Towards uncertainty quantification in high-throughput biology

Thomas Nok Hin Cheng[1,2*†], **Davy Deng**[1,2*†], **Martin Stražar**[3]

[1]Klarman Cell Observatory, Broad Institute, Cambridge, MA, USA; [2]Institute for Medical Engineering and Science (IMES), MIT, Cambridge, MA, USA; [3]Program of Infectious diseases and the Microbiome, Broad Institute, Cambridge, MA, USA

***For correspondence:**
nhcheng@mit.edu (TNHC);
davy@broadinstitute.org (DD)

[†]These authors contributed equally to this work

Abstract

Advances in biological technologies have facilitated the generation of large-scale datasets, profiling biological systems at an unprecedented scale and resolution. The high dimensionality and complexity of these datasets require machine learning (ML) approaches to extract meaningful and actionable biological insights. However, current ML models often lack interpretability for human practitioners, and their predictions typically do not include a measure of uncertainty. In this study, we apply Gaussian Process (GP) to two high-throughput biological datasets, single cell RNA sequencing (scRNA-seq) and mass spectrometry-based metabolomics (MS) for classification tasks. We further utilize the uncertainty estimates to propose new experiments aimed at improving the model's confidence. Collectively, our results indicate that GP is widely applicable in various stages of biological data analysis.

Introduction

High-throughput biology, with its capacity to generate large-scale datasets of biological systems, is a powerful tool that aids in understanding complex human diseases. However, this avalanche of data introduces new challenges in data analysis. To extract biological signals from these datasets, numerous groups have devised machine learning (ML) approaches for data analysis and low-dimensional representation (McInnes et al.)(family=Maaten and Hinton)(Maćkiewicz and Ratajczak). Large-scale datasets inherently exhibit variability due to biological, technical, and stochastic factors, making it critical to quantify the uncertainty in the data analysis process.

Uncertainty quantification assists in assessing whether a model prediction is under- or over-confident and provides a measure of the reliability of model predictions. This is particularly crucial in high-stakes applications like medical diagnosis, where prediction reliability is vital (Unc). In experimental biology, uncertainty quantification can help guide the design of new experiments. For instance, single-cell RNA sequencing (scRNA-seq), a powerful tool for profiling the transcriptome of individual cells, is prone to technical noise (Sin). While this noise can be reduced by increasing the sequencing depth, doing so is often infeasible due to cost and time constraints. In such cases, uncertainty quantification can help guide the design of new experiments by suggesting which cell types to prioritize next to maximize information gain.

In this study, we aim to quantify uncertainty in high-throughput biological datasets. We choose to focus on two primary data modalities: single-cell RNA sequencing (scRNA-seq) and mass spectrometry-based metabolomics (MS). Both data modalities are commonly used in biological research and are subject to technical noise and batch effects. We demonstrate that Gaussian Processes (GP) can be employed to quantify both uncertainties from data noise (aleatoric uncertainty) and from model uncertainty (epistemic uncertainty). We then utilize the uncertainty estimates to suggest new ex-

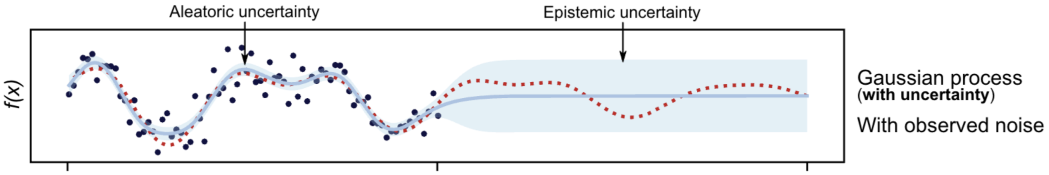

Figure 1. Two types of uncertainty in machine learning. Aleatoric uncertainty is due to noise in the data, while epistemic uncertainty is due to uncertainty in the model.

periments that enhance the model's confidence. Collectively, our results suggest that GP can be widely applicable in various types of biological data.

