

Genetic Programming and Artificial Neural Networks in a Chemical Process System

Warren I. Luzano, Ph.D.

Mindanao University of Science and Technology

luzano.warren@yahoo.com

Abstract

The main objective in this paper is to develop a mathematical model of the glucose to gluconic acid batch fermentation process utilizing novel techniques such as Genetic Programming and Artificial Neural Networks. Downloaded experimental data incorporating the effects of the substrate (glucose) and biomass concentrations, and the dissolved oxygen content have been used to model the fermenter. Three Genetic Programming variants: the GPLAB (a matlab toolbox), GP-OLS (a hybrid GP and Orthogonal Least Square method) and GP-Eureqa, as well as Multi-Layer Perceptron Neural Network have been used and compared.

Keywords: Symbolic Regression, Multi-Layer Perceptron, Genetic Programming, Artificial Neural Network

I. Rationale

The problem of discovering a mathematical expression that describes the operation of a physical system using empirically observed data measurements is an important problem in many scientific areas. Sometimes, there is no known mathematical way (or a solution in closed form) to explain relationships between involved input-output variables. This can lead to faulty mathematical expressions or models.

For control engineers, modeling of a dynamic process is crucial and important because the process implements knowledge about structure and dynamics which are the starting point for the design and configuration of controllers.

There are two different classes of process models, the structured mathematical model and the unstructured/ input-output models. The first type involves an assumption of a model which may be derived analytically and then determines parameters in order to carry out its operation (e.g. linear regressions,, linearized/non-linearized differential equations, transfer functions, etc).

In general, the creation of these mathematical models is difficult, time-consuming, and expensive. The second type requires relevant input and output variables only, and a computational search algorithm (e.g machine learning techniques and data mining). The more available data there is, the more accurate and efficient the model can be. Moreover, due to recent developments in evolutionary computation, these leaning techniques can be used to discover useful models even automatically.

There are numerous application of GP in the field of process engineering: Mckay et.al (1996) modeled two continuous tank reactors in series; Willis et.al (1997) on binary vacuum distillation and screw cooking extruder; Greef and Aldrich (1997) on acid pressure leaching of nickeliferous chromites, and uranium and radium liberation models; Grossman et.al (2002) on modeling of a mixing tank; Grosman and Lewin (2004) on modeling of a Karr liquid-liquid extraction column; Chen (2003) on simultaneous zinc and lead imperial smelting process; Madar (2005) on identification of nonlinear systems; Lew et.al (2006) on identification of response surface moldes; Cai et.al (2006) on heat transfer correlations with symbolic regression; Brezocnik (2006) on mathematical model to predict machinability of steel; Duda (2011) on tensile strength of cast iron and many others.

GP performs symbolic regression, determining both the structure and the complexity of the model during its evolution. This has the advantage that no a priori modelling assumptions have to be made. Moreover, the technique can identify and separate relevant and irrelevant process inputs - yielding economical model structures that accurately represent process characteristics.

This paper is an extension of a paper by Babu (2007). He noted that there are problems involved in modeling of glucose to gluconic acid bioprocess. The glucose to gluconic acid bioconversion using *A. niger* immobilized on the cellulosic micro fibrils involves reaction and mass transfer phenomena that leads to a complicated task since the physicochemical phenomena underlying the bioconversion and the associated kinetic and transport mechanisms are not well-understood. Aside from this, it has been observed that the process dynamics is nonlinear. The paper instead utilized genetic programming for symbolic regression, using GPLAB (a Matlab toolbox) and a hybrid GP and Orthogonal Least Square method (GP OLS). It was concluded in the paper that GP OLS had the good prediction accuracy (i.e. high fitness values and low mean square error) and had generated simple parsimonious expressions.

This paper will focus on the performance of Eureqa, a third party freeware software that utilizes Genetic Programming, and then compare its performance with that of two popular Matlab Genetic programming toolboxes, the GPLAB and GP-OLS utilized by the abovementioned paper. Another popular freeware software called WEKA which contains a collection of visualization tools and algorithms for data analysis and predictive modeling was also used in order to carry out the multi-layer perceptron (MLP), a feedforward artificial neural network model.

II. Theoretical Framework

2.1 Genetic Programming

Given a problem, the goal of genetic programming is to develop a computer program that can solve the particular problem. It starts with a population of randomly generated computer programs. Each program is run, and its performance is measured by a "fitness function", which reflects how good each program is at solving the problem. Then programs are chosen from the population to "reproduce, and the cycle continues, until a program with a certain level of fitness (i.e. a program that is good enough to solve the problem given) is found.

The programs are chosen randomly, but with a weighting mechanism that makes it more likely that the more "fit" programs are chosen. After a full new population of programs is formed, the old population is discarded, and each of the new programs is run and their fitness is measured. The word "until" implies that such a state will be reached. The basic GP algorithm is similar to any evolutionary algorithms and works as follows.

Algorithm GP Evolution:

- Step 1. Begin
- Step 2. Define pop-size as desired population size
- Step 3. Randomly initialize pop-size population
- Step 4. While (Ideal best found or certain number of generations met)
 - o Evaluate fitness
 - o While(number of children=population size)
 - o Select parents
 - o Apply evolutionary operators to create children
 - o End while
- Step 5. End While
- Step 6. Return Best solution
- Step 7. End

2.1.1 Solution Initialization

The innovation of GP lies in the variable sized solution representation which requires efficient initial population construction. Individuals are represented as trees constructed randomly from a primitive set. This primitive set contains functions and terminals. A tree's internal nodes are selected from the functions and leaf nodes are selected from the terminals. Initialization plays an important role in success of an evolutionary algorithm. A poor initial population can cause any good algorithm to get stuck in local optima. On the other hand a good initialization can make most of the algorithms work sufficiently well. Three initialization techniques popular in tree based GP are full method grow method ramped half and half method

2.1.2 Selection

The evolutionary operators are applied on individuals particularly selected for that operation. The individuals are selected using a particular selection mechanism. Two of such mechanisms are tournament and fitness proportionate selection.

2.1.3 Operators

The most common operators used for evolution of GP programs are crossover, mutation and reproduction.

Crossover operator works by selecting two parents from the population. Two random subtrees are selected from each parent and swapped to create children. Advancements have been made to pure random crossover operator

in order to make it more efficient and propagate good building blocks among generations. The second operator is called mutation which has three types. In "point mutation" a single node in parent tree is selected and replaced with a random node of same type. "Shrink mutation" selects a node randomly and the subtree rooted at that node is replaced by a single terminal node. Lastly, "grow mutation" selects a random node and a randomly generated subtree is replaced by the subtree rooted at that node.

Reproduction operator works by selecting an individual and copies it directly to the new generation without any changes or modifications to it.

