K-Means-Driven Gaussian Process Data Collection for Angle-Resolved Photoemission Spectroscopy

Charles N. Melton[1], Marcus M. Noack[2], Taisuke Ohta[3], Thomas E. Beechem[3], Jeremy Robinson[4], Xiaotian Zhang[1], Aaron Bostwick[1], Chris Jozwiak[1], Roland J. Koch[1], Petrus H. Zwart[2], Alexander Hexemer[1], and Eli Rotenberg[1,*]

[1]Advanced Light Source, Lawrence Berkeley National Laboratory. Berkeley, CA 94720
[2]Center for Advanced Mathematics for Energy and Research Applications, Lawrence Berkeley National Laboratory. Berkeley, CA 94720
[3]Sandia National Laboratories, Albuquerque, NM 87185
[4]U. S. Naval Research Laboratory, Washington, DC 20375
[*]Corresponding Author: erotenberg@lbl.gov

Abstract

We propose the combination of k-means clustering with Gaussian Process (GP) regression in the analysis and exploration of 4D angle-resolved photoemission spectroscopy (ARPES) data. Using cluster labels as the driving metric on which the GP is trained, this method allows us to reconstruct the experimental phase diagram from as low as 12% of the original dataset size. In addition to the phase diagram, the GP is able to reconstruct spectra in energy-momentum space from this minimal set of data points. These findings suggest that this methodology can be used to improve the efficiency of ARPES data collection strategies for unknown samples. The practical feasibility of implementing this technology at a synchrotron beamline and the overall efficiency implications of this method are discussed with a view on enabling the collection of more samples or rapid identification of regions of interest.

Keywords: ARPES, Gaussian Process, K-means clustering, Machine Learning, Spectra

1 Introduction

Machine Learning and Artificial Intelligence have fundamentally changed the way we conduct scientific experiments [1, 2], particularly when it comes to data collection and analysis [3, 4]. In a traditional experiment, scientists would either adjust measurement conditions based on intuition and past experience or by an exhaustive systematic search. Both approaches are wasting staff, device and computation time, and data storage. Uncertainty quantification within ML frameworks has allowed scientists to autonomously steer experiments, select measurements based on their information content, and avoid the acquisition of redundant information [5]. This way, a high-information-density dataset is collected. These methods liberate the human scientists from micromanaging decision making during an experiment and let them focus on the scientific meaning of the result.

Data collection and analysis focused on high-throughput materials discovery are being improved by a multitude of machine learning and statistical analysis techniques [6–10]. The ability to quickly analyze data as it is collected and make informative decisions on how to continue the experiment is currently being investigated heavily, using a variety of machine learning techniques and other statistical methods, such as Gaussian process (GP) regression [11–15]. As sample preparation rates accelerate towards high throughput experimentation, allocated experiment time must follow suit to ensure proper analysis of the increased number of samples. As detectors become more advanced and have higher resolution, more storage space

is needed to archive the data. Therefore, it becomes necessary to develop methods of conducting measurements in a fashion that maximizes information content per data set, allowing several experimental samples to be analyzed in the time frame previously allocated to just one sample.

Angle-resolved photoemission spectroscopy (ARPES) is a technique that is used to study properties of quantum materials and make inferred decisions for material synthesis. ARPES experiments are based on the photoelectric effect, and are used to study relationships between the energy and momentum of electrons in materials [16]. Specifically, this method can study band structure and lifetime of charged excitations in materials. In the past, only homogeneous samples have been studied, but more recently, apparatuses capable of resolving these properties in samples that are heterogeneous on the nanometer scale have become available [17]. Distinct electronic phases with abrupt boundaries can occur in samples because the momentum-energy relationships are strongly affected by interactions between electrons that are in turn affected by local variations in disorder, structure, and composition. Understanding the details of these phases provides clues on how to control the interactions to enhance properties such as superconductivity or to develop new exotic materials such as topological insulators [18–20].

The phase and electronic property information that is encoded in ARPES spectra is of interest for determining future material preparation and is encoded in 2D spectral images. These spectra can contain pure or mixed phase information depending on the location the sample was scanned and the relative sizes of the probe and sample features. Interesting physics lie in the mixed phase regions, and this information is sought when determining new samples to produce. It is therefore necessary to be able to decompose and group a collection of spectra to study the underlying physics. It has been shown that a useful method of showing which x-ray spectra correspond to different components of an experiment sample (i.e. chemical composition, particle spacing, etc.) is measurement decomposition. Decomposition methods such as Principle Component Analysis (PCA) [21] and Non-negative matrix factorization (NMF) [22] have successfully been used to obtain useful material information from spectra that are considered to be a linear combination of several different base spectra [23]. Another route is to identify areas of similarity on a sample, thus making a phase map of the sample; that is, identifying areas of a sample that hold similar spectral characteristics. From there we can identify phase borders and thus isolating phase regimes. One such technique to accomplish this, is k-means clustering. K-means clustering labels similar data points in a collected dataset with a cluster identifier, and has been shown to be successful at identifying and separating electronic properties of materials [24].

In this paper we demonstrate the potential of combining two machine learning techniques to autonomously steer measurements in order to quickly obtain the best representation of the overall sample in a time frame superior to standard grid scanning methods. By combining k-means clustering which identifies regions of similarity and using those metrics to drive the autonomous decision making, important information and sample space understanding is obtained at a rate superior to grid scanning and random scanning, and that the autonomous method consistently generates a superior model of the overall sample. From here on, we refer to this method as a K-Means-Driven Gaussian Process data collection, or KMGP.

This paper will follow the layout described here. Section 2 describes the machine learning algorithms being used, and the science they are applied to. Section 3 focuses on results. Section 4 presents the discussion, and Section 5 contains conclusions and future work.

2 Methods

2.1 Autonomous Experimentation

2.1.1 An Introduction to Autonomous Experimentation

Over the past few decades, the complexity of material's composition-processing-structure-property relations, being explored at experimental facilities across the globe, has risen to unprecedented levels. This development is partly driven by ever-increasing data-collection rates. However, the data-collection rates cannot keep pace with the increase in the number of dimensions of the parameter spaces that underlie materials discovery. This gives rise to methods that are trying to optimize the scientific value of each measurement. As early as the late 1800s, statisticians came up with rules to make experiments more efficient. The field became widely known as "Design of Experiments" (DOE). The work "Note on the Theory of the Economy of Research" by C.S. Peirce might have been the first work on optimal experimental design for general regression models. Work on optimal experiments for polynomial models even preceded Peirce's work [25]. A more recent, and very popular method of DOE is the method known as the latin-hyper-cube method, which is still widely used [26]. With the emergence of machine learning, a sub-field developed, called "active learning" [27]. Active learning is concerned with methods which optimize the data collection in order to reconstruct a model function as efficiently as possible. One particularly successful strategy is to collect data in regions where the uncertainty or the potential for information gain is high, shifting the focus to probability theory and statistics. To model the uncertainty, Gaussian processes (GPs) are commonly used due to their computational simplicity and analytical tractability. We will introduce GPs in the following section. As we will see, a GP, given some data points, provides a posterior mean function and an uncertainty. One strategy is to perform the next measurement where this uncertainty is a maximum. The scheme for such procedures is explained by Noack et al. [5, 28].