This manuscript is organized as follows: We first introduce the concept of uncertainty quantification and the Gaussian Process. We then apply the Gaussian Process to a real-world dataset and compare its performance with other machine learning methods. Finally, we discuss the advantages and disadvantages of using the Gaussian Process in omics data analysis.

Results

Gaussian Processes

Gaussian processes (GPs) are powerful and flexible tools in machine learning and statistics for modeling complex relationships between variables and predicting uncertain outcomes. They have been employed in a wide range of applications, including computer vision, robotics, bioinformatics, and finance (Kalaitzis and Lawrence)(Simek et al.). Fundamentally, GPs are a collection of random variables, any finite number of which have a joint Gaussian distribution. They can be used to model functions that map inputs to outputs, such as the relationship between the features of a dataset and their associated labels. Unlike traditional regression techniques that assume a fixed functional form for the relationship between input and output variables, GPs allow for a more flexible modeling approach capable of capturing complex nonlinear relationships.

One of the key strengths of GPs is their ability to provide uncertainty estimates alongside their predictions. This is particularly useful in situations where data is noisy or there is a high level of uncertainty in the underlying model. Figure 2 shows an example of a GP regression model with uncertainty estimates. The shaded region represents the 95% confidence interval, which is wider in regions with fewer data points and narrower in regions with more data points. The uncertainty estimates can be used to inform decision-making and improve prediction reliability.

Uncertainty quantification is crucial in numerous applications, such as medical diagnosis. Here, prediction accuracy is of utmost importance and must be accompanied by a measure of confidence in the results. To estimate uncertainty, GPs model the output variable as a Gaussian distribution with a mean and variance. The variance represents the uncertainty in the prediction, with larger variances indicating greater uncertainty. GPs can also be used to perform Bayesian inference, allowing for the incorporation of prior knowledge and the updating of beliefs as new data becomes available. Bayesian inference proves especially useful in situations with limited data or when the model is complex and challenging to estimate using traditional techniques.

GPs offer several advantages over other machine learning techniques, such as neural networks and support vector machines (Işık and Alptekin). They are non-parametric and do not assume a fixed functional form for the relationship between input and output variables, thereby allowing for greater flexibility and the modeling of complex relationships. GPs can also manage missing data and noisy measurements, which are common in biological sciences. Additionally, they provide a measure of uncertainty that can inform decision-making and enhance prediction reliability.