2.1.4 Solution Fitness

Fitness is the performance of an individual corresponding to the problem it is aimed to solve. It tells which elements or the regions of the search space are good. The fitness measure steers the evolutionary process towards better approximate solutions to the problem. Fitness of individuals in a population can be measured in many ways. It can be measure of error between the original and desired output of a solution. It can be compliance of the structure to the task it is required to solve based on a user specified criteria. The difference between fitness evaluation in GP and other evolutionary algorithms is that each individual of GP is a program which needs recursive execution of the nodes of the tree in precise manner. This adds overhead to the algorithm increasing its evolution time and required computational sources.

2.2 Artificial Neural Network

Artificial neural networks (ANN) have been developed as generalizations of mathematical models of biological nervous systems. A first wave of interest in neural networks emerged after the introduction of simplified neurons by McCulloch and Pitts (1943) also known as connectionist models. An Artificial Neural Network is a network of collections of very simple processors ("Neurons") each possibly having a (small amount of) local memory. The units operate only on their local data and on the inputs they receive via the connections or links which are unidirectional (Ajith, 2006). A network unit has a rule for summing the signals coming in and a rule for calculating an output signal that is then, sent to other network units. According to Callen the rules for calculating the output is known as the activation function (Zhi-Hua, 2007)

A neural network has three layers in its structure. First layer is input layer which is directly interact with external worlds; second layer is of hidden unit where computation is done according to function provided, the last layer is output layer from where we get output. Knowledge in neural networks is stored as synaptic weights between neurons. The network propagates the input data from layer to layer until the output data is generated.

If the networks is multilayer perceptron with Backpropagation algorithm and the output is different from the desire output, then an error is calculated and propagated backwards through the network. The synaptic weights are modified as the error is propagated (Fiona, 2001)

2.2.1 Types Of Learning In Neural Networks

The first type is called supervised learning. In supervised learning networks inputs as well as corresponding output is given to the networks. In forward the errors or discrepancies between the desired and actual response for each node in the output layer are found. These are then used to determine weight changes in the net according to the prevailing learning rule. The term supervised originates from the fact that the desired signals on individual output nodes are provided by an external teacher. Examples of supervised learning are the delta rule, and the perceptron rule (Fauset, 1996)

The second is unsupervised learning. In this type of learning external teacher is not present. So this kind of learning is based upon clustering technique. According to inputs patterns clusters are divided into different classes. This kind of learning is also called self organization. Typical examples are the Hebbian learning rule and the competitive learning rule. Unsupervised learning is much more important than supervised learning since it likely to be much more common in the brain than supervised learning. The kind of learning is determined by the way in which the changes to network parameters have done.

Lastly, we have what we called reinforcement learning. This kind of learning is based upon both supervised and unsupervised learning. Reinforcement learning is learning what to do how to map situations to actions so as to maximize a numerical reward signal. In this learning there is reward for correct outputs and penalty for wrong outputs. Reinforcement learning is also called learning with a critic as opposed to learning with a teacher (Sutton, 1999)

III. Methodologies

3.1. The Chemical Process

In the work of Babu (2007) for the development of a GP based model for the glucose to gluconic acid bioprocess, experimental input-output data from the fermenter were used.

A. Fermentation Medium for Immobilized Mycelia

Anhydrous purified glucose (100 g), $\text{MgSO}_4 \cdot 7\text{H}_2\text{O}$ (0.035g), KH_2PO_4 (0.05 g), and 0.1 g of $(\text{NH}_4)_2\text{HPO}_4$ were dissolved in 1 L of water. The pH of this medium was adjusted to 6.0 using $1\text{M H}_2\text{SO}_4$. A woven cellulosic fabric support (69 _ 8.5 _ 0.6 cm) with void volume of approximately 140 mL was sterilized at 15 psi for 60 min.

B. Submerged Fermentation

Submerged fermentation utilizing the immobilized culture was carried out in a modified locally fabricated batch fermenter. In the fermenter, the matrix with fully grown *A. niger* was folded in a spiral shape. For preventing mycelial recirculation, the upper end of the fixed bed was closed by the filter mesh. The batch reactor was drained after the substrate reached its lowest concentration.

C. Maintaining Oxygen Partial Pressure

A constant flow of air was used to maintain the oxygen partial pressure and a Dissolved Oxygen (DO) probe (Ingold, 170-ppm type DO amplifier) was used for measuring the dissolved oxygen concentration.

D. Glucose and Gluconic Acid Analyses

Feed and the unconverted glucose were analyzed by the dinitrosalicylic acid method and the gluconic acid concentration in the bioreactor was measured by titrating against 6 N NaOH.

The data set (see Table 1) comprises of glucose concentration (g/L), biomass concentration (g/L), and dissolved oxygen (DO) concentration (mg/L), and the corresponding process output variable, i.e., gluconic acid concentration (y).

| Batch | Glucose concentration (x1) | Biomass concentration (x2) | Dissolved Oxygen (x3) | Gluconic Acid concentration (y) |
|-------|----------------------------------|----------------------------------|-----------------------------|---------------------------------------|
| 1 | 100 | 1 | 10 | 6.416 |
| 2 | 150 | 2 | 10 | 48.015 |
| 3 | 200 | 2 | 15 | 27.1 |
| 4 | 150 | 2.5 | 15 | 57.946 |
| 5 | 150 | 3 | 15 | 57.389 |
| 6 | 120 | 2 | 25 | 36.262 |
| 7 | 120 | 2 | 30 | 45.02 |
| 8 | 150 | 2 | 30 | 94.424 |
| 9 | 150 | 3 | 25 | 80.486 |
| 10 | 150 | 2 | 40 | 128.907 |
| 11 | 150 | 2 | 45 | 146.036 |
| 12 | 150 | 2 | 50 | 154.23 |
| 13 | 180 | 2 | 50 | 175.525 |
| 14 | 150 | 3 | 40 | 129.006 |
| 15 | 150 | 2.5 | 50 | 154.36 |
| 16 | 150 | 2.5 | 55 | 152.44 |
| 17 | 150 | 2.5 | 60 | 148.94 |
| 18 | 160 | 2.5 | 60 | 163.067 |
| 19 | 175 | 3 | 55 | 176.49 |
| 20 | 160 | 3 | 60 | 162.42 |
| 21 | 180 | 3 | 60 | 172.598 |
| 22 | 150 | 3 | 60 | 151.28 |
| 23 | 100 | 3 | 60 | 21.803 |
| 24 | 100 | 2 | 10 | 6.67 |
| 25 | 120 | 2.5 | 10 | 22.952 |
| 26 | 100 | 2 | 15 | 7.829 |
| 27 | 150 | 2 | 15 | 57.261 |
| 28 | 120 | 2 | 20 | 31.486 |
| 29 | 150 | 2 | 20 | 66.9 |
| 30 | 150 | 2.5 | 20 | 67.449 |
| 31 | 150 | 3 | 20 | 67.328 |
| 32 | 150 | 2 | 35 | 111.328 |
| 33 | 150 | 2.5 | 30 | 95.988 |
| 34 | 150 | 3 | 30 | 94.707 |
| 35 | 150 | 2.5 | 40 | 129.93 |
| 36 | 150 | 3 | 35 | 111.604 |
| 37 | 150 | 2 | 60 | 152.43 |
| 38 | 120 | 2 | 60 | 73.502 |
| 39 | 150 | 3 | 45 | 144.651 |
| 40 | 180 | 2.5 | 55 | 179.064 |
| 41 | 150 | 3 | 50 | 152.89 |
| 42 | 180 | 2.5 | 60 | 174.483 |
| 43 | 150 | 3 | 55 | 154.23 |
| 44 | 166 | 3 | 60 | 169.45 |
| 45 | 165 | 3 | 60 | 167.91 |
| 46 | 162 | 3 | 60 | 164.87 |

