

ARTICLE OPEN

Analyzing machine learning models to accelerate generation of fundamental materials insights

Mitsutaro Umehara ^{1,2}, Helge S. Stein¹, Dan Guevarra¹, Paul F. Newhouse¹, David A. Boyd¹ and John M. Gregoire ¹

Machine learning for materials science envisions the acceleration of basic science research through automated identification of key data relationships to augment human interpretation and gain scientific understanding. A primary role of scientists is extraction of fundamental knowledge from data, and we demonstrate that this extraction can be accelerated using neural networks via analysis of the trained data model itself rather than its application as a prediction tool. Convolutional neural networks excel at modeling complex data relationships in multi-dimensional parameter spaces, such as that mapped by a combinatorial materials science experiment. Measuring a performance metric in a given materials space provides direct information about (locally) optimal materials but not the underlying materials science that gives rise to the variation in performance. By building a model that predicts performance (in this case photoelectrochemical power generation of a solar fuels photoanode) from materials parameters (in this case composition and Raman signal), subsequent analysis of gradients in the trained model reveals key data relationships that are not readily identified by human inspection or traditional statistical analyses. Human interpretation of these key relationships produces the desired fundamental understanding, demonstrating a framework in which machine learning accelerates data interpretation by leveraging the expertise of the human scientist. We also demonstrate the use of neural network gradient analysis to automate prediction of the directions in parameter space, such as the addition of specific alloying elements, that may increase performance by moving beyond the confines of existing data.

npj Computational Materials (2019)5:34; <https://doi.org/10.1038/s41524-019-0172-5>

INTRODUCTION

Machine learning has transformed several research fields^{1–6} and is increasingly being integrated into material science research.^{7–17} Motivated by the pervasive need to design functional materials for a variety of technologies, the machine learning models for materials science have primarily focused on establishment of prediction tools.^{7,8,12–14} A complementary effort in data science for materials involves knowledge extraction from large datasets to advance understanding of the present data.^{10,18,19} This strategy can be employed globally, as exemplified by the recent modeling of all known materials phases to generate classifications of the elements akin to the periodic table,²⁰ or locally to reveal the fundamental properties of a given materials system. For materials systems with low-dimensional parameter spaces, composition-property relationships can be directly mapped and represent the understanding of the underlying materials science.^{10,21} Composition-processing parameter spaces are often high dimensional, posing challenges for both experimental exploration of the spaces and the interpretation of the resulting data. Machine learning models such as neural networks excel at modeling complex data relationships but generally do not directly provide fundamental scientific insights, motivating our effort in the present work to analyze the models themselves to identify composition-property and composition-structure-property relationships that lead to fundamental materials insights.

The field of combinatorial materials science comprises an experimental strategy for materials exploration and establishment

of composition-structure-property relationships via systematic exploration of high-dimensional materials parameter spaces.^{18,22,23} High-throughput experimentation can be used to accelerate such materials exploration^{23–28} and enables generation of sufficiently large datasets to utilize modern machine learning algorithms. The dataset in the present work was generated using high-throughput synthesis, structural characterization, and photoelectrochemical performance mapping of BiVO₄-based photoanodes^{29,30} as a function of composition in Bi-V-A and Bi-V-A-B compositions spaces where A and B are chosen from a set of five alloying elements. Previous manual analysis and use of materials theory provided several scientific insights in this materials system, raising the question of whether the data-to-insights process can be accelerated via machine learning.

To explore that concept, we start by modeling of how raw composition and structural data relate to performance using a convolutional neural network (CNN). CNNs have been deployed in material science for tasks such as image recognition^{31–34} and property prediction.^{20,35} Analysis of gradients of the CNN model, which quantify how the predicted property varies with respect to each input dimension, can serve as a measure of the importance of each input dimension and can be further analyzed to interpret the data model,^{36–38} which is one approach to the broader effort of improving interpretability in machine learning.^{39–41} This approach has been used in materials science for classifying regions of micrographs based on their contribution to ionic conductivity,³⁴ and we demonstrate that CNN gradient analysis

¹Joint Center for Artificial Photosynthesis, California Institute of Technology, Pasadena, CA 91125, USA and ²Future Mobility Research Department, Toyota Research Institute of North America, Ann Arbor, MI 48105, USA

Correspondence: John M. Gregoire (gregoire@caltech.edu)

Received: 22 October 2018 Accepted: 6 February 2019

Published online: 08 March 2019

can provide a general framework for data interpretation and even automate the identification of composition-structure-property relationships in high-dimensional materials spaces. We demonstrate the use of CNN-computer gradients to visualize data trends, both locally in composition space and as a global representation of high-dimensional data relationships, which, in addition to aiding human understanding of the data, can provide guidance for design of new high performance materials. We then demonstrate automated identification and communication of composition-property and composition-structure-property relationships, a compact representation of the data relationships that need to be studied to attain a fundamental understanding of the underlying materials science. With this strategy, the machine learning algorithm accelerates science by directing the scientists to data relationships that are emblematic of the fundamental materials science.

RESULTS AND DISCUSSION

Neural network gradient analysis

The multi-dimensional dataset for CNN training was assembled from the high-throughput measurement of the PEC power density (P) and Raman signal for a series of BiVO_4 alloys, using methods described in detail previously.^{29,30} The map of P over the library of samples is shown in Fig. 1a, and select Raman spectra are shown in Fig. 1b. The dataset was compiled from the set of samples comprising $\text{Bi}_{1-x}\text{V}_x\text{O}_{2+\delta}$ compositions with $x = 0.48, 0.5, \text{ and } 0.52$;

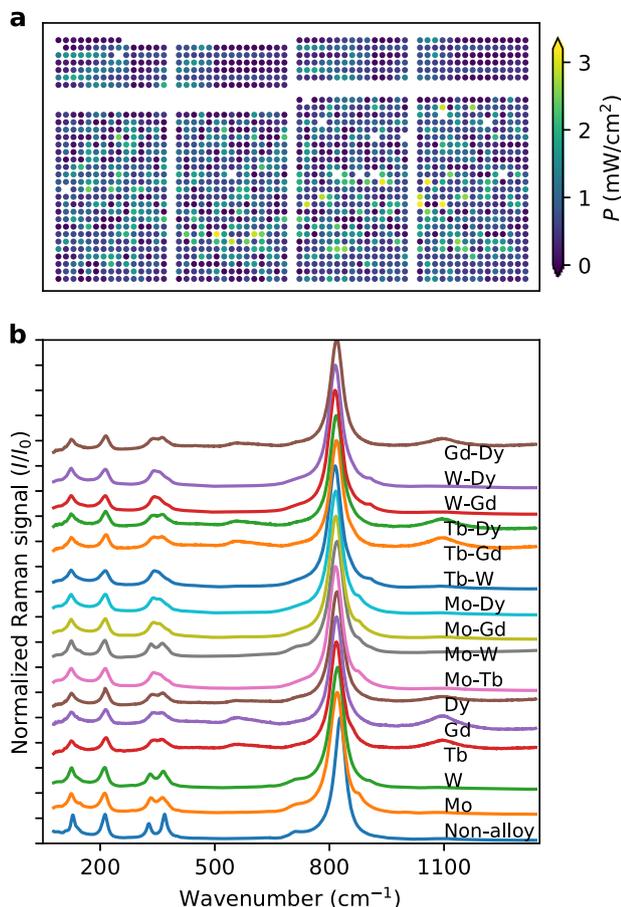


Fig. 1 **a** The map of measured photoelectrochemical power generation for the 1379 photoanode samples. Each sample is $\sim 1 \text{ mm}^2$ and arranged on a 2 mm grid. **b** Representative Raman signals, all normalized by the maximum intensity, for each of the 16 composition spaces in the dataset

