Descriptors of atoms and structure information for predicting properties of crystalline materials

Jonggul Lee, Jungho Shin, Tae-Wook Ko, Seunghee Lee, Hyunju Chang, Yunkyong Hyon

Materials Research Express

PAPER • **OPEN ACCESS**

Descriptors of atoms and structure information for predicting properties of crystalline materials

To cite this article: Jonggul Lee *et al* 2021 *Mater. Res. Express* **8** 026302

Materials Research Express

PAPER

Descriptors of atoms and structure information for predicting properties of crystalline materials

Jonggul Lee[1], **Jungho Shin**[2], **Tae-Wook Ko**[3], **Seunghee Lee**[4], **Hyunju Chang**[2] **and YunKyong Hyon**[3]

[1] INSERM, Sorbonne Université, Pierre Louis Institute of Epidemiology and Public Health, Paris, France
[2] Korea Research Institute of Chemical Technology, Daejeon 34114, Republic of Korea
[3] National Institute for Mathematical Sciences, Daejeon 34047, Republic of Korea
[4] Konyang University Hospital, Daejeon 35365, Republic of Korea

E-mail: hyon@nims.re.kr

Keywords: descriptors, crystalline materials, machine learning, material property

Supplementary material for this article is available online

OPEN ACCESS

RECEIVED
9 September 2020

REVISED
26 January 2021

ACCEPTED FOR PUBLICATION
3 February 2021

PUBLISHED
12 February 2021

Abstract

Machine learning (ML) has increasingly been of interest in the design of new materials. However, it is still challenging to exploit an ML model in this field because its performance highly depends on the representation of materials, its properties, and the amount of data. In this study, for the cases of prediction of properties of crystalline materials, we explore a systematic comparison of two state-of-the-art frameworks: Crystal Graph Convolutional Neural Networks (CGCNNs) and the Sure Independence Screening and Sparsifying Operator (SISSO). The common key advantage of these two models is the fact that painstakingly handcrafted descriptors from simple material properties are not required. The main differences between the two models are (1) the use of structure information in the arbitrary size of compounds (CGCNN) and (2) limited interpretability (CGCNN) but simple and analytic relations between descriptor-property (SISSO). Using these two ML algorithms we evaluate the prediction performance on the target properties, which are band gap, formation energy, and elasticity of crystalline compounds in the database of Materials Project (MP). Moreover, to improve prediction of the properties of the materials without human bias in the selection of initial atomic features for the CGCNNs, we use Atom2Vec that provides atom representation obtained in an unsupervised manner from the materials. We also perform the predictions with the different sizes of training set to investigate the data-size dependency of the predictive models. According to the amount of dataset, the use of structural information, and the ability to identify the best descriptor with its interpretability, these algorithms showed different prediction performances. This result will enable researchers in materials discovery to gain appropriate choices and insights in various attempts to improve the prediction performance of crystalline materials' properties.

1. Introduction

Accumulation of materials data and development of various machine learning (ML) algorithms are accelerating research on the prediction of material properties for the design of new materials. However, it is still challenging to predict the properties of crystalline compounds that have an arbitrary size with complex structure. To tackle this issue, the structures of crystalline compounds were limited to garnets and perovskites for the use of deep neural networks (Ye *et al* 2018), or fixed-length feature vectors (or descriptors) for atoms in the compounds using compositional properties are necessarily constructed for ML models (Zhou *et al* 2018). Finding appropriate descriptors that could enhance model performance is the key step in these frameworks, but it requires high-level domain knowledge and/or occasionally suffers from human bias.

Recently, Xie and Grossman have developed a deep learning (DL) framework that provides universal descriptors for crystalline compounds with the arbitrary size, called crystal graph convolutional neural network

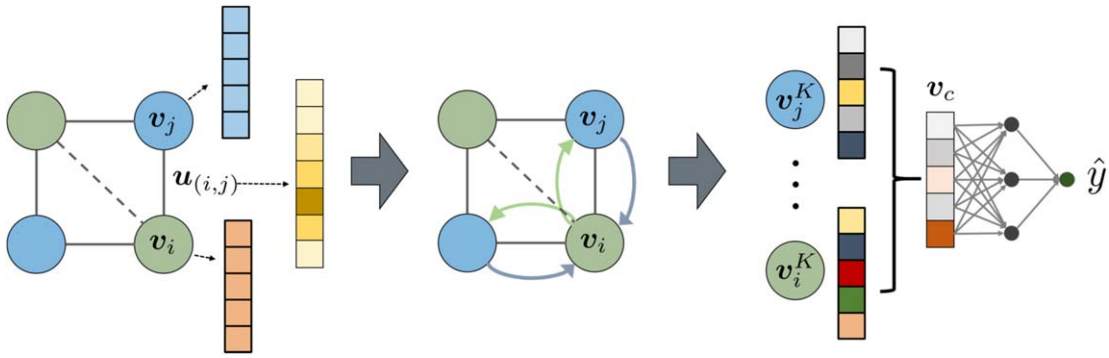

Figure 1. Structure of the crystal graph convolutional neural networks.

