Machine learning of optical properties of materials - predicting spectra from images and images from spectra

Helge S. Stein,[a,*] Dan Guevara [a], Paul F. Newhouse [a], Edwin Soedarmadji [a], John M. Gregoire [a,*]

[a] *Joint Center for Artificial Photosynthesis, California Institute of Technology, Pasadena, California 91125 (USA)*
*stein@caltech.edu, gregoire@caltech.edu

As the materials science community seeks to capitalize on recent advancements in computer science, the sparsity of well-labelled experimental data and limited throughput by which it can be generated have inhibited deployment of machine learning algorithms to date. Several successful examples in computational chemistry have inspired further adoption of machine learning algorithms, and in the present work we present autoencoding algorithms for measured optical properties of metal oxides, which can serve as an exemplar for the breadth and depth of data required for modern algorithms to learn the underlying structure of experimental materials science data. Our set of 180,902 distinct materials samples spans 78 distinct composition spaces, includes 45 elements, and contains more than 80,000 unique quinary oxide and 67,000 unique quaternary oxide compositions, making it the largest and most diverse experimental materials set utilized in machine learning studies. The extensive dataset enabled training and validation of 3 distinct models for mapping between sample images and absorption spectra, including a conditional variational autoencoder that generates images of hypothetical materials with tailored absorption properties. The absorption patterns auto-generated from sample images capture the salient features of ground truth spectra, and direct band gap energies extracted from these auto-generated patterns are quite accurate with a mean absolute error of 240 meV, which is the approximate uncertainty from traditional extraction of the band gap energy from measurements of the full transmission and reflection spectra. Optical properties of materials are not only ubiquitous in materials applications but also emblematic of the confluence of underlying physical phenomena that yield the type of complex data relationships that merit and benefit from neural network-type modelling.

Introduction

Recent advances in computer science[1-4] enable materials scientists to identify new descriptors, predict new properties,[2] generate entirely new materials,[5] and identify reaction pathways.[6] Illustrative examples of predictive models in materials science include the prediction of optical and electrical properties based on representations of crystal structures as fragments[7,8] and the prediction of materials with complex electronic structures such as thermoelectrics[9] or organic light emitting diodes.[10] These successful implementations of modern machine learning algorithms are mostly limited to theoretical (i.e. computational) data, leaving an open question as to whether such algorithms can be impactful in materials science experiments. A primary hurdle to applying machine learning in experiments is the general lack of datasets that contain the appropriate diversity of materials and accompanying breath of metadata for training machine learning models. High-throughput materials science[11-16] can help address these data scarcity issues, as demonstrated by Zakutayev et al.[17] with their utilization of high throughput experiment data to train a random forest model that predicts electrical resistivity from material composition.

The common use of random forest models is understandable given their predictive power, but the lack of interpretability of this and other machine learning models limit their ability to generate new materials knowledge. Design of materials with tailored properties is central to materials research, and the machine learning-based acceleration of materials design was demonstrated by Gomez-Bombarelli et al.[5] through development of a conditional variational autoencoders (cVAE) that predicts new organic molecules based on user-specified properties. Variational autoencoders (VAE)[18] and cVAEs[19] utilize neural networks, whose deployment in materials science can enable new modes of scientific discovery through exploration of the latent space, which can reveal new and previously unknown relationships.[20] Our quest to develop models that learn the underlying structure of experimental materials data has resulted in the development of a VAE and cVAE to predict optical absorption spectra from images of materials, and images from user-tailored absorption spectra.

Our focus on optical characterization data is motivated by the importance of optical properties for a broad span of technologies, from computer displays to solar energy utilization.[21] The training and validation of models is enabled by extracting a large optical measurement dataset from the Joint Center for Artificial Photosynthesis database, in particular a set of measurements where synthesis and characterization were performed using the same set of instruments.[22,23] Optical characterizations utilizing inexpensive commercial sensors are particularly amenable to high throughput

experimentation, making optical characterization of new materials more expedient by experiment than by computation, particularly due to the high computational expense for predicting optical properties like band gap energy at reasonable accuracy; state of the art hybrid functionals require several CPU hours per material to achieve a bandgap prediction RMSE of 0.74 eV for metal oxide materials.[24] The recently reported machine learning model by Oses et al.[7] achieves an RMSE of 0.51 eV for computationally-predicted band gaps, which accelerates band gap prediction but not band gap measurement. Recently published algorithms have automated the extraction of band gap energy from an ultraviolet-visible (UV-Vis) optical absorption spectrum,[25,26] leaving spectrum acquisition as the rate limiting step of band gap measurement. As a result, the prediction of absorption spectra from a higher throughput experimental technique would be quite impactful, and we demonstrate machine learning automation of this task by combining a VAE with a deep neural network, requiring only an image of the material as input. We also exploit machine learning of the relationships between image and absorption spectrum to create a predictive model for the image of a material with tailored optical absorption properties, which is the first generative model[5,19] trained exclusively from experimental materials data.

