THE JOURNAL OF PHYSICAL CHEMISTRY LETTERS
A JOURNAL OF THE AMERICAN CHEMICAL SOCIETY

pubs.acs.org/JPCL

Letter

Deep Learning-Based Prediction of Material Properties Using Chemical Compositions and Diffraction Patterns as Experimentally Accessible Inputs

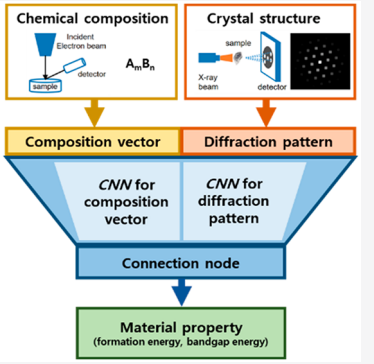
Jeongrae Kim,[†] Leslie Ching Ow Tiong,[†] Donghun Kim,[*] and Sang Soo Han[*]

Cite This: *J. Phys. Chem. Lett.* 2021, 12, 8376–8383

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: We report a deep learning (DL) model that predicts various material properties while accepting directly accessible inputs from routine experimental platforms: chemical compositions and diffraction data, which can be obtained from the X-ray or electron-beam diffraction and energy-dispersive spectroscopy, respectively. These heterogeneous forms of inputs are treated simultaneously in our DL model, where the novel chemical composition vector is proposed by developing element embedding with the normalized composition matrix. With 1524 binary samples available in the Materials Project database, the model predicts formation energies and band gaps with mean absolute errors of 0.29 eV/atom and 0.66 eV, respectively. According to the weighing test between these two inputs, the properties tend to be more influenced by the chemical composition than the crystal structure. This work intentionally avoids using inputs that are not directly accessible (e.g., atomic coordinates) in experimental platforms, and thus is expected to substantially improve the practical use of DL models.



Data/AI-driven research has emerged as a method accelerating the discovery of novel materials. In particular, deep learning (DL) techniques are proven to be effective characterization tools owing to their superior capabilities to disentangle complex structure−property relationships. The largest-scale, openly available material databases are mainly composed of results obtained from computer simulations rather than experiments. Examples include the Materials Project,[1] Novel Materials Discovery (NOMAD),[2] and Open Quantum Materials Database (OQMD).[3] As these computer-simulation-based databases provide constituent atom types and their coordinates as basic structural information, tremendous efforts have been put into developing DL models with these coordinates as key input descriptors. For example, Xie and Grossman[4] proposed crystal graph convolutional neural networks (CGCNNs), where graphs constructed from atomic positions are effective in predicting bulk material properties such as formation energy and band gap energy. Kim et al.[5] and Gu et al.[6] used the atomic coordinates of catalysts and adsorbates to predict their interaction energy during catalysis. Although these DL models excellently correlate with various material properties, the required input features (e.g., atomic coordinates) are not directly accessible during experimental studies; thus, an additional analysis or experiment is required to determine them. Accordingly, the practicality of these models can be unfortunately lacking from an experimental standpoint.

In this regard, to improve the practical use of DL models, it is critical to reformulate the problem using readily accessible inputs in our routine experimental platforms. For experimentalists to identify atomic structural information on a synthesized sample, two basic characterization methods are available: X-ray or electron diffraction techniques and energy dispersive spectroscopy (EDS). The former enables the determination of crystal symmetries (e.g., space group), and the latter provides chemical formulas and compositions (e.g., $A_XB_YC_Z$). For this reasoning, DL models in this work utilize the results obtained from these two basic characterization methods instead of the exact coordinates of constituent atoms, with the goal of predicting various material properties (formation energy, band gap energy, etc.).

Recent studies report the development of machine learning models accepting either diffraction results or chemical compositions, and indeed their treatments have quickly evolved. First, regarding diffraction data, Park et al.,[7] Vecsei et al.,[8] Wang et al.,[9] and Oviedo et al.[10] used one-dimensional (1D) powder X-ray diffraction (XRD) curves, for which information such as peak positions, intensities, and full widths at half-maximum (FWHM) were mainly treated as the key input values. In addition, Ziletti et al.,[11] Aguiar et al.,[12] and Kaufmann et al.[13] treated electron-beam-based 2D diffraction

Received: July 17, 2021
Accepted: August 23, 2021

https://doi.org/10.1021/acs.jpclett.1c02305
J. Phys. Chem. Lett. 2021, 12, 8376–8383

