Automation and Machine Learning: How a Robot Can Conduct Its Own Experiments Using Gaussian Processes

Miller Gruen and Adi Timin - Hauber Research Fellowship 2024
Advisor: Dr Mary Lowe, Physics Department, Loyola University Maryland

Introduction

Scientists often need to take large amounts of data and conduct numerous tests. Rather than taking many measurements that require time, resources, and money, it may be better to automate the task. Beyond automation, imagine if a device were able to do more than simply repeat a task, if it could make its own decisions and conduct an experiment wholly without needing constant human supervision. This could combat the complexity of research and optimize the taking of data, as well as greatly easing the burden on scientists in fields that require large amounts of testing, such as chemistry, physics, drug discovery, biology, and material synthesis. We can accomplish this using Gaussian processes applied to machine learning. Gaussian processes is a method of doing regression that enables you to find a quantitative model to describe the relationships in the data points. In our experiment, using a liquid handling robot, we applied Gaussian processes to determine the model and the sequence of data points to acquire, allowing the device to operate autonomously. The goal of this project was to understand both exploration and exploitation-based machine learning. We explored machine learning both experimentally and computationally. Our experiment could be viewed as proof of concept of a method that can be applied to many other disciplines, settings, and applications.

Theory and Background

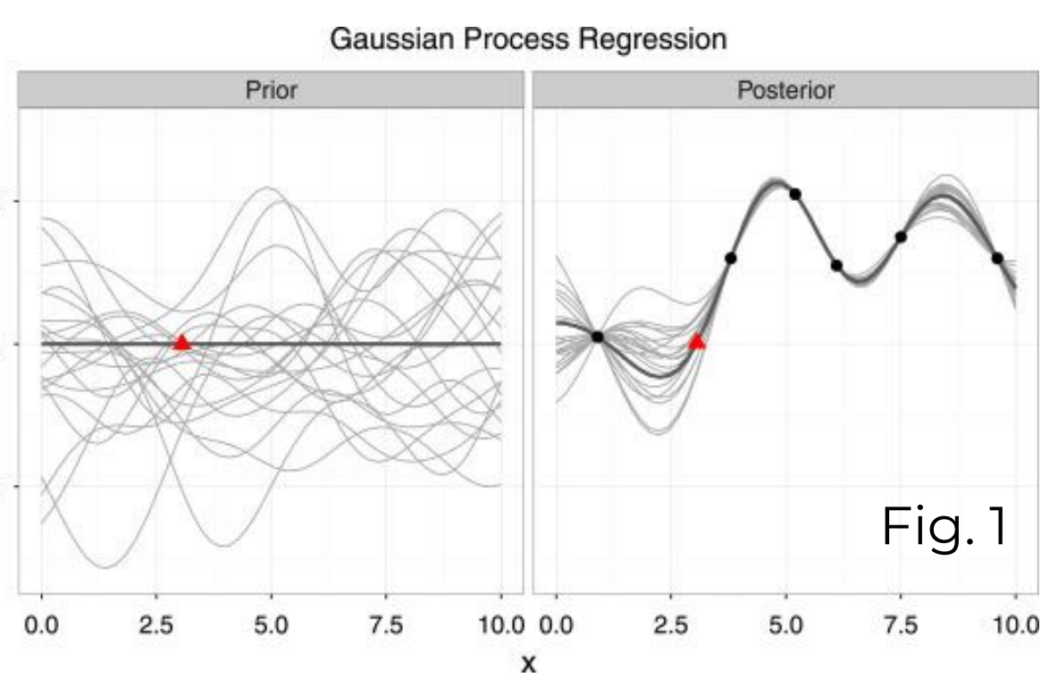

Fig. 1

Gaussian processes is a method of regression analysis used to develop a model that indicates which functions are most likely to describe the data obtained at any time during an experiment. When there are no data, the prior model includes all functions. Then as more data are observed, the number of possible functions decreases. The functions are non-parametric, which means there is no equation that describes the function being fitted; there is only a curve described by a set of points.

In essence, a Gaussian process is a collection of normally distributed random variables. Thus, for each value of the control parameter, there is a Gaussian probability distribution (also known as a Bell curve) as an output (Fig. 2). The variance in the distributions represents the uncertainty in the function. Where we have a better idea of the underlying function, the distributions are narrower. This variance includes both the uncertainty in the model due to the lack of data (epistemic uncertainty), and the noise in the data itself (aleatoric uncertainty). The uncertainty is plotted on Fig. 3 as the light-blue region (2 standard deviations). The darker-blue line on this graph is the surrogate function, the mean of the possible functions. A surrogate function is the curve that most likely represents the true curve. The underlying curve that we desire to find is shown by the black dotted line.