Results and Discussion

Design and Training of Machine Learning Models

At a high level, imaging a material with a standard sensor, such as a red-green-blue (RGB) complementary metal oxide semiconductor (CMOS) sensor, is a spatially resolved measurement of an optical property averaged over some spectral range, including some spectral overlap of the 3 color filters.[27] The optical property being measured is an unknown combination of reflection, absorption and transmission properties, which is complementary to a spectral optical absorption measurement that averages over a sample region (lower spatial resolution) but uses spectrometers to attain high energy resolution. The spectral absorption technique also employs distinct transmission and reflection measurements from which the spectral absorption can be modelled. The inability to derive a first-principles transformation between these types of data arises from the unknown relationship between the RGB image and absorption, and unknown mapping from the broad spectral response of each CMOS channel to the high energy resolution of an absorption spectrum, which in the present case is 220 energies between 1.31 and 3.1 eV. Deriving such a mapping would be facilitated by a low-parameter functional form for how absorption varies with energy in metal oxide materials, but such a model is not forthcoming due to the various types of absorption phenomena and the mixing of absorption signals from multiple phases in the typically mixed-phase metal oxide samples. Consequently, machine learning of the underlying data relationships is the only viable option. Predicting absorption spectra from images is thus only possible if a machine learning algorithm can exploit “hidden” information in the high spatial-resolution images, i.e. data structure unbeknownst to expert materials scientists. Our exploration of the ability of machine learning to model complex relationships in materials data proceeds through the development of 3 models (Figure 1) with training and validation data extracted from a set of 181,129 images and spectra, including 1,908 “blank” samples (nothing deposited on the substrate) and 179,221 metal oxide samples synthesized via inkjet printing of mixed elemental precursors followed by thermal processing in an O_2-containing atmosphere. The metal oxides samples contain various combinations of 1 to 4 cation elements along with various inkjet printing and thermal processing parameters, which are not used in the models describe herein.

Model 1 - Variational Autoencoder

To establish the appropriate methods for encoding images of metal oxides, we commence with the design and training of model 1, an autoencoder for flatbed scanner images of materials synthesized by the inkjet printing technique. In addition to assessing models according to their cross-entropy loss in reconstruction of the test set, we also made qualitative evaluations of the behavior of encoding methods based on visual inspection. For example, we found that models employing convolutional layers[19] excel at reconstructing sample morphology but often failed to recover the human-perceived color of the material. The better color-preservation performance of models with fully connected layers led to our focus on the development of a variational autoencoder with fully connected layers.

Model 2 - Prediction of UV-Vis spectra

The Absorption Spectra Prediction model (ASPM) builds upon the VAE of model 1 to predict a UV-Vis absorption spectrum (220 entries) from a coordinate in the 100-dimensional latent space of the VAE. Under the assumption that the image encoder captures various image properties such as the color, color variation, morphology, etc., we exploit the high information density of the latent space (100 dimensions compared to the 12,288 dimensions of the 64×64 RGB

image) for the construction of absorption spectra, in this case using a deep neural network model that is trained independently from model 1.

Model 3 - Conditional Variational Autoencoder

The conditional Variational Autoencoder (cVAE) follows the structure of the VAE with modified inputs for both the encoding and decoding algorithms. The encoder input is the concatenation of the flattened image and absorption spectrum, and the decoder input is the concatenation of the latent space coordinate and the conditional absorption spectrum so that the resulting image represents the latent space coordinate under the condition that the material exhibits the specified conditional absorption spectrum. During training, the same absorption spectrum is used in the encoder and decoder inputs as noted in Figure 1, and during application of the model the conditional absorption spectrum is user-specified.

Image autoencoding and spectral prediction

The VAE of model 1 converges within 18 epochs as shown in Figure S3. Using t-distributed stochastic neighbor embedding (t-SNE)[28], the 100-dimensional latent space of the can be visualized as shown in Figure 2a) for the 54,270 images of test set, where each sample point is plotted using its representative color (see figure caption). Even though the VAE was not supplied any spectral information, it inherently exploits spectral features during autoencoding as is evident from the black-brown to blue-purple color gradient from left to right. The apparent clustering of samples, particularly those with a similar representative color, reveals aspects of the latent space structure, with the empty spaces between sample clusters indicating some structure of latent space coordinates.

Example raw (I_i) and VAE-reconstructed ($\widetilde{I_i}$) images are shown in Figure 3, demonstrating that the general appearance and especially the perceived color of the materials is well reconstructed. Any spatially-resolved variation in the raw image, such as color distribution within a printed blob, is not apparent in the reconstructed image due to blurring that occurs in image autoencoding with dimensionality of the latent space well below that of the images.[18,19] Since an absorption spectrum is measured with illumination of the entire sample, yielding the spatially "averaged" absorption signal, this blurriness of the reconstructed images is not important for the present purposes, but it is worth noting that the presence of a so-called coffee ring in I_i typically results in a darker edge of the sample blob in $\widetilde{I_i}$. The VAE preservation of perceived color (Figure 3) and color-based clustering in the latent space (Figure 2a) indicate that the VAE successfully encodes spectral features even though the model was not supplied any spectral information, motivating the use of this model for predicting spectral absorption.

