ARTICLE   OPEN

Check for updates

Imaging mechanism for hyperspectral scanning probe microscopy via Gaussian process modelling

Maxim Ziatdinov[1,2], Dohyung Kim[3], Sabine Neumayer [1], Rama K. Vasudevan[1], Liam Collins[1], Stephen Jesse[1], Mahshid Ahmadi[3] and Sergei V. Kalinin[1 ✉]

We investigate the ability to reconstruct and derive spatial structure from sparsely sampled 3D piezoresponse force microcopy data, captured using the band-excitation (BE) technique, via Gaussian Process (GP) methods. Even for weakly informative priors, GP methods allow unambiguous determination of the characteristic length scales of the imaging process both in spatial and frequency domains. We further show that BE data set tends to be oversampled in the spatial domains, with ~30% of original data set sufficient for high-quality reconstruction, potentially enabling faster BE imaging. At the same time, reliable reconstruction along the frequency domain requires the resonance peak to be within the measured band. This behavior suggests the optimal strategy for the BE imaging on unknown samples. Finally, we discuss how GP can be used for automated experimentation in SPM, by combining GP regression with non-rectangular scans.

INTRODUCTION

Over the past three decades, scanning probe microscopy (SPM) has emerged as a primary tool for characterization of structure and functionality at the nanometer and atomic scales. Following the initial demonstration of topographic imaging in contact mode[1], the intermittent and non-contact topographic imaging modes, as well as SPM modes for electrical[2], magnetic[3], mechanical[4], and electromechanical imaging[5,6] and spectroscopy[7] followed. This rapid development of imaging modalities implemented on these benchtop tools have opened the nanoworld for exploration and was one of the key elements underpinning the explosive growth of nanoscience, nanotechnology, and many areas of fundamental and applied research[8].

All SPM modes rely on the concept of the local probe interacting with the sample surface. Scanning the probe allows sequential measurements of materials response over a spatial grid, giving rise to images or hyperspectral images, i.e., multidimensional SPMs or SPM spectroscopic imaging modes. The data acquisition process in SPM can hence be represented as the combination of two elements—the SPM engine, or the excitation/detection scheme, and the spectroscopic mode. The spectroscopic modes define the sampling of the parameter space of interest at each spatial location, for example, classical force–distance and current–voltage curve mappings in atomic force microscopy[9] and scanning tunneling microscopy, and complex time and voltage spectroscopies in piezoresponse force microscopy (PFM) and electrochemical strain microscopy[10–12]. In modern SPMs, the parameter space is usually sampled sequentially, albeit this limitation is not rigid.

The SPM engine defines the response at a single point in the parameter space, i.e., the nature of response at a single voxel. Classical SPM engines typically employ a combination of sinusoidal excitation and lock-in detection, e.g., as used in classical amplitude detection SPM, or a combination of sinusoidal excitation with the phase-locked loop detection in frequency detection schemes. These two engines yield (multimodal) scalar information, i.e., two response values per pixel. Examples of vector detection engines are the band excitation (BE)[13], exciting and detecting multiple frequencies in parallel, and harmonic inter-modulation methods[14] that detect the mixing harmonics between the two excitation signals. In both cases, the engine compresses the data stream from the detector to a single parameter or a set of parameters corresponding to response vector components. Finally, more complex detection schemes such as G-mode SPM[15–17] are based on detection and storage of the full data stream from the detector, obviating the data compression stage during scanning, after which in-depth data analysis can be performed. Both BE, and intermodulation and G-mode detection were extended to a broad variety of SPM modes, including topographic imaging, magnetic and electrostatic force imaging[18], PFM[19], and is expected to be universally applicable to all SPM modes.

The proliferation of BE, intermodulation, and G-Mode SPMs and their nascent adoption by the commercial vendors necessitates understanding basic image formation mechanisms in these techniques. To date, most of such analyses were based on the physics-based models, where the known (or postulated) physics of the imaging process was used to transform the high-dimensional data to a number of reduced, ideally material-specific (i.e., independent of imaging system), parameters. The examples of such analysis include the simple harmonic oscillator (SHO) fit in the BE methods[20] or reconstruction of the intermodulation harmonics in intermodulation atomic force microscopy[14,21]. Here the measured resonance frequencies and force–distance curves describing salient features of tip–surface interaction can then be transformed into effective Young moduli.

At the same time, of interest is the amount of information contained in the multidimensional data as determined from a purely information theory viewpoint, as well as approaches to compress and visualize it for exploratory data analysis. This

[1]Center for Nanophase Materials Sciences, Oak Ridge National Laboratory, Oak Ridge, TN 37831, USA. [2]Computational Sciences and Engineering Division, Oak Ridge National Laboratory, Oak Ridge, TN 37831, USA. [3]Joint Institute for Advanced Materials, Department of Materials Science and Engineering, University of Tennessee, Knoxville, TN 37996, USA. ✉email: sergei2@ornl.gov

information will provide both insight into the fundamentals of imaging mechanisms and tip–surface interaction, can be correlated with materials' structure to yield insight into materials' behaviors, and suggest strategies for automated experimentation. Previously, hyperspectral data sets were explored using multivariate statistical methods such as principal component analysis[22], more complex methods such as non-negative matrix factorization that allow for certain physics-based constraints[23], or non-linear autoencoders[24]. However, these methods are based exclusively on the analysis of the spectral dimensions, whereas spatial correlations are explicitly ignored. In other words, the endmembers of spectral unmixing do not depend on the relative positions of the spatial pixels. Correspondingly, analysis of spatial features in the loading maps was used as a way to infer the understanding of the system[23]. Alternatively, neural network-based algorithms were suggested as an approach to identify the data based on labeled examples[25] or to extract the parameters of a theoretical model[26,27]. However, in these cases as well analysis was essentially single-pixel based.

