Data Driven Methods for Chemical Process and Product Synthesis and Design

By

Sarah Elizabeth Davis

A dissertation submitted to the Graduate Faculty of
Auburn University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

Auburn, Alabama
December 15, 2018

Keywords: Computer Aided Molecular Design, Principal Component Analysis,
Surrogate Modeling

Mario R. Eden, Chair, Joe T & Billie Carole McMillan Professor, Chemical Engineering
Selen Cremaschi, Co-Chair, B. Redd Associate Professor, Chemical Engineering
Allan E. David, John W. Brown Associate Professor, Chemical Engineering
Steve E. Taylor, Associate Dean for Research, Biosystems Engineering

Abstract

Data driven methods for chemical process and product synthesis have become integrated in all aspects of design. The responsibly of the academic community should be to provide users with guidance when managing the ever-increasing amount of data and possible data analytics methods with a goal of utilizing these new design tools to ensure that their applications provides meaningful results.

Progressive model improvement will lead us to improve characterization techniques to better describe molecules, more advanced modeling methods provide more correct results, and uncertainty management will ensure that the results are more accurate. The methods presented in this work illustrate applications of data driven methods for chemical process and product synthesis and design with a focus on two specific tools computer aided molecular design and surrogate modeling.

Computer Aided Molecular Design is a framework that allows us to utilize data to design molecules specific to a process. This is important because it eliminates the need to alter the design to match the available inputs, rather the inputs are modified to match the design. Once issue with this method is that it is reliant on characteristic data for each molecule or building block. The work presented in this dissertation allows us to generate necessary data to apply to the framework thus expanding the possible molecules that can be utilized even further than the computer aided molecular design framework alone.

Surrogate modeling allows us to understand complex or unknown processes to provided understanding of the process and improve designs. The work presented in this dissertation provides information about the application of those models based on the surface shape and number of inputs. Additionally, it provides information about sampling methods and sizing. Basically, this information can help make an informed decision when selecting which surrogate model, sampling method and group for each type of application.

Both advances provide added depth to data analysis by enhancing current methodologies. This type of work is important because as the modern chemical engineer begins to implement data driven design techniques, the applications that are utilized will need to become more robust and accurate.

Acknowledgements

I would like to begin my acknowledgements with Dr. Mario R Eden. Thank you for allowing me this amazing opportunity and for your continued support as I pursued my doctoral program in the Chemical Engineering Department of Auburn University. Dr. Selen Cremaschi, thank you for your guidance through this process. Thank you to Dr. Allen David and Dr. Steven Taylor for serving on my dissertation committee and for your review and feedback which helped to improve and complete my dissertation. Thank you to Dr. Ruel Overfelt for serving as my university reader. Thank you to Dr. Subin Hada and Dr. Robert Herring for your advice.

Thank you, most of all, to my loving husband, John, without whom this would not have been possible. Without your unwavering confidence and support, I would not have been able to succeed. To my kids who have understood the journey our whole family has taken. Thank you to the strong women who raised me and taught me that I can be anything that I was willing to work for. I love you all!

Table of Contents

List of Tables

List of Figures

CHAPTER 1. Introduction

1.1. Extracting Value from Data

Over the broad spectrum of data analytics applications, one implicit goal is understanding. More specifically when applied to chemical engineering processes, we seek to confirm what we know about a process or learn from something unusual. Process design can be improved because we can use existing data to optimize, improve or control processes. Further extensions include prediction of process behavior prior to implementation and monitoring once a process is in place.

1.1.1. Types of Data

In the 1950's when industrial manufacturing and chemical engineering really began to develop, data was collected from existing process or bench models. Due to the manual nature, each variable measured was expensive, so only the most vital were collected. Classical visualization tools generally included scatter plots, time-series plots, exponentially weighted moving average charts for process monitoring, and multiple linear regression least-squares models. To provide a graphical illustration, X will represent any data set where each row contains values from a variable where each row includes the measurements or observations at a point in time, various properties of a final product or raw material. Columns represent the values recorded for each observation or variables and the number measured will be described as K.

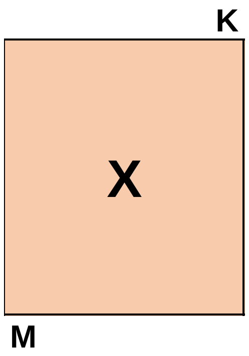


Figure 1. Basic Data Set, X, with M observations and K variables.

These data sets from the 1950's usually had more observations and fewer variables due to time and money constraints. As a result, the selection of variables was carefully considered, so variables were often independent with no or little correlated information and the variables were often measured in a controlled environment with a low amount of error. In this case, the financial limitations of the data created an enhancement. Present day engineers generally do not have the same limitations as we can now measure variables electronically, however some of the accuracy and independence of the data may be lost. Higher dimensional data sets are very common, so noise, correlated data and unnecessary data can create complicate the analysis.

1.1.2. Potential Issues with Data

As sample sizes have increased along with increased number of variables, data management becomes increasingly important. Additionally, larger data sets present a challenge for analytics because the data set must be treated in such a way as to learn from the relevant information and eliminate the irrelevant or incorrect data.

Independent variables are essential when analyzing a data set. In process modeling, many variables are dependent on one another, therefore lack of independence is a major concern. The balancing act between omitting valuable data and reducing the number of variables can have a major impact on the results.

Steady state, a goal of most engineering systems, provides challenges for data analysis. Data from such systems have very little signal and high noise as much of the data is from constant operations, noise, slow drift, or error. Finding the interesting signals in this routine data is the challenge (Denbig, 1951).

Data that was collected with error can skew the results of any experiment or observation. Assumptions are often made that the measuring equipment was calibrated, that the person taking the reading was correct, or many other variables that can cause error in data can lead to unintended or unobserved error. Missing data is very common in engineering applications due to any number of factors leading to a “missing” observation or data point.

1.2. Data Driven Methods in Chemical Engineering

Today’s data driven methods need to overcome the potential points where failure occurs. These methods should be able to extract relevant information and handle missing information from multidimensional data sets where the data may be stored in different locations while managing collinearity or measurement error in the recorded data. Latent property models are one example of this type of advanced data driven problem solving.

1.3. Data Driven Process Systems Engineering

The basis of process system engineering research is the development of computationally efficient prediction of the performance of a system of unit operations which are the essential stages of a process described by universal physical law such as fluid flow, heat and mass transfer, thermodynamic phase behavior, and reactions. Commonly, processes are a combination of process steps with many unit operations, resulting in highly complex systems. To understand those complex systems, simulation and prediction models have been and continue to be developed.

Over the last few decades a shift has taken place in the chemical engineering industry from gold standard commodity chemicals which were produced in bulk based on tried and true historical usage. The chemicals tended to have minimal variation among manufacturers and were molecules with simple architecture. Profits were proportional to increased production volume and improved process improvements. As the market demands more sophisticated chemicals which are ideal for a specific process, the industry has begun moving forward from traditional process optimization to product innovation. The obvious shift has taken place where expected in the biomedical arena, but in also more traditional chemical process plants. Traditionally, chemical engineering process and product design studies have been performed by physical experimentation based on hypotheses and tested in a laboratory to test the validity of a design as shown in Figure 2. This process has a significant shortcoming because the solution is limited to existing molecules with associated performance data.

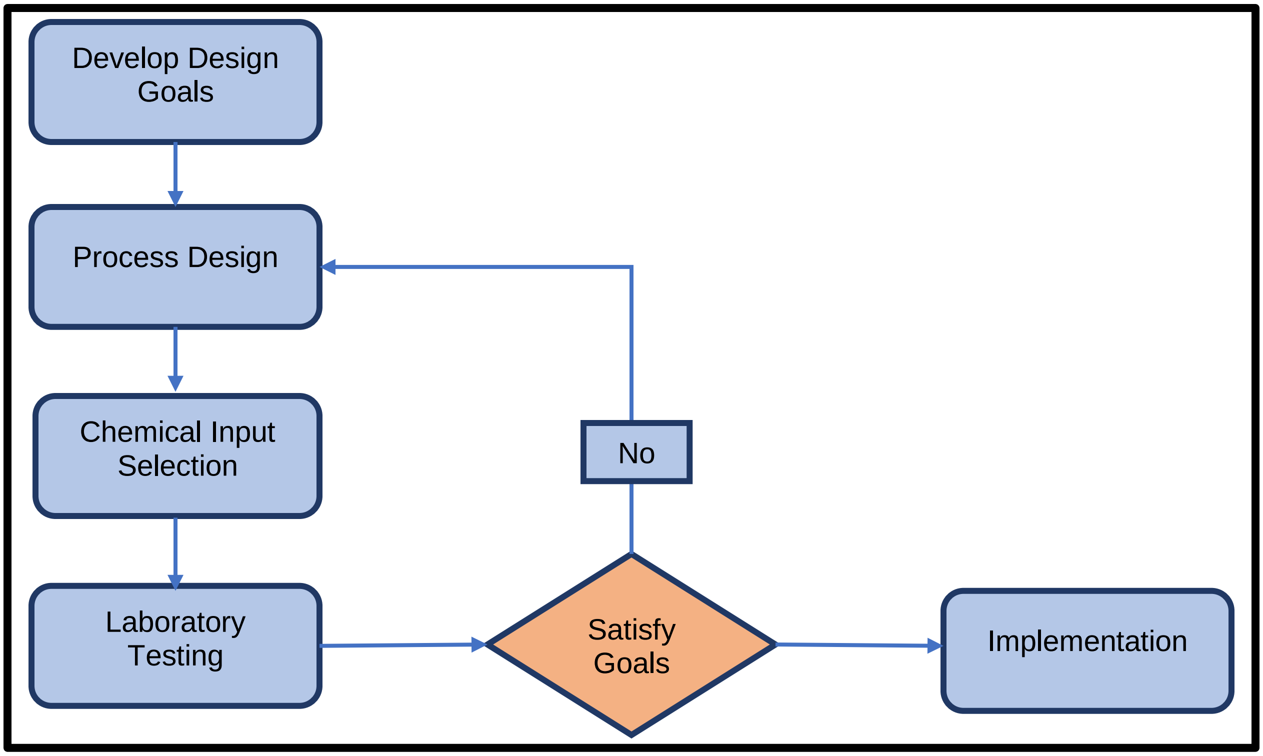


Figure 2. Traditional Chemical Engineering Process Design

Specialty chemicals moving to the forefront has broadened the audience of chemical engineers seeking to understand the vital relationship between the molecular architecture of a specialty chemical and its physical and chemical properties. Unlike traditional design of commodity chemicals, the design of specialty chemicals is not limited to raw material resources or defined molecular architectures. The mindset is reversed; the consumer is asked to define the attributes of the ideal chemical than selecting the closest match.

Advancement is occurring in the computational and data driven arena rather than the laboratory to explore both molecular design and process simulation. Over the last couple of decades, our ability to build models that simulate the behavior of complex phenomena has increased tremendously. Partially driven by powerful computing tools, we can now model phenomena at different spatial and temporal scales and combine them in multi-scale models to predict the performances more accurately of our product and process designs while managing computational burdens. Additionally, computer aided molecular design (CAMD) and the idea of reverse design allow the engineer to design the optimal chemical for the process which removes the traditional limitation of selecting chemicals that were readily available. These advancements have allowed us to improve our chemical engineering design through optimization of the processes and materials as shown in Figure 3.

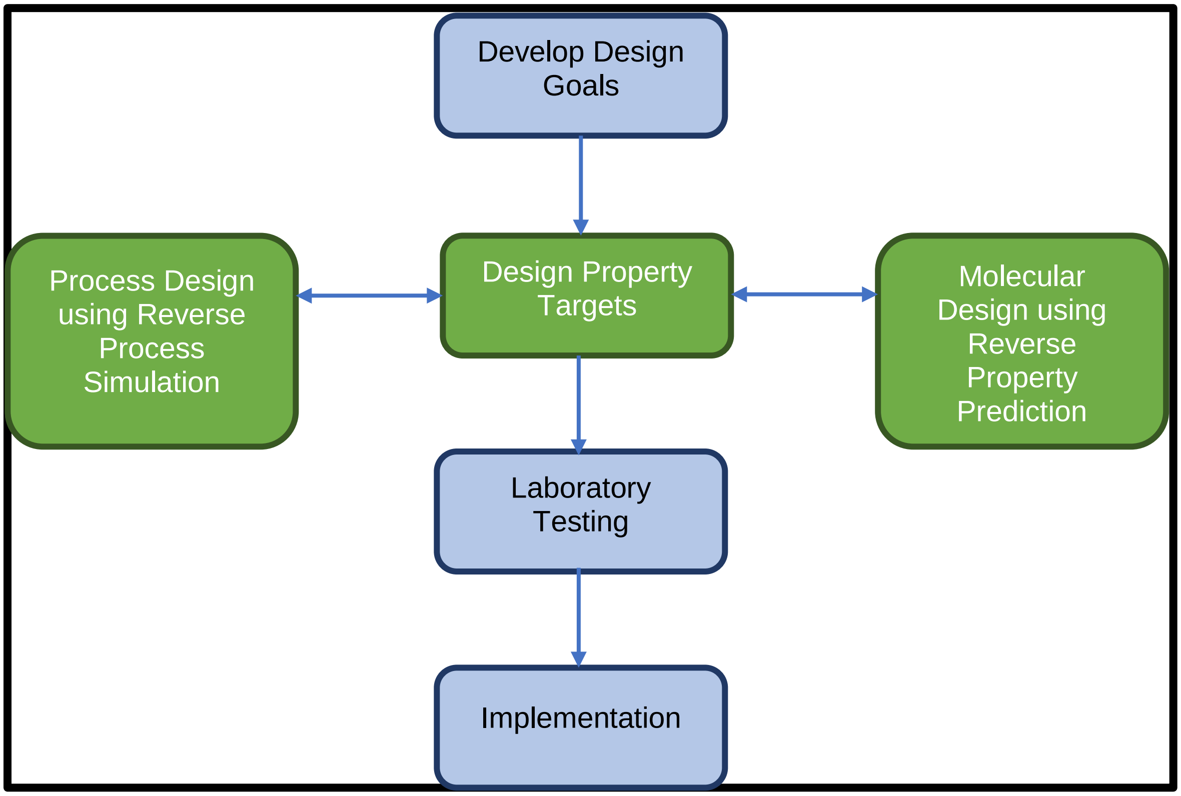

Develop Design Goals
Process Design using Reverse Process Simulation
Design Property Targets
Molecular Design using Reverse Property Prediction
Laboratory Testing
Implementation

Figure 3. Data Driven Chemical Engineering Design

1.3.1. Product Design

Molecular design can design the ideal molecule for a set of constraints. The tunable nature of ionic liquids and the estimation that 100 trillion possibilities may exist at room temperature lends itself well to molecular design (Turner et al, 2003).

Ionic liquids can provide environmentally benign solutions that can be tailored to specific process requirements (Wasserscheild & Welton, 2007). Due to the unique structure of ionic liquids, the cation, anion, and length of the alkyl chain can be varied to create a molecule with specific physical properties; however, experimental trial and error methods are impractical, so more advanced methods must be created determine the ideal combination for a specific set of process requirements.

A study was undertaken to demonstrate this method and sought to determine the ideal ionic liquid to capture CO_2. However, only a small percentage of the potential ionic liquids have been synthesized and tested for the solubility of CO_2. This innovative approach detaches the solution from further laboratory study and streamlines the search for the ideal candidate.

A characterization-based group contribution method is combined with density functional theory to determine the ionic liquid that can most effectively absorb CO_2. Infrared spectra data contains descriptor data that can be used to estimate properties of ionic liquids but does not exist for all ionic liquids. Density functional theory is used to create IR data based on a training set of experimental data. Principal component analysis and partial least square techniques are employed to reveal important features and patterns in the molecular architecture. A characterization-based group contribution method is used to estimate properties. The reverse design of potential ionic liquid molecules is completed by an exhaustive search of combinations with various cation, anions, and lengths of alkyl chains until a candidate molecule is found that provides the highest solubility of CO_2.

1.3.2. Process Modeling

One method that is often utilized to simulate a process is surrogate modeling which statistically relate input data to output data. Most often, these models are utilized when the relationship between input and output data is unknown, or when the relationship is highly complex and a simpler relationship with reasonable accuracy is desired. Many studies have developed and defined different surrogate model forms, but little work focuses on systematically comparing the abilities of these surrogate models to learn the response of the complicated models with different characteristics.

The goal of this study was to provide guidance when selecting surrogate model form. The performances of eight surrogate model forms, Artificial Neural Network, Automated Learning of Algebraic Models for Optimization, Extreme Learning Machines, Support Vector Regression, Radial Basis Function Networks, Gaussian Process Regression, Random Forests, Multivariate Adaptive Regression Splines, are compared via computational experiments. The training data was generated from thirty-five challenge functions using three different sampling methods, which are Latin Hypercube Sampling (LHS), Sobol Sequence and Halton Sequence. Six performance metrics, r-squared, r-squared adjusted, maximum absolute error (MaAE), root mean square error (RMSE), and Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC), were calculated for each challenge function and surrogate model combination. The results provide guidance for selecting the ideal surrogate model based on specifics of the problems including: surface shape of the challenge function, number of inputs of the challenge function, input sample generation method and input sample size.

1.4. Organization

Chapter 2 introduces the general theoretical background required for the development of the work presented in this dissertation. Section 2.1 describes the computer aided molecular design framework that was utilized to predict the ionic liquid properties including: reverse problem formulation, characterization-based group contribution technique using latent property parameters, density functional theory, chemometric techniques, and property clustering. Section 2.2 describes the surrogate models that were tested including: Artificial Neural Networks (ANN), Automated Learning of Algebraic Models for Optimization (ALAMO), Automated Learning of Algebraic Models for Optimization (ALAMO), Extreme Learning Machine (ELM), Support Vector Regression (SVR), Radial Basis Function Networks (RBF), Gaussian Process Regression (GPR), Random Forests (RF), Multivariate Adaptive Regression Splines (MARS). Chapter 3 provides a case study for the prediction of properties of ionic liquids with the intent to find the ideal ionic liquid to remove CO_2 from process flue gas. Chapter 4 provides a study of surrogate models with the intent to seek the surrogate model best suited to a situation. Chapter 5 presents the overall conclusions and details possible future work that would advance data driven chemical engineering process and product design.

CHAPTER 2. Product Design

Product design, the conversion of a conceptual idea into a tangible, manufactured object, enables manufactures to remain completive in a highly competitive and fluctuating market with global supply chains. Unlike overall process design, product design is based primarily on a set of consumer requirements or requests that are directly related to molecular architecture and the resulting physical-chemical properties. Work presented in this section was previously published in Computers and Chemical Engineering 34 (Davis, Hada, Herring III, & Eden, 2014) and Computers and Chemical Engineering 81 (Hada, Herring III, Davis, & Eden, 2015)

2.1. Computer Aided Molecular Design (CAMD)

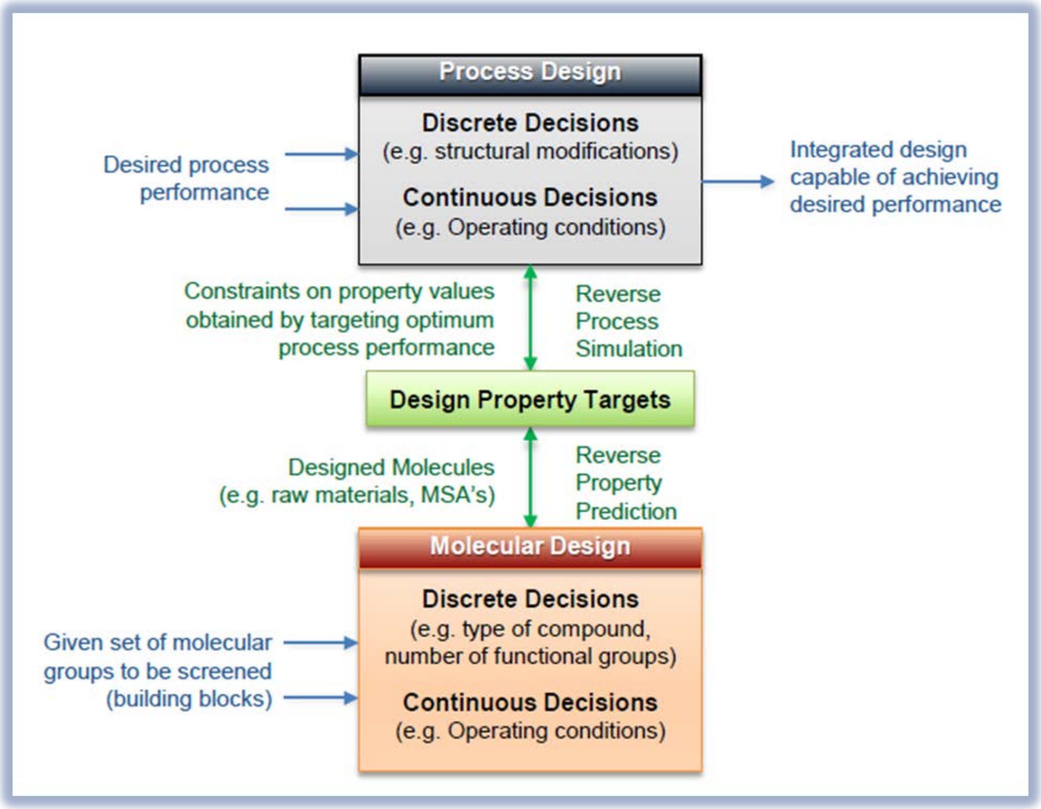
Computer Aided Molecular Design (CAMD) facilitates the utilization of algorithms to solve chemical product or process design formulations. Traditionally, chemical engineering design has seen two major roadblocks: the ability to predict chemical and physical properties and the ability to solve large scale optimization problems.

The work addressed in CHAPTER 3 takes advantage of the existing methods within the *computer-aided molecular design* (CAMD) framework. (Harper et al, 2000 Eljack et al, 2008 McLesse et al, 2010) A characterization-based method was combined with chemometric (Solvanson et al, 2011 Hada et al, 2011) and property clustering techniques (Shelley et al, 2000) in a reverse problem formulation (Eden et al, 2004) to develop a logical and systematic approach of selectively choosing a given ionic pair that matches a set of desired physical property targets. This work was previously published in Computer aided Chemical Engineering (Davis, Herring III, & Eden, 2016).

2.1.1. Reverse Problem Formulation

Reverse problem formulation helps circumvent the challenges posed by coupling of scales by bridging them through a property domain. Reverse problem formulations use the duality of linear programming to reformulate the design problem as a series of reverse problems solved in the property domain. This way, an immense computational cost associated with the hierarchical nesting across multiple-scales is relieved leading to a much more efficient solution achieved through reduction in the need for enumeration. (Eden et al, 2003)

Reverse problem formulation decomposes the conventional forward process-product design problem which are naturally iterative into two reverse problems linked by property targets. The first step defines the property targets which will satisfy the desired process performance and the second step selects the molecules which provide the property targets. This efficient process will identify the optimum solution and is pictorially displayed in Figure 4.


Process Design
Discrete Decisions
(e.g. structural modifications)
Continuous Decisions
(e.g. Operating conditions)
Desired process performance
Integrated design capable of achieving desired performance
Constraints on property values obtained by targeting optimum process performance
Reverse Process Simulation
Design Property Targets
Designed Molecules
(e.g. raw materials, MSA's)
Reverse Property Prediction
Molecular Design
Discrete Decisions
(e.g. type of compound, number of functional groups)
Continuous Decisions
(e.g. Operating conditions)
Given set of molecular groups to be screened (building blocks)

.

Figure 4. Reverse Process Design using Data Driven Methods

2.1.2. Prediction of Properties

Property prediction call pull from computationally complex models to data-dependent regression models and can be applied to any chemical system. Constantinou and Gani (Gani & Constantinou, 1994) classified property estimation methods into two groups. The first group referred to as approximate was further divided into empirical models and semi-empirical models. Empirical models included chemometrics, pattern matching, factor analysis, and quantitative structure property/activity relationships; whereas semi-empirical models included corresponding state theory, topology/geometry, group/atom/bond additivity. Opposite the approximate group, the reference group included fundamental models such as quantum mechanics, molecular mechanics, and molecular simulation.

Group Contribution theory employs quantitative structure property/activity relationships (QSPR/QSAR) to provide property estimations (Linusson et al, 2010). QSPR models relate the physical, mechanical, or chemical properties with the structure features of materials (Varnek et al, 2007). These models can be used for screening and optimization because they provide information on features that affect the physiochemical properties. One important consideration for this method is the requirement of a large training data set that includes the optimum molecules spanning the chemical and property space (Karmer, 1998).

2.1.3. Characterization Based Group Contribution Method

The group contribution method is based on the idea that each portion of the molecule contributes to the overall function of the molecule. The properties of the molecule are estimated by identifying fragments such as bonds, atoms or groups and summing all the contributions from each fragment to make the whole molecule (Chemmangattuvalappil, Eljack, & Eden, 2009). These estimations can be made utilizing only structural information. As shown in Equation 2.1 (Gani R. , 2004) a group contribution property model estimates the property function of the molecule as a linear combination of the fragment contributions.

$$f(x) = \sum_i N_i C_i + \sum_j M_j D_j + \sum_k O_k E_k \tag{2.1}$$

Where, N_i = the number of occurrences of first-order group *i*

C_i = the contribution from the first-order group *i*

M_j = the number of occurrences of second-order group *j*

D_i = the contribution from the second-order group *j*

O_k = the number of occurrences of third-order group *k*

E_k = the contribution from the third-order group *k*

Since they assume no interaction between groups, basic information is stored in the first order terms. Second order terms correct for the interactions between first order terms and are derived from the first order terms. Third order terms are estimated to correct for poly-functional compounds using the first and second order terms (Harper & Gani, Computer aided tools for design/selection of environmentally firnedsly substances, 2000).

While a good estimator, data is a limiting factor. Since the availably of atom or group type and bonding present is required to describe the structure and property contributions for the groups must be obtainable, the method is reliant on building blocks that have been synthesized and measured. For example, only a small portion of ionic liquids have been studied, so we are limited to molecules built from known building blocks that have been measured. Additionally, group contribution method cannot represent all possible atomic configurations leading to the need for an efficient method to design structured molecules. One method combines multivariate methods with decomposition techniques.

This framework employs infrared, near infrared spectroscopy or other multivariate characterization techniques to describe a set of samples. Then, decomposition techniques including principal component analysis and partial least squares to determine the underlying latent variables that will describe the molecule’s properties. Latent variables are characterized indirectly rather than being observed directly. The candidate molecules can then be identified by combining molecular fragments until the resulting properties match the targets.

2.1.4. Characterization Techniques

Characterization techniques are a class of experimental tools meant to describe chemical constituency and molecular structure plus the orientation and alignment of those molecules. These techniques provide large quantities of correlated data that can provide molecular architecture information (Marrero & Gani, 2001). Managing this complexity requires a systematic method for determination of which specific information will be useful in the modeling to reduce complexity. Common examples include infrared spectroscopy, nuclear magnetic resonance, and x-ray diffraction spectroscopy. This data is often applied to a training set of molecules defined by an experimental design used to understand a set of property attributes. Figure 5 shows general characterization and associated attributes.

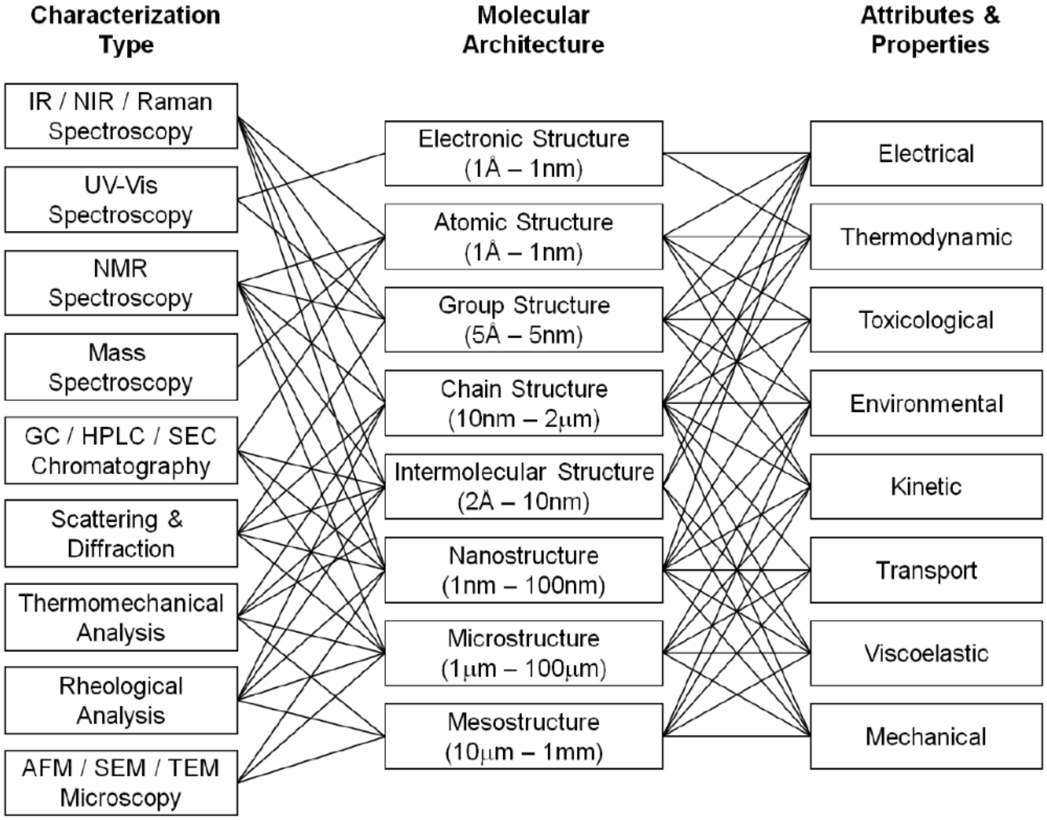

Characterization Type
Molecular Architecture
Attributes & Properties
IR / NIR / Raman Spectroscopy
UV-Vis Spectroscopy
NMR Spectroscopy
Mass Spectroscopy
GC / HPLC / SEC Chromatography
Scattering & Diffraction
Thermomechanical Analysis
Rheological Analysis
AFM / SEM / TEM Microscopy
Electronic Structure (1Å – 1nm)
Atomic Structure (1Å – 1nm)
Group Structure (5Å – 5nm)
Chain Structure (10nm – 2μm)
Intermolecular Structure (2Å – 10nm)
Nanostructure (1nm – 100nm)
Microstructure (1μm – 100μm)
Mesostructure (10μm – 1mm)
Electrical
Thermodynamic
Toxicological
Environmental
Kinetic
Transport
Viscoelastic
Mechanical

Figure 5. An Overview of the Interconnectivity of Characterization Techniques, Molecular Architecture and Physical Properties and Attributes of Chemical and Material Products. (Solvason, 2011)

Infrared and near infrared spectroscopy provides information concerning the electronic structure, atomic structure, chain structure and intermolecular structure. While other techniques could have provided similar results, spectroscopy was considered in this dissertation. Spectroscopy is the detection and analysis of the radiated energy absorbed or emitted by the architecture of a chemical species. (Atkins, 1998) The shape and size of relative intensities are indicators of molecular architecture because those intensities are primarily functions of the atom specific dipole changes caused by the vibrations of the corresponding bonds. The infrared absorbance frequencies and magnitudes of the functional groups' spectrums are listed below in Figure 6.

Methine Groups, -CH-

Band	**Wavelength Region $[cm^{-1}]$**			**Relative Intensity**	**% Transmittance**
	High	**Low**	**Average**		
Bending *(δ)*	1360	1320	1340	*w*	90
Stretching *(υ)*	2890	2880	2885	*w*	90

Methylene Groups, $-CH_2-$

Band	**Wavelength Region $[cm^{-1}]$**			**Relative Intensity**	**% Transmittance**
	High	**Low**	**Average**		
Scissoring Bend (δ_s)	1480	1440	1460	*m*	50
Symmetrical Stretching (υ_s)	2870	2840	2855	*m*	50
Asymmetrical Stretching (υ_a)	2940	2915	2928	*m-s*	30

Methyl Groups, -CH3

Band	**Wavelength Region $[cm^{-1}]$**			**Relative Intensity**	**% Transmittance**
	High	**Low**	**Average**		
Sym Bend (δ_s)	1390	1370	1380	*m-s*	30

Asym. Bend. (δ_a)	1465	1440	1453	*m*	50
Symmetrical Stretching (ν_s)	2885	2865	2875	*m*	50
Asymmetrical Stretching (ν_a)	2975	2950	2963	*m-s*	30

Tetramethyl Groups, $-C(CH_3)_3$

Band	Wavelength Region [cm^{-1}]			Relative Intensity	% Transmittance
	High	Low	Average		
C-C Skeletal Bend (δ_s)	930	925	928	*m*	50
C-C Skeletal Bend (δ_s)	1010	990	1000	*m-w*	70
C-C Skeletal Bend (δ_s)	1225	1165	1195	*m*	50
C-C Skeletal Bend (δ_s)	1255	1245	1250	*m*	50
$C-CH_3$ Sym. Bend. (δ_s)	1395	1350	1373	*m-s*	30
$C-CH_3$ Sym. Bend. (δ_s)	1420	1375	1398	*m*	50
$C-CH_3$ Asym. Bend. (δ_a)	1475	1435	1455	*m*	50
C-H Sym. Stretching (ν_s)	2885	2865	2875	*m*	50
C-H Asym. Stretching (ν_a)	2975	2950	2963	*m-s*	30

Aliphatic Methoxy Groups, $-O-CH_3$ (Special Methyl)

Band	Wavelength Region [cm^{-1}]			Relative Intensity	% Transmittance
	High	Low	Average		
C-O Def. Bend. (δ_d)	580	340	460	*m-w*	70
CH_3/CO Rocking Bend (δ_d)	1190	1100	1145	*m-w*	70
CH_3 Rock Bend (δ_d)	1235	1155	1195	*m-w*	70
CH_3 Sym Bend (δ_s)	1460	1420	1440	*M*	50
CH_3 Asym. Bend. (δ_a)	1475	1435	1455	*m*	50
CH_3 Asym. Bend. (δ_a)	1485	1445	1465	*m*	50
$C-H_3$ Sym. Str. (ν_s)	2880	2815	2848	*m*	50
$C-H_3$ Asym. Str. (ν_a)	2985	2920	2953	*m*	50
C-H Asym. Str. (ν_a)	3030	2950	2990	*m*	50

Vinyl Group, $-CH=CH_2$

Band	Wavelength Region [cm^{-1}]	Relative	%

	High	Low	Average	Intensity	Transmittance
C=C Tors. Bend (δ_T)	485	410	448	*m-s*	30
C=C Eth. Twist. Bend. (δ_t)	600	380	490	*m-s*	30
C=C Eth. Twist. Bend. (δ_t)	720	410	565	*w*	90
C-H_2 OoP Rock. Bend. (δ_r)	980	810	895	*s*	10
C-H OoP Bending. (δ_r)	1010	940	975	*s*	10
C-H IP Def. Bend. (δ_d)	1180	1010	1095	*m-w*	70
C-H_2 Def. Bend. (δ_d)	1330	1240	1285	*m*	50
C-H_2 Sci. Bend. (δ_s)	1440	1360	1400	*m*	50
C=C Stretching (ν)	1645	1640	1643	*m-w*	70
C-H_2 1st Overtone Bend (2δ)	1840	1820	1830	*v*	90
C-H 1st Overtone Bend (2δ)	1990	1970	1980	*v*	90
C-H_2 Sym. Stretch (ν_s)	3070	2930	3000	*M*	50
C-H Stretch (ν)	3110	2980	3045	*M*	50
C-H_2 Asym. Stretch (ν_a)	3150	3000	3075	*M*	50

Vinylidene Group, CH_2=C- -

Band	Wavelength Region [cm^{-1}]			Relative Intensity	% Transmittance
	High	Low	Average		
C=C Skeletal Stretch (ν)	470	435	453	*m-w*	70
C=C Skeletal Stretch (ν)	560	530	545	*s*	10
C=C Eth. Twist. Bend. (δ_t)	715	680	698	*w*	90
C-H_2 OoP Rock. Bend. (δ_r)	895	885	890	*s*	10
C-H_2 IP Def. Bend. (δ_d)	1320	1290	1305	*w*	90
C-H_2 Sci. Def Bend. (δ_s)	1420	1405	1413	*w*	90
C=C Stretching (ν)	1675	1625	1650	*m-w*	70
C-H_2 1st Overtone Bend (2δ)	1800	1750	1775	*w*	90
C-H_2 Sym. Stretch (ν_s)	2985	2970	2978	*m-w*	70
C-H_2 Asym. Stretch (ν_a)	3095	3075	3085	*m-w*	70

cis-Vinylene Group, -CH=CH-

Band	Wavelength Region [cm^{-1}]			Relative Intensity	% Transmittance
	High	Low	Average		
C-H Tors. Bend (δ_T)	490	320	405	*m-s*	30
C=C Skeletal Bend (δ_T)	500	460	480	*s*	10
-C=CH Def. Bend. (δ_d)	590	440	515	*m-s*	30
C=C Eth. Twist. Bend. (δ_t)	630	570	600	*s*	10
C-H Wag. Bend. (δ_w)	790	650	720	*m-s*	30
C-H Wag. Bend. (δ_w)	1000	850	925	*m-w*	70
C-H Def. Bend. (δ_d)	1295	1185	1240	*w*	90
C-H Def. Bend. (δ_d)	1425	1355	1390	*w*	90
C=C Stretching (ν)	1665	1630	1648	*m*	50
C-H Stretch (ν)	3040	2980	3010	*m*	50
C-H Stretch (ν)	3090	3010	3050	*m*	50

trans-Vinylene Group, -CH=CH-

Band	Wavelength Region [cm^{-1}]			Relative Intensity	% Transmittance
	High	Low	Average		
C-H Tors. Bend (δ_T)	490	320	405	*m-s*	30
C=C Skeletal Bend (δ_T)	500	480	490	*s*	10
-C=CH Def. Bend. (δ_d)	590	440	515	*m-s*	30
C=C Eth. Twist. Bend. (δ_t)	580	515	548	*m-s*	30
C-H Wag. Bend. (δ_w)	850	750	800	*m-w*	70
C-H Wag. Bend. (δ_w)	1000	910	955	*v*	90
C-H Def. Bend. (δ_d)	1305	1260	1282.5	*v*	90
C-H Def. Bend. (δ_d)	1340	1355	1347.5	*v*	90
C=C Stretching (ν)	1680	1665	1673	*m-w*	70
C-H Stretch (ν)	3050	3000	3025	*m*	50
C-H Stretch (ν)	3065	3015	3040	*m*	50

Hydroxyl Group, -OH (with intermolecular H-bonding)

Band	Wavelength Region [cm^{-1}]			Relative Intensity	% Transmittance
	High	Low	Average		

Bending (δ)	710	570	640	*m*	50
Stretching (ν)	3550	3230	3390	*m-s*	30

Primary Alcohol Group, -CH2OH (with intermolecular H-bonding)

Band	Wavelength Region [cm^{-1}]			Relative Intensity	% Transmittance
	High	Low	Average		
C-O Def. Bend (δ_d)	555	395	475	*m-w*	70
C-O IP. Def. Bend (δ_d)	500	440	470	*w*	90
O-H OoP. Def. Bending (δ_d)	710	570	640	*m-w*	70
C-CO Stretch (ν)	900	800	850	*m*	50
C-H_2 Twist Bend (δ_t)	960	800	880	*m-w*	70
C-C-O Stretch (ν)	1090	1000	1045	*S*	10
C-H_2 Twist. Bending (δ_t)	1300	1280	1290	*m-w*	70
C-H_2 Wag Bend (δ_w)	1390	1280	1335	*m-w*	70
O-H Def. Bend (δ_d)	1440	1260	1350	*m-s*	30
C-H_2Def Bend (δ_d)	1480	1410	1445	*m-w*	70
C-H_2 Sym. Stretch (ν_s)	2935	2840	2888	*m-w*	70
C-H_2 Asym. Stretch (ν_a)	2990	2900	2945	*m-w*	70
O-H Stretching (ν)	3550	3230	3390	*m-s*	30

Secondary Alcohol Group, - -CHOH (with intermolecular H-bonding)

Band	Wavelength Region [cm^{-1}]			Relative Intensity	% Transmittance
	High	Low	Average		
C-O OoP. Def. Bend (δ_d)	390	330	360	*m-w*	70
C-O IP. Def. Bend (δ_d)	500	440	470	*w*	90
O-H OoP. Def. Bending (δ_d)	660	600	630	*m-w*	70
C-CO Stretch (ν)	900	800	850	*m*	50
C-O Stretch (ν)	1150	1075	1113	*m-w*	70
C-H Def. Bending (δ_d)	1350	1290	1320	*s*	10
C-H Wag Bend (δ_w)	1400	1330	1365	*s*	10
O-H + C-H_2 Coup. Bend. (δ_c)	1430	1370	1400	*m-w*	70

O-H Def. Bend (δ_d)	1440	1260	1350	*m-w*	70
C-H Stretching (ν)	2890	2880	2885	*m-s*	30
O-H Stretching (ν)	3550	3230	3390	*m-w*	70

Aliphatic Ether Group, -O-

Band	**Wavelength Region [cm^{-1}]**			**Relative Intensity**	**% Transmittance**
	High	**Low**	**Average**		
C-O-C def vib (δ_d)	440	420	430	*w*	90
Sym C-O-C str (ν_s)	1140	820	980	*w*	90
Asym C-O-C Str (ν_a)	1150	1060	1105	*s*	10
Rocking vib	1200	1185	1193	*m-w*	70
Wagging vib	1400	1360	1380	*m*	50
Asym and Sym -CH_3 def. vib	1470	1435	1453	*m*	50
CH_2 def vib	1475	1445	1460	*m*	50
Sym CH_2 str	2880	2835	2858	*m*	50
Sym. -CH_3 Str	2900	2840	2870	*m*	50
Asym CH_2 str	2955	2920	2938	*m*	50
Asym. -CH_3 Str	2995	2955	2975	*m*	50

Alkyl Peroxide Group, -O-O-

Band	**Wavelength Region [cm^{-1}]**			**Relative Intensity**	**% Transmittance**
	High	**Low**	**Average**		
O-O Stretch (ν)	900	800	850	*w*	90
C-O Stretch (ν)	1150	1030	1090	*m-s*	30

Saturated Aliphatic Ester Group, -CO-O-

Band	**Wavelength Region [cm^{-1}]**			**Relative Intensity**	**% Transmittance**
	High	**Low**	**Average**		
C-O-C Sym. Stretch (ν_s)	1160	1050	1105	*s*	10
C-O-C Asym. Stretch (ν_a)	1275	1185	1230	*s*	10
C=O Stretch (ν)	1750	1725	1738	*s*	10
C=O 1st Overtone ($2\nu_s$)	3460	3440	3450	*w*	90

Saturated Aliphatic Methyl Ester Group, -CO-O-CH_3

Band	Wavelength Region [cm^{-1}]			Relative Intensity	% Transmittance
	High	Low	Average		
Unlisted	450	430	440	*m-s*	30
CO-O Rocking Bend (δ_r)	530	340	435	*w*	90
C-C-O Sym. Stretch (ν_s)	1160	1050	1105	*s*	10
C-O Stretch (ν)	1175	1155	1165	*s*	10
C-C-O Asym. Stretch (ν_a)	1275	1185	1230	*s*	10
O-CH_3 Stretch (ν)	1315	1195	1255	*s*	10
Unlisted	1370	1350	1360	*w*	90
CH_3 Sym. Def. Bend (δ_d)	1460	1420	1440	*m-s*	30
CH_3 Asym. Def. Bend (δ_d)	1465	1420	1443	*m-s*	30
CH_3 Asym. Def. Bend (δ_d)	1485	1435	1460	*m*	50
C=O Stretch (ν)	1750	1725	1738	*s*	10
CH_3 Sym. Stretch (ν)	3000	2860	2930	*m*	50
CH_3 Asym. Stretch (ν)	3030	2950	2990	*m-w*	70
CH_3 Asym. Stretch (ν)	3050	2980	3015	*m-w*	70
C=O 1st Overtone ($2\nu_s$)	3460	3440	3450	*w*	90

Saturated Aliphatic Ethyl Ester Group, -CO-O-CH_2CH_3

Band	Wavelength Region [cm^{-1}]			Relative Intensity	% Transmittance
	High	Low	Average		
C-O-C Def Bend (δ_d)	370	250	310	*m-w*	70
C-O-C Def Bend (δ_d)	395	305	350	*m-w*	70
CO-O Rocking Bend (δ_r)	485	365	425	*m-w*	70
CO OoP Rocking Bend (δ_r)	700	550	625	*w*	90
CH_2 Rocking Bend (δ_r)	825	775	800	*w*	90
C-C str (ν)	940	850	895	*w*	90
CH_3 Rock. Bend (δ_r)	1150	1080	1115	*w*	90
C-C-O Sym. Stretch (ν_s)	1160	1050	1105	*s*	10
CH_3 Rock. Bend (δ_r)	1195	1135	1165	*w*	90
C-C-O Asym. Stretch (ν_a)	1275	1185	1230	*s*	10

CH_2 Twist. Bend (δ_T)	1340	1325	1333	*m-w*	70
CH_2 Wag. Bend (δ_w)	1385	1335	1360	*m-w*	70
CH_3 Sym. Def. Bend (δ)	1390	1360	1375	*m-s*	30
CH_3 Asym. Def. Bend (δ)	1480	1435	1458	*m*	50
OCH_2 Def. Bend. (δ)	1490	1460	1475	*m-w*	70
C=O Stretch (ν)	1750	1725	1738	*s*	10
CH_3 Stretch (ν)	2920	2860	2890	*w*	90
CH_3 Sym. Stretch (ν_s)	2930	2890	2910	*w*	90
CH_3 Asym. Stretch (ν_a)	2995	2930	2963	*m*	50
C=O 1st Overtone ($2\nu_s$)	3460	3440	3450	*w*	90

Acrylate Ester Group, CH_2=CH-CO-O-

Band	**Wavelength Region [cm^{-1}]**			**Relative Intensity**	**% Transmittance**
	High	**Low**	**Average**		
C=C Tors. Bend (δ_T)	485	410	448	*m-s*	30
CO-O Rocking Bend (δ_r)	485	365	425	*m-w*	70
C=C Eth. Twist. Bend. (δ_t)	600	380	490	*m-s*	30
C-O-C Def Bend (δ)	675	660	668	*m*	50
CO OoP Rocking Bend (δ_r)	700	550	625	*w*	90
=CH2 Twist Bend (δ_t)	810	800	805	*m-s*	30
CH_2 Rocking Bend (δ_r)	825	775	800	*w*	90
C-C str (ν)	940	850	895	*w*	90
=CH2 Wag. Bend (δ_w)	970	960	965	*s*	10
C-H Def. Wag (δ_w)	990	980	985	*m*	50
C-H OoP Bending. (δ_r)	1010	940	975	*s*	10
C-C Skel. Bend (δ)	1070	1065	1068	*m*	50
CH_3 Rock. Bend (δ_r)	1150	1080	1115	*w*	90
C-C-O Sym. Stretch (ν_s)	1160	1050	1105	*s*	10
C-H IP Def. Bend. (δ_d)	1180	1010	1095	*m-w*	70
Unlisted	1200	1195	1198	*s*	10
C-C-O Asym. Stretch (ν_a)	1275	1185	1230	*s*	10
=CH Rock. Bend (δ_r)	1290	1270	1280	*m*	50
Unlisted	1290	1280	1285	*s*	10
=CH2 Def Bend (δ)	1420	1400	1410	*m*	50

C-H_2 Sci. Bend. (δ_s)	1440	1360	1400	*m*	50
C=C Stretch (ν)	1635	1615	1625	*m*	50
C=C Stretch (ν)	1650	1630	1640	*m-s*	30
C=O Stretch (ν)	1725	1710	1718	*s*	10
C-H_2 1st Overtone Bend (2δ)	1840	1820	1830	*w*	90
C-H 1st Overtone Bend (2δ)	1990	1970	1980	*w*	90
C-H_2 Sym. Stretch (ν_s)	3070	2930	3000	*m*	50
C-H Stretch (ν)	3110	2980	3045	*m*	50
C-H_2 Asym. Stretch (ν_a)	3150	3000	3075	*m*	50
C=O 1st Overtone ($2\nu_s$)	3460	3440	3450	*w*	90

Methacrylate Ester Group, $CH_2=C(CH_3)$-CO-O-

Band	Wavelength Region [cm^{-1}]			Relative Intensity	% Transmittance
	High	Low	Average		
C=C Skeletal Stretch (ν)	470	435	453	*m-w*	70
C=C Skeletal Stretch (ν)	560	530	545	*s*	10
C-O-C Def Bend (δ)	660	645	653	*m*	50
C=C Eth. Twist. Bend. (δ_t)	715	680	698	*w*	90
C-C Skel Bend (δ)	825	805	815	*m-s*	30
C-H_2 OoP Rock. Bend. (δ_r)	895	885	890	*s*	10
=CH2 Wag. Bend (δ_w)	950	935	942.5	*s*	10
C-C Skel. Bend (δ)	1010	990	1000	*m*	50
C-C Skel. Bend (δ)	1020	1000	1010	*m*	50
C-O-C Sym. Stretch (ν_s)	1160	1150	1155	*s*	10
C-O-C Asym. Stretch (ν_a)	1275	1185	1230	*s*	10
Unlisted	1310	1290	1300	*s*	10
C-H_2 IP Def. Bend. (δ_d)	1320	1290	1305	*w*	90
=CH Rock. Bend (δ_r)	1335	1315	1325	*m*	50
CH_3 Sym Bend (δ_s)	1390	1370	1380	*m-s*	30
=CH2 Def Bend (δ)	1420	1400	1410	*m*	50
CH_3 Asym. Bend. (δ_a)	1465	1440	1453	*m*	50
C=C Stretch (ν)	1650	1630	1640	*m*	50
C=O Stretch (ν)	1725	1710	1718	*s*	10

C-H$_2$ 1st Overtone Bend (2δ)	1800	1750	1775	*w*	90
CH$_3$ Sym. Stretching (ν_s)	2885	2865	2875	*m*	50
C-H$_2$ Sym. Stretch (ν_s)	2985	2970	2978	*m-w*	70
CH$_3$ Asym. Stretching (ν_a)	2975	2950	2963	*m-s*	30
C-H$_2$ Asym. Stretch (ν_a)	3095	3075	3085	*m-w*	70
C=O 1st Overtone ($2\nu_s$)	3460	3440	3450	*w*	90

o-Alkyl Phenol Group (With H-bonding)

Band	Wavelength Region [cm^{-1}]			Relative Intensity	% Transmittance
	High	Low	Average		
C-OH IP Bending (δ)	450	375	413	*w*	90
O-H OoP. Def. Bending (δ_d)	720	600	660	*s*	10
C-O Stretch (ν)	1260	1180	1220	*s*	10
O-H IP Bending (δ)	1410	1310	1360	*s*	10
COH bending vib	1330	1310	1320	*m*	50
O-H Stretching (ν)	3250	3000	3125	*m*	50
CO Str	1255	1240	1248	*s*	10
OH def and CO str vib	1175	1160	1168	*s*	10
OH def and CO str vib	760	740	750	*m*	50
OR substituted	3595	3470	3533	*m*	50

p-Alkyl Phenol Group (With H-bonding)

Band	Wavelength Region [cm^{-1}]			Relative Intensity	% Transmittance
	High	Low	Average		
C-OH IP Bending (δ)	450	375	413	*w*	90
O-H OoP. Def. Bending (δ_d)	720	600	660	*s*	10
C-O Stretch (ν)	1260	1180	1220	*s*	10
O-H IP Bending (δ)	1410	1310	1360	*s*	10
O-H Stretching (ν)	3250	3000	3125	*m*	50
CO Str	1260	1245	1253	*s*	10
OH def and CO str vib	1175	1165	1170	*s*	10

OH def and CO str vib	835	815	825	*m*	50
OR substituted	3595	3470	3533	*m*	50

Monosubstituted Benzenes

Band	Wavelength Region [cm^{-1}]			Relative Intensity	% Transmittance
	High	Low	Average		
Ring OoP def vib	560	415	488	*m-s*	30
Ring IP def vib	630	605	618	*m-w*	70
=C-H Ring OoP def vib	710	670	690	*s*	10
=C-H OoP def vib	820	720	770	*s*	10
=C-H OoP def vib	900	860	880	*m-w*	70
=C-H IP def vib	1010	990	1000	*w*	90
=C-H IP def vib	1040	1000	1020	*m-w*	70
=C-H IP def vib	1085	1050	1068	*m*	50
=C-H IP def vib	1175	1130	1153	*w*	90
=C-H IP def vib	1195	1165	1180	*m-w*	70
=C-H IP def vib	1250	1230	1240	*w*	90
-C=C- Str Vib	1625	1590	1608	*v*	90
=C-H Str. Vib	3105	3000	3053	*m*	50

1,2,4- Trisubstituted Benzene

Band	Wavelength Region [cm^{-1}]			Relative Intensity	% Transmittance
	High	Low	Average		
Ring OoP def vib	475	425	450	*m-s*	30
=C-H OoP def vib (2H)	740	690	715	*m-w*	70
=C-H OoP def vib (2H)	780	760	770	*s*	10
=C-H OoP def vib (2H)	860	840	850	*m-s*	30
=C-H OoP def vib (1H)	940	885	913	*m-s*	30
=C-H IP def vib	1040	1020	1030	*m-w*	70
=C-H IP def vib	1160	1140	1150	*m-w*	70
=C-H IP def vib	1220	1200	1210	*w*	90
-C=C- Str Vib	1625	1590	1608	*v*	90

=C-H Str. Vib	3105	3000	3053	*m*	50

Figure 6. IR Frequencies and Magnitudes of Functional Groups

When the molecule is bombarded with radiation, molecular spectra are generated based on the motion of atomic nuclei within the architecture, exploiting the fact that molecules absorb specific frequencies that are characteristic of their nature. Motions are either absorbed or emitted based on the surrounds and can be in straight line like symmetrical and asymmetrical stretching, and/or rotational like twisting, rocking, wagging, and scissoring. (Workman, 2008)

When group combination theory is applied to the design problem, the infrared spectroscopy data can be related to functional groups providing significant computational efficiency. Using the concept of symmetry, many atoms, and combinations of atoms in a symmetric molecule can be considered to be in the same chemical environment, and by extension, the vibrational motions and absorptions or infrared spectra will be identical. (Solvason, 2011) Figure 7 shows the infrared spectrum of butylated hydroxytoluene molecule and its molecular structure showing the first and second order GC groups.

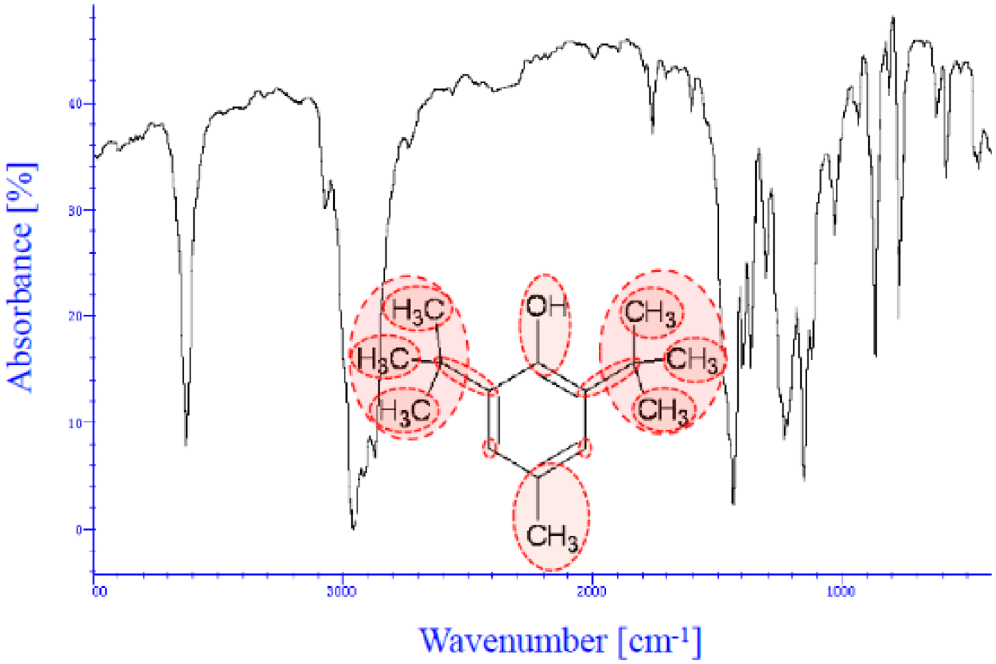


Figure 7. IR spectra of butylated hydroxytoluene molecule

This descriptor data provides information on the molecular architecture; however it is likely that descriptor variables will be correlated because they are linear functions of other variables (Matsuda, 2007). Therefore, it is important to build appropriate models to manage that complexity and capture the important features of the data. Multivariate statistical techniques will be used to decompose the information to be used in the initial training set.

2.1.5. Density Functional Theory

Most properties of a molecule are dependent on the behaviors of its electrons, and to model or predict them it is necessary to have an accurate method to compute the electronic structure. (Talaty, 2004) Density functional theory is a computational technique the can be used to predict the properties of molecules through the investigation of the electronic structure based on the electron density. It is a quantum theory in which the only input data are the atomic number of the constituent atoms and some initial structural information (Hall & Kier, 2001). Rather than calculating the implications of all electrons, they are replaced with an equivalent single electron calculation, also known as a functional, in which each electron is moving in an effective potential. (Carrera et al, 2005) The potential is a sum of the external potentials determined by the elemental composition of the system and the goal is to minimize the total energy function.

One class of energy functional forms used for estimation with density functional theory are hybrid functionals which combine aspects of density functional theory with the Hartree-Fock method. The Hartree-Fock method approximates the wave function and energy of a quantum body in a stationary state (Scott A. R., 1996). This combination solves problems inherent with each method. Hartree-Fock methods exactly treat exchange correlation but have difficulties recovering dynamic electron correlation while density functional theory has an exact for dynamic electronic correlation but since DFT is not quantum mechanical, it must approximate exchange correlation. These hybrid functionals are linear combinations of Hartree-Fock exchange functional shown in Equation 2.2.

$$E_x^{HF} = -\frac{1}{2}\sum_{i,j}\int\int \varphi_i^*(r_1)\varphi_j^*(r_1)\frac{1}{r_{12}}\varphi_i(r_2)\varphi_2 dr_1 dr_2 \qquad 2.2$$

One example of this class of hybrid functions is Becke, three-parameter, Lee-Yang-Parr exchange correlation functional, also known as B3LYP. It has become the most common hybrid method and the equation for this function is given in equation 2.3 which incorporates equation 2.2. (Scott & Random, 1996)

$$E_{xc}^{B3LYP} = E_x^{LDA} + a_0(E_x^{HF} - E_x^{LDA}) + a_x(E_x^{GGA} - E_x^{LDA}) + E_C^{LDA} + a_c(E_c^{GGA} - E_c^{LDA}) \qquad 2.3$$

The application of density functional theory as it relates to characterization techniques as a part of a computer aided molecular design allows for the simulation of infrared spectroscopy for molecules where experimental data is not available. For example, ionic liquids are a class that has the potential for many combinations that have not been synthesized. Simulated infrared spectroscopy will allow for the study of those liquids in different applications without the time and expense of synthesizing all possibilities. When infrared absorbance frequencies are seen in a certain wavelength range as shown in Figure 7, the presence of a certain functional group is implicit.

2.1.6. Latent Variable Modeling

In an effort to analyze the spectra data, decomposition techniques should be employed to consolidate the date and derive the latent variable relationships. The goal of the decomposition techniques is to describe the variation in the characterization technique data using the smallest number of variables in a process that transforms a *p*-dimensional property characterization structural descriptor data set of molecular architecture information into a low *m*-dimensional sub property space. (Eriksson, et al., 2006) By compressing the p property data to m principal components data using the variance-covariance structure, it guarantees that the property space is orthogonal and without collinearity that may exist in that attribute. Principal component analysis is the most common decomposition technique and utilizes least square as the fitting function (Gabrielsson, Lindberg , & Lundsteadt, 2002).

The main shortcoming of principal component analysis and other decomposition techniques is the susceptibility to large differences in scales and variance, so data should be standardized prior to analysis. (Solvason, 2011) If this pretreatment step is not conducted, then the resulting model runs the risk of not providing useful data. General practice calls for the property variables to be mean-centered and scaled prior to analysis. Since variance is directly related to the size of the numerical range and principal component analysis is a maximum variance projection method, it follows that variables with large variances are more likely to be expressed in the model than lower variance variables. This process ensures that all variables are considered equally as visually depicted in Figure 8. The bars represent each variable with the horizonal lines denoting the mean and length of the bar is equal to its standard deviation.

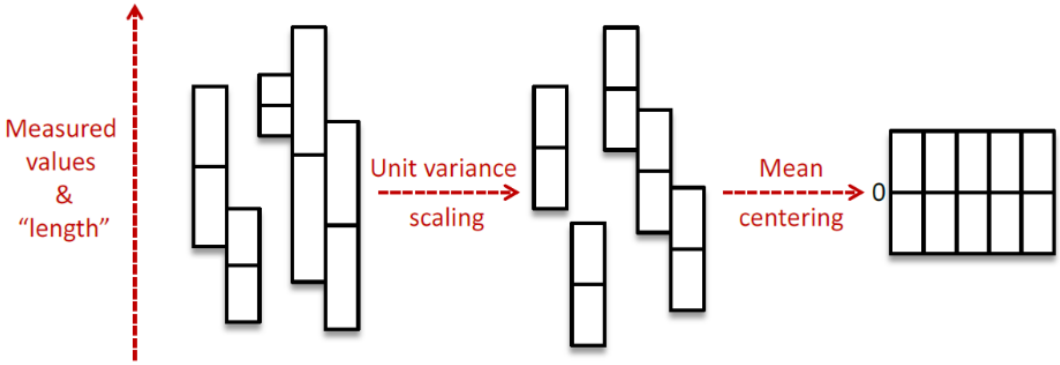


Figure 8. Visual representation of data pre-treatment for principal component analysis

The property descriptor data matrix X_{MxK} consisting of M observations described by K descriptors, is mean-centered and scaled as follows. The mean for each variable is calculated based on the entire sample and then subtracted from each measurement to mean center the data. The scaling process then divides the mean centered data by the standard deviation as shown in the following equations.

$$\bar{x}_j = \left[\frac{1}{M}\sum_{j=1}^{M} x_{i,j}\right] \qquad 2.4$$

$$s_j^2 = \frac{1}{M-1}\sum_{j=1}^{M}\left(x_{i,j} - \bar{x}_j\right)^2 \qquad 2.5$$

$$x_{ij} = \frac{x_{ij}-\bar{x}_j}{s_j} \qquad 2.6$$

This standardization ensures that the multiple sources of property data will be equally considered, and the data will be decomposed into a meaningful regression.

