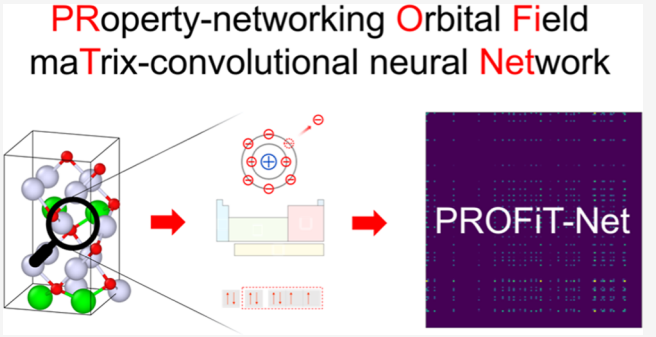
JACS
JOURNAL OF THE AMERICAN CHEMICAL SOCIETY
pubs.acs.org/JACS
Article

PROFiT-Net: Property-Networking Deep Learning Model for Materials

Se-Jun Kim, Won June Kim, Changho Kim, Eok Kyun Lee,* and Hyungjun Kim*

Cite This: J. Am. Chem. Soc. 2024, 146, 26000−26007
Read Online
ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: There is a growing need to develop artificial intelligence technologies capable of accurately predicting the properties of materials. This necessitates the expansion of material databases beyond the scope of density functional theory, and also the development of deep learning (DL) models that can be effectively trained with a limited amount of high-fidelity data. We developed a DL model utilizing a crystal structure representation based on the orbital field matrix (OFM), which was modified to incorporate information on elemental properties and valence electron configurations. This model, effectively capturing the interrelation between the elemental properties in the crystal, was coined the PRoperty-networking Orbital Field maTrix-convolutional neural Network (PROFiT-Net). Remarkably, PROFiT-Net demonstrated high accuracy in predicting the dielectric constant, experimental band gaps, and formation enthalpies compared with other leading DL models. Moreover, our model accurately identifies physical patterns, such as avoiding the prediction of unphysical negative band gaps and exhibiting a Penn-model-like trend while maintaining the scalability. We envision that PROFiT-Net will accelerate the development of functional materials.



1. INTRODUCTION

The development of artificial intelligence caused a paradigm shift in prediction and design of crystal materials.[1] The traditional methods of characterization and design of crystal materials often rely on laborious trial-and-error experimentation, guided by intuition and high-throughput computational screening using density functional theory (DFT).[2] This approach suffers from high costs, time inefficiencies, and the breadth of material design, which hinders the rapid development of advanced materials.

Based on the rapid data accumulation and the development of computing power, machine learning (ML) techniques enable the extraction of meaningful patterns and correlations from vast data sets comprising material properties.[3] By leveraging algorithms capable of learning from data, ML empowers researchers to predict material properties, identify promising candidates, and optimize parameters with unprecedented accuracy and efficiency. Particularly, deep learning (DL), a subset of ML, excels in handling intricate, high-dimensional data sets, offering advanced capabilities in feature extraction, pattern recognition, and predictive modeling.[4]

In materials science and computational chemistry, the accurate representation of molecular and atomic structures is essential for understanding and predicting their properties and behaviors. With the advent of ML techniques, the development of efficient and robust methods for encoding structural information into numerical descriptors that can be readily utilized by ML algorithms has drawn considerable attention. In this context, several innovative approaches have emerged, each with its own strengths and applications. Prominent examples include the Coulomb matrix,[5] which is based on a pairwise electrostatic representation; atom-centered symmetry functions,[6] which capture local structures; smooth overlap of atomic positions (SOAP),[7] which employs smooth rotationally and translationally invariant atomic density distributions; and many-body tensor representation,[8] which considers pairwise interactions extended to higher-order interactions involving multiple atoms. In addition, graphs, which are fundamental data structures used to represent the connections or relationships between pairs of objects using vertices (nodes) and edges (connections), are widely used to represent the atomic structures of molecules and materials.[9]

The development of crystal structure representations has frequently correlated with enhancements in DL models for predicting material properties. For example, the crystal graph convolutional neural network (CGCNN), introduced by Xie

Received: April 16, 2024
Revised: September 4, 2024
Accepted: September 5, 2024
Published: September 12, 2024

https://doi.org/10.1021/jacs.4c05159
J. Am. Chem. Soc. 2024, 146, 26000−26007

and Grossman[10] has demonstrated reliable accuracy in predicting properties including formation energy, absolute energy, and band gap. Chen et al. developed the materials graph network (MEGNet), which integrates atom, bond, and global state variables into a unified model while embedding elemental features.[11] Additionally, the orbital graph convolutional neural network (OGCNN) was recently introduced,[12] combining the orbital field matrix (OFM)[13] and CGCNN atomic features within the framework of CGCNN and an autoencoder.[12] However, these DL models have predominantly trained on databases at the DFT level, which inherently limits their ability to accurately predict properties such as band gaps. Furthermore, their fidelity-tolerance and performance with more accurate, albeit rarer, databases have not been thoroughly explored.

In this study, we developed a DL method, based on encoding pairwise intercorrelations between orbitals and elemental properties in a CNN. This approach is referred to as *PR*operty-networking *O*rbital *Fi*eld ma*T*rix (PROFiT)-Net, which is based on a generalization of the previously developed crystal representation of the OFM.[13] Using dielectric materials as an example, a popular material group common in semiconductor research, we trained PROFiT-Net to predict electronic dielectric constant (ε) for 1217 distinct systems. We demonstrated superior performance compared with other state-of-the-art techniques. Considering the importance of a large band gap (E_g) for high dielectric materials and a low formation enthalpy (ΔH_f) for thermodynamic stability, we extended the training of PROFiT-Net to predict E_g and ΔH_f. The model demonstrated significantly enhanced predictive power, especially for high-fidelity data sets. Subsequently, we found that our model correctly learns physical patterns from data such as the trend of DFT underestimating band gap without the information on data set fidelity and the Penn-model-like physical trend of using large-scale databases. Finally, we validated the scalability of our model by investigating a large database.

