Powder X-Ray Diffraction Pattern Is All You Need for Machine-Learning-Based Symmetry Identification and Property Prediction

Byung Do Lee, Jin-Woong Lee, Woon Bae Park, Joonseo Park, Min-Young Cho, Satendra Pal Singh, Myoungho Pyo, and Kee-Sun Sohn*

1. Introduction

Powder X-Ray diffraction (XRD) patterns are the most fundamental and frequently encountered type of structural data in inorganic materials science.\textsuperscript{11–9} We suggest that the powder XRD pattern should be used as a material descriptor for the symmetry identification and property prediction based on machine learning (ML) (or deep learning [DL]). ML (or DL), along with high-throughput experimentation and computation, has recently gained significant attention from material scientists and engineers.\textsuperscript{10–17}

For promising ML modeling, it is customary to define appropriate material descriptors for the systematic representation of materials. As far as conventional powder XRD patterns (experimental or theoretical) are available, it is unnecessary to extract knowledge-based material descriptors for utilization in ML (or DL). A full-profile powder XRD pattern could replace typical descriptors that require costly knowledge-based extraction procedures because the XRD pattern represents the material identity. Recently, there have been several ML (or DL) studies that are based on various XRD techniques (not necessarily powder XRD patterns),\textsuperscript{18–38} such as phase identification,\textsuperscript{18–22} symmetry identification,\textsuperscript{23–29} structure–property modeling,\textsuperscript{30–36} and novel materials discovery.\textsuperscript{26,37,38} Although the ML-driven symmetry recognition and phase identification from a powder XRD pattern have been widely developed,\textsuperscript{18–29} the property prediction directly from the pristine powder XRD pattern is still challenging.

It is also noted that all previous successes are achieved in a narrow range of specific materials. Most of the existing ML models for XRD analyses are specific-task-oriented in a narrow range of materials, lacking general application despite being more predictable. For instance, we have reported two DL-driven XRD analyses with approximately 100% accuracy for phase identification. One of these analyses was confined to the Sr–Al–Li–O quaternary system\textsuperscript{20} and the other to the Li–Zr–P–O system.\textsuperscript{21} It should be noted that there is no guarantee that these successful results can be transferred to other composition systems. Even in
property prediction based on crystal graph convolutional neural network (CGCNN), where no XRD data were involved, only a part of MP entries was used after removing many crystals, and outstanding regression results were obtained.

Developing a versatile ML (or DL) model covering all general inorganic materials for both symmetry identification and property prediction is challenging. The present investigation aims at developing an ML (or DL)-driven approach that is generally applicable for all experimental and theoretical (virtual) inorganic compounds. Therefore, our training dataset includes almost all entries registered in the inorganic crystal structure database (ICSD) and materials project (MP). However, some entries from other well-known databases such as NOMAD, AFLOW, and OQMD were left out. The DL models were trained for symmetry identification (crystal system, extinction group, and space group) and property prediction (bandgap, formation energy, and energy above convex hull) from the powder XRD patterns of nearly all the ICSD and MP entries (189,476 + 139,027 entries in total). This type of large-scale and diversely distributed data-based approach deserves attention.

We aim to use a full-profile XRD pattern as a descriptor for symmetry identification, property prediction, and low-dimensional embedding. The majority of ML (or DL) approaches in the materials research society have used conventional discrete materials descriptors, and some XRD-based ML (or DL) approaches have adopted discrete descriptors extracted directly from the XRD pattern based on handcrafted features engineering processes. However, neither the knowledge-based material descriptor extraction nor the XRD descriptor extraction resulted from the dimension reduction of full-profile XRD pattern are dealt with in our approach. One of the key points of our approach is to minimize human intervention, that is, no knowledge-based data selection or handcrafted feature engineering. Therefore, we used the full profile XRD data as inputs to all ML (or DL) models instead of the traditional discrete material descriptors.

With the full-profile XRD pattern as an input, a fully convolutional neural network (FCN) and a transformer encoder (T-encoder) were used for symmetry classification, and an FCN concatenated with multilayer perceptron (MLP) was used for property regression. The CGCNN was also used for the same analyses to compare the results. Finally, a variational autoencoder (VAE) consisting of an FCN and a fully transposed convolutional network (FTCN) was introduced along with various loss settings, accounting for a more reasonable reconstruction and regularization scheme. All ICSD and MP entries were embedded onto the VAE-driven 2D (or 3D) latent space in which clear clustering was observed according to the symmetry and property.