for each of these Bi:V stoichiometries, the dataset also included a series of alloys with 5 alloying elements (Mo, W, Dy, Gd, and Tb), as well as each of the 10 pairwise combinations of these alloying elements. The 5 single-alloy spaces include 10 alloy compositions up to approximately 8 at.% and 5 duplicate samples of each of these compositions. The 10 co-alloy spaces include 17 unique co-alloy concentrations with combined alloy concentration between approximately 2 at.% and 8 at.%. For each of the 1379 samples, the feature vector \mathbf{X}_j of j^{th} sample is the concatenation of the Bi-V-Mo-W-Dy-Gd-Tb composition and the normalized Raman spectrum. The dataset \mathbf{X} is a 1022×1379 array where rows 1 to 7 are the composition dimensions (abbreviated \mathbf{X}_{comp}) and the remaining rows are the Raman spectrum dimensions (abbreviated \mathbf{X}_{spec}), which are collectively used to train the CNN model of Fig. 2 to predict the PEC power density P from any coordinate in the M -dimensional parameter space:

$$\vec{\mathbf{x}} = (x_1, x_2, \dots, x_M)$$

$$P_j = f^{(n)}(\vec{\mathbf{x}})|_{\vec{\mathbf{x}}=\mathbf{x}_j}$$

where n is the model index corresponding to eight independent trainings from randomly-generated initializations of the CNN. Analysis is performed on the collection of these independently trained models to help ensure that the interpreted data relationships originate from the data itself and not the initialization of the CNN. While the Introduction motivates the use of a CNN model with regards to its established role in materials science, we additionally note that the gradient analysis functionality in the Keras⁴² makes it a practical choice for the present work. The use of gradient analysis as opposed to a prediction tool makes the results less sensitive to the detailed structure of the CNN as the model needs only to be sufficiently expressive to model the relationships in the data, and we discuss in the SI the considerations that led to the specific structure shown in Fig. 2, as well as the predictive power of the CNN model.

While a given model could be used to predict the performance of other compositions and/or Raman patterns, we instead explore the model itself through analysis of the gradients in performance with respect to each feature vector dimension. These gradients are readily evaluated at all feature vector positions, yielding an array of gradients akin to the partial derivative in the model for P with respect to the i^{th} dimension of the feature vector and evaluated at sample j :

$$G_{i,j}^{(n)} = \left. \frac{\partial f^{(n)}(\vec{\mathbf{x}})}{\partial x_i} \right|_{\vec{\mathbf{x}}=\mathbf{x}_j}$$

For the position in composition-Raman space corresponding to a given sample, this gradient provides the model prediction for how P will be impacted by a change in any composition variable or the intensity at any position in the Raman spectrum.^{36–38}

Local gradient analysis and moving beyond the existing data

To illustrate the gradient analysis of individual samples, we commence with a plot of the sample composition and Raman spectrum along with the respective model gradients for a $\text{Bi}_{0.5}\text{V}_{0.5}\text{Tb}_{0.014}$ sample with $P = 0.008 \text{ mW cm}^{-2}$, a very poor PEC performance (Fig. 3). For this sample, the range of gradient values obtained over the 8 model trainings is shown for each feature vector dimension. For the composition dimensions, the largest gradients are observed for Mo and W where the addition of 1 at.% of either of these elements is predicted to provide a large increase in P , which is commensurate with our extensive manual analysis of the data that identified inclusion of an electron donor (Mo or W) as the most important strategy for optimizing performance.^{29,30} The gradient analysis also indicates a benefit from increasing the Bi:V ratio and increasing the concentration of any of the rare earth elements. With regards to the Raman signal, the region with largest

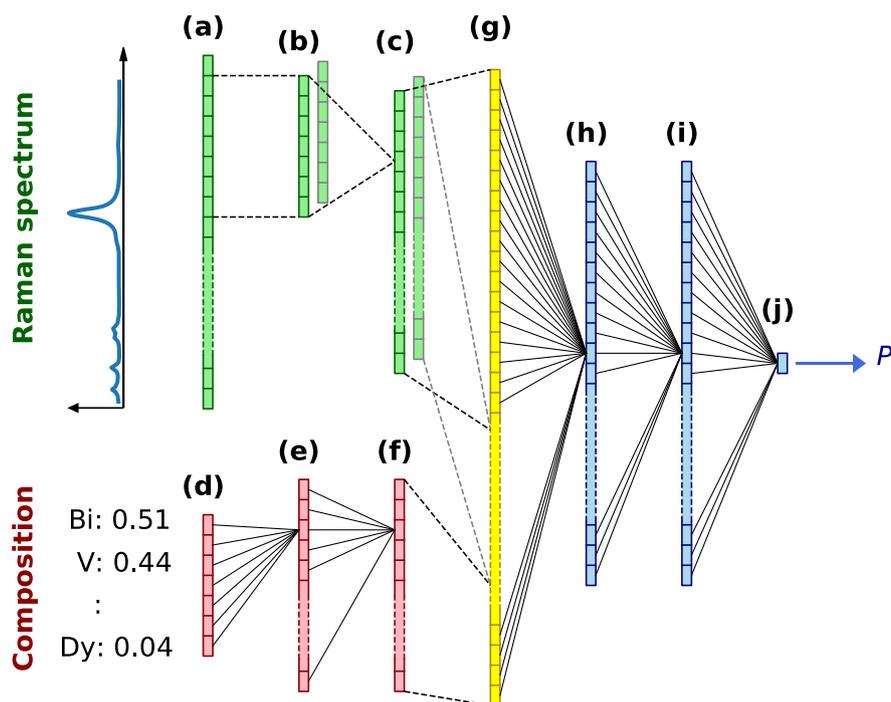


Fig. 2 Schematic of CNN model structure. The model takes the Raman spectrum and the composition as input to predict P . The differently colored layers correspond to red: dense layers acting on composition, green: convolutional 1D layers acting on spectra, yellow: flattening and concatenation layers, blue: dense layers acting on both the composition and spectral data. Each of the 10 layers of the CNN model are labelled **a** to **j**

gradients is the $340\text{--}360\text{ cm}^{-1}$ region where a doublet peak exists in the measured signal and the gradient analysis indicates that improved P can be obtained by increasing the intensity between the 2 peaks and decreasing the intensity on the outer shoulders of the doublet peak, which is akin to lowering the splitting between the 2 peaks.^{29,43,44} This is precisely the discovery featured in a previous publication wherein we identified that a lowered $m\text{-BiVO}_4$ distortion, which is manifested in the Raman signal by a lowered splitting of these peaks, leads to improved PEC performance.^{29,30}