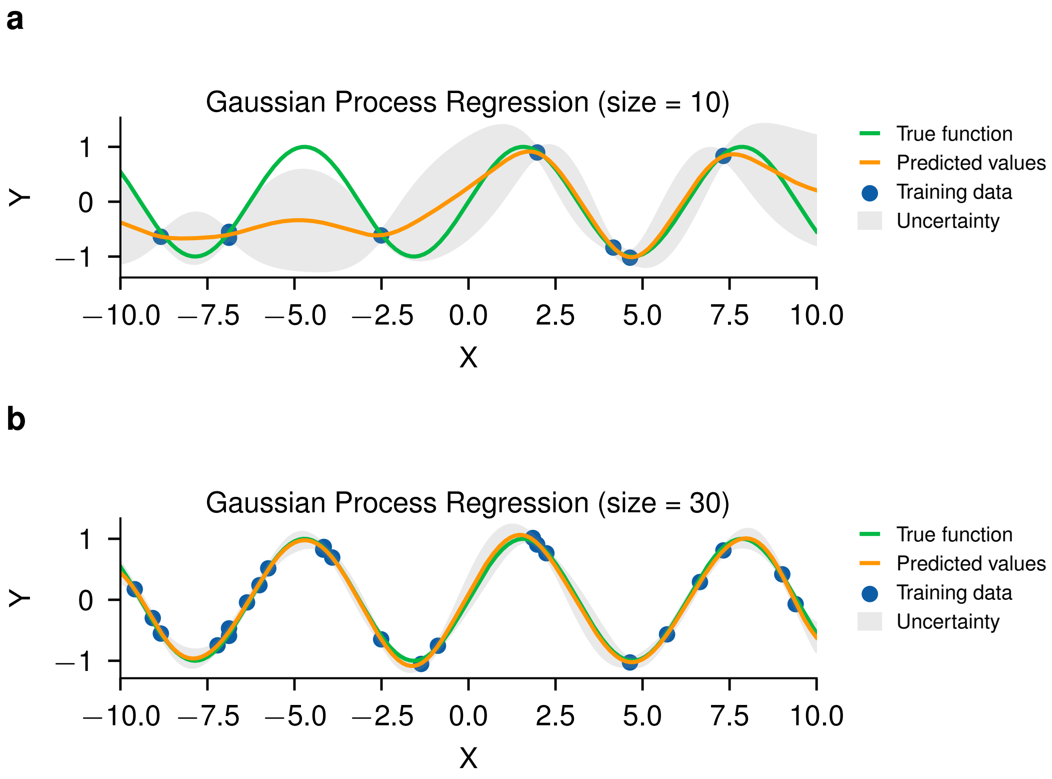

Figure 2. Gaussian Processes Regression on different number of test points.

Active learning

Active learning is a machine learning technique in which the algorithm determines which data to label for training, instead of passively receiving labeled data. This approach can be beneficial in scenarios where obtaining labeled data is expensive or time-consuming (Hemmer et al.). The goal is to minimize the amount of labeled data needed to train a model by intelligently choosing which data points to label next.

Active learning algorithms operate by iteratively selecting the most informative samples to label, based on a predefined criterion. Several criteria can be utilized for this purpose, including uncertainty sampling, query by committee, and expected model change. Uncertainty sampling involves choosing the data points for which the model is most uncertain about the correct label. Query by committee involves selecting the data points that are most controversial among a group of models. Expected model change involves selecting data points expected to have the greatest impact on the model's performance.

One method of implementing active learning is through Gaussian processes (Riis et al.). Gaussian processes are a flexible and potent probabilistic modeling technique that can be used for regression and classification tasks. In the context of active learning, Gaussian processes can be employed to model the uncertainty of the model's predictions (Riis et al.).

In Gaussian process-based active learning, the algorithm begins with an initial set of labeled data points and fits a Gaussian process model to these points. It then selects the data point with the highest uncertainty according to the Gaussian process model and requests its label from an oracle (i.e., a human expert). This labeled data point is then added to the training set, and the Gaussian process model is updated. The process is repeated iteratively until the desired level of accuracy is achieved or the labeling budget is exhausted.

One advantage of Gaussian process-based active learning is that it facilitates a principled approach to modeling uncertainty [10]. The Gaussian process model can be used to compute the

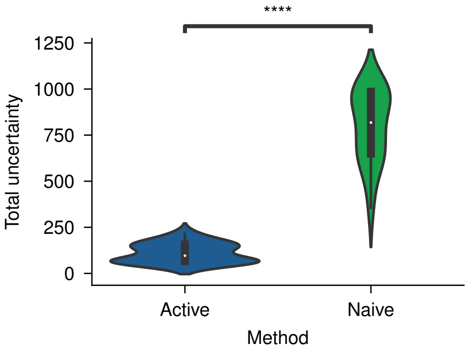

Figure 3. Active learning improves model confidence.

uncertainty of the model's predictions, which can then guide the selection of the most informative data points for labeling. Another advantage is that Gaussian processes are non-parametric models, meaning they can adapt to complex patterns in the data without making strong assumptions about the underlying distribution.

However, Gaussian process-based active learning also has some limitations. One limitation is that Gaussian processes can be computationally expensive to train and evaluate, especially for large datasets. Another limitation is that the performance of the Gaussian process model depends on the choice of kernel function and hyperparameters, which can be challenging to optimize (Krauth et al.).