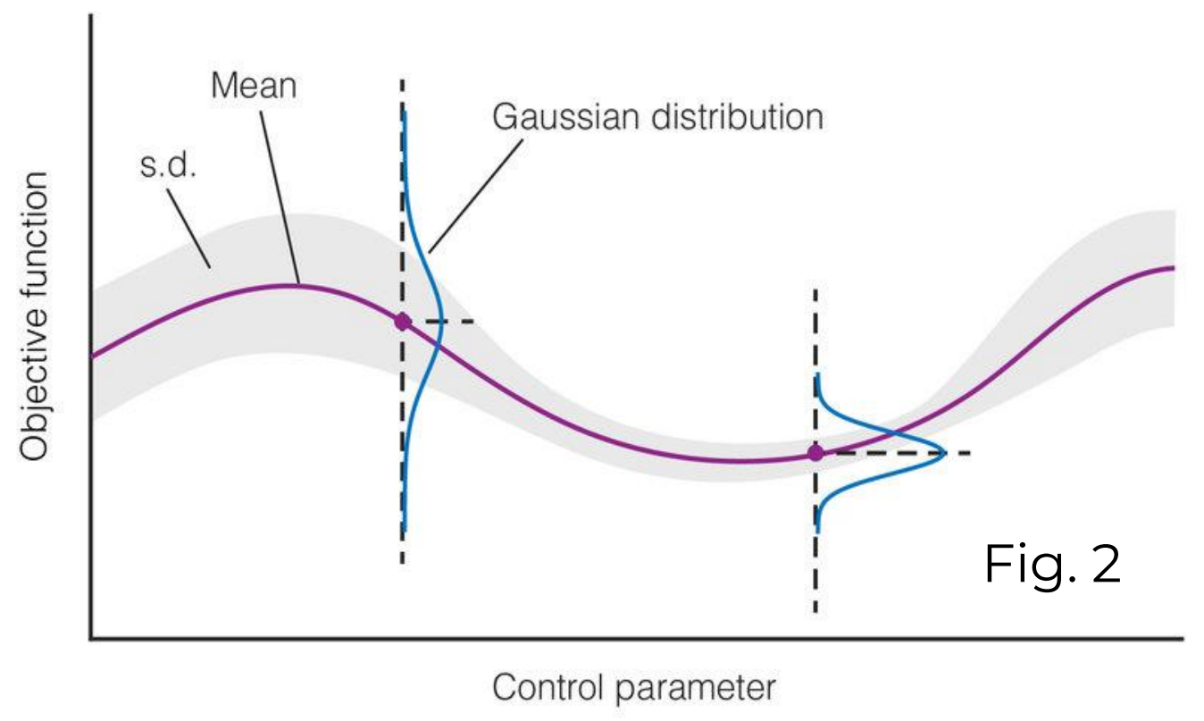

Fig. 2

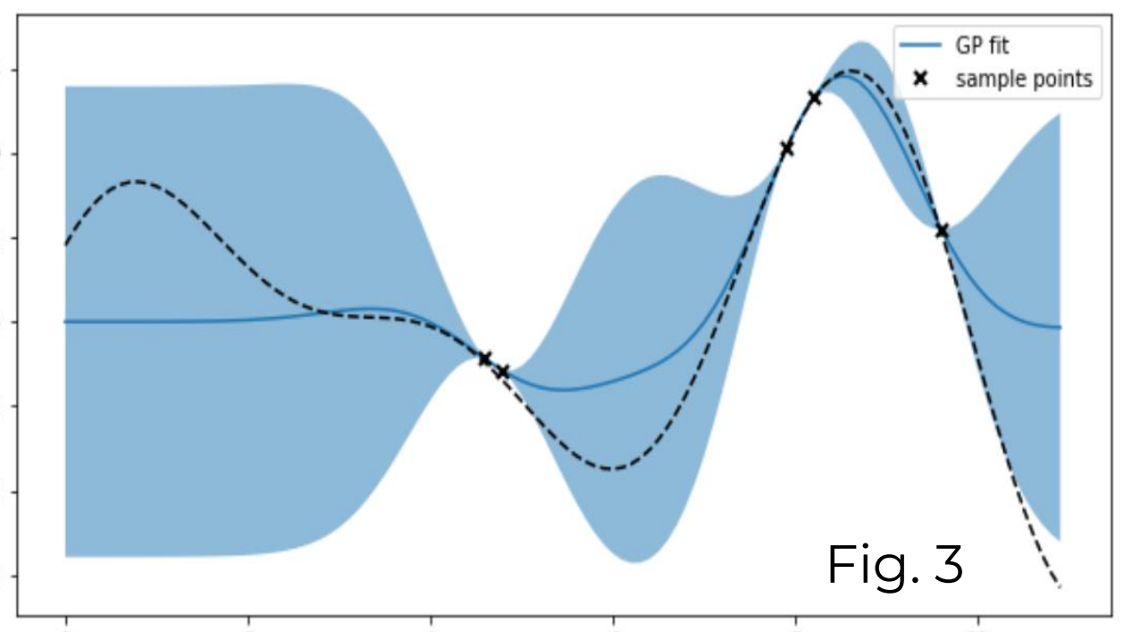

Fig. 3

$$\kappa_{\text{rbf}}(x_p, x_q) = \sigma_f^2 \exp\left(-\frac{(x_p - x_q)^2}{2\ell^2}\right) + \sigma_n^2 \delta_{pq}$$ Eq. 1