Continuing with analysis of individual samples, we turn to visualization of a sample with a high PEC power density. The gradient analysis of the $\text{Bi}_{0.5}\text{V}_{0.5}\text{Gd}_{0.024}\text{Mo}_{0.057}$ sample with $P = 3.2\text{ mW cm}^{-2}$ is shown in Fig. 4. While this sample is locally optimal with respect to its composition neighbors in this library, the nonzero gradients for this sample suggest that the global maximum lies beyond the extent of the present dataset, which is important from a materials design perspective as it provides guidance in the form of the direction in parameter space to modify the best samples to obtain an even higher performance material. With respect to composition, this sample has the highest Bi:V out of the three values in the dataset, and the model indicates further increase of this ratio would be beneficial. Concerning the alloying elements, the gradients indicate that higher rare earth concentrations would be beneficial and higher W concentration would be deleterious. The directions in parameter space of other samples are illustrated in Figure S1. Regarding the gradients in the Raman spectrum, in the $340\text{--}360\text{ cm}^{-1}$ region the variation in gradient with wavenumber is similar to that of Fig. 3, but with smaller magnitude due to the nearly-complete merging of the doublet peak in this sample. The Raman feature with negative gradient on the shoulder of the main peak, near 715 cm^{-1} , is commensurate with increasing P by lowering the monoclinic distortion, as this peak is the antisymmetric stretching mode of V-O bond that decreases in intensity as the monoclinic distortion vanishes to yield the tetragonal scheelite polymorph.^{43,44} The large positive gradients in the $400\text{--}600\text{ cm}^{-1}$ range don't

correspond to any detected features in the Raman patterns and are thus not immediately interpretable.

Gradient ensemble visualization

While Figs. 3 and 4 demonstrate gradient analysis of single samples, the ensemble of gradients from all sample provide additional insights into the most pertinent data relationships for understanding the underlying materials science. The gradients for each input dimension and sample are first averaged over the eight independently trained models, enabling analysis of the distribution of gradients for each dimension of \mathbf{X} as shown in Fig. 5a, b. There is considerable variation in the gradients for each composition dimension, and Mo and W gradients exhibit bimodal distributions, indicating that analysis of the average variation of P with any composition dimension will not sufficiently characterize the data relationships. For comparison, three different scalar metrics for the relationship of P to each dimension of \mathbf{X}_{comp} are provided in Fig. 5c: the feature importance for a random forest regression model (FI) trained with the same input data as the CNN, the maximal information coefficient (MIC), and the Pearson correlation coefficient. While all three of these metrics provide alternate perspectives on the data relationships, only the CNN gradient analysis is commensurate with the established conclusions regarding the elemental concentrations, which include the following composition design rules (in decreasing order of importance for maximizing P) and corresponding observations from Fig. 5a: (i) W or Mo should be included to increase electrical conductivity; the composition dimensions for the elements have the highest average gradient as well as gradient distributions that extend to the highest values. (ii) Once electronic conductivity is no longer limiting performance, adding a rare earth element improves P by increasing hole transport via crystal structure modulation; the three REs have near zero gradient for many samples but their distributions each extend to high values. (iii) Depending on the alloying elements, the highest P is observed

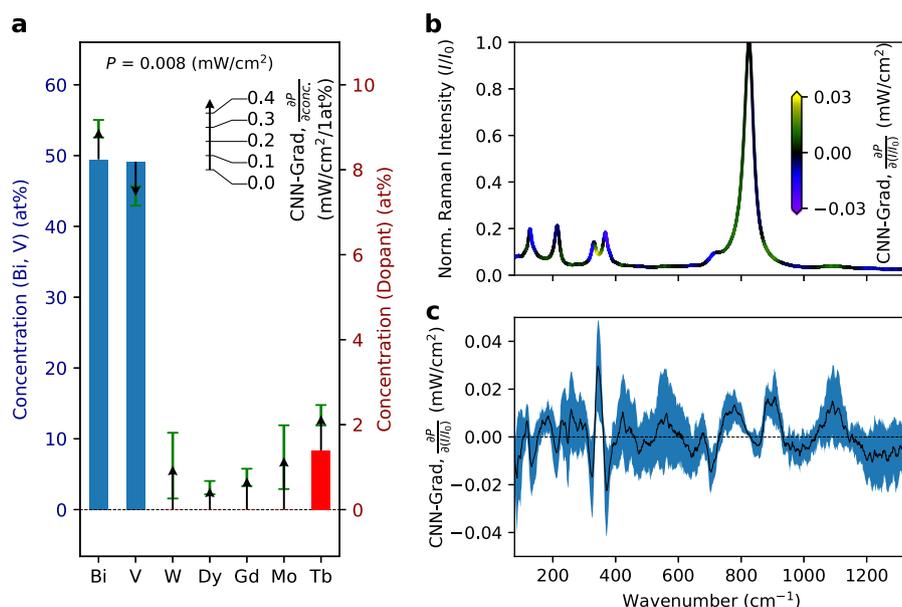


Fig. 3 **a** Composition of a poor-performing sample (bar plot) and corresponding gradients for the composition dimensions (arrows), where the legend provides the relationship between arrow length and gradient magnitude, and up and down arrows indicate positive and negative gradients, respectively. The green error bar for each arrow indicates the standard deviation of the respective gradient over the 8 independent models. **b** Since the Raman pattern has too many dimensions to create the same arrow representation of gradients, the plot of the Raman pattern is colored by the average gradient. **c** The average gradient is also plotted (black) with the wavenumber-specific standard deviation over the 8 models (blue)

with 1:1 Bi:V or the Bi-rich variant; the gradients for both Bi and V are mostly near 0 with a small distribution at positive values for Bi and negative values for V. While the CNN gradient analysis is commensurate with the established scientific interpretations, it is important to note that the details of these scientific interpretations cannot be derived from the gradient analysis. Instead, this summary of gradients provides a compact visualization of the data relationships for scientists to inspect and interpret.

Figure 5b, d contains a similar set of analyses for the Raman spectra, with the large dimensionality of \mathbf{X}_{spec} hindering visualization of the full gradient distributions, prompting our visualization of the variation in gradients by plotting the green shaded region corresponding to the ± 1 standard deviation in Fig. 5b. The main peak-like patterns in this sample-averaged gradient signal draws correspond to the doublet peak in the 340–360 cm⁻¹ region where, as discussed above, the positive gradient between the pair of peaks and the negative gradient on the outer shoulders of each peak corresponds to the improvement in P with merging of the doublet peak. This gradient analysis would have greatly accelerated the identification of the corresponding structural modulation that provides the PEC improvement, which was only identified after considerable manual inspection including the development of custom analysis algorithms for identifying the data relationships. This type of guidance is not forthcoming from the three scalar-based assessments (Rand.For, FI, MIC, Pearson) of the $\mathbf{X}_{\text{spec}}-P$ relationships (Fig. 5d), which each direct primary attention to the most intense Raman peak.