Here we explore the imaging mechanisms in the BE PFM using the Gaussian process (GP) regression[28–32]. This method allows exploration of the data structure from purely information theory perspective simultaneously in the spatial and parameter space. We apply this Bayesian machine learning approach to determine the characteristic length scale of the phenomena and information content in the hyperspectral images and suggest the strategies for automated experimentation based on exploiting sampling of space where maximum uncertainty is predicted.

RESULTS AND DISCUSSION

BE PFM on BiFeO₃ film

The BE PFM measurements were performed on an Asylum Cypher microscope using an in-house built BE controller. The BE PFM is based on exciting and monitoring the cantilever/sample response within a continuous frequency band (instead of a single frequency as used in a classical set-up) using parallel multifrequency excitation and detection. This allows quantitative measurements via decoupling the material response from the changes in resonance frequency of the tip–surface junction.

As a model system, we use an epitaxial $BiFeO_3$ thin film of 100-nm thickness. The characteristic surface topography is shown in Fig. 1a. The BE data can be fitted to a SHO model to extract amplitude, phase, resonance frequency, and Q-factor maps as shown in Fig. 1b–e. These images clearly indicate a ferroelectric domain structure as expected for this material. Figure 1f shows the zoomed-in amplitude image at the top left corner of Fig. 1b.

GP for reconstruction of BE PFM data

To explore the information distribution in the BE data set, we employ the GP method[31,32]. GP generally refers to an indexed collection of random variables, any finite number of which have a joint Gaussian distribution. More simply, a GP is a distribution over functions on a given domain and can be used for approximating continuous nonlinear functions. GP can be completely specified in terms of its mean function and covariance function (also referred to as kernel function). A common application of GP in machine learning is a GP regression analysis where one estimates an unknown function given noisy observations $\mathbf{y} = (y(\mathbf{x}_1), \ldots, y(\mathbf{x}_n))^T$ of the function at a finite number of points $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$. The marginal likelihood of the data is conditioned on the kernel hyperparameter θ as following:

$$\log p(\mathbf{y}|X, \theta) \propto -\frac{1}{2}\log|\mathbf{K}_\theta + \sigma^2 \mathbf{I}| - \frac{1}{2}\mathbf{y}^T(\mathbf{K}_\theta + \sigma^2 \mathbf{I})^{-1}\mathbf{y} \quad (1)$$

