THE JOURNAL OF PHYSICAL CHEMISTRY Letters

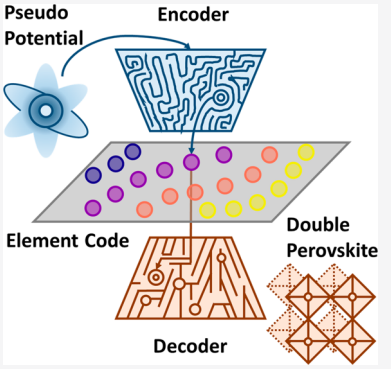
pubs.acs.org/JPCL Letter

Element Code from Pseudopotential as Efficient Descriptors for a Machine Learning Model to Explore Potential Lead-Free Halide Perovskites

Meng-Huan Jao, Shun-Hsiang Chan, Ming-Chung Wu,* and Chao-Sung Lai*

Cite This: *J. Phys. Chem. Lett.* 2020, 11, 8914–8921 Read Online

ABSTRACT: The rapid development of machine learning has proven its potential in material science. To acquire an accurate and promising result, the choice of descriptor plays an essential role in dictating the model performance. In this work, we introduce a set of novel descriptors, Element Code, which is generated from pseudopotential. Using a variational autoencoder to perform unsupervised learning, the produced Element Code is verified to contain representative information on elements. Attributed to the successful extraction of information from pseudopotential, Element Code can serve as the primary descriptor for the machine learning model. We construct a model using Element Code as the sole descriptor to predict the bandgap of a lead-free double halide perovskite, and an accuracy of 0.951 and mean absolute error of 0.266 eV are achieved. We believe our work can offer insights into selecting lead-free halide perovskites and establish a paradigm of exploring new materials.



Recently, the astonishing growth of machine learning (ML) has drawn enormous attention. Machine learning has successfully solved some issues in images, languages, and games,[1−4] for which it is challenging to define explicit functions to achieve the goal. The triumph of machine learning is attributed to a delicately designed neural network,[5−8] which can extract features at various levels and discern any obscure clues that may be indistinguishable to humans. In the field of material science, powerful machine learning techniques can be applied to the high-throughput prediction of material properties.[9,10] Nowadays, this kind of task is mainly carried out by density function theory (DFT) calculation, starting from pseudopotential to obtain the properties of targeted materials. However, it takes time and high-end hardware to execute DFT calculation.

On the other hand, the machine-learning-based method can provide a timely efficient strategy. Although the data set preparation and model building may be time-consuming, the prediction process is rather fast. In other words, once the model training process is complete, it only takes a few seconds to get the prediction done. Therefore, combining a machine-learning-based strategy with a DFT calculation can accelerate the search of new materials. Machine learning can be a powerful tool to carry out a preliminary screening before executing a DFT calculation and material synthesis. As a result, it provides another route to narrow down the candidate pool and facilitate the discovery of potential materials.

As a proof of concept, Wang et al. used an ML strategy to screen out several potential compounds as the active layer in perovskite solar cells from excessive possible combinations of elements.[11] The well-developed gradient boosting regression, supporting vector regression, and kernel ridge regression algorithms were employed to facilitate the precise prediction of material properties. A variety of similar works were reported to demonstrate how ML can facilitate the search of adequate materials.[12−16]

To further raise the accuracy and versatility of ML model, Xie et al. built a sophisticated neural network to learn chemical information and extract structural characteristics from crystalline compounds.[17] The architecture of their model (CGCNN) originated from the graph convolution network to portrait the relation of atoms and bondings in crystalline compounds. With the power of the graph convolution network, CGCNN was able to achieve an accurate prediction on numerous properties, spanning from thermodynamic, electronic, to mechanic. Later on, Ong et al. constructed a graph-convolution-network-based universal model (MEGNet) that can handle both the molecular and crystalline compounds.[18] The model was delicately designed so it can learn the chemical structure expression by updating parameters of atoms and bonds mutually. In addition, environmental conditions can also be taken into consideration. The MEGNet can perform a state-

Received: August 5, 2020
Accepted: September 29, 2020

https://dx.doi.org/10.1021/acs.jpclett.0c02393
J. Phys. Chem. Lett. 2020, 11, 8914–8921

of-the-art prediction on both molecular and crystalline materials.

Most of the works focused on the establishment of a workflow pipeline and neural network model. However, for an ML model to perform a reliable prediction, the input descriptors also play a critical role. For instance, Chang et al. found that when it comes to formation energy, choosing electronegativity and ionic radius would get better results while choosing the highest occupied atomic level and orbital radius as input descriptors can lead to better performance when focusing on bandgap.[19] The descriptors frequently selected for the prediction of material properties include atomic number, atomic radius, electron affinity, electronegativity, ionization energy, highest occupied atomic level, lowest unoccupied atomic level, and so on. One of the interesting things is that basically all those descriptors can be viewed as some kinds of morphisms of electron density distribution, which is the sole "descriptor" when processing a DFT calculation. Therefore, it would be interesting to study if the electron density distribution can serve as an efficient descriptor for an ML model to handle the material science tasks.

In this report, in order to transform electron density distribution into numerical values that can be subsequently fed into an ML model as a descriptor, we used an unsupervised ML neural network to automatically extract representative values from the electron density distribution of each element. With the construction of various models, values containing information with various patterns were generated. We generated four sets of representative values, named as Element Codes, by defining four different variational related loss functions in ML model. The generated Element Codes showed a strong correlation with the pattern of the periodic table. After the successful extraction of representative Element Codes, we evaluated their performance as descriptors to predict various material properties. Also, we demonstrated that highly accurate predictions were achieved using Element Code as the primary descriptor in an ML model. As a proof of the concept, a bandgap predicting ML model trained on inorganic halide double perovskites was established, obtaining an AUC score of 0.951 and an MAE of 0.266 eV. This paved a path to efficiently explore potential candidates in energy conversion application.

