

ARTICLE OPEN



Materials structure–property factorization for identification of synergistic phase interactions in complex solar fuels photoanodes

Dan Guevarra^{1,4}, Lan Zhou^{1,4}, Matthias H. Richter¹, Aniketa Shinde^{1,3}, Di Chen², Carla P. Gomes² and John M. Gregoire¹✉

Properties can be tailored by tuning composition in high-order composition spaces. For spaces with complex phase behavior, modeling the properties as a function of composition and phase distribution remains a formidable challenge. We present materials structure–property factorization (MSPF) as an approach to automate modeling of such data and identify synergistic phase interactions. MSPF is an interpretable machine learning algorithm that couples phase mapping via Deep Reasoning Networks (DRNets) to matrix factorization-based modeling of the representative properties of each phase in a dataset. MSPF is demonstrated for Bi–Cu–V oxide photoanodes for solar fuel generation, which contains 25 different phase combinations and correspondingly exhibits complex composition–structure–photoactivity relationships. Comparing the measured photoactivity to a learned model for non-interacting phases, synergistic phase interactions are identified to guide further photoactivity optimization and understanding. MSPF identifies synergistic interactions of a BiVO₄-like phase with both Cu₂V₂O₇-like and CuV₂O₆-like phases, creating avenues for understanding complex photoelectrocatalysts.

npj Computational Materials (2022)8:57; <https://doi.org/10.1038/s41524-022-00747-1>

INTRODUCTION

Enhancing materials research via integration with artificial intelligence (AI) comprises a recent transformation in the evolution of materials science^{1–6}. Such efforts span the research lifecycle from experiment planning⁷ to data analysis⁸, with early demonstrations of AI-assisted data processing naturally occurring in image analysis since many machine learning (ML) algorithms were initially developed to automate pattern recognition in images^{9–11}. Complementary examples of AI-assisted data analysis include automated identification of composition–structure–property relationships where the existence of a relationship is automatically identified, but often the labeling of the relationships and their scientific interpretation are relegated to human experts^{12,13}.

To further automate mapping of structure-dependent properties, the community has made a concerted effort in crystal structure phase mapping, wherein a collection of XRD patterns is used to identify the identity and distribution of each crystalline phase, typically within a 2 to 4-element composition space^{14–19}. We recently reported Deep Reasoning Networks (DRNets)^{20,21} to automate phase mapping for complex datasets where prior algorithms and extensive human analyses fail to fully solve the underlying phase behavior. While phase mapping is a route to accelerate generation of phase diagrams in high-order composition spaces, most immediately the results are needed to interpret variations in measured materials properties, i.e. the underlying composition–structure–property relationships. Herein we present materials structure–property factorization (MSPF) as an approach to automate modeling of the properties as a function of composition and phase distribution, enabling identification of synergistic phase interactions. While the MSPF model of the present work can in principle be coupled to any phase mapping algorithm, MSPF is most necessary when the number of phase

combinations and the associated variations in materials properties exceed the capabilities of traditional analysis, and such complex datasets are precisely where DRNets excel at phase mapping.

Phase mapping can seed a variety of further investigations, as recently demonstrated for experiment planning to discover optimal phase-change materials in a ternary composition library²². In that work, the CAMEO framework was introduced wherein phase mapping results are used to partition the composition library into phase field regions. Each region receives an independent active learning model so that the desired materials property is optimized within each phase field. In that case, the maximal performance was observed in a phase-pure material, whereas the present work focuses on the identification of phase mixtures that exhibit synergistic interactions. While CAMEO may be extended to compare active learners from different phase fields, the training of machine learning models in each phase field is impractical for increased phase diagram complexity where some phase fields may contain only several composition samples.

In the present work, we use comprehensive high throughput experimentation to provide the performance data that is interpreted in the context of the phase map. In the example dataset of the present work, the Bi–Cu–V oxide composition library contains 25 unique combinations of phases with an average of about 13 samples per phase field. As a result, there is little opportunity to lower the required number of samples via training of 25 independent active learners, highlighting that for complex materials systems, the most direct route to accelerating the generation of scientific knowledge is to extract knowledge via the application of data science to large datasets from combinatorial experimentation. Toward this goal, we introduce MSPF to model the average performance contribution of each phase as well as the interactions among phases, especially when the composition-

¹Division of Engineering and Applied Science and Liquid Sunlight Alliance, California Institute of Technology, Pasadena, CA, USA. ²Department of Computer Science, Cornell University, Ithaca, NY, USA. ³Present address: Nova Measuring Instruments Inc., Fremont, CA, USA. ⁴These authors contributed equally: Dan Guevarra, Lan Zhou.

✉email: gregoire@caltech.edu

structure–property relationships are too complex to be readily interpreted by manual analysis. MSPF is demonstrated herein for the identification of optimal solar fuels photoanodes via measurement of photoelectrochemical (PEC) performance. Metal oxide photoanodes comprise a class of complex, multi-functional materials whose properties are often optimized in high-order compositions, including mixed-phase compositions^{23–28}. High throughput photoelectrochemistry experiments have been effective at discovering complex photoanodes^{23,29,30}, and a goal of the present work is to provide a framework for automating interpretation of the resulting complex datasets.

In MSPF for metal oxide photoanodes, phase concentration maps provided by phase mapping are used to identify representative PEC behavior of each phase and subsequently analyze the phase combinations that provide better or worse performance compared to the underlying assumption that the performance contribution from each phase is proportional to phase concentration. In the broader effort to automate data analysis and interpretation, the contributions of the present work include the first deployment of DRNets for analyzing structure–property relationships, the integration of phase mapping and property factorization to provide a compact and interpretable analysis of complex data, and illustration of the materials design principle of exploiting synergistic interactions among multiple phases to amplify materials performance. This identification of emergent properties in complex, multi-phase materials is critical to accelerated exploration and understanding of high-performance materials in high-order composition spaces^{29–31}.

