ACS APPLIED MATERIALS & INTERFACES

www.acsami.org

Research Article

Prediction of Frequency-Dependent Optical Spectrum for Solid Materials: A Multioutput and Multifidelity Machine Learning Approach

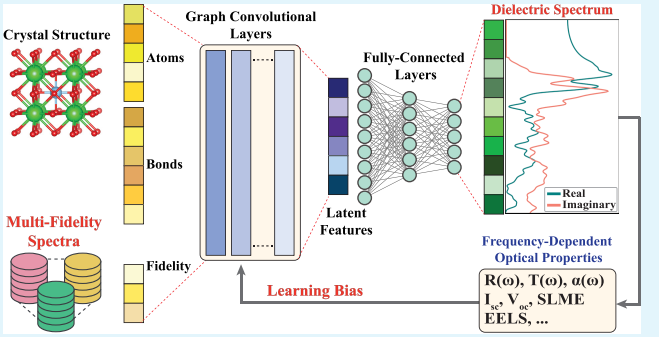
Akram Ibrahim and Can Ataca*

Cite This: https://doi.org/10.1021/acsami.4c07328

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: The frequency-dependent optical spectrum is pivotal for a broad range of applications from material characterization to optoelectronics and energy harvesting. Data-driven surrogate models, trained on density functional theory (DFT) data, have effectively alleviated the scalability limitations of DFT while preserving its chemical accuracy, expediting material discovery. However, prevailing machine learning (ML) efforts often focus on scalar properties such as the band gap, overlooking the complexities of optical spectra. In this work, we employ deep graph neural networks (GNNs) to predict the frequency-dependent complex-valued dielectric function across the infrared, visible, and ultraviolet spectra directly from the crystal structures. We explore multiple architectures for the spectral multioutput representation of the dielectric function and utilize various multifidelity learning strategies, such as transfer learning and fidelity embedding, to address the challenges associated with the scarcity of high-fidelity DFT data. Additionally, we model key solar cell absorption efficiency metrics, demonstrating that learning these parameters is enhanced when integrated through a learning bias within the learning of the frequency-dependent absorption coefficient. This study demonstrates that leveraging multioutput and multifidelity ML techniques enables accurate predictions of optical spectra from crystal structures, providing a versatile tool for rapidly screening materials for optoelectronics, optical sensing, and solar energy applications across an extensive frequency spectrum.



KEYWORDS: *graph neural networks, transfer learning, fidelity embedding, dielectric function, absorption coefficient, solar cells*

INTRODUCTION

Frequency-dependent optoelectronic properties provide essential insights critical for the design and optimization of a wide array of devices spanning various applications including photovoltaic (PV) cells,[1] light-emitting diodes,[2] transparent electronics,[3] optical sensors,[4] optical coatings,[5] chemical analysis,[6] and astrochemistry.[7] The capability to accurately and efficiently predict optical properties across a spectrum of frequencies is critical for integrating materials into cutting-edge optoelectronic devices. Computational approaches, mainly using DFT, can provide optical spectra with accuracy comparable to experiments more cost-effectively. Additionally, DFT optical spectra, generated with consistent calculation settings, can serve as benchmarks to identify influences beyond band-to-band transitions such as experimental setups or substrate effects. However, the vast array of candidate materials poses formidable computational challenges for DFT, necessitating the exploration of data-driven predictive models for preliminary screening.

Nevertheless, a gap persists in the literature concerning ML surrogate models capable of accurately predicting the frequency-dependent optical properties of solid materials. Previous studies have exclusively concentrated on predicting individual scalar properties, such as the band gap[8−10] and the static dielectric constant,[11,12] without accounting for the frequency dependence of optical properties. While the prediction of spectral properties has only recently emerged in materials science, multiple studies have explored multioutput learning for predicting the electronic and phononic density of states.[13−15] In the context of optical spectra, a hierarchical-correlation model was utilized to predict the absorption coefficient at different frequencies within the visible range, solely based on the chemical composition within a collection of 69 three-cation metal oxides.[16] Another study utilized a Gaussian process model to predict the dielectric constant of polymers using a data set of 1210 experimentally

Received: May 4, 2024
Revised: July 7, 2024
Accepted: July 15, 2024

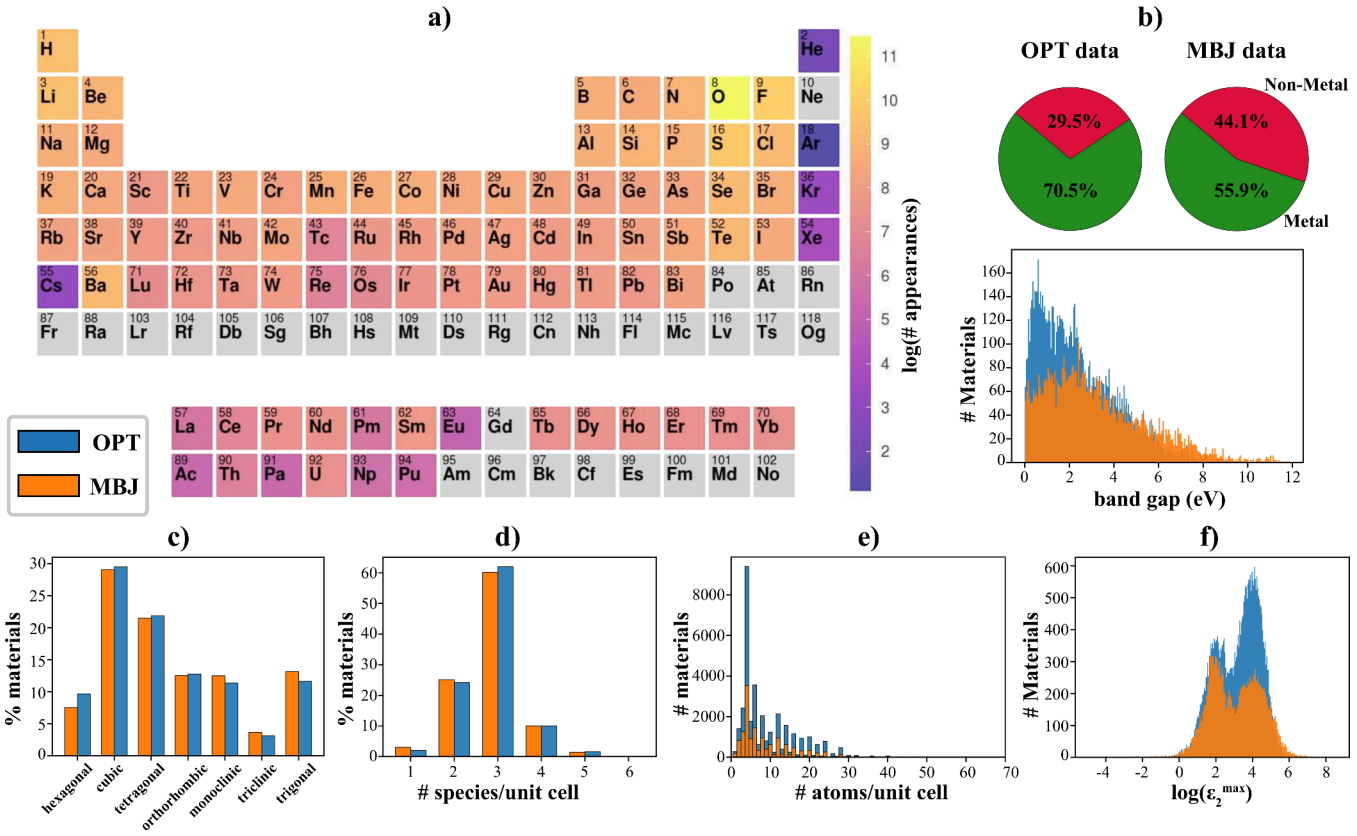


Figure 1. Summary of the frequency-dependent optical data set. (a) Heatmap illustrating the presence of the periodic table elements in the combined crystal structures of the OPT and MBJ data sets. Elements not present in the data set are marked in gray. (b) Proportions of metals versus nonmetals and distribution of band gap values for nonmetals in both data sets. (c–e) Distributions of crystal systems, number of distinct species in the unit cell, and atom count in the unit cell. (f) Histogram distribution of the logarithm of the imaginary dielectric function peak value.

measured values at different frequencies.[17,17] These studies, however, were confined to specific material chemical spaces, employing composition-based features while excluding crystal structure and constrained their predictions to particular discrete frequencies. To the best of our knowledge, no published work has yet utilized ML to directly predict the continuous, frequency-dependent dielectric function or absorption coefficient across a general chemical space of materials based on the crystal structure.

Early endeavors in predicting material properties relied on manual feature engineering from composition, crystal structure, and electronic band structure using featurization algorithms.[18] Currently, state-of-the-art predictive modeling of materials utilizes GNNs,[19−21] which adeptly generate latent feature representations from composition and structure, enabling automatic learning of features specific to the target property. In this work, we use GNNs to predict the frequency-dependent, complex-valued dielectric function of solid materials directly from crystal structure data. The dielectric function, a fundamental spectral output from ab initio calculations, determines the material's response to electromagnetic waves. It also enables the calculation of crucial practical frequency-dependent optical properties, such as the refractive index, electron energy loss spectra,[22] quality factors for localized surface plasmon resonances and surface plasmon polaritons,[23] and the quantum efficiency of optical sensors and PV cells.[24]

For material spectral properties such as phonon or electronic density of states, the full-energy density of occupied states is characterized by a known integral for each material, attributed to its atom or electron count. This facilitates modeling the spectrum as a probability distribution, simplifying learning by correlating increases in intensity in one range with decreases in another. However, the optical spectrum lacks this property, and the magnitude of the optical response can vary significantly among materials. Yet scaling the optical spectrum can still offer a way to establish a correlation within the predicted output. We find that proper spectrum scaling can lead to improved organization within the latent feature space, subsequently enhancing the model's performance.