In the DFT calculation, the concept of pseudopotential is introduced to alleviate the expensive cost of complex computation. The pseudopotential is created with the criteria to describe the most critical part of an atom accurately, which is the distribution of valence electrons. On the other hand, the distribution of core electrons is moderately simplified. A good pseudopotential can strike a balance between complexity and accuracy. In our experiment, we chose PseudoDojo, an open-source community-maintained repository for norm-conserving pseudopotential,[20,21] to extract representative element descriptors. The PseudoDojo provides 85 elements, containing H in the first period to Rn in the sixth period, with their systematically validated pseudopotentials. The pseudopotential for the GGA-PBE exchange-correlation function with standard accuracy was selected. Data provided by PseudoDojo, including charge density and potential versus radius, were used as the representative of the elements. The numerical element descriptors were extracted from charge density and locale potential by constructing a neural-network-based autoencoder.

In ML theory, the training task can generally be categorized into two types.[22] One is the supervised learning task, where we have a manually labeled target during the ML model training process. The other is the unsupervised learning, where we have no idea about the label for the training data set. We hope that the ML model can learn the label from data set by itself. Extracting the representative values from pseudopotential belongs to the unsupervised learning. To achieve our objective, we constructed a neural-network-based autoencoder. The working principle of autoencoder is similar to the encoding and decoding procedure.[23,24] For the encoding part, the model would receive original data as an input and distill the information by passing the data through a bottleneck-like structure. Afterward, the original data became compressed, containing limited information, often regarded as code. For the decoding part, the model would use the code as an input and manage to reconstruct the original data. Ideally, if the reconstructed data is identical to the original data, it can be assessed that the code must capture some essence of the original data. In other words, the compressed data can roughly represent the original data.

The schematic diagram of the autoencoder is shown in Figure 1. The data flew from left to right. The original data for

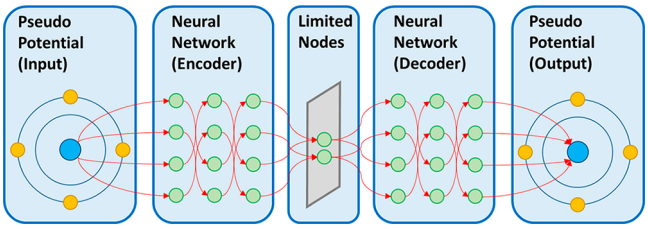


Figure 1. Schematic diagram of variational autoencoder (VAE) to extract Element Code from pseudopotential. The charge density and potential are taken as input data and fed into the encoder. The output is a two-dimensional Element Code, which serves as the input of the decoder. The objective of the decoder is to reconstruct charge density and potential as close as possible with respect to the input of the encoder.

each element included charge density and locale potential, comprising 1024 data points. The total data points could be viewed as one specific data point located in 1024-dimensional space. The encoder would consume the original data and transform it into another data point in low-dimensional space. In our model, for the sake of intuitive visualization, we set the target dimension as two. Then, the compressed two-dimensional data point was fed into the decoder, which was trained to reconstruct as close to the original data as possible. To improve the reproducibility and information extraction in various ways, we imposed constraints on the model so that the model needed to follow specific criteria to produce codes. This kind of autoencoder is called a "variational autoencoder" (VAE).[25]

Depending on the constraints, the code extracted from the original data can have completely different characteristics. Therefore, we generated four sets of codes from three distinguished VAEs, an evidence lower bound VAE (ELBO VAE),[25] a maximum mean discrepancy VAE (MMD VAE),[26,27] and a vector quantized VAE (VQ_VAE).[28] We wanted to compare the pattern of codes extracted from various VAEs and evaluated the performance when various codes served as descriptors for an ML model.

After a training process of 100 epochs, the reconstruction error from each VAE is below 0.007 per data point, suggesting

https://dx.doi.org/10.1021/acs.jpclett.0c02393
J. Phys. Chem. Lett. 2020, 11, 8914−8921

