SCIENTIFIC DATA

OPEN

DATA DESCRIPTOR

Received: 9 October 2018
Accepted: 11 February 2019
Published online: 27 March 2019

# Synthesis, optical imaging, and absorption spectroscopy data for 179072 metal oxides

Helge S. Stein, Edwin Soedarmadji, Paul F. Newhouse, Dan Guevarra & John M. Gregoire

**Optical absorption spectroscopy is an important materials characterization for applications such as solar energy generation. This data descriptor describes the to date (Dec 2018) largest publicly available curated materials science dataset for near infrared to near UV (UV-Vis) light absorbance, composition and processing properties of metal oxides. By supplying the complete synthesis and processing history of each of the 179072 samples from 99965 unique compositions we believe the dataset will enable the community to develop predictive models for materials, such as prediction of optical properties based on composition and processing, and ultimately serve as a benchmark dataset for continued integration of machine learning in materials science. The dataset is also a resource for identifying materials composition and synthesis to attain specific optical properties.**

## Background & Summary

The availability of scientific database systems[1], fast measurement instruments[2] and network infrastructures enable scientists to assemble ultra large datasets that enable to go beyond the answering of some original research question and gain fundamentally new knowledge via learning on all data collected[3]. Currently, fields such as organic chemistry[4], drug design[5–7], ab-initio materials science[8], and biology gain rapid pace through the availability of large datasets that enable predictive machine learning models but experimental materials science lacks such ultra large datasets (with the notable exception of the High-throughput Experimental Materials Database - HTEM[1]) as different synthesis procedures, processing conditions and analyses effectively block the assembly due to prohibitive inconsistencies in the data across experimental runs. Within the Joint Center for Artificial Photosynthesis, exploration of metal oxides for solar fuels generation included high throughput synthesis and optical characterization with tracking of synthesis and processing parameters. The exploration of the chemical space offered by the periodic table was not randomly or systematically explored as compositions spaces were chosen based on specific research directions.

Recently we published an algorithm paper that allows us to predict UV-Vis data based on a sample image[9] via a neural net machine learning model that effectively hyper scales the low energy resolution RGB image to optical absorbance values at 220 energies between 1.32 to 3.2 eV. The herewith published dataset contains all images and spectra used for this model.

This dataset[10] will enable materials scientists to continue developing algorithms that build upon recent advances including finding embeddings for materials composition[11,12], predicting optical properties[9] from composition, linking experimental findings to theory databases[8,13], and extracting band gap energy from UV-Vis spectra[14,15].

By making the dataset available as a hdf5[16] container we aim to make the dataset more amendable for scientists who are not fluent in database query languages as all data is organized in tabular format where every entry corresponds to the same sample. In this manuscript we will give some background about how the dataset was acquired and is structured.

Joint Center for Artificial Photosynthesis, California Institute of Technology, Pasadena, California, 91125, USA. Correspondence and requests for materials should be addressed to J.M.G. (email: gregoire@caltech.edu)

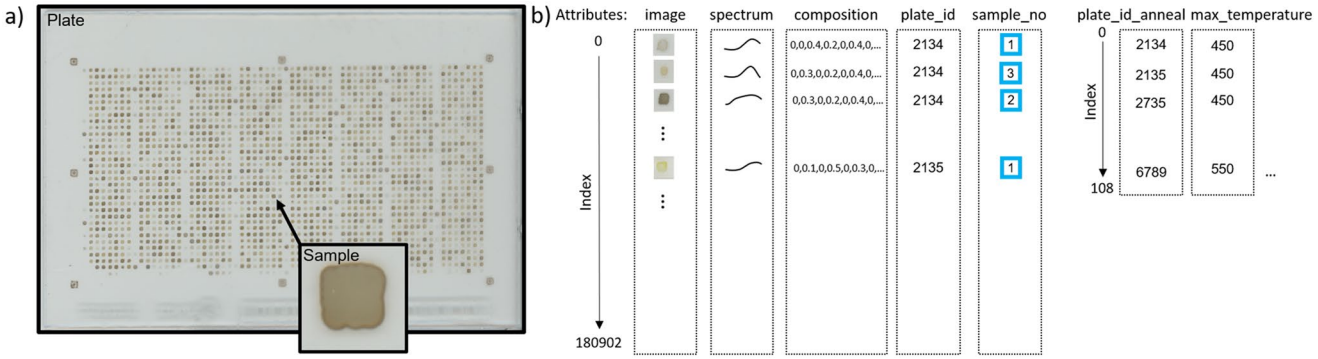


**Fig. 1** Data layout comparison between plate and data container. The logical layout is shown in (**a**) and the hdf5 container layout is shown in (**b**). Each plate can contain one or multiple composition spaces where each sample is uniquely defined by its sample ID, and plate ID pair. The logical layout is flattened in the hdf5 container such that all samples are placed along a single index.

