Title

Gaussian processes for autonomous data acquisition at large-scale synchrotron and neutron facilities

Permalink
https://escholarship.org/uc/item/4z5380qs

Journal
Nature Reviews Physics, 3(10)

ISSN
2522-5820

Authors
Noack, Marcus M
Zwart, Petrus H
Ushizima, Daniela M
et al.

Publication Date
2021-10-01

DOI
10.1038/s42254-021-00345-y

Peer reviewed

Gaussian Processes for Autonomous Data Acquisition at Large-Scale Synchrotron and Neutron Scattering Facilities

Marcus M. Noack[1,*], Petrus H. Zwart[1,2,3], Daniela M. Ushizima[1,4], Masafumi Fukuto[5], Kevin G. Yager[6], Katherine C. Elbert[7], Christopher B. Murray[7], Aaron Stein[6], Gregory S. Doerk[6], Esther H. R. Tsai[6], Ruipeng Li[5], Guillaume Freychet[5], Mikhail Zhernenkov[5], Hoi-Ying N. Holman[2,3], Steven Lee[2,3,8], Liang Chen[2,3], Eli Rotenberg[9], Tobias Weber[10], Yannick Le Goc[10], Martin Böhm[10], Paul Steffens[10], Paolo Mutti[10], and James A. Sethian[1,11]

[1]The Center for Advanced Mathematics for Energy Research Applications (CAMERA), Lawrence Berkeley National Laboratory, Berkeley, CA 94720
[2]Molecular Biophysics and Integrated Bioimaging Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720
[3]Berkeley Synchrotron Infrared Structural Biology, Lawrence Berkeley National Laboratory, Berkeley, CA 94720
[4]Bakar Institute, University of California, San Francisco, San Francisco, CA 94143
[5]National Synchrotron Light Source II, Brookhaven National Laboratory, Upton, NY 11973
[6]Center for Functional Nanomaterials, Brookhaven National Laboratory, Upton, NY 11973
[7]Department of Chemistry, University of Pennsylvania, Philadelphia
[8]Department of Physics, University of California, Berkeley
[9]Advanced Light Source, Lawrence Berkeley National Laboratory, Berkeley, CA 94720
[10]Institut Laue-Langevin (ILL), 71 Avenue des Martyrs, CS 20156, 38042 Grenoble Cedex 9, France
[11]Department of Mathematics, University of California, Berkeley
[*]MarcusNoack@lbl.gov

July 26, 2021

Abstract

The execution and analysis of ever more complex experiments are increasingly challenged by the vast dimensionality of the parameter spaces that underlie investigations in the biological, chemical, physical, and materials sciences. While an increase in data-acquisition rates should allow broader querying of the parameter space, the complexity of experiments and the subtle dependence of the model function on input parameters remains daunting due to the sheer number of variables. To meet these challenges, new strategies for autonomous data acquisition are rapidly coming to fruition, and are being deployed across a spectrum of scientific experiments. One promising direction that is being explored is the use of Gaussian process regression (GPR). GPR is a quick, non-parametric, and robust approximation and uncertainty quantification method that can directly be applied to autonomous data acquisition. In this work, we review and reformulate our most recent contributions to GPR-driven autonomous experimentation in more general terms, and illustrate the functionality of the techniques we present, using new, real-world examples from large experimental facilities in the United States (ALS and NSLS II) and France (ILL). We start by introducing the basics of a GPR-driven autonomous loop with a focus on Gaussian processes. We then shift the focus to the infrastructure that has to be built around GPR to create a closed-loop. Finally, our examples show that Gaussian-process-based autonomous data acquisition is a widely applicable method that can facilitate the optimal utilization of instruments and facilities by enabling the efficient acquisition of high-value datasets.

1 INTRODUCTION

Experimental facilities around the globe are facing a challenging task: while equipment is becoming increasingly advanced, leading to a steady rise in data acquisition rates, it is still outpaced by the increase in complexity of scientific questions and therefore unable to reliably answer them without parallel advances in experiment design. The rise in complexity leads to high-dimensional parameter spaces, spanned by the parameters describing the sample, the instrument, and the data acquisition protocol, embracing the full range of synthesis, processing, and environmental parameters that describe the sample and its characterization throughout the design-experiment loop. These parameter spaces have to be explored efficiently in search of new scientific discoveries. Traditional methods address the rise in dimensionality by checking more and more possible configurations, counting on an increase in data-acquisition rates. This "brute-force" approach takes full advantage of advanced computing facilities for computation and storage, and new fast detectors. In a standard approach, a Cartesian grid is often defined, which is then used to automatically control and schedule measurements. This approach is often referred to as scanning or raster scanning. Since little information about the model is known beforehand, the grid is often defined to be very fine, which leads to long periods of data collection, and a significant amount of redundant data is often collected, processed, and stored. Another approach is to perform a coarse-grid scan first and then, based on the practitioner's input, focus on sub-regions of the parameter space using a fine-grid scan. This method needs human intervention and potentially leads to bias, lost information, and overly redundant data. As the dimensionality of the parameter space increases, grid-based approaches become increasingly impractical, since the number of grid points scales with the power of the dimension of the space — even a "simple" problem inhabiting a ten-dimensional parameter space is prohibitively expensive under such a brute-force approach.

In the case of high-dimensional parameter spaces, practitioners often change their approach to an intuition-based technique, in which, after just a few measurements are acquired, the user attempts to make out patterns and trends in the data, which will then be used to steer the experiment. While data is collected more deliberately, this approach leads to highly-trained scientists micromanaging the experiment, while choosing measurements sub-optimally; after all, human brains are not well-equipped for pattern recognition and decision-making in high-dimensional spaces. This results in the need for constant vigilance and attention from the experiment designer, as well as the expectation that the user will be able to interpret and integrate the current results on the basis of all the previous measurements. Additionally, user-bias can creep in, in which the expectation that results should look a certain way, or that unexplored regions probably won't be interesting, can skew the investigation.

Many of the problems that come with high-dimensional parameter spaces can be solved by designing methods for improved decision-making during an experiment. Designing optimal experiments is not new, and can be traced back to the mid-1800s with the work of C.S. Peirce in his series "Illustrations of the Logic of Science" [46, 47]. The field became soon known as design of experiments (DOE), which still exists, and its best-known method, the latin-hyper-cube method is still widely used [33, 15]. With the rise of machine learning, DOE transitioned and was partly replaced by "active learning"; a collection of mathematical tools that allow a machine to choose where data should be collected [51, 26]. Many methods of active learning are rooted in Bayesian analysis in which a prior probability density function is placed over the entities of interest, and is then conditioned to infer knowledge. The advantage of Bayesian methods within machine learning is that uncertainty quantification is naturally included and can be used for decision-making and interpretation of the model. One particular interesting Bayesian model, due to its tractability and robustness, is the Gaussian process (GP), notable in part because all equations needed for inference are available in closed form; this makes training and prediction efficient enough to be used within an autonomous-data-acquisition loop. In Gaussian process regression (GPR), a Gaussian prior probability density function (PDF) is placed over all functions we deem to constitute possible regression models. As soon as some initial data is acquired, the prior PDF can be conditioned to yield a posterior mean and variance at each point of the model. The prior and the posterior can then be used to make autonomous decisions about future optimal measurements.

We have been pursuing this approach, developing autonomous-data-acquisition algorithms and applying them to a variety of experiments, especially at light source facilities [42, 43, 39, 44, 63]. The autonomous loop starts with as little as two data points, which are then used to train and condition the prior. The training is repeated regularly to account for changing information in the dataset. The prior and the posterior (mean and variance) are then used to decide where future data will be collected. Most often, the posterior mean and posterior variance are combined in a user-defined way to create an acquisition function, which is then maximized to obtain the next optimal points to perform measurements. For certain types of acquisition functions, the search resembles Bayesian optimization [17]. A schematic of the autonomous-experiment loop is depicted in Figure 1.

In this work, we review and reformulate, in more general terms, our formulation of GPR for autonomous experimentation, outlining various improvements to the core methods and demonstrating some of our techniques using four case studies (Table 1). Additionally, this paper introduces our GPR-driven, readily available software framework *gpCAM* [37] to the community. Table 1 summarizes the qualitative performance gains that were achieved in those case studies by employing the described approach.

Our aim in this perspective is to provide an overview of GPR-driven autonomous experimentation to designers of experimental facilities, users, instrument scientists, and support staff. The paper aims to provide overviews about the technical framework, the software tool, and case studies that illustrate examples of GPR for various experiment techniques, with the overall objective to enable more practitioners to run their experiments autonomously and accelerate scientific discovery.

The paper is organized as follows: First, we will show a birds-eye view on Gaussian processes, including some of the details that are particularly important for autonomously steered experiments. Second, we will discuss some improvements to the standard framework, the computational challenges we faced, and a few ways to address them. Third, we show several new, unpublished use-cases illustrating the flexibility, functionality, and success of the strategy.

Experiment	Dim. of Input Space	Benefit
Autonomous X-ray Scattering Mapping ([42, 43, 44]; Section 3.1)	2	surveying parameter-space landscape and identifying regions of interest quickly, beam utilization vastly increased (e.g., a six-fold reduction in the number of measurements required to reach a similar model quality compared to a grid scan [42])
Autonomous Synchrotron Infrared Mapping ([21]; (Section 3.2)	2	identifying regions of interest quickly without human biases, up to 50 times fewer data needed for interpretation, beam utilization increased, enable shorter timescale (higher time resolution), observation of transient chemical processes [21, 58]
Autonomous Discovery of Correlated Electron Phases (Section 3.3)	2	dataset size decreased to <10 % of original size
Inelastic Neutron Scattering (Section 3.4)	4	no human intervention, experiment time decreased from several days to one night, hit rate inside the ROIs about twice as high compared to random and grid-based measurement selection

Table 1: An overview of the benefit of using the proposed GP-driven autonomous data acquisition for our case studies. For a quantitative comparison of data-collection methods see [42, 43, 44, 34].

2 Gaussian Process Regression for Autonomous Data Acquisition

GP-driven autonomous data acquisition requires an entire automated workflow to successfully operate without any human intervention. The workflow is comprised of several steps: At the core of the algorithm, we need a flexible and fast Gaussian-process engine (Section 2.1). Due to rapidly advancing technology of instruments and detectors, data-acquisition rates are accelerating; the GP has to keep up with the data acquisition to avoid wasting valuable instrument time. Therefore, the training and the prediction have to be

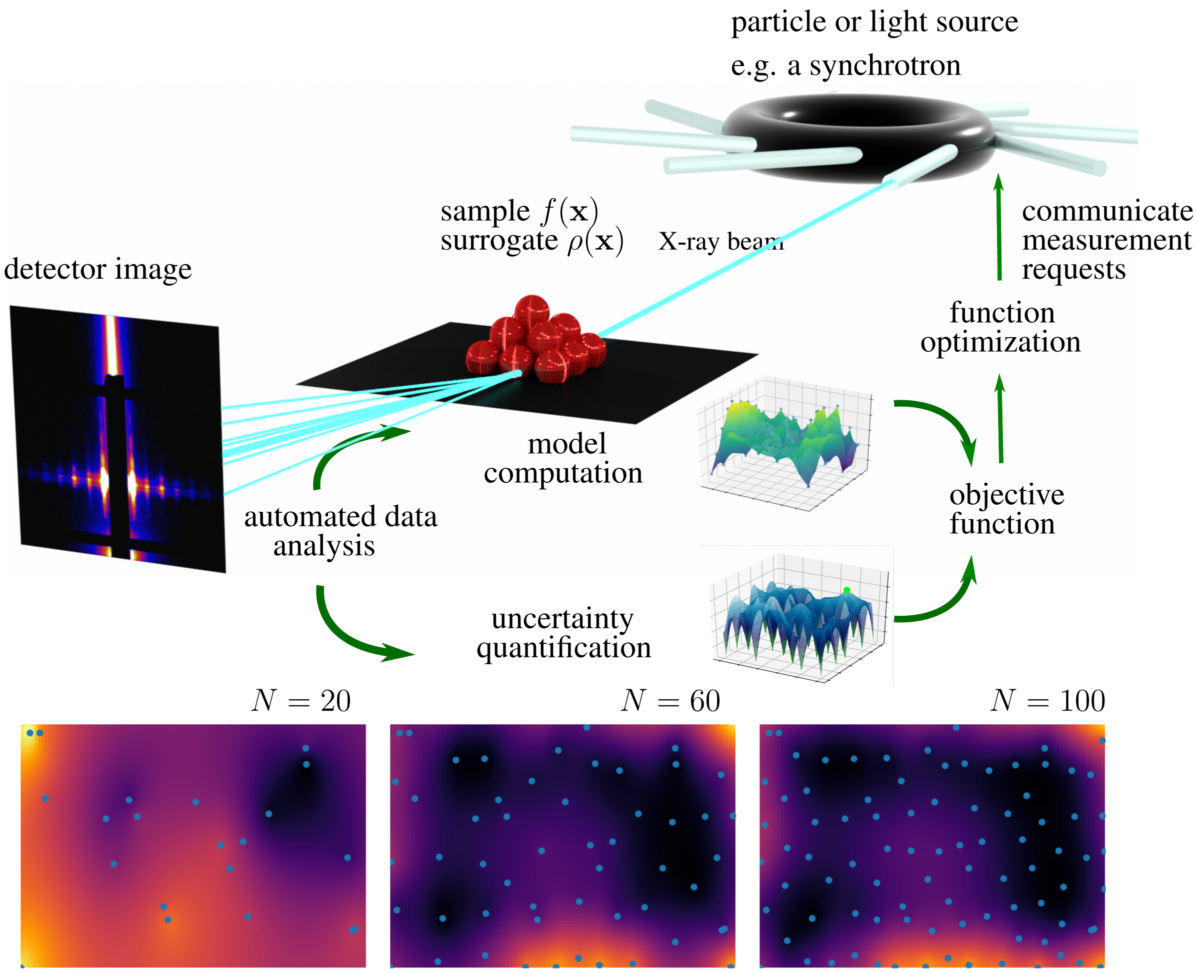


Figure 1: Schematic of an example autonomous experiment at a synchrotron radiation beam line. The measurement results depend on parameters $\mathbf{x}$. The raw data is sent through an automated data processing and analysis pipeline. This pipeline often includes featurization and dimensionality reduction techniques (Section 2.4). From the analyzed data, the Gaussian-process-based autonomous-experiment algorithm creates a prior probability density function by finding the possibly anisotropic variations of data with distance (Section 2.2). Conditioning the prior yields posterior mean and covariance functions which can be combined with the trained prior to define an acquisition function whose maxima (or minima) represent points of high-value measurements (Equation 11); they are found by employing tailored function optimization tools. The new measurement parameters $\mathbf{x}$ are then communicated to the data-acquisition device and the loop starts over. The surrogate model — most often the posterior mean $m(\mathbf{x})$ — approximates the unknown latent function $f(\mathbf{x})$. The measurements can be subject to heterogeneous (non-i.i.d) measurement noise. In the bottom row, the approximation of *Himmelblau's* function is shown. In this case, points are chosen based on the maximum posterior variance — a common choice for pure exploration — which can give the impression of a random distribution; however, in contrast to randomly chosen points, measurements are guaranteed to be placed in areas of high uncertainty and clustering is therefore naturally avoided. N indicates the number of performed measurements. Image courtesy of [44].

computed as efficiently as possible. The training process is predominantly composed of maximizing the marginal log-likelihood function (Section 2.2). The result of the GP, the trained prior and the posterior, are used in the acquisition-function definition (Section 2.3) which is passed through another function optimization to find the optimal measurement positions. As before, this optimization has to be concluded as rapidly as possible to avoid stalling the instrument. The acquisition-function optimization needs the repeated evaluation of the posterior mean and variance, which draws the connection to efficient GP computations (Section 2.5). To increase efficiency, the GP-driven autonomous data acquisition never operates on raw data; the raw data resulting from a measurement is passed through a dimensionality-reduction operator, e.g. statistical techniques such as PCA or NMF or machine learning techniques such as neural nets or clustering algorithms (Section 2.4). Recently, there has been a push towards domain and physics-aware machine learning; we will address this topic for GP-driven autonomous data acquisition in Section 2.6.