The way to find the surrogate function is by using a "kernel", or covariance function (Eq. 1). For any two given inputs (x_p, x_q), the kernel determines how similar the outputs are, and thus informs the shape of the surrogate function. For example, if the outputs are more similar, the function will be smoother. In the Gaussian process, the "hyperparameters" $(\sigma_f{}^2, l, \sigma_n{}^2)$ of the kernel are fitted to the data. The kernel is also used to construct a covariance matrix, which is used to calculate the mean and the variance of the function at each point (Fig. 3).[1]

This information is then used to determine the best data point to take next using an "acquisition function" that the scientist has specified. For instance, if the acquisition function is equal to the variance at each point (Fig. 4), a measurement will be taken where the uncertainty is highest, as shown by the vertical magenta line.

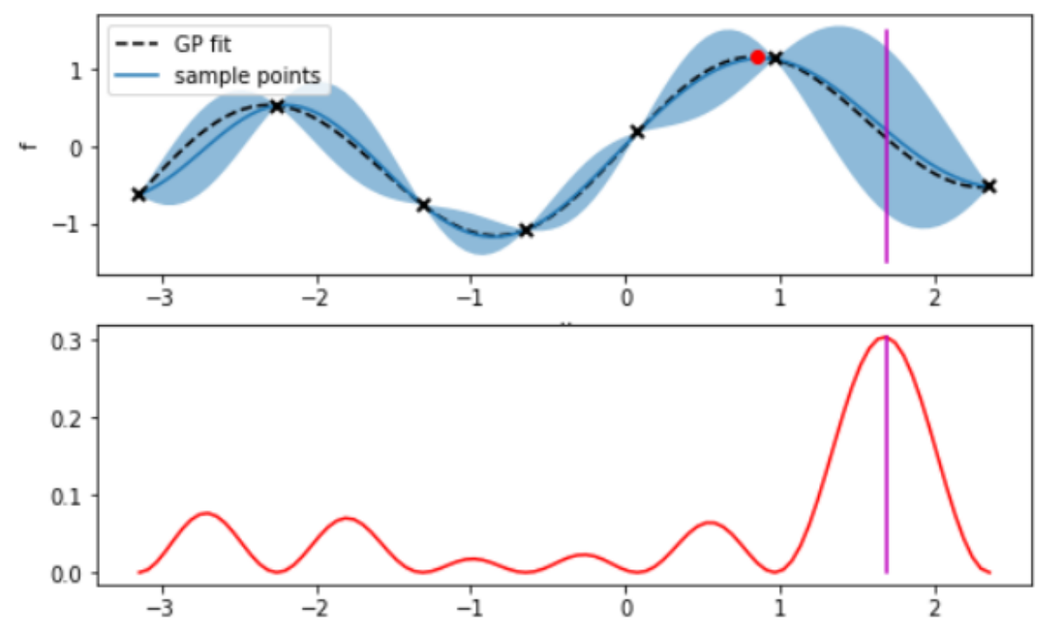

Fig. 4

Apparatus

LEGOLAS, which stands for LEGO-based low-cost autonomous scientist, is a robotic kit created by a team of students and faculty at the University of Maryland and the National Institute of Standards and Technology.[2] Often devices for material synthesis are expensive, and machine learning is being studied to reduce the number of costly experiments and make data collection more efficient. LEGOLAS was created to be a low-cost accessible apparatus enabling machine learning to be taught in any setting.

Legolas possesses:
- 2 Raspberry Pis and Build Hats to control motors
- 5 Lego motors
- 1 Arduino
- 1 pH sensor
- 1 pipette with a plunger

