ADVANCED MATERIALS
www.advmat.de

Why it is Unfortunate that Linear Machine Learning "Works" so well in Electromechanical Switching of Ferroelectric Thin Films

*Shuyu Qin, Yichen Guo, Alibek T. Kaliyev, and Joshua C. Agar**

Machine learning (ML) is relied on for materials spectroscopy. It is challenging to make ML models fail because statistical correlations can mimic the physics without causality. Here, using a benchmark band-excitation piezoresponse force microscopy polarization spectroscopy (BEPS) dataset the pitfalls of the so-called "better", "faster", and "less-biased" ML of electromechanical switching are demonstrated and overcome. Using a toy and real experimental dataset, it is demonstrated how linear non-temporal ML methods result in physically reasonable embedding (eigenvalues) while producing nonsensical eigenvectors and generated spectra, promoting misleading interpretations. A new method of unsupervised multimodal hyperspectral analysis of BEPS is demonstrated using long-short-term memory (LSTM) β-variational autoencoders (β-VAEs) . By including LSTM neurons, the ordinal nature of ferroelectric switching is considered. To improve the interpretability of the latent space, a variational Kullback–Leibler-divergency regularization is imposed . Finally, regularization scheduling of β as a disentanglement metric is leveraged to reduce user bias. Combining these experiment-inspired modifications enables the automated detection of ferroelectric switching mechanisms, including a complex two-step, three-state one. Ultimately, this work provides a robust ML method for the rapid discovery of electromechanical switching mechanisms in ferroelectrics and is applicable to other multimodal hyperspectral materials spectroscopies.

1. Introduction

Ferroelectric thin films are promising for next-generation electromechanical energy conversion, sensing, memory, and logic. Achieving maximum performance generally relies on perching materials near phase transitions where small external perturbations can drive large structural and associated property changes. For example, in $BiFeO_3$, epitaxial strain can cause phase competition between nearly degenerate rhombohedral and tetragonal-like phases.[1] A small electric field can drive the transition between these phases resulting in electromechanical actuation of >5%.[2] There has been growing interest in understanding collective switching mechanisms in highly correlated ferroelectric systems that exhibit novel responses.[3,4] For example, colossal electromechanical response is achievable at the morphotropic phase boundary,[5] in regions of phase competition,[2] and relaxor ferroelectrics.[6] One recent highlight was the discovery of highly tunable low-loss dielectrics near a novel region of phase competition in $BaTiO_3$.[7] This concept was extended to topological polar structures in oxide superlattices where the polarization state undergoes a continuous geometric transformation.[8,9] The result is highly correlated topological structures with a range of novel susceptibilities.

Understanding collective ferroelectric susceptibilities requires probing ferroelectric responses across length, time, and frequencies scales. Piezoresponse force microscopy (PFM) is the most common tool to image ferroelectric susceptibilities. In PFM, a conductive cantilever-mounted tip is scanned in contact with a surface while applying an oscillating drive voltage. The local piezoelectric response is qualitatively measured from the deflection of the cantilever. Various stroboscopic studies have been designed to measure collective transformations using a tip-induced or environmental change in voltage, temperature, humidity, or mechanical stress,[3] providing insights into collective transformations. This has been used to study ferroelastic transformations,[3] temperature-induced phase transitions,[10] jerky domain dynamics,[11] and much more.[12]

S. Qin, A. T. Kaliyev
Department of Computer Science and Engineering
Lehigh University
Bethlehem, PA 18015, USA
Y. Guo, J. C. Agar
Department of Materials Science and Engineering
Lehigh University
Bethlehem, PA 18015, USA
E-mail: jca92@drexel.edu
A. T. Kaliyev
College of Business
Lehigh University
Bethlehem, PA 18015, USA
J. C. Agar
Department of Mechanical Engineering and Mechanics
Drexel University
Philadelphia, PA 19104, USA

DOI: 10.1002/adma.202202814

Stroboscopic studies, however, impart minimal insight into transformation pathways and dynamics important for memory and logic, and energy conversion. To measure dynamic nanoscale ferroelectric transformation, multimodal scanning probe spectroscopies are required.[13] These techniques measure the electromechanical susceptibility using PFM while simultaneously driving ferroelectric switching. To improve the precision band-excitation (BE) is used to simultaneously excite the cantilever at a band of frequencies near the resonance frequency.[14] Digital signal processing and fitting extracts the cantilever resonance amplitude, phase, resonance frequency, and quality factor. BE, in addition to piezoresponse and phase, also acquires the resonance frequency—a qualitative measure of the elastic modulus, and the quality factor—a qualitative measure of the piezoelectric dissipation. A variety of BE modes have been developed by changing the transformation bias. For example, bipolar triangular waveforms can be used to measure ferroelectric switching (Figures S1 and S2, Supporting Information),[15] First-order reversal curves can measure switching dynamics,[16] and contact-Kelvin probe force microscopy to isolate ionic, electrostatic, ferroelectric contributions.[17–19] The challenge is extracting actionable information from data that spans many positions, frequency, voltage, cycle, and time dimensions and contains multichannel information. These challenges are exacerbated when searching for rare spatial-temporal phenomena.

Experimentalists have increasingly relied on statistical approaches based on machine learning to accelerate information extraction.[20] Depending on the objective, various machine-learning methods have emerged, including classification based on support vector machines[21] and deep neural networks,[22] dimensionality reduction based on principal component analysis,[23] non-negative matrix factorization,[24] and deep autoencoders,[25] segmentation methods based on U-Net architectures,[26] etc. Underpinning all these techniques are many simple mathematical operators optimized toward an objective. The only difference between empirical fitting and machine learning is the complexity and flexibility of the functions. Machine-learning models can approximate solutions to problems impossible using human-discovered empirical expressions. However, because machine-learning models are highly overparameterized they lack parsimony, and thus are susceptible to overfitting. There has been a push to include explicitly or learn underlying physical expressions using machine learning,[27,28] however, there are still open challenges when dealing with complex and noisy experimental data.

To combat overfitting, constraints are added to machine-learning models. These include soft constraints on the objective function (e.g., L_1, L_2, statistical, and custom regularization), damaging mechanisms (e.g., dropout[29]), bottleneck layers, and restrictive activation functions (e.g., linear, scaled-exponential linear units [SeLu],[30] etc.). Subtle changes to the model architecture can significantly alter mathematics, performance, and how physics is learned.

Open-source machine-learning packages has democratized these tools within materials science. Applying these tools as a "black box" creates justified concerns about their application in mission-critical tasks and scientific interpretability. Machine-learning models are masters of disguise; regardless of how ill-posed the model is, they tend to work remarkably well. This was highlighted in Geoffrey Hinton's famous talk "What is wrong with convolutional neural networks,[31]" where they highlight how translational invariance and max-pooling are antithetical to image understanding; however, convolutional neural networks (CNNs) which are bounded by these constraints still achieve top-5 accuracy on ImageNet Classification of >98%.[32] Maximizing interpretability requires designing model architecture that limit numerical approximations and aphysical model outcomes. For example, if the model has nonlinear trends, nonlinear models should be used. If the model has sequential information (e.g., time or voltage series), the ordinal nature should be considered. While true parsimony is usually impossible it is important that the interpretability of machine-learning methods be bounded by what can be validated.

Here, using a benchmark band-excitation piezoresponse force microscopy polarization spectroscopy (BEPS) dataset openly published by the authors in 2017[33] we demonstrate and overcome some of the common pitfalls of so-called "better", "faster", and "less biased" machine-learning-based discovery of electromechanical switching.[34] Using a toy and real experimental dataset, we observe how the application of linear nontemporal machine-learning methods to multimodal hyperspectral data results in the extraction of physically reasonable embedding (eigenvalues) while producing nonsensical eigenvectors and generated spectra promoting misleading interpretations.

We demonstrate and rigorous validate a new method of unsupervised multimodal hyperspectral analysis of BEPS using a long-short term memory (LSTM) β-variational autoencoders (β-VAE) that considers the experiments information content. By including recurrent LSTM neurons, we consider the ordinal nature of ferroelectric switching. To improve the interpretability of the latent space, we impose a variational Kullback–Leibler (KL)-divergence regularization term to the loss function. Finally, we leverage regularization scheduling of β as a disentanglement metric to reduce user bias. Combining these experiment-inspired modifications enables the automated detection of ferroelectric switching mechanisms, including a complex two-step, three-state ferroelastic switching mechanism. Ultimately, this work provides a robust machine-learning methodology for the rapid discovery of electromechanically switching mechanisms in ferroelectrics and is applicable to other multimodal hyperspectral materials spectroscopies. More broadly, this work highlights the necessity of considering the experimental methods, data, and validation methodologies when designing machine-learning methods for materials science.

