cm CHEMISTRY OF MATERIALS

Connecting Vibrational Spectroscopy to Atomic Structure via Supervised Manifold Learning: Beyond Peak Analysis

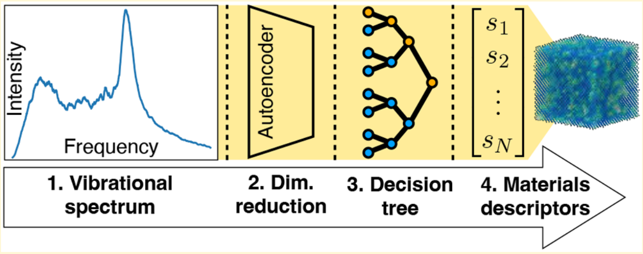
Daniel Vizoso, Ghatu Subhash, Krishna Rajan, and Rémi Dingreville*

Cite This: *Chem. Mater.* 2023, 35, 1186−1200

Read Online

ACCESS | Metrics & More | Article Recommendations | SI Supporting Information

ABSTRACT: Vibrational spectroscopy is a nondestructive technique commonly used in chemical and physical analyses to determine atomic structures and associated properties. However, the evaluation and interpretation of spectroscopic profiles based on human-identifiable peaks can be difficult and convoluted. To address this challenge, we present a reliable protocol based on supervised manifold learning techniques meant to connect vibrational spectra to a variety of complex and diverse atomic structure configurations. As an illustration, we examined a large database of virtual vibrational spectroscopy profiles generated from atomistic simulations for silicon structures subjected to different stress, amorphization, and disordering states. We evaluated representative features in those spectra via various linear and nonlinear dimensionality reduction techniques and used the reduced representation of those features with decision trees to correlate them with structural information unavailable through classical human-identifiable peak analysis. We show that our trained model accurately (over 97% accuracy) and robustly (insensitive to noise) disentangles the contribution from the different material states, hence demonstrating a comprehensive decoding of spectroscopic profiles beyond classical (human-identifiable) peak analysis.

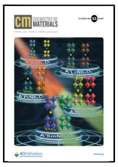


■ INTRODUCTION

The vibrational modes of atoms in a lattice dictate many transport properties including thermal expansion, heat and electric conductivities, and the Debye temperature. The vibrational density of states (VDoS) is a spectral property that is sensitive to variations in local atomic arrangements and the state of deformation of the material. For instance, instabilities in the VDoS can be associated with phase transitions as seen across many different materials.[1−6] Similarly, vibrational states can be correlated with elastic heterogeneity at the nanoscale in increasingly disordered crystals, but the VDoS becomes almost insensitive to disorder once an amorphization threshold has been reached.[7] Likewise, the vibrational modes of nanoparticles deviate from the corresponding bulk vibrational spectrum due to the increased contribution from undercoordinated surface atoms.[8−11]

Experimentally, the vibrational structure of a material can be probed using absorption-based (such as infrared (IR) spectroscopy) or scattering-based (such as inelastic neutron scattering (INS) or Raman spectroscopy) techniques which modify the vibrational state of a material through absorption or scattering, respectively. By quantifying changes in the measured spectra, such as peak shifts or peak broadening, these techniques have been used for instance to identify deformation states and phase transformations[12−15] or to measure pressure dependence in materials' properties.[16−18] However, the predictive power of interpreting profiles based on human-identifiable peaks suffers from inconsistencies depending on the width method being used.[19] As pointed out by Weidenthaler,[20] in many cases the evaluation and interpretation of the results from these 1D spectroscopic profiles can be difficult or erroneous, especially when the material being probed deviates from a pristine, defect-free state to being in a deformed and defected state. Even though spectroscopic measurements are often accompanied by density functional theory (DFT) or molecular dynamics (MD) simulations to facilitate their interpretation,[21−26] regressing changes in the vibrational properties to the state of the material remains a challenging problem, notably when multiple microstructural sources simultaneously affect changes in the vibrational spectrum. These challenges are associated with identifying key representative features in the spectral profile and relating their characteristics to the structural characterization. Currently, most spectroscopy practitioners focus on the positions and widths of human-identifiable peaks as representative features and compare the attributes of these peaks to known standard profiles in order to infer information

Received: October 21, 2022
Revised: January 11, 2023
Published: January 24, 2023



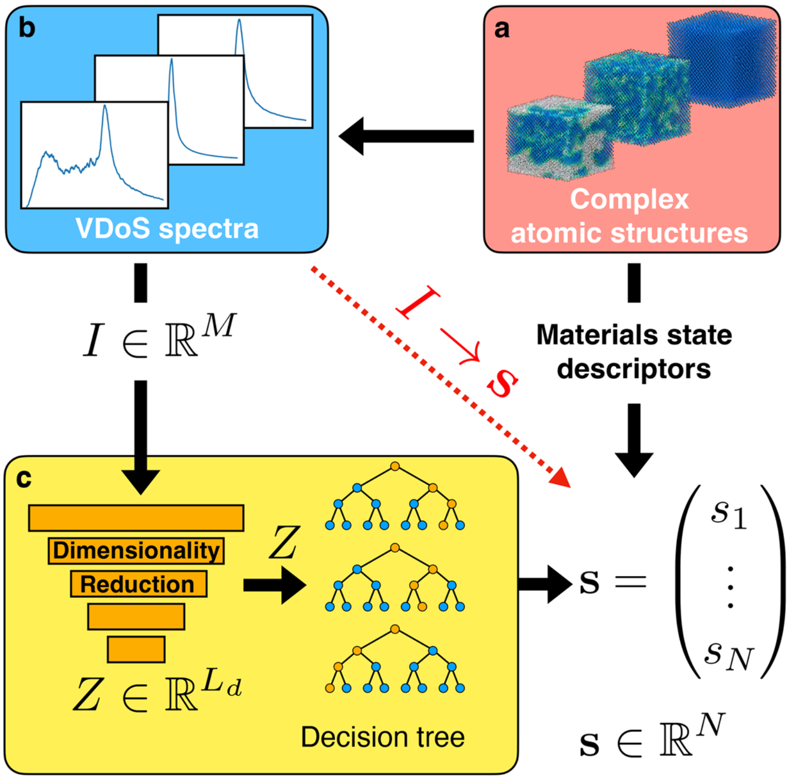

b
VDoS spectra
a
Complex atomic structures
$I \in \mathbb{R}^M$
$I \to \mathbf{s}$
Materials state descriptors
c
Dimensionality Reduction
Z
$Z \in \mathbb{R}^{L_d}$
Decision tree
$\mathbf{s} = \begin{pmatrix} s_1 \\ \vdots \\ s_N \end{pmatrix}$
$\mathbf{s} \in \mathbb{R}^N$

Figure 1. Schematic of the workflow utilized for this work. (a) Complex atomic structures are obtained by deforming a Si monocrystalline system under hydrostatic and uniaxial compression and by progressively introducing disorder. (b) VDoS database is computed from atomistic models. (c) Mapping $\mathcal{F}$ consists of reducing the dimensionality of the VDoS spectrum, I, into a latent variable, Z, of dimension L_d and using this latent representation to regress material state descriptors **s** via a gradient boosting decision tree model.

on the underlying atomic structure. Complications also arise due to experimental noise when doing these comparisons.

As an alternative, researchers have recently applied advances in machine learning techniques to successfully map atomic structures to spectral properties such as VDoS or diffractograms[27−37] and help with a systematic and unbiased interpretation of spectroscopic data. For instance, Lee et al.[31] trained a convolutional neural network (CNN) on a data set of synthetic X-ray diffraction (XRD) diffraction line profiles and used the trained model to predict the phase fractions of multiphase inorganic compound powders in experimental XRD data. Similarly, Kong et al.[35] employed a generative model for predicting *ab initio* phonon and electronic densities of states for thousands of pristine crystalline materials. Other research groups have used graph neural networks[36,37] to perform the same task. Most of these studies focus on perfect crystalline atomic structures and leverage crystallographic and lattice symmetries in the architecture of their algorithms to achieve good accuracy and prediction of the corresponding atomic structures. However, when these symmetries are degenerated or broken, as is the case for structures containing defects or for material systems undergoing a gradual phase transformation or being deformed for instance, the mapping from spectral properties to changes in the atomic structure remains difficult due to the output complexity and finite data volume.

In this work, we evaluate the connection between vibrational spectra and the state of the material when it deviates from a pristine, defect-free configuration by building a predictive model using machine learning models. We formulate this task as being equivalent to learning a mapping function $\mathcal{F}$: $I \in \mathbb{R}^M \to \mathbf{s} \in \mathbb{R}^N$, where I is a high-dimensional array of dimension $M \sim O(10000)$ describing the one-dimensional (1D) VDoS spectrum and the vector **s** describes the corresponding state of the atomic structure in terms of $N \sim O(10)$ variables representing the deformation and the defect states present in the material. To build this model, we represented the VDoS in a low-dimensional, latent space Z of dimension $L_d \sim O(10)$, $L_d \ll M$, using different dimensionality reduction techniques (Figure 1c). We then used this reduced representation of the VDoS to predict the state of the material from an observed VDoS using a decision tree regression model. The VDoS and corresponding material state databases were obtained from MD simulations we performed on a monocrystalline silicon (Si) system (Figure 1a). We selected Si as a model material system due to available information on spectral properties in the literature. We calculated the VDoS for this atomistic system undergoing hydrostatic and uniaxial compression and also when disorder was progressively introduced (Figure 1b) or both for compression and disorder combined. We selected these environments as they provide a large taxonomy of material states and associated changes in their spectral signatures. We performed several analyses to evaluate the performance of our model as a function of the dimensionality reduction and regression techniques used. Our supervised-learning protocol accurately (over 97% accuracy) and robustly (filtering out noise) predicts the state of the material from an observed VDoS and can deconvolve the contribution from the deformation state and that of the amorphization and disordering states, demonstrating that the vibrational spectra do contain the necessary information to represent the state of complex atomic structure configurations.

■ METHODS

The foundational methods supporting our predictive model rest on three elements: (i) atomistic simulations performed to generate the various deformed and defected atomic systems, (ii) calculations of the VDoS for these different atomic configurations, and (iii) dimensionality reduction techniques and regression models to regress material state descriptors representing the atomic system from the observed VDoS.

Atomistic Simulations. We generated a large database of monocrystalline Si atomic structures using the Large-scale Atomic/Molecular Massively Parallel Simulator (LAMMPS).[38] The interactions between pairs of Si atoms were described using a modified embedded atom method (MEAM) potential developed by Lenosky et al.[39] This interatomic potential was fit to force and energy data as well as the formation energies for various defects calculated by *ab initio* simulations for a wide range of Si polymorphs. This potential was found to accurately reproduce the elastic constants and phonon properties of various Si phases and polymorphs. The initial Si structure consisted of a cubic diamond structure with a lattice constant of 5.431 Å. The simulation cell for the compression simulations consisted of a 12 × 12 × 12 unit cell containing 13 824 atoms. The simulation cell for the disorder simulations consisted of a 19 × 19 × 19 unit cell containing 54 872 atoms. Prior to any compression or disorder insertion, we equilibrated the Si atomic structure at 300 K and zero pressure for 200 ps.

Compression simulations were performed under both uniaxial and hydrostatic loading conditions. Structures were oriented such that the ⟨100⟩ family of crystallographic directions was along the *x*, *y*, and *z* axes of the simulation box. All compression simulations were performed under a canonical (*NVT*) ensemble at a constant temperature of 300 K, with the compression of the cell being handled by the `fix deform` command in LAMMPS using atomic displacements. Both hydrostatic and uniaxial compression simulations were strain-controlled with constant engineering strain rates of 10^{-4} ps^{-1} (10^{8} s^{-1}). Uniaxial compression simulations were performed along the [001] direction while maintaining the dimensions of the simulation cell in the [100] and [010] directions. Hydrostatic compression simulations were performed by reducing the dimensions of the simulation cell along each of the ⟨100⟩ directions at strain rates computed with the following relation:

$$1 + \dot{\varepsilon}_{\text{tot}} = (1 + \dot{\varepsilon}_{xx})(1 + \dot{\varepsilon}_{yy})(1 + \dot{\varepsilon}_{zz}) \tag{1}$$

where $\dot{\varepsilon}_{\text{tot}}$ is the previously mentioned strain rate of 10^{-4} ps^{-1} and $\dot{\varepsilon}_{xx}$, $\dot{\varepsilon}_{yy}$, and $\dot{\varepsilon}_{zz}$ are the strain rates along each of the simulation cell axes. Simulations of hydrostatic and uniaxial compression were run to a final engineering strain, ε_{tot}, of 30%.