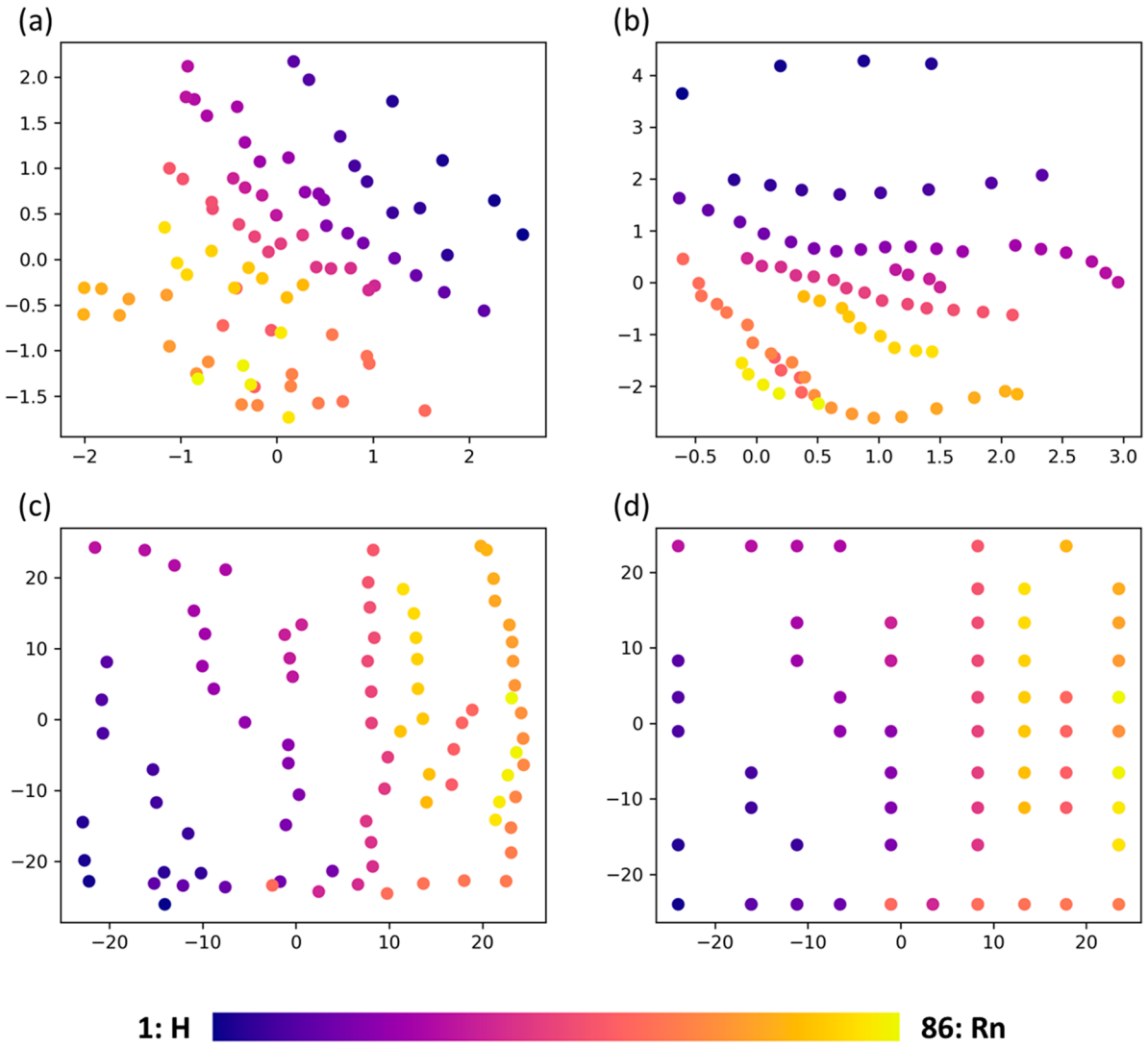


Figure 2. Element Codes derived from the same set of pseudopotential using various algorithms. (a) Evidence lower bound (ELBO), (b) maximum mean discrepancy (MMD), (c) vector quantized before embedding (VQ_o), (d) vector quantized after embedding (VQ_q). The color bar is used to represent elements with various atomic numbers.

the successful extraction of representative code. The obtained Element Codes concerning various VAEs are shown in Figure 2.

For ELBO VAE, the constraint was defined so that the distribution of Element Code was supposed to approximate a Gaussian distribution (Figure 2a), and the objective loss function was set as eq 1.[25]

$$L_{ELBO} = E_{pdata(x)}[-KL(q\phi(z|x)\|p(z))] + E_{pdata(x)}E_{q\phi(z|x)}[\log p\theta(x|z)] \quad (1)$$

For MMD VAE, the distribution of Element Code was assigned to approach specific distribution as well (Figure 2b), and eq 2 was used to assess the objective loss.[27]

$$L_{MMD-VAE} = MMD(q\phi(z)\|p(z)) + E_{pdata(x)}E_{q\phi(z|x)}[\log p\theta(x|z)] \quad (2)$$

For VQ VAE, instead of being a continuous distribution, the Element Code was limited to some quantized values, which were learned during training. The objective loss function was defined as eq 3.[28]

$$L_{VQ-VAE} = \left\|sg[z] - z_q\right\|^2 + \beta\left\|z - sg[z_q]\right\|^2 + E_{pdata(x)}E_{q\phi(z|x)}[\log p\theta(x|z)] \quad (3)$$

In the experiment, we designated a 10 by 10 quantized grid mesh for a total of 85 elements. However, when inspecting the results, there was some Element Code being allocated at the same grid point, which indicated the lack of ability to distinguish the difference between certain elements. Therefore, the values before quantization were also preserved as another set of Element Code, named as VQ_o (Figure 2c), compared to VQ_q (Figure 2d), the one after quantization.

Because we designated Element Codes as two-dimensional data points, we were able to observe the encoding results when various loss functions were used. The visualization outcomes are shown in Figure 2. We represented the data points using a deep purple to light yellow gradient color bar to correspond with the increasing atomic number from hydrogen to radon. It was noticed that the arrangement of Element Codes had a lot to do with the atomic number. For MMD- and VQ-based VAEs, the arrangement could even be considered as a periodic table mapped in two-dimensional space. The sequence of Element Codes not only followed the order of atomic number but also exhibited a periodic feature. The result was not surprising, since the pseudopotentials that were introduced to

https://dx.doi.org/10.1021/acs.jpclett.0c02393
J. Phys. Chem. Lett. 2020, 11, 8914−8921

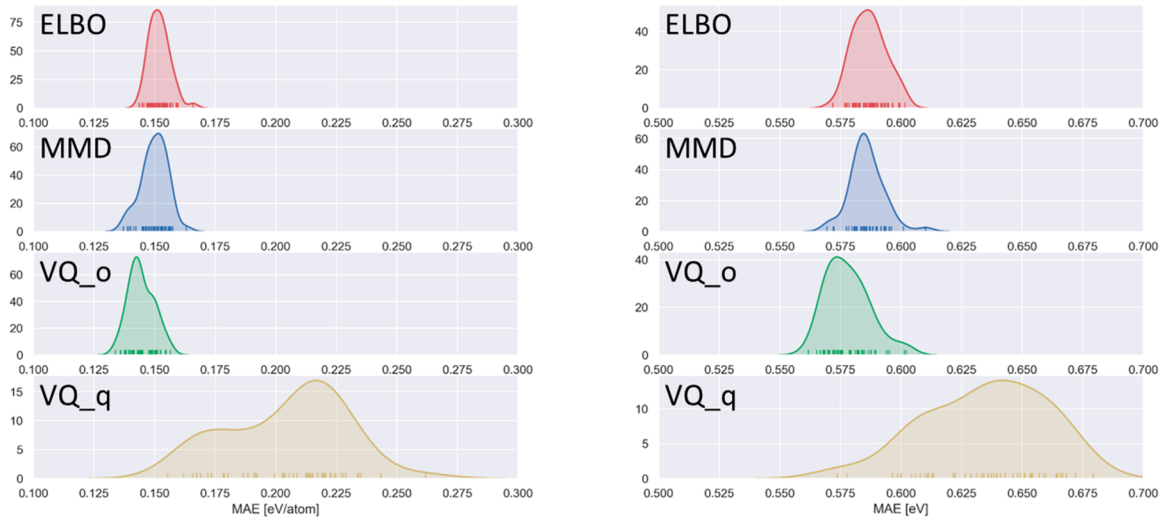


Figure 3. Performance of different Element Codes as descriptors to predict the formation energy (E_f) and band gap (E_g) of ABO_3 perovskite from the OQMD database.