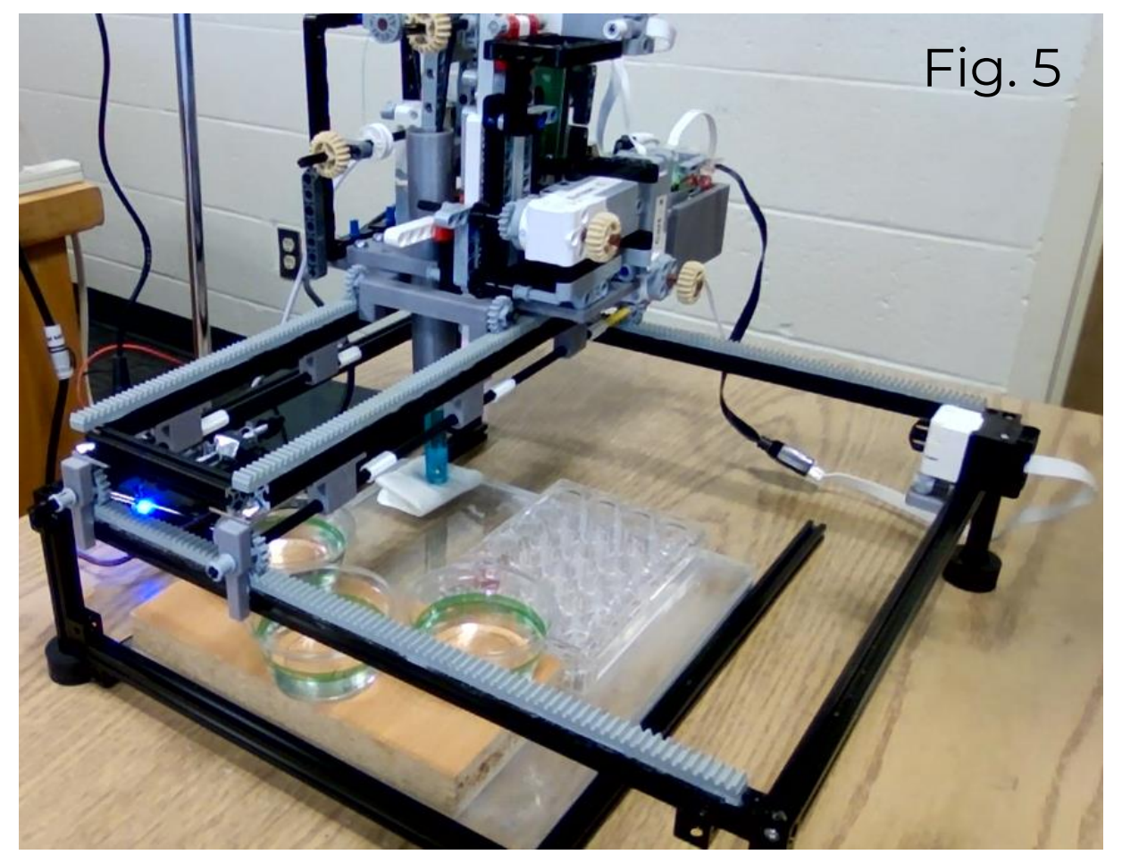

Fig. 5

The Raspberry Pis connect wirelessly via a local network (Fig. 6) to a computer running the code and run the Lego motors using a Build Hat.