2. Results and Discussion

We constructed a toy problem with known ground truth to compare machine-learning methods for hyperspectral data analysis. To do this, we established a mathematical basis based on three values which can be visualized by constructing a red-green-blue (RGB) images (**Figure 1**). For training, we generated 10 000 RGB values sampled from a uniform distribution between 0 and 1 ($R, G, B \in \mathbb{R}[0,1]$). The RGB values of this image were used to construct 10 000 spectra with 25 spectroscopic timesteps using the function

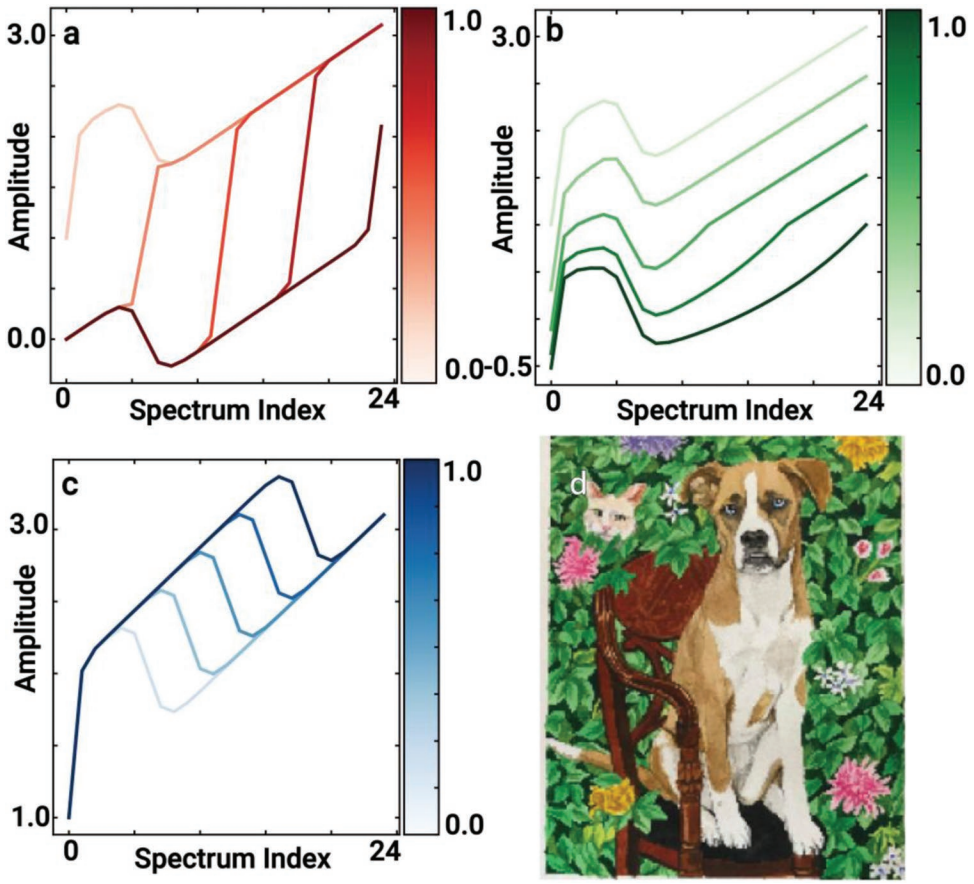

Figure 1. a–c) Generated spectra across an RGB color basis. d) Painting of Nala in the likeness of the Barak Obama Portrait. The painting is original art by Irene Dogmatic and is reproduced here with permission from the corresponding author.

$f(t,R,G,B) = \tanh(20[t - 2(R - 0.5)]) + \mathrm{SELU}(t - 2(G - 0.5)) + \mathrm{sigmoid}(-20[t - (B - 0.5)])$ where t represents the timestep. This toy dataset with a known ground truth mimics experiments of electromechanical switching of ferroelectrics. The mathematical form of the function has similar characteristics common in materials spectroscopy. Specifically, it is ordinal, has nonlinear trends, and has phase shifts. We show examples of spectra generated across the RGB color space (Figure 1a–c). Each generated spectra is drawn from a basis defined by a painting of Nala in the likeness of the Barak Obama Portrait (Figure 1d). Thus, the colors are akin to ferroelectric switching mechanisms. Building a dataset with a known ground-truth is a common strategy to benchmark machine-learning methods.[32,35,36]

We start by conducting analysis using a dictionary learning technique claimed to be "better", "faster", and "less-biased" than more computational complex deep-learning methods.[34] Dictionary learning learns a dictionary (D) of vectors and coefficients (γ) that describes a signal as a linear combination using as few columns as possible. This is achieved by solving, $\arg\min_{D,\gamma} \frac{1}{2}\| x - D\gamma \|_2^2 + \lambda \| \gamma \|_1$. In this equation, the first term represents the mean squared reconstruction loss, and the second term represents an ℓ_1-regularization term, controlled using the hyperparameter λ the degree of sparsity. When training a dictionary learning model, the user must determine or optimize the number of components (n) in the dictionary and λ. Since we know our ground-truth has three (3) endmembers representing the RGB color channels, we set $n = 3$. Before training, we scaled the data and standardized the features for each spectroscopic timestep by removing the mean value and scaling to unit variance $z = (x - u)/s$, where x represents the spectra in same spectroscopic timestep, u, and s represent the mean and standard deviation of all spectra in that spectroscopic timestep, z represents the new spectra after transformation. To optimize the hyperparameter λ, We conducted a grid-search where the lowest loss was found at $\lambda = 0.57$ (Figure S3, Supporting Information).

We validated our model by generating spectra using the RGB values from the painting of Nala. Since every spectrum is associated with a pixel position, the coefficients can be reshaped back into the original image to visualize the trends (**Figure 2**a–c) as an embedding map. This reconstruction shows that dictionary learning can deconvolute all the key features in the image. This is akin to identifying the spectroscopic differences in hyperspectral materials spectroscopy. The high-fidelity of the reconstruction map provides an illusion that dictionary learning is actually learning details of the spectroscopic data. We extracted the best, median, and worst mean squared reconstruction loss (Figure 2d–f). While dictionary learning can achieve moderate errors (MSE = 0.184) the spectroscopic trends are not learned.