where the first and second terms can be interpreted as the hyperparameter learning and the data fit, respectively[31]. Here σ^2 is noise variance. Because the application of GP models to large data sets is intractable, we adapt a sparse GP regression method for constructing an approximation using a subset of observations called inducing points[33]. The inducing points are optimized together with kernel hyperparameters during model training. Maximizing the number of inducing points generally yields more accurate results, albeit at the cost of computation time and memory.
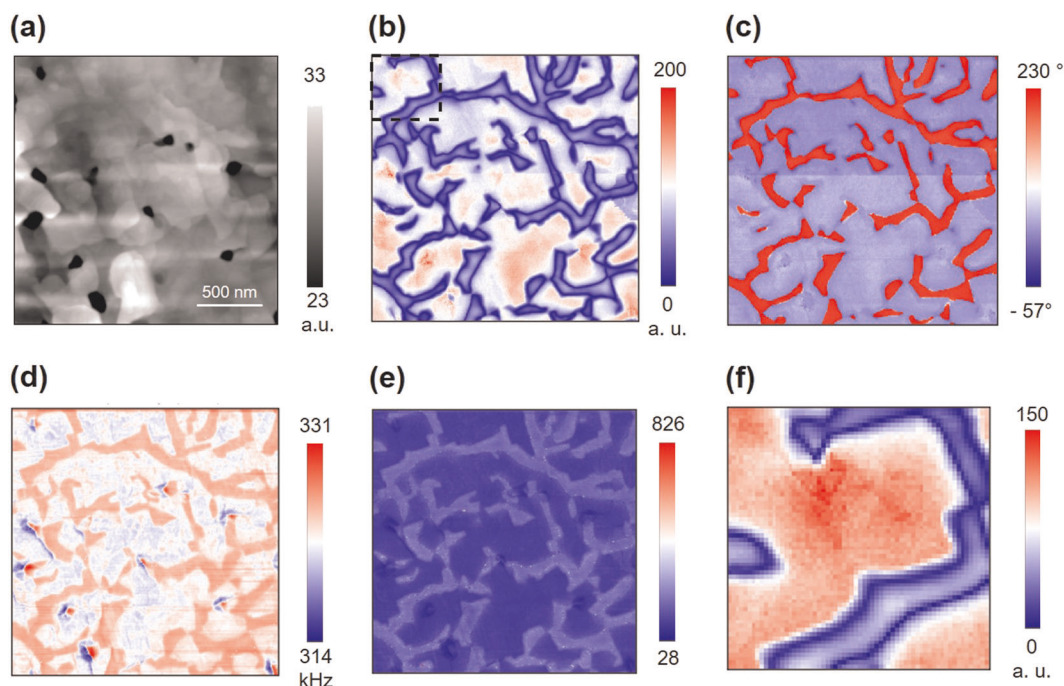
Fig. 1   BE-PFM images in the BiFeO₃ epitaxial film. a Topographic map. **b** BE-PFM amplitude, **c** phase, **d** resonance frequency, and **e** Q-factor maps. **f** Zoomed-in amplitude image at the top left corner of **b**.
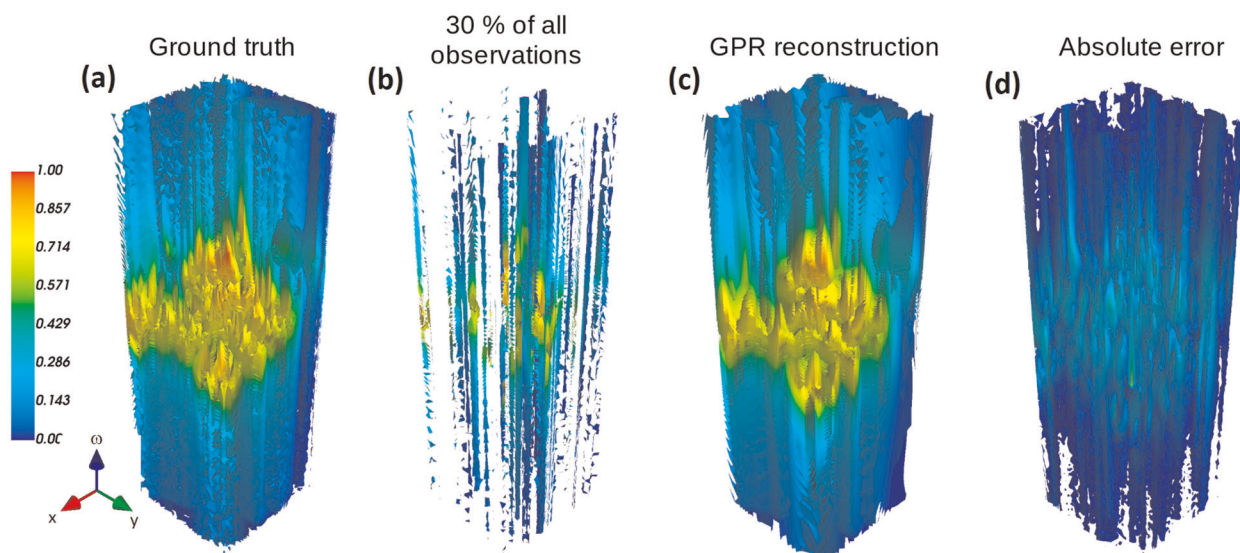
Fig. 2 Illustration of GP application to 3D spectroscopic data set. Original image (ground truth) in **a** is corrupted by removing 70% of observations (measured spectroscopic curves) as shown in **b**. The GP regression is then used to reconstruct the signal (**c**). The absolute error is shown in **d**. Note that the absolute error may be misleading when the ratio of signals of interest does not change (see Fig. 3). The data set dimensions are $32 \times 32 \times 102$.

It is important to know that the signature aspect of GP method is that observations at different locations are assumed to be linked via the kernel function, defining the connection between the dissimilar locations. The kernel function can be either defined a priori, i.e., from the known physics of the system or additional information, or can be treated as a hyperparameter. In the latter case, the functional form of the kernel is defined and the corresponding parameters are determined as a part of the fitting process. Here we note that the kernel parameters determined self-consistently as a part of the regression process should provide robust information on the image formation mechanism in the technique and explore this proposition below.

To explore the applicability of the GP processing in BE, we first demonstrate its potential for reconstruction from *partial* data and subsequently show how exploratory analysis using information about maximum uncertainty as a guide for selecting the next measurement point can be performed. We selected a $32 \times 32 \times 102$ subset of the original hyperspectral data (see Fig. 1f) in order to reduce computational time/costs and to make it easier to reproduce our results without a need to use high-performance computing. Our GP analysis code tailored toward the analysis of two- and three-dimensional (2D and 3D, respectively) image and spectroscopic data is based on Pyro probabilistic programming language[32]. We used a fixed number (1500) of inducing points (determined based on the limits of graphics processing unit (GPU) memory) for reconstruction of individual data sets, while for the "sample exploration" problem the number of inducing points was set to 5% of the overall data points. We also provide an executable Jupyter notebook for reproducing the paper's results (available at https://colab.research.google.com/github/ziatdinovmax/GP/blob/master/notebooks/GP_BEPFM.ipynb). The notebook can be executed either using a standard Google Cloud Platform virtual machine with NVIDIA's Tesla P100 GPU and 15 GB of RAM (running the notebook one time from top to bottom costs ~2 USD) or in Google Colab with NVIDIA's Tesla K80 GPU, which is free of charge but may require significantly longer computational times.