2. Results and Discussion

2.1. Symmetry Identification

The overall task flow is schematically described in Figure 1, illustrating crystal structure symmetry recognition and property prediction solely from the powder XRD pattern of inorganic compounds, and low-dimensional embedding of all the ICSD and MP entries. The symmetry recognition should be performed through classification models. We set up an FCN to perform the three classification tasks for the recognition of the crystal system (7-class classification), extinction group (101-class classification), and space group (230-class classification). The FCN result was similar to our previous convolutional neural network (CNN)-driven XRD symmetry classification results cited by Ryan et al. as the first example of DL for crystallography. The previous CNN architecture involved fully connected MLP layers, resulting in a considerable number of parameters. However, in the present work, we removed the MLP layers from the CNN architecture and introduced the FCN architecture. The kernel size was also significantly reduced in the FCN model. Figure 2a shows the finally adopted FCN architecture that adopts 13 convolution layers with maxpooling and dropout. By optimizing the hyperparameters, listed in Table S1, Supporting Information, the highest test accuracy for the holdout dataset was 92.12% for the crystal system identification when trained with the ICSD dataset. The training data used for the FCN model are larger and more general than those used for the previous CNN.

The T-encoder was also used for the same symmetry classification tasks that were performed by the FCN. The transformer
Figure 2. DL model architectures for symmetry recognition, property regression, and data embedding: architectures of a) FCN, b) T-encoder, c) FCN–MLP, and d) VAE.
was originally developed for natural language processing (NLP); however, more recently, it has been used for image recognition using a vision transformer (ViT). To the best of our knowledge, the conventional transformer model has never been used for full-profile XRD analysis. However, a simple combination of a CNN and a single-head attention has been very recently used for deep feature visualization from XRD patterns in the limited 2θ range of 20°–40°. This early attempt is far from complete and differs from a conventional transformer consisting of several mult HEAD self-attention blocks.

The selected T-encoder architecture for symmetry classification using the hyperparameter optimization process has two multihead self-attention blocks with six heads and a feed-forward hidden layer of size 1024. Attention can recognize important information by considering correlations in the data. Attention considers the patched XRD data as dictionary objects and obtains correlations using their inner products. The entire XRD profile was split into 64 patches, and the size of each patch was 128. In contrast to the typical ViT, we did not produce lower dimensional linear embeddings from the flattened patches, but the flattened
patch itself was considered as embedding. Positional embeddings were added and thereby the sequence was fed as an input to the transformer. An MLP that is terminated with softmax activation and cross entropy loss function was attached to the output layer of the transformer; therefore, 7-, 101-, and 230-class classifications were available. The final adopted T-encoder architecture is shown in Figure 2b. The hyperparameter optimization was implemented by screening all the hyperparameters in Table S2, Supporting Information, and the maximum test accuracy for the crystal system identification was 79.67%.

Figure 3a–c shows the test accuracy for the crystal system, extinction group, and space group identification. The FCN and T-encoder training were performed independently for the ICSD and MP datasets. The aforementioned (the highest) test accuracies for the crystal system identification (i.e., 92.12% for FCN and 79.67% for T-encoder) were obtained from the ICSD dataset. The test accuracy values for the MP dataset were slightly lower than those for the ICSD dataset. The one-top test accuracy for extinction and space group identification accuracies are deteriorated compared to those for the crystal system identification, regardless of whether the ICSD or MP dataset was dealt with. However, the 3-top and 5-top accuracies that are considered as reasonable accuracy metrics in the DL society are acceptable. It is a commonsense that 1D powder XRD pattern is not sufficient for the complete identification of space group in principle. Around the accuracy of 80% should be the upper bound when using well-established XRD analysis tools such as ITO,[49] TREOR,[50] DICVOL,[51] McMaille,[52] EXPo,[53] and X-CELL.[54] The complete identification of space group must be achieved by the ensuing procedures involving direct (or direct space) method and Rietveld refinement, which requires much more time and cost. It is worthwhile to put an emphasis on the fact that the DL-based XRD technique can never outperform the human expert in terms of the identification accuracy, but only a merit of the DL approach is the rapidity. The ≈80% accuracy for our DL-based space group identification is similar to Suzuki et al.[29] which would be an appropriate upper bound. The reason why we had a higher accuracy around 85% in our previous report[23] originates from the different data set, which was smaller and also more curated in comparison to the present non-handcrafted one.

The ICSD dataset exhibited higher accuracy than the MP dataset for all symmetry classification tasks. Presumably, the higher number of duplicated entries in ICSD might have produced a certain degree of information leakage between the training and test datasets. This could be one of reasons for the superiority of test accuracy for ICSD. The duplicated entries differ from one another in terms of lattice size, texture, and strain, among others. Therefore, the presence of such duplicates would not significantly help in improving the test accuracy. Although the reason for the superiority of the ICSD dataset to the MP dataset is not clear, we believe that the virtual (theoretical) structure entries influence test accuracy. ICSD involves a significantly smaller number of theoretical (virtual) entries (e.g., 922 virtual entries for ICSD collection code 1-100 000) compared to MP (more than 86 974 virtual entries). Most experimentally realized ICSD entries are more reliable than the theoretical entries. All the theoretical structures in the MP dataset are fully occupied; however, they would possibly form partially occupied structures, provided they are realized in the real world. The partially occupied (and nonordered) structure would lead to higher symmetry compared to its fully occupied counterpart. This indicates that the MP data distribution is skewed toward the lower symmetry side relative to the ICSD dataset, as shown in Figure 3d. This might be one of the reasons for the deteriorated test accuracy for the MP dataset.