2. METHODS

The OFM, as originally proposed by Lam Pham et al.,[13] is a 2D descriptor, in which the local environment of the central atom is described by the valence electron configurations of the central atom and neighboring atoms using one-hot vectors. To define the atomic feature, like a natural language model, a one-hot vector ($\vec{O}_{\text{atom}}$) consisting of 32 bins is assigned to each atom to represent the valence electron configuration of $D = \{s^1, s^2, p^1 \ldots f^{13}, f^{14}\}$ (n^m means that m electrons are occupied in the n-orbital). For instance, in the case of an oxygen atom with a valence electron configuration of $2s^22p^4$, the s^2 and p^4 bins are filled with ones, and the others are filled with zeros. The neighbor list of each atom is determined according to O'Keeffe's theory,[14] whereby Voronoi polyhedral are used to define coordinating atoms. This eliminates arbitrariness in defining the local environment, unlike many previous graph-based models that often define coordinating atoms based on an empirically chosen cutoff distance.[10,11]

To represent the local environment, OFM considers a multiplication of a transposed one-hot vector of a central atom p, denoted by $\overrightarrow{O_p}^{\mathrm{T}}$, with a one-hot vector of a coordinating atom k, denoted by $\overrightarrow{O_k}$, i.e., $\overrightarrow{O_p}^{\mathrm{T}} \times \overrightarrow{O_k}$. Since this leads to a degenerated representation for the pairs of atoms having the same valence electron configurations (e.g., NaCl vs KCl), it is further normalized by the bond length r_{pk} between p and k, resulting in a bond matrix.

$$\frac{1}{r_{pk}}(\overrightarrow{O_p}^{\mathrm{T}} \times \overrightarrow{O_k}) \tag{1}$$

Then, OFM defines a local environmental matrix $X^{(p)}$ of the central atom p by summing bond matrices (see eq 1) over the neighbor list of the central atom p with weights defined by the ratio of the solid angle $\theta_k^{(p)}$ of the atom k to the maximum value $\theta_{\max}^{(p)} = \max_{1 \le k \le n_p} \theta_k^{(p)}$ with n_p is the number of atoms coordinating the central atom p. Hence, $X^{(p)}$ and its component $X_{i,j}^{(p)}$ are expressed as

$$X^{(p)} = \sum_{k=1}^{n_p} \frac{1}{r_{pk}} \frac{\theta_k^{(p)}}{\theta_{\max}^{(p)}} (\overrightarrow{O_p}^{\mathrm{T}} \times \overrightarrow{O_k}) \tag{2}$$

$$X_{i,j}^{(p)} = \sum_{k=1}^{n_p} \frac{1}{r_{pk}} \frac{\theta_k^{(p)}}{\theta_{\max}^{(p)}} o_p(i) o_k(j) \tag{3}$$

where $o_p(i)$ is the ith element of $\overrightarrow{O_p}^{\mathrm{T}}$ and $o_k(j)$ is the jth element of $\overrightarrow{O_k}$. Finally, the entire representation of the crystal, denoted by F, is constructed by averaging X_p of all atoms ($p = 1, 2, \ldots, N$) with N being the total number of atoms in the unit cell

$$F = \frac{1}{N} \sum_{p=1}^{N} X_p \tag{4}$$

While valence electronic configurations are linked to various chemical properties of an atom, one cannot predict all essential atomic and material properties solely from an understanding of this configuration. For example, the nucleus-electron interaction plays a vital role in determining the ionization energy (IE) and electron affinity (EA) of an element. Notably, the average of IE and EA determines the atomic electronegativity (E), a key factor influencing bond dipole and bond ionicity (covalency), which in turn affect the band gap and dielectric constant of a material. This signifies the limitations associated with predicting properties solely using valence electron configurations.

Therefore, in our model, we concatenate elemental properties with the valence electron configuration, expanding the number of bins of $\vec{O}_{\text{atom}}$ to a total of 136. These include group/period numbers (GN/PN), E, covalent radius (CR), the number of valence electrons, first IE (FIE), EA, block (B), atomic volume, and atomic polarizability (P). How each feature is embedded in an atomic feature vector is described in Table S1. By adopting the concatenation method, the elemental properties can interact across bonded pairs during construction of $X^{(p)}$, facilitating the extraction of hidden patterns and relationships crucial for determining the chemophysical nature of materials. We thus coined our model PROFiT.

To construct the DL model using CNN, we flattened the two-dimensional array of eq 4 into one-dimensional array as an input for convolution layers and sequentially dense layers. The convolution layers and fully connected layers were sequentially connected by Tensorflow[15] and Keras,[16] and ReLU[17] was used as an activation function.

Our network consists of convolution layers and fully connected layers. The activation function, ReLU, follows each convolution layer to learn the nonlinear relationship between structure and property. In addition, the dropout technique was used to prevent overfitting after the convolution layers. After the initial three blocks (first−third blocks), our model learned the relationship only with the combination of a convolution layer and a dropout in the other blocks (fourth−fifth blocks). Thus, the convolution layer part of our model is constructed as the mixture of the first−third blocks where the ReLU function exists, and the fourth−fifth blocks, where only convolution and dropout layers are included to optimize the trade-off between simplifying the model and capturing the complexity of the patterns. Since pooling layers can improve computational efficiency by reducing the input size, the pooling layer follows the dropout layer for the max pooling layer, while the order is reversed for the average

https://doi.org/10.1021/jacs.4c05159
J. Am. Chem. Soc. 2024, 146, 26000−26007

pooling layer. For fully connected layers, dense layers and ReLU activation functions are alternatively repeated five times, with the sixth dense layer being the last one. Since our DL model is designed to predict a scalar value like ε, the final output from fully connected layers is consequently reduced into a single value. This output in a scalar form is used in calculating the loss, which here is the mean squared error (MSE) loss, and the weights are updated by minimizing this loss via the backpropagation algorithm. Details of the model are tabulated in Table S2.

The number of epochs was set as 500 during all training and the best model from the validation results was used to assess the test set. NVIDIA GeForce GTX 1080 was used for model learning. Hyperparameter optimization of our model was carried out with respect to the ε data set by changing the number of convolution layers, dense layers, kernel size, the number of kernels, dropout, and max or average pooling.