2.1 A Birds-Eye View on Gaussian Process Regression

A Gaussian process (GP) gets its name from a Gaussian prior probability density function that is defined over all conceivable model functions. The model functions are defined over our parameter space — the set of all combinations of parameters the measurement outcome depends on, e.g. stage position, temperature or other synthesis, processing, or environmental parameters — as linear combinations of, so-called, kernel functions. The kernel functions depend on hyperparameters, most often correlation length scales and signal variances. The Gaussian prior can be trained — which amounts to finding the hyperparameter values — and conditioned on the data, resulting in a posterior which is used in an acquisition function. The acquisition function can then be passed to an optimizer to find optimal next measurement positions. This high-level view of GP-driven autonomous data acquisition is sufficient to follow the rest of this paper and to use our software tool in its basic form. For advanced use and theory-oriented readers, we define a GP more rigorously in this section.

Defining a GP regression model from data $D = \{(\mathbf{x}_1, y_1), ..., (\mathbf{x}_N, y_N)\}$, where $y_i = f(\mathbf{x}_i) + \epsilon(\mathbf{x}_i)$, is accomplished in a Bayesian framework by placing a Gaussian probability density function

$$p(\mathbf{f}) = \frac{1}{\sqrt{(2\pi)^N |\mathbf{K}|}} \exp\left[-\frac{1}{2}(\mathbf{f}-\boldsymbol{\mu})^T \mathbf{K}^{-1}(\mathbf{f}-\boldsymbol{\mu})\right], \tag{1}$$

called the prior, over a function space, and condition it on the data. Due to noise in the data, one also defines a likelihood as

$$p(\mathbf{y}) = \frac{1}{\sqrt{(2\pi)^N |\mathbf{V}|}} \exp\left[-\frac{1}{2}(\mathbf{y}-\mathbf{f})^T \mathbf{V}^{-1}(\mathbf{y}-\mathbf{f})\right]. \tag{2}$$

$\boldsymbol{\mu} = [\mu(\mathbf{x}_1), ..., \mu(\mathbf{x}_N)]^T$ is the mean of the prior Gaussian probability density function, $\mathbf{f} = [f(\mathbf{x}_1), ..., f(\mathbf{x}_N)]^T$, $\mathbf{K}_{ij} = \mathbf{k}(\phi, \mathbf{x}_i, \mathbf{x}_j);\ \mathbf{x} \in \mathcal{X}$ is the covariance of the Gaussian process, with its covariance function or kernel $k(\phi, \mathbf{x}_i, \mathbf{x}_j)$, where ϕ is a set of hyperparameters, commonly length scales l and signal variance σ_s^2, and where $\mathbf{V}$ is the covariance of the non-i.i.d. observation noise, which has been shown to play an important role in autonomous data acquisition [44].

The kernel k is a symmetric and positive semi-definite function, such that $k : \mathcal{X} \times \mathcal{X} \to \mathcal{R}$. $\mathcal{X}$ is our parameter space which is commonly referred to as index set or input space in the literature. A common choice is the Matérn kernel class defined as

$$k(\mathbf{x}_i, \mathbf{x}_j) \ = \ \sigma_s^2 \frac{2^{(1-\nu)}}{\Gamma(\nu)} \left(\sqrt{2\nu}\frac{r}{l}\right)^{\nu} B_\nu \left(\sqrt{2\nu}\frac{r}{l}\right), \tag{3}$$

where B_ν is the Bessel function of second kind, Γ is the gamma function, $r \ = \ ||\mathbf{x}_i - \mathbf{x}_j||_{l_2}$ and ν is a parameter that controls the order of differentiability of the kernel and therefore of the posterior mean and covariance functions [64]. We can change the distance measure in $\mathcal{X}$ to give the GP the flexibility to learn anisotropic correlations between data points [44]. Kernels are an integral part of the theory of Gaussian process regression and have a major impact on the quality of the prediction. The hyperparameters ϕ, and if desired, the prior-mean function can be found by maximizing the marginal log-likelihood, i.e., solving

$$\underset{\phi\ \mu}{\arg\max} \Big(\log(L(D; \phi, \mu(\mathbf{x}))) \Big) \tag{4}$$

where

$$\log(L(D;\phi,\mu(\mathbf{x}))) \;=\; -\frac{1}{2}(\mathbf{y}-\boldsymbol{\mu})^T(\mathbf{K}(\phi)+\mathbf{V})^{-1}(\mathbf{y}-\boldsymbol{\mu}) \\ -\frac{1}{2}\log(|\mathbf{K}(\phi)+\mathbf{V}|)-\frac{\dim}{2}\log(2\pi). \tag{5}$$

In the isotropic case, we only have to optimize for one signal variance and one length scale (per kernel function). The prior-mean function $\mu(\mathbf{x})$, is often assumed to be constant in the standard literature; however it is a powerful tool to communicate domain knowledge in form of a trend to the GP (Section 2.6). The optimization problem can also be subject to constraints, which is another flexible way to use domain knowledge. With anisotropic distances and a non-constant prior-mean function, the number of hyperparameters, we have to find, can quickly increase, which gives rise to the need for efficient optimization methods (Section 2.2). Given a set of optimized hyperparameters, the joint prior is given as

$$p(\mathbf{f},\mathbf{f}_0)=\frac{1}{\sqrt{(2\pi)^{\dim}|\boldsymbol{\Sigma}|}}\exp\left[-\frac{1}{2}\Big(\begin{bmatrix}\mathbf{f}-\boldsymbol{\mu}\\ \mathbf{f}_0-\boldsymbol{\mu}_0\end{bmatrix}^T\boldsymbol{\Sigma}^{-1}\begin{bmatrix}\mathbf{f}-\boldsymbol{\mu}\\ \mathbf{f}_0-\boldsymbol{\mu}_0\end{bmatrix}\Big)\right], \tag{6}$$

where

$$\boldsymbol{\Sigma} \;=\; \begin{pmatrix}\mathbf{K} & \boldsymbol{\kappa}\\ \boldsymbol{\kappa}^T & \boldsymbol{\mathcal{K}}\end{pmatrix}, \tag{7}$$

where $\boldsymbol{\kappa}_i \;=\; k(\phi,\mathbf{x}_0,\mathbf{x}_i)$, $\boldsymbol{\mathcal{K}} \;=\; k(\phi,\mathbf{x}_0,\mathbf{x}_0)$ and, as a reminder, $\mathbf{K}_{ij} \;=\; k(\phi,\mathbf{x}_i,\mathbf{x}_j)$. $\mathbf{x}_0$ is the point of interest $\in\ \mathcal{X}$.

The predictive distribution is defined as

$$p(\mathbf{f}_0|\mathbf{y}) \;=\; \int p(\mathbf{f}_0|\mathbf{f},\mathbf{y})\,p(\mathbf{f},\mathbf{y})\,d\mathbf{f} \\ \propto \mathcal{N}(\boldsymbol{\mu}+\boldsymbol{\kappa}^T\,(\mathbf{K}+\mathbf{V})^{-1}\,(\mathbf{y}-\boldsymbol{\mu}),\boldsymbol{\mathcal{K}}-\boldsymbol{\kappa}^T\,(\mathbf{K}+\mathbf{V})^{-1}\,\boldsymbol{\kappa}) \tag{8}$$

and the predictive mean and the predictive variance are therefore respectively defined as

$$m(\mathbf{x}_0)=\boldsymbol{\mu}+\mathbf{k}^T(\mathbf{K}+\mathbf{V})^{-1}(\mathbf{y}-\boldsymbol{\mu}) \tag{9}$$

$$\sigma^2(\mathbf{x}_0)=k(\mathbf{x}_0,\mathbf{x}_0)-\mathbf{k}^T(\mathbf{K}+\mathbf{V})^{-1}\mathbf{k}, \tag{10}$$

which are the posterior mean and variance at $\mathbf{x}_0$, respectively.

The trained prior, and the GP posterior mean and variance can be combined to form the acquisition function that will be used to find the next optimal measurements in the GP-driven autonomous data acquisition loop (Section 2.3). However, any acquisition function is only as good as the function approximations coming out of the GP, which in turn depend on a well-trained prior. Therefore, the prior has to be trained in an intelligent way, which will be discussed in the next section.

2.2 Training a Gaussian Process Model

In the standard GP literature, there is often not much attention spent on the training of a Gaussian process, even though it the most important and costly step. This is due to the fact that GPs can provide results with just two hyperparameters if the parameter space is chosen to be isotropic and when standard kernels are used. Those two hyperparameters are the signal variance and the isotropic length scale. For a small number of hyperparameters, there is no need to spend much effort on finding the optimal method for the marginal-log-likelihood optimization (Equation (4)). However, if anisotropic distances are considered and more advanced kernel definition are used, the number of hyperparameters increases quickly beyond 10. When we want to consider non-stationary kernels or spectral kernels, this number scales with the dimensionality of the input space, and can easily exceed 100. In high-dimensional spaces, standard multi-start gradient optimizations will most likely fail to find a decent solution and global methods need too many function evaluations to converge. This problem is magnified for autonomous experimentation, where the hyperparameters have to be updated repeatedly while the dataset grows. We have developed a tool that can perform the training using HPC architecture and operate on a subset of the data without stalling the prediction. The method is inspired by a hybrid optimization scheme [40] which, while allowing for parallel execution on HPC architecture, also allows for multiple solutions to be found. Those multiple solutions translate into several interpretations of the same dataset, which in turn, allow for several “realities” to be explored. By overlapping training and prediction, the cost of the training can effectively be hidden which increases the overall performance of the autonomous experiment significantly.

The method works as follows: A set of walkers are placed inside the domain. Each walker performs a local optimization algorithm. The individual walks are entirely independent and can therefore be distributed across processes that run in parallel on different cores or nodes. When all walkers have converged, the solutions are communicated and a local deflation operator is applied to the gradient of the function where optima were found. The local deflation operator is based on the bump function and effectively removes optima from the function, avoiding that walkers find the same solution again. The fittest walkers are replaced by a global optimization to new locations where the procedure starts over. The result is a sorted list of optima, which can be queried whenever new solutions, hyperparameters in the case of the GP training, are needed. Since local optimizations can be done in parallel, we can compute and utilize the Hessian of the marginal log-likelihood, which provides additional information about the solution.

2.3 Customizing the GP-Driven Autonomous Data Acquisition

After the GP is trained and conditioned on the data, the prior and the posterior can be used to define an acquisition function, which encodes the main drive of the experiment in functional form. It can, for instance, measure the total uncertainty in the acquired signal over the full parameter space or favor specific morphological features, such as grain size or chemical composition. Once a suitable acquisition function is defined by the user, it is passed through a mathematical optimization to find the optimal next point for data acquisition. The definition of the acquisition function as a scalar function on $\mathcal{X}$

$$\begin{aligned} \mathcal{X} &\rightarrow \mathcal{R} \\ f_a(\mathbf{x}) &= f_a(m(\mathbf{x}), \sigma^2(\mathbf{x}), \boldsymbol{\mathcal{K}}, \boldsymbol{\Sigma}(\mathbf{x})) \end{aligned} \tag{11}$$

directly impacts where new measurement points are anticipated to have the highest impact — the locations of these points can be obtained by numerical optimization techniques. The simplest example of the acquisition function is

$$f_a(\mathbf{x}) = \sigma^2(\mathbf{x}), \tag{12}$$

whose maximization means placing points where the current estimated posterior variance is a maximum. Definition (11) allows for most Bayesian optimization techniques, such as the upper confidence bound

$$f_a(\mathbf{x}) = m(\mathbf{x}) + c\,\sqrt{\sigma^2(\mathbf{x})}, \tag{13}$$

and information-theoretical entities, like the Shannon-information gain [20]

$$f_a(\mathbf{x}) = Entropy(\boldsymbol{\mathcal{K}}) - Entropy(\boldsymbol{\Sigma}(\mathbf{x})) \tag{14}$$

to be computed and used for measurement-location suggestions. The acquisition function can also effectively be used to find regions of interest, e.g. a particular material property or behavior. This method will be used in case study 3.2, in which which the similarity to a reference spectrum is used to find regions of interest.