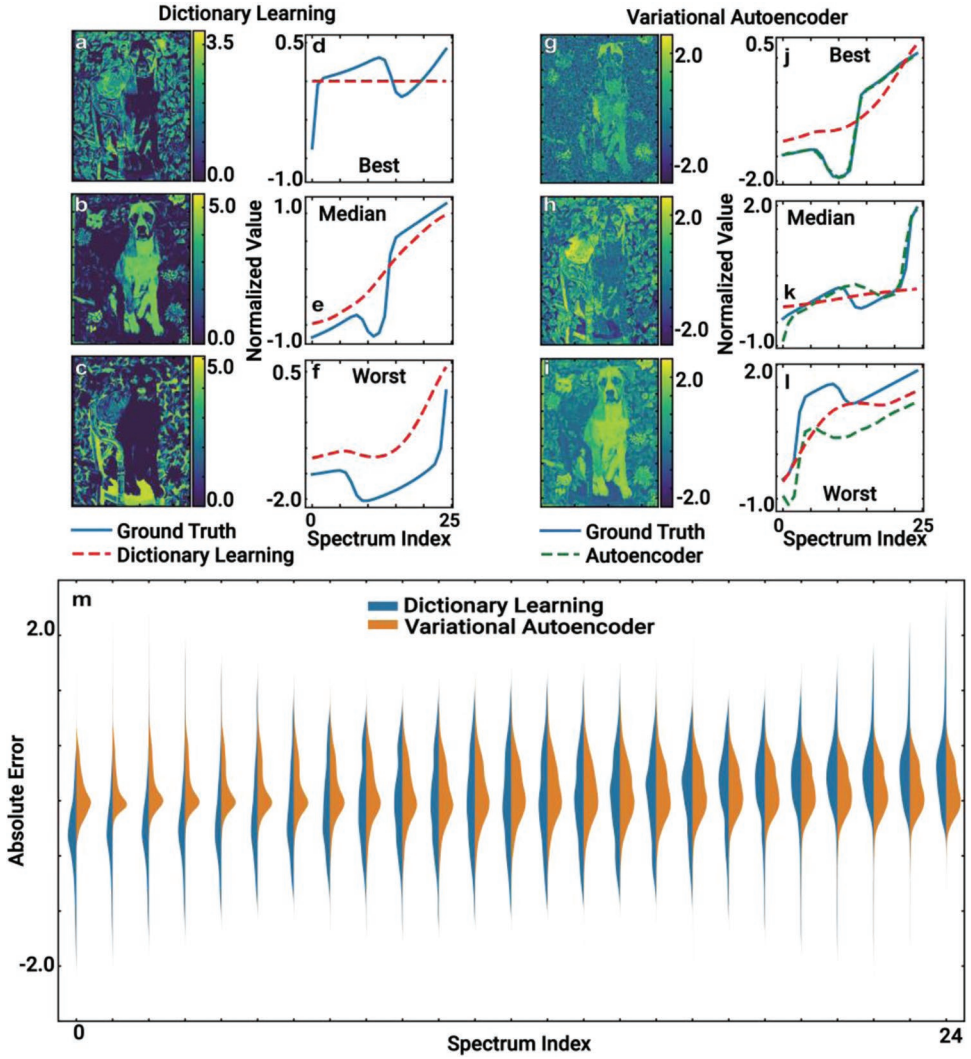

Figure 2. a–c) Dictionary learning coefficients (embeddings) extracted when validating model on spectra generated from the painting of Nala. d) Best, e) median, f) worst mean-squared error reconstruction based on dictionary learning within the validation dataset. g–i) β-VAE embeddings extracted when validating model on spectra generated from the painting of Nala. j) Best, k) median, l) worst mean-squared error reconstruction within the validation dataset. m) Violin plot comparing the absolute error of the validation predictions at each spectrum index in the sequence. The painting of the dog used in this figure is an original work by Irene Dogmatic and is reproduced here with permission of the corresponding author (see Figure 1).

Specifically, dictionary learning produces under-fit spectra that capture the global trend but misses important spectroscopic details. This is best visualized as an animation (Movie S1, Supporting Information) where we plot the generated spectra as we linearly increase each of the coefficients independently. The generated spectra in this movie show mild evidence phase shift but are utterly inept at capturing the known spectroscopic shape. This is an expected result, as the model does not consider each timestep ordinally. Thus, dictionary learning cannot explicitly include phase shift and must approximate the spectra using a linear combination of an under complete basis. This is akin to trying to make a circle out of the merger of *n*, in this case, three triangles. It is just not mathematically possible. However, as n becomes large, it can be coarsely approximated. Using an insufficient metric, the mean squared error, it is possible to get reasonable numerical result with an unacceptable practical result.

Having demonstrated that dictionary learning has unacceptable performance, we designed an alternative machine-learning approach for unsupervised spectral unmixing hyperspectral images. Our model is based on an autoencoder structure (**Figure 3**). Autoencoders consist of an encoder that learns a compact statistical representation of an input and a decoder that decodes this representation into the original spectra. The goal is to learn an identity function $f(x) = x$ that minimizes a loss function ($\mathcal{L}$). Generally, this is achieved by minimizing the mean-squared reconstruction error $\min \mathcal{L} = \frac{1}{n}\sum_{i=1}^{n}(x_i - f(x_i))^2$.

To improve the model, we made experiment-inspired architectural modifications. First, to consider the ordinal nature of spectroscopic data, the encoder and decoder are constructed of three (3) ResNet blocks, each composed of two (2) layers of 128 bidirectional long-short term memory (LSTM) neurons. LSTM neurons were chosen due to their success in various natural

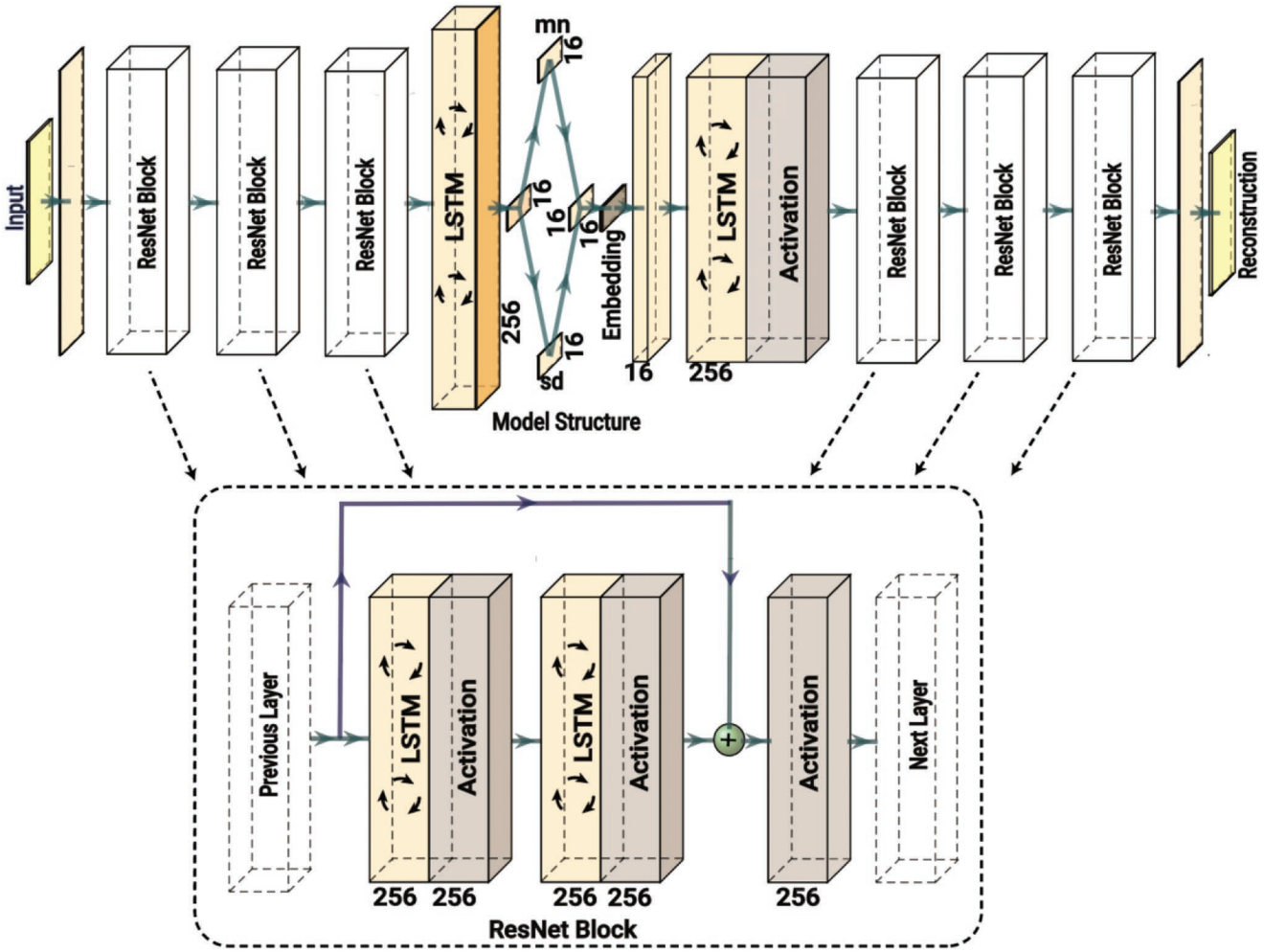

Figure 3. Schematic drawing of deep-recurrent β-variational autoencoder.

language processing and scientific tasks where the sequential nature of inputs is important.[37] LSTMs are ordinal since they process inputs recurrently through time. There is an internal learnable logic and message passing within each neuron that allows retention of observations on short- and long-time scales.[38] To minimize the effect of vanishing gradients common in LSTMs, we included residual skip layers in each block.[39]

Since the functions are nonlinear, we include nonlinearity in the model using rectified-linear (ReLu) activation functions following each layer. Between the encoder and decoder, we have an embedding block. This block takes the output from the last timestep of the encoder, as a compact representation of the entire time series. This reduces the dimensionality in the sequential ordinal domain from n to 1.