2.1.2 The Basics of GP Regression

Building a GP regression model from data $D = \{(\mathbf{x}_1, y_1), ..., (\mathbf{x}_N, y_N)\}$, where $y_i = f(\mathbf{x}_i) + \epsilon$, and $\mathbf{x} \in \mathcal{X}$, is done by defining a Gaussian probability density function as

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

where $\boldsymbol{\mu} = [\mu(\mathbf{x}_1), ..., \mu(\mathbf{x}_N)]^T$ is the mean, $\mathbf{f} = [f(\mathbf{x}_1), ..., f(\mathbf{x}_N)]^T$ contains the latent function values and $\mathbf{K} = \mathbf{K}(\boldsymbol{\phi}, \mathbf{x}_i, \mathbf{x}_j)$ is the prior covariance over the data with $\mathbf{x} \in \mathcal{X}$. $\boldsymbol{\phi}$ is a set of hyper-parameters and $\mathcal{X}$ is commonly referred to as the index set, or the input, sample or parameter space. The noise, given a function $\mathbf{f}$ as the mean, follows the density

$$p(\mathbf{y}) = \frac{1}{\sqrt{(2\pi)^k |\sigma^2 I|}} \exp[-\frac{1}{2}(\mathbf{y} - \mathbf{f})^T (\sigma_m^2 \mathbf{I})^{-1} (\mathbf{y} - \mathbf{f})], \tag{2}$$

where $\mathbf{I}$ is the identity matrix and σ_m^2 are the i.i.d. (i.e. homogeneous or input-independent) measurement variances. From equations (1) and (2), we can calculate $p(f(\mathbf{x}_0)|\mathbf{x_0}, D)$, i.e., the probability distribution for a measurement outcome at $\mathbf{x}_0$, given the dataset. The mean and variance of this distribution are

$$m(\mathbf{x}_0) = \boldsymbol{\mu} + \mathbf{k}^T (\mathbf{K} + \sigma_m^2 \mathbf{I})^{-1} (\mathbf{y} - \boldsymbol{\mu}) \tag{3}$$

$$\sigma^2(\mathbf{x}_0) = k(\mathbf{x_0}, \mathbf{x_0}) - \mathbf{k}^T (\mathbf{K} + \sigma_m^2 \mathbf{I})^{-1} \mathbf{k}, \tag{4}$$

respectively. The covariance matrix entries are defined by positive semi definite kernel functions $k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ such that

$$K_{ij}(\phi) = k(\mathbf{x}_i, \mathbf{x}_j; \phi) \; and \\ \mathbf{k} = k(\mathbf{x}_0, \mathbf{x}_j; \phi). \tag{5}$$

Popular choices of kernel functions are the Matérn kernels

$$k(\mathbf{x}_1, \mathbf{x}_2) = \frac{1}{2^{v-1}\Gamma(v)} \Big(\frac{\sqrt{2v}}{l} r \Big)^v B_v \Big(\frac{\sqrt{2v}}{l} r \Big) \tag{6}$$

and the exponential kernel

$$k(\mathbf{x}_1, \mathbf{x}_2) = \exp\Big(-\frac{r^2}{2l^2} \Big), \tag{7}$$

where $r = r(||\mathbf{x}_1 - \mathbf{x}_2||)$ is a function of some norm, l is the length scale, B_v is the modified Bessel function and v is a parameter controlling the differentiability of the kernel function. The exponential kernel function was used for this work. The mean function and the hyper-parameters are found by maximizing the marginal log-likelihood

$$\log(L(D; \boldsymbol{\phi}, \boldsymbol{\mu})) \; = \\ -\frac{1}{2}(\mathbf{y} - \boldsymbol{\mu})(\mathbf{K}(\boldsymbol{\phi}) + \sigma_m^2 \mathbf{I})^{-1}(\mathbf{y} - \boldsymbol{\mu}) \\ -\frac{1}{2}\log(|\mathbf{K}(\boldsymbol{\phi}) + \sigma_m^2 \mathbf{I}|) - \frac{\dim(\mathbf{y})}{2}\log(2\pi). \tag{8}$$

When applied to experimentation, the GP described above works by interpolating some function that is defined over a parameter space which is defined by the user. The parameter space can be composed of sample positional coordinates or experimental parameters, such as concentrations or temperatures. For this work, the function we are interpolating is a labeled phase diagram of a given ARPES sample. That phase diagram is constructed by dividing up the collected data by k-means clustering.

2.2 K-Means Clustering

K-Means clustering is an unsupervised machine learning method that seeks to segregate observations into a defined number of clusters. Given a set of n observables (in this work, spectra) $\mathbf{s} = (\mathbf{s}_1, \mathbf{s}_2, ..., \mathbf{s}_n)$, the objective is to divide the observables into K clusters: $\mathbf{C} = (\mathbf{C}_1, \mathbf{C}_2, ..., \mathbf{C}_K)$. This is done my minimizing the sum of squares

$$W(\mathbf{s}, \mathbf{C}) = \sum_{k=1}^{K} \sum_{i \in \mathbf{C}_k} ||\mathbf{s}_i - \mathbf{c}_k||^2 \tag{9}$$

where $\mathbf{c}_k = \frac{1}{n_k} \sum_{i \in \mathbf{C}_k} \mathbf{s}_i$ is the centroid of cluster $\mathbf{C}_k$, n_k is the number of points in cluster $\mathbf{C}_k$, and $|| \cdot ||$ is the Euclidean norm [29].

2.3 K-Means and GP Driven Autonomous Experiments

As discussed in the previous section, GP-driven autonomous experiments aim to interpolate collected data as a function of various experimental parameters. In this specific case, the autonomous decision making is based on ARPES spectra collected at various x-y coordinates in sample space. However, instead of

having the GP interpolate entire spectra across the sample, we are interpolating the cluster label of the data $D = \{(\mathbf{x}_1, y_1), ..., (\mathbf{x}_N, y_N)\}$ defined to each specific spectrum over the sample space.