(CGCNN) (Xie and Grossman 2018). In this framework, crystal structures transform into crystal graphs with nodes representing atoms and edges representing connections between atoms.

Although such graph-based models are generally less sensitive to the choice of atomic descriptors, one may not rule out the possibility of better performance by better selection of atom descriptors without any human bias. For finding unbiased atom descriptors and improving prediction accuracies of target material properties, one can consider the Atom2Vec workflow (Zhou *et al* 2018) which learns the basic properties of atoms by themselves in an unsupervised way and show effectiveness over simple empirical descriptors.

Despite growing interest in the ML methods mentioned above, the amount of data that can be directed used for the prediction is usually not enough since most of the related studies of material properties have quite a narrow scope. Therefore, only a very little portion of the published data has matched with the given specific scope. Recently, a new approach, called Sure Independent Screening Sparsifying Operator (SISSO) (Ouyang *et al* 2018), showed stable prediction performances in a relatively small dataset by identifying the intrinsic relationship which is immutable between a target property and physical quantities. The relationship among physical quantities can be expressed in a mathematical equation that is composed of several descriptors. It provides a model that can be used to predict unknown properties of materials of interest like ML has performed for the same reason, but analytic function can be obtained by SISSO, while a role among input features is completely hidden in a generated ML model.

As we mentioned above, such ML models achieved a remarkable improvement in the prediction of material properties. However, for material engineers who want to apply ML to their research but are not familiar with it, exploiting the ML models is still challenging because its performance highly depends on the amount of data to be prepared/preprocessed for predictive models, and representation of atoms in compounds for their features, and appropriate targets, which are the important material properties in its design.

In this study, we investigate the difference in model performance according to the choice of atomic features—selected by empirical knowledge or by the unsupervised way (Atom2vec)—as initial input values to the CGCNN model. Assuming that the amount of data for learning a model is limited, we also examine whether SISSO could show promising prediction performance. Both models predict the properties of crystalline compounds in the database of Materials Project (MP) and investigate which model is suitable depending on the circumstance.

2. Methods

2.1. CGCNN

CGCNN is a deep learning framework for predicting material properties of crystal structures represented by crystal graphs with nodes and edges corresponding to atomic information and bonding interaction between atoms, respectively. Figure 1 shows the structure of the CGCNN framework. Let $\mathbf{v}_i$ and $\mathbf{u}_{(i,j)_k}$ be a node feature vector for an atom in ith node and a bonding feature vector for kth edge connecting two atoms in node i and j.

As in the original CGCNN work (Xie and Grossman 2018), the nine properties are encoded in the atom feature in the CGCNN model: group number, periodic number, electronegativity, covalent radius, valence electrons, first ionization energy, electron affinity, block, and atomic volume. Then, the atom feature vector $\mathbf{v}_i$ is updated by passing through a *convolutional* layer on top of the crystal graph, as following:

$$\mathbf{v}_i^{(t+1)} = \mathbf{v}_i^{(t)} + \sum_{j,k} \sigma(\mathbf{z}_{(i,j)_k}^{(t)} \mathbf{W}_c^{(t)} + \mathbf{b}_c^{(t)}) \odot g(\mathbf{z}_{(i,j)_k}^{(t)} \mathbf{W}_s^{(t)} + \mathbf{b}_s^{(t)}), \quad (1)$$