generate Element Codes contained the fundamental properties of each element. This kind of elemental descriptor, showing a pattern that strongly coincides with the periodic table, has never been reported before.

The four kinds of Element Codes derived from pseudopotential were used as descriptors of an ML model to estimate their performance. Because the Element Codes belonged to the compositional descriptor, it would be rather straightforward to apply Element Codes on tasks where compositional degrees of freedom were restricted. We chose the inorganic perovskite oxides with the general formula of ABO_3 as the platform and built a simple model to make prediction of formation energy (E_f) and band gap (E_g). The material properties of inorganic perovskite oxides can be significantly influenced by the material composition and crystalline structure, making it a challenge to accurately predict E_f and E_g.[29] In the work reported by Morgan et al., 1900 samples were taken to train the model, and they achieved a root-mean-square error (RMSE) of 0.028 eV/atom using 70 element related descriptors.[30] In the recent work lead by Liu et al., a progressive learning method was introduced.[31] After the feature generation and feature reduction process, a total of 26 descriptors was fed into the model to achieve accuracy with a mean square error (MSE) of 0.268 eV, or mean absolute error (MAE) of 0.381 eV, on the prediction of E_g.

We acquired the inorganic perovskite oxides data set from the Open Quantum Material Database (OQMD), one of the abundant community-maintained online data sets.[31−33] The compounds consisting of elements beyond the scope of our Element Code were filtered out (details in the Supporting Information). The remaining 1400 compounds were split with a train-test ratio of 80:20. The gradient boosting strategy served as the algorithm.[34,35] We randomly generated 50 sets of Element Codes for each algorithm, using them as the sole descriptors to predict the material properties of ABO_3. In this specific system, we only explored the possible combinations of A and B site atoms. Therefore, when constructing the ML model, we simply needed four descriptors, the two-dimensional Element Codes for A and B site atoms. The preliminary result is shown in Figure 3. From the figure, it can be observed that VQ_q exhibited a higher error rate and less reproducible result. However, when using ELBO, MMD, and VQ_o to predict E_f and E_g, an MAE of 0.15 eV/atom and 0.57 eV were achieved, respectively. Among them, Element Code derived from VQ_o algorithm performed with the best accuracy, reaching a preliminary result with averaged MAE of 0.144 eV/atom and 0.578 eV when predicting E_f and E_g, respectively.

To further understand the predicting power of various Element Codes, we inspected the correlation matrix between each Element Code and some common properties of elements, as shown in Figure 4. It was found that the Element Codes

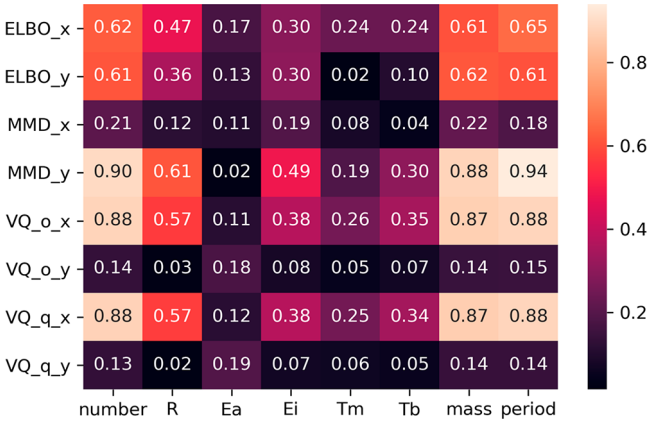


Figure 4. Correlation matrix between Element Codes and some basic element properties, including atomic number, atomic radius (R), electron affinity (Ea), ionization energy (Ei), melting temperature (Tm), boiling temperature (Tb), atomic mass, and period.

extracted from pseudopotential could easily capture the essence of some essential characteristics such as atomic number, radius, mass, density, and the period. For the Element Code derived using ELBO algorithm, the loss function encouraged the code to maximize the likelihood of a Gaussian distribution. Therefore, the information acquired from pseudopotential was evenly allocated to each dimension. On the contrary, for MMD- and VQ-based Element Codes, the information pertaining to basic element properties was stored in a certain dimension in a biased manner, leading to a high correlation factor in one dimension and a very low correlation factor in the other. Because both dimensions are required to

https://dx.doi.org/10.1021/acs.jpclett.0c02393
J. Phys. Chem. Lett. 2020, 11, 8914−8921

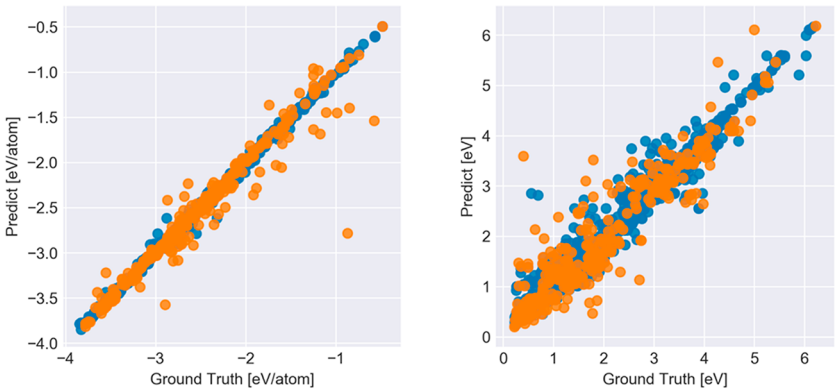


Figure 5. Training results using VQ_o-based Element Code as a descriptor after optimization of hyperparameters. The left figure shows the difference between the prediction and ground truth of E_f. The right one shows the prediction result of E_g. The blue-colored scatters belong to the training set, and the orange-colored ones belong to the testing set.