2.1.6.1. Principal Component Analysis (PCA)

Principal component analysis identifies patterns and visualizes multivariate data by using as few variables as possible by mapping the original data to a lower dimension. Principal component analysis compresses the size of the data and complexity of the data while capturing and analyzing the structure of the data. (Erisson & Johansson, 1996)

The process begins by relating a set of variables into principal components or linear combinations of the original variables ordered from greatest variance to the least variance. The principal components are orthogonal and have no correlations with other principal components (Muteki, 2006). Latent properties represent the relative distance to the projected values of each property on the eigenvector hyperplane. The eigenvalues measure how the properties are weighted for each principal component and their representation on the hyperplane. Eigenvalues which are uncorrelated have a value of 0 ranging up to -1 or 1 for highly correlated values. (Jackson, 1991)

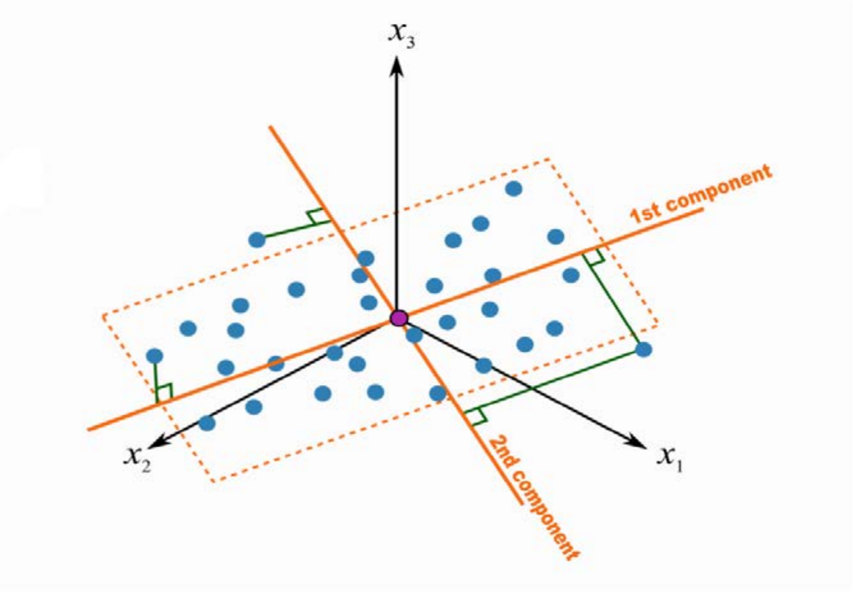


Figure 9. Projection of higher dimensional data onto a single hyperplane

When principal components are derived together, they define a plane, as seen in Figure 9, which is the best approximation of the data (Wold S. , 1995). The score matrix (T) represents the projection of the data onto the plane and each new coordinate along the principal components line represents the score (t_i). The loadings matrix (L) and each loading defines the orientation of the principal component plane with respect to the original variable and reveal both the magnitude and direction of the correlation. The data set of molecular architecture information known as X_{MxK} signifies M observations of K variables where T is the score matrix and P is the loading matrix both of mutually orthogonal columns and is detailed in the following equation (Jaeckle C. M., 1998).

$$X_{MxK} = \sum_{i=1}^{K} t_i p_i^T = T_{MxK} P_{KxK}^T \qquad 2.7$$

Most principal component analysis result in the first two or three principal components accounting for 80% to 90% of the variation in a domain. (Johnson, 2007)

The orthogonal latent property components are fitted to the data structure beginning with the latent properties with the highest eigenvalues first and continuing until no appreciable difference from one principal component to the next.

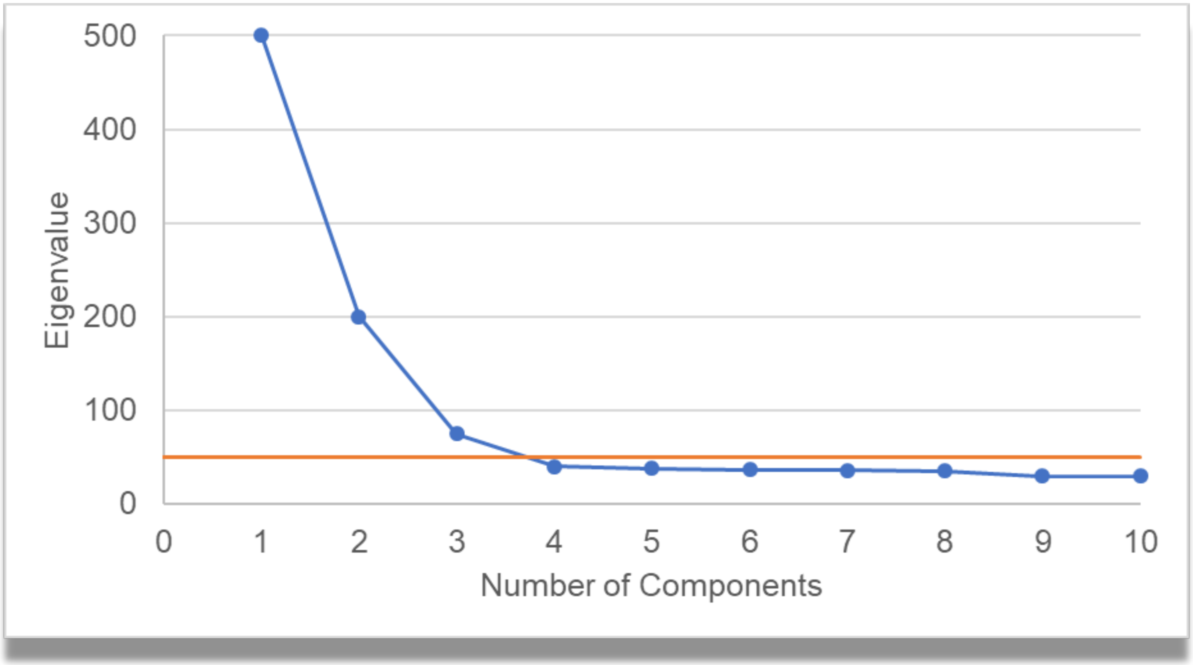


Figure 10. JMP Scree Plot of the Covariance

A scree plot, as shown in Figure 10, displays the magnitude of the eigenvectors of each principal component in decreasing magnitude. The appropriate number of principal components are determined by the bend in the graph where the shows that the remaining values are small and the same. The leveled portion of the graph displays the principal components that have no significant impact on the solution. In principal component analysis, it becomes important to understand that selecting enough principal components is important to describe all useful data. However, selecting too many can lead to overfitting the model, thus poor prediction capabilities.

Once the number of principal components is selected, the correlations between variables becomes relevant. One can understand how the variables relate to observations by examining the loading and scores plots shown in which display the level of correlation of the variable and information about the direction of the correlation either positive or negative.

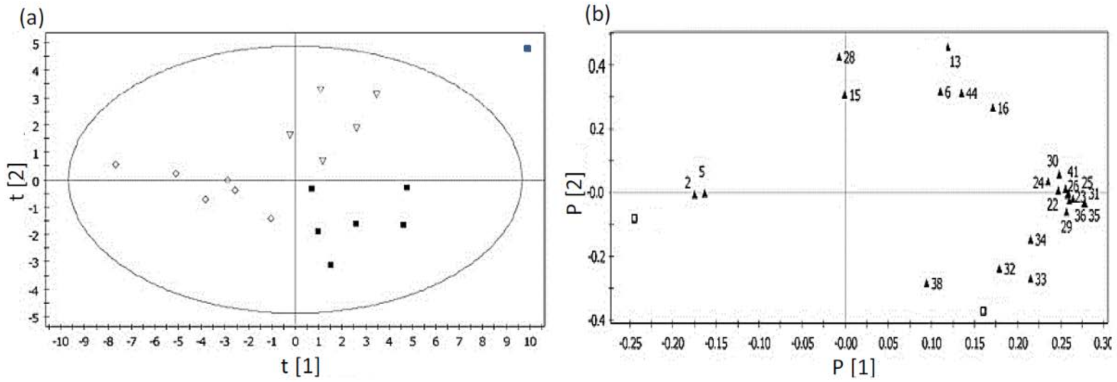


Figure 11. (a) Score Plot and (b) loading plot

2.1.6.2. Principal Component Regression (PCR)

Principal component regression is essentially a linear regression based on the results of principal component analysis and can be summarized by Figure 12.

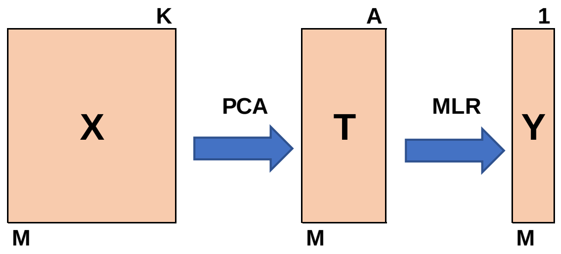


Figure 12. Diagram of Principal Component Regression

The prediction of a response variable (Y) from predictor variables (X) is obtained employing the PCR as shown in equation 2.8.

$$\hat{Y}_{MxL} = T_{MxA} * \hat{B}_{AxL} \qquad 2.8$$

$$where, \hat{B} = (T^T * T)^{-1} * T^T * Y \qquad 2.9$$

In order to solve a design problem in a single domain, all the physico-chemical attributes/properties are converted to *principal properties* (PP) by using the regression (B) coefficients from the calibration model shown in equation 2.10 (Jaeckle C. M., 1998).

$$(t_{IxA}^T)_{new} = (y_{IxL}^T)_{new} * \left(\hat{B}_{LxA}^T * \hat{B}_{AxL}\right)^{-1} * \hat{B}_{LxA}^T \qquad 2.10$$

2.1.6.3. Partial Least Squares (PLS)

Partial least squares is a regression method that is based on principal component analysis and multiple linear regression. It takes principal component regression one step future because it deals with both the descriptive information such as found with characterization property data and response information including attributes of physical-chemical property data (Kettanch-Wold, 1992). It is important to note that while partial least squares provides the best correlation for the matrices together, it is not necessarily the best description of the two individually (Joback & Reid, 1983).

To illustrate, assume that data matrix, **P**, contains molecular descriptors and data matrix, **Y**, contains attribute information and that partial least squares is modelling their relationship. The diagram below shows the individual principal component analysis for each matrix leading to the partial least squares regression where the descriptor scores, *u*, are plotted against the response scores, *t*.

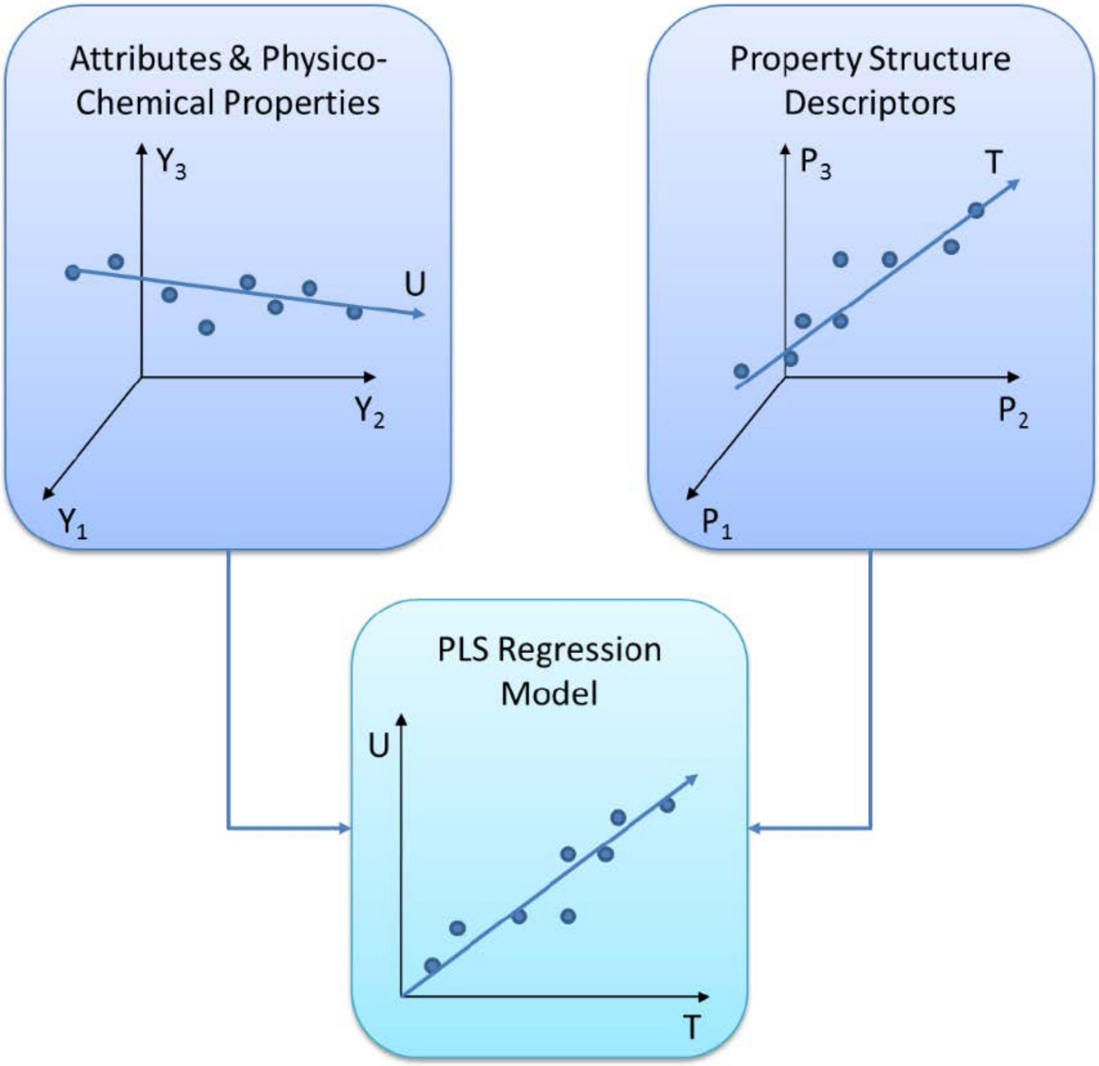

Attributes & Physico-Chemical Properties
Y3
U
Y2
Y1
Property Structure Descriptors
P3
T
P2
P1
PLS Regression Model
U
T

Figure 13. A Partial Least Squares Regression Performed on the P Descriptor and Y Response Variables (Gabrielsson, Lindberg , & Lundsteadt, 2002)

NIPALS is an iterative algorithm used to handle the development of these relationships. The method uses least squares regression for the mixture of related forms between P with Y and U with T. Loading weights are needed to maintain orthogonality. The general equation is shown in equation 2.11.

$$\hat{Y}_{nx\phi} = U_{nx1}W_{1x\phi} \qquad 2.11$$

where $\hat{Y}$ is the point estimate of the physical-chemical properties, U is a nx1 matrix of the latent variable scores and W is the 1xϕ matrix of the latent variable loadings. The latent variable representations of the attributes and property descriptors will have a lower dimensionality. Therefore, regressing the lower dimensional U_{nx1} attribute scores against the T_{nxm} property descriptor scores will result in Equation 2.11 where $B_{mx\phi}$ is set of $mx\phi$ regressors describing the latent attribute-property descriptor relationship.

$$U_{nx\phi} = T_{nxm}B_{mx\phi} \qquad 2.12$$

Iterative algorithms, like NIPALS, introduced above are described in detail in many sources including (Geladi & Kowalski, 1986), (Wold S. , 1995), (Macgregor & Muteki, 2007). It is generally considered standard practice to calibrate spectroscopic techniques with principal component regression and partial least squares because it works to avoid collinearity problems that often occur in multivariate regression (Jollieffe, 2002).

2.2. Process Modeling/Simulation

Process simulation seeks to optimize scenarios by computerized modelling of a process with the ability to modify the variables without the time and expense of laboratory testing. Despite advancement in computer technology, calculations related to optimization of processes can be quite cumbersome. Surrogate modelling is one method that can reduce computational burden. The input and output relationship of black box or complex relationships which only require simpler solutions can be statistically related through the use of surrogate models.

2.2.1. Surrogate Modeling

2.2.1.1. Artificial Neural Networks (ANN)

Artificial Neural Networks (ANN) were developed from studying the brain and the connections between neurons. Artificial neurons mimic biological neurons, which process signals, and weights mimic the synapses, which create the network between neurons. Biologically, those signal strengths are adaptive feedback loops. ANN training is the process of optimizing the weights and biases to find the lowest error between the outputs and the target data (Haykin et al, 2009).

2.2.1.2. Automated Learning of Algebraic Models for Optimization (ALAMO)

Automated Learning of Algebraic Models for Optimization (ALAMO) was developed to reduce surrogate model complexity while maintaining high accuracy. The approach is similar to polynomial regression and by extension, response surface methodology (RSM), because the final model is a summation of multiple basis functions. Unlike polynomial regression and RSM, ALAMO uses polynomial, multinomial, exponential, logarithmic, and trigonometric (sine and cosine) basis functions as is appropriate for the given data. Overfitting is generally not a problem, due to the nature of the ALAMO algorithm (Cozad et al, 2014).

2.2.1.3. Extreme Learning Machine (ELM)

A special type of single layer feedforward neural network (SLFN) is called an Extreme Learning Machine (ELM). In ELMs, the hidden layers' weights and biases are randomly assigned provided that the activation function is differentiable. The training of SLFN then becomes a linear equation system to solve. ELM has a few advantages over other methods because it does not require a learning rate, stopping criteria or validation, and no gradient descent learning methods for training. Those advantages are suggested to lead to faster training times with lower errors (Huang et al, 2015).

2.2.1.4. Support Vector Regression (SVR)

Support vector machines are another common method used to relate nonlinear input and output data (Jin et al, 2001). Support Vector Regression (SVR) is an application of support vector machines. This method transforms data (x_i, y_i) to an *m*-dimensional feature space and fits it to a linear model. The main advantage of SVR is its significantly faster training times compared to other models.

2.2.1.5. Radial Basis Function Networks (RBF)

Under the umbrella of ANNs lies the Radial Basis Function Networks (RBFs). The hidden layer in RBFs calculates the Euclidean distance between the input weights and the inputs, multiplies them with a bias vector, and passes them through the radial basis transfer function (Chen , Cowen, & Grant, 1991).

2.2.1.6. Gaussian Process Regression (GPR)

Developed as a machine learning technique, Gaussian Process Regression (GPR) is a surrogate model that generates the output as a linear combination of the inputs. A Gaussian process is a collection of random variables, a finite set of which have a joint Gaussian distribution (Mirbagheri, 2015). The random variables are considered to be the value of the function evaluated at *x*. An *a priori* distribution must first be assumed over the data. The parameters of the *a priori* distribution are a mean function $\mu(x)$ and a covariance function $K(x, x')$. The distribution is then updated with the training data to generate the posterior probability distribution.

2.2.1.7. Random Forests (RF)

Random Forests (RF) also stem from the machine learning field and have been found to be useful in many applications. RF models generate an ensemble of regression decision trees called forests using bootstrap aggregation and random feature selection to average the predictions (Breiman, 2001).

2.2.1.8. Multivariate Adaptive Regression Splines (MARS)

Multivariate Adaptive Regression Splines (MARS) was introduced by Jerome Friedman in 1991. The model is comprised of a linear summation of basis functions where the basis functions can be either spline functions or product of two or more spline functions. The model adds terms to intentionally cause overfitting, but then goes through a pruning process where the terms that contribute the least to the overall model are removed to generate the final surrogate model. One disadvantage to using MARS is that this model tends to perform poorly with small sample sizes (Freidman et al, 1991).

CHAPTER 3. Reverse Design of Ionic Liquids

3.1. Global Warming and CO_2 Emission Reduction

The “Greenhouse Effect” may seem like a modern buzz word used to describe our ever-warming environment, but it was first theorized by French mathematician and physicist, Jean-Baptiste Joseph Fourier. In 1827, Fourier published an article with the title translated to English as the temperature of the Earth and planetary spaces. He theorized that Earth’s atmosphere of gases was creating a warming insulation layer. Then in 1958, Charles David Keeling began studying the rapidly increasing levels of carbon dioxide (CO_2) in the atmosphere and recorded his findings on the Keeling curve. (EPA, 2015) Since this discovery, the world leaders have made recommendations about making changes to our processes and systems to reduce emissions. Scientists and politicians have been working to create awareness, encourage change and in some cases, require emission reduction. Newly constructed power plants are being regulated to operate more efficiently with lower emissions. However, the older facilities which have been allowed to remain in operation for decades are left without many options if they are forced to reduce emissions. In some cases, the CO_2 can be recycled based into the system to provide benefit and reduce the amount emitted in the flue gas. Generally, rebuilding a plant is not economically viable, so capture and sequestering of CO_2 may be the only option.

Traditional capture methods include separation using monoethanolamide (MEA)-based solvents. Amine based solvents have been widely used to treat acid gases, such as hydrogen sulfide and carbon dioxide, in gas streams since the beginning of the natural gas industry. (Hasib-ur-Rahman, Siaj, & Larachi, 2010) They are well tested and reliable.

However, amine-based solvents and other similar solvents can present an environmental and health risk to the area. Though these types are solvents have proven that they are effective solvents, their risk factors dictate that we look for viable options with a more benign impact.

3.2. Ionic Liquids

Ionic liquids are organic salts which are composed of a cation, anion and an alkyl chain, as shown in Figure 14.

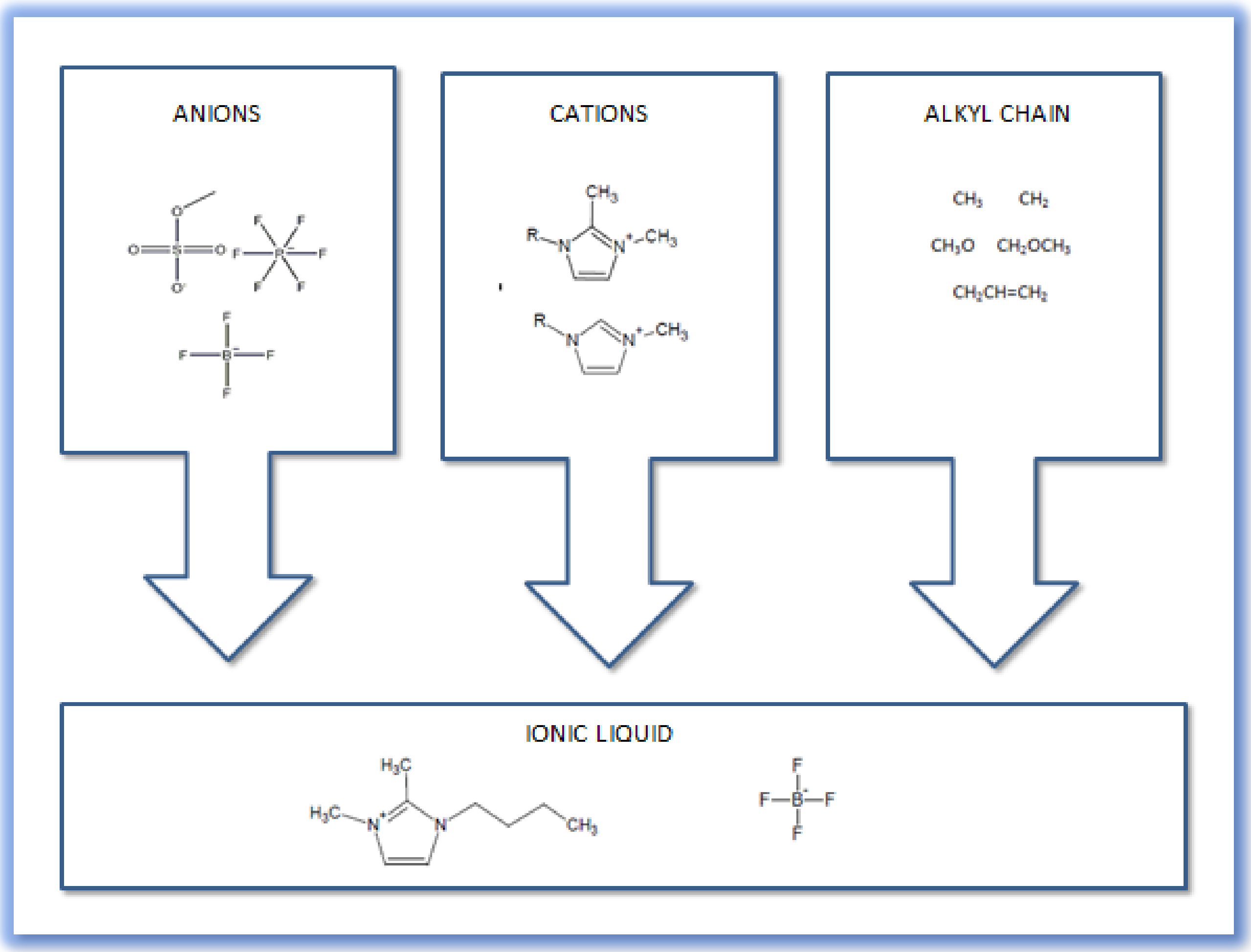


Figure 14. Building Blocks of Ionic Liquids

They can be advantageous from an environmental and health hazard perspective for many reasons. For example, low vapor pressures eliminate volatile organic carbon compounds (VOCs) which are closely monitored by the EPA and given off by most solvents and stability at high temperatures are just a couple of reasons ionic liquids may be valuable. (Holbrey & Seddon, 1999)

Ionic liquids have been used for centuries and known as molten salt; however more current research has been conducted since the 1980's. Though research has been ongoing for decades, there is still a relatively small amount of data compared to the number of possible ionic liquids (Ayala et al, 2006). Estimations have stated that there are over 10^{14} possible combinations of cation, anion and alkyl chain combinations that would be liquids at room temperature (Turner et al, 2003). The majority of these possibilities have not been synthesized.

One of the major benefits of ionic liquids is their unique capability to be customized by altering the combination of cation, anion and alkyl chain length to a specific situation including CO_2 removal and sequestration, but also energy storage, chemical separations and many other chemical processes. However, a method must be developed to more effectively determine the ideal ionic liquid, rather than the more conventional trial and error method where candidate ionic liquids are synthesized and tested for efficacy.

The goal of this work is to determine the most ideal candidate for CO_2 sequestration by calculating the highest Henry's law constant. A set of ionic liquids were selected from the IUPAC Ionic Liquids Database (Ionic Liquids Databased (ILThermo), 2014) for chemicals that had listed melting temperature and Henry's Law constant data. Table 1 lists the ionic liquids used in this work.

Ionic Liquid	**K_H**
3-(3-cyanopropyl)-1-methylimidazolium 1,1,1-trifluoro-N-[(trifluoromethyl)sulfonyl]methanesulfonamide	5,890
1-methyl-3-propylimidazolium hexafluorophosphate	5,200
1-methyl-3-octylimidazolium tetrafluoroborate	7,560
1-(3-cyanopropyl)-3-methylimidazolium cyanocyanamide	15,420
1-ethyl-2,3-dimethylimidazolium bis[(trifluoromethyl)sulfonyl]imide	6,050
1-butyl-2,3-dimethylimidazolium tetrafluoroborate	9,220
1-ethyl-3-methylimidazolium diethylphosphate	8,120
1-butyl-3-methylimidazolium hexafluorophosphate	7,280
1-butyl-2,3-dimethylimidazolium hexafluorophosphate	8,850
1,3-dimethylimidazolium dimethylphosphate	12,720
1-butyl-3-methylimidazolium bis[(trifluoromethyl)sulfonyl]imide	4,470
1-ethyl-3-methylimidazolium bis[(trifluoromethyl)sulfonyl]imide	4,630
3-(3-cyanopropyl)-1,2-dimethylimidazolium 1,1,1-trifluoro-N-[(trifluoromethyl)sulfonyl]methanesulfonamide	5,960
1-butyl-3-methylimidazolium tetrafluoroborate	9,160
1-ethyl-3-methylimidazolium dicyanamide	9,850
1-butyl-3-methylimidazolium hexadecanoate	7,550
1-butyl-3-methylimidazolium dibutylphosphate	5760
1-hexyl-3-methylimidazolium bis[(trifluoromethyl)sulfonyl]imide	3,500
1-methyl-3-propylimidazolium bis[(trifluoromethyl)sulfonyl]imide	3,700
1-ethyl-3-methylimidazolium trifluoromethanesulfonate	7,400
1-butyl-3-methylimidazolium octadecanoate	5,920

Table 1. Ionic Liquids Test Group and Associated Henry’s Law Constant in kPa @ 323K

3.3. Reverse Design of Ionic Liquids

A methodical approach to select the most effective combination of cation, anions and alkyl chains to meet the required physical and chemical property targets will be required to further the use of ionic liquids into mainstream industry. The characterization-based group contribution method (cGCM) builds on the theory that each group provides different contribution to the physical and chemical properties of the overall compound. This is particularly useful when researching ionic liquids because each cation, anion and alkyl chain can provide valuable design data. This method is combined with a reverse problem formulation to provide a method that can be tailored to a more diverse group of process needs. The following sections more specifically describe the developed method.

3.3.1. Infrared Spectroscopy Data

Infrared spectroscopy data (IR) contains information about each functional group within an ionic liquid to provide clues to its composition because different molecules absorb specific frequencies.

The group contribution method utilizes latent property parameters which are applied to the architecture of the overall molecule. For this reason, the initial IR data will determine the validity of the resulting solution. However, ample IR data has not been generated for the majority of ionic liquids. To remove this as a limiting factor, density functional theory was utilized to generate the needed IR data. Density functional theory is a molecular modelling method that can be used to predict the electronic structure of molecules.

To generate a starting point geometry for each ionic liquid pair, the structure was drawn in Avogadro (v1.1.1) (Hanwell, et al., 2012) and then optimized by minimization of the energy using Merck Molecular Force-Field (MMFF94). MMFF94 has been parameterized for a broad spectrum of organic chemicals including many charged ions (Halgren, 1996). Ionic liquids with phosphate anions have shown to be ill-suited for MMFF94. Those ionic liquids were optimized universal force field (UFF) (Linusson, Gottfires, Lingren, & Wold, 2000). Each optimization algorithm with 10,000 iterations with a convergence of 10^{-7} was rerun until a local energy minimum was reached. To provide a visual example of these calculations, a rendering of 1-butyl-2,3-dimethylimidazolium hexafluorophosphate is depicted in Figure 15.

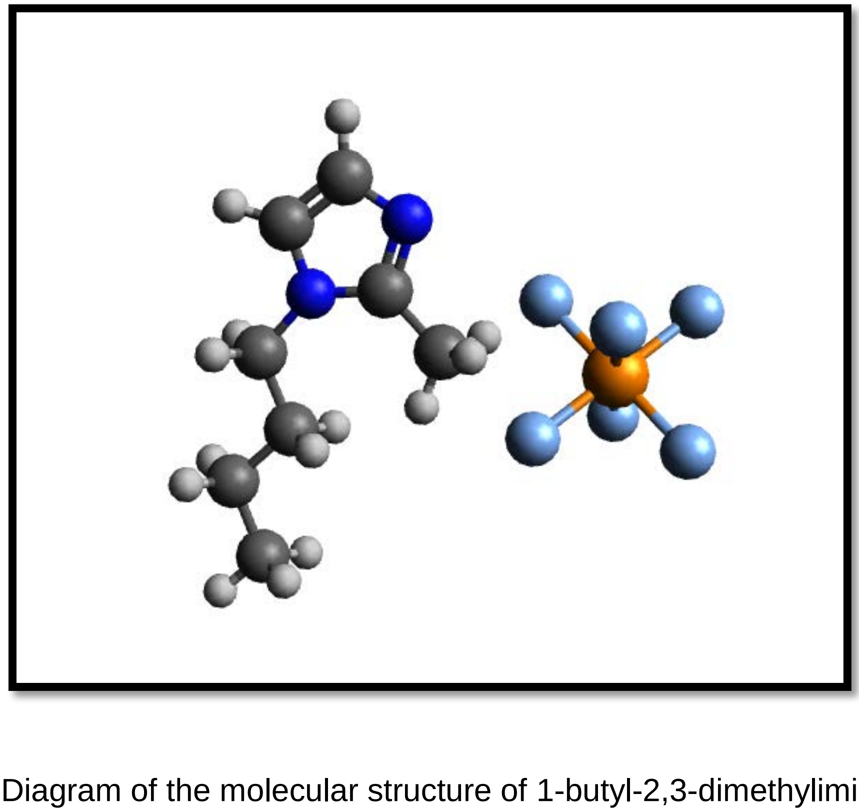


Figure 15. Diagram of the molecular structure of 1-butyl-2,3-dimethylimidazolium hexafluorophosphate

An input file requesting frequency optimization for Gaussian 09 was generated by the utility within Avogadro. The Gaussian input requires specification of the specific density functional theory method and B3-LYP with 6-31G(d) was found to the most applicable. B3-LYP, HF, MP2, and QCISD were compared and the lowest root-mean-square error for simulated fundamental molecular vibrations was found to be B3-LYP with 6-31G(d). (Scott et al, 1996) Additionally, experimental data was compared to estimated values for 1-ethyl-3-methyl imidazolium hexafluorophosphate and strong agreement was found. (Hada et al, 2015)

The files were then input into Gaussian 09 running through the Alabama Supercomputing Authority supercomputing cluster. Gaussview was then utilized to generate an IR spectrum graph which was digitized to create a table of the resulting data for each ionic liquid, an example of which is shown in Figure 16.

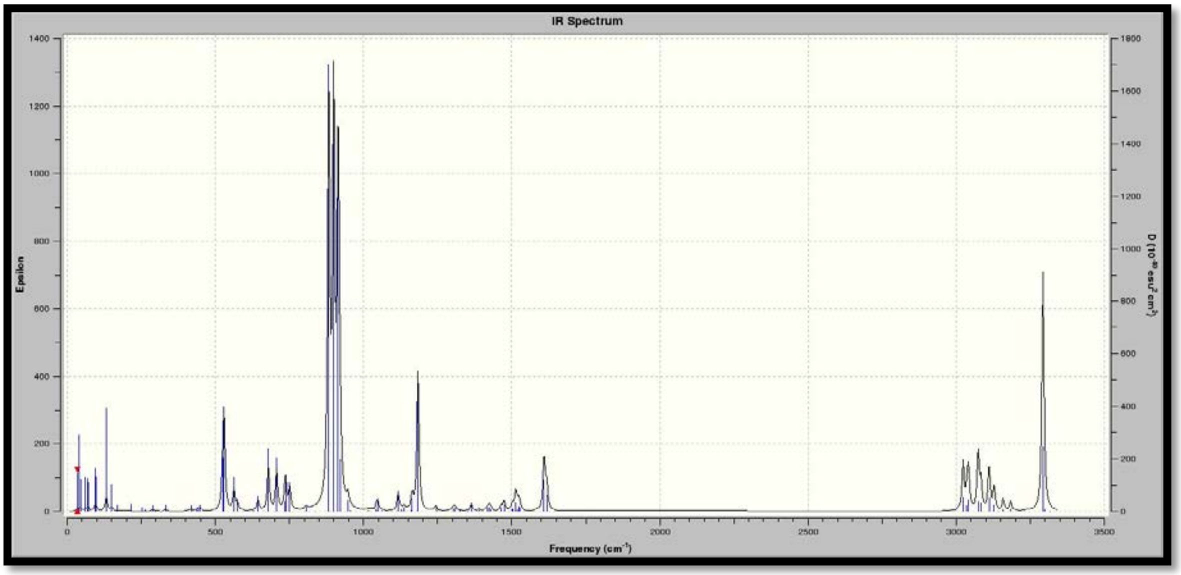


Figure 16. Generated IR Spectrum

3.3.2. Principal Component Analysis & Partial Least Squares Regression

The resulting data carries a wide range of complicated information but will need to be reduced to only the most important. Principal component analysis (PCA) seeks to identify patterns and important features in large quantities of information (Eriksson et al, 2006). This method was combined with partial least squares (PLS) to predict the structure property relationships by reducing the dimensions of the data to the most significant. JMP 11 (V11.0.0) was used to perform the analysis. Generally, three principal components will provide the majority of the variance. In this work, the first 3 components provided 63% of the data and a plot of the eigenvalue vs. number of components leads us to select the first three principal components, as shown in Figure 17.

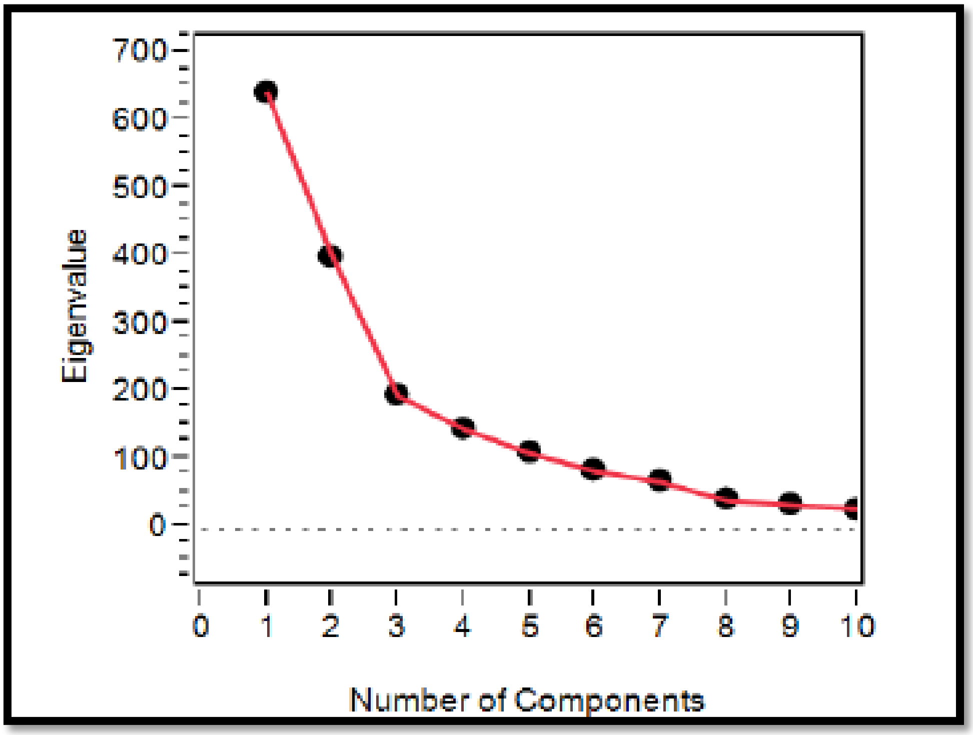


Figure 17. Scree Plot used in PCA

Loadings, *P*, describe the magnitude of correlation and the direction either positive or negative and how they contribute to the scores, *t*. Model coefficients were also determined using JMP and PCR along with qualitative structure property relationship (QSPR) models for melting temperature to ensure that the resulting molecules were liquid at room temperature and Henry's law constant shown in equations 3.1 and 3.2 which can be correlated with solubility (Duchowicz, Garro, & Castro, 2008).

$$e^{\left({}^{T_m}/_{T_{mo}}\right)} = \beta_0 + \sum_{i=1}^{3} \beta_0 t_i + \sum_{i<j}^{3} \sum_{j<i}^{3} \beta_{ij} t_i t_j + \sum_{i=1}^{3} \beta_{ii} t_i^2 \qquad 3.1$$

$$ln(K_H) = \beta_0 + \sum_{i=1}^{3} \beta_i t_i + \sum_{i=1}^{3} \beta_{ii} t_i^2 \qquad 3.2$$

3.3.3. Reverse Design of Ionic Liquids using QSPR and cGCM

While the number of possible ionic liquid combinations is great, the number of functional groups which serve as the building blocks are more manageable. The cation, anions and alkyl chains which were chosen to represent the ionic liquids in the test group are shown in Table 2.

Name	**Building Block Type**
bis[(trifluoromethyl)sulfonyl]amide	Anion
Dibutylphosphate	Anion
Dicyanamide	Anion
Diethylphosphate	Anion
Dimethylphosphate	Anion
Hexadecanoate	Anion
Hexafluorophosphate	Anion
Octadecanoate	Anion
Tetrafluoroborate	Anion
1,3-dimethylimidazolium	Cation
1-butyl-2,3-dimethylimidazolium	Cation
1-butyl-3-methylimidazolium	Cation

1-ethyl-2,3-dimethylimidazolium	Cation
1-hexyl-3-methylimidazolium	Cation
1-methyl-3-octylimidazolium	Cation
1-methyl-3-propylimidazolium	Cation
Cyanopropyl-3-methylimidazolium	Cation
Methyl	Alkyl Groups
Methylene	Alkyl Groups

Table 2. Ionic Liquid Building Blocks from Test Group

The reverse design of the ideal ionic liquid included an exhaustive search using an algorithm in the Python programming language (Hada S. , 2013) to generate all the possible ionic liquid possibilities that could be created using the above-mentioned anion, cation, and alkyl chain building blocks. For each ionic liquid, the melting point was calculated to determine if the resulting combination was liquid at room temperature. If the melting temperature was less than 298K, then the Henry's law constant was calculated. The program then repeated until all possible combinations were considered. The program then returned the top three results.

3.4. Results

The highest value calculated was for dimethylimidazolium dimethylphospahte at 13,660 kPa, methylimidazolium bis(trifluoromethylsulfonly)-amide at 12,340 kPa, and methylimidazolium tetrafluoroborate at 9,280 kPa.

While laboratory research is limited in this field, several studies have measured the solubility of CO_2 in room temperature ionic liquids. The results of a study which compared the solubility of CO_2 in nine ionic liquids using a quartz crystal at 298 K and 1 bar show that the imidazolium based ionic liquids have the highest solubility at these conditions (Baltus et al, 2004). As a result of several studies which showed a data supporting the theory that imidazolium based ionic liquids were the most effective solvents, one literature study was written which gathered a wide range of measured CO_2 in ionic liquid solubility data. The results pointed to methylimidazolium based ionic liquids for higher solubility (Candena et al, 2004). These papers support the results of the calculations performed in this work which also found that the highest three molecules were methylimidazolium based.

3.5. Conclusions

The existing laboratory data related to ionic liquids is inadequate and can be a limiting factor for researchers seeking new applications for ionic liquids. This information gap exists because it would simply be cost prohibitive to generate data for all the possible ionic liquids and study applications based on a traditional trial and error method. Researchers can utilize the computer-aided molecular design (CAMD) framework with reverse design methods and density functional theory to generate infrared spectroscopy data. This can serve as a platform for future research that will span many sectors of academia and can easily be extended to the industrial sector.

CHAPTER 4. Surrogate Modeling Optimization

Process simulation is one method that seeks to optimize scenarios by computerized modelling of a process with the ability to modify the variables without the time and expense of laboratory testing. Despite advancement in computer technology, calculations related to optimization of processes can be quite cumbersome. Surrogate modelling is one method that can reduce computational burden (Beck, 2015). The input and output relationship of black box or complex relationships which only require simpler solutions can be statistically related through the use of surrogate models. Many studies have been sought to understand the applications of various surrogate models (Dife & Diwekar, 2016); however, this case study seeks to provide guidance about the optimal application of surrogate models. This work has been published in Computer Aided Chemical Engineering 40 (Davis, Cremaschi, & Eden, 2017) and Computer Aided Chemical Engineering 44 (Davis, Cremaschi, & Eden, 2018).

4.1. Surrogate Models Compared in This Study

Surrogate models are a common optimization method employed to model a process by statistically relating the inputs with the outputs. These models can be used to optimize black box processes or complex relationship that can be modelled with a simpler relationship when reasonable accuracy is desired.

4.1.1. Artificial Neural Networks

Artificial Neural Networks were inspired by the biological neural networks of the brain. Just as a child learns to relate objects that they see to words, artificial neural networks progressively improve performance. Each encounter, or calculation, the network develops an identity to the information which will be used in later steps (Haykin, 2009). In this calculation schema, biological neurons are replaced with artificial neurons called nodes. Those nodes are connected by a set of weights which are the counterpart of the biological synapse. Learning is the process of optimizing the weights and biases to find the lowest error between the outputs and target data. Artificial Neural Networks are useful because they can be applied to many applications and is an approachable method (Akkoyunlu, 2010). As with biological learning, the learning process can take more computing time as compared to other models.

4.1.2. Automated Learning of Algebraic Models for Optimization

Automated Learning of Algebraic Models for Optimization utilizes a polynomial type regression to provide accurate results with a simpler model (Cozad & Miller, 2014). Polynomial, multinomial, exponential, logarithmic and trigonometric functions are chosen based on the given data set. Different variations of the regression are performed and an Akaike information criterion is calculated. The model is presented when the Akaike information criterion either decreases or stabilizes. Automated Learning of Algebraic Models for Optimization has the real advantage of creating a model of a real system that is simple while still providing increased accuracy over other models. However, this accuracy can increase the computational time requirements.

4.1.3. Extreme Learning Machines

Extreme learning machines are a type of single layer feedforward neural network, so the basis is similar to artificial neural networks. However, the primary difference between extreme learning machines and traditional artificial neural networks is the learning scheme. Rather than using weights in a differential equation, extreme learning machines transforms the function into a linear equation utilizing a least squares solution to determine the training error (Huang, Zhu, & Siew, 2015). The reduction in complication provides a system that can be solved in a fraction of the time without reduction of accuracy. Improved computation time provides solutions quickly, so it may be more useful as an industry computation.

4.1.4. Support Vector Regression

Support vector regression utilizes learning algorithms that classify data by mapping them as points in space through a kernel process. This classification of data seeks to separate the data into groups and the details gathered are used to further classify data in a process in which a hyperplane that divides the group is a solution function. Support vector regression is often a preferred because the resulting solution will be based on global minima rather than local minima that other surrogate models may provide. Additionally, there is no risk of overfitting due to the nature of the solver. (Jin, Chen, & Simpson, 2001)

4.1.5. Radial Basis Function Networks

Radial basis function networks fall under the umbrella of artificial neural networks and are executed as linear model regressions where the function becomes the summation of weighted basis functions. Radial basis functions are developed with a distance criterion with respect to a central point. One advantage of radial basis function networks is their ability find the global minima without discovery of a local minima. (Chen , Cowen, & Grant, 1991)

4.1.6. Gaussian Process Regression

Gaussian process regression utilizes a kernel method similar to support vector regression. However, the essential difference is that gaussian process regression is based on the estimation of probabilities (Mirbagheri, 2015). Those estimations are then used to predict the graphical representation of the data. This method has the same advantages as the support vector regression related to the minima and overfitting.

4.1.7. Random Forests

Random forests are based on the implementation of decision trees. This method utilizes the kernel method to remove trees to avoid overfitting the model. It can be used for regression and classification solutions. Random forests can provide applications for many situations, so the implementation is a good option. (Breiman, 2001)

4.1.8. Multivariate Adaptive Regression Splines

Multivariate adaptive regression splines is comprised of a linear summation of basis functions. The model adds terms to intentionally cause overfitting, but then goes through a pruning process where the terms that contribute the least to the overall model are removed to generate the final surrogate model. It is a preferred method because the model training process is efficient (Freidman, 1991). However, it has been found in studies that the model requires a larger data set to perform accurately.

4.2. Challenge Functions

To evaluate the efficacy of each surrogate model, challenge functions were used to calculate the functions and data sets. Those functions were divided by surface plot shape and number of inputs. The shapes reviewed included: multi-local minima, bowl, plate, valley and ridges/drops; and the numbers of inputs were two, three, four, five and ten. Thirty-Four functions within the optimization group were applied to this study. A summary of these groups is shown in Figure 18.

Challenge Function	Number of Inputs	Surface Shape
ackley	2	Multilocal Minima
bukin	2	Multilocal Minima
crossit	2	Multilocal Minima
drop	2	Multilocal Minima
egg	2	Multilocal Minima
holder	2	Multilocal Minima
langer	2	Multilocal Minima
levy13	2	Multilocal Minima
schaffer2	2	Multilocal Minima
schaffer4	2	Multilocal Minima
shubert	2	Multilocal Minima
levy	3	Multilocal Minima
greiwank	10	Multilocal Minima
rastr	10	Multilocal Minima
schwef	10	Multilocal Minima
booth	2	Plate
matya	2	Plate
mccorm	2	Plate
powersum	10	Plate
zakharow	10	Plate
boha1	2	Bowl
spheref	3	Bowl
sumsqu	3	Bowl
rothyp	2	Bowl
trid	5	Bowl
perm0db	10	Bowl
sumpow	4	Bowl
camel3	2	Valley
camel6	2	Valley
dixonpr	3	Valley
rosen	3	Valley
dejong5	2	Ridges & Drops
easom	2	Ridges & Drops
michal	5	Ridges & Drops

Figure 18. Challenge Functions with shape and number of inputs

These functions were obtained from Virtual Library of Simulation Experiments (Surjanovic, 2015) which is a resource created at Simon Fraser University. The goal of that work was to provide a tool to evaluate simulation methods. This resource provides the equation, a graphical representation of the function and a MATLAB implementation which is available for use. The following sections provides more information about the functions including the number of inputs, shape, equation and a graphical depiction of the surface.

4.2.1. Challenge Functions

4.2.1.1. Ackley Function has a shape of Multilocal Minima and has two inputs.

$$f(x) = -a * exp\left(-b\sqrt{\frac{1}{d}\sum_{i=1}^{d} x_i^2}\right) - \exp\left(\frac{1}{d}\sum_{i=1}^{d}\cos cx_i\right) + a + \exp(1)$$

Variable value recommendations when d = 2: a = 20, b = 0.2, and c = 2π.

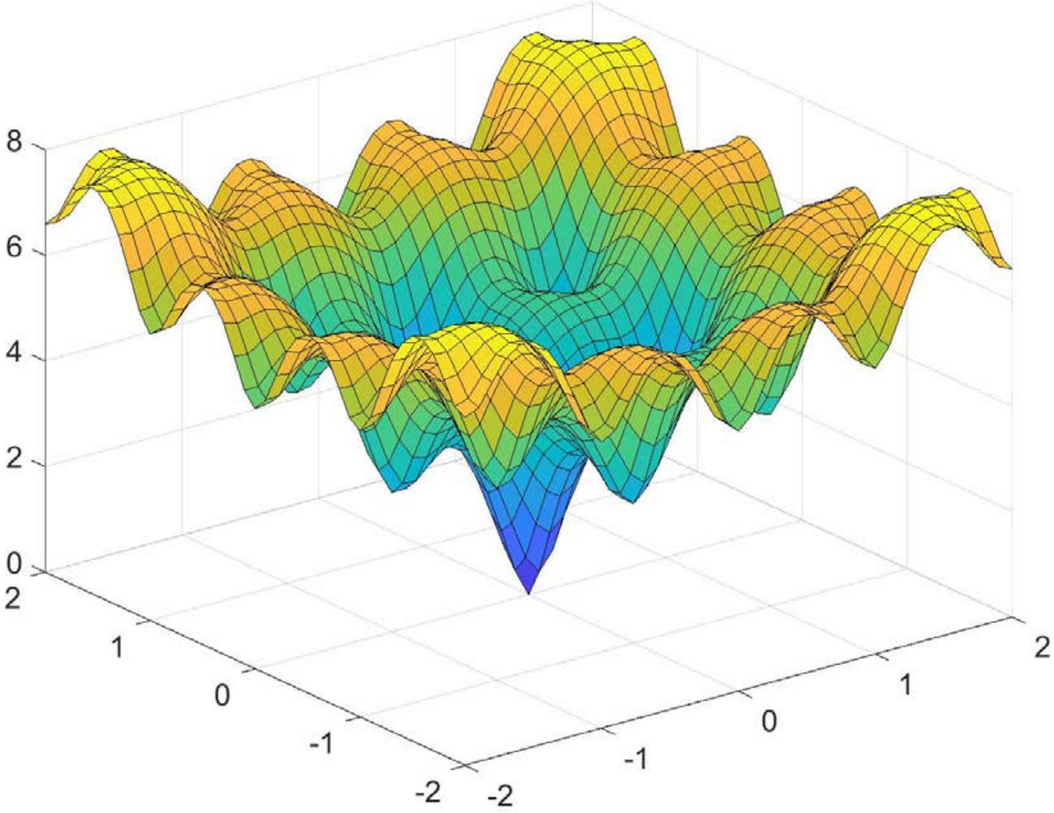


Figure 19. Surface of the Challenge Function: Ackley

4.2.1.2. Bukin Function N.6 has a shape of Multilocal Minima and has two inputs.

$$f(x) = 100\sqrt{|x_2 - 100x_1^2|} + 0.01|x_1 + 10|$$

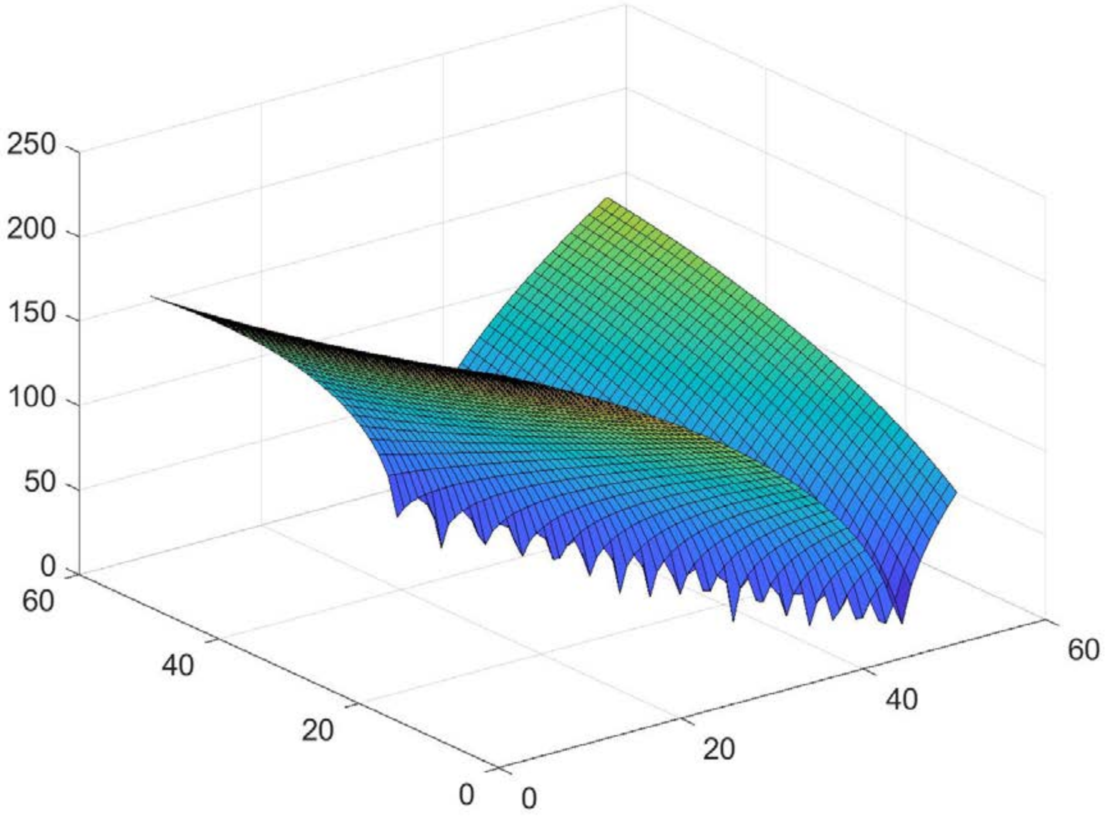


Figure 20. Surface of the Challenge Function: Bukin Function N.6

4.2.1.3. Cross-in-Tray Function has a shape of Multilocal Minima and has two inputs.

$$f(x) = -0.0001\left(\left|\sin(x_1)\sin(x_2)exp\left(\left|100 - \frac{\sqrt{x_1^2 + x_2^2}}{\pi}\right|\right)\right| + 1\right)^{0.1}$$

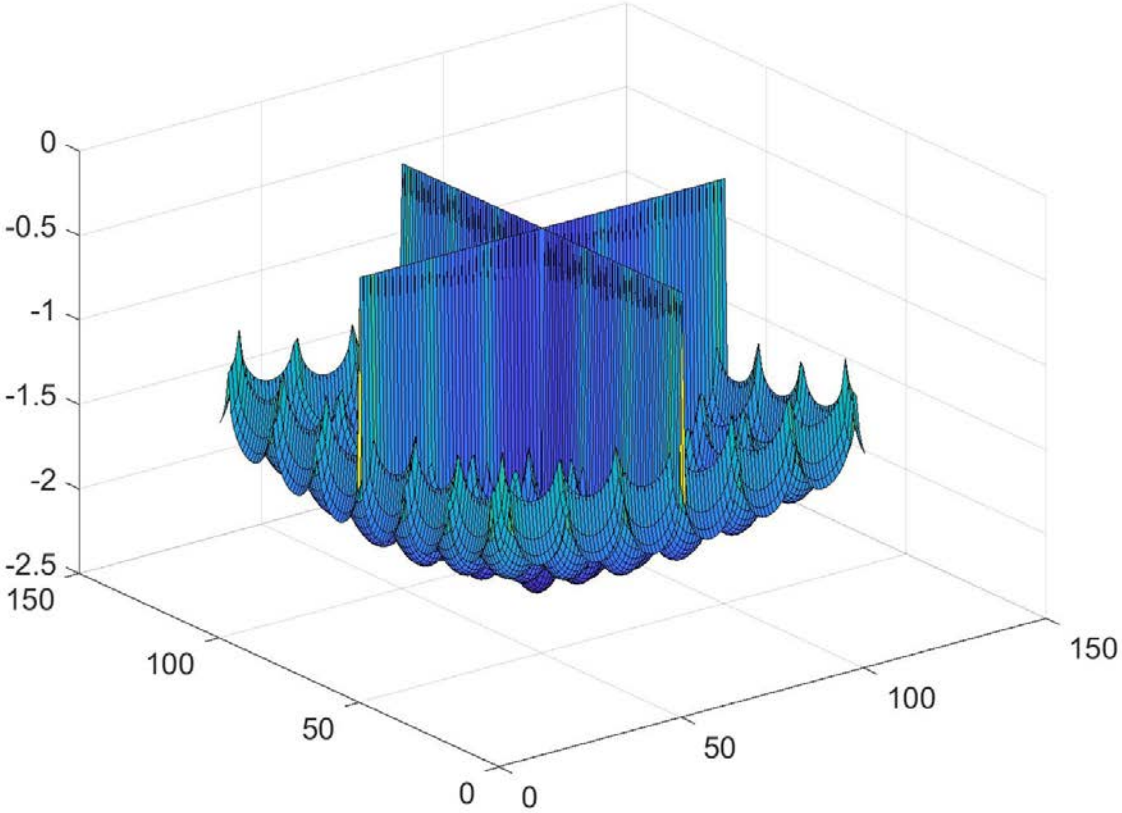


Figure 21. Surface of the Challenge Function: Cross-in-Tray Function

4.2.1.4. Drop-Wave Function has a shape of Multilocal Minima and has two inputs.

$$f(x) = \frac{1 + \cos\left(12\sqrt{x_1^2 + x_2^2}\right)}{0.5(x_1^2 + x_2^2) + 2}$$

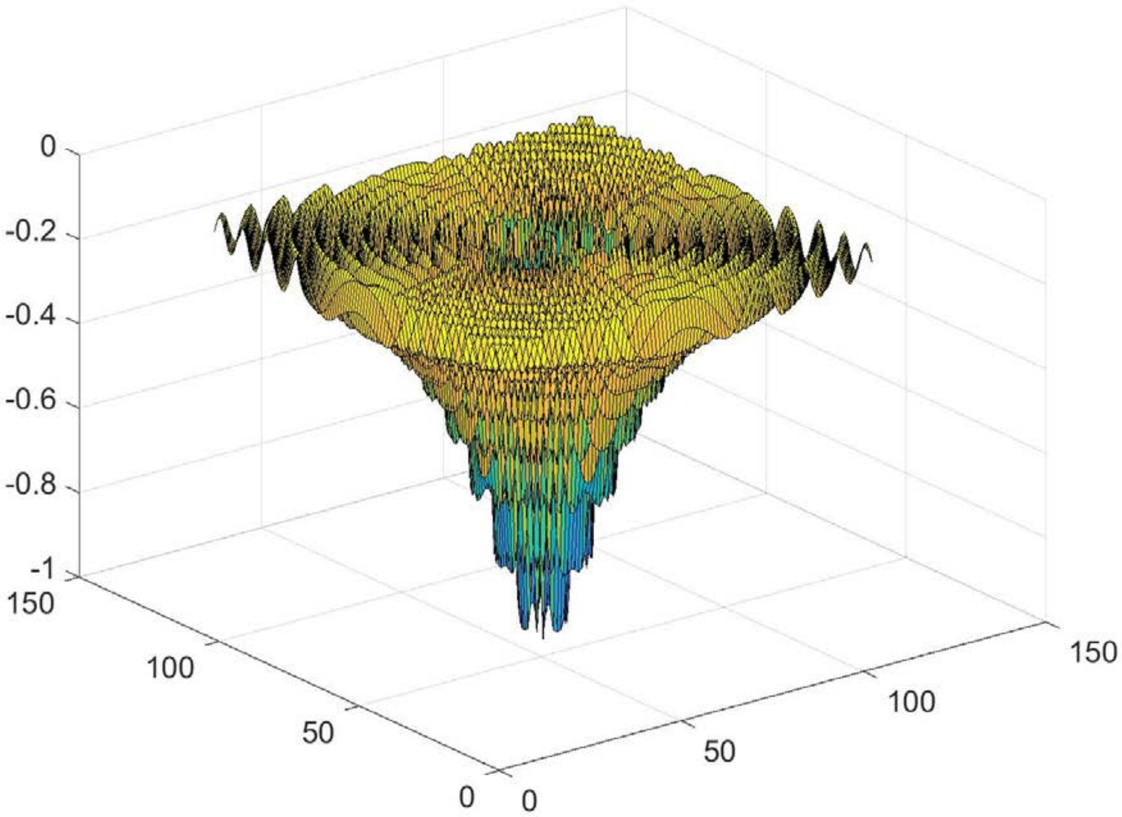


Figure 22. Surface of the Challenge Function: Drop-Wave Function

4.2.1.5. Eggholder Function has a shape of Multilocal Minima and has two inputs.

$$f(x) = -(x_2 + 47)\sin\left(\sqrt{\left|x_2 + \frac{x_1}{2} + 47\right|}\right) - x_1 \sin\left(\sqrt{|x_1 - (x_2 + 47)|}\right)$$