In addition to emphasizing particular characteristics of the prior or the posterior, the acquisition function can be used to assign costs to potential measurement locations and include them into the optimization. Cost functions are defined as

$$\begin{aligned} \mathcal{X} \times \mathcal{X} &\rightarrow \mathcal{R} \\ c &= c(\mathbf{x}_1, \mathbf{x}_2) \end{aligned} \tag{15}$$

and define the cost of the measurement itself but also the cost of moving within the parameter space $\mathcal{X}$, from $\mathbf{x}_1$ to $\mathbf{x}_2$ (e.g. sample-position movement [43]). Dividing the acquisition function by the cost function leads to the next measurement being suggested where it is the most efficient with respect to the cost function.

With growing dataset sizes, we have to be increasingly mindful when it comes to choosing the optimization tool to maximize the acquisition function. Finding the optimum of the acquisition function has to be done efficiently to avoid stalling the instrument. In our experience, for dataset sizes smaller than 1000 points and data acquisition-rates of more than 10 seconds per measurement, an efficient global optimization will suffice. For larger datasets or faster data acquisition, local and hybrid optimization schemes have to be applied to minimize the number of function evaluations. In addition, it is often requested that several local optima and ideally the global optimum are found. An optimization method as described in Section 2.2 [40] can help accelerating the optimization and provides several suggested optimal measurements.

2.4 A Note on Raw Data and GP-Based Autonomous Experimentation

Often times, raw data cannot be used in GP-based autonomous experimentation, either due to formatting of the raw data or the unfavorable worst-case scaling of exact GPs, which is $O(N^2)$ in storage and $O(N^3)$ in computation. Here, N is the number of data points. For instance, computing a dense GP over all data points of two 1000×1000 detector images would lead to a covariance matrix of size 4×10^{12}, which exceeds most storage capabilities and should not be inverted. Therefore, image data from detectors are more valuable to GP methods after preprocessing and/or using dimensionality reduction techniques to extract a high-value dataset. Popular techniques include automatic featurization using convolutional neural networks [30], principal component analysis [2], non-negative matrix factorization, and clustering [34].

Automated image featurization by means of an artificial neural network can extract valuable information out of very large inputs, providing relevant yet compact signatures about the raw data. Previous work, in Reference [31], employed deep learning techniques based on convolutional neural networks, able to transform disparaged data collections with millions of samples into searchable/query-enabled databases [2]. In order to fully exploit experimental data such as X-ray scattering patterns or spectra, one of the first steps is the characterization of dense datasets so that we can both easily represent them in a compact way and that allows organizing very large collections of patterns for comparison, sorting and retrieval.

The result of feature extraction and dimensionality reduction is a low-dimensional feature vector that serves as input to the GP. If this step is done correctly — meaning that the most informative features of a dataset (e.g. an image) are extracted successfully — it leads to inexpensive and information-rich approximations and steering. Generally speaking, the optimal feature extraction method depends on the experimental technique, the nature of the data, and the overall objective of the experiment. Despite dimensionality reductions and feature extractions, the GP itself has to be highly optimized to keep up with the data acquisition. This will be the topic of the next section.

2.5 Computational Aspects of Autonomous Data Acquisition

Rising data-acquisition rates pose a challenge for any technique that is attempted in order to achieve autonomous data acquisition. In a traditional framework, we would interpret every iteration of the experiment loop (Figure 1) as dependent on the last one, which means that the Gaussian process has to be retrained and the maximum of the acquisition function has to be found in every iteration. With a worst-case scaling of GPs of $O(N^3)$ — N being the number of data points — this is infeasible even for moderate acquisition rates and dataset sizes. However, there are many techniques to speed up the GP computation; they can generally be divided into exact and approximate techniques. A vast amount of work has been conducted in the past to find approximations of GPs that scale much more favorably [48]. The methods are often referred to as inducing-points-GP models and are normally based on using a subset of the data points or pseudo-data points. As the name alludes to, the methods are not exact and errors depend on the approximated function itself and the associated number of needed pseudo-data points. For highly nonlinear functions, the improvement in scaling practically vanishes. Additionally, the function values at the pseudo-data points are found by using standard interpolation techniques that only work in low-dimensional spaces.

Fortunately, there are exact methods for speeding up GPs and GP-based autonomous experimentation. The most costly computation of the GP is the inversion of the covariance matrix in Equations (5), (9) and (10). However, as commonly recommended in computational linear algebra, the inversion should be transformed into a solution of a linear system which is faster and more accurate. Combined with using sparse kernels, the GP training and prediction can be sped up significantly. Depending on the computing architecture and dataset size, several independent Gaussian processes, called Gaussian process experts, can be used and distributed across cores or nodes to reach practically linear scaling [9, 18]. A significant speedup has been achieved by reformulating all system solutions as matrix-vector products which can be computed very efficiently on GPU computer architecture. However, even this technique takes away from the flexibility of Gaussian process regression since it works best for certain well-behaved classes of kernels.

Given a computationally optimized GP, it is important to minimize the number of function evaluations of the marginal log-likelihood (5), the posterior mean (9), the posterior variance (10), and their derivatives.

All functions are evaluated within a function optimization to find the optimal hyperparameters or the next optimal measurement location, respectively. One way to minimize the total number of evaluations of the log-likelihood function and its derivatives is to perform global training from scratch only rarely when the dataset size has changed a significant amount and to update hyperparameters locally more often. A more advanced method is to decouple prediction and training entirely by performing the training asynchronously to the autonomous-experiment loop. In this case, the training cost can effectively be hidden (see also Section 2.2).

The next optimal measurement location is determined by maximizing the acquisition function. This does not have to be done by a global optimization in every iteration but can be performed locally depending on the characteristics of the acquisition function. Hybrid optimizers can be used to find an effective trade-off between costly global and inexpensive but potentially weak local optimizers (Section 2.2)[40]. Hybrid optimizers often take advantage of several local walkers which can be distributed over several cores or nodes on high-performance computing architecture. Combining exact methods for fast GPs with these novel optimization methods opens the possibility to perform autonomous data acquisition up to millions of data points.

2.6 Making GP-Driven Autonomous Data Acquisition Physics Aware

As scientific problems become more involved, the dimensionality of the associated parameter spaces is increasingly too high for them to be filled with information from data alone. As an example, imagine a function over a 100-dimensional parameter space $[0, 10]^{100}$ and assume that we need around 1 point per unit volume to define the model function with confidence. This scenario would result in 10^{100} needed measurements. Given today's increase in data acquisition rates, even far-future experiments would only be able to complete a fraction of the needed measurements within a human lifetime.

The solution comes from domain and, more specifically, physics knowledge of the practitioners. This knowledge, together with a limited amount of data can be used to propagate information to all regions of vast parameter spaces. The user should be aware: Including domain knowledge into a GP framework erases the agnosticism of the basic framework and may cause bias in the steering and analysis of the final model. However, it is possible to inject knowledge in parametric and probabilistic ways and let the algorithm learn, adapt, confirm or reject certain knowledge bases if the collected data warrants it. We want to draw attention to three ways of how domain knowledge can be embedded in GP-driven autonomous data acquisition. (1) Advanced kernel designs can be used to communicate hard constraints to the posterior mean. The model could, for instance, be constrained to functions of the type $f(\mathbf{x}) = \sum_{i=1}^{100} g(x_i)$ which would, given the problem described above, result in the same quality model after only $(100 \times 9) + 1$ measurements — with a correlation length of 1 we would need 9 points in every direction plus 1 at the origin. While this simplification might seem a little too optimistic, it illustrates the power of domain knowledge. Advanced Kernel designs are also a great tool to enforce periodicity, symmetry, or a given order of differentiability upon the model function [41]. Combined with anisotropic kernel design, these constraints can be defined per direction of the parameter space. These hard constraints have to be used carefully; once the constraint is active, the posterior mean will have to adhere, even if the data is inconsistent with the constraint. (2) The maximization of the marginal log-likelihood function can be subject to constraints. This is useful to inform the prior of other data sources from simulations or other experiments or to constrain the posterior or its derivatives within certain bounds. The position of the prior, the prior mean, can play the role of a known or partly known trend which is then used to inform the model function. Depending on the exact nature of the constraint on the marginal log-likelihood, the danger of imposing wrong constraints varies. A wrongly defined prior mean, for instance, will affect the steering efficiency and therefore the value of the collected data; however, when the extra data has been collected, the final model is unaffected. (3) The acquisition-function definition can be used to emphasize certain domain-informed characteristics of the surrogate model. For instance, one could define an acquisition function that emphasizes high or low gradient or curvature regions [42, 43] — the interest in those regions originates from domain knowledge. Examples of autonomous experiments using a targeted acquisition function are presented in Sections 3.1, 3.2, and 3.4. The introduced bias depends on the exact definition of the acquisition function. Some will fall back to exploration if the target cannot be found, some others will find the first region and get stuck. Users should generally use caution when defining acquisition functions. All the mentioned techniques to make autonomous data acquisition domain aware are implemented in the current version of the code and are ready to be used. More details on advanced kernel designs and the potential benefits can be found in Reference [41].

2.7 *gpCAM*

In Section 2, we outlined the basics of Gaussian-processes-driven autonomous data acquisition and highlighted certain computational and mathematical aspects that can be customized to increase the efficiency of the autonomous experiment and to inject domain knowledge into the GP. All mentioned customizations can be deployed by using our in-house implementation of a GP-driven autonomous experiment, called *gpCAM* [42, 43, 39, 44]. *gpCAM* is a light-weight, flexible and HPC-ready Python implementation that can be downloaded from the bitbucket repository [37] or installed via *pip install gpcam*. To set up *gpCAM* in the basic operational mode, we only need the parameter space $\mathcal{X}$ and some bounds for the hyperparameters. For more advanced use, cost, kernel, prior-mean and acquisition functions, as well as advanced optimization techniques and constraints, can easily be communicated to *gpCAM* as arguments in the API function calls. For data-acquisition rates in the order of seconds and a desired number of measurements of < 10000, a desktop computer is sufficient for the GP analysis — assuming that decision-making time should be less than the time needed for one measurement. Faster acquisition rates or more measurement points call for a more powerful computer architecture.

3 Case Studies

The following case studies demonstrate how the GP-driven autonomous data acquisition is used at large experimental facilities. The examples were chosen in order to cover a variety of different techniques and scientific objectives, with the ultimate goal to provide the reader with guidance on how to set up their own autonomous experimentation. The presented experiments use our software, *gpCAM*.

3.1 Autonomous X-ray Scattering Mapping

X-ray scattering is a powerful technique for quantifying material order at the atomic, molecular, and nanoscale [16, 13, 66]. The high-flux and small X-ray beam size available at modern synchrotrons provides the opportunity to measure samples in an imaging mode, where a scattering/diffraction pattern is collected at every (x, y) sample position. This provides a wealth of information across scales, since material structure and heterogeneity is imaged at the macro- and micro-scale, while analysis of the scattering pattern provides detailed information about the order at the nano and atomic scale. Mapping experiments are typically performed as grid scans, where a step-size and scan region are selected ahead of time. This has been used to image heterogeneous materials such as plant cell walls [29], tissues [45, 1, 28], and biominerals [60, 59]. However, grid scans have several shortcomings, especially for heterogeneous materials for which the characteristic length-scale(s) of ordering are not known ahead of the measurement. In particular, a grid scan requires the arbitrary selection of a step-size, which dictates the imaging resolution. Data acquisition proceeds progressively through the defined scan-region, potentially wasting valuable experiment time by measuring sample regions ultimately found to be irrelevant to the underlying scientific question of the study.

Autonomous experimental decision-making can be easily applied to the selection of measurement points in the two-dimensional imaging space — i.e. sample coordinates for X-ray scattering measurements. As discussed in Section 2.4, one can use an automated data-analysis pipeline to extract a set of features or "signals" of interest from the scattering pattern, and feed these into *gpCAM*, which then selects the next measurement point according to the defined objective. Using the default search behavior — where *gpCAM* is minimizing the uncertainty of the surrogate model (via Equation (12) in Section 2.3) — the progression of measurement points naturally begins by spreading points over the empty space, then by filling-in gaps, and finally by adding measurements throughout the space to efficiently improve resolution. This approach has several advantages. Data-taking proceeds by first generating a coarse, low-resolution view of the entire space, and then by progressively refining the image (with resolution improving as points are added) [42, 43]. This allows the experimenter to determine early on whether the selected sample region is appropriate, and adjust as necessary. The ultimate image resolution need not be prescribed ahead of time; simply allowing the autonomous mapping to proceed for longer will continually improve resolution (limited of course by the X-ray beam size). The experimenter can thus decide to end the experiment when sufficient resolution has been achieved. An even more powerful approach is to empower the autonomous algorithm to terminate the experiment, by providing it with a target model uncertainty to attain [42]. Since *gpCAM* has access to a continually improving model of the signal variation over the search space, it can rigorously determine when imaging resolution has reached the empirical spatial heterogeneity of the material being measured.

In landmark experiments at Brookhaven National Laboratory (BNL), the decision-making algorithm of *gpCAM* was deployed to control X-ray scattering imaging experiments [42, 43, 44]. These studies were performed through a collaboration between BNL's National Synchrotron Light Source II (NSLS-II), the Center for Functional Nanomaterials (CFN) and Berkeley Lab's Center for Advanced Mathematics for Energy Research Applications (CAMERA). Autonomous imaging was applied to a variety of heterogeneous materials problems, including polymer crystallization, nanoparticle film formation [42], electrospray deposition of polymer films [43], and additive manufacturing (3D printing). The same methodology can be applied to explore physical parameter spaces if one manufactures a sample library. Specialized synthesis methods enable the fabrication of one-dimensional or two-dimensional gradients of material properties, which thus enables creation of combinatorial sample libraries. For instance, one can use thermal gradients to generate continuous libraries of annealing temperature [36], flow-coating for thickness [55, 52, 11, 44] or composition [35] gradients, chemical gradients for libraries of surface chemistry [50, 6, 53, 44], or electrospray deposition to generate arbitrary patterns of film composition and thickness [57]. Gradient methods can frequently be combined to generate two-dimensional libraries where material parameters are varied continuously; the final sample thus represents an exhaustive slice through the high-dimensional space describing material synthesis/processing conditions. The outstanding challenge — to efficiently explore that slice — can be addressed using autonomous methods.