Table 1. Experimental Data Utilized for Building GP Model

3.4. Quality of Models

There exists a large variety of measure values characterizing models qualities. Criteria like (a) error measures in the estimation and validation period (e.g mean squared error, mean absolute error, mean absolute percentage error, mean error, mean percentage error), (b) residual diagnostics and goodness-of-fit tests (e.g. plots of residuals versus time, versus predicted values, and versus other variables; residual autocorrelation plots, and normal probability plots; coefficients of skewness and kurtosis; measures of extreme or influential observations, etc), and (c) qualitative considerations (e.g. appearance of forecast plots, intuitive reasonableness of the model, simplicity of the model, etc)

With so many plots and statistics and considerations to worry about, it's sometimes hard to know which comparisons are most important. If there is any one statistic that normally takes precedence over the others, it is the mean squared error (MSE) within the estimation period. This is the statistic whose value is minimized

during the parameter estimation process. The MSE equals the variance of the errors plus the square of the mean error (i.e: $MSE = VAR(E) + (ME)^2$). Hence, to minimize mean squared error, means implicitly minimizing the bias, as well as the variance of the errors.

Another famous and widely-used measure value is simply denoted by the so-called Mean Absolute Error (MAE). This is a common measure used to compare the prediction performance of different models. This criterion is generally referred to as predictability criterion. The MAE is the average magnitude of the errors in a set of forecasts over the verification sample of the absolute values of the differences between forecast and the corresponding observation.

IV. Results and Discussion

4.1. The Model

Table 2 below shows the Symbolic Regression models. The best symbolic regression function that was found on this data set is selected by choosing the function with the lowest Mean Square Error (MSE). This model has a Mean Square Error of 0.0012, as well as a Mean Absolute Error (MAE) of 0.0260. This chosen model is:

$$(\text{Gluconic Acid}) = 2.231 + 25.34 * (\text{Dissolved Oxygen})^2 + 6.231 * \text{Glucose}^2 + 2.646 * \text{Glucose} * (\text{Dissolved Oxygen}) - 6.228 * \text{Glucose} - 8.727 * (\text{Dissolved Oxygen}) - 1.807 * \text{Glucose}^3 - 26.05 * (\text{Dissolved Oxygen})^3$$

| Size | Model Equation | MSE | MAE |
|------|---|--------|--------|
| 43 | (Gluconic Acid) = 2.231 + 25.34*(Dissolved Oxygen)^2 + 6.231*Glucose^2 + 2.646*Glucose*(Dissolved Oxygen) - 6.228*Glucose - 8.727*(Dissolved Oxygen) - 1.807*Glucose^3 - 26.05*(Dissolved Oxygen)^3 | 0.0012 | 0.0260 |
| 39 | (Gluconic Acid) = 25.81*(Dissolved Oxygen)^2 + 2.814*Glucose*(Dissolved Oxygen) + 1.886*Glucose^2 - 0.6191 - 9.13*(Dissolved Oxygen) - 0.8328*Glucose^3 - 26.38*(Dissolved Oxygen)^3 | 0.0014 | 0.0273 |
| 37 | (Gluconic Acid) = 2.682*Glucose + 25.93*(Dissolved Oxygen)^2 + 2.894*Glucose*(Dissolved Oxygen) - 1.838 - 9.28*(Dissolved Oxygen) - 0.4067*Glucose^3 - 26.47*(Dissolved Oxygen)^3 | 0.0016 | 0.0285 |
| 35 | (Gluconic Acid) = 5.22*Glucose + 26.04*(Dissolved Oxygen)^2 + 2.966*Glucose*(Dissolved Oxygen) - 2.988 - 9.41*(Dissolved Oxygen) - 1.792*Glucose^2 - 26.56*(Dissolved Oxygen)^3 | 0.0018 | 0.0304 |
| 5 | (Gluconic Acid) = 1.781*Glucose*(Dissolved Oxygen) | 0.0465 | 0.1711 |

Table 2: The model equations according to the size with corresponding MSE, MAE, R-Square and No. of coefficients.

It can be noted that genetic programming gives more than one equation to a given set of input output data, with varying mean square errors. This is somehow necessary for it presents user more choices to consider. For example, in the Eureqa model of size = 5, the variables “dissolved oxygen” and “glucose” are being just being multiplied. The rest of the models contain the same two variables that are being raised to the second and third powers. It seems like the variable biomass concentration has nothing to do in the outcome of the gluconic acid concentration - as being suggested by the data.

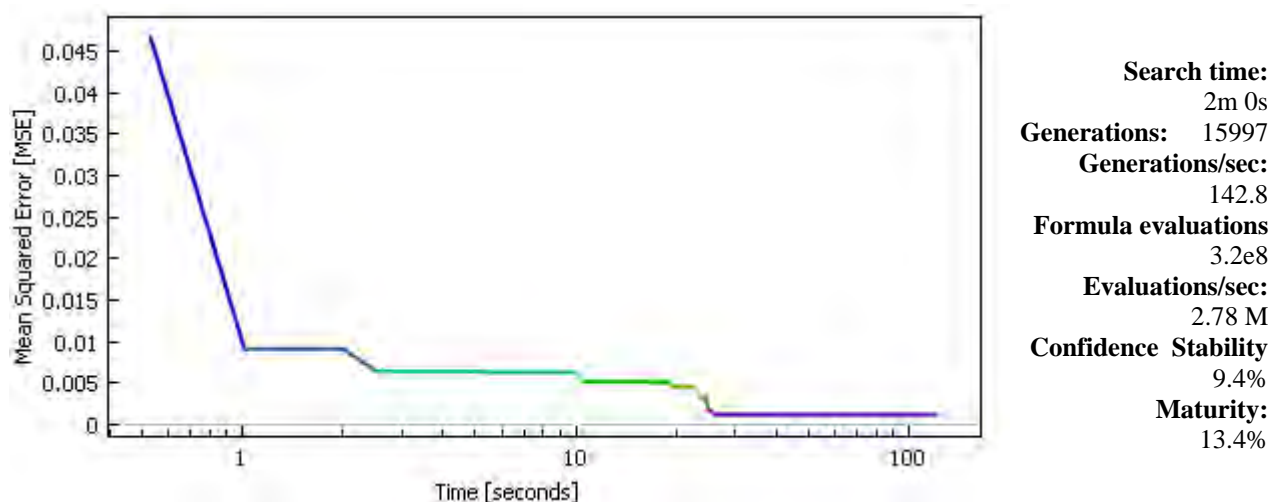


Figure 1: The mean absolute error (MAE) as a function of time, along with the number of generations, formula evaluations, confidence stability and maturity.