We introduced disorder in the Si crystalline structure by gradually inserting Frenkel pairs (i.e., vacancy−interstitial pairs) over time.[40−42] In these simulations, each disorder insertion step consisted of 50 randomly selected atoms being displaced from their original positions by a randomly sampled distance between 20 and 60 Å, creating 50 Frenkel pairs. These displacements were adjusted so that none of the displaced atoms would be within 2 Å of any other atom. After 50 atoms (<0.1% of the atoms within the structure) had been selected and displaced, the system was evolved under an isothermal−isobaric (*NPT*) ensemble held at 300 K and zero pressure for 0.5 ps with a time step of 0.1 fs and then for 2.5 ps with a time step of 0.5 fs. This process was repeated until the number of displaced atoms was equal to the number of atoms in the structure (i.e., on average every atom was displaced once).

Finally, we performed simulations that combined both disorder and compression. These simulations consisted of initially disordered atomic structures that were subsequently deformed under uniaxial or hydrostatic compression following the same procedure described above.

Material State Descriptors. For all the simulations performed, we extracted several material state descriptors s_i, namely, the average stress over the entire structure (σ_{tot}), the applied strain (ε_{tot}), and the phase fraction of disordered atoms (ϕ_{tot}). The average stress was obtained using the standard approach in LAMMPS for the computation of stresses, including both the kinetic energy contribution and the virial contribution.[43] The applied strain was known for all compressed states, with the strain for uniaxial compression corresponding to the applied strain along the *z*-direction and the strain for hydrostatic compression being computed using eq 1. The stress and strain values were recorded as system averages as well as broken into their respective *x*, *y*, and *z* components. The phase fraction of the disordered Si atoms was identified as atoms not in the perfect cubic diamond structure as defined by the "identify diamond-structure modifier"[44] in the OVITO software package.[45]

Additionally, when disorder is present, as in the case of the uniaxial compression or disorder simulations, we also calculated the characteristic length scale, l_ϕ, associated with the disorder parameter ϕ_{tot}. This length scale characterizes the morphology of the disordering within the atomic structure. We computed l_ϕ using a modified version of the FoamExplorer program.[46] The input to this calculation was the Si atomic structure with all atoms that were not in the perfect cubic diamond structure removed. FoamExplorer was then used to measure the size distributions of the voids created by removing these atoms. The characteristic length scale, l_ϕ, was extracted by averaging the entire set of length measurements produced by FoamExplorer. Microstructures that did not have any disorder present (such as the pristine microstructures under low strains or the entirety of the pristine hydrostatically compressed simulation) had their disorder length scale set to be the first nearest neighbor distance. For microstructures where no cubic diamond phase was identified, the characteristic length scale was set to the length scale of the simulation box. Note that additional descriptors could have been considered for a more comprehensive description of the state of materials. However, these would have to be materials specific and a decision would have to be made prior to the supervised-learning analysis for label definition. For instance, there is the possibility of an intermediate-range ordering in amorphous Si[47−49] that could potentially exist in our data.

For the single-output regression model, the material descriptor vector is a scalar *s* consisting of either $s = \sigma_{\text{tot}}$, $s = \varepsilon_{\text{tot}}$, or $s = \phi_{\text{tot}}$. For the multioutput regression model, we extended the prediction such that the material descriptor vector $\mathbf{s} = [s_i]$, $i = 6$, consists of the three components of the stress tensor: σ_{xx}, σ_{yy}, σ_{zz}; the total strain ε_{tot}; the disorder parameter ϕ_{tot}; and the characteristic length scale l_ϕ.

Vibrational Density of States. We measured the velocity autocorrelation function using the `compute vacf` command in LAMMPS, which performs the following operation:

$$\gamma(t) = \frac{1}{N_{\text{at}}} \sum_{i=1}^{N_{\text{at}}} v_i(t)\cdot v_i(0) \tag{2}$$

where N_{at} is the total number of atoms in the simulation cell, $v_i(t)$ is the velocity vector of atom *i* at time *t*, and $v_i(0)$ is the velocity vector of atom *i* at the start of the measurement period. The VDoS is then computed by taking the Fourier transform of $\gamma(t)$ to convert the data to the frequency domain

$$f(\omega) = \int_{-\infty}^{\infty} \gamma(t)\mathrm{e}^{-\mathrm{i}2\pi\omega t}\,\mathrm{d}t \tag{3}$$

The result of the Fourier transform contained both real and imaginary components; for the purpose of this work, only the real portion of the spectrum was used. We truncated the VDoS at a frequency of 900.09 cm^{-1}, resulting in a 1D VDoS vector with M = 10 794 values. The output of the Fourier transform contains a significant amount of noise, so we smoothed the VDoS using a Savitsky−Golay filter[50] prior to visualization or use in model training.

The form of the simulated VDoS profile is influenced by the simulation conditions, namely, the interatomic potential used, the size of the simulation volume, the time interval chosen for writing the output of the velocity autocorrelation function, and the number of measurements taken for the velocity autocorrelation function. We performed test simulations using several interatomic potentials of various functional styles.[39,51−55] The MEAM potential by Lenosky et al.[39] was selected due to its acceptable computational cost and the resemblance of the predicted VDoS to that predicted by density functional theory (DFT) (see the Supporting Information for the comparison between DFT and MD). We also performed sensitivity studies for all of our simulation conditions, and we chose values that balanced the computational cost of the simulations with consistency in the measurements of the VDoS. As illustrated in the Supporting Information, we verified and validated our VDoS implementation by comparing the VDoS for a pristine Si atomic structure computed by DFT and by MD. We note slight differences in peak shapes and positions; however, there is general agreement between the two techniques. Some of these differences, particularly differences in peak widths,[56] can be attributed to temperature effects due to the dynamic nature of the measurement of the VDoS via MD. In both simulated spectra, the peaks correspond to specific branches in the dispersion relation for Si.[56−58] For the simulated VDoS computed by MD, the distinct features include (i) the highest-intensity peak with a maximum at approximately 470 cm^{-1}; (ii) three lower-intensity peaks with maxima at roughly 385, 335, and 232 cm^{-1}; and (iii) a broad peak with no clear maximum between frequencies 130 and 200 cm^{-1}.

All simulations performed to compute VDoS were performed under the canonical (*NVT*) ensemble held at 300 K for a simulated time of 400 ps with the velocity autocorrelation value being recorded every 0.01 ps. These simulations were dynamic, and as such it is expected that materials that are in nonequilibrium states such as systems that have mobile defect structures or systems that are in the process of undergoing a phase change are expected to evolve while the velocity autocorrelation function is being measured. This likely impacts the measurement of the velocity autocorrelation function and through that the measured VDoS. Examinations of the structures pre- and post-simulation found that little to no change in the defect content occurred during the measurement of the velocity autocorrelation function, and changes in phase transformations progressed by less than 2% of the total system volume during the measurement period. In total, we generated 770 unique VDoS measurements for Si atomic structures under a variety of compressive and disorder states. Two different metrics were measured as an analysis of human-identifiable features: the frequency where the maximum intensity of the VDoS profile was measured and a full width at half-maximum (FWHM) for the primary peak of the spectrum. The maximum intensity of the VDoS profile always aligned with the high-frequency peak while that peak was present. The FWHM was measured as the peak width at half of the maximum intensity for that profile.

Dimensionality Reduction. To detect the underlying structure of the VDoS data and circumvent issues pertaining to its high dimensionality (dimension M = 10 794), we used dimensionality reduction methods. As illustrated in Figure 1, dimensionality reduction is the mathematical mapping $\mathcal{G}$: $I \in \mathbb{R}^M \rightarrow Z \in \mathbb{R}^{L_d}$, $L_d \ll M$, of high-dimensional data (in this case 1D VDoS profile I) into a meaningful representation of the intrinsic dimensionality (compact representation of VDoS Z). The intrinsic dimensionality is the lowest number of variables (L_d) that one can use to represent the true structure of the data and capture the most salient features of the atomic structure. Although we did not try to review all possible dimensionality reduction techniques, we tested linear and nonlinear embedding techniques to conclude which techniques do well in representing 1D VDoS spectral properties. Namely, we tried principal component analysis (PCA),[59] the isometric mapping method (Isomap),[60,61] and a convolutional autoencoder.[62]

PCA was chosen as an example of a linear embedding technique and because it is a widely used technique in materials and physical sciences. It consists of an orthogonal transformation of the VDoS profile into a vector of linearly, uncorrelated principal components that are ordered so that the first L_d components retain most of the original variability in the data. Generally, PCA achieves optimal reduction when the data lies on a linear manifold (e.g., a L_d-dimensional hyperplane). However, the VDoS data may lie on a nonlinear manifold (e.g., a L_d-dimensional hypersphere). Isomap embedding and an autoencoder were chosen as examples of nonlinear embedding techniques since PCA may not necessarily be suitable. Isomap embedding is a manifold-learning algorithm that constructs a neighborhood graph among all data points and computes geodesic distances between all these points. This graph is then used to compute the low-dimensional representation of the data by applying multidimensional scaling (MDS). The MDS algorithm determines the low-dimensional representation that best preserves the interpoint distances by minimizing the cost function of error between the pairwise geodesic distances in the low-dimensional and high-dimensional representations of the data. The convolutional autoencoder is a nonlinear, artificial neural network map that learns the "coding" of the data. The encoder compresses the input (VDoS) data to a (latent) code, Z, and the decoder reconstructs the VDoS data from that code. For all these dimensionality reduction techniques, the latent dimension L_d was selected as being 10.

Regression Model. With the dimensionality of the VDoS reduced using one of the dimensionality reduction techniques described above, the problem of predicting the condition of the state of the material from an observed VDoS becomes tractable. Surrogate models, $\mathcal{F}$: $Z = \mathcal{G}(I) \rightarrow \mathbf{s}$, which map the low-dimensional representation of the VDoS, Z, to the state of the material, $\mathbf{s}$, were created using two different approaches depending on the dimensionality of the $\mathbf{s}$ vector that was being predicted. Surrogate models that predicted single values were created using a gradient boosting decision tree model.[63] This approach belongs to the ensemble-learning method that combines regression trees with boosting and consists of an additive regression model in which individual terms are simple trees that are fitted in a forward, sequential, and stagewise manner. Models that predicted multiple outputs simultaneously were performed using a decision tree regressor without boosting. The implementations of the gradient boosting decision tree and the decision tree without boosting were taken from Scikit-learn,[64] with all default parameters used for both approaches except for the number of boosting stages for the gradient boosting decision tree, which was set to 10 000. For both regression approaches and each regression task, 10 different models were created by altering the distribution of the data in the training and validation sets. All regression models were trained and tested using a 70−30% test−train split of the complete VDoS data set.

The accuracy of the trained regression models was quantified by determining the coefficient of determination (R^2) score using the `r2_score` function from Scikit-learn. For single-output regression models trained using the latent spaces produced by PCA, Isomap embedding, or the 10 different autoencoder latent spaces, 10 different regression models were trained with 10 randomly selected test−train splits, such that a maximum, average, and standard deviation could be determined from that sample of 10 regression models per dimensionality reduction technique. The averages and standard deviations for the autoencoder regression models included all 100 regression models trained with autoencoder latent spaces. For multioutput regression models, averages and standard deviations of the R^2 scores were determined using the same approach as for the single-output regression models. The maximum R^2 scores for multioutput regression models is determined by taking an average of the R^2 scores for each component of the output vector. Maximum R^2 scores presented for the components of the output vector for multioutput regression models are those that were averaged to create the highest average R^2 score.