Gradient correlation analysis and automated detection of composition-structure-property relationships

While Fig. 5a, b demonstrate the utility of the gradient analysis for generating compact, human-readable summaries of high-dimensional data, there is no clear way to automate interpretation of these visualizations. Given the importance of composition-structure-property relationships in elucidating the fundamental origins of an observed variation in a property (in this case P), we focus the automation of data interpretation via gradient analysis

on reporting composition-structure-property relationships. Correlation analysis of gradients from different features (dimensions of \mathbf{X}) quantifies the extent to which these features similarly impact P . Performing this correlation analysis is facilitated by the ability to evaluate the gradient with respect to each feature dimension at any coordinate in the feature space. Since the grid of compositions in the dataset is based on a 6-dimensional composition space that is only explored in up to three dimensions at a time, this set of samples is not conducive to direct calculation of local partial derivatives for each input dimension. We performed the correlation analysis by calculating the correlation matrix (similar to covariance matrix with every value being a Pearson correlation coefficients of the respective pair of features) for each of the eight independently trained models and then averaging over the eight models. For this analysis, the V concentration dimension was ignored since the design of the composition library involves three different Bi:V values and thus V concentration is nearly linearly related to that of Bi, obscuring separate analysis of the covariance of these dimensions with any other dimension of \mathbf{X} . Pairwise plots of the gradients $\mathbf{G}_{i,j}^{(n)}$ and the correlation coefficients (averaged over the eight models) are shown in Fig. 6 for the set of six elements. Analysis of these correlation coefficients reveals sets of elements that similarly impact P . That is, from the collection of samples in the high-dimensional composition space, the model-predicted change in P with increasing concentration is correlated for elements whose functional role is similar. This correlation doesn't necessarily relate to similarity of the elements, only their similar alteration of the property of interest.

To automate identification and communication of these sets of elements with similar composition-property relationships, we choose a threshold correlation value (0.9 in this case) and find all sets of elements for which every pairwise correlation coefficient exceeds the threshold. Using the data in Fig. 6, the resulting sets are {Dy,Gd,Tb} and {W,Mo}. To provide some intuitive explanation of how the CNN encoded these commonalities, Figure S2 shows the activations of the seven dimensions of \mathbf{X}_{comp} in the first neural network layer (Fig. 2e), revealing that through training of the model, the best reconstructions of the P data were found by

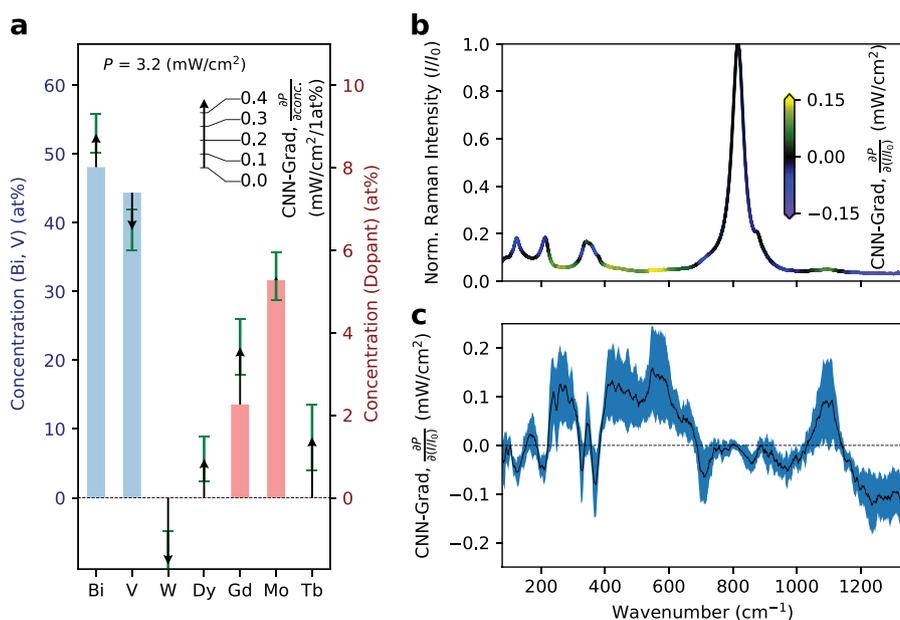


Fig. 4 **a** Composition of the highest performance sample (bar plot) and corresponding gradients for the composition dimensions (arrows), **b** Raman spectrum of the highest performance sample with heat map of gradient, and **c** averaged gradient (black solid line) and its standard deviation (blue filled region), in similar format to Fig. 3

activating the TMs similarly and the REs similarly in this first layer, resulting in similar functional modeling for the TMs and for the REs, which produces the observed correlations in the gradients. This pattern of activations is the model's "learning" of the similar composition-property relationships.

For each of these sets, we next automatically identify features in the Raman spectra that can elucidate composition-structure-property relationships. For the present work, we do not explore all composition-structure relationships, only those related to improving P . If the improvement in P upon increasing an elemental concentration is related to a structural feature in the Raman spectra, then the dimensions of \mathbf{X} corresponding to the structural feature will have gradients correlated with the concentration gradient, and this correlation coefficient could be positive or negative depending on whether the given Raman mode is increasing or decreasing in intensity or shifting to a different wavenumber. To automate detection of such relationships, for each of the element sets ($\{\text{Dy,Gd,Tb}\}$ and $\{\text{W,Mo}\}$ in this case), we identify each dimension of \mathbf{X}_{spec} whose gradient correlation coefficient with respect to each element in the set exceeds a threshold value (absolute value above 0.3 in this case). Subsequent identification of the contiguous ranges of wavenumbers that meet this criterion produces the list of Raman feature locations that represent the composition-structure-property relationships.

An automated report summary of these findings is illustrated in Table 1 where a list of 16 observations, each identifying a composition-property relationship or a Raman spectral region related to such a relationship, guide human investigation of the materials science. Where the human-derived materials science explanation for a given data observation has been identified or hypothesized, the materials science explanation is also summarized in the table. The observations commence with a report of the three elements whose gradients are most positive among the samples exhibiting highest P (in this case, above the 95th percentile of P), and while the table provides classification of data relationships, we note that quantification of the relationships is a powerful aspect of the gradient analysis. The next two observations are the elemental sets identified from the composition-property correlation analysis and are commensurate

with observations from the anecdotal examples in Figs. 3 and 4, that Mo and W similarly increase conductivity and Dy, Gd, and Tb similarly decrease the monoclinic distortion, with both phenomena leading to improvements in P . Of the 13 observations related to composition-structure-property relationships, 6 of them (#4–9) are explained by changes to the bending and stretching modes of $m\text{-BiVO}_4$ due to decreasing monoclinic distortion, which occurs with RE alloying and to an even greater extent in the RE-TM co-alloying spaces.

From the perspective of knowledge discovery, it is insightful to further inspect how the observations of Table 1 relate to those in the literature. We note that this system was chosen to validate the algorithms of the present work due to its years of research precedence and publication history that are imperative for establishing a set of ground truth observations against which the automatically-generated observations can be compared. In this regard, observations 1–6 and 8–9 are commensurate with the results of ref. ²⁹ with an important caveat that custom, not-broadly-applicable algorithms were developed to identify these relationships in that work. The automated gradient analysis also extends the observations of that work in two critical aspects, by quantifying the relationships and by identifying that Tb and Dy alloying elements follow the same relationships as Gd. Line 7 and 10–11 are observations related to the same underlying phenomenon but were not identified by the previous analysis and are thus discovered and quantified by the gradient analysis. Line 12 involves a spectral feature identifiable by previous Raman literature⁴⁴ for Mo but with no literature precedent for W, and identification and quantification of its relationship to photoelectrochemical performance is new to the present work. Lines 13–16 are also new to the present work and involve spectral features which have yet to be identified. Due to the size and dimensionality of the dataset, observation such as the similarity of the REs may be identifiable by manual inspection of the data, but quantification of the similar effect of RE alloying at all points in the high-dimensional space is uniquely enabled by the gradient analysis, and the automated identification of the composition-structure-property relationships are not forthcoming from manual human analysis.

