Research, 1987.

[29] R. H. Myers, D. C. Montgomery, G. Geoffrey Vining, C. M. Borror, and S. M. Kowalski, "Response Surface Methodology: A Retrospective and Literature Survey," *Journal of Quality Technology*, vol. 36, no. 1, pp. 53–78, 2004.

[30] Nuran, *The response surface methodology*. PhD thesis, Indiana University, 2007.

[31] A. I. Khuri and S. Mukhopadhyay, "Response surface methodology," *WIREs Computational Statistics*, vol. 2, no. 2, pp. 128–149, 2010.

[32] Cornell JA, "How to Apply Response Surface Methodology. The ASQC Basic References in Quality Control: Statistical Techniques," *ASQC*, vol. 8, 1990.

[33] R. H. Myers and D. C. Montgomery, "Response Surface Methodology," *IIE Transactions*, vol. 28, pp. 1031–1032, dec 1996.

[34] R. H. Myers, "Response Surface Methodology—Current Status and Future Directions," *Journal of Quality Technology*, vol. 31, pp. 30–44, jan 1999.

[35] R. Unal, R. Lepsch, W. Engelund, and D. Stanley, "Approximation model building and multidisciplinary design optimization using response surface methods," *American Institute of Aeronautics and Astronautics*, pp. 592–598, 2013.

[36] A. Morshedi and M. Akbarian, "Application Of Response Surface Methodology: Design Of Experiments And Optimization: A Mini Review," *Indian Journal of Fundamental and Applied Life Sciences*, vol. 4, no. 2002, pp. 2231–6345, 2014.

[37] S. A. Weissman and N. G. Anderson, "Design of Experiments (DoE) and Process Optimization. A Review of Recent Publications," *Organic Process Research and Development*, vol. 19, no. 11, pp. 1605–1633, 2015.

[38] Y. Liang, J. Liu, Q. Zhong, L. Shen, J. Yao, T. Huang, and T. Zhou, "Determination of major aromatic constituents in vanilla using an on-line supercritical fluid extraction coupled with supercritical fluid chromatography," *Journal of Separation Science*, vol. 41, pp. 1600–1609, apr 2018.

[39] V. K. Ky, C. D'Ambrosio, Y. Hamadi, and L. Liberti, "Surrogate-based methods for black-box optimization," *International Transactions in Operational Research*, vol. 13, no. 7 SUPPL., pp. 127–128, 2016.

[40] S. Koziel and L. Leifsson, *Surrogate-Based Modeling and Optimization: Applications in Engineering*. Springer Nature, aug 2013.

[41] P. I. Frazier, "A Tutorial on Bayesian Optimization," *Cornell University*, no. Section 5, pp. 1–22, 2018.

[42] L. Yang and A. Shami, "On hyperparameter optimization of machine learning algorithms: Theory and practice," *Neurocomputing*, vol. 415, pp. 295–316, 2020.

[43] T. Yu and H. Zhu, "Hyper-Parameter Optimization: A Review of Algorithms and Applications," *Cornell University*, pp. 1–56, 2020.

[44] A. I. Forrester, A. J. Keane, and A. J. Kanel., "Recent advances in surrogate-based optimization," *Progress in Aerospace Sciences*, vol. 45, no. 1-3, pp. 50–79, 2009.

[45] J. F. Dickson, J. M. Rosenberger, V. C. P. Chen, P. Kung, and A. Ronbinson, *Optimization to Select Energy Efficient Building Options*. PhD thesis, The University of Texas at Arlington, 2013.

[46] S. Wang, Y. Zhang, C. Wu, W. Darvas, and W. Chaovalitwongse, "Online Prediction of Driver Distraction Based on Brain Activity Patterns," *IEEE Transactions on Intelligent Transportation Systems*, vol. 16, pp. 136–150, 2014.

[47] R. N. Kacker, E. S. Lagergren, and J. J. Filliben, "Taguchi's Orthogonal Arrays Are Classical Designs of Experiments," *Journal of research of the National Institute of Standards and Technology*, vol. 96, no. 5, pp. 577–591, 1991.

