A Fourier-transformed feature engineering design for predicting ternary perovskite properties by coupling a two-dimensional convolutional neural network with a support vector machine (Conv2D-SVM)

Chenebuah, Ericsson Tetteh; Nganbe, Michel; Tchagang, Alain Beaudelaire

Publisher's version / Version de l'éditeur:
https://doi.org/10.1088/2053-1591/acb683
Materials Research Express, 10, 2, 2023-01-27

Materials Research Express

PAPER

A Fourier-transformed feature engineering design for predicting ternary perovskite properties by coupling a two-dimensional convolutional neural network with a support vector machine (Conv2D-SVM)

Ericsson Tetteh Chenebuah[1,2,*], **Michel Nganbe**[1] and **Alain Beaudelaire Tchagang**[2]

[1] Department of Mechanical Engineering, University of Ottawa, 161 Louis-Pasteur, Ottawa, ON, K1N 6N5, Canada
[2] Digital Technologies Research Centre, National Research Council of Canada, 1200 Montréal Road, Ottawa, ON, K1A 0R6, Canada
[*] Author to whom any correspondence should be addressed.

E-mail: echen013@uottawa.ca, mnganbe@uottawa.ca and alain.tchagang@nrc-cnrc.gc.ca

Keywords: perovskite, fourier transformation, convolutional neural network, energy bandgap, support vector machine

Supplementary material for this article is available online

OPEN ACCESS

RECEIVED
25 September 2022

REVISED
25 January 2023

ACCEPTED FOR PUBLICATION
26 January 2023

PUBLISHED
7 February 2023

Abstract

In computational material sciences, Machine Learning (ML) techniques are now competitive alternatives that can be used in determining target properties conventionally resolved by *ab initio* quantum mechanical simulations or experimental synthesization. The successes realized with ML-based techniques often rely on the quality of the design architecture, in addition to the descriptors used in representing a chemical compound with good target mapping property. With the perovskite crystal structure at the forefront of modern energy materials discovery, accurately estimating related target properties is even of high importance due to the role such properties may have in defining the functionalization. As a result, the present study proposes a new feature engineering approach that takes advantage of both the direct ionic features and the periodic Fourier transformed reciprocal features of a three-dimensional perovskite polyhedral. The study is conducted on about 27,000 ABX_3 perovskite structures with the stability energy, the formation energy, and the energy bandgap as targets. For accurate modeling, a feature-extracting two-dimensional convolutional neural network (Conv2D) is coupled with a prediction-enhancing Support Vector Machine (SVM) to form a hybridized Conv2D-SVM architecture. A comparison with previous benchmark evaluations reveals appreciable improvements in modeling accuracy for all target properties, particularly for the energy bandgap, for which the feature extraction approach yields 0.105 eV MAE, 0.301 eV RMSE, and 93.48% R^2. Besides, the proposed design is further demonstrated to out-perform other similar periodic feature engineering approaches in the Coulomb matrix, Ewald-sum matrix, and Sine matrix, all in their absolute eigenvalue forms. All preprocessed data, source codes, and relevant sample calculations are openly available at: github.com/chenebuah/high_dim_descriptor.

1. Introduction

Perovskites have evolved into versatile energy materials with multifunctional properties including superconductivity [1], piezoelectricity [2, 3], ferroelectricity [4–7], optoelectronics [8, 9], magnetoresistivity [10], and catalysis [11], among others. Their multifunctionality stems from their diverse stoichiometry and geometrical distortion, as several chemical elements across the periodic table can occupy distinctive ionic sites within the crystal structure. Some examples of common perovskite stoichiometries include the ternary ABX_3, double A-site, double B-site, and hybrid organic-inorganic, as well as antiperovskites. The ternary ABX_3 structure constitutes the most prevalent of perovskite compounds thanks to their well-defined ionic arrangement of constitutive chemical elements compared to their more complex stoichiometrical counterparts.

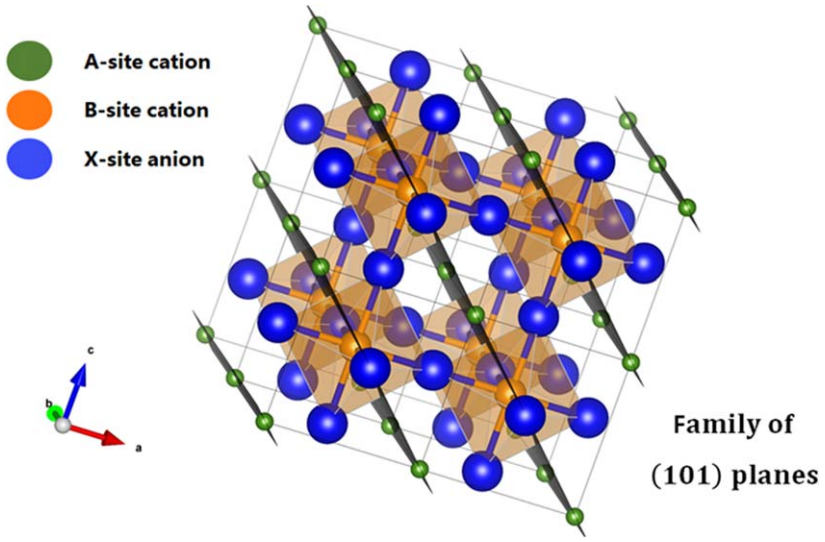

Figure 1. Cubic ABX_3 perovskite unit cell showing (101) planes in the crystal lattice. Elemental ionic property vectors are used to represent each constitutive ion in the unit cell. Unit cell structure was developed using VESTA software.

In the ideal cubic-ABX_3 configuration (figure 1), the B-site cation is coordinated by six X-site anions in a corner-sharing octahedron, whereas the A-site cation is situated at the center of a twelve-fold coordination system, all encompassed in a three-dimensional polyhedral [12]. Other derivative forms emerging from the ABX_3 stoichiometry predominantly exist in the non-idealized form. The possibility of such non-idealized geometries advantageously creates distortional complexities that are accompanied by specially tailored properties in diverse engineering applications.

Considering their great potential for a large variety of applications, estimating the properties of perovskites is currently of great interest to researchers. Conventionally, perovskite target properties are determined by experimental trials/synthetizations or via first principle (*ab initio*) quantum mechanical deterministic methods, such as Density Functional Theory (DFT) and Molecular Dynamics (MD) [13, 14]. In recent years however, Machine Learning (ML) techniques have emerged as suitable and inexpensive alternatives due to their proven predictive reliability. As applied in computational physical sciences, ML methods are generally based on solving a predefined target-forward challenge or generative-inverse problem, which could be in the form of supervised [15–21], semi-supervised [22, 23] or unsupervised learning [24, 25]. The successes realized with ML in target property prediction are often related to the quality of the distinguished features used to describe the material, in addition to the ML design architecture for accurate modeling. Among some relevant perovskite target properties of interest to scientists and engineers are the stability energy, formation energy, and energy bandgap. The stability energy (in some cases referred to as the energy above convex hull) indicates the thermodynamic state of a structure with respect to decomposition at the defined phase composition [26]. The formation energy is the energy required in forming a chemical structure from a disintegrated form and is necessary for developing phase diagrams [13, 26]. The energy bandgap quantifies the energy region between the valence band and the conduction band, and as such, is a useful indicator in characterizing the electronic state of a material (i.e. insulating, semiconducting or conducting material) [27]. Even though the aforementioned targets have been previously investigated and reported in existing ML literature, there are great needs and potentials for further predictive improvement, given the combined effect of newer descriptor designs and state-of-the-art ML techniques. In this regard, table 4 provides an outline of some recent target modeling results and reports on the respective modeling prediction technique. In a study conducted by Li *et al* [16], features that describe the bond-valence properties of ABO_3 compounds were used to predict the formation energy and bandgap at 0.087 eV/atom MAE and 0.384 eV MAE, respectively, as trained on a gradient boosting machine. Moreover, it was demonstrated in our previous work that modeling the formation energy of ABX_3 and $A_2BB'X_6$ perovskites using features that include the convex hull energy as an additional parameter can considerably improve the prediction capability up to 0.055 eV/atom MAE [17]. In a different study by Xie *et al* [18], a crystal graph convolutional neural network (CGCNN) was developed for predicting the formation energy and bandgap at 0.039 eV/atom MAE, and 0.388 eV MAE, respectively, as performed on a general inorganic crystal dataset. These are just a few examples of studies that continuously contribute in pushing the ML predictive boundaries towards first principle or experimental accuracy levels with the possibility for further modeling improvements emerging from newer methodologies.

2. Methods

2.1. Proposed modeling methodology

To contribute in further advancing modeling methodology and improving prediction accuracy, the current study proposes an advanced feature engineering method that combines two different machine learning algorithms. Building on the invertible Fourier Transformed Crystal Property (FTCP) representation [22], the developed descriptor design explores the Fourier transformed reciprocal lattice space of a periodic crystal structure and incorporates additional features for improved results. Although the case study used in the analysis focuses on the ABX_3 ternary perovskite structure, the proposed concept can be extrapolated to other forms of stoichiometrical inorganic compounds. The descriptor used for each perovskite compound is uniquely organized into a high-dimensional input image that is reminiscent of a gray-scale picture format in image recognition and object detection. The high-dimensional input image consists of both real/direct and Fourier transformed reciprocal properties that is based on demonstrated solid-state physics principles. Due to the imposition of the Fourier representations, the descriptor is therefore periodic in form and mimics the long-range atomic ordering in a crystal lattice. For accurate target modeling, a new feature engineering approach is architectured that involves a feature-extracting two-dimensional convolutional neural network (Conv2D) [28] and a prediction-enhancing Support Vector (Regression) Machine (SVM) [29, 30] model. On final training, the results reveal updated and improved performance scores on the stability energy, formation energy, and bandgap. The results are compared to some benchmark evaluation and to three other commonly used periodic descriptors in Coulomb-matrix [31, 32], Ewald-sum matrix [21, 33, 34], and Sine matrix [21], all represented in their periodic eigenvalue forms. In addition, the simulation exercise is re-performed using the original descriptor design that emerges from the FTCP elemental property matrix, thereby illustrating the pronounced effect of incorporating more features into the FTCP. The achieved results demonstrate the effectiveness of the proposed design for accurate target modeling of inorganic perovskite crystal structures.