Absorption Spectrum Prediction

The Absorption Spectra Prediction Model (ASPM) is trained validated using the VAE latent space coordinates, with the same train-test split used in model 1. The weights of the VAE are no longer trainable at this stage. Overall, there is good convergence for the ASPM across the energy range of the absorption spectra as shown in Figure 4. The residual density plot constitutes only small deviation between the predicted and ground truth signal, demonstrating the excellent absorption spectrum prediction from the VAE encoding of a material's image.

Detailed analysis of the absorption spectrum prediction for a span of representative samples is shown in Figure 5 that compares ground truth absorption spectra (green) from the test set and their prediction (black) from model 2. The figure includes a row of plots from each loss decile, with ten randomly selected samples in each row. For up to the 80th loss percentile, the predicted spectra appear in good agreement with the ground truth spectra. Impressively, the model reconstructs fine features of the absorption spectra such as local maxima that result from sub-band gap absorption or thin-film interference, even when these features occur over a spectral ranges much smaller than the sensitivity range of the original RGB sensor. Even an expert materials scientist cannot identify the presence of such features from inspection of an image, demonstrating the super-human analysis capabilities of the machine learning models.

The quality of the predicted UV-vis spectra allows their utilization for estimating band gap energy, which is typically a manual human analysis exercise but has recently been automated to identify a representative band gap energy for a given absorption spectrum.[25,26] As most of the herein studied materials are multiphase materials (due to their high computational order) it should be noted that the MARS algorithm employed here returns only a single representative

band gap energy without a measure of uncertainty. For sample i, the performance of the model 2 for band gap estimation is thus evaluated by comparing the MARS-identified direct band gap energy from the ground truth spectra S_i to that from the predicted spectra $\widetilde{S_i}$, as shown in Figure 6 for the test set. The mean band gap error is 74 meV (median 100 meV), mean squared error 96 meV^2 (RMSE 309 meV), and mean absolute error is 240 meV. The prediction of band gaps based on the latent space representation of images therefore outperforms the ab-initio calculations noted above by extracting knowledge from the coarse optical characterization data in the flatbed scanner images.

Conditional Variational Autoencoder

A complementary demonstration of the ability of machine learning models to encode materials properties is the development of a generative model that makes predictions of materials data from user-specified properties. For this purpose, the cVAE of model 3 is designed to predict how a printed sample should look like given a target absorption spectrum. From visual inspection of reconstructed images in Figure 3, the autoencoding of the cVAE is superior to the VAE of model 1, especially in terms of reconstructing the image color, which is primarily due to the use of the absorption spectrum both in the encoder input and as a condition in the latent space input to the decoder. Using the cVAE, coffee rings are more pronounced in the reconstruction, and the color of some samples appear more vibrant (e.g. the blue reconstructed image).

To generate conditional images, we arbitrarily chose a sample from the test set and identified its latent space coordinate $\widetilde{z_i}$. From this fixed point in the cVAE latent space, various tailored absorption spectra were applied as the conditional input to the decoder, resulting in cVAE-generated images of hypothetical materials as shown in Figure 7. To ascertain the sensitivity of the generated image to the starting latent space coordinate, Figure S3 shows a series of images using the conditional spectra from Figure 7 b) and 200 $\widetilde{z_i}$ values from randomly chosen samples. As expected, latent space coordinate impacts the apparent morphology of the material in the generated image but not its apparent color, which is primarily determined by the conditional absorption spectrum.

The series of sigmoidal absorption spectra shown in Figure 7a) spans a broad range of shapes by decreasing the inflection point energy (from top to bottom) and the slope (from left to right). This corresponds to a change in apparent band gap from about 1.36 to 2.9 eV according to the MARS model. To model different material thickness or maximum absorption coefficient, the sigmoid shapes are scaled to values of 0.84, 0.42, and 0.21 for image generation in Figures 7b, 7c, and 7d, respectively. The highest absorption factor measured in the test set was 0.75, making the generated images in Figure 7b an extension beyond the span of absorption spectra in the train set. Figure 7b) is commensurate with the general observation in materials science that a material with a high band gap should be yellowish-transparent (e.g. $BiVO_4$ with 2.5 eV band gap), a material with an intermediate band gap is red-brown (e.g. Fe_2O_3 with 2.2 eV band gap), and a materials with very low band gap appear blue-grey (e.g. Si with 1.2 eV band gap). The apparent transparency and saturation of the generated images is also quite intuitive as high absorption values and low sigmoid slopes that correspond to absorption of a broad spectral range lead to high opacity and low color saturation. With lower maximum absorption the center part of each image tends to become grayer, which is assumed to be the model's simulation of transparency given the gray appearance of the substrate/background in the flatbed scanner images. An interesting feature of image generator is that the blob size in images with a very high conditional absorption slope tend to be minimally bigger than those with a lower one. This is likely due to the network trying to match the condition by making the absorbing part of the image larger, an unintended but interesting mechanism by which the conditional spectrum impacts shapes in the generated image.