All data sets used in this work, as well as their references and the number of structures in data sets are listed in Table S3. Note that the sizes of data sets employed in our study are slightly different from the reported sizes in the references as indicated in parentheses, which is due to the update of the Materials Project.[18]

The ε data set as calculated by density functional perturbation theory was composed of 1217 structures that are queried via the Materials Project[18] from ref 19 (1364). The PBE(+U)-level DFT band gap (E_g^{PBE}) and PBE(+U)-level DFT formation enthalpy per atom (ΔH_f^{PBE}) data sets were also queried through the Materials Project[18] from ref 10 (46,744), leading to 36,837 crystal structures. This data set was composed of 15,614 metallic systems and 21,223 semiconducting or insulating materials. Only nonmetallic structures are used for the E_g^{PBE} data set, while all structures are used for the ΔH_f^{PBE} data set. For HSE06-level DFT band gaps (E_g^{HSE06}) calculated by Heyd−Scuseria−Ernzerhof06 (HSE06) hybrid functional,[20] we utilized ref 21 (10,481) and removed the compounds which contain deuterium, resulting in 10,388 structures. The experimental band gaps (E_g^{exp}) data set was composed of 465 structures was obtained from ref 22 (472) while experimental formation enthalpies (ΔH_f^{exp}) consisting of 1127 structures were constructed using ref 23 (1143). The calculation data sets of ε, E_g^{PBE}, E_g^{HSE06}, and ΔH_f^{PBE} were cleaned in previous works, while no additional cleaning was carried out for the experimental data sets of E_g^{exp} and ΔH_f^{exp}. The data set was split into 6:2:2 as training, validation, and test sets, respectively. All materials data were preprocessed with Pymatgen[24] and Matminer[25] Python packages. Voronoi tessellation was performed using the Pymatgen code by setting the tolerance parameter to 0.

3. RESULTS & DISCUSSION

To benchmark the accuracy level of PROFiT-Net, we compared it with other DL models including CNNs employing the conventional OFM (OFMCNN), CGCNN,[10] MEGNet,[11] and OGCNN.[12] For OFMCNN, all parameters and model architecture were set to be the same as PROFiT-Net, except that the pairwise intercorrelations between elemental properties were absent in the OFM. For CGCNN, MEGNet, and OGCNN, which are state-of-the-art-level graph-based CNN models, the default settings in their GitHub repositories were used to conserve the originality of each model, except for the cutoff of MEGNet and the definition of the neighbor list, which was slightly adjusted only when errors occurred due to the existence of isolated atoms. Details of CGCNN, MEGNet, and OGCNN are provided in Tables S4−S6, with a comparison of the number of model parameters shown in Table S7. PROFiT-Net has 14 M model parameters, thus exhibiting substantial flexibility at the expense of a relatively long computation time (Figure S1).

Using $LiAg_2O_3$ as a representative example, Figure 1 illustrates the flowchart used to obtain the square feature matrix F, see eqs 2−4. The matrix demonstrates sparsity due to its construction by the multiplication of atomic feature vectors. Diagonal components capture interactions between identical properties of neighboring atom pairs; for instance, $F_{1,1}$ represents the interaction involving the half-filled s-orbital of

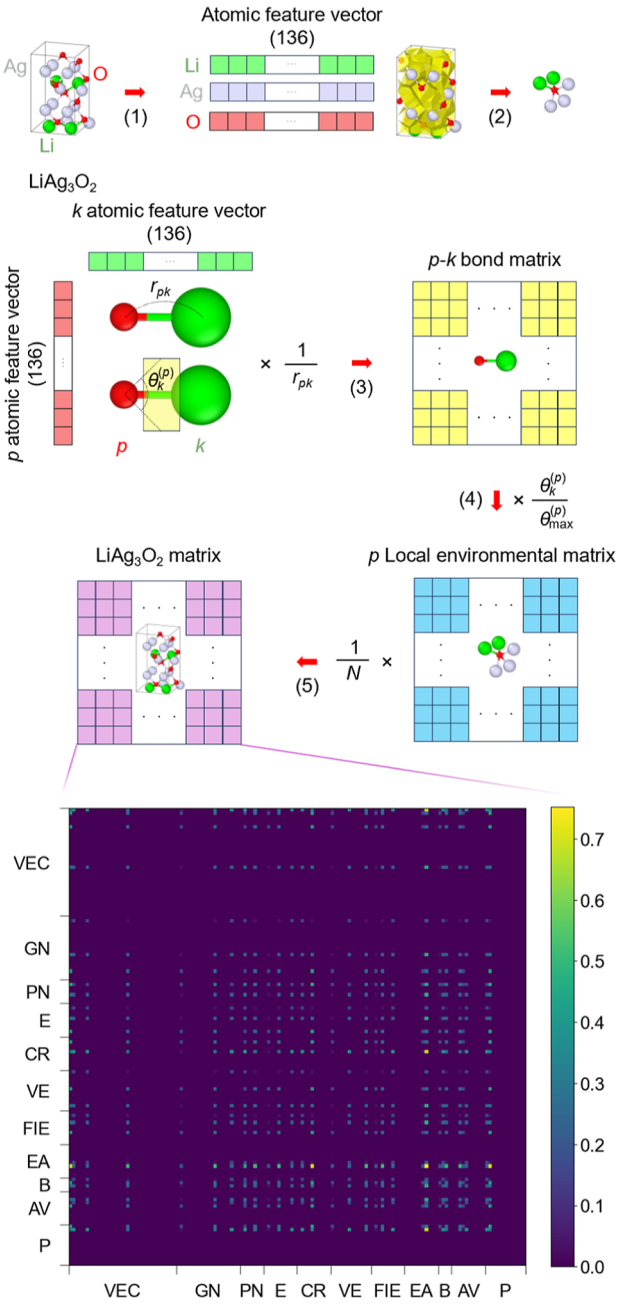


Figure 1. Scheme of crystal structure representation for $LiAg_3O_2$. (1) The atomic feature vector is defined for each atom in a unit cell, with the number in parentheses indicating the vector size. (2) The neighbor list of each atom is determined through the three-dimensional Voronoi tessellation. (3) The bond matrix is constructed by multiplying the transposed atomic feature vector of a central atom with that of the coordinating atom. (4) Each bond matrix is weighted by the solid angle and merged into the local environmental matrix. (5) The local environmental matrices of all atoms in a unit cell are combined and normalized by the number of atoms in the unit cell (N), resulting in the matrix representing $LiAg_3O_2$. The color map of the matrix based on the magnitude of each component shows the sparsity of the matrix. Abbreviations of atomic features are used (see Table S1).

https://doi.org/10.1021/jacs.4c05159
J. Am. Chem. Soc. 2024, 146, 26000−26007

an atom with those of its neighbors. On the other hand, off-diagonal components highlight interactions between different properties. As an illustration, $F_{50,100}$ denotes the interaction between the group number 17 of an atom and the first ionization energies of its neighboring atoms. In this manner, our model encodes pairwise property-property relationships beyond the orbital interactions of the original OFM.[19] This sets apart it unique to previous ones.