[48] V. C. P. Chen, "Measuring the goodness of orthogonal array discretizations for stochastic programming and stochastic dynamic programming," *SIAM Journal on Optimization*, vol. 12, no. 2, pp. 322–344, 2002.

[49] N. Butler, "Optimal and Orthogonal Latin Hypercube Designs for Computer Experiments," *Biometrika*, vol. 88, no. 3, pp. 847–857, 2001.

[50] T. M. Cioppa and T. W. Lucas., "Efficient Nearly Orthogonal and Space-filling Latin Hypercubes," *Technometrics*, vol. 49, pp. 45–55, 2007.

[51] I. M. Sobol, "Uniformly distributed sequences with an additional uniform property," *USSR Computational Mathematics and Mathematical Physics*, vol. 16, no. 5, pp. 236–242, 1976.

[52] I. M. Sobol, "The Distribution of Points in a Cube and the Approximate Evaluation of Integrals," *U. S. S. R. Computational Mathematics and Mathematical Physics*, vol. 7, pp. 86–112, 1967.

[53] S. Hosder, L. Watson, B. Grossman, W. Mason, H. Kim, R. Haftka, and S. Cox, "Polynomial Response Surface Approximations for the Multidisciplinary Design Optimization of a High Speed Civil Transport," *Optimization and Engineering*, vol. 2, no. 4, pp. 431–452, 2001.

[54] M. Kaufman, V. Balabanov, A. A. Giunta, B. Grossman, W. H. Mason, S. L. Burgee, R. T. Haftka, and L. T. Watson, "Variable-complexity response surface approximations for wing structural weight in HSCT design," *Computational Mechanics*, vol. 18, no. 2, pp. 112–126, 1996.

[55] Y. Lian and M. S. Liou, "Multiobjective optimization using coupled response surface model and evolutionary algorithm," *AIAA Journal*, vol. 43, no. 6, pp. 1316–1325, 2005.

[56] T. Goel, R. Vaidyanathan, R. T. Haftka, W. Shyy, N. V. Queipo, and K. Tucker, "Response surface approximation of Pareto optimal front in multi-objective optimization," *Computer Methods in Applied Mechanics and Engineering*, vol. 196, no. 4-6, pp. 879–893, 2007.

[57] G. G. Wang, Z. Dong, and P. Aitchison, "Adaptive response surface method - A global optimization scheme for approximation-based design problems," *Engineering Optimization*, vol. 33, no. 6, pp. 707–733, 2001.

[58] T. W. Simpson, D. J. Cappelleri, B. Wilson, and M. Frecker, "Efficient Pareto Frontier Exploration using Surrogate Approximations," *Optimization and Engineering*, no. 2, pp. 31–50, 2001.

[59] T. Malvić, J. Ivšinović, J. Velić, and R. Rajić, "Kriging with a small number of data points supported by Jack-Knifing, a case study in the Sava depression (Northern Croatia)," *Geosciences (Switzerland)*, vol. 9, no. 1, 2019.

[60] M. Emmerich, A. Giotis, A. Ozdemir, T. Back, and K. Giannakoglou., "Meta-model Assisted Evolution Strategies," in *Parallel Problem Solving from Nature*, pp. 362–370, Springer-Verlag, 2002.

[61] E. Tresidder, Y. Zhang, and A. I. J. Forrester, "Acceleration of building design optimisation through the use of Kriging surrogate models," *BSO12 - Building Simulation and Optimization Conference*, no. September, pp. 1–8, 2012.

[62] B. Gaspar, A. P. Teixeira, and C. G. Soares, "Assessment of the efficiency of Kriging surrogate models for structural reliability analysis," *Probabilistic Engineering Mechanics*, vol. 37, pp. 24–34, 2014.

[63] R. G. Regis, "Stochastic radial basis function algorithms for large-scale optimization involving expensive black-box objective and constraint functions," *Computers & Operations Research*, vol. 38, no. 5, pp. 837–853, 2011.