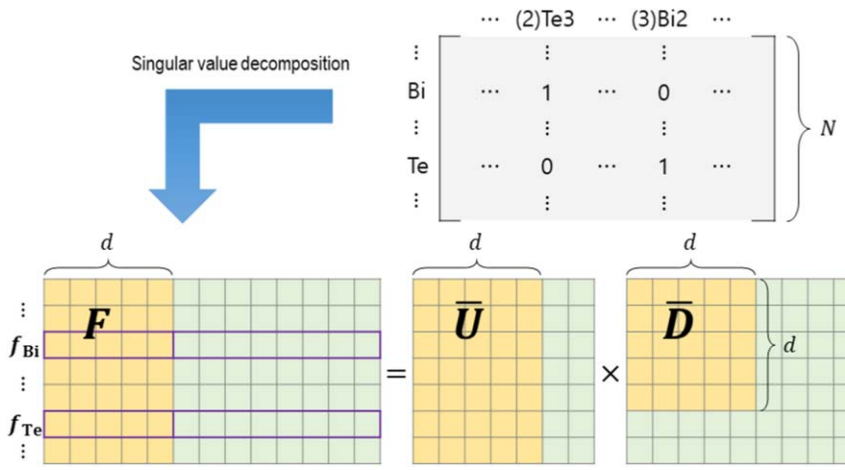

Figure 2. Schematic learning algorithm for Atom2vec. An atom-environment matrix is generated from materials database. Then, using singular-value decomposition (SVD), we can obtain row vectors in the subspace of the $\boldsymbol{d}$ largest singular values which determines the dimension of encoded feature vectors for atoms: $\mathbf{F} = [\boldsymbol{f}_1,\ldots,\boldsymbol{f}_N] = \bar{\boldsymbol{U}}\bar{\boldsymbol{D}}$ where $\bar{\boldsymbol{U}}$ is $\boldsymbol{N} \times \boldsymbol{d}$ matrix corresponding to the $\boldsymbol{d}$ largest singular values from an orthogonal matrix, and $\bar{\boldsymbol{D}}$ is $\boldsymbol{d} \times \boldsymbol{d}$ matrix from a diagonal matrix.

where $\mathbf{z}_{(i,j)_k}^{(t)} = \mathbf{v}_i^{(t)} \oplus \mathbf{v}_j^{(t)} \oplus \mathbf{u}_{(i,j)_k}$ is a feature vector concatenated all vectors of the focal node $\mathbf{v}_i$, its connecting node $\mathbf{v}_j$ and their kth bond $\mathbf{u}_{(i,j)_k}$ allowing the periodicity of the unit cell. The nonlinear convolution functions σ and g carry out convolution with weight matrix and bias, $\mathbf{W}_c^{(t)}$, $\mathbf{W}_s^{(t)}$, $\mathbf{b}_c^{(t)}$ and $\mathbf{b}_s^{(t)}$, respectively. Note that σ function is a sigmoid function to make learning deeper networks.

After updating, to get a global vector representation of the crystal $\mathbf{v}_{\mathcal{G}}$, the normalized summation of every node features is used as a *pooling* layer, as follow:

$$\mathbf{v}_c = \frac{1}{N}\sum_i \mathbf{v}_i. \tag{2}$$

Finally, the global vector representation of the crystal goes through a fully connected layer to obtain the material property ($\hat{y}$).

2.2. Atom2Vec

Atom2Vec introduced in (Zhou *et al* 2018) is unsupervised learning of atoms from a database and provides atomic descriptors that capture well the similarities and properties of atoms in a vector space and show their enhanced effectiveness over simple empirical descriptors in ML problems for materials discovery. Atom2Vec lets machines learn atom representation from only the existence of compounds (or *environments*) in a materials database. To represent how atoms are bond together to form their environments, atom-environment pairs are generated from the chemical composition of a compound: Each atom is selected as a target, and the counts of all remaining atoms are represented as a corresponding environment to the target one, and of itself. For example, *Bi*2*Te*3 generates two atom-environments pairs: "*Bi*"-"(2) *Te*3", and "*Te*"-"(3) *Bi*2". Then, from a materials database, we build atom-environment matrix $\mathbf{X}$ whose $(i,\ j)$ entry represents the number of pairs where the ith atom and the jth environment appear together. The authors in (Zhou *et al* 2018) proposed two types of learning algorithms for Atom2Vec, one is *model-free* and the other *model-based* machines. Nevertheless, to exclude any probability model to describe connections between atom and environment, we here only consider a model-free machine. In the model-free machines, the normalization of the atom-environment matrix is applied to its row vector to overcome imbalanced atom distribution in the environment. Then, using singular-value decomposition (SVD) (Sapper and Hinderliter 2013), we can obtain row vectors in the subspace of the d largest singular values which determines the dimension of encoded feature vectors for atoms and is fixed as $d = 20$ in this work (figure 2).