Active learning in improving confidence

Following the example in Figure 2, an uncertainty-guided active learning approach can be deployed to acquire new data points that enhance the model's confidence. Figure 3 illustrates an instance of active learning in a GP regression model. The model is initialized with a small number of data points and iteratively selects new data points to acquire based on the uncertainty estimates. These estimates are used to select data points that are likely to boost the model's confidence. This approach is contrasted with a random sampling method, where data points are selected randomly. The results demonstrate that the active learning approach is capable of improving the model's confidence more rapidly than the random sampling method, given the same number of data points.

Application in scRNA-seq data

We then sought to apply the Gaussian Process (GP) to a single-cell RNA sequencing (scRNA-seq) dataset derived from peripheral blood mononuclear cells (PBMCs). PBMCs represent a heterogeneous population of immune cells, including T cells, B cells, natural killer (NK) cells, monocytes, and dendritic cells, among others. In a typical scRNA-seq experiment, PBMCs are first isolated from the blood and then subjected to droplet-based or plate-based single-cell capture. Here, individual cells are encapsulated into microfluidic droplets or wells. The cells are then lysed, and the RNA is reverse transcribed into complementary DNA (cDNA). This cDNA is subsequently amplified, and the resulting library is sequenced using high-throughput sequencing technologies, typically Illumina sequencing. The sequencing data obtained provides information about the gene expression profiles of each individual cell.

Gaussian Processes reveal epistemic uncertainty in cell type classification

We applied Gaussian Processes to a classification problem to predict cell types based on gene expression (Figure 4a). We set aside 50% of the data for testing and used the remaining 50% for training. To simulate a real-world scenario where the training data has limited coverage compared to a clinical query dataset, we additionally held out all B cells (Figure 4b). Encouragingly, the GP model was able to accurately predict the cell type of the test data with high accuracy (AUC $\approx$ 1)