RESULTS AND DISCUSSION

The data-driven search for synergies in multi-phase photoanodes commences with the deposition of four composition libraries, the Bi–Cu–V oxide library along with each of the 2-cation subspaces as shown in Fig. 1a. XRD measurements on 353 unique compositions produce diffraction patterns (see Fig. 1b) that collectively contain information about the phase behavior in this composition system but individually are not readily interpretable due to the presence of 2 or 3-phase mixtures from a set of 99 candidate phases. This set of phases is described in Supplementary Table 1 and was determined by searching the International Crystallography Diffraction Database (ICDD) for entries containing oxygen and any combination of Bi, Cu, and V. This set of phases contains many diffraction patterns with overlapping signal, especially when considering experimental measurement of grains with sub-100 nm crystallites due to the resulting broadening of diffraction peaks. The consequence is that a single XRD pattern may be well explained by dozens of unique combinations of phases from the 99 candidates. Even if each XRD pattern has only 2 alternate interpretations, the corresponding number of phase map solutions would be 2^{353} , which is larger than 10^{106} . Consequently, individually analyzing each XRD pattern and subsequently stitching together a solution from all the possible combinations of phases is intractable. Finding the best global solution to the collection of XRD patterns requires collectively demixing the phases in each XRD pattern while applying constraints that guide the demixing process to an optimal solution. The constraints are based on the thermodynamics of phase diagrams and include a maximum number of co-existing phases per Gibbs' phase rule, which is taken to be 2 for the 2-cation compositions and 3 for the 3-cation compositions under the approximation that the thin films reach equilibrium with the O₂ in the annealing atmosphere at the anneal temperature of 550 °C. Additional rules include the generation of phase fields that are connected within the composition graph, and the algorithm models the presence of alloying by identifying composition-dependent shifting of the XRD patterns indicative of lattice constant modulation. Since alloying comprises at least 1 thermodynamic degree of freedom, its

presence results in a reduction in the maximum number of phases to 1 for 2-cation and 2 for 3-cation compositions.

While several phase mapping algorithms have been developed to incorporate these rules^{15,16,18}, they have not been fully integrated into the pattern demixing to date since the enforcement of rules inherently involves computational reasoning, whereas the pattern demixing inherently involves computational learning. Deep Reasoning Networks^{20,21} are a framework for seamlessly integrating reasoning and learning, enabling phase mapping of XRD datasets whose complexity cannot be harnessed by manual analysis or prior phase mapping algorithms. The resulting DRNets phase map solution reveals the challenges in interpreting the measured XRD patterns, as the 353 XRD patterns collectively contain 21 different phases (see Supplementary Fig. 1 and Supplementary Table 2) in 33 unique combinations when considering phase concentrations above 1%, which is the approximate detectability limit. The distribution of phase concentrations for the primary 14 phases is shown in Fig. 1c.

Some of the 21 phases appear only as dilute phases and/or only appear in a few compositions, making the presence of the phase in the library insufficient for learning its properties. Considering phases that appear in at least 10 compositions, the composition libraries contain 11 phases, and the samples with these phases include 25 of the 33 phase combinations. Some Cu–V and Bi–V oxide phases are omitted by this criterion, which results in several compositions in these 2-cation spaces being removed from the below analysis, although since these phases are not prominent in the 3-cation composition region (as indicated by their lack of appearance in at least 10 compositions), they are relatively unimportant for the present focus on identifying multi-phase synergies. The resulting dataset contains 335 different compositions (data samples).

PEC characterization of the composition library reveals a range of shapes of the photocurrent vs. electrochemical potential (I – E) curves as well as overall photoactivity. Understanding the composition-photoactivity trends in the context of the 11 phases and 25 phase combinations requires assistance from machine learning, motivating our development of MSPF. We commence with an approximation for the expected behavior of a photoelectrode containing a simple mixture of phases. If the photoelectrode contained a combination of non-interacting phases, e.g. by piecing together a set of phase-pure electrodes, then the photocurrent of the mixed-phase electrode will be the sum of the photocurrent of each phase scaled by the respective fraction of the electrode occupied by that phase. While some mixed-phase photoelectrodes may behave according to this weighted-sum model, heterogeneous grain boundaries can have deleterious (e.g. recombination sites) or beneficial (e.g. junction that facilitates charge separation) effects. While discovering the deleterious effects is important for guiding future synthesis efforts by establishing the phase mixtures that should be avoided when scaling up photoelectrodes, the opportunities for photoanode optimization lie in the discovery of phase mixtures whose photoactivity exceeds that of the weighted-sum approximation. The strategy of the present work is to analyze deviations from the weighted-sum model and reserve analysis of the underlying causes of the deviations to future work. We note that alternatively the weighted-sum model could be replaced by a model for the specific type of physics under investigation, e.g. elasticity, ionic transport, or scattering processes, to directly learn parameters of interest from the phase map and performance data. Such an approach may be suitable when the physics of interacting phases is well understood, which is not the case for the mixed-phase photoelectrodes of the present work. Understanding this physics will require detailed investigation of exemplar systems, and the methods of the present work automate the identification of those exemplars.

