Breaking the Aristotype: Featurization of Polyhedral Distortions in Perovskite Crystals

Kazuki Morita, Daniel W. Davies, Keith T. Butler, and Aron Walsh

Published version information

Citation: K Morita et al. Breaking the Aristotype: Featurization of Polyhedral Distortions in Perovskite Crystals. Chem Mater 34, no. 2 (2022): 562-573

DOI: 10.1021/acs.chemmater.1c02959

Breaking the aristotype: featurisation of polyhedral distortions in perovskite crystals

Kazuki Morita,[†] Daniel W. Davies,[‡] Keith T. Butler,[*,¶,§] and Aron Walsh[*,†,‖]

[†]*Department of Materials, Imperial College London, London SW7 2AZ, United Kingdom*

[‡]*Research Computing Service, Information & Communication Technology, Imperial College London, London SW7 2AZ, United Kingdom*

[¶]*SciML, Scientific Computer Division, Rutherford Appleton Laboratory, Harwell OX11 0QX, United Kingdom*

[§]*Department of Chemistry, University of Reading, Reading, RG6 6AD, UK*

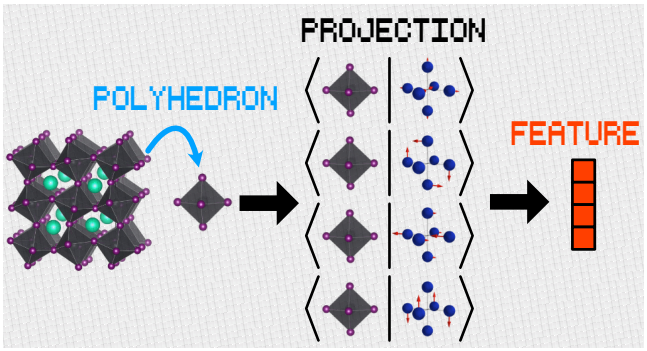
[‖]*Department of Materials Science and Engineering, Yonsei University, Seoul 03722, Korea*

E-mail: keith.butler@stfc.ac.uk; a.walsh@imperial.ac.uk

Abstract

While traditional crystallographic representations of structure play an important role in materials science, they are unsuitable for efficient machine learning. A range of effective numerical descriptors have been developed for molecular and crystal structures. We are interested in a special case, where distortions emerge relative to an ideal high-symmetry parent structure. We demonstrate that irreducible representations form an efficient basis for the featurisation of polyhedral deformations with respect to such an aristotype. Applied to a dataset of 552 octahedra in ABO_3 perovskite-type materials, we use unsupervised machine learning with irreducible representation descriptors to identify four distinct classes of behaviour, associated with predominately corner, edge, face, and mixed connectivity between neighbouring octahedral units. Through this

analysis, we identify $SrCrO_3$ as a material with tuneable multiferroic behaviour. We further show, through supervised machine learning, that thermally activated structural distortions of $CsPbI_3$ are well described by this approach.



Introduction

Materials informatics has grown into a substantial field, supported by the surge in the development of machine learning (ML) techniques.[1–4] Although classical ML and deep neural networks have shown success in fields such as image and natural language processing, their efficiency for material structure inputs are still limited. The problem originates from the difficulty in encoding domain knowledge of material science onto ML training. In other words, the crystallographic information stored in materials datasets is not fully used. To improve this, intense efforts have been made to design efficient material representations to featurise the high structural degrees of freedom into a compact size.[2,5–7]

Unless specially tailored ML models are used,[8–10] a number of criteria exist for crystal features. Firstly, a feature must not depend on the permutation of symmetry equivalent atoms, because atomic indices are only defined for convenience and they have little physical meaning.[10] Secondly, it should not depend on the choice of the unit cell orientation, that is it should not depend on translation or rotation of the axes. Lastly, it must have a suitable size, with the optimal size depending on the problem of interest. If the target properties are complicated, it will require more dimensions to describe it, whereas if the feature is unnecessarily large, more data will be required to train the ML model due to the "curse of dimensionality".[11] Additionally, physical transparency is favourable since it is becoming possible to relate model predictions with the feature(s) responsible.[12]

Material structures would have been easier to represent if we were able to apply a filter to smear atomistic properties in a mean-field manner. Although such coarse-graining has been studied,[13,14] it is often the case that the local structural properties of a material could induce a non-negligible effect on macroscopic properties. For example, in the perovskite structure type, slight displacement of B-site cation could induce both a local electric dipole, as well as macroscopically observable ferroelectric behaviour.[15,16] Another example in a recent study revealed that for the spin-orbit coupling induced Dresselhaus effect, local inversion symmetry, rather than the global crystal symmetry, is responsible.[17] Other interesting phenomena such

as Jahn-Teller distortions, orbital orderings, and magnetic disorders are known, and their coexistence has been reported.[18–20] Given this importance in local structure, many analysis methods have been developed.

There are numerous ways of obtaining a structural feature including Voronoi decomposition, radial distribution functions, nearest neighbours, and electrostatic Ewald summation.[21] Some efforts have been put into the development of calculating coordination numbers. Although coordination number is an intuitive concept, several different approaches have been suggested for a quantitative definition.[22–25] One advanced method is to analyse the connectivity of atoms and use the polygon created by the bonds to categorise the environment.[26,27] Other methods such as Smooth Overlap of Atomic Positions (SOAP), Coulomb matrix, Many-Body Tensor Representations (MBTR), or minimum bounding ellipsoid (MBE) has been suggested, which are based on atomic positions and do not rely on knowledge of the bonding network.[28–31]

Group theory serves as an important tool to interpret the underlying symmetry relations of crystal structures.[32–37] For example, Howard and Stokes exhausted the space groups accessible from cubic perovskites through rigid octahedral tilting.[38] More recently, Wagner et al. analysed density functional theory (DFT) results using a combination of group theory and statistical correlation analysis and showed the efficacy of combining these techniques.[39] Not only in theoretical analyses, but also in experiments, group theoretical techniques are used to determine crystal structures, which is a challenging task in some cases, and therefore being an actively developing research field.[40]

In this paper, we take advantage of established techniques in group theory and use them to encode polyhedron shapes. In particular, we projected the distortions onto the basis vectors of the irreducible representations (irreps) to obtain a physically intuitive decomposition of the distortions. The obtained expression is atomic permutation invariant, axis invariant, minimum length, and physically transparent, meeting all criteria for a suitable material representation for training statistical models. Although our method is applicable to any

type of polyhedron, we chose octahedra inside oxide perovskite-type materials as a model system, as it is well studied.[41–47] We show that our approach when applied to these classes of materials, not only rediscovers intuitively understandable behaviour, but is also capable of capturing trends that originate from subtle differences in an octahedral geometry.