The total calculations for the model of size 33 took a search time of 120 seconds. In this span of time, about 3.2×10^7 formula evaluations seconds were made with an MSE of 0.0465 (see Fig. 1.).

4.2. Comparison with the other GP variants

Table 3 below shows how the eureqa model competes with that of two variants of matlab toolboxes GPLAB and GP-OLS and with Multi Layer Perceptron (MLP).

| Model | Model Equation $y = f(x_1, x_2, x_3)$ | MSE | Time (secs) |
|-------------------|---|---------|-------------|
| GPLAB | $y = \text{mypower}(\text{times}(\text{times}(\text{mydivide}(X1, \text{mypower}(\text{times}(\text{mypower}(\text{times}(X1, X3), \text{minus}(0.82359, \text{minus}(X1, \text{mypower}(X3, X1))))), \text{times}(\text{minus}(X1, \text{minus}(0.47932, X3)), X3)), \text{times}(\text{times}(\text{mydivide}(X1, \text{mypower}(\text{times}(\text{mypower}(\text{mypower}(\text{times}(X1, X3), \text{minus}(0.82359, \text{minus}(X1, X3))), \text{minus}(0.82359, \text{minus}(\text{times}(X3, X2), X3))), \text{times}(\text{minus}(X1, \text{minus}(0.47932, X3)), X3)), \text{times}(X1, \text{times}(\text{minus}(0.82359, \text{minus}(0.47932, X3)), X1))), X3), \text{times}(\text{minus}(0.82359, \text{minus}(0.47932, X3)), X1))), X3), \text{times}(\text{minus}(0.82359, \text{minus}(0.47932, X3)), X1)), 0.38498)$ | – | – |
| GP-OLS (2) | $y = -2.276953 * (x_3) + 2.323131 * ((x_3) - \exp((x_1) - \exp((x_1) * \exp(x_3)))) + 0.888878$ | 0.03206 | 205 |
| GP-OLS (1) | $y = 0.517192 * ((x_1) - \exp(((x_3) * \exp((x_3) * \exp(x_3)))) * (((x_3) - \exp(x_1)) * (x_3)))) + 0.517702$ | 0.02759 | 288 |
| Eureqa (1) | $y = 2.231 + 25.34 * (x_3)^2 + 6.231 * x_1^2 + 2.646 * x_1 * (x_3) - 6.228 * x_1 - 8.727 * (x_3) - 1.807 * x_1^3 - 26.05 * (x_3)^3$ | 0.0012 | 200 |
| MLP -ANN | – | 27.22 | 16 |

Table 4: The results of the three variants of GP (GP-Eureqa, GPLAB and GP-OLS) and Multi Layer Perceptron (MLP-ANN)

The first model, GPLAB has a very complicated equation and has no practical use. Two equations of the GP-OLS are less complex however; their mean square errors are 0.03206 and 0.02759 respectively. Comparatively, the eureqa model has the lowest mean square error stabilized at 0.0012. Since the mean squared error is comprised of the variance of the estimator and the squared bias; this means that the smaller the variance, the more precise the estimator is; and the smaller the bias, the more accurate the estimator is. Overall, the eureqa model is the most accurate and precise among the three GP variants.

The multi layer perceptron model have a mean square error of 27.22, suggesting that it is not a good model in this particular chemical process.

Moreover, looking at the running time process, eureka model have already come up with the best equation already after 120 seconds, compared to 205 and 288 for the GPLAB and GP-OLS respectively.

V. Conclusion

The aim of this study was to compare the three GP variants: GPLAB, GP-OLS and GP-Eureka, together with a model created by an artificial neural network, as they attempt to create a mathematical model for the glucose and gluconic acid chemical process using the downloaded experimental data.

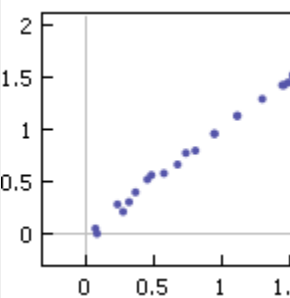
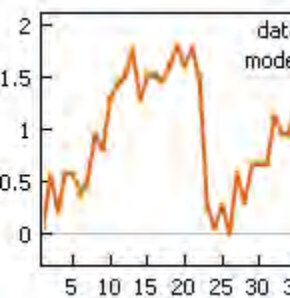
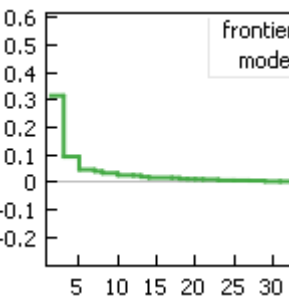
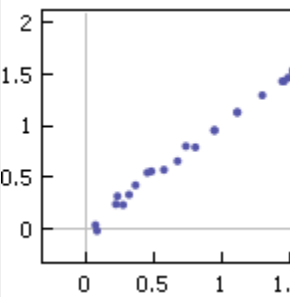
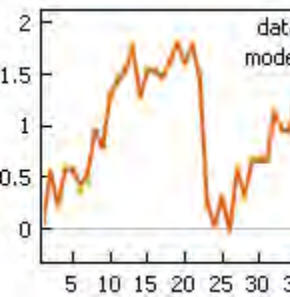
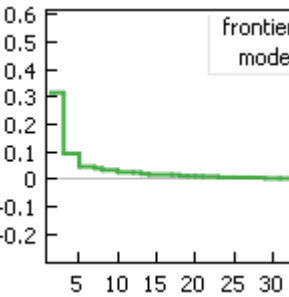
The quality and reliability of the four models were evaluated via performance indicators (MSE and MAE). Assessment of model performance indicated that the eureka model is the most accurate and precise among the four models.