Here the original BE data set is considered to be the "known" ground truth. A part of the data is removed, creating the artificial data set. The GP regression is used to reconstruct the full data set, and the reconstruction error is evaluated. This process is illustrated in Fig. 2, showing the original data set as a response in the x, y, and frequency space (Fig. 2a), the reduced data set (Fig. 2b), the reconstructed data set (Fig. 2c), and the absolute error (Fig. 2d). Here the "Matern52" kernel is used[31]. The reconstruction error shown in Fig. 2d generally does not exceed ~20% despite the fact that 70% of the data was eliminated.

To explore the robustness of this approach, we explore the veracity of reconstruction for the various degrees of image reduction. The 2D representation of the reconstruction of the BE data set for the full data set and for the data sets with removal of 70%, 90%, and 99% of the original data is shown in Fig. 3. Even for the data reduction by 90%, the reconstructed data sets maintain the characteristic features of the response, including both the general domain configuration and the behavior of the amplitude–frequency curves. Note that GP process yields not only estimated response values but also the confidence intervals at the given point in the parameter space, thus allowing for the formulation of optimal strategies for experiment automation, as will be explored later. Overall, we conclude that the BE data are strongly oversampled, and even without strong physics-specific priors, the GP process allows reconstruction of data from partial observations.

We further note that the key element of GP process is that it yields the insight into the structure of the data via the kernel parameters. Here we use the weakly informative 3D Matern kernel, with the characteristic length scales determined as a part of regression process. These length scales hence define the characteristic resolution in the spatial and frequency domains and do so in robust (with respect to noise) fashion. These behaviors are illustrated in Fig. 4a–d for the full data set and for the partially reduced data sets discussed in Fig. 3. Note that, for all cases discussed in Fig. 3, the frequency length scale converges to the similar value given by the width of the BE peak. For spatial length scales, the analysis of the full data set allows to establish characteristic spatial resolution as the kernel length scale. This length scale in turn provides robust estimate of the characteristic length scale of tip–surface interactions. We also showed trajectories of inducing points for each case, which were selected as a subset of the input data points by taking every nth point from the original data set. Here n was adjusted such that the total number of inducing points remained the same (1500, based on
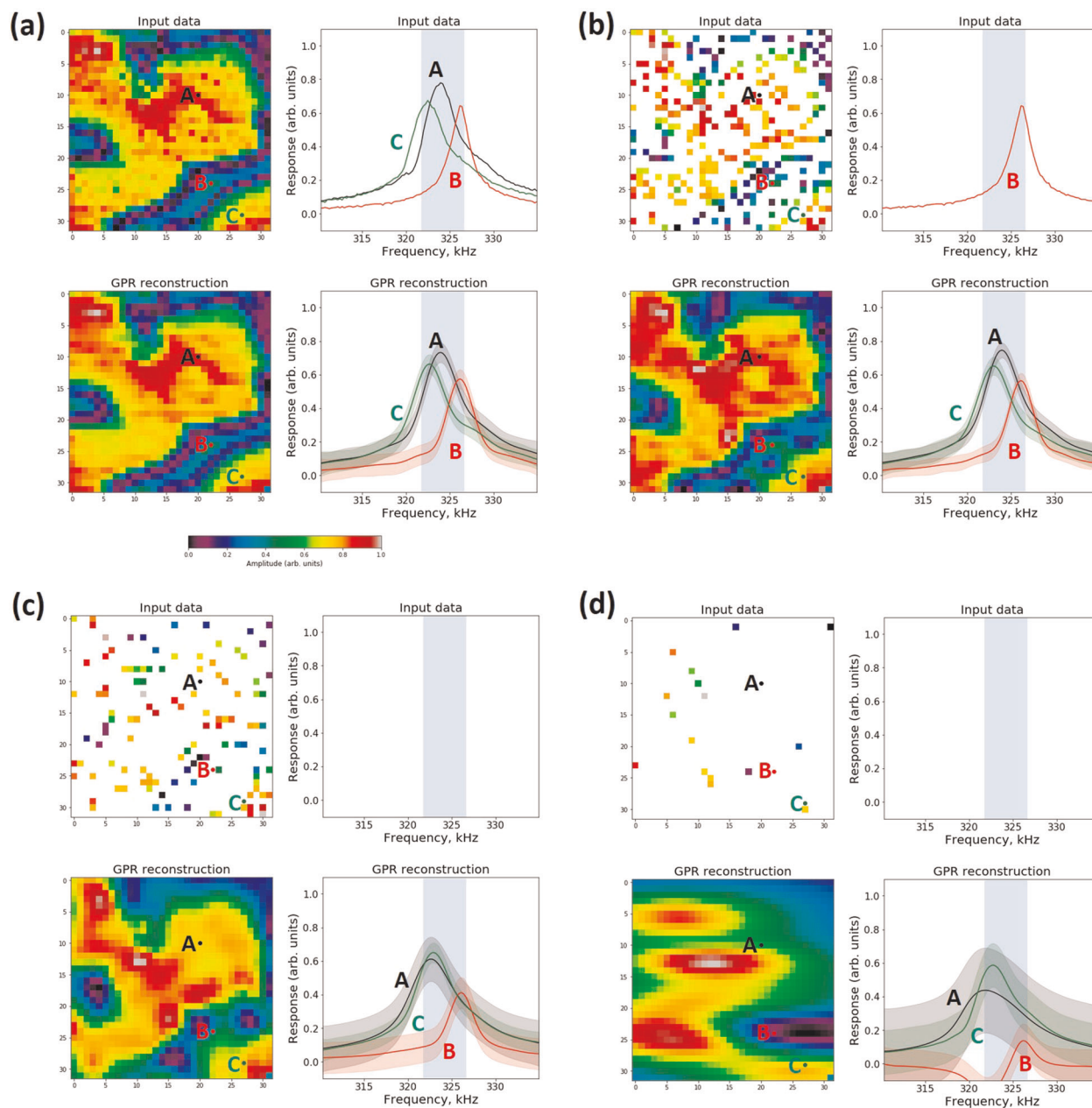
Fig. 3 Illustration of GP reconstruction on selected 2D data slices. Selected 2D representations corresponding to the application of GP regression to original (full) data (**a**) and data corrupted by removing 70% (**b**), 90% (**c**), and 99% (**d**) of observations. In each panel, the top row shows the input data (2D slices of hyperspectral data average over selected frequency range and spectroscopic curves extracted at A, B, and C points) and the bottom row shows the model output (the same averaged slices and the reconstructed curves from the same locations as in the input data). The vertical blue span in the plots indicate a slice used for 2D image plots (the same in all four panels). The spatial resolution in all the images is ~9.76 nm/px. The full 3D data sets are available from the notebook at https://colab.research.google.com/github/ziatdinovmax/GP/blob/master/notebooks/GP_BEPFM.ipynb.