The ability to generate simulated data for a "coarse" measurement based on a desired fundamental property may enable rapid screening for desired materials, but more foundationally the cVAE of model demonstrates the successful training of a generative model using only experimental data. To provide an estimate of the data size required to train these types of models, models 1 and 2 were trained with 0.1%, 1%, 10%, and 50% of the train images and evaluated using a static test set, yielding RMSE values for bandgap prediction of 437, 400, 403, and 389 meV, respectfully (100% train set yielded 309 meV), corresponding to approximately 50 meV improvement in RMSE for every 10-fold increase in training set size (see SI for additional details). Given that the full training set produces an RMSE value comparable to the uncertainty in the band gap extraction algorithm, further enlarging the dataset would likely not be impactful, but the need for hundreds of thousands of samples to push this limit highlights the challenges of applying machine learning

techniques in materials science. Building more experimental materials databases of this size requires a revolution in data and metadata management. To date, computational materials datasets have been more amenable to machine learning due to relative ease in integration of data across research groups, whereas variations in experimental instruments and the lack of a framework to encode differences between instruments and experimental techniques limits assembly of large experimental databases. Consequently, the machine learning demonstrations in experimental materials science, namely the work by Zakutayev et al.[17] and the present work, utilize specific types of data acquired within a single research organization, and we believe these demonstrations lay the foundation for future generation of more broadly-applicable machine learning models[3,29,30] in experimental materials science.

Conclusion

Empowered by an unprecedented dataset of optical characterizations of metal oxide materials, we train a series of machine learning models employing convolutional and deep neural networks. A materials image autoencoder was developed by training a VAE using images of thin film materials acquired with a commercial flatbed scanner. The VAE, even though not trained with spectral information, encodes spectral characteristics in its information-rich latent space, enabling the development of a DNN model for predicting the full UV-Vis absorption spectrum of a material from only its image. Band gap energies extracted from the predicted spectra match the uncertainty from the extraction algorithm and supersedes common ab-initio methods for phase-pure materials. An additional model predicts the image of a hypothetical material based on its user-specified absorption pattern, providing the first example of a cVAE model trained exclusively of experimental materials data. This study has been enabled by the construction of a database of over 10^5 materials, demonstrating the utility of high throughput experiments with rigorous data management for further adoption of machine learning in experimental materials science.

Methods

Samples were synthesized using ink-jet printing of precursors salts, typically metal nitrates, that are subsequently annealed to form metal oxides.[31] Optical absorption spectra were recorded using an on-the-fly scanning UV-Vis dual-sphere spectrometer as described elsewhere.[22] Sample images were acquired using a commercially available flatbed scanner (EPSON Perfection V600) in reflection configuration as described elsewhere.[23] The scanner acquired 1200 dpi images are a rate of 2.0 $cm^2\ s^{-1}$, corresponding to 0.019 s per sample with our library design of approximately 1 mm^2 samples on a square grid with 2 mm pitch.

All calculations were performed on an Alienware Aurora R7 workstation equipped with an Intel i7-8700K@3.70 GHz CPU, 32 GB RAM, a Nvidia GTX1080Ti GPU with 12 GB dedicated GPU memory. Software used was Python version 3.6.4, Keras version 2.1.5, and TensorFlow version 1.1.0. The test-train split was 30% test, 70% train.

Machine Learning model descriptions

Model 1: The input images I_i were flattened, batch normalized and passed to a dense layer with 2048 output dimensions and *tanh* activation. The output of this layer is again batch normalized and passed to two layers $\boldsymbol{\mu}$ and σ with 100 output dimensions (length of latent space embedding) and linear activation. The output of these was passed to a sampling layer z that samples the latent space via:

$$z = \mu + a * \epsilon * e^{\sigma/2}$$

where ϵ is a random normal tensor of the same shape as $\boldsymbol{\mu}$ with zero mean and unit variance. During training the constant a is set to one, otherwise zero. The model until here is the Encoder E_{VAE} as shown in Figure 1. The output of this layer is fed to a Dense layer with 12288 output dimensions with sigmoid activation. The output of this layer is reshaped to match the dimensions of the input/output image with 64x64x3 pixels (12288 values). This transformation from latent space coordinate to sample image is the decoder, D_{VAE}. The model is trained using the Adam optimizer with early stopping monitoring the validation loss improvement over 3 epochs. The Loss is the sum of the Kullback-Leibler divergence and the binary cross entropy which is multiplied by the number of pixels in the output image (12288). The scaling of the binary cross entropy ensures good convergence of both the KL-loss and the image reconstruction as both are equally weighted during training. When the reconstruction loss was not weighted correctly the KL-loss converges but images are not reconstructed well. The model converges after 18 epochs. The complete VAE model has about 26 million trainable parameters of which 25 million are in the first dense layer that encodes the flattened images.