Furthermore, it is noteworthy that the matrix size, which determines the input size of the CNN, is independent of the type or size of the material. It is given by a square of the length of the atomic feature vector, i.e., 136 × 136, irrespective of the system size. Considering that the neighbor list is constructed using O'Keeffe's theory, our representation, which preserves translational and rotational invariances, is empiricism-free and size-invariant while uniquely integrating the chemical natures and connectivity of adjacent atoms concurrently within a single square matrix.

We then trained PROFiT-Net (see its structure in Figure 2) against the ε data set consisting of 1217 different materials.

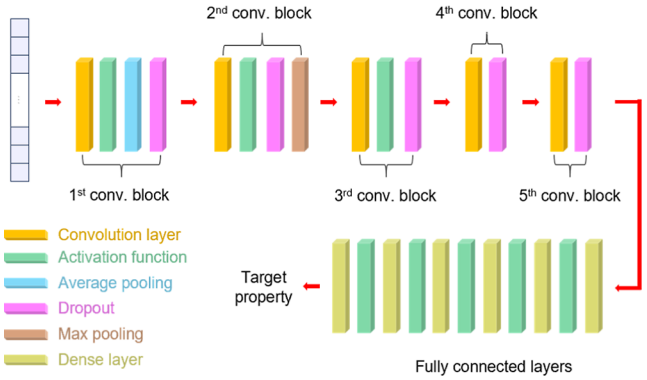


Figure 2. Architecture of PROFiT-Net. The descriptor matrix, which is constructed according to the scheme outlined in Figure 1, is flattened into a one-dimensional array and fed as input to PROFiT-Net. In the architecture diagram, the convolution layer, activation function, average pooling, dropout, max pooling, and dense layers are depicted in orange, green, blue, purple, brown, and yellow, respectively. Relevant parameters values are given in Table S2.

The distribution of this data set is illustrated in Figure S2a, with values ranging from 1.63 to 26.53. Figure 3a demonstrates the predictive power of PROFiT-Net for ε in terms of mean absolute error (MAE). Our model (MAE: 0.493) exhibits almost a 2-fold improvement compared to the OFMCNN model (MAE: 0.898). This suggests that concatenating additional elemental properties into the one-hot vector is crucial for accurately predicting ε.

When compared to CGCNN, MEGNet, and OGCNN, our model exhibits significantly improved performance. PROFiT-Net shows a 12−15% reduction in MAE compared to the other models, with MAE values for CGCNN, MEGNet, and OGCNN being 0.568, 0.554, and 0.582, respectively. Using MSE as an error metric, our model demonstrates a significantly reduced error (0.825) more than two to three times as small as CGCNN (MSE: 1.924), MEGNet (MSE: 1.686), OGCNN (MSE: 1.953) as well as OFM (MSE: 2.922) as shown in Figure 3b. Note that MSE is more sensitive to outliers than MAE, the characterization of which is critical for the wide distribution of dielectric constants. Particularly, the population of high dielectric materials is much smaller than that of

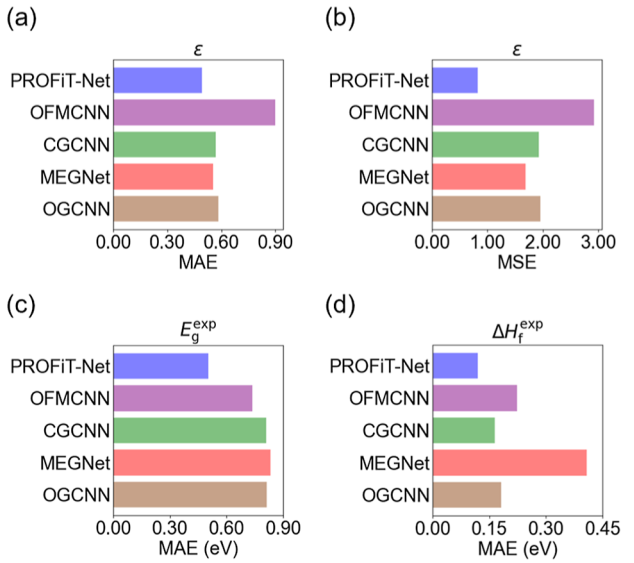


Figure 3. (a) MAEs (b) and MSEs of dielectric constant prediction (ε). MAEs of (c) band gap prediction (E_g^{exp}) and (d) formation enthalpy prediction (ΔH_f^{exp}) targeted to predict the experimental values. All results represent error metrics evaluated on the test sets. Corresponding scatter plots are displayed in Figure S3. PROFiT-Net, OFMCNN, CGCNN, MEGNet, and OGCNN are depicted in blue, purple, green, red, and brown, respectively.

moderate dielectric materials, it is particularly important to reliably predict outlying values. Therefore, PROFiT-Net exhibits tolerance to outliers and thus can be an effective tool for exploring unknown high dielectric materials.

Next, we assessed the predictive capability of PROFiT-Net for the band gap, E_g. We trained PROFiT-Net on three distinct data sets, namely E_g^{PBE}, E_g^{HSE06}, and E_g^{exp}. Note that the data fidelity increases progressively in the order of E_g^{PBE}, E_g^{HSE06}, E_g^{exp}. Figure S2b illustrates the data distributions of E_g^{PBE}. PROFiT-Net exhibits a 0.443 eV MAE, whereas the OFMCNN shows a 0.615 eV MAE (Table S8), suggesting again that the inclusion of elemental properties in the atomic feature vector is crucial for improving predictive performance. However, other graph-based models such as CGCNN, MEGNet, and OGCNN demonstrate comparable or slightly reduced MAE values of 0.421, 0.389, and 0.390 eV, respectively. It is noteworthy that our model shows a positive mean error (ME) value of 0.024 eV, whereas the MEs of CGCNN, MEGNet, and OGCNN shows negative values of −0.004, −0.012, and −0.033 eV, respectively. Given that PBE(+U)-level DFT typically underestimates the band gap, the overestimation trend of our model in predicting E_g^{PBE} can be viewed as a corrective tendency of PROFiT-Net.

