MIT Open Access Articles

An Open Combinatorial Diffraction Dataset Including Consensus Human and Machine Learning Labels with Quantified Uncertainty for Training New Machine Learning Models

The MIT Faculty has made this article openly available. ***Please share*** how this access benefits you. Your story matters.

As Published: https://doi.org/10.1007/s40192-021-00213-8

Publisher: Springer International Publishing

Persistent URL: https://hdl.handle.net/1721.1/136823

Version: Author's final manuscript: final author's manuscript post peer review, without publisher's formatting or copy editing

Terms of use: Creative Commons Attribution-Noncommercial-Share Alike

AUTHOR ACCEPTED MANUSCRIPT

An Open Combinatorial Diffraction Dataset Including Consensus Human and Machine Learning Labels with Quantified Uncertainty for Training New Machine Learning Models

Cite this article as: Jason R. Hattrick-Simpers, Brian DeCost, A. Gilad Kusne, Howie Joress, Winnie Wong-Ng, Debra L. Kaiser, Andriy Zakutayev, Caleb Phillips, Shijing Sun, Janak Thapa, Heshan Yu, Ichiro Takeuchi and Tonio Buonassisi, An Open Combinatorial Diffraction Dataset Including Consensus Human and Machine Learning Labels with Quantified Uncertainty for Training New Machine Learning Models, Integrating Materials and Manufacturing Innovation https://doi.org/10.1007/s40192-021-00213-8

This Author Accepted Manuscript is a PDF file of an unedited peer-reviewed manuscript that has been accepted for publication but has not been copyedited or corrected. The official version of record that is published in the journal is kept up to date and so may therefore differ from this version.

Terms of use and reuse: academic research for non-commercial purposes, see here for full terms. https://www.springer.com/aam-terms-v1

Author accepted manuscript

© 2021 The Minerals, Metals & Materials Society.

AUTHOR ACCEPTED MANUSCRIPT

An Open Combinatorial Diffraction Dataset Including Consensus Human and Machine Learning Labels with Quantified Uncertainty for Training New Machine Learning Models

Jason R. Hattrick-Simpers[1*], Brian DeCost[1], A. Gilad Kusne[1], Howie Joress[1], Winnie Wong-Ng[1], Debra L. Kaiser[1], Andriy Zakutayev[2], Caleb Phillips[2], Shijing Sun[3], Janak Thapa[3], Heshan Yu[4], Ichiro Takeuchi[4], Tonio Buonassisi[3]

[1] Materials Measurement Science Division, National Institute of Standards and Technology, Gaithersburg, MD 20899, USA
[2]Materials, Chemical, and Computational Science Directorate, National Renewable Energy Laboratory, Golden, Colorado,
United States
[3]Department of Mechanical Engineering Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA
[4]Department of Materials Science and Engineering University of Maryland, College Park, MD 20742, USA

*** corresponding author: jason.hattrick-simpers@nist.gov**

Abstract

Modern machine learning and autonomous experimentation schemes in materials science rely on accurate analysis of the data ingested by these models. Unfortunately, accurate analysis of the underlying data can be difficult, even for domain experts, complicating the training of the models intended to drive experiments. This is especially true when the goal is to identify the presence of weak signatures in diffraction or spectroscopic data sets. In this work, we examine a set of as-obtained diffraction data that track the phase transition from monoclinic to tetragonal in a Nb doped VO_2 film as a function of temperature and dopant concentration. We then task a set of domain experts and a set of machine learning experts with identifying which phase is present in each diffraction pattern manually and algorithmically, respectively; in both cases the labels can vary dramatically especially at the phase boundaries. We use the mode of the labels and the Shannon entropy as a method to capture, preserve, and propagate consensus labels and their variance. Further we use the expert labels as a benchmark and demonstrate the use of Shannon entropy weighted scoring to test the performance of machine learning generated labels. Finally, we propose a materials data challenge centered around generating improved labeling algorithms. This real-world data set curated with expert labels can act as test bed for new algorithms. The raw data, annotations, and code used in this study are all available online at data.gov, and the interested reader is encouraged to replicate and improve the existing models

Author accepted manuscript

© 2021 The Minerals, Metals & Materials Society.

AUTHOR ACCEPTED MANUSCRIPT

Introduction

The past 5 years to 10 years are distinguished by a marked increase in the number of publications using data-driven methods (e.g., machine learning (ML)/artificial intelligence (AI)) by researchers in materials science, condensed matter physics and chemistry.[1–4] ML and AI are used often colloquially to signify the same concept, however here we use ML to signify an algorithm that learns and adapts through experience. Whereas an AI is a machine capable of simulating human thinking via learning. More recently a confluence between ML and experimental automation, referred to as autonomous science has emerged.[5,6] This new paradigm seeks to create robotic agents capable of planning studies, conducting experiments, and making decisions on the next iteration of experiments as new information is gathered. To date, several autonomous systems have been demonstrated to expedite the discovery of organic hole-transport materials[7], explore the toughness of additively manufactured plastic structures[8] and optimize the synthesis of carbon nanotubes[9]. Central to the application of ML to either autonomous or conventional experiments is the need for ground truth datasets to evaluate model performance. This is a need regardless of whether supervised or unsupervised methods are used.