Evidently, the FCN outperforms the T-encoder for both the ICSD and MP datasets. This finding contradicts the recent general trend in the field of vision recognition, that is, the ViT and BEIT have recently replaced state-of-the-art (SOTA) records achieved by CNN-based models in the field of image (or vision) recognition.[46,47] It should be noted that the number of parameters (weights and biases) was confined between 1 300 000 and 1 700 000 when optimizing the hyperparameters in our T-encoder models. The number of parameters (the model size) in our T-encoder models was arbitrarily smaller than that, for the FCN models, because the number of transformer blocks was not allowed to exceed two to prevent overfitting. More than two transformer blocks significantly enhanced the number of parameters, leading to extreme overfitting and drastic impairing of the test accuracy. In other words, the dearth of training data inevitably led to parameter number restrictions.

The small model size of the T-encoder along with the data paucity should result in marginal performance. It is certain that a greater amount of data suitable for a greater number of transformer blocks would definitely improve the accuracy of the T-encoder. In fact, conventional transformers used for both NLP and visual recognition are pretrained with a large training dataset containing even corrupted (masked) text and images, which is called self-supervised learning, and then fine-tuned with a smaller dataset.[46,47,55] It should be noted that the size of our ICSD and MP datasets is significantly below the conventional transformer standard. Therefore, the conventional transformer training process consisting of pretraining and fine-tuning was infeasible for our XRD dataset. The SOTA architecture contains 12 transformer blocks with 12 attention heads and the intermediate size of feed-forward networks (3072).[46] Thus, the T-encoder architecture is significantly smaller than the SOTA architecture. The test accuracy (79.67%) obtained from the from-scratch training for the crystal system identification for the ICSD dataset is reasonable.

The CGCNN result, as a baseline reference, indicates that the CGCNN test accuracy for the crystal system identification for the MP dataset was always lower than those of the FCN and T-encoder. Figure 3a shows that the FCN and T-encoder test accuracies reach 92.12% and 79.67% for the ICSD dataset and 82.17% and 69.01% for the MP dataset, respectively. However, the CGCNN test accuracy for the MP dataset (61.56%) is significantly lower, as denoted by an asterisk in Figure 3a. We used the CGCNN architecture (hyperparameters) optimized by Xie and Grossman.[15] The inclusion of many elemental traits as node features and bond traits as edge features in the CGCNN does not affect crystal symmetry in principle. The graph formulation incorporates a local structure (symmetry) to a certain extent, but does not seem to consider the long-range periodicity. This is the reason for the severely deteriorated CGCNN test accuracy for symmetry classification compared to the test accuracies of our XRD-pattern-based FCN and T-encoder approaches.
Figure 3. Test accuracies for a) crystal system, b) extinction group, and c) space group identification for FCN and T-encoder for both ICSD and MP datasets. The test accuracy for crystal system identification for CGCNN for MP dataset is provided as a comparative reference. d) Data distribution for the ICSD and MP datasets with respect to the crystal system. The plots for predicted versus real data for the MP holdout test dataset, resulting from e) formation energy, f) energy above convex hull, and g) bandgap regressions. The red data points represent the FCN–MLP regression result. The blue data points represent the CGCNN regression result.
2.2. Property Regression

Regression models were developed for the prediction of the DFT-calculated (MP-provided) bandgap ($E_g$), formation energy ($E_f$), and energy above the convex hull ($E_h$) from the XRD pattern. The FCN alone predicted the $E_g$, $E_f$, and $E_h$, although the result was not as promising as that of successful symmetry identification. These FCN regression results are not presented here for brevity. Therefore, to attain a more promising regression DL model, we set up an improved architecture with two channels, that is, the FCN and MLP channels. Full-profile XRD patterns (treated as 8192-dimensional vectors) were input to the FCN, and the composition vectors to the MLP were processed in parallel. The composition vector dimension was 100, which is equal to the total number of elements appearing in the ICSD and MP datasets. The normalized fraction of each constituent element in a certain compound was assigned to a corresponding slot in the composition vector, and the remaining slots were assigned to zero. The XRD pattern vector and the composition vector were independently processed in two separate channels at the initial stages; thereafter, these vectors merged (concatenated) at a certain layer and continued to flow to another MLP. When concatenated, the FCN side output vector size was 64, and the MLP side output vector size was 16. The schematic architecture of this FCN–MLP model is shown in Figure 2c. Such parallel multichannel DL model has been widely used in other domains.\cite{56,57}

The architectures (hyperparameters) of the $E_g$, $E_f$, and $E_h$ regression models were optimized by screening all the hyperparameters listed in Table S3, Supporting Information. The FCN–MLP model training was implemented in terms of mean absolute error (MAE). The MAE along with the coefficient of determination ($R^2$) for both the FCN–MLP and CGCNN is given in Table S4, Supporting Information. All the MAE and $R^2$ values in Table S4, Supporting Information, are based on the holdout test dataset. We used MAE rather than mean square error (MSE) to make a reasonable comparison with the previous result.\cite{15} We also validated the FCN–MLP model for $E_h$ prediction using Bhutani et al.’s data\cite{58} which is given in Table S5, Supporting Information.