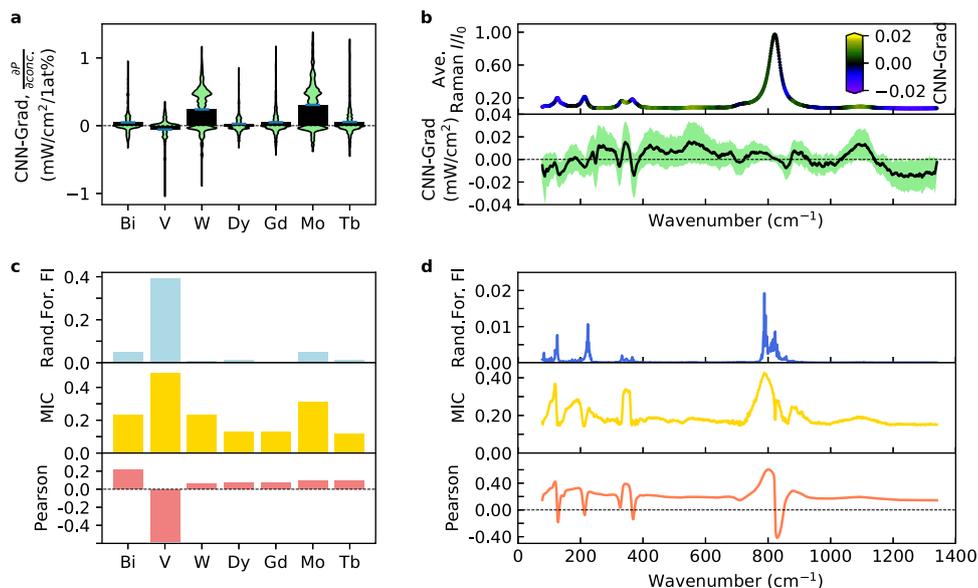


Fig. 5 **a** Gradients of each element with violin plots showing the distribution of values over all samples and bar plot showing the average value of each of these distributions. **b** Averaged Raman signal colored by the sample-averaged gradients (top panel), and the sample-averaged gradients are also plotted in the bottom panel (black line) with the respective ± 1 standard deviation (green area) representing variation over the sampled parameter space. **c, d** The relationship between P and each composition (**c**) and spectrum (**d**) dimension of the source data, as quantified by Random Forest feature importance (Rand.For. FI), Maximal information coefficient (MIC), and Pearson correlation coefficient (Pearson)

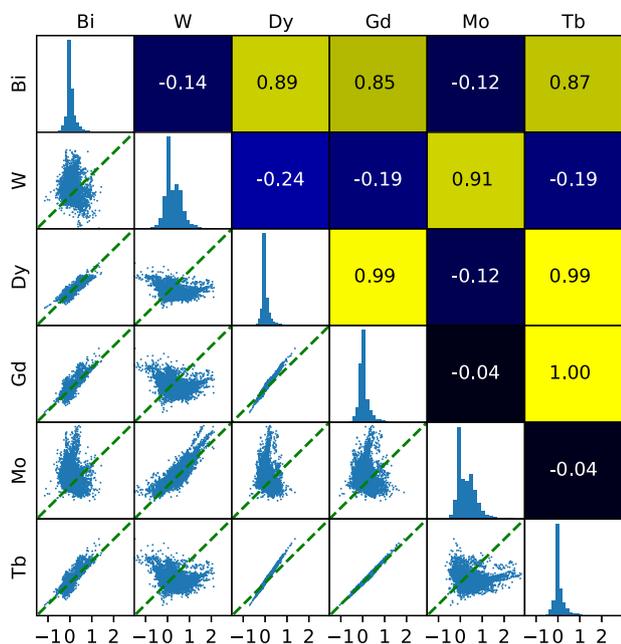


Fig. 6 Pairwise correlation analysis of gradients for 6 composition dimensions of the input data. V is excluded due to its inherent inverse correlation with Bi, and each data point in the bottom-left correlation plots represents the pair of gradients for single sample over the 8 models. Each plot on the diagonal is the histogram of gradients for the respective element, and the numbers in each box in the upper-right portion of the figure show the Pearson correlation coefficient averaged over 8 models for the respective correlation plot (correlation coefficient of gradients over the sample set)

While this anecdotal example of automated identification of key data relationships demonstrates that this analysis would have greatly accelerated the understanding of the fundamental materials science in this class of photoanodes, it is important to

note limitations on the generality of the present techniques and of machine learning-based data interpretation. For the automated report generation (Table 1), we assert that there is considerable generality to the concept of analyzing CNN gradients to identify the data relationships that are critical for understanding the fundamental science, as described with the right-most column in Table 1, but the criteria for enumerating data observations (including threshold values and criteria noted above) were user-chosen in the present case and will likely need to be altered for analyzing other datasets. Other than excluding V from the gradient correlation analysis, we did not discuss methods for mitigating the influence of correlations in the set of materials (used for CNN training) in the gradient analysis. This issue is perhaps not critical for the present dataset because each alloy and co-alloy composition space was sampled with the same grid of compositions, but generalization of these techniques will require further inspection of how correlations in the source data impact CNN gradients.^{45,46} Finally, the CNN model has no concept of TM vs. RE classification of elements and did not “learn” anything about the chemistry of these elements, only that when it comes to alloying-based improvements to P , the TM and RE families of elements each have a characteristic data relationship whose identification enables the scientist to learn something fundamental about the underlying materials science. Consequently, this machine learning-based identification of key data relationships augments but does not replace human interpretation of scientific discoveries.

To leverage the ability of CNNs to model complex data relationships in high-dimensional spaces, we trained a CNN model to predict photoelectrochemical performance of BiVO_4 -based photoanodes from the composition and Raman spectrum of 1379 photoanode samples containing various 3 and 4-cation combinations from a set of 7 elements. Gradients calculated from the CNN model, akin to partial derivatives of the performance with respect to each input variable, enabled effective visualization of data trends at specific locations in the materials parameter space as well as collectively for the entire dataset. Automated analysis of

Table 1. An example report of observations related to further materials optimization (1), composition-property relationships (2–3), and composition-structure-property relationships (4–16)