## Methods

These methods are expanded versions of descriptions in our related work, which is referenced below for each technique. All samples in this dataset were synthesized via ink-jet printing of precursor salts with subsequent thermal processing to form metal oxides[17]. Mostly this synthesis involves printing metal nitrate salts on a glucose coated FTO/Glass substrate. The general assumption is that any chosen metal precursor salt, e.g. $Mn(NO_3)_2$, will thermally decompose under oxidizing conditions into a metal oxide, e.g. Mn oxide, via removal of the precursor's anion as a gas, e.g. $NO_2$. A typical thermal processing is annealing at 500 °C for 1 h in air or synthetic air. Some compositions, especially pure elemental oxides, are duplicated many times in the dataset, which can be readily identified via the composition table.

**Sample image generation.** All sample images were taken using a commercially-available consumer flatbed scanner (EPSON Perfection V600) in reflection configuration at 1200 dpi corresponding to a rate of 2.0 $cm^2\,s^{-1}$ or 0.019 s per sample as described elsewhere[18]. We assumed no lamp drift over time as the scanner is equipped with LED light sources. The scanner takes an images of a complete plate that is diced into 2.1 mm × 2.1 mm or 101 × 101 pixels with 24 bit color depth. Dicing of images was done semi automatically as scientists told the algorithm where fiducials for alignment were subsequent to scanning. To reduce the data size all images were rescaled to 64 × 64 pixels via the python image library (pillow) with anti-aliasing. Sample images typically have a colored region in the center corresponding to the printed material surrounded by grey area that is the background signal of the glass in the scanner bed. Some images appear darker at the edge of the printed material due to the so-called coffee ring that forms during drying of the printed solutions.

**UV-Vis spectra measurement.** All optical absorption spectra were measured using an on-the-fly scanning UV-Vis dual-sphere spectrometer as described elsewhere[19]. Since the spectral range over which the data was acquired varied, we interpolated on the smallest common energy range, 1.31 to 3.1 eV, which we discretize into 220 photon energies. We report fractional optical absorbance, which is the product of the absorption coefficient $\alpha$ and effective material thickness $L$, calculated via measurements of the fractional total reflectance R and total transmittance T:

$$\alpha L = -\ln \frac{T}{1-R}.$$

**Composition calculation.** All samples are labelled with their intended metals composition. Various quality control methods, which are not annotated in the dataset, were employed to omit samples whose composition is believed to differ from the intended composition. These methods include optical inspection and X-ray fluorescence measurements of the elemental loadings. The oxygen concentration results from thermal processing and is unknown. To enable researchers to study thickness effects of materials the loading as well as atomic fractions are reported. The total loading is the sum of loadings for each sample from which the atomic fractions were calculated. Loadings are calculated from ink concentration and known deposited volumes.

## Code Availability

Custom code for handling the dataset is available at https://github.com/helgestein/materials-images-spectra/. This python code enables users to easily download the dataset, pull specific or random images and accompanying spectra as well as processing and composition data. The code is intended to enable easy exploration of the dataset and to provide templates for use in machine learning models. The code requires python version 3.6.4 or higher with the following packages: h5py >= 2.7.1, numpy >= 1.15.2, tqdm >= 4.23.0.

| Dataset | Content Description | Data Range | Data Size | Physical Units | Method |
|---|---|---|---|---|---|
| Images | Sample images | 0–1 for every channel | (64,64,3,180902) | Color values for RGB | platebead scanner |
| spectra | fractional optical absorbance spectrum | 0–ca. 0.5 | (220,180902) | fractional absorb. coefficient | dual-sphere optical spectrometer |
| loadings | loading of each element | 0–1 | (43,180902) | nmol | calculated from loading and ink concentration |
| atfrac | Atomic fractions | 0–1 | (42,180902) | fractions | calculated from loadings |
| plate_id | Identifier index for plate | integer | (1,180902) | none | assigned |
| sample_id | Identifier index for each sample | integer | (1,180902) | none | assigned |
| energy_eV | Energy axis for spectra | float | (220,1) | Electron Volt (eV) | measured by spectrometer |
| loading_keys | Identifier index for loading element | String starting with Element | String list 180902 entries | Element names | assigned |
| atfrac_keys | Identifier index for loading element | String starting with Element | String list 180902 entries | Element names | assigned |
| substrate | Substrate used | string | String list 108 entries | none | assigned |
| plate_id_anneal | Maximum temperature during anneal | integer | (1,108) | none | assigned |
| max_temperature | Maximum temperature the plate was annealed at | float | (1,108) | Celcius | anneal recipe |
| soak_time_at_max_temperature | Time at maximum temperature | float | (1,108) | minutes | anneal recipe |
| nominal_pressure | Nominal pressure at maximum temperature | float | (1,108) | Torr | anneal recipe |
| gas_composition_string | Composition of the annealing gas | string | 108 Strings | none | anneal recipe |
| intended_element | Element intended to be added during anneal | string | 108 Strings | none | anneal recipe |

**Table 1.** Summary of all attributes in the hdf5 container accompanying this manuscript. All attributes contain arrays of the tuple shape given in the data size column.