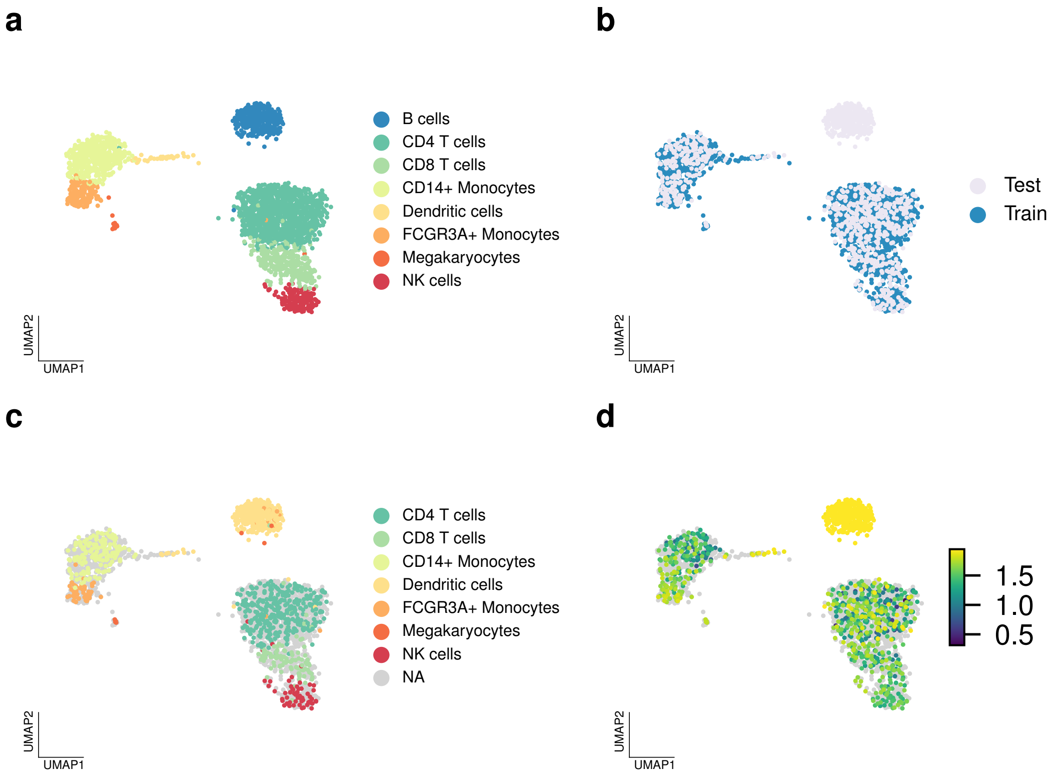

Figure 4. scRNA-seq dataset of 3000 peripheral blood mononuclear cell.

(Figure 4c). As anticipated, the model was unable to predict the held-out B cells and incorrectly classified the cells as dendritic cells, likely due to their biological similarity. This type of error is known as epistemic uncertainty, which arises from the lack of training data, leading to uncertainty in the model. This kind of classification is known as a label transfer task, and it's widely used in annotating scRNA-seq data. Misannotation is, therefore, a common pitfall in scRNA-seq analysis and can lead to erroneous biological conclusions.

In the application of GP, we calculated the uncertainty of the prediction as the Shannon entropy of the prediction probability associated with each classification (Figure 4d). As expected, the uncertainty of the prediction was highest among the held-out B cells. Surprisingly, GP was also able to assign high uncertainty accurately to rare cell types such as Megakaryocytes and Dendritic cells.

Analogous to the sinusoidal function example above, we again used active learning to iteratively select new data points to acquire based on the uncertainty estimates. In the earliest iteration, model uncertainty appears to increase across cell types (Figure 5a). This is due to the introduction of B cells into the new training set, which increases the alphabet size for entropy calculation. Initially, B cells exhibited the highest uncertainty, but as more data points were acquired, the uncertainty decreased (Figure 5b). In the final iterations, all cell types had lower uncertainty than in the initial iterations. This is because the model, having seen more data points, had increased confidence in its predictions, even for cell types initially included in the training set. Upon examining the acquired data, we observed that the model prioritized acquiring B cells in the beginning. As the uncertainty of B cells became comparable to other cell types, the model began to acquire other cell types (Figure 5d). As a comparison, we also performed a random data acquisition approach, analogous to generating the same scRNA-seq dataset experimentally. While the random approach also improved the model's confidence, it did so at a much slower rate than the uncertainty-guided approach (Figure 5c). This is because the random approach does not consider the model's uncertainty and is thus unable to prioritize data points likely to improve the model's confidence.