To date, much of the proof-of-concept work has relied on validating ML models against idealized datasets. However, this is insufficient for the generation of robust algorithms that can handle the autonomous exploration of novel material systems which often occurs far from ideal conditions. Unfortunately, in experimental materials science, identification of a universally accepted ground truth is seldom easily attained. Ground truth is inferred through observations of material interactions with external stimuli and analyzing it in the context of fundamental physical laws. Still, new understandings can emerge (e.g., the existence of quasicrystals[10]) and disagreements can continue indefinitely (e.g., structures of liquid and solid H_2O[11,12]) about what correct ground truth is, even when interpreting the same data. Identifying ground truth can be further complicated by measurement noise, background, and sample quality. In some cases, it is not possible to clearly identify the ground truth. In these cases, including the work described in this article, the ground truth can be approximated by community agreed upon (consensus) values or labels for a given material property or response that will be used to train or evaluate a ML model.

For instance, in autonomous studies of new materials, the protocols for synthesis, processing, and characterization have not been optimized. But the search space is large, so the risk of pursuing a new model-suggested material in unexplored regions of chemical/processing space is balanced by the time lost optimizing the model uncertainty for previous measurements. To reasonably achieve this balance, it is important to develop and benchmark models with datasets that accurately reflect the data to be experimentally generated. It is also vital that training and evaluation methods be developed to validate model performance on consensus ground truths that reflect measurement and annotation uncertainties. This is a known issue in the computer science literature, although most work focuses on removing crowd-sourcing bias and adversarial annotations, with relatively little work done on inferring ground truth based on a small set of expert annotations as is often the case in materials science.[13–16]

In this work we focus on a critical materials science task: validating the discovery of a

Author accepted manuscript

© 2021 The Minerals, Metals & Materials Society.

AUTHOR ACCEPTED MANUSCRIPT

new material through an understanding of its underlying crystal structure. Several recent studies have demonstrated autonomous x-ray diffraction platforms[17,18] with the goal of automating this task. A specific application is the use of temperature-dependent phase mapping to identify the onset and completion of phase transformations as a function of composition and temperature. This is a non-trivial task: diffraction patterns even from a single phase can contain numerous peaks, each of which can broaden, shift, or vanish due to microstructural effects and signal-to-noise ratio. If a second phase is present, then identifying that phase may be complicated by peak overlap and SNR. Thus, it is not unreasonable that a group of experts provided with the same set of diffraction data will disagree about the positions of the phase boundaries. An additional complication is that context is important for interpreting diffraction patterns; comparison of patterns close in composition and temperature is often used to estimate the position of boundaries. Context is also an issue for the technical literature where the primary data used to make the “ground truth” judgement have often been lost.[19]

The first goal of this paper is to use a difficult-to-interpret diffraction dataset as a test case for investigating variability in human and machine labeling. The diffraction data were generated by performing temperature-dependent diffraction measurements on a $V_{1-x}Nb_xO_{2-y}$ “combinatorial” composition-spread sample discussed in detail in reference [[20]]. We provided the dataset to 5 experts in the interpretation of diffraction data and four materials data science experts and challenged them with labeling the phase(s) of each composition-temperature combination. The second goal is to demonstrate the use of statistical tools to identify the consensus label (e.g., human vs. human and ML vs. ML) for each spectrum and to quantify the label’s uncertainty. In general, the human labels agreed with one another, except for in the specifics of where the phase boundaries were positioned, while the ML algorithms had substantially greater variance. The final goal is to demonstrate a statistical means of benchmarking new ML labeling techniques to the consensus human labels (with variance). Finally, using these data, we propose an open materials challenge to humans and computers alike. The dataset, the human and machine-generated labels, and the code used to generate the machine-generated labels are all available via data.gov [[21]]. We encourage the submission of new human and ML labels of the dataset which will be curated and added to the online data set.

Experimental Procedure

The details of the synthesis method for the film used in this study and the associated deposition tool have been described elsewhere[20,22]. Briefly, the films were deposited as layer-by-layer $V_{1-x}Nb_xO_{2-y}$ composition spreads using combinatorial pulsed laser deposition. The targets were 25.4 mm V_2O_5 and Nb_2O_5 disks that were ablated by a KrF laser at 10 Hz with energies between 200 mJ/pulse and 230 mJ/pulse. The films were deposited on 76.2 mm diameter silicon substrates in an oxygen partial pressure of 0.65 Pa to a maximum thickness of 297 nm. The substrate was maintained at 793 K during the deposition and was cooled back to room temperature in vacuum afterwards.

For diffraction versus temperature measurements, a strip measuring 76.2 mm x 10 mm was cleaved from the center of the wafer. Compositional measurements were performed on this strip via x-ray photoelectron spectroscopy on a regular grid and the values were interpolated to

Author accepted manuscript

© 2021 The Minerals, Metals & Materials Society.