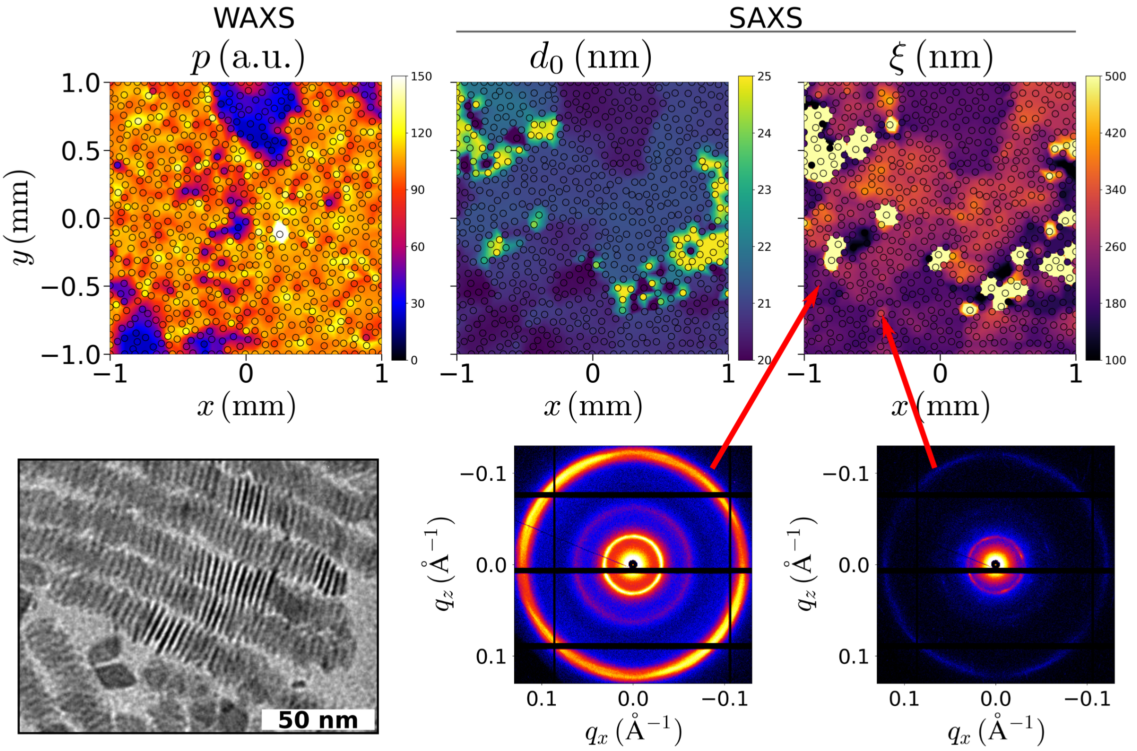


Figure 2: Autonomous mapping of a heterogeneous sample. A layer of nanoscale platelets self-assemble into stacks and superlattices (electron microscopy image in lower-left). Synchrotron X-ray scattering is measured as a function of position on the sample surface. Both small-angle (SAXS) and wide-angle (WAXS) X-ray scattering date are collected simultaneously. Multiple signals were extracted in real-time from collected data (top row), including scattering intensity of the WAXS peak associated with the nanoplate unit cell (p), the repeat-spacing of the nanoplate packing (d_0) and the grain size of the assembled domains (ξ). The combination of multiple signals simultaneously probe different aspects of heterogeneity. *gpCAM* selected the position of measurement points so as to minimize model uncertainty; the corresponding reconstruction provides information at multiple scales, and allows identification of small regions of distinct ordering that might otherwise be overlooked (example SAXS images provided in lower-right). As pointed out in Figure 1, while the point distribution appears similar to random, every single measurement is placed optimally to reduce uncertainty, and clustering is therefore naturally avoided.

As an example experiment benefiting from these methods, we studied the self-assembly of nanoscale platelets [14]. Self-assembled superlattices inevitably exhibit a variety of domains of differing ordering, size and orientation. Simple ensemble measurements yield incomplete information, averaging over the distribution of ordering motifs. Spatially resolving these distinct domains provides both quantification of the relative frequency of ordering states, as well as imaging of the size-scale and microstructure of this heterogeneity. The goal in such an experiment is to reconstruct the heterogeneity using a minimal number of measurements. This can be accomplished within the GP paradigm by defining an acquisition function that minimizes the total model uncertainty. In a typical experiment (Figure 2), both small-angle (SAXS) and wide-angle (WAXS) X-ray scattering data were collected at each position, using the Complex Materials Scattering (CMS, 11-BM) beamline at NSLS-II. The combination of SAXS and WAXS provides a wealth of information across scales. However, not all recorded signals need to be considered by the autonomous algorithm. Indeed, there is a strong advantage to dimensionally-reduce the raw data via automated analysis [27] down to a small set (e.g. 1-5) of independent meaningful signals. This deployment of the GP-based autonomous methodology has several advantages. Reconstruction of a low-error model occurs more rapidly than with simpler approaches (e.g. grid) since data is collected exactly where model uncertainty is higher [42]. Moreover, data is collected in a progressive mode, where one quickly obtains a coarse image of the entire space, which is continually and optimally refined as data-taking proceeds. For heterogeneous materials the required imaging resolution is often not known a priori. The presented autonomous approach avoids an arbitrary selection of resolution, instead allowing the experiment to be terminated (manually or automatically) when desired data-quality is achieved. By minimizing the model uncertainty (via Equation (12)), *gpCAM* efficiently explores the space, while the simultaneous data-analysis pipeline outputs maps of diverse signals (Figure 2, top row).

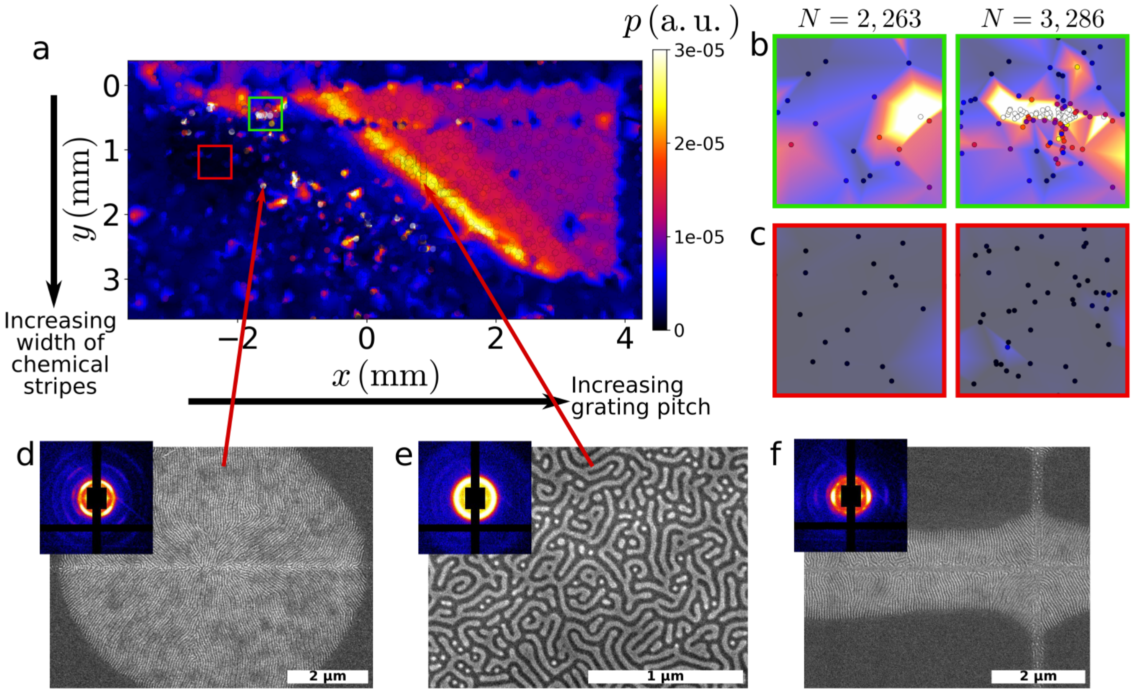


Figure 3: Autonomous mapping of a process parameter space. Small-angle X-ray scattering (SAXS) images are collected as a function of (x, y) position coordinates across a sample surface. The sample consists of an underlying chemical grating template that directs the ordering of a block copolymer film cast on top. The pitch of the guiding grating increases in the x-direction, while the width of the chemical stripes increases in the y-direction. The resultant map of block copolymer scattering intensity (at the q and angular position associated with lamellar-like order aligned with the grating), p, demonstrates substantial variation (a). Autonomous mapping was used to initially explore this process-space, and then to focus on regions with surprising behavior (bright spots). This targetted acquisition focuses data in areas of interest (green box, b) and away from unimportant areas (red box, c). This allowed subsequent scanning electron microscopy (SEM) investigation of select regions (d, e, f), which in turn enabled identification of manufacturing/processing defects and optimization of the fabrication protocol.

As an example of using *gpCAM* to explore a processing parameter space, we generated a two-dimensional combinatorial library to study the self-assembly of block copolymers [12]. In particular, we cast a thin film of a block copolymer blend [65] on top of a chemical pattern that can guide the ordering in the film [56]. We used electron-beam lithography to generate a dense array of chemical templates, where the grating pitch varies in one direction ($+x$) while the width of the chemical stripe increases in the orthogonal direction ($+y$). The combinatorial sample enables exhaustive study of this slice through the high-dimensional overall processing space. Using autonomous microbeam SAXS mapping (Figure 3), performed at the Soft Matter Interfaces (SMI, 12-ID) beamline at NSLS-II, we were then able to efficiently explore this space, and rapidly identify a host of interesting features. The *gpCAM* algorithm was first tasked with exploring the space, by defining an acquisition function that minimizes model uncertainty (via Equation (12)) for the scattering intensity signal (Figure 3a). This mode provides an efficient exploration, or survey, of the overall space, and allowed experimenters to identify several regions of unexpected ordering — localized regions with high scattering intensity and/or unexpected symmetry. In order to understand these features, the algorithm was retasked to focus on small regions exhibiting unusually high scattering intensity, by defining an acquisition function that maximizes the value of the selected signal, $f_a(x, y) = \phi m(x, y) + \sqrt{\sigma^2(x, y)}$ (see Equation (13)), where m and σ^2 are the predictive mean and variance of the signal (Equation (9) and (10)), respectively, and ϕ is a unitless weighting factor (see Equation (11) in Reference [43]). Importantly, the GP algorithm leverages the surrogate model (Equation (9)) obtained from the exploration phase in order to make high-quality predictions about where to find such regions, which accelerates the measurement sequence [43]. Indeed *gpCAM* efficiently localized these measurements in regions with high scattering intensity (Figure 3b) rather than wasting them in uninteresting regions (Figure 3c). Follow-up measurements using scanning electron microscopy (SEM) could then be conducted at the points of interest identified by autonomous X-ray mapping, including investigating unexpected four-fold symmetry (Figure 3d). This enabled identification of processing errors in sample fabrication, which in turn allowed the fabrication of new samples with optimized processing conditions (in other words, follow-up experiments were moved to another slice in the high-dimensional processing space). This autonomous workflow enabled efficient refinement of fabrication strategy, since it is able to rapidly identify small regions exhibiting unexpected ordering. Conventional fabrication based on select process conditions would (statistically) miss such regions, while grid scans are not guaranteed to converge/focus on regions of interest to the scientist. In this case, autonomous mapping enabled a dramatic qualitative improvement of the optimization workflow. Autonomous mapping methods are well-suited to cases where one is searching for rare features of interest in large and complex parameter spaces.

3.2 Towards Autonomous Synchrotron Infrared Mapping

Scanning synchrotron radiation-based Fourier transform Infrared absorption spectroscopy (SR-FTIR) is a powerful non-invasive microprobe technique that can provide spatially resolved chemical information about biogeochemical and environmental samples without *a priori* knowledge [25]. However, a major challenge in using the SR-FTIR technique for real-time characterization of time-dependent biochemical processes is the substantial spectral-image acquisition time. This poor time resolution is due to the high dimensionality of the generated dataset, which contains not only spatial but also spectral information. Furthermore, many samples, especially those from biological or biogeochemical systems, are often heterogeneous with many localized “hotspots” in chemical, biological, or physical properties with high spatial gradient.

Given the computational constraints associated with the high dimensionality in spectral space, current scanning SR-FTIR experiments are performed using the following strategy: Experimenters perform initial sample measurements at a number of points based on intuition. After acquiring these initial measurements, the experimenters identify spectral features of interest in the data. These spectral features are then used to identify candidate imaging locations. Presently, experimenters define a Cartesian grid over the imaging area with a pixel size ranging from 0.5 to 5.0 μm [22, 23, 32, 24, 19, 3, 4, 5, 67, 49, 58]. Each map is further limited by time constraints to fewer than a total of 2,500 pixels. The acquisition time for such a single spectral image can range from 6 to over 12 hours, depending on the signal to noise ratio. Unfortunately, this substantially limits the timescales of measuring the spatial distributions of transient biogeochemical processes. Furthermore, while the design of the data acquisition is chosen carefully based on the experimenter’s intuition and experience, it is unlikely that the experimenter is able to optimally perform pattern recognition and decision-making in the high-dimensional SR-FTIR spectral spaces; and similar to concerns raised in the X-ray scattering case study, using the uniform grid (UG) sampling method will lead to oversampling in information-poor regions while undersampling in information-rich (e.g. high gradient) regions. To address these challenges and advance SR-FTIR towards being an efficient and unbiased real-time spatiochemical imaging tool, we turn to autonomous grid-less data-acquisition approaches. We recently demonstrated the capabilities of a grid-less adaptive sampling approach that combines 2D barycentric linear interpolation with Voronoi tessellation (LIV) [21]. Based on a gradient-pursuing approach, LIV could substantially decrease imaging acquisition time while increasing sampling density in regions of steeper physico-chemical gradients without experimenter intervention.

To further expand algorithmic flexibility beyond the gradient-pursuit mode, we are turning our second-generation attempt at grid-less adaptive sampling to the more sophisticated and versatile *gpCAM* algorithm. We use *gpCAM*'s custom-defined acquisition functions (Section 2.3) to improve the SR-FTIR imaging speed and optimize sampling density in areas of interests. For this pilot study, we began with a well-characterized microbialite sample from Bacalar, a karst coastal oligosaline lagoon system (Figure 4 a (top) [58]). The microbialite sample lies upon a 205 μ meter by 140 μ meter grid where each grid point has an associated 1,738 dimensional spectra. Our goal in using GPR in the form of *gpCAM* is to reconstruct the high-dimensional mid-infrared spectral map accurately using as few measured sampling points as possible.