2.2. Dataset generation and preprocessing

The present study harnesses data from the Open Quantum Materials Database (OQMD). The OQMD platform openly provides over a million entries of proven DFT calculated thermodynamic and structural properties of inorganic crystal structures [13, 35]. The OQMD was primarily chosen for core experimentation due to the robust number of trainable ABX_3 compounds available on the platform, which is needful for deep learning models. Approximately 28,000 ABX_3 structures were initially generated and extracted from the database. The dataset consists of both International Crystal Structure Database (ICSD) compounds and DFT generated compounds. The first preprocessing step involves the removal of data samples having incomplete or wrongful entries. As such, samples with missing information as related to the formation energy, bandgap and/or stability energy were ejected. The next screening process removes certain entries which possess the unfavorable potential of obscuring the modeling accuracy. The threshold used to screen such samples was set at formation energy and/or stability energy with values more than 5 eV/atom. Besides, these entries are highly impossible to synthesize given their critically unstable state. In the final screening process, only compounds with number of atoms in all crystallographic sites no more than twenty (i.e. $\boldsymbol{n_{atoms}} \leqslant 20$) are selected. The search space is limited to 20 atoms due to the relatively lower number of available samples beyond this value. Overall, the data cleaning process resulted into 27,587 ABX_3 compounds with a diverse mix of different A-, B-, and X-site chemical elements, all spanning across the periodic table, including lanthanides and actinides. The dataset also contains a good proportion of ionic-swapping inverse- or anti-perovskites [36]. Figures 2(A)–(C) are histogram plots illustrating the occurrences of contributing chemical elements, as it relates to site-1 cations, site-2 cations, and site-3 anions, respectively, from the three distinctive ionic sites in the ABX_3 compound. As can be observed, the cationic chemical elements that are associated with sites-1 and -2 are fairly distributed among the dataset. For anionic site-3 however, oxide-perovskites dominate with about 70% of all samples in the dataset. Other typical anions present in the dataset include halides such as fluorine, chlorine and bromine, and they occur relatively less at 3.8%, 3.2% and 1.8%, respectively. To illustrate the range of values for target predictive properties, Kernel Density Estimate (KDE) curves, are provided in figures 2(D), (E) and (F), showing the frequencies of the stability energy, formation energy, and bandgap, respectively. The data analytical process confirms about 6% of all samples as perfectly stable ($\boldsymbol{E_s} = 0$), 91.8% have negative formation energies ($\boldsymbol{E_f} < 0$), and 71.8% have infinite metallic bandgaps ($\boldsymbol{E_g} = 0$).

2.3. Fourier-transformed reciprocal crystal space

A crystal system that is periodic in real space with periodicity $\boldsymbol{p}$, is also periodic in the k-momentum space with periodicity $2\boldsymbol{\pi}/\boldsymbol{p}$. The k-momentum space is also referred to as the reciprocal space, and can be derived from the real lattice by implementing a Fourier transform [37]. Based on the conservation of crystal momentum, the Fourier transformation emerges into a spatial periodicity of atomic arrangements in the reciprocal space, which

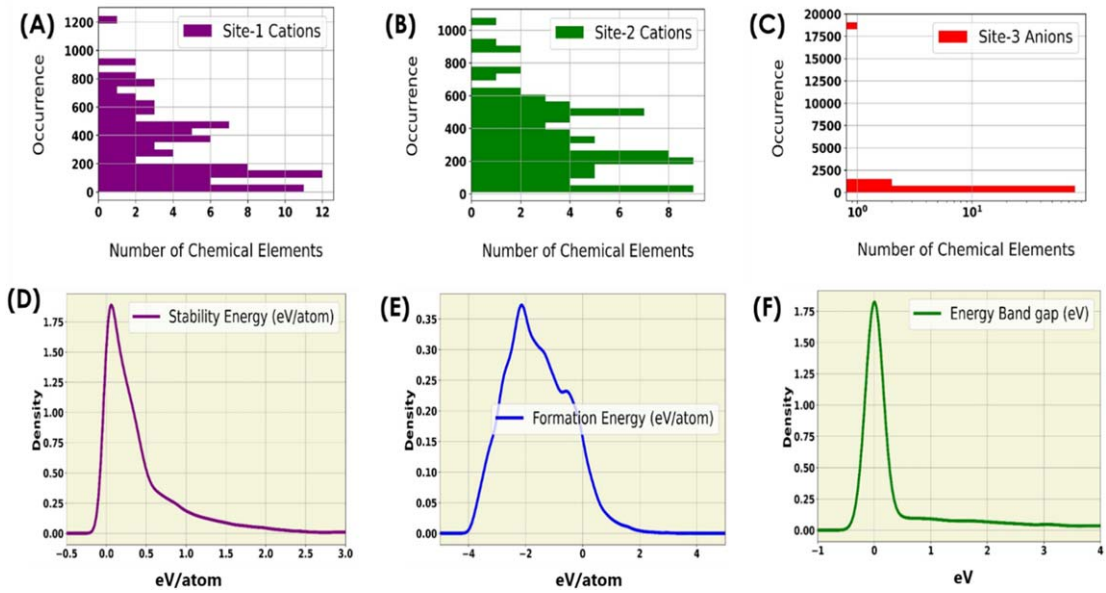

Figure 2. Histograms of contributing chemical elements for typical ionic sites in a three dimensional polyhedral ABX_3 setting: (A) cationic site-1; (B) cationic site-2; and (C) anionic site-3. Kernel Density Estimates (KDE) curves for target properties: (D) stability energy; (E) formation energy; and (F) bandgap.

results in the expression of a defined structure factor given as [38]:

$$S(G) = \int_{\substack{unit \\ cell}} e^{iGX}\, V(X).dX \tag{1}$$

G and X are integer points in the reciprocal and real lattice, respectively. $S(G)$ is the structure factor as a function of the reciprocal space, and $V(X)$ is the periodic potential or electronic density in a unit cell to be summed over all atoms. Moreover, the periodic potential can be approximated by using a periodic property that belongs to the crystal lattice. Emerging from the FTCP representation [22], the periodic scattering potential is addressed by replacing the conventional atomic form factor with the discrete properties of the ionic elements that occupy the A-, B- and X-sites in a perovskite polyhedral setting. On substituting the periodic potential with the elemental properties for a specific hkl crystal plane, the structure factor given in equation 1 can be rewritten in the form:

$$S_{hkl} = \sum_{j} Q_j e^{2\pi i(hx_j + ky_j + lz_j)} \tag{2}$$

Q_j is the elemental ionic vector to be transformed in the reciprocal space, and (x, y, z) are the fractional atomic coordinates for atom j. The ionic vector quantities used in the current study are thirteen in total (table 1) and they describe the physicochemical characteristics affiliated with each ionic element in a perovskite compound. Moreover, a pre-factor formulated as $-0.5i\ \ln[e^{2\pi i\Omega}]$is used to simplify equation 2 by eliminating the imaginary unit in the exponential term [22]**.** As a result, the structure factor is reduced to the form given in equation 3:

$$S_{hkl} = \sum_{j} Q_j \pi(\Omega) \tag{3}$$

Where $\Omega = hx_j + ky_j + lz_j$. The structure factor, as formulated using equation-3, is therefore used as inputs in the reciprocal space. A brief sample calculation based on equation 3 is demonstrated in supplementary (**section S2**) for KNO_3 perovskite structure with five atoms in the unit cell. Moreover, three other reciprocal features are introduced in the proposed descriptor design: (1) the reciprocal lattice vectors; (2) the magnitude of the reciprocal vector $|G_{min}|$normal to a crystallographic plane; and (3) the shortest distance between similar planes, d_{hkl}. The real lattice vectors are transformed into the reciprocal lattice vectors by applying and satisfying the Kronecker delta function (δ_{ij})from solid-state physics [38]. Similar crystal structure descriptors that also impose periodic conditions include: Ewald sum matrix, sine matrix, and modified coulomb matrix. In this research, the predictive performance of the developed model is equally compared with the aforementioned matrices.

Table 1. All discretized ionic features used in the present study to describe a perovskite sample in the direct space feature (DSF) block. For each one-hot encoded feature vector of an atom in a unit cell, Fourier transformation is used to project the direct/real space property into the periodic reciprocal crystal space.

Ionic chemical property	Unit	Range	Number of Bins
Group number	—	1,2,…,18	18
Row number	—	1,2,…,9	9
Pauling electronegativity [49]	Pauling	0.7–3.98	10
Covalent radius [50]	Angstrom	0.28–2.6	10
Valence	—	1,2,…,9	9
First ionization energy (log scale) [49]	eV	3.89–24.59	10
Electron affinity [49]	eV	−2.33–3.61	10
Block	—	s,p,d,f	4
Molar volume (log scale) [51]	cm^3	4.39–70.94	10
Average ionic radius [51]	Angstrom	0–1.94	10
Static average electric dipole Polarizability [49]	$10^{-24}\,cm^3$	0.21–59.42	10
Specific heat (log scale) [49]	KJ/kg.K	0.06–14.3	10
Thermal conductivity (log scale) [52]	W/m.K	0.0036–430	10

2.4. ML training via convolutional neural network (Conv2D) coupled with an auxiliary model

The convolutional neural network learns the regressive forward problem by performing non-linear operations in convolutions on the high-dimensional input image [28]. In this study, we train a deep Conv2D model to preliminarily predict the target properties from the input image. By doing so, the hidden layers connected within the architecture of the network are effectively optimized using back propagation as the respective weights are continuously updated to minimize the cost function. We therefore extract the low-dimensional features prior to the final target layer and use them as input for further training in a different algorithm. The feature extraction process is effected by customizing a callable layer within a functional *keras* API [39] on a *tensorflow* backend [40] of the Conv2D model. In general, a neuron-like processing unit can be described in the form [41]:

$$\boldsymbol{E} = \varnothing\left(\sum_{\boldsymbol{i}=1}^{N} \boldsymbol{w_i X_i} + \boldsymbol{b}\right) \tag{4}$$

Where $\boldsymbol{E}$ is the pre-trained target property; $\boldsymbol{w}$ are the updated weights that are associated with each hidden layer $\boldsymbol{i}$; $\boldsymbol{X}$ represent the input features to the unit; $\boldsymbol{b}$ is a bias; and $\varnothing$ is the activation function. For the concerned hidden neuron (feature extraction layer) in perspective, the activation function is thus non-linearly effected by introducing a *tanh* function given the equation 5:

$$\boldsymbol{h} = \tanh\left(\boldsymbol{w_h X} + \boldsymbol{b_h}\right) \tag{5}$$

$$\text{whereby, from back propagation, } \boldsymbol{w_h^{new}} = \boldsymbol{w_h^{old}} - \boldsymbol{\eta}\left[\frac{\partial \boldsymbol{\mathcal{L}}}{\partial \boldsymbol{w_h}}\right] \tag{6}$$

$\boldsymbol{h}$ represents the feature extracted layer, which is optimized by enabling a gradient descent with respect to a loss function $\boldsymbol{\mathcal{L}}$ and learning rate $\boldsymbol{\eta}$ over several epochs, as the weights are constantly updated by back-propagation.

Considering the learned target property in particular, the extracted unit possesses the high-quality attributes evolving from the high-dimensional parent image. Figure 4(C) reveals an example of the extracted low-dimensional feature representation for $NdGaO_3$, which was obtained from the pre-training process on the formation energy. The low-dimensional feature is one-dimensional (first-rank tensor) in size with vector length ($\mathbb{R}^{10}$: 1×10) and is bounded with values between 1 and −1 due to the non-linear effect of the hyperbolized tangent function (*tanh*) [42] used in activating the extracted hidden layer. For better modeling accuracy, the extracted unit is further analysed using five different auxiliary enhancement models that are comparatively studied on the Conv2D. For measuring the accuracy of each evaluation process, regressive metrics are used in the Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Coefficient of Determination (R^2). For more details on ML architecture, see section **S3** in supplementary.