Furthermore, while training ML models on high-fidelity data yields more accurate results, securing a sufficient volume of such data for effective training presents a notable challenge. A viable solution is adopting multifidelity learning frameworks that integrate data from both low-fidelity and high-fidelity sources. The larger data sets employing low-fidelity DFT functionals can enhance GNN models' ability to learn better encodings of crystal structures, consequently boosting performance in learning high-fidelity data. In this study, we investigate two multifidelity learning approaches: "transfer learning" and "fidelity embedding",[25−27] demonstrating that multifidelity learning effectively addresses the bottleneck caused by the scarcity of high-fidelity optical spectra.

Finally, we assess the potential to enhance the prediction accuracy of physical features of interest alongside the overall spectrum prediction by incorporating a physical learning bias during training. Focusing on PV cell absorption metrics (short-circuit current, reverse saturation current, and the spectro-

https://doi.org/10.1021/acsami.4c07328
ACS Appl. Mater. Interfaces XXXX, XXX, XXX−XXX

scopic limit of maximum efficiency (SLME), which represents the theoretical maximum photoconversion efficiency of a single p−n junction PV cell), we demonstrate that learning these properties within the context of frequency-dependent absorption coefficient learning through learning biases is more effective than directly learning them as standalone target properties.

METHODS

Data Set. The data set was obtained from the JARVIS-DFT Vienna ab initio simulation package (VASP) raw files on Figshare as of January 2024, encompassing all publicly available optical spectra data.[28−30] It comprised 34 327 calculations employing the OptB88vdW (OPT) functional and 14 560 calculations utilizing the meta-GGA modified Becke−Johnson (MBJ) potential,[31−33] following data cleansing. The complex dielectric tensor, $\varepsilon(E) = \varepsilon_1(E) + i\varepsilon_2(E)$, is provided for each material in the data set across a 5000-point energy grid that spans the entire range between the minimum and maximum Kohn−Sham eigenvalues. To ensure uniformity in the energy values at which $\varepsilon(E)$ is evaluated for all materials within the data set, we employ cubic interpolation and then uniformly extract the function values within the 0−12.0 eV range with a resolution of 0.04 eV. This energy range encompasses the infrared (IR), visible, and ultraviolet (UV) spectral regions. Notably, the employed model frameworks can be readily adapted to cover broader or different spectral ranges.

The $\varepsilon(E)$ output from VASP is represented by a 3 × 3 tensor computed for the primitive unit cell of each material. To simplify the problem, we diagonalize $\varepsilon(E)$, yielding three eigenvalues, which correspond to eigenvectors oriented along the principal crystallographic axes. Our GNN models are then trained using the mean of these eigenvalues. It is important to note that predicting the optical spectra along a specific symmetry axis follows the same formalism. For example, in the context of van der Waals layered materials, the model can be trained using either the eigenvalues of the in-plane or the out-of-plane axis. To streamline learning, we utilize the imaginary part of the dielectric function ($\varepsilon_2(E)$) and a reduced form of the absorption coefficient, rather than the real part ($\varepsilon_1(E)$), to model the complex optical spectra. This choice is made due to commonalities shared by the imaginary part and the absorption coefficient, such as their non-negativity and zero values for energies within the band gap. The reduced adsorption coefficient is defined as

$$\alpha(E) = -\varepsilon_1(E) + \sqrt{\varepsilon_1^2(E) + \varepsilon_2^2(E)} \quad (1)$$

The total absorption coefficient is given by $2\sqrt{2}\pi(E/hc)\sqrt{\alpha(E)}$, where h and c represent Planck's constant and the speed of light in vacuum, respectively. Notably, the real part can be readily derived from both the imaginary part and the absorption coefficient. Figure 1 illustrates a statistical distribution of all materials within the extracted OPT and MBJ data sets, categorized according to the frequency of elements they contain, band gap values, crystal systems, diversity of constituent species, number of atoms in the unit cell, and ε_2 peak magnitudes. More details about the data set preparation are provided in the Supporting Information.

Graph Neural Network Formalism. Figure 2a illustrates a general architectural overview of the employed GNN models. We utilize the MEGNet graph convolutional layers proposed by C. Chen et al.[19] Initially, each crystal structure is transformed into a graph characterized by node features, edge features, and global features, corresponding to atoms, bonds, and the overall state of the structure, respectively. For atom (node) features, we use only the atomic numbers of the constituent elements, which are then mapped to an embedding layer to learn elemental embeddings. Bond (edge) features are expressed through expanding interatomic bond distances using a Gaussian basis with 180 centers, each 0.5 Å wide, uniformly distributed between bonded atoms. Atoms are considered bonded if their interatomic distance is less than a cutoff radius of 5.5 Å, which encompasses not only the nearest neighbors but also interactions with more distant atoms. For state (global) features in fidelity-embedding GNN models, which are jointly trained on multifidelity data, the fidelity level of the DFT functional (OPT or MBJ) associated with each data point is represented by an integer (0 or 1). This is followed by a trainable embedding layer utilized to learn encodings for the fidelity levels. In contrast, for single-fidelity (SF) learning (where the model is trained on data from a single DFT functional) or in transfer learning (where the model is trained on multifidelity data sequentially) only two placeholder nodes, without embedding, are employed to facilitate global information exchange.

The input features undergo a preprocessing step through dense layers before being forwarded to the consecutive graph convolutional layers. These convolutional layers execute a sequence of update operations through convolution and pooling layers, transforming an input graph $G = (\mathbf{e}, \mathbf{v}, \text{and } \mathbf{u})$ into an output graph $G' = (\mathbf{e}', \mathbf{v}', \text{and } \mathbf{u}')$, where $\mathbf{e}$, $\mathbf{v}$, and $\mathbf{u}$ represent the edge, vertex, and global features, respectively. A stack, comprising three repetitions of the preprocessing dense layers and graph convolutional layers as shown in Figure 2a), is utilized to enhance model flexibility and indirectly access information beyond the 5.5 Å cutoff radius, enabling the model to capture more intricate long-range interactions.[19] A dropout layer follows the final graph convolution layer to mitigate overfitting. Rather than padding the structure graphs to uniformize the sizes of their atomic features, they are assembled into a single, large disjoint union of graphs for training. In the final stages of the model, a readout operation is performed on both the atom and bond feature vectors using an order-invariant set2set model.[19,34] The set2set layer combines atom and bond feature vectors with vectors denoting the indices of these atoms and bonds within the disjoint union graph. After the readout process, the atom, bond, and state feature vectors are concatenated to form the latent feature vector (LFV), which is subsequently processed through a series of dense layers to generate the multioutput prediction representing the discretized optical spectrum over the considered frequency range. Further details concerning the GNN architecture, hyperparameters, and MEGNet graph convolutional layers are available in the Supporting Information.

Model Training. The model construction and training were executed though employing the Keras API with the TensorFlow backend.[35,36] Our data set was partitioned into three segments, with 80% allocated for training, 5% for validation, and 15% for testing. The selection of hyperparameters, including the size of atom and bond features, dimensions of hidden layers, batch size, and dropout rate, was meticulously chosen through Bayesian hyperparameter optimization facilitated by Optuna[37] (see the Supporting Information). The best-performing models were selected based on their performance on the validation set and subsequently evaluated using the test set.

Training on multifidelity data can be approached through various methods. Figure 2b presents a schematic of the two multifidelity frameworks considered in this study: (1) transfer learning (TL) and (2) fidelity embedding (FE). In the TL framework, multifidelity learning progresses sequentially: the GNN model is initially trained on all of the OPT data, and then the LFV is obtained. Following this, the set of dense layers after the LFV is further optimized to accommodate additional layers and neurons and then trained on the MBJ data set using an 80/5/15 train/validation/test split. This means that the weights of the layers before the LFV remain frozen and only the later dense layers of the GNN are trained on the high-fidelity MBJ data. In the FE framework, multifidelity learning is approached jointly. The fidelity level (MBJ or OPT) for each data point is encoded as an integer and input into the GNN model through a trainable fidelity-embedding matrix, serving as the input state feature vector. Optimal validation results for ε_2 and α were achieved with fidelity-embedding vector lengths of 20 and 16, respectively. The 80/5/15 train/validation/test split applies exclusively to the MBJ data set, while the entire OPT data set is used for training. Further details regarding the architecture of the TL and FE GNNs are available in the Supporting Information.

https://doi.org/10.1021/acsami.4c07328
ACS Appl. Mater. Interfaces XXXX, XXX, XXX−XXX