AUTHOR ACCEPTED MANUSCRIPT

provide a set of x,y positions in roughly 1 at.% increments. The XPS measurements were accurate to within 0.1 at.% but uncertainties in the interpolation and sample positioning result in an uncertainty closer to 0.3 at.%. X-ray diffraction measurements were performed using a Bruker D8 Discover[1] powder diffractometer. The system was equipped with a movable XYZ stage and a 2-D detector, and the sample was irradiated with a Cu k-alpha micro-source (nominal beam size 500 microns). The sample was mounted to a hotplate using silver paste. Diffraction measurements were taken isothermally for all compositions between temperature ramping to avoid hysteresis effects at 296 K, 303 K, 309 K, 318 K, 323 K, 328 K, 335 K, and 341 K. At each temperature, the sample was allowed to equilibrate for 10 minutes prior to performing a diffraction measurement and the sample was measured in ~1 at.% increments. Each diffraction pattern was taken using a fixed geometry with an incident angle of 14° and the detector covering a 2θ range of 18.00º to 37.20º. Each pattern was integrated for 10 min.

Procedure for Labeling

The full grid of diffraction patterns for every composition and temperature combination is available in the CombiView[23] format at Data.gov [[21]]. The diffraction and ML experts were given the complete dataset and asked to label each diffraction pattern as exhibiting monoclinic, tetragonal or mixed phases. In other words, the human and ML experts were provided with 352 diffraction patterns (2-theta versus intensity) structured in the CombiView format, as described in the SI. Each diffraction pattern was explicitly linked to a Nb concentration, temperature, and a wafer position. They were told to limit the composition range considered: to be below 20 at.% to 25 at.%; a sufficient range to fully capture the expected phase transformation. For comparison of the sets of labels, we considered diffraction patterns to samples with less than 25 at.% Nb for human labels and 20 at.% for ML labels. All human and ML labelers were told which material system the patterns came from (V, Nb containing oxide), provided the Nb composition, the measurement temperature for each data point, and asked to sort the XRD data into three classes. The process for generating human labels was not specified. The ML labelers were constrained to work in an unsupervised mode, without access to any pre-labeled data.

An anonymized, detailed write-up for how each human labeled the diffraction data is contained in the supplementary material. Generally, the methods employed by humans can be grouped into two classes, those that tracked the breadth of the diffraction peaks and those that tracked the intensity of the peaks. In the first class, various methods for approximating the full width half maximum (FWHM) were employed: (1) peak fitting with commercial diffraction software, (2) monitoring the position at ⅓ of the max intensity of a peak, and (3) using interactive data visualization software (CombiView) to track the peak breadth. In the second class, the diffraction data was plotted via a contour map and the intensities as a function of spectrum number and 2θ were used to manually cluster the data into the three regions. In both instances, trends in either the FWHM or intensity as a function of composition and temperature were used to distinguish between the different phase regions.

Author accepted manuscript

[1] *Certain commercial equipment, instruments, or materials are identified in this paper in order to specify the experimental procedure adequately. Such identification is not intended to imply recommendation or endorsement by the National Institute of Standards and Technology, nor is it intended to imply that the materials or equipment identified are necessarily the best available for the purpose.*

© 2021 The Minerals, Metals & Materials Society.

AUTHOR ACCEPTED MANUSCRIPT

The four ML labelers each used slightly different versions of spectral clustering with three clusters in their analysis. The specific implementation of spectral clustering was different depending upon the labeler, though all used some variant of the radial basis function (RBF)[24] kernel to define the affinity matrix. The first ML labeler employed the default scaling factor (sigma = 1) with a cosine distance-based[25] variant of the RBF affinity. The second ML labeler used the same cosine RBF affinity variant with a local scaling factor[26]. The third ML labeler used the standard RBF variant with the default scaling factor with an underlying distance metric that combines the pairwise cross-correlation between diffraction data with a compositional term. The final ML labeler first employed a variational autoencoder (VAE)[27] to learn two-dimensional representations of the diffraction data, and then applied standard spectral clustering to the latent diffraction representations, using a scaling factor related to the number of VAE latent features. The tools used are summarized in Table 1.

Table 1: Comparison of spectral clustering preprocessing and hyperparameters.

ML Labeling Technique	**Cosinel**	**Cosine-Local-Scaling**	**Comp-Distance**	**VAE**
Preprocessing	Linear background subtraction, truncation from 26.5º – 29º	Linear background removal, 2θ truncation 27º - 28.5º	None	Baseline subtraction, normalization, truncation to 27º - 28.75º and VAE
Composition range	Nb ≤ 24 at.%	Nb ≤ 25 at.%	Nb ≤ 20 at.%	Nb ≤ 21 at.%
Distance Metric	Cosine	Cosine	Cross-correlation + Composition term	Euclidean
Scaling Factor	Sigma = 1	Local scaling[26]	Sigma = 1	1/(number of VAE features/2)^(½)

Author accepted manuscript

Quantifying Consensus

To quantify the degree of consensus and uncertainty of the human and ML labels, we computed the Shannon entropy $H(X) = -\sum_{i=1}^{n} P(x_i) log(P(x_i))$ for the label distribution of each composition-temperature point. The entropy quantifies how informative the observation of a given label is when all observations are averaged over all possible outcomes. If all the labels for given data point agree the entropy is 0, and the value of the entropy increases as more labelers disagree. The mode of each label was used to generate the consensus label for each diffraction pattern while the Shannon entropy was used to represent the label certainty. In the event of an even split, here we default to labeling the range as being mixed-phase.