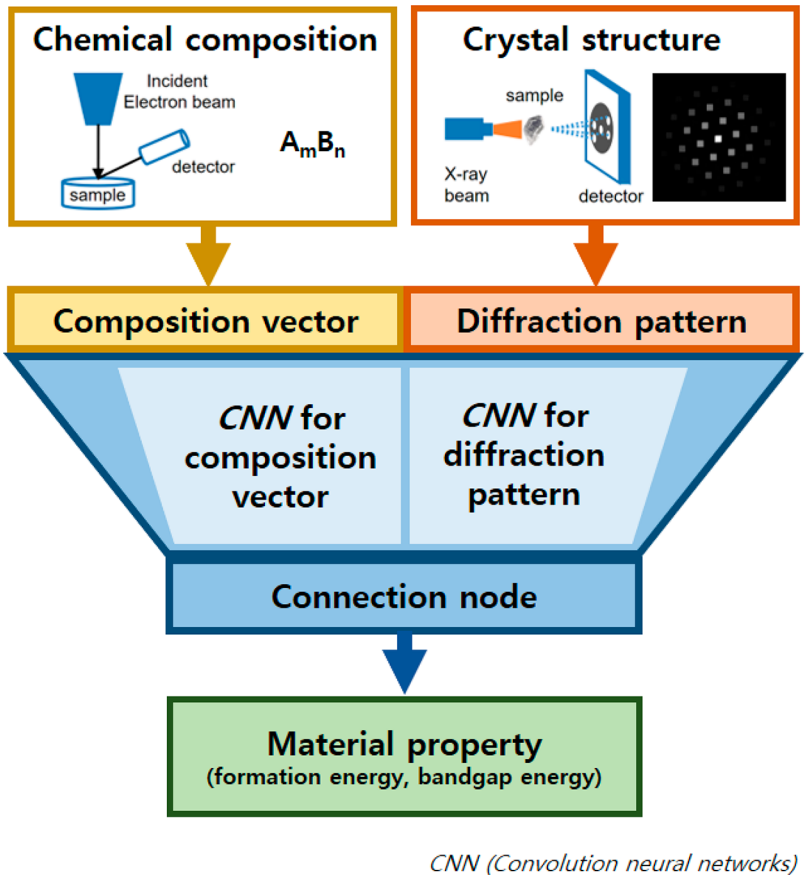


Figure 1. Schematic structure of the proposed DL model. The model accepts the inputs of the composition vector and diffraction pattern. Two independent CNNs apply to each input, and the results are connected at the connection node stage. The output properties of interest are the formation energy and band gap energy.

patterns in a raw image format. Second, for chemical composition data, Zhou et al.[14] created the *atom2vec* vector representing the chemical formula of a material using the atom embedding method. Tshitoyan et al.[15] similarly introduced elemental embedding using the *word2vec* technique when processing the Abstract section of a large volume of literature (over 3.3 million articles). Despite the large contributions made by these studies, the development of DL models accepting heterogeneous types of both chemical composition and diffraction patterns is highly limited today. The only attempt thus far, to our understanding, is the recent report by Aguiar et al. where a DL model that concatenates a diffraction module and a chemistry module is presented.[16] However, this study is still limited to a crystallographic prediction purpose, and the effectiveness of the combined modules has yet to be validated for other various material properties.

With the aim of predicting various material properties, we propose a practical DL model that accepts heterogeneous types of chemical compositions (texts) and diffraction patterns (images), both of which are readily accessible on experimental platforms (EDS measurement or diffraction equipment). We develop the unique chemical composition vector using a newly proposed embedding method named element embedding with the normalized composition matrix (EENCM). We also use 2D diffraction images where the results of three orthogonal beam axes are superimposed. These two heterogeneous data are simultaneously treated in the model. With 1528 binary oxide, sulfide, fluoride, and nitride samples available in the Materials Project database,[1] our model predicts the formation energies and band gaps with mean absolute errors (MAEs) of 0.29 eV/atom and 0.66 eV, respectively. We note that these accuracies are not as good as the CGCNN model, which reports MAEs of as low as 0.04 eV/atom and 0.39 eV for the same samples.[4] However, in contrast to their model, ours avoids using the exact coordinates of constituent atoms, which substantially improves the practicality of the method. Additionally, the weighing tests between two different types of inputs reveal that the investigated properties tend to be much more influenced by chemical compositions than by crystal structures.

Figure 1 shows the schematic structure of the proposed DL model for predicting material properties. The model accepts two inputs: chemical composition and diffraction data. The combination of these two data sets is generally sufficient to assign a unique material in inorganic crystal databases. Two independent convolutional neural networks (CNNs)[17−20] apply to each input, and the results are connected at the connection node stage. The inputs of chemical composition and diffraction images are apparently heterogeneous in terms of both dimensions and value ranges; hence, they should be treated carefully when merged. For example, these two inputs should be appropriately normalized so that the components can have similar ranges of values (0−0.5) and tested with different weights for optimized learning results. Although the output of the model can be any material property, we consider the formation energy and band gap as our targets in this study, as they are representative thermodynamic and electronic properties of a material, respectively. Details on the CNN

https://doi.org/10.1021/acs.jpclett.1c02305
J. Phys. Chem. Lett. 2021, 12, 8376−8383