3. Results and discussion

3.1. High-dimensional input image (descriptor)

The high-dimensional input image used to describe each perovskite material is structured into two separate blocks: (1) Direct Space Features (DSF) and (2) Periodic Reciprocal Features (PRF). Both blocks are adjoined along their columns to form a two-dimensional matrix representation. The DSF block contains both continuous (real-values) and discretized (one-hot encoded) attributes. The DSF block is designed using direct features in the

real space that characterizes the complete crystal structure as a whole and the individual atoms occupying the ionic sites. The continuous features, as originally provided by the database include: lattice base vectors ($\vec{R}_a$, $\vec{R}_b$, $\vec{R}_c$), calculated inter-axial angles from the base vectors (α, β, γ), and the fractional atomic coordinates of the constitutive atoms in the unit cell. Based on the present design, the maximum number of rows allocated to the fractional atomic coordinate in the DSF block coincide with the maximum number of atoms considered in the study (i.e. $n_{atoms} \leqslant 20$). For the discretized attributes in the DSF, some modifications are made to the original Fourier Transformed Crystal Property (FTCP) representations. Namely, unlike the FTCP that considers only the ionic chemical properties similarly applied in the Crystal Graph Convolutional Neural Network (CGCNN) model [18], the present descriptor design infuses new features into the DSF discrete block that fit the set objective and available resources. The newly introduced features are: average ionic radius, polarizability, specific heat, and thermal conductivity. They essentially address crucial thermo-chemistry qualities in a crystal [43], which are missing in the standard CGNN. Moreover, appending these new properties ensures that functional properties, wholly used to define some specific applications in a perovskite material, are not omitted. Table 1 summarizes all ionic elemental features transformed into discretized one-hot encoded vectors. Considering the PRF block in particular, all aforementioned discrete ionic features are projected into the reciprocal space using equation 3. Given a unit cell, the constitutive ions are Fourier-transformed using their respective fractional atomic coordinates on a specific set of crystallographic ***hkl***Miller indices. In addition to the Fourier-transformed ionic features, we further project the reciprocal lattice vectors and angles, the magnitude of the reciprocal vector $|G_{min}|$ normal to a crystal plane, and the shortest distance between similar crystallographic planes d_{hkl} , which are also missing in the standard FTCP representation. By including the reciprocal lattice vectors in the high-dimensional image, we infuse analogous variables used as inputs in first-principle calculations, and also in the analytical determination of the actual atomic form/scattering factor, as approximated by a sum of Gaussians [44–46]. Overall, 57 maximum crystallographic planes are considered in the feature projection process (i.e. $n_{max} \leqslant 57$) with the absolute summation of all ***hkl*** plane integers no more than three. By combining both DSF-PRF block arrays, the high-dimensional input image becomes complete for modeling investigation. The maximum dimension the input image can be organized into is a ($154 \times 60 \times 1$) gray-scale picture format, corresponding to (image height $\times$ image width $\times$ number of channels). Figure 3 illustrates the stacking arrangement of each feature in the high-dimensional image (for more details, see sections **S1 and S2** in supplementary). A general example is provided in figure 4, as specific to the $NdGaO_3$ (NGO) perovskite compound. The rare Earth based NGO material is a well-studied paramagnetic insulator and is extensively utilized as substrates in the fabrication of high-temperature superconducting thin-films [47, 48]. Figure 4(A) is a ball-and-stick model, displaying the three-dimensional geometry and interatomic bonding of the NGO unit cell. Emerging from the proposed descriptor design, the NGO compound is represented high-dimensionally using figure 4(B). Bright intensities, as observed on the input image, represent higher pixel values from the normalization of features.

It should be noted that representing entry samples using the proposed descriptor form is unique to a particular perovskite structure due to the imposition of the distinctive fractional atomic coordinates used in constructing the Fourier-transformed reciprocal space. Given the experimented OQMD dataset for instance, only about 8000 of all considered samples are non-duplicate entries, whereas the remaining are crystal polymorphs. In as much as polymorphic structures may have identical chemical formulas (or stoichiometry) with similar perovskites, their fractional coordinates, lattice vectors, and number of atoms in the unit cell are entirely different. With these key differences, the high-dimensional input image exclusively describes a particular perovskite sample based on their distinctive crystallographic features. Furthermore, the descriptor concept, as applied in this study, can be broadly expanded to other forms of general inorganic crystalline materials that comes with different stoichiometries (e.g. quaternary compounds). Unlike in the investigated case with the DSF column size reflecting the distinctive ionic sites of the ABX_3 setting, the descriptor used in a different stoichiometric case will have to account for the total number of original ionic site positions in a crystal lattice, in addition to the maximum number of constitutive atoms in the unit cell. Based on the strong effect of the PRF space, the target mapping quality, as obtained from the modified descriptor form, is expected to reproduce similar results as presented in the current study.

3.2. Results and discussion from the preliminary training exercise using the Conv2D model

In order to extract the high-quality attributes from the input image, the Conv2D model is first trained to predict the respective target property. As such, for all pre-training purposes, the data is split into three sets: 60% training, 20% validation (network optimization), and 20% for testing on a traditional holdout set. Table 2 reports the standardized errors on all targets as evaluated on the testing set. In general, as the number of ***hkl*** planes increases in the Fourier-transformed space, the modeling accuracy considerably improves. By comparing the error evaluations between cases without planes and with maximum planes, the accuracy in stability energy improves

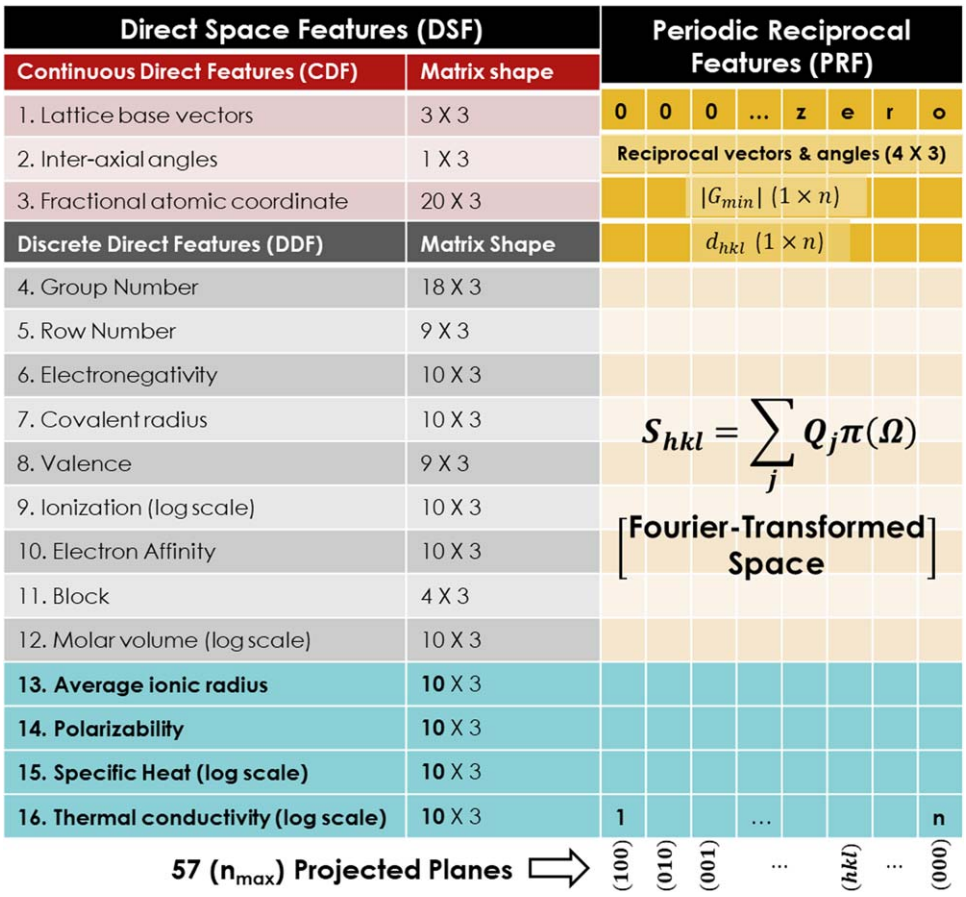

Figure 3. Structural layout of the high-dimensional input image (descriptor) used to describe a perovskite material. The input features are broadly separated into Direct Space Features (DSF) and Periodic Reciprocal Features (PRF). The PRF space imposes *descriptor periodicity*, as it relates to crystal structures.

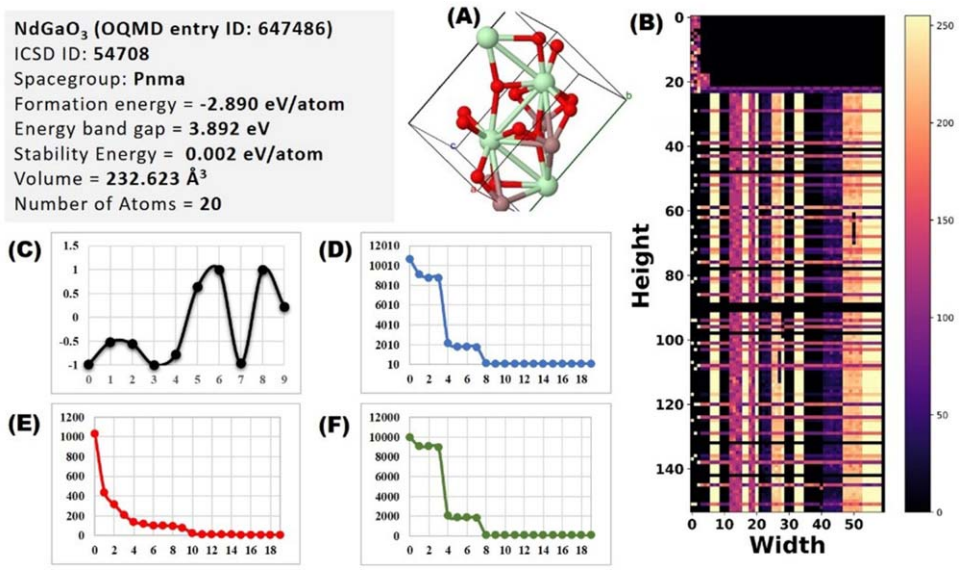

Figure 4. Periodic representations for $NdGaO_3$ (rare-Earth NGO perovskite material): (A) Ball-and-stick model; (B) Normalized high-dimensional input with image width extending to the maximum number of planes. Bright intensities on input image indicate higher pixel values; (C) low-dimensional (high-quality) feature extraction of the input-image, as pertinent to formation energy pre-training; (D) eigenvalues of the Coulomb matrix modified with periodic boundary conditions; (E) eigenvalues of Ewald-sum matrix; (F) eigenvalues of Sine matrix, all corresponding to the described NGO compound.

Table 2. Standard accuracy measurements for the stability energy, formation energy, and bandgap, as evaluated on the test set. The table reports the performance accuracy with increasing number of crystal planes/periodicity. In general, the modeling accuracy on all properties improves with the addition of more crystal planes used in the Fourier-transformation of the direct space features.