When PROFiT-Net was trained on the E_g^{HSE06} data set with improved fidelity, as depicted in Figure S2c, it demonstrated the lowest MAE (0.513 eV) compared with the other models; OFMCNN shows a 0.657 eV MAE, whereas the MAEs of CGCNN, MEGNet, and OGCNN are 0.546, 0.534, and 0.513 eV, respectively (Table S8).

A more remarkable improvement was observed when our model was trained on the highest-fidelity E_g^{exp} data set (data distribution is shown in Figure S2d). Our model surpassed all other models by more than 0.24 eV in terms of MAE (Figure 3c)—the MAE of our model is 0.504 eV, whereas those of OFMCNN, CGCNN, MEGNet, and OGCNN are 0.738, 0.811, 0.832, and 0.814 eV, respectively. Considering these

https://doi.org/10.1021/jacs.4c05159
J. Am. Chem. Soc. 2024, 146, 26000−26007

results, we believe that our model shows continuously improved predictive ability as the fidelity of the training data increases, a trend attributed to the use of accurate elemental property features, reinforcing the aforementioned discussion. Additionally, it is interesting to observe that OFMCNN also shows a better performance compared with CGCNN, MEGNet, and OGCNN in predicting E_g^{exp}. This highlights the efficacy of employing valence electron interactions in the accurate prediction of electronic properties over graph-based representations.

Our final focus was predicting formation enthalpy, ΔH_f. For this purpose, we trained our model using ΔH_f^{PBE} and ΔH_f^{exp} data sets. The distributions of ΔH_f^{PBE} and ΔH_f^{exp} as shown in Figure S2e,f, respectively, display a predominance of negative values indicating that these data sets primarily consist of thermodynamically stable structures. In the prediction of ΔH_f^{PBE}, our model achieves a 0.053 eV MAE, representing a 3-fold reduction in error compared with OFMCNN (MAE: 0.159 eV). While this improvement is substantial, the performance is nearly on par with the graph-based models (Table S8): CGCNN records a 0.059 eV MAE, MEGNet demonstrates a 0.041 eV MAE, and OGCNN exhibits a 0.048 eV MAE.

However, when predicting the high-fidelity data of ΔH_f^{exp}, our model once again showcases a significantly smaller MAE (0.119 eV) in comparison with the other models (Figure 3d). This improvement is not only notable when contrasted with the MAE of OFMCNN (0.224 eV), but also when compared with the MAEs of CGCNN (0.165 eV), MEGNet (0.407 eV), and OGCNN (0.182 eV).

To further assess the tolerance to variation in data set fidelity, we compared the MAEs of various fidelities related to the prediction of band gaps (Figure 4a) and formation

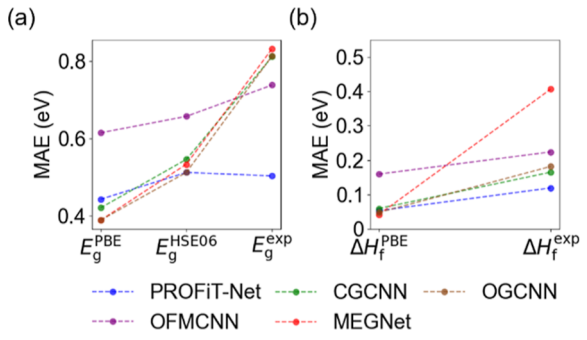


Figure 4. (a) MAEs of band gap prediction targeted to predict PBE values (E_g^{PBE}), HSE06 values (E_g^{HSE06}), and experimental values (E_g^{exp}). (b) MAEs of formation enthalpy per atom targeted to predict PBE values (ΔH_f^{PBE}) and experimental values (ΔH_f^{exp}). PROFiT-Net, OFMCNN, CGCNN, MEGNet, and OGCNN are depicted in blue, purple, green, red, and brown, respectively.

enthalpies (Figure 4b). Notably, our model demonstrated that PROFiT-Net outperforms the other models as the fidelity increases toward high-fidelity data sets, while in lowest fidelity data set the accuracy is on par with the state-of-the-art graph neural network (GNN). Thus, our model exhibits a high degree of tolerance to variations in data set quality, alleviating the need for reoptimization or the search for optimal hyperparameters when applying the model to a new data set.

It is noteworthy that PROFiT-Net demonstrates an accuracy level in predicting E_g^{exp} and ΔH_f^{exp} that either surpasses or is comparable to the accuracy achieved by DFT. When evaluated against the same experimental band gap data set, the MAEs of HSE and PBE0 hybrid-functional DFTs were recorded as 0.5 and 0.6 eV, respectively,[22] which are comparable to the MAE of PROFiT-Net (0.504 eV). In predicting ΔH_f^{exp}, the MAE of PBE(+U)-level DFT is 0.136 eV,[26] indicating that PROFiT-Net, with a smaller MAE of 0.119 eV, outperforms the PBE(+U)-DFT. This suggests that PROFiT-Net can reliably predict the band gap and thermodynamic stability of materials with similar or improved accuracy compared to DFT.

To explore the potential influence of hyperparameters on model performance, we reoptimized the hyperparameters of other graph-based CNN models using the ε data set used for the hyperparameter optimization of PROFiT-Net (Table S9 and Figures S4−S7). Even after this reoptimization, PROFiT-Net showed the best performance for predicting ε. Interestingly, the reoptimization improved the performance of MEGNet and CGCNN for predicting E_g^{HSE06}. However, PROFiT-Net still showed the best performance for predicting the highest fidelity data of E_g^{exp}, and the main conclusion therefore remained unchanged. We additionally benchmarked the CNN models using open databases comprising DFT- or DFPT-level data (Table S10), such as MatBench[27] and JARVIS-Leaderboard,[28] where PROFiT-Net still shows a comparable performance for the low-fidelity data sets.

For a deeper understanding of the origin of the improved performance, we analyzed feature importance using the Shapley additive explanation (SHAP) method.[29] A detailed procedure for quantifying the feature importance of PROFiT-Net is provided in the Supporting Information and Figure S8. Notably, PROFiT-Net is based on the pairwise interaction between two properties, defining feature importance for pairs of properties, which results in an 11 × 11 matrix (Figure S9).