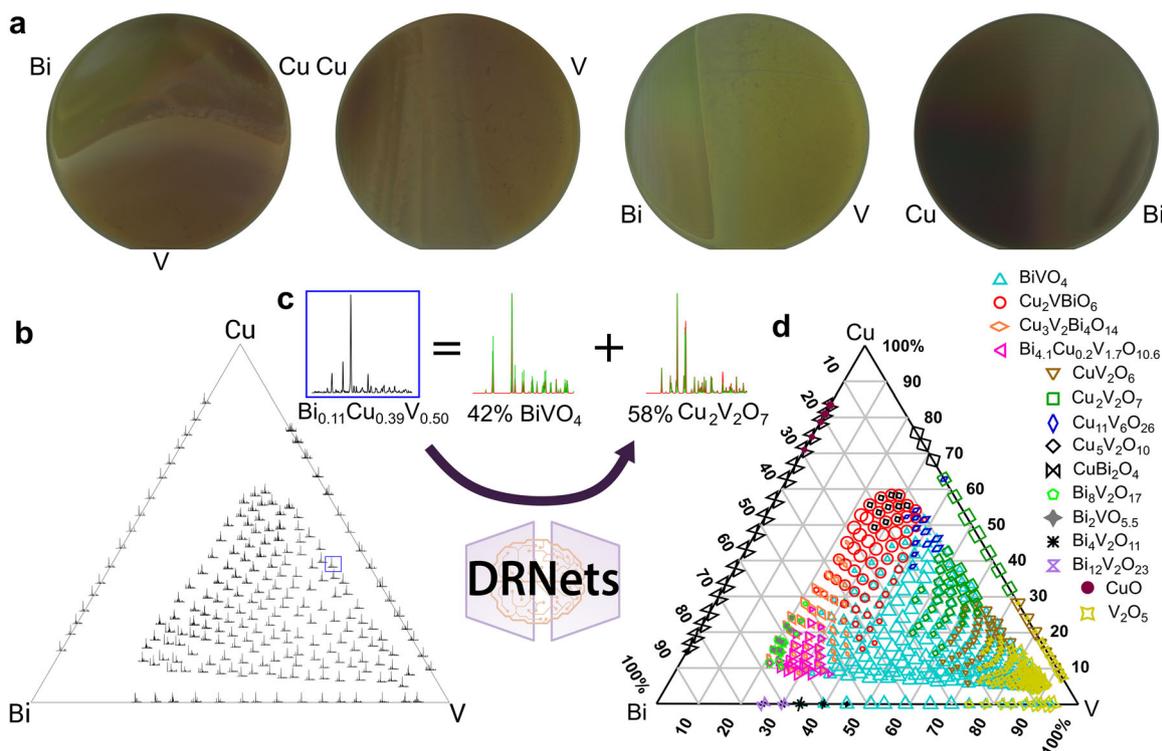


Fig. 1 Phase mapping in the Bi-Cu-V oxide system. **a** Photographs of 4 sputter-deposited composition libraries in Bi-Cu-V oxide system. **b** Representative x-ray diffraction patterns collected on the composition libraries, where each XRD pattern is plotted at the location of the composition in the Bi-Cu-V composition graph. **c** The DRNets solution for one composition, showing that the XRD pattern of the $\text{Bi}_{0.11}\text{Cu}_{0.39}\text{V}_{0.50}$ oxide sample is composed of 42% BiVO_4 and 58% $\text{Cu}_2\text{V}_2\text{O}_7$. **d** The resulting phase map of the primary 14 of the 21 phases identified by DRNets, where each phase has a unique symbol whose point size indicates phase concentration in the composition graph. Multi-phase samples have multiple overlapping symbols.

MSPF applies the weighted-sum model to the entire dataset via matrix factorization wherein the set of measured $I-E$ data is compiled in a matrix **A**. The weighted-sum approximation can be conceptualized as modeling the measured PEC data using a single $I-E$ curve for each phase, multiplying these curves by the phase concentrations, and summing over the phases. This process is captured by MSPF as illustrated in Fig. 2. **A** contains the photocurrent at 15 electrochemical potentials for the 335 compositions, resulting in a 15×335 matrix. This matrix is then approximated as the product of the basis pattern matrix and the phase concentration matrix. The matrix of basis patterns is a 15×11 matrix that contains the representative $I-E$ data for each of the 11 phases. Note that many of the phases are not observed as phase-pure materials, and even phase-pure materials may appear at different compositions due to alloying, so the data does not directly tell us the representative $I-E$ basis pattern for any of the phases, motivating the learning of basis patterns from the data. The concentration of the 11 phases is provided for each of the 335 compositions by the DRNets phase mapping solution, resulting in a matrix of size 11×335 .

We solve the matrix factorization problem piecewise by first considering the 2-cation compositions with the assertion that the basis pattern for 2-cation phases are best learned from the corresponding 2-cation data provided that the phase exists in sufficient phase purity in the 2-cation data. Using a minimum phase concentration threshold of 80%, the following 6 of the 11 phases have their basis components learned via matrix factorization with 33 compositions from the 2-cation data: BiVO_4 , CuBi_2O_4 , $\text{Cu}_2\text{V}_2\text{O}_7$, $\text{Cu}_5\text{V}_2\text{O}_{10}$, CuV_2O_6 , and V_2O_5 . These patterns are then frozen while using all 335 compositions to learn the basis patterns from the remaining 5 phases: $\text{Bi}_2\text{Cu}_2\text{VO}_7$, $\text{Bi}_4\text{Cu}_3\text{V}_2\text{O}_{14}$, $\text{Bi}_{4.1}\text{Cu}_{0.2}\text{V}_{1.7}\text{O}_{10.6}$, $\text{Cu}_{11}\text{V}_6\text{O}_{26}$, and $\text{Bi}_8\text{V}_2\text{O}_{17}$. This latter matrix

factorization step was evaluated for robustness against random removal of 20% of the 3-cation data, resulting in a positive assessment as detailed in Supplementary Fig. 2.