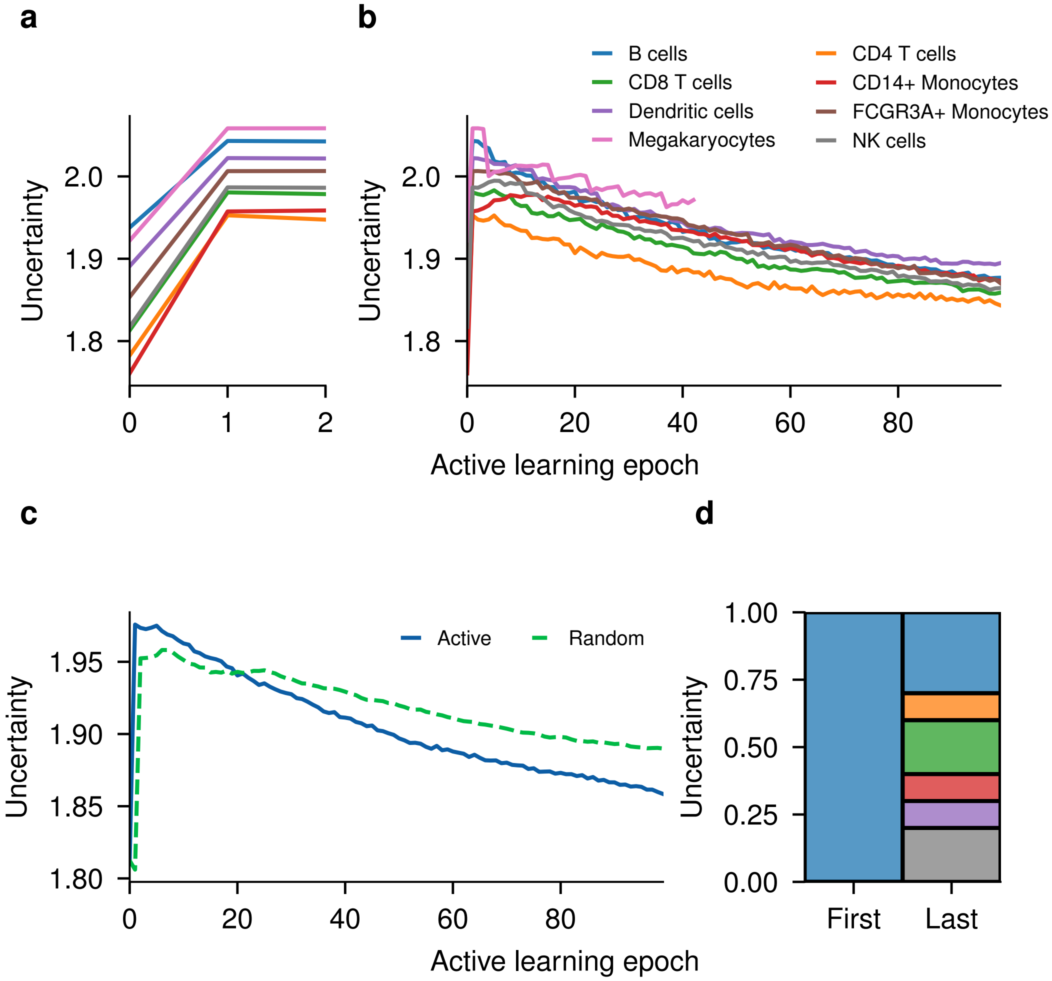

Figure 5. Active learning increase model confidence across cell types.

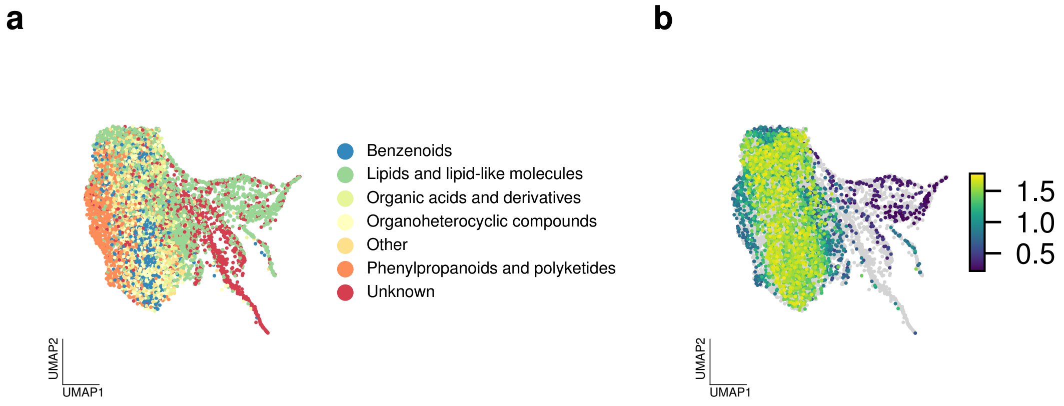

Figure 6. Gaussian processes reveal aleatoric uncertainty in molecular class prediction.