Methodology

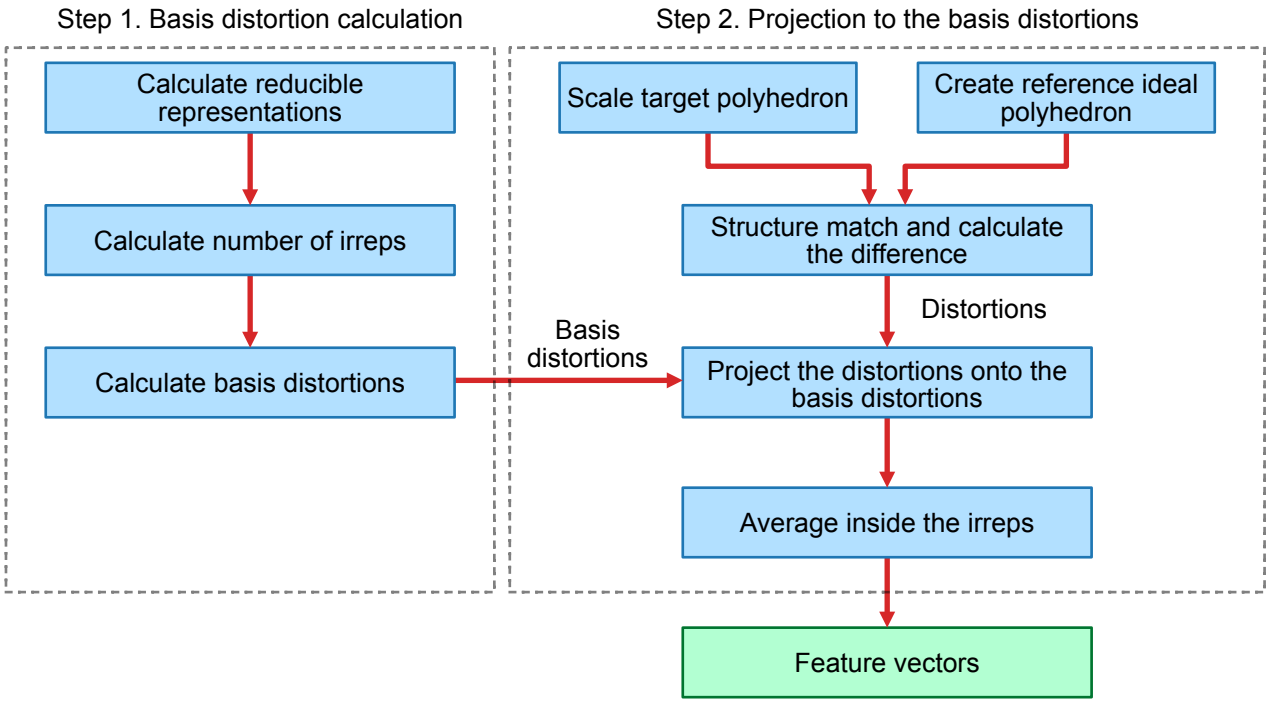


Figure 1: Flowchart of featurisation process. The coordinates of the polyhedron are taken as an input and the feature vectors are returned.

The overview of the featurisation process is schematically presented in Figure 1. The process is largely split into two parts: the basis distortion calculation and the projection of the distortions onto the basis. In the first step, basis distortions corresponding to the irreps are computed using group theoretical techniques. Next, the distortions of the target polyhedra are calculated with respect to the ideal aristotype polyhedron and the distortions are projected onto the basis distortions calculated in the previous step. The explicit isolation

of the first step allows it to be pre-computed, thus minimising the overall computational cost when applied to large datasets.

Basis distortion calculation

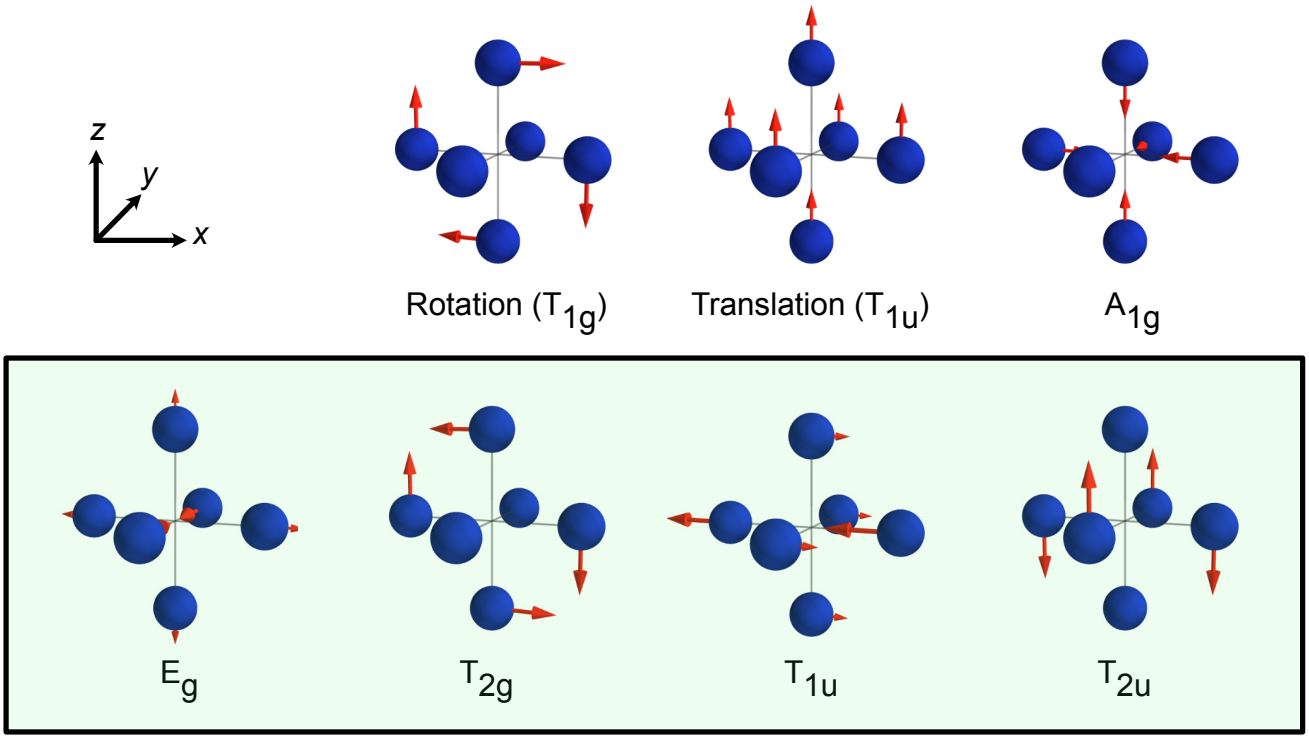


Figure 2: Basis set distortions for the irreducible representations of a six atom octahedron as found in a cubic perovskite. For multi-dimensional irreducible representations, only one distortion is shown. For the actual projection, we have used the four distortions presented in the bottom row. The full list is presented in Figure S1 and S2.