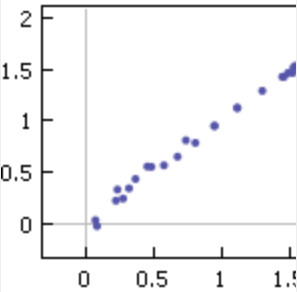
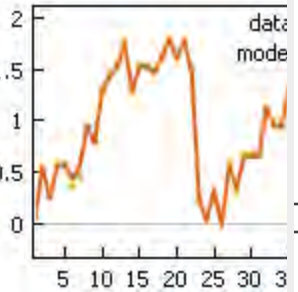
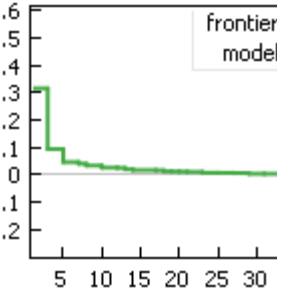
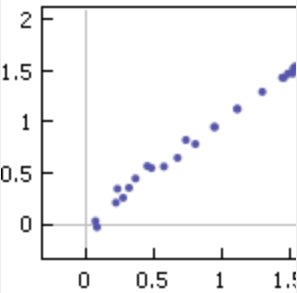
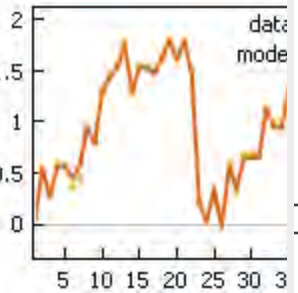
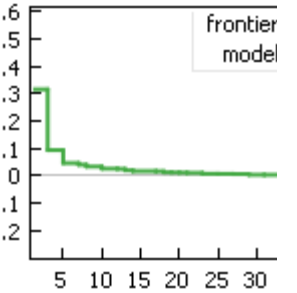
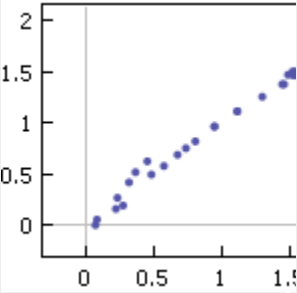
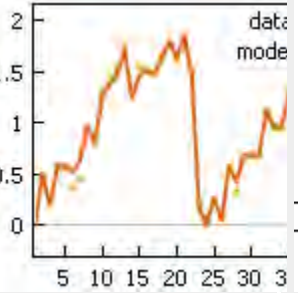
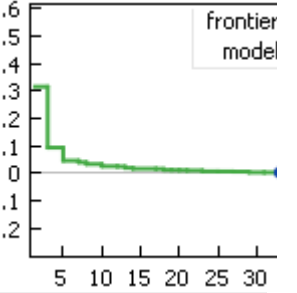
Aside from this, the eureka model was able to converge to a very minimum mean square error less than two minutes time, suggesting further that the computational running time is much faster compared to GPLAB and GP-OLS matlab toolboxes.

References

- [1] Babu, V. B., "Genetic Programming for Symbolic Regression of Chemical Process Systems", Engineering Letters, 2007 Vol 14 Issue: 2 Pages: 42-55
- [2] M Brezocnik, Kovacic, M, "Prediction of steel machinability by genetic programming", Journal of Achievement in Materials and Manufacturing Engineering, 2006 Vol 16 Issue: 1-2 Pages: 107-113
- [3] Duda, J. Stawowy, A., "Genetic Programming for the prediction of tensile strength of cast iron", Archives of Foundry Engineering, 2011 Vol 11 Issue: 4 Pages: 31-34
- [4] Paterlini, S; Minerva, T.; "Regression Model Selection Using Genetic Algorithms" Recent Advances in Neural Networks, Fuzzy Systems and Evolutionary Computing
- [5] Taneja, A.; Chauhan, R. "A Performance Study of Data Mining Techniques: Multiple Linear Regression vs. Factor Analysis" International Journal Of Innovative Technology and Creative Engineering, April 2011
- [6] Antonoaie, V; Antonoaie, C; "FORECASTING THE SELLING PRICE OF TIMBER AUCTIONS", Bulletin of the Transilvania University of Braşov Vol. 5, No. 1 – 2012
- [7] Karaboga, D., Ozturk C., "Artificial bee colony programming for symbolic regression" Journal of Information Sciences, Volume 209, 20 November 2012, Pages 1–15

Appendix 1. Eureka Summary Report

| Model | Observed vs. Predicted | Output vs. Row | Error/Complexity Pareto |
|--|---|--|---|
| Model: $\begin{aligned} \text{(Gluconic Acid)} = & a + \\ & b * (\text{Dissolved Oxygen})^2 + \\ & c * \text{Glucose}^2 + \\ & d * \text{Glucose} * (\text{Dissolved Oxygen}) - \\ & e * \text{Glucose} - \\ & f * (\text{Dissolved Oxygen}) - \\ & g * \text{Glucose}^3 - \\ & h * (\text{Dissolved Oxygen})^3 \end{aligned}$ |  |  |  |
| Model: $\begin{aligned} \text{(Gluconic Acid)} = & a * (\text{Dissolved Oxygen})^2 + \\ & b * \text{Glucose} * (\text{Dissolved Oxygen}) + \\ & c * \text{Glucose}^2 - d - \\ & e * (\text{Dissolved Oxygen}) - \\ & f * \text{Glucose}^3 - \\ & g * (\text{Dissolved Oxygen})^3 \end{aligned}$ |  |  |  |

| Model | Observed vs. Predicted | Output vs. Row | Error/Complexity Pareto |
|--|---|--|---|
| Model: $(\text{Gluconic Acid}) = a * \text{Glucose} + b * (\text{Dissolved Oxygen})^2 + c * \text{Glucose} * (\text{Dissolved Oxygen}) - d - e * (\text{Dissolved Oxygen}) - f * \text{Glucose}^3 - g * (\text{Dissolved Oxygen})^3$ |  |  |  |
| Model: $(\text{Gluconic Acid}) = a * \text{Glucose} + b * (\text{Dissolved Oxygen})^2 + c * \text{Glucose} * (\text{Dissolved Oxygen}) - d - e * (\text{Dissolved Oxygen}) - f * \text{Glucose}^2 - g * (\text{Dissolved Oxygen})^3$ |  |  |  |
| Model: $(\text{Gluconic Acid}) = a * \text{Glucose}^2 + b * \text{Glucose} * (\text{Dissolved Oxygen})^2 - c - d * \text{Glucose}^3 - e * (\text{Dissolved Oxygen})^4$ |  |  |  |

Appendix 2. Log Output

| TIME | Fitness | Equation |
|-----------|---------|---|
| 2:31:27PM | 0.0466 | $(\text{Gluconic Acid}) = 1.781 * \text{Glucose} * (\text{Dissolved Oxygen})$ |
| 2:31:27PM | 0.0940 | $(\text{Gluconic Acid}) = 2.713 * (\text{Dissolved Oxygen})$ |
| 2:31:28PM | 0.0092 | $(\text{Gluconic Acid}) = 7.824 * \text{Glucose} + 4.071 * \text{Glucose} * (\text{Dissolved Oxygen}) + 0.1979 * \text{Biomass}^2 - 4.995 - 0.04695 * \text{Biomass} - 3.951 * (\text{Dissolved Oxygen}) - 0.6557 * \text{Glucose} * \text{Biomass} - 281 * \text{Glucose}^2$ |