If cross-task correlations are considered, the scaling of GPR is $O(Outputs^3 * N^3)$ (compare to Section 2.4), where N is the number of measurements, and "Outputs" is the dimension of the measurement which corresponds to 1,738 in our sample. To circumvent this computational bottleneck, it is necessary to reduce the dimensionality of the dataset while still retaining the critical spectral features. There are at least ten popular approaches for data-dimensionality reduction. Here we started with the Principal Components Analysis (PCA) to compress our high-dimensional spectral data [10]. PCA is a dimensionality reduction technique that finds an orthonormal basis; typically retaining only the first few basis vectors preserves the majority of the variance of the dataset while substantially reducing data dimensionality.

We tested *gpCAM* on the microbialite sample in Figure 4 a (top), and followed the autonomous experiment loop paradigm as illustrated in Figure 1. The steps of the autonomous loop are as follows: First, high-dimensional spectra at a number of randomly-chosen initial (x, y) positions are collected, an example spectrum is shown in Figure 4 d (left). Second, we performed PCA on the dataset to reduce its dimensionality. Although any number of components (up to the number of sampled wavenumbers) can be chosen, in practice we found that the first three PCA components were necessary and sufficient for both modeling and feature extraction as they could cumulatively explain over 90% of the data variance. Projecting each of the collected spectra onto the three PCA components, as shown in Figure 4 b (top-right, bottom-left, bottom-right), reduced the dimensionality of our data from 1,738 to 3. Third, these three projection coefficients associated with each (x, y) are supplied to *gpCAM* in lieu of the entire spectrum. Fourth, using these three projection coefficients, *gpCAM* calculates the most likely coefficient values, the posterior mean m (Equation (9)), and their associated uncertainties, the posterior variance σ^2 (Equation (10)), across the (x, y) space. As mentioned in Section 3.1, reconstruction using a minimal number of measurements within the GPR framework can be obtained by defining an acquisition function that minimizes the total model uncertainty or posterior variance $f_a(\mathbf{x}) = \sigma^2(\mathbf{x})$ (see Equation 11 and Section 2.3). Fifth, this acquisition function is optimized and the location (x, y) with the greatest uncertainty is estimated by *gpCAM* and gets sampled next. This loop is repeated until a set number of points or the uncertainty defined is below a pre-specified tolerance value.

To measure the performance of the *gpCAM* algorithm using the $f_a(\mathbf{x}) = \sigma^2(\mathbf{x})$ acquisition function (Equation 11 and henceforth referred to as GP sampling) , we used the Euclidean distance between the ground truth and the interpolated spectra as a quantitative metric. We compared performances across three different sampling schemes: uniform grid, random and GP. The Euclidean distance is defined as $\sqrt{\sum_{i,j}(u_{i,j} - v_{i,j})^2}$ where the indices i, j reference (x, y) coordinates, $u_{i,j}$ is the ground truth spectrum at point i, j, and $v_{i,j}$ is the interpolated spectrum at the same position. The reconstructed spectrum at each (x, y) is obtained by performing the inverse PCA transform on the three PCA component projection coefficients corresponding to that position; if the spectrum at that position has not been measured yet, we use GPR, in the form of *gpCAM*, to interpolate the three PCA projection coefficients — *gpCAM* has the capability to interpolate in all three sampling schemes. The results are shown in Figure 4 c. We can see that using the $\sigma^2(\mathbf{x})$ acquisition function for sampling is superior to both random sampling and the commonly-used grid scanning in terms of the average reconstruction error at all points. The overarching reason is that grid scanning and random sampling do not make use of any information gained from previous measurements, while GP sampling continuously tries to sample in the region of greatest uncertainty given previously-measured data. Furthermore, GP sampling outperforms grid scanning because GP sampling is adept at finding the correct spatial length scales that are fixed in the grid scanning scheme. GP sampling also outperforms random sampling because the latter tends to form measurement-point clusters, whereas GP sampling is inherently self-avoiding.

To further illustrate the capabilities of *gpCAM*, we demonstrated how feature-finding can be implemented, see Figure 4 d. The reference spectrum at locations associated with the Si-Bonded organics that we wish to find is shown in Figure 4 d (left). To enable the feature-finding capability, we require the autonomous experiment to sample with a higher density where the targeted molecules are located. We reached this goal in *gpCAM* by defining a customized acquisition function that considers the correlation coefficient between the reference spectrum in Figure 4 d and the reconstructed spectrum in real time (see Section 2.3) via

$$f_a(\mathbf{x}) = \tilde{\sigma}(\mathbf{x}) \left(\frac{1}{2} + \frac{1}{2} \tanh \left[\beta \left(\mathrm{r}\left(u(\mathbf{x}), v(\mathbf{x})\right) - \alpha\right)\right]\right) \tag{16}$$

where β is a scaling factor, α is the correlation coefficient threshold, $u(\mathbf{x_{ref}})$ is the ground truth spectrum at the reference point, $v(\mathbf{x})$ is the corresponding reconstructed spectrum at $\mathbf{x}$, $r(\mathbf{x_{ref}}, \mathbf{x})$ is the correlation coefficient function between the reference spectrum and the reconstructed spectrum, and $\tilde{\sigma}(\mathbf{x})$ is the posterior standard deviation of the predicted spectrum at $\mathbf{x}$. This acquisition function identifies regions where the correlation coefficients between the reference spectrum and the reconstructed spectrum are above a pre-determined threshold value α (usually $>$ 0.8). The correlation coefficients in the identified regions are then non-linearly transformed using a *tanh* function and multiplied by the posterior uncertainty (Equation (10)) — the posterior uncertainty is included here to avoid sampling near positions that have already been sampled. Note that the reconstructed spectrum is an approximation of the ground truth spectrum, but the error of reconstruction has negligible impact on the feature finding as recognized in Figure 4 d. To initialize the feature-finding experiment, we first sampled 10 randomly chosen positions, followed by GP sampling using the $\sigma^2(\mathbf{x})$ acquisition function for an additional 50 sampling points. Then we used the GP sampling with feature-finding acquisition function for selecting the final 100 sampling points as shown in Figure 4 d (right). These 100 sampling points through feature finding process are localized around an area that matches the Si-Bonded organics "hotspot" in the ground truth map.

The results of our preliminary study demonstrated the potential of *gpCAM* to model the scanning SR-FTIR map of a complex biogeochemical sample with improved efficiency. *gpCAM* gave an estimate of the sample map with far fewer sampling points than the conventional grid scanning method. In addition, with the help of a reference spectrum, *gpCAM* was able to concentrate sampling points around a particular region of interest where the experimenter may desire to have high sampling density. We believe these smart and autonomous sampling strategies offered by *gpCAM* are powerful tools at experimenters' disposal to get the most scientific value out of their precious beam time.

3.3 Towards Autonomous Discovery of Correlated Electron Phases

Angle-resolved photoemission spectroscopy (ARPES) is one of the fundamental probes of the propagation of charges in materials. The ARPES process consists of the emission of high-energy electrons ("photoelectrons") induced when a source of ultraviolet or X-ray photons illuminates a material. These photoelectrons are emitted in a fixed set of directions, dependent upon their kinetic energy. As such, there is a fundamental resemblance between photoelectron emission patterns and X-ray scattering and diffraction patterns discussed in Section 3.1. Like diffraction patterns, the photoelectrons are emitted in a symmetrical way that is closely related to the symmetry of their atomic lattices.

But unlike X-rays, which have a simple relationship between emission angle and X-ray energy, the dispersion relationship between electronic energy and direction (which is mapped in a simple way to electron momentum) is much more complex and provides an interesting and distinct "fingerprint" of the material. Figure 5a shows a representative example of the momentum-dependent photoelectron spectrum of graphene, and two-dimensional honeycomb lattice of carbon atoms that is the basis of many new proposed new electronic schemes. These patterns are indicative of many material properties such as optical appearance, magnetic structure, and electronic mobility. For graphene, the electrons have only the freedom to move in a plane, so that maps like Figure 5a can be a complete representation of the dispersion relations. Other materials can have an additional momentum degree of freedom in the third direction, so that, in general, the electronic dispersion maps are sampled in a hypercubic subset of the full 4-dimensional space.

Improved efficiency of traversing the two-dimensional parameter space to identify new electronic phases, and to correlate these phases with material composition, structure, performance, and sample morphology is greatly desired. We have taken the approach to use *gpCAM* to more efficiently identify the distinct electronic phases wherever they occur in this hyper-phase space as a first positive step.

Although autonomous identification of electronic phases was not yet achieved, an autonomous experiment was conducted as a demonstration of the technique's potential. Experiments were performed on 4D spatially-mapped ARPES datasets [34] in which 2D ARPES maps were acquired in a fine grid mesh across inhomogenous materials. The data could then be sampled as though the measurement was conducted point-by-point under autonomous control. The convergence of the autonomous data towards the ground truth was evaluated by a simple metric, the mean absolute percentage error, which was evaluated for an interpolation of the partial data chosen by *gpCAM*.

As discussed in Section 2, the reduction of the data dimensionality is essential for *gpCAM* to be efficient. Traditionally, we might intuit a particular discrete feature whose intensity or position could be extracted as the interpolated quantity explored by *gpCAM*. But this would introduce a significant bias into the measurement because we do not generally know which features in the high-dimensional space are important. Therefore, we adopted an *a priori* approach, which was to use a clustering algorithm to assign each new data point a cluster number, and let this number be interpolated across the sample by the Gaussian process [34]. K-means clustering, an unsupervised machine learning method, was found to be the most effective way to cluster the data. Compared to identification of components through PCA and NMF (compare to Section 2.4), it offers the major advantage that the returned data clusters are all physically observable spectra. Component identification through PCA and NMF can lead to a basis set of spectra that while mathematically well-defined representations of the spectral components in the dataset, are not necessarily characteristic of an experimentally realizable spectrum. The disadvantage of clustering, though, is that it assumes that the probe size is smaller than any compositional variation in the sample. If measuring with a finite size probe astride the sharp boundary between phases, PCA and NMF will correctly identify the resulting spectrum as a linear combination of components on either side of the boundary, while clustering will identify the boundary spectrum as a new cluster member. Therefore, clustering is more suitable for phase diagrams with large homogeneous regions, separated by sharp boundaries. If the boundaries become to diffuse then clustering will return too many components than are present in the material. A further, and essential requirement of an *ab initio* approach when using clustering in concert with a Gaussian process is a method to autonomously choose the number of clusters present; one option is to use the silhouette score [34].

Results for one sample are shown in Figure 5 for a "twisted" stack of two graphene layers studied at the MAESTRO beamline of the Advanced Light Source. The lower layer is a continuous, homogeneous graphene sheet. The upper layer is also graphene, but it is a sample broken up into distinct regions, each with a different azimuthal rotation, separated by sharp boundaries. The spectrum at any given point resembles at first glance the incoherent sum of two data cubes of Figure 5a, in which one is azimuthally rotated by some random angle θ. It was recently observed that for certain "magic" values of θ, the material unexpectedly becomes superconducting. The aim of the experiment is to survey the sample to find domains near the magic angle as efficiently as possible. The physical interest lies in the fine spectral details observable where the energy bands cross, that correlate to the observed superconductivity.

The results of the autonomous experiment are shown in Figure 5 [34]. The primary findings are that the algorithm successfully and efficiently returns the number of clusters, and that for this and other datasets studied, the *gpCAM* with K-means clustering outperforms a random search and significantly outperforms regular grid-based searches. Overall, we found that often $< 10\%$ of the dataset was adequate to acquire an accurate description of the number of phases present together with their boundaries (Table 1).

The critical question of when to stop collecting data depends very much on the particular experiment's goals, the total time allocation on the apparatus, as well as cost functions such as facility hourly charges, time to move motors, cost of fabricating the sample, etc. We can attempt to define a stopping point for a few use cases. (i) When one is searching for a known target phase, one can stop when the phase is found. (ii) If the target is more general, i.e. the experimenter wants to know "complete" phase distribution in a two-dimensional subspace, exemplified here, the stopping point is not well-defined until every pixel of size equal to the probe beam is sampled — at which point one might as well use a grid search — since new phases could exist at any unsampled pixel. Under cost constraints, the stopping point will clearly depend on the cost functions mentioned in Section 2.3, which are experiment-dependent. In many cases the costs are low, and getting "complete" information in a two-dimensional subspace, as exemplified here, will not be prohibitive and the experimenter will tend to oversample the subspace at modest cost. The advantage of GP-driven autonomous control is that maximal available information is obtained whatever the stopping point. (iii) As additional scan variables are added (such as sample temperature, or fabrication degrees of freedom), the dimensionality of the scanned subspace increases correspondingly, making "complete" information less likely to be achievable in practice. The experiment will typically run until the total resources allocated to the project are expended or until a certain model certainty is reached. *gpCAM* can guarantee that the maximum information has been obtained.

The results point to challenges to improve the efficiency further. Under a critical eye, the efficiency of *gpCAM* over a random search can be modest. The *gpCAM* algorithm employed was customized for exploration — using the acquisition function defined in Equation (12) — not optimization which is one of its strengths. After the initial survey, the domain closest to the magic angle (teal color in Figure 5) could be identified, and similarity to this spectrum could then be chosen as the interpolant for a GP-based optimization across larger regions of the sample. This is a short-term vision of a far more efficient combination of scans than can be presently performed. In the longer term, we expect to achieve more informative results by defining and using more advanced acquisition functions. In addition, the classification method used was not very sensitive to subtle changes in spectral signature. Better classification methods, perhaps based on deep-learning (see Section 2.4), could reveal a larger number of components, with assurance that the difference between found components are meaningful.