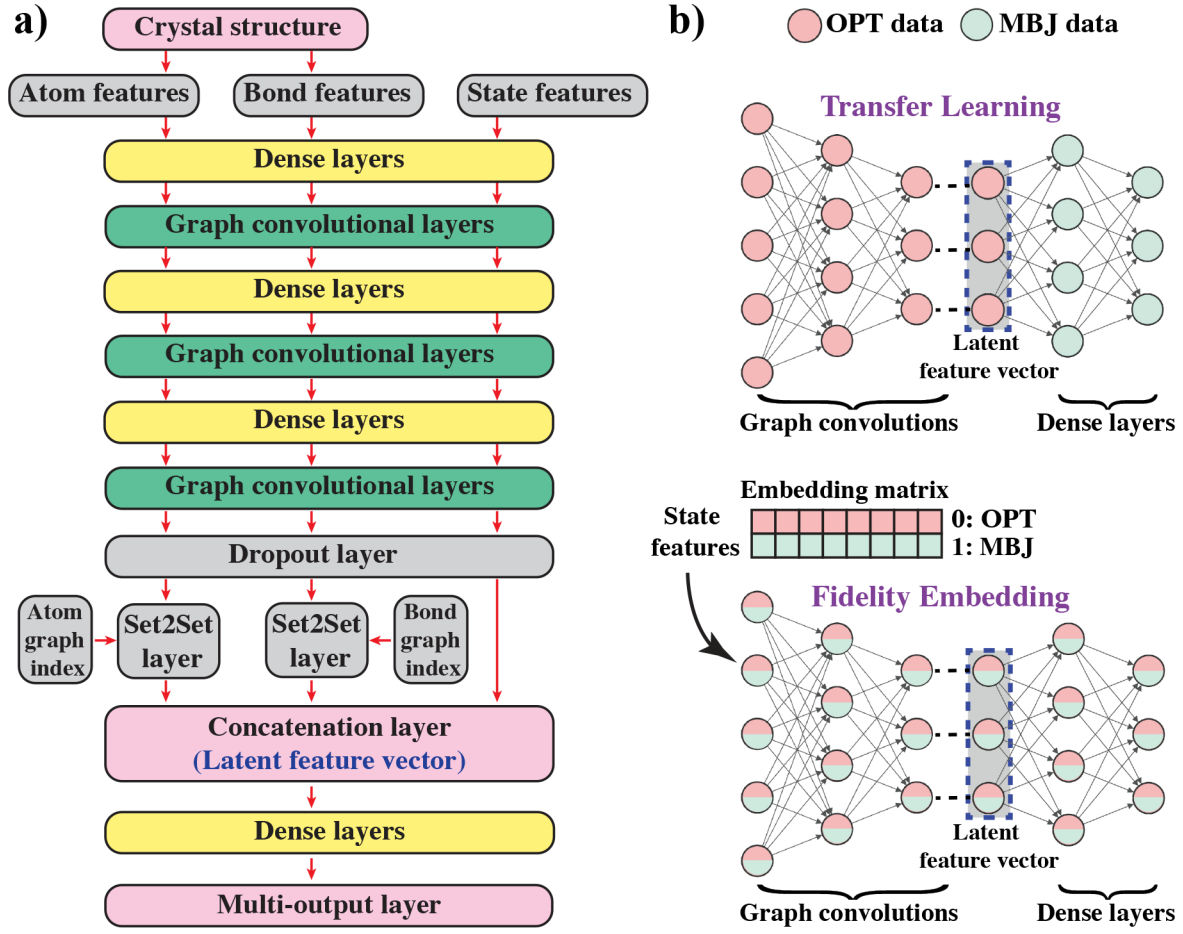


Figure 2. Overview of the GNN architecture and multifidelity learning frameworks. (a) Schematic of the GNN architecture, illustrating input structures with embedded atomic numbers, Gaussian-expanded bond distances, and state features. The multioutput layer outputs the predicted discretized optical spectrum over the considered frequency range. (b) Diagram of the multifidelity learning frameworks. Transfer learning sequentially trains on the OPT data across the entire GNN, then employs the learned LFV as input to the dense layers for subsequent MBJ data learning while freezing the graph convolutional layers. Fidelity-embedding jointly learns the OPT and MBJ data throughout the entire GNN, using a trainable embedding matrix as the input state feature to encode the DFT functional fidelity level (MBJ or OPT) for each crystal structure.

RESULTS AND DISCUSSION

Spectrum Multioutput Architecture. In this section, we evaluate the performance of different multioutput GNN architectures for representing the optical spectrum, which involve various combinations of data scaling schemes and training loss functions. All GNN models discussed in this section are solely trained and evaluated by utilizing MBJ data. Since we have no ground truth regarding a physical scaling feature, such as electron/atom number for the electronic/phonon density of states, we explore both scaling to a maximum value of 1 (MaxNorm) and normalization to a cumulative sum of 1 (AvgNorm), alongside the unnormalized spectrum (UnNorm).[14,15] For the MaxNorm and AvgNorm models, the loss function is formulated as $\mathcal{L} = \mathcal{L}_N + w\mathcal{L}_S$, where $\mathcal{L}_N$ pertains to the error in the norm, $\mathcal{L}_S$ to the error in the normalized spectrum, and w is a hyperparameter that denotes the relative weight of the two loss components during training, and its optimal value is determined through a grid search process on the validation set (see Supporting Information). The mean absolute error (MAE) loss function is utilized for the UnNorm model, for $\mathcal{L}_N$ in both MaxNorm and AvgNorm models and for $\mathcal{L}_S$ in the MaxNorm model. In the AvgNorm model, where the normalized spectrum represents a probability density function (PDF), we experiment with training $\mathcal{L}_S$ using two distinct loss functions: MAE and the Kullback–Leibler (KL) divergence loss.[38] The four models {UnNorm, MaxNorm, AvgNorm (KL), and AvgNorm (MAE)} share the same architecture, differing only in the output layer and/or the loss function. In the UnNorm model, the output layer comprises a dense layer of 300 neurons (representing the considered 12 eV range at 0.04 eV resolution) and features a rectified linear unit (ReLU) activation. For the remaining three models, the output layer consists of 301 neurons, including an additional neuron representing the norm with a ReLU activation. For MaxNorm, a sigmoid activation function is employed for the normalized spectrum (S), specifically applied to $(S - S_{max}/2)$. For the two AvgNorm models, the normalized spectrum utilizes a softmax activation function applied to $(S - S_{max})$. The radar plots in Figure 3a) summarize the performance of the four models on the ε_2 and α MBJ test sets. The benchmarking metrics cover various indicators to evaluate performance across different facets of the spectrum. These metrics include, in order, the overall MAE and Pearson correlation for the unnormalized spectrum (ε_2 and α), MAE for the maximum and average values, along with MAE, KL divergence, Wasserstein distance (WD),[39] and first derivative MAE for the PDF-normalized

https://doi.org/10.1021/acsami.4c07328
ACS Appl. Mater. Interfaces XXXX, XXX, XXX−XXX

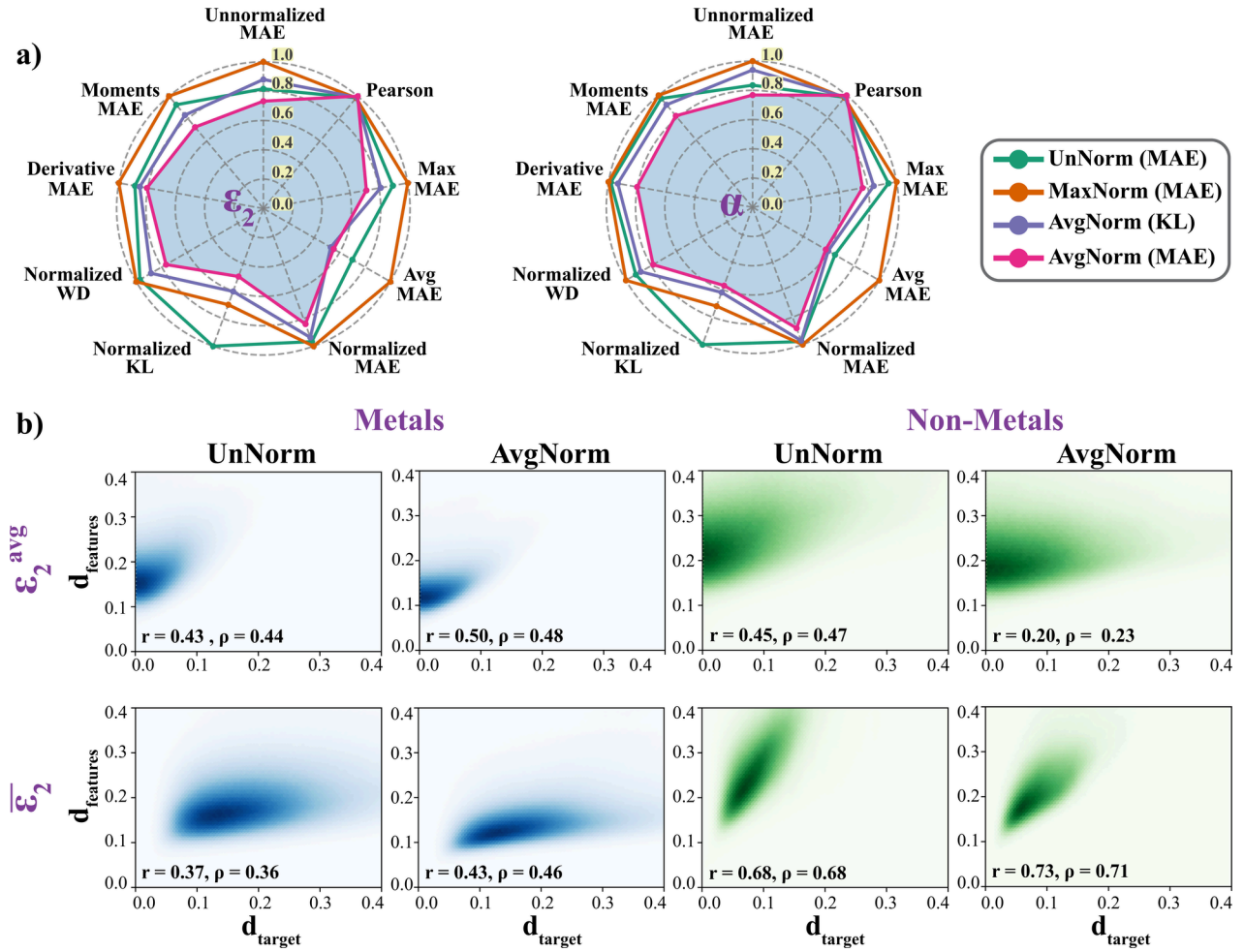


Figure 3. Effect of the multioutput architecture on model performance. (a) Radar charts illustrating the median errors, normalized to a maximum of one, for the four considered multioutput GNN architectures. These GNN models differ in their spectrum scaling method and/or the loss function. The shown performance evaluations are conducted using the MBJ test set, detailing various metrics for both the imaginary part (ε_2) and the reduced absorption coefficient (α). Numeric values of the median error and the IQR are outlined in Tables S1 and S2. (b) Pearson (r) and Spearman (ρ) correlation coefficients between the normalized Euclidean pairwise distances of ε_2 targets (d_{target}) (which include average norms ($\varepsilon_2^{\text{avg}}$) and PDF-normalized spectra ($\overline{\varepsilon}_2$)) and the corresponding normalized Euclidean pairwise distances of latent feature vectors (d_{features}) are depicted for both metals and nonmetals in the whole MBJ data set.

spectrum ($\overline{\varepsilon}_2$ and $\overline{\alpha}$). Additionally, we define another metric for aggregated statistical moments, for instance, for $\overline{\varepsilon}_2$, as $\sum_n (E^n \cdot \overline{\varepsilon}_2)^{(1/n)}$, encompassing the first four moments. Numeric values of the radar plots, including uncertainties represented by the interquartile range (IQR), are detailed in Tables S1 and S2.