Training of the Machine Learned Models. The fitting operations for PCA and Isomap were performed with the entire data set at once; i.e., there was no training−validation split for either PCA or Isomap. Both PCA and Isomap were fit using their respective functions from the Scikit-learn[64] Python library, with default settings being used for both PCA and Isomap with the exception of the `svd_solver` setting for PCA, which was set to `full`. As illustrated in Figure 2a, dimensionality reduction through PCA was

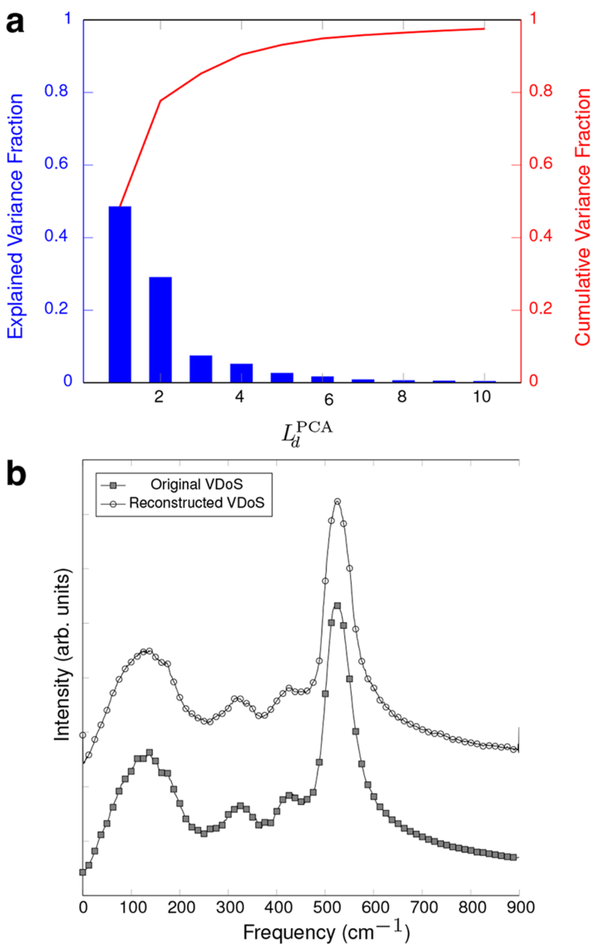


Figure 2. Training of dimensionality reduction techniques. (a) Explained variance and cumulative variance as a function of the intrinsic dimensionality, L_d. Note that L_d = 10 captures over 97.5% of the variance in our data set. (b) Quality of the VDoS reconstruction when using an autoencoder with L_d = 10. Error between the original and (autoencoder) reconstructed VDoSs is less than 3%.

able to capture over 90% of the total variance of the data set with the first four PC scores, with over 97.5% of the variance of the data set being captured when L_d^{PCA} = 10.

Training for the purpose of dimensionality reduction was also performed with an autoencoder. The encoder is composed of five convolutional layers, each with a kernel size of 10 and a stride of 3. Schematically, the encoder architecture can be described as follows: $\mathrm{Conv}^{\mathrm{ReLU}}_{8\times 3595} \times \mathrm{Conv}^{\mathrm{ReLU}}_{16\times 1196} \times \mathrm{Conv}^{\mathrm{ReLU}}_{32\times 396} \times \mathrm{Flat}_{1\times 12672} \times \mathrm{Linear}_{1\times L_d=10}$. The nomenclature "Conv" describes a convolutional layer, "Linear" describes a linear layer, and "Flat" describes a flatten operation. The superscript represents the activation function; the subscript represents the dimensionality after the layer with the first number indicating the number of channels and the second the size of the layer. The decoder architecture is composed of the inverse structure and was used during the model training process. Each autoencoder was trained for a total of 300 epochs with a batch size of 32 VDoSs and a learning rate of 0.001.

The autoencoders were trained with a training−validation split of 70% training and 30% validation. This training−validation split was created using the `random_split` function from the Pytorch Python library.[65] Ten different training−validation splits were created by altering the random seed prior to performing the split. These 10 different training−validation splits were then used to train 10 different autoencoders using identical training conditions. The loss function used during the training process was the mean squared error function, and the optimization of the parameters of the autoencoder was performed using the Adam optimizer from the Pytorch library. After each epoch, the losses computed from the training data set and the validation data set were compared to ensure that the model was not being overfit to the training data. A comparison between an input VDoS and its reconstruction from the decoder is shown in Figure 2b, with that particular reconstruction having a reconstruction error of less than 3%.

RESULTS AND DISCUSSION

Dependence of Vibrational Modes on the State of the Material. We first analyze the VDoS data based on a classical (human-identifiable) peak analysis to describe observed trends and correlations between the material state and changes in VDoS data. As illustrated in Figure 3, vibrational modes are sensitive to the global and local states of the atomic structure. Under hydrostatic compression (Figure 3a), the Si structure remains in the cubic-diamond phase throughout the entirety of the compression process, with no observed defect formation or phase change. In this case, the human-identifiable main peaks of the VDoS shift to higher frequencies with increasing pressure (or equivalently with increasing strain). The evolution of the high-amplitude peak mirrors trends observed in the 520 cm^{-1} peak seen in the Raman spectra of compressed Si.[66,67] The evolution of the VDoS under uniaxial compression (Figure 3b) is very similar to that of the hydrostatically compressed case for strain values below 15% strain. However, unlike the hydrostatic case, under uniaxial compression, the Si structure undergoes some structural transformations. Once the structural transformations begin, we note that the evolution of the VDoS starts diverging from that of the hydrostatic case, with the onset of these structural transformations aligning with the gradual splitting of the high-intensity peak in the VDoS starting at 15% strain as seen in Figure 3b.

Figure 4a illustrates the corresponding Si structure during and after the phase transformation. Different short and intermediate ordering can also be observed. The radial distribution function for atomic positions in Figure 4 shows the splitting of the second nearest neighbor peak during the phase transformation. The stress at which this phase transition begins aligns well with reported literature values for the phase transition of Si from cubic diamond to the β-SN structure[67,68] under compression. An amorphous shear band forms at around 23% compressive strain, and the system completely amorphizes at around 28% strain and a corresponding stress of 46.6 GPa. The VDoS profile in Figure 3b at 30% strain does differ substantially from the expected experimental VDoS for compressed amorphous Si.[69] This can likely be attributed to the choice of interatomic potential. However, for the purposes of this work, the primary interest is in training models that can predict the state of the material based on the profile of the VDoS, and while the inability of the chosen interatomic potential to reproduce the expected VDoS for compressed amorphous Si is regrettable, it does not impede the process of dimensionality reduction and model training.

As can be seen in Figure 4b, the gradual insertion of disorder resulted in the formation, growth, and coalescence of small amorphous domains uniformly distributed throughout the Si atomic system. Figure 3c shows that the formation and growth of these amorphous domains do not cause large shifts in peak positions in the VDoS like those observed in the uniaxial compression case, but rather that they result in a gradual peak

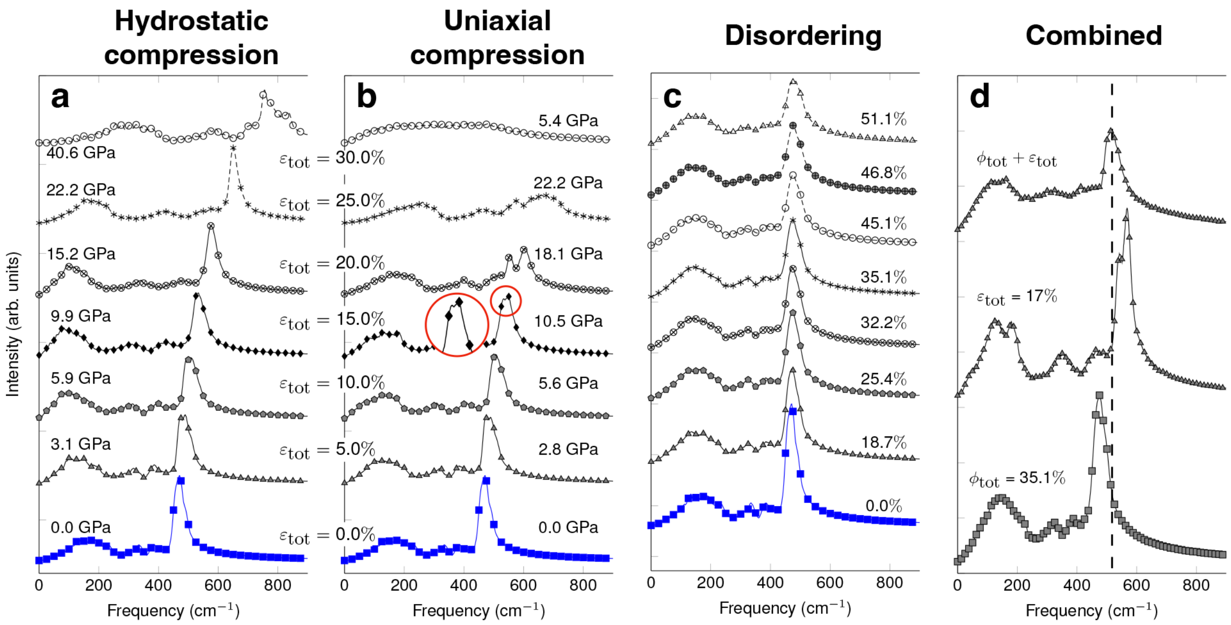


Figure 3. Changes in the VDoS in Si as a function of the state of the material for (a) hydrostatic compression, (b) uniaxial compression, (c) disorder, and (d) combined disorder and compression. For (a) and (b) the corresponding stress and strain measurements are provided next to their respective VDoS profiles. The circled area in (b) zooms in on the primary peak at 15% for the uniaxial compression case and shows the start of peak splitting. For (d), the bottom VDoS corresponds to disorder only, the middle VDoS corresponds to uniaxial compression only, and the top VDoS corresponds to disorder and compression combined. The vertical dashed line indicates the location of the human-identifiable peak for combined conditions. Profiles highlighted in blue indicate the VDoS for the pristine atomic configuration. Vertical offset is provided for clarity.

broadening and intensity softening of the spectra. In other words, as more disorder is introduced in the system, the ratio between the first and second peaks decreases, while the higher frequency peaks broaden at the same time. As seen in Figure 3d, when compression and disorder are combined, the spectral signal becomes convoluted and complex. Indeed, we observed a shift in the primary peak that lies between the primary peak when disorder alone is present with no stress which has a lower frequency than that of the primary peak when uniaxial compression is solely applied which has a higher frequency. In addition, we note a peak broadening and reduction of the intensity of the primary peak for combined effects, but no peak splitting as observed in the uniaxial compression case, even though disorder domains are present in the atomic system. These results show that the VDoS spectrum is not a simple superposition of the disorder and compression cases. They illustrate the complexity of interpreting the microstructure state from the VDoS spectrum.

The presence of disorder in the atomic structure, as is the case in the uniaxial compression case, the disorder case, or when disorder and compression are combined, also results in the emergence of an internal characteristic length scale, l_ϕ, which changes with the state of the material. The differences of this characteristic length scale as a function of the configuration of the atomic structures are illustrated in Figure 4c. As expected, in the case of the hydrostatic case, the internal characteristic length scale is constant and relates to the first nearest neighbor by construction. In the uniaxial compression case, the internal characteristic length scale only emerges when the amorphization starts and corresponds to the peak splitting observed in the VDoS (Figure 3b). The drop in l_ϕ at 23% strain corresponds to the formation of the amorphous shear band previously discussed. However, no obvious feature in the VDoS corresponds to this nonmonotonic change of l_ϕ. For the disorder configurations, we observe a progressive increase in the internal characteristic length scale with increasing disorder. This gradual change seems to correlate with the progressive broadening and reduction in intensity of the main human-identifiable peak in the VDoS (Figure 3c). When disorder and compression are combined, we observe that the preexisting presence of disorder (at zero strain) corresponds to a nonzero internal characteristic length scale and then this characteristic length scale increases with increasing strain. In this case, it is hard to infer any obvious qualitative correlation between changes in the internal characteristic length scale and changes in the VDoS.