Figure 5 shows the top five property pairs with the highest feature importance for each target material property. Overall, properties (X) paired with B, i.e., X-B, had significant contributions. This behavior was attributed to the general notion that material properties depend on their characteristic parameters, e.g., alkali/alkaline earth metals (s block), transition metals (d and f blocks), and nonmetals (p block).

Furthermore, we revealed the important contributions of E and PN to determining ε and E_g. One can reasonably expect a periodic trend of constituting the element's properties, which in turn affect the material's properties. Moreover, bond covalency, largely determined by E, is known to be strongly correlated with ε and E_g.[30,31] Interestingly, the importance of PN gradually increased as data set fidelity improved for E_g. Additionally, the importance of E and CR, both of which are vital physical quantities determining the bond characteristics and, consequently, thermodynamic stability of materials, became critical in predicting ΔH_f.

To investigate the scalability of our model, we expanded the chemical space from the given data sets to a large-scale database. We used PROFiT-Net to predict ε, E_g^{exp} and ΔH_f^{exp} for 154,718 materials from Materials Project database as well as CGCNN, MEGNet, and OGCNN. Figure 6a shows the correlation between ε and E_g^{exp} when using PROFiT-Net. This overall shape is similar to the Penn model where ε is inversely proportional to the square of E_g^{exp}.[32] This implies that our model can learn the physical trend of material properties.

Interestingly, PROFiT-Net does not predict negative band gap materials, whereas the other models predict unphysical negative band gaps. For instance, CGCNN, MEGNet, and

https://doi.org/10.1021/jacs.4c05159
J. Am. Chem. Soc. 2024, 146, 26000−26007

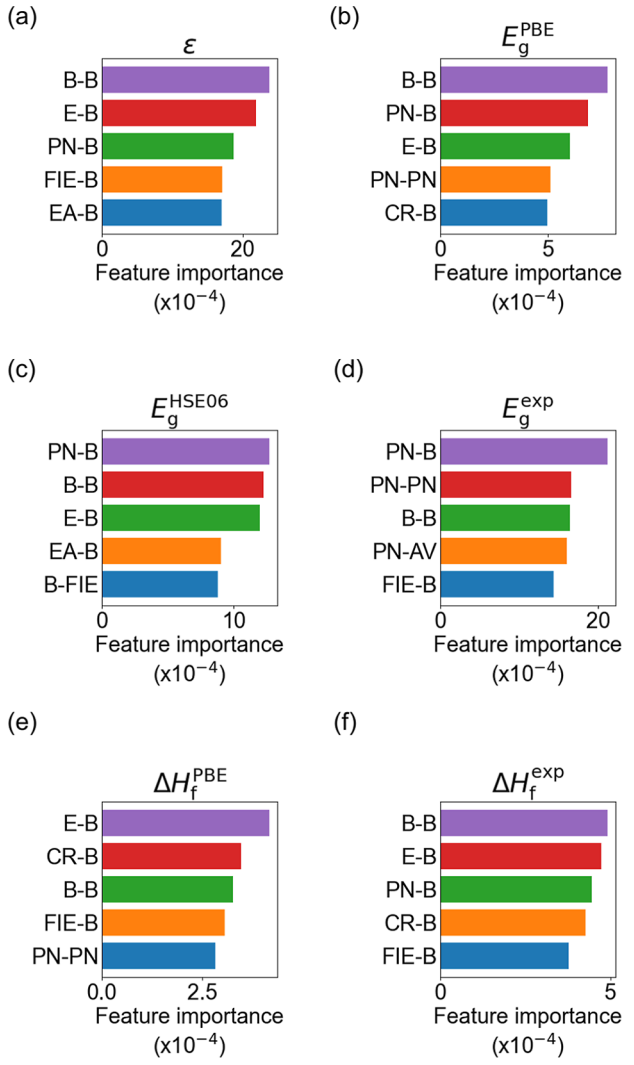


Figure 5. Top five property pairs with the highest feature importance for (a) ε, (b) PBE(+U)-level DFT band gaps (E_g^{PBE}), (c) HSE06-level DFT band gaps (E_g^{HSE06}), (d) experimental band gaps (E_g^{exp}), (e) PBE(+U)-level DFT formation enthalpies per atom (ΔH_f^{PBE}), and (f) experimental formation enthalpies (ΔH_f^{exp}).

OGCNN predicted 577, 1739, and 2613 materials with negative band gaps, respectively. Remarkably, this trend correlates with the trend of ME observed in predicting the band gap trained against the PBE-level data set (as discussed earlier); the model exhibiting more negative ME values in predicting the low-fidelity data tends to produce more negative band gaps when predicting the high-fidelity data.

Next, we investigated the pairwise intercorrelations between PROFiT-Net and other models, computing Pearson correlation coefficients (R) for ε, E_g^{exp}, and ΔH_f^{exp}. In predicting ε, PROFiT-Net correlated strongly with other models (Figure 6b); R values with CGCNN, MEGNet, and OGCNN are 0.935, 0.816, and 0.929, respectively. This mutual validation supports the reliability of predicting ε using various DL models. However, weaker correlations among the various ML models were observed in predicting the other properties, i.e., ΔH_f^{exp} and E_g^{exp}. For instance, the correlation between PROFiT-Net and MEGNet in predicting ΔH_f^{exp} demonstrated a reduced R value of 0.636 (Figure 6c). In predicting E_g^{exp}, the correlations of CGCNN, MEGNet, and OGCNN with