#	Data observations from analysis of CNN models	Human-derived materials science explanation
	Adding <A> in the high-P samples increase P by an average of mW/cm ² /at.%. <A> 	
1	Mo,Tb,Gd 0.25,0.10,0.10 <A> have similar concentration-P relationships with average pairwise correlation of . <A> 	Synergy of both Mo-Tb and Mo-Gd co-alloying lead to beneficial structure modulations
2	{Dy, Gd, Tb} 0.99	All 3 REs have the same role of decreasing mono. dist.
3	{W, Mo} 0.91 For increasing conc. of <A>, Raman intensity in <C> cm ⁻¹ has similar influence on P. <A> <C>	Both TMs have same role of increasing conductivity and decreasing mono. dist. when coalloyed with REs
4	Dy,Gd,Tb Decreasing 331.4–332.7	Intensity at left and right edges of the VO ₄ ⁻³ bending mode doublet peak decrease and centroid increases due to peak merging with decreasing monoclinic distortion
5	Dy,Gd,Tb Decreasing 359.1–361.7	
6	W,Mo Increasing 339.3–357.7	
7	Dy,Gd,Tb Decreasing 698.9–728.8	Intensity of antisymmetric V-O stretch (715 cm ⁻¹) decreases with decreasing monoclinic distortion
8	W,Mo Increasing 791.9–796.9	Intensity at left/right edge increases/decreases as the symmetric V–O stretching peak shifts to lower wavenumber due to increased symmetry of VO ₄ ⁻³ tetrahedra
9	W,Mo Decreasing 834.9–838.5	
10	W,Mo Decreasing 127.3–143.6	Possibly related to VO ₄ ⁻³ external translation mode
11	W,Mo Decreasing 212.6–224.7	Possibly related to VO ₄ ⁻³ external rotation mode
12	W,Mo Increasing 880.0–884.8	Possibly increasing Mo-O stretching mode intensity
13	W,Mo Increasing 302.2–303.5	Unknown
14	W,Mo Increasing 667.6–668.9	Unknown
15	W,Mo Increasing 675.1–683.9	Unknown
16	Dy,Gd,Tb Increasing 919.9–921.1	Unknown

For each type of observation, a human-interpretable sentence is provided with fill-in values noted for each enumerated observation. The right-most column is the human-generated explanation of the underlying materials science that gave rise to each data relationship, as discussed in our previous work^{29,30} and using literature identification of Raman modes^{44,47–50}

the gradients provides guidance for research, including how to move beyond the confines of the present dataset to further improve performance. To accelerate generation of fundamental scientific understanding, correlations in the gradients are analyzed to identify the key data relationships whose interpretation by a human expert can provide comprehensive understanding of the composition-property and composition-structure-property relationships in the materials system. This approach to interpreting machine learning models accelerates scientific understanding and illustrates avenues for continued automation of scientific discovery.

METHODS

Experimental

The details of the materials synthesis, photoelectrochemical measurements, and Raman measurements are described elsewhere.^{29,30} Briefly, two duplicate thin-film materials libraries were prepared by ink-jet printing using Bi, V, Mo, W, Tb, Gd, and Dy metal-nitrate inks on SnO₂:F (FTO) coated glass. Each library was calcined at 565 °C in O₂ gas for 30 min, after which one was used for Raman measurements and the other for photoelectrochemical measurements. The photoelectrochemical measurements included, for each material sample, a cyclic voltammogram (CV) using a Pt counter electrode and Ag/AgCl reference

electrode in a 3-electrode cell setup. Aqueous electrolyte with potassium phosphate buffer (50 mM each of monobasic and dibasic phosphate) was used with 0.25 M sodium sulfate as a supporting electrolyte (pH 6.7). CVs were acquired for each sample on the ML at chopped illumination using a 455 nm light emitting diode (LED). Maximum photoelectrochemical power generation (*P*) is calculated as a figure-of-merit for photoanode performance from CV for each sample. Raman spectroscopy spectrum of each sample was acquired by averaging Raman spectra mapping of whole library with a resolution of 75 μm × 75 μm using Renishaw inVia Reflex. Composition of each sample was determined by the printed amount of ink-jet printer.

Gradient analysis for visualization

To analyze the CNN model, we used a visualization method similar to the previously reported method,^{36–38,51} and repeated the analysis eight times using randomly initialized models. The CNN model (*f*) is a function of input vector of spectrum (\vec{x}_{spec}) and composition (\vec{x}_{comp}) with output of power generation performance $Y_{\text{predicted}i}$:

$$\begin{aligned}
 Y_{\text{predicted}}^{(n)} &= f^{(n)}(\vec{x}) = f^{(n)}(\vec{x}_{\text{comp}}, \vec{x}_{\text{spec}}) \\
 \vec{x} &= (\vec{x}_{\text{comp}}, \vec{x}_{\text{spec}}) = (x_1, x_2, x_3, \dots, x_M) \\
 \vec{x}_{\text{comp}} &= (x_1, x_2, x_3, \dots, x_7) \\
 \vec{x}_{\text{spec}} &= (x_8, x_9, x_{10}, \dots, x_{1022})
 \end{aligned}$$

where n indicates the n -th run of the analysis ($n = 1..8$), and M indicates the input vector dimension ($=7 + 1015$). The input dataset X is a matrix of each j -th input vector;

$$X = (X_1, \dots, X_N) = \begin{pmatrix} X_{1,1} & \dots & X_{1,N} \\ \vdots & \ddots & \vdots \\ X_{M,1} & \dots & X_{M,N} \end{pmatrix} = \begin{pmatrix} X_{\text{comp}} \\ X_{\text{spec}} \end{pmatrix}$$

$$X_{\text{comp}} = \begin{pmatrix} X_{1,1} & \dots & X_{1,N} \\ \vdots & \ddots & \vdots \\ X_{7,1} & \dots & X_{7,N} \end{pmatrix}$$

$$X_{\text{spec}} = \begin{pmatrix} X_{8,1} & \dots & X_{8,N} \\ \vdots & \ddots & \vdots \\ X_{1022,1} & \dots & X_{1022,N} \end{pmatrix}$$

where X_j indicate the inputs vector of j -th sample, and N indicates the total number of samples ($=1379$). We defined gradient matrix G as a partial derivative in output with respect to the input value in input vectors;

$$G_{ij}^{(n)} = \left. \frac{\partial f^{(n)}}{\partial x_i} \right|_{\vec{x}=X_j}$$

where j indicates j -th sample and i indicates i -th value in input vector. Also, we calculated average and standard deviation of gradient form eight runs;

$$G_{ij}^{\text{ave}} = \frac{1}{8} \sum_{n=1}^8 G_{ij}^{(n)}$$

$$G_{ij}^{\text{std}} = \sqrt{\frac{1}{8} \sum_{n=1}^8 (G_{ij}^{(n)} - G_{ij}^{\text{ave}})^2}$$

where G^{ave} is averaged gradient of 8 models, and G^{std} is standard deviation in 8 models. These gradients indicate how much impact the input value has on the output; if the gradient is positive, then the input value has positive influence, and if the gradient is negative, then the input value has negative influence on the output.

The Pearson correlation coefficient matrix C is defined as follows;

$$C = \frac{1}{8} \sum_{n=1}^8 C^{(n)}$$

$$C_{jk}^{(n)} = \text{Pearson}(G_i^{(n)}, G_k^{(n)}) = \frac{\text{cov}(G_i^{(n)}, G_k^{(n)})}{\sigma_i^{(n)} \sigma_k^{(n)}}$$

$$G_i = (G_{i1}, G_{i2}, \dots, G_{iN})$$

where G_i is gradient vector with respect to i -th parameter in input vector, $\text{cov}(G_i, G_k)$ is covariance of G_i and G_k , σ_{G_i} is standard deviation of G_i . G has 1022 (input vector dimension = 1015 + 7) \times 1379 (sample number) dimension, and C has 1022 \times 1022 dimension.