Fig. 6

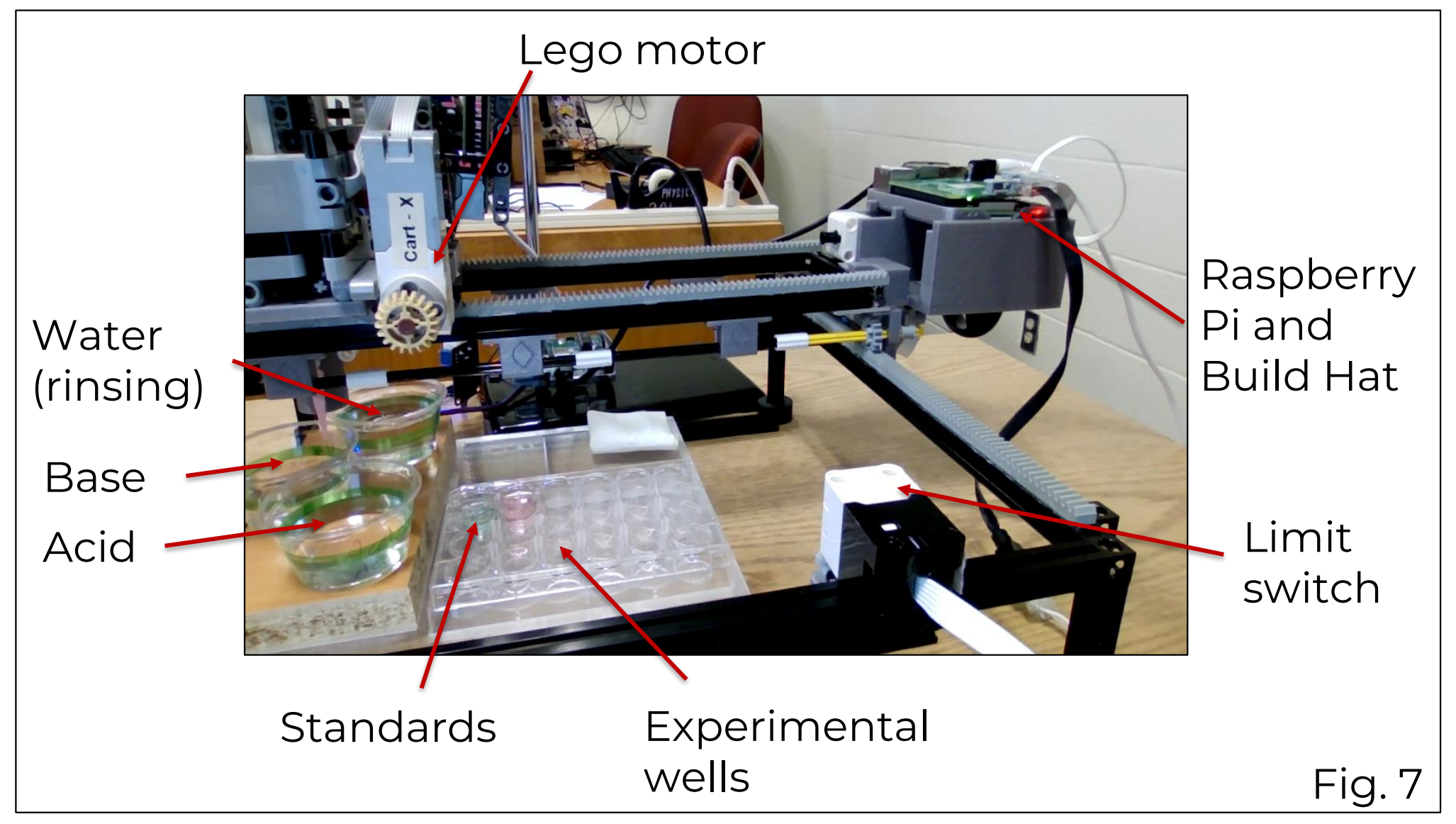

Fig. 7

The code is written in Python and runs completely on a separate computer where the machine learning and Gaussian processes calculations are completed. The initial code was written by UMD personnel and edited by us. We used a Jupyter Notebook to test the robot and moved the code to a Python script for data collection.
The code is able to:
- Establish communication between the computer and Raspberry Pis
- Apply machine learning and regression
- There are software functions for:
  - Moving the motors in the x and y directions
  - Pipetting liquid
  - Moving the pH sensor vertically
  - pH measurement
  - Rinsing and drying the pH sensor

Experimental Results

Our goal was to discover a quantitative model for the shape of the pH curve that results when mixing an acid and a conjugate base. We chose to test the relationship between pH and the ratio r of acid to conjugate base because the relationship is governed by a known equation: the Henderson-Hasselbalch equation (Eq. 2).