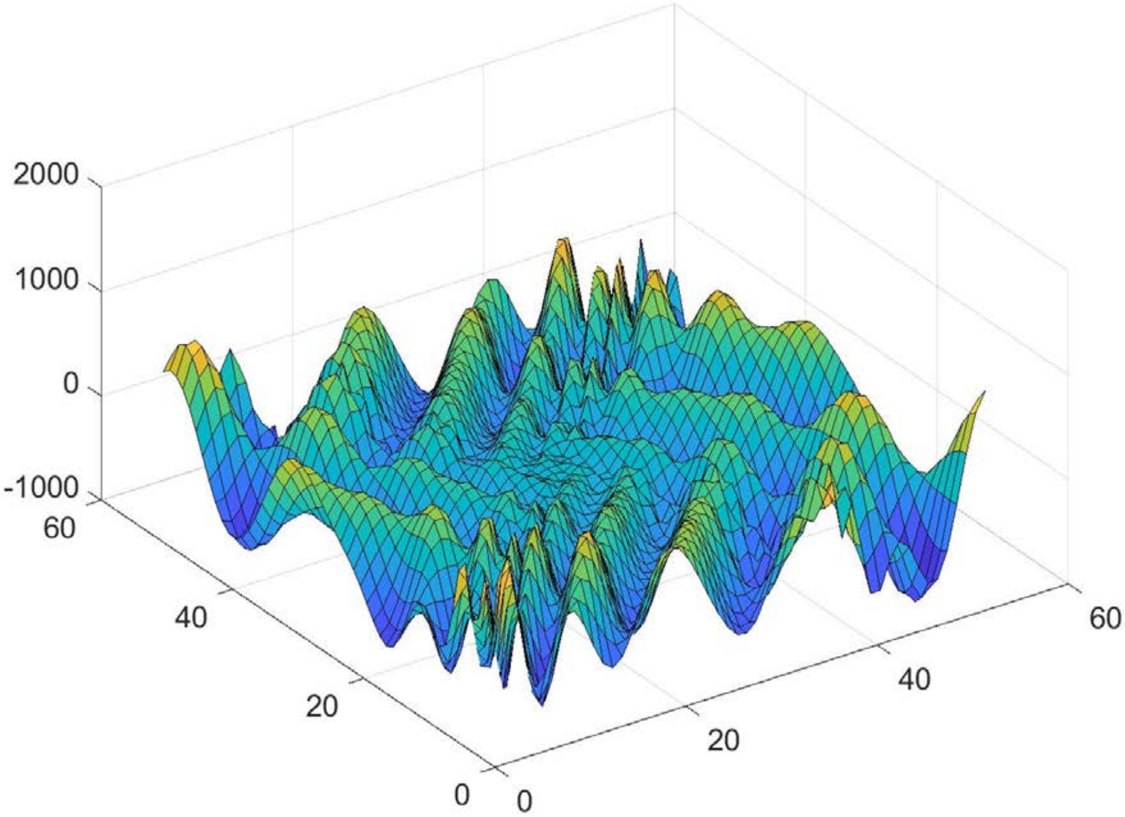


Figure 23. Surface of the Challenge Function: Eggholder Function

4.2.1.6. Holder Table Function has a shape of Multilocal Minima and has two inputs.

$$f(x) = -\left|\sin(x_1)\cos(x_2)exp\left(\left|1 - \frac{\sqrt{x_1^2 + x_2^2}}{\pi}\right|\right)\right|$$

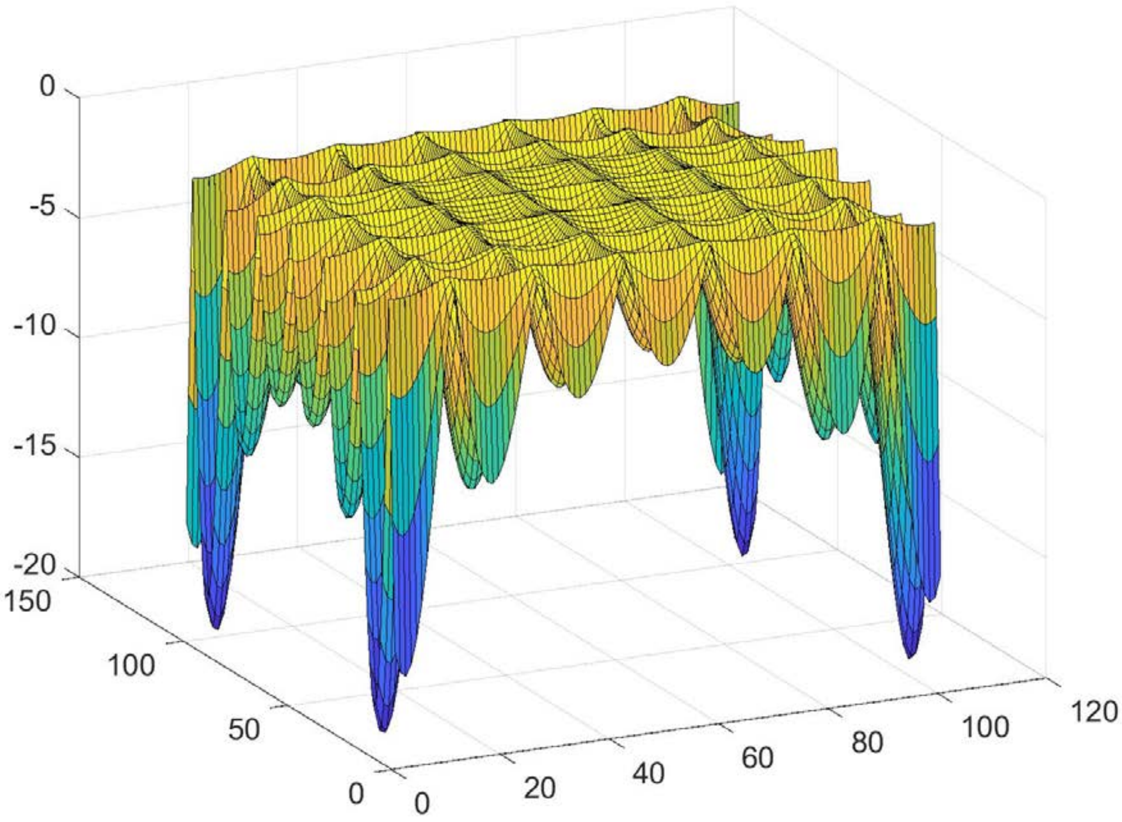


Figure 24. Surface of the Challenge Function: Holder Table Function

4.2.1.7. Langermann Function has a shape of Multilocal Minima and has two inputs.

$$f(x) = \sum_{i=1}^{m} c_i * exp\left(-\frac{1}{\pi}\sum_{j=1}^{d}\left(x_j - A_{ij}\right)^2\right) * \cos\left(\pi\sum_{j=1}^{d}\left(x_j - A_{ij}\right)^2\right)$$

Variable value recommendations when d = 2:

$$m = 5; c = (1, 2, 5, 2, 3)\ and\ A = \begin{bmatrix} 3 & 5 \\ 5 & 2 \\ 2 & 1 \\ 1 & 4 \\ 7 & 9 \end{bmatrix}$$

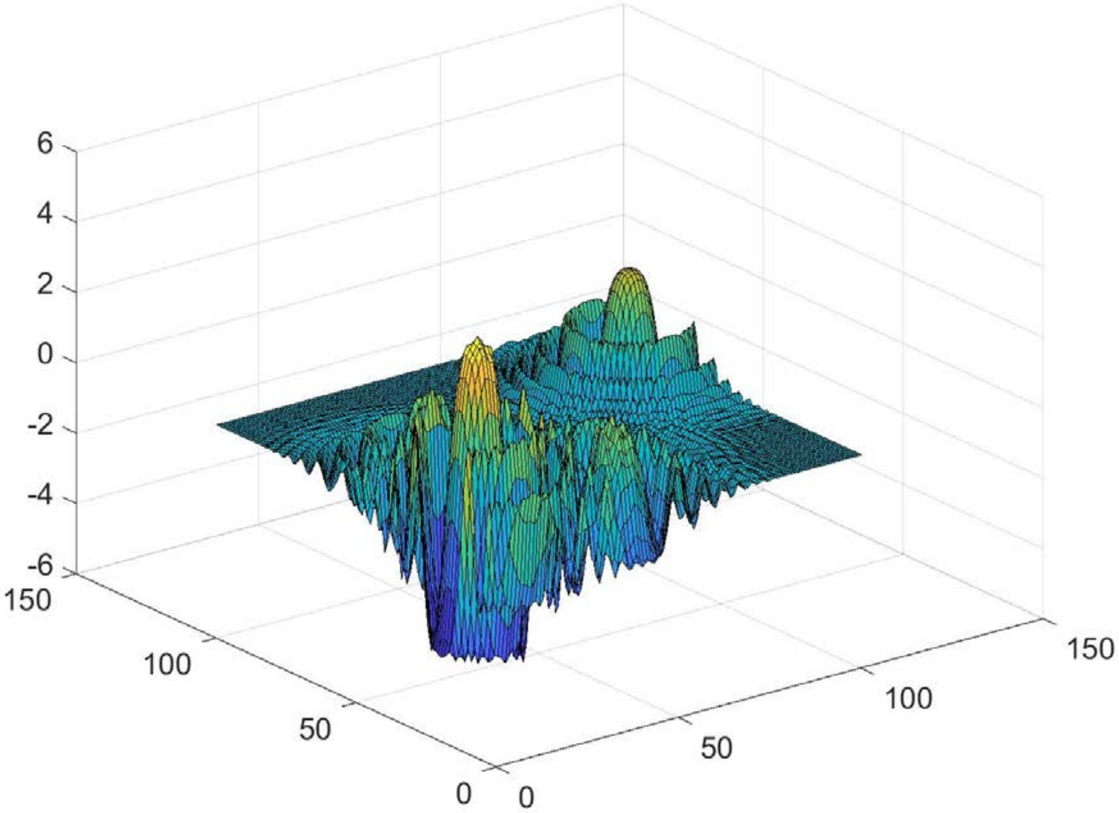


Figure 25. Surface of the Challenge Function: Langermann Function

4.2.1.8. Levy Function N.13 has a shape of Multilocal Minima and has two inputs.

$$f(x) = \sin^2(3\pi x_1) + (x_1 - 1)^2[1 + \sin^2(3\pi x_2)] + (x_1 - 1)^2[1 + \sin^2(2\pi x_2)]$$

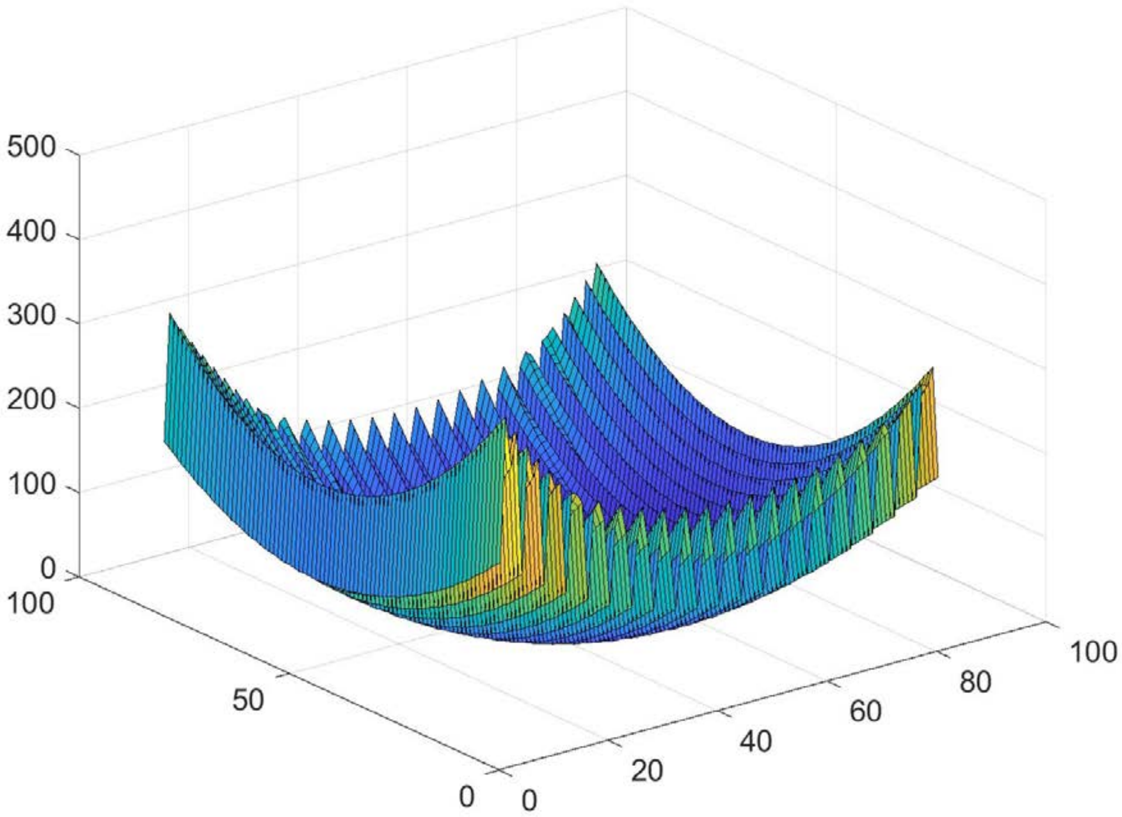


Figure 26. Surface of the Challenge Function: Levy Function N. 13

4.2.1.9. Schaffer Function N.2 has a shape of Multilocal Minima and has two inputs.

$$f(x) = 0.5 + \frac{\sin^2(x_1^2 - x_2^2) - 0.5}{[1 + 0.001(x_1^2 + x_2^2)]^2}$$

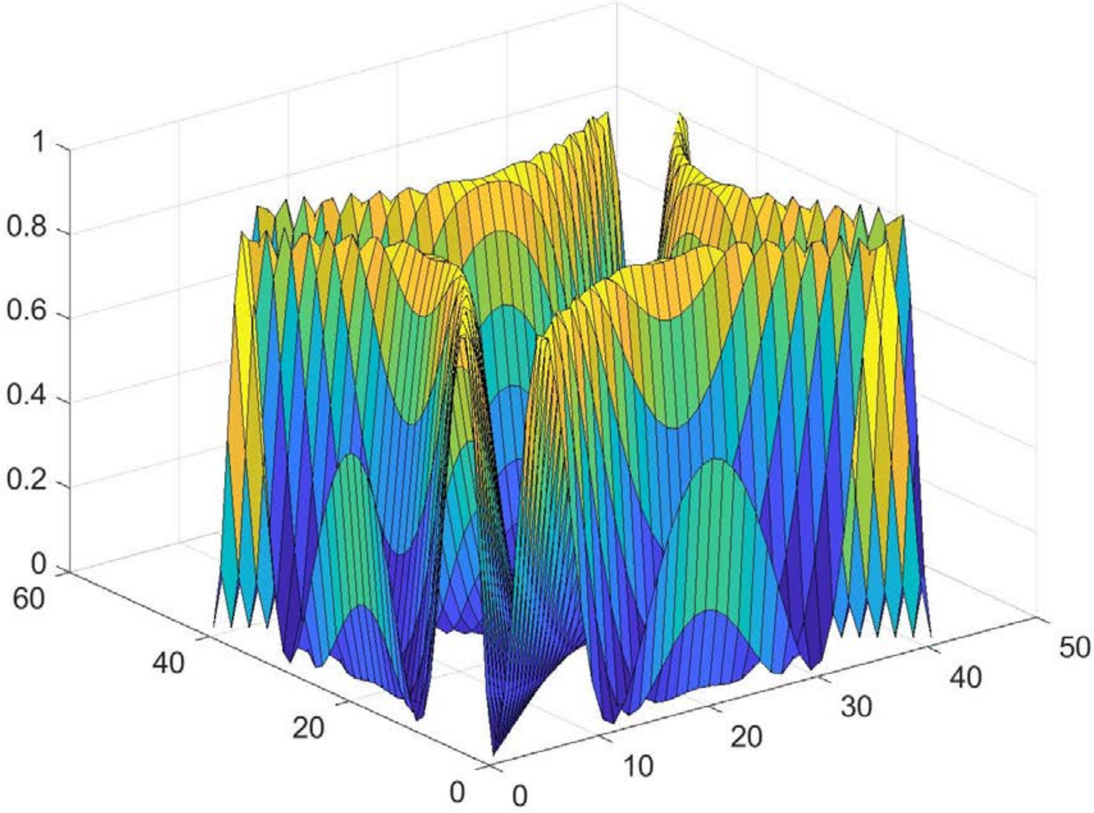


Figure 27. Surface of the Challenge Function: Schaffer Function N. 2

4.2.1.10. Schaffer Function N. 4 has a shape of Multilocal Minima and has two inputs.

$$f(x) = 0.5 + \frac{\cos(\sin(|x_1^2 - x_2^2|)) - 0.5}{[1 + 0.001(x_1^2 + x_2^2)]^2}$$

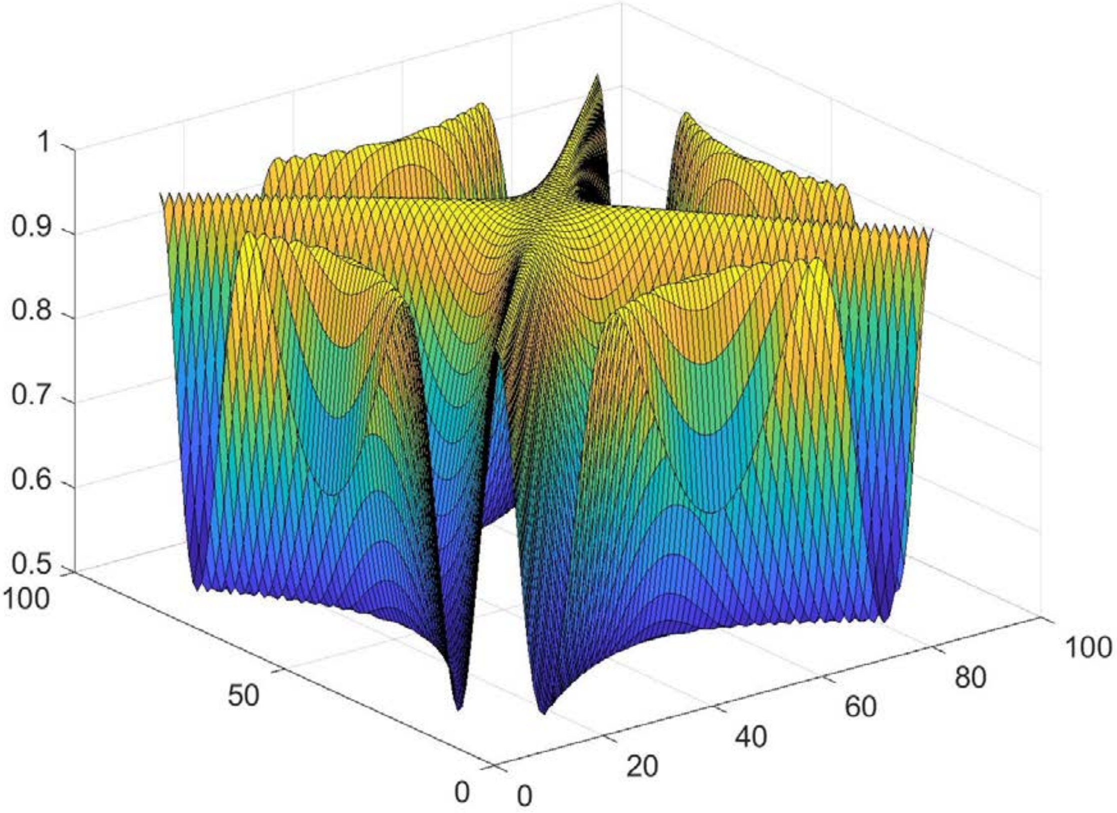


Figure 28. Surface of the Challenge Function: Schaffer Function N. 4

4.2.1.11. Shubert Function has a shape of Multilocal Minima and has two inputs.

$$f(x) = \left(\sum_{i=1}^{5} i\cos\big((i+1)x_1 + i\big)\right)\left(\sum_{i=1}^{5} i\cos\big((i+1)x_2 + i\big)\right)$$

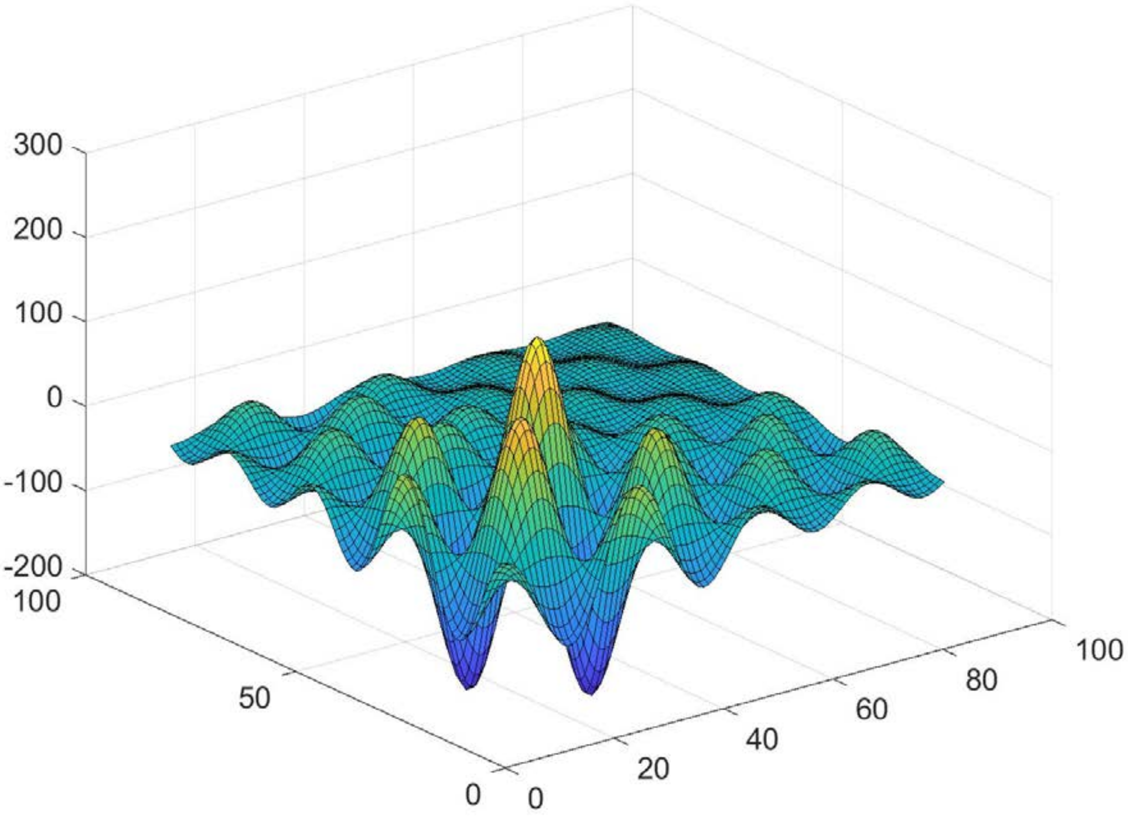


Figure 29. Surface of the Challenge Function: Shubert Function

4.2.1.12. Levy Function has a shape of Multilocal Minima and has three inputs.

$$f(x) = \sin^2(\pi w_1) + \sum_{i=1}^{3}(w_i - 1)^2[1 + 10\sin^2(\pi w_i + 1)] + (w_3 - 1)^2[1 + \sin^2(2\pi w_3)]$$

$$Where: w_i = 1 + \frac{x_i - 1}{4}$$

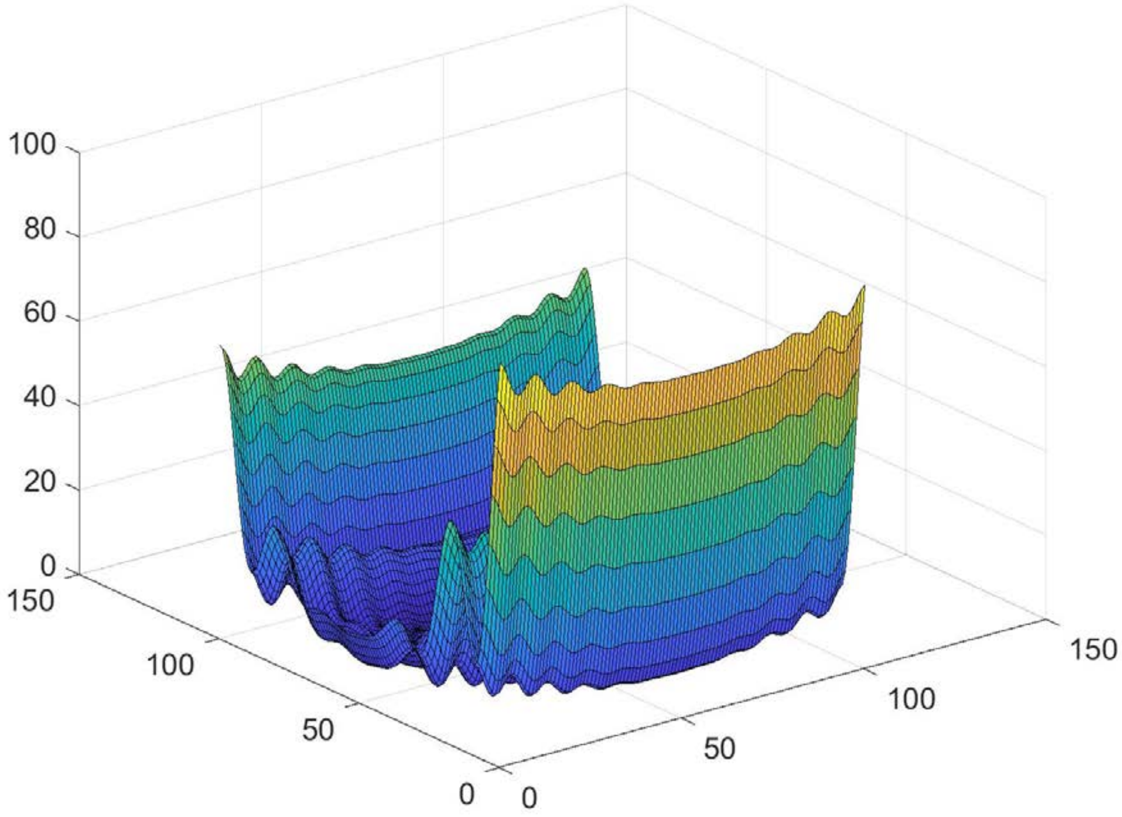


Figure 30. Surface of the Challenge Function: Levy Function

4.2.1.13. Griewank Function has a shape of Multilocal Minima and has ten inputs.

$$f(x) = \sum_{i=1}^{10} \frac{x_i^2}{4000} - \prod_{i=1}^{10} \cos\left(\frac{x_i}{\sqrt{i}}\right) + 1$$

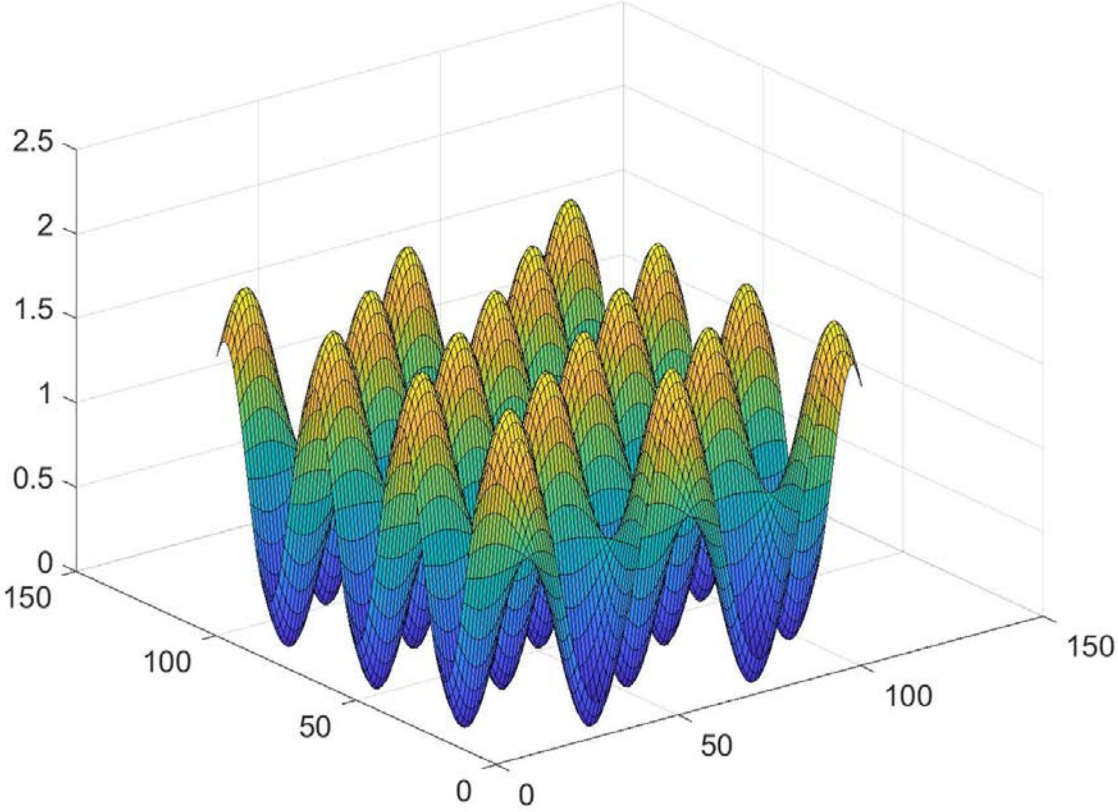


Figure 31. Surface of the Challenge Function: Grenwank Function

4.2.1.14. Rastrigin Function has a shape of Multilocal Minima and has ten inputs.

$$f(x) = 100 + \sum_{i=1}^{10} [x_i^2 - 10\cos(2\pi x_i)]$$

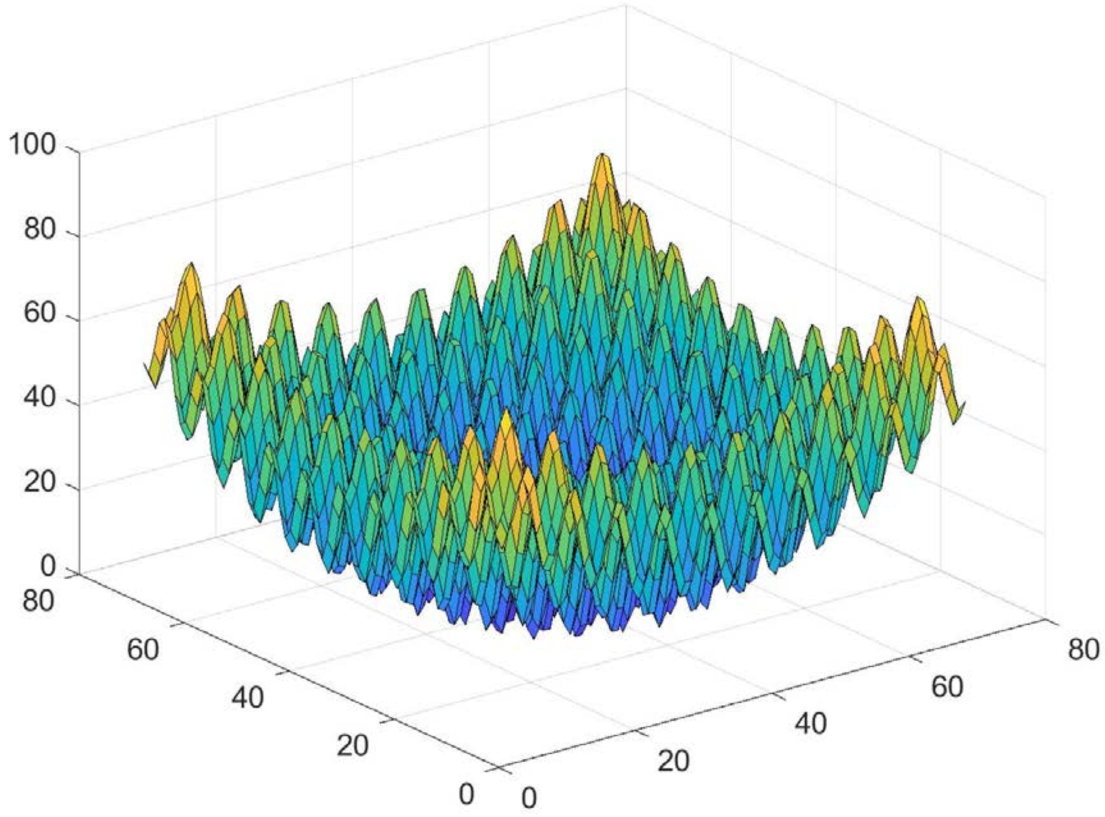


Figure 32. *S*urface of the Challenge Function: Rastrigin Function

4.2.1.15. Schwefel Function has a shape of Multilocal Minima and has ten inputs.

$$f(x) = 4189.829 - \sum_{i=1}^{10} x_i \sin\left(\sqrt{|x_i|}\right)$$

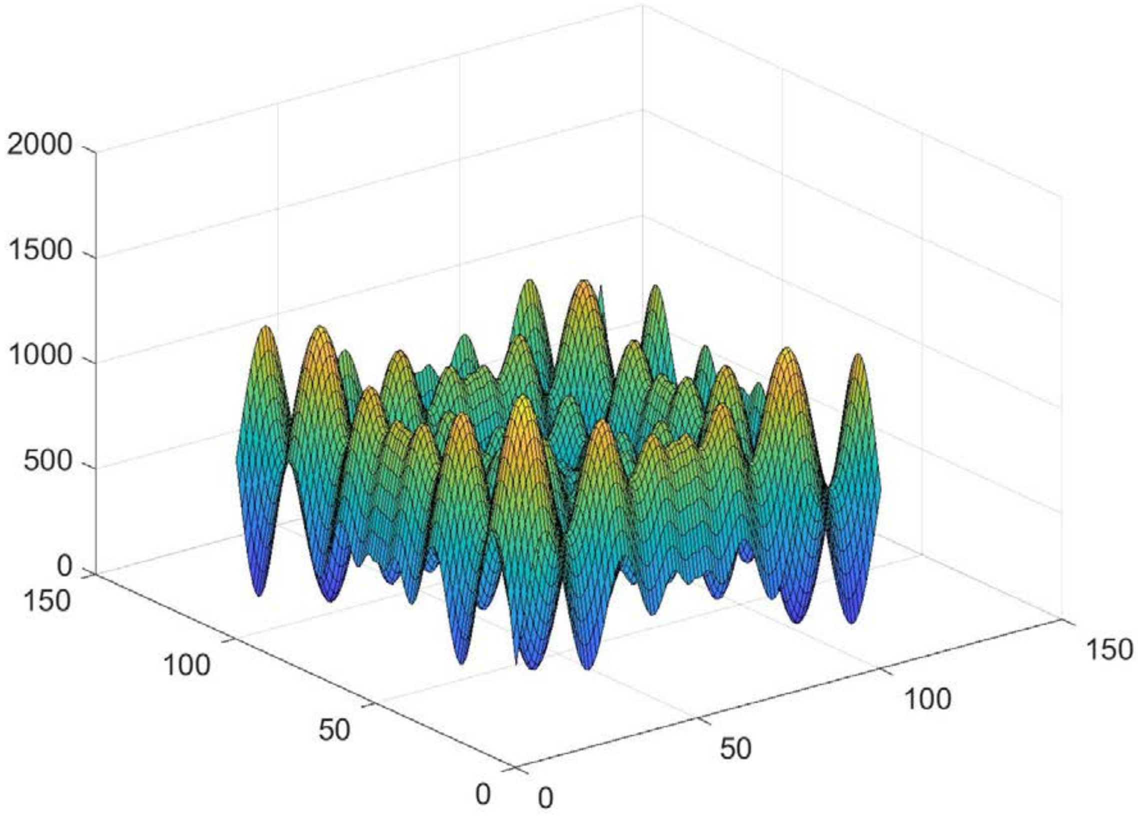


Figure 33. Surface of the Challenge Function: Schwefel Function

4.2.1.16. Bohachevsky Function has a shape of bowl and has two inputs.

$$f(x) = x_1^2 + 2x_2^2 - 0.3\cos(3\pi x_1) - 0.4\cos(4\pi x_2) + 0.7$$

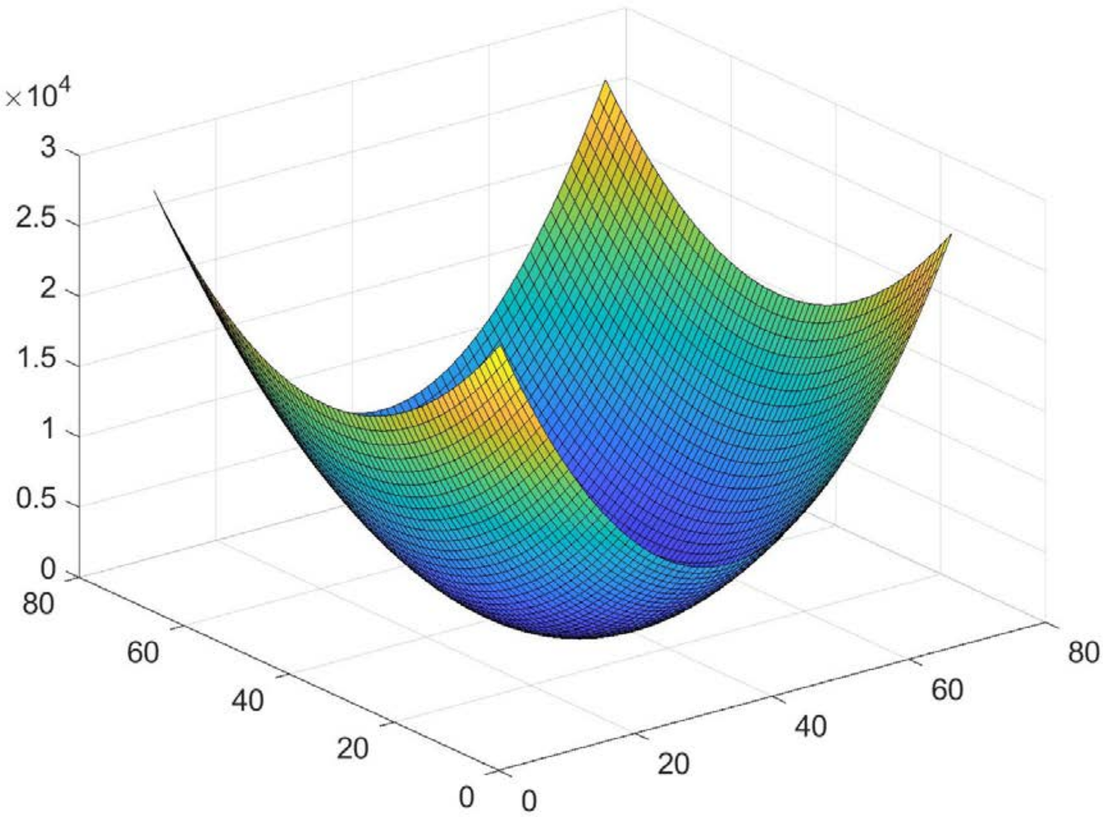


Figure 34. Surface of the Challenge Function: Bohachevsky Function

4.2.1.17. Sphere Function has a shape of bowl and has two inputs.

$$f(x) = x_1^2 + x_2^2$$

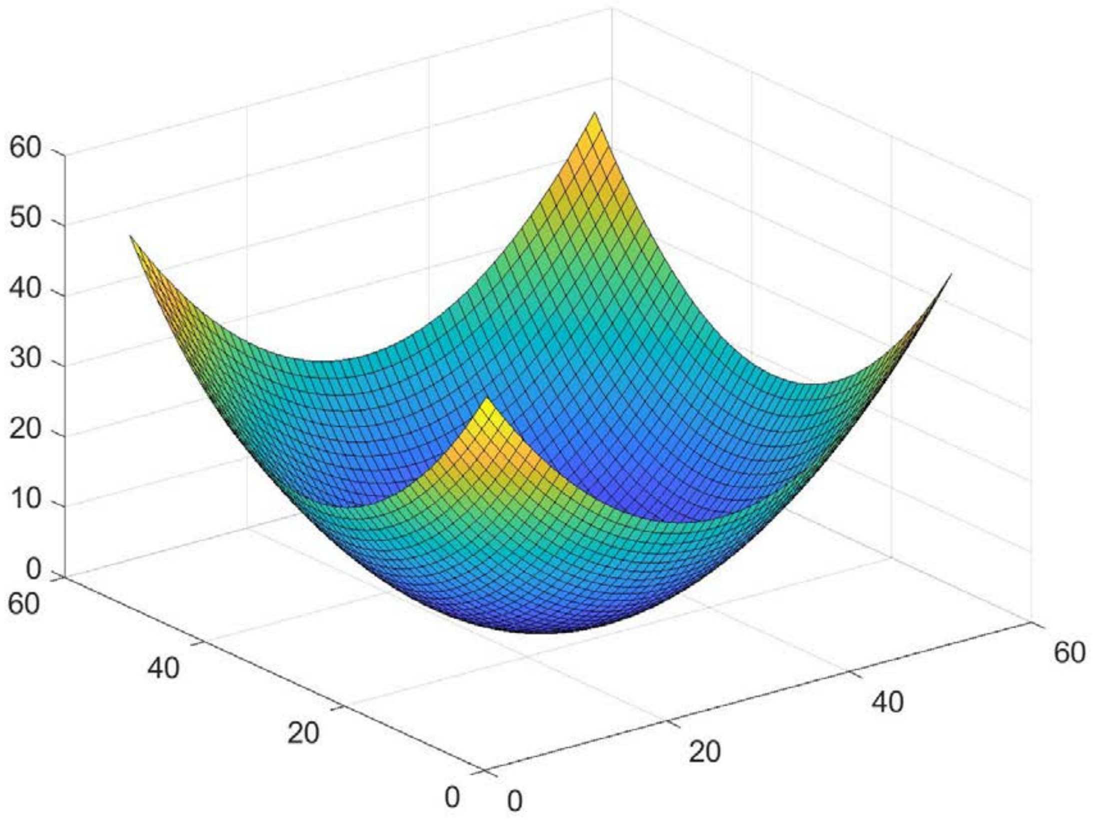


Figure 35. Surface of the Challenge Function: Sphere Function

4.2.1.18. Rotated Hyper-Ellipsoid Function has a shape of bowl and has two inputs.

$$f(x) = \sum_{i=1}^{2} \sum_{j=1}^{i} x_j^2$$

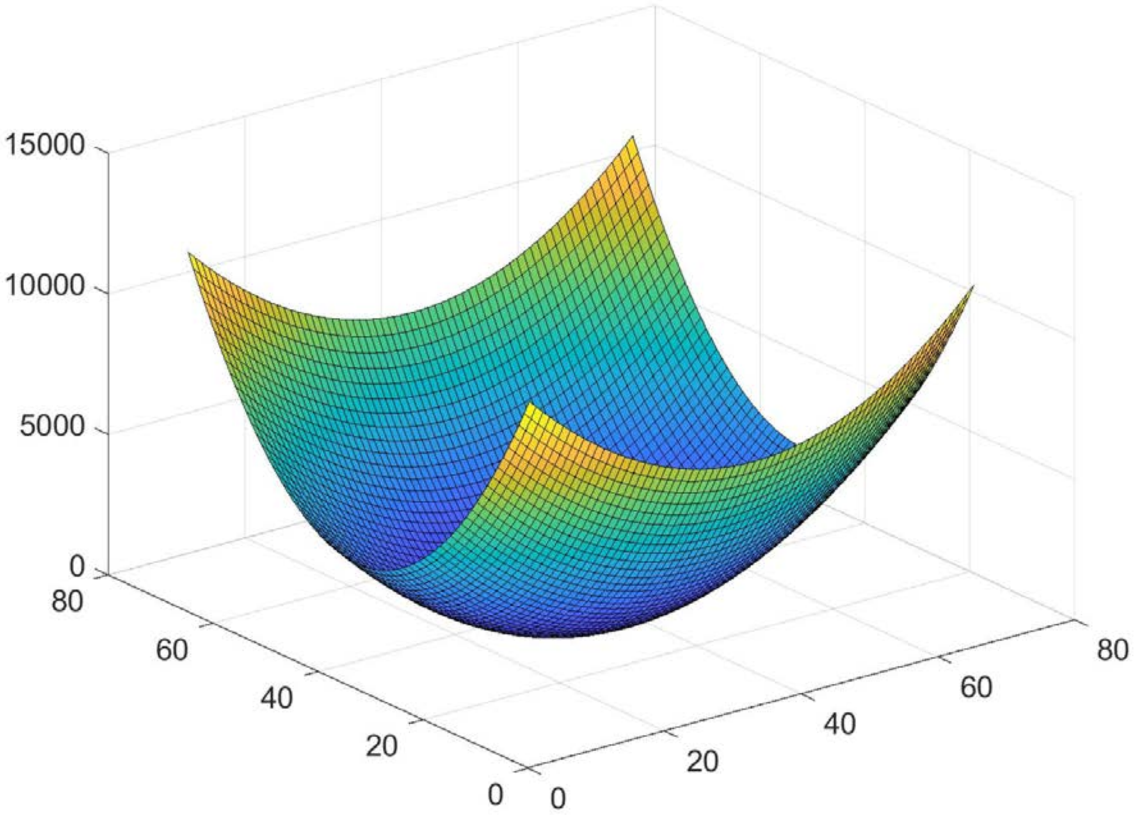


Figure 36. Surface of the Challenge Function: Rotated Hyper-Ellipsoid Function

4.2.1.19. Sum Squares Function has a shape of bowl and has three inputs.

$$f(x) = x_1^2 + 2x_2^2 + 3x_3^2$$

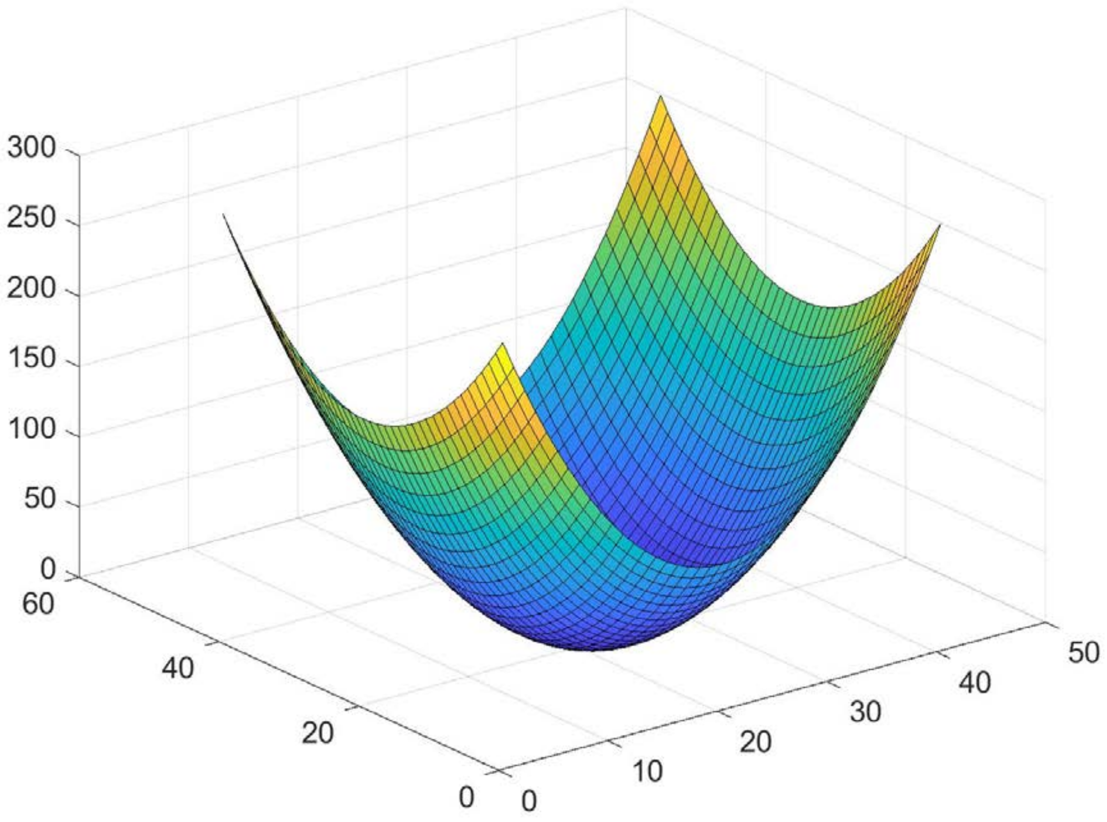


Figure 37. Surface of the Challenge Function: Sum Squares Function

4.2.1.20. Sum of Different Powers Function has a shape of bowl and has four inputs.

$$f(x) = \sum_{i=1}^{4} |x_i|^{i+1}$$

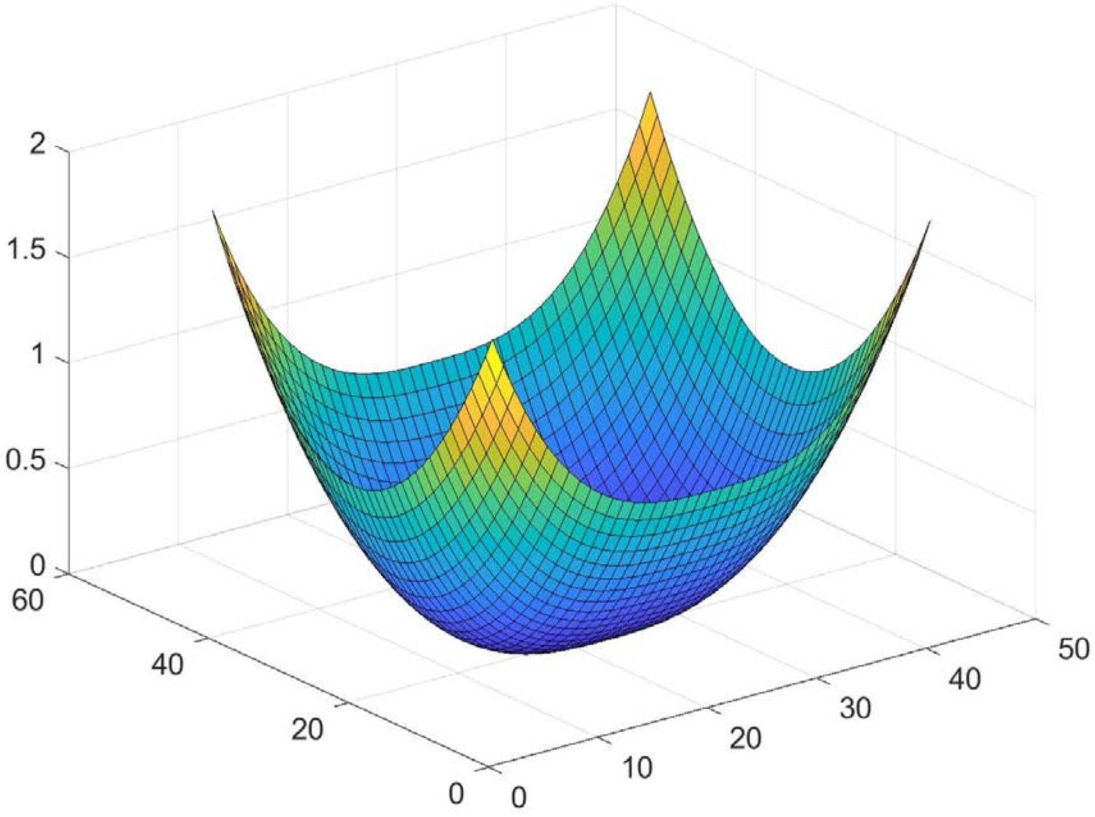


Figure 38. Surface of the Challenge Function: Sum of Different Powers Function

4.2.1.21. Trid Function has a shape of bowl and has five inputs.

$$f(x) = \sum_{i=1}^{5}(x_i - 1)^2 - \sum_{i=2}^{5} x_i x_{i-1}$$

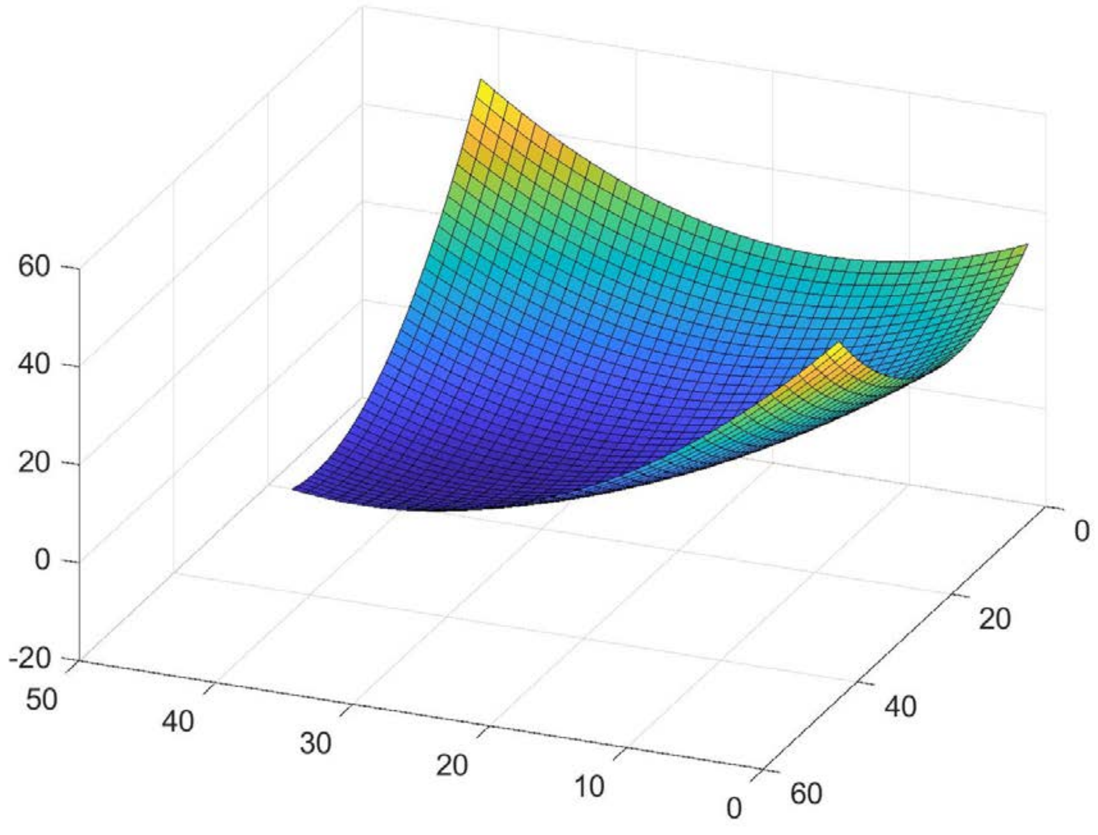


Figure 39. Surface of the Challenge Function: Trid Function

4.2.1.22. Perm Function 0, D, Beta has a shape of bowl and has ten inputs.

$$f(x) = \sum_{i=1}^{10} \left(\sum_{j=1}^{10} (j + \beta) \left(x_j^i - \frac{1}{j^i} \right) \right)^2$$