GPU memory limits) for all 4 scenarios. Notice that most inducing points in Fig. 4e–g stop their motion after ~500 stochastic variational inference iterations and that for data set with 99% of observations removed all the inducing points remain mostly at their original location. While the inducing points do not have a physical meaning, understanding their evolution during the optimization process is important for applying GP models to larger hyperspectral data sets and to the experiment automation.

We further explored the robustness of the reconstruction with respect to the choice of the frequency range. In this case, the high-quality reconstruction is possible when the local resonance frequency is within the measurement range but start to

deteriorate rapidly once outside (Supplementary Figs. 1 and 2). These observations suggest that introduction of GP in the imaging workflow favors the mapping over large frequency ranges but at the reduced spatial point densities. However, we expect that incorporation of physics-based kernels containing the information on the SHO model will allow to expand this approach further.

Uncertainty-guided sample exploration

We finally demonstrate that GP regression can be used for guiding the actual measurements. Here we start with just a few measured points along each edge in the *xy* plane (~1% of all the observations). A single exploration step consists of (i) performing
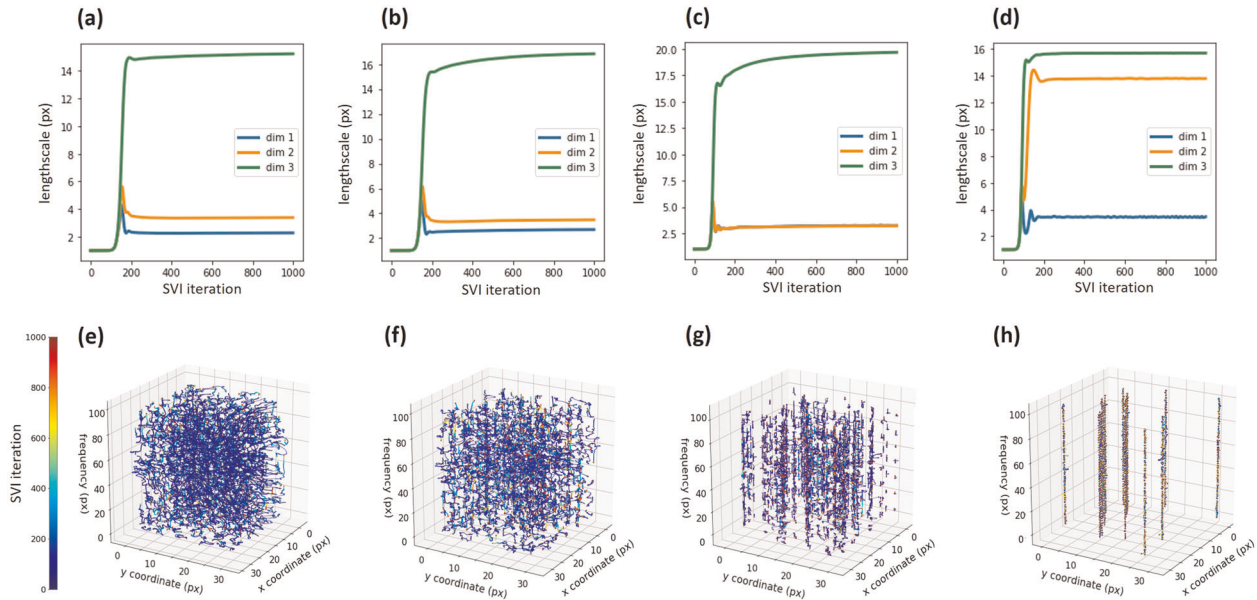
Fig. 4 Evolution of GP parameters during training. Evolution of kernel length scales (**a–d**) and inducing points (**e–h**) during the stochastic variational inference (SVI)-based model training for full data set (**a, e**), and data corrupted by removing 70% of observations (**b, f**), 90% of observations (**c, g**), and 99% of observations (**d, h**). **a** The first two dimensions in **a–d** (dim 1 and dim 2) correspond to x and y coordinates, whereas the third dimension (dim 3) corresponds to frequency. The kernel length scales define the spatial resolution of the technique (assuming atomically thin domain wall width) in the spatial domain and the width of the resonance peak in the frequency domain.
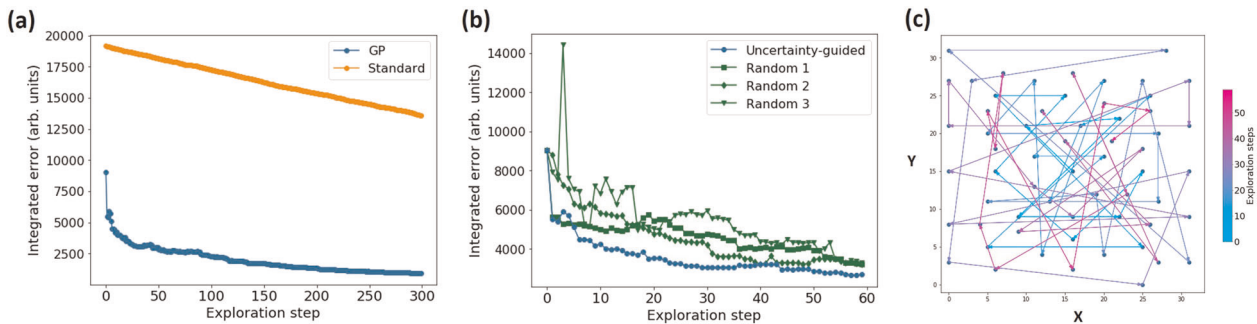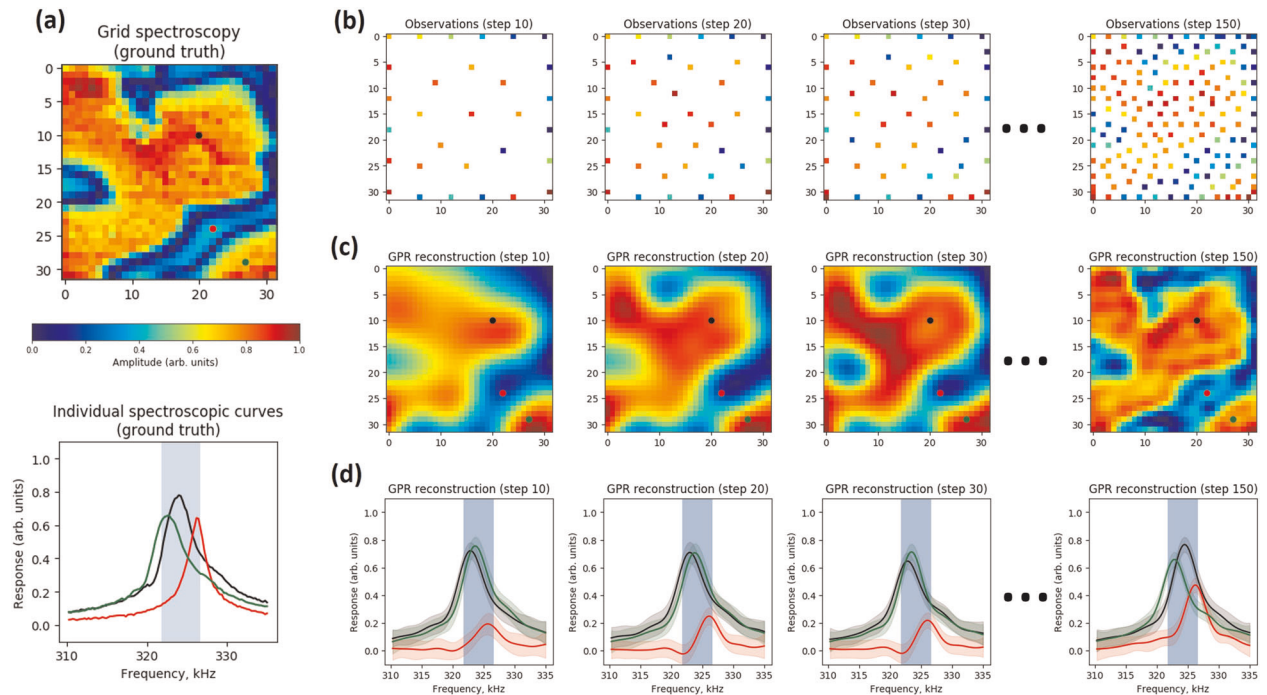

Fig. 5 Simulation of the GP-guided experimental measurements. The data are real (experimental). **a** The integrated absolute error versus exploration steps for measurements with (GP) and without (standard) GP regression-based reconstruction. **b** Comparison of the integrated absolute error when the next measurement point is selected using the maximal uncertainty in GP reconstruction versus when the next measurement point is selected randomly (for three different pseudo-random seeds). **c** Exploration path in xy coordinates for the first 60 steps.

a GP regression, (ii) doing a single "measurement" in a point with maximum uncertainty, and (iii) updating the previous input data with the data points associated with this measurement. The GP part of this process can be viewed as a "black box" that is effectively separated from the BE PFM acquisition software. It accepts the datacube of a fixed size containing the BE PFM response values (the size is determined by the scan size and frequency range) and the associated grid indices as the input and outputs the indices of the next point to be measured based on the GP uncertainty estimation. The points that are not yet measured are represented as NaNs in the datacube and are updated with the BE PFM response values once measured. The optimal frequency range (the third dimension of the datacube) is selected based on a prior knowledge of an operator and remains fixed during each measurement cycle.