Accuracy	0 PLANES	10 PLANES	20 PLANES	30 PLANES	40 PLANES	50 PLANES	57 PLANES	57 PLANES (FTCP)
				Stability Energy				
MAE (eV/atom)	0.164	0.101	0.091	0.089	0.087	0.087	0.084	0.094
RMSE (eV/atom)	0.293	0.194	0.189	0.182	0.171	0.169	0.166	0.200
R^2 (%)	68.67	86.32	86.91	87.03	88.62	88.92	89.25	85.44
				Formation Energy				
MAE (eV/atom)	0.258	0.120	0.114	0.109	0.108	0.104	0.098	0.114
RMSE (eV/atom)	0.407	0.224	0.220	0.212	0.211	0.193	0.185	0.227
R^2 (%)	86.42	95.88	96.05	96.33	96.42	96.92	97.19	95.76
				Energy bandgap				
MAE (eV)	0.309	0.235	0.218	0.202	0.197	0.194	0.183	0.202
RMSE (eV)	0.669	0.538	0.518	0.496	0.481	0.479	0.463	0.470
R^2 (%)	66.93	79.74	82.13	83.67	83.69	83.92	84.85	84.41

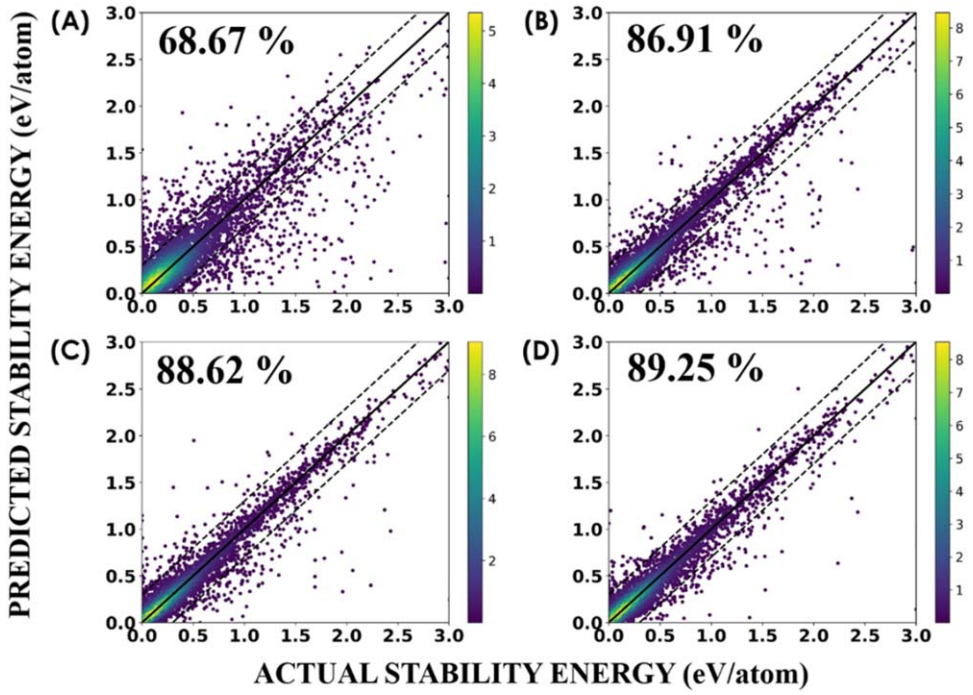

Figure 5. Test set regression fittings on the stability energy, as produced from the pre-training process on the Conv2D model: (A) zero periodicity (i.e. no Fourier-transformation with planes); (B) 20 crystal planes, (C) 40 crystal planes, and (D) 57 (maximum) crystal planes used in the transformation of direct space features. Dashed lines surrounding fitted y = x solid line are deviated at ± 0.3 eV/atom.

by 48.8% from 0.164 eV/atom to 0.084 eV/atom in MAE, and by 43.3% from 0.293 eV/atom to 0.166 eV/atom in RMSE values. Likewise, the accuracy of the formation energy improves by 62% from 0.258 eV/atom to 0.098 eV/atom MAE, and by 54.5% from 0.407 eV/atom to 0.185 eV/atom RMSE. As for the bandgap, the changes are from 0.309 eV to 0.183 eV MAE, and from 0.699 eV to 0.463 eV RMSE, which corresponds to a 40.8% and 33.8% improvement in MAE and RMSE values, respectively. On taking a closer look into the regression fittings in figures 5 and 6, the best fitting performance is realized at 97.19% R^2 on the formation energy when the maximum 57 crystal planes are all used in the periodic projection of direct features. Similarly, for the stability energy, the highest accuracy is obtained at 89.25% R^2. For the bandgap however, higher marginal errors with R^2 at 84.85% can be observed when compared to their other target property counterparts. Upon reproducing the preliminary results based on the same maximum number of projected crystal planes (i.e. 57) and the original ionic property features (from the FTCP descriptor) yields slightly lower accuracy in standardized measurements across all target properties. As previously explained, the original FTCP descriptor design does not include properties such as the average ionic radius, polarizability, specific heat, thermal conductivity, reciprocal lattice vectors and angles, and the magnitude of the reciprocal vector $|\boldsymbol{G_{min}}|$ normal to a crystal plane. As reported using table 2, the FTCP ionic makeup predicts perovskite targets at 0.094 eV/atom, 0.114 eV/atom and 0.202 eV, corresponding to about 12%, 16% and 10.4% in MAE prediction inaccuracies for the stability energy, formation energy and bandgap, respectively. This recognizes the considerable effect of the additional features incorporated into the high-dimensional image, and recommends their usage for better target-modeling result.

Generally, accurately estimating the bandgap has been a major challenge to researchers, which may in part be due to the obstacle of bandgap undervaluation from deterministic DFT simulation [53]. In the present case study, an attempt is made to further improve the bandgap's predictive capability by hybridizing the Conv2D with a coupled auxiliary model in the sequel feature-extraction arm. By doing so, even deeper trends that are associated with bandgap distribution can be further analyzed. In addition, the capability of the used descriptor to accurately classify bandgap functional properties is demonstrated for two popular classes of importance in electronic applications: (1) metallic/infinite ($\boldsymbol{E_g} = 0$) and (2) non-metallic/finite ($\boldsymbol{E_g} > 0$) bandgap perovskites. Such pre-classification tasks could assist in streamlining bandgap-targeted materials for further experimental investigation [16–18]. Figure 7 (A) reveals the results of the classification analysis. The Receiver Operating Characteristic (ROC) curve displays the proportion of correctly classified metallic perovskites (true positive rate) against the incorrectly classified non-metallic (false positive rate) at different threshold settings. The classifier performs at 99.4%, corresponding to Area Under Curve (AUC) measurements.

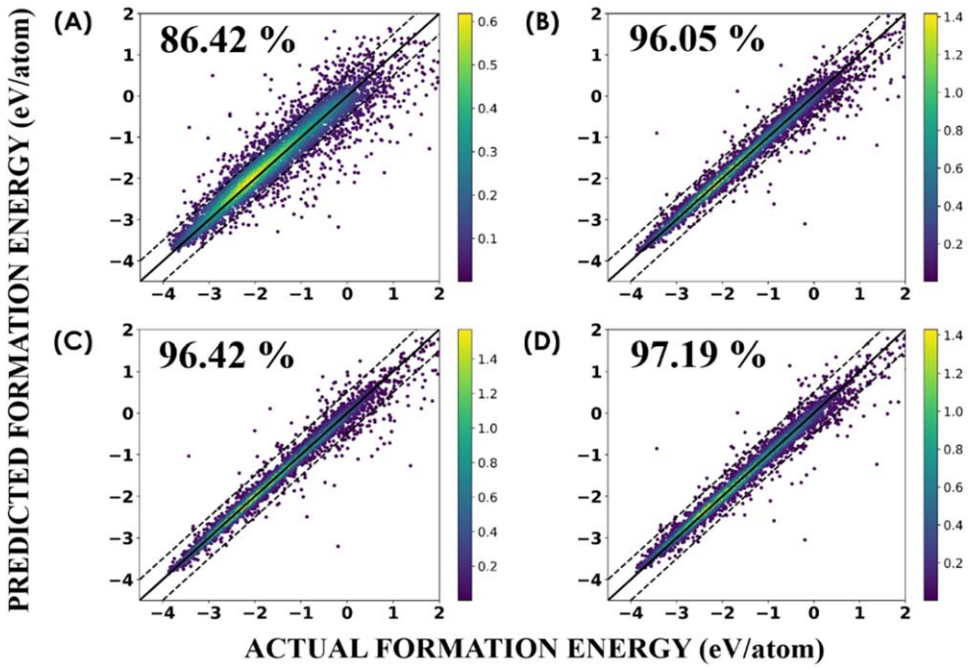

Figure 6. Test set regression fittings on the formation energy, as produced from the pre-training process on the Conv2D model: (A) zero periodicity (i.e. no Fourier-transformation with planes); (B) 20 crystal planes, (C) 40 crystal planes, and (D) 57 (maximum) crystal planes used in the transformation of direct space features. Dashed lines surrounding fitted y = x solid line are deviated at ± 0.5 eV/atom.

3.3. Results and discussion from the coupled feature extraction approach

For further training, the extracted low-dimensional feature attributes of the parent input image is used as input into auxiliary ML models. The dataset is reorganized into a $M \times N = 27,587\ (perovskites) \times 10\ (features)$ matrix; ten features corresponding to the high-quality attributes upon successful extraction from the Conv2D dense (hidden) layer. The modeling accuracy is broadly evaluated on a five-fold cross validation exercise. At first, five auxiliary models were selected and compared based on their relative performances. The considered models include: Gradient Boosting Regression (GBR) [54], Light Gradient Boosting Regression (LGB) [55], Random Forest Regression (RFR) [56], Support Vector (Regression) Machine (SVM) [29, 30], and eXtreme Gradient Boosting (XGB) [57]. Upon comparison based on standardized accuracy scores among all models, SVM is identified as the preferred option for predicting target variables. As illustrated in table 3 and figure 8(A), hybridizing the two-dimensional convolutional neural network with the auxiliary support vector machine (i.e. Conv2D-SVM) out-performs its peers. On looking into the average cross-validated scores for stability energy, formation energy, and bandgap, the MAE results for Conv2D-SVM are optimized at 0.05 eV/atom, 0.058 eV/atom and 0.105 eV, respectively. Likewise, RMSE is updated at 0.116 eV/atom, 0.133 eV/atom and 0.301 eV in the same order. Moreover, the Conv2D-SVM is a well-established fusion design as related to image-based recognition [58, 59]. In such analysis, the Conv2D model is attributed to provide strong featuristic representation of the embedded pixels, whereas SVM systematically analyzes the shallow structures emerging from the feature-extracted layers. The present study therefore points to the application potential of the Conv2D-SVM architecture in the field of computational materials science, and highlights its advanced predictive capability on target property prediction. Considering the bandgap in particular, all re-evaluated standardized metrics can be seen to greatly improve. As illustrated in figure 7, the updated regression fitting based on feature extraction for the bandgap improves to 93.48% from the previous 84.85%. Table 4 shows some benchmark results for formation energy and bandgap prediction for general inorganic crystalline materials. It can be seen that predicting the band gap (in particular) using the Conv2D-SVM model provides improved accuracies compared to other descriptor forms, highlighting on the present study contribution in the field.