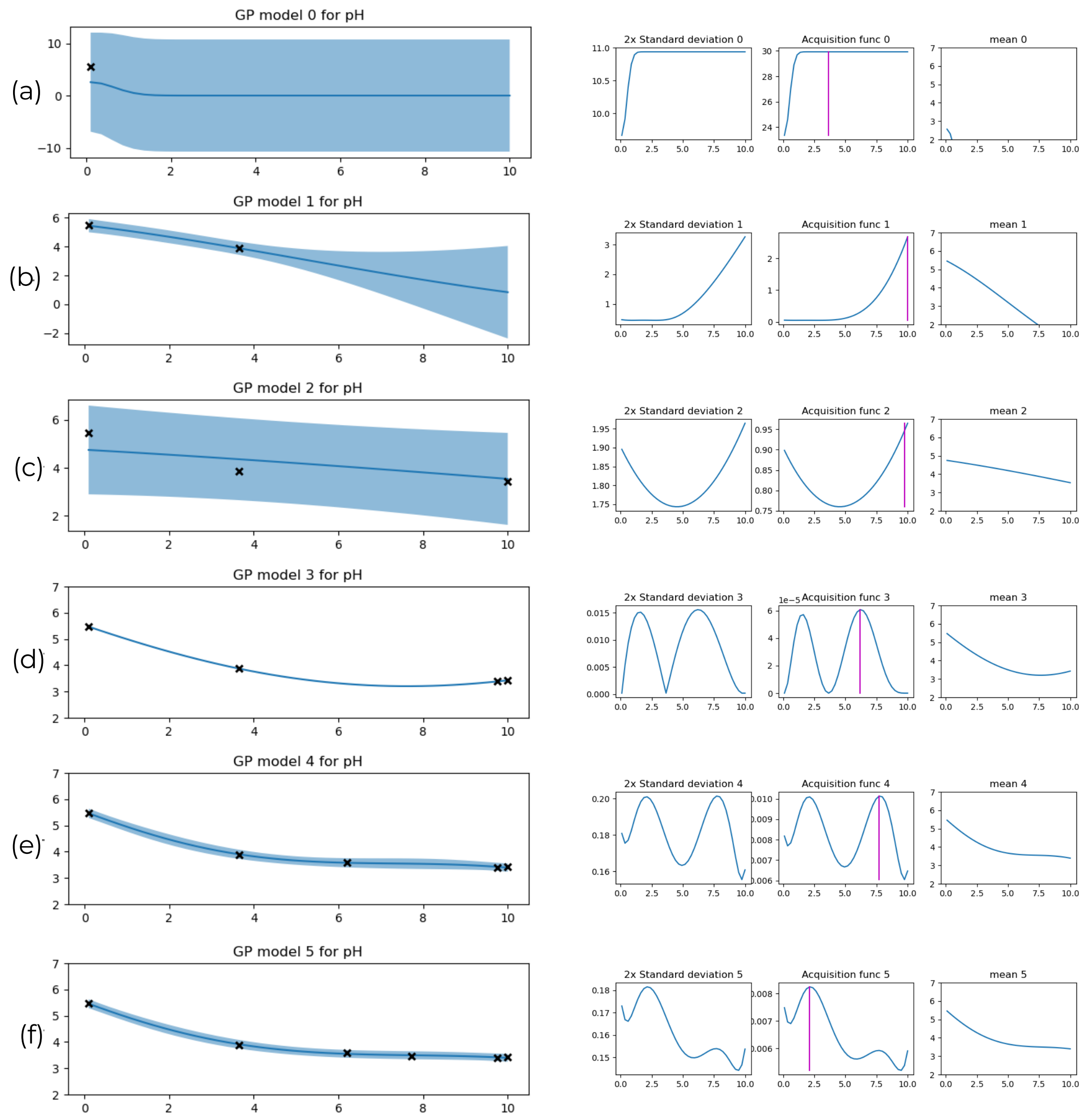

Fig. 8: Code Outputs;
vertical axis – *pH*, horizontal axis - *ratio*

$$pH = pKa - \log\frac{Concentration\ of\ acid}{Concentration\ of\ conjugate\ base}$$ Eq. 2

We selected a known relationship in order to understand the concepts of machine learning and to ensure its correct application

We used **0.1M Acetic Acid** and **0.1M Sodium Acetate**.
When there are no data, using Gaussian processes, the mean function begins as a line at 0 with variance 1. We began data collection by taking a measurement at *r* = 0.1 and obtained a new curve (Fig. 8a). The acquisition function (Fig. 8 right column) was defined to be the variance. The vertical magenta line indicates the point where the next measurement was taken. After two data points were acquired, (Fig. 8b) the variance in the vicinity of the data points is reduced significantly. Data collection continued until the mean and variance became stable (Fig. 8e,f).

Fig. 9 shows the comparison between the Henderson-Hasselbalch equation and the data points acquired in our experiments. While the shape of our data is similar to the Henderson-Hasselbalch results, they consistently have lower pH values. This is likely due to the formation of carbonic acid from CO_2 in the atmosphere.[3]

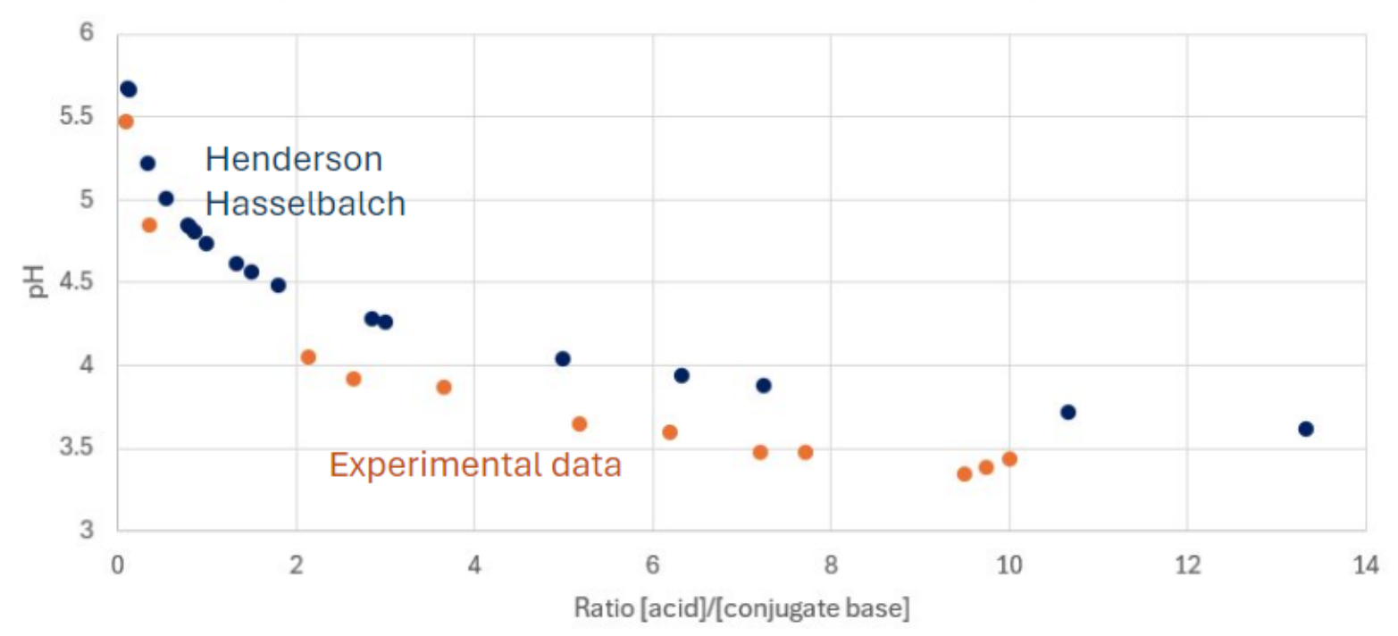

Fig. 9

Computational & Experimental Results with a New Acquisition Function (α)