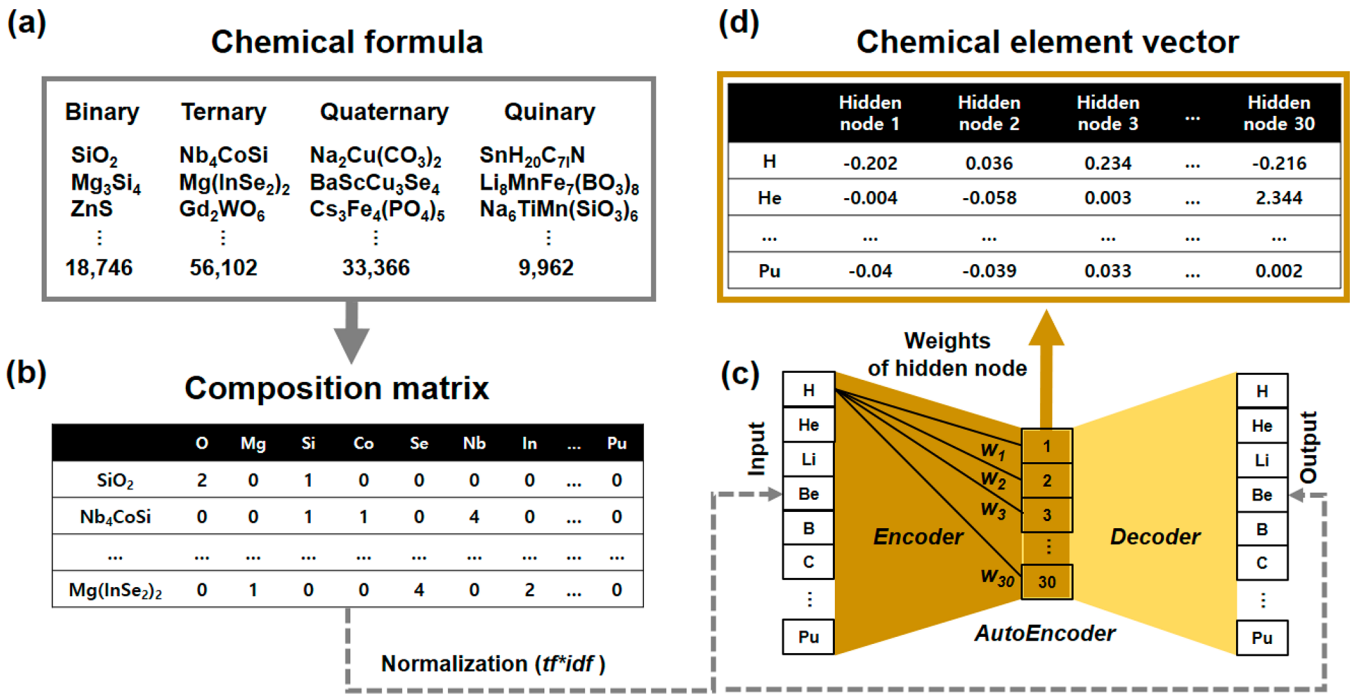


Figure 2. Architecture of the EENCM for the creation of chemical element vectors. The EENCM creates a unique chemical element vector by applying the chemical formula of a material to the AutoEncoder. The chemical formula (a) is transformed into a composition matrix (b) to train the chemical composition information on the material. The composition matrix (b) is normalized by the term frequency-inverse document frequency (*tf***idf*) technique for training the AutoEncoder (c). The trained network weights of the AutoEncoder (c) include information on the learning characteristics of the elements constituting the material. The EENCM uses the weights of the AutoEncoder (c) as the chemical element vector (d).

architecture shown in Figure 1 are additionally explained in the Supporting Information.

To develop the novel composition vector, we first create a chemical element vector, which represents a chemical element, through element embedding. For our new element embedding method, several studies, such as Zhou et al.,[14] Tshitoyan et al.,[15] and Herr et al.[21] have been referenced. In particular, Zhou et al.[14] generated one-hot encoded data sets consisting of elements and environments of the material formula and embedded elements through single-value decomposition and a probability model. Tshitoyan et al.[15] extracted elemental information through natural language processing of published papers to generate a one-hot encoded data set and embedded elements through *word2vec*.[22] We created the data set to clearly include the chemical composition information on a material and try to prevent bias that may occur in the case of one-hot encoded data during the training process for embedding. To develop the unique composition vector, we propose a novel chemical element embedding method, namely, the element embedding with the normalized composition matrix (EENCM). EENCM is a method to represent each element as a unique chemical element vector by training a large volume of chemical formula data (e.g., $A_XB_YC_Z$). In the EENCM shown in Figure 2, the chemical formula data in the Materials Project database[1] were used as source data to learn the properties of chemical bonding between elements. Out of a total of 120 612 materials, 118 176 (~98%) are binary (2 elements) to quinary (5 elements) materials, and only 2% are either unary or composed of more than 5 elements (Figure S1, Supporting Information). The 98% portion of samples (binary to quinary materials) are used in the EENCM training process.

In Figure 2b, the composition matrix is described, where the rows contain the chemical formula of each material (118 176 rows) and the columns contain the elements constituting the chemical formula (87 columns). With SiO_2 as an example to explain how to fill the matrix elements, 1 and 2 are put into the Si and O columns, respectively, and all other parts are filled with a zero. In contrast to previous similar efforts,[14,15] the term frequency-inverse document frequency (*tf***idf*) technique,[23] which is widely used in the text mining field, is implemented for data normalization in the EENCM method to prevent biased training for specific elements and materials. Details about implementing the *tf***idf* technique are available in the Supporting Information.

The composition matrix after the normalization process is used in the AutoEncoder[24,25] is shown in Figure 2c. This AutoEncoder trains the characteristics of the elements that make up the formula of the material. Then, the 30 trained weight parameters of the encoder part were defined as each chemical element vector, as shown in Figure 2d. The dimension (nodes of the hidden layer in EENCM) was set to 30 after extensive testing of different numbers ranging from 3 to 40 (Figure S2, Supporting Information). Once each chemical element vector was ready based on the above, the composition vector for the material A_mB_n can be defined as follows:

$$\text{Composition vector} = \left[\frac{m}{m+n} \times \vec{A}, \frac{n}{m+n} \times \vec{B}\right] \quad (1)$$

where $\vec{A}$ and $\vec{B}$ represent the chemical element vectors of elements A and B. As the element vectors have dimensions of 1 × 30, the composition vectors for binary and ternary materials have dimensions of 1 × 60 and 1 × 90, respectively. To

https://doi.org/10.1021/acs.jpclett.1c02305
J. Phys. Chem. Lett. 2021, 12, 8376−8383