3.4. Comparison with other periodic forms of crystal structure representations

Imposing periodicity with crystal structure descriptors has been found important in accounting for the long-range atomic order in crystalline materials [21, 64]. The descriptors used in the present study models the periodicity of an inorganic crystal by implementing a Fourier-transform on unit-cell ionic properties in the reciprocal lattice space. However, other forms of periodic representations exist and have been applied in past studies for predicting target properties. Therefore, the modeled performance obtained are compared to other

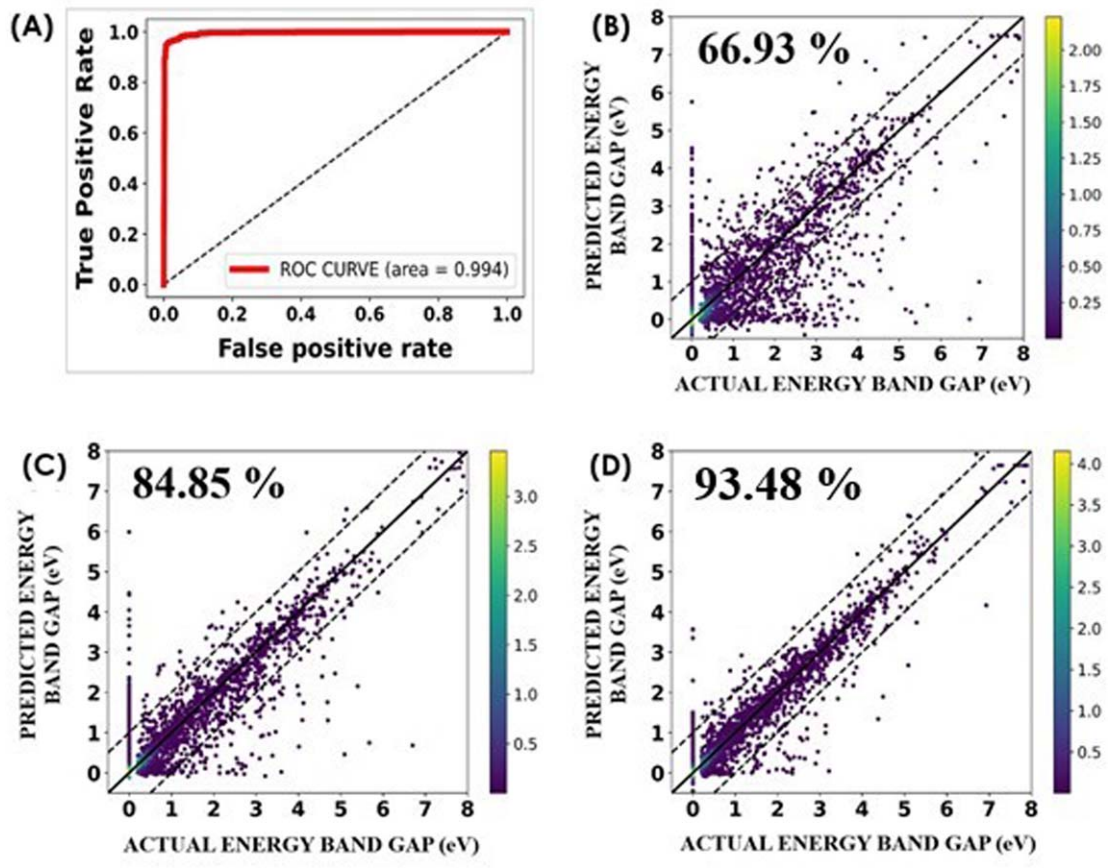

Figure 7. Energy bandgap evaluations: (A) Receiver Operating Characteristic (ROC) curve showing classification measurements for metallic and non-metallic bandgaps at different threshold settings. Area Under Curve (AUC) is reported at 99.4%; (B) Regression fitting with no planes in descriptor; (C) Regression fitting with the 57 crystal planes; (D) Regression fitting based on the feature extraction approach (Conv2D-SVM). Dashed lines surrounding fitted y = x solid line are deviated at ±1.0 eV.

Table 3. Standardized error evaluation of coupled auxiliary models to the Conv2D used for prediction enhancement.

	Stability Energy			Formation Energy			Energy Bandgap		
Auxiliary Model	MAE (eV/atom)	RMSE (eV/atom)	R^2 (%)	MAE (eV/atom)	RMSE (eV/atom)	R^2 (%)	MAE (eV)	RMSE (eV)	R^2 (%)
Conv2D-SVM	0.050	0.116	88.61	0.058	0.133	98.52	0.105	0.301	93.48
Conv2D-XGB	0.063	0.131	88.13	0.077	0.151	98.07	0.129	0.329	92.33
Conv2D-LGB	0.068	0.133	84.57	0.081	0.151	98.13	0.128	0.320	92.89
Conv2D-RFR	0.069	0.133	85.66	0.090	0.173	97.43	0.126	0.334	92.10
Conv2D-GBR	0.071	0.132	85.69	0.094	0.165	97.65	0.130	0.330	92.29
Conv2D only	0.084	0.166	89.25	0.098	0.185	97.19	0.183	0.463	84.85

commonly used *global periodic descriptors* for crystal structures. The descriptors considered for comparison include: (1) Coulomb matrix; (2) Ewald-sum matrix; and (3) Sine matrix, all with self-imposed periodic boundary conditions. The aforementioned matrices are strictly analogous to first principle Schrödinger equations due to their atomistic structure-property relation, which is typical in DFT computations. For instance, the Coulomb matrix [31] encodes the constitutive atoms and inter-atomic separations of a finite-system into a two-dimensional square array using identical equations that finds root in solving the Coulomb potential [32]. The Coulomb matrix representation can be generalized in its extended periodic form to account for the electrostatic interaction between neighboring unit cells [21, 64]. The Ewald-sum representation is an extension of the Coulomb matrix for periodic systems, and models the electrostatic interaction between constitutive atoms of a crystal, by eliminating inter-dependence between interatomic distances [21, 33, 34]. The Sine matrix rather encodes the properties of a periodic system from the respective coulombic interaction between atoms using a sine function [21]. Therefore, all perovskite samples in the dataset are described in this study using the

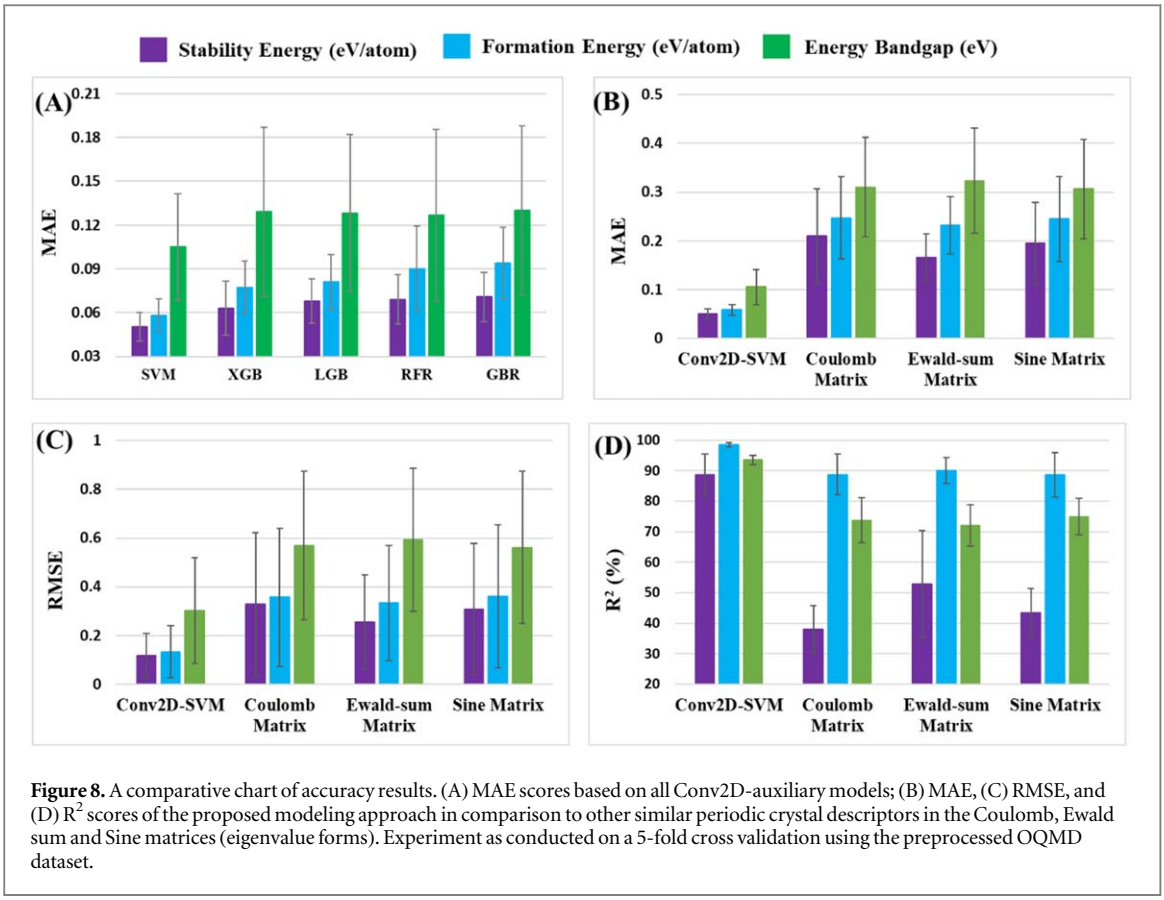

Figure 8. A comparative chart of accuracy results. (A) MAE scores based on all Conv2D-auxiliary models; (B) MAE, (C) RMSE, and (D) R^2 scores of the proposed modeling approach in comparison to other similar periodic crystal descriptors in the Coulomb, Ewald sum and Sine matrices (eigenvalue forms). Experiment as conducted on a 5-fold cross validation using the preprocessed OQMD dataset.

Table 4. Some examples of benchmark evaluation from past studies as related to formation energy and bandgap predictions. The higher accuracy of the feature extraction approach can be seen, particularly for the bandgap.

References	Prediction Technique	Accuracy evaluation	No. of training data
	Formation Energy (eV/atom)		
[18]	Crystal Graph Convolutional Neural Networks (CGCNN)	0.039 MAE	28,046
[13]	Generalized Gradient Approximation (GGA + U) OQMD	0.081–0.136 MAE DFT	292,070
[16]	Bagging (Bond-valence sum)	0.087 MAE	606
[15]	Decision Forest	0.088 MAE	228,676
[17]	Support Vector Regression (generalized features without stability energy)	0.114 MAE	1,308
[60]	Light Gradient Boosting Machine + Efficient Global Optimization (EGO)	0.160 MAE	1,250
[21]	Kernel Ridge Regression (KRR) using periodic Sine-matrix descriptor	0.370 MAE	3,000
[61]	Deep Neural Network (DNN)	0.180 RMSE, 80% R^2	510
	Feature extraction approach (Hybrid Conv2D-SVM)	**0.058 MAE**	**16,552**
	Energy Bandgap (eV)		
[62]	Random Forest	0.149 MAE	432
[63]	Materials Graph Network (MEGNet)	0.280 MAE	10,000
[16]	Gradient Boosting Regression (GBR)	0.384 MAE	606
[18]	Crystal Graph Convolutional Neural Networks (CGCNN)	0.388 MAE	16,458
[17]	Support Vector Regression with Radial Bias Function (SVR-RBF)	0.462 MAE	1,308
[14]	Generalized Gradient Approximation (GGA + U)	0.6 MAE DFT	80,000
	Feature extraction approach (Hybrid Conv2D-SVM)	**0.105 MAE**	**16,552**

eigenvalue representations of the Coulomb, Ewald sum, and Sine matrix forms. Besides, representing crystals in their differentiable eigenvalue form is suggested to be atomically invariant to translation, rotation and symmetry of neighboring atomic positions, at the expense of uniqueness [31, 65, 66]. Figures 4(D), (E) and (F) show an example of the one-dimensional eigenvalue spectra on the Coulomb-matrix, Ewald-sum matrix, and Sine matrix, respectively, for the NGO compound. Relative to the proposed feature engineering approach, the eigenvalues are now substituted as the input variables for ML training in the hybridized convolutional neural network - SVM setup. However, the convolutional network used for this training purpose is remodeled to be one-dimensional (i.e. Conv1D), given that the eigen-representative inputs are simply first order tensors of real

Table 5. Standard target accuracy measurements from the feature extraction process compared to other forms of periodic descriptors for crystal structure representation. All measurements are reported based on the average scores from a 5-fold cross-validation process.