In the embedding layer, we impose regularization to prevent overfitting and control the learning process. Our autoencoder is based on a β-variational autoencoder (VAE).[] VAEs control the latent distribution to be well behaved by encouraging it to match a statistical distribution, in this case, a Gaussian distribution $f(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$, where $f(x)$ is probability density function, σ is standard deviation, and μ is mean value. This is achieved by encoding an input as a distribution over the latent space using two equal-sized dense layers representing the mean and variance. From these layers, a random point is sampled from this latent distribution and passed to a single fully connected layer (n = 16) with L_1 activity regularization to reduce the output magnitude. To optimize VAEs, the mean squared reconstruction error is augmented with a KL divergence regularization term $\beta\mathrm{KL}[N(\mu_n,\theta_n),N(0,1)]$. The KL divergence measures how a probability distribution differs from a reference distribution. Thus, this regularization term encourages the latent space to be Gaussian. To regulate the importance of the reconstruction, L_1 and KL-divergence loss terms have linked hyperparameters λ and β, respectively. λ controls the penalty for large embedding activations, while the β term regulates the learning capacity with large values imposing entirely statistical learning (i.e., belonging to a Gaussian distribution). The β term also acts as a disentanglement factor separating spectral modes into a set of characteristic Gaussian distributions on each neuron. VAEs have received significant applications in materials microscopy.[41–46]

The model is optimized using adaptive momentum estimation (ADAM),[47,48] an improvement on stochastic gradient, to updating the weights from each minibatch based on the gradient to minimize the $\text{loss} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \widehat{Y}_1)^2 + \lambda\sum_i |a_i| + \beta\text{KL}[N(\mu_n, \theta_n), N(0,1)]$, where a_i is the activation of the embedding layer, $N(\mu_n, \theta_n)$ represents the distribution of n features, μ_n is the mean value of n features, θ_n is the standard deviation of n features, and $N(0, 1)$ represents the normal reference distribution.

We tested our model by training it on the same randomly generated toy spectral dataset for 6500 epochs (complete cycles through the entire dataset), with a batch size of 300, and a learning rate of 3×10^{-5}. We used regularization scheduling to control the linked λ and β hyperparameters during the training process. We primed the model by training with λ and β equal to 0 for 2000 epochs. The initial value of β and the schedule rate was determined such that the $\beta\text{KL}[N(\mu_n, \theta_n), N(0,1)]$ represents 60% of the loss at epoch 2000. Subsequently, we increase β by this value 5×10^{-3} and λ by $1 \times 10^{-5} \times 10^{\beta}$ every 100 epochs. We note that the model was relatively insensitive to reasonable hyperparameters.

We validated the model using the spectra generated from the painting of Nala and preprocessed as previously described. Since the embedding size of 16 is overparameterized based on the ground truth of 3 colors, we visualized the embedding maps through the training process. We selected the β value where only 3 channels reflected the original image, and the rest were Gaussian noise (Movie S2, Supporting Information). This process minimizes user bias as the number of active channels is learned through disentanglement and interactively explored following training rather than being set prior to training.

The results from embedding show all key characteristics of the painting (Figure 2g–i). This is akin to identifying the spectroscopic differences in hyperspectral materials spectroscopy. We extracted examples of the best, median, and worst mean squared reconstruction loss to validate the performance. The results (Figure 2j–l) show that the β-VAE can capture significantly more of the essential spectroscopic detail missed by dictionary learning. Even in the case of the worst fit, the neural network can capture the general spectroscopic shape despite being vertically shifted. This demonstrates a potential flaw of using the mean-squared error as a metric when the fine structure is more important than the absolute value. All told, the β-VAE reproduces the spectroscopic shape with significantly better fidelity than dictionary learning.

The models' performance can be best visualized by constructing a movie (Movie S3, Supporting Information) using the decoder as a generator. This movie is built by continuously linearly sampling across an embedding distribution and generating the spectroscopic response from the mean embedding of the 100-nearest neighbor pixels. The results clearly show the important spectroscopic phase shifts and are characteristic mixtures of the original functions.

To compare the β-VAE to dictionary learning, we constructed violin plots that show the order-dependent error distribution (Figure 2m). This plot shows that the error distributions for dictionary learning are uniformly larger and more dispersed than the β-VAE. More importantly, examples in the beginning and end of the generated spectra are Gaussian distributed but offset from the 0 value. This indicates that dictionary learning is unable to learn or lacks the capacity to discover a reasonable solution. We note that the better performance of the neural network is not merely the result of having more parameters such that it overfits the training data. We validate the model by generating spectra across the full RGB color gamut (Figure S4, Supporting Information). The predicted MSE reconstruction error for the β-VAE is only 0.104, much lower than achieved using dictionary learning (MSE = 0.186). These results demonstrate how in hyperspectral image analysis, nonlinear, ordinal, and statistically regularized machine-learning methods, in this case, LSTM β-VAE can achieve improved performance and generalizability with less human imposed biased than linear, non-ordinal machine-learning methods such as dictionary learning.

While these studies do not directly inform our understanding of electromechanical switching of ferroelectrics developing and testing a toy model with a known ground truth and similar characteristics to the scientific objective can elucidate advantages, disadvantages, and limitations of machine-learning methods. Generating an exemplar toy dataset improves confidence in the application and interpretability of machine-learning models applied to real experimental data.

We applied these models to hyperspectral images of electromechanical switching. We chose a benchmark sample, and hyperspectral dataset originally synthesized and collected by Agar et al. in 2015,[49] and published using the open-source Creative Commons Attribution 4.0 International license.[25] This dataset was collected from a 400 nm-thick $PbZr_{0.2}Ti_{0.8}O_3$/30 nm $SrRuO_3$/$GdScO_3$ (110) film grown using pulsed-laser deposition. These and similar heterostructures have been extensively studied, and have been found to have a mixed $c/a/c/a$ and a_1/a_2 domain structure that emerges due to a strain-induced-spinodal instability.[50] This sample has interesting regions of strong vertical and lateral PFM response, exhibits both classical and unique two-step, three-state switching mechanisms,[50] has collective switching behavior,[3] and has domain structure geometries that alters switching mechanism[51] and electromechanical energy conversion by forming charged domain walls.[49] PFM-based BE polarization spectroscopy (BEPS; Figures S1 and S2, Supporting Information) was acquired to measure the nanoscale switching mechanisms. BEPS measures the voltage-dependent piezoresponse at a band of frequencies near the cantilever resonance. Fitting the cantilever resonance to a simple-harmonic oscillator model allows the extraction of the amplitude (A), phase (φ), resonance frequency (ω), and quality factor (Q) of the cantilever resonance which are qualitative measures of the piezoresponse, polarization direction, elastic modulus, and energy absorption. This dataset was obtained by scanning a 2 μm region in a 60×60 grid. Measurements were conducted through two piezoelectric switching cycles, each with 96 voltage steps. All subsequent analysis was conducted on the second switching cycle in the off-field state. Further details are provided in the original manuscript where this dataset was first published.[49] The dataset's interesting characteristics, quality, and availability have made it a benchmark BEPS dataset studied in three peer-reviewed articles on machine learning in electromechanical switching of ferroelectrics.[34,49,51]

We start by exploring the efficacy of dictionary learning in capturing the important physics of electromechanical switching in ferroelectrics. This approach is derived from a method claiming to achieve "better", "faster", and "less-biased" machine learning of electromechanical switching in ferroelectric thin-films[34] with a few noted exceptions. The piezoelectric hysteresis loops were computed before training using $P = A\cos(\phi + \theta)$ where θ is the optimum rotation angle. The piezoelectric hysteresis and resonance response loops were normalized using a standard scalar $z = (x - u)/s$. So-called dimensional stacking was used to concatenate the piezoresponse and resonance data into a single 192-length vector. Notably, our preprocessing steps exclude any anomaly suppression and mean subtraction used in the referenced work[34] as this imposes too much user bias for automated use and has the potential to obscure essential physics. Following fitting, using the best hyperparameter of $\lambda = 1$ and $n = 5$ were selected and are consistent with prior studies.[34]

We computed the atoms for each spectrum and used their values to reconstruct the real-space images (**Figure 4**a–c). For brevity, we show only 3 out of 5 images (additional images are provided Figure S5, Supporting Information). We can see the key features of the domain structure are identified. The first coefficient map (Figure 4a) identifies a region of $c/a/c/a$ domains. The contrast is most pronounced along the right-hand boundary that corresponds to the peak in the topography (Figure S6, Supporting Information). The regions subtracted from the c domains appear in the second coefficient map (Figure 4b). These two regions has been attributed to changes in switching mechanisms associated with charge domain wall formation,[51] that manifests as an increase in the electromechanical contact resonance. The third map (Figure 4c), highlights the response in the a domains with a preference toward the left-hand boundary that corresponds to the peak in the topography. This boundary is associated with asymmetric ferroelastic switching based on geometrically necessary charged domain wall formation. These results can be viewed as a movie where we linearly sample across the latent space (Movie S4, Supporting Information). Overall, this shows is the dictionary learning model can disentangle the primary differences in the domain structure and switching mechanisms.