The incorporation of the composition vector significantly promoted the fitting quality compared to the conventional FCN alone. The MAE of the holdout dataset test for the $E_g$, $E_f$, and $E_h$ predictions from the XRD pattern and the composition were 0.34, 0.09, and 0.07, respectively. The relationships between the predicted and real data for the holdout test dataset are shown in Figure 3c–g. It should be noted that only a simple composition addition to the existing XRD pattern significantly improves the regression performance. The fitting quality (predictability) for the FCN–MLP model is as good as the CGCNN result, which exhibits a test MAE of 0.41, 0.07, and 0.06 for $E_g$, $E_f$, and $E_h$, respectively. The CGCNN with no hyperparameter change was adopted because it was already optimized by Xie and Grossman.\cite{135} Our XRD-based FCN–MLP outperformed CGCNN for $E_h$ prediction. On the other hand, the performance of CGCNN was slightly higher than FCN–MLP for the $E_f$ and $E_h$ predictions. It is reconfirmed that, for the prediction of more symmetry-related properties such as bandgap, our XRD-based approach outperforms the CGCNN. Moreover, for the symmetry recognition (crystal structure, extinction group, and space group), the performance of our XRD-based approach is even more superior than the CGCNN, as shown in Figure 3a. The symmetry information, particularly the periodicity (translational symmetry) information, is not well considered in the CGCNN approach. Instead, the CGCNN more systematically involves the chemical and physical information on the constituent elements and their bonding, thereby outperforming the XRD-based FCN–MLE model for $E_f$ and $E_h$ prediction. We also tested the $E_g$, $E_f$, and $E_h$ prediction model without the XRD pattern using the above-described composition vector only. As a result, the symmetry-related $E_g$ prediction was deteriorated but the $E_f$ and $E_h$ prediction quality never changed. Antunes et al.\cite{59} very recently revealed that it was successful to predict various material properties from chemical formulas only by employing brilliant atomic representations (e.g., SkipAtom, Atom2Vec, Bag-of-Atom, etc.). Nonetheless, we found that the XRD plays a significant role for the symmetry-related $E_g$ prediction in the present investigation.

As mentioned previously, the MP data used for the regression were not preprocessed. Some selected downsized data, obtained by eliminating all the metals from the $E_g$ regression and all the zero-$E_h$ entries from the $E_h$ regression, would enhance the fitting quality. However, we believe that the DL approaches should be applied to noncurated, pristine data without incorporating any knowledge-based features and data engineering to secure generality to a certain extent. Alternatively, the CGCNN involves a brilliant knowledge-based feature selection, that is, some elemental features (e.g., atomic number, group number, period number, electronegativity, covalent radius, ionization energy, and electron affinity). In addition, some edge information, such as bond length, is fully incorporated in the CGCNN.\cite{32–34} In contrast to the CGCNN requiring meticulous feature selection based on human knowledge, our FCN–MLP approach only requires the compound composition information in addition to the pristine XRD pattern. The powder XRD pattern is a type of data that elucidate the electron density in the 1D projected reciprocal space, which is condensed from the 3D electron density in the real space. It is very difficult to perfectly describe the actual structure using only a powder XRD pattern because considerable information loss is unavoidable as the 3D electron density distribution is condensed into a 1D powder XRD pattern. Thus, it is impossible to identify the space group from only a powder XRD pattern, even when the conventional-rule-based structural determination process is adopted.\cite{1–9}

It should be noted that the powder XRD pattern that originates from the symmetry-associated electron density distribution eventually leads to the core concept of DFT calculations. Thus, the DFT-calculated properties should be implicated in XRD. Our understanding is that it would be unnecessary to separately incorporate such miscellaneous elemental and bonding information as far as the XRD pattern is adopted as an integrated input to any ML model. The symmetry information was also definitely involved in the XRD pattern. Thus, the XRD pattern could be a promising supplementary input feature for any type of ML-based material property prediction. Moreover, the practicality and accessibility of the powder XRD pattern is outstanding, as it has been used as a representative material analysis tool for...
a long time. A conventional powder XRD pattern, along with a simple composition vector, could replace several knowledge-based handcrafted descriptors.