| | | |
|-----------|--------|---|
| 2:31:28PM | 0.0253 | (Gluconic Acid) = $1.702 \cdot \text{Glucose} \cdot (\text{Dissolved Oxygen}) + 0.9731 \cdot \text{Glucose} \cdot \text{Biomass} - 0.03697 - 0.08839 \cdot \text{Biomass} - 0.4798 \cdot (\text{Dissolved Oxygen}) - 0.2677 \cdot \text{Biomass}^2 - 0.6239 \cdot \text{Glucose}$ |
| 2:31:29PM | 0.0092 | (Gluconic Acid) = $7.827 \cdot \text{Glucose} + 4.105 \cdot \text{Glucose} \cdot (\text{Dissolved Oxygen}) + 0.1925 \cdot \text{Biomass}^2 - 5.041 - 4.002 \cdot (\text{Dissolved Oxygen}) - 0.6697 \cdot \text{Glucose} \cdot \text{Biomass} - 2.276 \cdot \text{Glucose}^2$ |
| 2:31:29PM | 0.0442 | (Gluconic Acid) = $\text{Glucose} \cdot \sin(2 \cdot (\text{Dissolved Oxygen}))$ |
| 2:31:29PM | 0.0348 | (Gluconic Acid) = $1.089 \cdot \text{Glucose} \cdot \sin(2 \cdot (\text{Dissolved Oxygen}))$ |
| 2:31:29PM | 0.0340 | (Gluconic Acid) = $1.013 \cdot \text{Glucose} \cdot \sin(2.292 \cdot (\text{Dissolved Oxygen}))$ |
| 2:31:29PM | 0.0342 | (Gluconic Acid) = $\text{Glucose} \cdot \sin(2.306 \cdot (\text{Dissolved Oxygen}))$ |
| 2:31:29PM | 0.0309 | (Gluconic Acid) = $0.3508 \cdot \text{Biomass} + 0.1543 \cdot \text{Glucose} + 2.75 \cdot \text{Glucose} \cdot (\text{Dissolved Oxygen}) - 0.4612 - 2.011 \cdot (\text{Dissolved Oxygen}) - 0.005714 \cdot \text{Glucose} \cdot \text{Biomass} - 0.06158 \cdot \text{Biomass}^2$ |
| 2:31:29PM | 0.0064 | (Gluconic Acid) = $0.04485 \cdot \text{Biomass} + 16.36 \cdot \text{Glucose} \cdot (\text{Dissolved Oxygen}) - 11.76 \cdot (\text{Dissolved Oxygen}) - 0.1472 \cdot \text{Glucose}^2 - 3.337 \cdot \text{Glucose}^2 \cdot (\text{Dissolved Oxygen}) - 1.737 \cdot \text{Glucose}^2 \cdot (\text{Dissolved Oxygen})^2$ |
| 2:31:30PM | 0.0105 | (Gluconic Acid) = $6.274 \cdot \text{Glucose} + 3.037 \cdot \text{Glucose} \cdot (\text{Dissolved Oxygen}) + 0.04708 \cdot \text{Biomass}^2 - 3.903 - 0.2643 \cdot \text{Biomass} - 2.443 \cdot (\text{Dissolved Oxygen}) - 2.148 \cdot \text{Glucose}^2$ |
| 2:31:30PM | 0.0104 | (Gluconic Acid) = $6.417 \cdot \text{Glucose} + 0.2176 \cdot \text{Biomass} + 3.431 \cdot \text{Glucose} \cdot (\text{Dissolved Oxygen}) - 4.443 - 3.022 \cdot (\text{Dissolved Oxygen}) - 0.1768 \cdot \text{Glucose} \cdot \text{Biomass} - 2.112 \cdot \text{Glucose}^2$ |
| 2:31:31PM | 0.0253 | (Gluconic Acid) = $1.765 \cdot \text{Glucose} \cdot (\text{Dissolved Oxygen}) + 0.9477 \cdot \text{Glucose} \cdot \text{Biomass} - 0.1207 - 0.5729 \cdot (\text{Dissolved Oxygen}) - 0.2781 \cdot \text{Biomass}^2 - 0.613 \cdot \text{Glucose}^2$ |
| 2:31:31PM | 0.0106 | (Gluconic Acid) = $6.194 \cdot \text{Glucose} + 3.15 \cdot \text{Glucose} \cdot (\text{Dissolved Oxygen}) - 4.108 - 2.606 \cdot (\text{Dissolved Oxygen}) - 0.02872 \cdot \text{Glucose} \cdot \text{Biomass} - 2.116 \cdot \text{Glucose}^2$ |
| 2:31:32PM | 0.0312 | (Gluconic Acid) = $0.157 + 3.142 \cdot \text{Glucose} \cdot (\text{Dissolved Oxygen}) + 0.053 \cdot \text{Glucose} \cdot \text{Biomass} - 2.577 \cdot (\text{Dissolved Oxygen}) - 0.06336 \cdot \text{Glucose}^2$ |
| 2:31:32PM | 0.0299 | (Gluconic Acid) = $6.18 \cdot \text{Glucose} + 2.017 \cdot (\text{Dissolved Oxygen}) - 4.867 - 0.0146 \cdot \text{Glucose} \cdot \text{Biomass} - 1.763 \cdot \text{Glucose}^2$ |
| 2:31:32PM | 0.0063 | (Gluconic Acid) = $0.1564 + 16.6 \cdot \text{Glucose} \cdot (\text{Dissolved Oxygen}) - 12.05 \cdot (\text{Dissolved Oxygen}) - 0.1765 \cdot \text{Glucose}^2 - 3.33 \cdot \text{Glucose}^2 \cdot (\text{Dissolved Oxygen}) - 1.76 \cdot \text{Glucose}^2 \cdot (\text{Dissolved Oxygen})^2$ |
| 2:31:33PM | 0.0165 | (Gluconic Acid) = $6.174 \cdot \text{Glucose} + 1.436 \cdot \text{Glucose} \cdot (\text{Dissolved Oxygen}) - 4.516 - 0.02891 \cdot \text{Glucose} \cdot \text{Biomass} - 1.92 \cdot \text{Glucose}^2$ |
| 2:31:33PM | 0.0108 | (Gluconic Acid) = $5.944 \cdot \text{Glucose} + 3.126 \cdot \text{Glucose} \cdot (\text{Dissolved Oxygen}) - 3.966 - 2.607 \cdot (\text{Dissolved Oxygen}) - 2.055 \cdot \text{Glucose}^2$ |
| 2:31:34PM | 0.0323 | (Gluconic Acid) = $0.2192 + 3.19 \cdot \text{Glucose} \cdot (\text{Dissolved Oxygen}) - 2.572 \cdot (\text{Dissolved Oxygen}) - 0.01722 \cdot \text{Glucose}^2$ |
| 2:31:34PM | 0.0300 | (Gluconic Acid) = $6.053 \cdot \text{Glucose} + 1.999 \cdot (\text{Dissolved Oxygen}) - 4.792 - 1.733 \cdot \text{Glucose}^2$ |
| 2:31:35PM | 0.0168 | (Gluconic Acid) = $5.922 \cdot \text{Glucose} + 1.412 \cdot \text{Glucose} \cdot (\text{Dissolved Oxygen}) - 4.373 - 1.859 \cdot \text{Glucose}^2$ |
| 2:31:35PM | 0.0318 | (Gluconic Acid) = $0.1378 \cdot \text{Glucose} + 2.806 \cdot \text{Glucose} \cdot (\text{Dissolved Oxygen}) - 2.045 \cdot (\text{Dissolved Oxygen})$ |
| 2:31:37PM | 0.0457 | (Gluconic Acid) = $0.06562 + 1.693 \cdot \text{Glucose} \cdot (\text{Dissolved Oxygen})$ |
| 2:31:37PM | 0.0052 | (Gluconic Acid) = $3.696 \cdot \text{Glucose} + 6.874 \cdot \text{Glucose} \cdot (\text{Dissolved Oxygen}) - 2.379 - 5.347 \cdot (\text{Dissolved Oxygen}) - 1.463 \cdot \text{Glucose}^2 - 1.657 \cdot \text{Glucose}^2 \cdot (\text{Dissolved Oxygen})^2$ |
| 2:31:39PM | 0.0323 | (Gluconic Acid) = $0.1851 + 3.086 \cdot \text{Glucose} \cdot (\text{Dissolved Oxygen}) - 2.426 \cdot (\text{Dissolved Oxygen})$ |
| 2:31:39PM | 0.0204 | (Gluconic Acid) = $3.321 \cdot \text{Glucose} \cdot (\text{Dissolved Oxygen}) - 0.1878 - 3.843 \cdot (\text{Dissolved Oxygen})^2$ |
| 2:31:40PM | 0.0108 | (Gluconic Acid) = $5.944 \cdot \text{Glucose} + 3.126 \cdot \text{Glucose} \cdot (\text{Dissolved Oxygen}) - 3.966 - 2.607 \cdot (\text{Dissolved Oxygen}) - 2.055 \cdot \text{Glucose}^2$ |