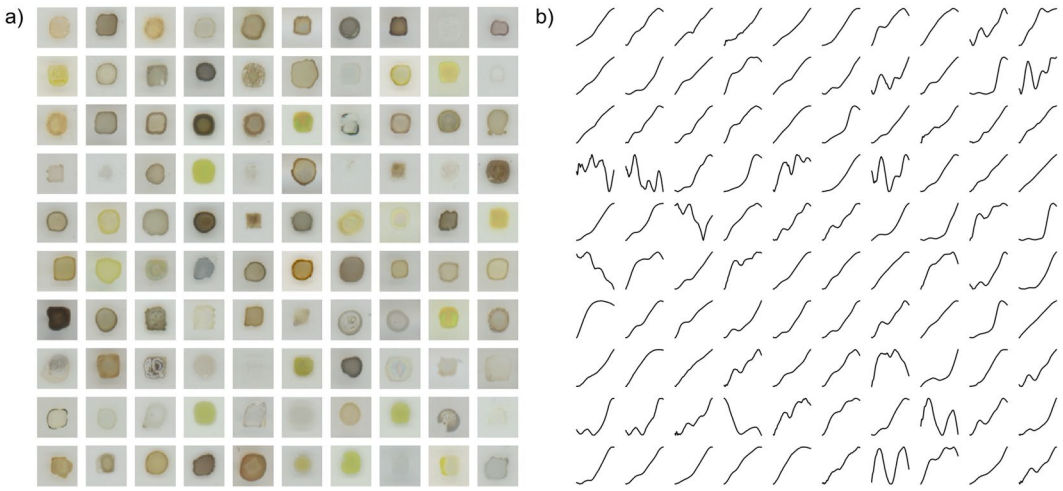


**Fig. 2** Comparison of materials images and their spectra. (**a**) Example images from the dataset with their corresponding (**b**) fractional optical absorbance spectra. The energy range for all spectra is 1.32 eV (left end) to 3.1 eV (right end).

## Data Records

During preparation of the hdf5 container we used the h5py library version 2.7.1 on a Windows 10 workstation. Images and spectra are compressed using the gzip option during creation of the file. The container has several attributes (see Fig. 1) that will be briefly described and are summarized in Table 1. The largest attribute in terms of data amount is the images that are $64 \times 64$ pixel containing each 3 colors corresponding to red, green, blue. All color values are floating point values between 0 and 1. In the spectrum dataset all spectra are placed in the same order as images. The composition of each sample is stored in the composition dataset as an array of concentrations for 42 elements in the dataset (most concentration values are zero). It should be noted that not all compositions sum to unity due to rounding error. The element labels (loadings and normalized atomic fractions) are stored separately as a string dataset in the "loadings" and "atfrac" datasets. The loading array contains 1 additional dimension for the total loading. Tracking indices for each library plate and each sample within a plate are stored in the correspondingly named attributes. Other information such as the anneal conditions are described in the last 5 rows of Table 1.

There are 180902 discrete samples, 1830 of which are "reference" samples where no material was deposited on the substrate, leaving 179072 materials samples. Due to duplication of compositions to enable exploration of different synthesis conditions, provide internal standards, and evaluate reproducibility, various compositions appear multiple times in the database, sometimes with variation in the synthesis conditions. Rounding to the nearest 1 at.% (although composition intervals are typically 5 at.%), there are 99965 unique compositions. The total number of plates is 108, each containing about 2000 samples.

## Technical Validation

Each sample in the dataset is part of a library plate that was visually inspected for printing quality during the materials synthesis phase. Detailed validation of the composition and other properties of individual samples have been performed on a small subset of the samples, with the only present availability of this data being journal publications describing specific libraries[14,18,20–22]. The array of materials in a library plate are indexed with sample location determined in each measurement using printed fiducials.

Standard data analysis software like the open source hdf5 library for python (https://www.h5py.org/) can read the container.

Example images and corresponding spectra are shown in Fig. 2.