[64] R. G. Regis, "Constrained optimization by radial basis function interpolation for high-dimensional expensive black-box problems with infeasible initial points," *Engineering Optimization*, vol. 46, pp. 218–243, feb 2014.

[65] H.-M. Gutmann, "A Radial Basis Function Method for Global Optimization," *Journal of Global Optimization*, vol. 19, pp. 201–227, 2001.

[66] J. M. Parr, A. J. Keane, A. I. J. Forrester, and C. M. E. Holden, "Infill sampling criteria for surrogate-based optimization with constraint handling," *Engineering Optimization*, vol. 44, pp. 1147–1166, oct 2012.

[67] S. Kitayama, J. Srirat, M. Arakawa, and K. Yamazaki, "Sequential approximate multi-objective optimization using radial basis function network," *Structural and Multidisciplinary Optimization*, vol. 48, no. 3, pp. 501–515, 2013.

[68] R. Datta and R. G. Regis., "A Surrogate-assisted Evolution Strategy for Constrained Multi-Objective Optimization," *Expert Systems with Applications*, vol. 57, pp. 270–284, 2016.

[69] R. G. Regis, "Multi-objective constrained black-box optimization using radial basis function surrogates," *Journal of Computational Science*, vol. 16, pp. 140–155, 2016.

[70] R. G. Regis and S. M. Wild, "CONORBIT: constrained optimization by radial basis function interpolation in trust regions," *Optimization Methods and Software*, vol. 32, pp. 552–580, may 2017.

[71] V. N. Vapnik, "The nature of statistical learning theory ," 1995.

[72] K. Miller and A. J. Smola, "Using SVM for Time Series Prediction," *International Conference on Artificial Neural Networks*, pp. 1–12, 2000.

[73] A. J. SMOLA and B. SCHOLKOPF, "A tutorial on support vector regression," *Statistics and Computing*, vol. 14, pp. 199–222, 2004.

[74] P. Zhu, F. Pan, W. Chen, and S. Zhang, "Use of support vector regression in structural optimization: Application to vehicle crashworthiness design," *Mathematics and Computers in Simulation*, vol. 86, no. 800, pp. 21–31, 2012.

[75] H. Xiang, Y. Li, H. Liao, and C. Li, "An adaptive surrogate model based on support vector regression and its application to the optimization of railway wind barriers," *Structural and Multidisciplinary Optimization*, vol. 55, no. 2, pp. 701–713, 2017.

[76] J. H. Friedman, "Multivariate adaptive regression splines," *The Annals of Statistics*, pp. 1–67, 1991.

[77] S. Crino and D. E. Brown, "Global optimization with multivariate adaptive regression splines," *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 37, no. 2, pp. 333–340, 2007.

[78] M. Costas, J. Diaz, L. Romera, and S. Hernandez, "A Multi-objective Surrogate-based Optimization of the Crashworthiness of a Hybrid Impact Absorber," *International Journal of Mechanical Sciences*, vol. 88, pp. 46–54, 2014.

[79] V. L. Pilla, J. M. Rosenberger, V. C. P. Chen, N. Engsuwan, and S. Siddappa, "A Multivariate Adaptive Regression Splines Cutting Plane Approach for Solving a Two-Stage Stochastic Programming Fleet Assignment Model," *European Journal of Operational Research*, vol. 216, pp. 162–171, 2012.

[80] S. Siddappa, D. Gunther, J. M. Rosenberger, and V. C. P. Chen, "Refined Experimental Design and Regression Splines Method for Network Revenue Management," *Journal of Pricing and Revenue Management*, vol. 6, no. 3, pp. 188–199, 2007.

[81] D. L. Martinez, *Variants of Multivariate Adaptive Regression Splines (MARS): Convex vs. Non-convex, Piecewise-linear vs. Smooth and Sequential Algorithms*. PhD thesis, The University of Texas at Arlington, 2013.

[82] N. M. Martinez, H. Anahideh, J. M. Rosenberger, D. L. Martinez, V. C. P. Chen, and B. P. Wang., "Global Optimization of Non-convex Piecewise Linear Regression Splines," *Journal of Global Optimization*, vol. 68, pp. 563–586, 2017.