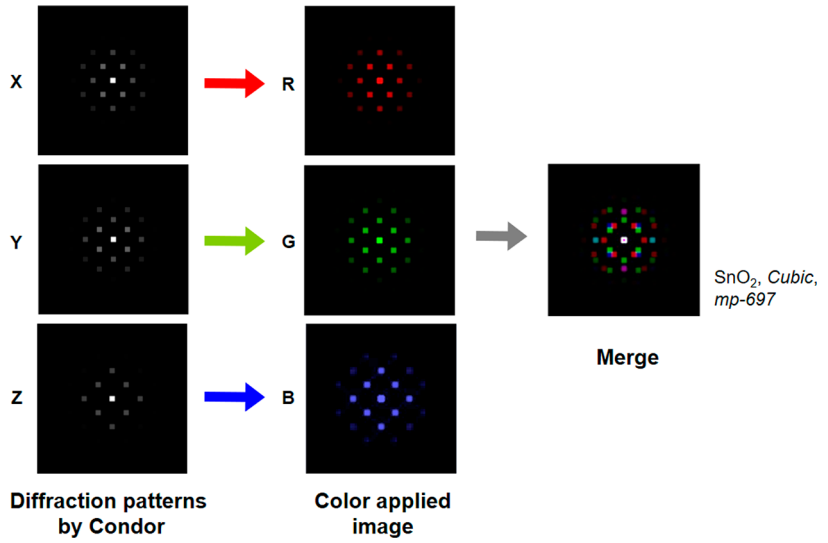


Figure 3. Preprocessing diffraction patterns. Diffraction pattern images containing structural information are generated by Condor software.[26] Incident beams are injected in three directions for each material to generate diffraction patterns in three directions as images of the *x*-, *y*-, and *z*-axes. We apply red, green, and blue colors to recognize the directionality and re-express three images into one image. This sample is a *cubic* structure of SnO_2 (Materials Project ID is *mp-697*).

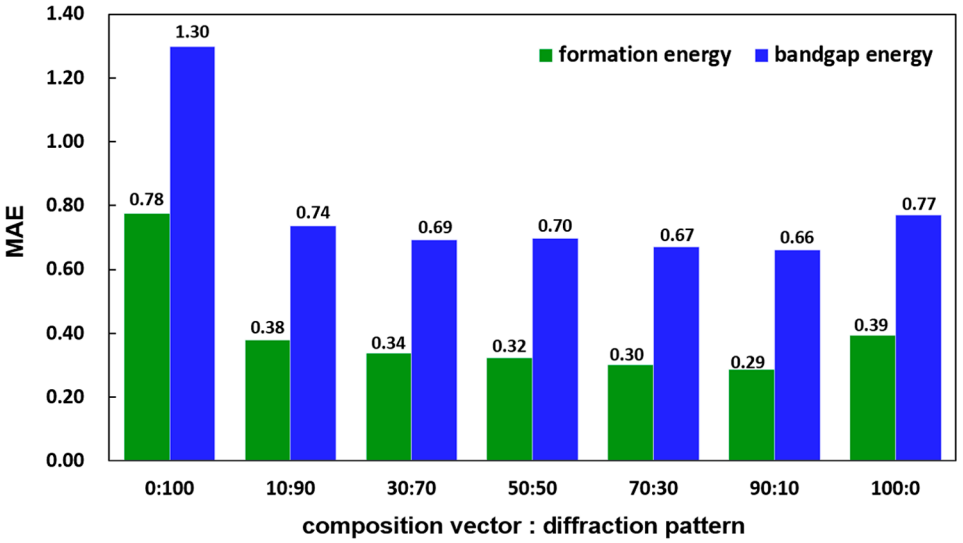


Figure 4. Performance of our DL models. This figure shows the performance for predicting the formation energies (green) and band gaps (blue) of materials with the ratio of the composition vector and diffraction pattern on the connection node stage in our DL architecture.

determine whether our composition vector performs better than previous approaches in property prediction, we carried out the following tests. For 1492 materials in an *elpasolite* structure,[26] we developed a deep learning method using the standard CNN model to predict the formation energy. Regarding the formation energy data, our deep learning model employing EENCM achieves a mean absolute error (MAE) of 0.054 ± 0.019 eV/atom (Figure S3, Supporting Information), which is lower than the comparison models (0.15 eV/atom of Zhou et al.[14]) and comparable to the 0.056 eV/atom of Tshitoyan et al.[15]

Along with the aforementioned composition vector, a diffraction pattern is the other equally important input in our DL model. The diffraction pattern has been mainly studied to analyze information on the structure of a crystalline material; thus, it is used for training along with the composition vector as crystal structure information. We used the simulator introduced in the literature to overcome the difficulty of generating an actual diffraction pattern image data set; thus, Condor software[27] was used to simulate the diffraction pattern images. We preprocessed the diffraction pattern images by benchmarking the related study of Ziletti et al.[11] Figure 3 helps explain image preprocessing. To recognize the three orthogonal zone axes of *x*, *y*, and *z*, grayscale images were converted to red (R), green (G), and blue (B), taking into account the degree of brightness. Then, using a merging method, the diffraction patterns in the three directions were converted into a single image (128 × 128 dimensions). The Condor settings are described in the Supporting Information.