The first goal is to calculate complete and orthogonal basis distortions (basis vectors) of the irreps. The irreps fulfill the "great orthogonality theorem",[48]

$$\sum_R \Gamma^{(i)}(R)_{\mu\nu}\Gamma^{(j)}(R)_{\alpha\beta} = \frac{h}{l_i}\delta_{ij}\delta_{\mu\alpha}\delta_{\nu\beta}. \tag{1}$$

Here, $\Gamma^{(i)}(R)_{\mu\nu}$ is a μ, ν matrix element of operator R in the irrep i, h is number of group elements, and l_i is the dimensionality of $\Gamma^{(i)}$. We cannot directly use this, however, because the specific elements of Γ are unknown *a priori*. Therefore, throughout the section, we make

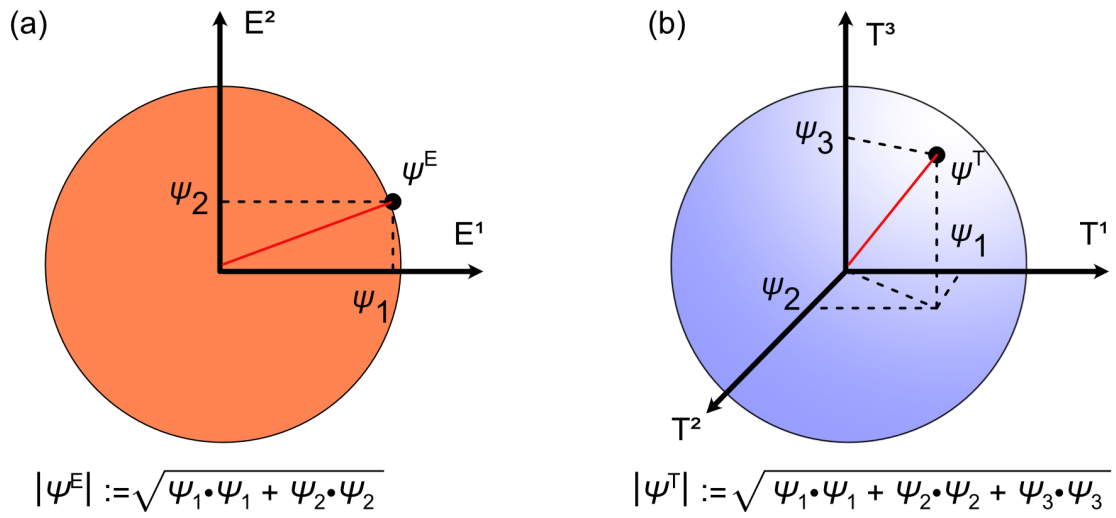


Figure 3: Illustration of how amplitudes are averaged within a multi-dimensional irreducible representation. (a) a two-dimensional irreducible representation (E_g), (b) three-dimensional irreducible representations (T_{2g}, T_{1u} and T_{2u}).

use of their trace or their character, which are readily available from standard character tables. We will use the six-atom octahedron geometry as an example, but our method is applicable to all symmetric coordination environments. The notation follows Ref 48.

Firstly, we need to calculate 18-dimensional reducible representations, which is a direct product between a six-dimensional atomic site and three-dimensional vector representations. The three-dimensional representations $\tilde{\Gamma}^{(3)}(R)$ (tilde indicating a reducible representation) are readily available from previous studies, in which we have adopted them from the *phonopy* package.[49] On the other hand, six-dimensional representations $\tilde{\Gamma}^{(6)}(R)$ depend on specific problems, therefore we have calculated them by applying three-dimensional representation $\tilde{\Gamma}^{(3)}(R)$ to atomic coordinates and keeping track which atoms transformed to which atomic sites. The final 18-dimensional representations $\tilde{\Gamma}^{(18)}(R)$ were constructed by taking a tensor product between three and six-dimensional representation $\tilde{\Gamma}^{(3)}(R) \otimes \tilde{\Gamma}^{(6)}(R)$.

Secondly, we calculate the number of irreps hidden within the 18-dimensional reducible

representation $\tilde{\Gamma}^{(18)}(R)$. To do this, we use the following equation,

$$\sum_R \chi^{(i)}(R)\chi^{(j)}(R) = \frac{h}{l_i}\delta_{ij} \tag{2}$$

Here $\chi^{(i)}$ is a character of irrep $\Gamma^{(i)}$, which is calculated by taking a trace. Although this relation is simply derived by taking the trace of Eqn. 1, it is useful in our case, since it does not require knowledge of specific elements of irreps, while the characters are known (Table S1). Since, $\tilde{\chi}^{(18)}(R) = \sum_i a_i \chi^{(i)}(R)$ where a_i is the number of irrep i in 18-dimensional representation, equation 2 could extract a_i. The calculated result for an octahedron is shown in Table 1. We can see that there are one A (one-dimensional), one E (two-dimensional), and four T's (three-dimensional), which add up to the 18 total degrees of freedom in the system.

Table 1: Number of irreducible representations in 18-dimensional reducible representation in O_h symmetry.

A_{1g}	A_{2g}	E_g	T_{1g}	T_{2g}	A_{1u}	A_{2u}	E_u	T_{1u}	T_{2u}
1	0	1	1	1	0	0	0	2	1

Finally, we calculated the basis vectors. To do so, we have used the "basis-function generating machine",[48] which is defined as

$$\mathscr{P}^{(i)}_{\lambda\kappa} := \frac{l_i}{h}\sum_R \Gamma^{(i)}(R)_{\lambda\kappa} P_R, \tag{3}$$

where P_R is the projection operator of symmetry operator R. The useful property of $\mathscr{P}$ is that when it is operated on an arbitrary function

$$F := \sum_i \sum_\kappa^{l_i} f^{(i)}_\kappa, \tag{4}$$

it could take out $f_\kappa^{(i)}$, the κ-th element within irrep i of the function F

$$\mathscr{P}_{\kappa\kappa}^{(i)} F = f_\kappa^{(i)}. \tag{5}$$

Again, a problem arises due to a lack of knowledge on $\Gamma^{(i)}(R)$. Analogously to the relation between equation 1 and 2, there is a slightly restricted version,[50] which is

$$\mathscr{P}^{(i)} := \frac{l_i}{h} \sum_R \chi^{(i)}(R) P_R \tag{6}$$

$$\mathscr{P}^{(i)} F = \sum_\kappa f_\kappa^{(i)}. \tag{7}$$

The difference is that we could only resolve up to an irrep and components inside an irrep κ remains degenerate. Our approach for deciding the basis set inside multi-dimensional irreps was to generate arbitrary vectors within an irrep and use Gram-Schmidt orthogonalisation to decompose them into orthogonal basis vectors.

Specifically, for each irrep within Table 1, we arbitrarily chose a vector residing on an atom and subsequently applied all the symmetry operators and multiplied the character corresponding to the irrep. The projected results were then added, which resulted in a basis set, as in equation 6. This step was repeated three times with unit vectors in x, y, and z directions. Although the number of trial initial vectors is arbitrary, this choice is the minimum number required to generate all irreps. We then removed duplicates, zero vectors, and further applied Gram-Schmidt orthogonalisation,

$$\psi_\kappa^{(i)} = \psi'^{(i)} - \sum_{\lambda \neq \kappa}^{l_i} (\psi'^{(i)} \cdot \psi_\lambda^{(i)}) \psi_\lambda^{(i)}, \tag{8}$$

where $\psi'^{(i)}$ is an unorthogonalised vector residing in irrep i, and λ runs over other basis set within irrep i that is not κ. Lastly, we have normalised the vectors such that their inner product with themselves equal unity.