reach a low reconstruction loss when generating Element Codes from VAE, we speculated that some hidden features were embedded in the other dimension for MMD- and VQ-based Element Codes. Efficiently storing required information in both dimensions made MMD- and VQ_o-based Element Codes to perform as superior descriptors as compared to the ELBO-based one.

It is worth noting that using pseudopotential derived Element Codes to correctly depict electron affinity was challenging. For MMD-based Element Code, the correlation factors of both dimensions were 0.11 and 0.02, while VQ-based Element Code had a correlation higher than 0.1 in both dimensions. Successfully grasping the nature in the weak properties caused VQ-based Element Code to possess better predicting power than the MMD-based Element Code. When comparing VQ_o and VQ_q derived Element Code, it was intuitive to understand the inferior performance of VQ_q due to its many-to-one relation for some elements, making it impossible for VQ_q Element Code to distinguish any difference between those elements.

After demonstrating VQ_o as an effective algorithm to generate Element Code, we used VQ_o as the descriptor and fine-tuned the hyperparameters of the model to further enhance predicting power. We attempted to use as less descriptors as possible to predict the material properties. The powerful gradient boosting regression method was chosen to perform the training. The hyperparameters, such as learning rate, number of estimators, and maximum depth, were explored to minimize the loss. After training on data set of 1400 ABO_3 compounds, a decent prediction could be made on both E_f and E_g. The results are summarized in Figure 5 and Table 1. When predicting the E_f and E_g of ABO_3, an MAE of 0.077 eV/atom and 0.349 eV were achieved respectively, using five descriptors (Element Code of A and B site atoms and volume). Compared to other literature, where such a precise prediction can only be obtained using more than 10 descriptors, our result strongly implied that using the Element Code as a descriptor can facilitate the rapid search of mapping between descriptor and target material properties.

Table 1. Performance of ML Model When Predicting E_f and E_g of ABO_3 Compounds[a]

property	unit	ref[31]	this work
E_f	eV/atom	0.087 (24)	0.077 (4 + 1)
E_g	eV	0.381 (26)	0.349 (4 + 1)

[a]The MAE of the testing set is shown in the table. The number in parentheses indicates the amount of descriptors used to achieve such a performance. In our work, we use Element Code as the main descriptors (total amount of four is needed to describe A and B site atoms by Element Code). Using this strategy, only one additional descriptor is provided to achieve high accuracy.

After verifying Element Codes as efficient descriptors to achieve a highly accurate prediction, our final step was to apply them as the sole descriptors to explore possible candidates for the application in lead-free halide perovskite photovoltaics. Lead halide perovskite (chemical formula of ABX_3, with B being Pb and X composed of Cl, Br, or I) photovoltaics have been demonstrated as a promising technique to harvest energy from solar irradiance.[36−39] In a short period of one decade, vast studies of lead halide perovskites have made unprecedented progress of power conversion efficiency, from 3.8 to more than 25%.[40−43] However, there are some unsolved issues prohibiting the commercialization of lead halide perovskite solar cells. One of the largest obstacles is possible contamination from toxic lead-based compounds. To tackle this issue, research topics about searching for a suitable substitution for lead become more and more popular.[44−46] Recently, a new approach has been proposed to achieve a lead-free halide perovskite by introducing a double perovskite.[47,48] The chemical formula of a double perovskite is $A_2(B1)(B2)X_6$, featuring an alternative arrangement of corner-sharing $(B1)X_6$ and $(B2)X_6$ octahedral units. Here, we used Element Code as the only descriptor and trained a model consisting of a classifier and a regressor to predict the bandgap of inorganic double perovskite. The data set of an inorganic double perovskite was obtained from the result of a high-throughput simulation with DFT in a previous work.[49] A total number of 2400 compounds was used to train and evaluate the model, with 20% reserved as the testing set (details in the Supporting Information). The model performance on the testing set is shown in Figure 6. First, a classifier was trained to distinguish if a compound belongs to a metal (no finite bandgap) or semiconductor (with finite bandgap). This was achieved using gradient boosted decision trees as the classification algorithm. The effectiveness of classifier can be examined by the receiver operator characteristic (ROC) curve. The area under the ROC curve (AUC) reached 0.951 (Figure 6a), indicating successful classification of a semiconductor with Element Code. Afterward, a regressor was trained to predict the bandgap of those

https://dx.doi.org/10.1021/acs.jpclett.0c02393
J. Phys. Chem. Lett. 2020, 11, 8914−8921