Figure 3a shows that the AvgNorm (MAE) model consistently outperforms the other three models across all metrics. This superiority can be attributed to several factors. First, the model benefits from extracting the average norm and transforming the optical spectrum into a PDF. Figure 1f indicates that the peak values in ε_2 span approximately 6 orders of magnitude in our data set. Thus, this scaling approach enables the model to effectively capture trends in the normalized spectra, thereby preventing materials with high average values from disproportionately influencing the training process. Second, the utilization of the softmax activation function boosts the model's performance by facilitating improved learning of the spectral distribution by constraining the sum of predicted values of the normalized spectrum to unity. This constraint introduces a correlation along the predicted spectrum, where a high probability in some range of the spectrum automatically leads to a decrease in the probability in other ranges. Furthermore, the exponential function in softmax activation responds to lower stimulations (present in the lower regions of the normalized spectrum) with a more uniform distribution while exponentially amplifying higher stimulations, such as peaks and near-peak regions of the normalized spectrum. This mechanism ensures that large probabilities predominate while still incorporating information from lower probabilities within the spectrum.

Training the PDF using KL divergence loss is observed to yield a similar performance, albeit slightly lower than when employing MAE loss, which surprisingly results in a lower KL error on the test set compared to training directly with KL divergence loss. This can be attributed to KL divergence quantifying discrepancies between PDFs logarithmically, helping to enhance sensitivity to low-probability areas but simultaneously reducing it in high-probability regions due to the logarithmic function's dampening effect on large values. Thus, KL divergence may overlook finer details around peak areas of the optical PDF.

https://doi.org/10.1021/acsami.4c07328
ACS Appl. Mater. Interfaces XXXX, XXX, XXX−XXX

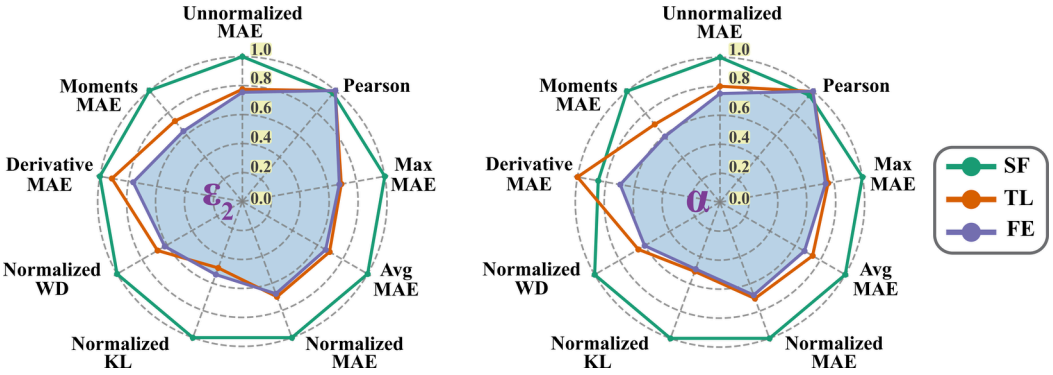


Figure 4. Multifidelity learning. Radar charts depicting the median errors, normalized to a maximum of one, for three GNNs: the single-fidelity (SF) model (trained exclusively on MBJ data), the transfer learning (TL) model (initially trained on all OPT data, followed by further training of the dense layers post-LFV on 80% of MBJ data), and the fidelity-embedding (FE) model (trained jointly on all OPT data and 80% of MBJ data). The performance evaluations shown here are conducted using the higher-fidelity MBJ data test set, detailing various metrics for both the imaginary part (ε_2) and the reduced absorption coefficient (α). Numeric values of the median error and the IQR are outlined in Tables S3 and S4.

We also experimented with the WD loss for the PDF-normalized spectrum within the AvgNorm architecture but observed that the predicted spectra became impractically noisy. This likely stems from WD penalizing errors in the cumulative distribution function (CDF) rather than the PDF,[40] leading to precise resonant peak positioning but significant noise in regions such as the baseline (ranges of no interaction with electromagnetic fields) or peak tails. This occurs because the CDF loss permits fluctuations between underestimation and overestimation at successive spectral points, resulting in a lower CDF error comparable to a prediction that consistently under- or overestimates. Given the necessity for accurate, noise-free spectral predictions for analyzing the dielectric function, WD was deemed unsuitable as a standalone loss function for the normalized spectrum and was therefore excluded from our analysis.

On the other hand, extracting the maximum value and normalizing the spectrum to a maximum of one result in inferior performance compared to the unnormalized case. This decline can be partly attributed to the use of the sigmoid function with the MaxNorm model. Although the sigmoid function offers a smooth exponential form that confines the normalized spectrum within the range of 0−1, unlike softmax, it does not impose constraints that can correlate the spectrum points. The sigmoid activation for $(S - S_{max}/2)$ treats values around the half-maximum almost linearly but rapidly saturates values deviating upward from the half-maximum to 1 and values deviating downward from the half-maximum to zero. This saturation effect reduces the resolution between points with higher values in the spectrum, thereby contributing to the observed decrease in performance. We also conducted experiments in which the application of the sigmoid function in the MaxNorm architecture was omitted, opting instead for a linear activation function with clamping between 0 and 1. However, the resulting output did not yield spectra that were deemed to be plausibly smooth. Further discussion providing more insights into the reasons behind the observed effects of spectrum scaling on GNN performance is discussed in the "Latent Space Visualization" section.

The efficacy of the AvgNorm model architecture implies that representing the optical spectrum of arbitrary crystal structures through an average norm, alongside a PDF, learned via uniform error penalization without weighting any part of the distribution more heavily than another while using softmax activation enhances the model's capacity to discern fundamental patterns in optical spectra across diverse materials. Consequently, the AvgNorm (MAE) architecture is utilized in subsequent analyses. Notably, the ramifications of this scaling extend beyond mere postprocessing of the output, profoundly influencing the arrangement of materials in the latent feature space. This phenomenon is illustrated in Figure 3b), where we detail the distribution of normalized Euclidean pairwise distances for ε_2 targets, including average norms (ε_2^{avg}) and PDF-normalized spectra ($\bar{\varepsilon}_2$), versus those of the LFVs across all materials in the MBJ data set, correlating them using both Pearson (r) and Spearman (ρ) correlation coefficients. In the case of metals, the AvgNorm model is observed to increase r between features and both ε_2^{avg} and $\bar{\varepsilon}_2$ by about 16%. Conversely, for nonmetals, while the AvgNorm model increases r between features and $\bar{\varepsilon}_2$ by around 7%, it concurrently significantly reduces the correlation with ε_2^{avg} by about 56%. These alterations in the correlation of features with targets underscore the profound influence of optical spectrum scaling on the structural organization of the latent feature space. In essence, the AvgNorm model orchestrates a rearrangement of the materials within the latent space, fostering proximity among materials exhibiting similar PDF-normalized spectra and simultaneously boosting/attenuating the arrangement based on the average norms for metals/nonmetals. Further discussion on the organization of materials in the latent space is presented in the "Latent Space Visualization" section.

Learning from Multifidelity Data. Ideally, the ML model training should rely on experimental or high-fidelity computational data. However, given the inherent scarcity of such high-fidelity data, practical limitations necessitate leveraging the ample low-fidelity data to enhance predictive power through improved feature learning.