| | | |
|-----------|--------|--|
| 2:31:40PM | 0.0078 | $(\text{Gluconic Acid}) = 4.522 * \text{Glucose} + 3.163 * \text{Glucose} * (\text{Dissolved Oxygen}) - 3.299 - 1.584 * \text{Glucose}^2 - 3.529 * (\text{Dissolved Oxygen})^2$ |
| 2:31:41PM | 0.0077 | $(\text{Gluconic Acid}) = 4.678 * \text{Glucose} + 3.145 * \text{Glucose} * (\text{Dissolved Oxygen}) - 3.36 - 0.02666 * \text{Biomass} - 1.632 * \text{Glucose}^2 - 3.464 * (\text{Dissolved Oxygen})^2$ |
| 2:31:41PM | 0.0075 | $(\text{Gluconic Acid}) = 4.773 * \text{Glucose} + 3.408 * \text{Glucose} * (\text{Dissolved Oxygen}) - 3.356 - 0.8382 * (\text{Dissolved Oxygen}) - 1.695 * \text{Glucose}^2 - 2.913 * (\text{Dissolved Oxygen})^2$ |
| 2:31:45PM | 0.0178 | $(\text{Gluconic Acid}) = 3.032 * \text{Glucose} * (\text{Dissolved Oxygen}) - 0.2346 - 5.103 * (\text{Dissolved Oxygen})^3$ |
| 2:31:46PM | 0.0124 | $(\text{Gluconic Acid}) = 6.675 * \text{Glucose} + 0.9115 * \text{Glucose}^2 * (\text{Dissolved Oxygen}) - 4.595 - 2.257 * \text{Glucose}^2$ |
| 2:31:46PM | 0.0075 | $(\text{Gluconic Acid}) = 2.304 * \text{Glucose} + 3.121 * \text{Glucose} * (\text{Dissolved Oxygen}) - 2.293 - 3.468 * (\text{Dissolved Oxygen})^2 - 0.3643 * \text{Glucose}^3$ |
| 2:31:46PM | 0.0045 | $(\text{Gluconic Acid}) = 2.82 * \text{Glucose} + 7.91 * \text{Glucose} * (\text{Dissolved Oxygen}) - 1.677 - 7.058 * (\text{Dissolved Oxygen}) - 1.189 * \text{Glucose}^2 - 0.9981 * \text{Glucose}^3 * (\text{Dissolved Oxygen})^2$ |
| 2:31:47PM | 0.0067 | $(\text{Gluconic Acid}) = 16.27 * \text{Glucose} * (\text{Dissolved Oxygen}) - 0.09693 - 10.22 * (\text{Dissolved Oxygen}) - 3.354 * (\text{Dissolved Oxygen})^3 - 4.551 * \text{Glucose}^2 * (\text{Dissolved Oxygen})$ |
| 2:31:47PM | 0.0066 | $(\text{Gluconic Acid}) = 4.544 * \text{Glucose} + 5.031 * \text{Glucose} * (\text{Dissolved Oxygen}) - 3.084 - 2.967 * (\text{Dissolved Oxygen}) - 1.684 * \text{Glucose}^2 - 2.212 * \text{Glucose} * (\text{Dissolved Oxygen})^2$ |
| 2:31:48PM | 0.0063 | $(\text{Gluconic Acid}) = 4.651 * \text{Glucose} + 3.399 * \text{Glucose} * (\text{Dissolved Oxygen}) - 3.223 - 1.566 * (\text{Dissolved Oxygen}) - 1.653 * \text{Glucose}^2 - 3.07 * (\text{Dissolved Oxygen})^3$ |
| 2:31:50PM | 0.0045 | $(\text{Gluconic Acid}) = 2.681 * \text{Glucose} + 4.757 * (\text{Dissolved Oxygen})^2 + 4.242 * \text{Glucose} * (\text{Dissolved Oxygen})^2 - 2.446 - 0.364 * \text{Glucose}^3 - 13.32 * (\text{Dissolved Oxygen})^3$ |
| 2:31:50PM | 0.0042 | $(\text{Gluconic Acid}) = 0.413 + 18.59 * (\text{Dissolved Oxygen})^2 + 16.11 * \text{Glucose} * (\text{Dissolved Oxygen}) - 15.96 * (\text{Dissolved Oxygen}) - 4.518 * \text{Glucose}^2 * (\text{Dissolved Oxygen}) - 20.4 * (\text{Dissolved Oxygen})^3$ |
| 2:31:50PM | 0.0034 | $(\text{Gluconic Acid}) = 3.085 * \text{Glucose} + 6.682 * \text{Glucose} * (\text{Dissolved Oxygen}) - 1.827 - 6.294 * (\text{Dissolved Oxygen}) - 1.258 * \text{Glucose}^2 - 0.6046 * \text{Glucose}^4 * (\text{Dissolved Oxygen})^3$ |
| 2:31:50PM | 0.0046 | $(\text{Gluconic Acid}) = 2.812 * \text{Glucose} + 7.791 * \text{Glucose} * (\text{Dissolved Oxygen}) - 1.695 - \text{Glucose}^3 * (\text{Dissolved Oxygen})^2 - 6.833 * (\text{Dissolved Oxygen}) - 1.182 * \text{Glucose}^2$ |
| 2:31:51PM | 0.0041 | $(\text{Gluconic Acid}) = 4.027 * \text{Glucose} + 20.24 * (\text{Dissolved Oxygen})^2 + 14.85 * \text{Glucose} * (\text{Dissolved Oxygen}) - 1.869 - 17.48 * (\text{Dissolved Oxygen}) - 1.889 * \text{Glucose}^2 - 15.88 * \text{Glucose} * (\text{Dissolved Oxygen})^2$ |
| 2:31:51PM | 0.0046 | $(\text{Gluconic Acid}) = 2.801 * \text{Glucose} + 7.772 * \text{Glucose} * (\text{Dissolved Oxygen}) - 1.712 - \text{Glucose}^3 * (\text{Dissolved Oxygen})^2 - 6.805 * (\text{Dissolved Oxygen}) - 1.169 * \text{Glucose}^2$ |
| 2:31:52PM | 0.0018 | $(\text{Gluconic Acid}) = 5.22 * \text{Glucose} + 26.04 * (\text{Dissolved Oxygen})^2 + 2.966 * \text{Glucose} * (\text{Dissolved Oxygen}) - 2.988 - 9.41 * (\text{Dissolved Oxygen}) - 1.792 * \text{Glucose}^2 - 26.56 * (\text{Dissolved Oxygen})^3$ |
| 2:31:52PM | 0.0016 | $(\text{Gluconic Acid}) = 2.682 * \text{Glucose} + 25.93 * (\text{Dissolved Oxygen})^2 + 2.894 * \text{Glucose} * (\text{Dissolved Oxygen}) - 1.838 - 9.28 * (\text{Dissolved Oxygen}) - 0.4067 * \text{Glucose}^3 - 26.47 * (\text{Dissolved Oxygen})^3$ |
| 2:31:52PM | 0.0341 | $(\text{Gluconic Acid}) = \text{Glucose} * \sin(2.345 * (\text{Dissolved Oxygen}))$ |
| 2:31:52PM | 0.0046 | $(\text{Gluconic Acid}) = 2.814 * \text{Glucose} + 7.793 * \text{Glucose} * (\text{Dissolved Oxygen}) - 1.71 - \text{Glucose}^3 * (\text{Dissolved Oxygen})^2 - 6.838 * (\text{Dissolved Oxygen}) - 1.178 * \text{Glucose}^2$ |
| 2:31:53PM | 0.0014 | $(\text{Gluconic Acid}) = 25.81 * (\text{Dissolved Oxygen})^2 + 2.814 * \text{Glucose} * (\text{Dissolved Oxygen}) + 1.886 * \text{Glucose}^2 - 0.6191 - 9.13 * (\text{Dissolved Oxygen}) - 0.8328 * \text{Glucose}^3 - 26.38 * (\text{Dissolved Oxygen})^3$ |
| 2:31:53PM | 0.0046 | $(\text{Gluconic Acid}) = 2.806 * \text{Glucose} + 7.764 * \text{Glucose} * (\text{Dissolved Oxygen}) - 1.719 - \text{Glucose}^3 * (\text{Dissolved Oxygen})^2 - 6.816 * (\text{Dissolved Oxygen}) - 1.165 * \text{Glucose}^2$ |
| 2:31:53PM | 0.0012 | $(\text{Gluconic Acid}) = 2.231 + 25.34 * (\text{Dissolved Oxygen})^2 + 6.231 * \text{Glucose}^2 + 2.646 * \text{Glucose} * (\text{Dissolved Oxygen}) - 6.228 * \text{Glucose} - 8.727 * (\text{Dissolved Oxygen}) -$ |