Although this method is systematic, one arbitrary choice is the initial vectors for equation 6. In principle, we could use three unit vectors in different directions and still obtain irrep. We will later show that we decided to average over dimensions, and such averaging is necessary even if we have used the full basis set generating machine in Eqn. 3. Following this procedure produces a complete and orthogonal basis set for the irreps which describe all the possible displacement of atoms in an octahedron. The representative distortions are presented in Figure 2 (full list in Figure S1 and S2).

Projection to the basis distortions

The projection of an arbitrary structure on this basis set was performed in three steps: normalisation, structure matching, and distortion amplitude averaging.

If we simply project two distorted octahedra with the same shape but different sizes, we will obtain different distortion amplitudes. This is not favourable in the context of analysing the shape of the octahedra. Therefore, some kind of normalisation of the input octahedron is necessary. Our approach was to scale the distorted octahedron such that the average bond length is 1.0 Å and obtain the distortion vector by comparing it against the ideal octahedron with a bonding length of 1.0 Å. By applying this scaling, the resulting distortion amplitudes for octahedra of the same shape, but different sizes became identical.

Although our method is permutation invariant, practically, we have to label atoms within the code. Therefore, to calculate the distortions the atomic indices of the distorted and the ideal octahedron must be matched. This structure matching requires $\mathcal{O}(N!)$ computational cost, if calculated rigorously by brute-force algorithm, but we found that this is too slow for high-throughput applications. To make the computational cost feasible, we employed the Hungarian algorithm, as implemented in the *pymatgen* package.[51,52] We confirmed that this algorithm works well in perovskites and perovskite-related materials, which typically have well-defined octahedra, however, for geometry with large variation in bonding length, brute force algorithms are likely to be favoured. After matching the structure, the distortion vectors

were calculated and were projected onto basis vectors presented in Figure 2. Furthermore, we have validated the quality of this basis by reconstructing the original distortion from the projection and confirmed that the error is negligible (Figure S3).

It is tempting to use the amplitudes we have obtained above directly, however, the raw values encompass the aforementioned arbitrariness within the multi-dimensional irreps, which originates from the usage of equation 6 rather than equation 3. Taking a closer look, the choice of basis vectors within a single irrep follows a rotational group or special orthogonal group. Since, the actual configuration of an input octahedron inside a crystal may be rotated in any possible direction, even if we have used the full basis set generating machine (Eqn. 3), the resulting amplitudes of the basis vectors would have had a dependence on the choice of the axis. For example, if the T_{1u} distortion in Figure 2 is rotated 90° about the x axis, the amplitude obtained by projection onto the original T_{1u} distortion and the transformed T_{1u} will be different. This situation is encountered in all the distortions except for A_{1g}, which has no multiplicity and is thus rotational invariant. Therefore, the arbitrariness due to a dependence of rotation is a problem that exists regardless of whether or not we use equation 6. Since one of the purposes of this analysis is to obtain ML-friendly features, rotational variance is not favourable, especially because for a typical ML model, learning a permutation is a challenging task.[53]

Our approach was to use the total length spanned by vectors within the irreps. As shown in Figure 3, we have calculated the length of the vectors in two- or three-dimensional space using the Euclidean norm,

$$\Phi^{(i)} = \sqrt{\sum_{\kappa} \psi_{\kappa}^{(i)} \cdot \psi_{\kappa}^{(i)}}. \tag{9}$$

Here the summation is over the dimension inside irrep (i). Just like the Euclidean distance of a given point from the origin remains the same under rotations about the origin, this expression is invariant under any orthogonal transform. Another interpretation of this approach is that we are rotating the axis in Figure 3, so that one of the axes is aligned with

the amplitude vector and then reading the value off that axis.

Through the above procedures, we were able to obtain a scalar value for each irrep for any distorted octahedron. Lastly, translation, rotation, and scaling distortions (A_{1g}) were discarded, since they do not have information regarding the shape of the octahedron. We note that it is possible to encode information such as rigid shifting, rigid tilting or octahedron size into these irreps, but it will require modification to the structure matching procedure and are likely to introduce additional complexity in the algorithms. Therefore, we report four scalar values each corresponding to E_g, T_{2g}, T_{1u}, and T_{2u} for rest of the work.

Lastly, it is worth mentioning about extension to other polyhedra. One of the simpler, but yet often encountered geometry in materials science is the tetrahedron. Since a tetrahedron has four vertices, the total degrees of freedom will be 12, of which seven belong to translation, rotation, and scaling distortions. The remaining five are E and T_2 distortions, which are averaged within the irrep and used as a feature. In this case, the feature may be too small to be used to train an ML model solely, but at the same time, this indicates that the variation in the shape of a tetrahedron is residing in much smaller dimensions compared to the case of an octahedron. Another abundant geometry is cuboctahedron, which is seen for example in the A-site of perovskites. The total degree of freedom is 36 and removing translation, rotation, and scaling distortions will retain: A_{2g}, two E_g, T_{1g}, two T_{2g}, A_{2u}, E_u, two T_{1u}, and two T_{1u} distortions. Depending on the application, this dimensionality may be directly used to train ML models, but in practice, if the analysis is restricted to certain classes of materials, we speculate that some of the distortions would not be present and could be removed by using feature selection techniques, such as *k highest score*.[54]

Dataset processing

To apply the projection, we have obtained 46,048 materials from the *Materials Project* database (accessed on 27/06/2020) through the API in the *pymatgen* package.[52,55,56] We then applied the *CrystalNN* algorithm to obtain coordination number and environment for

all the atomic sites. We have selected ABO_3 stoichiometry materials containing six-fold coordinated cations. In theory, there may be six-fold coordinated atomic sites without an octahedron geometry, but we did not observe such a case within our curated dataset. For materials containing multiple symmetry inequivalent octahedra, we have detected them according to their Wyckoff positions with the *spglib* library.[57] The inequivalent sites were treated as independent data, which resulted in 552 distinct octahedra in total. For a given composition, there are multiple structures and we have not explicitly taken into account their thermodynamic stability. Therefore, our analysis contains structures that may not have been synthesised to date, but represent local minima on DFT potential energy landscapes.