Interpreting how the learned embeddings relate to the switching mechanisms and physics requires the interpretation of the spectral reconstructions. This can be calculated by taking the dot product of the learned dictionary and atoms. While the model's overall performance resulted in a MSE of (0.36), it is important to validate the models' accuracy in reproducing the individual spectra. We show examples of the best, median, and worst reconstruction in the dataset based on the dictionary learning model (Figure 4d–j). For comparative purposes, we include results on the same data point generated using the β-VAE discussed later. Overall, the results indicate that the dictionary learning model cannot capture the essential spectroscopic details. Even for the best example (Figure 4d,h), the reconstruction cannot capture the depth of the softening that occurs during ferroelectric switching. On the example with median error (Figure 4e,i), we observe that the generated piezoelectric hysteresis loop has a lower-than-expected piezoresponse. Even more concerning, the generated resonance response has almost zero correspondence to the actual spectra. Instead, the spectra look like a classical resonance behavior during ferroelectric switching. Unsurprisingly, the worst reconstruction (Figure 4f,j) has nearly no correlation with the actual ferroelectric hysteresis loop or resonance behavior. Comparatively, the β-VAE reconstructions are markedly better in all cases. The poor quality of the dictionary learning reconstructions makes them unsuitable for physics-based interpretations.

We conducted a similar analysis using the β-VAE as previously described. The raw data was preprocessed in the same way as for dictionary learning. Instead of appending the piezoresponse and resonance spectra, these were joined as two independent dimensions. The model was trained for 20 000 epochs using ADAM optimizer with a learning rate of (3×10^{-5}). Regularization scheduling was used to increase the magnitude of β and λ by $n \times 0.0025$ and $1 \times 10^{-5} \times 10^{n\beta}$ every 1000 epochs, respectively. The β scheduling rate was determined such that it contributed to 60% of the loss when first changed from 0.

Following training, we extracted the activations from the low dimensional embedding layer and reconstructed the values as images for each epoch. By watching this movie (Movie S5, Supporting Information), we can observe how disentangles features. We used this to select a model where the important features are sufficiently disentangled. For consistency with the dictionary learning, we selected the model at 13 977 epochs where the embedding condensed to five channels, with the rest being Gaussian noise (Figure S7, Supporting Information). For brevity, we show only 3 out of 5 learned embeddings (Figure 4k–m) additional maps (Figure S7, Supporting Information). The first embedding (Figure 4k) identifies regions with a $c/a/c/a$–a_1/a_2 domain structure. The second embedding (Figure 4l) identifies areas within the a_1/a_2 band, with a preference toward the peak c/a boundary similar to what was identified with dictionary learning (Figure 4I). Finally, the third embedding (Figure 4m) highlights the a_1/a_2 band near the valley c/a boundary. This region has been previously shown to undergo a symmetric three-state, two-step, ferroelastic switching process.[51]

To evaluate the model's performance, we once again plot example reconstructions with the lowest, median, and highest mean-squared reconstruction error for the VAE (Figure 4n–s). The results show that the best and median reconstructions (Figure 4n,o,q,r) capture all the key spectroscopic details of the piezoresponse and resonance spectra. All the essential details of the spectroscopic curvature related to the switching mechanism can be captured with the VAE. The fit results are uniformly better than the dictionary learning model on the same spectra. Finally, unsurprisingly, the generated spectra with the highest error (Figure 4p,s) do not correspond well to the raw data; this is likely just an outlier data point of minimal significance. This demonstrates that our model is well regularized and not merely overfitting the data.

To further compare the performance of the dictionary learning and VAE model, we computed the mean-squared reconstruction error for each spectrum and plotted it as an image of the domain structure. Both plots are visualized on an identical scale from 0 to 0.5. Starting with the dictionary learning model (Figure 4t), we observe a large MSE highly correlated with the domain structure. This means that the model

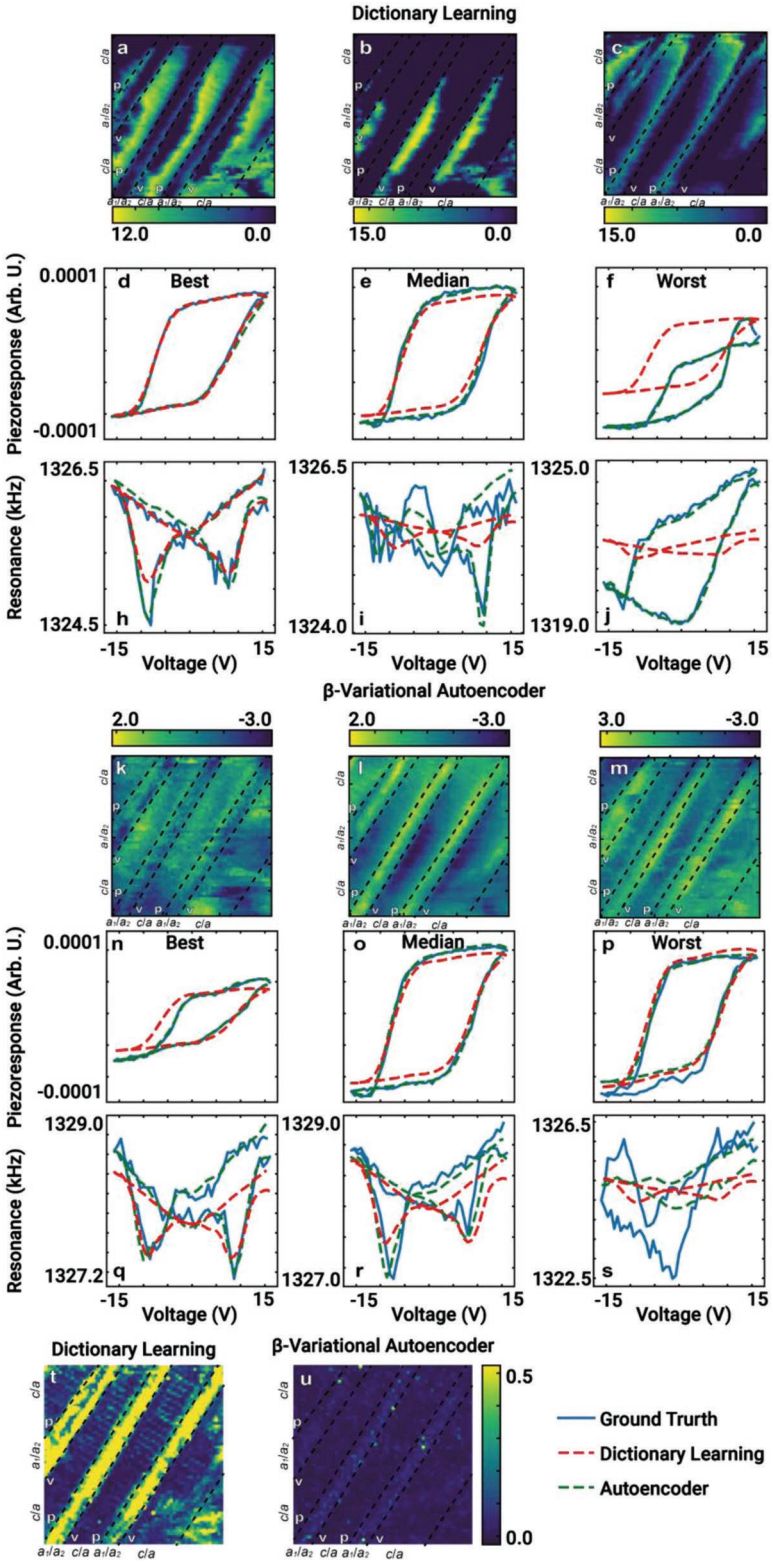

Figure 4. a–c) 3 out of 5 learned embeddings (atoms) obtained from dictionary learning model trained on dimensionality stacked piezoresponse and resonance hysteresis loops. a–c,h–j) Dictionary learning best (a,h), median (b,i), and worst (c,j), reconstruction errors of the piezoresponse and resonance response, respectively. k–m) 3 out of 5 learned embeddings obtained from β-VAE model trained on piezoresponse and resonance hysteresis loops. n–s) β-VAE best (n,q), median (o,r), and worst (p,s), reconstruction error of the piezoresponse and resonance response, respectively. t,u) Per pixel mean squared reconstruction error of the combined piezoresponse and resonance response for dictionary learning and β-VAE. The errors are plotted on the same color scale. All maps are 2 µm.