The radar plots in Figure 4 illustrate the performance metrics for the SF GNN, which is trained solely on MBJ data, as well as for the TL and FE GNNs for both ε_2 and α (numeric values of median error and IQR are detailed in Tables S3 and S4). All three models employ the AvgNorm (MAE) architecture discussed in the previous section. The results indicate a noticeable decrease across all error metrics and an increase in correlation with the DFT (MBJ) spectra upon the incorporation of the lower-fidelity OPT data. Notably, the FE model demonstrates a higher improvement, with the median

https://doi.org/10.1021/acsami.4c07328
ACS Appl. Mater. Interfaces XXXX, XXX, XXX−XXX

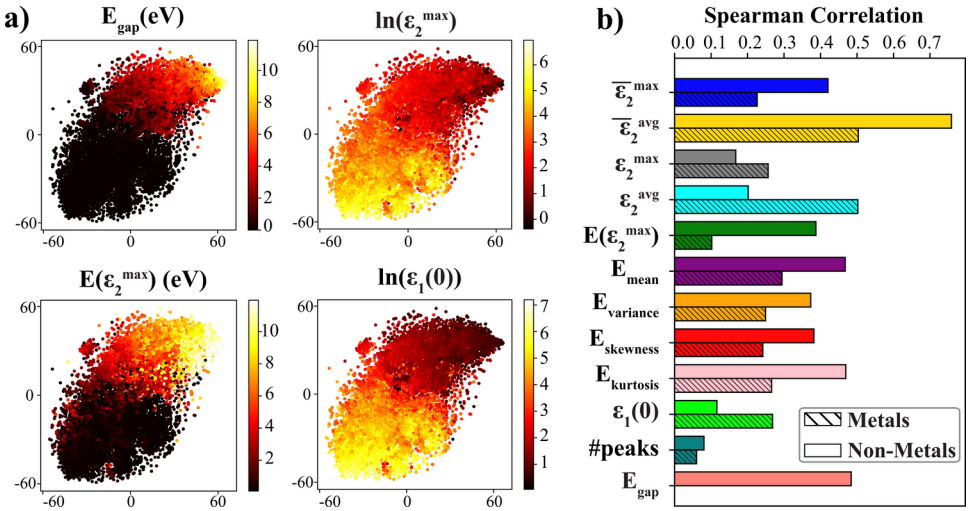


Figure 5. Significance of optical spectrum properties in organizing the latent feature space of crystal structures. (a) Two-dimensional t-SNE projection of the latent features of all the structures in the MBJ data set (perplexity = 150). The color maps represent DFT (MBJ) values for various scalar properties of the optical spectrum. The latent feature vectors are derived from the AvgNorm FE model optimized for predicting ε_2. (b) Spearman correlation coefficients between the normalized Euclidean pairwise distances of latent feature vectors and the corresponding normalized Euclidean pairwise distances of different properties extracted from the optical spectrum for both metals and nonmetals.

MAE of the unnormalized spectrum decreasing by 24.6% and 25.0% for ε_2 and α, respectively, compared to reductions of 22.5% and 20.0% for TL. Both FE and TL exhibit comparable improvements in the Pearson correlation, with ε_2 increases of 2.5% for FE versus 2.3% for TL and with α increases of 3.9% for FE versus 3.7% for TL. The superior performance of the FE framework can be attributed to its joint learning strategy, which simultaneously integrates low- and high-fidelity data during training, in contrast to the sequential learning approach utilized in TL. In TL, only the dense layers post-LFV can detect the nuanced differences between the OPT and MBJ data. Conversely, FE empowers the entire GNN to fine-tune its weights for fitting both OPT and MBJ data, thus achieving a broader optimization scope. Moreover, while the larger size of the OPT data set enables the TL model to glean a more robust LFV compared to SF models trained solely on the smaller-sized MBJ data, this approach inherently restricts the LFV learning to patterns of the lower-fidelity data set, potentially overlooking insights that could be gained from holistic learning involving both data fidelities simultaneously. Therefore, the joint learning strategy of the FE framework enables the GNN to exploit the most extensive data set resulting from the amalgamation of OPT and MBJ data sets, thereby accessing a broader range of information and achieving a more refined LFV compared to TL.

Latent Space Visualization. In essence, the enhanced performance of a GNN model in predicting material properties implies an improved ordering of materials within the latent feature space. While this concept is relatively straightforward in single-target prediction scenarios, where an improved latent space representation should manifest a more correlated ordering of latent feature vectors with the target scalar, complexities emerge in spectral multioutput prediction problems. In such cases, numerous scalar and vectorial features derived from the spectrum can potentially organize the latent space. The pertinent question then becomes which feature holds greater significance in organizing the latent space.

To gain deeper insights into the latent feature space, Figure 5a illustrates a two-dimensional projection via t-distributed stochastic neighbor embedding (t-SNE)[41] of the latent features of all structures within the MBJ data set, accompanied by heatmaps for certain physical scalar properties derived from the optical spectrum. The latent feature vectors are derived from the best-performing AvgNorm FE GNN model optimized for predicting ε_2. Further quantitative insights into the latent feature space are obtained by calculating the Spearman rank-order correlation between the normalized Euclidean pairwise distances of latent features and the corresponding normalized Euclidean pairwise distances of various target properties extracted from the optical spectrum (0−12 eV range), as illustrated in Figure 5b).

The band gap (E_{gap}) heatmap t-SNE plot in Figure 5a) reveals that the GNN model proficiently segregates metals from nonmetals into two distinctly discernible clusters within the latent space, thereby showcasing the model's adept comprehension of the distinctive optical characteristics of these two material categories. Moreover, a prominent gradient in E_{gap} is evident within the nonmetal cluster, capturing the diversity in band gaps among semiconductors and insulators. This is demonstrated by a Spearman correlation of ρ = 0.48 between E_{gap} and the latent features, as illustrated in Figure 5b). In a similar vein, the model exhibits structured organization within the latent space regarding additional dielectric properties, such as the logarithm of the peak value of the imaginary dielectric function, denoted by $\ln(\varepsilon_2^{max})$, and the corresponding energy at this peak, denoted as $E(\varepsilon_2^{max})$. A pronounced gradient in $\ln(\varepsilon_2^{max})$ is evident within the metal cluster, characterized by markedly elevated peaks at its boundary, distant from the interface between metal and nonmetal clusters. This pattern reveals that metals distant from the metal/nonmetal cluster interface display elevated optical conductivity peaks, which diminish progressively toward the interface, running counter to the band gap gradient direction within the nonmetal cluster; this is consistent with physical expectations, as one would expect optical conductivity to exhibit patterns that are opposite to those of the band gap. Conversely, within the nonmetal cluster, a less pronounced gradient is observed for $\ln(\varepsilon_2^{max})$. This can be attributed to the

https://doi.org/10.1021/acsami.4c07328
ACS Appl. Mater. Interfaces XXXX, XXX, XXX−XXX

fact that the imaginary dielectric function of semiconductors and insulators exhibits a more complex dependence on band structure, reflecting the intricate probabilities of valence to conduction interband transitions. In contrast, the simpler intraband transitions of free electrons in metals, particularly at lower energies, can be effectively characterized using bulk properties such as electron density and scattering rate, as described by the Drude model.[42] Therefore, the optical spectra of nonmetals are more challenging to represent with simple dielectric magnitudes. This is highlighted by the lower correlation coefficients of both maximum and average values (ε_2^{max} and ε_2^{avg}) for nonmetals compared to metals, as demonstrated in Figure 5b), where nonmetals exhibit correlations of 0.17 and 0.20, compared to 0.26 and 0.50 for metals.

Furthermore, a gradient is evident in $E(\varepsilon_2^{max})$, with metals generally manifesting lower values as indicated in Figure 5a), indicative of their lower natural frequencies due to the presence of delocalized electrons. Conversely, semiconductors and insulators tend to exhibit notably higher $E(\varepsilon_2^{max})$ values owing to the increased energies required to excite their tightly bound valence electrons to the conduction band, necessitating the overcoming of the band gap before observing the peak. Interestingly, within the metal cluster, the latent space of the imaginary part displays a noticeable gradient in the logarithm of the static dielectric constant, $\ln(\varepsilon_1(0))$, which is related to the real part. This suggests that the GNN effectively captures the physical linkage between the imaginary and real parts of the dielectric function via the Kramers−Kronig relation.[43] In contrast, a less pronounced gradient is observed among nonmetals, again highlighting their more intricate dependence on band structures rather than simple dielectric magnitudes. This is reflected by a lower $\varepsilon_1(0)$ correlation value of 0.12 for nonmetals compared to 0.27 for metals, as shown in Figure 5b.

To address our question regarding which feature from the optical spectrum is most influential in structuring the latent space, an examination of Figure 5b indicates that for nonmetals, the average-normalized spectrum ($\overline{\varepsilon}_2^{avg}$), interpreted as a PDF, emerges as the property most strongly correlated with the latent features, exhibiting a high ρ of 0.76, compared to the other evaluated properties, none of which surpass 0.48. Notably, the pronounced correlation of $\overline{\varepsilon}_2^{avg}$ is not a consequence of training the FE GNN model with the AvgNorm architecture. Evidence supporting this observation is illustrated in Figure S6, which shows $\overline{\varepsilon}_2^{avg}$ maintaining its status as the most correlated property for nonmetals across all SF models, regardless of the employed output scaling architecture. Thus, the PDF-normalized $\overline{\varepsilon}_2^{avg}$ can be deemed as a fundamental descriptor for learning the optical spectra of semiconductors and insulators, reinforcing our earlier observations regarding the superior performance of the AvgNorm model compared with the UnNorm model. Moreover, the lower correlation of features with the max-normalized spectrum ($\overline{\varepsilon}_2^{max}$), with ρ = 0.42 for nonmetals, bolsters the previously observed superior performance of the AvgNorm architecture over the MaxNorm. E_{gap} and statistical scalar attributes, including E_{mean}, $E_{variance}$, $E_{skewness}$, $E_{kurtosis}$, and $E(\varepsilon_2^{max})$ (indicating the mode), rank as secondary in significance for structuring nonmetals' latent space, exhibiting ρ values between approximately 0.37 and 0.48. Nonetheless, these attributes are comprehensively integrated within the normalized spectrum $\overline{\varepsilon}_2^{avg}$. Additional scalar properties, including dielectric magnitudes such as the static dielectric constant ($\varepsilon_1(0)$), the average (ε_2^{avg}) and maximum (ε_2^{max}) imaginary dielectric values, and other generic features like the number of spectral peaks (defined as the largest local maxima exceeding 0.25 of the highest peak and separated by at least 2 eV), exhibit the lowest correlation scores with the latent features, all with $\rho \leq 0.2$. This underscores their relatively minor role in shaping the latent space of nonmetals.