Density functional theory calculations

Although we have largely applied the method to openly available from the *Materials Project* database,[52,55,56] for validation we performed some calculations with stricter conditions. The plane-wave DFT calculations within projector-augmented wave scheme were performed using the *VASP*.[58–60] The input file was automatically generated via *VISE* package,[61] resulting in cut-off energy of 520 eV and the reciprocal space sampling of at least $2\pi \times 0.05$ Å^{-1}. Using the structures in the *Materials Project* as an initial input, the cell size and the atomic coordinates were fully relaxed using HSE06 exchange-correlation functional.[62,63] The visualisation of structures was done using *VESTA*.[64]

Predicting potential energy

To demonstrate applicability of our featurisation procedure towards supervised machine learning of materials, we trained an ML model to predict potential energy generated by Born-Oppenheimer molecular dynamics (BOMD) of $CsPbI_3$. The unit cell was expanded into a $2\times 2\times 2$ pseudo-cubic supercell following previous studies.[65,66] Using *VASP*, 300K NVT ensemble calculation was computed with cut-off energy of 400 eV, the reciprocal space sampling of $3 \times 3 \times 3$, the time step of 0.4 fs and with PBEsol exchange-correlation functional.[62,67] The

system was equilibrated for 15,000 steps and a production run was performed for 39,000 steps with sampling taken every 50 steps, which resulted in 780 data points.

For each snapshot, octahedral distortions were featurised, normalised with robust scaler, and shuffled randomly. The preprocessed features were fed into support vector regression (SVR), as implemented in the *scikit-learn* library, and trained to predict the potential energy.[54] For the kernel, the radial basis function (RBF) kernel was employed. Five-fold cross-validation was used to optimise hyperparameters resulting in kernel coefficient γ of 0.1 and regularization parameter (variable C in *scikit-learn*) of 1.0.

Results and discussion

Projection onto normal distortions

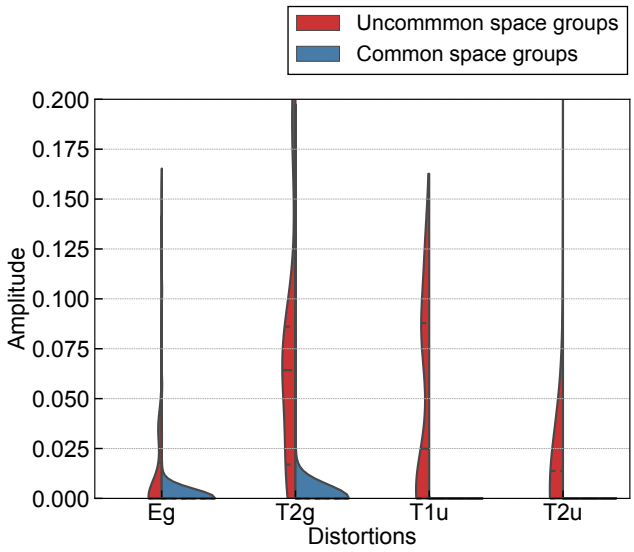


Figure 4: Distortion amplitudes for 552 ABO_3 octahedra from 492 materials. The blue and red shading refers to materials that belong or do not belong to the common space groups for distorted corner-sharing perovskites, respectively.

The distribution of distortion amplitudes for all 552 materials is presented in Figure 4. The materials are categorised by whether or not they belong to the common corner-shared perovskite space group (cubic $Pm\bar{3}m$, tetragonal P4mm, tetragonal P4/mmm, tetrago-

nal P4/mbm, tetragonal I4/mcm, orthorhombic Pnma, orthorhombic Amm2, orthorhombic Cmcm, monoclinic $P2_1/m$, rhombohedral R3m, rhombohedral R3c, and rhombohedral $R\bar{3}c$).[38] The number of materials in common and uncommon space groups were 443 and 109, respectively. From Figure 4, differences in the distributions are clearly noticeable for the two classes of materials. For the common space groups, the vast majority had little or no distortion and the number of materials decay monotonically with increasing amplitudes. In contrast, for less common space groups, the distribution exhibited a wider spread and the larger portion of materials had larger amplitudes. Additional peaks are clearly seen for T_{2g} and T_{1u} around 0.075 and 0.100, respectively. Accounting for the fact that there were no clear chemical trends (Figure S4∼S7), this result suggests a strong relationship between the crystal structure and the local distortions of the octahedra.

Connectivity analysis

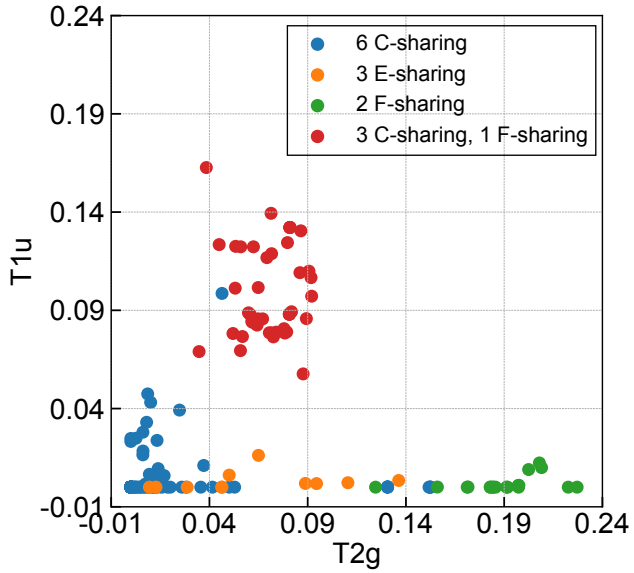


Figure 5: Relation between T_{1u} distortion against the T_{2g} distortion. Each point represents an octahedron site and is coloured according to its connectivity with other octahedra. Blue points are connected with via six corner-sharing (6 C-sharing), orange points are connected with via three edge-sharing (3 E-sharing), red points are connected with via three corner-sharing and one face-sharing (3 C-sharing and 1 F-sharing), and green points are connected with via two face-sharing (2 F-sharing).

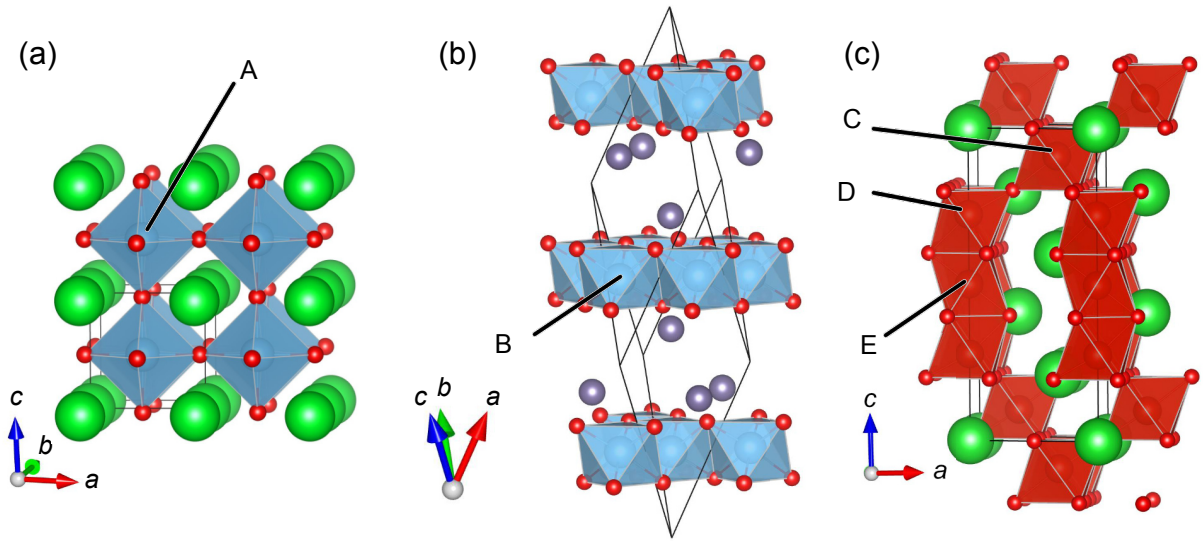


Figure 6: Structures of (a) cubic $SrTiO_3$, (b) rhombohedral $GeTiO_3$, and (c) hexagonal $BaVO_3$. Octahedra are composed of TiO_6, TiO_6, and VO_6, respectively. Distinct Wyckoff positions are labelled by A to E.