Results & Discussion

Dataset Introduction

© 2021 The Minerals, Metals & Materials Society.

AUTHOR ACCEPTED MANUSCRIPT

VO_2 is known to undergo a phase transformation from monoclinic to tetragonal phase transformation near 340 K.[28] This structural transition is associated with a change in the opacity of the films owing to a related metal-insulator phase transformation.[29] Due to the proximity of the transition to room temperature, there has been significant interest in VO_2 as a smart window coating that would preferentially reject infrared red light on warm sunny days. The addition of heavy elements such as W, Nb, and Mo is known to rapidly suppress the transformation temperature below room temperature, often also impacting the width of the transition region.[30–32]

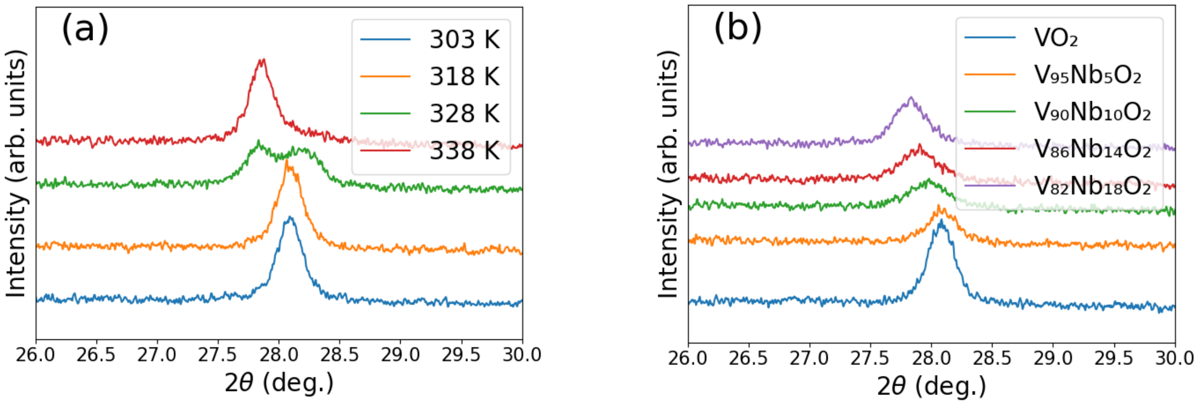

Figure 1 (a) Diffraction versus temperature for a V0.99Nb0.01O2 sample in the library, illustrating the clear peak splitting of a martensitic like transformation from monoclinic to tetragonal structures. (b) Diffraction versus composition at constant temperature for a series of VxNb1-xO2 films illustrating a less clear transition from the monoclinic to tetragonal structures.

Fig. 1 (a) presents a sample of the diffraction data, showing the temperature dependence of the diffraction pattern for a $V_{0.99}Nb_{0.01}O_2$ film from 303 K to 338 K from reference [[20]]. Within this temperature range, a single peak is observed at a 2θ value of 28.1º, corresponding to monoclinic VO_2. At 328 K, a clear splitting of the peak is observed with a new peak appearing as a shoulder at 2θ = 27.9º, corresponding to tetragonal VO_2. The 1st order nature of the transformation permits the co-existence of both phases possible within the transformation region. Finally, by 338 K the tetragonal peak is the most intense one. Note that in this last scan, it is difficult to determine whether the asymmetry of this diffraction peak is instrumental noise or the presence of some residual monoclinic phase.

Using these attributions and previous measurements of the optical properties[20], it is possible to identify the onset temperatures of the transitions (e.g. the boundaries between the observed phase regions). From Fig. 1(a), the onset temperature for the start of the metal-insulator transition for $V_{0.99}Nb_{0.01}O_2$ is between 318 K and 328 K, which is consistent with the previous report on this sample.[20] The clear bifurcation of the diffraction peak means it is likely that all the labels would exhibit strong consensus in identifying phase boundaries for such diffraction patterns.

Fig. 1 (b) presents the compositional variation of the diffraction pattern as Nb is substituted into VO_2 at room temperature. Up to 5 at.% Nb, the position (2θ) of the monoclinic peak is relatively unimpacted, although the peak intensity decreases. Between 5 at.% Nb and 10 at.% Nb the intensity of this peak decreases, the FWHM increases, and the peak shifts to a lower angle. Similar broadening behavior of the primary diffraction peak as a function of

© 2021 The Minerals, Metals & Materials Society.

AUTHOR ACCEPTED MANUSCRIPT

composition has been reported previously by Yiang et. al[33]. Finally, between 10 at.% Nb and 18 at.% Nb, the peak is observed to sharpen and increase in intensity. The peak position in this region starts at 27.8º, which corresponds to the tetragonal phase, but continues to shift to lower angle with increasing Nb concentration. The lack of peak splitting makes the determination of the borders of the transition region very challenging, as was noted in the original manuscript.[20]

Human Labels of Diffraction Data

Figure *2* shows a summary of the human diffraction data labels. Each circle represents a single composition and temperature and the symbol shape represents the Shannon entropy of the human labels, or the amount of disagreement between the labels. The small circles represent points with perfect consensus, with large circles representing some disagreement, squares more disagreement, and the triangles complete disagreement. The colored backgrounds are generated by the mode of the consensus class labels. The dashed lines represent the estimated phase boundaries between the monoclinic, multiphase, and tetragonal phase regions.