The first step is to define how many clusters the collected spectra are divided into. While this is usually done with *a priori* knowledge, there are several metrics in place to quantify this number. We use the silhouette score in this work. Starting with a collection of initial spectra and measurement locations, an initial clustering from $k = 2$ to $k = 10$ is performed. For each value of k, a silhouette score is calculated for all n points, which is defined as

$$\text{sil} = \frac{1}{n} \sum_{i=1}^{n} \frac{b\,(i) - a\,(i)}{\max\{a\,(i)\,, b\,(i)\}}, \tag{10}$$

where $a\,(i)$ is the mean distance between spectrum i and all other spectra in a cluster, and $b\,(i)$ is the smallest distance between spectrum i and all other points in other clusters which do not contain spectrum i. The silhouette score is a normalized measure of the maximum distance between points in one cluster and all the other points of another cluster. The silhouette score is calculated for each value of k, and the optimal number of clusters is given by the maximum score.

Using the determined optimal number of clusters, the data is then clustered and given a label. These labels are used to identify different "phases" based on the calculated centroids. The remaining phase labels are then interpolated via the GP. Then following standard GP practices, the point of the interpolated phase map that has the highest variance is measured next. Once the new spectrum is collected, the clustering algorithm and GP is then repeated until the desired number of points have been scanned.

2.4 ARPES Experiment and its Applicability to Materials Discovery

The datasets considered here derive from spatial-, energy- and momentum-resolved angle resolved photoemission (ARPES) data acquired at the MAESTRO beamline at the Advanced Light Source. ARPES measures the probability that a focused soft x-ray photon beam ejects a bound electron from a material into the surrounding space at a given angle θ and kinetic energy. These quantities can be used to infer the initial state of the electron, namely its binding energy ω and momentum $\vec{k}$ (in units where Planck's constant $\hbar$ is set equal to 1). The momentum is a generally a vector in three dimensional (x, y, z)-space, and energy ω adds a fourth measured quantity. The measured ARPES spectral function, called $A(\mathbf{k}, \omega)$ [16], can be represented as a four-dimensional image, with each voxel corresponding to a measurement of the probability of detecting an electron at coordinate $(\mathbf{k}, \omega)$.

For many materials of interest, the electrons are completely, or approximately, confined to an (x, y) plane, and therefore the spectral function $A(\mathbf{k}, \omega)$ spans a three dimensional space that can be readily visualized. An example for electrons in graphene, a single monolayer of carbon atoms arranged in a honeycomb lattice is shown Fig. 1 [30]. The bright features correspond to the allowed energy bands of the material, which are eigenvalues of the Schrödinger equation for carbon valence electrons on a honeycomb lattice [31]. All conventional properties of materials, such as its optical appearance dielectric response, metallic / insulating character, elastic moduli, etc, can be related back to these states.

In a mathematical sense, these solutions should lead to infinitely sharp features along allowed contours in $(\mathbf{k}, \omega)$ space, but the reality is different: The states are considerably broader than expected, with detectable, diffuse intensity throughout the probed volume. Furthermore, deviations in the location of features are typical. These effects arise from interactions among the electrons in the material, as well as between the electrons and lattice defects, including atomic vibrations.

These many-body interactions are of tremendous interest because they drive materials into useful new ground state electronic phases such as magnetism and superconductivity. Since the many-body interactions can be influenced by materials structure and composition, the search for new materials with better ground state properties (for example, superconductors that work at room temperature). Subtle changes in the many

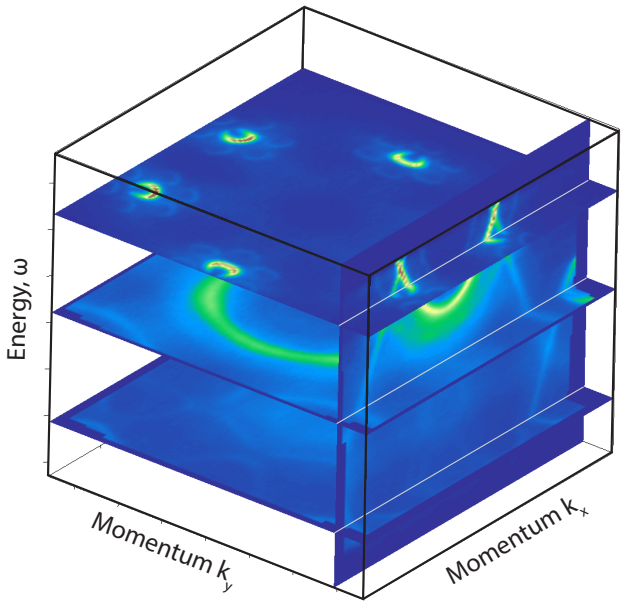

Figure 1: ARPES spectral function for the electrons in graphene, a single monolayer of carbon atoms arranged in a honeycomb lattice [30]. The bright features seen in the Figure are the allowed energy bands of the material. These allowed energy bands are acquired from eigenvalues of the Schrödinger equation. In ARPES, spectra are collected as a function of position. In this paper, we are clustering all collected spectra; the cluster labels are then fed into a Gaussian process to calculate uncertainties in the labels which are then used to dictate where the next experiment should take place

body interactions that lead to improved properties are therefore encoded in the ARPES spectra, but lacking accurate theoretical models, their interpretation is often qualitative.

In modern ARPES experiments, data such as in Figure 1, or two-dimensional cuts through the $(\mathbf{k}, \omega)$ volume, are acquired as a function of position (x, y) across the surface of heterogenous samples. The goal in such experiments is to quickly identify the interesting phases, "zoom in" on the interesting features and study them in further detail or with other probes in order to understand how the electronic structure in Figure 1 is related to the local material properties.

A more ambitious goal would be to sample the ARPES data not in real-space coordinates for heterogeneous samples, but in an abstract space that spans the "fabricational" degrees of freedom (DOFs)–structural and compositional–that are available to materials designers. But with a large number of available chemical elements, and a practically infinite way to arrange these atoms on many length scales, it is a daunting challenge to effectively find new material phases in such a large, high-dimensional space, especially when the indicators of these new phases are composed of large multidimensional ARPES datasets without easily quantifiable descriptors. Furthermore, development of a fully autonomous materials discovery workflow incorporating ARPES would require exorbitant allocations of time on shared user facilities such as synchrotrons where ARPES is typically conducted. Our aim is to develop machine learning techniques that maximize the use of expensive resources by judiciously sampling the DOFs in order to collect the most information with the least number of points sampled.