To analyse the underlying material trends in more detail, we have plotted the T_{1u} distortion against the T_{2g} distortion and categorised each site according to their connectivity with neighbouring octahedra (Figure 5). The connectivity was obtained by counting how many oxygen atoms are shared with neighbouring octahedra, accounting for periodic boundary conditions. The four connectivities in Figure 5 are: six corner-sharing (6 C-sharing, A and C in Figure 6), three edge-sharing (3 E-sharing, B in Figure 6), two face-sharing (2 F-sharing, E in Figure 6), and three corner-sharing and one face-sharing (3 C-sharing and 1 F-sharing, D in Figure 6).

A cluster of distortion amplitudes are distinguishable about (T_{2g},T_{1u}) = (0.075,0.100). Two interesting observations could be made from this clustering. The first is that T_{1u} distortion amplitude of over 0.05 is only present in this cluster. This suggests that the large amplitude of T_{1u} distortions could only exist when T_{2g} distortions coexist. The behaviour is analogous to improper ferroelectrics where the coexistence of two distortions creates a ferroelectric distortion.[68–70] Secondly, this cluster is composed mostly of three corner-sharing and one face-sharing connectivity. This type of octahedral connectivity is realised in hexagonal perovskite polytypes where a 1D chain of face-sharing octahedra terminates as in Figure 6(c).

Accounting for the fact that three corner-sharing and one face-sharing octahedron were not seen outside of this cluster, this result indicates that hexagonal phases could support distortions much larger than that seen in corner-shared perovskites. The one fully-point-shared outlier in the cluster was $BiFeO_3$, which exhibited an unusually large distortion. The possible origin of the large distortion is the steroechemical activity of the Bi lone pair, as suggested by previous studies.[71–73]

Outside of this cluster, the T_{1u} distortion was generally small. Most of the fully corner-shared octahedra and fully edge-sharing octahedra possess an ideal structure, which made the data points to be scattered around the zero amplitude point. Two face-sharing octahedra interestingly, had a very large T_{2g} distortion but lacked T_{1u}. Since this connectivity occurs in the middle of a 1D chain in hexagonal phases as in site E in Figure 6, the uniaxial strain due to being sandwiched by neighbouring octahedra is likely to have caused the compression of the octahedron.

Clustering analysis

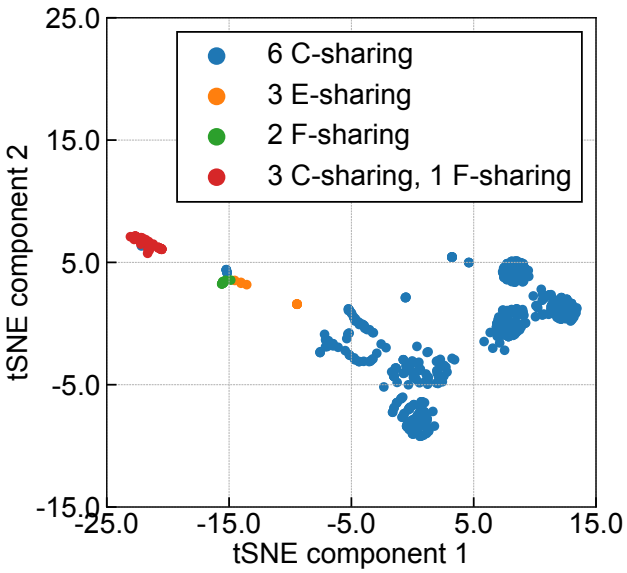


Figure 7: The clustering of different octahedron connectivity plotted on the dimensionally reduced axis was obtained through t-distributed stochastic neighbour embedding (t-SNE).

Up to here, we have made discussions based on the trends in Figure 5, however, such a discussion may be overlooking trends in higher dimensions. Therefore, we performed dimensionality reduction analysis to understand the clustering of different octahedral connectivities in higher dimensions. We employed t-distributed stochastic neighbor embedding (t-SNE) to perform nonlinear reduction from four to two dimensions.[74,75] The result is shown in Figure 7. Different octahedral connectivities are clearly separated. This result is fortuitous since it indicates that the shape of octahedra is largely determined by their connectivities with neighbouring octahedra. In other words, a geometrical network of bonds dominantly determines the shape of the octahedra rather than the chemical property of individual bonds. The two fully corner sharing (6 C-sharing) outliers near the face sharing (2 F-sharing) cluster were $TeCoO_3$ and $TeMnO_3$ (filed as $CoTeO_3$ and $MnTeO_3$ in the *Materials Project*, respectively). The large distortions in these materials are realised by covalent interaction between tellurium and oxygen and due to strong tellurium lone pairs.[76] $BiFeO_3$ seen in Figure 5 also appears again as an outlier within 3 C-sharing and 1 F-sharing cluster.

We next perform a clustering analysis in the full four-dimensional space to see if there is additional information to be obtained. The multi-dimensional clustering was analysed by a Gaussian mixture model (GMM).[75] GMM requires a number of clusters to be set *a priori*, therefore, we calculated the minimum number of clusters needed to account for the data using the information criteria analysis and selected nine clusters to be adequate (Figure S12). The obtained nine clusters are presented in Figure 8 (plot against all axes are shown in Figure S13). It should be noted that in GMM, a data point could only belong to a single cluster. In Figure 8(a), a clear ellipsoid of cluster 0 can be distinguished. This cluster corresponds to the three corner-sharing and one face-sharing in Figure 5 at (T_{2g},T_{1u}) = (0.075,0.100). A closer look reveals that there is a subset of materials within the ellipsoid that belong to cluster 5. Their difference is not distinguishable from Figure 8(a), but plotting against the T_{2u} distortion axis in Figure 8(b) reveals that cluster 5 is displaced from cluster 0 in the T_{2u} distortion axis. Cluster 0 had no T_{2u} distortions, whereas cluster 5 had about