A CGCNN framework requires knowledge-based features of atoms, for instance, group number, period number, electronegativity, etc. These features convert to the atom representation by one-hot encoding. Even though the properties are simple and enable to use of atom coordinate without complex transformation, the initial atom representation might have high feature dimensions or highly-collated features. Besides, feature selection plays a role in the enhancement of prediction performance, but it could be varied relying on researchers' selection of atom properties by insights or domain knowledge. Here we use Atom2Vec to provide atom representation learning from the various structures of crystalline materials and feed the initial atom representation into CGCNNs for comparison with the empirical atomic features.

2.3. SISSO

SISSO is an algorithm that can be used to discover the intrinsic relationship which is immutable between a target property and physical quantities based on the compressed-sensing method (Ouyang *et al* 2018). Since its performance is not dependent on the size of an input dataset, it has been known to provide stable results even though the training set is relatively small.

In this paper, SISSO has been employed to obtain the mathematical equations composed of atomic features and to predict unknown materials' properties. We have used thirty-six kinds of physical quantities of each atom as the input features; atomic radius, atomic volume, atomic weight, and so on (see all in figure S1 in SI (available online at stacks.iop.org/MRX/8/026302/mmedia)). To find the relationship between such atomic features and material properties, SISSO investigates their combinations which means combining more than one atomic features by the given mathematical operator set which is defined as

$$H^{(m)} \equiv \{+, -, \times, \div, \exp, \log, |-|, \sqrt{}, \hat{}(-1), \hat{}2, \hat{}3\}[\phi\_1, \phi\_2],$$

where ϕ_1 and ϕ_2 are objects in Φ and, the superscript (m) denotes that physically meaningful operations are only allowed. While Φ_0 is the set of raw atomic features excluding any operation process, Φ_1 is the set of collections of the equations composed of ϕ_1, ϕ_2 and the single operator from the operator set. The feature space is recursively defined with an increment of n as follows:

$$\Phi_n \equiv \bigcup_{i=1}^{n} \hat{H}^{(m)}[\phi_1, \phi_2], \ \forall\, \phi_1, \phi_2 \in \Phi_{i-1}$$

All feature space of Φ_n for $n = 1, 2, 3$, have tested to find the best descriptors (d_{nD}) to predict the target property through combinatorial optimization which means the given features are handled to minimize the errors between the target values and the predictive values in an iterative way; once the best descriptor is chosen from the allowed set of features, SISSO goes to detect another descriptor that can be used to make the best equation describing target property by the linear combination of two descriptors and coefficients multiplied for each.

$$\mathrm{P} \equiv \sum_{i=1}^{n} (d_{nD} \cdot c_{nD}) \equiv d_{1D} \cdot c_{1D} + d_{2D} \cdot c_{2D} + \cdots + d_{nD} \cdot c_{nD}$$

The first descriptor is called as a 1D descriptor (d_{1D}), and the second one is a 2D descriptor (d_{2D}). This multidimensional equation can be extended to more than 3D to find the more exact equation. Finally, the best performance model is chosen and used for the prediction of the test set. In this process, selecting the most effective atomic features is very critical, following the considerable effort of careful feature engineering.

2.4. Dataset

The Material Projects (MP) database, including a variety of computed properties such as crystal structure, electronic band structure, and energy, was used for training and validation of the CGCNN models: inorganic crystalline compounds (Jain *et al* 2018). The MP database (pymatgen = 2019.3.27) has 124,515 inorganic compounds with 89 elements and 227 space groups, and almost 9 out of 10 compounds are binary, ternary, and quarternary.

We focused on predicting four basic properties of crystalline materials: formation energy (E_f), band gap (E_g), bulk modulus (K_{VRH}), and shear modulus (G_{VRH}). Before training the predictive models for the target properties, we prescreened the data obtained from poorly converged calculation and multivalent cathode projects in MP. Note that only binary and ternary compounds were considered in this study because the feature space in SISSO rapidly grows with the elements in a compound: The total number of binary and ternary compounds was 000 for formation energy, 000 for band gap, and 000 for bulk and shear moduli (figure 2). To investigate the data-size dependency of the predictive models, a thousand of crystalline materials was randomly set apart from the whole dataset. Each model was trained with 80% of the data and then tested with 20% of the data.

Note that Atom2Vec can learn the representation of atoms from a massive database, and the learned vectors can be used as general descriptors in ML problems with a different database in materials science. This fact enables us to independently learn atom representations from the whole MP database and to use the learned vectors in the CGCNN model.