2.5 The Simulated Beamline Experiment

To develop a testbed for the proposed technique, we focus on spatial (x, y) scans of heterogeneous materials, which serve as a simple surrogate problem. We acquire ARPES data for heterogenous samples, where the

goal is to quickly scan the materials spatially in order to locate regions of interest that can be distinguished by their electronic structure. Until now, hyperspectral ARPES maps are acquired by uniform sampling of $\sim$ 10 000 points in an (x, y) grid spanning the sample. We seek to acquire the equivalent information with far fewer data points, sampling non-uniformly by points chosen algorithmically as data are collected. To model the performance of the KMGP algorithm for such real-world data, we use the uniformly sampled gridded ARPES data that has been previously collected and treat it as unknown data to be sampled sequentially by our algorithms. As the data set is sampled point-by-point, the collected spectra are fed into the KMGP algorithm which decides where to scan next. For this work we compare the efficiency of the KMGP algorithm to both random and grid scanning.

Angle-resolved photoemission intensity maps used for ground truth comparisons were recorded on a Scienta R4000 analyzer at the MAESTRO beamline using a focused synchrotron x-ray beam at the Advanced Light Source. A Fresnel zone plate was used as a focusing element, which allows a minimum spot size of 100 nm. For ARPES, the photon energy was set to 100 eV and the sample was held at room temperature. The samples were annealed under vacuum at 200 °C for 30 min right before the measurement for removal of surface adsorbates and containment

Two different datasets were used for this study. ARPES measurements for twisted graphene and tungsten disulfide (WS_2) were collected using experimental conditions as stated above. The data, originally collected at a very high resolution, was downsized to an image size of 128×128 and cropped accordingly to remove artifacts that may have been present due to electron collection at the detector edge. The twisted graphene dataset consisted of 8 281 collected spectra, and the WS_2 had 40 401 spectra. To aid in computation time, a 50×50 section of the WS_2 dataset was used for the simulated experiment. The spectra were also standardized to have a mean of 0 and a variance of 1 to enhance faint features that may be present, yet physically significant.

The metric chosen to determine the quality of the measured data points for various acquisition methods is the mean absolute percent error (MAPE). The different scanning methods all begin with 4 spectra from the corners of the sample. The remaining spectra across the sample space are then linearly interpolated. The MAPE is then calculated between the true dataset and the interpolated dataset. After each new spectrum is measured, the dataset we are building is interpolated again, and a new MAPE is calculated. This process is repeated for every measured point for every scanning method.

This method reflects the quality of the knowledge of the collected data. As more data points are collected, the interpolated points will begin to converge to their known values. We hypothesize that the KMGP process will lead to a better understanding of the sample in a much quicker time frame compared to grid scanning and random scanning methods.

3 Results

3.1 Twisted Graphene Dataset

Figure 2 shows an example of tuning the number of clusters examined as a function of total points measured on the sample of twisted graphene. Initially, the preferred number of clusters as dictated by the maximum silhouette score is low. This is expected as initially there are a low number of spectra collected, and they look similar in terms of their features. As more spectra are collected, the difference in features becomes more apparent and so the maximum silhouette score occurs at higher values of k. In the case of a new silhouette score predicting a lower optimal value of k after a new spectrum was collected, the previous value for k was used instead. At roughly 30 spectra collected, the unsupervised learning algorithm settled on a consistent value of k.

Figure 3 shows the evolution of the clustered data points as more points are scanned. At just over collected 100 points, region borders have yet to be defined but patterns have begun to appear in the clustering.

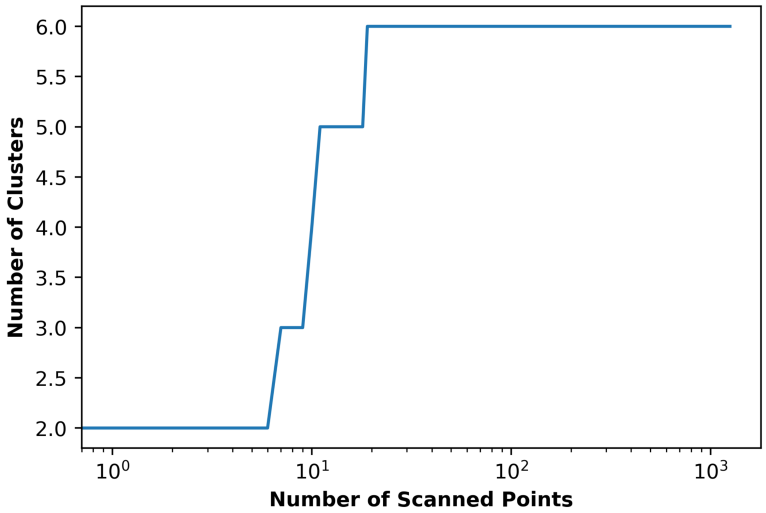

Figure 2: Decided optimal number of clusters vs number of measured spectra on the sample. The number of clusters used for the k-means clustering is determined by the maximum silhouette score of a possible range from 2 to 10 clusters. In the case of a lower optimal cluster number is determined compared to the previous collected spectra, the last maximum cluster number is used instead.

At 250 data points, the regions are visible, and by 500 data points borders between phases have successfully been defined. It should be noted that this diagram contains just one-sixteenth of the entire dataset.

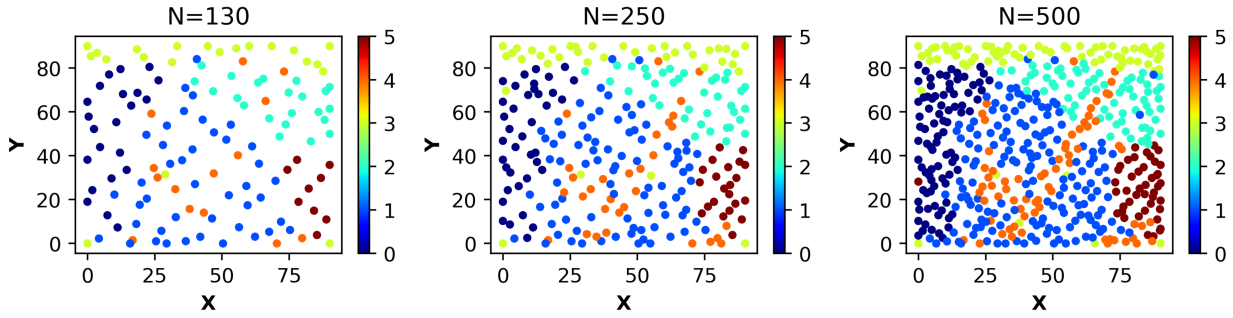

Figure 3: Clustered spectra after an increasing number of collected points (N). Following KMGP data collection, distinct phase regions can already be determined when the spectra are clustered. At just 500 collected spectra, major phase areas are identified. The x-y axes are the indices of each spectrum in the dataset.