When individual conditions are applied (i.e., compression without the introduction of disorder or conversely disorder without compression), it is conceivable that simple metrics such as shifts and/or broadening of human-identifiable peaks could be used to create models that predict individual material conditions such as applied strain, stress, or the level of disorder. These simple correlations are evidenced for example in Figure 5 (additional analyses are provided in the Supporting Information), for which we observe some correlations between stress and peak position in the case of the hydrostatic and uniaxial compression (Figure 5a,b) or between the disorder parameter and FWHM in the case of the disordered configuration (Figure 5c). However, this approach is nonetheless fraught with biases and potential errors when selecting and measuring the appropriate human-identifiable features[19,20,34] and does not necessarily reflect the nonmonotonic behavior observed for the material descriptors. For instance, the selection of a human-identifiable peak in the case of the uniaxial case when amorphization starts is not obvious, or as shown with the lack of correlations between the disordering parameter and peak location in Figure 5c. Additionally, when complex atomic configurations are considered together, as

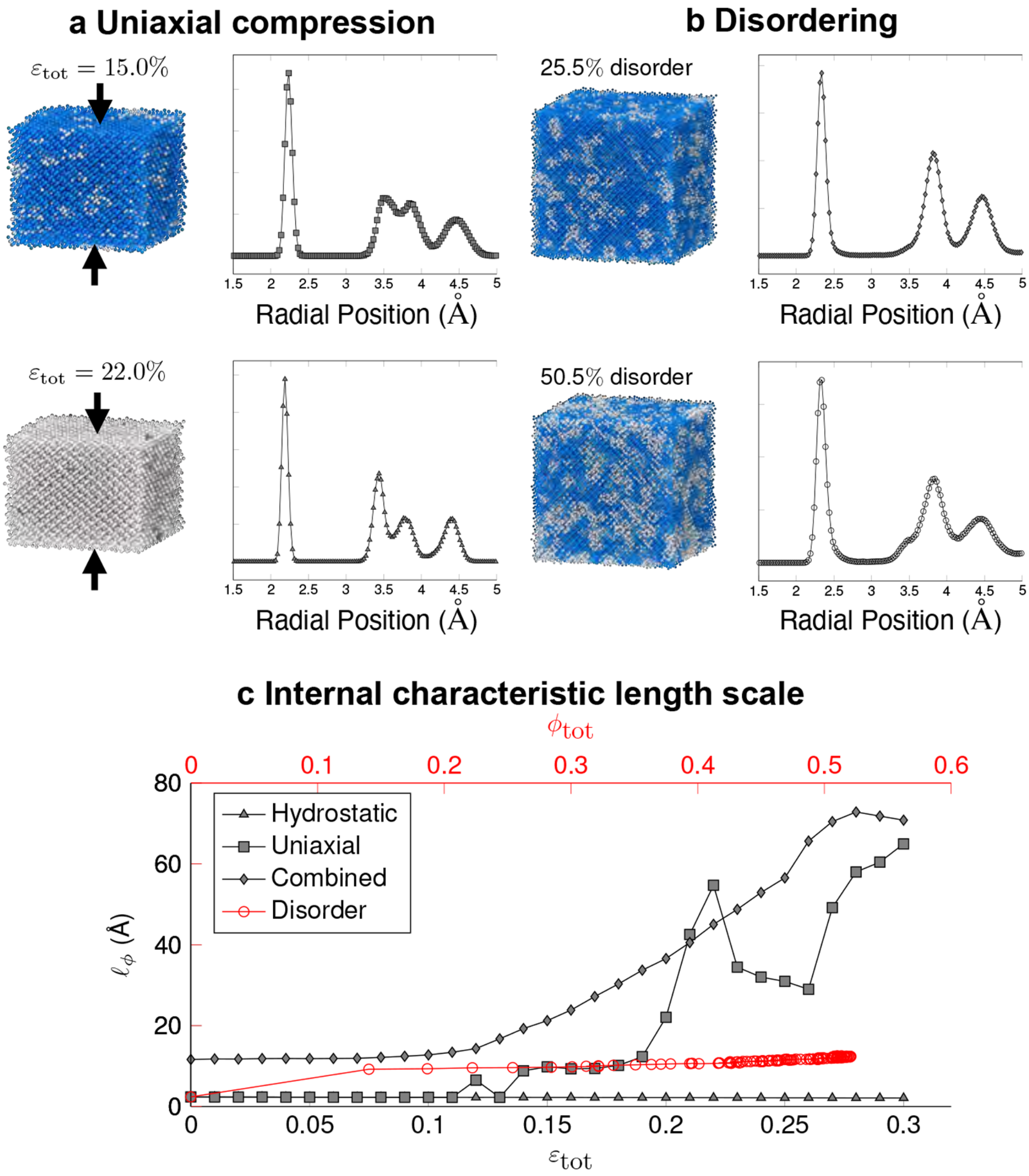


Figure 4. Change of the Si atomic structure (a) under uniaxial compression or (b) with the gradual insertion of disorder. The radial distribution functions ($g(r)$) quantify the change from ordered structures to amorphous/disordered structures. Blue atoms indicate atoms in the cubic diamond configuration; white atoms indicate atoms not in the cubic diamond phase. (c) Evolution of the internal characteristic length scale, l_ϕ, for different states of the atomic structure as a function of the total strain (x_1-axis for hydrostatic and uniaxial compressions and when disorder and compression are combined) and as a function of the disorder parameter (x_2-axis when only disorder is present).

illustrated in Figure 5d, or when material conditions become more complicated, the ability to use human-identifiable features to produce predictive models becomes less feasible. We posit that these spectral profiles actually do contain such subtleties and that sensitivities to the state of the material can be captured through manifold learning.[70]

Low-Dimensional Representations of Vibrational Spectra. As such, we now turn our attention to how the changes observed in the VDoS spectra are captured in their low-dimensional representations. Figure 6 shows representative results for the distribution of the full 770 VDoS data set (i.e., it comprises all of the hydrostatic and uniaxial compressions, disordered, and combined disordered + compression configurations) for the first four latent dimensions produced through the various dimensionality reduction techniques we tested. For PCA (Figure 6a), the latent dimensions are ordered such that the first dimension captures the greatest amount of variance in the data set, and each following dimension captures less variance than the previous one. As shown previously in Figure 2, the first four dimensions capture 90.4% of the total variance in the data set. Taking a PCA latent dimension $L_d^{\text{PCA}} = 10$ captures over 97%. For this linear embedding technique, we observe that for the VDoS data there is a strong, nonlinear association among the various dimensions. We make a similar observation when the VDoS data is represented via Isomap (Figure 6b) with an Isomap latent dimension $L_d^{\text{Isomap}} = 10$. Conversely, the latent representation of the VDoS data using the autoencoder ($L_d^{\text{autoencoder}} = 10$) shows a more compact representation in latent space. This comparison illustrates that, in the cases of PCA and Isomap, the VDoS data is more spread

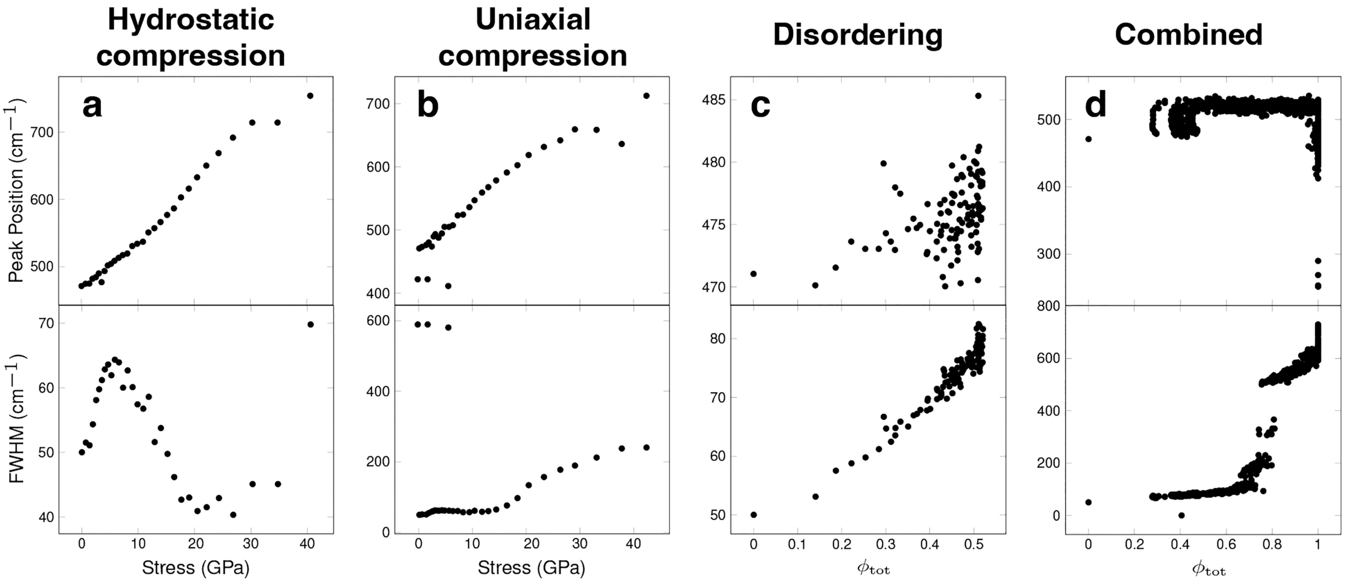


Figure 5. Peak analysis (peak location and FWHM) of the VDoS as a function of the state of the material for (a) hydrostatic compression, (b) uniaxial compression, (c) disorder, and (d) combined disorder and compression. For (a) and (b), the analysis is plotted against the average stress, σ_{tot}. For (c) and (d), the analysis is plotted against the disorder parameter, ϕ_{tot}.