[83] J. Xinglong, *Multivariate Adaptive Regression Splines Knot Optimization and Global Optimization*. PhD thesis, The University of Texas at Arlington, 2019.

[84] N. Sakhavand, *New Algorithms for Stochastic Power Systems Planning and Operations Problems*. PhD thesis, The University of Texas at Arlington, 2020.

[85] X. Song, L. Lv, J. Li, W. Sun, and J. Zhang, "An Advanced and Robust Ensemble Surrogate Model: Extended Adaptive Hybrid Functions," *Journal of Mechanical Design*, vol. 140, feb 2018.

[86] J. Zhang, X. Yue, J. Qiu, M. Zhang, and X. Wang, "A unified ensemble of surrogates with global and local measures for global metamodelling," *Engineering Optimization*, pp. 1–22, mar 2020.

[87] T. Goel, R. T. Haftka, W. Shyy, and N. V. Queipo, "Ensemble of surrogates," *Structural and Multidisciplinary Optimization*, vol. 33, no. 3, pp. 199–216, 2007.

[88] Q. Ouyang, W. Lu, T. Miao, W. Deng, C. Jiang, and J. Luo, "Application of ensemble surrogates and adaptive sequential sampling to optimal groundwater remediation design at DNAPLs-contaminated sites," *Journal of Contaminant Hydrology*, vol. 207, no. November, pp. 31–38, 2017.

[89] D. R. Jones, M. Schonlau, and W. J. Welch, "Efficient Global Optimization of Expensive Black-Box Functions," *Journal of Global Optimization*, vol. 13, no. 4, pp. 455–492, 1998.

[90] B. J. Williams, T. J. Santner, and W. I. Notz, "Sequential design of computer experiments to minimize integrated response functions," *Statistica Sinica*, vol. 10, pp. 1133–1152, oct 2000.

[91] K. Crombecq, E. Laermans, and T. Dhaene, "Efficient space-filling and non-collapsing sequential design strategies for simulation-based modeling," *European Journal of Operational Research*, vol. 214, no. 3, pp. 683–696, 2011.

[92] N. C. Xiao, M. J. Zuo, and C. Zhou, "A new adaptive sequential sampling method to construct surrogate models for efficient reliability analysis," *Reliability Engineering and System Safety*, vol. 169, no. 2006, pp. 330–338, 2018.

[93] D. R. Jones, M. Schonlau, and W. J. Welch., "Efficient Global Optimization of Expensive Black-Box Functions," *Journal of Global Optimization*, vol. 13, no. 4, pp. 455–492, 1998.

[94] J. Knowles, "ParEGO: A hybrid algorithm with on-line landscape approximation for expensive multiobjective optimization problems," *IEEE Transactions on Evolutionary Computation*, vol. 10, no. 1, pp. 50–66, 2006.

[95] N. Nezami and H. Anahideh, "An Empirical Review of Model-based Adaptive Sampling for Global Optimization of Expensive Black-box Functions." 2022.

[96] R. G. Regis and C. A. Shoemaker, "Combining radial basis function surrogates and dynamic coordinate search in high-dimensional expensive black-box optimization," *Engineering Optimization*, vol. 45, no. 5, pp. 529–555, 2013.

[97] T. Krityakierne, T. Akhtar, and C. A. Shoemaker., "SOP: Parallel Surrogate Global Optimization with Pareto Center Selection for Computationally Expensive Single Objective Problems," *Journal of Global Optimization*, vol. 66, pp. 417–437, 2016.

[98] D. Eriksson, D. Bindel, and C. A. Shoemaker, "pySOT and POAP: An event-driven asynchronous framework for surrogate optimization," *Cornell University*, no. 1, pp. 1–19, 2019.

[99] T. Akhtar and C. A. Shoemaker, "Combining local surrogates and adaptive restarts for global optimization of moderately expensive functions," *AIP Conference Proceedings*, vol. 2070, no. February, 2019.

[100] D. Zhan and H. Xing, "Expected improvement for expensive optimization: a review," *Journal of Global Optimization*, vol. 78, no. 3, pp. 507–544, 2020.