Figure 4a shows the MAPE between the real dataset and the interpolated dataset for various scanning methods. Each method was repeated 10 times. The solid lines are the mean of the runs and the shaded regions around each curve are the 95% confidence intervals. Grid scanning, random scanning, and KMGP are highlighted in the figure. Grid scanning eventually converges to random scanning at 500 data points measured (the total number of points in the grid). This result highlights that both KMGP and the random method allow for a faster rate of data collection compared to grid scanning, which KMGP providing a better representation of the data collected.

By utilizing the GP to aid in data collection, we show the ability to collect data at a higher rate than standard grid data collection. Both KMGP and random scanning produce more knowledge of the dataset faster than grid scanning, but this is expected due to the fact that while grid scanning is restricted to sequential scan points predefined by the user, random and KMGP scan points are not subjected to such a limit. One possible method of improving grid data collection is to use a finer grid. Figure 4b shows the MAPE for each

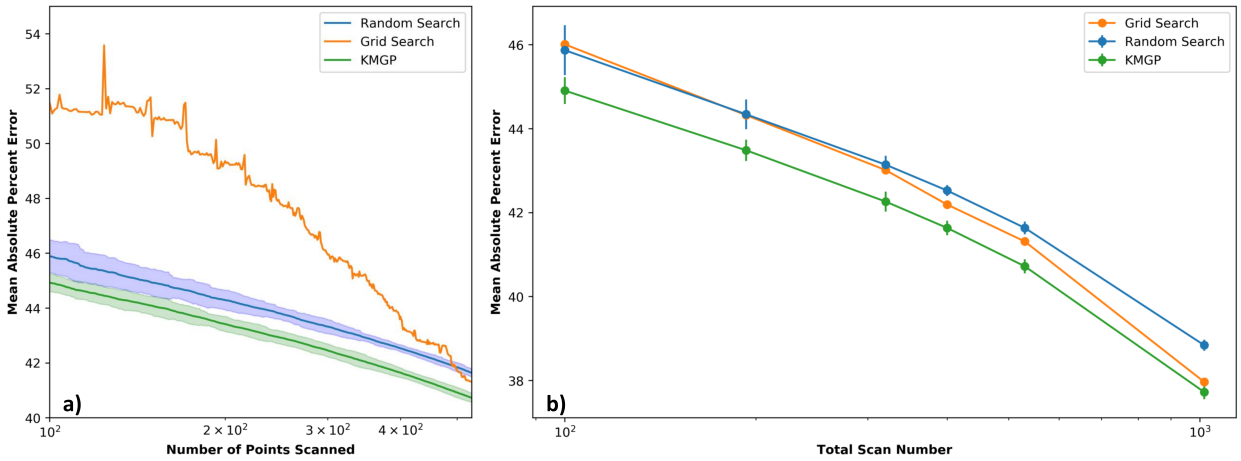

Figure 4: Figure a) shows the interpolated dataset error as a function of total scan number shown on semi-log plot. In a comparison experiment with a grid of 530 spectra, collected in a traditional fashion, KMGP provides a better model of the entire dataset. b) shows the interpolated dataset error for increasing grid collection schemes, from a 100 pt grid to a 1000 pt grid. Errorbars are 95% confidence interval.

method with a progressively finer grid being used. Consistently, KMGP produces a more reliable model of the data, up to 1 000 data points.

Figure 5 shows the evolution of an interpolated spectrum from the sample as more points are scanned by the KMGP method. As more spectra are collected, finer features become apparent in the interpolation. To numerically determine convergence to the ground truth spectra we used the Pearson correlation coefficient, whose values take on $[-1, 1]$, with -1 corresponding to fully anti-correlated, and 1 being fully correlated. As 1 000 spectra are collected, the correlation value reaches approximately 0.99. This highlights that a majority of the sample knowledge is collected with just $\sim$12% of the original data.

3.2 Tungsten Disulfide Sample

Figure 6 shows the cluster count change as a function of the number of the collected spectra. Again, this was determined by calculating silhouette scores and determining which maximum to used the same criteria as the twisted graphene cluster determination. For WS_2 the optimal number of clusters was obtained after 100 scans. This region held a majority of similar spectra. Several differences were present, such as a boron-nitride background and spectra that exhibit systematic energy shifts with position. The inset shows the final clustering of the spectra as the simulated experiment was completed.

Figure 7 shows the improvement of the interpolation of a spectrum from a defined region of the WS_2 sample as more points are chosen by the KMGP method. As more points are measured, we again see the spectrum improve and approach the ground truth solution, which is highlighted by the increasing Pearson correlation coefficient. As 1 000 collected spectra are approached (2.5% of the original data set), the Pearson coefficient approaches 0.96.

4 Discussion

In this paper we have proposed a method of autonomous data collection for ARPES which combines both supervised and unsupervised machine learning algorithms, namely k-means clustering and GP regression.