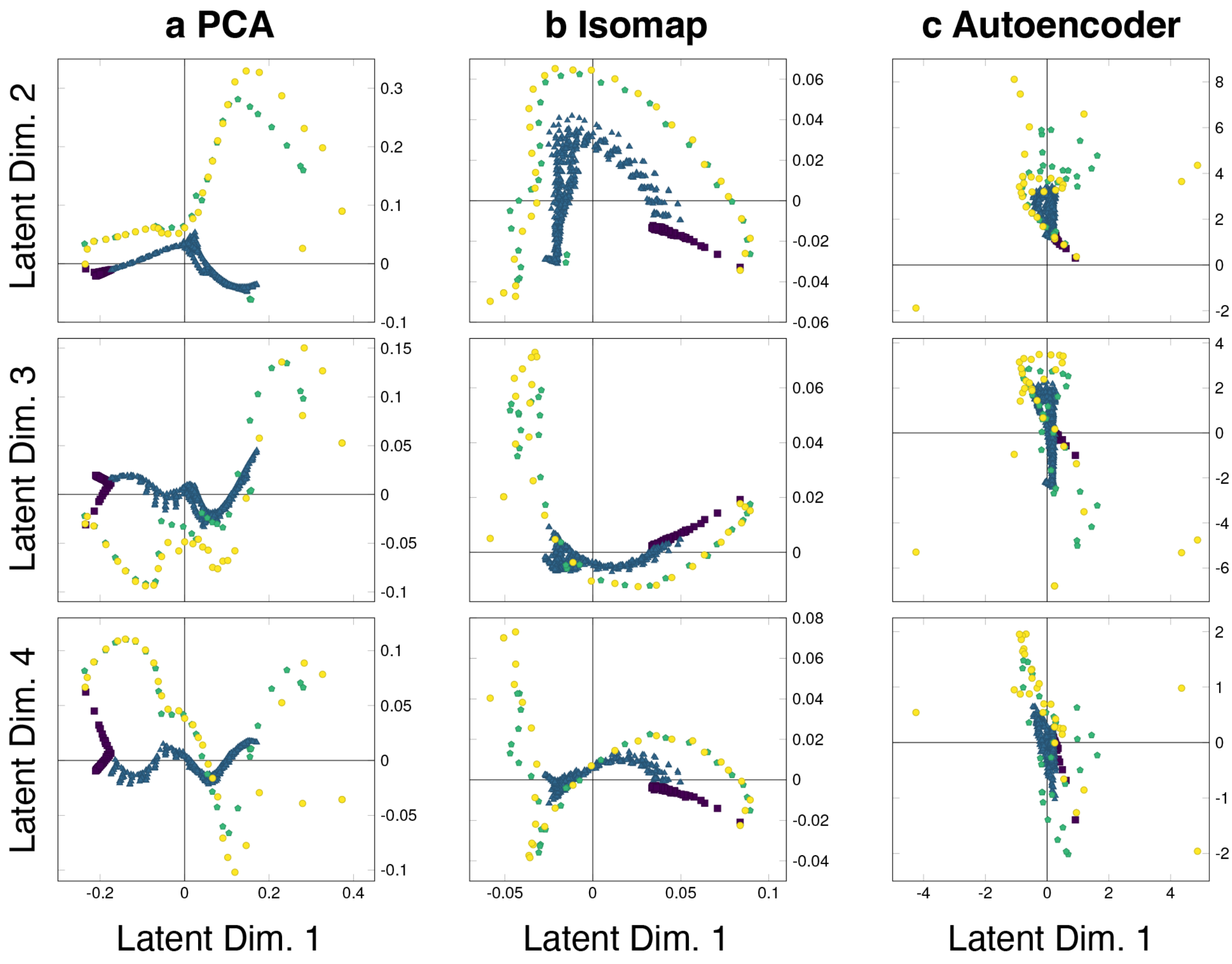


Figure 6. Latent representations of VDoS using (a) PCA, (b) Isomap, and (c) an autoencoder. Only the first four dimensions (out of 10) are represented. Scales of latent dimensions are unique per latent representation and cannot be compared. Symbol nomenclature is as follows: circles (yellow) are for the hydrostatic data set, pentagons (green) are for the uniaxial data set, squares (dark blue) are for the disorder insertion data set, and triangles (light blue) are for the combined data set. Points are colored according to the peak position in the VDoS profile.

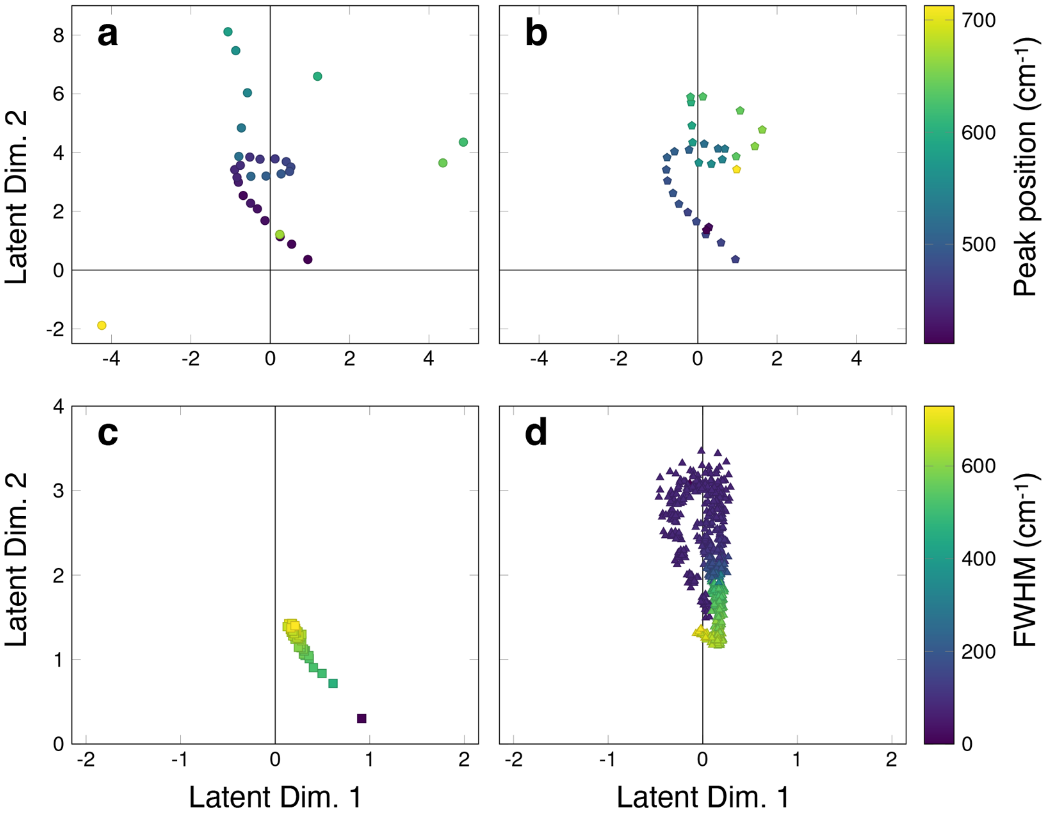


Figure 7. Latent representations of VDoS using autoencoder projection (first two dimensions) for (a) hydrostatic compression (circles), (b) uniaxial compression (pentagons), (c) disordering (squares), and (d) combined disordering and compression (triangles). Points for (a) and (b) are colored according to peak position, and points for (c) and (d) are colored according to FWHM.

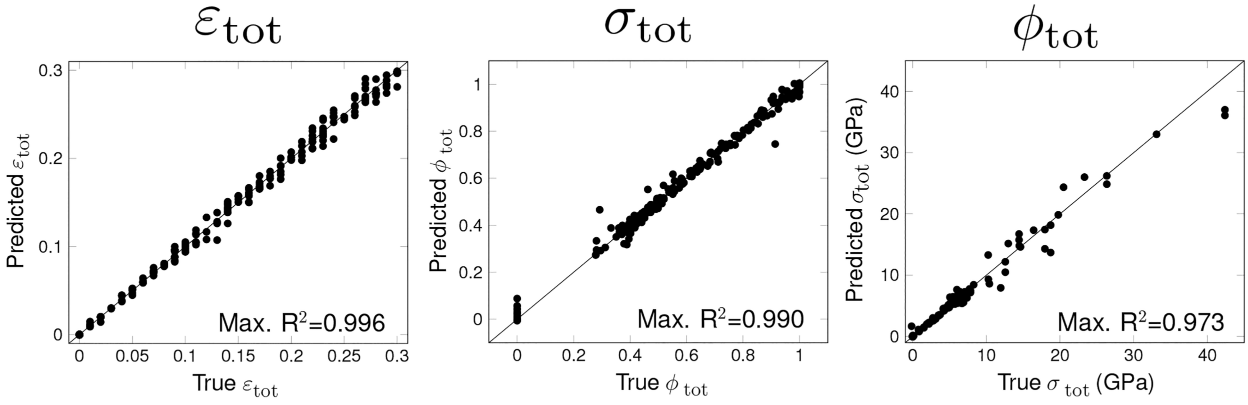


Figure 8. Parity plots for the single-output regression models trained using data reduced via the autoencoder. Parity plots are for models trained to predict the following material conditions: total strain (ε_{tot}), stress (σ_{tot}), and fraction of disordered atoms (ϕ_{tot}).

out than in the case of the autoencoder. Additionally, we note that, when using PCA or Isomap, it can be hard to pick up the distinction among the different atomic configurations (hydrostatic compression versus uniaxial compression) when the latent dimension is colored and labeled as a function of the material state (hydrostatic in yellow circles and uniaxial in green pentagon symbols). However, the representations of the different configurations are more distinct when using an autoencoder.

Figure 7 provides further insights into the interpretation of the low-dimensional representation of the VDoS when separated by the configuration of the atomic structure. Panels a and b of Figure 7 capture the changes observed in the VDoS for the hydrostatic and uniaxial compressions, while panel c represents changes due to disordering and panel d shows the changes due to the combined effects of disordering and uniaxial compression. We observe in Figure 7a,b that the reduced representations between hydrostatic and uniaxial compressions are similar and nonlinear for strains ε_{tot} at or below 15% deformation and that the data is ordered as a function of the human-identifiable features such as peak position, for instance. This observation is expected since we could not distinguish noticeable differences in the human-identifiable peaks in the actual VDoS. The deviation between the two representations for hydrostatic and uniaxial compressions starts for a deformation state above 15% when coincidentally the atomic system starts to amorphize and develops an internal characteristic length scale in the case of the uniaxial compression. In the case of disordering (Figure 7c), we observe a compact and linear representation of the VDoS that seems to correlate with the gradual broadening of the human-identifiable peaks in the actual VDoS. When disordering and compression are combined, we also note a more compact representation as compared to the compression cases in Figure 7a,b. However, the low-dimensional representation of the VDoS for the combined case appears to be more complex than a simple superposition of the disordering (Figure 7c) and compression cases (Figure 7a,b), indicating complex correlations between the atomic configuration and its VDoS spectral signature.

Predictions of Material Conditions from Latent Representations. With the data prepared through dimensionality reduction, the task of creating predictive models for connecting the VDoS spectra to material conditions can be performed. Figure 8 shows parity plots of single-output regression models trained on the latent space produced by the autoencoder. These plots represent the models that had the highest R^2 scores across all the single-output models trained on the latent spaces produced via the different autoencoders we tested. These parity plots show that our regression models are accurate when a single-output material descriptor is predicted from an observed VDoS spectrum. This statement is especially true when predicting the applied strain or fraction of disordered atoms. However, we also note that a single-output regression model is less accurate to predict the average stress given the noted uncertainty in the intermediate stress regime above 10 GPa. Table 1 further reinforces these observations by reporting the average R^2 scores and their standard deviations for the single-output regression models trained on data reduced through different dimensionality reduction techniques. In Table 1 we note that, regardless of the dimensionality reduction technique used, the prediction of the average stress is more difficult as reflected by the lower R^2 values. Note that the resolution of the VDoS (i.e., the number of points M in the VDoS) affects the accuracy of our regression (see the analysis provided in the Supporting Information). However, one could circumvent this issue by fitting the potentially downsampled VDoS with Gaussian and Lorentzian functions to get a reasonable VDoS resolution.

Table 1. Predictions from the Single-Output Regression Models for Different Dimensionality Reduction Techniques

dim. red.	target descriptor	avg R^2	std dev	max R^2
PCA	ε_{tot}	0.9944	0.00073	0.9953
	σ_{tot}	0.8479	0.1419	0.9583
	ϕ_{tot}	0.9590	0.0215	0.9877
Isomap	ε_{tot}	0.9910	0.0023	0.9932
	σ_{tot}	0.8747	0.0635	0.9530
	ϕ_{tot}	0.8717	0.0683	0.9617
autoencoder	ε_{tot}	0.9921	0.0063	0.9960
	σ_{tot}	0.8334	0.1520	0.9730
	ϕ_{tot}	0.9212	0.0473	0.9904

Despite their good performance, these single-output regression models are incapable of making distinctions between the different loading conditions, as the information related to the loading condition is not represented within a single material condition. We therefore trained multi-output regression models that allow for the prediction of material descriptors reflecting such information on the state of the atomic structure. When these multi-output regression models are trained on the *xx*, *yy*, and *zz* components of the stress tensor, the model can distinguish between the hydrostatic loading and uniaxial loading that was applied to the Si atomic structures. In addition, this multi-output model also predicts the deformation state, ε_{tot}, and the two metrics associated with disorder of the structure, namely ϕ_{tot} and l_ϕ.