Model 2: The input to the first layer was batch normalized. The first layer was a fully-connected layer with 512 output dimensions, tanh activation, and 25% dropout. The first layer output is batch normalized an passed its output to a second fully-connected layer sequence identical to the first. The output of the second dense layer was again batch

normalized then passed to a dense layer with 220 output dimensions and sigmoid activation to predict the optical absorption spectra. The ASPM is trained using the Adam optimizer and MSE loss with early stopping monitoring the validation loss improvement over 50 epochs. Training of the ASPM is done after training the VAE, hence the VAE has not seen any spectral information.

Model 3: The conditional Variational Autoencoder (cVAE) followed the structure of the VAE except for the concatenation of the flattened image and absorption spectrum before the first dense layer and the concatenation of the output of the sampling layer z and the absorption spectrum. The spectra were not scaled or transformed.

Conflicts of interest

There are no conflicts to declare.

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).

Notes and references

1 R. Ramprasad, R. Batra, G. Pilania, A. Mannodi-Kanakkithodi and C. Kim, *npj Computational Materials*, 2017, **3**, 54.

2 L. Ward and C. Wolverton, *Current Opinion in Solid State & Materials Science*, 2017, **21**, 167–176.

3 S. K. Suram, M. Z. Pesenson and J. M. Gregoire, in *Information Science for Materials Discovery and Design*, eds. T. Lookman, F. J. Alexander and K. Rajan, Springer International Publishing, Chambridge, 2016, vol. 225, pp. 271–300.

4 K. Rajan, *Annu. Rev. Mater. Res.*, 2015, **45**, 153–169.

5 R. Gómez-Bombarelli, J. N. Wei, D. Duvenaud, J. M. Hernández-Lobato, B. Sánchez-Lengeling, D. Sheberla, J. Aguilera-Iparraguirre, T. D. Hirzel, R. P. Adams and A. Aspuru-Guzik, *ACS Central Science*, 2018, **4**, 268–276.

6 Z. W. Ulissi, A. J. Medford, T. Bligaard and J. K. Nørskov, *Nature Communications*, 2017, **8**, 14621.

7 C. Oses, C. Toher, E. Gossett, A. Tropsha, O. Isayev and S. Curtarolo, *Nature Communications*, 2017, **8**, 1–12.

8 K. T. Schütt, H. Glawe, F. Brockherde, A. Sanna, K. R. Müller and E. K. U. Gross, *Phys. Rev. B*, 2014, **89**, 1875.

9 J. Carrete, W. Li, N. Mingo, S. Wang and S. Curtarolo, *Phys. Rev. X*, 2014, **4**, 18.

10 R. Gómez-Bombarelli, J. Aguilera-Iparraguirre, T. D. Hirzel, D. Duvenaud, D. Maclaurin, M. A. Blood-Forsythe, H. S. Chae, M. Einzinger, D.-G. Ha, T. Wu, G. Markopoulos, S. Jeon, H. Kang, H. Miyazaki, M. Numata, S. Kim, W. Huang, S. I. Hong, M. Baldo, R. P. Adams and A. Aspuru-Guzik, *Nature Materials*, 2016, **15**, 1120–1127.

11 M. L. Green, C. L. Choi, J. R. Hattrick-Simpers, A. M. Joshi, I. Takeuchi, S. C. Barron, E. Campo, T. Chiang, S. Empedocles, J. M. Gregoire, A. G. Kusne, J. Martin, A. Mehta, K. Persson, Z. Trautt, J. Van Duren and A. Zakutayev, *Appl. Phys. Rev.*, 2017, **4**, 011105.

12 W. Setyawan and S. Curtarolo, *Comp. Mat. Sci.*, 2010, **49**, 299–312.

13 F. Ren, L. Ward, T. Williams, K. J. Laws, C. Wolverton, J. Hattrick-Simpers and A. Mehta, *Sci Adv*, 2018, **4**, eaaq1566.

14 A. Ludwig, R. Zarnetta and S. Hamann, *J. Mater. Chem. A*, 2008, **99**, 1144–1149.

15 M. Woodhouse and B. A. Parkinson, *Chem. Soc. Rev.*, 2008, **38**, 197–210.

16 2013, 1–54.

17 A. Zakutayev, J. Perkins, M. Schwarting, R. White, K. Munch, W. Tumas, N. Wunder and C. O. Phillips, 2017.

18 D. P. Kingma and M. W., *ICLR 2016*, **stat.ML**, 1312.6114v10.

19 A. Radford, L. Metz and S. C., *ICLR 2016*, 1511.06434v2.

20 E. O. Pyzer-Knapp, K. Li and A. Aspuru-Guzik, *Adv. Funct. Mater.*, 2015, **25**, 6495–6502.

21 H. Döscher, J. F. Geisz, T. G. Deutsch and J. A. Turner, *Energy & Environ. Sci.*, 2014, **7**, 2951–2956.

22 S. Mitrovic, E. W. Cornell, M. R. Marcin, R. J. R. Jones, P. F. Newhouse, S. K. Suram, J. Jin and J. M. Gregoire, *Rev. Sci. Instrum.*, 2015, **86**, 013904.

23 S. Mitrovic, E. Soedarmadji, P. F. Newhouse, S. K. Suram, J. A. Haber, J. Jin and J. M. Gregoire, *ACS Comb. Sci.*, 2015, **17**, 176–181.