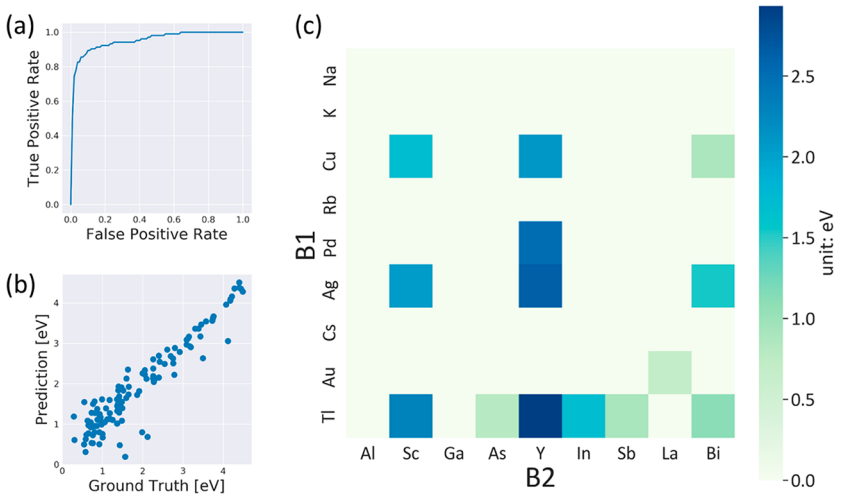


Figure 6. Performance of $A_2(B1)(B2)X_6$ bandgap prediction ML model using VQ_o Element Code as the only descriptor. (a) Receiver operating characteristic curve of trained classifier; an AUC score of 0.951 was obtained on the testing set. (b) Bandgap prediction results of trained regressor; an averaged mean absolute error of 0.266 eV was achieved. (c) The bandgap mapping of $Cs_2(B1)(B2)I_6$, with B1 and B2 composed of nine different selected elements.

regarded as semiconductors. A train-test split ratio of 80% was used. After training, the mean absolute error of 0.266 eV was obtained on the testing set (Figure 6b). Compared to the mean absolute error between density function theory calculation and experimental values, which can be as high as 0.6 eV,[50] our result strongly suggested that an efficient and accurate prediction was made based on Element Code. Finally, a survey of bandgap was done on $Cs_2(B1)(B2)I_6$ compounds by the well-trained model. The potential inorganic double perovskites can be created by replacing Pb with monovalent and trivalent cations. Therefore, we selected representative monovalent and trivalent cations as B1 and B2, respectively, and examined every possible combination to explore possible candidates. The result is shown in Figure 6c. We fixed the A and X elements as Cs and I, respectively, and focused on the potential substitution for Pb. According to the prediction, compounds with B1 consisting of Cu, Ag, and Tl and B2 consisting of Sc, Y, and Bi have a high possibility to form the desired double perovskite exhibiting finite bandgap, which can serve in the photovoltaic devices.

When setting up an ML model, an appropriate selection and engineering of the descriptor are paramount to the final predicting performance. In material science, there are many general descriptors, including atomic number, radius, ionization energy, electronegativity, electron affinity, and so on. It usually takes 10 or more descriptors for the model to achieve satisfactory accuracy. In our work, we derived a new set of descriptors, named as Element Codes, by feeding a pseudopotential through a variational autoencoder. Various algorithms were applied on a variational autoencoder to generate Element Codes with various patterns, and their performance as descriptors was compared. Using the best performing vector quantized Element Code as the sole descriptor, we constructed an ML model showing high accuracy with much fewer descriptors than reported in the literature. This can be attributed to the effectiveness of Element Code to encompass fundamental element features. Therefore, we believe Element Code generated from pseudopotential can be served as an efficient descriptor for ML model to solve a wide range of material science tasks.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jpclett.0c02393.

Experimental methods including the preparation and augmentation of the pseudopotential data set for Figure S1, the preparation of the perovskite oxide data set, the preparation of the double perovskite data set, the construction of the variational autoencoder for Figure S2, and the discussion about the pattern of Element Codes for Figure S3 (PDF)

■ AUTHOR INFORMATION

Corresponding Authors

Ming-Chung Wu − *Green Technology Research Center, Artificial Intelligent Research Center, and Department of Chemical and Materials Engineering, Chang Gung University, Taoyuan 33302, Taiwan; Division of Neonatology, Department of Pediatrics, Chang Gung Memorial Hospital, Taoyuan 33302, Taiwan;* orcid.org/0000-0002-3584-3871; Email: mingchungwu@cgu.edu.tw

Chao-Sung Lai − *Green Technology Research Center, Artificial Intelligent Research Center, Department of Electronic Engineering, and Biosensor Group, Biomedical Engineering Research Center, Chang Gung University, Taoyuan 33302, Taiwan; Department of Nephrology, Chang Gung Memorial Hospital, Taoyuan 33305, Taiwan; Department of Materials Engineering, Ming Chi University of Technology, New Taipei City 24301, Taiwan;* orcid.org/0000-0002-2069-7533; Email: cslai@mail.cgu.edu.tw

Authors

Meng-Huan Jao − *Green Technology Research Center and Artificial Intelligent Research Center, Chang Gung University, Taoyuan 33302, Taiwan*

Shun-Hsiang Chan − *Green Technology Research Center and Department of Chemical and Materials Engineering, Chang Gung University, Taoyuan 33302, Taiwan*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jpclett.0c02393

https://dx.doi.org/10.1021/acs.jpclett.0c02393
J. Phys. Chem. Lett. 2020, 11, 8914−8921

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The authors appreciate Dr. Ming-Tao Lee (BL-13A1) and Dr. Jyh-Fu Lee (BL-17C1) at the National Synchrotron Radiation Research Centre for useful discussion and suggestions. The financial support from the Ministry of Science and Technology, Taiwan (Project No. 106-2221-E-182-057-MY3, 108-2119-M-002-005, 109-2221-E-182-059, and 109-3116-F-002-002-CC2), Chang Gung University (QZRPD181), and Chang Gung Memorial Hospital, Linkou (CMRPD2H0163 and BMRPC74) are highly appreciated.

■ REFERENCES

(1) Vinyals, O.; Babuschkin, I.; Czarnecki, W. M.; Mathieu, M.; Dudzik, A.; Chung, J.; Choi, D. H.; Powell, R.; Ewalds, T.; Georgiev, P.; et al. Grandmaster Level in Star Craft II Using Multi-Agent Reinforcement Learning. *Nature* **2019**, *575*, 350−354.