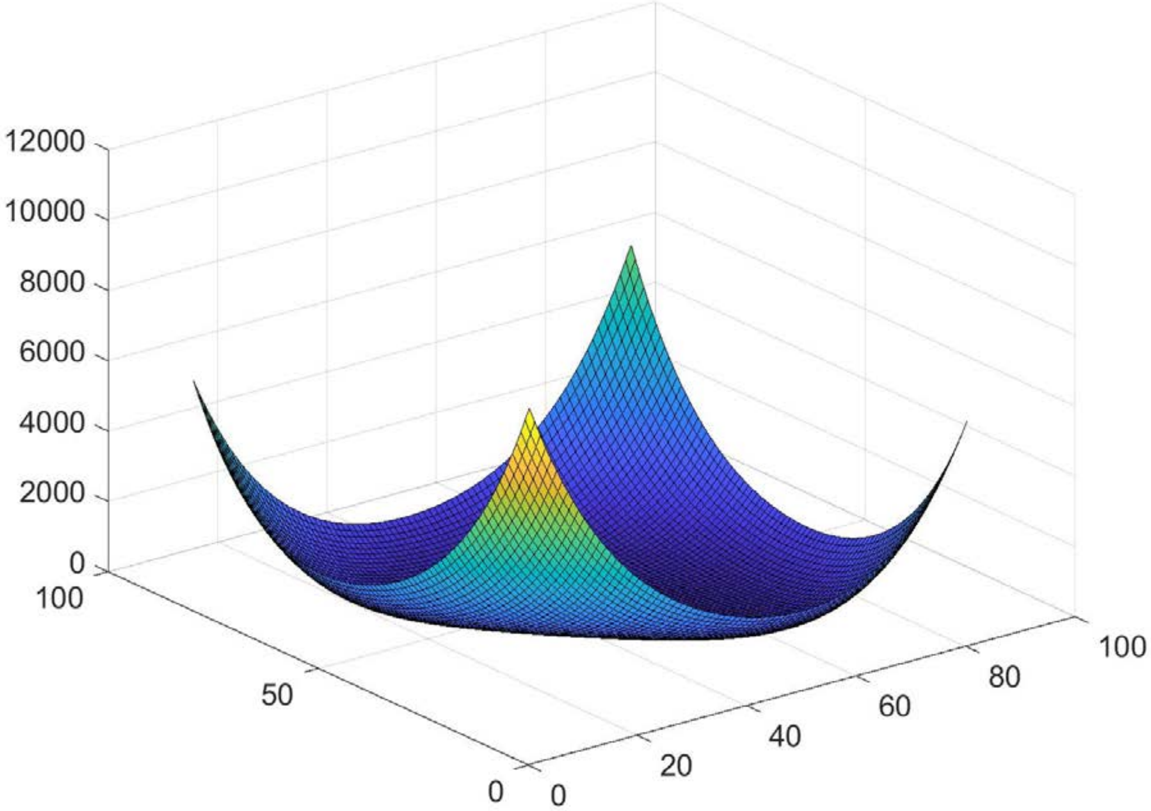


Figure 40. Surface of the Challenge Function: Perm Function 0, D, Beta

4.2.1.23. Booth Function has a shape of plate and has two inputs.

$$f(x) = (x_1 + 2x_2 - 7)^2 + (2x_1 + x_2 - 5)^2$$

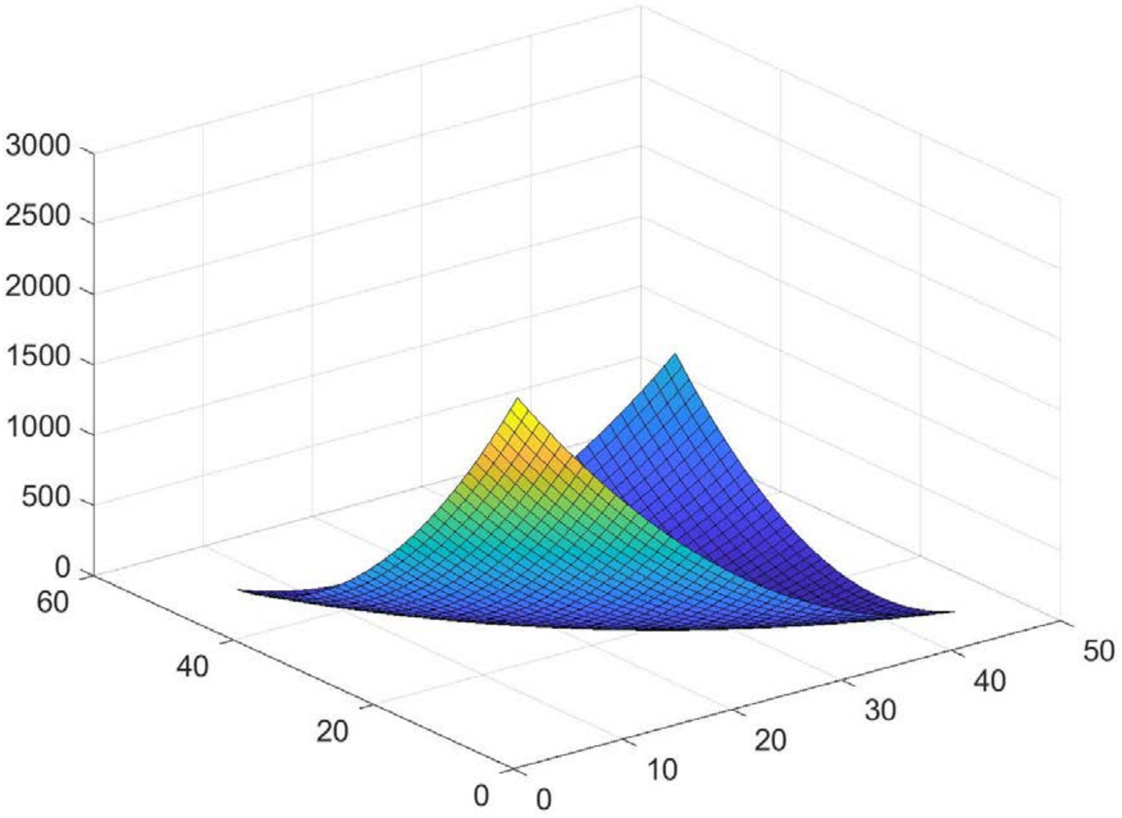


Figure 41. Surface of the Challenge Function: Booth Function

4.2.1.24. Matyas Function has a shape of plate and has two inputs.

$$f(x) = 0.26(x_1^2 + x_2^2) - 0.48x_1x_2$$

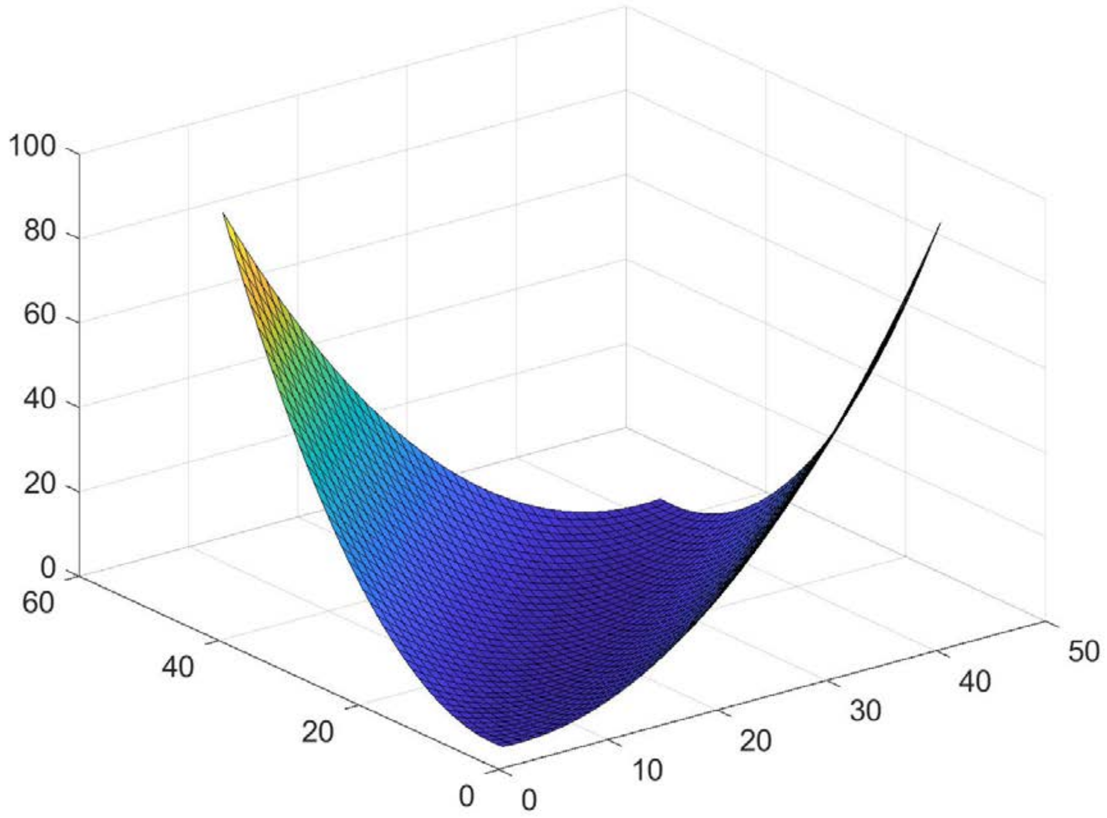


Figure 42. Surface of the Challenge Function: Matyas Function

4.2.1.25. McCormick Function has a shape of plate and has two inputs.

$$f(x) = \sin(x_1 + x_2) + (x_1 - x_2)^2 - 1.5x_1 + 2.5x_2 + 1$$

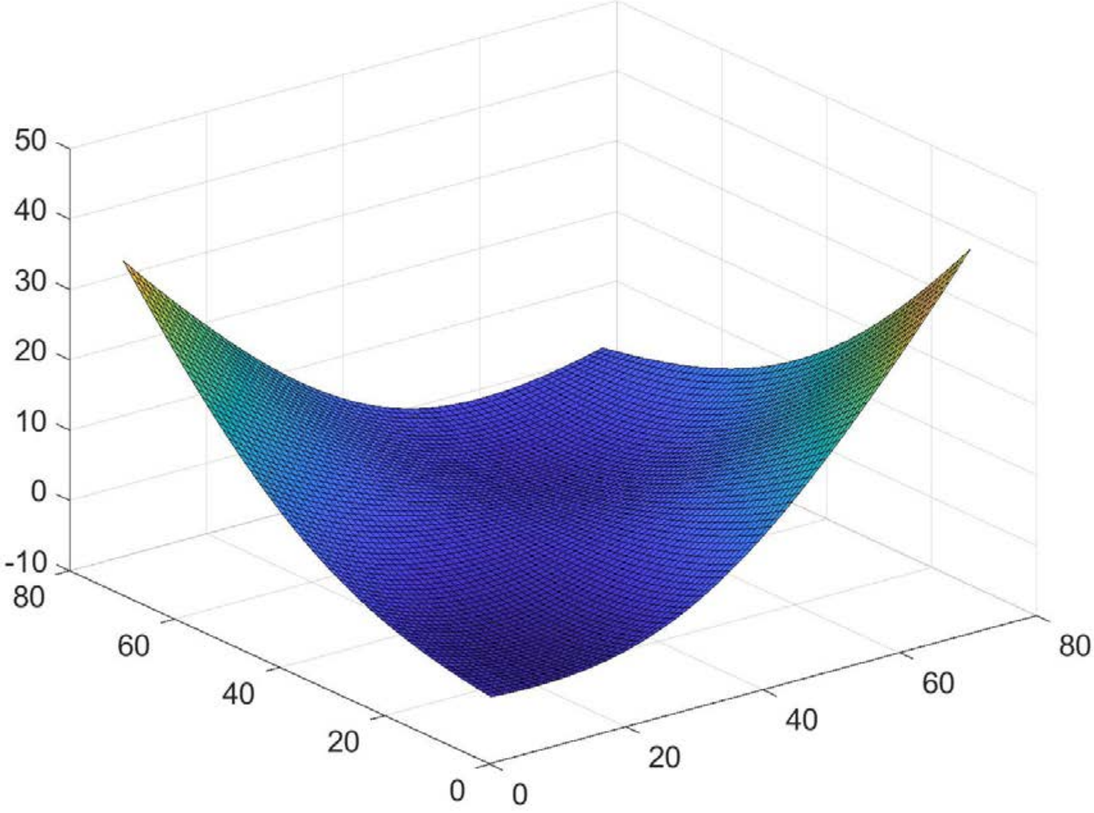


Figure 43. Surface of the Challenge Function: McCormick Function

4.2.1.26. Power Sum Function has a shape of plate and has ten inputs.

$$f(x) = \sum_{i=1}^{10} \left[\left(\sum_{j=1}^{10} x_j^i \right) - b_i \right]^2$$

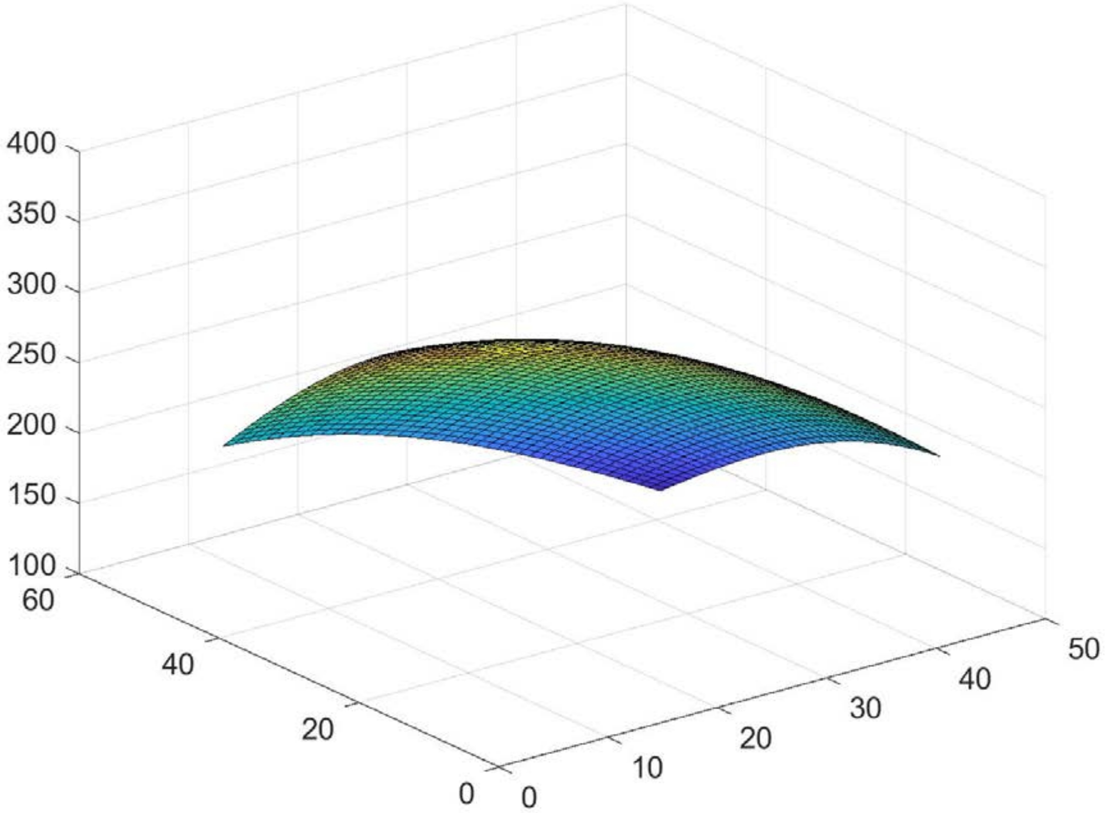


Figure 44. Surface of the Challenge Function: Power Sum Function

4.2.1.27. Zakharov Function has a shape of plate and has two inputs.

$$f(x) = \sum_{i=1}^{10} x_i^2 + \left(\sum_{i=1}^{10} 0.5ix_i\right)^2 + \left(\sum_{i=1}^{10} 0.5ix_i\right)^4$$

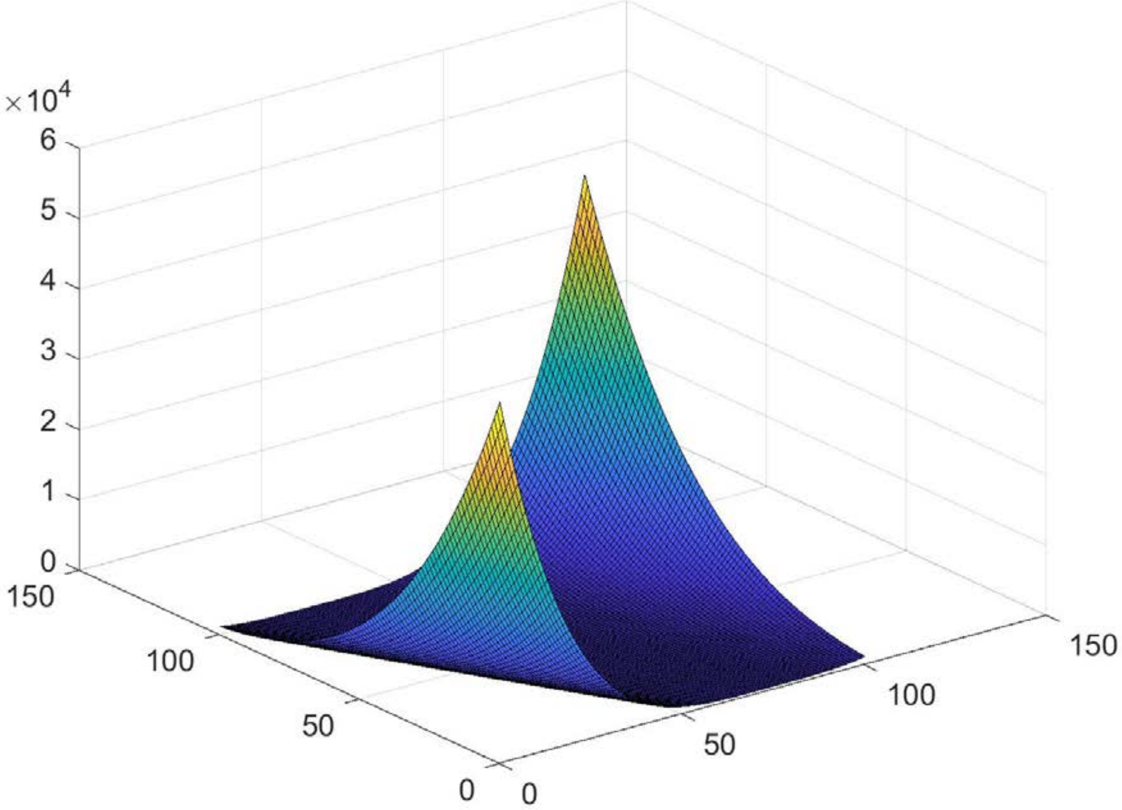


Figure 45. Surface of the Challenge Function: Zakharov Function

4.2.1.28. Three-Hump Camel Function has a shape of valley and has two inputs.

$$f(x) = 2x_1^2 - 1.05x_1^4 + \frac{x_1^6}{6} + x_1x_2 + x_2^2$$

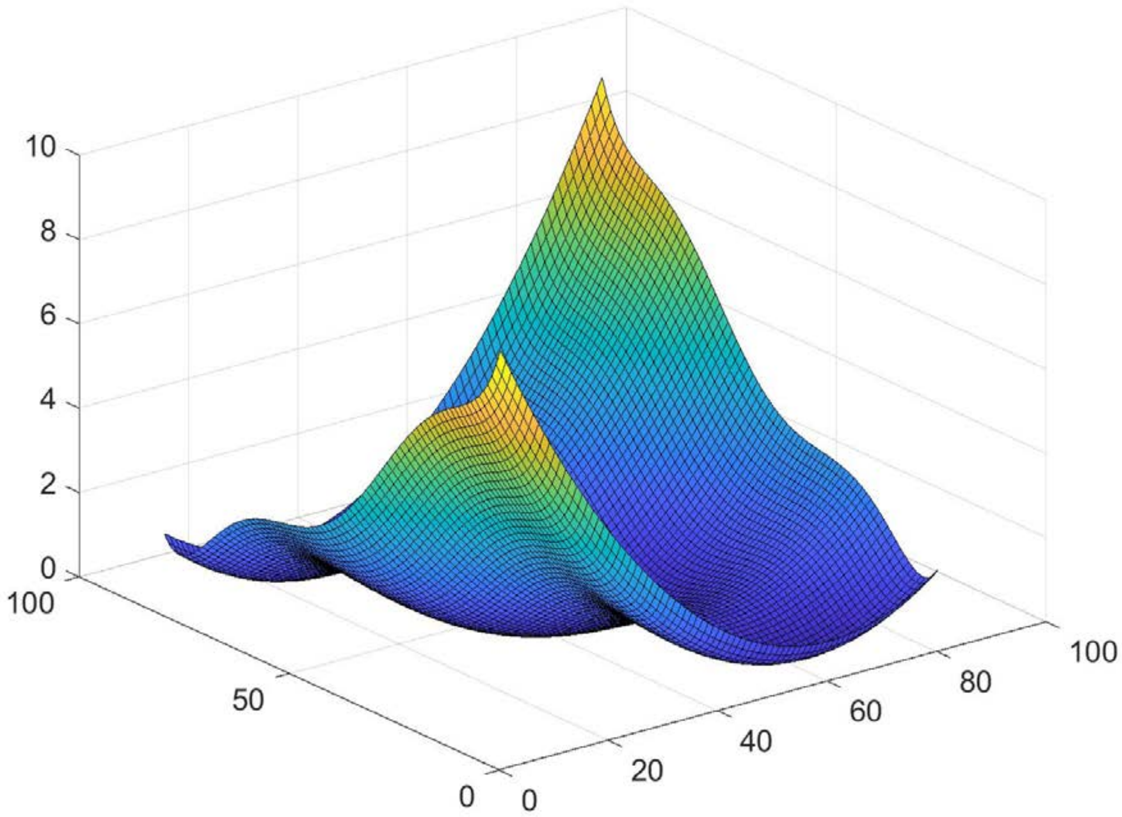


Figure 46. Surface of the Challenge Function: Three-Hump Camel Function

4.2.1.29. Six-Hump Camel Function has a shape of plate and has two inputs.

$$f(x) = \left(4 - 2.1x_1^2 + \frac{x_1^4}{3}\right)x_1^2 + x_1x_2 + (-4 + 4x_2^2)x_2^2$$

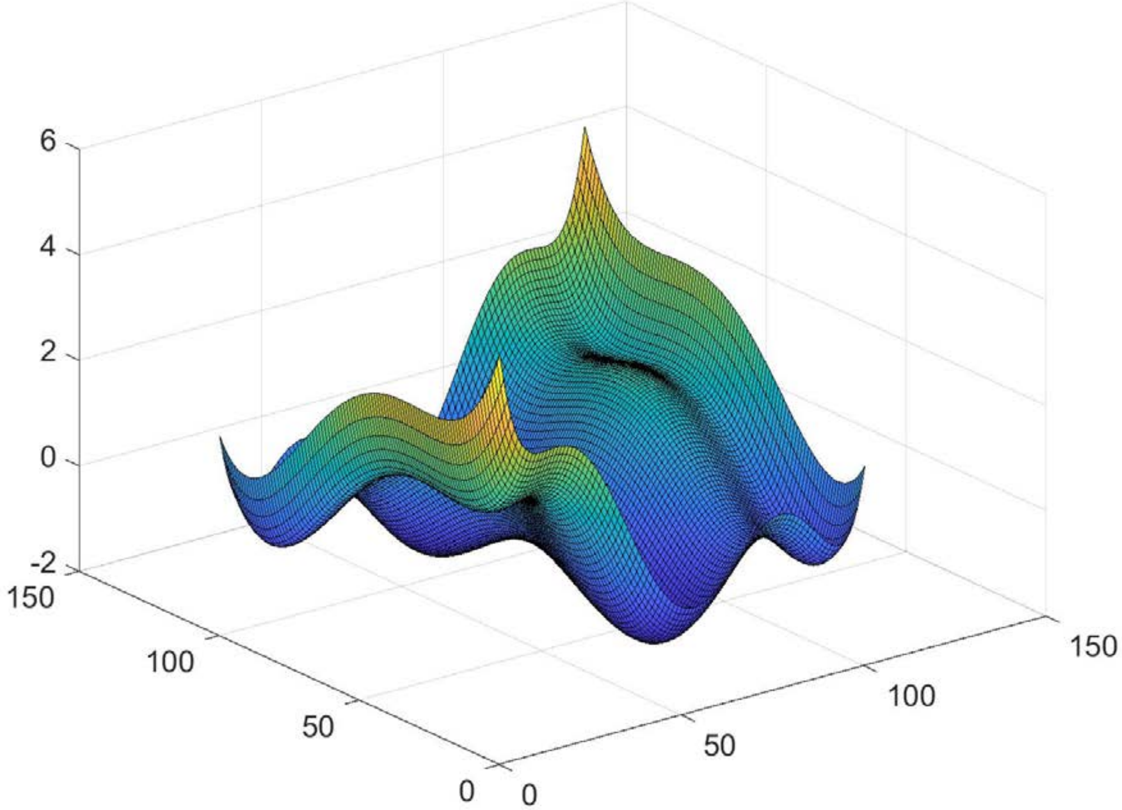


Figure 47. Surface of the Challenge Function: Six-Hump Camel Function

4.2.1.30. Dixon-Price Functions has a shape of plate and has three inputs.

$$f(x) = (x_1 - 1)^2 + \sum_{i=2}^{3} i(2x_i^2 - x_{i-1})^2$$

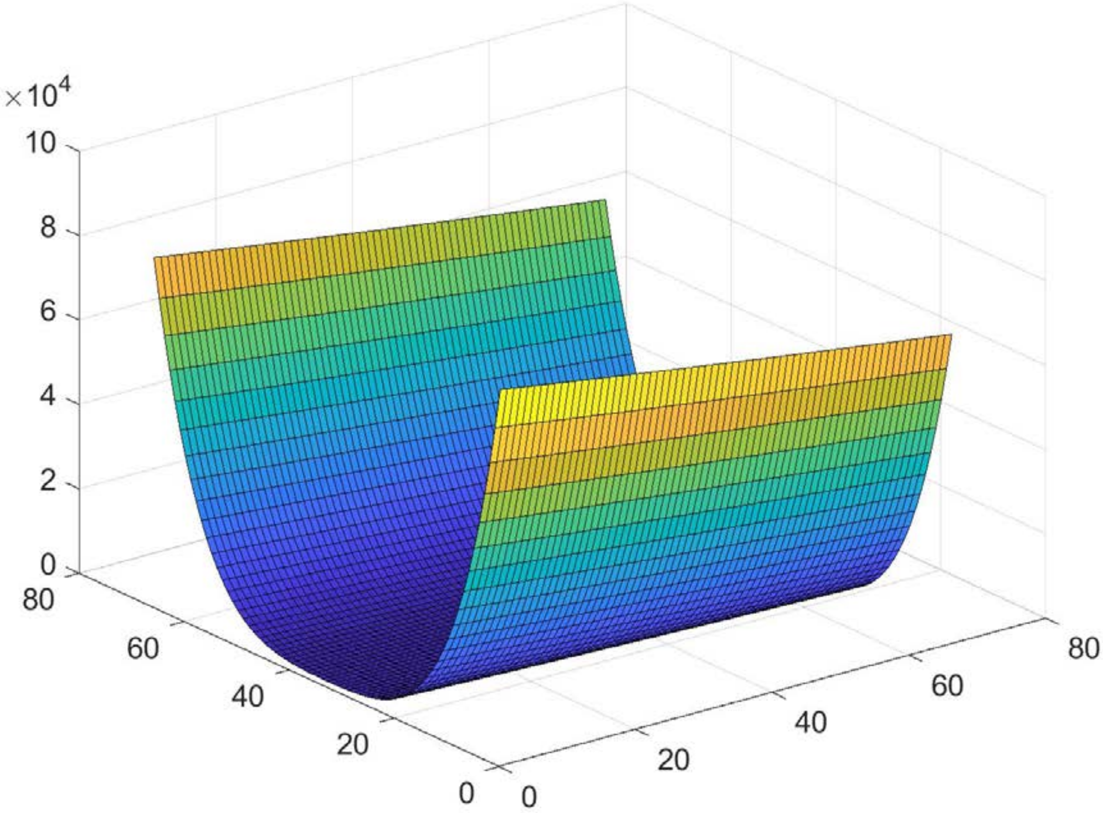


Figure 48. Surface of the Challenge Function: Dixon-Price Function

4.2.1.31. Rosenbrock Function has a shape of plate and has three inputs.

$$f(x) = [100(x_2 - x_1^2)^2 + (x_1 - 1)^2] + [100(x_3 - x_2^2)^2 + (x_2 - 1)^2]$$

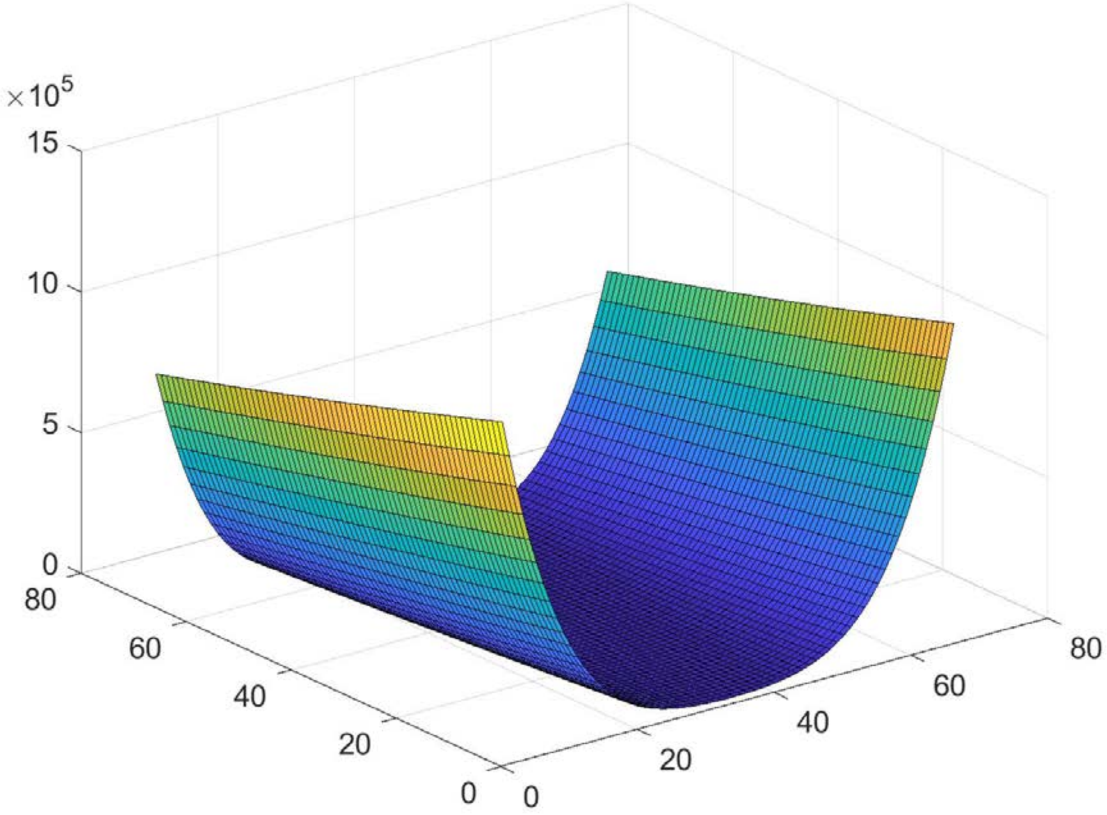


Figure 49. Surface of the Challenge Function: Rosenbrock Function

4.2.1.32. De Jong Function N. 5 has a shape of ridges and drops and has two inputs.

$$f(x) = \left(0.002 + \sum_{i=1}^{25} \frac{1}{i + (x_1 a_{1i})^2 + (x_2 - a_{2i})^6}\right)^{-1}$$

$$where\ a = \begin{pmatrix} -32 & -16 & 0 & 16 & 32 & -32 & \cdots & 0 & 16 & 32 \\ -32 & -32 & -32 & -32 & -32 & -16 & \cdots & 32 & 32 & 32 \end{pmatrix}$$

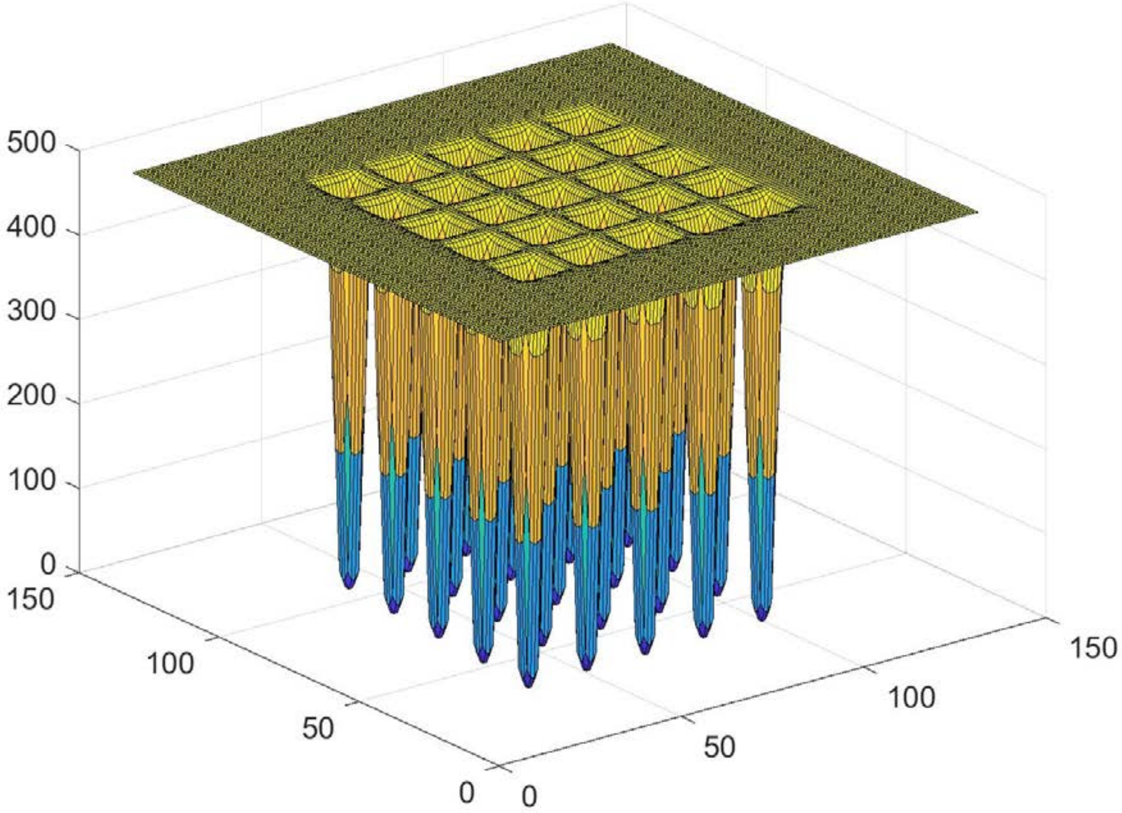


Figure 50. Surface of the Challenge Function: De Jong Function N.5

4.2.1.33. Easom Function has a shape of ridges and drops and has two inputs.

$$f(x) = -\cos(x_1)\cos(x_2)exp(-(x_1-\pi)^2-(x_2-\pi)^2)$$

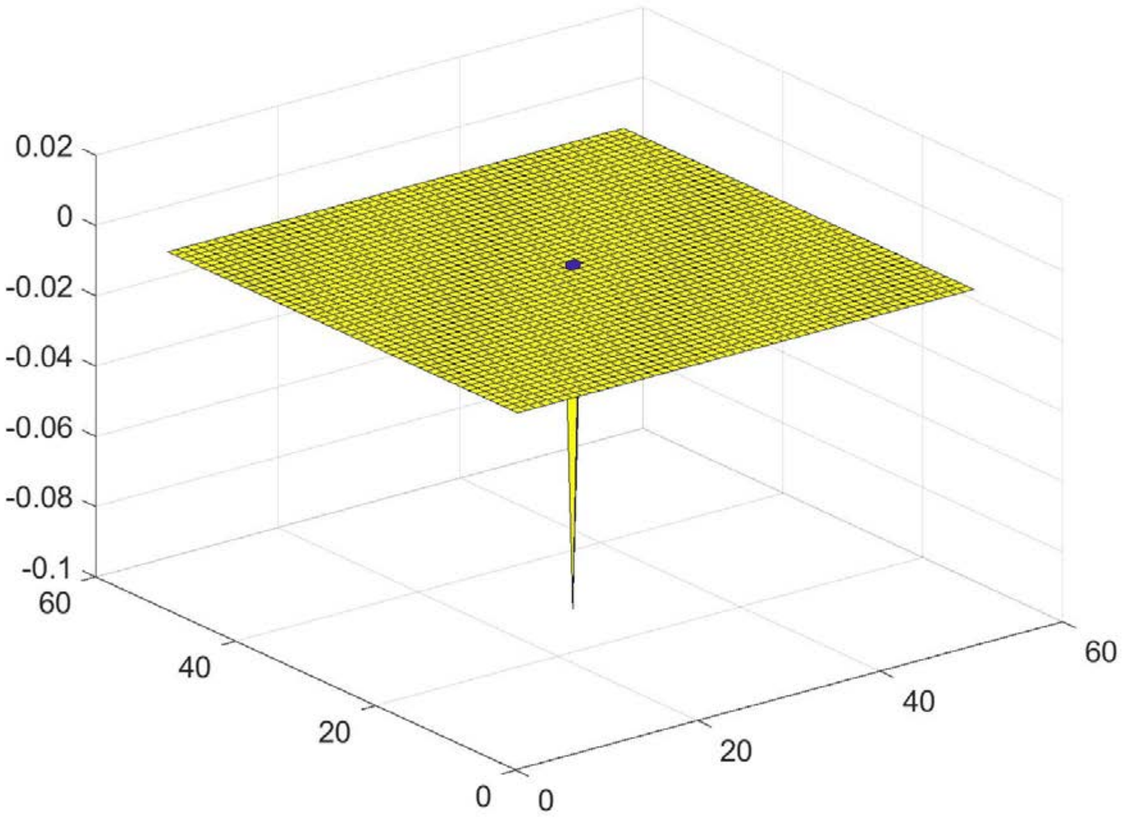


Figure 51. Surface of the Challenge Function: Easom Function

4.2.1.34. Michalewicz Function has a shape of ridges and drops and has five inputs.

$$f(x) = -\sum_{i=1}^{5} \sin(x_i)\sin^{2m}\left(\frac{ix_i^2}{\pi}\right)$$

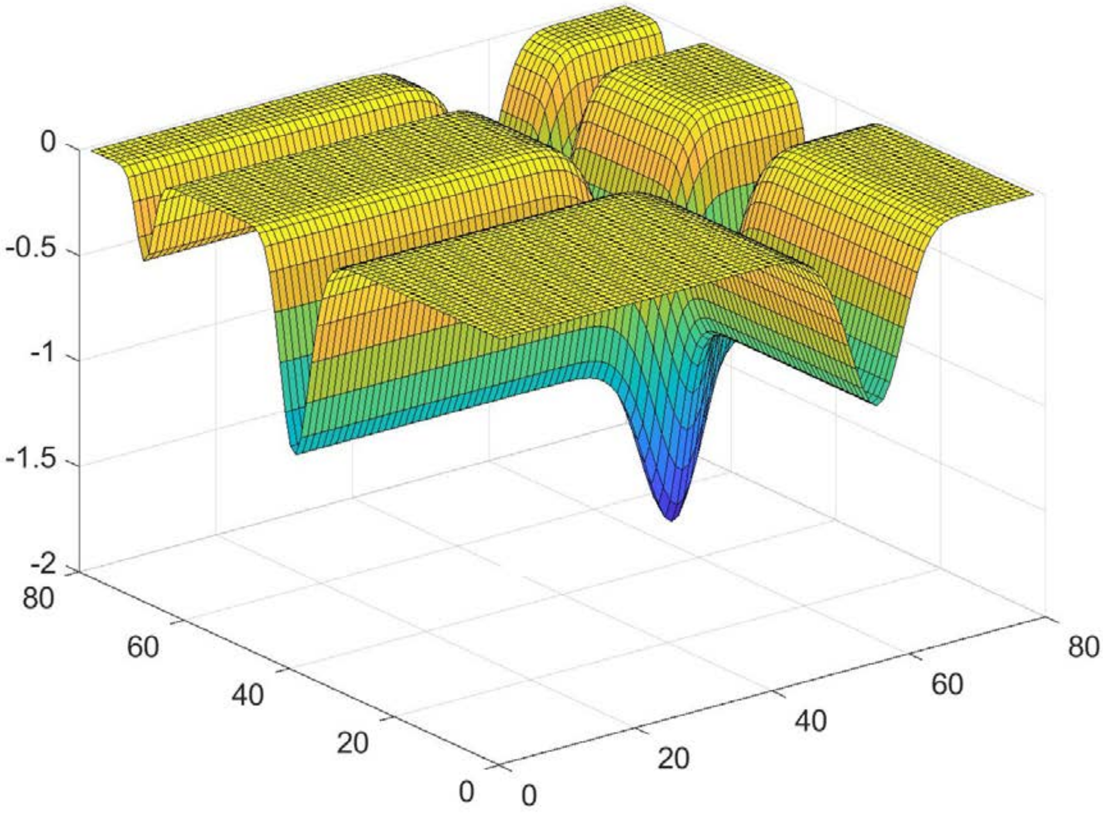


Figure 52. Surface of the Challenge Function: Michalewicz Function

4.3. Data Set Sampling Methods

To evaluate the surrogate models, data sets were created to train and test the models. Three statistical methods were employed to generate the data sets and include: Latin hypercube sampling, Sobol sequence and Halton sequence. The use of these three methods provided additional guidance for surrogate model selection by eliminating the variation based on the data generation method.

Latin hypercube sampling is based on a Latin square design. The square is divided into smaller squares and a single sample is assigned to each row and column. This extension allows for more dimensions to be calculated per square. This method allows the data set to be spread over the interval. One disadvantage is that the number of intervals must be chosen prior to the application (McKay, 1979). Figure 53 shows an example of a data set generated in MATLAB using this method with 500 data points.

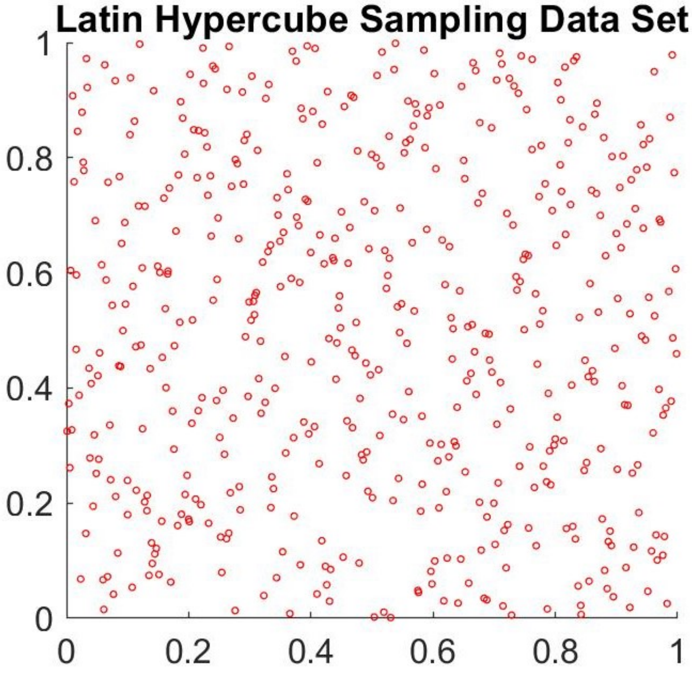


Figure 53. 500 datapoint set generated using Latin Hypercube Sampling Method

Sobol is a low discrepancy sequence which work by filling the larger gaps between the pervious numbers in the sequence. Sobol is executed by using a base of 2 to generate the data (Navid, 2018). Figure 54 shows and example of a data set generated in MATLAB using this method with 500 data points.

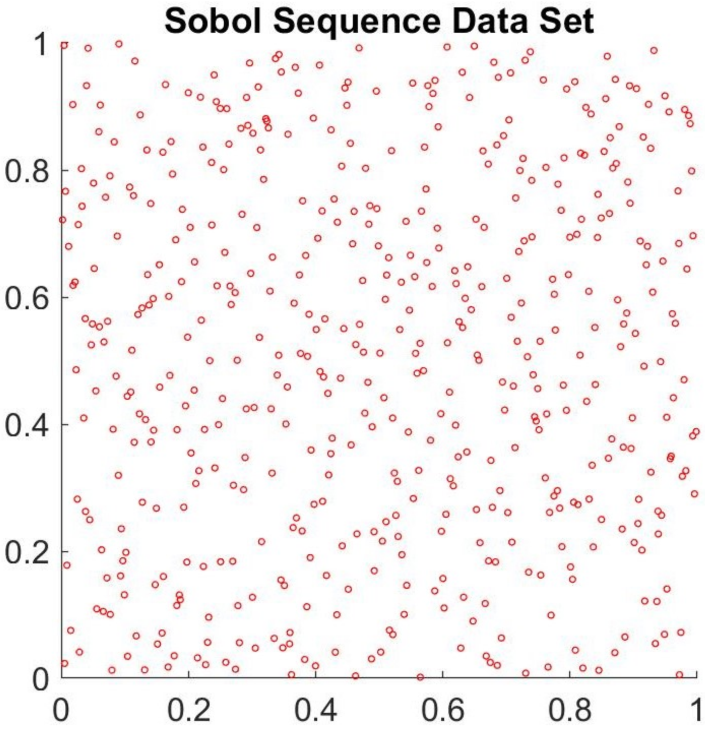


Figure 54. 500 datapoint set generated using Sobol Sequencing Method

Halton is similar to Sobol in that they are both low discrepancy sequences and work by filling the larger gaps between the previous numbers of the sequence. The execution is slightly different in Halton operates on prime number bases (Chi, 2005). Figure 55 shows an example of a data set generated in MATLAB using this method with 500 data points.

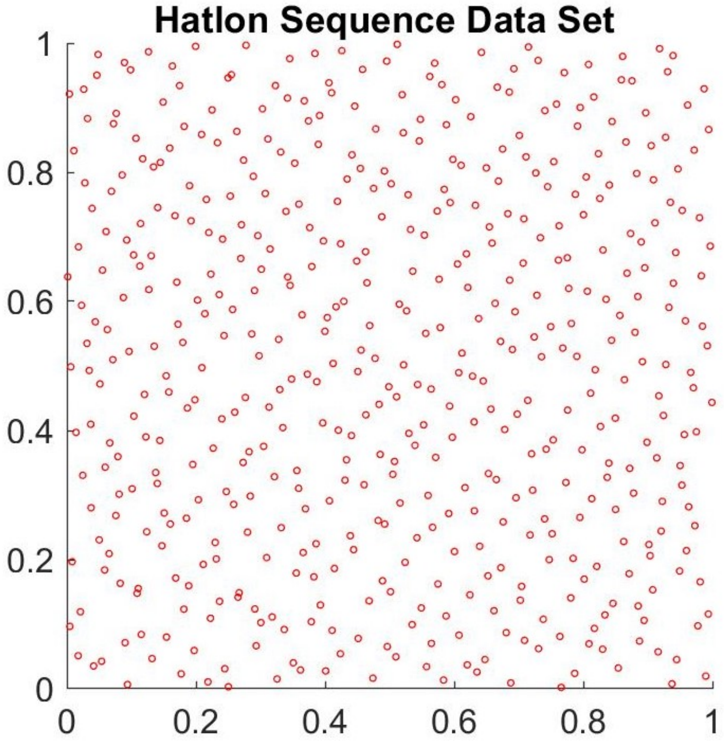


Figure 55. 500 datapoint set generated using Halton Sequencing Method

4.4. Error Calculation Methods

Error calculation methods were used in this work to compare the predicted value to the actual value. The six methods include R-squared, R-squared adjusted, Akaike Information Criterion, Bayesian Information Criterion, root mean square error and maximum absolute error.

The coefficient of determination, R-squared, is defined as the proportion of the variance in the dependent variable that is predictable from the independent variable (Tjur, 2009). Essentially, R-squared provides a measure of the accuracy of the regression model and can be calculated using the following equations 4.1, 4.2, and 4.3.

$$SS_{res} = \sum_i (y_i - \hat{y}_i)^2 \qquad 4.1$$

$$SS_{tot} = \sum_i \left(y_i - \frac{1}{n} \sum_{i=1}^{n} y_i \right)^2 \qquad 4.2$$

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}} \qquad 4.3$$

Where: n = Number of data points

$\hat{y}_i$= Predicted value

y_i= Actual Value

R-squared adjusted in based on the original R-squared, however takes into account the impact of a model with multiple predictors. It can be used to measure the useful variables in a model. As variables are added to a model, R-squared will only increase. R-squared adjusted may increase or decrease with additional variables depending on their worth (Montgomery et al, 2009). The calculation is shown in 4.4, 4.5, 4.6.

$$\bar{R}^2 = 1 - \frac{SS_{res}/_{df_e}}{SS_{tot}/_{df_t}} \qquad 4.4$$

$$df_e = n - p - 1 \qquad 4.5$$

$$df_t = n - 1 \qquad 4.6$$

Where: n = Number of data points

p = Number of variables in the model

Akaike Information Criterion can be used to perform model comparisons. This method is based on in-sample fit to estimate the likelihood of a model to predict the values ((Mohammed & Far, 2015). When using Akaike Information Criterion, a comparison may be made across a group of models by pinpointing the minimum value. The Akaike Information Criterion is calculated using equation 4.7.

$$AIC = -2 * \ln L + 2 * k \qquad 4.7$$

Where: L = Value of the Likelihood

k = Number of estimated parameters

Bayesian Information Criterion is very similar to the Akaike Information Criterion in form and application, however their assumptions are very different. Bayesian Information Criterion considers the number of recorded measurements (Busemeyer & Diederich, 2014) as seen in equation 4.8.

$$BIC = -2 * \ln L + 2 * \ln N * k \qquad 4.8$$

Where: L = Value of the Likelihood

N = Number of recorded measurements

K = Number of estimated parameters

Root mean square error represent the standard deviation of the residuals, or the distance from the regression line to the calculated point. Root mean square error is generally helpful to evaluate the regression quality (Hyndman & Koehler, 2006). The equation used for calculation is shown in equation 4.9.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{n}} \qquad 4.9$$

Where: n = Number of data points

$\hat{y}_i$= Predicted value

y_i= Actual Value

Maximum absolute error represents the maximum absolute value of the difference of the predicted and observed values. This method can provide valuable information related to the predicted points, whereas root mean square error can provide information about the fit of the regression (Willmott & Matsuura, 2005). The maximum absolute error can be calculated using equation 4.10.

$$MAE = \max|y_i - \hat{y}_i| \qquad 4.10$$

4.5. Computational Experiments

Input-output pairs were generated using Latin hypercube sampling, Sobol Sequence and Halton Sequence over nine sample sizes including 50, 100, 200, 400, 800, 1600, 3200, 6400, and 12800. Those data sets were generated for all 34 challenge functions and then used to train the eight surrogate models. Then a data set of 100,000 points created using Sobol Sequencing was generated to evaluate the models.

The effectiveness was compared through six error methods including R-squared, R-squared adjusted, Akaike Information Criterion, Bayesian Information Criterion, root mean square error and maximum absolute error based on the different between the output of the dataset and the calculated output. The results of these error calculations validate the findings across the board to provide more accurate description of the results. Additionally, the training time was recorded during the phase of the program where the model was trained, and evaluation time was recorded during the phase when the final outputs were generated. All computations were carried out on a HP Spectre X360 X64-based PC with 16 GB RAM using MATLAB 2017b.

4.6. Results

4.6.1. Results Based on Sampling Method

To establish a baseline understanding of the impact of sample size on the combination of surrogate model with sampling method, the R-squared Adjusted was calculated as a function of increasing sample size for each surrogate model when the data was generated using Sobol, Halton and LHS sampling methods and is displayed in Figure 56.

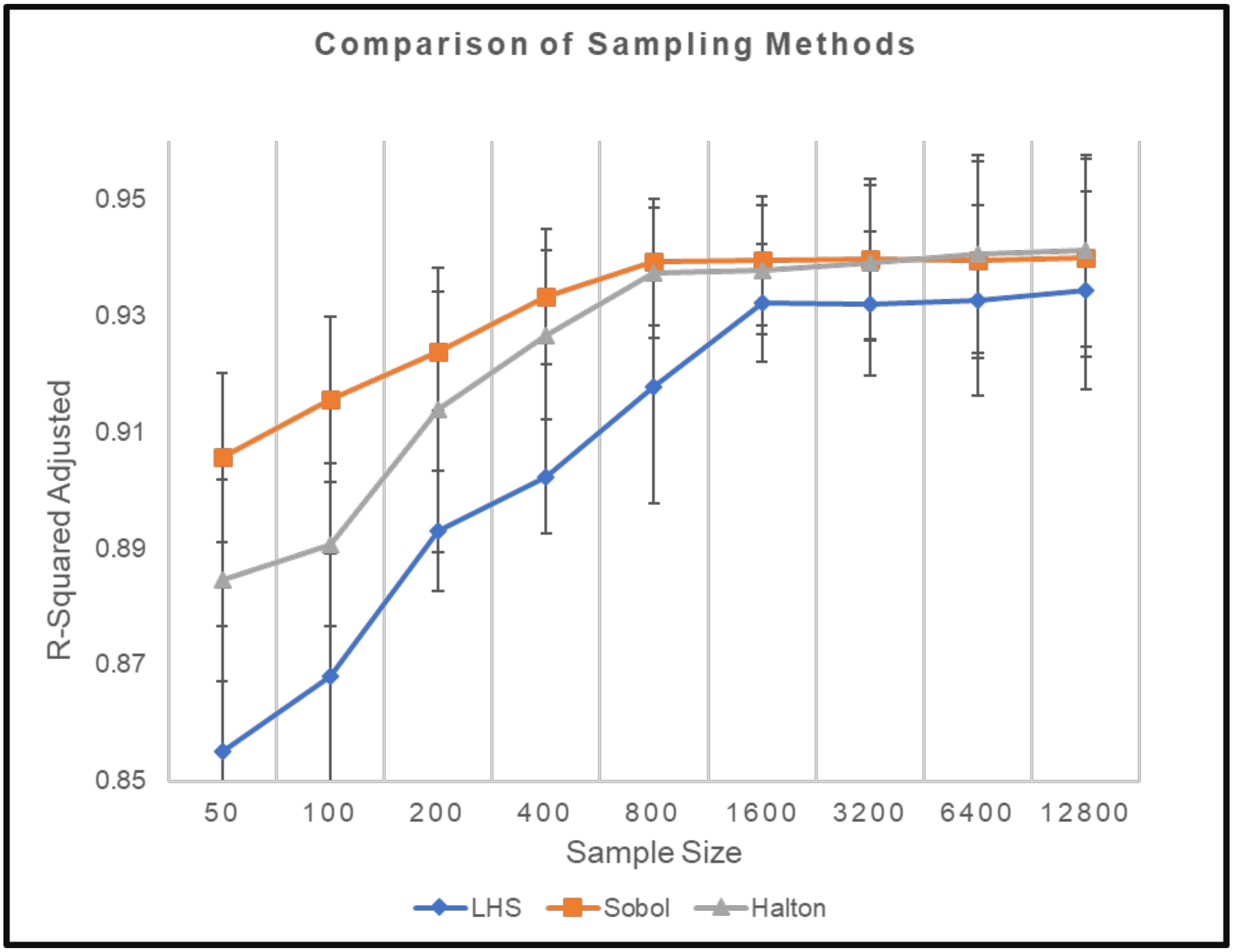


Figure 56. Comparison of Latin Hypercube Sampling, Sobol Sequence and Halton Sequence as a comparison of R-Squared Adjusted

Figure 57 compares the surrogate model results when the data set was generated using Halton sequence. Radial Basis Function seemed to be least affected by the small sample sizes. However, the performance of Multivariate Adaptive Regression Splines remained the least accurate until it began to rise when the sample size increased to 400. Artificial Neural Networks, Extreme Learning Machines and Automated Learning of Algebraic Models for Optimization performed similarly well after reaching a sample size of 1600.

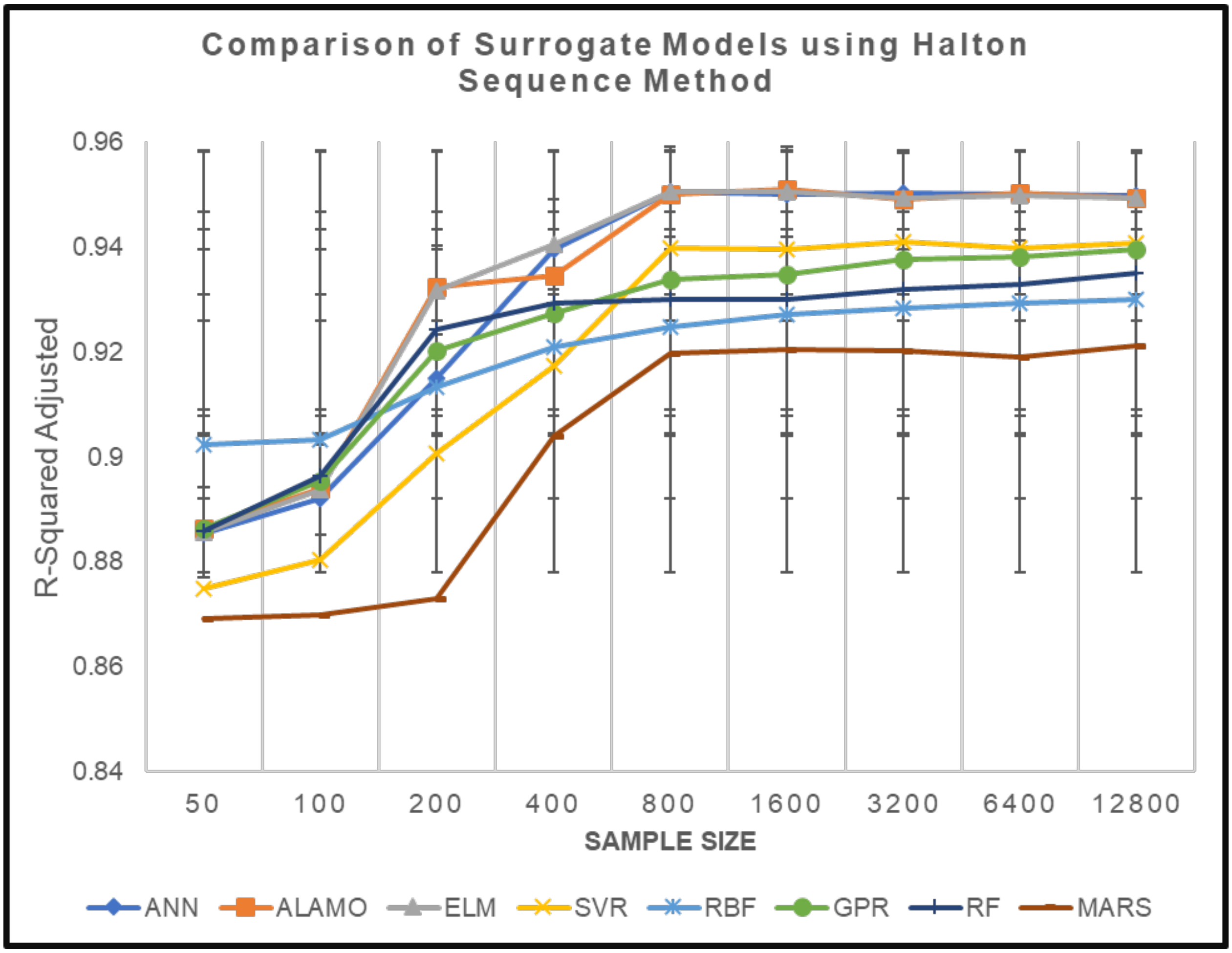


Figure 57. R-Squared Adjusted calculated for each surrogate model using Halton sampling method with respect to sample size

Figure 58. compares the surrogate model results when the data set was generated using Sobol sequence. Gaussian Process Regression and Random Forests showed little impact of the sample size on the values. The performance of Multivariate Adaptive Regression Splines remained the least accurate until it began to rise when the sample size increased to 200. Artificial Neural Networks, Extreme Learning Machines and Automated Learning of Algebraic Models for Optimization performed similarly well after reaching a sample size of 1600.

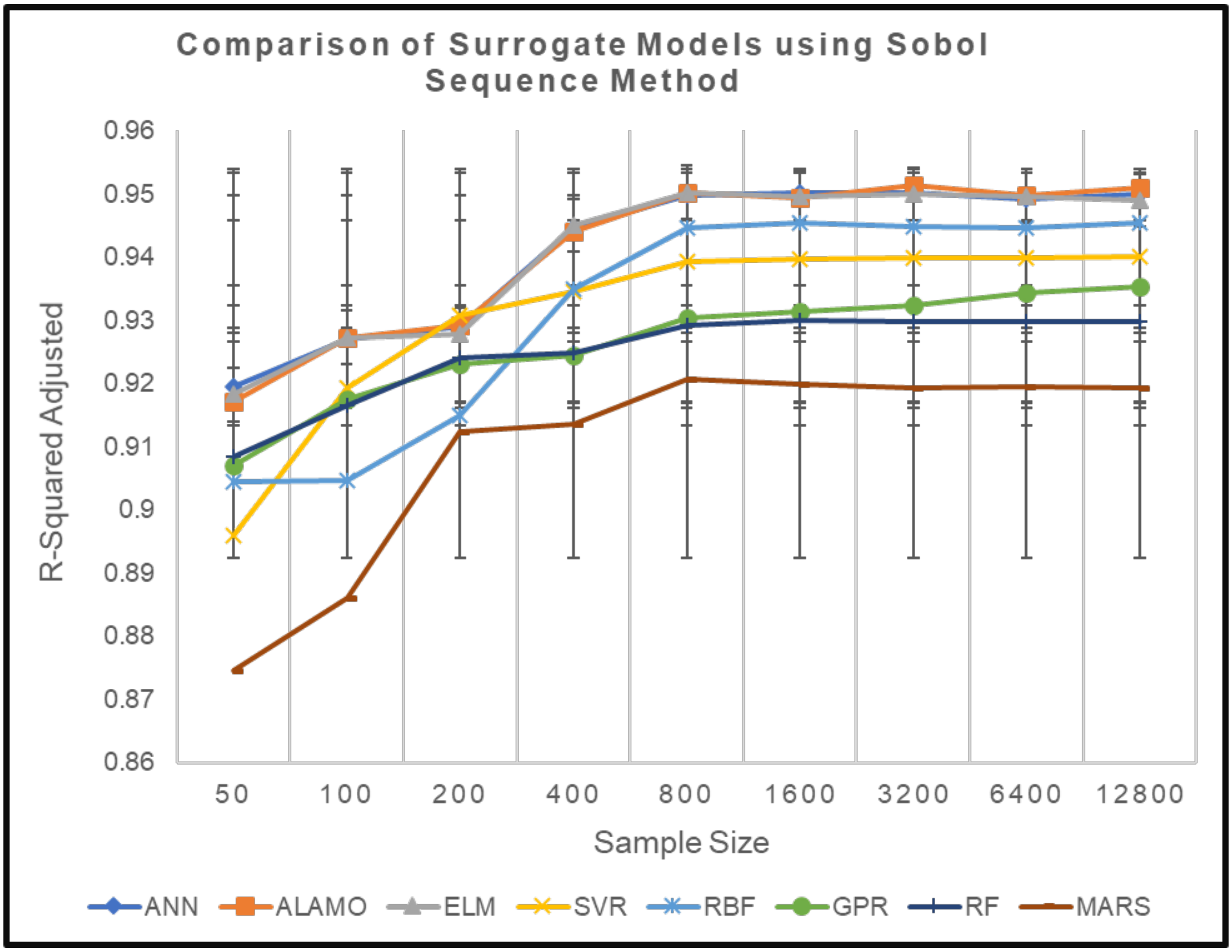


Figure 58. R-Squared Adjusted for each surrogate model using Sobol sampling method with respect to sample size

Figure 59 compares the surrogate model results when the data set was generated using Latin Hypercube Sampling. All surrogate models showed low R-Square adjusted values with small sample sizes and increase steadily as sample size increases to 1,600. Gaussian Process Regression and Random Forests showed little impact of the sample size on the values. Artificial Neural Networks, Extreme Learning Machines and Automated Learning of Algebraic Models for Optimization performed similarly well after reaching a sample size of 1600.

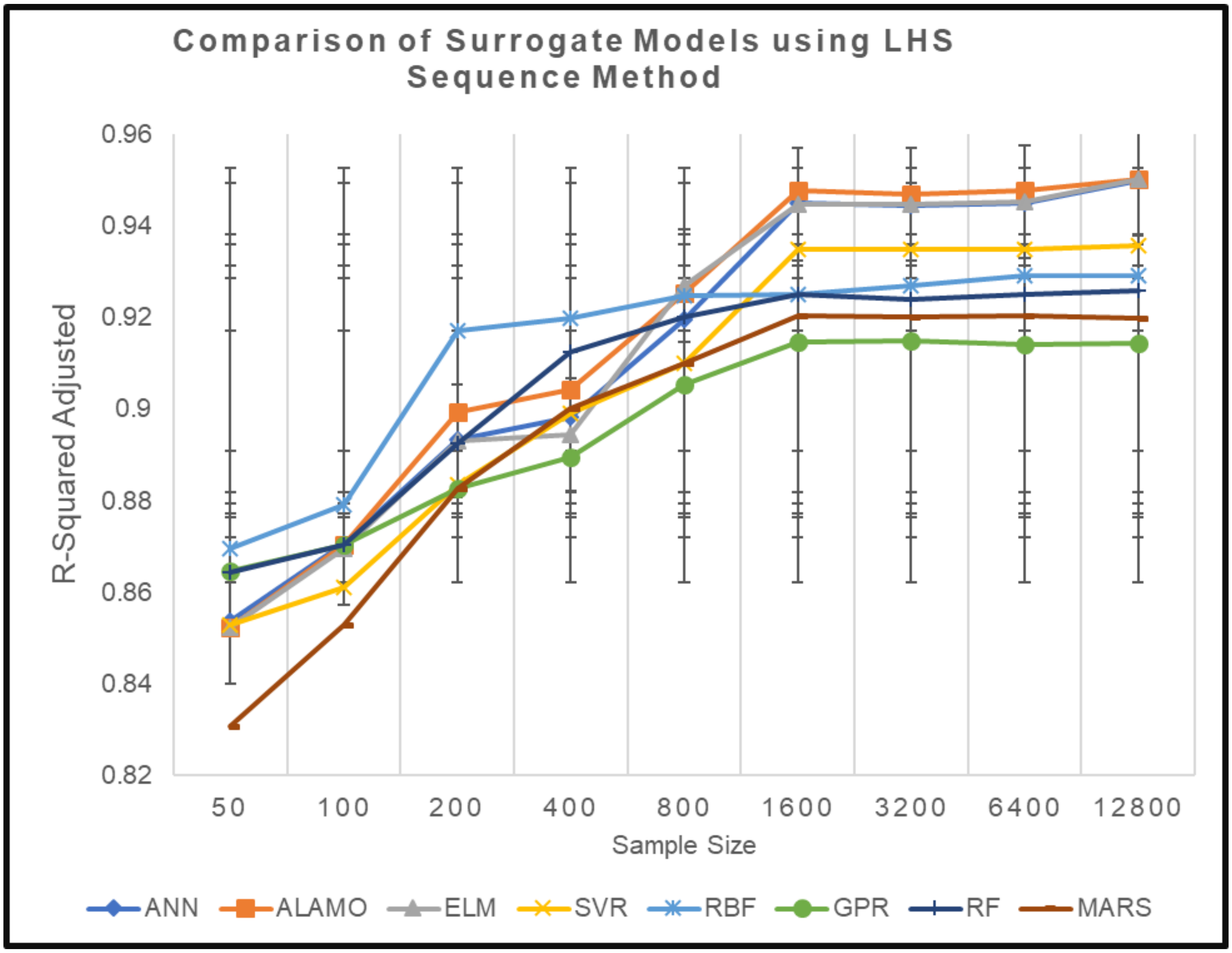


Figure 59. R-Squared Adjusted calculated for each surrogate model using Latin Hypercube Sampling method with respect to sample size

For all sampling methods and sizes, the previous graphs show that as the sample size increased from 50 to 1,600, the R-squared adjusted increases. As sample size increases from 1,600 to 12,800, the values remain close to constant.