In summary, our use of GP for label transfer in single-cell RNA-seq data demonstrated high accuracy and the ability to quantify prediction uncertainty. From the example dataset, we highlighted uncertainty associated with both held-out cell types and rare cell types. This understanding can help guide future experimental design to selectively enrich for these cell types, in order to improve representation in the model.

Application in metabolomic data

Mass spectrometry (MS) is a powerful analytical technique used to identify and characterize molecules based on their mass-to-charge ratio. The basic principle of mass spectrometry is the generation of ions from a sample, which are then separated based on their mass-to-charge ratio using a combination of electric and magnetic fields. The ions are then detected and the resulting signal is analyzed to determine the mass and abundance of the ions present in the sample. The chemical structure of the molecule can be identified by analyzing the mass spectrum of the molecular ion peak, which represents the intact molecule without any fragmentation. The mass of the molecular ion provides information about the molecular weight of the compound. The fragmentation patterns of the molecule provide information about the chemical bonds within the molecule and can be used to reconstruct the molecular structure. These fragmentation patterns can be analyzed using software tools that compare the observed mass spectrum to a database of known spectra, such as the MassBank database. This comparison can then help identify potential molecular structures that match the observed fragmentation pattern. However, identifying unknown compounds or compounds from a mixture of samples using mass spectra remains challenging. Over 90% of compounds from a typical sample are unknown, and quantifying uncertainty in the identification of these compounds is critical, both for the curation of a representative database and for future experimental design.

Gaussian Processes reveal aleatoric uncertainty in molecular class prediction

We transformed spectral intensities into feature vectors using a spectral featurizer, and trained a GP classifier with 50% of the data, withholding the remaining 50% for testing. Interestingly, even without any held-out molecular class, there was substantial uncertainty in the prediction (Figure 6a, b). Comparing the average uncertainty between molecular classes, we observed that in a UMAP representation of the data, the uncertainty is highest where there is greater label mixing within a neighborhood. Molecular classes such as lipids exhibited the lowest uncertainty, as they were well separated from other classes. This shows that GP is able to capture intrinsic uncertainty due to noise in the data, known as aleatoric uncertainty.

In order to be considered a viable option for routine MS analysis, GP must demonstrate a classification performance that is on par with or superior to state-of-the-art methods. We conducted

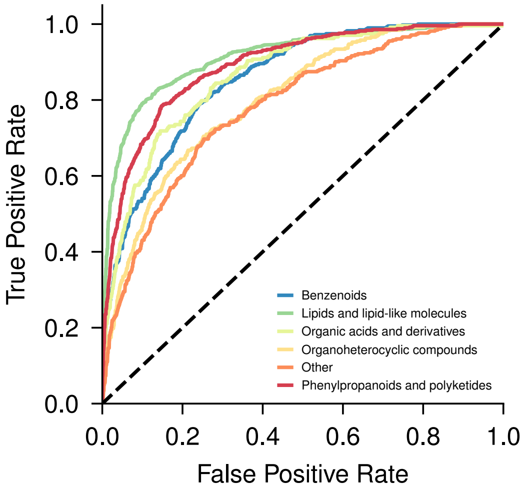

Figure 7. Gaussian processes classifier has good classification performance relative to state of art methods.

a comparison of the performance of GP to a variety of these methods, including random forest and support vector machine algorithms. Our findings revealed that GP offers comparable performance to these methods (Figure 7, Data not shown). This suggests that GP could serve as a suitable replacement for these methods, with the added advantage of being able to quantify the level of uncertainty in the prediction. This ability to estimate uncertainty could be particularly beneficial in areas such as metabolomics, where a large proportion of compounds remain unidentified and the ability to quantify the confidence in compound identification can provide valuable information for future experimental design and database curation.