The results of this calculation are the MSPF model \mathbf{A}_{MSPF} of the PEC data as well as the basis $I-E$ patterns of each phase, which are shown for some phases in Fig. 2 and for all phases in Supplementary Fig. 3. These latter results indicate that CuBi_2O_4 has cathodic photocurrent with a turn on potential near 1 V vs RHE, in agreement with recent reports³². BiCu_2VO_6 shows appreciable photoanodic activity, which is also in agreement with prior work^{26–28}. The other 3-cation phases show small negative values in their basis patterns, which is due to a combination of small cathodic photocurrents and the general decrease in photocurrent when these phases are mixed with more photoactive phases. The 4 phases with highest photoactivity are BiVO_4 , $\text{Cu}_2\text{V}_2\text{O}_7$, $\text{Cu}_5\text{V}_2\text{O}_{10}$, and CuV_2O_6 . In Supplementary Fig. 4, the MSPF model for the photoactivity of each phase is compared to the measured photoactivity for the 2 samples closest in composition to the respective formula unit and with high-phase purity of the respective phase. This analysis shows that the MSPF model successfully captures the representative photoactivity of each phase.

The photocurrent residual matrix, $\mathbf{R} = \mathbf{A} - \mathbf{A}_{\text{MSPF}}$, is representative of measured photocurrent beyond that explained by the weighted-sum MSPF model. The residuals can be analyzed as a function of phase concentrations, composition, and/or electrochemical potential, and we commence analysis by calculating the average residuals for 3 different ranges of electrochemical potential and plotting them as a function of composition in Fig. 3. This analysis reveals 3 distinct composition regions with respect to the MSPF residuals. The region with substantially negative residuals is the 3-cation region with V concentration between

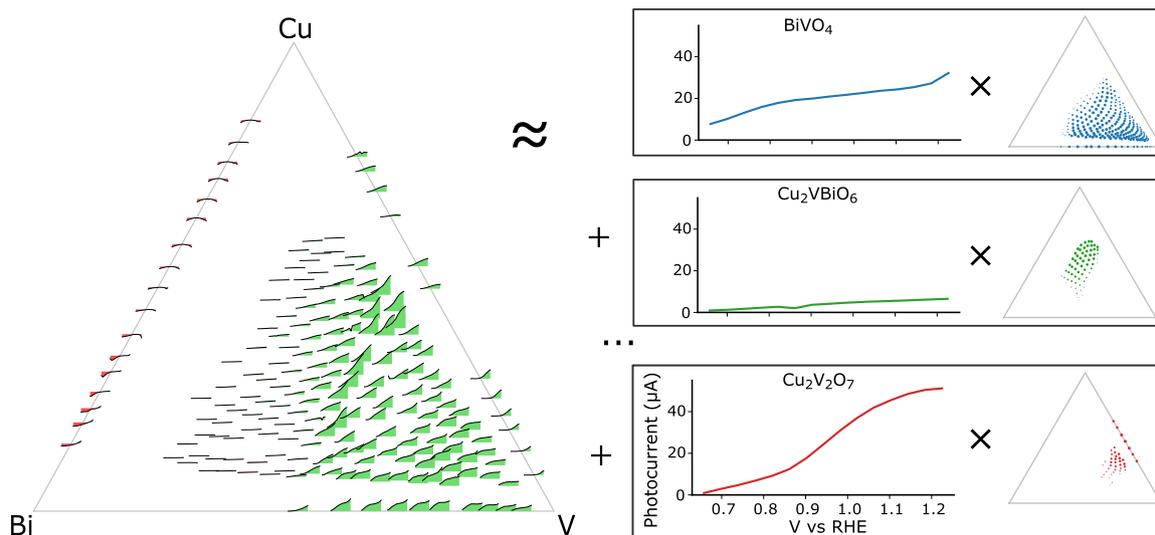


Fig. 2 Illustration of materials structure–property factorization. Representative measured I – E patterns in the Bi–Cu–V oxide composition space are shown with red and green fill color indicating negative and positive photocurrent, respectively. This set of photocurrent data is modeled as the sum of contributions for each phase, where the composition map of the contribution of each phase is determined by the product of a representative I – E pattern and the respective phase concentration map from DRNets.

30 and 50% and Cu concentration less than 25%. The primary phase field in this region is a 2-phase combination of BiVO_4 and $\text{Bi}_4\text{Cu}_3\text{V}_2\text{O}_{14}$. Given that BiVO_4 has the highest demonstrated photoanode performance out of all observed phases³³, adding a small amount of Cu to the $\text{Bi}_{0.5}\text{V}_{0.5}$ composition would be an intuitive strategy for optimizing photoactivity in the Bi–Cu–V system. The results of Fig. 3 reveal that this is an ineffective strategy. The MSPF analysis as well as inspection of the compositions with highest phase concentration of $\text{Bi}_4\text{Cu}_3\text{V}_2\text{O}_{14}$ indicate that this phase exhibits little photoactivity. The dilution of the photoactive BiVO_4 phase in the phase mixture with this 3-cation phase is expected to lower the photoactivity compared to the 2-cation BiVO_4 samples, yet the considerably negative values of R demonstrate that the decrease in photoactivity is even more drastic due to deleterious interactions of the 2 phases. While this analysis clearly identifies the deleterious phase interaction, it does not address the underlying physical mechanism.

The two composition regions with large positive values of R are (i) compositions with V concentration in excess of 70% and (ii) a smaller composition region with Cu concentration between 45 and 55% and Bi concentration below 20%. Figure 4 shows the measured photoactivity and its corresponding MSPF model for the maximum-residual composition in each of these regions. In Supplementary Fig. 5, the same measured photoactivity for the 2 select compositions are shown alongside the highest photoactivity observed with high-phase-purity photoelectrodes from the 2-cation composition space, revealing that in the respective potential regions, the select 3-cation composition outperforms not only the expectation from the MSPF model but also the best 2-cation compositions.

In the first composition region of interest, represented by $\text{Bi}_{0.07}\text{Cu}_{0.17}\text{V}_{0.76}$, the primary phase is CuV_2O_6 , which is a known photoanode phase (in the Cu–V 2-cation composition space) but with limited photoactivity, especially at low electrochemical potential^{34,35}. Figure 4a shows that this low electrochemical potential region is precisely where the phase mixture excels compared to the MSPF model, with measured photocurrents well above those of the MSPF model for potentials below 0.9 V vs RHE. Even though the sample of Fig. 4a is only 33% BiVO_4 in the DRNets phase map, the photocurrent in the low electrochemical potential region exceeds that of the best high-phase-purity BiVO_4 samples on the Bi–V line.