Fig. 2 (a) presents the consensus phase maps for composition versus temperature for all five sets of human labels. One labeler assumed diffraction from a first order phase transition could not exhibit the presence of multiple phases and thus created a set of labels quite distinct from the rest of the group. However, it is well known in the study of such transformations that both the high and low temperature phases can coexist over a range of temperatures and therefore Fig. 2 (b) removes this set of labels. In both panels, the dark blue range (monoclinic) has low Shannon entropy away from the phase boundary. In both panels, the Shannon entropy increases near the monoclinic to multiphase transition (dark blue to teal) with Fig 2(b) showing a maximum entropy for samples on the boundary. Within the multiphase region Fig 2 (a) there are few points with zero entropy, upon removing one set of human labels (Fig 2(b)) a considerable zero entropy region is observed. The entropy increases again near the boundary between the multiphase and tetragonal phase regions (teal to yellow) and it is notable that the range of

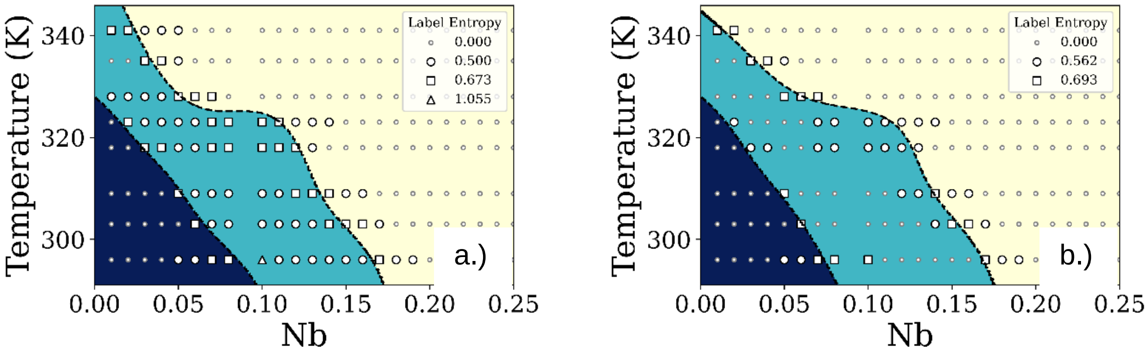

Figure 2 (a) consensus phase maps as a function of Nb concentration and temperature including all human labels. (b) human consensus labels with a single set of labels removed. Background represents the consensus phase in each region as calculated using the mode of the consensus labels: monoclinic (dark blue), mixed phase (teal), and tetragonal (yellow). The symbol shape represents the amount of Shannon entropy (measure of disagreement) between labelers.

Accepted manuscript

© 2021 The Minerals, Metals & Materials Society.

AUTHOR ACCEPTED MANUSCRIPT

compositions and temperatures with high entropy is much larger than in the monoclinic to multiphase transition. The tetragonal range (away from the boundaries) shows perfect agreement (zero entropy) between labelers. Overall, the majority of the labels had an entropy of 0 (e.g. all labels were identical) when considering four out of the five sets of human labels.

In terms of the positioning of the boundaries, the data sets with and without the fifth set of labels are very similar as would be expected with a large ⅘ consensus. There are two notable exceptions. First, the monoclinic to two-phase transformation at room temperature, where the addition of the 5th set of labels moves the boundary to > 10 at.% versus < 8 at.%. Second, although the general trend of the multiphase to tetragonal boundary is consistent, for low Nb concentrations is absolute placement varies by a few data points.

Focusing the discussion on Fig. 2 (b), for the lowest Nb concentrations, the expert labels agree up until the highest temperature measurement. As was expected, the largest entropies are observed near the boundaries between the different phase regions. Interestingly, even in the absence of peak splitting, the human labels for the transition from monoclinic to two-phase (Nb < 8 at.%) were consistent. In total, the monoclinic to two-phase boundary accounted for 10 of the 43 overall points with non-zero entropies and 5 of those disagreements can be found at the room temperature boundary. In a notable departure from the established literature on the VO_2 transition, four of the general diffraction experts evenly split as to whether the transformation completed by 341 K for the lowest Nb content sample. This is attributable to difficulties in distinguishing background and diffuse scattering effects from the presence of a small volume fraction of a residual phase.

Author accepted manuscript

ML Labels of Diffraction Data

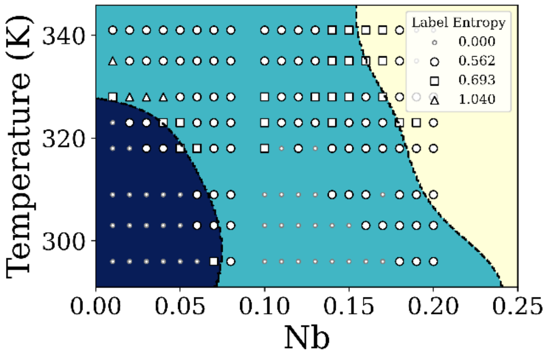

Figure 3 Shannon entropy and phase attribution results performed on the four spectral clustering results. Background represents the consensus phase in each region: monoclinic (dark blue), mixed phase (teal), and tetragonal (yellow). The symbol shape represents the amount of Shannon entropy (measure of disagreement) between labelers.