3.4 Autonomous Measurements of Magnons Through Inelastic Neutron Scattering

Inelastic neutron scattering (INS) — via classical triple-axis spectrometers — explores, in a point-by-point manner, the shape and intensity distribution of the so-called scattering function $S(\mathbf{Q}, E)$, a hyper-surface over a four-dimensional parameter space, spanned by the three coordinates of momentum transfer $\mathbf{Q}$ and the energy transfer E of the neutron. The scattering function is the Fourier transform in space and time of the time-dependent pair-correlation function, which correlates the position (and/or spin state) of particles at different times [54]. Inelastic neutron scattering is thus an important tool in solid state physics to probe the dynamics of correlated systems on atomic length scales and picoseconds time scales. The scattering probability, or cross-section, between neutrons and individual atoms is relatively weak which allows neutrons to deeply penetrate into the materials, but also induces counting times of the order of minutes for every chosen coordinate $(\mathbf{Q}, E)$. The ultimate aim for steering autonomous data acquisition on a three-axis spectrometer is therefore a good approximation of the scattering function with a minimum amount of measured data points.

Typically, the measuring strategy of physicists closely resembles grid-scanning, choosing trajectories along E or along reciprocal high-symmetry directions, moving iteratively through the scattering function. The number of measurements is a limiting factor, even with prior knowledge on the scattering systems, such as crystal symmetry, magnetic ordering or relevant energy scales.

For the work presented here, we tested *gpCAM* for exploring the reciprocal space of a magnetically ordered system, without providing any further prior physical knowledge in the form of symmetries to the algorithm. As a test sample, we chose a rare-earth-containing compound with an intense, dispersive-magnon branch around 3.5 to 4.5 meV energy transfer and a weak non-dispersive doublet around 2 meV, typical for a crystal field excitation (see Figure 6 d). We instructed the acquisition function, Equation (11), to favor high function values and avoid clustering via

$$f_a(\mathbf{x}) = \sigma(\mathbf{x}) + 3\, m(\mathbf{x})\, \sigma(\mathbf{x}). \tag{17}$$

We also added a cost function, in order to put a penalty on time expensive motor movements (see Section 2.3). The positioning along the energy-transfer axis in Figure 6 implies slow movements of the whole instrument, while the horizontal axis (**Q**=(h,0,0)) only involves fast motors, which rotate the crystal and the axis defining the scattering angle. For evaluating the total time between the starting and end points in reciprocal space, we interfaced *gpCAM* with a module of *Takin* [61, 62], a software that simulates instrument movements. As kernel function, we used an anisotropic Matérn kernel (Equation (3)) of first order differentiability (ν=3/2). The hyperparameters after 850 iterations were σ=86.6, l_E=0.005 and l_h=0.03. *gpCAM* finds a length scale l_h of the order of the instrumental resolution, while l_E is one order of magnitude lower, which is related to the inherent anisotropy of the model function.

For getting an estimate for the efficiency of *gpCAM* under the conditions presented above, we simply compared the number of points close to the relevant $S(\mathbf{Q}, E)$ regions ('signals' within regions of interest ROIs) with the number of measured background points, and bench-marked the result with a random-point selection and a grid-scanning mode over the same regions in the parameter space.

After initialization of some randomly-selected points, *gpCAM* shows a hit rate inside the ROIs about twice as high compared to random and grid-based measurement selection (see Table 1). Beyond about 500 points, the data density for all methods becomes the controlling factor and the differences reduce — for comparison, a grid-based method over this area with satisfying instrumental resolution necessitates about 1000 measurement points. A second important outcome was the optimization of the motor moving time, as illustrated in Figure 6. We also emphasize that the algorithm could be easily connected to the instrument control system and reliably steered the instrument movements without any intervention from day one of the implementation. While this experiment has already shown a gain in efficiency for exploring a parameter space without any prior knowledge on the physical system (except the physics-informed acquisition function), we currently foresee a much higher potential by combining *gpCAM* with physical models as outlined in Section 2.6.

4 Discussion and a Look into the Future of GP-driven Autonomous Data Acquisition

In this paper, we have demonstrated that Gaussian-process(GP)-based autonomous data acquisition can lead to the effective and efficient acquisition of high-value datasets at large experimental facilities — namely X-ray and Neutron scattering facilities — with minimal or even without any human intervention. We introduced the basic notion of autonomous data acquisition, Gaussian processes, and the techniques that are needed for data analysis, dimensionality reduction, and function optimization, and presented four examples of autonomously steered experiments at large experimental user facilities. By choosing case studies from a wide variety of experimental techniques we have shown that our framework based on GPs is agnostic to the details of the experiment. It can therefore be applied whenever a model function, defined over some parameter space, is to be explored in search of new scientific discoveries. The presented software tool is readily available for download [37], or can be installed by typing “pip install gpcam” in a terminal. More information can be found on the *gpCAM* website [38].

The presented case studies span across different situations, both in the dimensionality of the input space and the resulting data structure. For high-dimensional measurement outputs, we have shown how dimensionality-reduction techniques are used to enable an efficient GP for steering. The case studies have shown that the main benefits of the GP-driven autonomous data acquisition are: (1) Autonomy, scientists are not required to spend valuable time micro-managing the sequence of measurements. This allows more time for big-picture decisions and interpretations of the model. (2) Only high-value data is collected which saves staff and instrument time, and storage and computing capabilities. By using cost functions, the focus can easily be changed to optimizing other quantities, such as the use of expensive materials, or ray or particle exposure. (3) The experimenter always has access to a complete model that can be used for interpretation and termination of the experiment when a scientific conclusion is reached. This is in contrast to grid-scan methods that deliver only a partial result if interrupted or queried before successful termination. (4) Noise is naturally included in the GP framework when it comes to uncertainty and model prediction. This also means that the final model is the most likely model given all the data including measurement variances. Exact interpolation techniques (e.g., linear, cubic) will lead to artifacts when used for interpolating noisy data. (5) No training data is required from previous experiments. The model is based exclusively on data from the current experiment and potentially carefully-chosen domain knowledge; however, if desired, the domain knowledge can depend on other datasets. This is especially important when compared to techniques like reinforcement learning where training data from other experiments are needed but it is unclear how the training data affects the steering and the final result. No dependency on previous data also means that bias is avoided. (6) Transparency and interpretability; while the details of a GP are complicated, the important values (e.g., hyperparameters, surrogate models, uncertainties, kernels) are always accessible for interpretation and supervision by the experimenter. This is again in contrast to many other techniques that are less understood and often criticized for a lack of interpretability. (7) Uncertainty quantification is naturally included in the GP framework which is important for successful steering and interpretation of the final results. (8) There are several ways of including domain knowledge into a statistical framework such as a GP. And finally, as a result of all mentioned benefits follows (9), the GPs generality and flexibility; the framework makes no assumptions about the nature of the parameters defining the parameter space or the signals. This allows it to be applied to any definable search for any problem-space, including multi-modal experiments that combine techniques, or hybrid experimental/simulation approaches.

A major objective for autonomous data acquisition in the future is to expand on domain awareness. As mentioned in Section 2.6, for GP-driven autonomous data acquisition we see three major paths to achieve domain awareness that can be summarized as: advanced kernel designs, constraints on the marginal log-likelihood optimization and acquisition-function design. We have shown how one class of domain awareness was used in practice, namely the customized acquisition function. In the case of infrared spectroscopy, what was known was a reference spectrum that can be used to define the acquisition function and target the measurement process (Section 3.2). Other variations of the acquisition function were used in some of the other case studies we presented. While option three, the informative acquisition function, is inexpensive and potentially very beneficial to efficient autonomous data acquisition, the advanced kernel designs and the constrained likelihood optimization will have a major impact in the future of the field. Both techniques require a large number of hyperparameters that, contrary to standard GPs, scales with the dimensionality of the input space which, due to increasingly complex scientific questions, rises rapidly. Therefore, the GP-driven autonomous data acquisition will increasingly become a high-performance-computing optimization problem. This motivates the authors to focus on the development and employment of efficient and HPC-ready optimization tools. Another focus area are the constraints that can be communicated by advanced kernel designs.

Full domain awareness and HPC readiness of Gaussian-process driven autonomous data acquisition will lead to the acceleration of scientific discovery in biological, chemical, physical and materials sciences.

Acknowledgments

The work was partly funded through the Center for Advanced Mathematics for Energy Research Applications (CAMERA), which is jointly funded by the Advanced Scientific Computing Research (ASCR) and Basic Energy Sciences (BES) within the Department of Energy's Office of Science, as well as by Laboratory Directed Research and Development Program of Lawrence Berkeley National Laboratory, under U. S. Department of Energy Contract No. DE-AC02-05CH11231. This research used resources of the Center for Functional Nanomaterials and the National Synchrotron Light Source II, which are U.S. DOE Office of Science Facilities, at Brookhaven National Laboratory under Contract No. DE-SC0012704. This research also used resources of the Berkeley Synchrotron Infrared Structural Biology (BSISB) Imaging Program funded by the U.S. Department of Energy, Office of Biological and Environmental Research under Contract No. DE-AC02-05CH11231. The Advanced Light Source is supported by the Director, the Office of Science, and the Office of Basic Energy Sciences. Both were supported through Contract DE-AC02-05CH11231. K.C.E. and C.B.M. acknowledge support from the Office of Naval Research Multidisciplinary University Research Initiative Award ONR N00014-18-1-2497. K.C.E. acknowledges support from the NSF Graduate Research Fellowship Program under Grant No. DGE-1321851. This work is based on experiments performed at the Institut Laue-Langevin (ILL) in Grenoble, France, the collected datasets have the DOIs 10.5291/ILL-DATA.TEST-3123 and in parts 10.5291/ILL-DATA.4-01-1643. We wish to thank E. Villard, P. Chevalier and J. Locatelli for technical support at *Thales*.

Author Contributions Statement

M.M.N. wrote the initial drafts of Sections 1 and 2, devised the used algorithm, formulated the required mathematics, and implemented the computer codes (*gpCAM*). P.H.Z. designed, coordinated and collaborated on the development of basic computational strategies in gpCAM and on its use in SFTIR microscopy and ARPES experiments and took part in writing and editing this manuscript. D.U. designed, configured, and implemented codes associated to convnets for reverse image search and wrote the related section. M.F. and K.G.Y. planned, supervised, and coordinated experiments at NSLS-II, and wrote the related section. M.F., K.G.Y., E.H.R.T., R.L., G.F. and M.Z. performed X-ray scattering experiments at NSLS-II, including beamline operation and data analytics. K.C.E. and C.B.M. prepared nanoplatelet materials. A.S. and G.S.D. prepared chemical templates and self-assembled films. E. R. planned and led the ARPES measurements at ALS, and wrote the related section. C. N. M. (author of [34]) performed the K-Means cluster based GP collection simulations. H.Y.H. led the SR-FTIR measurements, coordinated the simulations and wrote the initial draft of the related section, S.L. designed and performed the PCA based GP collection simulations and wrote the related section. L.C. designed the simulations and wrote the related section. Y.L.G. and T.W. customized *gpCAM* for use at *Thales*. T.W. developed and performed preparatory simulations with *gpCAM* using theoretical dynamical structure factor models for neutron scattering. T.W. planned and T.W., M.B., P.S., and P.M. performed the first autonomous commissioning experiment at *Thales* measuring the magnons in the chiral magnet MnSi. M.B. proposed and M.B., T.W., P.S., and P.M. performed the second autonomous commissioning experiment at *Thales* whose results are shown in Figure 6. The sample for the first experiment (MnSi) was provided by A. Bauer, the sample for the second autonomous commissioning experiment was provided by M.B. T.W. analyzed the data of the first experiment (MnSi, not shown), M.B. analyzed the data of the second experiment (Figure 6). M.B. and T.W. wrote the text of the corresponding section 3.4 to equal parts. J.A.S. supervised the development of the mathematics and the implementation of the code, and revised and improved the manuscript significantly.

All authors commented on the manuscript and revised it repeatedly.