(2) Silver, D.; Hubert, T.; Schrittwieser, J.; Antonoglou, I.; Lai, M.; Guez, A.; Lanctot, M.; Sifre, L.; Kumaran, D.; Graepel, T.; et al. A General Reinforcement Learning Algorithm That Masters Chess, Shogi, and Go through Self-Play. *Science* **2018**, *362*, 1140−1144.

(3) Lee, J.; Yoon, W.; Kim, S.; Kim, D.; Kim, S.; So, C. H.; Kang, J. BioBERT: A Pre-Trained Biomedical Language Representation Model for Biomedical Text Mining. *Bioinformatics* **2019**, *36*, 1234−1240.

(4) He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. B. Mask R-CNN. *arXiv*, March 20, 2017, arXiv:1703.06870.

(5) 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.

(6) LeCun, Y.; Bengio, Y.; Hinton, G. Deep Learning. *Nature* **2015**, *521* (7553), 436−444.

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

(8) Hochreiter, S.; Schmidhuber, J. Long Short-Term Memory. *Neural Comput* **1997**, *9*, 1735−1780.

(9) Butler, K.; Davies, D.; Cartwright, H.; Isayev, O.; Walsh, A. Machine Learning for Molecular and Materials Science. *Nature* **2018**, *559*, 547−555.

(10) Smith, J. S.; Isayev, O.; Roitberg, A. E. ANI-1: An Extensible Neural Network Potential with DFT Accuracy at Force Field Computational Cost. *Chem. Sci.* **2017**, *8*, 3192−3203.

(11) Wu, T.; Wang, J. Global Discovery of Stable and Non-Toxic Hybrid Organic-Inorganic Perovskites for Photovoltaic Systems by Combining Machine Learning Method with First Principle Calculations. *Nano Energy* **2019**, *66*, 104070.

(12) Ward, L.; Agrawal, A.; Choudhary, A.; Wolverton, C. A General-Purpose Machine Learning Framework for Predicting Properties of Inorganic Materials. *Npj Comput. Mater.* **2016**, *2*, 16028.

(13) Yu, Y.; Tan, X.; Ning, S.; Wu, Y. Machine Learning for Understanding Compatibility of Organic−Inorganic Hybrid Perovskites with Post-Treatment Amines. *ACS Energy Lett.* **2019**, *4*, 397−404.

(14) Jin, H.; Zhang, H. J.; Li, J. W.; Wang, T.; Wan, L. H.; Guo, H.; Wei, Y. D. Discovery of Novel Two-Dimensional Photovoltaic Materials Accelerated by Machine Learning. *J. Phys. Chem. Lett.* **2020**, *11*, 3075−3081.

(15) Mansouri Tehrani, A.; Oliynyk, A. O.; Parry, M.; Rizvi, Z.; Couper, S.; Lin, F.; Miyagi, L.; Sparks, T. D.; Brgoch, J. Machine Learning Directed Search for Ultraincompressible, Superhard Materials. *J. Am. Chem. Soc.* **2018**, *140*, 9844−9853.

(16) Deringer, V. L.; Caro, M. A.; Csányi, G. Machine Learning Interatomic Potentials as Emerging Tools for Materials Science. *Adv. Mater.* **2019**, *31*, 1902765.

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

(18) Chen, C.; Ye, W.; Zuo, Y.; Zheng, C.; Ong, S. P. Graph Networks as a Universal Machine Learning Framework for Molecules and Crystals. *Chem. Mater.* **2019**, *31*, 3564−3572.

(19) Im, J.; Lee, S.; Ko, T.-W.; Kim, H. W.; Hyon, Y.; Chang, H. Identifying Pb-Free Perovskites for Solar Cells by Machine Learning. *Npj Comput. Mater.* **2019**, *5*, 37.

(20) van Setten, M. J.; Giantomassi, M.; Bousquet, E.; Verstraete, M. J.; Hamann, D. R.; Gonze, X.; Rignanese, G.-M. The Pseudo Dojo: Training and Grading a 85 Element Optimized Norm-Conserving Pseudopotential Table. *Comput. Phys. Commun.* **2018**, *226*, 39−54.

(21) Hamann, D. R. Optimized Norm-Conserving Vanderbilt Pseudopotentials. *Phys. Rev. B: Condens. Matter Mater. Phys.* **2013**, *88*, 085117.

(22) Chen, C.; Zuo, Y.; Ye, W.; Li, X.; Deng, Z.; Ong, S. P. A Critical Review of Machine Learning of Energy Materials. *Adv. Energy Mater.* **2020**, *10*, 1903242.

(23) Hinton, G. E.; Salakhutdinov, R. R. Reducing the Dimensionality of Data with Neural Networks. *Science* **2006**, *313*, 504−507.

(24) Eraslan, G.; Simon, L. M.; Mircea, M.; Mueller, N. S.; Theis, F. J. Single-Cell RNA-Seq Denoising Using a Deep Count Autoencoder. *Nat. Commun.* **2019**, *10*, 390.

(25) Kingma, D. P.; Welling, M. Auto-Encoding Variational Bayes. *arXiv*, December 20, 2013, arXiv:1312.6114, https://arxiv.org/abs/1312.6114.

(26) Gretton, A.; Borgwardt, K.; Rasch, M.; Schölkopf, B.; Smola, A. A Kernel Method for the Two-Sample Problem. *arXiv*, May 15, 2008, arXiv:0805.2368, https://arxiv.org/abs/0805.2368.

(27) Zhao, S.; Song, J.; Ermon, S. InfoVAE: Balancing Learning and Inference in Variational Autoencoders. *Proc. AAAI Conf. Artif. Intell.* **2019**, *33*, 5885−5892.

(28) van den Oord, A.; Vinyals, O.; Kavukcuoglu, K. Neural Discrete Representation Learning. *arXiv*, November 2, 2017, arXiv:1711.00937, https://arxiv.org/abs/1711.00937.

(29) Ye, W.; Chen, C.; Wang, Z.; Chu, I.-H.; Ong, S. P. Deep Neural Networks for Accurate Predictions of Crystal Stability. *Nat. Commun.* **2018**, *9*, 3800.