significantly underfits the data and lacks the capacity to describe the physics. Comparatively, the VAE has a uniformly lower MSE with only minimal correlation to the domain structure (Figure 4u), indicating the model has the capacity to learn the data. This suggests that the VAE can learn a latent manifold of the same dimensionality reduction as dictionary learning with a much higher capacity to encode the underlying physics. This improved performance is mainly due to its ability to learn nonlinear functions and consider the ordinal nature of data. To further analyze the differences in the models' learning potential, we constructed violin plots of the mean reconstruction error at each spectral index (Figure S8, Supporting Information). Visualizing the reconstruction error in this way reveals that the dictionary learning model, compared to the VAE model, has a much larger error distribution, is less Gaussian, has more outliers, and is systematically shifted from the zero-axis indicating that it is underfitting and thus cannot accurately model the observed responses.

Having proven that both models can, at first pass, provide statistical descriptors of the observed responses, we sought to explore how the choice of the model influences physics interpretations. Starting with the dictionary learning model, generation of the spectra as we increase the value of the embedding (atom) identified primarily within the $c/a/c/a$ domains reveals increasing square hysteresis loops and evidence of increased resonant frequency from the peak-to-valley boundary (**Figure 5**a). The corollary to this response closer to the valley boundary shows a continuation of the hardening process (Figure 5b). This observation is not attributed to charge injection as it is highly correlated to the domain structure, and is along the fast-scan direction. The third embedding extracted (Figure 5c), highlights the response in the a domains with a preference toward the left-hand boundary corresponding to the topography peak. The extracted responses look like a classical piezoelectric hysteresis loop with reduced (enhanced) piezoresponse at the a domains (c domain) regions. We note that the generated spectrum shows no evidence of intermediate concavities and hardening processes associated with three-step, two-state ferroelastic switching mechanisms that have been observed using manual and more robust machine-learning analysis approaches. Further analysis of the learned features and spectral representations is provided (Figure S5, Supporting Information).

We conducted a similar analysis of the VAE embedding and learned spectroscopic behavior (Movie S6, Supporting Information). Starting with the response in the a_1/a_2 band, with a preference toward the peak c/a boundary (Figure 5d), we observe three characteristic trends. The piezoresponse hysteresis loop goes from large and square to reduced and slightly tilted. Furthermore, the suppressed piezoresponse loops (yellow) show intermediate concavities and resonance hardening when switching under positive and negative bias, associated with the three-step, two-state ferroelastic switching mechanism.[51] Finally, we observe resonance softening as we move toward the peak boundary. The second discussed embedding (Figure 5e), highlights the a_1/a_2 band near the valley c/a boundary. In the dark regions, the generated spectra reveal classical c-like hysteresis loops and resonance behavior; however, in the light regions (yellow) toward the valley a_1/a_2 band, we see suppressed piezoresponse hysteresis loops, intermediate concavities in the piezoresponse, and hardening on positive bias switching. This behavior was experimentally attributed to and confirmed with phase-field modeling to be related to asymmetric charge domain wall formation during switching due to the domain geometry.[51] The dictionary learning model obscured this important physical mechanism. The final embedding discussed (Figure 5f) shows the response in the a domains and along the c/a-a_1/a_2 boundary. As the embedding moves from dark to light, we observe slightly suppressed piezoresponse and increased tilt of the piezoelectric hysteresis loops. The resonance softening becomes less pronounced. We attribute these spectroscopic changes to the increased ferroelastic switching. We note that the suppression of the piezoresponse is less pronounced than in the a_1/a_2 band because of the large volume fraction of c-like character. We provide further discussion of the learned embedding (Figure S10, Supporting Information).

All told, unsurprisingly, similar to the toy dataset, the dictionary learning model can identify the key characteristic regions associated with the domain structure, and the switching mechanism but is unable to represent the spectroscopic details correctly. This is confirmed both by the low quality of the spectral reconstruction and the significant deviation of the learned representation from the known physics. Thus, interpretation of the learned dictionaries and reconstructions is highly likely to result in misleading conclusions. Conversely, the VAE has a much lower reconstruction error and thus is more adept at reconstructing and identifying important switching mechanisms. This is not to say that the VAE has no failure modes, just fewer. Any conclusion aided by machine-learning methods must be rigorously validated mathematically, empirically, and theoretically.

The significant improvement in the aptitude of the VAE model is related to the model's consideration of the structure and information content in the data. Unlike dictionary learning, the VAE includes nonlinearity and can consider the ordinal nature of the data. The VAE includes significant structural improvements on our prior work, a deep recurrent neural network.[51] The key advantages are that the VAE uses regularization scheduling to, without bias, disentangle features. Furthermore, the imposition of the KL-divergency constraint regularizes the latent space to be defined by a Gaussian distribution making the latent space more interpolatable and interpretable. While training dictionary learning models is generally faster this is somewhat of a moot point since it is nonperformant. We highlight the key differences of dictionary learning and VAEs in analysis of electromechanical switching (**Table 1**), for further details (see Supporting Information).

3. Conclusion

We demonstrate using toy and experimental benchmark datasets some of the challenges in applying machine-learning models to ferroelectric switching and scientific data broadly. We show that while nearly any machine-learning model can be used to categorize responses, this can be purely statistical without any correlation to physics. Specifically, we develop a VAE that supports data fusion of resonance and piezoresponse data, which we train using regularization scheduling to disentangle spectroscopic features into a nearly Gaussian latent space