Accuracy	Conv2D-SVM	Coulomb matrix	Ewald sum matrix	Sine matrix
		Stability Energy		
MAE (eV/atom)	0.050	0.209	0.166	0.196
RMSE (eV/atom)	0.116	0.329	0.255	0.309
R^2 (%)	88.60	37.94	52.78	43.418
		Formation Energy		
MAE (eV/atom)	0.058	0.247	0.232	0.245
RMSE (eV/atom)	0.133	0.357	0.333	0.361
R^2 (%)	98.52	88.77	90.15	88.584
		Energy Bandgap		
MAE (eV)	0.105	0.310	0.323	0.306
RMSE (eV)	0.301	0.570	0.593	0.562
R^2 (%)	93.48	73.72	71.95	74.924

numbers and not second order. Prior to training moreover, the eigenvalue features are combined with a new set of *generalized features*. The purpose of combining both feature sets is to improve the predictive performance of the periodic representation. The generalized features build on the real physicochemical properties that are associated with the ABX_3 perovskite. Moreover, they have been previously used in past literature as target-mapping descriptors [16, 17]. By combining the generalised features with the eigenvalue representations, the descriptor used to represent a crystal structure is practically periodic. Therefore, the new input dataset for experimentation is organized into:

$$\boldsymbol{X} \oplus \boldsymbol{G} = \begin{bmatrix} \boldsymbol{x}_{11} & \cdots & \boldsymbol{x}_{1N} \\ \vdots & \ddots & \vdots \\ \boldsymbol{x}_{M1} & \cdots & \boldsymbol{x}_{MN} \end{bmatrix} \oplus \begin{bmatrix} \boldsymbol{G}_{11} & \cdots & \boldsymbol{G}_{1P} \\ \vdots & \ddots & \vdots \\ \boldsymbol{G}_{M1} & \cdots & \boldsymbol{G}_{MP} \end{bmatrix}$$
$$\oplus\text{:Concatenation operator along the column axis} \tag{7}$$

Where **X** is the size of an eigenvalue representation and **G** is the size of the generalized features. **M** is the total number of perovskites in the dataset (i.e. 27,587 samples); **N** is the total number of eigenvalue features, corresponding to $\boldsymbol{n_{atoms}} \leqslant 20$; and **P** is the total number of generalized input features, which is 21 in the current analysis. By concatenating both **X** and **G** along their column axis, the periodic input features used in predicting all target properties is formulated into $27,\ 587\ (\textit{perovskites}) \times 41\ (\textit{features})$ matrix.

The result of the comparative analysis is reported in table 5 and is performed on five-fold cross-validation. As can be observed from the table, representing crystal structures based on the developed feature extraction process produces better results when compared to their periodic eigenvalue counterparts. A comparative display of the mean and standard deviation from the cross-validation result is illustrated in figure 8. On the formation energy for example, MAE scores are estimated at 0.247 eV/atom, 0.232 eV/atom and 0.245 eV/atom for Coulomb, Ewald-sum and Sine periodic forms, respectively. Moreover, in a similar study by F Faber *et al* [21], MAE generalization error values on the formation energy were reported at 0.64 eV/atom, 0.49 eV/atom and 0.37 eV/atom for Coulomb-like matrix, Ewald-sum matrix and Sine function matrix, respectively. The better performance achieved in the current study is primarily due to the concatenation with the generalized features. This highlights the considerable capability of the used Fourier-transformed reciprocal features in periodically describing crystal structures, in addition to the novel feature extraction approach. The computations used to determine the eigenvalue periodic matrices were enabled by *DScribe* [64], which is an open-source library of descriptors for machine learning in materials science. All calculations related to the atomic coordinates were done in Atomic Units. The source codes for reproducing the calculations, in addition to a breakdown of the generalized features, are made openly available in supplementary (see sections **S4 and S5**).

3.5. Pixel-importance based on *hkl* crystallographic planes in the Fourier-transformed space

In this section, the individual crystallographic planes used in the Fourier-transformation and their relative importance in predicting all considered target properties are further analysed, allowing to investigate the crucial role some crystallographic orientations may play with respect to perovskite formability and electrical behavior. For this purpose, specific Miller indices (***hkl*** planes) are assessed with respect to their ability to provide correlations on the stability energy, formation energy and bandgap. Figure 9 shows the pixel-importance inspection on the projected crystal planes for the different targets as conducted on a gradient boosting machine. Hot pixel intensities denote the superiority of that feature scalar in mapping target properties. As expected, all images are similar in terms of their pixel contrasts regardless of target property. This is because all investigated

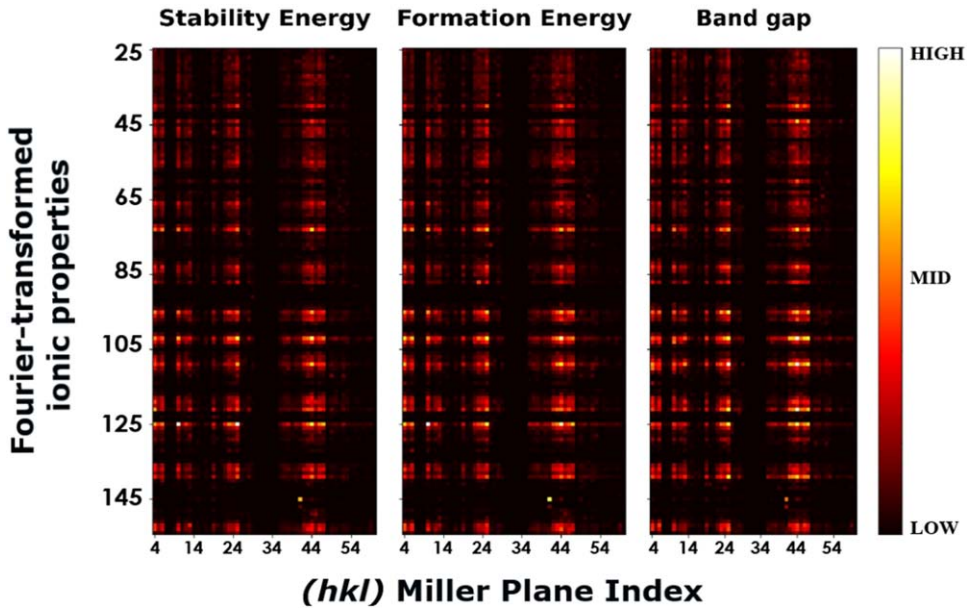

Figure 9. Pixel-importance on the Fourier-transformed ionic properties in the reciprocal crystal space, as it relates to the prediction of target properties. The *y*-axis (image-height) describes the transformed ionic features, whereas the *x*-axis (image width) shows the specific Miller indices, in the same pattern as the high-dimensional image (illustrated in figure 3). Hot intensities on the image suggests the high relevance of that scalar property in predicting a target.

Table 6. Miller indices of the most important crystal planes corresponding to identified pixel-importance clusters given in figure 9. For the full ordered list of all 57 crystal planes used in the Fourier-transformation, see table **S1** in supplementary.

Image width	Miller indices	Image width	Miller indices
4	(100)	**24**	$(1\bar{1}1)$
5	(010)	**25**	$(11\bar{1})$
6	(001)	**42**	$(2\bar{1}0)$
10	(101)	**43**	$(20\bar{1})$
12	(011)	**44**	$(\bar{1}02)$
13	$(\bar{1}01)$	**45**	$(\bar{1}20)$
22	(111)	**46**	$(02\bar{1})$
23	$(\bar{1}11)$	**47**	$(0\bar{1}2)$

properties are physically inter-related. For instance, estimating the stability energy via thermodynamic calculations has been shown to be possible by constructing a convex hull around the formation energy, or by simply relating the convex hull distance to the formation energy [26, 67]. Moreover, it has been reported that using the formation energy as an additional parameter for ML training will positively influence bandgap prediction [16, 17]. As a result, four important clusters can be identified where specific ***hkl*** planes are of most significance. The clustered image widths are identified as ***4–6***, ***10–13***, **22–25**, **&** ***42–47***, corresponding to the *X*-axis of the high-dimensional descriptor image. The respective Miller indices of the identified cluster ordering are outlined in table 6. Overall, the pixel-importance examination highlights 16 planes out of the initial 57 with good correlation to the targets. The identified crystal planes can be presumed as the general regions in the high-dimensional descriptor image where the Fourier-transformation of discretized ionic features is of crucial importance.

In retrospect, the rationale behind the identified planes emerges from the structure-factor calculations. As previously discussed, the assumed ansatz for ionic feature projection is analogous to the periodic scattering potential, which can be approximated as the atomic form-factor. Based on this analogy, the defining planes identified in this study may considerably share some similar characteristics to physical experiments. For example, the amplified intensities in XRD plots indicate which planes are more coherent with Bragg's law, and inform on the type of crystal system in x-ray diffraction experiments [38]. To bridge the gap between theory and experiment, it is important to further investigate the identified planes with respect to potential valuable practical

Table 7. Significance of some identified crystal planes with overall good mapping qualities to the considered targets, as related to ABX_3 formability, growth, stability, failure mechanism and functionality.

Miller Indices	Description
(100)	1. Preferred orientation for epitaxial stable growth of $(Ba,Sr)TiO_3$ [68].
	2. Preferred orientation for synthesizing $Pb(Zr,Ti)O_3$ thin films with good piezoelectric qualities [2].
	3. Plane orientation facilitates electron transfer in organic-inorganic lead iodide perovskite for solar cell applications [69].
	4. Plane orientation is susceptible to plastic deformation in polycrystalline $MgGeO_3$ perovskites, due to compressive stresses at lower mantle pressure [70].
(010)	1. Plane orientation aligns perpendicularly to compressive stresses in $CaIrO_3$ perovskites upon varying shortening cycle and thermal conditions [71].
	2. Plane orientation constrains polarization vector at high misfit strains for ferroelectric $BiFeO_3$ [4].
	3. Plane exhibits strong exchange interaction with neutron scattering techniques for antiferromagnetic $KCuF_3$ [5].
(001)	1. Preferred growth orientation for Perovskite Single Crystals (PSCs) exhibiting ultrahigh photoresponsivity and detectivity [8].
	2. Plane exhibits high photoresponsivity qualities compared to other planes for $CsPbBr_3$ isotropic single crystals in optoelectronic applications [9].
(101)	3. Plane orientation exhibits major twin-defect failure in the formation of pure and doped $YAlO_3$ single crystals by Czochralski growth [72].
	4. Preferred growth orientation in the sol-gel fabrication of $PbTiO_3$ perovskite along gel fiber axis [73].
(011)	1. Considerable increase in Zr-O covalency near the CaZrO3 (**011**) plane, which leads to different surface energies [74].
($\bar{1}$01)	1. A-type antiferromagnetic ordering along ferromagnetic plane-like direction in triclinic $TlMnO_3$ [6].
(111)	1. Single crystal $Pb(Zr,Ti)O_3$ thin-films annealed at 606°C possess maximum polarizability with excellent ferroelectric properties at the specific plane orientation [7].
	2. Orientation develops microstructural unstable holes in epitaxially grown ultrathin $PbTiO_3$ single-crystals [75].
	3. $Pb(Yb,Nb,Ti)O_3$ thin films grown at specific plane orientation exhibit extrinsic contribution to the dielectric behavior of such perovskite [76].
	4. Higher polarization characteristics was obtained on (**111**) oriented planes for $BiFeO_3$ (BFO) pure perovskite films [77].

connections in the current ABX_3 case study. The results are summarized in table 7 with focus on significance in perovskite growth/formation, stability, failure mechanism and functionality. For example, the growth of epitaxially stable $(Ba,Sr)TiO_3$ and $Pb(Zr,Ti)O_3$ thin-film perovskites proves to preferably take place in the (100) plane [2, 68]. Moreover, the (001) plane exhibits high photoresponsivity qualities compared to other planes for $CsPbBr_3$ isotropic single crystals in optoelectronic applications [9]. As outlined in table 5, the pixel-importance examination also identifies both (100) and (001) planes as determinant. Therefore, the current study suggests the newly identified planes as promising research focus for future experimental studies. This is specifically the case for crystal planes of the form $|\boldsymbol{h}| + |\boldsymbol{k}| + |\boldsymbol{l}| = 3$ (e.g. $(0\bar{1}2)$ image width: 47) that remain so far only scarcely reported in existing literature across multidisciplinary perovskite or inorganic solid-state research.