3. Results and discussion

3.1. CGCNN with the initial atom embeddings

We compared atom vectors from Atom2Vec with the atom representation from the empirical and random features and investigated which atom vector is more effective in use for supervised learning tasks of formation

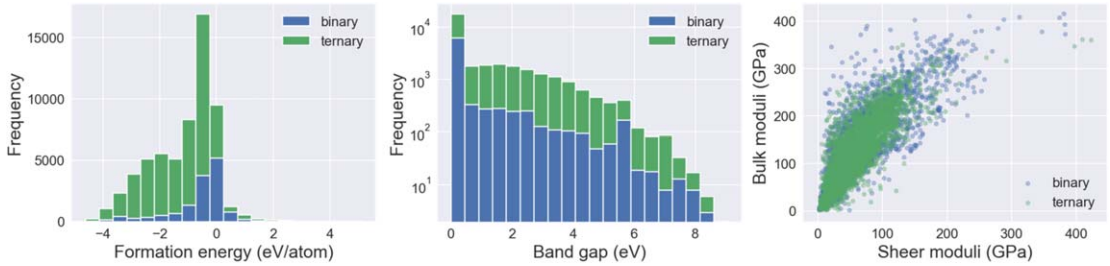

Figure 3. Statistics of the MP. Formation energy per atom, band gap, and elasticity distributions are displayed in the left, middle, and right panel, respectively.

Table 1. MAEs through the CGCNN algorithm.

Property	Unit	MAE_{small}				MAE_{whole}			
		Binary		Ternary		Binary		Ternary	
		Atom2vec	Empirical	Atom2vec	Empirical	Atom2vec	Empirical	Atom2vec	Empirical
Band gap	*eV*	0.5374	0.5168	0.7448	0.7253	0.2561	0.2663	0.3439	0.3446
Formation energy	*eV/atom*	0.2435	0.2418	0.2440	0.1779	0.0925	0.0842	0.0565	0.0492
Bulk moduli	log(*GPa*)	0.2224	0.2351	0.2768	0.2779	0.1783	0.1900	0.1800	0.1872
Shear moduli	log(*GPa*)	0.3112	0.3206	0.2812	0.2842	0.2555	0.2697	0.2122	0.2212

Table 2. R^2 scores through the CGCNN algorithm.

Property	Unit	R^2_{small}				R^2_{whole}			
		Binary		Ternary		Binary		Ternary	
		Atom2vec	Empirical	Atom2vec	Empirical	Atom2vec	Empirical	Atom2vec	Empirical
Band gap	*eV*	0.3541	0.3766	0.4231	0.4896	0.8313	0.8208	0.8613	0.8616
Formation energy	*eV/atom*	0.8585	0.8524	0.8877	0.9460	0.9710	0.9728	0.9929	0.9947
Bulk moduli	log(*GPa*)	0.8419	0.8118	0.7368	0.7580	0.8719	0.8552	0.8696	0.8647
Shear moduli	log(*GPa*)	0.7978	0.7820	0.7587	0.7470	0.8420	0.8300	0.8557	0.8466

energy, band gap, and elasticity predictions. For comparison, we fixed the architecture of the CGCNN models in all the dataset: three convolution layers with atom representation of length sixty-four, one fully-connected layer with one hundred twenty-eight hidden units after the pool layers. For training, Adam optimizer (Kingma and Ba 2015) with a learning rate of 0.001 was used. Then, we fed the three different types of atom representations in the CGCNN model and examined the prediction performance for the targets, formation energy, band gap, and elasticity in inorganic compounds.

Model performances in the predictions utilizing the two atom representations are summarized in figure 3, tables 1 and 2 for mean absolute error (MAE), and R^2 score, respectively. Overall, the two different atom representations have similar R^2 scores in all target properties; the difference of the averaged R^2 scores in the atomic featurizations are less than 0.2. The initial atom features generated by empirical featurization led to the better performance of band gap and formation energy prediction than the initial atom features from Atom2vec for both the small and whole dataset, except band gap prediction for the whole dataset. In particular, CGCNN with empirical atom representation to predict formation energy outperformed those with Atom2vec atom representation for the ternary compounds in the small dataset. Meanwhile, CGCNN with Atom2vec achieved slightly better performances of elasticity prediction for the binary and ternary compounds in the small and whole dataset. In particular, R^2 scores of band gap prediction almost doubled for both the binary and ternary compounds. Note that the length of the initial atom vector by the unsupervised learning of atom in Atom2vec is much less than those of the empirical one. Nevertheless, the atom representation from the Atom2vec showed comparable or better performances to the empirical one, independent of the size of the database, the type of

Table 3. MAEs through the SISSO algorithm.