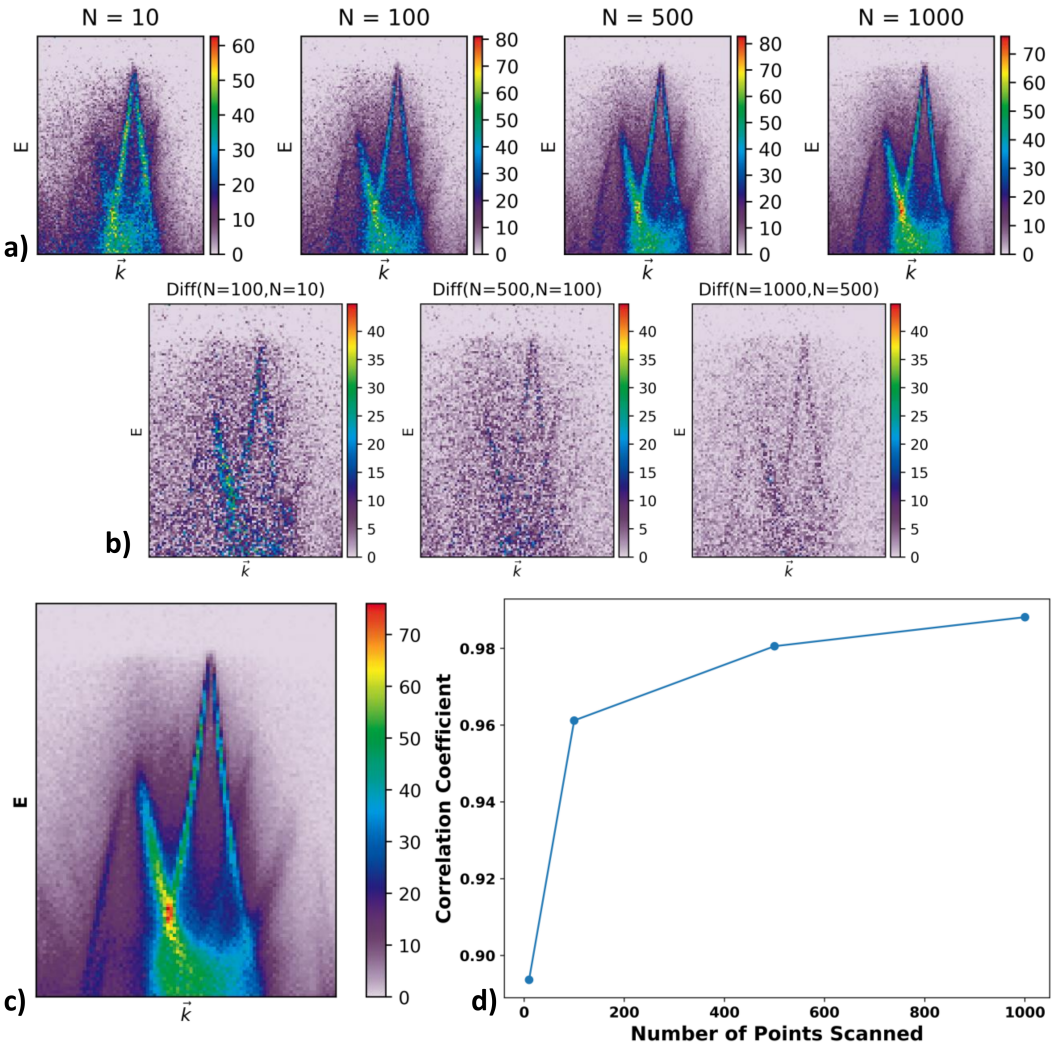

Figure 5: Interpolated spectra from twisted graphene as more points are collected. Figure a) shows the interpolated spectrum improving as more measured points are collected. Figure b) shows the difference between each iteration of interpolation in a). As more spectra are collected with the KMGP method, the overall change in the interpolated spectrum decreases. Figure c) shows the spectrum from the same region in the ground truth dataset. Figure d) shows the Pearson correlation improving as more points are collected.

The materials examined using ARPES can have phase diagrams constructed by k-means clustering which groups similar spectra together and calculated a centroid which in this case is a dominant spectrum for different regions. Since k-means clustering requires a-priori knowledge to determine the proper number of clusters for a dataset, we use the silhouette score metric. Figure 2 shows how the method chooses to cluster the data after each new measurement. Over several decades the tuning settled on a consistent number of $k = 6$. This highlights the effectiveness of using a GP to drive the data collection based on clustering of spectra. At just over 100 spectra collected, the GP had found the major cluster areas and continued to scan those areas in a smart fashion, which again was highlighted in Figure 3.

The KMGP method of data collection works in principal, but we need to highlight its advantages to traditional grid scanning methods and even a random scan of the sample. GPs come with high computational

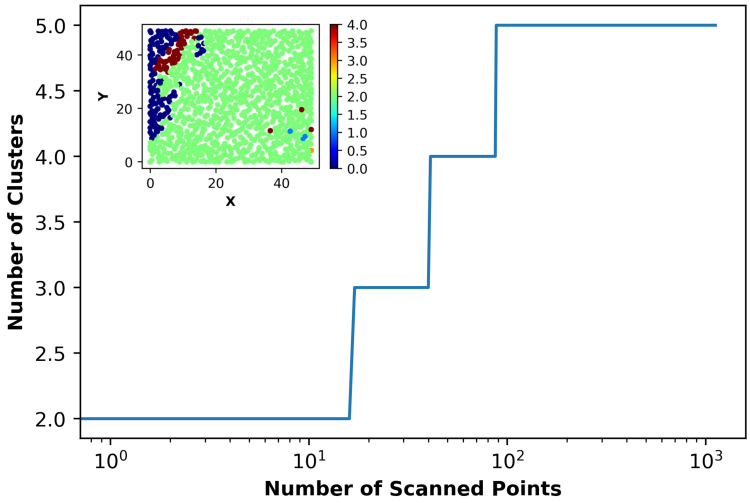

Figure 6: Decided optimal number of clusters vs number of measured spectra on the sample. The inset highlights the final clustered areas. Minute differences in the WS_2 were detected, as well as the boron-nitride background (dark blue) and a large defect due to a folding of the material (dark red). The x-y axes are the indices of each spectrum in the dataset.

costs as more data is collected, so we show that both the rate of collecting useful information is increased, as well as the quality of that information. Figures 4 and 5 highlight these points. In terms of data collection rate, both KMGP and random scans collect useful data in a quicker fashion than a grid scan of the same number of sampled spectra. However, KMGP outperforms random data collection due to the algorithm avoiding redundant information that may be acquired when simply choosing random points to scan. Figure 5 specifically shows how the knowledge of the spectra in a given area of the sample improves drastically as more points are scanned. As we approached just 12% of the original data collected, we see a very accurate representation of the ground truth. Even faint features in the spectra are visible, even though the size of the data set has been significantly reduced. This provides evidence that the KMGP method is able to aid in collecting useful sample information in much fewer measurements compared to traditional methods.

Along with an improved rate of sample knowledge comes a more accurate representation of the sample space. To verify this, we compared the errors for various grid scan sizes, starting at 100 points distributed across the sample to over 1000 points. Again, the error was calculated between our ground-truth dataset and an interpolated dataset based on the measured points from each different scan type. Grid data collection at 100 points gave us a low quality interpolation of the data, followed by 100 random points, and the highest quality was produced by KMGP. As grid scanning becomes finer, it overtakes random scanning in acquiring more knowledge of the sample space. However, even up to over 1 000 points scanned, KMGP performs better than the grid scan, with an MAPE dropping below 40%. Again, this is attributed to the fact the the GP places measurements in areas of large estimated errors, thus avoiding areas which are already well defined. We applied these analysis criteria to the WS_2 dataset and saw similar results. The KMGP method led to an overall accurate representation of the sample spectra after just 500 measured points.

When evaluating the errors on the plots in Figure 4, one has to keep in mind that we are approximating a function based on a minimum number of function evaluations. This should not be confused with error commonly associated with over-exhaustive grid based methods. The errors (i.e. standard deviations) can often be large, starting at infinite for zero measured points, and will continue to be large when a minimal number of data points are collected. However, the value of the knowledge obtained should not be underestimated.