[101] J. A. Caballero and I. E. Grossmann, "An Algorithm for the Use of Surrogate Models in Modular Flowsheet Optimization," *AIChE Journal*, vol. 54, no. 10, 2008.

[102] A. Chaudhuri and R. T. Haftka, "A Stopping Criterion for Surrogate Based Optimization using EGO," in *10th World Congress on Structural and Multidisciplinary Optimization*, (Orlando,), 2013.

[103] X. Wan, J. F. Pekny, and G. V. Reklaitis, "Simulation-based optimization with surrogate models - Application to supply chain management," *Computers and Chemical Engineering*, vol. 29, no. 6 SPEC. ISS., pp. 1317–1328, 2005.

[104] V. Ivan and K. Andy, "Multi-objective Optimization Using Surrogates," in *Computational Intelligence in Optimization*, ch. 7, pp. 155–175, Computational Intelligence in Optimization, 2010.

[105] S. D. Manjare and K. Dhingra, "Supercritical fluids in separation and purification: A review," *Materials Science for Energy Technologies*, vol. 2, no. 3, pp. 463–484, 2019.

[106] M. Zoccali, D. Giuffrida, P. Dugo, and L. Mondello, "Direct online extraction and determination by supercritical fluid extraction with chromatography and mass spectrometry of targeted carotenoids from red Habanero peppers (Capsicum chinense Jacq.)," *Journal of Separation Science*, vol. 40, pp. 3905–3913, oct 2017.

[107] D. Giuffrida, M. Zoccali, A. Arigò, F. Cacciola, C. O. Roa, P. Dugo, and L. Mondello, "Comparison of different analytical techniques for the analysis of carotenoids in tamarillo (Solanum betaceum Cav.)," *Archives of Biochemistry and Biophysics*, vol. 646, pp. 161–167, may 2018.

[108] A. P. Wicker, D. D. Carlton, K. Tanaka, M. Nishimura, V. Chen, T. Ogura, W. Hedgepeth, and K. A. Schug, "On-line supercritical fluid extraction—supercritical fluid chromatography-mass spectrometry of polycyclic aromatic hydrocarbons in soil," *Journal of Chromatography B*, vol. 1086, pp. 82–88, jun 2018.

[109] M. Sakai, Y. Hayakawa, Y. Funada, T. Ando, E. Fukusaki, and T. Bamba, "Development of a practical online supercritical fluid extraction–supercritical fluid chromatography/mass spectrometry system with an integrated split-flow method," *Journal of Chromatography A*, vol. 1592, pp. 161–172, may 2019.

[110] M. Zoccali, P. Donato, and L. Mondello, "Recent advances in the coupling of carbon dioxide-based extraction and separation techniques," jul 2019.

[111] A. P. Wicker, K. Tanaka, M. Nishimura, V. Chen, T. Ogura, W. Hedgepeth, and K. A. Schug, "Multivariate approach to on-line supercritical fluid extraction – supercritical fluid chromatography - mass spectrometry method development," *Analytica Chimica Acta*, vol. 1127, pp. 282–294, 2020.

[112] L. Murray, R. L. Mason, R. F. Gunst, and J. L. Hess, *Statistical Design and Analysis of Experiments: With Applications to Engineering and Science.*, vol. 85. Journal of the American Statistical Association, 1990.

[113] M. A. Bezerra, R. E. Santelli, E. P. Oliveira, L. S. Villar, and L. A. Escaleira, "Response surface methodology (RSM) as a tool for optimization in analytical chemistry," *Talanta*, vol. 76, no. 5, pp. 965–977, 2008.

[114] A. Y. Aydar, "Utilization of Response Surface Methodology in Optimization of Extraction of Plant Materials," in *Statistical Approaches With Emphasis on Design of*

Experiments Applied to Chemical Processes, Statistical Approaches With Emphasis on Design of Experiments Applied to Chemical Processes, 2018.