Fig. 3 presents the Shannon entropy and phase classification results performed on the four ML (spectral clustering) results. All four sets of labels clearly identify the monoclinic phase of VO_2. The ML consensus monoclinic (dark blue) to mixed phase (teal) boundary of the four models closely mirrors that of the human experts, especially when accounting for the human label entropy. The ML consensus boundary between the two-phase region (teal) and the tetragonal (yellow) phase is located above 20 at.% Nb at room temperature. In terms of the

© 2021 The Minerals, Metals & Materials Society.

AUTHOR ACCEPTED MANUSCRIPT

label entropy, the models largely agree in the monoclinic range, otherwise for most other points the labels disagreed with one another. There are two other regions of agreement, below 313 K between 10 at.% Nb and 17 at.% Nb, and the top right corner of the diagram.

It is notable that none of the four models detected the full monoclinic to two-phase to tetragonal phase transformation for any sample, particularly those with low Nb content. As was the case for the human labels, the onset of the monoclinic to multiphase transformation for the sample with the least Nb content was a source of labeling entropy for the ML labeled sets. A key differentiation between the ML and human labeling was that for the ML labels the maximum peak intensity drove the cluster assignments rather than the shape of the peak. Conversely, the majority of human labelers focused more on peak shape (either by eye or through fitting) in their labeling. In fact, no set of ML labels matched the human labels qualitatively.

Method for Comparing Model Effectiveness

Key to enabling more automated ML analysis of experiments is to be able to benchmark the ML labels against the human consensus and variance. This allows for an equitable comparison where models are not overly penalized for “incorrect” predictions where experts disagree. In order to quantitatively compare labeling by ML to the consensus of the human experts, we calculated a raw and confidence-weighted accuracy score for each ML technique considered here. We evaluated the scores for all samples with composition <= 20 at.% Nb. If the score is 1, then the ML label is consistent with a zero-entropy (*i.e.* unanimous) human consensus; a score of 0 indicates that the ML disagrees with a zero-entropy human consensus. If there was disagreement about a label (non-zero entropy) then the score is down-weighted such that `weight = 1 - H(x) / $H_{uniform}$`. Where H(x) is the Shannon entropy of the human label distribution for the instance, and $H_{uniform}$ is the Shannon entropy for a uniform distribution.

In order to generate an overall view of the effectiveness of each ML model as measured against the human labels, the score for each model were summed for all compositions and temperatures and normalized to the number of spectra (Figure 5). For the sake of comparison, an unweighted accuracy is also included. The dashed line represents the maximum possible score, given the uncertainties from the human labels. Although in this instance the relative performance of the techniques is preserved, the overall score is reduced so that 3 out of the 4 techniques have weighted accuracies of less than 50%.

As discussed in the methods section, although the ML practitioners were given the freedom to evaluate Nb concentrations between 20 at.% and 25 at.%, the actual comparisons were performed on the composition region of maximal overlap (up to 20 at.%). When the results of the models were thus confined, VAE spectral clustering did have a slightly better overall score. However, we found that the quality of the cluster assignments from each of these algorithms is sensitive to the composition threshold and will discuss this in greater detail in a subsequent paper.

Author accepted manuscript

© 2021 The Minerals, Metals & Materials Society.

AUTHOR ACCEPTED MANUSCRIPT

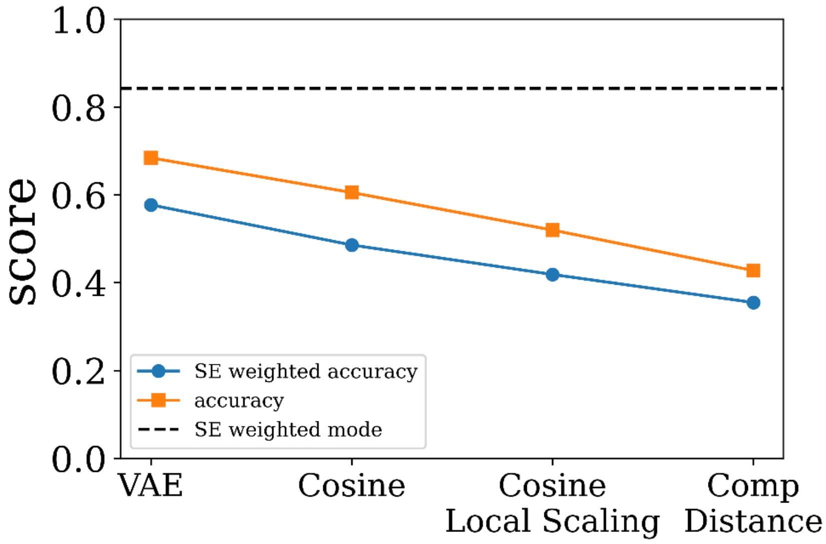

Figure 4 Plot of the accuracy and SE weighted accuracy vs all ML techniques generated within this study. The dashed line represents the maximum possible score given the variance in the expert labels.