CNN model

CNN model was constructed in python using Keras package with Tensorflow backend, a schematic model description is shown in Fig. 2. There are two input vectors and one output value in this model; a spectrum input vector \vec{x}_{spec} , a composition input vector \vec{x}_{comp} , and output value Y . The spectrum input vector is 1015 dimensions-length with a range from 300 to 1400 cm^{-1} wavenumbers of each sample. Each spectrum is normalized by the main of the peak value at around 825 cm^{-1} , which is attributed to V-O symmetric stretching vibration mode of BiVO_4 . The composition input vector is 7-length vector of atomic fraction of elements (Bi, V, Mo, W, Tb, Gd, and Dy), which has 0–1 value so that the sum of values in a vector equals unity ($\text{Bi} + \text{V} + \text{Mo} + \text{W} + \text{Tb} + \text{Gd} + \text{Dy} = 1$). The output value Y of this model is standardized maximum photoelectrochemical power generation P ; $Y_i = (P_i - \mu)/\sigma$, $i = 0, 1, \dots, N$, where μ is mean value of P , σ is standard deviation of P , i indicates the i -th sample, and N is the total number of samples. These input vectors are fed into the first layers as shown in Fig. 2a, d. The first layer for the spectrum input vector is an input layer for following CNN layer (See Fig. 2a). This first layer has a dropout with dropout rate of 0.25 (not shown in Fig. 2). The second layer, Fig. 2c, is a CNN layer and the kernels of this layer is shown in Fig. 2b. The kernel size is 7 and the number of filters is 2. Exponential Linear Unit (ELU) is used as an activation function of this layer. This layer does not have any pooling layer. It is worth to mention that we found the prediction

performance of the model without pooling layer is better than that with pooling layer, which is attributed to the peak position sensitiveness of the model without pooling layer. This layer also has a dropout with dropout rate of 0.25. The output of this layer is flattened and fed into the next concatenated layer, Fig. 2g. The first layer for composition input vector is an input layer for following neural network layer (See Fig. 2d). The next layer, Fig. 2e, is a neural network layer with 16 units, which activation function is ELU. This layer does not have dropout. The next layer, Fig. 2f has 16 units, and activation function is ELU. In next layer, Fig. 2g, the output of the CNN layer (Fig. 2c) is flattened and concatenated with the output of the composition layer (Fig. 2f), and this 2034-length output ($1009 \times 2 + 16$) is fed into the following neural network layer, Fig. 2h, with 32 units and ELU activation. The output of this layer is then fed into the next neural network layer, Fig. 2i, with 32 units and ELU activation, followed by output layer, Fig. 2j, with one unit and linear activation, which predict the output value Y .

Code availability

The authors declare that the code used to perform the analysis is provided at https://github.com/johnmgreigoire/CNN_Gradient_Analysis.

DATA AVAILABILITY

The authors declare that the data supporting the findings of this study are available within the paper and its supplementary information files.

ACKNOWLEDGEMENTS

This study is based upon work performed by the Joint Center for Artificial Photosynthesis, a DOE Energy Innovation Hub, supported through the Office of Science of the U.S. Department of Energy (Award No. DE-SC0004993). Development of the algorithm for automating the model interpretation (J.M.G. and H.S.S.) was funded by Toyota Research Institute through the Accelerated Materials Design and Discovery program.

AUTHOR CONTRIBUTIONS

M.U. performed model training and gradient analysis. H.S.S. and D.G. assisted with design of the model and comparisons to other techniques. P.F.N., D.G. and D.A.B. performed all experiments. M.U., H.S.S., D.G. and J.M.G. interpreted model outputs and created data visualization schemes. J.M.G. created algorithm for automated relationship identification with assistance from M.U. and H.S.S. M.U., H.S.S. and J.M.G. were the primary authors of the manuscript.

ADDITIONAL INFORMATION

Supplementary information accompanies the paper on the *npj Computational Materials* website (<https://doi.org/10.1038/s41524-019-0172-5>).

Competing interests: The authors declare no competing interests.

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

REFERENCES

- Hinton, G. et al. Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups. *IEEE Signal Process. Mag.* **29**, 82–97 (2012).
- Krizhevsky, A., Sutskever, I. & Hinton, G. E. ImageNet Classification with Deep Convolutional Neural Networks. In *Proc. Advances In Neural Information Processing Systems* 1097–1105 (Curran Associates/Red Hook, NY, USA, 2012).
- Simonyan, K. & Zisserman, A. *Very deep convolutional networks for large-scale image recognition*. <https://arxiv.org/abs/1312.6034> (2014). Accessed 10 Apr 2015.
- Jurafsky, D. & Martin, J. H. *Speech and Language Processing: An Introduction to Natural Language Processing*. In *Computational Linguistics and Speech Recognition* (Pearson Education, London, 2000).
- Silver, D. et al. Mastering the game of Go with deep neural networks and tree search. *Nature* **529**, 484–489 (2016).
- Levinson, J. et al. Towards fully autonomous driving: Systems and algorithms. In *Proc. IEEE Intelligent Vehicles Symposium* (Curran Associates/Red Hook, NY, USA, 2011).