2.3. Classification and Regression Result Analysis

We examined the test result in more detail. The crystal system identification result from the FCN classification and the $E_p$ prediction result from the FCN–MLP regression were rearranged in terms of the crystal system. Table S6, Supporting Information, shows the rearranged test result for the crystal system identification and $E_p$ prediction. It is evident that inorganic compounds with the lowest symmetry (Triclinic) exhibit the worst test result for both the crystal system identification and $E_p$ prediction. This result is coincident with Suzuki et al.\[^{29}\] In particular, the crystal system identification (classification) accuracy is obviously deteriorated for the Triclinic symmetry, which is due to the lack of symmetry. Figure S1, Supporting Information, shows the $E_p$ regression test result rearranged according to each of the crystal systems. Although the MAE values in Table S6, Supporting Information, differ from one another and also show a trend that lower symmetries exhibit slightly higher MAE values, the schematic representation is not discernable, as evidence in Table S6, Supporting Information.

2.4. Mapping XRD Patterns into a Low-Dimensional Embedding Space

Because the FCN model outperformed any other models for symmetry identification, the FCN-based VAE should be investigated to visualize low-dimensional embedding in the latent hidden space. Banko et al.\[^{27}\] recently reported a very interesting VAE-driven XRD analysis. Chen et al.\[^{18}\] also recently reported deep reasoning networks (DRNs) that are similar to VAE and realized an unsupervised phase matching in the Bi–Cu–V oxide system. In the DRN approach, they employed a reconstruction loss of the input data and a reasoning loss that captured the domain constraints to achieve excellent phase identification of mixtures.

Our VAE differs from those of Banko et al.\[^{27}\] and Chen et al.\[^{18}\] in several aspects. We covered all the ICSD/MP entries in 230 space groups. On the other hand, Banko et al.\[^{27}\] implemented a limited model system with a small number of samples ($\approx 15,000$) from three high-symmetry space groups showing a few peaks. Chen et al.\[^{18}\] used a few hundred samples in specific ternary oxide systems. However, there is a different opinion that this type of limited amount of data for the same level of performance is rather advantageous. We adopted an FCN-based, significantly deeper VAE architecture instead of the simple MLP architecture. More importantly, the FCN part in the encoder side preserved the previous supervised learning (symmetry classification) result by introducing the parameters that were optimized in other previous classification/regression tasks as initial parameters for the VAE training. The cluster boundary in our VAE-embedded low-dimensional latent space was clearly created without the assistance of other classifier algorithms, such as K-nearest neighbors. In addition, both symmetry and property clustering were achieved in the low-dimensional latent space in the present investigation. Finally, we adopted various loss function settings by considering the real shape of the XRD data distribution.

Figure 2d shows the VAE architecture consisting of encoder and decoder. The FCN followed by an MLP block constitutes the encoder, and a FTCN constitutes the decoder. The tandem encoder architecture consists of an FCN block used for the crystal system identification and the property regression, and an MLP block connected to the FCN. Thus, the MLP output layer constitutes a 4D (or 6D) embedding layer (means and variances for 2D (or 3D) $z$ variable). The FCN parameters that are already trained for the crystal system identification and the property regression were used as initial values for the VAE training. This could be an economical approach if the embedded data distribution in the low-dimensional latent space is examined by symmetry- and property-based clustering (i.e., crystal system, space group, bandgap, formation energy, etc.). It is unnecessary to waste the computational resources for learning what has already been learned from previous supervised learning implementations.

In fact, the latent space embedding layer does not designate latent variables ($z$). However, it designates parameters (means and variances for Gaussian distribution and rate constants for exponential distribution) accounting for the distribution of $z$. The input to the decoder ($z$) is sampled from the distribution, which can be formulated using the parameters. The evidence lower bound (ELBO), also known as variational lower bound or negative variational free energy, is a loss for VAE training (Equation (1)).\[^{60,61}\] The first term of Equation (1) represents the reconstruction loss—the so-called data likelihood—and the minus sign was introduced to minimize the loss during the VAE training. The second term represents the regularization term, formulated as Kullback–Leibler divergence ($D_{KL}$), which elucidates the similarity between distributions; $D_{KL}$ is zero when these distributions are the same.

Minus ELBO Loss

$$-E_{z \sim q_{θ}(z|x)}(log(p_{θ}(x|z))) + D_{KL}(q_{θ}(z|x)\|p(z)) \tag{1}$$

where $q_{θ}(z|x)$ represents the encoder and $p_{θ}(x|z)$ designates the decoder, and $Φ$ and $Θ$ represent the parameters (weights and biases) for the encoder and decoder, respectively. In general, $p(z)$ is simply a standardized normal distribution ($N(0, 1)$), in which $q_{θ}(z|x)$ should be approximated by minimizing the $D_{KL}$. The second loss term ($D_{KL}$) results in grouping of the embedded data clusters, and it simultaneously makes each cluster discriminative from one another in the embedding space. It is customary to assume that the input variable ($x$) is in a Gaussian distribution, and the latent variable ($z$) is also in a Gaussian distribution when the VAE is trained. However, it should be noted that the XRD data distribution is apparently not a Gaussian distribution, but more like an exponential (decaying) distribution. The 1D exponential-like distributions are clearly detected at some 2θ points for the ICSD dataset (Figure S2, Supporting Information). Both the reconstruction and regularization losses should be altered, and the sampling process should be changed in accordance with the exponential distribution. In this case, $p(z)$ was not set as $N(0, 1)$, but as an exponential distribution with the rate constant ($\lambda$) equal to 1. In other words, the mean ($1/\lambda$) and variance ($1/\lambda^2$) in the
exponential distribution were 1. The detailed derivation process for both the reconstruction and regularization losses for the multivariate exponential distribution is described in the Supporting Information.