Figure 9 shows parity plots produced for each element of the material descriptor vector **s** for the multi-output regression model with the highest R^2 value trained on a latent space produced by an autoencoder. Comparing the R^2 values for ε_{tot}, ϕ_{tot}, and the different components of the stress tensor from Figure 9 to those shown in Figure 8, we note that for ε_{tot} and ϕ_{tot} we obtained a slight reduction in the value of the R^2 scores, while the scores for the stress components are comparable to the total stress score from the single-output regression model. Examining the distribution of the data points in Figure 9 to those in Figure 8, it can be seen that the noted decrease in the R^2 scores for comparable conditions seems to be primarily driven by the presence of several outliers in the predictions of the multi-output regression model, while the majority of the points are distributed similarly between the two approaches. Table 2 reports the average R^2 scores for multi-output regression models trained with different dimensionality reduction techniques to predict the full material descriptor vector **s**, reinforcing the observations made by comparing Figures 8 and 9, where expanding the regression model to predict multiple material conditions results in a decrease in the accuracy of the predictions for ε_{tot} and ϕ_{tot} while maintaining the accuracy of the predictions of the stress. Predictions of the characteristic length scale l_ϕ are quite good from the multi-output regression models, with the error increasing as l_ϕ increases beyond 80 Å, corresponding with atomic structures that have neared complete amorphization.

Robustness and Sensitivity to Noise. In Figure 10, we illustrate the accuracy and robustness of our trained model by systematically increasing the noise levels in the VDoS data and compare the predictions to their true values. We randomly selected a VDoS spectrum from our validation set for which we added Gaussian white noise to our VDoS data with zero mean and a standard deviation of ranging from 20 to 40%. We also considered impulsive noise by randomly selecting 100−300 points in our VDoS data with no noise, adding spikes to those points resulting in 1−3% spikes (appearing as sharp vertical lines in Figure 10) and then adding 40% Gaussian white noise on top of it. In the selected VDoS example in Figure 10, the Si atomic structure underwent combined disorder and compression loading. As seen in Table 3, our autoencoder-based protocol (i.e., when we use the autoencoder as the dimensionality reduction technique) proves to be exceptionally resilient to noise and is almost insensitive to noise up to 40% Gaussian white noise with and without impulse noise. Indeed, we note that the errors between the true and predicted values for all the material descriptors do not change noticeably when noise is added to the VDoS data (~5% relative error), with roughly the same relative error as compared to predictions without noise. The predictions deteriorate a little when impulse noise is present, especially for the disorder and internal characteristics length, but they nonetheless remain relatively accurate (~10% relative error). Such denoising

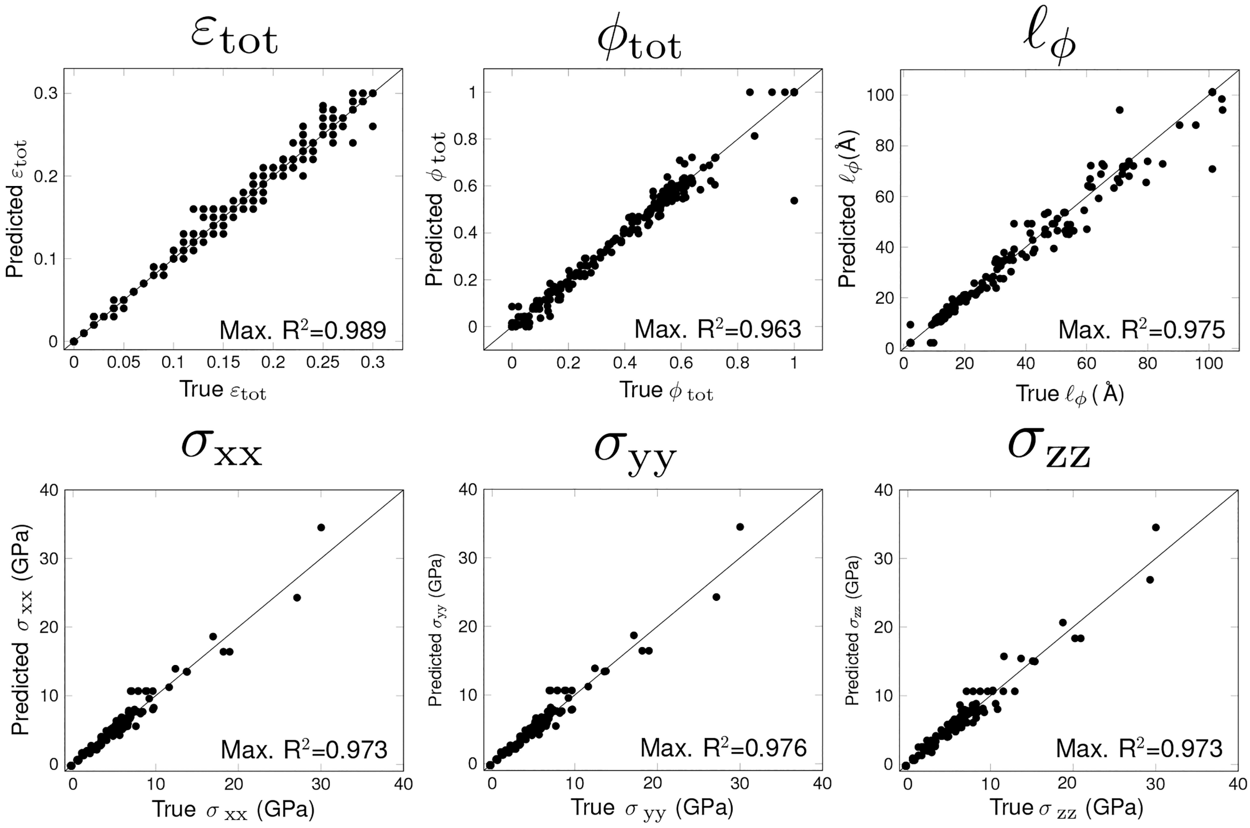


Figure 9. Parity plots for the multi-output regression models trained using data reduced via the autoencoder. Parity plots are for models trained to predict the following material conditions: total strain (ε_{tot}), disorder parameter (ϕ_{tot}), internal characteristic length scale (l_ϕ), and the stress tensor (σ_{ii}, i = x, y, or z).

Table 2. Predictions from the Multi-Output Regression Models for Different Dimensionality Reduction Techniques

dim. red.	target descriptor	avg R^2	std dev	max R^2
PCA	ε_{tot}	0.9787	0.0240	0.9887
	σ_{xx}	0.8540	0.1971	0.9740
	σ_{yy}	0.8530	0.1987	0.9752
	σ_{zz}	0.8448	0.1949	0.9757
	ϕ_{tot}	0.8494	0.0846	0.9844
	l_ϕ	0.9186	0.0388	0.9505
	avg score	0.8831	0.1140	0.9748
Isomap	ε_{tot}	0.9837	0.0055	0.9871
	σ_{xx}	0.8717	0.1215	0.9590
	σ_{yy}	0.8730	0.1224	0.9611
	σ_{zz}	0.8723	0.1108	0.9550
	ϕ_{tot}	0.7821	0.1037	0.9274
	l_ϕ	0.9332	0.0187	0.9621
	avg score	0.8860	0.0771	0.9586
autoencoder	ε_{tot}	0.9800	0.0165	0.9895
	σ_{xx}	0.8479	0.1082	0.9730
	σ_{yy}	0.8480	0.1093	0.9757
	σ_{zz}	0.8470	0.1075	0.9727
	ϕ_{tot}	0.8199	0.0968	0.9630
	l_ϕ	0.9109	0.0578	0.9746
	avg score	0.8756	0.0726	0.9748

capability is a known attribute of autoencoders.[71,72] Using an autoencoder as the dimensionality reduction technique does exactly that by filtering out the noise in the VDoS data and only retaining its dominant features in the latent space. As tabulated in Table 3, we note however that, when using PCA as the dimensionality technique, the predictions are not as good as the ones with the autoencoder for high noise level or when impulse noise is added to the white noise VDoS data. This observation is true across the board with 20−100% relative error on the stress and disorder predictions. By comparing the performance of the PCA-based protocol with that of the autoencoder-based protocol, we exemplify the importance of the choice of dimensionality reduction technique in the robustness of our protocol. Overall, the performance of the autoencoder-based protocol demonstrates the robustness to interferences (noise, spikes, baseline drift) potentially occurring during the acquisition of spectroscopic data which could lead to errors in the subsequent analysis of the spectra when using human-identifiable peaks, for instance. This denoising capability also bypasses the need for spectral preprocessing methods[73,74] meant to clean up the spectra prior to any analysis, limiting yet another source of error in the interpretation of those spectra.

■ CONCLUSIONS

Throughout this work, we presented a simple, reliable, and robust protocol that enables an extended mapping from vibrational spectra to a variety of complex configurations of atomic structures. The models connecting spectra to the state of the materials are based on two elements of supervised manifold learning: the representation of the VDoS spectra via dimensionality reduction techniques and a (decision-tree)

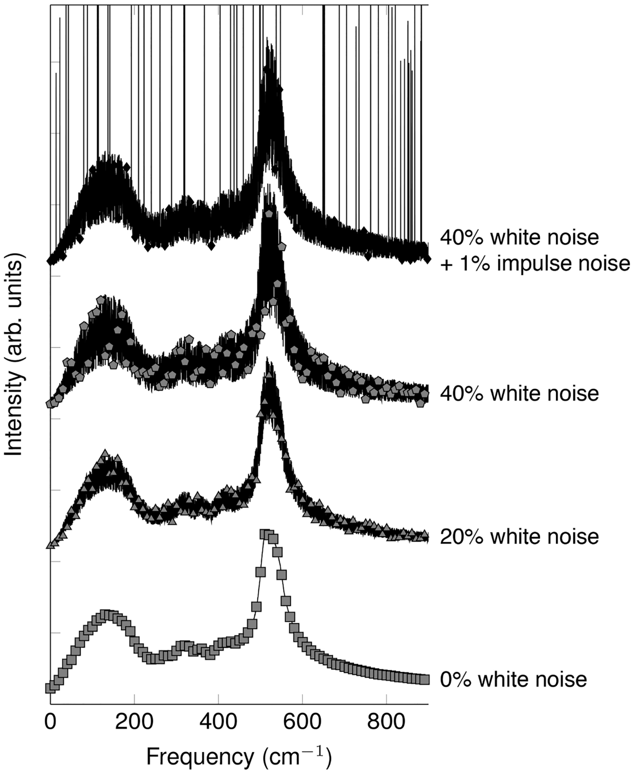


Figure 10. Example of VDoS spectra with increasing white and impulse noise. VDoS with 0% white noise is also plotted as a reference. The vertical lines showing up for the 40% white noise + 1% impulse noise case correspond to the impulse noise added on top of the white noise. For all VDoS spectra plotted, the corresponding Si atomic system underwent disorder (starting disorder was 49.5% prior to compression) followed by 10% uniaxial compression.

regression model that uses the reduced representation of the VDoS to decode structural information unavailable through classical human-identifiable peak analysis. The combination of these techniques results in a model that takes an observed VDoS spectrum as input and predicts a vector of material state descriptors characteristic of the atomic structure as output. These models were trained on over 700 simulated VDoS spectra for Si atomic systems undergoing various deformation and disordering states. Despite the fact that some spectra can be very similar for very different configurations, we show that our trained models accurately and robustly disentangle the contribution from the different material states (e.g., hydrostatic versus uniaxial compression) with an accuracy of over 97% even in the presence of white noise, hence demonstrating that these spectroscopic profiles do contain comprehensive information on the state of the atomic structure beyond our own (subject matter expert) cognition. We show that when using an autoencoder as the technique to provide a low-dimensional representation of the VDoS, the protocol is robust to noise present in the spectroscopic profile and maintains a good accuracy. The overall protocol is fast and easy to use and can assist spectroscopy practitioners to quickly identify complex atomic structure configurations, such as those measured during pressure-induced instabilities in materials. While we demonstrated our approach on Si as a model material system, this work can be generalized to a broader class of materials. Going forward, we note that our framework could be extended to other structural descriptors, other loading conditions, other materials, and other 1D spectroscopic techniques. In the work presented here, we assessed the local chemical environmental information in crystalline and amorphous Si, but the same protocol could be equally applied if we had extended our list of material descriptors (for instance, by including the addition of an intermediate-ordering descriptor) or if we had applied it to a different material system with a different crystalline structure. In the latter case, the practitioner would need to predefine a comprehensive list of material state descriptors representative of that specific material. For example, in a face-centered-cubic (fcc) system, in addition to the strain and stress descriptors, we may want to consider dislocation density and other characteristics of the dislocation network as descriptors. In a porous material, void density and percolation could be defined as representative descriptors, or similarly in a nanocrystalline material, the grain size and grain size distribution are natural descriptors. Finally, while we used simulations to illustrate our concept and serve as ground truths, this needs not to be the case. Experimental data (such as surface topography or stress measurements) can also be used. However, the generation and collection of multimodal