The correlation landscape for the latent space of metals exhibits a distinct profile, with both $\overline{\varepsilon}_2^{avg}$ and ε_2^{avg} playing equally significant roles in structuring the latent space, each exhibiting a correlation value of ρ = 0.50. As discussed, the spectral dielectric response of metals, characterized by predominant intraband transitions at low energies within the IR range, allows a simple dielectric magnitude, specifically ε_2^{avg}, to serve as a key descriptor for metals, emphasizing their optical conductivity. However, similar to nonmetals, $\overline{\varepsilon}_2^{avg}$ remains crucial for defining the latent space and accurately representing the spectral energy distribution. This consideration becomes particularly relevant given that metals may exhibit additional peaks at higher frequencies due to potential interband transitions in the late-visible and UV ranges. Consequently, ε_2^{avg} and $\overline{\varepsilon}_2^{avg}$ are identified as two pivotal descriptors for effectively learning metals' optical spectra.

Prediction of Frequency-Dependent Optical Spectra and Solar Absorption Efficiency. Figure 6 depicts predictions for both ε_2 and α for materials selected from the MBJ test set, chosen within an error margin of ±5% around the median MAE in the unnormalized spectrum. These predictions are generated using the optimized AvgNorm (MAE) FE GNN models. The shown median performance highlights the efficacy of the GNN models, enhanced by optimized spectrum scaling and multifidelity learning, in accurately capturing the nuances in DFT optical spectra at the meta-GGA MBJ level. Notably, the models exhibit proficiency in precisely identifying peak values and their respective positions across the entire spectrum, including the IR, visible, and UV regions, spanning metals, semiconductors, and insulators. From ε_2 and α, ε_1 can be derived, yielding the complex dielectric function from which various significant frequency-dependent optical properties can be calculated.

Having presented GNN models capable of generalizing across a broad range of frequencies and diverse materials, it is noteworthy that specific practical applications often necessitate the focus on particular material groups within narrower spectral regions. For example, in solar cell applications, the emphasis is on semiconductors with absorbance profiles that efficiently capture incident solar irradiation. The SLME represents the theoretical maximum photoconversion efficiency of a single p−n junction solar cell. While SLME is a scalar property directly learnable through a single-output ML model, gaining insight into the spectral absorbance characteristics contributing to this SLME value is valuable. This deeper understanding can clarify why a material exhibits a lower or higher SLME, suggesting potential modifications or exploration paths. Therefore, predicting SLME by forecasting the absorption coefficient proves advantageous.

We demonstrate that a multioutput GNN model, trained to predict the frequency-dependent absorption coefficient, can accurately forecast SLME at levels comparable to, or even surpassing, those of a single-output GNN model specifically designed for SLME learning, while maintaining precise spectrum prediction capabilities. By restricting our OPT and MBJ data to materials with band gaps in the range of 0.1−4.5

https://doi.org/10.1021/acsami.4c07328
ACS Appl. Mater. Interfaces XXXX, XXX, XXX−XXX

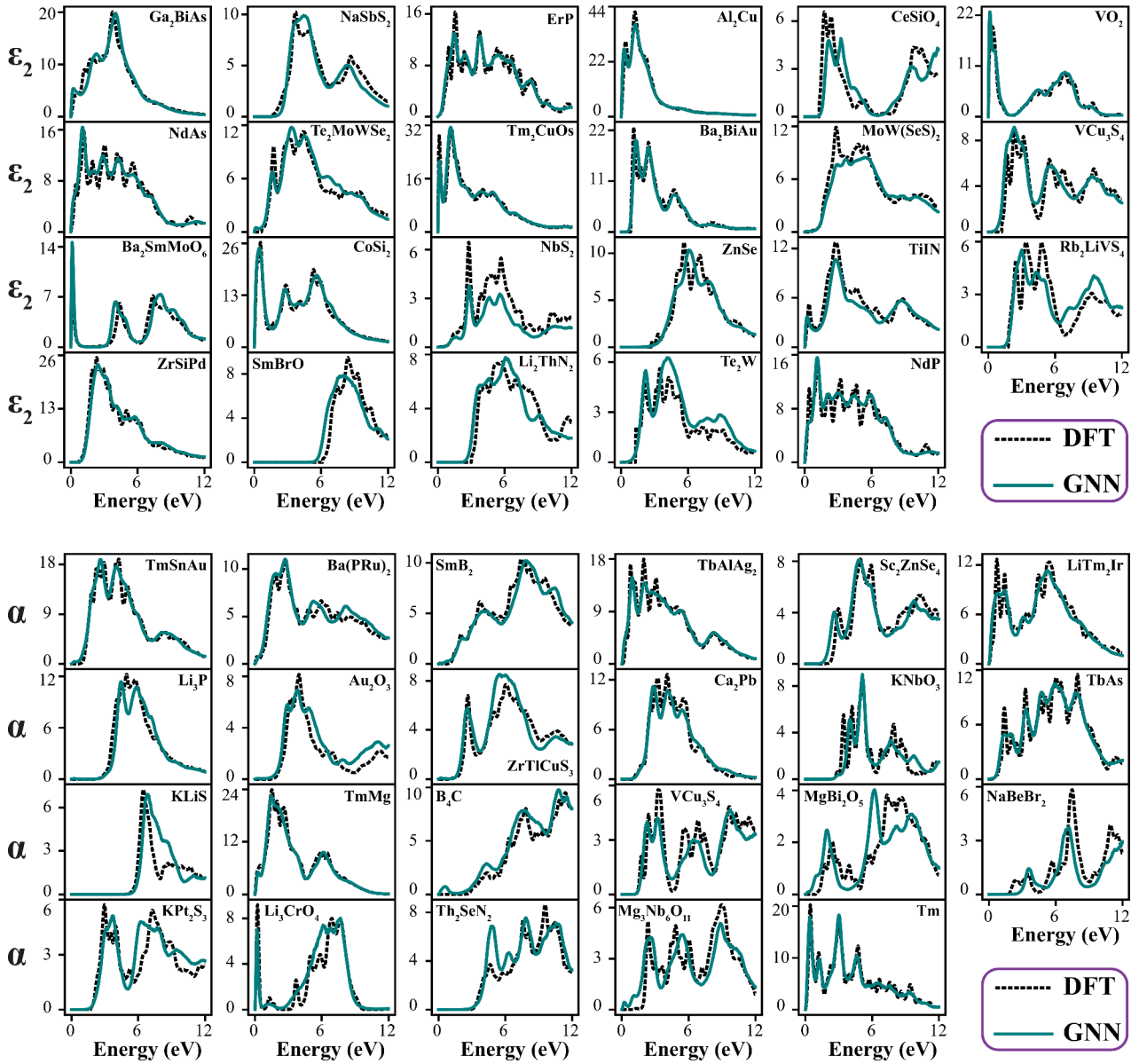


Figure 6. Graph neural network predictions of optical spectra. The interpolated predictions for the dielectric function's imaginary part (ε_2) and the reduced absorption coefficient (α) are shown against the DFT interpolations for samples of materials obtained from the MBJ test set within ±5% around median MAE in the unnormalized spectrum. The predictions are generated using the optimized AvgNorm (MAE) FE models trained on all OPT data and 80% of the MBJ data.

eV (solar irradiation range), we retrain the AvgNorm FE GNN model for predicting α as before but now with two additional neurons in the output layer for predicting the short-circuit current (J_{sc}) and the logarithm of the reverse saturation current ($\log(J_0)$). The loss function is further regularized with two additional loss terms for both J_{sc} and $\log(J_0)$. By training the model using all OPT and 80% of the MBJ data, we evaluate its performance on the remaining MBJ data, as shown in Figure 7. Utilizing the values of J_{sc} and $\log(J_0)$, we calculate SLME following the procedure proposed by K. Choudhary et al. to maximize the power density output from a solar cell,[44]

$$\text{SLME} = \frac{\max\{(J_{sc} - J_0(e^{eV/(kT)} - 1))V\}_V}{\int_0^\infty EI_{sun}(E)\,dE} \quad (2)$$

where $J_{sc} = e\int_0^\infty a(E)\,I_{sun}(E)\,dE$ with I_{sun} representing the AM1.5G solar spectrum.[44,45] The absorbance, $a(E)$, is determined from the absorption coefficient and the film thickness (L) as $a(E) = 1 - \exp(-2(2\sqrt{2}\,\pi(E/hc)\sqrt{\alpha(E)})L)$. $J_0 = e\pi\int_0^\infty a(E)\,I_{bb}(E,T)\,dE$ accounts for the radiative component of the electron−hole recombination current at equilibrium in darkness, with I_{bb} signifying the blackbody irradiation. We assumed thin films with a thickness of 50 nm operating at 300 K for all materials.

The efficacy of the GNN model in forecasting solar efficiency parameters is demonstrated in Figure 7, where it achieves an MAE of 2.04 and a coefficient of determination (R^2) exceeding 0.75 for SLME prediction on unseen materials. This performance demonstrates superiority over traditional non-graph-based ML models. To benchmark this, we trained several non-graph-based ML models, including random forests from scikit-Learn[46] and gradient boosting decision trees from XGBoost and LightGBM packages,[47,48] on the MBJ data set

https://doi.org/10.1021/acsami.4c07328
ACS Appl. Mater. Interfaces XXXX, XXX, XXX−XXX

Table 1. Comparison of Various GNN and Non-Graph-Based ML Models for Predicting Solar Absorption Efficiency Parameters[a]

model (learnables)	J_{sc} (R^2/MAE)	$\log(J_0)$ (R^2/MAE)	SLME (R^2/MAE)
GNN ($\alpha(E)$)	0.87/33.00	−1.86/60.26	−0.15/5.04
GNN ($\{J_{sc}, \log(J_0)\}$)	0.87/25.87	0.81/11.63	0.70/2.18
GNN ($\alpha(E)$ + $\{J_{sc}, \log(J_0)\}$ bias)	0.90/22.42	0.87/10.05	0.76/2.04
Random forest ($\{J_{sc}, \log(J_0)\}$)	0.71/50.59	0.69/17.60	0.60/3.10
XGBoost ($\{J_{sc}, \log(J_0)\}$)	0.71/46.95	0.71/15.96	0.59/2.93
LightGBM ($\{J_{sc}, \log(J_0)\}$)	0.75/44.18	0.74/15.18	0.63/2.84

[a]The ML-predicted short-circuit current (J_{sc}), natural logarithm of the reverse saturation current ($\log(J_0)$), and the spectroscopic limit of maximum efficiency (SLME) are evaluated on the DFT (MBJ) test set. The GNN models differ in their learning approach: they are either estimating the solar parameters from a learned frequency-dependent absorption coefficient, learning the solar parameters directly, or applying a learning bias for these parameters while concurrently learning the absorption coefficient. J_{sc} and J_0 are measured in amperes.