4.6.2. Results Based on Training Time

Model training time which is the time required to prepare the model for evaluation should be included when selecting a surrogate model. Figure 60 displays the training time as sample size is increased for each surrogate model. Extreme Learning Machine calculations were performed the quickest and did not increase significantly over increased sample size. Automated Learning of Algebraic Models for Optimization show to have the longest training time and increased from 3.57 to 21.1 CPUs as sampling size increased from 50 to 12,800.

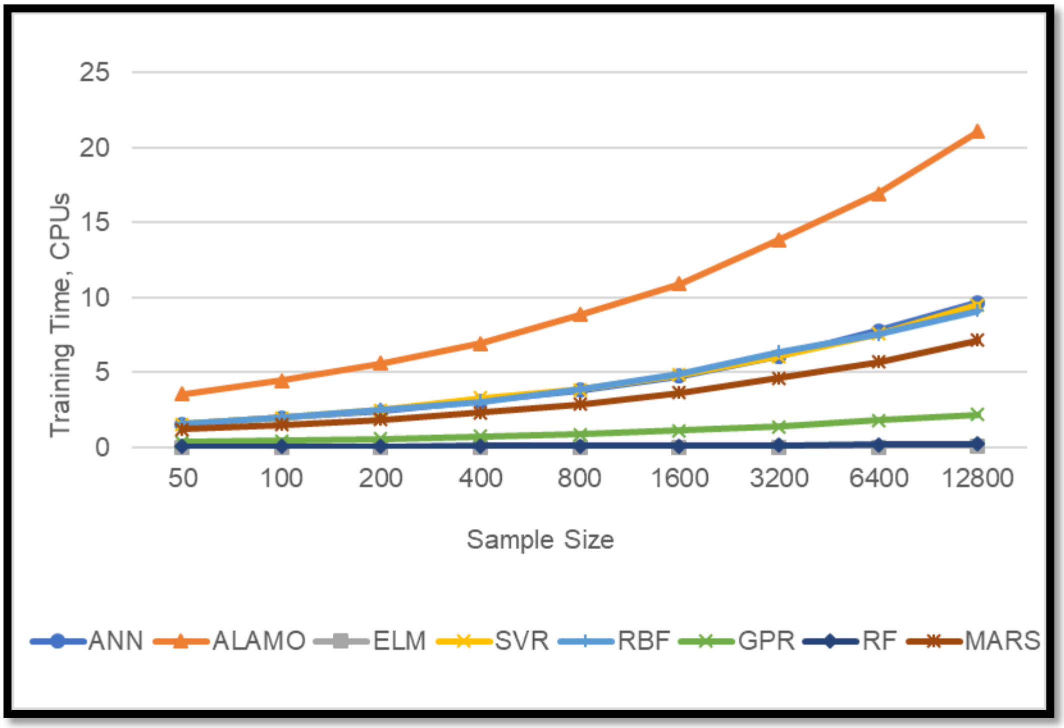


Figure 60. Surrogate Models Training Time Based on Sample Size

Figure 61 represents the evaluation time for each surrogate model over increasing sample sizes. It shows that increasing the sample size does not relate to the evaluation time. However, it does mirror the training time displayed in Figure 60 because the shortest time was ELM while the longest was ALAMO.

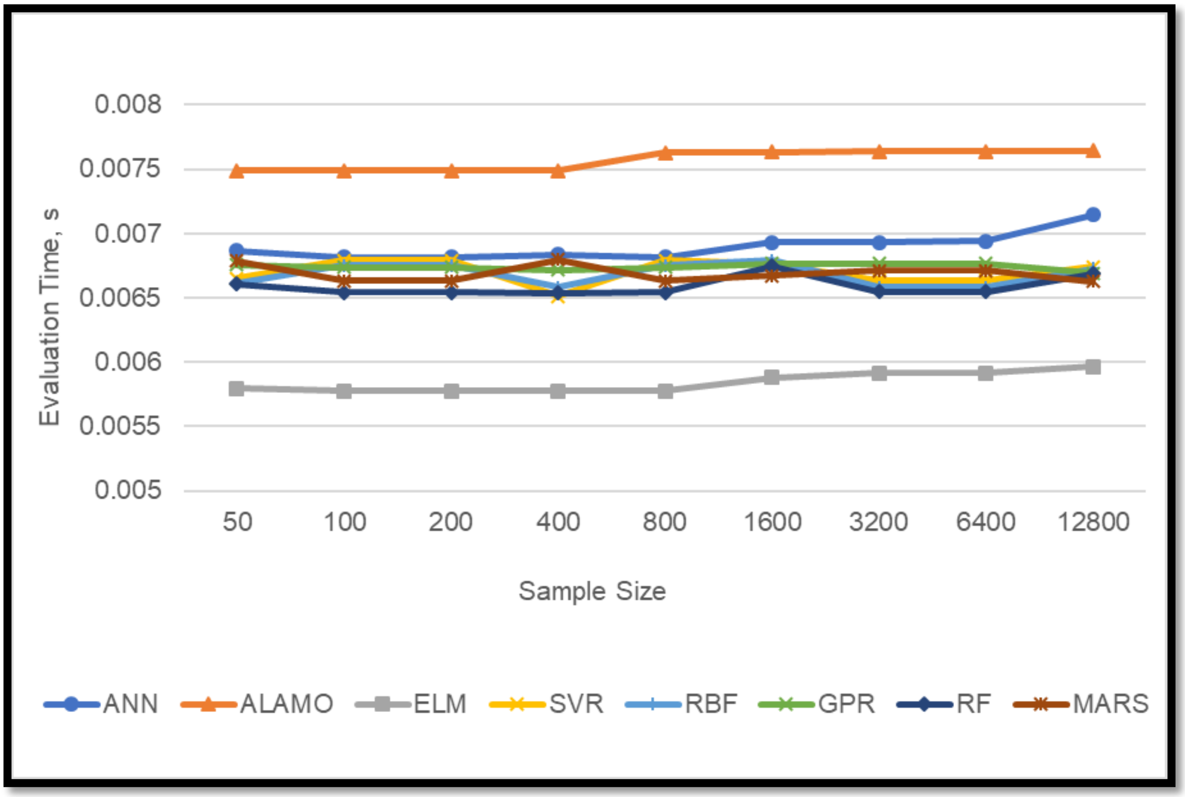


Figure 61. Surrogate Models Evaluation Time Based on Sample Size

4.6.3. Results Based on Challenge Function

4.6.3.1. Challenge Functions with Two inputs

Challenge functions with two inputs is the largest group with twenty functions. The R-squared values for the surrogate models over the range of sample sizes are shown in Figure 62. The graph shows that as the sample size increase 50 to 1,600 the R-square increases. For sample sizes of 1,600 and above, the values seem to remain constant and the error bars decrease in size showing that there is decreased variation.

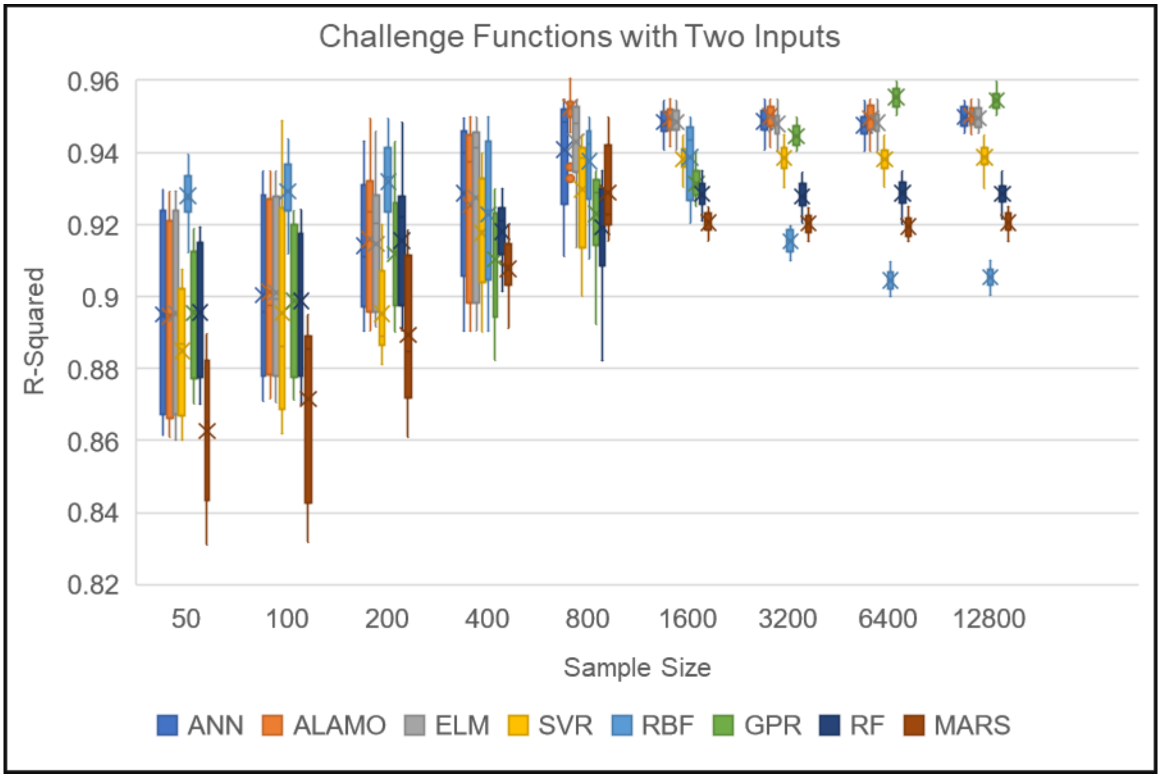


Figure 62. R-Squared Values of Challenge Functions with Two Inputs for all Surrogate Models over the range of Sample Sizes

The R-squared adjusted values for the surrogate models over the range of sample sizes are shown in Figure 63. The graph shows that as the sample size increase 50 to 1,600 the R-squared adjusted increases. For sample sizes of 1,600 and above, the values seem to remain constant and the error bars decrease in size showing that there is decreased variation.

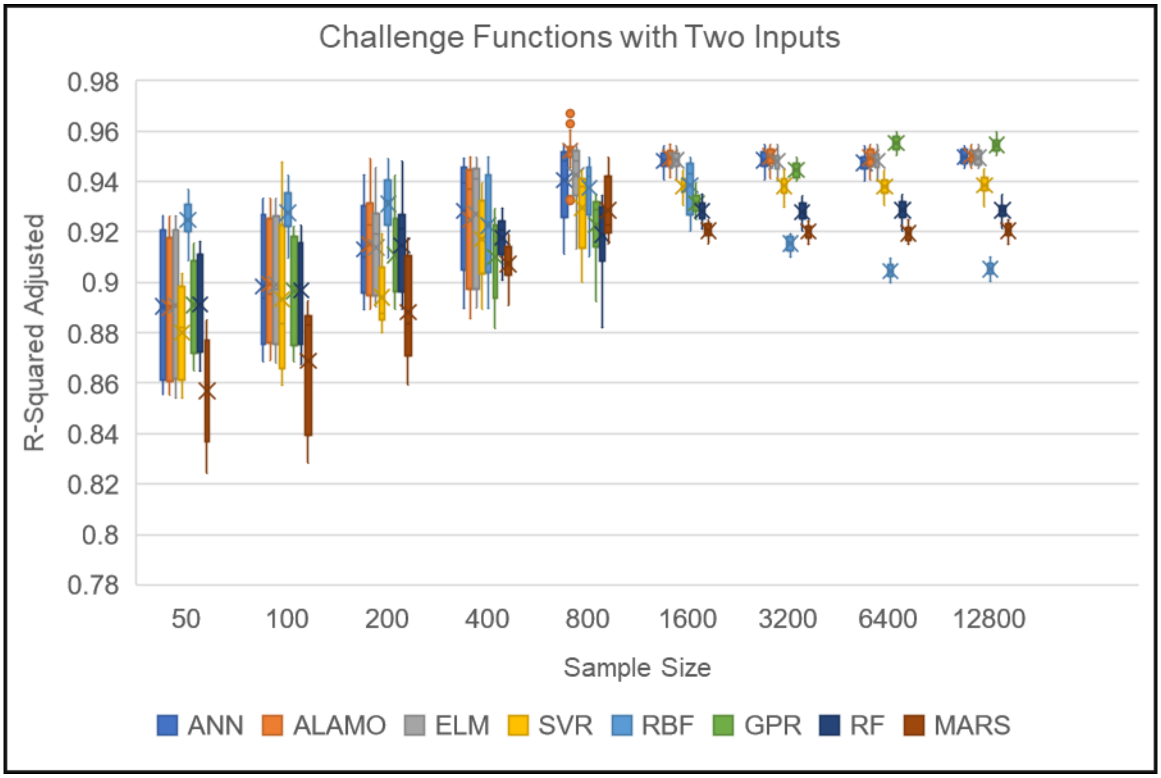


Figure 63. R-Squared Adjusted Values of Challenge Functions with Two Inputs for all Surrogate Models over the range of Sample Sizes

The Akaike Information Criterion for the surrogate models over the range of sample sizes are shown in Figure 64. The graph shows that when sample size increases over 1,600, the error bars and variation decreases. Akaike Information Criterion is useful for comparing different models. When considering the models with a sample size of 1,600 or greater, the values remain constant.

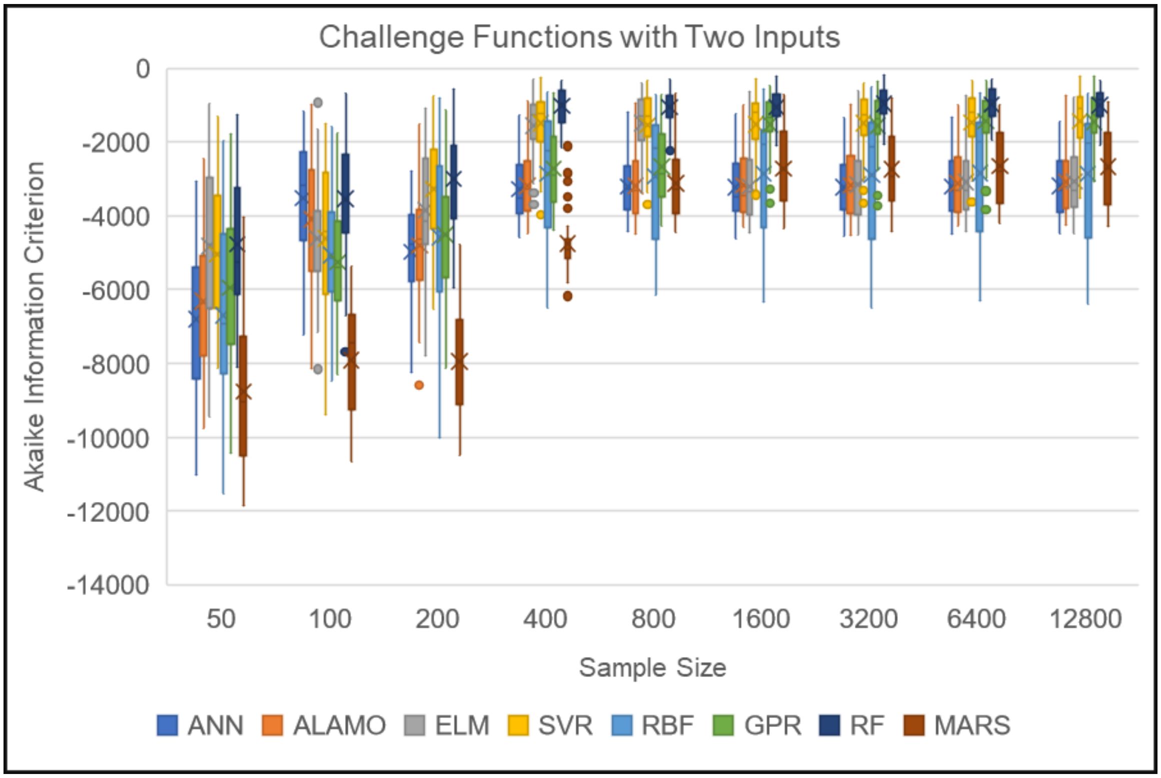


Figure 64. Akaike Information Criterion of Challenge Functions with Two Inputs for all Surrogate Models over the range of Sample Sizes

The Bayesian Information Criterion for the surrogate models over the range of sample sizes are shown in Figure 65. The graph shows that when sample size increases over 1,600, the error bars and variation decreases. Bayesian Information Criterion is useful for comparing different models. When considering the models with a sample size of 1,600 or greater, the values remain constant.

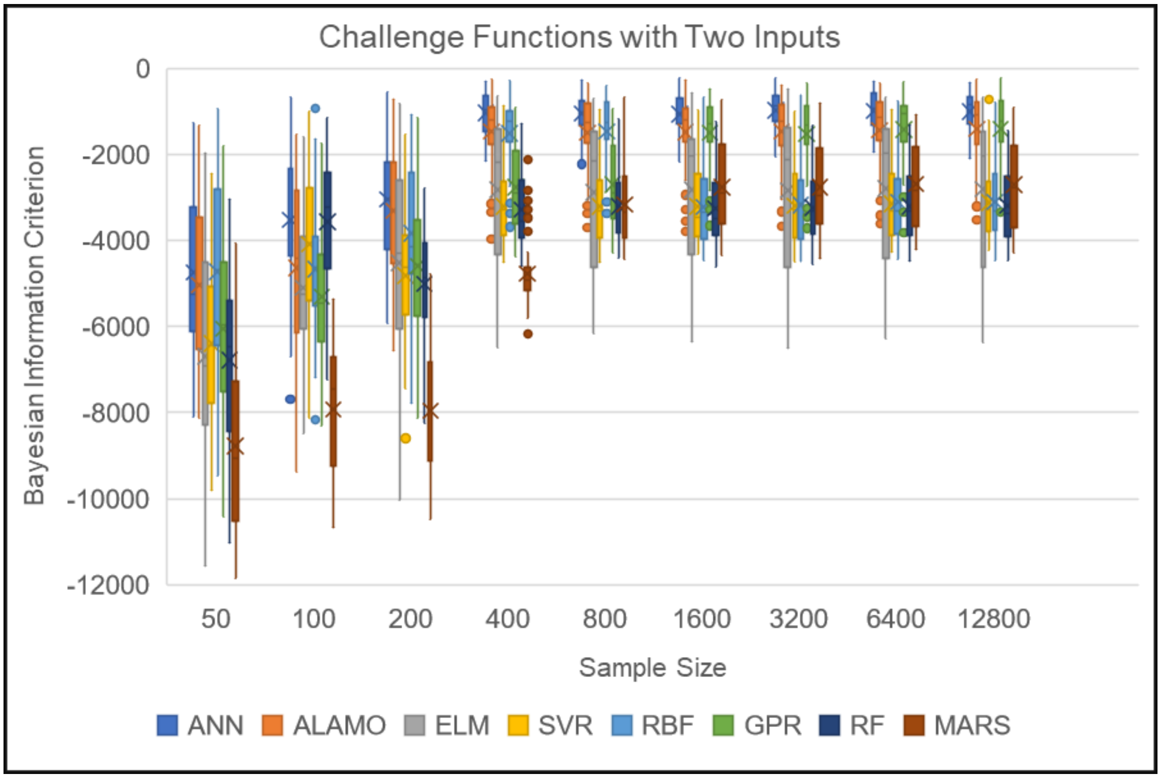


Figure 65. Bayesian Information Criterion of Challenge Functions with Two Inputs for all Surrogate Models over the range of Sample Sizes

The root mean square error for the surrogate models over the range of sample sizes are shown in Figure 66. The graph shows that when sample size increases over 400, the error values decrease. Additionally, the error bars decrease.

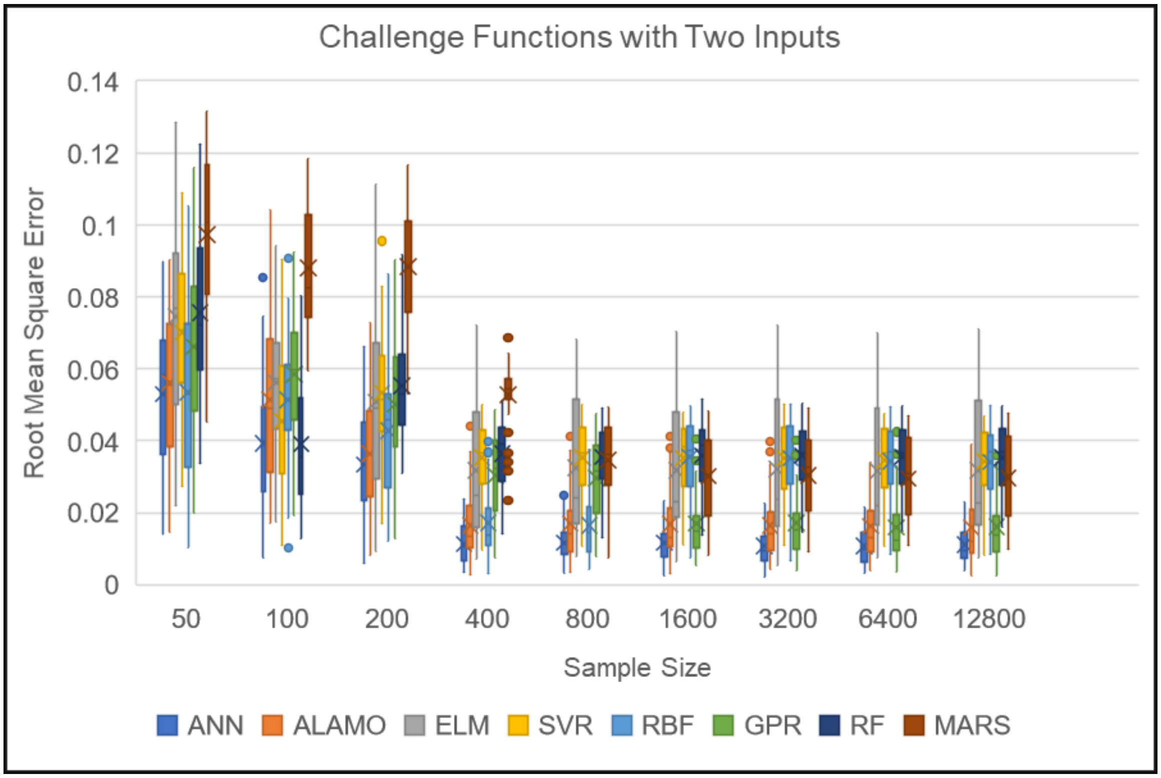


Figure 66. Root Mean Square Error of Challenge Functions with Two Inputs for all Surrogate Models over the range of Sample Sizes

The maximum absolute error for the surrogate models over the range of sample sizes are shown in Figure 67. The graph shows that when sample size increases over 400, the error values decrease, but not as dramatically as with the other error calculation methods. Additionally, the error bars do not seem to decrease over the range of sample sizes.

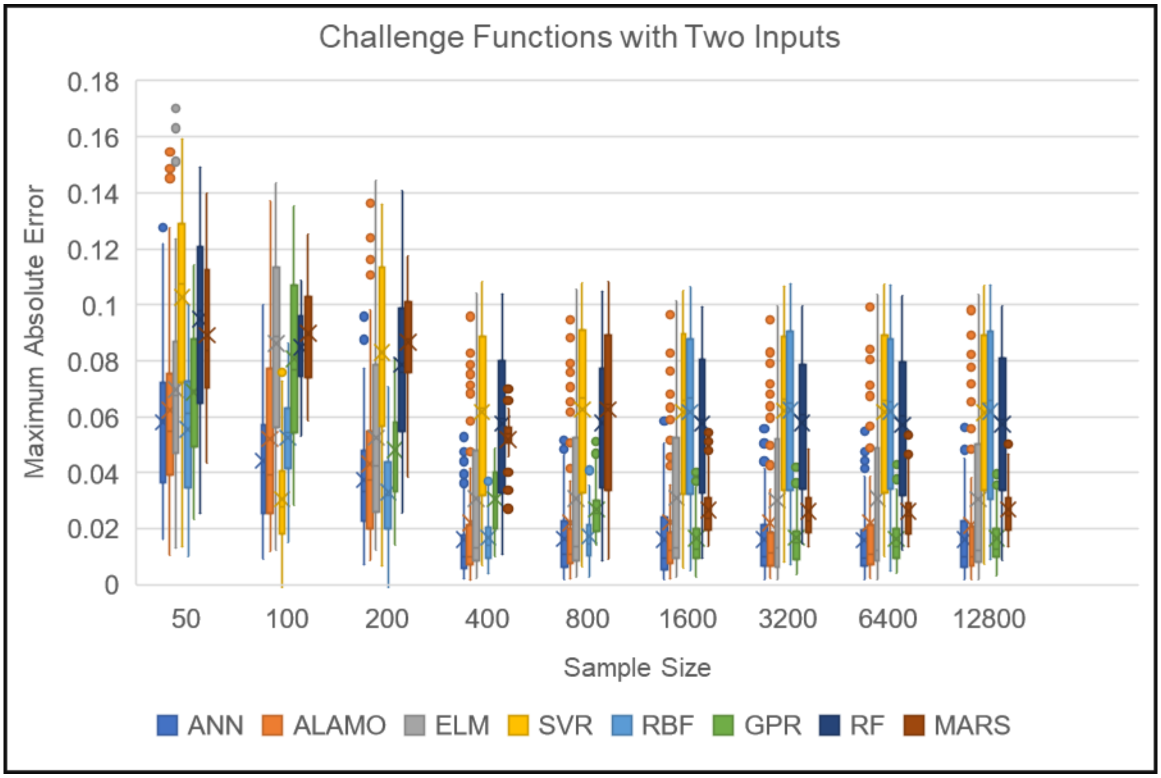


Figure 67. Maximum Absolute Error of Challenge Functions with Two Inputs for all Surrogate Models over the range of Sample Sizes

Based on the discussion above, a sample size of 1,600 balances the increased accuracy and decreased training time. To provide a more detailed view of the impact of surrogate model, the Akaike Information Criterion is displayed in Figure 68. The best performers were Artificial Neural Network, Automated Learning of Algebraic Models for Optimization, and Extreme Learning Machines which all provided very similar levels. Radial Basis Function reached the same value, but with a larger value range. Multivariate Adaptive Regression Splines showed the next best performance.

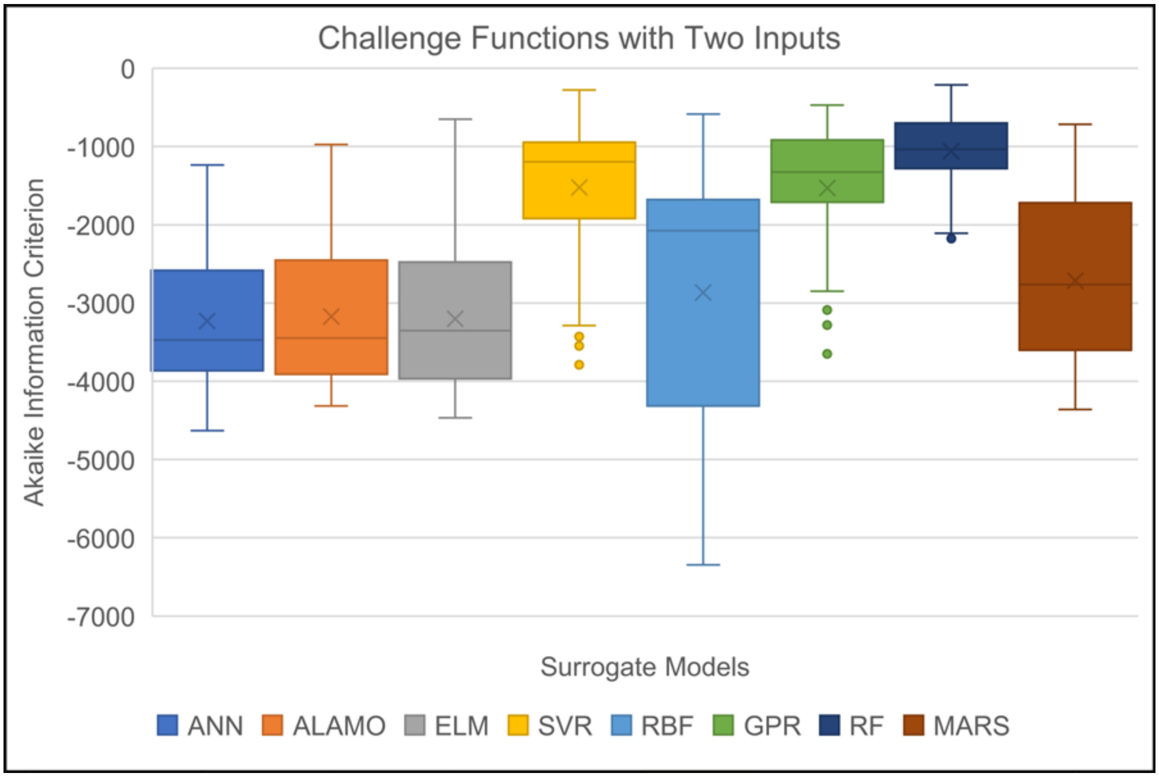


Figure 68. Akaike Information Criterion for Challenge Functions with Two Inputs

Based on the discussion in the previous section, the following sections will provide a detailed view of the surrogate model performance. The graphs will portray the Akaike Information Criterion for a sample size of 1,600 to provide a true comparison of the surrogate models. Additional error calculation graphs were calculated and are included in Appendix A.

4.6.3.2. Results for Challenge Functions with Three Inputs

Challenge functions with three inputs form a group of five functions. Figure 69 shows that that Artificial Neural Network, Automated Learning of Algebraic Models for Optimization, and Extreme Learning Machines provided well and with about the same value. Radial Basis Function, Support Vector Regression and Gaussian Process Regression showed large error bars indicating variation.

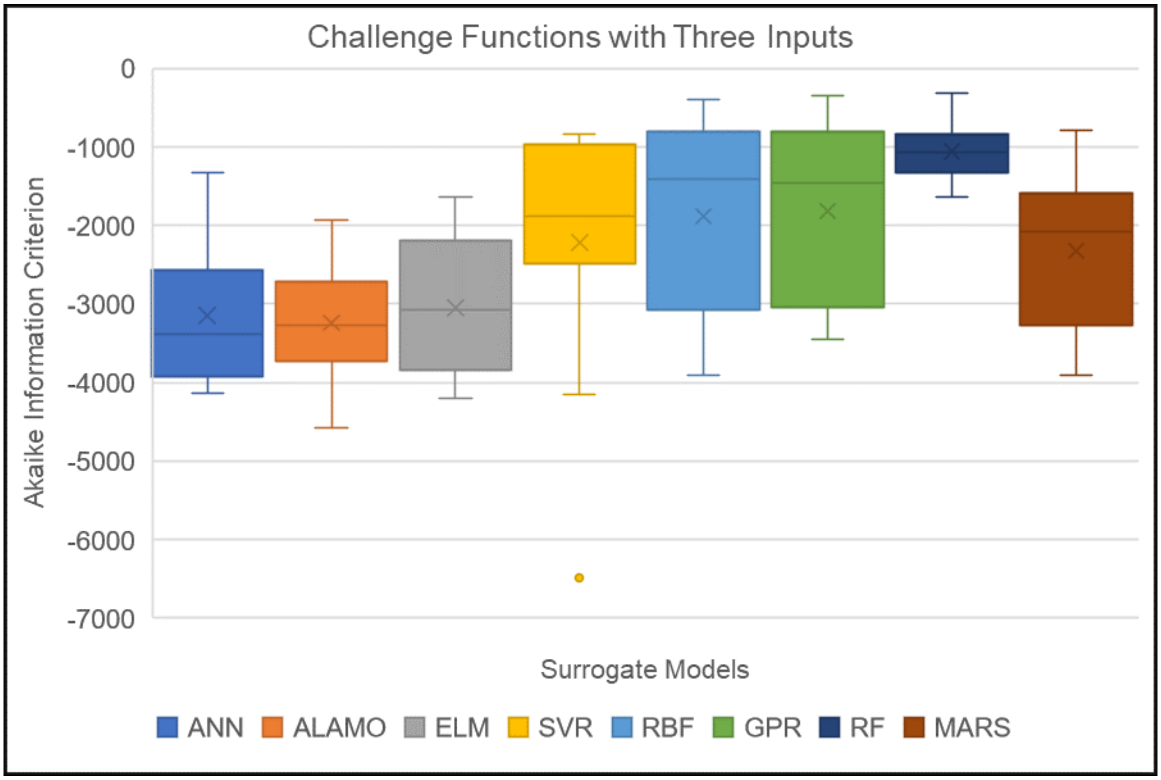


Figure 69. Akaike Information Criterion calculations for Challenge Functions with Three Inputs for all Surrogate Models

4.6.3.3. Challenge Functions with Four Inputs

In the group of challenge functions with four inputs, there are three challenge functions. Figure 70 shows that that Gaussian Process Regression did reach the lowest value but showed the largest range. Artificial Neural Network showed the least variation among the data and showed the lowest average.

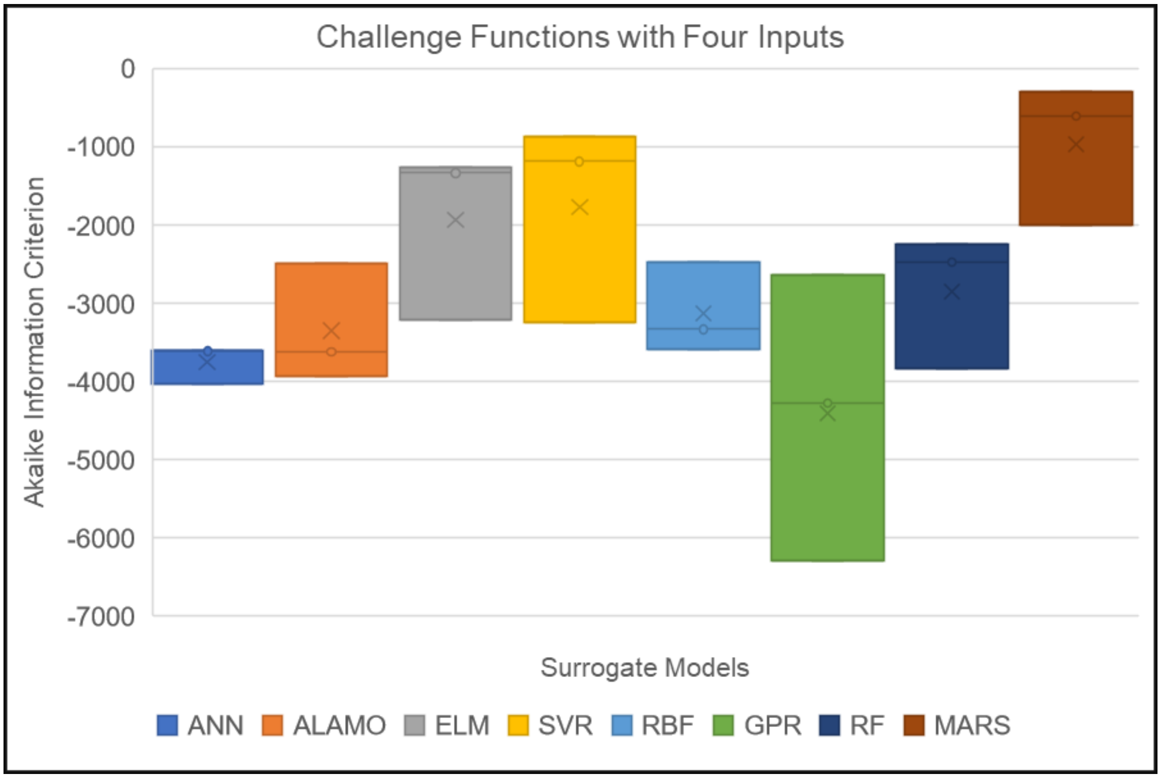


Figure 70. Akaike Information Criterion calculations for Challenge Functions with Four Inputs for all Surrogate Models

4.6.3.4. Challenge Functions with Five Inputs

In the group of challenge functions with five inputs, there are two challenge functions. The Akaike Information Criterion is displayed in Figure 71. Gaussian Process Regression did reach the lowest value but showed the largest range. Artificial Neural Network, Automated Learning of Algebraic Models for Optimization, Support Vector Regression and Multivariate Adaptive Regression Splines provided very similar levels.

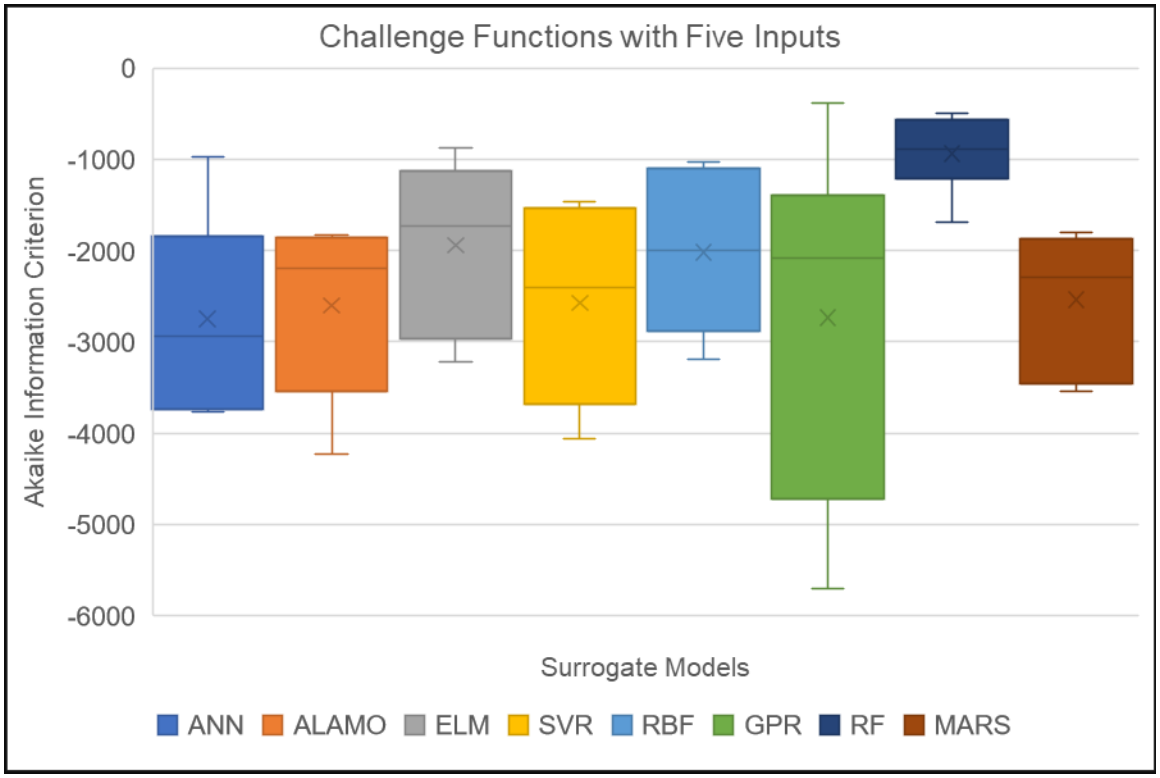


Figure 71. Akaike Information Criterion calculations for Challenge Functions with Five Inputs for all Surrogate Models

4.6.3.5. Challenge Functions with Ten Inputs

In the group of challenge functions with ten inputs, there are six challenge functions. To provide a more detailed view of the impact of surrogate model, the Akaike Information Criterion is displayed in Figure 72. Artificial Neural Network, Automated Learning of Algebraic Models for Optimization, and Gaussian Process Regression provided very similar levels at a low value. Radial Basis Function reached the same value, but with a larger value range. Multivariate Adaptive Regression Splines showed the next best performance.

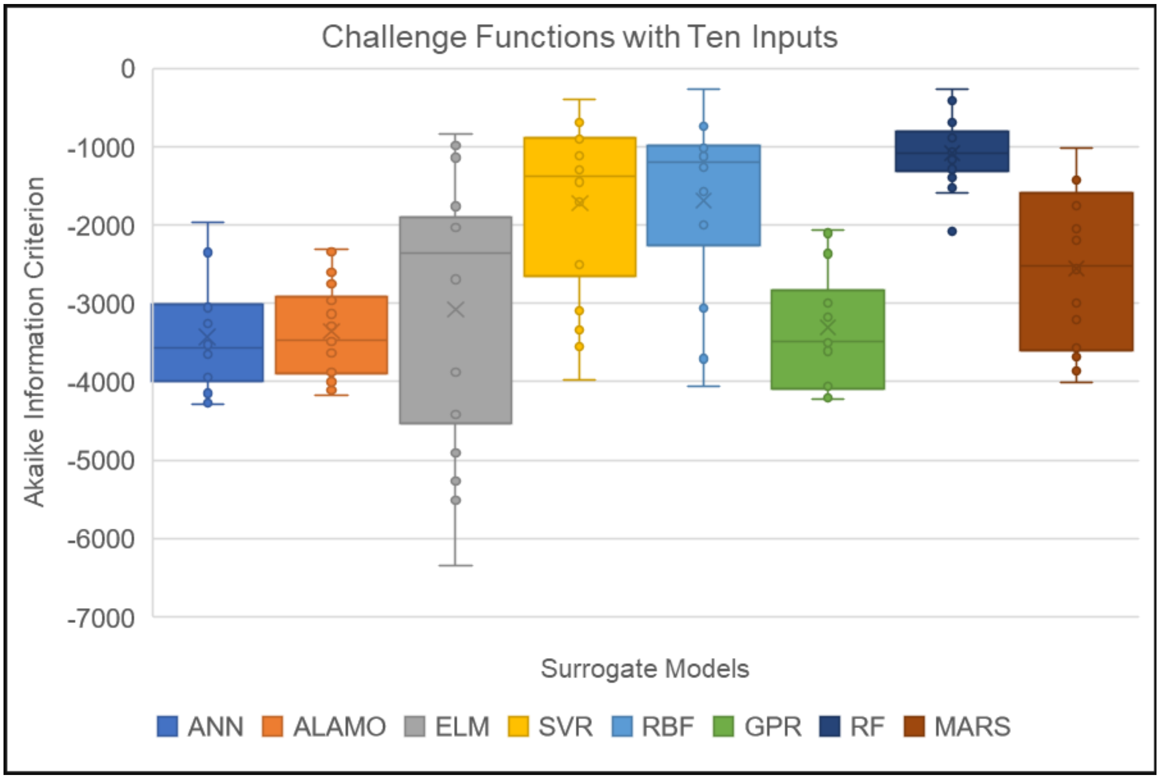


Figure 72. Akaike Information Criterion calculations for Challenge Functions with Ten Inputs for all Surrogate Models

4.6.3.6. Multilocal Minima Challenge Functions

Challenge function whose surface shape has Multilocal minima was the largest group with challenge functions fifteen. Figure 73 portrays the Akaike Information Criterion. Artificial Neural Network, Automated Learning of Algebraic Models for Optimization, Gaussian Process Regression, Radial Basis Function and Multivariate Adaptive Regression Splines provided very similar levels at a low value. However, Radial Basis Function and Mutlivariate Regression Splines reached the same value, but with a larger value range.

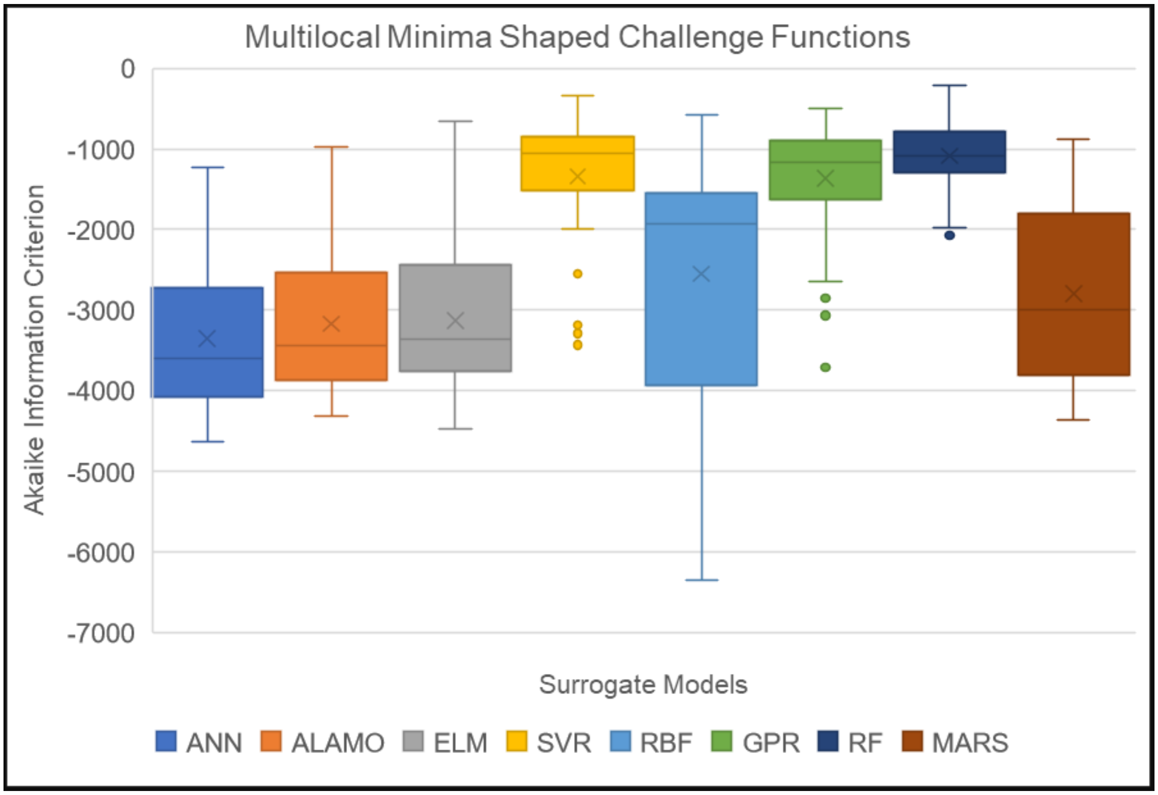


Figure 73. Akaike Information Criterion calculations for Multilocal Minima Shaped Challenge Functions for all Surrogate Models

4.6.3.7. Plate Shaped Challenge Functions

In the group of challenge functions with a plate shaped shape, there are 5 challenge functions. The Akaike Information Criterion is depicted in Figure 74. Automated Learning of Algebraic Models for Optimization and Support Vector Regression both reached the lowest value, however the range and length of the error bars is long. Artificial Neural Network, Random Forests and Multivariate Adaptive Regression Splines provided very similar average values, but with a smaller range.

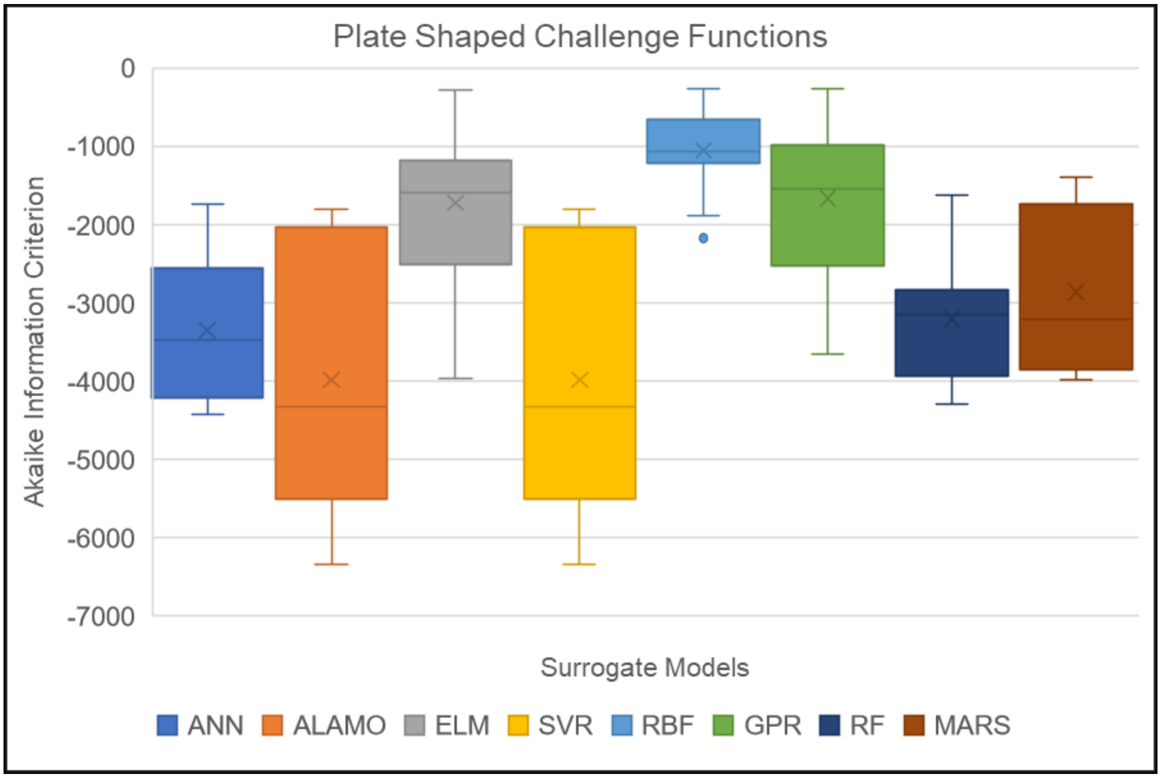


Figure 74. Akaike Information Criterion calculations for Plate Shaped Challenge Functions for all Surrogate Models

4.6.3.8. Ridges and Drops Shaped Challenge Functions

Challenge function whose surface shape has ridges and drops has three challenge functions. Figure 75 shows the Akaike Information Criterion. Artificial Neural Network and Extreme Learning Machines reached the lowest values and their performance was alike. With values only slightly larger, Gaussian Process Regression and Multivariate Adaptive Regression Splines were the next be models.

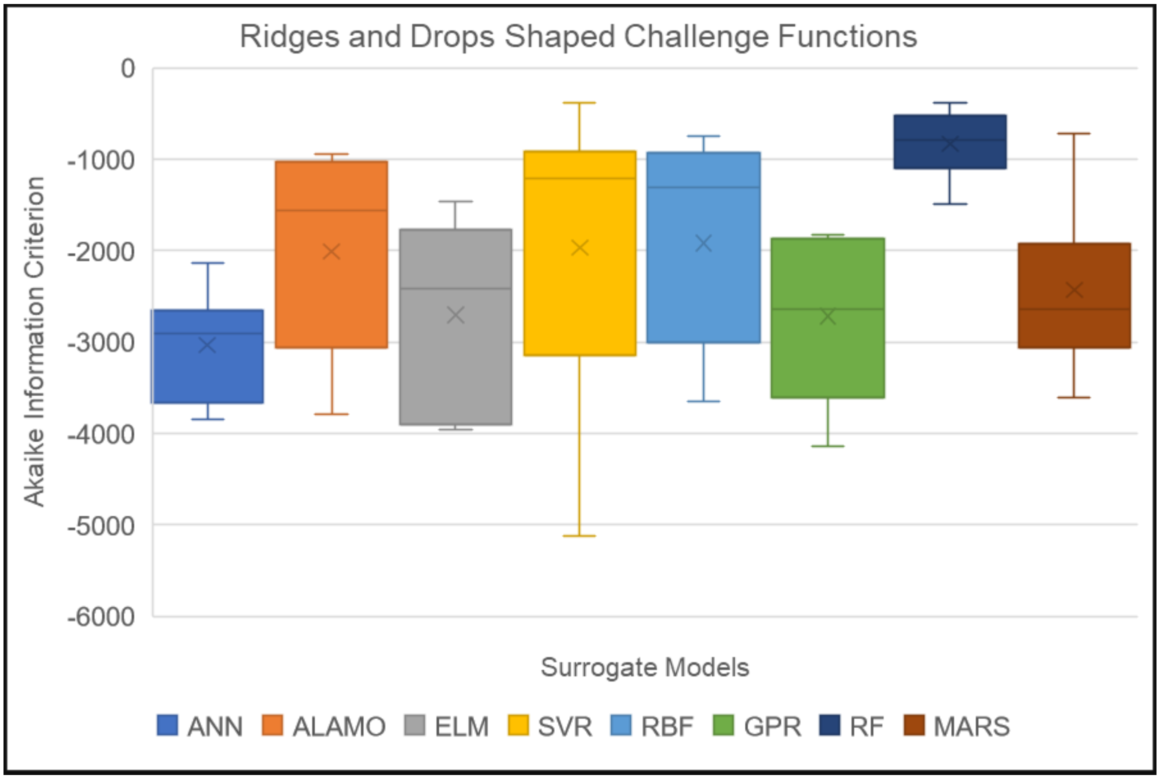


Figure 75. Akaike Information Criterion calculations for Ridges and Drops Shaped Challenge Functions for all Surrogate Models

4.6.3.9. Valley Shaped Challenge Functions

In the group of challenge functions with a plate shaped shape, there are 4 challenge functions. The Akaike Information Criterion is depicted in Figure 76. Automated Learning of Algebraic Models for Optimization, Extreme Learning Machines and Radial Basis Functions all perfumed best with about the same values.

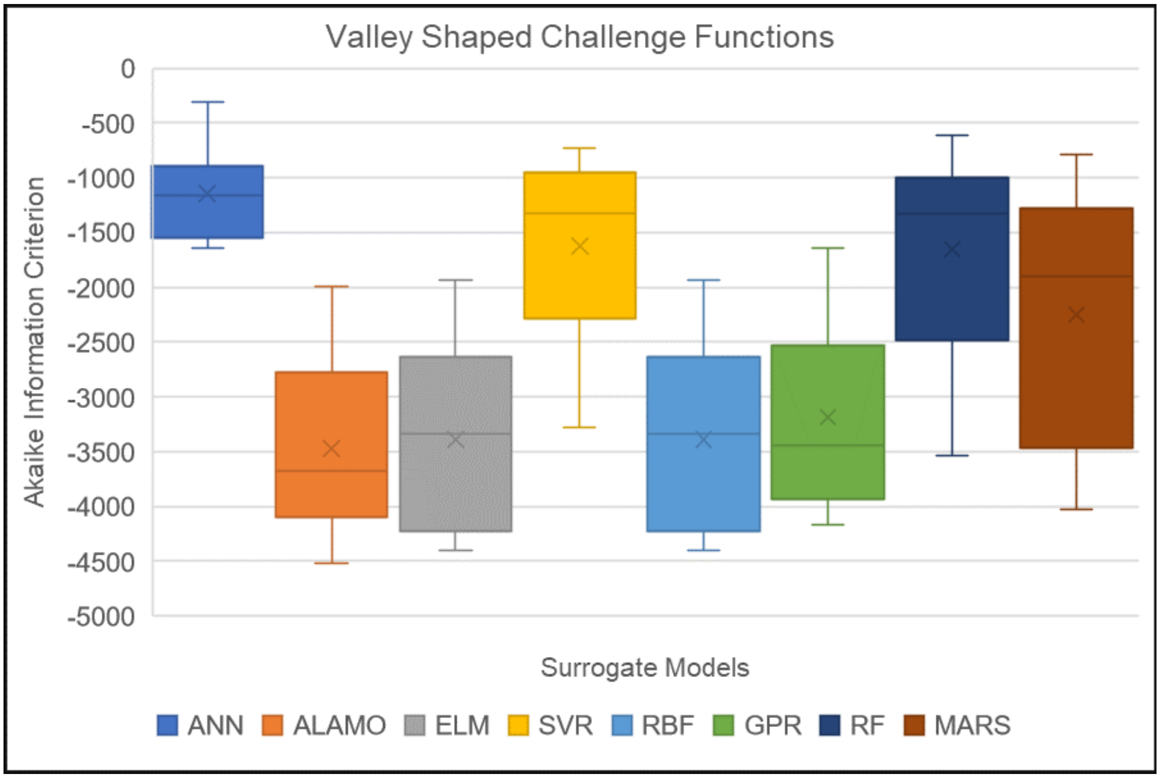


Figure 76. Akaike Information Criterion calculations for Valley Shaped Challenge Functions for all Surrogate Models

4.6.3.10. Bowl Shaped Challenge Functions

In the group of challenge functions with a bowl-shaped shape, there are 7 challenge functions. Figure 77 portrays the Akaike Information Criterion. Extreme Learning Machines shows to have the lowest value, but with the largest range and error bars. Artificial Neural Network and Automated Learning of Algebraic Models for Optimization were the next best performers with similar values.

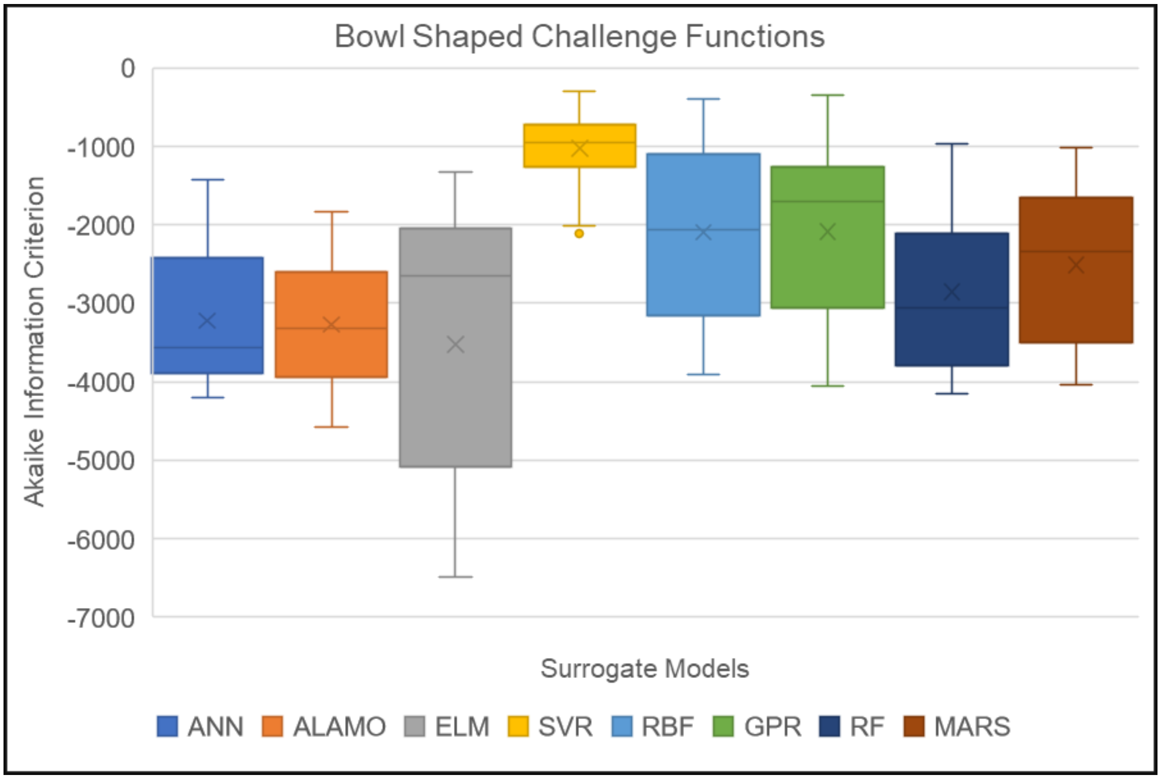


Figure 77. Akaike Information Criterion calculations for Bowl Shaped Challenge Functions for all Surrogate Models

4.7. Conclusions

Larger sample sizes performed more favourably over the smaller sample sizes. Accuracy increased exponentially as sample size increased from 50 to 1,600, however sample sizes of 1,600 or more performed equally as well as those with sample sizes of 12,800. To increase accuracy without increasing complexity, a sample size 1,600 was proven optimal.

For all permutations, Extreme Learning Machines provided the shortest training time while Automated Learning of Algebraic Models for Optimization required the longest training time. Due to relatively short training and evaluation time of all surrogate models, this should not be considered a limiting factor.

The six-performance metrics, R-squared, R-squared adjusted, Akaike Information Criterion, Bayesian Information Criterion, root mean square error and maximum absolute error all provided similar results. Akaike Information Criterion values were used to portray the results, but the other error calculations validated the findings. See Appendix A for detailed error results for each error calculation divided by challenge function shape and number of inputs.

Table 3 displays the first, second and third best performing models based on challenge function surface shape and number of inputs. Artificial Neural Networks performed best for the majority of the challenge functions with Multilocal minima and ridges and drops shape and inputs of two, five and ten. The next best performing model was ALAMO which ranked first for the challenge functions with plate and valley shape and inputs of three. For the challenge functions with a bowl shape, Extreme Learning Machines provided the best fit. Gaussian Process Regression provided the best fit for challenge functions with four inputs. Overall, each type of challenge function did provide a decent fit, however Artificial Neural Networks and ALAMO provided the best fit for the majority of challenge functions.

Average Akaike Information Criteria				
Challenge Function Type	**Number in Group**	**First**	**Second**	**Third**
Mulitlocal Minima Shape	15	ANN	ALAMO	ELM
Plate Shape	7	ALAMO	SVR	ANN
Ridges & Drops Shape	3	ANN	GPR	ELM
Valley Shape	4	ALAMO	ELM	RBF
Bowl Shape	7	ELM	ALAMO	ANN
2 Inputs	20	ANN	ELM	ALAMO
3 Inputs	5	ALAMO	ANN	ELM
4 Inputs	1	GPR	ANN	ALAMO
5 Inputs	2	ANN	GPR	ALAMO
10 Inputs	6	ANN	ALAMO	GPR

Table 3. Akaike Information Criterion Results Sorted by Challenge Function Surface Shape and Number of Inputs

CHAPTER 5. Conclusions and Future Work

5.1. Conclusions

Data driven methods for chemical process and product synthesis have become integrated in all aspects of design. The responsibly of the academic community should be to provide users with guidance when managing the ever-increasing amount of data and possible data analytics methods with a goal of utilizing these new design tools to ensure that their applications provides meaningful results.