Table 3. Impact of Noise Level on Accuracy of Multi-Output Regression Model When Using the Autoencoder or PCA as a Dimensionality Reduction Technique

		predicted value (% relative error)				
target descriptor	true	no noise	20% noise	40% noise	40% noise + 1% spike	40% + 3% spike
			Autoencoder			
ε_{tot}	0.1	0.09 (10%)	0.09 (10%)	0.09 (10%)	0.12 (20%)	0.11 (10%)
σ_{xx}(GPa)	5.4249	4.9892 (8.03%)	4.9892 (8.03%)	4.9892 (8.03%)	5.6403 (3.97%)	5.4836 (1.08%)
σ_{yy}(GPa)	5.4334	5.0166 (7.67%)	5.0166 (7.67%)	5.0166 (7.67%)	5.6595 (4.16%)	5.4608 (0.50%)
σ_{zz}(GPa)	7.5614	6.9431 (8.18%)	6.9431 (8.18%)	6.9431 (8.18%)	7.8371 (3.65%)	7.6275 (0.87%)
ϕ_{tot}	0.4745	0.4701 (0.93%)	0.4701 (0.93%)	0.4701 (0.93%)	0.5298 (11.66%)	0.5263 (10.92%)
l_{ϕ}(Å)	13.8242	13.8079 (0.12%)	13.8079 (0.12%)	13.8079 (0.12%)	15.97 (15.50%)	15.38 (11.22%)
			PCA			
ε_{tot}	0.1	0.09 (10%)	0.09 (10%)	0.09 (10%)	0.12 (20%)	0.12 (20%)
σ_{xx}(GPa)	5.4249	4.9892 (8.03%)	4.9623 (8.53%)	4.9623 (8.53%)	6.0658 (11.81%)	7.0963 (30.81%)
σ_{yy}(GPa)	5.4334	5.0166 (7.67%)	4.9761 (8.42%)	4.9761 (8.42%)	6.0621 (11.57%)	7.0828 (30.36%)
σ_{zz}(GPa)	7.5614	6.9431 (8.18%)	6.8917 (8.86%)	6.8917 (8.86%)	9.0083 (19.14%)	7.105 (6.03%)
ϕ_{tot}	0.4745	0.4701 (0.93%)	0.4718 (0.58%)	0.4718 (0.58%)	0.0176 (96.30%)	0.0 (100%)
l_{ϕ}(Å)	13.8242	13.8079 (0.12%)	13.4853 (2.45%)	13.4853 (2.45%)	6.5195 (52.84%)	2.245 (83.76%)

(experimental) data to be used to construct material state descriptors can be potentially cumbersome and costly.

ASSOCIATED CONTENT

Data Availability Statement

VDoS data and Jupyter Notebook for the entire workflow can be requested from the corresponding author.

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.chemmater.2c03207.

Validation of VDoS calculations by comparison of density functional theory vs molecular dynamics; peak analysis as a function of materials descriptors; effect of VDoS resolution on regression accuracy (PDF)

AUTHOR INFORMATION

Corresponding Author

Rémi Dingreville − *Center for Integrated Nanotechnologies, Sandia National Laboratories, Albuquerque, New Mexico 87185, United States;* orcid.org/0000-0003-1613-695X; Email: rdingre@sandia.gov

Authors

Daniel Vizoso − *Center for Integrated Nanotechnologies, Sandia National Laboratories, Albuquerque, New Mexico 87185, United States;* orcid.org/0000-0002-7733-6392

Ghatu Subhash − *Department of Mechanical and Aerospace Engineering, University of Florida, Gainesville, Florida 32611, United States;* orcid.org/0000-0002-5996-0909

Krishna Rajan − *Department of Materials Design and Innovation, University at Buffalo, Buffalo, New York 14260, United States;* orcid.org/0000-0001-9303-2797

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.chemmater.2c03207

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

The authors would like to thank J. Startt from Sandia National Laboratories for his help with the validation of the VDoS by DFT. The authors would also like to thank N. Trask from Sandia National Laboratories for insightful comments and review of this work. Computational and modeling resources are supported by the Center for Integrated Nanotechnologies, an Office of Science User Facility operated for the U.S. Department of Energy. K.R. acknowledges the support from NSF Award No. 1640867-DIBBs: EI: Data Laboratory for Materials Engineering. This article has been authored by an employee of National Technology & Engineering Solutions of Sandia, LLC, under Contract No. DE-NA0003525 with the U.S. Department of Energy (DOE). The employee owns all right, title, and interest in and to the article and is solely responsible for its contents. The United States Government retains and the publisher, by accepting the article for publication, acknowledges that the United States Government retains a nonexclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this article or allow others to do so, for United States Government purposes. The DOE will provide public access to these results of federally sponsored research in accordance with the DOE Public Access Plan (https://www.energy.gov/sites/prod/files/2014/08/f18/DOE_Public_Access%20Plan_FINAL.pdf).

REFERENCES

(1) Karki, B. B.; Wentzcovitch, R. M. Vibrational and quasiharmonic thermal properties of CaO under pressure. *Phys. Rev. B* **2003**, *68*, 224304.

(2) Li, Y.; Zhang, L.; Cui, T.; Ma, Y.; Zou, G.; Klug, D. D. Phonon instabilities in rocksalt AgCl and AgBr under pressure studied within density functional theory. *Phys. Rev. B* **2006**, *74*, 054102.

(3) Lukačević, I.; Gupta, S. K.; Jha, P. K.; Kirin, D. Lattice dynamics and Raman spectrum of rutile TiO_2: The role of soft phonon modes in pressure induced phase transition. *Mater. Chem. Phys.* **2012**, *137*, 282−289.

(4) May, A. F.; Delaire, O.; Niedziela, J. L.; Lara-Curzio, E.; Susner, M. A.; Abernathy, D. L.; Kirkham, M.; McGuire, M. A. Structural phase transition and phonon instability in $Cu_{12}Sb_4S_{13}$. *Phys. Rev. B* **2016**, *93*, 064104.

(5) Lin, Y.-C.; Erhart, P.; Bettinelli, M.; George, N. C.; Parker, S. F.; Karlsson, M. Understanding the interactions between vibrational modes and excited state relaxation in $Y_{3-x}Ce_xAl_5O_{12}$: Design principles for phosphors based on 5 d-4 f transitions. *Chem. Mater.* **2018**, *30*, 1865−1877.

(6) Jarry, A.; Walker, M.; Theodoru, S.; Brillson, L. J.; Rubloff, G. W. Elucidating structural transformations in $Li_xV_2O_5$ electrochromic thin films by multimodal spectroscopies. *Chem. Mater.* **2020**, *32*, 7226−7236.

(7) Mizuno, H.; Mossa, S.; Barrat, J.-L. Elastic heterogeneity, vibrational states, and thermal conductivity across an amorphisation transition. *EPL (Europhys. Lett.)* **2013**, *104*, 56001.

(8) Meyer, R.; Lewis, L. J.; Prakash, S.; Entel, P. Vibrational properties of nanoscale materials: From nanoparticles to nanocrystalline materials. *Phys. Rev. B* **2003**, *68*, 104303.

(9) Meyer, R.; Comtesse, D. Vibrational density of states of silicon nanoparticles. *Phys. Rev. B* **2011**, *83*, 014301.

(10) Sauceda, H. E.; Salazar, F.; Pérez, L. A.; Garzón, I. L. Size and shape dependence of the vibrational spectrum and low-temperature specific heat of Au nanoparticles. *J. Phys. Chem. C* **2013**, *117*, 25160−25168.

(11) Carles, R.; Benzo, P.; Pécassou, B.; Bonafos, C. Vibrational density of states and thermodynamics at the nanoscale: the 3D-2D transition in gold nanostructures. *Sci. Rep.* **2016**, *6*, 1−10.

(12) Phelan, D.; Millican, J. N.; Thomas, E. L.; Leão, J. B.; Qiu, Y.; Paul, R. Neutron scattering measurements of the phonon density of states of $FeSe_{1-x}$ superconductors. *Phys. Rev. B* **2009**, *79*, 014519.

(13) Trachet, A.; Subhash, G. Microscopic and spectroscopic investigation of phase evolution within static and dynamic indentations in single-crystal silicon. *Mater. Sci. Eng. A* **2016**, *673*, 321−331.

(14) Raeliarijaona, A.; Cohen, R. E. First-principles calculations of Raman and infrared spectroscopy for phase identification and strain calibration of hafnia. *Appl. Phys. Lett.* **2022**, *120*, 242903.

(15) Balasubramaniam, R.; Kumar, A. V. R. Characterization of Delhi iron pillar rust by X-ray diffraction, Fourier transform infrared spectroscopy and Mössbauer spectroscopy. *Corros. Sci.* **2000**, *42*, 2085−2101.

(16) Agarwal, A.; Tomozawa, M. Correlation of silica glass properties with the infrared spectra. *J. Non-Cryst. Solids* **1997**, *209*, 166−174.

(17) Mittal, R.; Chaplot, S. L.; Schober, H.; Kolesnikov, A. I.; Loong, C.-K.; Lind, C.; Wilkinson, A. P. Negative thermal expansion in cubic $ZrMo_2O_8$: Inelastic neutron scattering and lattice dynamical studies. *Phys. Rev. B* **2004**, *70*, 214303.

(18) Antonov, V. E.; Fedotov, V. K.; Ivanov, A. S.; Kolesnikov, A. I.; Kuzovnikov, M. A.; Tkacz, M.; Yartys, V. A. Lattice dynamics of high-pressure hydrides studied by inelastic neutron scattering. *J. Alloys Compd.* **2022**, *905*, 164208.

(19) Kunka, C.; Boyce, B. L.; Foiles, S. M.; Dingreville, R. Revealing inconsistencies in X-ray width methods for nanomaterials. *Nanoscale* **2019**, *11* (46), 22456−22466.

(20) Weidenthaler, C. Pitfalls in the characterization of nanoporous and nanosized materials. *Nanoscale* **2011**, *3*, 792−810.

(21) Daniel, C.*Density-Functional Methods for Excited States*; Springer: 2015; pp 377−413. DOI: 10.1007/128_2015_635

(22) Yaguchi, M.; Uchida, T.; Motobayashi, K.; Osawa, M. Speciation of adsorbed phosphate at gold electrodes: a combined surface-enhanced infrared adsorption spectroscopy and DFT study. *J. Phys. Chem. Lett.* **2016**, *7*, 3097−3102.

(23) Stewart, J. A.; Brookman, G.; Price, P.; Franco, M.; Ji, W.; Hattar, K.; Dingreville, R. Characterizing single isolated radiation-damage events from molecular dynamics via virtual diffraction methods. *J. Appl. Phys.* **2018**, *123*, 165902.

(24) Awasthi, A. P.; Subhash, G. High-pressure deformation and amorphization in boron carbide. *J. Appl. Phys.* **2019**, *125*, 215901.

(25) Yin, Y.; Wang, B.; E, Y.; Yao, J.; Wang, L.; Bai, X.; Liu, W. Raman spectra and phonon structures of $BaGa_4Se_7$ crystal. *Commun. Phys.* **2020**, *3*, 34.

(26) Mishra, A.; Kunka, C.; Echeverria, M. J.; Dingreville, R.; Dongare, A. M. Fingerprinting shock-induced deformations via diffraction. *Sci. Rep.* **2021**, *11*, 1−12.