24 Á. Morales-García, R. Valero and F. Illas, *J. Phys. Chem. C*, 2017, **121**, 18862–18866.

25 M. Schwarting, S. Siol, K. Talley, A. Zakutayev and C. Phillips, *Materials Discovery*, 2017, **10**, 43–52.

26 S. K. Suram, P. F. Newhouse and J. M. Gregoire, *ACS Comb. Sci.*, 2016, **18**, 673–681.

27 G. Agranov, V. Berezin and R. H. Tsai, *IEEE Transactions on Electron Devices*, 2003, **50**, 4–11.

28 L. van der Maaten, *J. of Mach. Learn. Res.*, 2014, **15**, 3221–3245.

29 Y. Xue, J. Bai, R. Le Bras, B. Rappazzo, R. B., J. Bjork, L. Longpre, S. Suram, R. B. van Dover, J. Gregoire and C. Gomes, *aaai.org,* **IAAI-17**, 4635–4642.

30 H. S. Stein, S. Jiao and A. Ludwig, *ACS Comb. Sci.*, 2017, **19**, 1–8.

31 X. Liu, Y. Shen, R. Yang, S. Zou, X. Ji, L. Shi, Y. Zhang, D. Liu, L. Xiao, X. Zheng, S. Li, J. Fan and G. D. Stucky, *Nano Lett.*, 2012, **12**, 5733–5739.

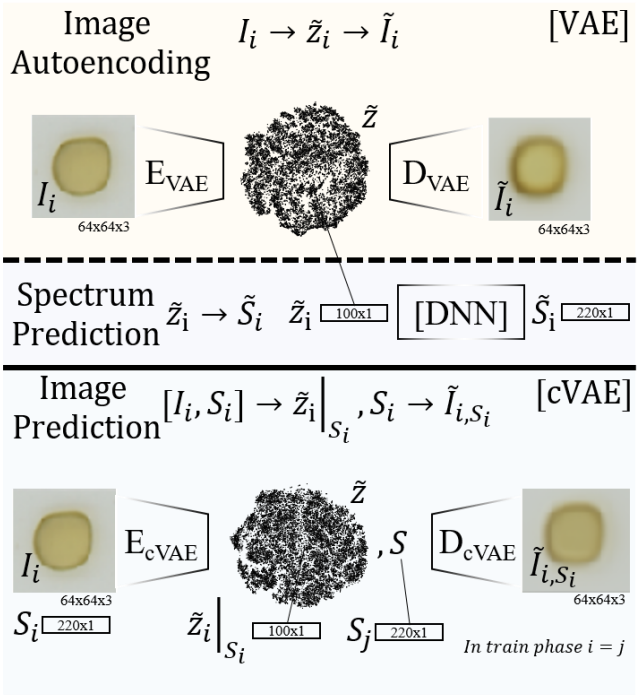

Figure 1: Schematic visualization of the 3 types of learning models for optical properties of materials. The first algorithm (top) illustrates the variational autoencoder (VAE) that autoencodes images $\tilde{I}_i$from I_i via a latent space representation $\tilde{z}_i$. The encoder performing the mapping I_i to $\tilde{z}_i$ is called E_{VAE}, the decoder performing the mapping from $\tilde{z}_i$ to $\tilde{I}_i$ is called D_{VAE}. The second model employs the latent space of the VAE but decodes $\tilde{z}_i$ into an absorption spectrum $\tilde{S}_i$ (instead of an image) using a deep neural net (DNN), producing an image to spectrum prediction model. The cVAE reconstructs images $\tilde{I}_i$ from images I_i and spectra S_i such that the latent space vector $\tilde{z}_i$ encodes image and spectral information, which is decoded in conjunction with a specific absorption spectrum S_i to yield an image $\tilde{I}_{i,S_i}$ that is predicted to exhibit the specified absorption properties.