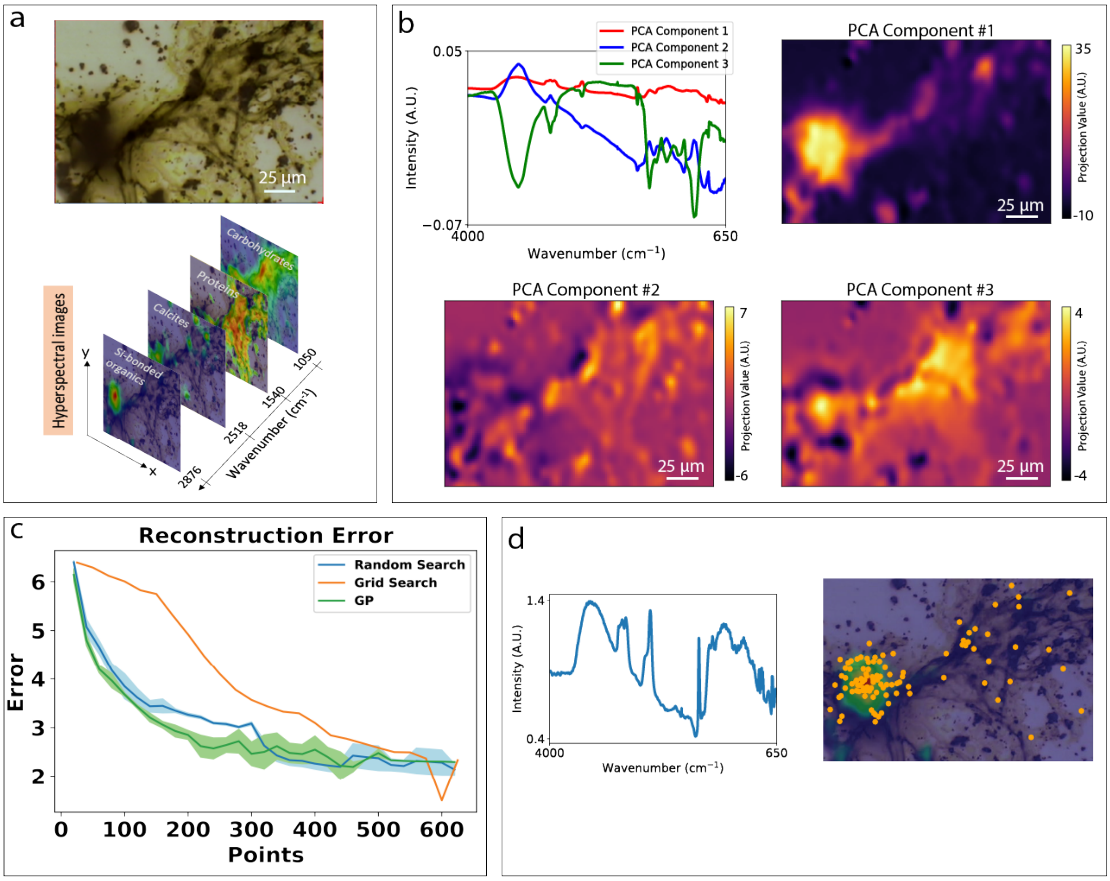
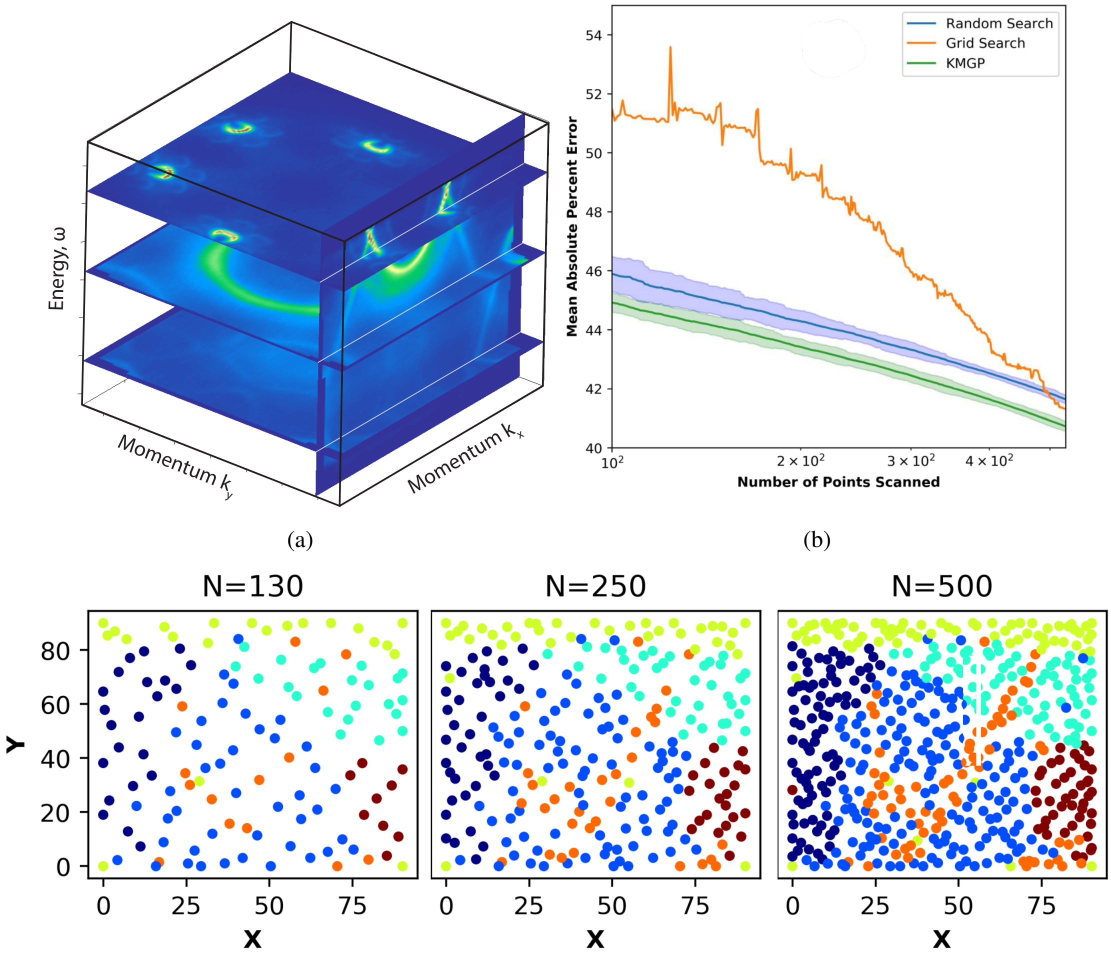


Figure 4: Key elements for a successful application of *gpCAM* to infrared mapping. (a) Bacalar lagoon microbialites used in the pilot study. A bright field micrograph (205 by 140 μmeters) showing a Bacalar lagoon microbialite used in the autonomous mapping pilot study (top). Distribution heat maps of the Si-bonded organics, calcites, proteins and carbohydrate extracted from a 1,738-spectral-dimensions dataset (bottom) (from [58]). (b) Dimensionality reduction of the microbialite sample spectra and the subsequent learned models. SR-FTIR spectra were reduced from 1,738 dimensions to thee principal components while retaining 90% of the chemical information (top-left). Heat maps of spectral projections onto the three PCA basis vectors: PCA1 (top-right), PCA2 (bottom left), and PCA3 (bottom-right), providing important information regarding the location of edges, the sample's microstructure and localized chemistry in the infrared domain as seen in 4 a (bottom). (c) Comparative reconstruction error for different sampling schemes and using *gpCAM*. The rather small difference between random sampling and GP-driven sampling stems from the fact that the interpolation was done by a GP in both cases. (d) Feature-finding in the microbialite sample using a known reference spectrum associated with the Si-Bonded Organics (left) and rendered feature-finding sampling locations (orange dots). The measurement locations mainly focus on the Si-Bonded organics "hotspot" in the ground truth map. In this case, the region of interest was identified and sufficiently-well resolved after 200 measurements, compared to circa 10000 measurements needed for a common full grid scan, leading to an estimated 50-fold improvement (see Table 1).



(a)

(b)

(c)

Figure 5: Autonomous identification of sample phases, and convergence towards the ground-truth for a surrogate sample measurement with ground truth spectrum consisting of 8000 points acquired on a uniform mesh. (a) ARPES measurement of electronic states in graphene, the single layer of carbon atoms that is the building block for many allotropes of carbon such as graphite, nanotubes, and buckeyballs [8]. Shown are planar slices through the acquired volume dataset. The bright features seen in the figure are the allowed energy bands of the material. These allowed energy bands are acquired from eigenvalues of the Schrödinger equation. A typical ARPES measurement is a 2D, 3D, or even 4D image that is a subset of the available 4D energy-momentum space (from Reference [34]). (b) The convergence of the interpolated dataset's mean absolute percent error (deviation from the ground truth data) for GP with K-means clustering (KMGP), a random search, and uniform-grid searches with progressively finer sampling. Keep in mind, while the distribution of points might look random for the eye, every measurement point is carefully chosen to decrease the uncertainty of the model (from Reference [34]). The reason for the seemingly small improvement of the GP sampling scheme compared to the random search and the grid search is the fact that the data for all schemes was interpolated using the GP. Also, the objective here was pure exploration which leads to a homogeneous point distribution. Most importantly, the interesting metric for comparison is the number of measurement points to achieve a certain model quality, not the difference in the uncertainty. (c) Spatial maps as a function of number of points N collected (dimensions are in microns) acquired using nanoARPES at the Advanced Light Source. Each data point is identified by the cluster number (by color), where the algorithm chose 6 clusters total. At $N = 130$ all clusters had been successfully identified. See [34] for more details.

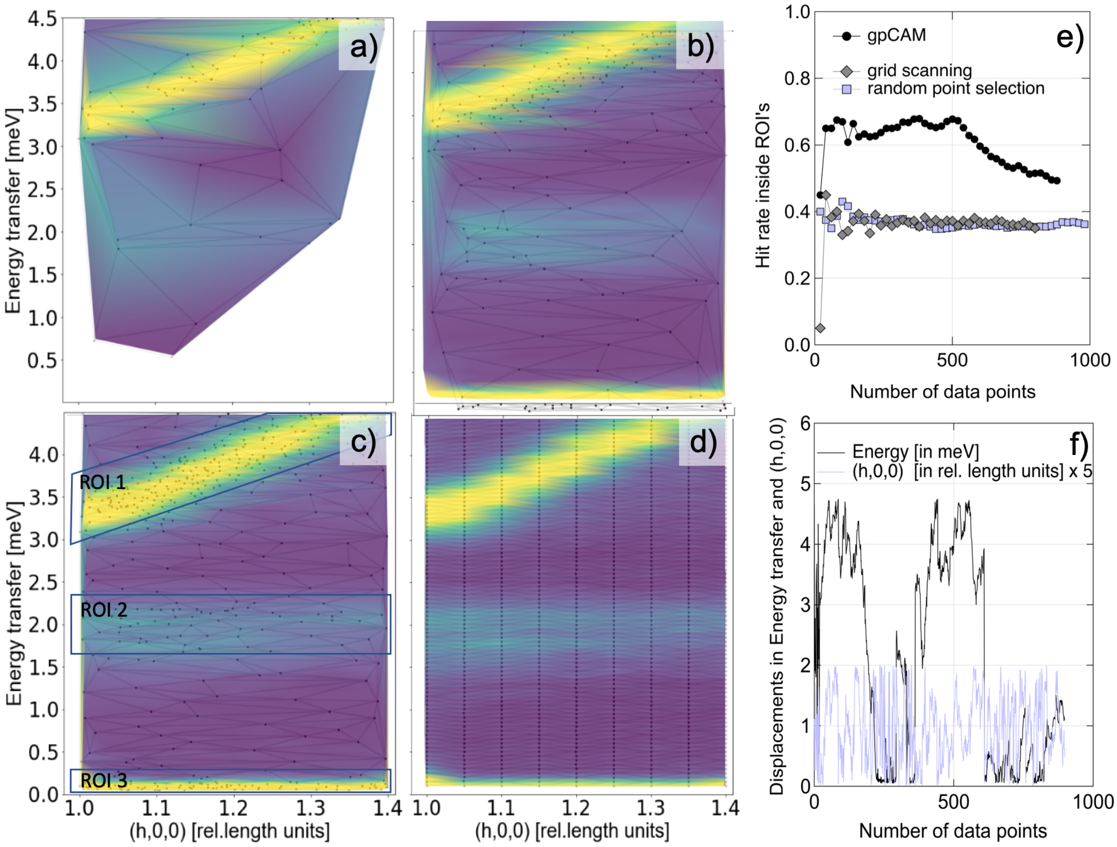


Figure 6: Commissioning run of *gpCAM* at the triple-axis spectrometer *Thales* [7] at the Institut Laue-Langevin (ILL) in Grenoble, France. *gpCAM* iteratively probed magnetic excitations at energies of approximately 2 meV (ROI2) and 3.5-4.5 meV (ROI1). ROI3 covers the so-called elastic line without information on the magnetic system. Shown are the experimental results after performing (a) 100, (b) 250 and (c) 500 measurements. d) Experimental results from the same sample on ThALES obtained with grid scanning (800 points in total). In the graphs a) to d), data points have been connected using Delaunay triangulation. e) Comparison of the hit-rate inside the region of interests (ROI's), defined in (c), for the three different methods: *gpCAM* (black circles), grid scanning (grey diamonds) and random scanning (purple squares). (f) Comparison of instrument displacements either along the energy transfer or the momentum transfer direction (h,0,0). The cost function puts a penalty on the energy transfer which inhibits frequent jumps over longer energy transfers.

REFERENCES

[1] Hossein Aghamohammadzadeh, Richard H. Newton, and Keith M. Meek. X-ray scattering used to map the preferred collagen orientation in the human cornea and limbus. *Structure*, 12:249–256, 2004. doi: 10.1016/j.str.2004.01.002.

[2] Araujo, Silva, Medeiros, Parkinson, Hexemer, Carneiro, and Ushizima. Reverse image search for scientific data within and beyond the visible spectrum. *Expert Systems with Applications*, 109:35–48, Nov 2018.

[3] Jacob Bælum, Sharon Borglin, Romy Chakraborty, Julian L Fortney, Regina Lamendella, Olivia U Mason, Manfred Auer, Marcin Zemla, Markus Bill, Mark E Conrad, et al. Deep-sea bacteria enriched by oil and dispersant from the deepwater horizon spill. *Environmental microbiology*, 14(9):2405–2416, 2012.

[4] Liane G Benning, VR Phoenix, Nathan Yee, and KO Konhauser. The dynamics of cyanobacterial silicification: an infrared micro-spectroscopic investigation. *Geochimica et Cosmochimica Acta*, 68 (4):743–757, 2004.

[5] Liane G Benning, VR Phoenix, Nathan Yee, and MJ Tobin. Molecular characterization of cyanobacterial silicification using synchrotron infrared micro-spectroscopy. *Geochimica et Cosmochimica Acta*, 68(4):729–741, 2004.

[6] Brian C. Berry, Christopher M. Stafford, Mayur Pandya, Leah A. Lucas, Alamgir Karim, and Michael J. Fasolka. Versatile platform for creating gradient combinatorial libraries via modulated light exposure. *Review of Scientific Instruments*, 78:072202, 2007. doi: 10.1063/1.2755729.

[7] M. Böhm, P. Steffens, J. Kulda, M. Klicpera, S. Roux, P. Courtois, P. Svoboda, J. Saroun, and V. Sechovsky. ThALES – Three Axis Low Energy Spectroscopy for highly correlated electron systems. *Neutron News*, 26(3):18–21, 2015. doi: 10.1080/10448632.2015.1057050.

[8] A. Bostwick, T. Ohta, J.L. L McChesney, T. Seyller, K. Horn, and E. Rotenberg. Band structure and many body effects in graphene. *The European Physical Journal Special Topics*, 148(1):5–13, sep 2007. ISSN 1951-6355. doi: 10.1140/epjst/e2007-00220-x.

[9] Samuel Cohen, Rendani Mbuvha, Tshilidzi Marwala, and Marc Peter Deisenroth. Healing products of gaussian process experts.

[10] T Davies and Tom Fearn. Back to basics: the principles of principal component analysis. *Spectroscopy Europe*, 16(6):20, 2004.

[11] Raleigh L. Davis, Sahana Jayaraman, Paul M. Chaikin, and Richard A. Register. Creating controlled thickness gradients in polymer thin films via flowcoating. *Langmuir*, 30:5637–5644, 2014. doi: 10.1021/la501247x.

[12] Gregory S. Doerk and Kevin G. Yager. Beyond native block copolymer morphologies. *Molecular Systems Design & Engineering*, 2:518–538, 2017. doi: 10.1039/c7me00069c.

[13] P. Dubcek. Nanostructures as seen by the saxs. *Vacuum*, 80:92–97, 2005. doi: 10.1016/j.vacuum.2005.07.045.