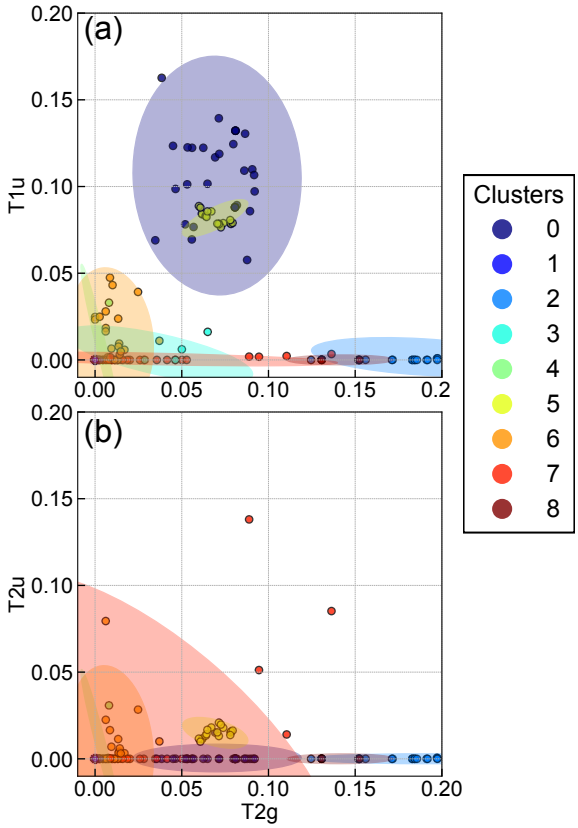


Figure 8: Clusters obtained by a Gaussian mixture model shown in the axis of (a) T_{2g} and T_{1u}, and (b) T_{2g} and T_{2u}. The dots are coloured differently depending on which of the seven different clusters the point belongs to. The shading shows the extent of the multivariate Gaussian distribution defined for each cluster.

0.02 T_{2u} distortion. This separation is not trivial from Figure 5 and highlights the value of clustering analysis in the high dimensional space. We will discuss specific constituent materials of cluster 5 next.

Analysis of specific materials

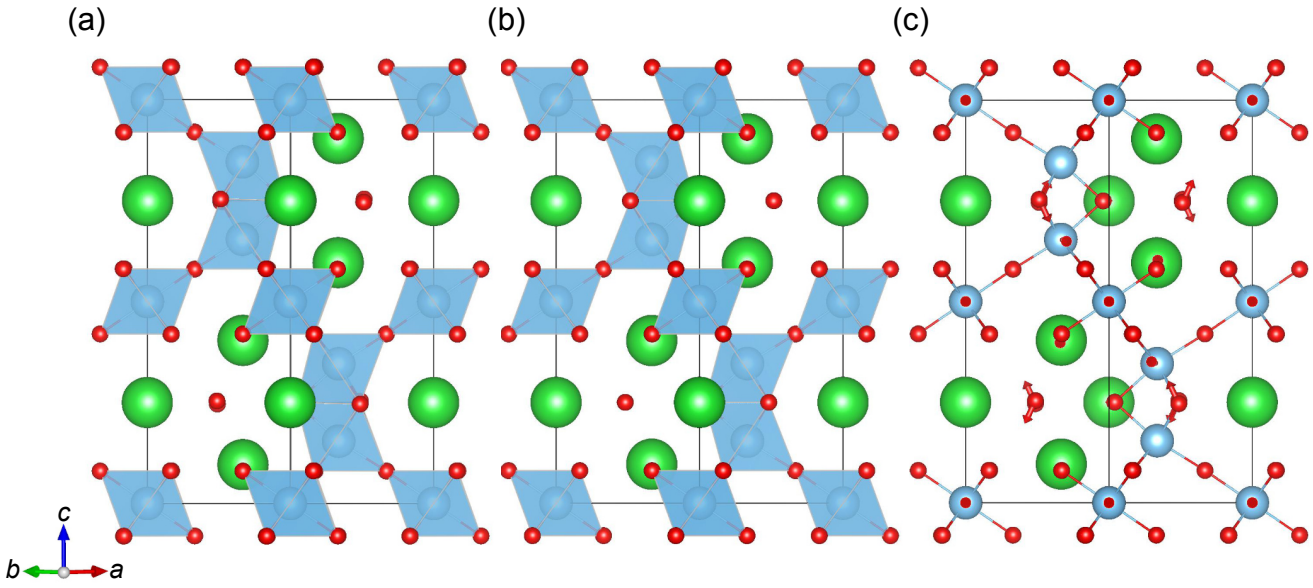


Figure 9: Structure of $BaTiO_3$ in (a) $C222_1$ and (b) $P6_3/mmc$ phase. (c) The atomic displacement to transform from $C222_1$ to $P6_3/mmc$. The blue, green, and red spheres are Ba, Ti, and O atoms respectively. The blue shading represents the TiO_6 octahedron.

Cluster 5 in Figure 8 is mainly composed of $BaTiO_3$ and different polymorphs of $SrCrO_3$. We find that the distortions in $BaTiO_3$ were typical for hexagonal phases. Within our dataset, there were two polymorphs of hexagonal $BaTiO_3$, the $C222_1$ phase and the $P6_3/mmc$ (Figure 9 (a) and (b), respectively). Experimentally, the $C222_1$ is stable in the range of about 70$\sim$220 K, where it transforms into the $P6_3/mmc$ phase at 220K.[77,78] The low temperature $C222_1$ phase has the T_{2u} distortions, but they are averaged out and are absent in the high temperature $P6_3/mmc$ phase. The structural difference between the $C222_2$ and the $P6_3/mmc$ phase is presented in Figure 9(c).

To confirm whether the absence of the T_{2u} distortions in other ABO_3 is due to the lack