The second composition region of interest, represented by $\text{Bi}_{0.11}\text{Cu}_{0.39}\text{V}_{0.50}$, exhibits large positive values of R and contains the primary phase $\text{Cu}_2\text{V}_2\text{O}_7$, another known photoanode that has recently received considerable attention^{34,36–38}. Here, the photoactivity is also enhanced compared to the MSPF model at all potentials, with the most substantial increase observed above 0.9 V vs. RHE, where the photocurrent far exceeds the maximum values observed on the Bi–V and Cu–V 2-cation composition lines.

The results of the MSPF residual analysis strongly suggest beneficial interactions of the respective pairs of phases in the complex 3-cation photoelectrodes of Fig. 4. While the intention of the present work is to identify these exemplars of synergistic phase interactions to set the stage for future inquiry, we additionally seek to confirm that the composition map of photoactivity residuals (Fig. 3) is not governed by changes in morphology. Using cross-section transmission electron microscopy (TEM) and plan-view scanning electron microscopy (SEM) imaging, the morphology of the 2 primary compositions of interest is characterized and compared to a range of other compositions in Figs. S6–S9. For the $\text{Bi}_{0.07}\text{Cu}_{0.17}\text{V}_{0.76}$ oxide sample, the film thickness varies between 300 and 500 nm with ~200 nm wide BiVO_4 domains embedded in a CuV_2O_6 -like matrix with small domains of Cu-rich precipitates that were not detected in the XRD-based phase map. For the $\text{Bi}_{0.11}\text{Cu}_{0.39}\text{V}_{0.50}$ oxide sample, the film thickness also varies between 300 and 500 nm with a broader size range of BiVO_4 domains (10–200 nm) embedded in a $\text{Cu}_2\text{V}_2\text{O}_7$ -like matrix that exhibits variability in both microstructure and Cu:V composition. The substantial difference in morphology between these 2 compositions is representative of the substantial range of morphology observed over the broader range of compositions, including the substantial variations observed among high-phase-purity BiVO_4 samples. These results indicate that morphology alone does not account for the substantial residuals from the MSPF analysis. The MSPF analysis demonstrates that for this photoanode synthesis method, specific mixtures of phases perform well, for which the underlying mechanism of photoactivity improvement may involve a combination of electronic, catalytic, and morphological effects. These may be driven by a combination of modulations of the composition within each phase, alterations to the crystallization during synthesis, and/or electronic junctions formed between 2 phases. MSPF identifies the specific phase synergies whose

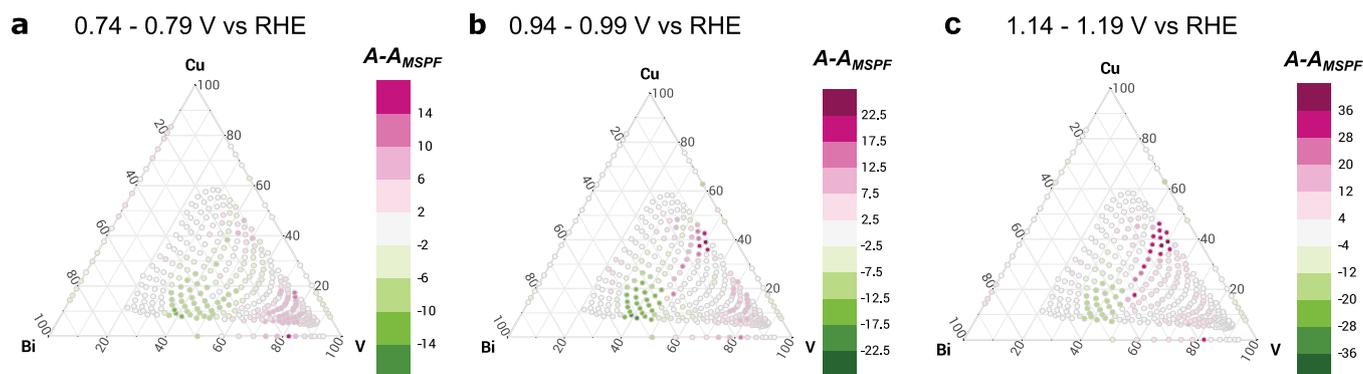


Fig. 3 Phase interaction detection. The residual from the factorization model of Fig. 2 is averaged over three different ranges of photoelectrochemical potential to ascertain whether a given composition outperforms (positive residual) or underperforms (negative residual) the expectation from materials structure–property factorization model at (a) 0.74–0.79 V vs RHE; (b) 0.94–0.99 V vs RHE; and (c) 1.14–1.19 V vs RHE.