To test our DL model for the prediction of material properties from the composition vectors and the diffraction pattern images, we considered the formation energies and band gaps of inorganic semiconducting materials as the representative thermodynamic and electronic properties of materials,

https://doi.org/10.1021/acs.jpclett.1c02305
J. Phys. Chem. Lett. 2021, 12, 8376−8383

Table 1. Formation Energies (eV/atom) (a) and Band Gap Energies (eV) (b) of Various Materials[a]

	(a) Formation Energy [eV/atom]							
	triclinic	monoclinic	orthorhombic	tetragonal	trigonal	hexagonal	cubic	SD
TiO_2(34)	−3.48(1)	−3.38(6)	−3.40(15)	−3.48(3)	−3.39(4)	−3.31(2)	−3.28(3)	0.08
ZrO_2(18)	-	−3.79(3)	−3.79(10)	−3.80(4)	-	-	−3.77(1)	0.01
Al_2O_3(20)	−3.31(5)	−3.39(7)	−3.34(4)	-	−3.32(3)	-	−3.42(1)	0.05
RbS(4)	-	−0.23(1)	−1.14(2)	-	-	−1.22(1)	-	0.55
ScF_3(7)	-	−3.71(2)	−3.81(2)	-	−4.32(2)	-	−4.32(1)	0.33
NaN_3(5)	-	0.29(3)	-	-	−0.38(2)	-	-	0.48
SD	0.12	1.87	1.12	0.22	1.71	1.47	0.46	0.25/1.00
	(b) Band Gap [eV]							
	triclinic	monoclinic	orthorhombic	tetragonal	trigonal	hexagonal	cubic	SD
TiO_2(34)	3.31(1)	2.14(6)	2.47(15)	2.13(3)	2.57(4)	1.60(2)	1.41(3)	0.63
ZrO_2(18)	-	3.85(3)	3.51(10)	3.59(4)	-	-	3.13(1)	0.29
Al_2O_3(20)	3.61(5)	4.37(7)	4.66(4)	-	4.77(3)	-	5.22(1)	0.59
RbS(4)	-	0.35(1)	1.21(2)	-	-	1.59(1)	-	0.64
ScF_3(7)	-	3.52(2)	3.77(2)	-	6.08(2)	-	6.08(1)	1.41
NaN_3(5)	-	1.88(3)	-	-	4.09(2)	-	-	1.56
SD	0.21	1.50	1.32	1.03	1.46	0.01	2.10	0.86/1.09

[a]Here, the property values indicate the averaged values of the materials classified by the chemical compositions and crystal structures. The standard deviations (SDs) of the averaged values are also included. The values in parentheses indicate the material numbers.

respectively. Here, 1524 binary systems such as metal oxides (839), sulfides (328), fluorides (192), and nitrides (165) were considered, whereas materials with zero band gaps were ignored (Figure S4, Supporting Information). In evaluating the proposed DL model, the data were divided into a training set:test set = 90:10, and the model was verified using k-fold cross-validation (k = 10). The prediction performance of the proposed DL model is shown in Figure 4, in which the MAEs for the formation energy and band gap are presented with the ratio of the composition vector and diffraction pattern on the connection node stage of our DL architecture. The relevant details are explained in the Supporting Information. Through this approach, we investigated the relative contributions of the crystal structures (diffraction patterns) and compositions (composition vectors) of materials to their properties.

In Figure 4, the DL model considering only the diffraction pattern images shows high MAE values (1.30 eV for band gaps and 0.78 eV/atom for formation energies). On the other hand, the DL model considering only the composition vectors significantly lowers the MAE values (0.77 eV for band gaps and 0.39 eV/atom for formation energies), implying that material properties are likely more affected by the chemical composition than by the crystal structure. However, if two materials have identical chemical compositions but different crystal structures, the DL model considering only the composition vectors provides the same property values for the two materials because it does not reflect the crystal structure information. Because of this, we observe horizontally or vertically distributed data in the scatterplots showing comparisons of the predicted and actual values in Figure S5, Supporting Information. From these results, it is obvious that the DL model considering only the composition vectors or the diffraction patterns is limited in its prediction of material properties. Accordingly, we consider the contributions of the two input features (composition vector and diffraction pattern) in the DL model and investigate the performance of the DL model as a function of the ratio between the input features (composition vector:diffraction pattern = 10:90, 30:70, 50:50, 70:30, and 90:10). Irrespective of the ratio, the DL models considering the two features show superior performance over those considering only one feature. The best performance is observed with the composition vector:diffraction pattern ratio as 90:10, which shows MAE values of 0.66 eV for the band gaps and 0.29 eV/atom for the formation energies; these values are an improvement of 49.23% for the band gap and 62.82% for the formation energies in comparison to the DL model considering only the diffraction patterns. This result also reveals that the material properties are determined by both the crystal structures and compositions; however, it is more influenced by the composition than by the crystal structure.

We also investigated the prediction accuracy over seven types of Bravais lattice systems (*triclinic*, *monoclinic*, *orthorhombic*, *tetragonal*, *trigonal*, *hexagonal*, and *cubic*) to understand any specific gains and losses in each crystal family. Table S2 in the Supporting Information shows the results of the extended analyses over seven crystal families. One of the most prominent observations is that the prediction accuracy (MAE) is the worst for the *cubic* systems. This is true for both formation energy (0.06 eV/atom larger for the *cubic* than for the overall) and band gap energy (0.18 eV larger for the cubic than for the overall). This is likely attributed to the highest crystal symmetry of the cubic systems, which will cause more visual resemblance of diffraction patterns. It may be difficult for the model to learn the structural subtle differences between materials within the *cubic* systems. As a result, the DL model performance is found the worst for the *cubic* system whereas it is relatively better for the lower symmetry systems such as *triclinic*, *trigonal*, and *hexagonal* systems.