Progressive model improvement will lead us to improve characterization techniques to better describe molecules, more advanced modeling methods provide more correct results, and uncertainty management will ensure that the results are more accurate. In addition, improvements to the modeling process, inclusion of additional process design that study the impact on society and sustainability will provide richer results compared to traditional economic and technical solutions. The methods presented in this work illustrate applications of data driven methods for chemical process and product synthesis and design with a focus on two specific tools computer aided molecular design and surrogate modeling.

Computer Aided Molecular Design is a framework that allows us to utilize data to design molecules specific to a process. This is important because it eliminates the need to alter the design to match the available inputs, rather the inputs are modified to match the design. Once issue with this method is that it is reliant on characteristic data for each molecule or building block. The work presented in this dissertation allows us to generate necessary data to apply to the framework thus expanding the possible molecules that can be utilized even further than the computer aided molecular design framework alone.

Surrogate modeling allows us to understand complex or unknown processes to provided understanding of the process and improve designs. Though many studies have sought to test and provide guidance for the application of surrogate models, the results tend to cover only one or two surrogate models. The work presented in this dissertation provides information about the application of those models based on the surface shape and number of inputs. Additionally, it provides information about sampling methods and sizing. Basically, this information can help make an informed decision when selecting which surrogate model, sampling method and group for each type of application.

Both advances provide added depth to data analysis by enhancing current methodologies. This type of work is important because as the modern chemical engineer begins to implement data driven design techniques, the applications that are utilized will need to become more robust and accurate. Progressive model improvement requires that each stage of design should be examined looking for way to improve.

5.2. Progressive Model Improvement

Once a framework, such as computer aided molecular design, has been developed, future study should seek to progressively improve the processes within each step. For example, multiway modeling could provide additional data through the replacement of infrared spectroscopy with excitation-emission fluorescence data or by selecting a more sophisticated modeling method by selecting a surrogate model over a linear model.

5.2.1. Multiway Modeling

Multiway modeling takes the matrix of two-way data with a single value for each variable or dimension one step further by adding additional dimensions. For example, a matrix where data is collected over time. (Bro & Kiers, 2003) Multiway arrays can be decomposed using parallel factor analysis, known as PARAFAC, which is essentially an extension of principal component analysis for two-way to three-way. Essentially, principal component analyses rely on a score and loading matrix; whereas the PARAFAC uses a score matrix and two loading matrices.

When considering the computer aided molecular design problem with ionic liquids presented, improvements from infrared or near infrared to excitation-emission matrix (EEM) fluorescence could provide more descriptive information. (Thygesen & Van Den Berg, 256-270) EEM fluorescence provides a three-dimensional plot of excitation wavelength versus emission wavelength versus fluorescence intensity, an example of which is shown in Figure 78.

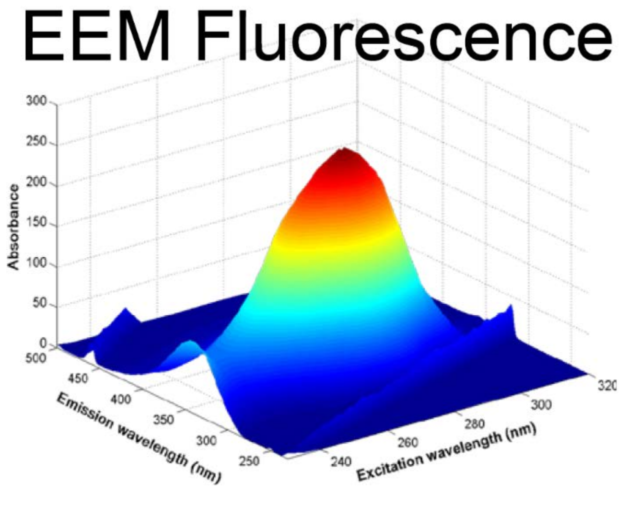


Figure 78. EEM Fluorescence

5.2.2. Modeling Method

In addition to increased descriptor data found in the EEM fluorescence plots or other new characterization techniques, data modeling can be improved. Since linear regression such as principal component analysis or PARAFAC are not valid for all datasets, systems with non-linear and complex relationships can be modeled using surrogate models. The surrogate modeling study presented in this dissertation provides guidance for the selection of the appropriate surrogate model for a given situation. (Bro & Kiers, 2003)

5.2.3. Uncertainty Management

A side effect of all data analysis is uncertainty. This uncertainty can be introduced in every step from data collection through final conclusions. Those uncertainties can have a major impact because they accumulate throughout model calibrations. Future work should focus on including of uncertainty analysis techniques within the modeling frameworks.

5.3. Process Design to Include Social Impact

Technical and economics are the two main considerations for traditional process or product design. However, modern process design is moving towards the inclusion of ecological and social sustainability as illustrated in Figure 79.

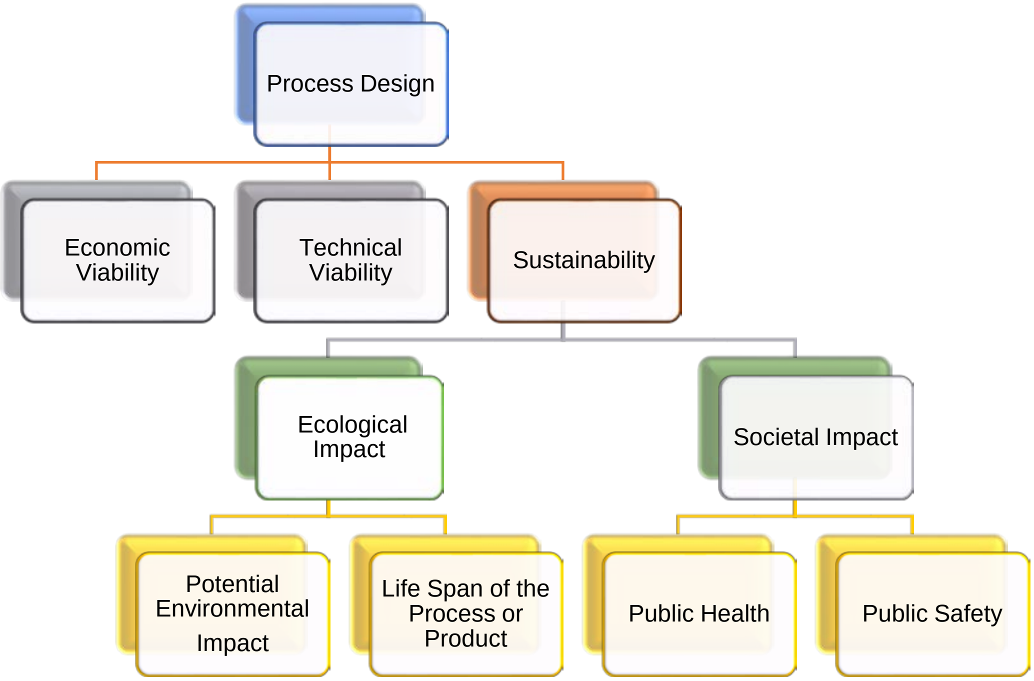


Figure 79. Process Design

When these factors such as human toxicity or environmental impact are included within the model as a part of the target properties, then the resulting design is truly the most ideal from all aspects of design.

CHAPTER 6. References

Akkoyunlu, A. (2010). A Neural Network-Based Approach for the Prediction of Urban SO2 Concentrations in the Istanbul Metropolitan Area. *International Journal of Environment and Pollution*, 301-321.

Ayala, A. E., Simoni, L. D., Lin, Y., & Brennecke, J. F. (2006). Process Design Using Ionic Liquids: Physical Property Modeling. *Computer Aided Chemial Engineering 21*, 463-468.

Baltus, R. E., Culbertson, B. H., Dai, S., Luo, H., & Depaoli, D. (2004). Low Pressure Solubility of Carbon Dioxide in Room Temperature Ionic Liquids Measurement with Quartz Crystal Microbalance. *Journal of Physical Chemistry Edition 108*, 721-272.

Beck, J., Friedrich, D., Brandani, S., & Farga, E. (2015). Multi-Objective Optimization using Surrogate Models for the Design of VPSA Systems. *Computers and Chemical Engineering*, 318-329.

Breiman, L. (2001). Random Forests. *Machine Learning 45.1*, 5-32.

Bro, R., & Kiers, H. (2003). A New Efficient Method for Determining the Number of Components in PARAFAC Models. *Journal of Chemometrics*, 274-286.

Busemeyer, J., & Diederich, A. (2014). Estimation and Testing of Computational Psychological Models. *Neuroeconomics (Second Edition)*.

Candena, C., Anthony, J. L., & Shah, J. K. (2004). Why is CO2 so solubile in Imidazolium-Based Ionic Liquids? *Journal of American Chemical Society*, 5300-5308.

Carrera, G., & Aires-de-Sousa, J. (2005). Estimation of melting points of pyridinum bromide ionic liquids with decision trees and neural networks. *Green Chemistry*, 20-27.

Chemmangattuvalappil, N. G., Eljack, F. T., & Eden, M. R. (2009). A novel algoritm for molecular synthesis using enhanced property operators. *Computers and Chemical Engineering 33(3)*, 636-634.

Chen , S., Cowen, C., & Grant, P. (1991). Orthogonal least squares learning algorithm for radial basis function networks. *IEEE Transactions on Neural Network*, 302-309.

Chi, H., Mascagni, M., & Warnock, T. (2005). On the optimal Halton Sequence. *Mathematics and Computers in Simulation Volume 70 Issue 1*, 9-21.

Cozad, S., & Miller, N. (2014). Learning Surrogate Models for Simulation-Based Optimization. *AIChE Journal 60.6*, 2211-2227.

Davis, S. E., Cremaschi, S., & Eden, M. R. (2017). Efficient Surrogate Model Development: Optimium Model Form Based on Input Function Characteristics. *Computer Aided Chemical Engineering Vol 40*, 457-462.

Davis, S. E., Hada, S., Herring III, R. H., & Eden, M. R. (2014). Characterization Based Reverse Design of Ionic Liquids. *Computer Aided Chemical Engineering*, 285-290.

Davis, S. E., Herring III, R. H., & Eden, M. R. (2016). Reverse Design of Ionic Liquids for the Absoption of CO2. *Computer Aided Chemical Engineering 38*, 1117-1122.

Davis, S., Cremaschi, S., & Eden, M. (2018). Efficient Surrogate Model Development: Impact of Sample Size and Underlying Model Dimensions. *Computer Aided Chemical Engineering, 44*, 979-984.

Davis, S., Cremaschi, S., & Eden, M. (2018). Efficient Surrogate Model Development: Impact of Sample Size and Underlying Model Dimensions. *Computer aided Chemical Engineering, 44*, 979-984.

Denbig, K. (1951). *The Thermodynamics of Steady State.* London: Methuen's Monographs on Chemical Subjects.

Dife, N., & Diwekar, U. M. (2016). Novel Sampling Technique for High Dimensional Stochastic Optimization/Stocastic Programming Problem. *AICHE Annual Meeting.* San Francisco.

Duchowicz, P., Garro, J., & Castro, E. (2008). QSPR study of the Henry's Law Constant for Hydrocarbons. *Chemometrics and Intelligent Laboratory Systems*, 133-140.

Eden, M. (2003). Property Based Process and Product Synthesis and Design. *CAPEC, Department of Chemical Engineering, Technical University of Denmark, Doctoral Dissertation*.

Eden, M. R., Jorgensen, S. B., Gani, R., & El-Halwagi, M. (2004). A Novel Framwork for Simultaneous Separation Process and Product Design. *Chemical Engineering and Processing 43(5)*, 595-608.

Eljack, F. T., Solvanson, C. C., Chemmangattuvalappil, N., & Eden, M. R. (2008). A Property Based Approach for Simultaneous Process and Molecular Design. *Chinese Journal of Chemical Engineering 16(3)*, 424-434.

EPA. (2015, January 1). Retrieved from EPA: http;//www.epa.gov/cleanpowerplan//fact-sheet-overview-clean-power-plan

Eriksson, L., Byrne, T., Johansson, E., Trygg, J., & Vikstrom, C. (2006). Multi and Megavariate Data Analysis: Basic Principles and Applications. *Umetrics Academy*.

Eriksson, L., Johansson, E., Kettaneh-Wold, N., Trygg, J., Wikstrom, C., & Wold, S. (2006). Multi-and Megavariate Data Analysis: Basic Principles and Applications (Part I). *Umetrics Academy*, 425-460.

Erisson, L., & Johansson, E. (1996). Multivariate Design and Modeling in QSAR. *Chemometrics and Intelligent Laboratory Systems 34(1)*, 1-19.

Freidman, J. (1991). Multivariate Adaptive Regression Splines. *Annals of Statistics 19.1*, 1-67.

Gabrielsson, J., Lindberg , N. O., & Lundsteadt, T. (2002). Multivariate Methods in Pharmaceutical Applications. *Journal of Chemometrics 16(3)*, 141-160.

Gani, R. (2004). Chemical product design: challanges and opporunties. *Computers and Chemical Engineering 28(12)*, 2441-2457.

Gani, R., & Constantinou, L. (1994). New group contribution method for estimating properties of pure compounds. *AICHE Journal 40(10)*, 1697-1710.

Gaussian 09 Revision. (n.d.).

Geladi, P., & Kowalski, B. R. (1986). Partial Least-Squares Regression: A Tutorial. *Analytica Chama Acta 185*, 1-17.

Hada, S. (2011). Molecular Desing of Biofuel Additives for Optimization of Fuel Characteristics. *21st European Symposium on Computer Aided Process Engineering*, 1633-1637.

Hada, S. (2013). Chemical Product Formulation through Multivariate Characterization, Modeling, and Design in Property Cluster Space. *Doctoral Disseration, Auburn University*.

Hada, S., Herring III, R. H., Davis, S. E., & Eden, M. R. (2015). Multivariate characterization, modeling and design of ionic liquid molecules. *Computers and Chemical Engineering*, 245-259.

Halgren, T. (1996). Merck Molecular Force Field. *Journal of Computational Chemistry 17(5-6)*, 490-519.

Halgren, T. (1996). Merck Molecular Force Field: Basis, form, scope parameterization and performance of MMFF94. *Journal of Computational Chemistry*, 490-519.

Hall, L. H., & Kier, L. B. (2001). Issues in Representation of Molecular Structure: The Development of Molecular Connectivity. *Journal of Molecular Graphics and Modelling*, 4-18.

Hanwell, M. D., Curis, D. E., Lonie, D. C., Vandermeesch, T., Zurek, E., & Hutchinson, G. R. (2012). Avogadro: An Advanced Semantic Chemical Editor, Visulization, and Analysis Platform. *Journal of Chemoinformatics*, 17-25.

Harper, P. M., & Gani, R. (1999). Computer-aided molecular design with combined molecular modeling and group contribution. *Fluid Phase Equilibria*, 337-347.

Harper, P. M., & Gani, R. (2000). Computer aided tools for design/selection of environmentally firnedsly substances. *Process Design Tools for Environment*, 371-404.

Hasib-ur-Rahman, M., Siaj, M., & Larachi, F. (2010). Ionic Liuiqds for CO2 Capture: Development and Process. *Chemical Engineering and Processing: Process Intensification 49(3)*, 313-322.

Haykin, S. (2009). *Neural Networks and Learning Machines.* Prentice Hall PTR.

Holbrey, J. D., & Seddon, K. D. (1999). Ionic Liquids. *Clean Producs and Processes Volume 1*, 223-236.

Huang, G. B., Zhu, Y., & Siew, K. (2015). Extreme Learning Machine: Theory and Applications. *Process Safety and Environmental Protection*, 111-124.

Hyndman, R., & Koehler, A. (2006). Another look at measures of forcast accuracy. *International Journal of Forecasting*, 679-688.

Ionic Liquids Databased (ILThermo). (2014, December 1). *National Instituate of Standards and Technology*. Retrieved from NIST Standard Reference databased #147: http://ilthermo.boulder.nist.gove/ILThermo/mainmenu.uix

Jackson, J. (1991). A User's Guide to Principal Components. *Jounral of the Operational Research Society Volume 43*, 641-661.

Jaeckle, C. M. (1998). Product Design through multivariate statisical analysis of process data. *AICHE Journal 44(5)*, 1105-1118.

Jaeckle, C. M., & Macgregor, J. F. (1998). Product Design through Multivariate Statistical Analysis of Process Design. *AICHE Journal 44(5)*, 1105-1118.

Jin, R., Chen, W., & Simpson, T. (2001). Comparative Studies of Metamodeling Techniques under Multiple Modeling Critera. *Structural and Multidisciplinary Optimzation 23.1*, 1-13.

JMP Version 9.0. (1989-2012). SAS Institute Inc.

Joback, K. G., & Reid, R. C. (1983). Estimation of Pure Component Properties from Group Contributions. *Chemical Engineering Communication 57*, 57-233.

Johnson, R. (2007). *Applied Multivariate Statistical Analysis 6th Edition pg 800.* New York: Pearson.

Jollieffe, I. (2002). *Principal Component Analysis.* New York: Springer.

Karmer, R. (1998). Chemometrics Techniques for Quanitative Analysis. *CRC*, 203.

Kettanch-Wold, N. (1992). Analysis of Mixture Data with Partial Least Squares. *Chemometrics and Intelligent Laboratory Systems*, 57-69.

Kiers, H., & Bro, R. (2003). A new efficient method for determining the number of components in PARAFAC models. *Journal of Chemometrics 17(5)*, 274-286.

Linusson, A., Elofsson, M., Anderson, I. E., & Dehlgren, M. K. (2010). Statistical Molecular Desing of Balanced Compound Libraries for QSAR modeling. *Current Medicinal Chemistry 17(19)*, 2001-2016.

Linusson, A., Gottfires, J., Lingren, F., & Wold, S. (2000). Statistical Molecular Design of Building Blocks for Combinatorial Chemistry. *Journal of Medicinal Chemistry*, 1320-1328.

Macgregor, J., & Muteki, K. (2007). Multi-BLock PLS Modeling for L-Shape Data Structures with Applications to Mixture Modleing. *Chemometrics and Intelligent Laboratory Systems*, 186-194.

Marrero, J., & Gani, R. (2001). Group-Contribution Based Estimation of Pure Component Properties. *Fluid Phase Equilibria*, 183-208.

Matsuda, H., Yamaoto, H., Kurihara, K., & Tochigi, K. (2007). Computer-aided Reverse Design for Ionic Liquids by QSPR using descriptors of group contribution type for ionic conductivies and viscocities. *Fluid Phase Equilibria 261(1-2)*, 434-443.

McKay, M., Beckman, R., & Conover, W. (1979). Comparision of three moethods for selecting valuves of in put variables in the analysis of output from a computer code. *Technometrics 21.2*, 239-245.

McLeese, S. E., Eslick, J. C., Hoffmann, N. J., Scurto, A. M., & Camarada, K. V. (2010). Design of Ionic liquids via computation molecular design. *Computers and Chemical Engineering*, 1476-1480.

Mirbagheri, S. (2015). Evaluation and Prediction of Membrane Fouling in a Submerged Membrane Bioreactor with Simultaneous Upward and Downward Aeration using Artificial Neural Network Genetic Algorithm. *Process Safety and Environmental Protection 96*, 111-124.

Mohammed, E., & Far, B. (2015). Emerging Buisiness Intelligence Framework for a Clinical Laboratory Throught Big Data Analytics. *Emergency Trends in Computation Biology, Bioinformatics, and Systems Biology*.

Montgomery, D., Runger, G., & Hubele, N. (2009). *Engineering Statistics.* J Wiley and Sons.

Muteki, K. (2006). Mixture Product Design Using Laten Variable Methods (Doctoral Dissertation). *Department of Chemical Engineering McMaster University*.

Navid, A., Khalilarya, S., & Abbasi, M. (2018). Disel engineer optimization with multi-objective performance characteristics by non-evolutionoary Nelder-Mead algorithm: Sobal Sequence and Latin Hypercube sample methods comparision in DoE process. *The Science and Technology of Fuel*, 349-367.

Scott, A. R. (1996). Harmonic Vibrational Frequencies: An Evaluations of Hartree-Fock, Moller-Plesset, Quadratic Configuations Interacation, Density Function Theory and Semiemperial Scale Factors. *Journal of Physical Chemistry*, 16502-16513.

Scott, A., & Random, L. (1996). Harmonic Vibrational Frequencies: An Evaluation of Hartree-Fock, Moller-Plesset, Quadratic Configuration Interaction, Density Functional Theory, Semiemperical Scale Factors. *Journal of Physical Chemistry*, 16502-16513.

Shelley, M. D., & El-Halwagi, M. M. (2000). Compenent-less design of recovery and allocation systems: a functionality-based clustering approach. *Computers and Chemical Engineering*, 2081-2091.

Socrates, G. (2001). Infrared and Raman Characteristic Group Frequencies: Tables and Charts. *Journal of Raman Spectoroscopy Vol 35*, 347-356.

Solvason, C. (2011). Integrated Multiscale Product Design using Property Clustering and Decomposotion tehcniques in a Reverse Problem Formulation (Doctoral Dissertation). *Department of Chemical Engineering, Auburn University, Auburn, Alabama*.

Surjanovic, B. (2015). *http://www.sfu.ca/~ssurjano*. Retrieved from Virtual Library of Simulatioin Experiements: Test Functions and Datasets.: http://www.sfu.ca/~ssurjano

Talaty, E. R., Raja, S., Storhaug, V. J., Do, A., & Carper, W. R. (2004). *Raman and Infrared Spectra and ab Initio Calculations of C 2-4 MIM Imidazolium Hexafluorophosphate Ionic Liquids.* Journal of Physical Chemistry 108: 13177-13184.

Thygesen, J., & Van Den Berg, F. (256-270). Calibration transfer for excitation-emmission fluorescence data with multiway analysis methods and artificial neural networks: an operations tool for improved drinking water treatment. *Environmetrics 22(3)*.

Tjur, T. (2009). Coefficients of determination in logistic regression models - A new proposal: The Coefficient of discrimination. *The American Statistician 63.4*, 366-372.

Turner, E. A., Pye, C. C., & Singer, R. D. (2003). Use of ab Initio Calculations toward the Rational Design of Room Temperature Ionic Liquids. *Journal of Physical Chemistry*, 2277-2288.

Turner, E. P. (2003). Use of ab Initio calculations toward the Rational Design of Room Temperature Ionic Liquids. . *Journal of Physical Chemistry*, 2277-2288.

Varnek, A., Kireeve, N., Tetko, I. V., Baskin, I. I., & Solov'ev, V. P. (2007). Exhasutive QSPR studies of a large diverse set of ionic liquids: How accuartely can we predict melting points? *Journal of Chemical Information and Modeling 47(3)*, 1111-1122.

Wasserscheild, P., & Welton, T. (2007). *Ionic Liquids in Synthesis 2nd Edition.* Wiley.

Willmott, C., & Matsuura, K. (2005). Advantages of the absolute error over the root mean square error in assessing average model performance. *Climate Research*, 79-82.

Wold, S. (1995). Chemometrics: What do we mean with it and what do we want from it? *Chemometrics and Intelligent Laboratry Systems 30(1)*, 109-115.

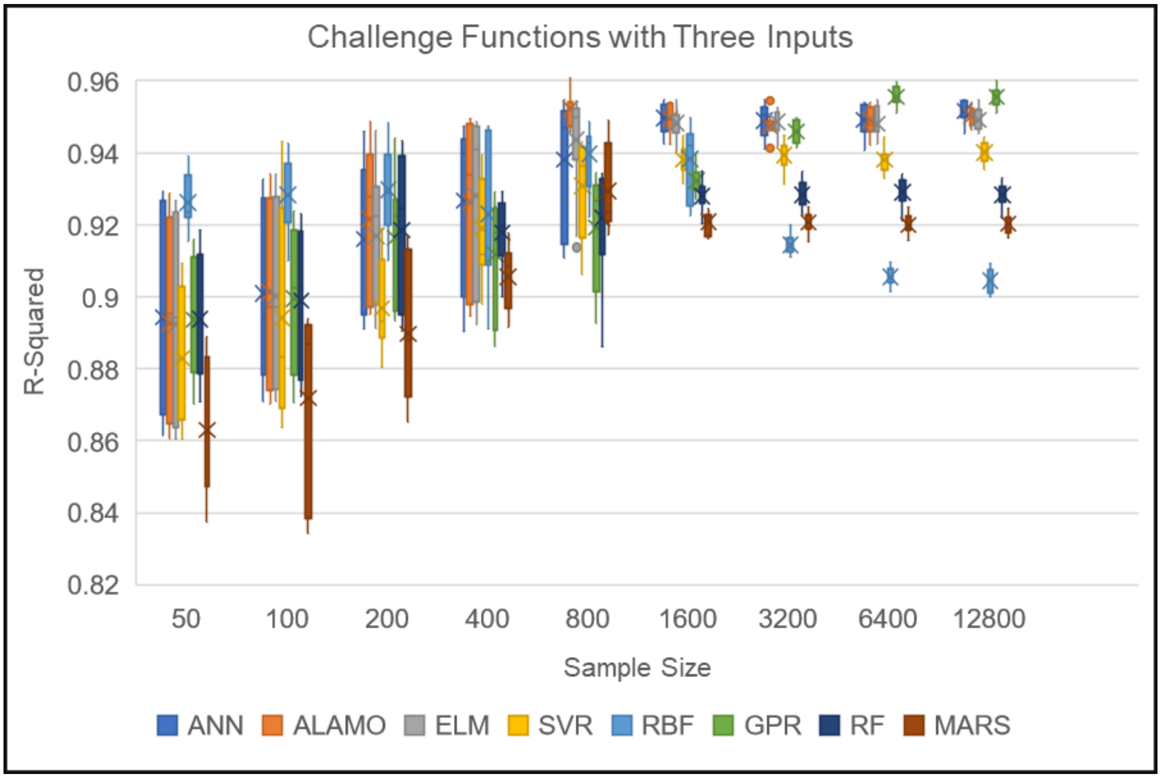
Wold, S., Kettaneh, N., & Tjessem, K. (1996). Hierachical Multiblock PLS and PC Modesl for Easier Model Interpretations and as an Alternative to Variable Selection. *Journal of Chemometrics 10*, 463-482.

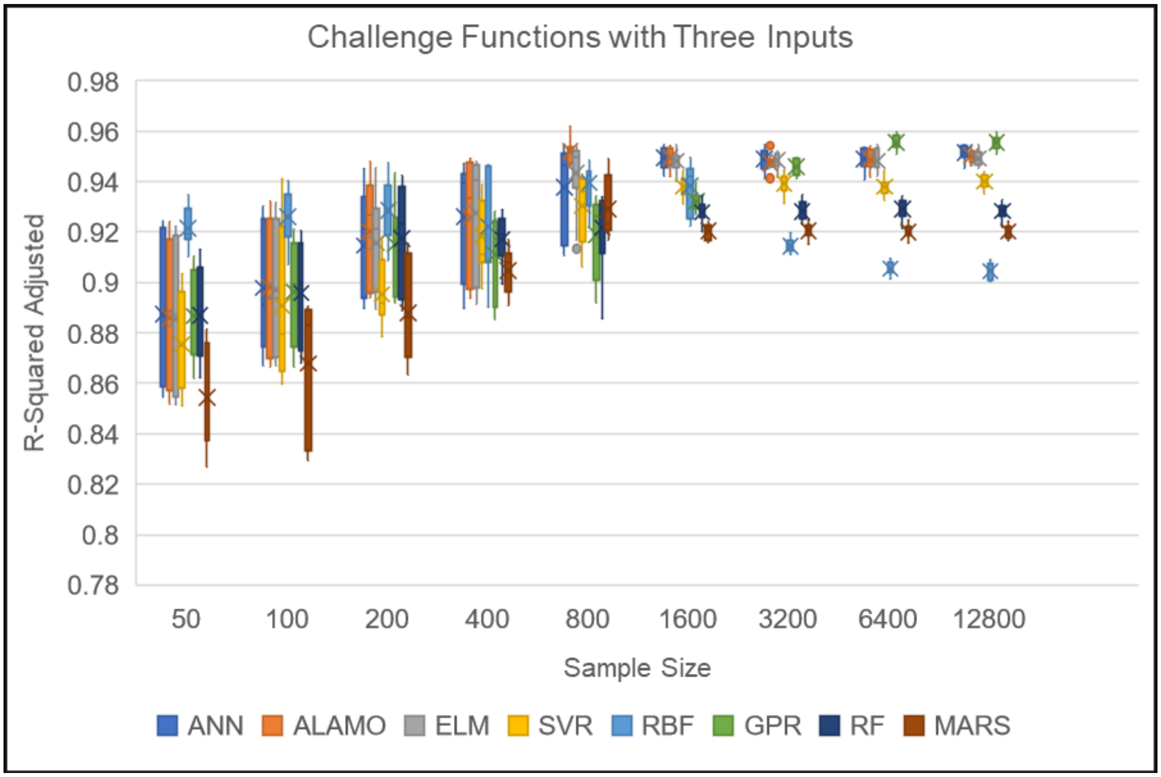
Workman, J. W. (2008). Practical Guide to Interpretive Near Infared Spectroscopy. *CRC Press 1st Edition*, 344-357.

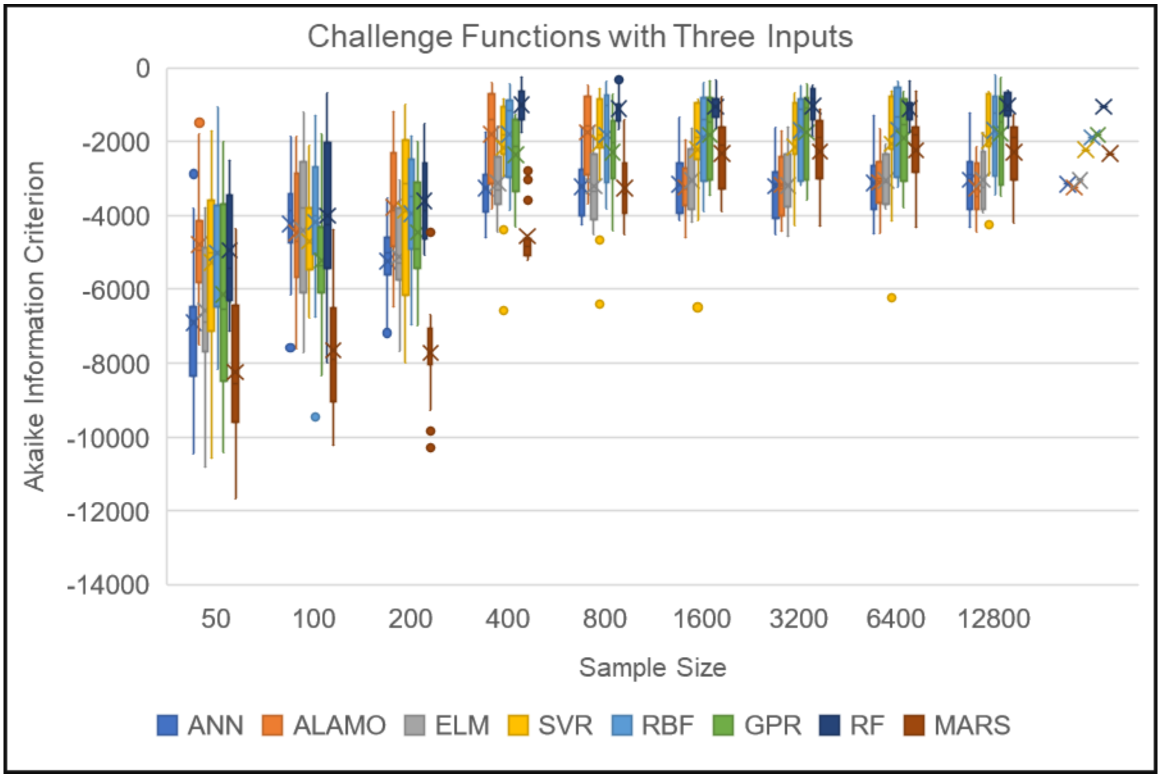
CHAPTER 7. Appendix A

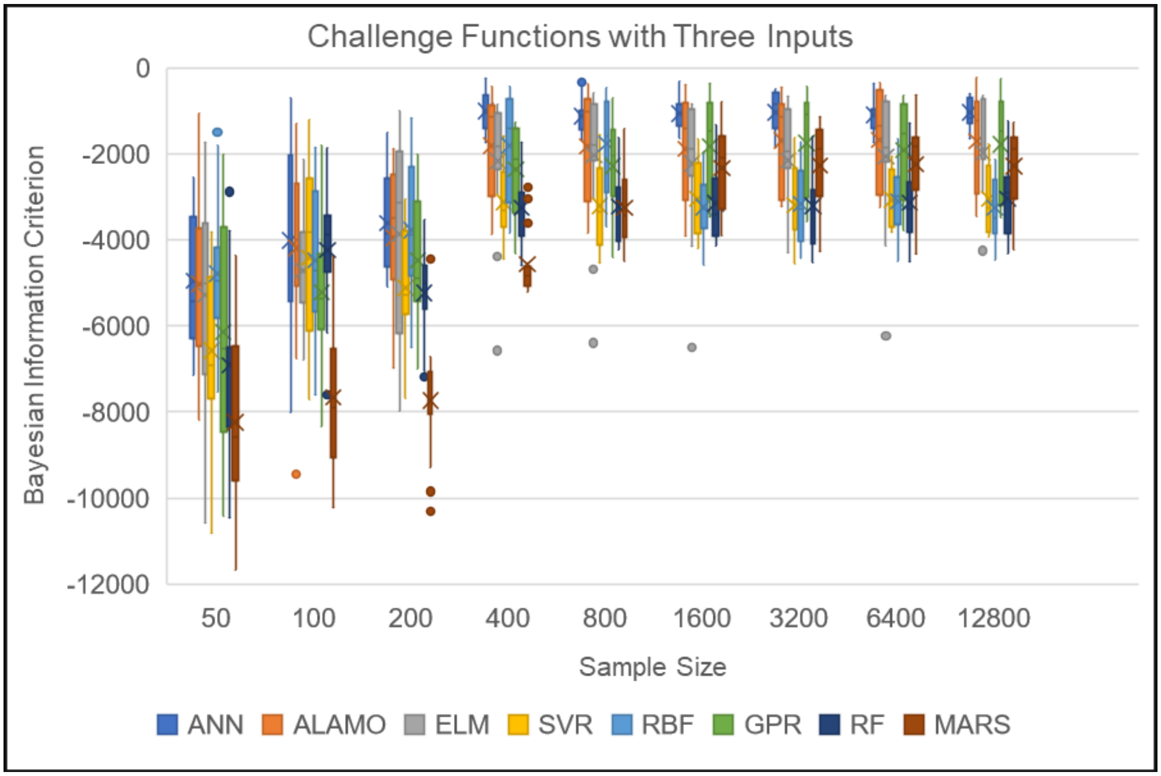
The following graphs show the results of each group of challenge functions over the range of sample sizes for each surrogate model. This in-depth view of the results is presented in Section 4.6.3.1 for challenge functions with two inputs. These graphs provided insight that a sample size of 1,600 was optimal. Additionally, these graphs provided similar results regardless of the error calculation method. As a result of these observations, the remaining challenge function groups were discussed based on a plot of Akaike Information Criterion with a sample size of 1,600. The comparisons for the remaining groups are presented in the section for reference.

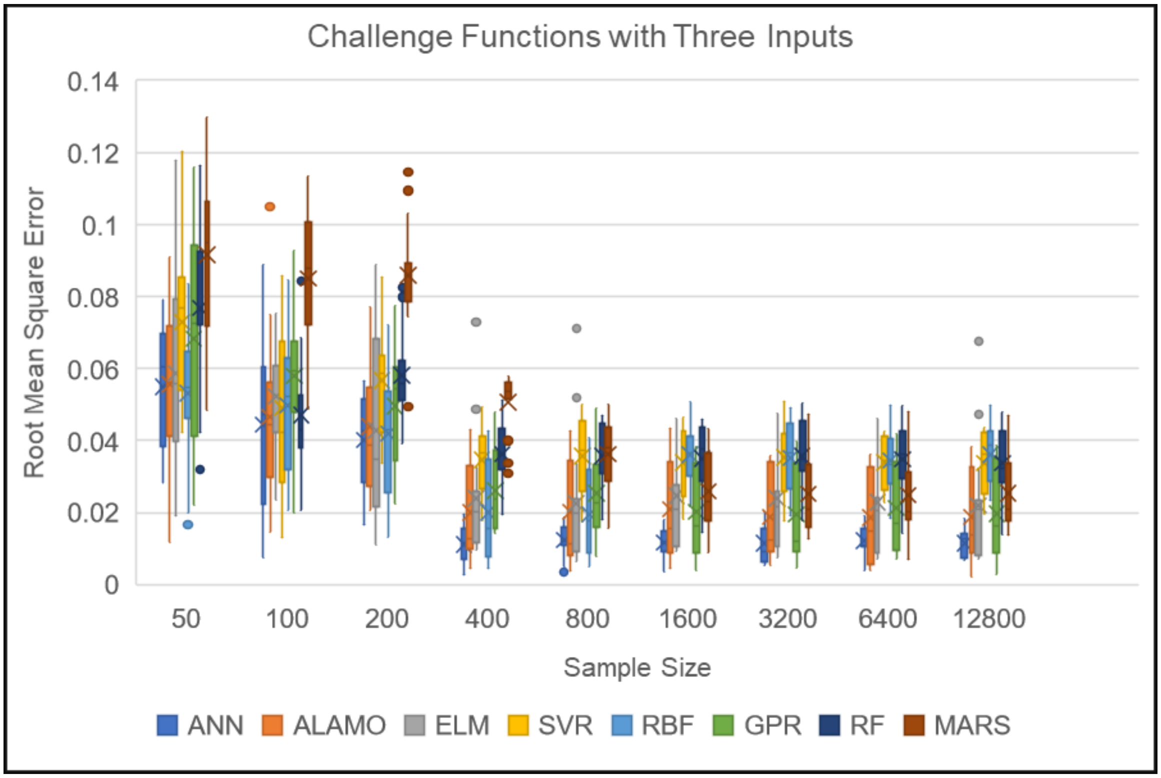
7.1. Challenge Functions with Three Inputs

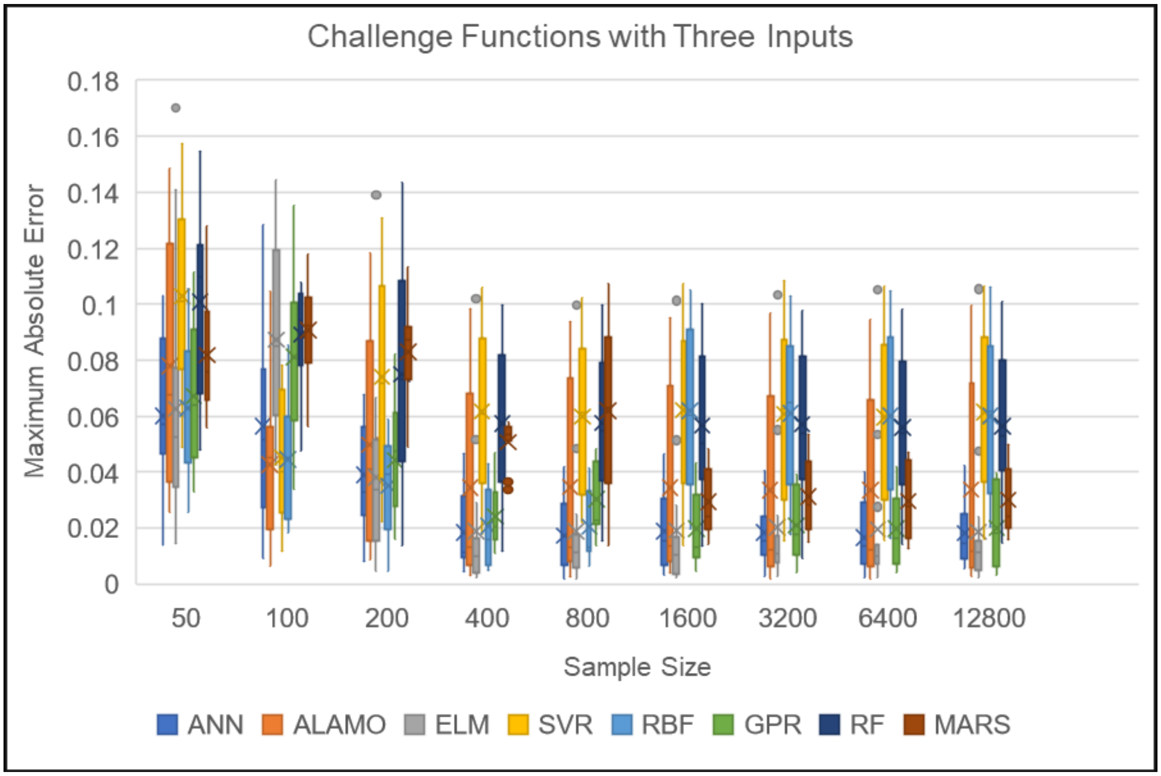




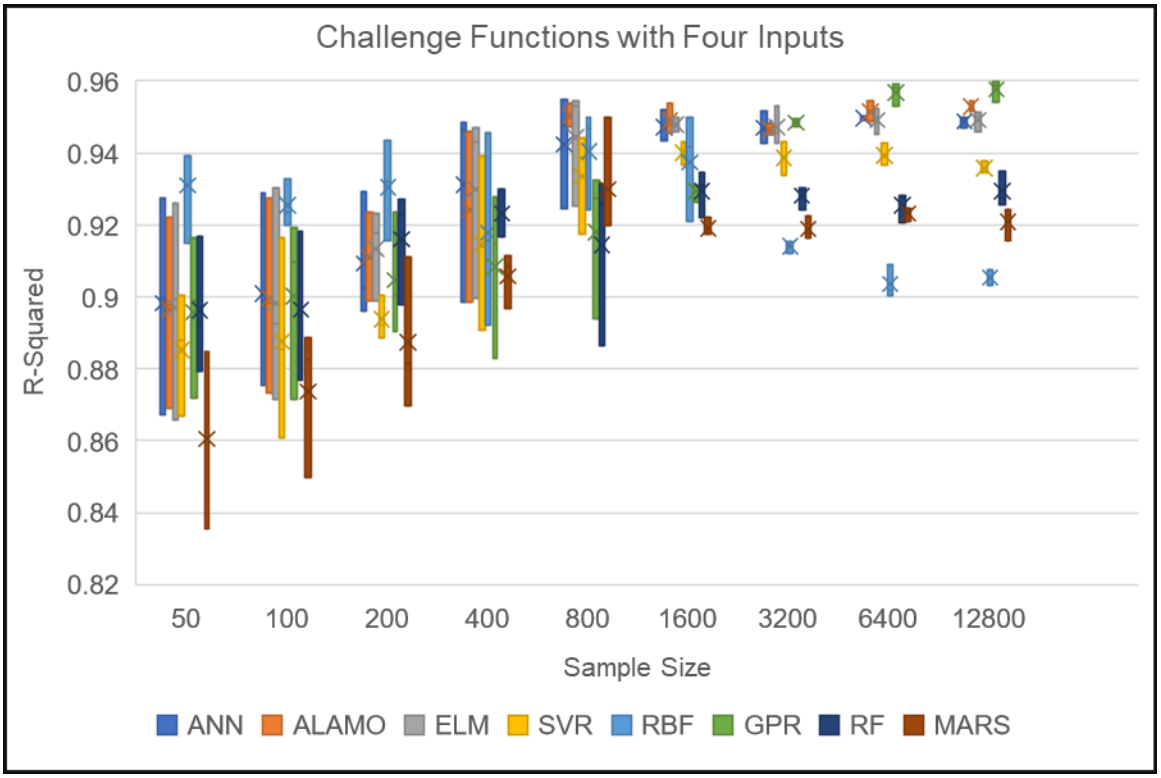

Challenge Functions with Three Inputs
Akaike Information Criterion
0
-2000
-4000
-6000
-8000
-10000
-12000
-14000
50
100
200
400
800
1600
3200
6400
12800
Sample Size
ANN
ALAMO
ELM
SVR
RBF
GPR
RF
MARS

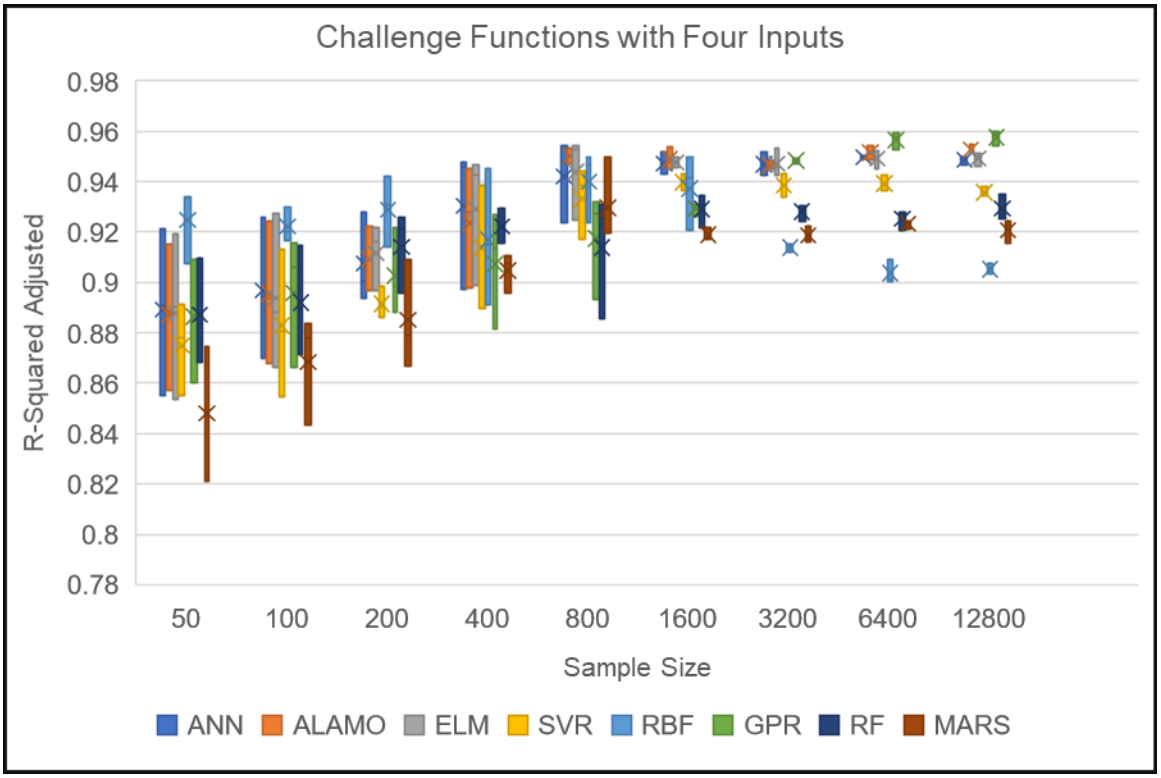

Challenge Functions with Three Inputs
Bayesian Information Criterion
0
-2000
-4000
-6000
-8000
-10000
-12000
50
100
200
400
800
1600
3200
6400
12800
Sample Size
ANN
ALAMO
ELM
SVR
RBF
GPR
RF
MARS

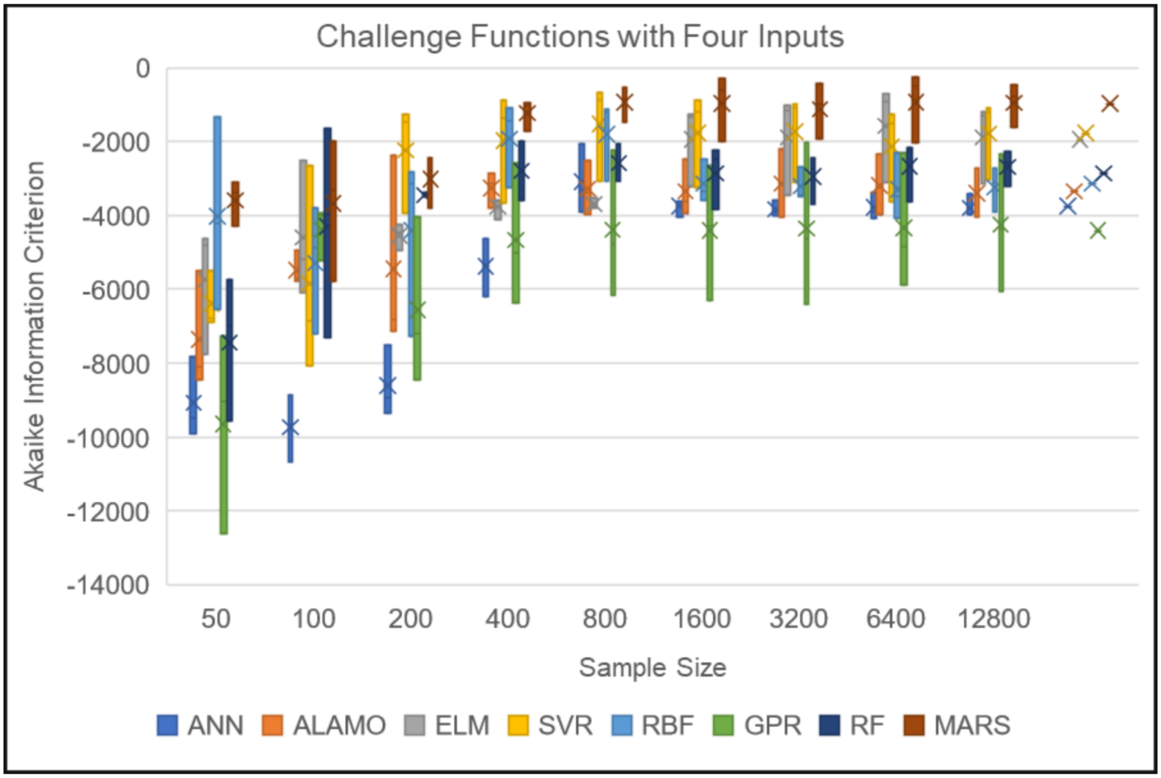

Challenge Functions with Three Inputs
Root Mean Square Error
0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
50
100
200
400
800
1600
3200
6400
12800
Sample Size
ANN
ALAMO
ELM
SVR
RBF
GPR
RF
MARS

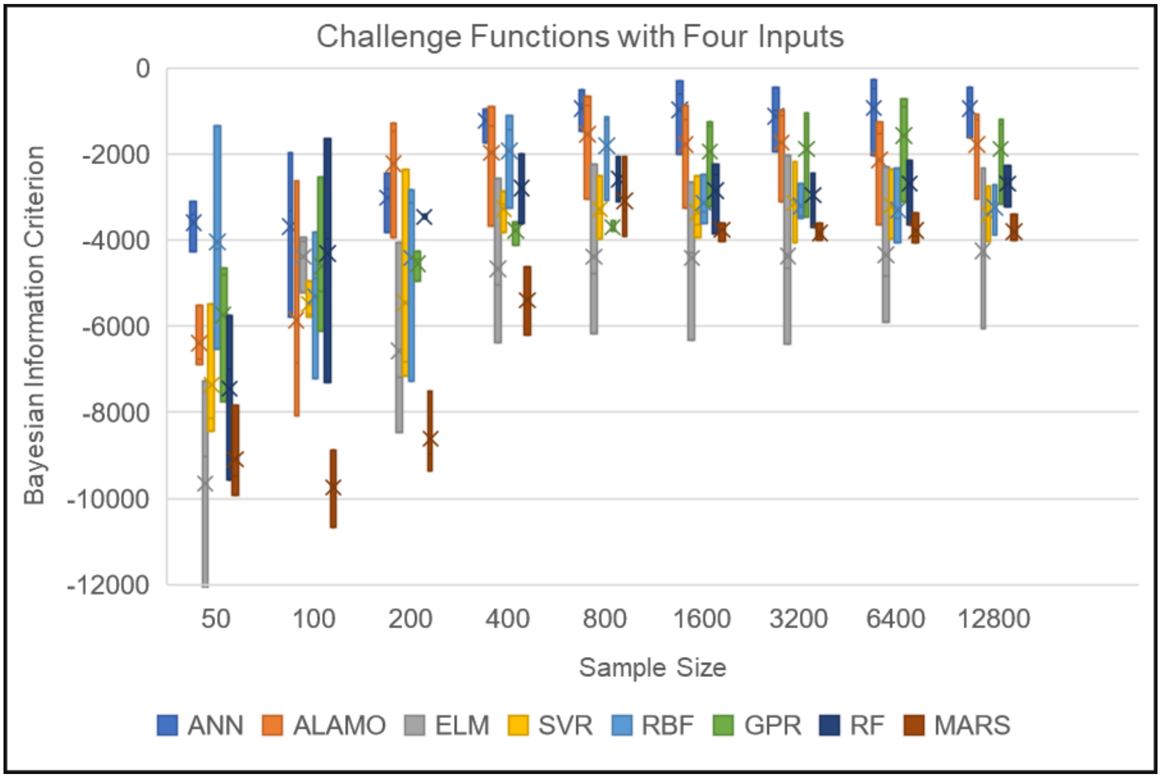

Challenge Functions with Three Inputs
Maximum Absolute Error
0.18
0.16
0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
50
100
200
400
800
1600
3200
6400
12800
Sample Size
ANN
ALAMO
ELM
SVR
RBF
GPR
RF
MARS

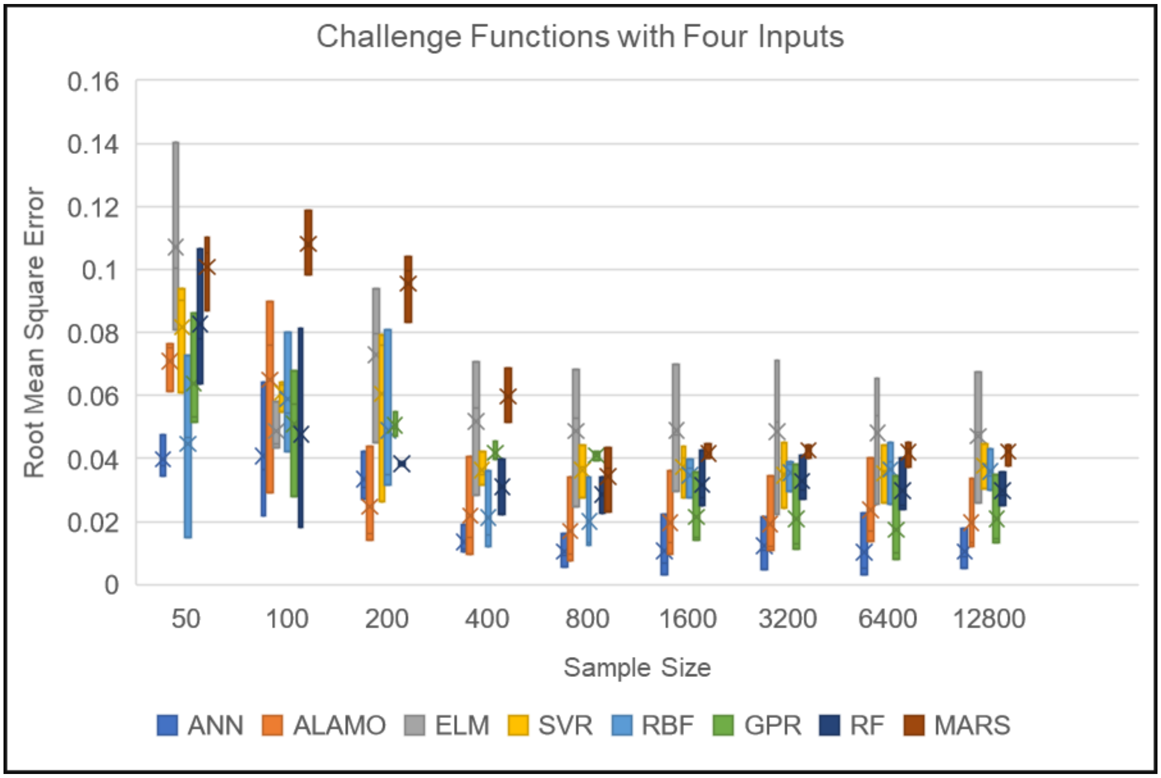
7.2. Challenge Functions with Four Inputs

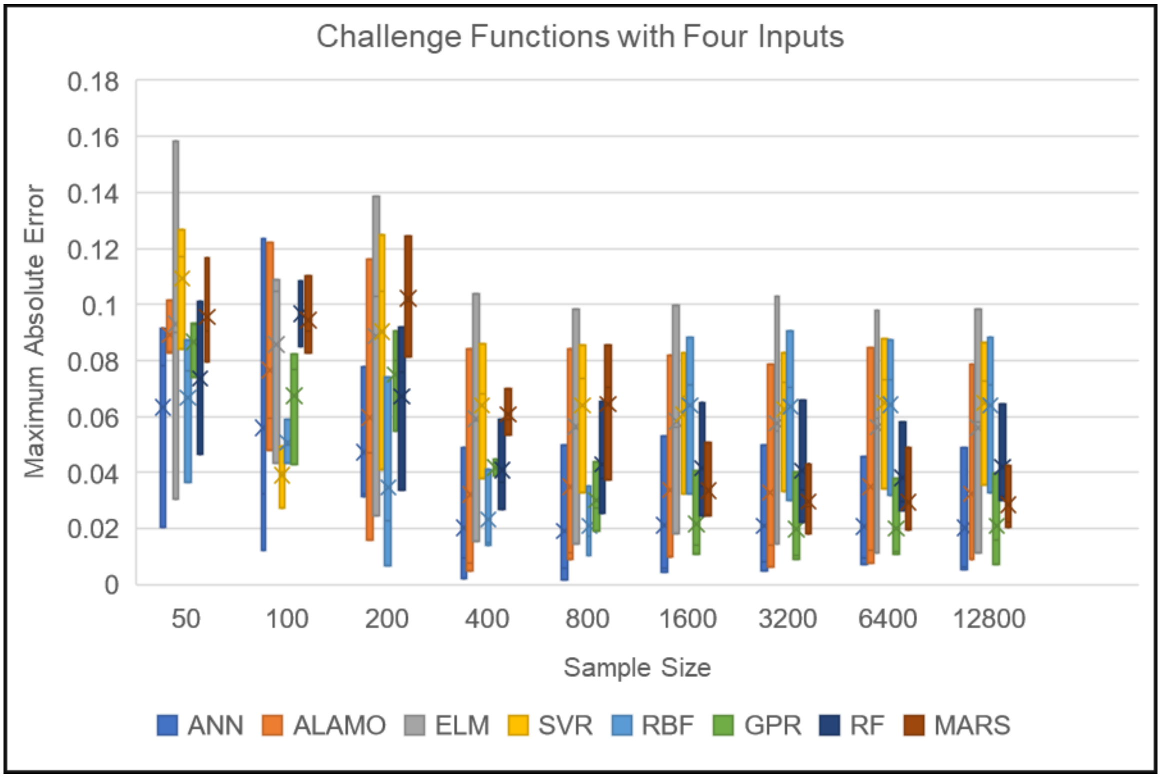




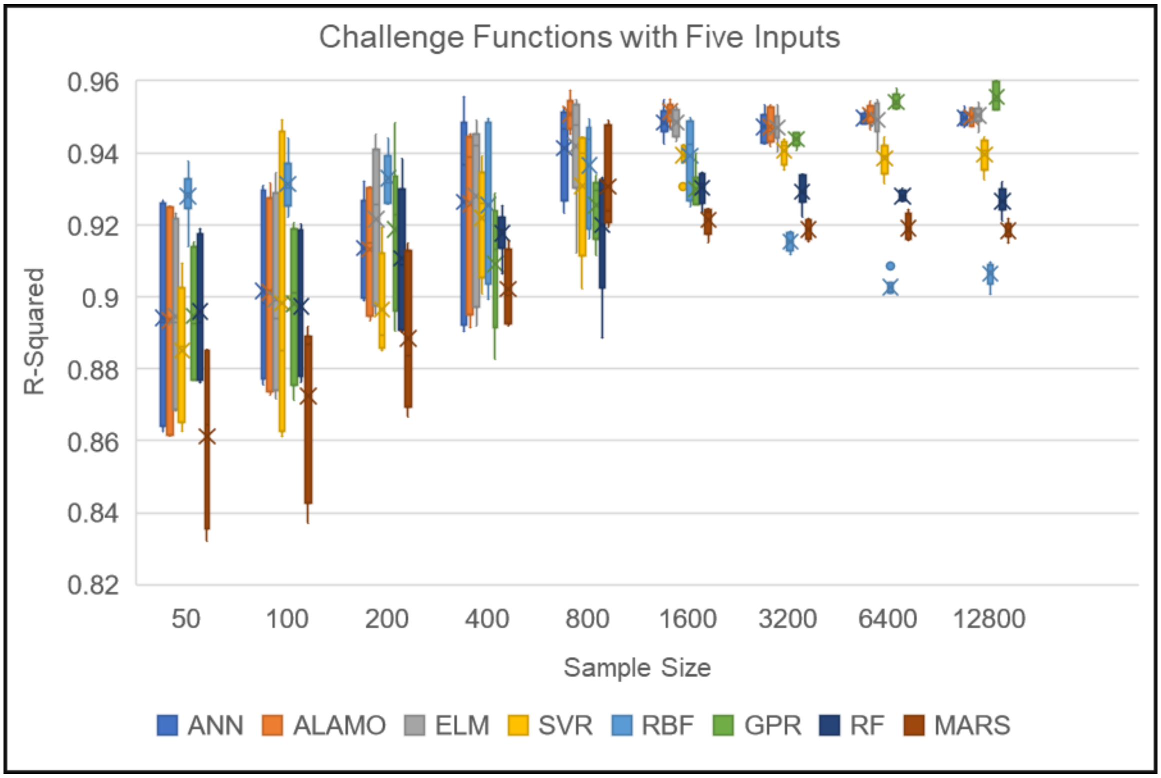

Challenge Functions with Four Inputs
Akaike Information Criterion
0
-2000
-4000
-6000
-8000
-10000
-12000
-14000
50
100
200
400
800
1600
3200
6400
12800
Sample Size
ANN
ALAMO
ELM
SVR
RBF
GPR
RF
MARS