Figure 4 shows the latent space embedding for the ICSD dataset, and that for the MP dataset is presented in Figure 5. The embedding results from all loss settings are shown in Figure 4 and 5. The reconstruction loss includes L2 loss, cross entropy, Gaussian likelihood, and exponential likelihood. Furthermore, the regularization loss includes the KL divergence between the exponential distributions on the top of the conventional KL divergence between the Gaussian distributions. Every loss setting resulted in a clear clustering according to the symmetry, and an obvious symmetry direction was clearly detected in the embedding space. The combination of the exponential likelihood (as a reconstruction loss term) and the exponential KL divergence (as a regularization loss term) has never been attempted before; however, it provides interesting embedding results in the present investigation. Although all the other loss settings involving the Gaussian KL divergence loss exhibited mostly Gaussian-type cluster shapes (rounded island shape), the exponential-distribution-based loss setting produced narrow-shape clusters, as shown in Figure 4e. Although the same clustering shape was not reproducible, every run ended up with this type of narrow distribution. This is natural considering the fact that we involved the standard exponential distribution in the KL divergence loss.

Although the actual XRD distribution is similar to the exponential distribution, other reconstruction loss options still work out because the distinction between actual high-dimensional data distributions is not sharp within the exponential distribution family, such as Gaussian, gamma, exponential, etc. Moreover, it is not necessary to stringently apply the exponential distribution for the x-distribution formulation, although the z distribution was approximated to an exponential distribution. This is because it would never be awkward to transform an exponential distribution to a Gaussian distribution through the encoder. In addition, if we adopt either L2 loss or...
cross-entropy loss, the type of data distribution for $x$ data would be insignificant. It is thus suggested that every loss setting can be used. Although a sort of theoretical absurdity would arise from some of the adopted loss settings, we acquired acceptable embeddings for every loss setting.

Figure 4 and 5 show the embedded data clustering in the 2D latent space represented with different colors in terms of the crystal system. Despite a certain degree of overlap, seven crystal systems were successfully separated from one another, and the embedding quality (distinction clarity between clusters) was better than that of MNIST digit data embedding.\textsuperscript{[63]} as can be seen in Figure S3, Supporting Information. In addition, the embedding quality is much better in comparison to the well-known t-SNE result.\textsuperscript{[29]} Well-separated embedding resulting from a small-scale, task-specified problem with a restricted dataset has been reported in previous studies.\textsuperscript{[27,34]} However, we obtained a better embedding despite using general data involving almost all inorganic compounds with generally interesting labels such as crystal system, space group, bandgap, and others. We believe that rather than an inordinately manipulated data curation to obtain a fancy-looking embedding result, generalized knowledge should be extracted from real-world general data in the pristine state. Although we could also present a well-separated embedding result (with no overlap) using only three arbitrarily selected space groups, as shown in Figure 4 and 5, this sort of fancy-looking result would make no sense at all in view of the data generality.

Our primary focus is general data embedding, as shown in Figure 4 and 5. Despite the slight overlap between the crystal systems, we could detect a clear symmetry trend (direction), as indicated by the arrows in Figure 4. In addition to the 2D latent space embedding, we also performed 3D latent space embedding in exactly the same manner, and a similar result was obtained in the 3D plot (Figure 6). Although a slight overlap between different symmetry clusters (i.e., different crystal systems) is inevitably observed in the 2D and 3D embedding spaces, a clear symmetry direction from the triclinic to cubic was detected for every loss setting. Detection of such promising embedding results is reasonable because a very high accuracy was obtained for the crystal system recognition using the same FCN architecture employed as an encoder in the VAE and preserving the fully learned parameters as initial parameters for the VAE training process. It should be noted that the crystal system recognition was finally processed in the 7D output layer; however, the embedding space was maximum 2D or 3D. If the embedding had been performed in the 7D hyperspace, significantly well-separated clustering would have been successfully detected in association with the high accuracy of the crystal system classification. However, it is unfortunate that there is no way to visualize the 7D hyperspace convincingly.