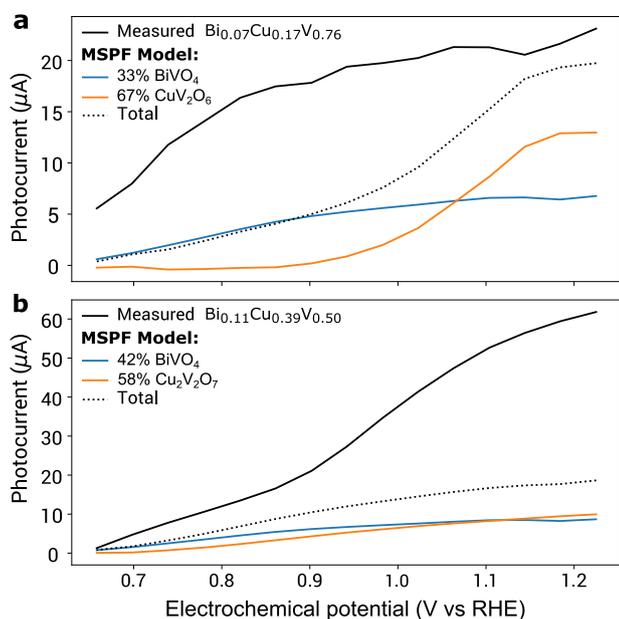


Fig. 4 The measured photoactivity and its corresponding property factorization model for the maximum-residual compositions in two composition regions shown in Fig. 3. a The $\text{Bi}_{0.07}\text{Cu}_{0.17}\text{V}_{0.76}$ photoanode composed of 33% BiVO_4 and 67% $\text{Cu}_2\text{V}_2\text{O}_6$; b The $\text{Bi}_{0.11}\text{Cu}_{0.39}\text{V}_{0.50}$ photoanode composed of 42% BiVO_4 and 58% $\text{Cu}_2\text{V}_2\text{O}_7$. Note that these phase concentrations are labeled by prototype formula unit, although the composition within each phase may be different due to alloying.

mechanistic underpinnings must be elucidated via further study. In the present case of studying photocurrent as a function of applied potential, we observe that the 2-phase mixtures of interest exhibit distinct potential-dependence variations in the photocurrent residuals, which suggests that the underlying mechanism of the synergy differs in each case and motivates further study of synergistic interactions of BiVO_4 with both $\text{Cu}_2\text{V}_2\text{O}_7$ -like and CuV_2O_6 -like phases.

Materials structure–property factorization provides a framework for identifying synergistic and deleterious phase interactions with respect to desired material properties. This framework is particularly needed when a given dataset includes many combinations of phases that exhibit properties of interest. Consequently, MSPF commences with phase mapping, in the present work using the recently developed Deep Reasoning

Networks. Our approach leverages the intuitive interpretability of matrix factorization, which aggregates the properties of each phase in a multi-phase sample under an assumption of non-interacting phases. The analysis of the data deviations from the matrix factorization model is insightful, revealing compositions and phase mixtures of greatest scientific interest since the associated properties cannot be explained by the “non-interacting” model. We also note that other non-linear methods could easily overfit the data, especially given the limited number of data samples (335), in addition to potentially being less interpretable. Applying this methodology to solar fuels photoanodes in the Bi–Cu–V oxide system reveals specific phase combinations with synergistic interactions that elevate the photoactivity and provide avenues for further photoelectrode development. These results demonstrate that the intuitive approach of adding small amounts of copper vanadates to the best known photoanode phase (BiVO_4) is far less effective than the strategy discovered in the present work, wherein the relatively low-performance copper vanadate photoanodes are dramatically improved upon the addition of a small amount of BiVO_4 . MSPF identifies the phase combinations that merit further investigation to reveal the underlying mechanism of performance enhancement.

METHODS

Physical vapor deposition (PVD) library synthesis

The four continuous composition libraries of Bi–Cu–V oxide system were synthesized by radio-frequency (RF) magnetron co-sputtering from elemental sources (Bi, Cu, and V) arranged symmetrically with respect to the 100-mm-diameter glass substrate with SnO_2/F conducting layer in a custom-designed combinatorial sputtering system³⁹, and post-deposition annealed in a conventional box oven at 550 °C in air for 3 h, followed by natural cooling to room temperature. The reactive working atmosphere composed of 4.8 mTorr of inert Ar gas and 1.2 mTorr of reactive O_2 gas, and the substrate was kept without intended heating. The RF powers on the elemental sources were adjusted for the relative deposition rates to obtain the designed composition in the substrate center, and the non-confocal geometry of sputter sources provided a continuous composition gradient on the order of 1 at.% mm^{-1} . These methods have been validated in prior work to produce metal oxide films—in particular oxides based on Bi, Cu, and V—that are suitable for photoelectrochemical evaluation^{31,40,41}.

Composition and structure measurements

The library compositions were obtained by x-ray fluorescence (XRF) measurements using an EDAX Orbis Micro-XRF system with an x-ray beam approximately 2 mm in diameter. The Bi L, V K, and Cu K XRF peak intensities (counts per second) were extracted from the Orbis software, then converted to the metal loadings (nmol cm^{-2}) and normalized compositions using the sensitivity factor for each element calibrated by commercial XRF standards (MicromatterTM).

The crystal structures of composition libraries were analyzed by x-ray diffraction (XRD) with a Bruker DISCOVER D8 diffractometer with Cu K α radiation from a Bruker I μ S source. The x-ray spot size was limited to a 2 mm length scale, over which the composition is constant to within approximately 1–2 at.%. Diffraction images were collected using a two-dimensional VANTEC-500 detector and integrated into one-dimensional patterns using DIFFRAC.SUITE™ EVA software.

Scanning Droplet Cell (SDC) photoelectrochemistry (PEC)

Scanning droplet cell experiments⁴² were performed on Bi-Cu-V oxide thin films from 4 composition libraries in the O₂-bubbled pH 9 sulfite-containing electrolyte. A 0.4 mm diameter optical fiber coupled to the SDC provided 388 nm wavelength LED illumination (3.2 ± 0.05 eV) combined with a white light (W 3500 K, Doric LEDC2-385/W35) such that toggling provided a large 388 nm peak and broad white light centered at 600 nm with a smaller peak at 450 nm. Current driver settings were 0.4 A for both sources, with white light irradiance at approximately 80 mW cm⁻². Illumination area was 1.13 mm² with sampling done on the same 4.5 mm grid as XRF and XRD measurements across the composition gradient.

The photocurrent measurements resulted from a linear potential sweep from 1.246 to 0.27 V vs RHE at 0.02 V s⁻¹ with illumination turned on for 1.344 s and turned off for 0.672 s, creating a series of illumination cycles from which the photocurrent was modeled as described previously³¹.

Electronic microscopy

Scanning electron microscopy (SEM) were obtained with a FEI Nova NanoSEM 450 microscope at 5 kV acceleration voltage and an Everhart-Thornley secondary electron detector.