7. Hautier, G., Fischer, C., Ehrlicher, V., Jain, A. & Ceder, G. Data mined ionic substitutions for the discovery of new compounds. *Inorg. Chem.* **50**, 656–663 (2011).
8. Xue, D. et al. Accelerated search for materials with targeted properties by adaptive design. *Nat. Commun.* **7**, 11241 (2016).
9. Welborn, M., Cheng, L. & Miller, T. F. Transferability in machine learning for electronic structure via the molecular orbital basis. *J. Chem. Theory Comput.* **14**, 4772–4779 (2018).
10. Lookman, T., Alexander, F. J. & Rajan, K. *Information science for materials discovery and design. Springer Series in Materials Science.* (Springer International Publishing, Switzerland, 2016).
11. Bartók, A. P., Kondor, R. & Csányi, G. On representing chemical environments. *Phys. Rev. B* **87**, 1–16 (2013).
12. Ward, L., Agrawal, A., Choudhary, A. & Wolverton, C. A general-purpose machine learning framework for predicting properties of inorganic materials. *NPJ Comput. Mater.* **2**, 16028 (2016).
13. Hatrick-Simpers, J. R., Choudhary, K. & Corgnale, C. A simple constrained machine learning model for predicting high-pressure-hydrogen-compressor materials. *Mol. Syst. Des. Eng.* **3**, 509–517 (2018).
14. Stanev, V. et al. Machine learning modeling of superconducting critical temperature. *NPJ Comput. Mater.* **4**, 29 (2018).
15. Nikolaeov, P. et al. Autonomy in materials research: a case study in carbon nanotube growth. *npj Comput. Mater.* **2**, 16031 (2016).
16. Carleo, G. & Troyer, M. Solving the quantum many-body problem with artificial neural networks. *Science* **355**, 602–606 (2017).
17. Alberi, K. et al. The 2019 materials by design roadmap. *J. Phys. D: Appl. Phys.* **52**, 013001 (2018).
18. Hatrick-Simpers, J. R., Gregoire, J. M. & Kusne, A. G. Perspective: composition–structure–property mapping in high-throughput experiments: turning data into knowledge. *APL Mater.* **4**, 53211 (2016).
19. Rajan, K. Combinatorial materials sciences: experimental strategies for accelerated knowledge discovery. *Annu. Rev. Mater. Res.* **38**, 299–322 (2008).
20. Zhou, Q. et al. Learning atoms for materials discovery. *Proc. Natl Acad. Sci. USA* **115**, E6411–E6417 (2018).
21. Dorenbos, P. Systematic behaviour in trivalent lanthanide charge transfer energies. *J. Phys. Condens. Matter* **15**, 8417–8434 (2003).
22. Green, M. L., Takeuchi, I. & Hatrick-simpers, J. R. Applications of high throughput (combinatorial) methodologies to electronic, magnetic, optical, and energy-related materials. *J. Appl. Phys.* **113**, 231101 (2013).
23. Kusne, A. G., Keller, D., Anderson, A., Zaban, A. & Takeuchi, I. High-throughput determination of structural phase diagram and constituent phases using GRENDEL. *Nanotechnology* **26**, 444002 (2015).
24. Van Dover, R. B., Schneemeyer, L. F. & Fleming, R. M. Discovery of a useful thin-film dielectric using a composition-spread approach. *Nature* **392**, 162–164 (1998).
25. Wang, J. et al. Identification of a blue photoluminescent composite material from a combinatorial library. *Science* **279**, 1712 (1998).
26. Reddington, E., Sapienza, A., Gurau, B., Viswanathan, R. & Sarangapani, S. Combinatorial electrochemistry: a highly parallel, optical screening method for discovery of better electrocatalysts. *Science* **280**, 1735–1737 (1998).
27. Yan, Q. et al. Solar fuels photoanode materials discovery by integrating high-throughput theory and experiment. *Proc. Natl Acad. Sci. USA* **114**, 3040–3043 (2017).
28. Suram, S. K. et al. Automated phase mapping with AgileFD and its application to light absorber discovery in the V-Mn-Nb oxide system. *ACS Comb. Sci.* **19**, 37–46 (2017).
29. Newhouse, P. F. et al. Combinatorial alloying improves bismuth vanadate photoanodes via reduced monoclinic distortion. *Energy Environ. Sci.* **11**, 2444–2457 (2018).
30. Newhouse, P. F. et al. Multi-modal optimization of bismuth vanadate photoanodes via combinatorial alloying and hydrogen processing. *Chem. Commun.* **55**, 489–492 (2018).
31. Ling, J. et al. Building data-driven models with microstructural images: generalization and interpretability. *Mater. Discov.* **10**, 19–28 (2017).
32. Ziatdinov, M., Maksov, A. & Kalinin, S. V. Learning surface molecular structures via machine vision. *npj Comput. Mater.* **3**, 31 (2017).
33. Ziatdinov, M. et al. Deep learning of atomically resolved scanning transmission electron microscopy images: chemical identification and tracking local transformations. *ACS Nano* **11**, 12742–12752 (2017).
34. Kondo, R., Yamakawa, S., Masuoka, Y., Tajima, S. & Asahi, R. Microstructure recognition using convolutional neural networks for prediction of ionic conductivity in ceramics. *Acta Mater.* **141**, 29–38 (2017).
35. Kajita, S., Ohba, N., Jinnouchi, R. & Asahi, R. A universal 3D voxel descriptor for solid-state material informatics with deep convolutional neural networks. *Sci. Rep.* **7**, 1–9 (2017).
36. Simonyan, K., Vedaldi, A. & Zisserman, A. *Deep inside convolutional networks: visualising image classification models and saliency maps.* <http://arxiv.org/abs/1312.6034> (2013). Accessed 19 Apr 2014.
37. Zeiler, M. D. & Fergus, R. Visualizing and understanding convolutional networks. In *Proc. European conference on computer vision* 818–833 (Springer/Cham, Switzerland, 2014).
38. Springenberg, J. T., Dosovitskiy, A., Brox, T. & Riedmiller, M. *Striving for simplicity: the all convolutional net.* <http://arxiv.org/abs/1412.6806> (2014). Accessed 13 Apr 2015.
39. Mascharka, D., Tran, P., Soklaski, R. & Majumdar, A. Transparency by design: closing the gap between performance and interpretability in visual reasoning. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition* 4942–4950 (Curran Associates/Red Hook, NY, USA, 2018).
40. Zhou, S.-M. & Gan, J. Q. Low-level interpretability and high-level interpretability: a unified view of data-driven interpretable fuzzy system modelling. *Fuzzy Sets Syst.* **159**, 3091–3131 (2008).
41. Wachter, S., Mittelstadt, B. & Floridi, L. Transparent, explainable, and accountable AI for robotics. *Sci. Robot.* **2**, eaan6080 (2017).
42. Chollet, F. & others. Keras. <https://keras.io> (2015).
43. Gutkowsky, R. et al. Unraveling compositional effects on the light-induced oxygen evolution in Bi(V–Mo–X)O₄ material libraries. *Energy Environ. Sci.* **10**, 1213–1221 (2017).
44. Zhou, D., Pang, L., Wang, H., Guo, J. & Randall, C. A. Phase transition, Raman spectra, infrared spectra, band gap and microwave dielectric properties of low temperature firing (Na_{0.5x}Bi_{1–0.5x})(MoxV_{1–x})O₄ solid solution ceramics with scheelite structures. *J. Mater. Chem.* **21**, 18412–18420 (2011).
45. Ancona, M., Ceolini, E., Oztireli, C. & Gross, M. Towards better understanding of gradient-based attribution methods for Deep Neural Networks. In *Proc. 6th International Conference on Learning Representations (ICLR, Zurich, 2018)*.
46. Sundararajan, M., Taly, A. & Yan, Q. Axiomatic attribution for deep networks. <https://arxiv.org/abs/1703.01365> (2017). Accessed 13 Jun 2017.
47. Yao, W., Iwai, H. & Ye, J. Effects of molybdenum substitution on the photocatalytic behavior of BiVO₄. *Dalt. Trans.* **11**, 1426–1430 (2008).
48. Gotić, M., Musić, S., Ivanda, M., Šoufek, M. & Popović, S. Synthesis and characterisation of bismuth (III) vanadate. *J. Mol. Struct.* **744**, 535–540 (2005).
49. Hardcastle, F. D., Wachs, I. E., Eckert, H. & Jefferson, D. A. Vanadium (V) environments in bismuth vanadates: a structural investigation using Raman spectroscopy and solid state ⁵¹V NMR. *J. Solid State Chem.* **90**, 194–210 (1991).
50. Merupo, V. I., Velumani, S., Oza, G., Makowska-Janusik, M. & Kassiba, A. Structural, electronic and optical features of molybdenum-doped bismuth vanadium oxide. *Mater. Sci. Semicond. Process.* **31**, 618–623 (2015).
51. Chollet, F. How convolutional neural networks see the world. <https://blog.keras.io/how-convolutional-neural-networks-see-the-world.html> (2016). Accessed 30 Jan 2016.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2019