The property-wise embedding of the MP dataset was also implemented using the FCN–MLP encoder, which was used for property regression (Figure S4, Supporting Information). The fully learned parameters were also inherited from the preceding property regression and used as initial parameter values for the VAE training. The $E_g$, $E_c$, and $E_f$ values are categorized into binary classes, for example, $E_g = 0$ (metallic) or $E_g > 0$ (nonmetal), $E_f < 0$ (stable) or $E_f > 0$ (unstable), and $E_c = 0$ (indecomposable) or $E_c > 0$ (decomposable). A certain degree of distinction between the binary clusters can be observed (Figure S4, Supporting Information), although considerable overlap was unavoidable. The property-wise embedding quality was not as good as that of the symmetry-wise embedding; however, a clear distinction between the binary clusters for $E_g$, $E_c$, and $E_f$ was detected in the 2D latent space. This relatively poor embedding quality for the property-wise clustering (Figure S4, Supporting Information) should be due to the continuous nature of the properties but in part might be due to the unbalanced incorporation of the composition; that is, the composition vector was involved in the encoder side, but removed from the decoder side reconstruction procedures.

Despite the success in the VAE embedding, the various loss settings are yet to be fully stabilized, which is why the convergence differed slightly for every run. This implies that the embedding results never reached the global optimum, but loitered around many local optima. Another problem resides

![Figure 6](image.png)
in the reconstruction side in spite of the promising embedding. The trained VAE could generate plausible XRD patterns except for the Gaussian likelihood loss. Although the generated XRD pattern appears similar to conventional XRDs, it would not be possible to further use them for appropriate XRD generation that could lead to a novel structure discovery. Further works based on the U-net-based VAE were made to reinforce the reconstruction side.

The U-net involves an architecture that consists of a contracting path to capture context and a symmetric expanding path that enables precise localization. We maintained the same FCN architecture in the encoder side (contracting path) and the FCN in the decoder side (expanding path). Some decoder features are concatenated with the encoder features (Figure S5, Supporting Information). While other types of U-net-based VAE approaches have more complicated architectures, we simplified the architecture such that the concatenation was not applied for every layer but only a half of the decoder layers was concatenated with the corresponding encoder layers, as marked with arrows in Figure S5, Supporting Information. Only a loss setting involving Gaussian KL divergence (regularization loss) and cross-entropy loss function (reconstruction loss) was adopted for the U-net-based VAE approach. The initial parameters for the encoder were inherited from the previous symmetry classification, and the initial decoder parameters from the preliminary autoencoder training. The straight autoencoder without the sampling in the encoder output layer was employed to provide initial decoder parameters for the U-net-based VAE. The U-net-based VAE gave an acceptable embedding and reconstruction quality. While the embedding quality was maintained at a similar to that for the pristine VAE, the reconstruction quality was greatly improved such that no distinction could be detected between the original and reconstructed XRD patterns (Figure S6, Supporting Information).

3. Experimental Section

3.1. Dataset Preparation

We collected all the inorganic structures registered in ICSD and MP, whose lattice volume was below 10 000 Å³. The XRD input vector size was 8192, spanning the 2θ range from 5° to 86.91°. Because the lower bound of 2θ is 5°, this lattice volume restriction is reasonable. In fact, as the XRD input vector size is 10 000, spanning the 2θ range from 5° to 105° was also feasible with almost similar test accuracies. We secured 189 476 ICSD and 139 027 MP entries in cif format. We used pymatgen for the cif extraction along with the property extraction from the MP database. The synthetic powder XRD pattern was computed from the crystal structure solution of these entries. It was crucial to determine the appropriate parameters that were essential for XRD pattern simulation. The adjustable parameters were the Lorentz polarization factor, preferred orientation, background, and peak profile. The polarization correction was set for laboratory XRD in the Bragg–Bragno geometry fitted with a graphite monochromator. The preferred orientation was not considered. The background was varied randomly using sixth-order polynomial functions. The key issue was not achieving synthetic XRD patterns that are close to the real ones as possible, but using the XRD pattern as an informative material descriptor to input to ML models. In this context, the peak profile was set by fixed mixing parameters, as well as Caglioti parameters. We have already confirmed that the synthetic XRD pattern is indistinguishable from the experimentally measured in our previous reports, where all parameters were randomly sampled from meticulously chosen parameter ranges in the synthetic XRD production process.

The structure solution of some ICSD and MP entries exhibits various origin settings despite having the same structure. All of the different origin settings were considered. In fact, the setting that Fullprof adopted as a standard was considered and several entries in the other settings were all excluded in our previous report. However, all of them were included in the present investigation. Thus, the total number of ICSD entries selected for the present investigation increased significantly to 189 476 from 150 000 entries. The total number of selected ML entries was 139 027. Both the ICSD and MP datasets were split into training and holdout test datasets with 80:20 fractions. In fact, considering the domain routine, the 20% holdout test dataset percentage is a bit high, but we chose this percentage to make a reasonable comparison with the previous CGCNN result. In addition, we adopted a fourfold cross-validation strategy in which the 20% validation dataset was instantaneously split from the training dataset during the training while the 20% holdout test dataset was separately preserved. All the accuracy and MAE values appearing in the present investigation were based on the holdout test dataset.