Property	Unit	MAE_{small}		MAE_{whole}		MAE_{gain} (%)	
		Binary	Ternary	Binary	Ternary	Binary	Ternary
Band gap	*eV*	0.68	0.89	0.51	0.93	+33.3	−4.3
Formation energy	*eV/atom*	0.36	0.39	0.31	0.38	+16.1	+2.6
Bulk moduli	log(*GPa*)	0.44	0.44	0.32	0.37	+25.7	+18.9
Shear moduli	log(*GPa*)	0.41	0.49	0.41	0.40	0.0	+22.5

Table 4. R^2 scores through the SISSO algorithm.

Property	Unit	R^2_{small}		R^2_{whole}		R^2_{loss}	
		Binary	Ternary	Binary	Ternary	Binary	Ternary
Band gap	*eV*	0.16	0.41	0.66	0.44	−0.50	−0.03
Formation energy	*eV/atom*	0.65	0.76	0.80	0.78	−0.15	−0.02
Bulk moduli	log(*GPa*)	0.52	0.54	0.63	0.62	−0.11	−0.08
Shear moduli	log(*GPa*)	0.56	0.48	0.62	0.58	−0.05	−0.10

compound, and the task. This result is line with which Zhou *et al* also showed that the effectiveness of Atom2vec in use for a supervised learning task for formation energy prediction of elpasolite crystals ABC_2D_6.

The best descriptors to predict target properties extracted from SISSO output and are presented with the evaluated correlation to the target values in Table S1 (in Supplementary). All the correlation values of the formation energies are systematically higher than the others, thus more reliable. From those results, it is remarkable that only the atomic feature of arr (Atomic radius by Rahm *et al*) is related all every descriptor for the formation energies in the table, and we propose it could be regarded as the most important feature for the given property. This relationship between atomic radius and formation energy was found by numerous comparisons among the presented atomic features and mathematical combinations of them.

3.2. Atom representation in SISSO

We have employed the SISSO algorithm to predict the properties listed above by using the same dataset as that of ML. (atomic features from Mendeleev, Rung = 2, 3D descriptors) (Supplementary Info.: about input atomic features, and details of running SISSO, version: SISSO 1.0) The MAE outcomes from SISSO results are systematically larger than that of ML due to two reasons. It seems to be mostly related to the lack of structural information of the materials in the input features. Another reason is the MAEs have not been fully minimized since high accuracy setting has not been considered. Therefore, the direct comparison of the performances between SISSO and ML methods presented below (table 3) could not be fair.

On the other hand, we could discuss the comparison of two different results by only SISSO: the difference due to the change of the dataset size. For the comparison, we introduced MAE^{gain} that means the increment of the MAEs in percentage when changing the input dataset to the small one from the whole one, and it is given by

$$MAE^{gain}(\%) = \frac{MAE^{small} - MAE^{whole}}{MAE^{whole}} \times 100.$$

In the comparison, ML prediction results show that MAE is changed dramatically depending on the size of the dataset. On the contrary, SISSO results systematically maintain the similar accuracies while the size of the dataset is reduced; especially the MAE of E_f of binary and G_{VRH} of ternary is increased slightly, even though quite a smaller data subset has been used. In the case of E_g of ternary, the MAE is decreased at the smaller dataset, surprisingly. The similar trend found in the R^2 scores presented in table 4: R^2_{loss} of ternary compounds shows that the model performance between small and whole dataset is not too large. However, in the case of the binary compounds in the small dataset, R^2 scores are seriously increased and its behavior is similar to those of the ML results. It seems that the intrinsic relationship between atomic features and target values has been lost in the test set of binary compounds, while the importance of the crystal geometry on the ML methods are strongly dependent on the change of the dataset size.

Next, we consider only ten atomic features, while a total of thirty-six atomic features is available from the results of the additional feature engineering process described in SI. Its prediction results are presented in tables 3 and 4 for MAEs and R^2scores, respectively. To investigate such larger feature space, additional steps have been performed for the formation energy of the binary compounds of the *small* dataset by using such 36 atomic