3.6. Experimental impact of modeling approach and future study

The results presented in study clearly demonstrates the appreciable modeling effect of the proposed Fourier-transformed feature engineering approach used in predicting deterministic perovskite target properties. Accurately estimating such targets can be invaluable to material scientists due to the role they play in defining the formability and functionalization. Customarily, determining these properties requires sophisticated first principle calculations or experimental analysis that are computationally laborious and expensive. In contrast, the proposed Conv2D-SVM model offers a reliable and cost-effective alternative that can be used to determine these properties. Moreover, the study highlights the considerable target-modeling effect of using periodically developed descriptors that are based on the reciprocal lattice space of a crystal. Fourier projecting direct features onto the reciprocal space (Brillouin zones) ensures that analogous features used in first-principle quantum mechanical simulations for obtaining crucial ground state properties are likewise introduced in the present descriptor design. For instance, DFT simulates target properties of many-body systems by approximating a solution to the Schrödinger equation in a self-consistent cycle. The basis sets used to store DFT Hamiltonian charge densities are traditionally resolved using plane waves in the Brillouin point mesh of the reciprocal lattice [78]. As such, priori-supplying surrogate ML models with similar DFT inputs can effectively assist in linking first-principle theories with supervisory machine learning. Comparing the presented results to other contemporary descriptor design (table 4) further demonstrates the impact of the present periodic descriptor design.

Finally, the developed descriptor design could potentially serve in applications related to inverse design simulations for accelerating the discovery of unknown perovskites. Considering the addition of labelled atomic numbers into the Direct Space Feature (DSF) block of the input image, novel perovskites with variational

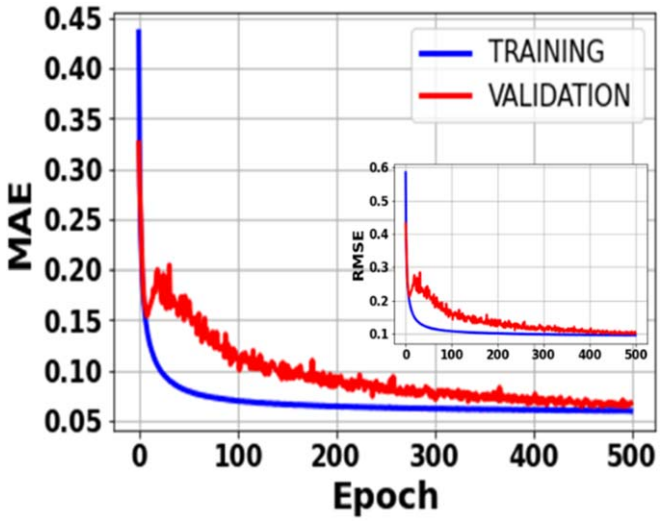

Figure 10. Learning curves on the autoencoding process of feature embedding associated with the high-dimensional input image (RMSE as figure inset).

properties can be identified in the latent space of a generative autoencoder. Moreover, the invertible descriptor design can adapt to discover other forms of perovskite (e.g. hybrid organic-inorganic perovskites), consequential to the reorganization of the DSF block to accommodate such stoichiometrical changes. Using a 2D convolutional autoencoder, figure 10 graphs preliminary learning curves on training and validation sets, as it relates to the reconstruction of all normalized feature embedding associated with the present input image. Upon training for 500 epochs, the decoded perovskite images are demonstrated to be recovered within acceptable error range at 0.066 MAE and 0.103 RMSE, based on standardized evaluation. Table **S5** in supplementary provides a full breakdown of constitutive feature error on the reconstruction process of the high-dimensional input image.

4. Conclusion

The present study demonstrates the effectiveness of the high-dimensional feature engineering approach for higher accuracy in target modeling, with primary focus on the stability energy, formation energy and bandgap of perovskites. The descriptor is inspired by the Fourier transformed Crystal Property (FTCP) representation for invertible material discovery, and is modified to incorporate additional features including: the reciprocal lattice vectors, the magnitude of the reciprocal vector normal to a crystallographic plane, the shortest distance between similar planes, the average ionic radius, the static average electric dipole polarizability, the specific heat, and the thermal conductivity. For target modeling, the descriptor is pre-trained using a two-dimensional convolutional neural network (Conv2D) that is optimized on target properties by back-propagation. The extracted feature is then used as input to a coupled Support Vector Machine (SVM) for further analysis. The following main conclusions can be drawn:

(1) The greatest regressive accuracy improvement is achieved for bandgap prediction, with scores of 0.105 eV MAE, 0.301 eV RMSE, and 93.48% R^2, which represent substantial improvement when compared to existing benchmark evaluations.

(2) The best accuracy scores for the stability energy are 0.050 eV/atom MAE, 0.116 eV/atom RMSE, and 88.60% R^2, and for the formation energy 0.058 eV/atom MAE, 0.133 eV/atom RMSE, and 98.52% R^2.

(3) Compared to the most precise eigenvalue representation obtained from three commonly used periodic descriptors (i.e. Coulomb, Ewald-sum, and Sine matrices in their eigenvalue representations), the proposed Conv2D-SVM model yields about 70%, 75%, and 66% greater MAE accuracy on the stability energy, formation energy, and bandgap, respectively.

(4) Sixteen crystallographic planes are demonstrated to best correlate with the targets, and therefore, to be potentially key for modeling the growth/formation, stability, failure mechanism and functionality of perovskites.

Acknowledgments

This research was supported by the Natural Sciences and Engineering Research Council of Canada [NSERC Discovery Grant number: 210487-180599-2001]; and the National Research Council of Canada (NRC) through its Artificial Intelligence for Design Program led by the Digital Technologies Research Centre. The authors wish to acknowledge the anonymous reviewers whose comments helped improve the quality of the paper.

Data availability statement

The data that support the findings of this study are openly available at the following URL/DOI: https://github.com/chenebuah/high_dim_descriptor.

Authorship contribution statement

ETC: Conceptualization, Methodology, Software, Writing - original draft, Data curation, Investigation. MN: Writing - review & editing, Supervision, Resources. ABT: Writing -review & editing, Validation, Supervision, Resources.

Additional information statement

Supplementary information accompanies this paper.

Ethical statement

The authors declare no competing interest.

ORCID iDs

Ericsson Tetteh Chenebuah https://orcid.org/0000-0003-2878-3484
Michel Nganbe https://orcid.org/0000-0002-2240-9099

References

[1] Choi S *et al* 2017 Correlation of fe-based superconductivity and electron-phonon coupling in an FeAs/oxide heterostructure *Phys. Rev. Lett.* **119** 107003
[2] Du X-H, Belegundu U and Uchino K 1997 Crystal orientation dependence of piezoelectric properties in lead zirconate titanate: theoretical expectation for thin films *Jpn. J. Appl. Phys.* **36** 5580–7
[3] Huang C, Cai K, Wang Y, Bai Y and Guo D 2018 Revealing the real high temperature performance and depolarization characteristics of piezoelectric ceramics by combined *in situ* techniques *J. Mater. Chem.* **6** 1433–44
[4] Chen Z *et al* 2012 Study of strain effect on in-plane polarization in epitaxial $BiFeO_3$ thin films using planar electrodes *Phys. Rev.* **86** 235125
[5] Hutchings M T, Milne J M and Ikeda H 1979 Spin wave energy dispersion in $KCuF_3$: a nearly one-dimensional spin-1/2 antiferromagnet *J. Phys. C: Solid State Phys.* **12** 739–44
[6] Khalyavin D D, Manuel P, Yi W and Belik A A 2016 Spin and orbital ordering in $TlMnO_3$: neutron diffraction study *Phys. Rev.* **94** 134412
[7] Li X-S, Yamashita K, Tanaka T, Suzuki Y and Okuyama M 2000 Structural and electrical properties of highly oriented $Pb(Zr,Ti)O_3$ thin films deposited by facing target sputtering *Sens. Actuator A Phys.*, **82** 265–9
[8] Liu Y *et al* 2019 Surface-tension-controlled crystallization for high-quality 2D perovskite single crystals for ultrahigh photodetection *Matter* **1** 465–80
[9] Zhang P, Zhang G, Liu L, Ju D, Zhang L, Cheng K and Tao X 2018 Anisotropic optoelectronic properties of melt-grown bulk CsPbBr3 single crystal *J. Phys. Chem. Lett.* **9** 5040–6
[10] Jin S, Tiefel T H, McCormack M, Fastnacht R A, Rameshand R and Chen L H 1994 Thousandfold change in resistivity in magnetoresistive La-Ca-Mn-O films *Science* **264** 413–5
[11] La O' G J, Ahn S J, Crumlin E, Orikasa Y, Biegalski M D, Christen H M and Shao-Horn Y 2010 Catalytic activity enhancement for oxygen reduction on epitaxial perovskite thin films for solid-oxide fuel cells *Angew. Chem. Int. Ed.* **49** 5344–7
[12] Johnsson M and Lemmens P 2007 Crystallography and Chemistry of Perovskites *Handbook of Magnetism and Advanced Magnetic Materials* ed H Kronmüller, S Parkin, M Coey, A Inoue and H Kronmüller (Wiley) (https://doi.org/10.1002/9780470022184.hmm411)
[13] Kirklin S, Saal J E, Meredig B, Thompson A, Doak J W, Aykol M, Rühl S and Wolverton C 2015 The open quantum materials database (OQMD): assessing the accuracy of DFT formation energies *NPJ Comput. Mater.* **1** 15010
[14] Jain A, Hautier G, Moore C J, Ong S P, Fischer C C, Mueller T, Persson K A and Ceder G 2011 A high-throughput infrastructure for density functional theory calculations *Comput. Mater. Sci.* **50** 2295–310