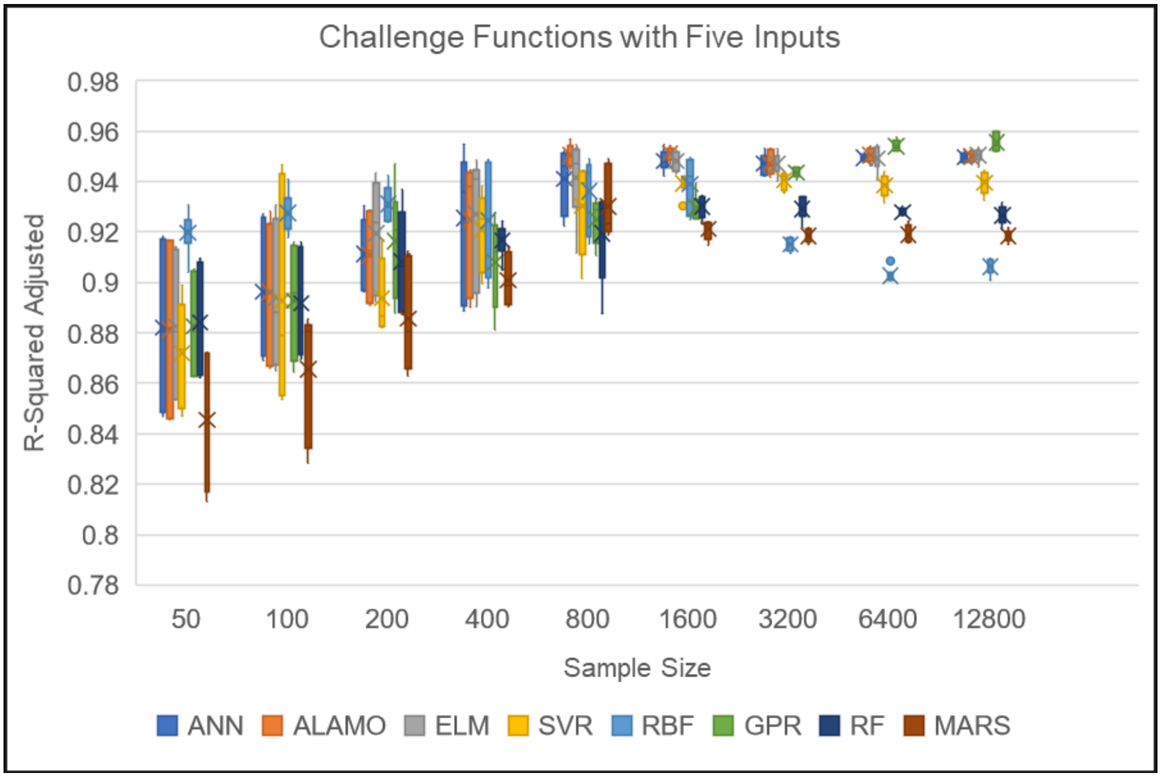

Challenge Functions with Four Inputs
Bayesian Information Criterion
0
-2000
-4000
-6000
-8000
-10000
-12000
50
100
200
400
800
1600
3200
6400
12800
Sample Size
ANN
ALAMO
ELM
SVR
RBF
GPR
RF
MARS

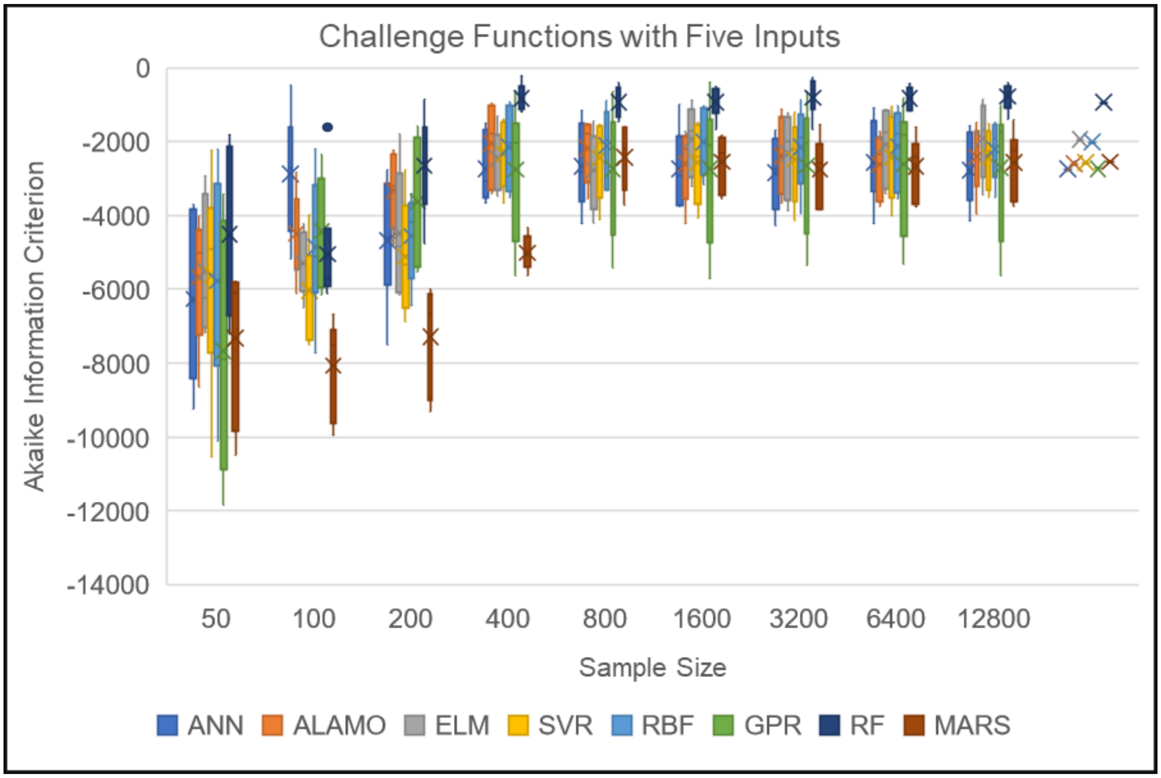

Challenge Functions with Four Inputs
Root Mean Square Error
0.16
0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
50
100
200
400
800
1600
3200
6400
12800
Sample Size
ANN
ALAMO
ELM
SVR
RBF
GPR
RF
MARS

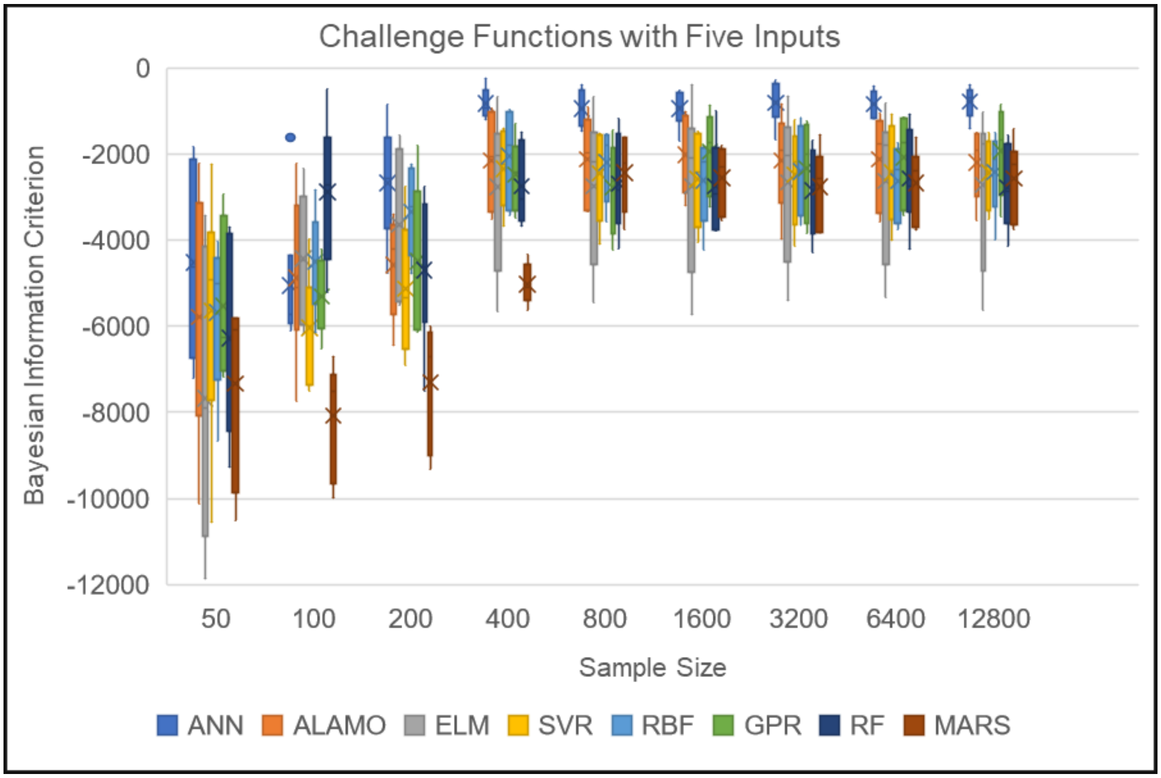

Challenge Functions with Four Inputs
Maximum Absolute Error
0.18
0.16
0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
50
100
200
400
800
1600
3200
6400
12800
Sample Size
ANN
ALAMO
ELM
SVR
RBF
GPR
RF
MARS

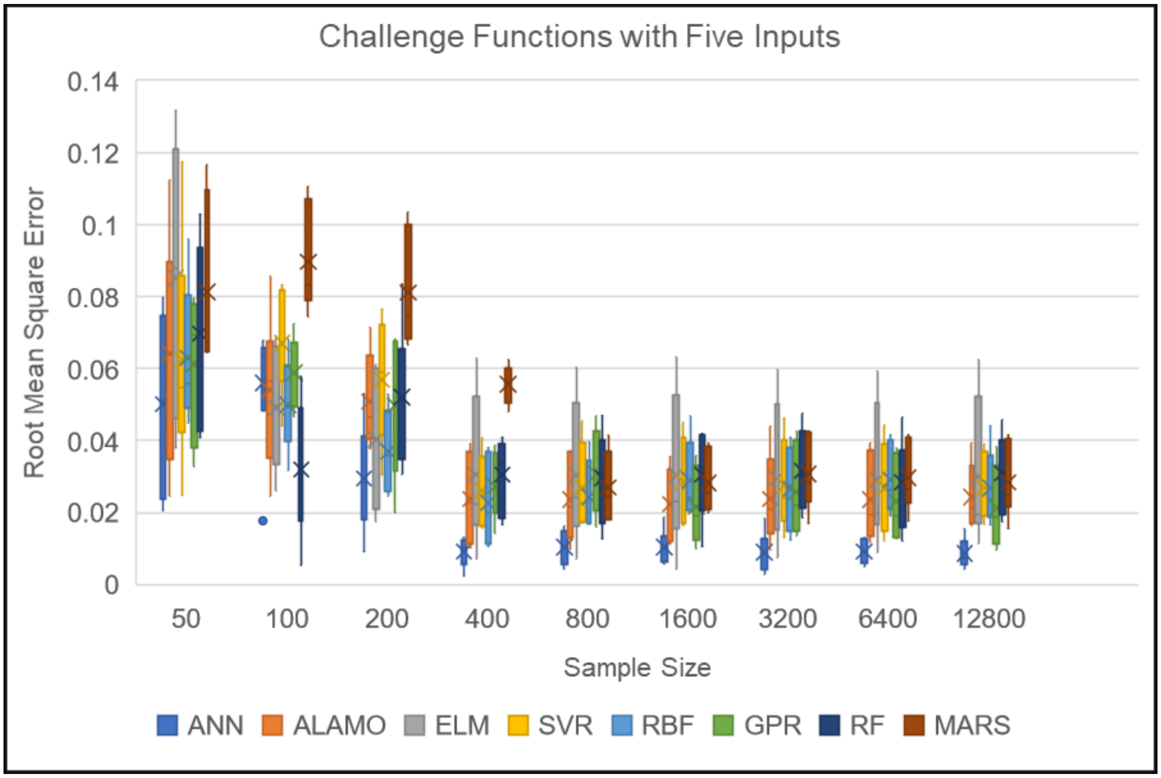
7.3. Results for Challenge Functions with Five Inputs

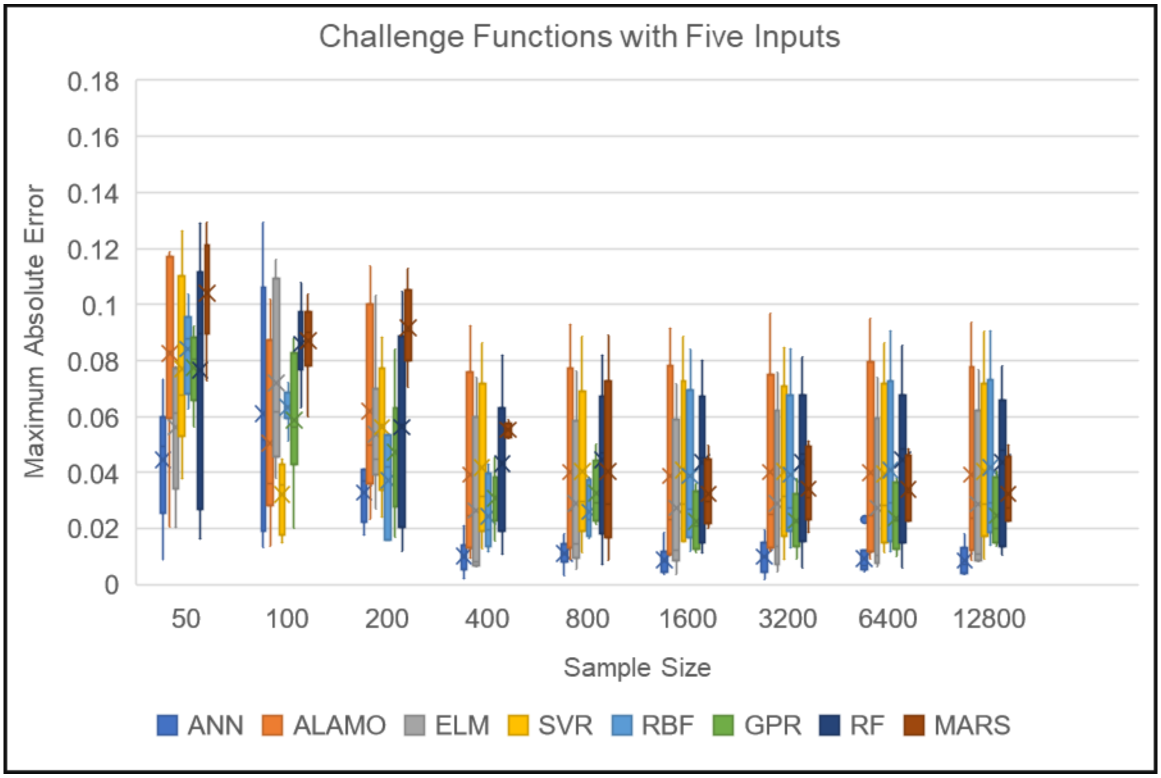




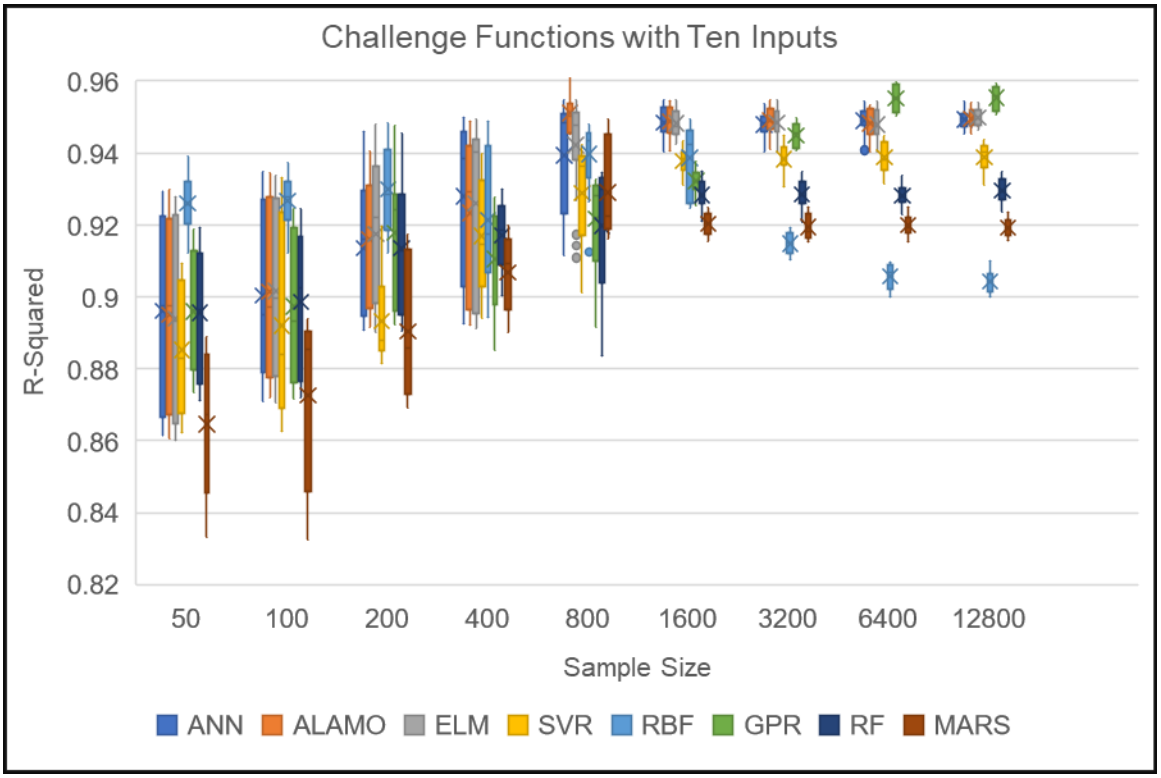

Challenge Functions with Five Inputs
Akaike Information Criterion
0
-2000
-4000
-6000
-8000
-10000
-12000
-14000
50
100
200
400
800
1600
3200
6400
12800
Sample Size
ANN
ALAMO
ELM
SVR
RBF
GPR
RF
MARS

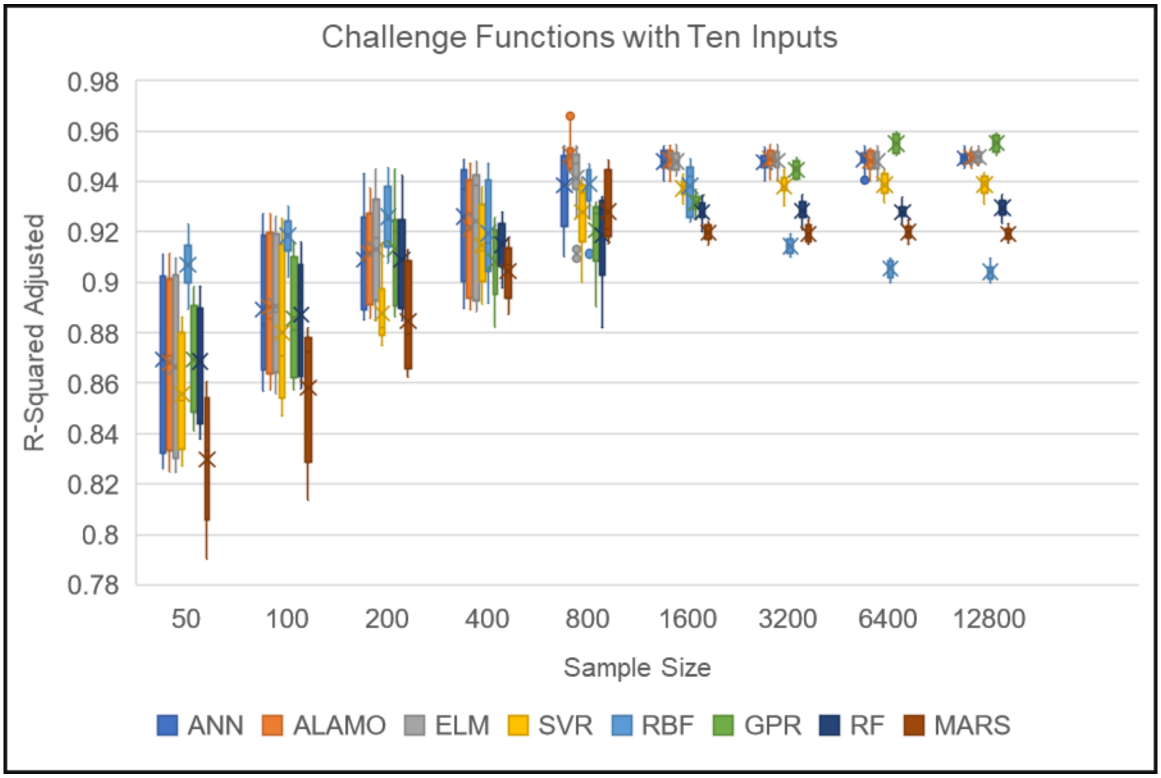

Challenge Functions with Five Inputs
Bayesian Information Criterion
0
-2000
-4000
-6000
-8000
-10000
-12000
50
100
200
400
800
1600
3200
6400
12800
Sample Size
ANN
ALAMO
ELM
SVR
RBF
GPR
RF
MARS

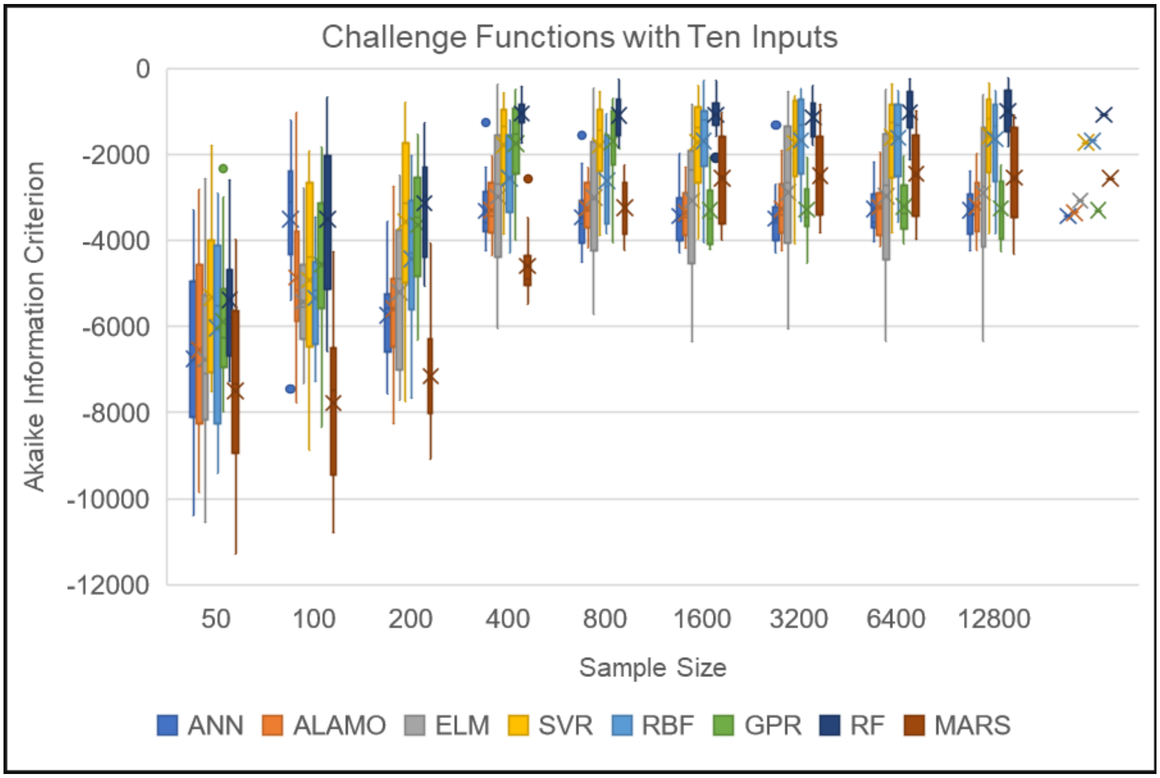

Challenge Functions with Five Inputs
Root Mean Square Error
0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
50
100
200
400
800
1600
3200
6400
12800
Sample Size
ANN
ALAMO
ELM
SVR
RBF
GPR
RF
MARS

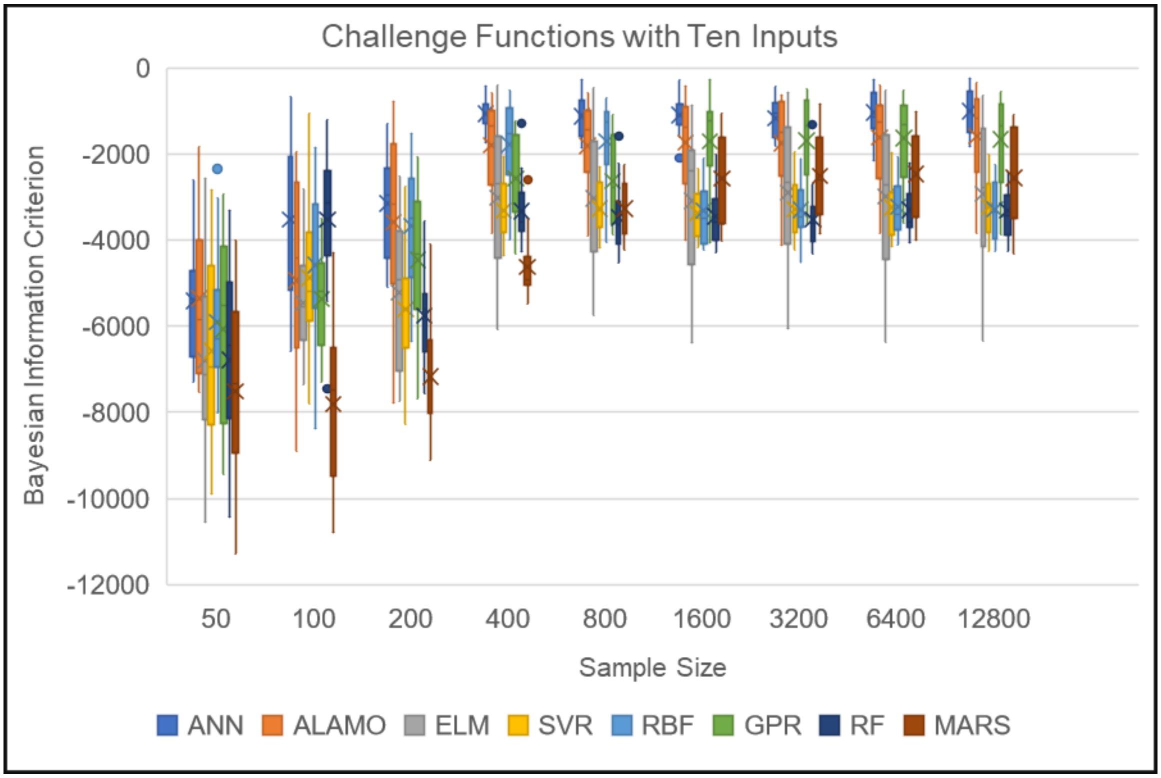

Challenge Functions with Five Inputs
Maximum Absolute Error
0.18
0.16
0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
50
100
200
400
800
1600
3200
6400
12800
Sample Size
ANN
ALAMO
ELM
SVR
RBF
GPR
RF
MARS

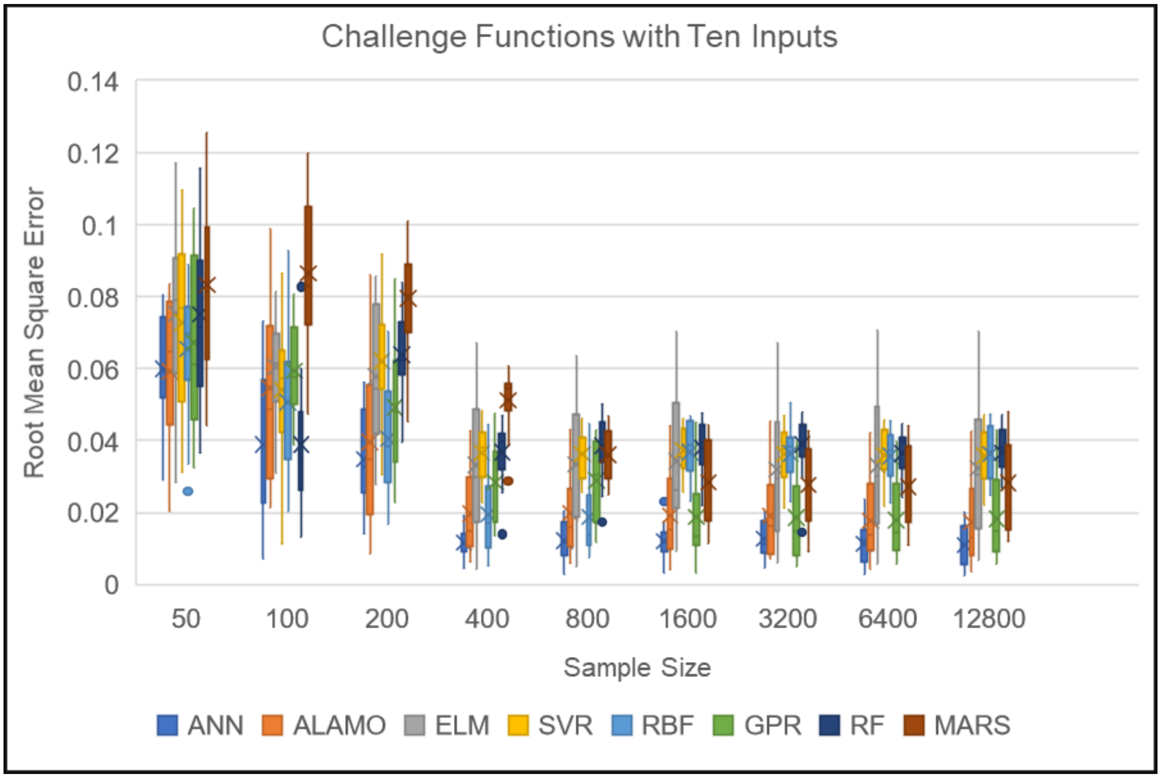
7.4. Results for Challenge Functions with Ten Inputs

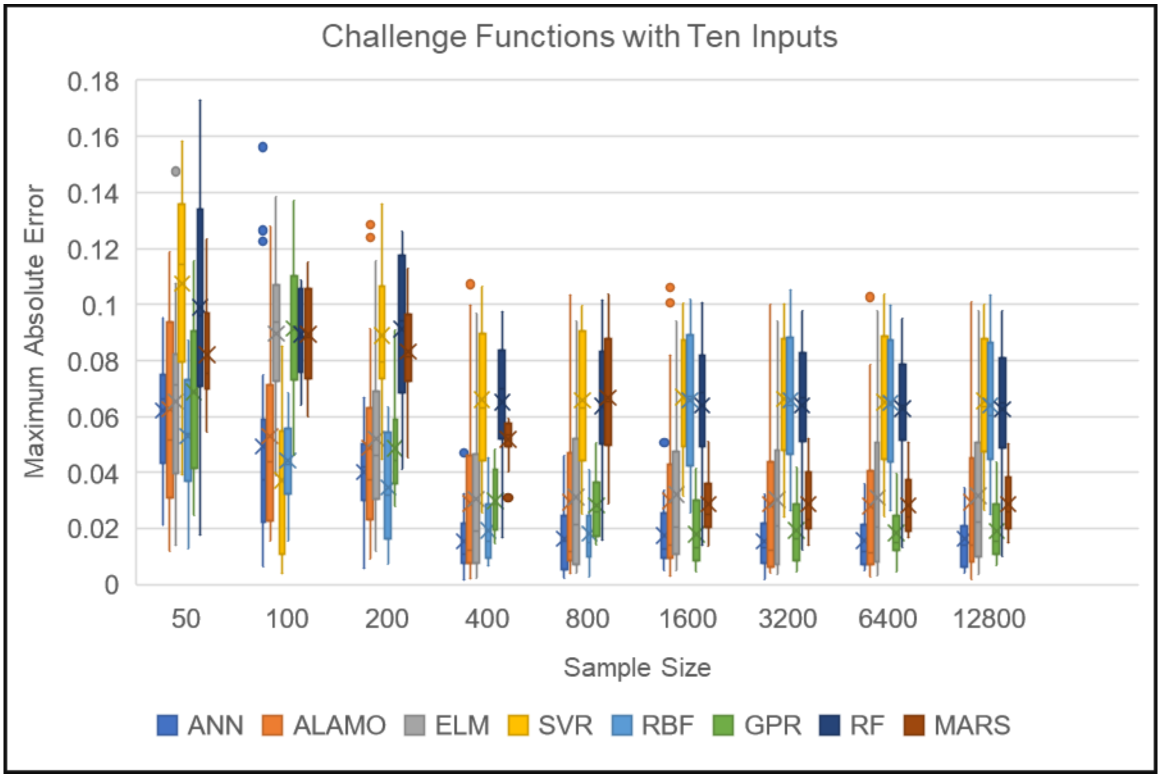




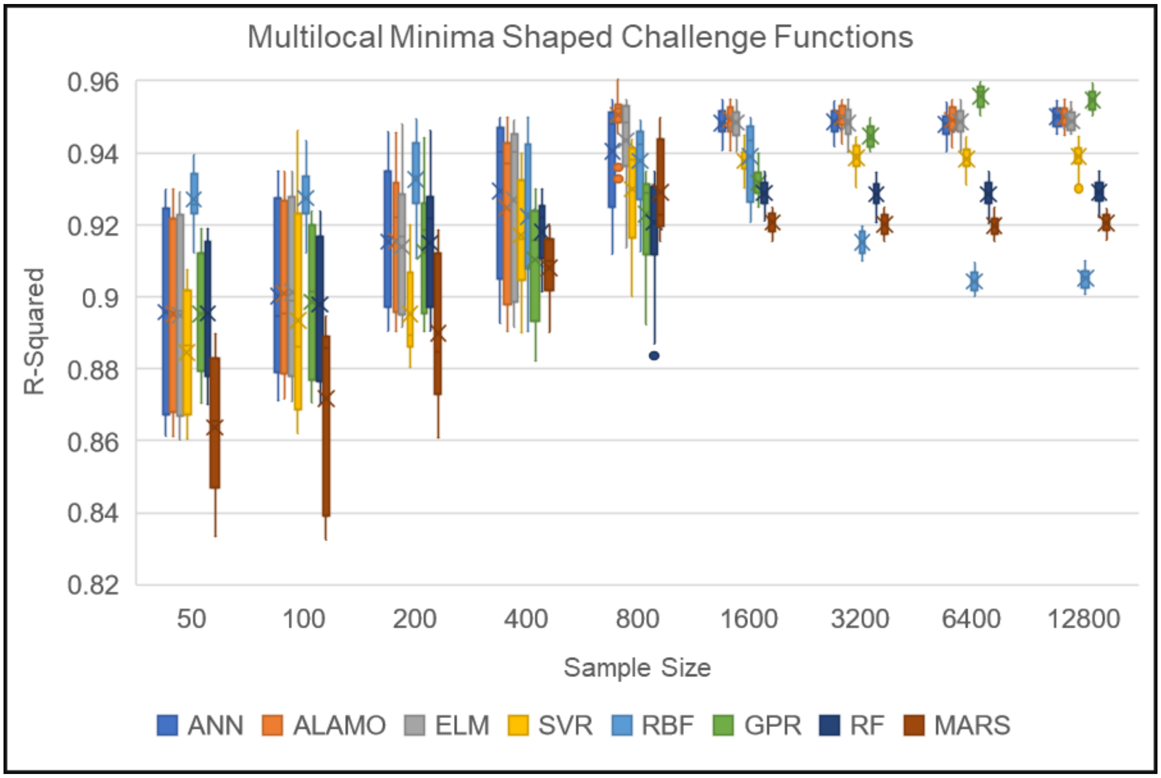

Challenge Functions with Ten Inputs
Akaike Information Criterion
0
-2000
-4000
-6000
-8000
-10000
-12000
50
100
200
400
800
1600
3200
6400
12800
Sample Size
ANN
ALAMO
ELM
SVR
RBF
GPR
RF
MARS

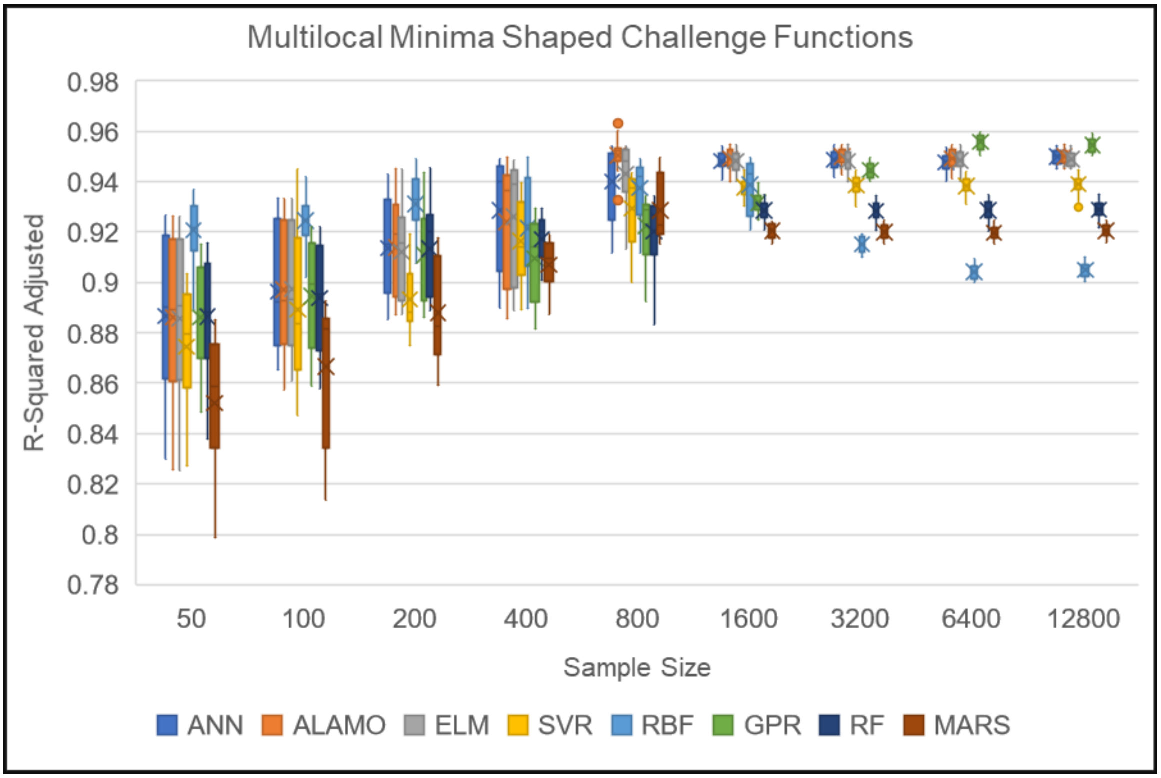

Challenge Functions with Ten Inputs
Bayesian Information Criterion
0
-2000
-4000
-6000
-8000
-10000
-12000
50
100
200
400
800
1600
3200
6400
12800
Sample Size
ANN
ALAMO
ELM
SVR
RBF
GPR
RF
MARS

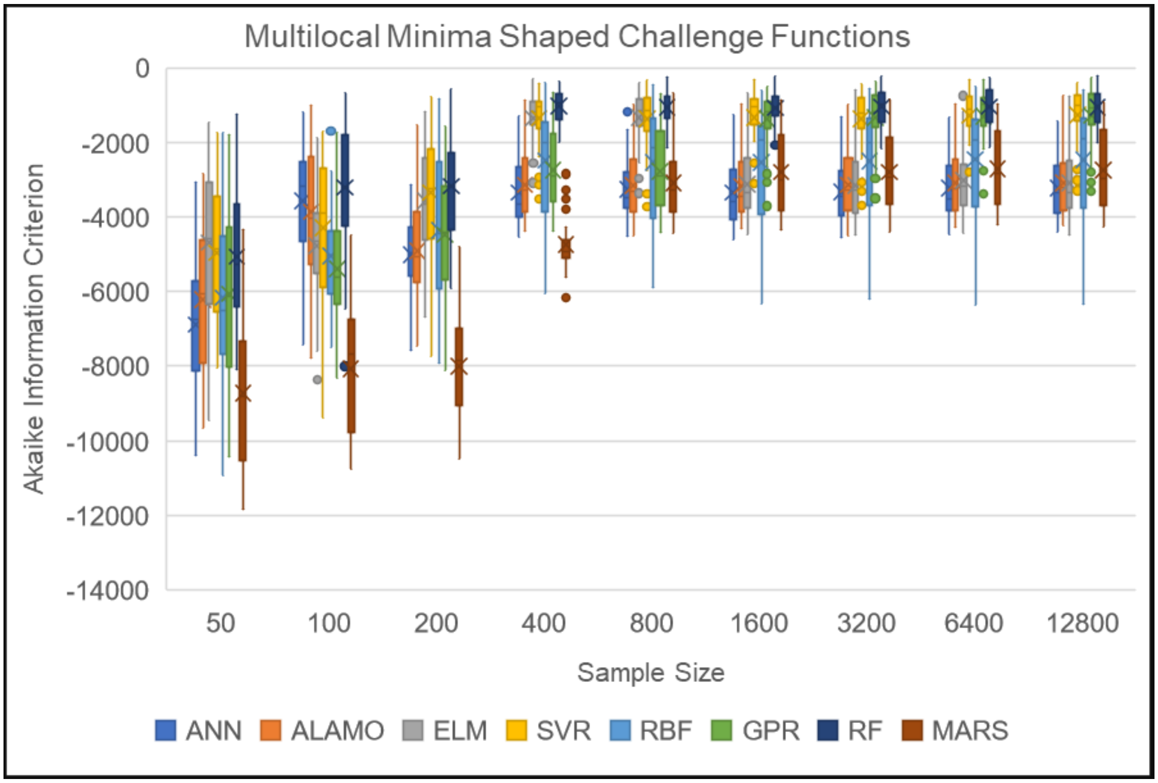

Challenge Functions with Ten Inputs
Root Mean Square Error
0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
50
100
200
400
800
1600
3200
6400
12800
Sample Size
ANN
ALAMO
ELM
SVR
RBF
GPR
RF
MARS

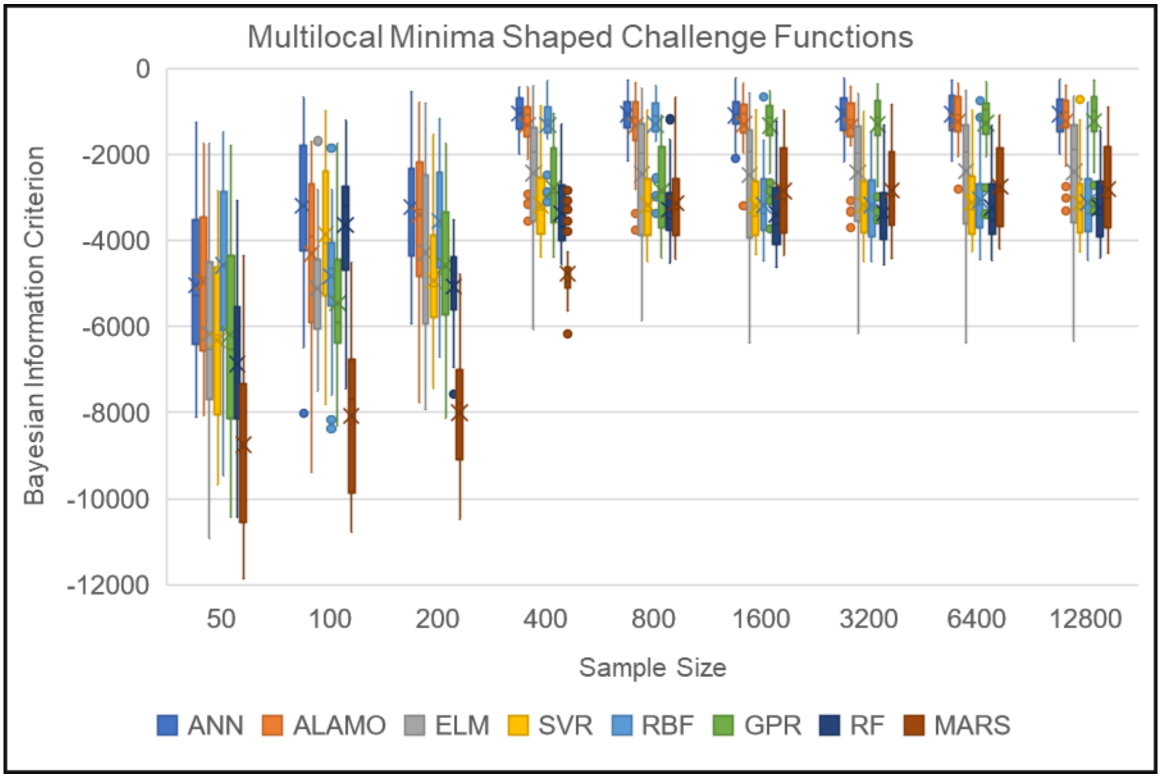

Challenge Functions with Ten Inputs
Maximum Absolute Error
0.18
0.16
0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
50
100
200
400
800
1600
3200
6400
12800
Sample Size
ANN
ALAMO
ELM
SVR
RBF
GPR
RF
MARS

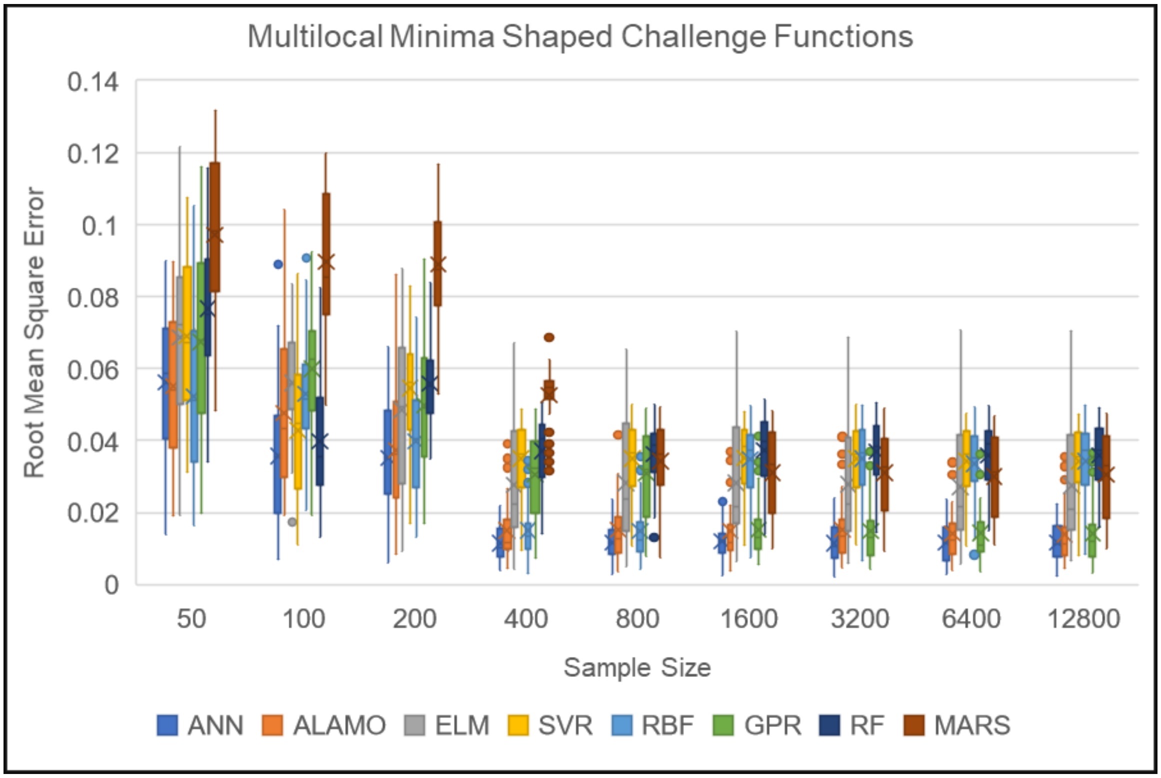
7.5. Results for Challenge Functions with a Multilocal Minima Shape

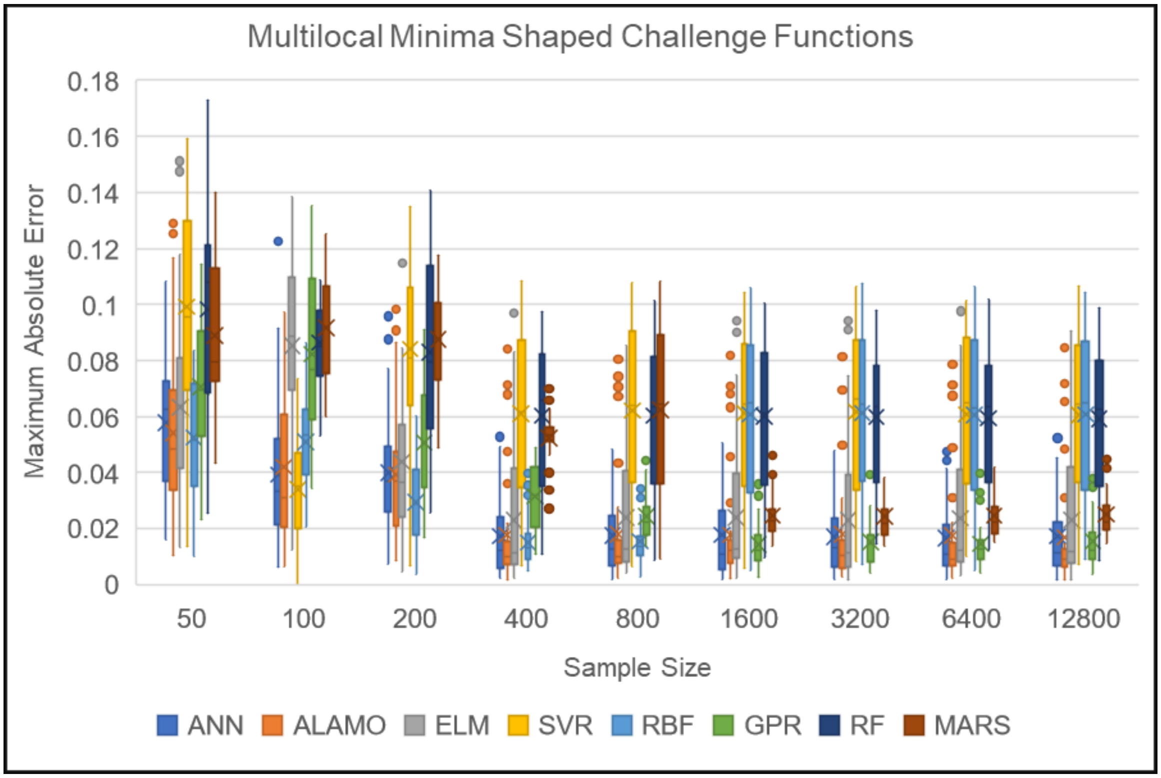




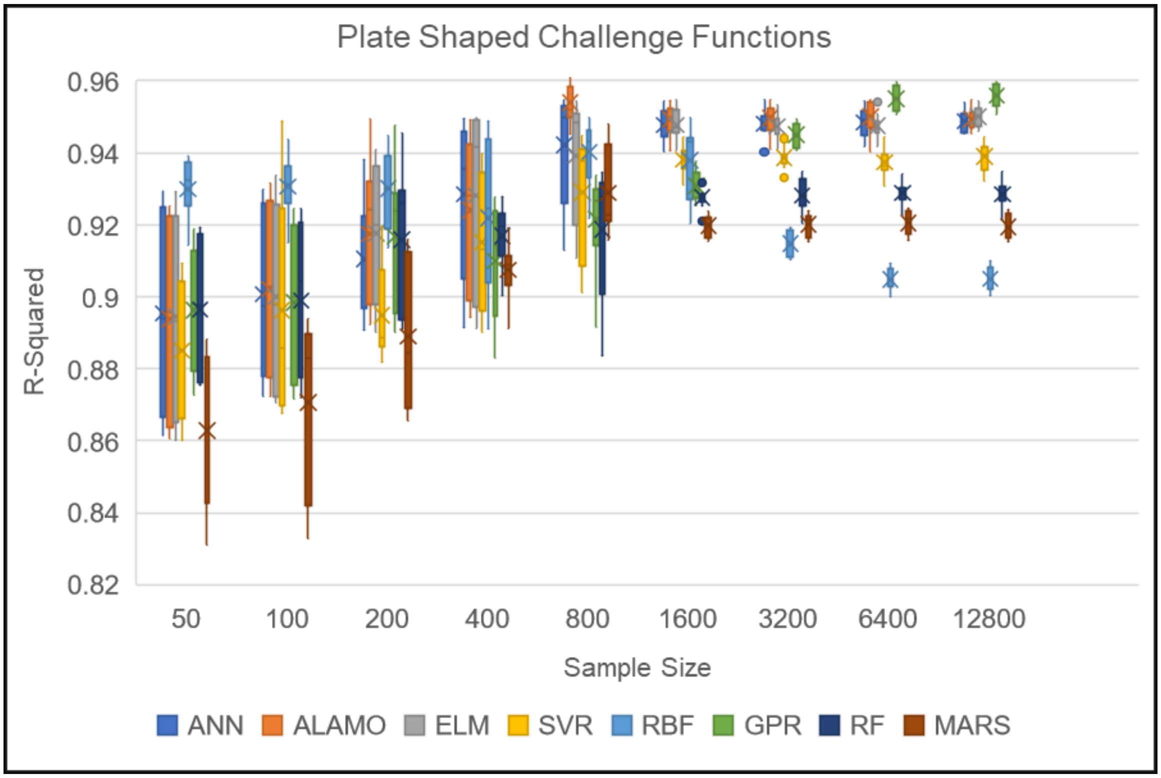

Multilocal Minima Shaped Challenge Functions
0
-2000
-4000
-6000
-8000
-10000
-12000
-14000
Akaike Information Criterion
50
100
200
400
800
1600
3200
6400
12800
Sample Size
ANN
ALAMO
ELM
SVR
RBF
GPR
RF
MARS

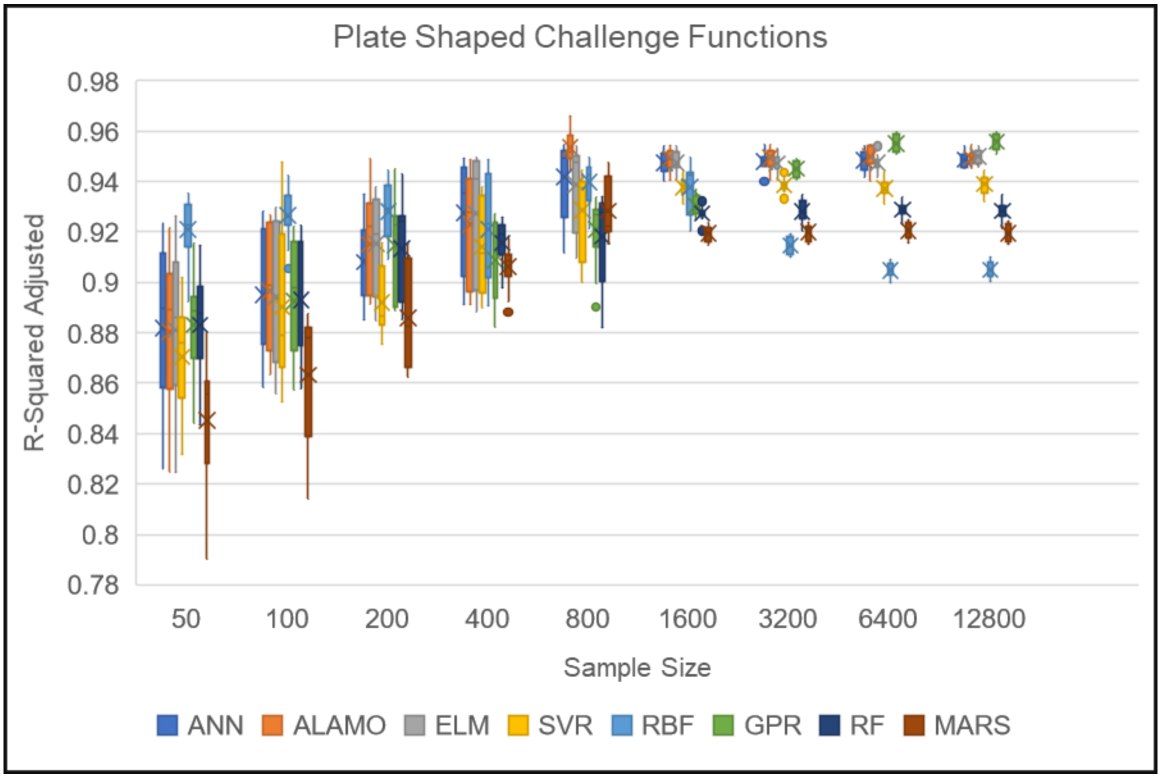

Multilocal Minima Shaped Challenge Functions
0
-2000
-4000
-6000
-8000
-10000
-12000
Bayesian Information Criterion
50
100
200
400
800
1600
3200
6400
12800
Sample Size
ANN
ALAMO
ELM
SVR
RBF
GPR
RF
MARS


Multilocal Minima Shaped Challenge Functions
Root Mean Square Error
0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
50
100
200
400
800
1600
3200
6400
12800
Sample Size
ANN
ALAMO
ELM
SVR
RBF
GPR
RF
MARS


Multilocal Minima Shaped Challenge Functions
Maximum Absolute Error
0.18
0.16
0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
50
100
200
400
800
1600
3200
6400
12800
Sample Size
ANN
ALAMO
ELM
SVR
RBF
GPR
RF
MARS

7.6. Results for Challenge Functions with a Plate Shape





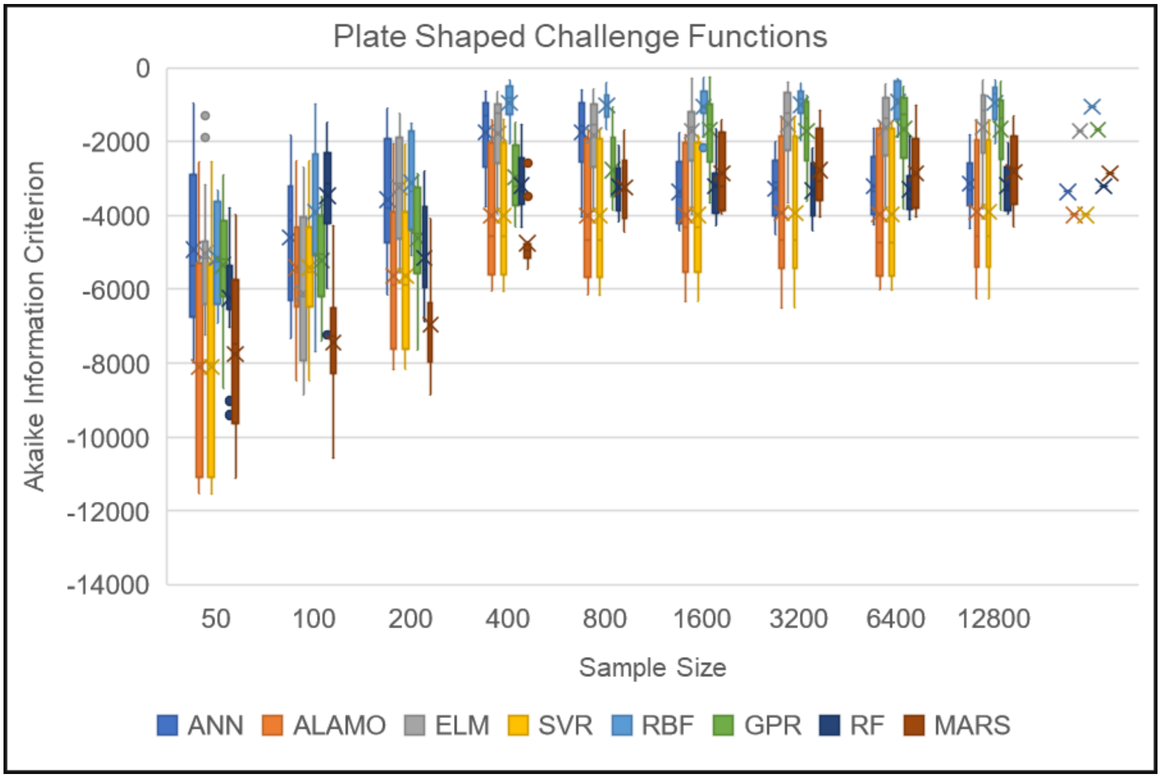


Plate Shaped Challenge Functions
Akaike Information Criterion
0
-2000
-4000
-6000
-8000
-10000
-12000
-14000
50
100
200
400
800
1600
3200
6400
12800
Sample Size
ANN
ALAMO
ELM
SVR
RBF
GPR
RF
MARS