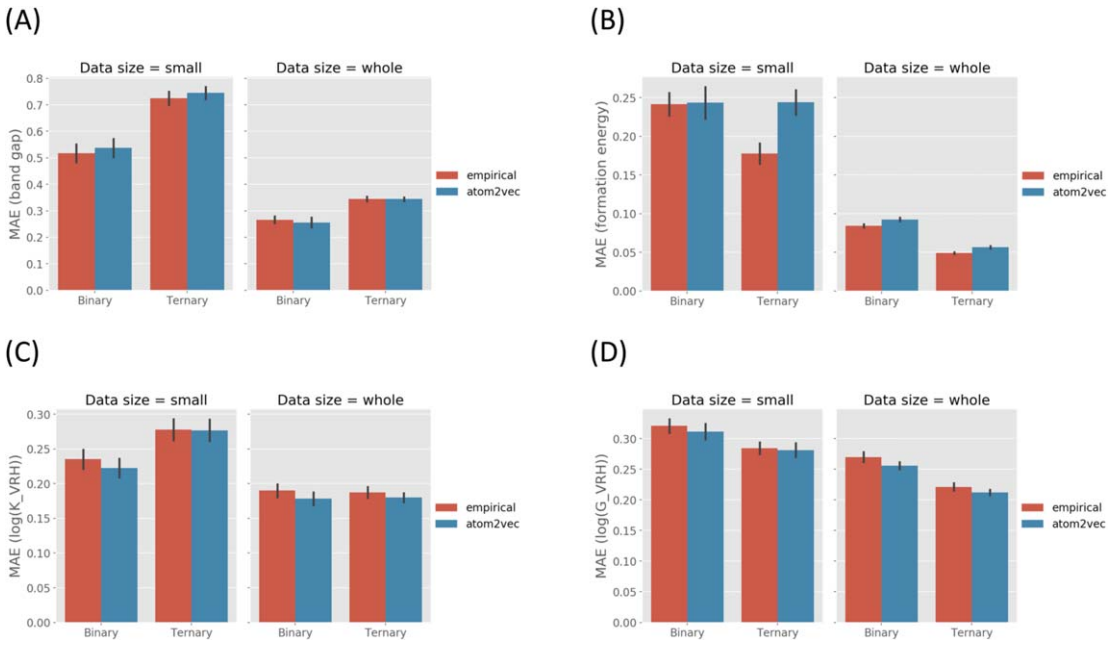

Figure 4. Mean absolute error of the band gap (A), formation energy per atom (B), and bulk (C) and shear moduli (D) for the different size (small or whole) and compound (binary or ternary) of the dataset. Atom features from the empirical combination (red bar) and Atom2vec (blue bar) were used to feed into the CGCNNs.

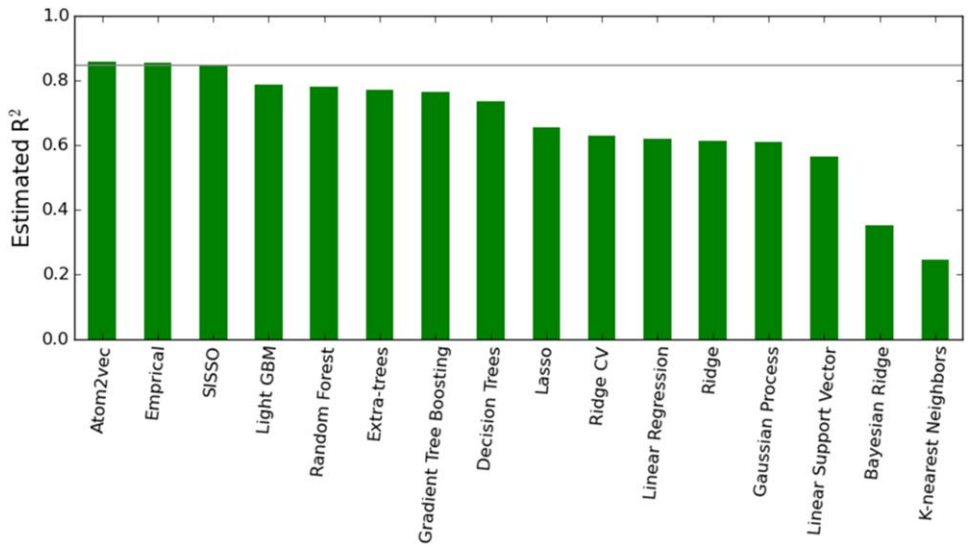

Figure 5. R^2 scores through CGCNN and SISSO (36 atomic features) compared with various ML algorithms.

features as presented in figure 4. It allows more combinations of input features and then the possibility of high-end performance of SISSO in the prediction of the target properties has been checked. Finally, the obtained R^2 score with 37 features for the formation energy is 0.846, which is comparable to that of the state-of-the-art approach based on crystal graphs as a structural representation such as CGCNN when the data set of the restricted size (*small* one) was used. The advantage of the SISSO combined with the *Mendeleev* Python library is its high availability that does not require any pre-defined structural information for the target compounds. This allows us to skip the long time process of DFT calculations. As a result, it could be used as an alternative of CGCNN methods when the given dataset has deficient information about atomic topologies and has only compositional information of the target compounds.