Here we demonstrate this approach using a "synthetic" experiment, that is, we use the data from the actual but already completed experiment, with 99% of all the observations removed, which allows us to estimate the absolute error at each step. As one can observe from Fig. 5a, the absolute error is rapidly decreasing during the first ~30 exploration steps. The error is much higher

when the next measurement point is selected just randomly, as demonstrated in Fig. 5b. We noticed that, if the edge points are not "opened" in the beginning, the algorithm typically spends the first ~10 iterations (for this particular data, dimensions) on measuring the edges since this is where the kernel diverges (and not because this is related to any sample properties) and then moves to the regions deeper inside the field of view. The selected inputs and outputs for this exploration range are shown in Fig. 6. Interestingly, one can get a good understanding of the sample domain structure already after 20–30 steps of such an autonomous experiment, which suggests that this approach can significantly reduce the data acquisition time. The remaining steps are typically spent on "refining" the uncovered structures (compare, for example, the third and the last column in Fig. 6b–d). The current approach can be improved by introducing a "cost function" determining which objects are of real physical interest (that, is "worth exploring"), in addition to a pure uncertainty-based exploration.

To summarize, we have explored the applicability of the GP regression with weakly informative priors for the analysis of the BE PFM data. Here we explored the signal in the 3D (x, y, frequency)

Fig. 6 Simulation of the GP-guided experimental measurements. The data are real (experimental). **a** Ground truth data in 2D representation for the selected frequency range and the individual spectroscopic curves from the selected locations. **b** Input data updated with a new single measurement at each step. **c** GP regression-based reconstruction of the entire field of view, integrated over a frequency range shown by a vertical blue span in Fig. 3. **d** Reconstructed individual spectroscopic curves from the locations shown by black, red, and green dots in **c**.

parameter space. Even for the weakly informative priors, the GP approach allows to unambiguously determine the characteristic length scales of the imaging process both in spatial and frequency domains. We further show that BE data set tend to be oversampled, with ~30% of original data set sufficient for high-quality reconstruction.

We further note that this analysis points at strong potential of GP for the development of the automated experiments, where the measurement points are chosen based on the results of previous measurements[34–36]. Here the Bayesian uncertainty along with the target-driven criteria can be used for balancing of the exploratory and exploitation activity. Furthermore, we believe that the analysis can be strongly improved with the addition of physics-based priors to reconstruction, where the kernel function incorporates the partially known physics of material (e.g., atomically resolved periodic features, sharp boundaries, Green's function for known geometric domain, etc.) or imaging process (resolution function).

DATA AVAILABILITY

CODE AVAILABILITY

REFERENCES

1. Binnig, G., Quate, C. F. & Gerber, C. Atomic force microscope. *Phys. Rev. Lett.* **56**, 930–933 (1986).
2. Nonnenmacher, M., Oboyle, M. P. & Wickramasinghe, H. K. Kelvin probe force microscopy. *Appl. Phys. Lett.* **58**, 2921–2923 (1991).
3. Martin, Y. & Wickramasinghe, H. K. Magnetic imaging by force microscopy with 1000-A resolution. *Appl. Phys. Lett.* **50**, 1455–1457 (1987).
4. Rabe, U. & Arnold, W. Acoustic microscopy by atomic-force microscopy. *Appl. Phys. Lett.* **64**, 1493–1495 (1994).
5. Franke, K., Besold, J., Haessler, W. & Seegebarth, C. Modification and detection of domains on ferroelectric Pzt films by scanning force microscopy. *Surf. Sci.* **302**, L283–L288 (1994).
6. Kolosov, O., Gruverman, A., Hatano, J., Takahashi, K. & Tokumoto, H. Nanoscale visualization and control of ferroelectric domains by atomic force microscopy. *Phys. Rev. Lett.* **74**, 4309–4312 (1995).
7. Jesse, S., Baddorf, A. P. & Kalinin, S. V. Switching spectroscopy piezoresponse force microscopy of ferroelectric materials. *Appl. Phys. Lett.* **88**, 062908 (2006).
8. Gerber, C. & Lang, H. P. How the doors to the nanoworld were opened. *Nat. Nanotechnol.* **1**, 3–5 (2006).
9. Garcia, R. & Perez, R. Dynamic atomic force microscopy methods. *Surf. Sci. Rep.* **47**, 197–301 (2002).
10. Vasudevan, R. K., Jesse, S., Kim, Y., Kumar, A. & Kalinin, S. V. Spectroscopic imaging in piezoresponse force microscopy: new opportunities for studying polarization dynamics in ferroelectrics and multiferroics. *MRS Commun.* **2**, 61–73 (2012).
11. Balke, N. et al. Decoupling electrochemical reaction and diffusion processes in ionically-conductive solids on the nanometer scale. *Acs Nano* **4**, 7349–7357 (2010).
12. Jesse, S., Maksymovych, P. & Kalinin, S. V. Rapid multidimensional data acquisition in scanning probe microscopy applied to local polarization dynamics and voltage dependent contact mechanics. *Appl. Phys. Lett.* **93**, 112903 (2008).
13. Jesse, S., Kalinin, S. V., Proksch, R., Baddorf, A. P. & Rodriguez, B. J. The band excitation method in scanning probe microscopy for rapid mapping of energy dissipation on the nanoscale. *Nanotechnology* **18**, 435503 (2007).
14. Platz, D., Tholen, E. A., Pesen, D. & Haviland, D. B. Intermodulation atomic force microscopy. *Appl. Phys. Lett.* **92**, 3 (2008).
15. Belianinov, A., Kalinin, S. V. & Jesse, S. Complete information acquisition in dynamic force microscopy. *Nat. Commun.* **6**, 6550 (2015).
16. Collins, L. et al. G-mode magnetic force microscopy: separating magnetic and electrostatic interactions using big data analytics. *Appl. Phys. Lett.* **108**, 193103 (2016).
17. Collins, L. et al. Full data acquisition in Kelvin probe force microscopy: mapping dynamic electric phenomena in real space. *Sci. Rep.* **6**, 30557 (2016).
18. Collins, L. et al. Breaking the time barrier in Kelvin probe force microscopy: fast free force reconstruction using the G-mode platform. *ACS Nano* **11**, 8717–8729 (2017).