(27) Kusne, A. G.; Gao, T.; Mehta, A.; Ke, L.; Nguyen, M. C.; Ho, K.-M.; Antropov, V.; Wang, C.-Z.; Kramer, M. J.; Long, C.; Takeuchi, I. On-the-fly machine-learning for high-throughput experiments: search for rare-earth-free permanent magnets. *Sci. Rep.* **2014**, *4*, 1437−1451.

(28) Sun, S.; et al. Accelerated development of perovskite-inspired materials via high-throughput synthesis and machine-learning diagnosis. *Joule* **2019**, *3* (6), 1437−1451.

(29) Cherukara, M. J.; Zhou, T.; Nashed, Y.; Enfedaque, P.; Hexemer, A.; Harder, R. J.; Holt, M. V. AI-enabled high-resolution scanning coherent diffraction imaging. *Appl. Phys. Lett.* **2020**, *117*, 044103.

(30) Kaufmann, K.; Zhu, C.; Rosengarten, A. S.; Maryanovsky, D.; Harrington, T. J.; Marin, E.; Vecchio, K. S. Crystal symmetry determination in electron diffraction using machine learning. *Science* **2020**, *367*, 564−568.

(31) Lee, J.-W.; Park, W. B.; Lee, J. H.; Singh, S. P.; Sohn, K.-S. A deep-learning technique for phase identification in multiphase inorganic compounds using synthetic XRD powder patterns. *Nat. Commun.* **2020**, *11*, 86.

(32) Chen, Z.; Andrejevic, N.; Smidt, T.; Ding, Z.; Xu, Q.; Chi, Y.-T.; Nguyen, Q. T.; Alatas, A.; Kong, J.; Li, M. Direct prediction of phonon density of states with Euclidean neural networks. *Adv. Sci.* **2021**, *8*, 2004214.

(33) Chen, Z.; Andrejevic, N.; Drucker, N. C.; Nguyen, T.; Xian, R. P.; Smidt, T.; Wang, Y.; Ernstorfer, R.; Tennant, D. A.; Chan, M.; Li, M. Machine learning on neutron and x-ray scattering and spectroscopies. *Chem. Phys. Rev.* **2021**, *2*, 031301.

(34) Kunka, C.; Shanker, A.; Chen, E. Y.; Kalidindi, S. R.; Dingreville, R. Decoding defect statistics from diffractograms via machine learning. *npj Comput. Mater.* **2021**, *7*, 67.

(35) Kong, S.; Ricci, F.; Guevarra, D.; Neaton, J. B.; Gomes, C. P.; Gregoire, J. M. Density of states prediction for materials discovery via contrastive learning from probabilistic embeddings. *Nat. Commun.* **2022**, *13*, 1−12.

(36) Fung, V.; Ganesh, P.; Sumpter, B. G. Physically informed machine learning prediction of electronic density of states. *Chem. Mater.* **2022**, *34*, 4848−4855.

(37) Kaundinya, P. R.; Choudhary, K.; Kalidindi, S. R. Prediction of the electron density of states for crystalline compounds with Atomistic Line Graph (ALIGNN). *JOM* **2022**, *74*, 1395−1405.

(38) Thompson, A. P.; Aktulga, H. M.; Berger, R.; Bolintineanu, D. S.; Brown, W. M.; Crozier, P. S.; in't Veld, P. J.; Kohlmeyer, A.; Moore, S. G.; Nguyen, T. D.; Shan, R.; Stevens, M. J.; Tranchida, J.; Trott, C.; Plimpton, S. J. LAMMPS - a flexible simulation tool for particle-based materials modeling at the atomic, meso, and continuum scales. *Comput. Phys. Commun.* **2022**, *271*, 108171.

(39) Lenosky, T. J.; Sadigh, B.; Alonso, E.; Bulatov, V. V.; de la Rubia, T. D.; Kim, J.; Voter, A. F.; Kress, J. D. Highly optimized empirical potential model of silicon. *Modell. Simul. Mater. Sci. Eng.* **2000**, *8*, 825.

(40) Chartier, A.; Meis, C.; Crocombette, J.-P.; Weber, W. J.; Corrales, L. R. Molecular dynamic simulation of disorder induced amorphization in Pyrochlore. *Phys. Rev. Lett.* **2005**, *94*, 025505.

(41) Chartier, A.; Onofri, C.; Van Brutzel, L.; Sabathier, C.; Dorosh, O.; Jagielski, J. Early stages of irradiation induced dislocations in urania. *Appl. Phys. Lett.* **2016**, *109*, 181902.

(42) Chen, E. Y.; Deo, C.; Dingreville, R. Reduced-order atomistic cascade method for simulating radiation damage in metals. *J. Phys.: Condens. Matter* **2020**, *32*, 045402.

(43) Thompson, A. P.; Plimpton, S. J.; Mattson, W. General formulation of pressure and stress tensor for arbitrary many-body interaction potentials under periodic boundary conditions. *J. Chem. Phys.* **2009**, *131*, 154107.

(44) Maras, E.; Trushin, O.; Stukowski, A.; Ala-Nissila, T.; Jonsson, H. Global transition path search for dislocation formation in Ge on Si(001). *Comput. Phys. Commun.* **2016**, *205*, 13−21.

(45) Stukowski, A. Visualization and analysis of atomistic simulation data with OVITO- the Open Visualization Tool. *Modell. Simul. Mater. Sci. Eng.* **2010**, *18*, 015012.

(46) Aparicio, E.; Millán, E. N.; Ruestes, C. J.; Bringa, E. M. FoamExplorer: Automated measurement of ligaments and voids for atomistic systems. *Comput. Mater. Sci.* **2020**, *185*, 109942.

(47) Voyles, P. M.; Zotov, N.; Nakhmanson, S. M.; Drabold, D. A.; Gibson, J. M.; Treacy, M. M. J.; Keblinski, P. Structure and physical properties of paracrystalline atomistic models of amorphous silicon. *J. Appl. Phys.* **2001**, *90*, 4437−4451.

(48) Voyles, P. M.; Abelson, J. R. Medium-range order in amorphous silicon measured by fluctuation electron microscopy. *Sol. Energy Mater. Sol. Cells* **2003**, *78*, 85−113.

(49) Borisenko, K. B.; Haberl, B.; Liu, A. C. Y.; Chen, Y.; Li, G.; Williams, J. S.; Bradby, J. E.; Cockayne, D. J. H.; Treacy, M. M. J. Medium-range order in amorphous silicon investigated by constrained structural relaxation of two-body and four-body electron diffraction data. *Acta Mater.* **2012**, *60*, 359−375.

(50) Savitzky, A.; Golay, M. J. E. Smoothing and differentiation of data by simplified least squares procedures. *Anal. Chem.* **1964**, *36*, 1627−1639.

(51) Tersoff, J. Empirical interatomic potential for silicon with improved elastic properties. *Phys. Rev. B* **1988**, *38*, 9902.

(52) Justo, J. F.; Bazant, M. Z.; Kaxiras, E.; Bulatov, V. V.; Yip, S. Interatomic potential for silicon defects and disordered phases. *Phys. Rev. B* **1998**, *58*, 2539.

(53) Kumagai, T.; Izumi, S.; Hara, S.; Sakai, S. Development of bond-order potentials that can reproduce the elastic constants and melting point of silicon for classical molecular dynamics simulation. *Comput. Mater. Sci.* **2007**, *39*, 457−464.

(54) Du, Y. A.; Lenosky, T. J.; Hennig, R. G.; Goedecker, S.; Wilkins, J. W. Energy landscape of silicon tetra-interstitials using an optimized classical potential. *Phys. Status Solidi B* **2011**, *248*, 2050−2055.

(55) Zuo, Y.; Chen, C.; Li, X.; Deng, Z.; Chen, Y.; Behler, J.; Csányi, G.; Shapeev, A. V.; Thompson, A. P.; Wood, M. A.; Ong, S. P. Performance and cost assessment of machine learning interatomic potentials. *J. Phys. Chem. A* **2020**, *124*, 731−745.

(56) Kim, D. S.; Smith, H. L.; Niedziela, J. L.; Li, C. W.; Abernathy, D. L.; Fultz, B. Phonon anharmonicity in silicon from 100 to 1500 K. *Phys. Rev. B* **2015**, *91*, 014507.

(57) Nilsson, G.; Nelin, G. Study of the homology between silicon and germanium by thermal-neutron spectrometry. *Phys. Rev. B* **1972**, *6*, 3777.

(58) Wei, S.; Chou, M. Y. Phonon dispersions of silicon and germanium from first-principles calculations. *Phys. Rev. B* **1994**, *50*, 2221.

(59) Mead, A. Review of the development of multidimensional scaling methods. *J. R. Stat. Soc.: Ser. D (The Statistician)* **1992**, *41*, 27−39.

(60) Tenenbaum, J. B.; De Silva, V.; Langford, J. C. A global geometric framework for nonlinear dimensionality reduction. *Science* **2000**, *290*, 2319−2323.

(61) Lee, J. A.; Verleysen, M.*Nonlinear Dimensionality Reduction*; Springer: New York, NY, 2007; Vol. *1*. DOI: 10.1007/978-0-387-39351-3.

(62) Hinton, G. E.; Salakhutdinov, R. R. Reducing the dimensionality of data with neural networks. *Science* **2006**, *313*, 504−507.

(63) Friedman, J. H. Greedy function approximation: a gradient boosting machine. *Ann. Stat.* **2001**, *29*, 1189−1232.

(64) Pedregosa, F.; et al. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825−2830.

(65) Paske, A.; et al.*Advances in Neural Information Processing Systems 32*; Curran Associates, Inc.: 2019; pp 8024−8035.

(66) Wermelinger, T.; Spolenak, R. Correlating Raman peak shifts with phase transformation and defect densities: a comprehensive TEM and Raman study on silicon. *J. Raman Spectrosc.* **2009**, *40*, 679−686.

(67) Ovsyannikov, S. V.; Korobeinikov, I. V.; Morozova, N. V.; Misiuk, A.; Abrosimov, N. V.; Shchennikov, V. V. "Smart" silicon: Switching between p- and n-conduction under compression. *Appl. Phys. Lett.* **2012**, *101*, 062107.

(68) Weinstein, B. A.; Piermarini, G. J. Raman scattering and phonon dispersion in Si and GaP at very high pressure. *Phys. Rev. B* **1975**, *12*, 1172.

(69) Daisenberger, D.; Deschamps, T.; Champagnon, B.; Mezouar, M.; Quesada Cabrera, R.; Wilson, M.; McMillan, P. F. Polyamorphic amorphous silicon at high pressure: Raman and spatially resolved X-ray scattering and molecular dynamics studies. *J. Phys. Chem. B* **2011**, *115*, 14246−14255.

(70) Fefferman, C.; Mitter, S.; Narayanan, H. Testing the manifold hypothesis. *J. Am. Math. Soc.* **2016**, *29*, 983−1049.

(71) Vincent, P.; Larochelle, H.; Bengio, Y.; Manzagol, P.-A.Extracting and composing robust features with denoising autoencoders. *ICML'08: Proceedings of the 25th International Conference on Machine Learning*; Association for Computing Machinery: 2008; pp 1096−1103..

(72) Vincent, P.; Larochelle, H.; Lajoie, I.; Bengio, Y.; Manzagol, P.-A.; Bottou, L. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *J. Mach. Learn. Res.* **2010**, *11*, 3371−3408.

(73) Mishra, P.; Biancolillo, A.; Roger, J. M.; Marini, F.; Rutledge, D. N. New data preprocessing trends based on ensemble of multiple preprocessing techniques. *TrAC Trends Anal. Chem.* **2020**, *132*, 116045.

(74) Valensise, C. M.; Giuseppi, A.; Vernuccio, F.; De la Cadena, A.; Cerullo, G.; Polli, D. Removing non-resonant background from CARS spectra via deep learning. *APL Photonics* **2020**, *5*, 061305.