Conclusions and Open Materials Data Science Challenge

Materials scientists are increasingly using ML/AI in their scientific workflows, however, such studies are conducted using idealized datasets without any attempt to capture differences in data interpretation between individuals. Here we show that even for relatively simple analysis tasks, labeling phases from diffraction data, there can be substantial variance between experts and ML/AI systems.

We demonstrate that, for a classification task, the Shannon Entropy can used to provide confidence assessments on multiple label assignments, obtained either from a panel of scientists or from a suite of automatic cluster analysis algorithms. We use these tools to show that, as expected, human experts disagree the most along the boundaries when a new phase appears, or an old phase vanishes. We also show that the currently considered group of ML models can correctly label the “easy” problem of looking for the transformation from monoclinic to multiphase but were not able to identify the more nuanced multiphase to tetragonal transition. Finally, we show that using these tools one can quantify the effective performance of a ML model when measured against the community consensus. Effectively this allows (1) ML models to be optimized in a manner that prioritizes data points with high consensus versus data points with large variance and (2) an even footing method of comparing ML models in the light of expert uncertainty.

The data from this manuscript including the raw diffraction patterns as a function of composition and temperature, the anonymized human and ML labels, and the human and ML consensus labels are made available at Data.gov. The interested reader is encouraged to send to the corresponding author of this study human, ML, and human-ML generated labels for this

© 2021 The Minerals, Metals & Materials Society.

Author accepted manuscript

AUTHOR ACCEPTED MANUSCRIPT

dataset. New human labels and ML labels will be added to the consensus data set which will be updated periodically. A condition for incorporation of new ML models will be the comparability of the new model's performance with a series of additional diffraction datasets on similar materials systems. This will help avoid the issue of overfitting to the available dataset. The highest model scores will be maintained in a separate file (the user can choose if they want to be anonymized).

Conflict of Interest Statement: On behalf of all authors, the corresponding author states that there is no conflict of interest.

Acknowledgements: The authors thank Marcus Mendenhall and Kamal Choudhary for their careful reading of our manuscript and helpful suggestions for its improvement. We also kindly thank our anonymous reviewer who worked with us to improve the focus of the original manuscript. Diffraction measurements were performed at the University of Maryland X-ray Crystallographic Center. The work at the National Renewable Energy Laboratory (NREL), operated by Alliance for Sustainable Energy LLC for the US Department of Energy (DOE) under Contract No. DE-AC36-08GO28308, was funded by the Laboratory Directed Research and Development (LDRD) Program. The views expressed in the article do not necessarily represent the views of the DOE or the US Government.

Author accepted manuscript

References

1. Schmidt, J., Marques, M. R. G., Botti, S. & Marques, M. A. L. Recent advances and applications of machine learning in solid-state materials science. *npj Computational Materials* vol. 5 1–36 (2019) doi:10.1038/s41524-019-0221-0.
2. Maksov, A. *et al.* Deep learning analysis of defect and phase evolution during electron beam-induced transformations in WS2. *npj Comput. Mater.* **5**, 1–8 (2019) doi:10.1038/s41524-019-0152-9.
3. Zhang, L., Lin, D. Y., Wang, H., Car, R. & Weinan, E. Active learning of uniformly accurate interatomic potentials for materials simulation. *Phys. Rev. Mater.* **3**, 023804 (2019) doi:10.1103/PhysRevMaterials.3.023804.
4. Li, W., Field, K. G. & Morgan, D. Automated defect analysis in electron microscopic images. *npj Comput. Mater.* **4**, 36 (2018) doi:10.1038/s41524-018-0093-8.
5. Aspuru-Guzik, A. & Persson, K. A. *Materials Acceleration Platform*. *Mission Innovation - Innovation Challenge 6* (2018).
6. Montoya, J. H. *et al.* Autonomous intelligent agents for accelerated materials discovery. *Chem. Sci.* **11**, 8517–8532 (2020) doi:10.1039/d0sc01101k.
7. MacLeod, B. P. *et al.* Self-driving laboratory for accelerated discovery of thin-film materials. *Sci. Adv.* **6**, eaaz8867 (2020) doi:10.1126/sciadv.aaz8867.
8. Gongora, A. E. *et al.* A Bayesian experimental autonomous researcher for mechanical design. *Sci. Adv.* **6**, eaaz1708 (2020) doi:10.1126/sciadv.aaz1708.
9. Nikolaev, P. *et al.* Autonomy in materials research: a case study in carbon nanotube growth. *npj Comput. Mater.* **2**, 16031 (2016) doi:10.1038/npjcompumats.2016.31.
10. Cahn, J. W., Gratias, D. & Shechtman, D. Pauling's model not universally accepted. *Nature* **319**, 102–103 (1986) doi:10.1038/319102a0.
11. Brini, E. *et al.* How Water's Properties Are Encoded in Its Molecular Structure and Energies. *Chem. Rev.* **117**, 12385–12414 (2017) doi:10.1021/acs.chemrev.7b00259.
12. Smart, A. G. The war over supercooled water. *Phys. Today* (2018) doi:10.1063/pt.6.1.20180822a.