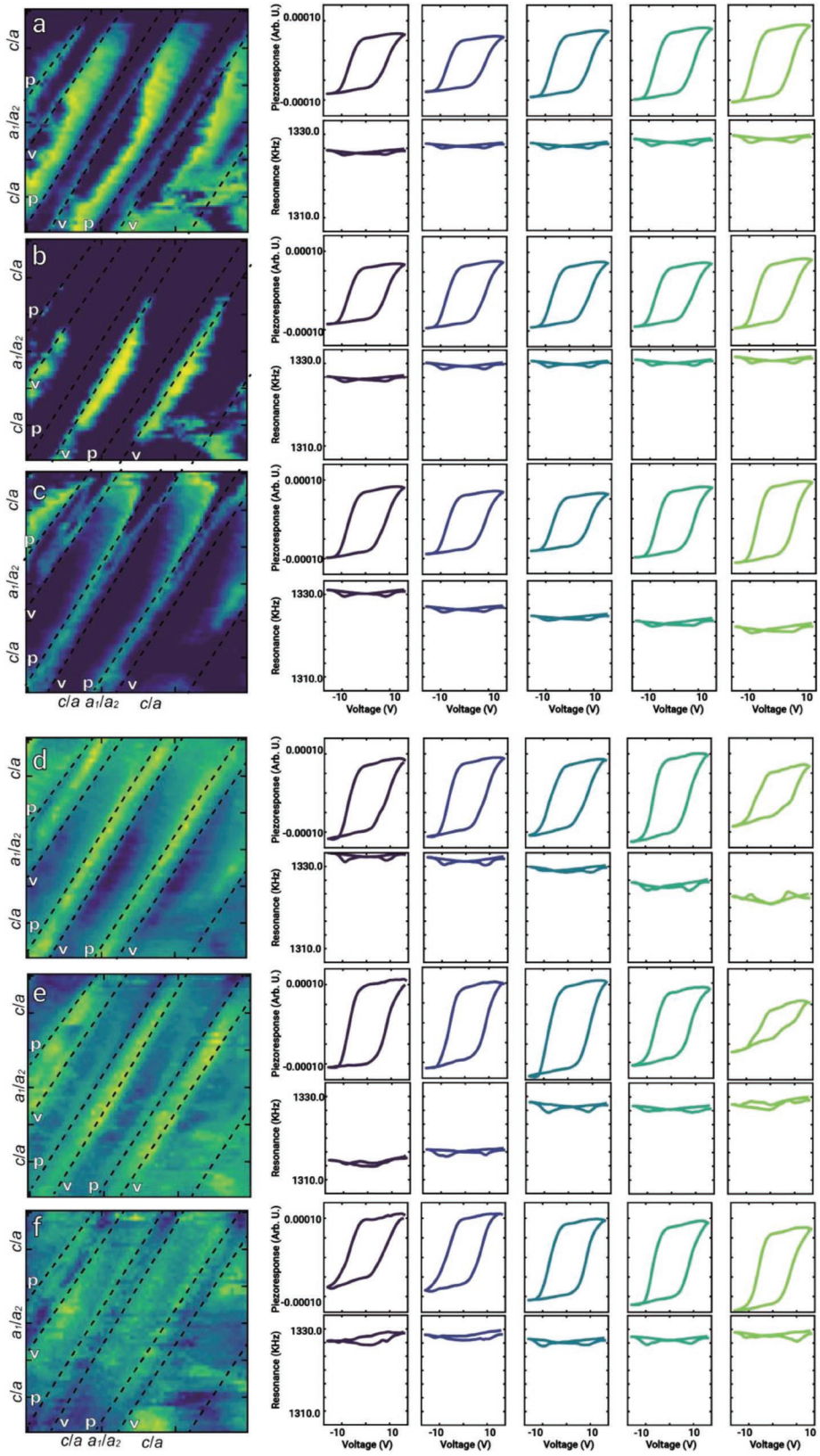

Figure 5. Extracted embeddings and generated spectra obtained from a–c) Dictionary learning model, and d–f) β-VAE. The colors on the maps on the left reflect the colors on the spectra on the right. Generated spectra are created from the mean of the 100 nearest neighbors when linearly sampling across the embedding space. All maps are 2 μm.

Table 1. Comparison of dictionary learning and VAE in the analysis of electromechanical switching.

	Dictionary learning	VAE
Reconstruction quality	Poor	Good
Ordinal	No	Yes
Nonlinear	No	Yes
Training time	Minutes	Hours
Inference time	Minutes	Minutes
Influence of hyperparameters	Sensitive	Less sensitive

automatically. We compare this model to a recently lauded linear, nonordinal dictionary learning model and show that the VAE can capture the essential spectroscopic features important for unraveling classic ferroelectric and three-state, two-step ferroelectric switching mechanisms observed in $PbZr_{0.2}Ti_{0.8}O_3$ with hierarchical $c/a–a_1/a_2$ domain structures. The dictionary learning model cannot. Ultimately, this shows when analyzing nonlinear hyperspectral data, it is essential to have models capable of learning nonlinear ordinal trends. Finally, this work highlights the importance and complexities of rigorous model validation to avoid deriving misleading physical conclusions based on the interpretation of machine-learned results.

4. Experimental Section

Data Acquisition: All samples, and data were obtained from prior studies published by Agar and co-workers.[49–51] The data used for this manuscript was available open source associated with these publications. Details regarding the experimental methods should refer to the original works.

Machine-Learning Methods: All computational analysis was completed in Python using open-source packages. Throughout analysis randomly selected random seeds were used to ensure reproducibility. Dictionary learning methods were derived from the open source packages and code associated with ref. [34]. All analysis codes were provided as a Jupyter Notebook. All code blocks derived from code originating from ref. [34] was identified by the open source license which it was derived. Design and development of the VAE model was conducted using TensorFlow. Utility functions to simplify deployment and implementation were provided in the M3-Learning group research package DeepMatter.[46] This package is released open source and is pip installable.

All deep-learning models were trained on a Lambda Labs deep-learning workstation with two TitanRTX GPUs. All codes, analysis, and data visualization was provided as an reproducible and interactive Jupyter Notebook capable of running using free remote services on Google Collab. A video demonstrating the use of this notebook is provided (Movie S7, Supporting Information).

Supporting Information

Supporting Information is available from the Wiley Online Library or from the author.

Acknowledgements

The original data was collected at UC-Berkeley and the Center for Nanophase Materials Sciences at Oak Ridge National Laboratory. Oak Ridge National Laboratory is a US DOE Office of Science User Facility. S.Q. acknowledges primary support from the National Science Foundation under grant TRIPODS+X:RES-1839234 and DOE Data Reduction for Science award Real-Time Data Reduction Codesign at the Extreme Edge for Science. Y.G. and J.C.A. acknowledge primary support from Army Research Laboratory Collaborative for Hierarchical Agile and Responsive Materials. A.T.K. acknowledges support from ORAU University Partnerships and the Lehigh Presidential Initiative on Nanohuman Interfaces. All data and analysis codes are made openly available under the BSD 3-Clause License. Data is provided on Zenodo,[52] which is a complied version derived from the original release of this data.[25,33] The analysis codes are provide within the M3-Learning GitHub package DeepMatter. The specific analysis conducted in this work is available as an excitable Jupyter Notebook, which can be run on Google Collab or a local Python instance. The painting of the dog in the table of contents image is original art by Irene Dogmatic and is reproduced in the table of contents entry with permission of the correspondin author.

Conflict of Interest

The authors declare no conflict of interest.

Data Availability Statement

The data that support the findings of this study are openly available in Zenodo at https://doi.org/10.5281/ZENODO.6321172, reference number 52.

Keywords

deep learning, dimensionality reduction, ferroelectric switching, machine learning, multimodal hyperspectral imaging, unsupervised learning

Received: March 27, 2022
Revised: July 7, 2022
Published online: October 17, 2022

[1] R. J. Zeches, M. D. Rossell, J. X. Zhang, A. J. Hatt, Q. He, C.-H. Yang, A. Kumar, C. H. Wang, A. Melville, C. Adamo, G. Sheng, Y.-H. Chu, J. F. Ihlefeld, R. Erni, C. Ederer, V. Gopalan, L. Q. Chen, D. G. Schlom, N. A. Spaldin, L. W. Martin, R. Ramesh, *Science* **2009**, *326*, 977.
[2] A. R. Damodaran, C.-W. Liang, Q. He, C.-Y. Peng, L. Chang, Y.-H. Chu, L. W. Martin, *Adv. Mater.* **2011**, *23*, 3170.
[3] X. Lu, Z. Chen, Y. Cao, Y. Tang, R. Xu, S. Saremi, Z. Zhang, L. You, Y. Dong, S. Das, H. Zhang, L. Zheng, H. Wu, W. Lv, G. Xie, X. Liu, J. Li, L. Chen, L.-Q. Chen, W. Cao, L. W. Martin, *Nat. Commun.* **2019**, *10*, 3951.
[4] R. Xu, S. Liu, S. Saremi, R. Gao, J. J. Wang, Z. Hong, H. Lu, A. Ghosh, S. Pandya, E. Bonturim, Z. H. Chen, L. Q. Chen, A. M. Rappe, L. W. Martin, *Nat. Commun.* **2019**, *10*, 1282.
[5] F. Xu, S. Trolier-McKinstry, W. Ren, B. Xu, Z.-L. Xie, K. J. Hemker, *J. Appl. Phys.* **2001**, *89*, 1336.
[6] F. Li, S. Zhang, T. Yang, Z. Xu, N. Zhang, G. Liu, J. Wang, J. Wang, Z. Cheng, Z.-G. Ye, J. Luo, T. R. Shrout, L.-Q. Chen, *Nat. Commun.* **2016**, *7*, 13807.
[7] Z. Gu, S. Pandya, A. Samanta, S. Liu, G. Xiao, C. J. G. Meyers, A. R. Damodaran, H. Barak, A. Dasgupta, S. Saremi, A. Polemi, L. Wu, A. A. Podpirka, A. Will-Cole, C. J. Hawley, P. K. Davies,