3.2. DL Models

We unique coded the FCN and FCN–MLP models; however, the T-encoder code contained significant implementation from the Keras example library for text classification with transformers and image classification with vision transformers. VAE models with various loss settings were coded based on the Keras variational autoencoder. The CGCNN code was downloaded from Xie and Grossman’s GitHub page, and only the classification code was slightly altered; that is, the original binary classification was altered to 7-class classification. All the CGCNNs for symmetry classification and property regression were implemented using the hyperparameters that Xie and Grossman optimized. All the codes used in the present study are available at our GitHub page.

The model architecture and hyperparameters were optimized in the range, as shown in Table S1–S3, Supporting Information. Rather than tuning miscellaneous individual hyperparameter, we made significant efforts to maintain the total number of parameters (weights and biases) at a certain level (1 300 000–1 700 000). This level was defined such that below or above this level, the training was not satisfactory. Moreover, within this level, the fitting quality was good regardless of the meticulous hyperparameter choice. The architecture for finally adopted DL models is schematically described in Figure 2 and more details are given in Figure S7, Supporting Information.
4. Conclusions

Many efforts have been made to describe inorganic materials using simple descriptors.\cite{10-17} Descriptor extraction is mostly based on human knowledge of the physical, chemical, and structural traits of materials. The CGCNN\cite{32-36} is a brilliant model that depends on the knowledge-based selection of atomic and bonding features. In contrast, the main idea of our approach is based on human-intervention-free feature (descriptor) extraction. The full-profile XRD pattern, which is one of the most well-known traditional material analysis utensils, can act as a material descriptor. There is no reason to discard such a well-developed material description that has long been used in various materials science areas. Thus, it would be unreasonable to waste such useful existing resources (i.e., the XRD pattern), even for the ML approach. The XRD pattern innately rooted in the electron density could implicate the DFT-calculated material properties. We confirmed that the XRD pattern plays a pivotal role as a material descriptor for use in ML models such as FCN, FCN-MLP, T-encoder, and VAE.

We obtained a higher test accuracy for symmetry classification when the ICSD dataset was used rather than the MP dataset. We discussed this issue by mentioning duplicated entries and virtual (theoretical) structure entries with awkward full occupancy. In addition, the FCN-involved models marginally outperformed the T-encoder, contradicting the recent trend in the vision recognition area. The incomplete T-encoder performance is discussed in relation to the dearth of our training data. Although we focus more on the generality based on the use of almost all ICSD/MP entries, we are still suffering from the data paucity problem when dealing with transformer models. Even self-supervised learning was inaccessible to the dataset size that we used.

We introduced CGCNN as a baseline reference because it exhibits a recent SOTA in ML-driven material property prediction.\cite{12-36} It is noted that the powder-XRD-pattern-based FCN and T-encoder significantly outperformed the CGCNN in the symmetry identification. The FCN–MLP model incorporating the powder XRD pattern together with the composition vector also exhibited a promising property regression performance, which reached the CGCNN SOTA level. It was evident that the CGCNN does not describe the long-range structure (translational symmetry) of inorganic materials, whereas it exhibited good accuracy in the material property prediction.

All the ICSD/MP data were embedded in the 2D (and 3D) latent space, and a clear symmetry-based clustering was detected along with a noticeable symmetry-increasing direction, from triclinic to cubic structure. The embedding led to a well-separated clustering according to the symmetry and the embedding quality (distinction clarity between clusters) was better than that of MNIST digit data embedding that KERAS provided.\cite{62} The various loss settings were successful regardless of the actual data distribution type. The other property-based embeddings were also partially successful in assessing the bandgap, formation energy, and energy above the convex hull. As a result, the low-dimensional embedding of the XRD pattern represents the original XRD pattern; thus, the low-dimensional embedding can also be viewed as a simplified material descriptor.

The validity of the suggested XRD-based DL models was confirmed by the fact that they outperformed the well-established CGCNN and the embedding quality was superior to the well-known MNIST digit data embedding.

Supporting Information

Supporting Information is available from the Wiley Online Library or from the author.

Acknowledgements

B.D.L., J.-W. L and W.B.P. contributed equally to this work. This research was supported by the Creative Materials Discovery Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT, and Future Planning (grant no. 2021R1A2C1009144) and (grant no. 2021M3A7C2089778). Every source code as well as the fully trained weights and the used dataset will be available on our GitHub page [https://github.com/socoolblue/XRD_is_All_You_Need].

Note: This article was amended on July 19, 2022 to add a second affiliation of the author S. Pal Singh.

Conflict of Interest

The authors declare no conflict of interest.

Data Availability Statement

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Keywords

depth learning, X-ray diffraction (XRD) pattern, machine learning, property regression, symmetry classification, X-Ray diffraction (XRD) pattern