Computational trials allow us to generate data artificially and to test Gaussian Processes more thoroughly and quickly than experimental work is able to do. Computation allows us to vary aspects such as the kernel, the acquisition function (α), and the hyperparameters. An example is shown in Figure 10, in which the lengthscale parameter (l) was fixed to a value of 0.5. The variance oscillates greatly as the short lengthscale value implies that the pH values of neighboring points are not necessarily similar. When running the test trials, the data were generated from the Henderson-Hasselbalch equation. Noise was added to the data to imitate the behavior of the physical system. In most trials, the standard deviation was set to 0.05 by using the following code:

```
pH_mu = 4.74 - np.log10(ratio)
pH = np.random.normal(loc=pH_mu, scale=0.05, size=None)
```

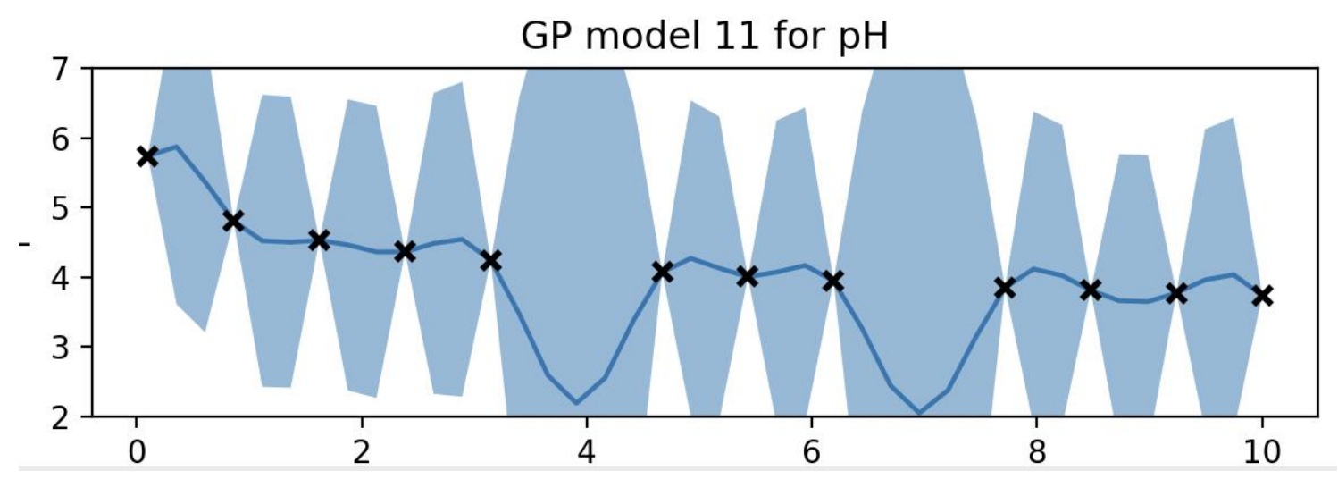

Fig. 10: lengthscale = 0.5 (computation)

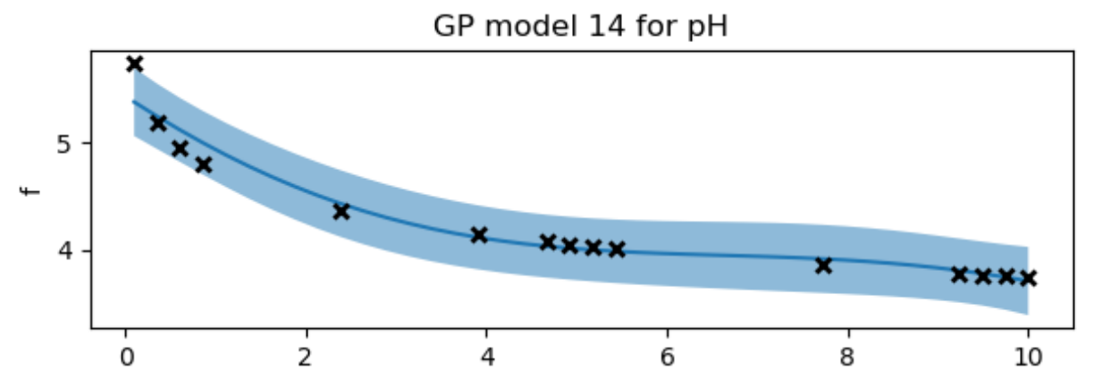

Fig. 11: α = variance (computation)

Computational tests indicated the leftmost data points are not accurately described by the surrogate function when the acquisition function is given by Eq. 3 (Fig. 11). We developed a new acquisition function by incorporating the slope of the curve into α and taking more measurements where the slope is greater. Our method was to use two different forms of α: Eq. 3 and Eq. 4, where n is the iteration number and σ^2 is the variance. Measurements were initially taken using Eq. 3, until the variance reached a sufficiently low value, at which point α became Eq. 4. This method would be useful in many instances when a function has a region where it changes drastically and requires more points to capture those changes.