R. A. York, I. Grinberg, L. W. Martin, J. E. Spanier, *Nature* **2018**, *560*, 622.
[8] R. Ramesh, *Microsc. Microanal.* **2016**, *22*, 1246.
[9] P. Shafer, P. García-Fernández, P. Aguado-Puente, A. R. Damodaran, A. K. Yadav, C. T. Nelson, S.-L. Hsu, J. C. Wojdeł, J. Íñiguez, L. W. Martin, E. Arenholz, J. Junquera, R. Ramesh, *Proc. Natl. Acad. Sci. USA* **2018**, *115*, 915.
[10] S. M. Neumayer, L. Collins, R. Vasudevan, C. Smith, S. Somnath, V. Y. Shur, S. Jesse, A. L. Kholkin, S. V. Kalinin, B. J. Rodriguez, *ACS Appl. Mater. Interfaces* **2018**, *10*, 42674.
[11] P. Tückmantel, I. Gaponenko, N. Caballero, J. C. Agar, L. W. Martin, T. Giamarchi, P. Paruch, *Phys. Rev. Lett.* **2021**, *126*, 117601.
[12] A. Gruverman, M. Alexe, D. Meier, *Nat. Commun.* **2019**, *10*, 1661.
[13] A. Belianinov, A. V. Ievlev, M. Lorenz, N. Borodinov, B. Doughty, S. V. Kalinin, F. M. Fernández, O. S. Ovchinnikova, *ACS Nano* **2018**, *12*, 11798.
[14] S. Jesse, S. V. Kalinin, *J. Phys. D: Appl. Phys.* **2011**, *44*, 464006.
[15] K. P. Kelley, S. V. Kalinin, M. Ziatdinov, O. Paull, D. Sando, V. Nagarajan, R. K. Vasudevan, S. Jesse, *Appl. Phys. Lett.* **2021**, *119*, 132902.
[16] E. Strelcov, Y. Kim, S. Jesse, Y. Cao, I. N. Ivanov, I. I. Kravchenko, C.-H. Wang, Y.-C. Teng, L.-Q. Chen, Y. H. Chu, S. V. Kalinin, *Nano Lett.* **2013**, *13*, 3455.
[17] D. Kim, R. K. Vasudevan, K. Higgins, A. Morozovska, E. A. Eliseev, M. Ziatdinov, S. V. Kalinin, M. Ahmadi, *J. Phys. Chem. C: Nanomater. Interfaces* **2021**, *125*, 12355.
[18] C. Sohn, X. Gao, R. K. Vasudevan, S. M. Neumayer, N. Balke, J. M. Ok, D. Lee, E. Skoropata, H. Y. Jeong, Y.-M. Kim, H. N. Lee, *Sci. Adv.* **2021**, *7*, eabd7394.
[19] M. Ziatdinov, D. Kim, S. Neumayer, L. Collins, M. Ahmadi, R. K. Vasudevan, S. Jesse, M. H. Ann, J. H. Kim, S. V. Kalinin, *J. Appl. Phys.* **2020**, *128*, 055101.
[20] S. V. Kalinin, E. Strelcov, A. Belianinov, S. Somnath, R. K. Vasudevan, E. J. Lingerfelt, R. K. Archibald, C. Chen, R. Proksch, N. Laanait, S. Jesse, *ACS Nano* **2016**, *10*, 9068.
[21] M. Rashidi, R. A. Wolkow, *ACS Nano* **2018**, *12*, 5185.
[22] E. Kazuma, Y. Kim, *Phys. Chem. Chem. Phys.* **2019**, *21*, 19720.
[23] S. Jesse, S. V. Kalinin, *Nanotechnology* **2009**, *20*, 085714.
[24] R. Kannan, A. V. Ievlev, N. Laanait, M. A. Ziatdinov, R. K. Vasudevan, S. Jesse, S. V. Kalinin, *Adv. Struct. Chem. Imaging* **2018**, *4*, 6.
[25] J. C. Agar, B. Naul, S. Pandya, S. van der Walt, R. Yao, J. Maher, J. Neaton, S. Kalinin, R. Vasudevan, Y. Cao, J. Bloom, L. Martin, **2018**, https://doi.org/10.5281/ZENODO.1482091 (accessed: July 2022).
[26] S. Farley, J. E. A. Hodgkinson, O. M. Gordon, J. Turner, A. Soltoggio, P. J. Moriarty, E. Hunsicker, *Mach. Learn.: Sci. Technol.* **2020**, *2*, 015015.
[27] K. Kaheman, J. N. Kutz, S. L. Brunton, *Proc. Math. Phys. Eng. Sci.* **2020**, *476*, 20200279.
[28] B. M. de Silva, K. Champion, M. Quade, *arXiv:2004.08424* **2020**.
[29] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov, *J. Mach. Learn. Res.* **2014**, *15*, 1929.
[30] G. Klambauer, T. Unterthiner, A. Mayr, S. Hochreiter, *Adv. Neur. Inf. Process. Syst.* **2017**, *30*, 972.
[31] Geoffrey Hinton talk 'What is wrong with convolutional neural nets? **2017**, https://www.youtube.com/watch?v=rTawFwUvnLE.
[32] ImageNet Classification Leaderboard, https://kobiso.github.io/Computer-Vision-Leaderboard/imagenet.html
[33] J. C. Agar, Zenodo, https://doi.org/10.5281/zenodo.997588 **2017**.
[34] L. A. Griffin, I. Gaponenko, N. Bassiri-Gharb, *Adv. Mater.* **2020**, *32*, 2002425.
[35] MNIST handwritten digit database, Yann LeCun, Corinna Cortes and Chris Burges, https://yann.lecun.com/exdb/mnist/
[36] COCO—common objects in context, can be found under https://cocodataset.org/
[37] Z. C. Lipton, J. Berkowitz, C. Elkan, *arXiv:1506.00019* **2015**.
[38] K. S. Tai, R. Socher, C. D. Manning, in *Proc. 53rd Annu. Meeting of the Association for Computational Linguistics and the 7th International Joint Conf. on Natural Language Processing (Volume 1: Long Papers)*, Association For Computational Linguistics, Stroudsburg, PA, USA **2015**.
[39] K. He, X. Zhang, S. Ren, J. Sun, in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, Piscataway, NJ, USA **2016**, pp. 770–778.
[40] C. P. Burgess, I. Higgins, A. Pal, L. Matthey, N. Watters, G. Desjardins, A. Lerchner, *arXiv:1804.03599* **2018**.
[41] Y. Liu, R. Proksch, C. Y. Wong, M. Ziatdinov, S. V. Kalinin, *Adv. Mater.* **2021**, *33*, 2103680.
[42] S. V. Kalinin, J. J. Steffes, Y. Liu, B. D. Huey, M. Ziatdinov, *Nanotechnology* **2021**, *33*, 055707.
[43] Y. Liu, B. D. Huey, M. A. Ziatdinov, S. V. Kalinin, *arXiv:2105.11475* **2022**.
[44] S. V. Kalinin, O. Dyck, S. Jesse, M. Ziatdinov, *Sci. Adv.* **2021**, *7*, eabd5084.
[45] M. Ziatdinov, A. Ghosh, S. Kalinin, *Machine Learning: Sci. Technol.* **2022**, *3*, 015003.
[46] R. Ignatans, M. Ziatdinov, R. Vasudevan, M. Valleti, V. Tileli, S. V. Kalinin, *Adv. Funct. Mater.* **2022**, *32*, 2100271.
[47] D. P. Kingma, J. Ba, *arXiv:1412.6980* **2014**.
[48] G. Goh, *Distill* **2017**, *2*, e6.
[49] J. C. Agar, Y. Cao, B. Naul, S. Pandya, S. van der Walt, A. I. Luo, J. T. Maher, N. Balke, S. Jesse, S. V. Kalinin, R. K. Vasudevan, L. W. Martin, *Adv. Mater.* **2018**, *30*, 1800701.
[50] A. R. Damodaran, S. Pandya, J. C. Agar, Y. Cao, R. K. Vasudevan, R. Xu, S. Saremi, Q. Li, J. Kim, M. R. McCarter, L. R. Dedon, T. Angsten, N. Balke, S. Jesse, M. Asta, S. V. Kalinin, L. W. Martin, *Adv. Mater.* **2017**, *29*, 1702069.
[51] J. C. Agar, B. Naul, S. Pandya, S. van der Walt, J. Maher, Y. Ren, L.-Q. Chen, S. V. Kalinin, R. K. Vasudevan, Y. Cao, J. S. Bloom, L. W. Martin, *Nat. Commun.* **2019**, *10*, 4809.
[52] J. Agar, S. Qin, Zenodo, https://doi.org/10.5281/ZENODO.6321172 **2022**.