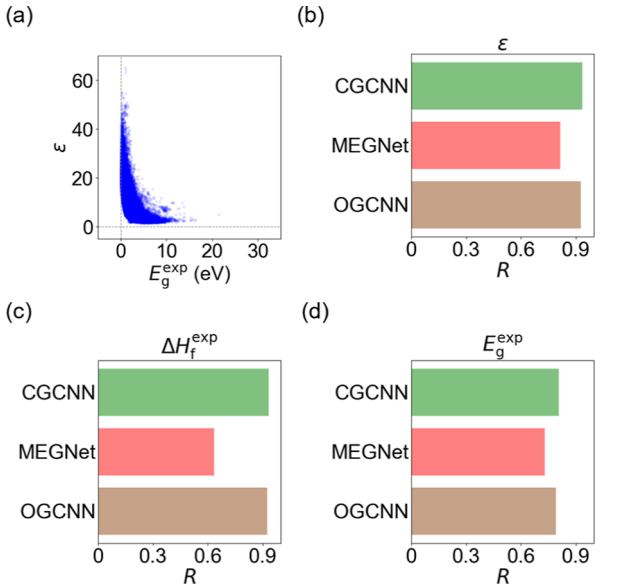


Figure 6. (a) Scatter plots of dielectric constants (ε) versus experimental band gaps (E_g^{exp}) predicted by using PROFiT-Net. Those of CGCNN, MEGNet, and OGCNN are represented in Figure S10. Pearson correlation coefficient (R) of predicted (b) ε, (c) ΔH_f^{exp}, and (d) E_g^{exp} between PROFiT-Net and other DL models including CGCNN (green), MEGNet (red), and OGCNN (brown). A total of 154,718 different materials were queried in the Materials Project.[18] Each scatter plot is shown in Figure S11. The slopes, intercepts, and Pearson correlation coefficients of the linear fits are tabulated in Table S11.

PROFiT-Net are also weakened, resulting in R values of 0.809, 0.730, and 0.793, respectively (Figure 6d). This clearly illustrates the distinctiveness of PROFiT-Net, particularly when it is trained using the high-fidelity data set, compared with the other DL models.

4. CONCLUSION

We developed PROFiT-Net, a DL model based on CNNs, capable of accurately predicting various material properties such as the dielectric constant, band gap, and formation enthalpy. This model encodes the interrelation between atomic and electronic properties of neighboring atoms, which is incorporated into the conventional CNN architecture comprising of sequentially arranged convolution layers and dense layers. PROFiT-Net exhibits improved performance over various GNN models, especially in data sets with high fidelity, and successfully reflects physical trends like predicting non-negative DFT band gaps and Penn-like models.

Currently, PROFiT-Net includes only interactions between neighboring atoms and uses a simple sequential CNN model. However, it can be further enhanced by increasing the dimension of the bond matrix to include many-body interactions and/or by refining the CNN architecture (such as customizing the convolution layer or autoencoder). These are areas we plan to develop in the future. Our model also maintains consistent accuracy across data sets of different sizes, suggesting that techniques like multifidelity or transfer learning could further enhance its performance. PROFiT-Net is a versatile tool capable of predicting various material properties represented by single scalar values. It exhibits favorable behaviors such as fidelity tolerance and scalability. Thus, we envisage that it can be widely used to predict new functional

https://doi.org/10.1021/jacs.4c05159
J. Am. Chem. Soc. 2024, 146, 26000−26007

materials, leading to advancements in material design and discovery.

■ ASSOCIATED CONTENT

Data Availability Statement

PROFiT-Net and data set are available at https://github.com/sejunkim6370/PROFiT-Net.

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/jacs.4c05159.

Application of SHAP method to PROFiT-Net, list of atomic features, detailed structures of PROFiT-Net with relevant parameter values, data set information, address, version, hyperparameters, and atomic features of other models, total number of model parameters, MAE and MSE values of test results without and with hyperparameter optimizations, MAE and MSE values of test results using open databases, linear fit results, prediction time, histograms, scatter plots, hyperparameter optimization results, the process to obtain the feature importance, and the feature importance analysis (PDF)

■ AUTHOR INFORMATION

Corresponding Authors

Eok Kyun Lee − *Department of Chemistry, Korea Advanced Institute of Science and Technology (KAIST), Daejeon 34141, South Korea*; Email: eklee@kaist.ac.kr

Hyungjun Kim − *Department of Chemistry, Korea Advanced Institute of Science and Technology (KAIST), Daejeon 34141, South Korea*; orcid.org/0000-0001-8261-9381; Email: linus16@kaist.ac.kr

Authors

Se-Jun Kim − *Department of Chemistry, Korea Advanced Institute of Science and Technology (KAIST), Daejeon 34141, South Korea*

Won June Kim − *Department of Biology and Chemistry, Changwon National University, Changwon-si, Gyeongsangnam-do 51140, South Korea*; orcid.org/0000-0001-6421-9237

Changho Kim − *Department of Applied Mathematics, University of California, Merced, California 95343, United States*; orcid.org/0000-0002-4064-8237

Complete contact information is available at: https://pubs.acs.org/10.1021/jacs.4c05159

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work was supported by the National Research Foundation of Korea (NRF), grants funded by the Korean government (MSIT) (nos. RS-2024-00450836 and RS-2024-00435493).

■ ABBREVIATIONS

AI	artificial intelligence
CNN	convolution neural network
CGCNN	crystal graph convolutional neural network
DFT	density functional theory
DL	deep learning
MAE	mean absolute error
ML	machine learning
OFM	orbital field matrix
OGCNN	orbital field matrix convolutional neural network
PROFiT-Net	PRoperty-networking Orbital Field maTrix-convolutional neural Network

■ REFERENCES

(1) Pyzer-Knapp, E. O.; Pitera, J. W.; Staar, P. W. J.; Takeda, S.; Laino, T.; Sanders, D. P.; Sexton, J.; Smith, J. R.; Curioni, A. Accelerating materials discovery using artificial intelligence, high performance computing and robotics. *npj Comput. Mater.* **2022**, *8* (1), 84.

(2) Maier, W. F.; Stöwe, K.; Sieg, S. Combinatorial and high-throughput materials science. *Angew. Chem., Int. Ed.* **2007**, *46* (32), 6016−6067.

(3) Schmidt, J.; Marques, M. R. G.; Botti, S.; Marques, M. A. L. Recent advances and applications of machine learning in solid-state materials science. *npj Comput. Mater.* **2019**, *5* (1), 83.

(4) Choudhary, K.; DeCost, B.; Chen, C.; Jain, A.; Tavazza, F.; Cohn, R.; Park, C. W.; Choudhary, A.; Agrawal, A.; Billinge, S. J. L.; et al. Recent advances and applications of deep learning methods in materials science. *npj Comput. Mater.* **2022**, *8* (1), 59.

(5) Rupp, M.; Tkatchenko, A.; Müller, K. R.; von Lilienfeld, O. A. fast and accurate modeling of molecular atomization energies with machine learning. *Phys. Rev. Lett.* **2012**, *108* (5), 058301.