In addition, we also investigated whether or not the contributions of chemistry factors (i.e., composition vector) substantially differ by crystal families. Supporting Information Table S2 shows the results by the ratio of the composition vector and diffraction pattern vector (CV:DP). We find that the optimal CV:DP ratio is largely unaffected by the types of crystal families and found around 90:10, although for only a few cases either 70:30 or 100:0 slightly outperforms. With these decomposition studies, we confirm once again that the material properties are much more determined by the chemical compositions rather than crystal structural information.

https://doi.org/10.1021/acs.jpclett.1c02305
J. Phys. Chem. Lett. 2021, 12, 8376−8383

Unfortunately, the performance of our DL model is not as good as that of the previous CGCNN model (0.39 eV for the band gaps and 0.04 eV/atom for the formation energies).[4] However, unlike the CGCNN model, our DL model does not require the exact coordinates of the constituent atoms that are not directly accessible experimentally. Rather, the DL model requires two inputs, composition information and diffraction pattern images, that are all directly accessible experimentally. The data set used in this study is from the Materials Project library,[1] indicating that the band gaps and formation energies are DFT values. Compared with experiments, DFT calculations themselves have MAEs of 0.6 eV for the band gaps and 0.08−0.14 eV/atom for the formation energies.[4,28] In this regard, although our DL model provides only slightly higher MAEs for the prediction of material properties than those of DFT calculations, our model is enough useful.

As already mentioned, it is found that the material properties (band gaps and formation energies) are more influenced by the chemical compositions of materials than by the crystal structures. To further support this, we additionally performed statistical analyses for the band gaps and formation energies with several chemical formulas (TiO_2, ZrO_2, Al_2O_3, RbS, ScF_3, and NaN_3), in which the materials were considered as representative oxides, sulfides, fluorides, and nitrides. In Table 1, the formation energies and band gaps of each material are averaged for 7 crystal structures (*triclinic, monoclinic, orthorhombic, tetragonal, trigonal, hexagonal, and cubic* structures), and the standard deviation (SD) values of the averaged properties are included. For example, there are 34 TiO_2 crystals in the Material Project library,[1] and the 34 TiO_2 crystals are classified into 1 triclinic, 6 monoclinic, 15 orthorhombic, 3 tetragonal, 4 trigonal, 2 hexagonal, and 3 cubic structures. Here, the formation energies for each crystal structure are averaged. The formation energies for TiO_2 crystals are −3.48 eV/atom for triclinic, −3.38 eV/atom for monoclinic, −3.40 eV/atom for orthorhombic, −3.48 eV/atom for tetragonal, −3.39 eV/atom for trigonal, −3.31 eV/atom for hexagonal, and −3.28 eV/atom for cubic. As a result, the SD for the averaged formation energies of TiO_2 is 0.08. Similarly, the SD values for ZrO_2, Al_2O_3, RbS, ScF_3, and NaN_3 are calculated as 0.01, 0.05, 0.55, 0.33, and 0.48, respectively (see the last column in Table 1a). These values represent the degree of variation in material property variations when the crystal structures are modified and the chemical formulas are fixed. In contrast, to understand the opposite case where the crystal structures are fixed and the chemical formulas are modified, we also calculated the SDs of the averaged formation energies of TiO_2, ZrO_2, Al_2O_3, RbS, ScF_3, and NaN_3 crystals for a given crystal structure. The results reveal that SDs are generally much larger: 0.12 for triclinic, 1.87 for monoclinic, 1.12 for orthorhombic, 0.22 for tetragonal, 1.71 for trigonal, 1.47 for hexagonal, and 0.46 for cubic (see the last row in Table 1a). Comparing the SDs in the last row and column in Table 1a clearly shows that the formation energy variations are induced much more by the chemical compositions of the materials than by their crystal structures. In Table 1b, we also find similar behavior for the band gap property; thus, detailed explanations are omitted.

To further support the argument that the variations in material properties are mainly due to the chemical composition rather than crystal structure, we provide a concrete example comparing TiO_2 and ZrO_2. TiO_2 and ZrO_2 are chosen because they are oxides and have the same stoichiometry (metal:oxygen = 1:2). In the Materials Project library,[1] they have the four same crystal structures (space group numbers of 14, 61, 136, and 141), and their formation energies and band gaps are summarized in Tables S3 and S4, Supporting Information. Regarding TiO_2, the space group numbers of 14, 61, 136, and 141 have formation energies of −3.46, −3.50, −3.47, and −3.52 eV/atom, respectively, leading to an SD of 0.03 for the formation energies. Regarding ZrO_2, the space group numbers of 14, 61, 136, and 141 have formation energies of −3.84, −3.83, −3.79, and −3.82 eV/atom, respectively, leading to an SD of 0.02. However, for the same space group, the formation energies of TiO_2 and ZrO_2 show much larger SD values (0.27 for space group no. 14, 0.23 for no. 61, 0.23 for no. 136, and 0.21 for no. 141). Likewise, for the band gaps, similar behaviors are observed. These results clearly reveal that the formation energies and band gaps are more influenced by the chemical composition than by the crystal structure.