[15] Ward L, Agrawal A, Choudhary A and Wolverton C 2016 A general-purpose machine learning framework for predicting properties of inorganic materials *NPJ Comput. Mater.* **2** 16028
[16] Li C, Hao H, Xu B, Zhao G, Chen L, Zhang S and Liu H 2020 A progressive learning method for predicting the band gap of ABO_3 perovskites using an instrumental variable *J. Mater. Chem.* **8** 3127–36
[17] Chenebuah E T, Nganbe M and Tchagang A B 2021 Comparative analysis of machine learning approaches on the prediction of the electronic properties of perovskites: a case study of ABX_3 and $A_2BB'X_6$ *Mater. Today Commun.* **27** 102462
[18] 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
[19] Van Herck W, Fisher J and Ganeva M 2021 Deep learning for x-ray or neutron scattering under grazing-incidence: extraction of distributions *Mater. Res. Express* **8** 045015
[20] Deng Y, Zeng H, Jiang Y, Chen G, Chen J and Sun L 2018 Ridge regression for predicting elastic moduli and hardness of calcium aluminosilicate glasses *Mater. Res. Express* **5** 035205
[21] Faber F, Lindmaa A, von Lilienfeld O A and Armiento R 2015 Crystal structure representations for machine learning models of formation energies *IJQC* **115** 1094–101
[22] Ren Z *et al* 2022 An invertible crystallographic representation for general inverse design of inorganic crystals with targeted properties *Matter* **5** 314–35
[23] Noh J, Kim J, Stein H S, Sanchez-Lengeling B, Gregoire J M, Aspuru-Guzik A and Jung Y 2019 Inverse design of solid-state materials via a continuous representation *Matter* **1** 1370–84
[24] Kim E *et al* 2020 Inorganic materials synthesis planning with literature-trained neural networks *J. Chem. Inf. Model.* **60** 1194–201
[25] Zhou G, Chu W and Prezhdo O V 2020 Structural deformation controls charge losses in $MAPbI_3$: unsupervised machine learning of nonadiabatic molecular dynamics *ACS Energy Lett.* **5** 1930–8
[26] Emery A and Wolverton C 2017 High-throughput DFT calculations of formation energy, stability and oxygen vacancy formation energy of ABO_3 perovskites *Sci. Data* **4** 170153
[27] Hu Z, Lin Z, Su J, Zhang J, Chang J and Hao Y 2019 A review on energy band-gap engineering for perovskite photovoltaics *Sol. RRL* **3** 1900304
[28] O'Shea K and Nash R 2015 An introduction to convolutional neural networksarXiv.1511.08458 [cs.NE]
[29] Drucker H, Burges C J C, Kaufman L, Smola A and Vapnik V 1996 Support vector regression machines *NIPS'96* (Denver: MIT Press) 155–61
[30] Vapnik V N 1998 *Statistical Learning Theory* (New York: Wiley)
[31] Rupp M, Tkatchenko A, Müller K-R and von Lilienfeld O A 2012 Fast and accurate modeling of molecular atomization energies with machine learning *Phys. Rev. Lett.* **108** 058301
[32] Ramakrishnan R, Hartmann M, Tapavicza E and von Lilienfeld O A 2015 Electronic spectra from TDDFT and machine learning in chemical space *J. Chem. Phys.* **143** 084111
[33] Ewald P P 1921 Die Berechnung optischer und elektrostatischer Gitterpotentiale *Ann. Phys.* **369** 253–87
[34] Toukmaji A Y and Board J A Jr 1996 Ewald summation techniques in perspective: a survey *Comput. Phys. Commun.* **95** 73–92
[35] Saal J E, Kirklin S, Aykol M, Meredig B and Wolverton C 2013 Materials design and discovery with high-throughput density functional theory: the open quantum materials database (OQMD) *JOM* **65** 1501–9
[36] Zheng J, Perry B and Wu Y 2021 Antiperovskite superionic conductors: a critical review *ACS Mater. Au.* **1** 92–106
[37] Bracewell R N 1986 *The Fourier transform and its applications.* 31999 (New York: McGraw-Hill)
[38] Simon S H 2013 *The Oxford Solid State Basics* (Oxford: Oxford University Press)
[39] Gulli A and Pal S 2017 *Deep learning with keras* (Packt Publishing Ltd)
[40] Abadi M *et al* 2016 TensorFlow: a system for large-scale machine learning arXiv.1605.08695 [cs.DC]
[41] Haykin S 1994 *Neural networks: a comprehensive foundation* (Prentice Hall PTR)
[42] Nwankpa C, Ijomah W, Gachagan A and Marshall S 2018 Activation functions: comparison of trends in practice and research for deep learning arXiv.1605.08695 [cs.LG]
[43] Suryanarayana C 1994 Structure and properties of nanocrystalline materials *Bull. Mater. Sci.* **17** 307–46
[44] Lobato I and Dyck D V 2014 An accurate parameterization for scattering factors, electron densities and electrostatic potentials for neutral atoms that obey all physical constraints *Acta Cryst.* **A70** 636–49
[45] Smith G H and Burge R E 1962 The analytical representation of atomic scattering amplitudes for electrons *Acta Cryst.* **15** 182–6
[46] Doyle P A and Turner P S 1968 Relativistic Hartree–Fock x-ray and electron scattering factors *Acta Cryst.* **A24** 390–7
[47] Ghosh S, Saha S, Liu Z *et al* 2016 Origin and quenching of novel ultraviolet and blue emission in $NdGaO_3$: concept of super-hydrogenic dopants *Sci. Rep.* **6** 36352
[48] Luis F, Kuz'min M D, Bartolomé F, Orera V M, Bartolomé J, Artigas M and Rubín J 1998 Magnetic susceptibility of $NdGaO_3$ at low temperatures: a quasi-two-dimensional Ising behavior *Phys. Rev.* **58** 798–804
[49] Lide D 2004 *CRC Handbook of Chemistry and Physics* 85th edn (Taylor & Francis)
[50] Cordero B, Gómez V, Platero-Prats A E, Revés M, Echeverría J, Cremades E, Barragán F and Alvarez S 2008 Covalent radii revisited *Dalton Trans.* **21** 2832–8
[51] Ong S P, Richards W D, Jain A, Hautier G, Kocher M, Cholia S, Gunter D, Chevrier V L, Persson K A and Ceder G 2013 Python materials genomics (pymatgen): a robust, open-source python library for materials analysis *Comput. Mater. Sci.* **68** 314–9
[52] Ho C Y, Powell R W and Liley P E 1972 Thermal conductivity of the elements *J. Phys. Chem. Ref. Data* **1** 279–421
[53] Perdew J P 1985 Density functional theory and the band gap problem *Int. J. Quantum Chem.* **28** 497–523
[54] Friedman J H 2001 Greedy function approximation: a gradient boosting machine *Ann. Statist.* **29** 1189–232
[55] Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, Ye Q and Liu T-Y 2017 LightGBM: a highly efficient gradient boosting decision tree *Adv. Neural Inf. Process. Syst* **30** 3146–54
[56] Breiman L 2001 Random Forests *Mach. Learn.* **45** 5–32
[57] Chen T and Guestrin C 2016 XGBoost: a scalable tree boosting system arXiv.1603.02754 [cs.LG]
[58] Ahlawat S and Choudhary A 2020 Hybrid CNN-SVM classifier for handwritten digit recognition *Procedia Comput. Sci.* **167** 2554–60
[59] Agarap A F 2019 An Architecture Combining Convolutional Neural Network (CNN) and Support Vector Machine (SVM) for Image Classification arXiv.1712.03541 [cs.CV]
[60] Min K and Cho E 2020 Accelerated discovery of potential ferroelectric perovskite via active learning *J. Mater. Chem. C* **8** (23) 7866–72
[61] Cherukara M J and Mannodi-Kanakkithodi A 2022 Deep learning the properties of inorganic perovskites *Modelling Simul. Mater. Sci. Eng.* **30** 034005

[62] Guo Z and Lin B 2021 Machine learning stability and band gap of lead-free halide double perovskite materials for perovskite solar cells *Sol. Energy* **228** 689–99
[63] Omprakash P, Manikandan B, Sandeep A, Shrivastava R and Panemangalore V P D B 2021 Graph representational learning for bandgap prediction in varied perovskite crystals *Comput. Mater. Sci.* **196** 110530
[64] Himanen L, Jäger M O J, Morooka E V, Canova F F, Ranawat Y S, Gao D Z, Rinke P and Foster A S 2020 DScribe: library of descriptors for machine learning in materials science *Comput. Phys. Commun.* **247** 106949
[65] Tchagang A B, Tewfik A H and Valdés J J 2020 Molecular Design Using Signal Processing and Machine Learning: Time-Frequency-like Representation and Forward Design arXiv.2004.10091 [physics.chem-ph]
[66] Moussa J E 2012 Comment on fast and accurate modeling of molecular atomization energies with machine learning *Phys. Rev. Lett.* **109** 059801
[67] Ishikawa T and Miyake T 2020 Evolutionary construction of a formation-energy convex hull: Practical scheme and application to a carbon-hydrogen binary system *Phys. Rev. B* **101** 214106
[68] Schwartz R W, Clem P G, Voigt J A, Byhoff E R, Van Stry M, Headley T J and Missert N A 1999 Control of microstructure and orientation in solution-deposited $BaTiO_3$ and $SrTiO_3$ thin films *J. Am. Ceram. Soc.* **82** 2359–67
[69] Yin J, Cortecchia D, Krishna A, Chen S, Mathews N, Grimsdale A C and Soci C 2015 Interfacial charge transfer anisotropy in polycrystalline lead iodide perovskite films *J. Phys. Chem. Lett.* **6** 1396–402
[70] Merkel S, Kubo A, Miyagi L, Speziale S, Duffy T S, Mao H-K and Wenk H-R 2006 Plastic deformation of $MgGeO_3$ post-perovskite at lower mantle pressures *Science* **311** 644–6
[71] Miyagi L, Nishiyama N, Wang Y, Kubo A, West D V, Cava R J, Duffy T S and Wenk H-R 2008 Deformation and texture development in $CaIrO_3$ post-perovskite phase up to 6 GPa and 1300 K *Earth Planet. Sci. Lett.* **268** 515–25
[72] Neuroth G and Wallrafen F 1999 Czochralski growth and characterisation of pure and doped $YAlO_3$ single crystals *J. Cryst. Growth* **198-199** 435–9
[73] Toyoda M, Hamaji Y and Tomono K 1997 Fabrication of $PbTiO_3$ ceramic fibers by Sol-Gel processing *J. Sol-Gel Sci. Technol.* **9** 71–84
[74] Eglitis R I, Kleperis J, Purans J, Popov A I and Jia R 2020 *Ab initio* calculations of $CaZrO_3$ (011) surfaces: systematic trends in polar (011) surface calculations of ABO_3 perovskites *J. Mater. Sci.* **55** 203–17
[75] Seifert A, Vojta A, Speck J S and Lange F F 1996 Microstructural instability in single-crystal thin films *J. Mater. Res.* **11** 1470–82
[76] Gharb N B and Trolier-McKinstry S 2005 Dielectric nonlinearity of $Pb(Yb_{1/2}Nb_{1/2})O_3$-$PbTiO_3$ thin films with {100} and {111} crystallographic orientation *J. Appl. Phys.* **97** 064106
[77] Lee Y-H, Liang C-S and Wu J-M 2005 Crystal growth and characterizations of highly oriented $BiFeO_3$ thin films *Electrochem. Solid-State Lett.* **8** F55
[78] Giannozzi P, Baroni S, Bonini N, Calandra M, Car R, Cavazzoni C, Ceresoli D, Chiarotti G L, Cococcioni M and Dabo I 2009 QUANTUM ESPRESSO: a modular and open-source software project for quantum simulations of materials *J. Phys. Condens. Matter* **21** 395502