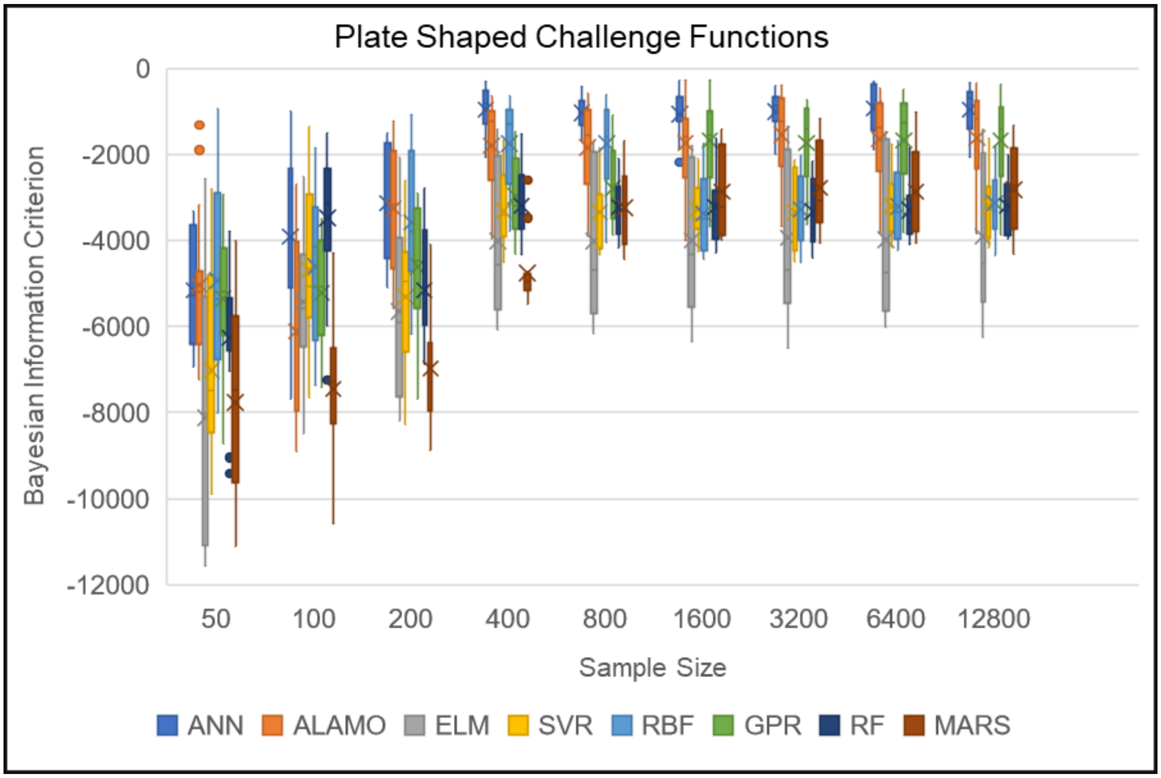


Plate Shaped Challenge Functions
Bayesian Information Criterion
0
-2000
-4000
-6000
-8000
-10000
-12000
50
100
200
400
800
1600
3200
6400
12800
Sample Size
ANN
ALAMO
ELM
SVR
RBF
GPR
RF
MARS

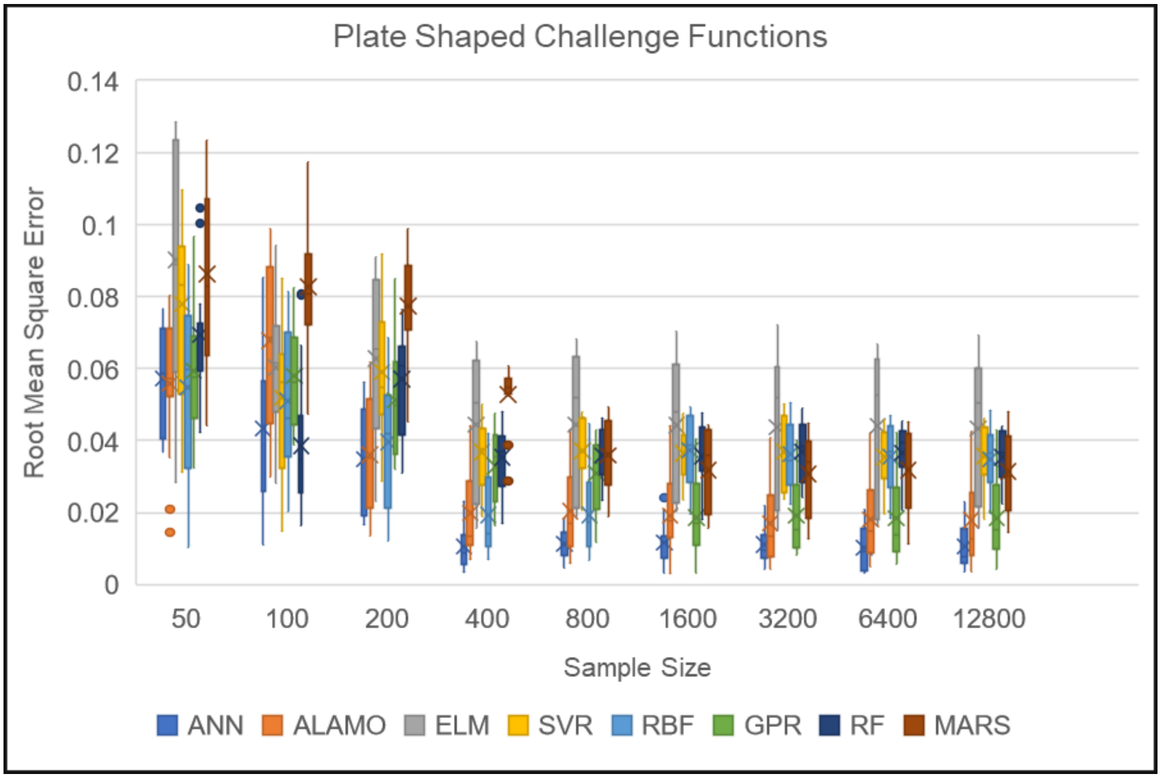


Plate Shaped Challenge Functions
Root Mean Square Error
0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
50
100
200
400
800
1600
3200
6400
12800
Sample Size
ANN
ALAMO
ELM
SVR
RBF
GPR
RF
MARS

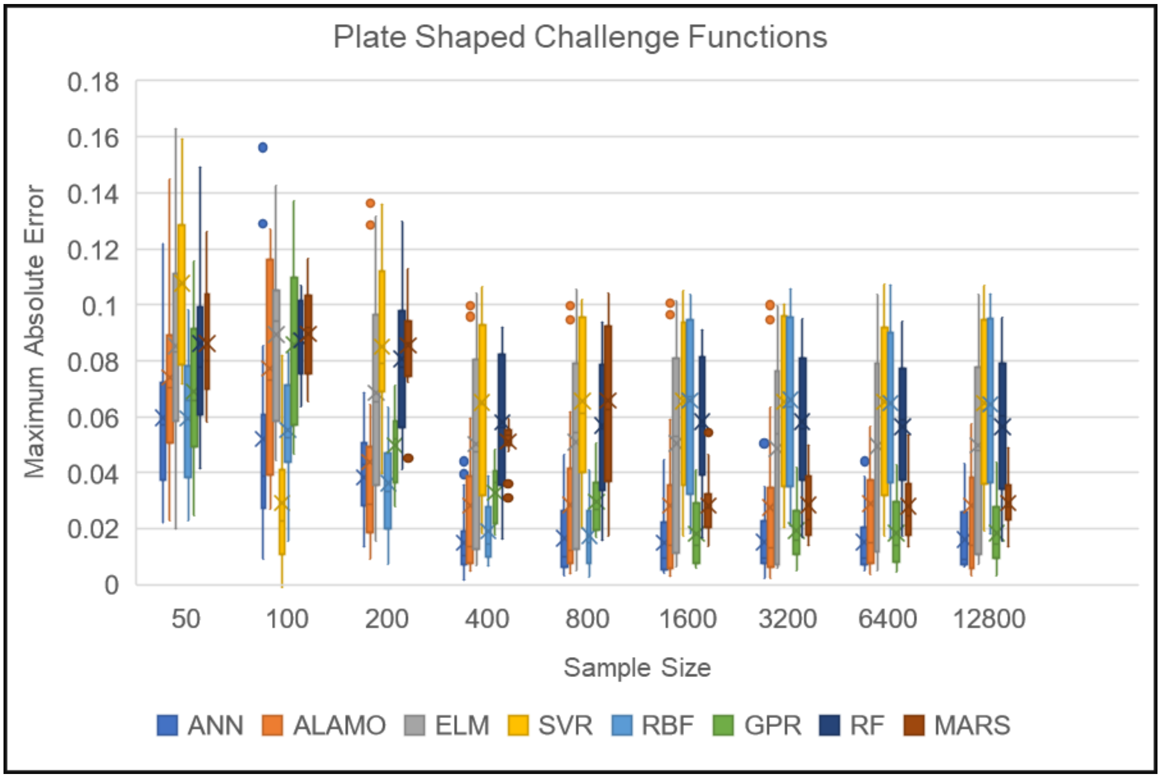
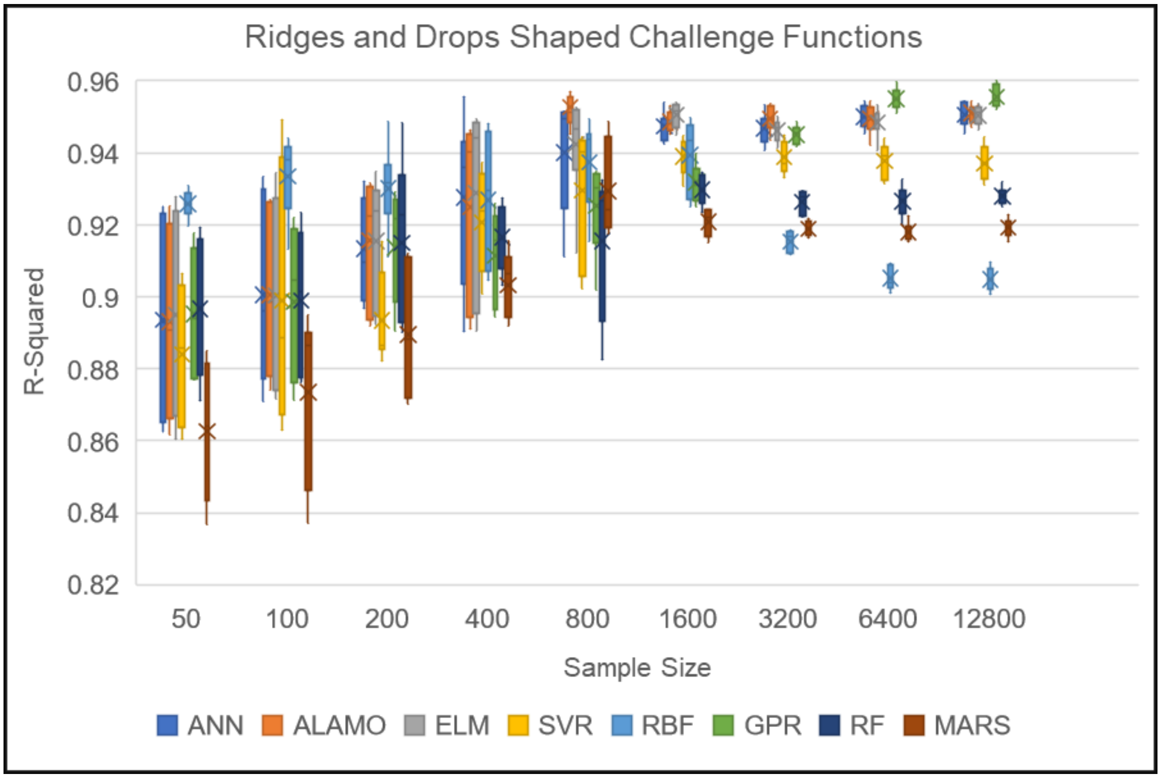
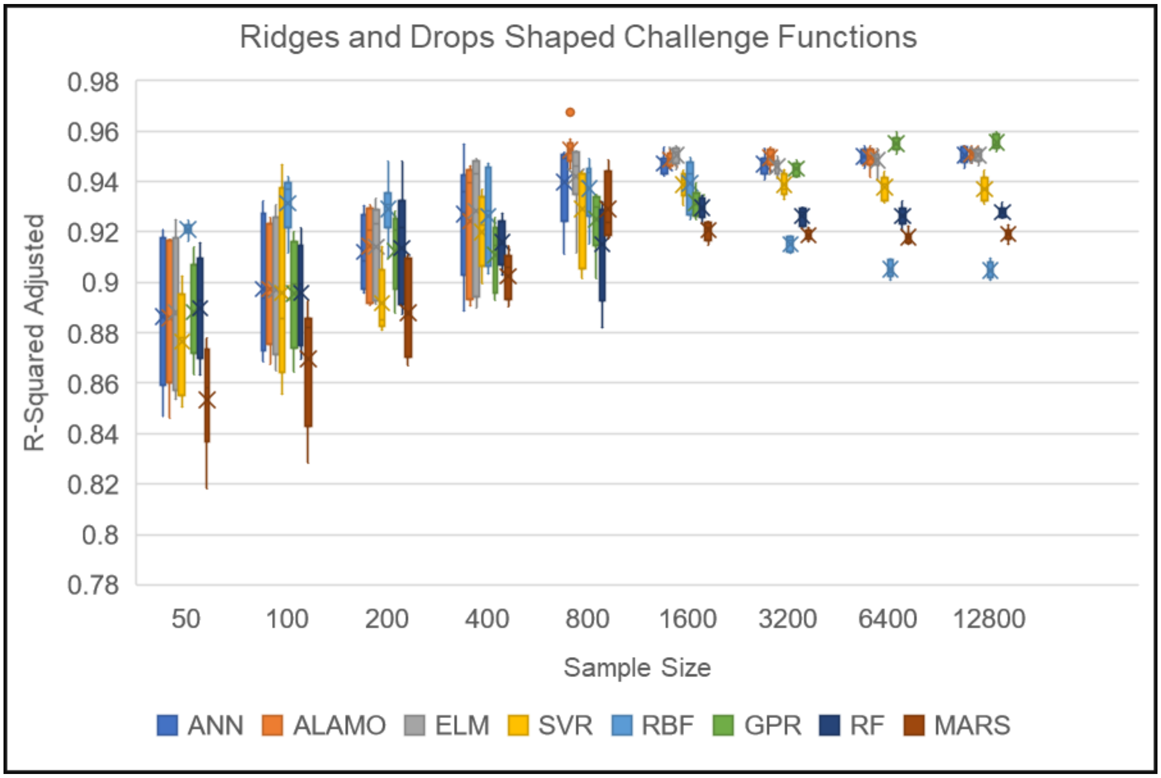
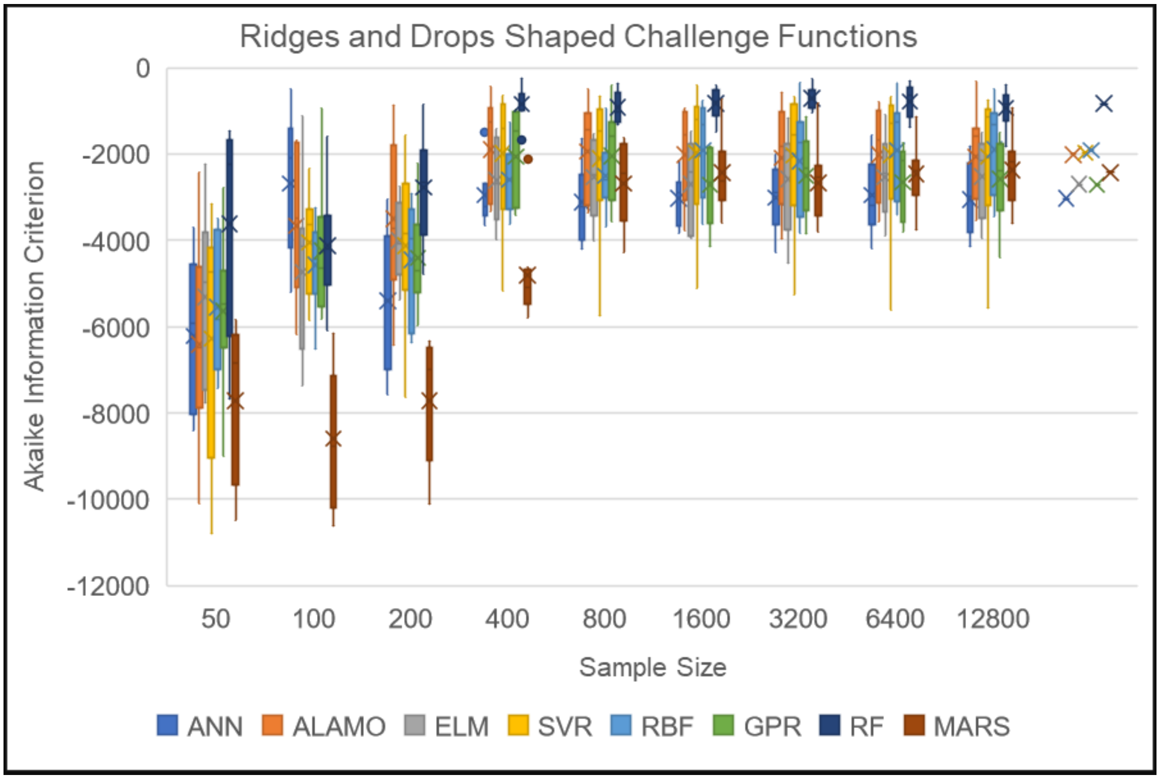


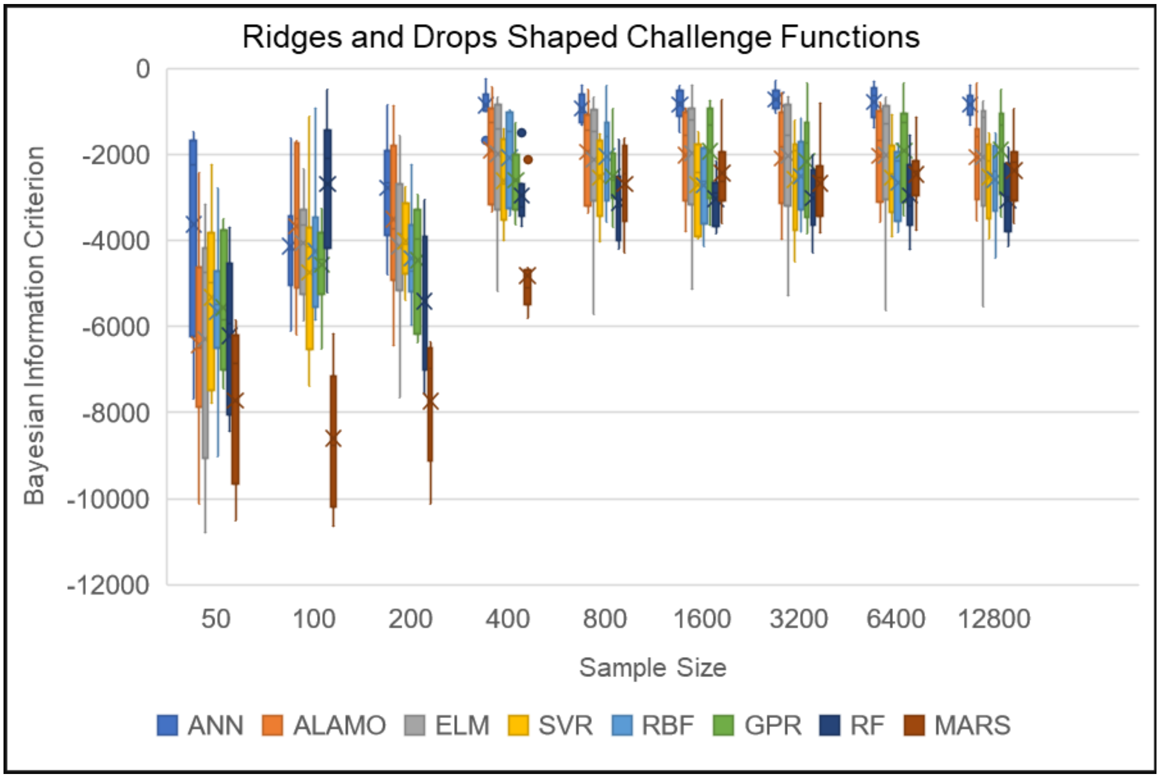
Plate Shaped Challenge Functions
Maximum Absolute Error
0.18
0.16
0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
50
100
200
400
800
1600
3200
6400
12800
Sample Size
ANN
ALAMO
ELM
SVR
RBF
GPR
RF
MARS

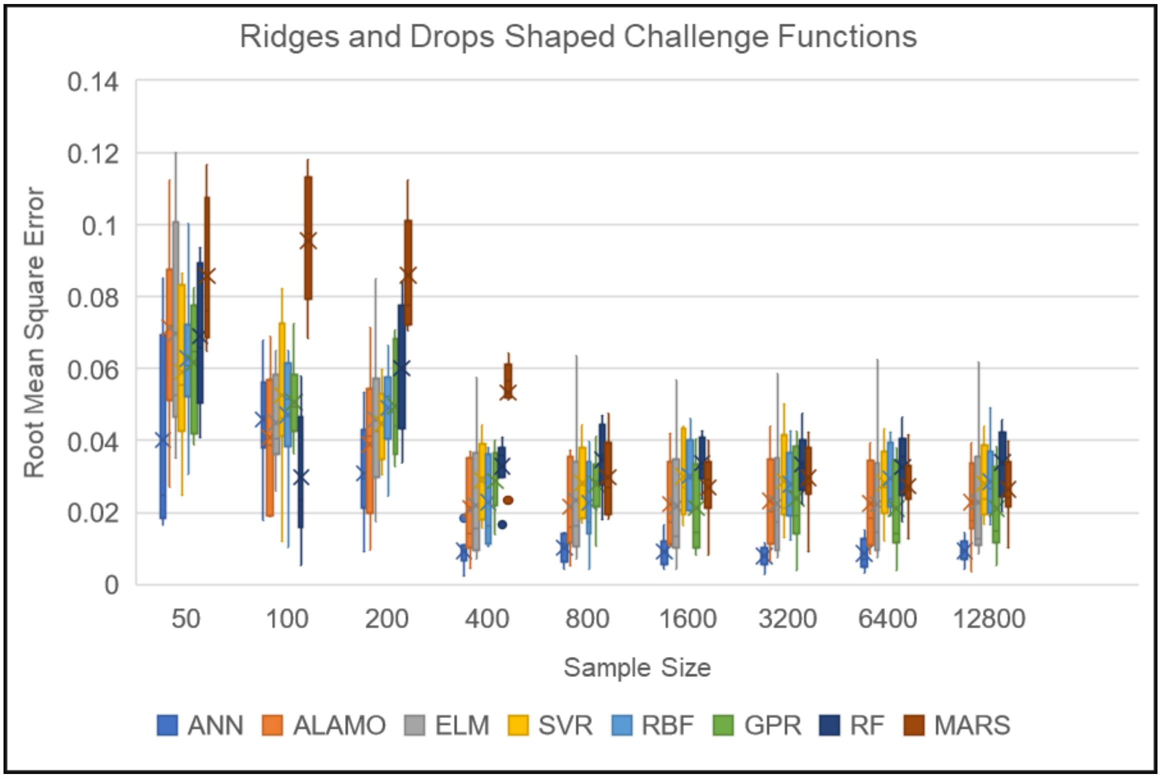
7.7. Results for Challenge Functions with a Ridges & Drops Shape

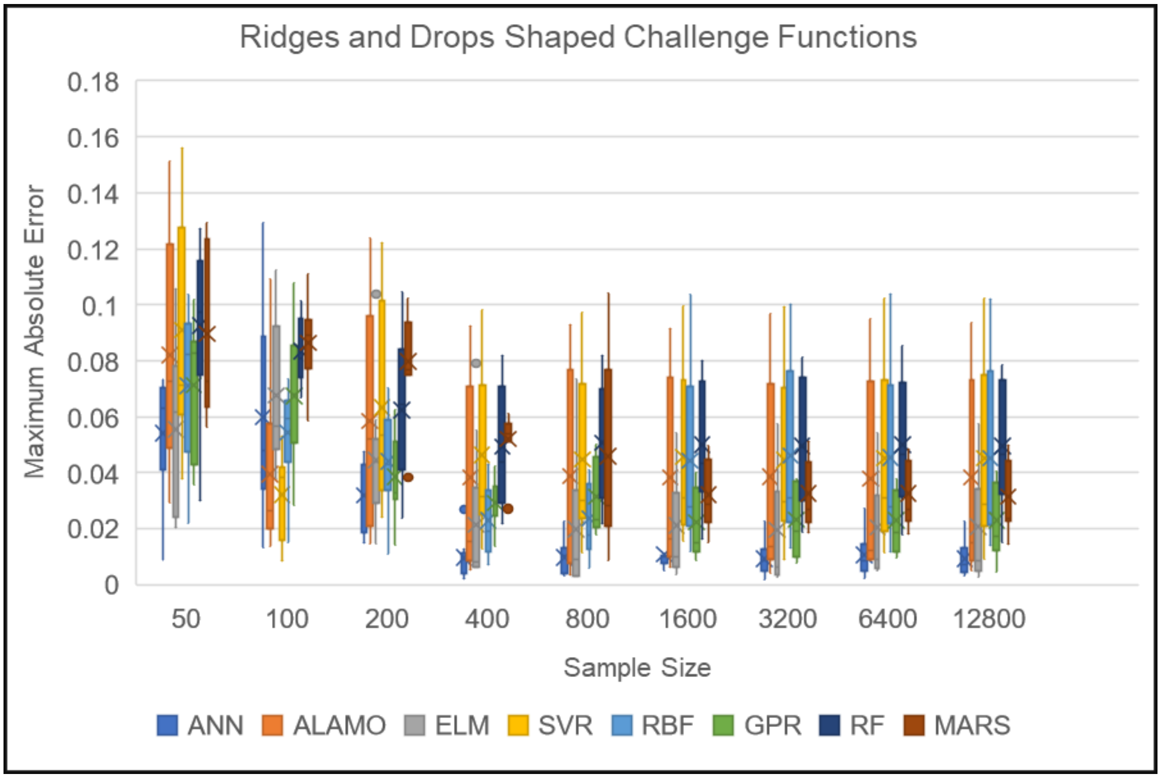




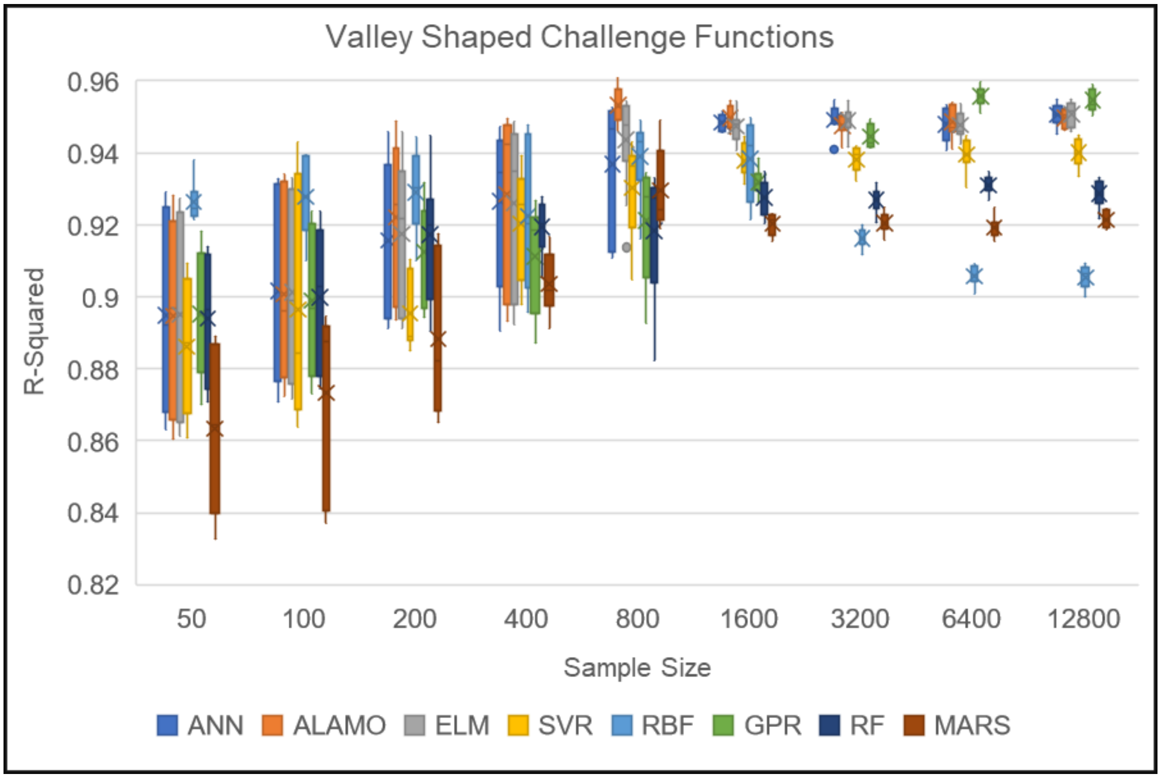

Ridges and Drops Shaped Challenge Functions
Akaike Information Criterion
0
-2000
-4000
-6000
-8000
-10000
-12000
50
100
200
400
800
1600
3200
6400
12800
Sample Size
ANN
ALAMO
ELM
SVR
RBF
GPR
RF
MARS

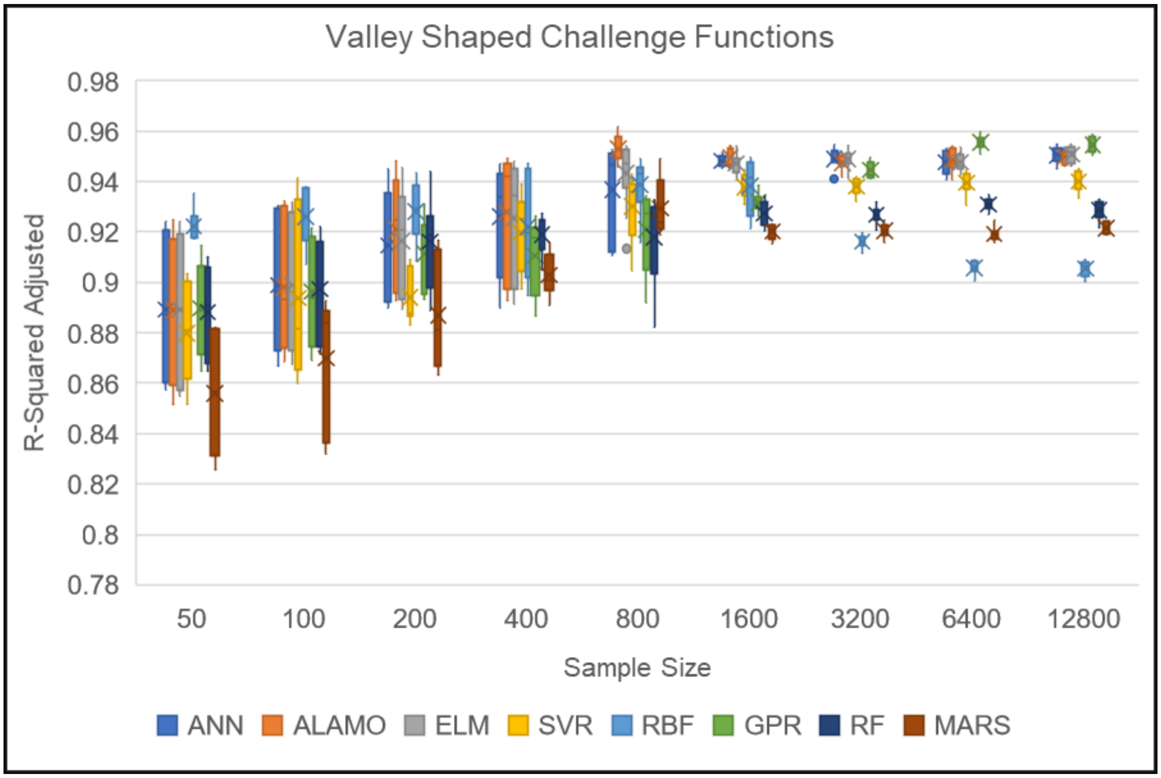

Ridges and Drops Shaped Challenge Functions
Bayesian Information Criterion
0
-2000
-4000
-6000
-8000
-10000
-12000
50
100
200
400
800
1600
3200
6400
12800
Sample Size
ANN
ALAMO
ELM
SVR
RBF
GPR
RF
MARS

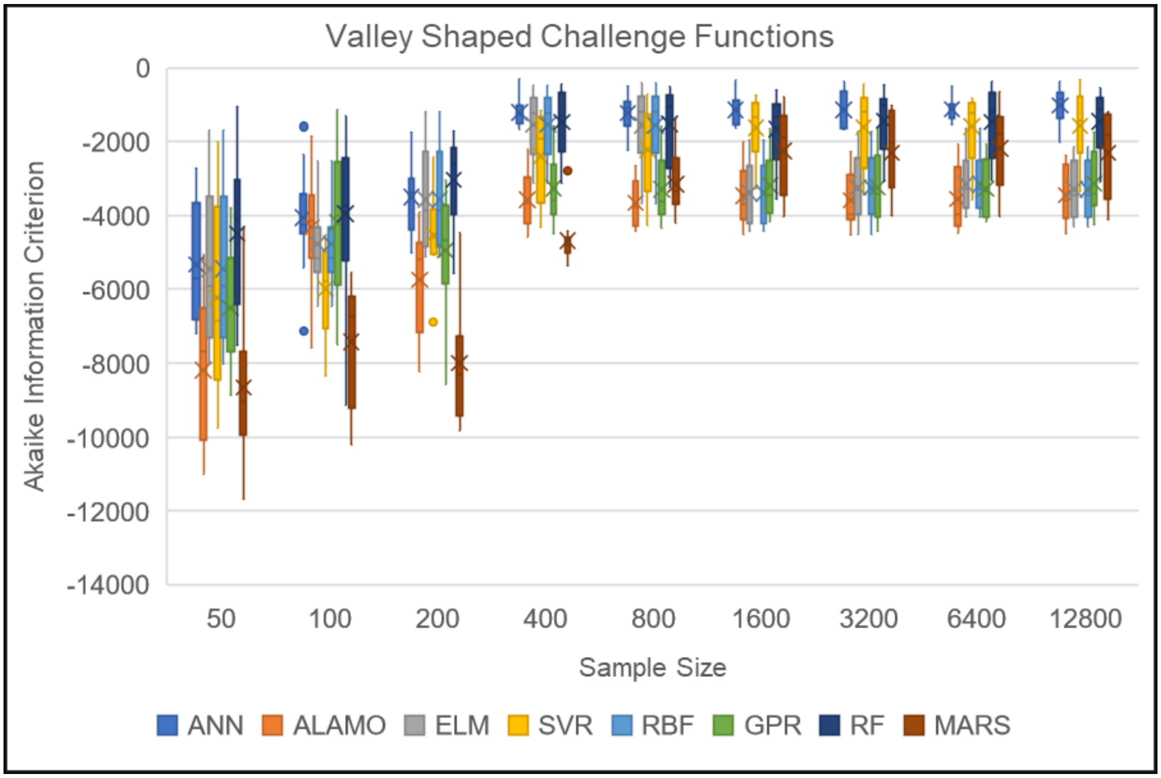

Ridges and Drops Shaped Challenge Functions
Root Mean Square Error
0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
50
100
200
400
800
1600
3200
6400
12800
Sample Size
ANN
ALAMO
ELM
SVR
RBF
GPR
RF
MARS

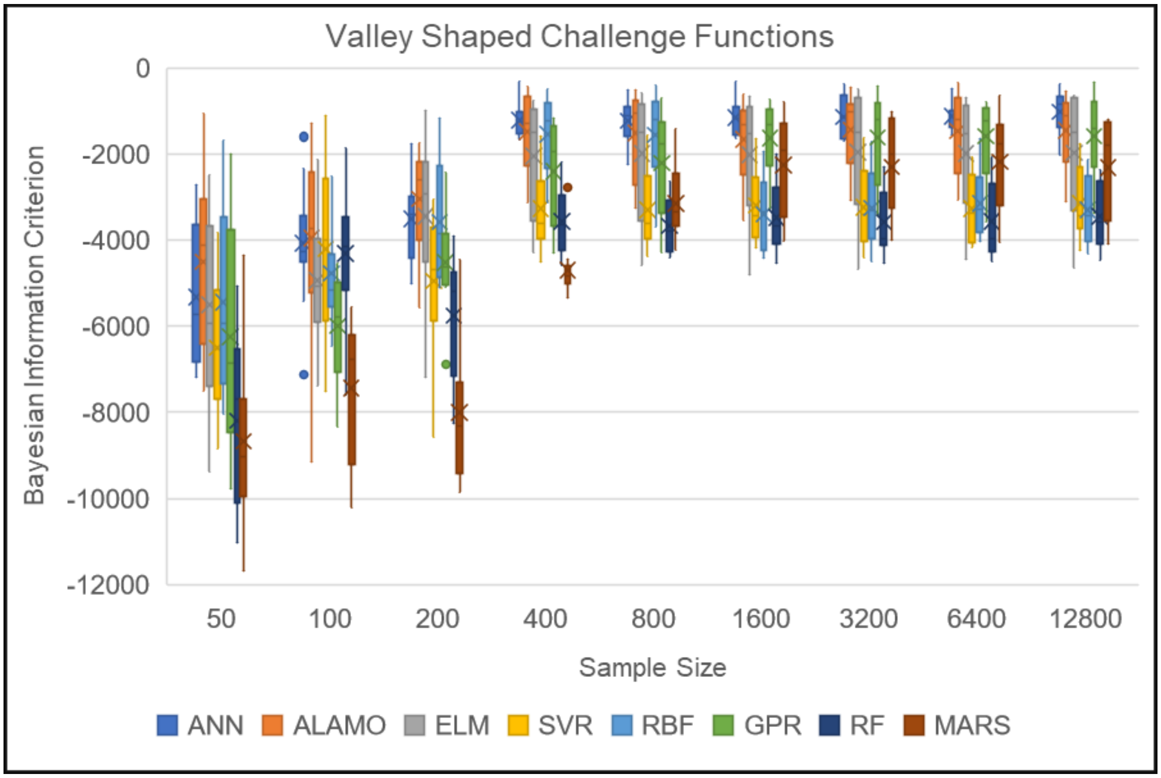

Ridges and Drops Shaped Challenge Functions
Maximum Absolute Error
0.18
0.16
0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
50
100
200
400
800
1600
3200
6400
12800
Sample Size
ANN
ALAMO
ELM
SVR
RBF
GPR
RF
MARS

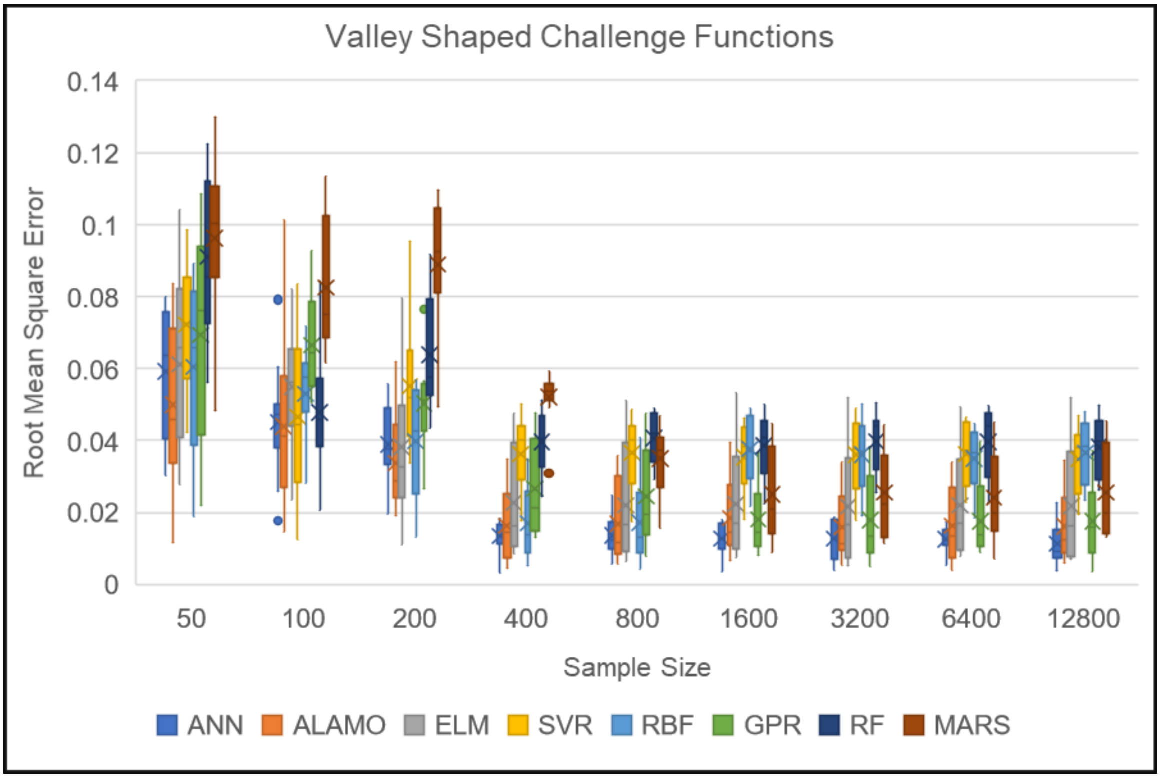
7.8. Results for Challenge Functions with a Valley Shape

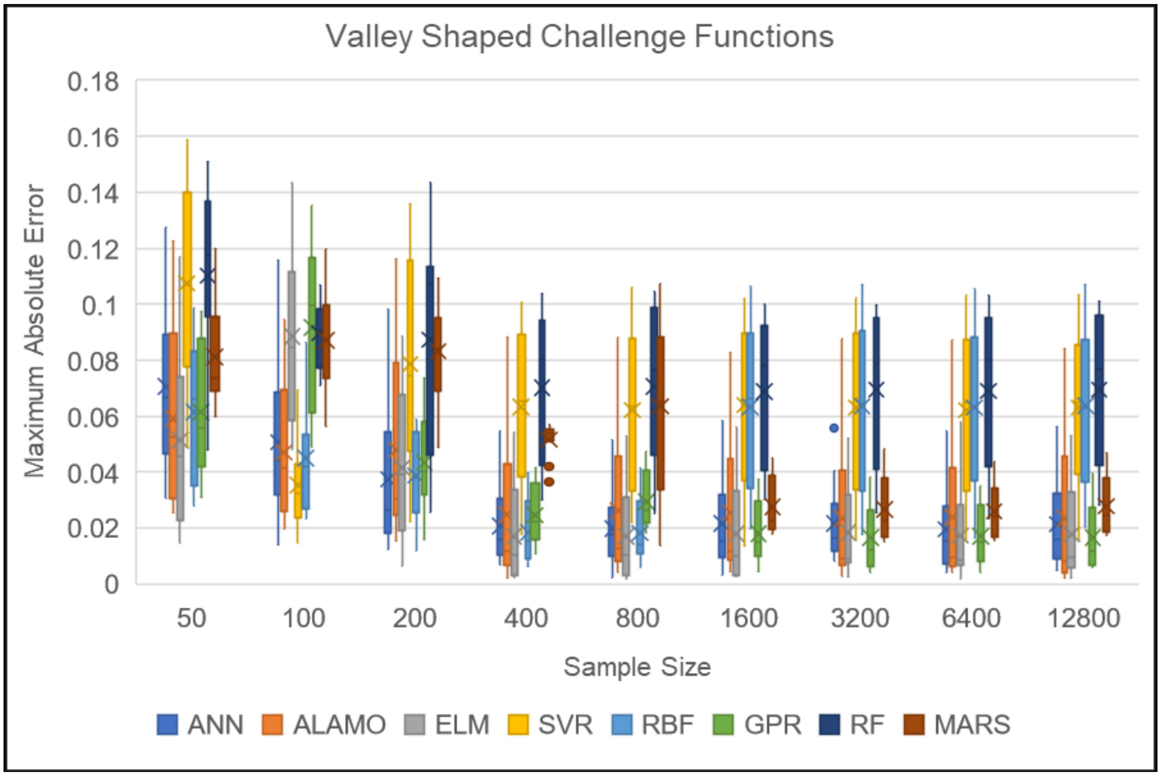




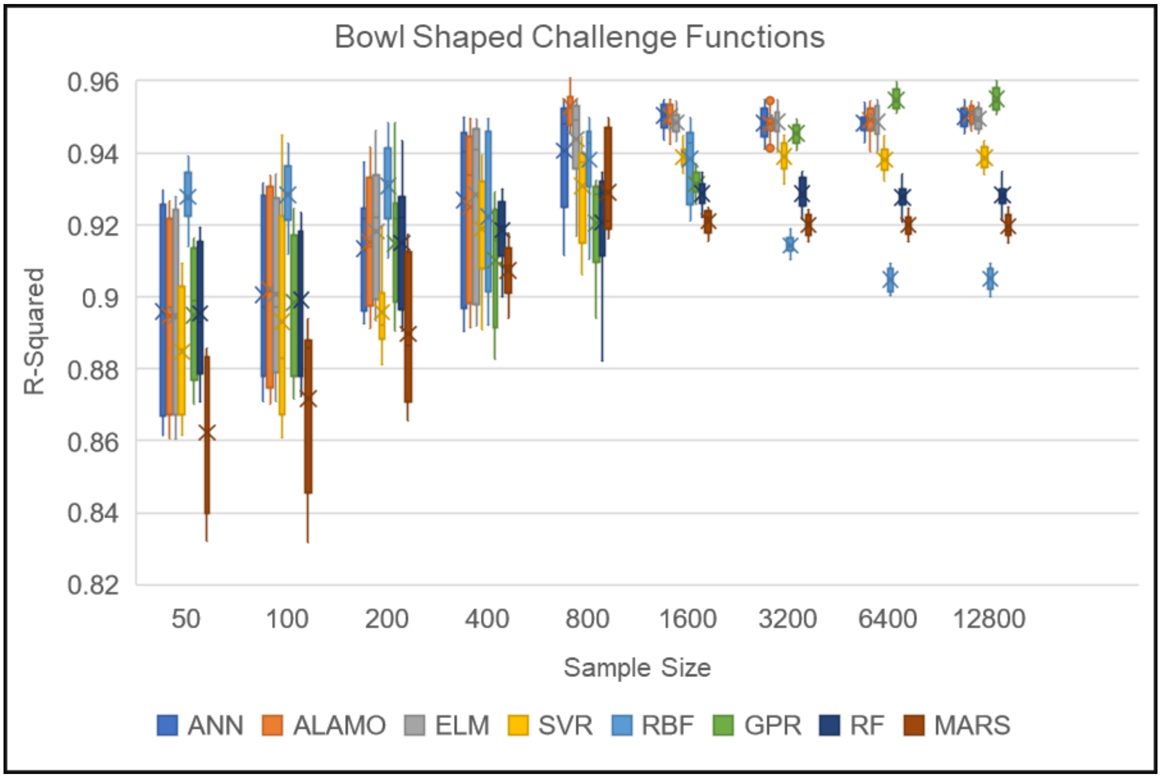

Valley Shaped Challenge Functions
0
-2000
-4000
-6000
-8000
-10000
-12000
-14000
Akaike Information Criterion
50
100
200
400
800
1600
3200
6400
12800
Sample Size
ANN
ALAMO
ELM
SVR
RBF
GPR
RF
MARS

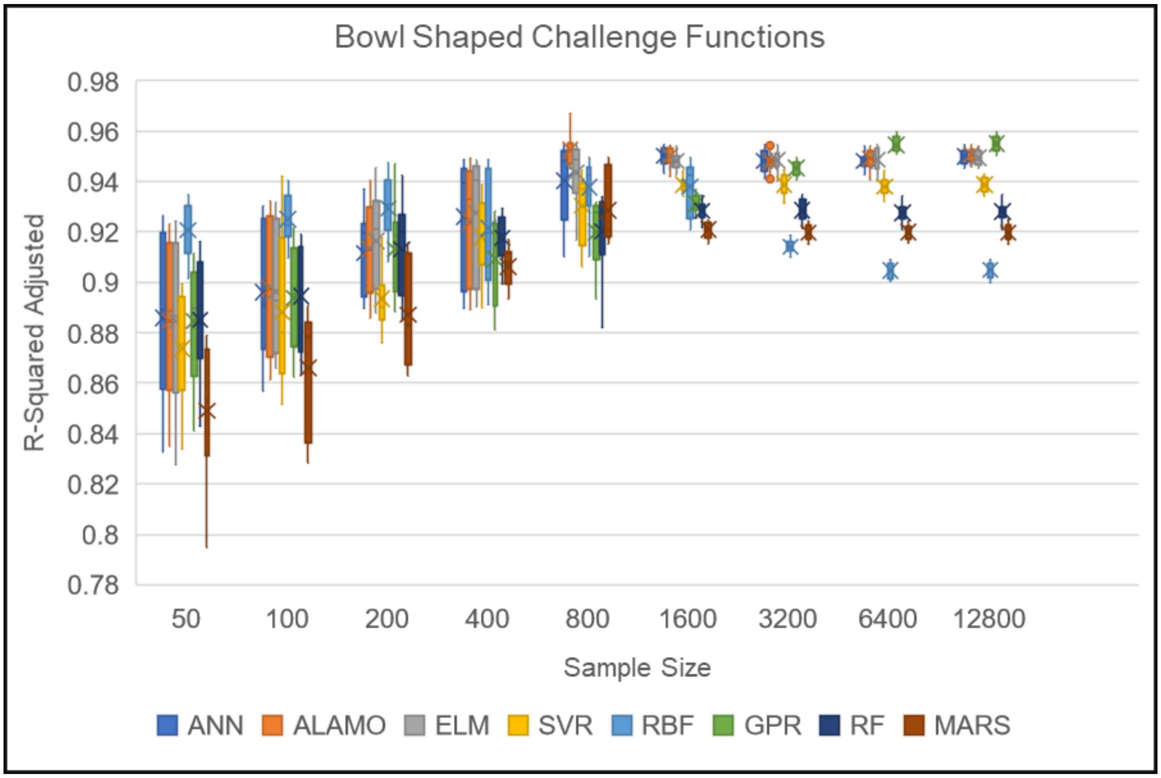

Valley Shaped Challenge Functions
0
-2000
-4000
-6000
-8000
-10000
-12000
Bayesian Information Criterion
50
100
200
400
800
1600
3200
6400
12800
Sample Size
ANN
ALAMO
ELM
SVR
RBF
GPR
RF
MARS

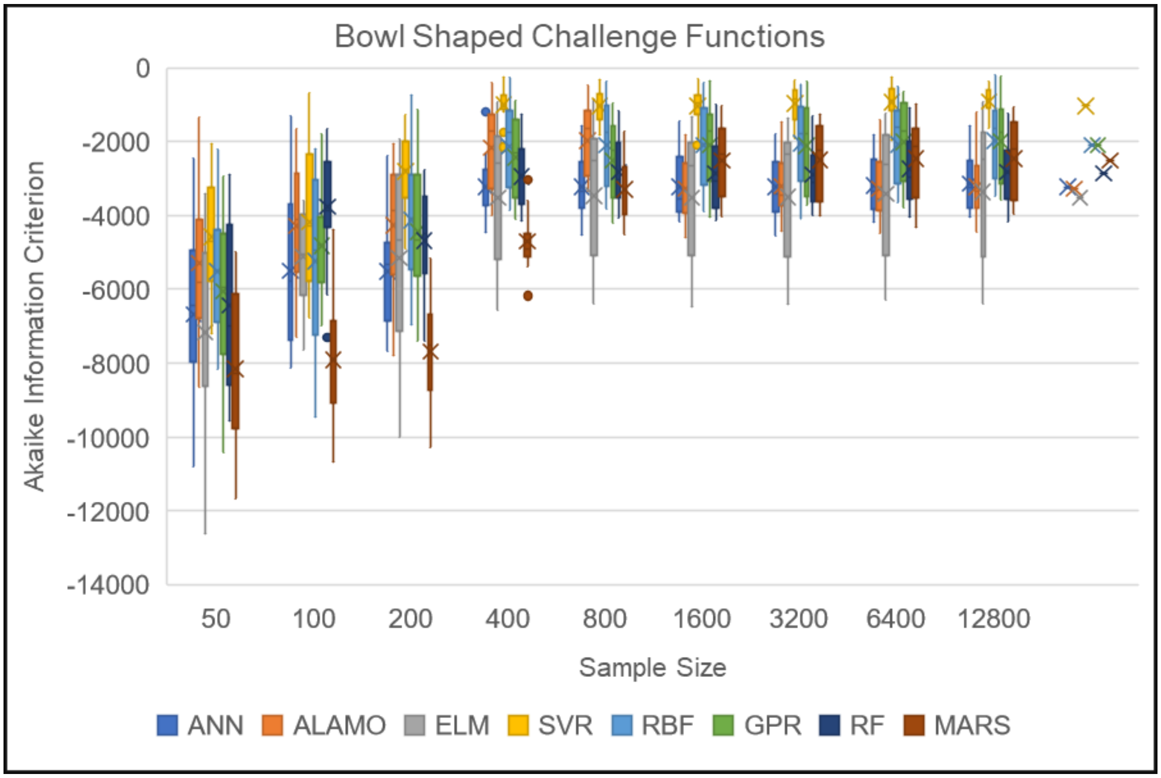

Valley Shaped Challenge Functions
Root Mean Square Error
0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
50
100
200
400
800
1600
3200
6400
12800
Sample Size
ANN
ALAMO
ELM
SVR
RBF
GPR
RF
MARS

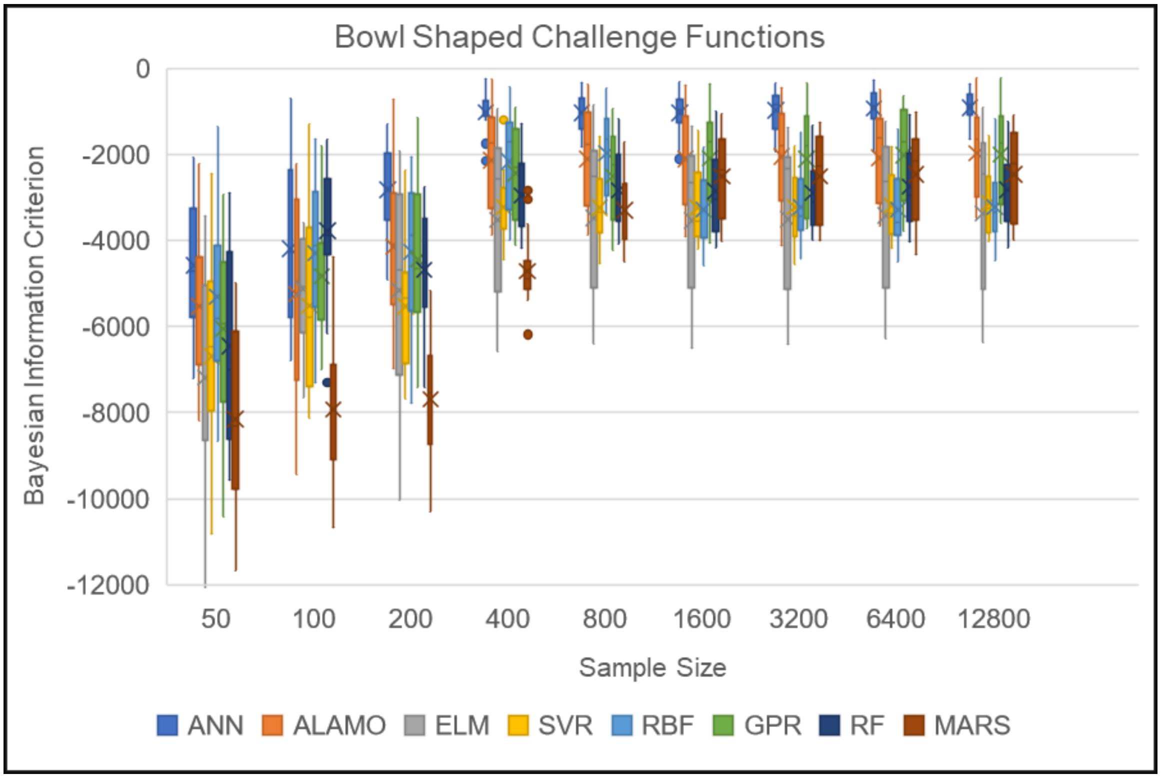

Valley Shaped Challenge Functions
Maximum Absolute Error
0.18
0.16
0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
50
100
200
400
800
1600
3200
6400
12800
Sample Size
ANN
ALAMO
ELM
SVR
RBF
GPR
RF
MARS

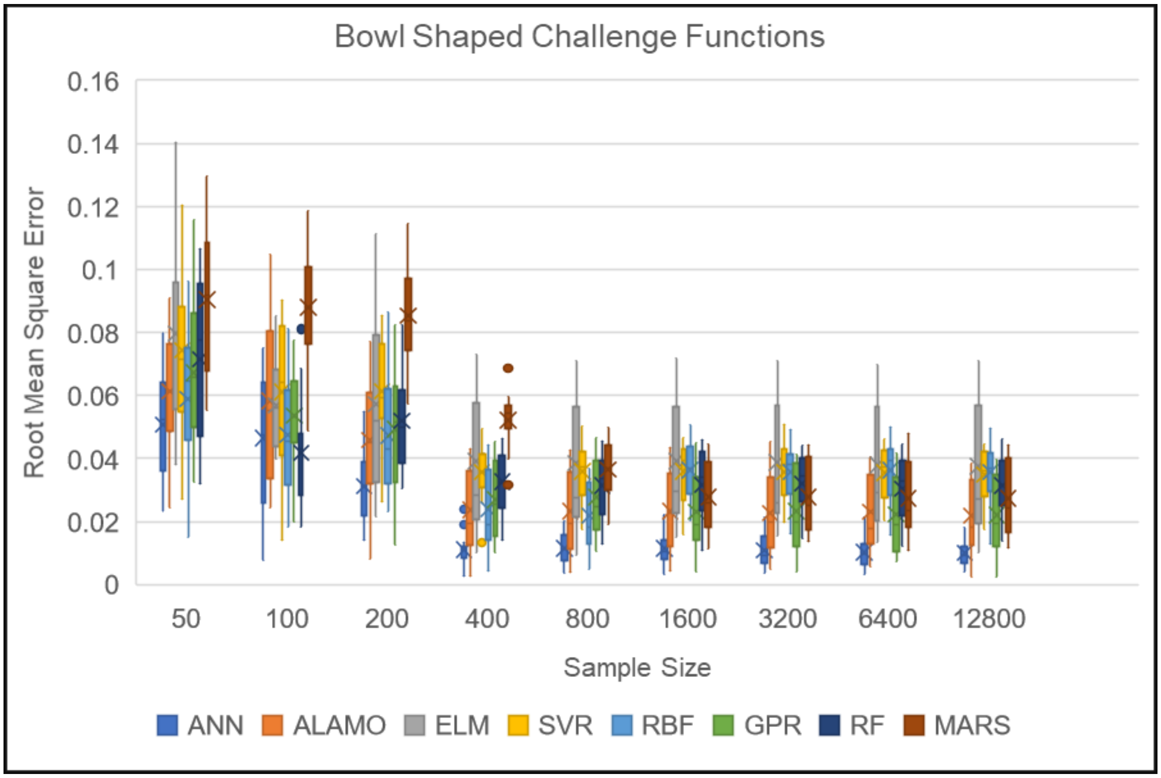
7.9. Results for Challenge Functions with a Bowl Shape

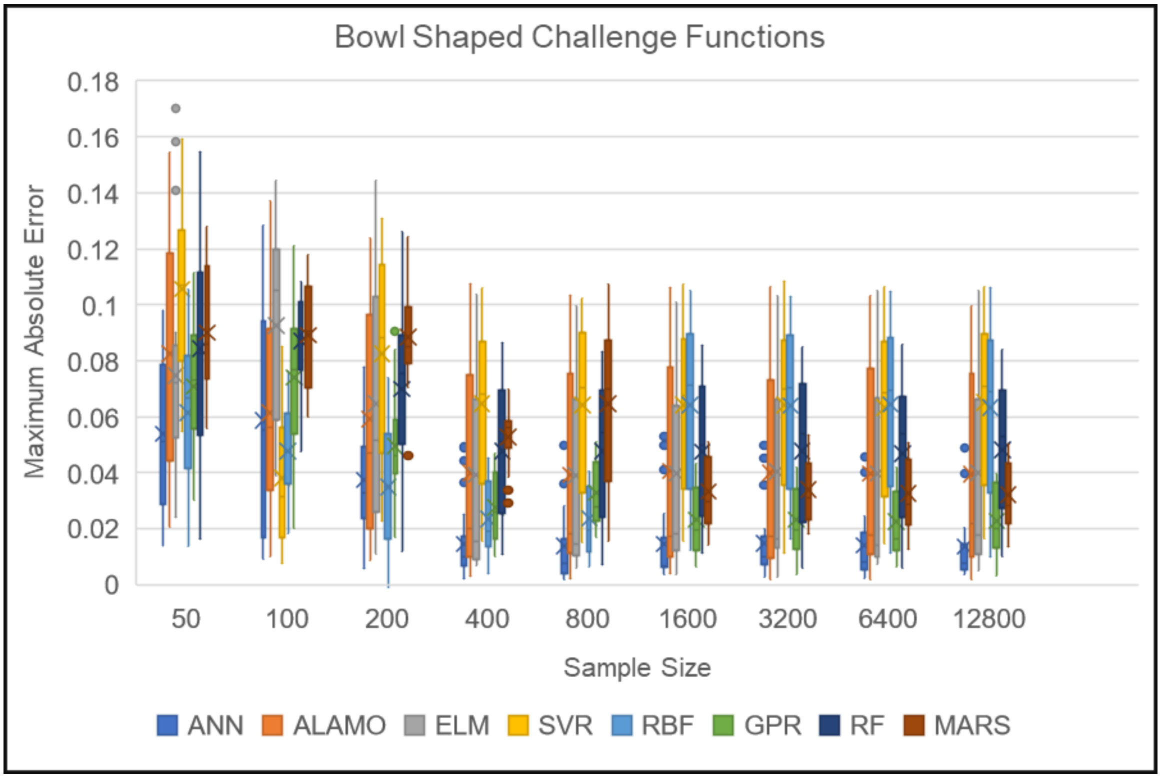





Bowl Shaped Challenge Functions
Akaike Information Criterion
0
-2000
-4000
-6000
-8000
-10000
-12000
-14000
50
100
200
400
800
1600
3200
6400
12800
Sample Size
ANN
ALAMO
ELM
SVR
RBF
GPR
RF
MARS


Bowl Shaped Challenge Functions
Bayesian Information Criterion
0
-2000
-4000
-6000
-8000
-10000
-12000
50
100
200
400
800
1600
3200
6400
12800
Sample Size
ANN
ALAMO
ELM
SVR
RBF
GPR
RF
MARS


Bowl Shaped Challenge Functions
Root Mean Square Error
0.16
0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
50
100
200
400
800
1600
3200
6400
12800
Sample Size
ANN
ALAMO
ELM
SVR
RBF
GPR
RF
MARS


Bowl Shaped Challenge Functions
Maximum Absolute Error
0.18
0.16
0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
50
100
200
400
800
1600
3200
6400
12800
Sample Size
ANN
ALAMO
ELM
SVR
RBF
GPR
RF
MARS