## References


1. Zakutayev, A. *et al*. High Throughput Experimental Materials Database. https://doi.org/10.7799/1407128 (2017)
2. Hattrick-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**, 53211–53212 (2016).
3. Ramprasad, R., Batra, R., Pilania, G., Mannodi-Kanakkithodi, A. & Kim, C. Machine learning in materials informatics: recent applications and prospects. *Npj Comput. Mater* **3**, 54 (2017).
4. Gómez-Bombarelli, R. *et al*. Automatic chemical design using a data-driven continuous representation of molecules. *ACS Cent. Sci.* **4**, 268–276 (2018).
5. Campbell, S. J. *et al*. Visualizing the drug target landscape. *Drug Discov. Today* **15**, 3–15 (2010).
6. Jordan, A. M. & Roughley, S. D. Drug discovery chemistry: a primer for the non-specialist. *Drug Discov. Today* **14**, 731–744 (2009).
7. Yosipof, A. *et al*. Data mining and machine learning tools for combinatorial material science of all-oxide photovoltaic cells. *Mol. Inform.* **34**, 367–379 (2015).
8. Jain, A. *et al*. Commentary: The materials project: a materials genome approach to accelerating materials innovation (2013).
9. Stein, H. S., Guevarra, D., Newhouse, P. F., Soedarmadji, E. & Gregoire, J. Machine learning of optical properties of materials - predicting spectra from images and images from spectra. *Chem. Sci* **10**, 47–55 (2019).
10. Stein, H. S., Soedarmadji, E., Newhouse, P. F., Guevarra, D. & Gregoire, J. M. Synthesis, optical imaging, and absorption spectroscopy data for 179072 metal oxides. *figshare* https://doi.org/10.6084/m9.figshare.7502207 (2019).
11. Sołtys, M., Jaroszewicz, S. & Rzepakowski, P. Ensemble methods for uplift modeling. *Data Min. Knowl. Discov.* **29**, 1–29 (2015).
12. Ward, L. & Wolverton, C. Atomistic calculations and materials informatics: A review. *Curr. Opin. Solid State Amp Mater. Sci* **21**, 167–176 (2017).
13. Kirklin, S. *et al*. The Open Quantum Materials Database (OQMD): assessing the accuracy of DFT formation energies. *Npj Comput. Mater.* **1**, 15010 (2015).
14. Suram, S. K., Newhouse, P. F. & Gregoire, J. M. High throughput light absorber discovery, part 1: an algorithm for automated tauc analysis. *ACS Comb. Sci* **18**, 673–681 (2016).
15. Schwarting, M., Siol, S., Talley, K., Zakutayev, A. & Phillips, C. Automated algorithms for band gap analysis from optical absorption spectra. *Mater. Discov* **10**, 43–52 (2017).
16. Folk, M., Heber, G., Koziol, Q., Pourmal, E. & Robinson, D. An overview of the HDF5 technology suite and its applications. *Proceedings of the EDBT/ICDT 2011 Workshop on Array Databases - AD '11* 36–47 ACM Press, https://doi.org/10.1145/1966895.1966900 (2011).
17. Liu, X. *et al*. Inkjet Printing Assisted Synthesis of Multicomponent Mesoporous Metal Oxides for Ultrafast Catalyst Exploration. *Nano Lett.* **12**(11), 5733–5739 https://doi.org/10.1021/nl302992q (2012).
18. Mitrovic, S. *et al*. Colorimetric screening for high-throughput discovery of light absorbers. *ACS Comb. Sci.* **17**, 176–181 (2015).
19. Mitrovic, S. *et al*. High-throughput on-the-fly scanning ultraviolet-visible dual-sphere spectrometer. *Rev. Sci. Instrum.* **86**, 13904 (2015).
20. Newhouse, P. F. *et al*. Discovery and characterization of a pourbaix-stable, 1.8 eV direct gap bismuth manganate photoanode. *Chem. Mater.* **29**, 10027–10036 (2017).
21. Newhouse, P. F. *et al*. Solar fuel photoanodes prepared by inkjet printing of copper vanadates. *J. Mater. Chem. A* **4**, 7483–7494 (2016).
22. Newhouse, P. F. *et al*. Combinatorial alloying improves bismuth vanadate photoanodes via reduced monoclinic distortion. *Energy Environ. Sci.* **11**, 2444–2457 (2018).


## 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). We thank Kevin Kan for processing materials libraries.

## Author Contributions

H.S.S. and J.M.G. conceived the project and wrote the majority of code and manuscript. E.S. maintained the database backend and generated composition information. P.F.N. synthesized libraries and collected spectra. D.G. curated processing information and helped in generating the h5 container. J.M.G. supervised the research project.

## Additional Information

**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.