A FEI DualBeam Focused Ion Beam/scanning electron microscope (FIB/SEM) was used to prepare transmission electron microscopy (TEM) sample, which was capped carbon/I-C prior to milling. TEM experiments were carried out in a FEI Tecnai Osiris FEG/TEM operated at 200 kV in bright-field (BF) and high-resolution (HR) TEM mode. The energy-dispersive X-ray spectroscopy (EDX) elemental mapping was acquired using Bruker Quantax. This characterization was performed by Eurofins EAG Precision TEM in Santa Clara, California.

Crystal structure phase mapping (DRNets)

DRNets for phase mapping were described previously²⁰ and used here on the union of XRD patterns from the 4 composition libraries. The 99 ICDD entries in Supplementary Table 1, curated using the quality rating to choose the entry to represent a respective formula unit and space group, were used to determine the presence of the 21 phases in Supplementary Table 2 in the dataset. Additionally, SnO₂ (ICDD 04-003-0649) was used to model the crystalline underlayer whose diffraction signal is present in each diffraction pattern but with variable intensity based on the attenuation of the photoanode sample. DRNets used candidate phase patterns normalized to a maximum intensity of 1. Supplementary Table 2 also provides the relative intensities of the 21 phases calculated from Eq. 5 of ref. ⁴³, which were used to renormalize the DRNets pattern concentrations to molar phase concentrations using Eq. (6) of ref. ⁴³.

Matrix factorization

Assuming a linear weighting of phase-specific properties in a mixed-phase signal, a signal A_i for sample i is modeled as the inner product of the vector of phase concentrations for sample i (H_i) and the matrix of representative signals or “basis patterns” for each phase (W). Given the set of samples with different phase mixtures, this becomes a matrix factorization problem where the phase concentration matrix H is fixed at the values obtained from phase mapping: $A \cong W \cdot H$. Given the PEC measurements A , the matrix W is solved by gradient descent using Tensorflow Keras with random initialization, RMSE loss, and RMSprop optimizer (learning rate 1e-4, momentum 0.9). The optimal solution was found after 1000 iterations.

DATA AVAILABILITY

Phase mapping and photocurrent data (inputs) and the basis patterns and modeled photocurrent (outputs) are available at <https://data.caltech.edu/records/1983> (<https://doi.org/10.22002/D1.1983>).

CODE AVAILABILITY

Source code is available at <https://data.caltech.edu/records/1983> (<https://doi.org/10.22002/D1.1983>).

Received: 3 June 2021; Accepted: 8 March 2022;
Published online: 05 April 2022

REFERENCES

1. DeCost, B. L. et al. Scientific AI in materials science: a path to a sustainable and scalable paradigm. *Mach. Learn.: Sci. Technol.* **1**, 033001 (2020).
2. Aykol, M. et al. The materials research platform: defining the requirements from user stories. *Matter* **1**, 1433–1438 (2019).
3. Alberi, K. et al. The 2019 materials by design roadmap. *J. Phys. D: Appl. Phys.* **52**, 013001 (2019).
4. Tabor, D. P. et al. Accelerating the discovery of materials for clean energy in the era of smart automation. *Nat. Rev. Mater.* **3**, 5–20 (2018).
5. Gomes, C. P., Selman, B. & Gregoire, J. M. Artificial intelligence for materials discovery. *MRS Bull.* **44**, 538–544 (2019).
6. Sha, W. et al. Artificial intelligence to power the future of materials science and engineering. *Adv. Intell. Syst.* **2**, 1900143 (2020).
7. Jensen, K. F., Coley, C. W. & Eyke, N. S. Autonomous discovery in the chemical sciences part I: progress. *Angew. Chem. Int. Ed.* <https://doi.org/10.1002/anie.201909987> (2019).
8. Agrawal, A. & Choudhary, A. Perspective: materials informatics and big data: realization of the “fourth paradigm” of science in materials science. *APL Mater.* **4**, 053208 (2016).
9. DeCost, B. L. & Holm, E. A. A computer vision approach for automated analysis and classification of microstructural image data. *Comput. Mater. Sci.* **110**, 126–133 (2015).
10. Stein, H. S., Guevarra, D., Newhouse, P. F., Soedarmadji, E. & Gregoire, J. M. Machine learning of optical properties of materials – predicting spectra from images and images from spectra. *Chem. Sci.* **10**, 47–55 (2018).
11. Agrawal, A. & Choudhary, A. Deep materials informatics: applications of deep learning in materials science. *MRS Commun.* **9**, 779–792 (2019).
12. Umehara, M. et al. Analyzing machine learning models to accelerate generation of fundamental materials insights. *npj Comput. Mater.* **5**, 1–9 (2019).
13. Vasudevan, R. K. et al. Materials science in the artificial intelligence age: high-throughput library generation, machine learning, and a pathway from correlations to the underpinning physics. *MRS Commun.* **9**, 821–838 (2019).
14. 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**, 053211 (2016).
15. 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).
16. Kusne, A. G. et al. On-the-fly machine-learning for high-throughput experiments: search for rare-earth-free permanent magnets. *Sci. Rep.* **4**, 6367 (2014).
17. Iwasaki, Y., Kusne, A. G. & Takeuchi, I. Comparison of dissimilarity measures for cluster analysis of X-ray diffraction data from combinatorial libraries. *npj Comput. Mater.* **3**, 4 (2017).
18. Gomes, C. P. et al. CRYSTAL: a multi-agent AI system for automated mapping of materials’ crystal structures. *MRS Commun.* **9**, 600–608 (2019).
19. Oviedo, F. et al. Fast and interpretable classification of small X-ray diffraction datasets using data augmentation and deep neural networks. *npj Comput. Mater.* **5**, 1–9 (2019).
20. Chen, D. et al. Deep reasoning networks for unsupervised pattern de-mixing with constraint reasoning. In: Daumé III, H. & Singh, A. (eds) *International Conference on Machine Learning* 1500–1509 (PMLR, 2020).
21. Chen, D. et al. Automating crystal-structure phase mapping by combining deep learning with constraint reasoning. *Nat. Mach. Intell.* **3**, 812–822 (2021).
22. Kusne, A. G. et al. On-the-fly closed-loop materials discovery via Bayesian active learning. *Nat. Comm.* **11**, 5966 (2020).
23. Zhou, L. et al. Successes and opportunities for discovery of metal oxide photoanodes for solar fuels generators. *ACS Energy Lett.* **5**, 1413–1421 (2020).
24. Lee, D. K., Lee, D., Lumley, M. A. & Choi, K.-S. Progress on ternary oxide-based photoanodes for use in photoelectrochemical cells for solar water splitting. *Chem. Soc. Rev.* **48**, 2126–2157 (2019).
25. Park, Y., McDonald, K. J. & Choi, K.-S. Progress in bismuth vanadate photoanodes for use in solar water oxidation. *Chem. Soc. Rev.* **42**, 2321–2337 (2013).
26. Liu, H., Nakamura, R. & Nakato, Y. Bismuth-copper vanadate BiCu₂VO₆ as a novel photocatalyst for efficient visible-light-driven oxygen evolution. *Chemphyschem* **6**, 2499–2502 (2005).