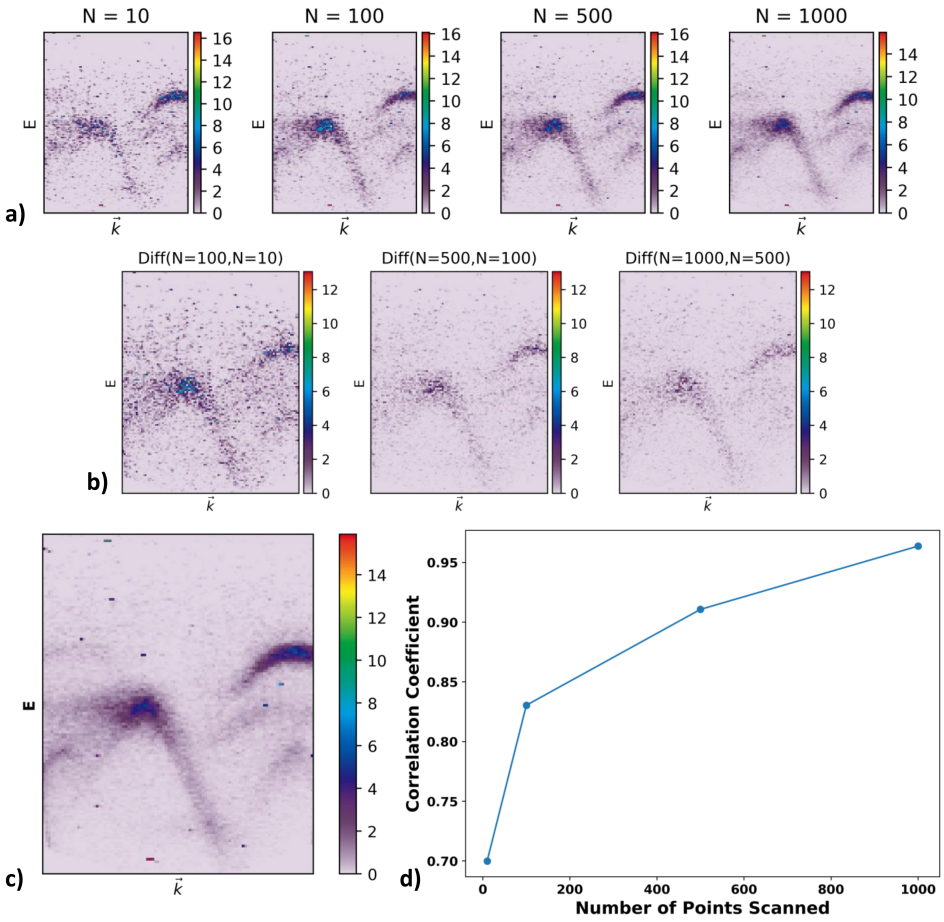

Figure 7: Interpolated spectra from WS_2 as more points are collected. Figure a) shows the interpolated spectrum improving as more scanned points are collected. Figure b) shows the difference between each iteration of interpolation in a). We can see as more spectra are collected with the KMGP method, the overall change in the interpolated spectrum decreases. Figure c) shows the spectrum from the same region in the ground truth dataset. Figure d) shows the Pearson correlation improving as more points are collected.

It is a great achievement of GPs to be able to report 40% error after a few hundred measured points, and an even greater achievement to stay below the errors of both grid and random scanning approaches. The presence of many faint features in ARPES data also attributes to this error percentage. It is expected that leaving out so many spectra will allow the reconstruction to be successful up to a point, but this does not diminish the value of this technique to scan a new, unknown sample and quickly define phases and areas of interest.

One glaring point of conflict in implementing an involved technique such as this, is the overhead the GP calculations will bring. In the standard setting, GPs involve the storage and inversion of a covariance matrix. Storage requirements therefore grow with complexity $O(N^2)$ while compute complexity grows at $O(N^3)$. This unfavorable behavior can become problematic when approaching numbers of measurements (N) in the order of 10^4. Motor controls are able to move from point to point in seconds, making the GP analysis time a potential bottleneck. Fortunately, this limitation can easily be circumvented. First, we have shown that even with relatively few measurement points, the model accuracy is high, compared to the competing methods. Second, there is a vast variety of methods to improve upon the speed and scaling of the GP analysis. Taking advantage of high performance computer architecture and the innerOur workings of GP

training and prediction, we can push the limit of the number of measurements towards 10^6, which is well above the capability of today's experiment setup. As the limit of the GP calculations is approached, beam line scientists can take over and investigate areas based on the calculated model. It is even possible switch to new criteria for GP regression that may focus on scanning interesting areas such as phase boundaries.

There are several spatially resolved measurement techniques that KMGP can be applied to to aid in data collection. Scanning probe microscopy is a technique that uses several probing methods to examine material, such as electrical, mechanical, and band excitation [12]. A second technique that happens to be similar to nanoARPES is nano Ultrafast Electron Diffraction (nanoUED) [32]. In nanoUED, electrons are used to identify crystal grain boundaries and map out crystallographic domains. Both of these techniques relate a response in a spatially resolved measurement and produce some form of hyperspectral image. Hyperspectral images from a single sample can be categorized into various domains, so KMGP has the potential to be applied to aid in data collection for these techniques.

5 Conclusion

In conclusion, we have investigated the feasibility of applying a combination of k-means clustering and GP regression to collect ARPES data in a smart fashion. By exploring a sample's 4D (x, y, energy, momentum) parameter space, the GP successfully maps out major areas of the sample with under 12% of the original data points collected. Collected knowledge of the sample was determined by linear interpolation of the spectra across the sample space and comparing that to the collected spectra. The GP consistently allowed for more accurate interpolations of the data compared to a random scan and a traditional grid scan. This continued to hold true for finer grid scans. The ability to gain useful sample information in as few collected spectra as possible in an autonomous fashion combined with physical insight provided by beam line scientists will allow for acquisition of less redundant data and shift the focus of scans to physically-interesting areas and more samples scanned compared to previous experiment times.

6 Acknowledgements

This work was supported by the 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 Advanced Light Source, which is a DOE Office of Science User Facility under contract no. DE-AC02-05CH11231. We are grateful to R. Guild Copeland and Anthony McDonald for their help in the sample preparation (twisted bilayer graphene). The work at SNL was supported by Sandia LDRD, the U.S. DOE Office of Basic Energy Sciences (BES), Division of Materials Science and Engineering, and the CINT user program. Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC., a wholly owned subsidiary of Honeywell International, Inc., for the U.S. Department of Energys National Nuclear Security Administration under contract DE-NA0003525. The work at NRL was funded by the Office of Naval Research through Base Programs at NRL. The authors thank Jyoti Katoch and Sren Ulstrup for help with sample preparation and nanoARPES measurements (WS_2 on BN).

7 Conflicts of Interest

The authors declare no conflict of interests.

8 Code Availability

The algorithms used in this work are available to all in academic and research fields who request it from the authors.