Lastly, we would like to discuss the effect of the possible uncertainty in the input feature preparations. The uncertainty may exist in the input preparation stage, and their effects on the prediction reliability are worth being addressed. For our DL model, two input features (composition vector and diffraction pattern vector) exist, and the composition vector turned out to be much more influential; thus, we focus on the composition vector solely in the analysis. The measurements of the composition cannot be errorless. For an arbitrary example of CeO_2, multiple measurement methods could lead to different results (e.g., method1 = Ce:31.67, O:70.00; method2 = Ce:35.00, O:63.33; method3 = Ce:36.67, O:60.00).

Supporting Information Figure S6 shows the results of the relationship between the composition measurement error (%) in the metal elements and the increase in the resultant errors ($\varepsilon = \Delta MAE/MAE_0$, %) of the output properties (both formation energy and band gap energy). Here, MAE_0 denotes the MAE at zero compositional error, and ΔMAE denotes the increase in MAE upon the change in the composition values. It is not surprising to see that ΔMAE monotonically increases with the increasing errors in composition measurements (regardless of the direction). As a guideline, we identified the threshold value in the compositional error that is expected to cause less than 3% in ε. For the formation energy, when the composition measurement error is within 5% (−5 ∼ +5%), ΔMAE is very small, less than only 2%. For the band gap energy, on the other hand, when the composition measurement error is within 10% (−10 ∼ +10%), ε is less than only 3%. The DL model is relatively robust with the larger errors in the composition measurements for the bandgap energy than the formation energy. This difference in the sensitivity is likely attributed to the fact that the absolute values of the formation energies are smaller (mostly between −2 to +2 eV/atom) than those of the band gap energies (mostly between 0 to +10 eV); hence, the ε is computed larger.

In conclusion, we developed a CNN-based DL model for the direct prediction of material properties (formation energy and band gap) from experimentally accessible input features (chemical compositions and diffraction patterns) to improve the practical use of DL models. Although the two inputs are heterogeneous in terms of the types (texts and images) and vector sizes, our DL model readily treats the heterogeneous features simultaneously, in which the novel chemical composition vector is proposed by developing a method named element embedding with the normalized composition matrix. Here, we intentionally avoid using inputs that are not

https://doi.org/10.1021/acs.jpclett.1c02305
J. Phys. Chem. Lett. 2021, 12, 8376−8383

directly accessible (e.g., exact coordinates of constituent atoms) in our experimental platform, which is expected to substantially improve the practical use of DL models in material property predictions. As a result of weighting tests between two feature types (chemical compositions and crystal structures) in the DL model, we find that the material properties are more dominantly determined by the chemical composition than by the crystal structure, which provides an important guideline for future studies of the inverse design of materials. That being said, it would be a more efficient strategy to first determine the chemical compositions of materials and subsequently predict their crystal structures for a given composition in performing the inverse design of materials.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jpclett.1c02305.

Detailed explanation on EENCM, diffraction pattern generation, and network architecture of the proposed DL model; the data distributions; and additional results (PDF)

■ AUTHOR INFORMATION

Corresponding Authors

Sang Soo Han − *Computational Science Research Center, Korea Institute of Science and Technology, Seoul 02792, Republic of Korea*; orcid.org/0000-0002-7925-8105; Email: sangsoo@kist.re.kr

Donghun Kim − *Computational Science Research Center, Korea Institute of Science and Technology, Seoul 02792, Republic of Korea*; orcid.org/0000-0003-0326-5381; Email: donghun@kist.re.kr

Authors

Jeongrae Kim − *Computational Science Research Center, Korea Institute of Science and Technology, Seoul 02792, Republic of Korea*

Leslie Ching Ow Tiong − *Computational Science Research Center, Korea Institute of Science and Technology, Seoul 02792, Republic of Korea*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jpclett.1c02305

Author Contributions

†(J.K. and L.C.O.T.) These authors contributed equally.

Notes

The authors declare no competing financial interest.

The codes and data are shared in this repository (https://github.com/ceright1/Prediction-material-property.git) so that the proposed model can be applied to other studies.

■ ACKNOWLEDGMENTS

This work was supported by the Samsung Research Funding & Incubation Center of Samsung Electronics under Project Number SRFC-MA1801-03.

■ REFERENCES

(1) 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*, 011002.

(2) Draxl, C.; Scheffler, M. The NOMAD Laboratory: From Data Sharing to Artificial Intelligence. *J. Phys. Mater.* **2019**, *2*, 036001.

(3) Saal, J. E.; Kirklin, S.; Aykol, M.; Meredig, B.; Wolverton, C. Materials Design and Discovery with High-Throughput Density Functional Theory: The Open Quantum Materials Database (OQMD). *JOM* **2013**, *65*, 1501−1509.

(4) Xie, T.; Grossman, J. Crystal Graph Convolutional Neural Networks for Accurate and Interpretable Prediction of Material Properties. *Phys. Rev. Lett.* **2018**, *120*, 145301.

(5) Kim, M.; Yeo, B. C.; Park, Y.; Lee, H. M.; Han, S. S.; Kim, D. Artificial Intelligence to Accelerate the Discovery of N_2 Electroreduction Catalysts. *Chem. Mater.* **2020**, *32*, 709−720.

(6) Gu, G. H.; Noh, J.; Kim, S.; Back, S.; Ulissi, Z.; Jung, Y. Practical Deep-Learning Representation for Fast Heterogeneous Catalyst Screening. *J. Phys. Chem. Lett.* **2020**, *11*, 3185−3191.