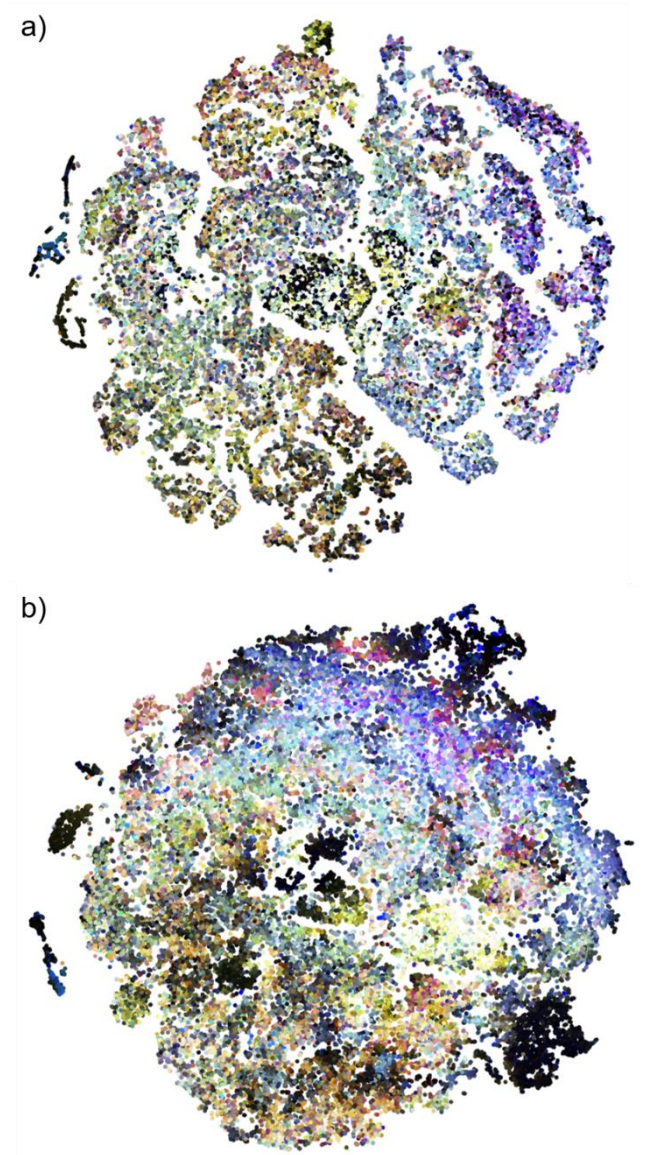

Figure 2: t-SNE visualization of the a) VAE, b) cVAE latent space (the cVAE latent space does not include conditional absorption spectra) of all 20165 images from the test set. Each image of a material appears as a point colored according to the quantile uniform transformed 1-alpha absorption at 1.48 eV (red), 1.7 eV (green) and 2.5 eV (blue). For example, points appearing white are mostly transparent and samples colored green absorb light particularly strongly at 1.7 eV. Although the VAE model was not supplied information about spectra, the apparent clustering of points by color is representative of the inherent structuring of the latent space with respect to optical absorption properties. Less color-specific clustering is observed in the cVAE since chromatic features of the optical absorption are additionally modelled by the latent space decoding that is conditional on an absorption pattern.

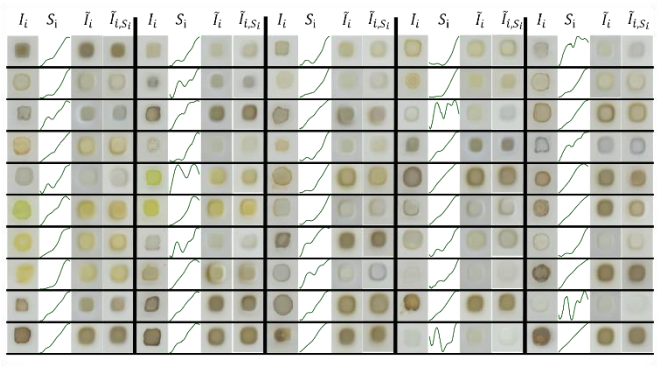

Figure 3: Reconstruction comparison from VAE and cVAE of randomly selected images from the test set. While both models successfully reconstruct the general color of the original image, the cVAE model performs better with respect to color preservation (by visual inspection), image sharpness, and aspects of the morphology (such as presence of a "coffee ring").

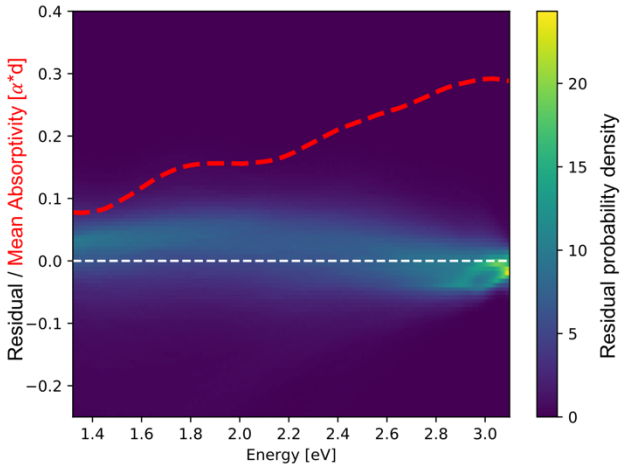

Figure 4: Plot of the mean predicted absorption spectrum ($mean\ \widetilde{S}_i$, red) residual probability density in the background that is calculated via a histogram at each energy. Predicted spectra are not thickness corrected.