(30) Li, W.; Jacobs, R.; Morgan, D. Predicting the Thermodynamic Stability of Perovskite Oxides Using Machine Learning Models. *Comput. Mater. Sci.* **2018**, *150*, 454−463.

(31) Li, C.; Hao, H.; Xu, B.; Zhao, G.; Chen, L.; Zhang, S.; Liu, H. A Progressive Learning Method for Predicting the Band Gap of ABO_3 Perovskites Using an Instrumental Variable. *J. Mater. Chem. C* **2020**, *8*, 3127−3136.

(32) 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.

(33) 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.

(34) Lu, S.; Zhou, Q.; Ouyang, Y.; Guo, Y.; Li, Q.; Wang, J. Accelerated Discovery of Stable Lead-Free Hybrid Organic-Inorganic Perovskites via Machine Learning. *Nat. Commun.* **2018**, *9*, 3405.

(35) Kailkhura, B.; Gallagher, B.; Kim, S.; Hiszpanski, A.; Han, T. Y.-J. Reliable and Explainable Machine-Learning Methods for Accelerated Material Discovery. *Npj Comput. Mater.* **2019**, *5*, 108.

(36) Correa-Baena, J.-P.; Saliba, M.; Buonassisi, T.; Graetzel, M.; Abate, A.; Tress, W.; Hagfeldt, A. Promises and Challenges of Perovskite Solar Cells. *Science* **2017**, *358*, 739−744.

https://dx.doi.org/10.1021/acs.jpclett.0c02393
J. Phys. Chem. Lett. 2020, 11, 8914−8921

(37) Oranskaia, A.; Yin, J.; Bakr, O. M.; Bredas, J. L.; Mohammed, O. F. Halogen Migration in Hybrid Perovskites: The Organic Cation Matters. *J. Phys. Chem. Lett.* **2018**, *9*, 5474−5480.
(38) Qiu, L.; He, S.; Ono, L. K.; Liu, S.; Qi, Y. Scalable Fabrication of Metal Halide Perovskite Solar Cells and Modules. *ACS Energy Lett.* **2019**, *4*, 2147−2167.
(39) Jones, T.; Osherov, A.; Alsari, M.; Sponseller, M.; Duck, B.; Jung, Y.-K.; Settens, C.; Niroui, F.; Brenes, R.; Stan, C.; et al. Lattice Strain Causes Non-Radiative Losses in Halide Perovskites. *Energy Environ. Sci.* **2019**, *12*, 596−606.
(40) Kojima, A.; Teshima, K.; Shirai, Y.; Miyasaka, T. Organometal Halide Perovskites as Visible-Light Sensitizers for Photovoltaic Cells. *J. Am. Chem. Soc.* **2009**, *131*, 6050−6051.
(41) Li, C.; Yin, J.; Chen, R.; Lv, X.; Feng, X.; Wu, Y.; Cao, J. Monoammonium Porphyrin for Blade-Coating Stable Large-Area Perovskite Solar Cells with > 18% Efficiency. *J. Am. Chem. Soc.* **2019**, *141*, 6345−6351.
(42) Alsalloum, A.; Türedi, B.; Zheng, X.; Mitra, S.; Zhumekenov, A.; Lee, K. J.; Maity, P.; Gereige, I.; Alsaggaf, A.; Roqan, I.; et al. Low Temperature Crystallization Enables 21.9% Efficient Single-Crystal $MAPbI_3$ Inverted Perovskite Solar Cells. *ACS Energy Lett.* **2020**, *5*, 657−662.
(43) Zhu, H.; Liu, Y.; Eickemeyer, F.; Pan, L.; Ren, D.; Ruiz-Preciado, M.; Carlsen, B.; Yang, B.; Dong, X.; Wang, Z.; et al. Tailored Amphiphilic Molecular Mitigators for Stable Perovskite Solar Cells with 23.5% Efficiency. *Adv. Mater.* **2020**, *32*, 1907757.
(44) Liang, L.; Gao, P. Lead-Free Hybrid Perovskite Absorbers for Viable Application: Can We Eat the Cake and Have It Too? *Adv. Sci.* **2018**, *5*, 1700331.
(45) Igbari, F.; Wang, Z.-K.; Liao, L. S. Progress of Lead-Free Halide Double Perovskites. *Adv. Energy Mater.* **2019**, *9*, 1803150.
(46) Giustino, F.; Snaith, H. Toward Lead-Free Perovskite Solar Cells. *ACS Energy Lett.* **2016**, *1*, 1233−1240.
(47) Igbari, F.; Wang, R.; Wang, Z.-K.; Ma, X.-J.; Wang, Q.; Wang, K.-L.; Zhang, Y.; Liao, L.-S.; Yang, Y. Composition Stoichiometry of $Cs_2AgBiBr_6$ Films for Highly Efficient Lead-Free Perovskite Solar Cells. *Nano Lett.* **2019**, *19*, 2066−2073.
(48) Xiao, Z.; Song, Z.; Yan, Y. From Lead Halide Perovskites to Lead-Free Metal Halide Perovskites and Perovskite Derivatives. *Adv. Mater.* **2019**, *31*, 1803792.
(49) Nakajima, T.; Sawada, K. Discovery of Pb-Free Perovskite Solar Cells via High-Throughput Simulation on K Computer. *J. Phys. Chem. Lett.* **2017**, *8*, 4826−4831.
(50) Jain, A.; Hautier, G.; Moore, C. J.; Ping Ong, S.; Fischer, C. C.; Mueller, T.; Persson, K. A.; Ceder, G. A High-Throughput Infrastructure for Density Functional Theory Calculations. *Comput. Mater. Sci.* **2011**, *50*, 2295−2310.

https://dx.doi.org/10.1021/acs.jpclett.0c02393
J. Phys. Chem. Lett. 2020, 11, 8914−8921