(7) Park, W. B.; Chung, J.; Jung, J.; Sohn, K.; Singh, S. P.; Pyo, M.; Shin, N.; Sohn, K.-S. Classification of Crystal Structure Using a Convolutional Neural Network. *IUCrJ* **2017**, *4*, 486−494.

(8) Vecsei, P. M.; Choo, K.; Chang, J.; Neupert, T. Neural Network Based Classification of Crystal Symmetries from X-Ray Diffraction Patterns. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2019**, *99*, 245120.

(9) Wang, H.; Xie, Y.; Li, D.; Deng, H.; Zhao, Y.; Xin, M.; Lin, J. Rapid Identification of X-Ray Diffraction Patterns Based on Very Limited Data by Interpretable Convolutional Neural Networks. *J. Chem. Inf. Model.* **2020**, *60*, 2004−2011.

(10) Oviedo, F.; Ren, Z.; Sun, S.; Settens, C.; Liu, Z.; Hartono, N. T. P.; Ramasamy, S.; DeCost, B. L.; Tian, S. I. P.; Romano, G.; Gilad Kusne, A.; Buonassisi, T. Fast and Interpretable Classification of Small X-ray Diffraction Datasets Using Data Augmentation and Deep Neural Networks. *npj Comput. Mater.* **2019**, *5*, 60.

(11) Ziletti, A.; Kumar, D.; Scheffler, M.; Ghiringhelli, L. Insightful Classification of Crystal Structures Using Deep Learning. *Nat. Commun.* **2018**, *9*, 2775.

(12) Aguiar, J. A.; Gong, M. L.; Unocic, R. R.; Tasdizen, T.; Miller, B. D. Decoding Crystallography from High-Resolution Electron Imaging and Diffraction Datasets with Deep Learning. *Sci. Adv.* **2019**, *5*, eaaw1949.

(13) Kaufmann, K.; Zhu, C.; Rosengarten, A. S.; Maryanovsky, D.; Harrington, T. J.; Marin, E.; Vecchio, K. S. Crystal Symmetry Determination in Electron Diffraction Using Machine Learning. *Science* **2020**, *367*, 564−568.

(14) Zhou, Q.; Tang, P.; Liu, S.; Pan, J.; Yan, Q.; Zhang, S.-C. Learning Atoms for Materials Discovery. *Proc. Natl. Acad. Sci. U. S. A.* **2018**, *115*, E6411−E6417.

(15) Tshitoyan, V.; Dagdelen, J.; Weston, L.; Dunn, A.; Rong, Z.; Kononova, O.; Persson, K. A.; Ceder, G.; Jain, A. Unsupervised Word Embeddings Capture Latent Knowledge from Materials Science Literature. *Nature* **2019**, *571*, 95−98.

(16) Aguiar, J. A.; Gong, M. L.; Tasdizen, T. Crystallographic Prediction from Diffraction and Chemistry Data for Higher Throughput Classification Using Mahcine Learning. *Comput. Mater. Sci.* **2020**, *173*, 109409.

(17) Lecun, Y.; Bottou, L.; Bengio, Y.; Haffner, P. Gradient-Based Learning Applied to Document Recognition. *Proc. IEEE* **1998**, *86*, 2278−2324.

(18) Krizhevsky, A.; Sutskever, I.; Hinton, G. E. ImageNet Classification with Deep Convolutional Neural Networks. *Commun. ACM* **2017**, *60*, 84−90.

(19) Kim, Y. Convolutional Neural Networks for Sentence Classification. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing* **2014**, 1746−1751.

(20) Chen, L. C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A. L. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully *Connected* CRFs. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *40*, 834−848.

(21) Herr, J. E.; Koh, K.; Yao, K.; Parkhill, J. Compressing Physics with an Autoencoder: Creating an Atomic Species Representation to

Improve Machine Learning Models in the Chemical Sciences. *J. Chem. Phys.* **2019**, *151*, 084103.
(22) Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G.; Dean, J. Distributed Representations of Words and Phrases and Their Compositionality. *Adv. Neural Inf. Process. Syst.* **2013**, 3111−3119.
(23) Jing, L.-P.; Huang, H.-K.; Shi, H.-B. Improved Feature Selection Approach TFIDF in Text Mining. In *Proceedings. International Conference on Machine Learning and Cybernetics*; November 4−5, 2002, Beijing, China; IEEE: 2002; Vol. *2*, pp 944−946, DOI: 10.1109/ICMLC.2002.1174522..
(24) Hinton, G. E.; Salakhutdinov, R. R. Reducing the Dimensionality of Data with Neural Networks. *Science* **2006**, *313*, 504−507.
(25) Bengio, Y.; Yao, L.; Alain, G.; Vincent, P. Generalized Denoising Auto-encoders as Generative Models. *arXiv*, 2013, 1305.6663, https://arxiv.org/abs/1305.6663.
(26) Faber, F. A.; Lindmaa, A.; von Lilienfeld, O. A.; Armiento, R. Machine Learning Energies of 2 Million Elpasolite (ABC_2D_6) Crystals. *Phys. Rev. Lett.* **2016**, *117*, 135502.
(27) Hantke, M. F.; Ekeberg, T.; Maia, F. R. N. C. Condor: A Simulation Tool for Flash X-Ray Imaging. *J. Appl. Crystallogr.* **2016**, *49*, 1356−1362.
(28) 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*, 15010.

https://doi.org/10.1021/acs.jpclett.1c02305
J. Phys. Chem. Lett. 2021, 12, 8376−8383