[115] T. Belwal, P. Dhyani, I. D. Bhatt, R. S. Rawal, and V. Pande, "Optimization extraction conditions for improving phenolic content and antioxidant activity in Berberis asiatica fruits using response surface methodology (RSM)," *Food Chemistry*, vol. 207, pp. 115–124, 2016.

[116] Y. Zhou, J. Z. Song, F. F. K. Choi, H. F. Wu, C. F. Qiao, L. S. Ding, S. L. Gesang, and H. X. Xu, "An experimental design approach using response surface techniques to obtain optimal liquid chromatography and mass spectrometry conditions to determine the alkaloids in Meconopsi species," *Journal of Chromatography A*, vol. 1216, no. 42, pp. 7013–7023, 2009.

[117] E. S. Hecht, A. L. Oberg, and D. C. Muddiman, "Optimizing Mass Spectrometry Analyses: A Tailored Review on the Utility of Design of Experiments," *Journal of the American Society for Mass Spectrometry*, vol. 27, no. 5, pp. 767–785, 2016.

[118] R. G. Regis and C. A. Shoemaker., "A Constrained Multi-objective Surrogate-based Optimization Algorithm," *Engineering Optimization*, vol. 45, pp. 529–555, 2013.

[119] R. Put, Q. S. Xu, D. L. Massart, and Y. Vander Heyden, "Multivariate adaptive regression splines (MARS) in chromatographic quantitative structure–retention relationship studies," *Journal of Chromatography A*, vol. 1055, no. 1, pp. 11–19, 2004.

[120] A. Cozad, S. Nikolaos V., and D. C. Miller, "Learning Surrogate Models for Simulation-Based Optimization," *AIChE Journal*, vol. 60, no. 6, pp. 2211–2226, 2012.

[121] N. Quirante, J. Javaloyes, R. Ruiz-Femenia, and J. A. Caballero, "Optimization of Chemical Processes Using Surrogate Models Based on a Kriging Interpolation," *Computer Aided Chemical Engineering*, vol. 37, no. June, pp. 179–184, 2015.

[122] C. Nentwich and S. Engell, "Application of surrogate models for the optimization and design of chemical processes," *Proceedings of the International Joint Conference on Neural Networks*, vol. 2016-Octob, pp. 1291–1296, 2016.

[123] A. M. Schweidtmann, A. D. Clayton, N. Holmes, E. Bradford, R. A. Bourne, and A. A. Lapkin, "Machine learning meets continuous flow chemistry: Automated optimization towards the Pareto front of multiple objectives," *Chemical Engineering Journal*, vol. 352, no. April, pp. 277–282, 2018.

[124] S. Surjanovic and D. Bingham, "Virtual Library of Simulation Experiments: Test Functions and Data sets." http://www.sfu.ca/~ssurjano, mar 2022.

[125] L. N. Jeong, R. Sajulga, S. G. Forte, D. R. Stoll, and S. C. Rutan, "Simulation of elution profiles in liquid chromatography—I: Gradient elution conditions, and with mismatched injection and mobile phase solvents," *Journal of Chromatography A*, vol. 1457, pp. 41–49, 2016.

[126] D. R. Stoll, R. W. Sajulga, B. N. Voigt, E. J. Larson, L. N. Jeong, and S. C. Rutan, "Simulation of elution profiles in liquid chromatography - II: Investigation of injection volume overload under gradient elution conditions applied to second dimension separations in two-dimensional liquid chromatography.," *Journal of chromatography. A*, vol. 1523, pp. 162–172, nov 2017.

[127] Shimadzu, "Basic Instrumentation of a Mass Spectrometer." https://sisu.ut.ee/heritage-analysis/52-mass-spectrometry, apr 2022.

[128] "Schematic diagram of Electro Spray Ionization process." https://www.chem.pitt.edu/facilities/mass-spectrometry/mass-spectrometry-introduction, apr 2022.

[129] M. A. Raji and K. A. Schug, "Chemometric study of the influence of instrumental parameters on ESI-MS analyte response using full factorial design," *International Journal of Mass Spectrometry*, vol. 279, no. 2-3, pp. 100–106, 2009.