(6) Behler, J. Atom-centered symmetry functions for constructing high-dimensional neural network potentials. *J. Chem. Phys.* **2011**, *134*, 074106.

(7) De, S.; Bartók, A. P.; Csányi, G.; Ceriotti, M. Comparing molecules and solids across structural and alchemical space. *Phys. Chem. Chem. Phys.* **2016**, *18* (20), 13754−13769.

(8) Huo, H.; Rupp, M. Unified representation of molecules and crystals for machine learning. *Mach. Learn.: Sci. Technol.* **2022**, *3* (4), 045017.

(9) Reiser, P.; Neubert, M.; Eberhard, A.; Torresi, L.; Zhou, C.; Shao, C.; Metni, H.; van Hoesel, C.; Schopmans, H.; Sommer, T.; et al. Graph neural networks for materials science and chemistry. *Commun. Mater.* **2022**, *3* (1), 93.

(10) Xie, T.; Grossman, J. C. Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. *Phys. Rev. Lett.* **2018**, *120* (14), 145301.

(11) 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* (9), 3564−3572.

(12) Karamad, M.; Magar, R.; Shi, Y.; Siahrostami, S.; Gates, I. D.; Barati Farimani, A. Orbital graph convolutional neural network for material property prediction. *Phys. Rev. Mater.* **2020**, *4* (9), 093801.

(13) Lam Pham, T.; Kino, H.; Terakura, K.; Miyake, T.; Tsuda, K.; Takigawa, I.; Chi Dam, H. Machine learning reveals orbital interaction in materials. *Sci. Technol. Adv. Mater.* **2017**, *18* (1), 756−765.

(14) O'Keeffe, M. A proposed rigorous definition of coordination number. *Acta Cryst. A* **1979**, *35* (5), 772−775.

(15) Abadi, M.; Agarwal, A.; Barham, P.; Brevdo, E.; Chen, Z.; Citro, C.; Corrado, G. S.; Davis, A.; Dean, J.; Devin, M.; et al. TensorFlow: Large-scale machine learning on heterogeneous systems. https://www.tensorflow.org (accessed 01 30, 2021).

(16) Chollet, F. Keras. https://keras.io (accessed 01 30, 2021).

(17) Agarap, A. F. Deep learning using rectified linear units (ReLU). *arXiv (Computer Science/Neural and Evolutionary Computing)* **2018**, arXiv:1803.08375.

(18) Jain, A.; Ong, S. P.; Hautier, G.; Chen, W.; Richards, W. D.; Dacek, S.; Cholia, S.; Gunter, D.; Skinner, D.; Ceder, G.; Persson, K. A. Commentary: The Materials Project: A materials genome approach to accelerating materials innovation. *APL Mater.* **2013**, *1* (1), 011002.

(19) Morita, K.; Davies, D. W.; Butler, K. T.; Walsh, A. Modeling the dielectric constants of crystals using machine learning. *J. Chem. Phys.* **2020**, *153* (2), 024503.
(20) Krukau, A. V.; Vydrov, O. A.; Izmaylov, A. F.; Scuseria, G. E. Influence of the exchange screening parameter on the performance of screened hybrid functionals. *J. Chem. Phys.* **2006**, *125* (22), 224106.
(21) Kim, S.; Lee, M.; Hong, C.; Yoon, Y.; An, H.; Lee, D.; Jeong, W.; Yoo, D.; Kang, Y.; Youn, Y.; Han, S. A band-gap database for semiconducting inorganic materials calculated with hybrid functional. *Sci. Data* **2020**, *7* (1), 387.
(22) Borlido, P.; Aull, T.; Huran, A. W.; Tran, F.; Marques, M. A. L.; Botti, S. Large-scale benchmark of exchange−correlation functionals for the determination of electronic band gaps of solids. *J. Chem. Theory Comput.* **2019**, *15* (9), 5069−5079.
(23) Gong, S.; Wang, S.; Xie, T.; Chae, W. H.; Liu, R.; Shao-Horn, Y.; Grossman, J. C. Calibrating DFT formation enthalpy calculations by multifidelity machine learning. *JACS Au* **2022**, *2* (9), 1964−1977.
(24) Ong, S. P.; Richards, W. D.; Jain, A.; Hautier, G.; Kocher, M.; Cholia, S.; Gunter, D.; Chevrier, V. L.; Persson, K. A.; Ceder, G. Python Materials Genomics (pymatgen): A robust, open-source python library for materials analysis. *Comput. Mater. Sci.* **2013**, *68*, 314−319.
(25) Ward, L.; Dunn, A.; Faghaninia, A.; Zimmermann, N. E. R.; 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.
(26) Kirklin, S.; Saal, J. E.; Meredig, B.; Thompson, A.; Doak, J. W.; Aykol, M.; Rühl, S.; Wolverton, C. The Open Quantum Materials Database (OQMD): assessing the accuracy of DFT formation energies. *npj Comput. Mater.* **2015**, *1* (1), 15010.
(27) 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* (1), 138.
(28) Choudhary, K.; Wines, D.; Li, K.; Garrity, K. F.; Gupta, V.; Romero, A. H.; Krogel, J. T.; Saritas, K.; Fuhr, A.; Ganesh, P.; et al. JARVIS-Leaderboard: A large scale benchmark of materials design methods. *npj Comput. Mater.* **2024**, *10* (1), 93.
(29) Lundberg, S. M.; Lee, S.-I. A unified approach to interpreting model predictions. *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 2017; pp 4768−4777.
(30) Duffy, J. A. Trends in energy gaps of binary compounds: an approach based upon electron transfer parameters from optical spectroscopy. *J. Phys. C: Solid State Phys.* **1980**, *13* (16), 2979−2989.
(31) Reddy, D. R. R.; Nazeer Ahammed, Y. Relationship between refractive index, optical electronegativities and electronic polarizability in alkali halides, III−V, II−VI group semiconductors. *Cryst. Res. Technol.* **1995**, *30* (2), 263−266.
(32) Penn, D. R. Wave-number-dependent dielectric function of semiconductors. *Phys. Rev.* **1962**, *128* (5), 2093−2097.

https://doi.org/10.1021/jacs.4c05159
J. Am. Chem. Soc. 2024, 146, 26000−26007