[14] Katherine C. Elbert, Thi Vo, Nadia M. Krook, William Zygmunt, Jungmi Park, Kevin G. Yager, Russell J. Composto, Sharon C. Glotzer, and Christopher B. Murray. Dendrimer ligand directed nanoplate assembly. *ACS Nano*, 13:14241–14251, 2019. doi: 10.1021/acsnano.9b07348.

[15] Ronald A Fisher. The arrangement of field experiments. In *Breakthroughs in statistics*, pages 82–91. Springer, 1992.

[16] Peter Fratzl. Small-angle scattering in materials science - a short review of applications in alloys, ceramics and composite materials. *Journal of Applied Crystallography*, 36:397–404, 2003. doi: 10.1107/S0021889803000335.

[17] Peter I Frazier. A tutorial on bayesian optimization. *arXiv preprint arXiv:1807.02811*, 2018.

[18] Yinghua Gao, Naiqi Li, Ning Ding, Yiming Li, Tao Dai, and Shu-Tao Xia. Generalized local aggregation for large scale gaussian process regression. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2020.

[19] Terry C Hazen, Eric A Dubinsky, Todd Z DeSantis, Gary L Andersen, Yvette M Piceno, Navjeet Singh, Janet K Jansson, Alexander Probst, Sharon E Borglin, Julian L Fortney, et al. Deep-sea oil plume enriches indigenous oil-degrading bacteria. *Science*, 330(6001):204–208, 2010.

[20] Arthur Hobson and Bin-Kang Cheng. A comparison of the shannon and kullback information measures. *Journal of Statistical Physics*, 7(4):301–310, 1973.

[21] E Holman and M Fang. Autonomous adaptive data acquistion for scanning hyperspectral imaging. *Communication Biology*, 2020. (in press).

[22] Hoi-Ying N Holman, Dale L Perry, Michael C Martin, Geraldine M Lamble, Wayne R McKinney, and Jennie C Hunter-Cevera. Real-time characterization of biogeochemical reduction of cr (vi) on basalt surfaces by sr-ftir imaging. *Geomicrobiology Journal*, 16(4):307–324, 1999.

[23] Hoi-Ying N Holman, Karl Nieman, Darwin L Sorensen, Charles D Miller, Michael C Martin, Thomas Borch, Wayne R McKinney, and Ronald C Sims. Catalysis of pah biodegradation by humic acid shown in synchrotron infrared studies. *Environmental science & technology*, 36(6):1276–1280, 2002.

[24] Hoi-Ying N Holman, Eleanor Wozei, Zhang Lin, Luis R Comolli, David A Ball, Sharon Borglin, Matthew W Fields, Terry C Hazen, and Kenneth H Downing. Real-time molecular monitoring of chemical environment in obligate anaerobes during oxygen adaptive response. *Proceedings of the National Academy of Sciences*, 106(31):12599–12604, 2009.

[25] Hoi-Ying N Holman, Hans A Bechtel, Zhao Hao, and Michael C Martin. Synchrotron ir spectromicroscopy: chemistry of living cells, 2010.

[26] Anita Krishnakumar. Active learning literature survey. *Tech. rep., Technical reports, University of California, Santa Cruz.*, 42, 2007.

[27] Brookhaven National Laboratory. Scianalysis. `https://github.com/CFN-softbio/SciAnalysis`, 2015.

[28] Jiliang Liu, Isabel Costantino, Nagarajan Venugopalan, Robert F. FIschetti, Bradley T. Hyman, Matthew P. Frosch, Teresa Gomez-Isla, and Lee Makowski. Amyloid structure exhibits polymorphism on multiple length scales in human brain tissue. *Scientific Reports*, 6:33079, 2016. doi: 10.1038/srep33079.

[29] Jiliang Liu, Jeong I. Kim, Joanne C. Cusumano, Clint Chapple, Nagarajan Venugopalan, Robert F. FIschetti, and Lee Makowski. The impact of alterations in lignin deposition on cellulose organization of the plant cell wall. *Biotechnology for Biofuels*, 9:126, 2016. doi: 10.1186/s13068-016-0540-z.

[30] Shuai Liu, Jie Li, Kochise C. Bennett, Brad Ganoe, Tim Stauch, Martin Head-Gordon, Alexander Hexemer, Daniela Ushizima, and Teresa Head-Gordon. Multiresolution 3d-densenet for chemical shift prediction in nmr crystallography. *The Journal of Physical Chemistry Letters*, 10(16):4558–4565, 2019. doi: 10.1021/acs.jpclett.9b01570. PMID: 31305081.

[31] Shuai Liu, Charles N Melton, Singanallur Venkatakrishnan, Ronald J Pandolfi, Guillaume Freychet, Dinesh Kumar, Haoran Tang, Alexander Hexemer, and Daniela M Ushizima. Convolutional neural networks for grazing incidence x-ray scattering patterns: thin film structure identification. *MRS Communications*, pages 1–7, 2019.

[32] Olivia U Mason, Terry C Hazen, Sharon Borglin, Patrick SG Chain, Eric A Dubinsky, Julian L Fortney, James Han, Hoi-Ying N Holman, Jenni Hultman, Regina Lamendella, et al. Metagenome, metatranscriptome and single-cell sequencing reveal microbial response to deepwater horizon oil spill. *The ISME journal*, 6(9):1715–1727, 2012.

[33] Michael D McKay, Richard J Beckman, and William J Conover. Comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics*, 21 (2):239–245, 1979.

[34] Charles Nathan Melton, Marcus Noack, Taisuke Ohta, Thomas E Beechem, Jeremy Robinson, Xiaotian Zhang, Aaron Bostwich, Chris Jozwiak, Roland J Koch, Petrus Zwart, et al. K-means-driven gaussian process data collection for angle-resolved photoemission spectroscopy. *Machine Learning: Science and Technology*, 2020.

[35] J. Carson Meredith, Alamgir Karim, and Eric J. Amis. High-throughput measurement of polymer blend phase behavior. *Macromolecules*, 33:5760–5762, 2000. doi: 10.1021/ma0004662.

[36] J. Carson Meredith, Archie P. Smith, Alamgir Karim, and Eric J. Amis. Combinatorial materials science for polymer thin-film dewetting. *Macromolecules*, 33:9747–9756, 2000. doi: 10.1021/ma001298g.

[37] Marcus Noack. gpcam v6. [Computer Software] `https://bitbucket.org/MarcusMichaelNoack/gpcam`, jan 2021. URL `https://doi.org/10.11578/dc.20210217.5`.

[38] Marcus Noack. gpcam, 2021. URL `https://www.gpcam.lbl.gov`.

[39] Marcus Noack and Petrus Zwart. Computational strategies to increase efficiency of gaussian-process-driven autonomous experiments. In *2019 IEEE/ACM 1st Annual Workshop on Large-scale Experiment-in-the-Loop Computing (XLOOP)*, pages 1–7. IEEE, 2019.

[40] Marcus M Noack and Simon W Funke. Hybrid genetic deflated newton method for global optimisation. *Journal of Computational and Applied Mathematics*, 325:97–112, 2017.

[41] Marcus M Noack and James A Sethian. Advanced stationary and non-stationary kernel designs for domain-aware gaussian processes. *arXiv preprint arXiv:2102.03432*, 2021.

[42] Marcus M Noack, Kevin G Yager, Masafumi Fukuto, Gregory S Doerk, Ruipeng Li, and James A Sethian. A kriging-based approach to autonomous experimentation with applications to x-ray scattering. *Scientific Reports*, 9:11809, 2019.

[43] Marcus M Noack, Gregory S Doerk, Ruipeng Li, Masafumi Fukuto, and Kevin G Yager. Advances in kriging-based autonomous x-ray scattering experiments. *Scientific Reports*, 10:1325, 2020.

[44] Marcus M Noack, Gregory S Doerk, Ruipeng Li, Jason K Streit, Richard A Vaia, Kevin G Yager, and Masafumi Fukuto. Autonomous materials discovery driven by gaussian process regression with inhomogeneous measurement noise and anisotropic kernels. *Scientific Reports*, 10:17663, 2020.

[45] Oskar Paris. From diffraction to imaging: New avenues in studying hierarchical biological tissues with x-ray microbeams (review). *Biointerphases*, 3:FB16, 2008. doi: 10.1116/1.2955443.

[46] Charles Sanders Peirce. The fixation of belief (1877). *The Essential Peirce*, 1, 1877.

[47] Charles Sanders Peirce. How to make our ideas clear. *The nature of truth: Classic and contemporary perspectives*, 2001:193–209, 1878.

[48] Geoff Pleiss. *A SCALABLE AND FLEXIBLE FRAMEWORK FOR GAUSSIAN PROCESSES VIA MATRIX-VECTOR MULTIPLICATION*. PhD thesis, Cornell University, 2020.

[49] Alexander J Probst, Hoi-Ying N Holman, Todd Z DeSantis, Gary L Andersen, Giovanni Birarda, Hans A Bechtel, Yvette M Piceno, Maria Sonnleitner, Kasthuri Venkateswaran, and Christine Moissl-Eichinger. Tackling the minority: sulfate-reducing bacteria in an archaea-dominated subsurface biofilm. *The ISME journal*, 7(3):635–651, 2013.

[50] Sonya V. Roberson, Albert J. Fahey, Amit Sehgal, and Alamgir Karim. Multifunctional tof-sims: combinatorial mapping of gradient energy substrates. *Applied Surface Science*, 200:150–164, 2002. doi: 10.1016/S0169-4332(02)00887-5.

[51] Burr Settles. Active learning literature survey. Technical report, University of Wisconsin-Madison Department of Computer Sciences, 2009.

[52] Archie P. Smith, Jack F. Douglas, J. Carson Meredith, Eric J. Amis, and Alamgir Karim. High-throughput characterization of pattern formation in symmetric diblock copolymer films. *Journal of Polymer Science Part B: Polymer Physics*, 39:2141–2158, 2001. doi: 10.1002/polb.1188.

[53] Archie P. Smith, Amit Sehgal, Jack F. Douglas, Alamgir Karim, and Eric J. Amis. Combinatorial mapping of surface energy effects on diblock copolymer thin film ordering. *Macromolecular Rapid Communications*, 24:131–135, 2003. doi: 10.1002/marc.200390001.

[54] G. L. Squires. *Introduction to the Theory of Thermal Neutron Scattering*. Cambridge University Press, 2012. ISBN 978-1-107-64406-9.

[55] Christopher M. Stafford, Kristen E. Roskov, Thomas H. Epps III, and Michael J. Fasolka. Generating thickness gradients of thin polymer films via flow coating. *Review of Scientific Instruments*, 77: 023908, 2006. doi: 10.1063/1.2173072.

[56] A. Stein, G. Wright, K. G. Yager, G. S. Doerk, and C. T. Black. Selective directed self-assembly of coexisting morphologies using block copolymer blends. *Nature Communications*, 7:12366, 2016. doi: 10.1038/ncomms12366.

[57] Kristof Toth, Chinedum O. Osuji, Kevin G. Yager, and Gregory S. Doerk. Electrospray deposition tool: Creating compositionally gradient libraries of nanomaterials. *Review of Scientific Instruments*, 91:013701, 2020. doi: 10.1063/1.5129625.

[58] Patricia M Valdespino-Castillo, Ping Hu, Martín Merino-Ibarra, Luz M López-Gómez, Daniel Cerqueda-García, González-De Zayas, Teresa Pi-Puig, Julio A Lestayo, Hoi-Ying Holman, Luisa I Falcón, et al. Exploring biogeochemistry and microbial diversity of extant microbialites in mexico and cuba. *Frontiers in microbiology*, 9:510, 2018.

[59] Qianqian Wang, Michiko Nemoto, Dongsheng Li, James C. Weaver, Brian Weden, John Stegemeier, N. Bozhilov, Krassimir, Leslie R. Wood, Garrett W. Milliron, Christopher S. Kim, Elaine DiMasi, and David Kisailus. Phase transformations and structural developments in the radular teeth of cryptochiton stelleri. *Advanced Functional Materials*, 23:2908–2917, 2013. doi: 10.1002/adfm.201202894.

[60] James C. Weaver, Garrett W. Milliron, Ali Miserez, Kenneth Evans-Lutterodt, Steven Herrera, Isaias Gallana, William J. Mershon, Brook Swanson, Pablo Zavattieri, Elaine DiMasi, and David Kisailus. The stomatopod dactyl club: A formidable damage-tolerant biological hammer. *Science*, 336:1275–1280, 2012. doi: 10.1126/science.1218764.

[61] T. Weber. Takin 2 (software). 2014 – 2021. doi: 10.5281/zenodo.4117437. URL `https://code.ill.fr/scientific-software/takin`.

[62] Tobias Weber. Update 2.0 to "Takin: An open-source software for experiment planning, visualisation, and data analysis", (PII: S2352711016300152). *SoftwareX*, 14:100667, 2021. ISSN 2352-7110. doi: 10.1016/j.softx.2021.100667.

[63] L Wiegart, GS Doerk, M Fukuto, S Lee, R Li, G Marom, MM Noack, CO Osuji, MH Rafailovich, JA Sethian, et al. Instrumentation for in situ/operando x-ray scattering studies of polymer additive manufacturing processes. *Synchrotron Radiation News*, 32(2):20–27, 2019.

[64] Christopher KI Williams and Carl Edward Rasmussen. *Gaussian processes for machine learning*, volume 2. MIT press Cambridge, MA, 2006.

[65] Kevin G. Yager, Erica Lai, and Charles T. Black. Self-assembled phases of block copolymer blend thin films. *ACS Nano*, 8:10582–10588, 2014. doi: 10.1021/nn504977r.

[66] Kevin G. Yager, Yugang Zhang, Fang Lu, and Oleg Gang. Periodic lattices of arbitrary nano-objects: modeling and applications for self-assembled systems. *Journal of Applied Crystallography*, 47:118–129, 2014. doi: 10.1107/S160057671302832X.

[67] Nathan Yee, Liane G Benning, Vernon R Phoenix, and F Grant Ferris. Characterization of metal-cyanobacteria sorption reactions: a combined macroscopic and infrared spectroscopic investigation. *Environmental Science & Technology*, 38(3):775–782, 2004.