© 2021 The Minerals, Metals & Materials Society.

AUTHOR ACCEPTED MANUSCRIPT

13. Krause, J. *et al.* Grader Variability and the Importance of Reference Standards for Evaluating Machine Learning Models for Diabetic Retinopathy. *Ophthalmology* **125**, 1264–1272 (2018) doi:10.1016/j.ophtha.2018.01.034.
14. Sheng, V. S., Provost, F. & Ipeirotis, P. G. Get Another Label? Improving Data Quality and Data Mining Using Multiple, Noisy Labelers. in *Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD 08* (ACM Press, 2008).
15. Raykar, V. C. *et al. Learning From Crowds 1. Supervised Learning From Multiple Annotators/Experts*. *Journal of Machine Learning Research* vol. 11 https://www.mturk.com. (2010) doi:10.5555/1756006.1859894.
16. Wauthier, F. L. & Jordan, M. I. Bayesian Bias Mitigation for Crowdsourcing. in *Proceedings of the 24th International Conference on Neural Information Processing Systems* 1–9 (2011).
17. Kusne, A. G. *et al.* On-the-fly machine-learning for high-throughput experiments: Search for rare-earth-free permanent magnets. *Sci. Rep.* **4**, 1–7 (2014) doi:10.1038/srep06367.
18. Noack, M. M. *et al.* A Kriging-Based Approach to Autonomous Experimentation with Applications to X-Ray Scattering. *Sci. Rep.* **9**, 1–19 (2019) doi:10.1038/s41598-019-48114-3.
19. Joress, H. *et al.* A High-Throughput Structural and Electrochemical Study of Metallic Glass Formation in Ni-Ti-Al. *ACS Comb. Sci.* **22**, 330–338 (2020) doi:10.1021/acscombsci.9b00215.
20. Barron, S. C., Gorham, J. M., Patel, M. P. & Green, M. L. High-Throughput Measurements of Thermochromic Behavior in V $_{1-x}$ Nb $_x$ O $_2$ Combinatorial Thin Film Libraries. *ACS Comb. Sci.* **16**, 526–534 (2014) doi:10.1021/co500064p.
21. *The data and code for generating the figures has been uploaded to data.gov pending final approval for release. They have been provided as a ZIP file addendum to the manuscript.*
22. Bassim, N. D., Schenck, P. K., Otani, M. & Oguchi, H. Model, prediction, and experimental verification of composition and thickness in continuous spread thin film combinatorial libraries grown by pulsed laser deposition. *Rev. Sci. Instrum.* **78**, 072203 (2007) doi:10.1063/1.2755783.
23. Long, C. J. CombiView. https://sourceforge.net/projects/xrdsuite/ (2013).
24. Buhmann, M. D. *Radial Basis Functions: Theory and Implementations*. *Cambridge Monographs on Applied and Computational Mathematics* (Cambridge University Press, 2003). doi:DOI: 10.1017/CBO9780511543241.
25. Cosine Distance. https://www.itl.nist.gov/div898/software/dataplot/refman2/auxillar/cosdist.htm.
26. Zelnik-Manor, L. & Perona, P. Self-tuning spectral clustering. in *Proceedings of the 17th International Conference on Neural Information Processing Systems* 1601–1608 (2004).
27. Kingma, D. P. & Welling, M. Auto-encoding variational bayes. in *2nd International Conference on Learning Representations, ICLR 2014 - Conference Track Proceedings* (International Conference on Learning Representations, ICLR, 2014).
28. Warwick, M. E. A. & Binions, R. Advances in thermochromic vanadium dioxide films. *J. Mater. Chem. A* **2**, 3275–3292 (2014) doi:10.1039/c3ta14124a.
29. Kozen, A. C. *et al.* Structural Characterization of Atomic Layer Deposited Vanadium Dioxide. *J. Phys. Chem. C* **121**, 19341–19347 (2017) doi:10.1021/acs.jpcc.7b04682.
30. Nishikawa, M., Nakajima, T., Kumagai, T., Okutani, T. & Tsuchiya, T. Adjustment of thermal hysteresis in epitaxial VO2 films by doping metal ions. *J. Ceram. Soc. Japan* **119**, 577–580 (2011).
31. Gomez-Heredia, C. L. *et al.* Measurement of the hysteretic thermal properties of W-doped and undoped nanocrystalline powders of VO2. *Sci. Rep.* **9**, 1–14 (2019)

Author accepted manuscript

© 2021 The Minerals, Metals & Materials Society.

doi:10.1038/s41598-019-51162-4.

32. Miyazaki, K., Shibuya, K., Suzuki, M., Wado, H. & Sawa, A. Correlation between thermal hysteresis width and broadening of metal-insulator transition in Cr- and Nb-doped VO2 films. *Jpn. J. Appl. Phys.* **53**, 71102 (2014) doi:10.7567/JJAP.53.071102.

33. Liang, Y. G. *et al.* Tuning the hysteresis of a metal-insulator transition via lattice compatibility. *Nat. Commun.* **11**, 1–8 (2020) doi:10.1038/s41467-020-17351-w.

Author accepted manuscript

© 2021 The Minerals, Metals & Materials Society.