Discussion

Our study has demonstrated the utility of Gaussian Processes (GP) in quantifying prediction uncertainty in both single-cell RNA-seq and metabolomic data. For the single-cell RNA-seq analysis, we were able to identify the levels of uncertainty in predictions for held-out cell types as well as rare cell types. With metabolomic data, GP was able to quantify uncertainty within the prediction of molecular classes. Importantly, in both scenarios, we showed that GP can be used to measure the uncertainty of predictions for held-out data points, which can provide valuable insights to help guide the design of future experiments, with the aim of enriching data for these particular points and thus enhancing their representation within the model.

In the current version of our implementation, we largely utilize a paired cosine similarity kernel for GP. However, there is potential for enhancing performance and scalability by exploring other kernels, such as the Radial Basis Function (RBF) kernel. While this study focused mainly on the application of GP as a drop-in replacement for standard analysis tasks within biological datasets, future research could benefit from integrating uncertainty quantification into earlier stages of the analysis process, such as spectral featurization.

Given the often large scale of data in high throughput biology, it can present challenges for the application of GP. To address this, it may be beneficial to explore performance engineering techniques in Julia, such as parallelization and GPU computing, to enhance the performance of GP.

Despite these challenges, we have clearly demonstrated the value of GP in quantifying prediction uncertainty within biological datasets. We anticipate that GP will prove to be a valuable tool in perturbation experiments, such as perturb-seq and chemical perturbation experiments, where experiments are not easily scalable and can be expensive to conduct. By quantifying uncertainty, we can more effectively guide the design of future experiments to selectively enrich for the most informative data, ultimately improving our understanding of complex biological systems.

Methods and Materials

Computational analysis

Code used in this study is available at https://github.com/nhcheng/Xavier_MS_Active_Learning_notebook. All analyses is performed using both the Julia programming language and Python. Unless otherwise specified, all gaussian processes analyses is performed using the *scikit-learn* package in Python. All analysis is performed on a 44-core Intel Xeon CPU computer with 88 GB of RAM.

Single-cell analysis

The *pbmc-3k* standard processed dataset is used throughout. All single-cell analysis is performed using the *scanpy* package in Python.

Metabolomic analysis

A pre-trained Siamese neural network MS2DeepScore to predict the structural similarity between a pair of spectra. For molecular class prediction, Classyfire is used for automated chemical classification of the molecules.

References

[Sin] Single-cell RNA-seq: Advances and future challenges | Nucleic Acids Research | Oxford Academic.

[Unc] Uncertainty of Measurement in Quantitative Medical Testing - PMC.

[family=Maaten and Hinton] family=Maaten, given=Laurens, p. d. u. and Hinton, G. Visualizing Data using t-SNE. 9(86):2579–2605.

[Hemmer et al.] Hemmer, P., Kühl, N., and Schöffer, J. DEAL: Deep Evidential Active Learning for Image Classification.

[Işık and Alptekin] Işık, K. and Alptekin, S. E. A benchmark comparison of Gaussian process regression, support vector machines, and ANFIS for man-hour prediction in power transformers manufacturing. In *Procedia Computer Science*, volume 207, pages 2567–2577.

[Kalaitzis and Lawrence] Kalaitzis, A. A. and Lawrence, N. D. A Simple Approach to Ranking Differentially Expressed Gene Expression Time Courses through Gaussian Process Regression. 12(1):180.

[Krauth et al.] Krauth, K., Bonilla, E. V., Cutajar, K., and Filippone, M. AutoGP: Exploring the Capabilities and Limitations of Gaussian Process Models.

[Maćkiewicz and Ratajczak] Maćkiewicz, A. and Ratajczak, W. Principal components analysis (PCA). 19(3):303–342.

[McInnes et al.] McInnes, L., Healy, J., and Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction.

[Riis et al.] Riis, C., Antunes, F., Hüttel, F. B., Azevedo, C. L., and Pereira, F. C. Bayesian Active Learning with Fully Bayesian Gaussian Processes.

[Simek et al.] Simek, K., Palanivelu, R., and Barnard, K. Branching Gaussian Processes with Applications to Spatiotemporal Reconstruction of 3D Trees.