As we have especially focused on two types of AI-based predictive tools of CGCNN and SISSO which have recently developed in this area, their performances were presented as the top 3 results in comparison with that of

various ML algorithms in figure 5. The default hyperparameter setting of the SciKit Learn Python module has been used for all the ML algorithms.

4. Conclusions

In this study, we introduced the two state-of-art ML frameworks, CGCNN and SISSO, to predict crystalline materials' properties such as band gap, formation energy, and bulk and shear moduli. CGCNN outperformed SISSO regardless of the size and type of the dataset, and its prediction tasks as other DL approaches have been done for the case in the field of materials science, recently (Gilmer *et al* 2017, Sutton *et al* 2018). CGCNN is also able to have the interpretability of an embedding vector in a specified structure such as perovskites. However, CGCNN needs some domain knowledge to avoid the human bias of atomic characteristics for vector embedding. Meanwhile, SISSO showed consistent results in prediction performance even in the small dataset. Also, SISSO could provide the best descriptor out of a large space of mathematical combinations of simple atomic features, and identify the relationship between atomic descriptors and property in terms of an analytical equation (Ouyang *et al* 2018).

To avoid human bias on feature selection for the initial atomic descriptor, we adopted Atom2vec. The atom representation from Atom2vec showed comparable or better performances to the empirical one, regardless of the size of the database, the type of compound, and the task. This showed the effectiveness in use for the supervised learning task of materials properties over empirical atomic features. Therefore, Atom2vec for the initial atom embedding in graph-based neural networks would be a better choice for researchers who want to apply it to the general task for materials discovery than the empirical one.

Although ML for materials design and discovery has been improving, it is still challenging to develop a model with high-throughput computation and precise prediction for most properties of crystalline compounds; researchers looking to introduce an ML method for this field are forced to spend a lot of time and effort in finding the appropriate model and descriptors according to the given dataset and the target properties. We assumed the various situations that these researchers might experience, such as limitation of the amount of training data, representation of atomic features, and the availability of structural information, and compared the performance of the state-of-art machine learning models for the target properties. We expect that this result will enable researchers in materials discovery to gain appropriate choices and insights in various attempts to improve the prediction performance of crystalline materials' properties.

Data availability statement

The data generated and/or analysed during the current study are not publicly available for legal/ethical reasons but are available from the corresponding author on reasonable request.

ORCID iDs

YunKyong Hyon https://orcid.org/0000-0002-5995-9748

References

Gilmer J, Schoenholz S S, Riley P F, Vinyals O and Dahl G E 2017 Neural Message Passing for Quantum Chemistry *ICML 2017 Proceedings of the 34th International Conference on Machine Learning* **70** 1263–72 (http://arxiv.org/abs/1704.01212)

Jain A *et al* 2018 The materials project: accelerating materials design through theory-driven data and tools *Handbook of Materials Modeling : Methods: Theory and Modeling* ed W Andreoni and S Yip (Springer Nature Switzerland: Springer International Publishing) pp 1–34

Kingma D P and Ba J L 2015 Adam: a method for stochastic gradient descent *ICLR: Int. Conf. on Learning Representations* (https://arxiv.org/pdf/1412.6980.pdf)

Ouyang R, Curtarolo S, Ahmetcik E, Scheffler M and Ghiringhelli L M 2018 SISSO: a compressed-sensing method for identifying the best low-dimensional descriptor in an immensity of offered candidates *Phys. Rev. Mater.* **2** 1–11

Sapper E and Hinderliter B 2013 Computational tools and approaches for design and control of coating and composite color, appearance, and electromagnetic signature *Coatings* **3** 59–81

Sutton C, Ghiringhelli L M, Yamamoto T, Lysogorskiy Y, Blumenthal L, Hammerschmidt T, Golebiowski J, Liu X, Ziletti A and Scheffler M 2018 *NOMAD 2018 Kaggle Competition: Solving Materials Science Challenges Through Crowd Sourcing* (http://arxiv.org/abs/1812.00085)

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

Ye W, Chen C, Wang Z, Chu I H and Ong S P 2018 Deep neural networks for accurate predictions of crystal stability *Nat. Commun.* **9** 1–6

Zhou Q, Tang P, Liu S, Pan J, Yan Q and Zhang S-C 2018 Learning atoms for materials discovery *Proc. Natl Acad. Sci.* **115** E6411–7