$$1.807 * \text{Glucose}^{26.05} * (\text{Dissolved Oxygen})^3$$

| | | |
|-----------|---------|---|
| 2:31:53PM | 0.0264 | (Gluconic Acid) = $\text{Glucose} * \sin(\text{Glucose}^2 * (\text{Dissolved Oxygen}))$ |
| 2:31:53PM | 0.0255 | (Gluconic Acid) = $1.026 * \text{Glucose} * \sin(\text{Glucose}^2 * (\text{Dissolved Oxygen}))$ |
| 2:31:54PM | 0.0041 | (Gluconic Acid) = $5.236 * \text{Glucose} + 4.764 * \text{Glucose} * (\text{Dissolved Oxygen})^2 - 3.575 - 1.719 * \text{Glucose}^2 - 11.65 * (\text{Dissolved Oxygen})^4$ |
| 2:31:54PM | 0.0038 | (Gluconic Acid) = $2.807 * \text{Glucose} + 4.716 * \text{Glucose} * (\text{Dissolved Oxygen})^2 - 2.465 - 0.3929 * \text{Glucose}^3 - 11.52 * (\text{Dissolved Oxygen})^4$ |
| 2:31:55PM | 0.0037 | (Gluconic Acid) = $2.009 * \text{Glucose} + 4.667 * \text{Glucose} * (\text{Dissolved Oxygen})^2 - 1.921 - 0.1317 * \text{Glucose}^4 - 11.4 * (\text{Dissolved Oxygen})^4$ |
| 2:31:55PM | 0.0036 | (Gluconic Acid) = $1.976 * \text{Glucose}^2 + 4.662 * \text{Glucose} * (\text{Dissolved Oxygen})^2 - 1.176 - 0.8422 * \text{Glucose}^3 - 11.38 * (\text{Dissolved Oxygen})^4$ |
| 2:32:09PM | 0.0341 | (Gluconic Acid) = $\text{Glucose} * \sin(2.334 * (\text{Dissolved Oxygen}))$ |
| 2:32:11PM | 0.0341 | (Gluconic Acid) = $\text{Glucose} * \sin(2.339 * (\text{Dissolved Oxygen}))$ |
| 2:32:15PM | 0.0457 | (Gluconic Acid) = $\text{Glucose} + 2.079 * (\text{Dissolved Oxygen}) - 1.213$ |
| 2:32:21PM | 0.0415 | (Gluconic Acid) = $0.189 + \text{Glucose}^2 * (\text{Dissolved Oxygen})$ |
| 2:32:23PM | 0.0174 | (Gluconic Acid) = $0.0008281 + 4.783 * \sin(\text{Glucose} * (\text{Dissolved Oxygen})) - 3.799 * (\text{Dissolved Oxygen})$ |
| 2:32:48PM | 0.0118 | (Gluconic Acid) = $0.1335 * \text{Glucose} + 10.23 * \text{Glucose} * (\text{Dissolved Oxygen}) * \sin(\text{Glucose}) - 8.58 * \text{Glucose} * (\text{Dissolved Oxygen})$ |
| 2:32:55PM | 0.0408 | (Gluconic Acid) = $0.1768 + \text{Glucose}^2 * (\text{Dissolved Oxygen})$ |
| 2:32:55PM | 0.0130 | (Gluconic Acid) = $0.16 + 10.72 * \text{Glucose} * (\text{Dissolved Oxygen}) * \sin(\text{Glucose}) - 9.005 * \text{Glucose} * (\text{Dissolved Oxygen})$ |
| 2:33:27PM | STOPPED | |