19. Somnath, S., Belianinov, A., Kalinin, S. V. & Jesse, S. Full information acquisition in piezoresponse force microscopy. *Appl. Phys. Lett.* **107**, 263102 (2015).

20. Jesse, S. & Kalinin, S. V. Band excitation in scanning probe microscopy: sines of change. *J. Phys. D Appl. Phys.* **44**, 464006 (2011).

21. Forchheimer, D., Platz, D., Tholen, E. A. & Haviland, D. B. Model-based extraction of material properties in multifrequency atomic force microscopy. *Phys. Rev. B* **85**, 7 (2012).

22. Jesse, S. & Kalinin, S. V. Principal component and spatial correlation analysis of spectroscopic-imaging data in scanning probe microscopy. *Nanotechnology* **20**, 085714 (2009).

23. Kannan, R. et al. Deep data analysis via physically constrained linear unmixing: universal framework, domain examples, and a community-wide platform. *Adv. Struct. Chem. Imaging* **4**, 20 (2018).

24. Su, Y. et al. DAEN: deep autoencoder networks for hyperspectral unmixing. *IEEE Trans. Geosci. Remote Sens.* **57**, 4309–4321 (2019).

25. Nikiforov, M. P. et al. Functional recognition imaging using artificial neural networks: applications to rapid cellular identification via broadband electromechanical response. *Nanotechnology* **20**, 405708 (2009).

26. Ovchinnikov, O. S., Jesse, S., Bintacchit, P., Trolier-McKinstry, S. & Kalinin, S. V. Disorder identification in hysteresis data: recognition analysis of the random-bond-random-field Ising model. *Phys. Rev. Lett.* **103**, 157203 (2009).

27. Kumar, A. et al. Spatially resolved mapping of disorder type and distribution in random systems using artificial neural network recognition. *Phys. Rev. B* **84**, 024203 (2011).

28. Lambert, B. *A Student's Guide to Bayesian Statistics* 1st edn (SAGE Publications Ltd, 2018).

29. Kruschke, J. *Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan* 2nd edn (Academic Press, 2014).

30. Gelman, A. et al. *Bayesian Data Analysis (Chapman & Hall/CRC Texts in Statistical Science)* 3rd edn (Chapman and Hall/CRC, 2013).

31. Rasmussen, C. E. & Williams, C. K. I. *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)* (The MIT Press, 2005).

32. Bingham, E. et al. Pyro: deep universal probabilistic programming. *J. Mach. Learn. Res.* **20**, 973–978 (2019).

33. Quiñonero-Candela, J. & Rasmussen, C. E. A unifying view of sparse approximate Gaussian process regression. *J. Mach. Learn. Res.* **6**, 1939–1959 (2005).

34. Kim, C., Chandrasekaran, A., Jha, A. & Ramprasad, R. Active-learning and materials design: the example of high glass transition temperature polymers. *MRS Commun.* **9**, 860–866 (2019).

35. Lookman, T., Balachandran, P. V., Xue, D. & Yuan, R. Active learning in materials science with emphasis on adaptive sampling using uncertainties for targeted design. *npj Computational Mater.* **5**, 21 (2019).

36. Kusne, A. G. et al. On-the-fly machine-learning for high-throughput experiments: search for rare-earth-free permanent magnets. *Sci. Rep.* **4**, 6367 (2014).

AUTHOR CONTRIBUTIONS

S.V.K. proposed the concept and led the paper writing. M.Z. wrote the code for GP-based analysis of hyperspectral SPM data, performed the analysis, and co-wrote the paper. D.K. collected the experimental data, with assistance from S.N. and L.C. S.J. and L.C. configured the original experimental set-up. R.K.V., M.A., and S.J. assisted with data interpretation and paper writing.

COMPETING INTERESTS

The authors declare no competing interests.

ADDITIONAL INFORMATION

Supplementary information is available for this paper at https://doi.org/10.1038/s41524-020-0289-6.

Correspondence and requests for materials should be addressed to S.V.K.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.