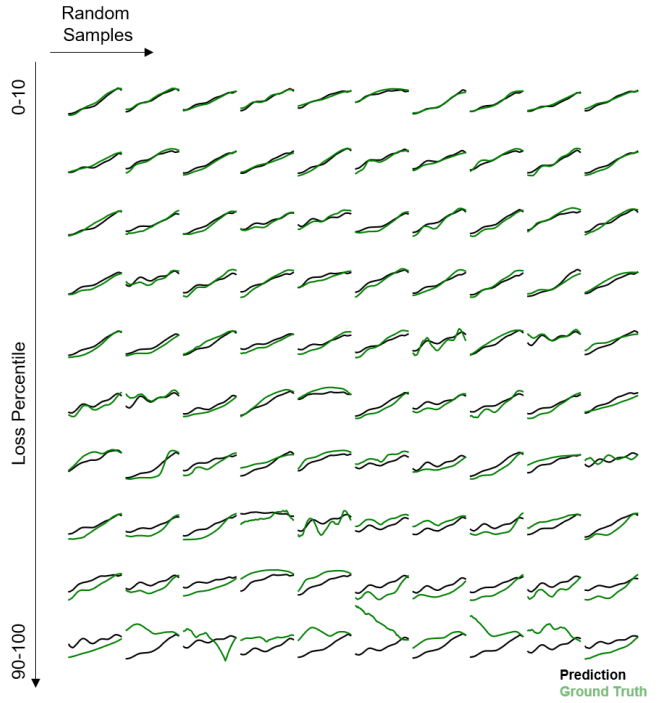

Figure 5: Randomly chosen ground truth (green) UV-Vis absorption spectra from the test set and those predicted (black) by model 2 using only the image of each material. Ten examples are provided from each of the MSE loss deciles (low loss reconstructions on top, high loss reconstructions on bottom). Most reconstructed patterns, in particular those below the 80th percentile of MSE loss, contain not only the general shape of the ground truth pattern but also finer details such as the presence of local maxima in absorption.

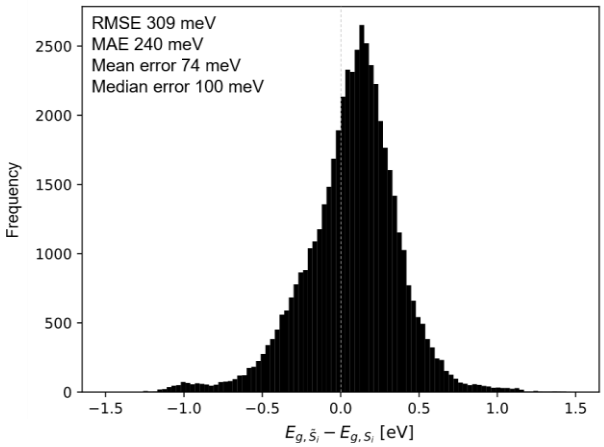

Figure 6: Difference between the band gap energy extracted from $\widetilde{S_i}$, which is predicted from model 2 using only the image of each material, and the experimentally measured Si. For both types of spectra, the band gap energy extraction was performed using the recently-reported MARS based segmentation model[22]. The mean error is 74 meV (median 100 meV), mean absolute error is 240 meV, root mean squared error is 309 meV.

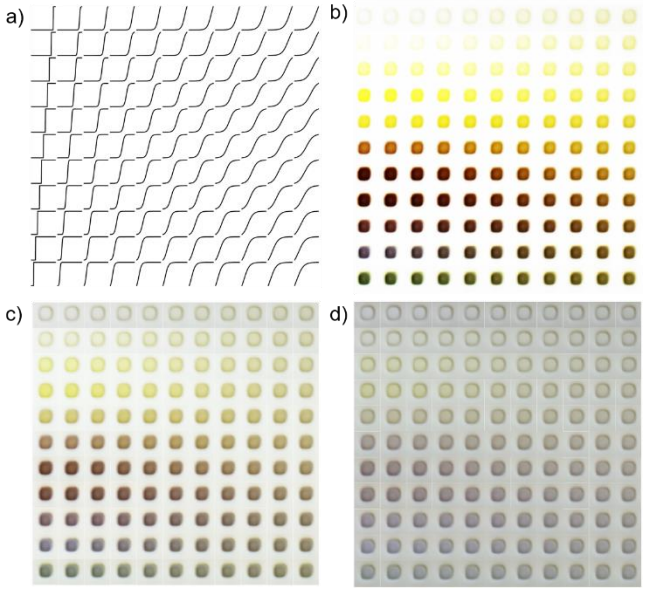

Figure 7: Demonstration of image prediction from an absorption spectrum using model 3 (cVAE). The array of synthetic absorption patterns are generated from the sigmoid function with transition energy increasing from bottom row (1.36 eV) to top (2.9 eV) row and transition width increasing from left column to right column as shown schematically in a). Each spectrum is additionally modified to obtain maximum absorption values of b)2.0, c) 0.75 d) 0.25 in the absorption spectrum S_i to simulate thicker (strongly absorbing) to thinner (weakly absorbing) materials. Each predicted material image originated from the same randomly chosen latent space coordinate (see Figures SX! for different latent space coordinates) and thus differ only by the respective conditional spectrum provided to the decoder D_{cVAE}. Conditionally predicted images based on a single latent space point using the conditional absorption spectra shown in a). Going down in the matrix corresponds to a lower band gap, going right to a lower slope of the sigmoid or less steep absorption factor. Given the gray appearance of the substrate (see Figures 3 and S1) samples with lower values in S_j appear more gray (less color saturation). Given the general rule of thumb of yellow-transparent high bandgap materials and brown-red-blue lower bandgap materials a reasonable prediction is achieved.