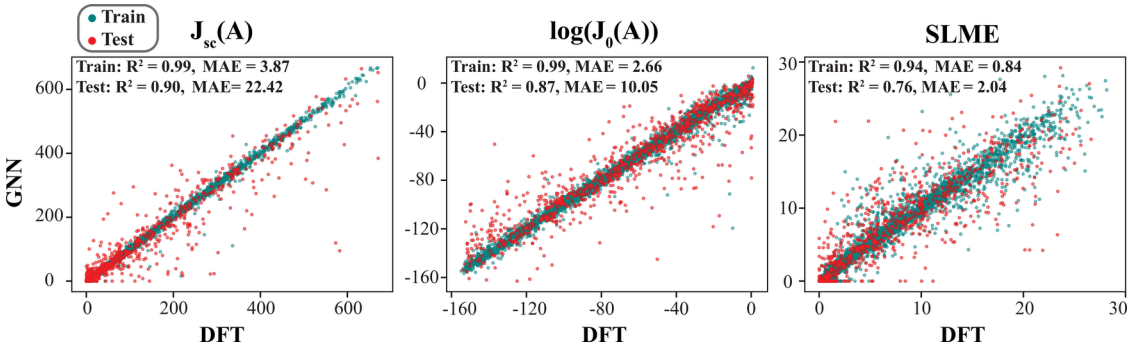


Figure 7. Graph neural network prediction of solar energy absorption efficiency. Predictions of the short-circuit current (J_{sc}), natural logarithm of the reverse saturation current ($\log(J_0)$), and the spectroscopic limit of maximum efficiency (SLME) are validated for materials with band gaps ranging from 0.1 to 4.5 eV against DFT (MBJ) values. J_{sc} and J_0 are expressed in amperes (A). The employed GNN predicts the frequency-dependent absorption coefficient with an extra learning bias to emphasize learning J_{sc} and $\log(J_0)$.

using features derived from chemical composition and crystal structure provided by the automatminer package.[49] The LightGBM model achieves the best performance on the test set among non-graph-based ML models, as outlined in Table 1. Compared to the performance metrics of the solar-biased GNN model presented in Figure 7 and denoted as GNN ($\alpha(E)$ + $\{J_{sc}, \log(J_0)\}$ bias) in Table 1, the GNN model demonstrates superior efficacy. This is manifested by the MAE for the best non-graph-based LightGBM model being approximately 97%, 51%, and 39% higher for J_{sc}, $\log(J_0)$, and SLME, respectively, compared to the solar-biased GNN model. Further details about the GNNs, as well as the features and hyperparameters used for the non-graph-based ML models, are provided in the Supporting Information.

Furthermore, a GNN model with the same architecture, yet with an output layer of only two neurons and a loss function tailored exclusively for predicting J_{sc} and $\log(J_0)$, denoted by GNN ($\{J_{sc}, \log(J_0)\}$) in Table 1, yields inferior performance compared to the solar-biased GNN, with the MAE higher by approximately 15%, 16%, and 7% for J_{sc}, $\log(J_0)$, and SLME, respectively. Thus, we can notice that leveraging a multioutput GNN initially designed for learning the absorption spectrum, when further refined with a learning bias to emphasize solar parameters' learning, can yield improved predictive performance in forecasting solar parameters compared to directly learning them as singular targets. This improvement can be attributed to the fact that these solar parameters result from convolution integrals with the frequency-dependent absorption coefficient. By synergizing the learning of the absorption coefficient with the learning of these solar parameters, the latent feature space of the GNN becomes enriched with information on the absorption spectrum, thus enhancing the model's predictive accuracy regarding solar parameters. On the other hand, a solar-unbiased GNN model with the same architecture trained solely for predicting the absorption coefficient, denoted by GNN ($\alpha(E)$) in Table 1, achieves MAE = 0.384 and r = 0.948 for the unnormalized spectrum of α. In comparison, the solar-biased GNN model demonstrates an almost identical performance, with MAE = 0.386 and r = 0.949. However, the solar-unbiased model yields suboptimal results for the solar parameters, as outlined in Table 1. The MAE increases by 47%, 50%, and 147% for J_{sc}, $\log(J_0)$, and SLME, respectively, compared with the solar-biased model. This suggests that incorporating a learning bias via simple regularization terms in the loss function, aimed at emphasizing specific practical physical properties, can aid in distributing the error of the multioutput prediction in a way that substantially enhances the learning of these physical properties without compromising the overall accuracy of predicting the absorption spectrum.

CONCLUSION

We have developed multioutput GNN models capable of predicting the frequency-dependent imaginary dielectric function and absorption coefficient with accuracy on par with meta-GGA DFT. This enables the derivation of the complete dielectric response of any arbitrary material using only its atomic structure as input. We considered a spectrum spanning a 12 eV range (from IR to UV), yet the employed GNN formalism offers easily adaptable spectrum ranges. We investigated the effect of spectrum scaling on the formation of the latent feature space and the GNNs' predictive capacity by

https://doi.org/10.1021/acsami.4c07328
ACS Appl. Mater. Interfaces XXXX, XXX, XXX−XXX

comparing various scaling schemes, including UnNorm, MaxNorm, and AvgNorm models. Our findings highlight that the AvgNorm GNN model, treating the optical spectrum of any material as an average norm and a probability distribution function, along with a softmax activation and a loss function with an evenly weighted spectrum, exhibits superior performance. Furthermore, our GNN models integrate multifidelity learning schemes, such as transfer learning and fidelity embedding, and utilize the whole low-fidelity OPT and high-fidelity MBJ optical spectra available in the JARVIS DFT database. We observe a notable decrease across all error metrics and an increase in correlation with the DFT (MBJ) spectra upon the incorporation of the lower-fidelity OPT data, with the fidelity embedding approach achieving moderately superior accuracy to transfer learning attributed to its broader multifidelity optimization scope.

We also demonstrated that the prediction of the frequency-dependent absorption coefficient by GNNs can be fine-tuned to emphasize specific scalar optoelectronic properties without compromising the overall spectrum multioutput prediction, through simple learning biases applicable to any property of interest. As a proof of concept, we optimized the learning of solar cell performance parameters (short-circuit current, reverse saturation current, and the corresponding solar cell efficiency SLME), showing that integrating these properties into the multioutput learning of the absorption coefficient leads to enhanced prediction of solar properties compared to learning them separately. This synergy of multioutput and multifidelity learning, along with the flexibility to apply target-specific learning biases, presents a versatile tool for the rapid screening of solid materials for a wide range of frequency-dependent optical device applications involving metals, semiconductors, or insulators.

■ ASSOCIATED CONTENT

Data Availability Statement

To enhance reproducibility, the data sets and codes for model design and result analysis are available on GitHub at https://github.com/UMBC-STEAM-LAB/GNN-frequency-dependent-optical-properties.

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acsami.4c07328.

Data set preparation, model architecture, hyperparameter optimization, and error distributions (PDF)

■ AUTHOR INFORMATION

Corresponding Author

Can Ataca − *Department of Physics, University of Maryland Baltimore County, Baltimore, Maryland 21250, United States;* orcid.org/0000-0003-4959-1334; Email: ataca@umbc.edu

Author

Akram Ibrahim − *Department of Physics, University of Maryland Baltimore County, Baltimore, Maryland 21250, United States;* orcid.org/0009-0008-7311-7062

Complete contact information is available at: https://pubs.acs.org/10.1021/acsami.4c07328

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The authors acknowledge funding from the National Science Foundation (NSF) under Grant NSF DMR-2213398 and from the Department of Energy (DOE) under Grant DE-SC0024236.

■ REFERENCES

(1) Day, J.; Senthilarasu, S.; Mallick, T. K. Improving spectral modification for applications in solar cells: A review. *Renewable Energy* **2019**, *132*, 186−205.

(2) Ren, A.; Wang, H.; Zhang, W.; Wu, J.; Wang, Z.; Penty, R. V.; White, I. H. Emerging light-emitting diodes for next-generation data communications. *Nat. Electron.* **2021**, *4*, 559−572.

(3) Won, D.; Bang, J.; Choi, S. H.; Pyun, K. R.; Jeong, S.; Lee, Y.; Ko, S. H. Transparent electronics for wearable electronics application. *Chem. Rev.* **2023**, *123*, 9982−10078.

(4) Li, M.; Cushing, S. K.; Wu, N. Plasmon-enhanced optical sensors: a review. *Analyst* **2015**, *140*, 386−406.

(5) Hu, C.; Guo, K.; Li, Y.; Gu, Z.; Quan, J.; Zhang, S.; Zheng, W. Optical coatings of durability based on transition metal nitrides. *Thin Solid Films* **2019**, *688*, 137339.