Eq. 3 $\alpha = \sigma^2$

Eq. 4 $\alpha = \sqrt{n} * \sigma^2 + |\text{slope}|$

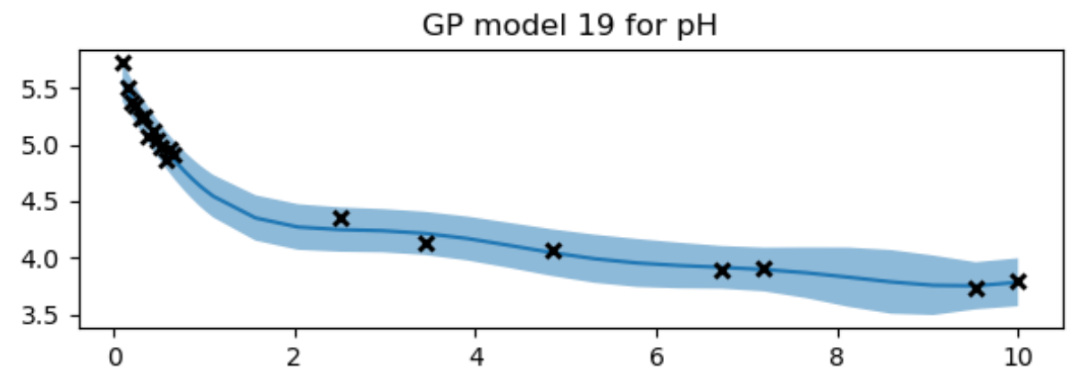

Fig. 12: new α (computation)

The computational results yielded a better fit to the leftmost points (Fig. 12). We then implemented this method experimentally with LEGOLAS with a maximum of 20 data points (corresponding to the 20 available wells). The results are shown in Fig. 13. The steep part of the curve is not fitted as well as the computational results, likely due to the limited number of measurements and the noise in the data. However, the mapping of that part of the graph is significantly more complete. Nonetheless, the computational result show that there is promise in this method.

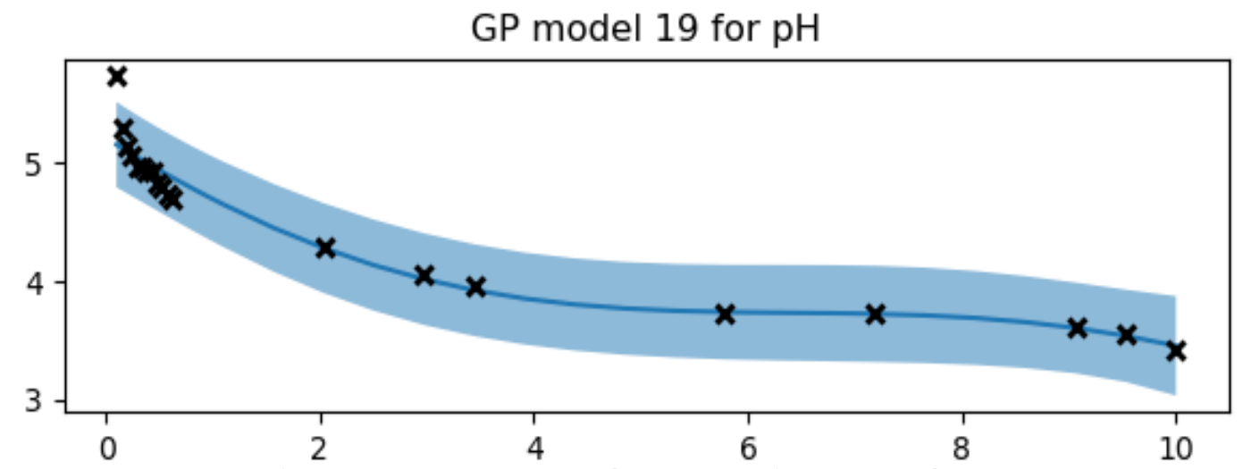

Fig. 13: new α (experiment)

Acknowledgements

We wish to acknowledge Dr. Mary Lowe, the Loyola University physics department, the Hauber Research Fellowship Program, George Hall, Gilad Kusne, and the UMD and NIST team working on Legolas.

References

1. Görtler, J., Kehlbeck, R., & Deussen, O. (2021, December 17). A visual exploration of Gaussian processes. Distill. https://distill.pub/2019/visual-exploration-gaussian-processes/#MargCond
2. Saar, L., Liang, H., Wang, A., McDannald, A., Rodriguez, E., Takeuchi, I., & Kusne, A. G. (2022). The LEGOLAS Kit: A low-cost robot science kit for education with symbolic regression for hypothesis discovery and validation. MRS Bulletin, 47(9), 881–885. https://doi.org/10.1557/s43577-022-00430-2
3. Schmitz, G. (2002). pH of sodium acetate solutions. Journal of Chemical Education, 79(1), 29. https://doi.org/10.1021/ed079p29.1