27. Ding, Z., Fu, Y., Xie, Z. & Li, Z. A polymeric complex method to nanocrystalline BiCu₂VO₆ with visible light photocatalytic activity. *Mater. Lett.* **65**, 460–463 (2011).
28. Nakabayashi, Y., Nishikawa, M. & Nosaka, Y. Fabrication of bismuth copper vanadate electrodes through feasible chemical solution method for visible light-induced water oxidation. *J. Appl. Electrochem.* **46**, 9–16 (2016).
29. Woodhouse, M. & Parkinson, B. A. Combinatorial approaches for the identification and optimization of oxide semiconductors for efficient solar photoelectrolysis. *Chem. Soc. Rev.* **38**, 197–210 (2009).
30. Kumari, S. et al. Structural and photoelectrochemical properties in the thin film system Cu–Fe–V–O and its ternary subsystems Fe–V–O and Cu–V–O. *J. Chem. Phys.* **153**, 014707 (2020).
31. Zhou, L. et al. Quaternary oxide photoanode discovery improves the spectral response and photovoltage of copper vanadates. *Matter* **3**, 1614–1630 (2020).
32. Zhang, Z. et al. Fermi level engineering of passivation and electron transport materials for p-type CuBi₂O₄ employing a high-throughput methodology. *Adv. Funct. Mater.* **30**, 2000948 (2020).
33. Abdi, F. F. & Berglund, S. P. Recent developments in complex metal oxide photoelectrodes. *J. Phys. D: Appl. Phys.* **50**, 193002–193002 (2017).
34. Guo, W. et al. Synthesis and characterization of CuV₂O₆ and Cu₂V₂O₇: two photoanode candidates for photoelectrochemical water oxidation. *J. Phys. Chem. C* **119**, 27220–27227 (2015).
35. Newhouse, P. F. et al. Solar fuel photoanodes prepared by inkjet printing of copper vanadates. *J. Mater. Chem. A* **4**, 7483–7494 (2016).
36. Song, A. et al. Evaluation of copper vanadate (β-Cu₂V₂O₇) as a photoanode material for photoelectrochemical water oxidation. *Chem. Mater.* **32**, 2408–2419 (2020).
37. Guo, W. et al. Facile growth of β-Cu₂V₂O₇ thin films and characterization for photoelectrochemical water oxidation. *Mater. Lett.* **258**, 126842 (2020).
38. Kim, M. et al. Electrospayed copper hexaaxodivanadate (CuV₂O₆) and pyrovanadate (Cu₂V₂O₇) photoanodes for efficient solar water splitting. *J. Alloys Compounds* **708**, 444–450 (2017).
39. Suram, S. K. et al. Combinatorial thin film composition mapping using three dimensional deposition profiles. *Rev. Sci. Instrum.* **86**, 033904 (2015).
40. Yan, Q. et al. Solar fuels photoanode materials discovery by integrating high-throughput theory and experiment. *Proc. Nat. Acad. Sci.* **114**, 3040–3043 (2017).
41. Zhou, L. et al. High throughput discovery of solar fuels photoanodes in the CuO–V₂O₅ system. *Adv. Ener. Mat.* **5**, 1500968 (2015).
42. Gregoire, J. M., Xiang, C., Liu, X., Marcin, M. & Jin, J. Scanning droplet cell for high throughput electrochemical and photoelectrochemical measurements. *Rev. Sci. Instrum.* **84**, 024102 (2013).
43. 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).

ACKNOWLEDGEMENTS

This study is based upon work supported by the U.S. Department of Energy, Office of Science, Office of Basic Energy Sciences, under Award DE-SC0020383. Experiments

were additionally supported by the Liquid Sunlight Alliance, which is supported by the U.S. Department of Energy, Office of Science, Office of Basic Energy Sciences, Fuels from Sunlight Hub under Award Number DE-SC0021266.

AUTHOR CONTRIBUTIONS

D.G. implemented the MSPF algorithm. D.G. and L.Z. processed and analyzed the experimental data. L.Z. M.R., and A.S. collected experimental data. D.C. and C.G. performed phase mapping with assistance from L.Z. and J.G.; D.G., L.Z., C.G., and J.G. wrote the manuscript. J.G. conceptualized and supervised the research.

COMPETING INTERESTS

The authors declare no competing interests.

ADDITIONAL INFORMATION

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41524-022-00747-1>.

Correspondence and requests for materials should be addressed to John M. Gregoire.

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

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



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) 2022