(6) Babbe, F.; Sutter-Fella, C. M. Optical absorption-based in situ characterization of halide perovskites. *Adv. Energy Mater.* **2020**, *10*, 1903587.

(7) Materese, C. K.; Gerakines, P. A.; Hudson, R. L. Laboratory Studies of Astronomical Ices: Reaction Chemistry and Spectroscopy. *Acc. Chem. Res.* **2021**, *54*, 280−290.

(8) Wang, T.; Zhang, K.; Thé, J.; Yu, H. Accurate prediction of band gap of materials using stacking machine learning model. *Comput. Mater. Sci.* **2022**, *201*, 110899.

(9) Rajan, A. C.; Mishra, A.; Satsangi, S.; Vaish, R.; Mizuseki, H.; Lee, K.-R.; Singh, A. K. Machine-learning-assisted accurate band gap predictions of functionalized MXene. *Chem. Mater.* **2018**, *30*, 4031−4038.

(10) Zhuo, Y.; Mansouri Tehrani, A.; Brgoch, J. Predicting the band gaps of inorganic solids by machine learning. *journal of physical chemistry letters* **2018**, *9*, 1668−1673.

(11) Shimano, Y.; Kutana, A.; Asahi, R. Machine learning and atomistic origin of high dielectric permittivity in oxides. *Sci. Rep.* **2023**, *13*, 22236.

(12) Takahashi, A.; Kumagai, Y.; Miyamoto, J.; Mochizuki, Y.; Oba, F. Machine learning models for predicting the dielectric constants of oxides based on high-throughput first-principles calculations. *Phys. Rev. Mater.* **2020**, *4*, 103801.

(13) Ben Mahmoud, C.; Anelli, A.; Csányi, G.; Ceriotti, M. Learning the electronic density of states in condensed matter. *Phys. Rev. B* **2020**, *102*, 235130.

(14) Fung, V.; Ganesh, P.; Sumpter, B. G. Physically informed machine learning prediction of electronic density of states. *Chem. Mater.* **2022**, *34*, 4848−4855.

(15) Kong, S.; Ricci, F.; Guevarra, D.; Neaton, J. B.; Gomes, C. P.; Gregoire, J. M. Density of states prediction for materials discovery via contrastive learning from probabilistic embeddings. *Nat. Commun.* **2022**, *13*, 949.

(16) Kong, S.; Guevarra, D.; Gomes, C. P.; Gregoire, J. M. Materials representation and transfer learning for multi-property prediction. *Appl. Phys. Rev.* **2021**, *8*, No. 021409.

(17) Chen, L.; Kim, C.; Batra, R.; Lightstone, J. P.; Wu, C.; Li, Z.; Deshmukh, A. A.; Wang, Y.; Tran, H. D.; Vashishta, P.; et al. Frequency-dependent dielectric constant prediction of polymers using machine learning. *npj Comput. Mater.* **2020**, *6*, 61.

(18) Ward, L.; Dunn, A.; Faghaninia, A.; Zimmermann, N. E.; Bajaj, S.; Wang, Q.; Montoya, J.; Chen, J.; Bystrom, K.; Dylla, M.; et al. Matminer: An open source toolkit for materials data mining. *Comput. Mater. Sci.* **2018**, *152*, 60−69.

(19) Chen, C.; Ye, W.; Zuo, Y.; Zheng, C.; Ong, S. P. Graph networks as a universal machine learning framework for molecules and crystals. *Chem. Mater.* **2019**, *31*, 3564−3572.

(20) Choudhary, K.; DeCost, B. Atomistic line graph neural network for improved materials property predictions. *npj Comput. Mater.* **2021**, *7*, 185.
(21) Xie, T.; Grossman, J. C. Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. *Physical review letters* **2018**, *120*, 145301.
(22) Lewis, N. R.; Jin, Y.; Tang, X.; Shah, V.; Doty, C.; Matthews, B. E.; Akers, S.; Spurgeon, S. R. Forecasting of in situ electron energy loss spectroscopy. *npj Comput. Mater.* **2022**, *8*, 252.
(23) Shapera, E. P.; Schleife, A. Discovery of New Plasmonic Metals via High-Throughput Machine Learning. *Adv. Opt. Mater.* **2022**, *10*, 2200158.
(24) Chander, S.; Purohit, A.; Nehra, A.; Nehra, S.; Dhaka, M. A study on spectral response and external quantum efficiency of monocrystalline silicon solar cell. *Int. J. Renewable Energy Res.* **2015**, *5*, 41−44.
(25) Jha, D.; Choudhary, K.; Tavazza, F.; Liao, W.-k.; Choudhary, A.; Campbell, C.; Agrawal, A. Enhancing materials property prediction by leveraging computational and experimental data using deep transfer learning. *Nat. Commun.* **2019**, *10*, 5316.
(26) Hoffmann, N.; Schmidt, J.; Botti, S.; Marques, M. A. Transfer learning on large datasets for the accurate prediction of material properties. *Digital Discovery* **2023**, *2*, 1368−1379.
(27) Chen, C.; Zuo, Y.; Ye, W.; Li, X.; Ong, S. P. Learning properties of ordered and disordered materials from multi-fidelity data. *Nature Computational Science* **2021**, *1*, 46−53.
(28) Wines, D.; Gurunathan, R.; Garrity, K. F.; DeCost, B.; Biacchi, A. J.; Tavazza, F.; Choudhary, K. Recent progress in the JARVIS infrastructure for next-generation data-driven materials design. *Appl. Phys. Rev.* **2023**, *10*, No. 041302.
(29) Choudhary, K.; Garrity, K. F.; Reid, A. C.; DeCost, B.; Biacchi, A. J.; Hight Walker, A. R.; Trautt, Z.; Hattrick-Simpers, J.; Kusne, A. G.; Centrone, A.; et al. The joint automated repository for various integrated simulations (JARVIS) for data-driven materials design. *npj Comput. Mater.* **2020**, *6*, 173.
(30) Choudhary, K.; Zhang, Q.; Reid, A. C.; Chowdhury, S.; Van Nguyen, N.; Trautt, Z.; Newrock, M. W.; Congo, F. Y.; Tavazza, F. Computational screening of high-performance optoelectronic materials using OptB88vdW and TB-mBJ formalisms. *Sci. Data* **2018**, *5*, 180082.
(31) Klimeš, J.; Bowler, D. R.; Michaelides, A. Chemical accuracy for the van der Waals density functional. *J. Phys.: Condens. Matter* **2010**, *22*, No. 022201.
(32) Becke, A. D.; Johnson, E. R. A simple effective potential for exchange. *J. Chem. Phys.* **2006**, *124*, 221101.
(33) Tran, F.; Blaha, P. Accurate band gaps of semiconductors and insulators with a semilocal exchange-correlation potential. *Physical review letters* **2009**, *102*, 226401.
(34) Vinyals, O.; Bengio, S.; Kudlur, M. Order Matters: Sequence to sequence for sets. *arXiv* **2015**, 1511.06391 https://arxiv.org/abs/1511.06391.
(35) Abadi, M.; et al.. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems, 2015. https://www.tensorflow.org/.
(36) Chollet, F.; et al. Keras, 2015. https://keras.io.
(37) Akiba, T.; Sano, S.; Yanase, T.; Ohta, T.; Koyama, M. Optuna: A next-generation hyperparameter optimization framework. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*; SIGKDD, 2019; pp 2623−2631.
(38) Joyce, J. M. In *International Encyclopedia of Statistical Science*; Lovric, M., Ed.; Springer: Berlin, 2011; pp 720−722.
(39) Frogner, C.; Zhang, C.; Mobahi, H.; Araya-Polo, M.; Poggio, T. Learning with a Wasserstein Loss. *arXiv* **2015**, 1506.05439 https://arxiv.org/abs/1506.05439.
(40) Ramdas, A.; García Trillos, N.; Cuturi, M. On Wasserstein two-sample testing and related families of nonparametric tests. *Entropy* **2017**, *19*, 47.
(41) van der Maaten, L.; Hinton, G. Visualizing Data using t-SNE. *J. Mach. Learn. Res.* **2008**, *9*, 2579−2605.
(42) Hilfiker, J. N.; Tiwald, T. Dielectric function modeling. *Spectroscopic Ellipsometry for Photovoltaics: Volume 1: Fundamental Principles and Solar Cell Characterization* **2018**, *212*, 115−153.
(43) Kuzmany, H.; Kuzmany, H. The dielectric function. *Solid-State Spectroscopy: An Introduction* **1998**, 101−120.
(44) Choudhary, K.; Bercx, M.; Jiang, J.; Pachter, R.; Lamoen, D.; Tavazza, F. Accelerated discovery of efficient solar cell materials using quantum and machine-learning methods. *Chemistry of materials* **2019**, *31*, 5900−5908.
(45) Collins, D. G.; Blättner, W. G.; Wells, M. B.; Horak, H. G. Backward Monte Carlo calculations of the polarization characteristics of the radiation emerging from spherical-shell atmospheres. *Appl. Opt.* **1972**, *11*, 2684−2696.
(46) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825−2830.
(47) Chen, T.; Guestrin, C. Xgboost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*; SIGKDD, 2016; pp 785−794.
(48) Ke, G.; Meng, Q.; Finley, T.; Wang, T.; Chen, W.; Ma, W.; Ye, Q.; Liu, T.-Y. Lightgbm: A highly efficient gradient boosting decision tree. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 3146−3154.
(49) Dunn, A.; Wang, Q.; Ganose, A.; Dopp, D.; Jain, A. Benchmarking materials property prediction methods: the Matbench test set and Automatminer reference algorithm. *npj Comput. Mater.* **2020**, *6*, 138.

https://doi.org/10.1021/acsami.4c07328
ACS Appl. Mater. Interfaces XXXX, XXX, XXX−XXX