References

[1] Fang Ren, Logan Ward, Travis Williams, Kevin J. Laws, Christopher Wolverton, Jason Hattrick-Simpers, and Apurva Mehta. Accelerated discovery of metallic glasses through iteration of machine learning and high-throughput experiments. *Science Advances*, 4(4), 2018.

[2] Yuki K. Wakabayashi, Takuma Otsuka, Yoshiharu Krockenberger, Hiroshi Sawada, Yoshitaka Taniyasu, and Hideki Yamamoto. Machine-learning-assisted thin-film growth: Bayesian optimization in molecular beam epitaxy of srruo3 thin films. *APL Materials*, 7(10):101114, 2019.

[3] B. Wang, K. Yager, D. Yu, and M. Hoai. X-ray scattering image classification using deep learning. In *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 697–704, 2017.

[4] 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*, 9(2):586592, 2019.

[5] 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(1):11809, Aug 2019.

[6] Ghanshyam Pilania, Chenchen Wang, Xun Jiang, Sanguthevar Rajasekaran, and Ramamurthy Ramprasad. Accelerating materials property predictions using machine learning. *Scientific Reports*, 3(1):2810, Sep 2013.

[7] Evgheni Strelcov, Alexei Belianinov, Ying-Hui Hsieh, Stephen Jesse, Arthur P. Baddorf, Ying-Hao Chu, and Sergei V. Kalinin. Deep data analysis of conductive phenomena on complex oxide interfaces: Physics from data mining. *ACS Nano*, 8(6):6449–6457, Jun 2014.

[8] Anubhav Jain, Shyue Ping Ong, Geoffroy Hautier, Wei Chen, William Davidson Richards, Stephen Dacek, Shreyas Cholia, Dan Gunter, David Skinner, Gerbrand Ceder, and Kristin A. Persson. Commentary: The materials project: A materials genome approach to accelerating materials innovation. *APL Materials*, 1(1):011002, 2013.

[9] Sergei V. Kalinin, Bobby G. Sumpter, and Richard K. Archibald. Big–deep–smart data in imaging for guiding materials design. *Nature Materials*, 14(10):973–980, Oct 2015.

[10] Lei Wang. Discovering phase transitions with unsupervised learning. *Phys. Rev. B*, 94:195105, Nov 2016.

[11] 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(1):1325, Jan 2020.

[12] Maxim Ziatdinov, Dohyung Kim, Sabine Neumayer, Rama K. Vasudevan, Liam Collins, Stephen Jesse, Mahshid Ahmadi, and Sergei V. Kalinin. Imaging mechanism for hyperspectral scanning probe microscopy via gaussian process modelling. *npj Computational Materials*, 6(1):21, Mar 2020.

[13] Dhiren K. Pradhan, Shalini Kumari, Evgheni Strelcov, Dillip K. Pradhan, Ram S. Katiyar, Sergei V. Kalinin, Nouamane Laanait, and Rama K. Vasudevan. Reconstructing phase diagrams from local measurements via gaussian processes: mapping the temperature-composition space to confidence. *npj Computational Materials*, 4(1):23, Apr 2018.

[14] B. Meredig, A. Agrawal, S. Kirklin, J. E. Saal, J. W. Doak, A. Thompson, K. Zhang, A. Choudhary, and C. Wolverton. Combinatorial screening for new materials in unconstrained composition space with machine learning. *Phys. Rev. B*, 89:094104, Mar 2014.

[15] Eric Schulz, Maarten Speekenbrink, and Andreas Krause. A tutorial on gaussian process regression with a focus on exploration-exploitation scenarios. *bioRxiv*, 2017.

[16] A Damascelli. Probing the Electronic Structure of Complex Systems by ARPES. *Physica Scripta*, T109:61–74, 2004.

[17] Eli Rotenberg and Aaron Bostwick. microARPES and nanoARPES at diffraction-limited light sources: opportunities and performance gains. *Journal of Synchrotron Radiation*, 21(5):1048–1056, Sep 2014.

[18] Donghui Lu, Inna M. Vishik, Ming Yi, Yulin Chen, Rob G. Moore, and Zhi-Xun Shen. Angle-resolved photoemission studies of quantum materials. *Annual Review of Condensed Matter Physics*, 3(1):129167, Jan 2012.

[19] Haifeng Yang, Aiji Liang, Cheng Chen, Chaofan Zhang, Niels B. M. Schroeter, and Yulin Chen. Visualizing electronic structures of quantum materials by angle-resolved photoemission spectroscopy. *Nature Reviews Materials*, 3(9):341353, 2018.

[20] Baiqing Lv, Tian Qian, and Hong Ding. Angle-resolved photoemission spectroscopy and its application to topological materials. *Nature Reviews Physics*, 1(10):609626, 2019.

[21] Chris Ding and Xiaofeng He. K-means clustering via principal component analysis. In *Proceedings of the Twenty-First International Conference on Machine Learning*, ICML 04, page 29, New York, NY, USA, 2004. Association for Computing Machinery.

[22] Rachel Mak, Mirna Lerotic, Holger Fleckenstein, Stefan Vogt, Stefan M. Wild, Sven Leyffer, Yefim Sheynkin, and Chris Jacobsen. Non-negative matrix analysis for effective feature extraction in x-ray spectromicroscopy. *Faraday Discuss.*, 171:357–371, 2014.

[23] C. J. Long, D. Bunker, X. Li, V. L. Karen, and I. Takeuchi. Rapid identification of structural phases in combinatorial thin-film libraries using x-ray diffraction and non-negative matrix factorization. *Review of Scientific Instruments*, 80(10):103902, 2009.

[24] Maxim Ziatdinov, Artem Maksov, Li Li, Athena S Sefat, Petro Maksymovych, and Sergei V Kalinin. Deep data mining in a real space: separation of intertwined electronic responses in a lightly doped BaFe2as2. *Nanotechnology*, 27(47):475706, oct 2016.

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

[26] Ching-Shui Cheng. Orthogonal arrays with variable numbers of symbols. *The Annals of Statistics*, pages 447–453, 1980.

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

[28] 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. *arXiv preprint arXiv:2006.02489*, 2020.

[29] Trupti Kodinariya and P.R. Makwana. Review on determining of cluster in k-means clustering. *International Journal of Advance Research in Computer Science and Management Studies*, 1:90–95, 01 2013.

[30] 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.

[31] P R Wallace. The Band Theory of Graphite. *Phys. Rev.*, 71(9):622–634, 1947.

[32] F. Ji, D. B. Durham, A. M. Minor, P. Musumeci, J. G. Navarro, and D. Filippetto. Ultrafast relativistic electron nanoprobes. *Communications Physics*, 2(1):54, May 2019.