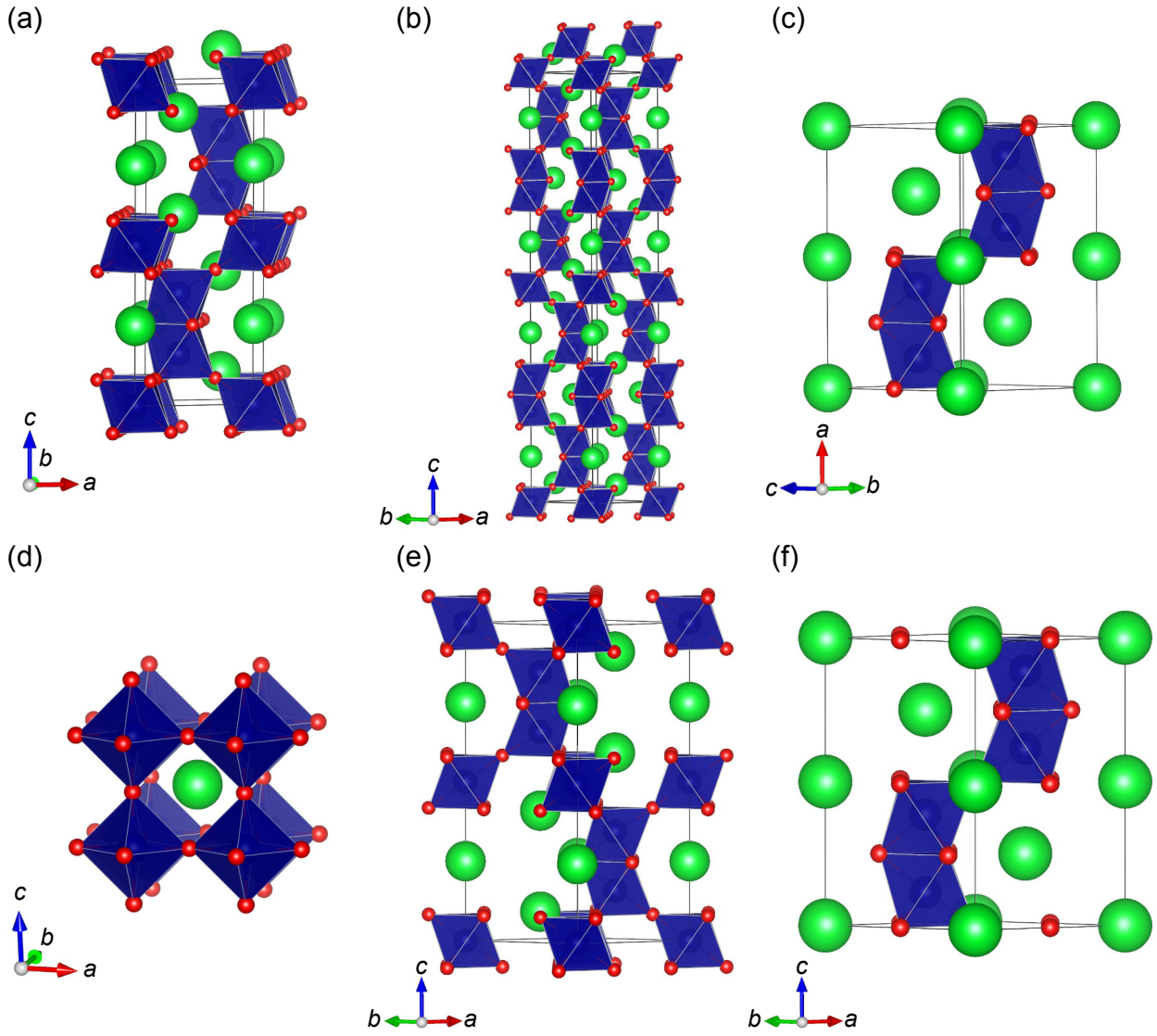


Figure 10: Structures of different $SrCrO_3$ polymorphs. The details are summarised in Table S2.

Table 2: Calculated relative stability (DFT/HSE06) of the low temperature $C222_1$ phase compared to the high temperature $P6_3/mmc$ (see illustration in Fig. 9).

Compound	$E_{P6_3/mmc} - E_{C222_1}$ (meV/atom)
$CaTiO_3$	18.61
$CaCrO_3$	61.56
$CaMnO_3$	30.08
$SrTiO_3$	3.88
$SrCrO_3$	-16.11
$SrMnO_3$	5.18
$BaTiO_3$	0.72
$BaVO_3$	37.31
$BaCrO_3$	-6.29
$BaMnO_3$	10.34
$BaRuO_3$	-1.11
$BaRhO_3$	4.68

of data or due to different phase stability, we have compared the energies of $P6_3/mmc$ and the $C222_1$ phases in 11 additional compounds (Table 2). We found that in most compounds $C222_1$ phase was stable suggesting it to be the lower temperature phase, thus showing that the $BaTiO_3$ with finite T_{2u} is not exceptional, but rather a property of hexagonal phase materials. The exceptions were $SrCrO_3$, $BaCrO_3$, and $BaRuO_3$. The energy difference in $BaCrO_3$, and $BaRuO_3$ were subtle, but $SrCrO_3$ had clearly higher stability of the $P6_3/mmc$ phase.

$SrCrO_3$ is an interesting case that has an interplay of metallicity, ferroelectricity and magnetic order. In cubic $SrCrO_3$, there have been reports on multiferroicity, which are induced by orbital ordering.[20,79,80] Since this material has been suggested to be internally strained,[81] we believe this is the reason for the distinct distortion behaviour of this material. For hexagonal polytypes of $SrCrO_3$ (Figure 10), which have not been reported to the best of our knowledge, we note that the formation energy predicted by DFT is smaller than the known cubic phase (Table S2) which suggests that they should be accessible. Interestingly, within the hexagonal phases, the Ama2 phases (Figure 10(a), (b), and (c)) were calculated to be metallic, whereas $P6_3/mmc$ phases (Figure 10(e) and (f)) were insulators (Table S2). For polymorphs without a band gap, we have confirmed their metallicity with

our DFT/HSE06 calculations. Since, the ratio of corner-shared and face-shared connectivities could be controlled by the stacking sequence, we speculate that through the tuning of the polytype order, metallicity/insulating, ferroelectricity/paraelectricity, and ferromagnetic/paramagnetic behaviour could be accessed. Furthermore, like orbital ordering observed for the cubic phase, coupling of different behaviours are also expected here.

Application to supervised learning

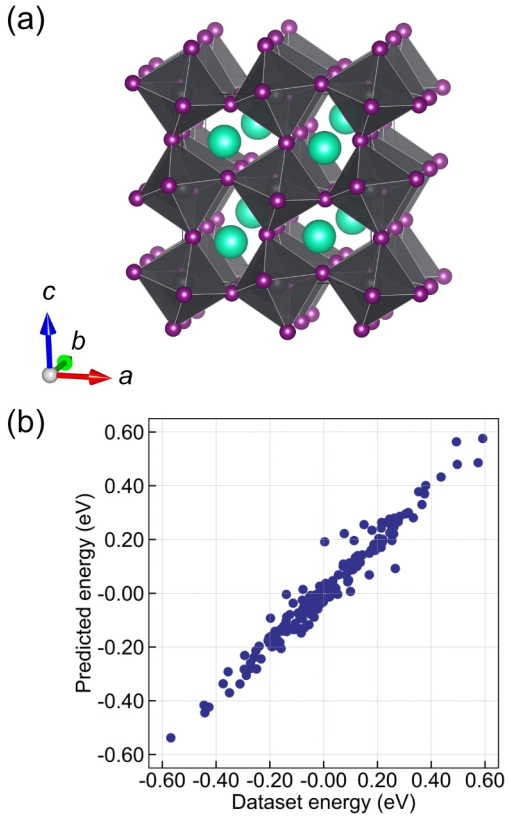


Figure 11: (a) Pseudo-cubic $CsPbI_3$ $2 \times 2 \times 2$ supercell used for the molecular dynamics simulation. The grey octahedra with purple vertices are PbI_6 and the teal spheres are Cs. (b) Comparison between the dataset and the support vector regression prediction of the potential energy test data.

In addition to the material discovery problem demonstrated above, our featurisation method could also be applied to thermodynamic problems. We generated a dataset of potential energy for snapshots obtained from the BOMD calculation of $CsPbI_3$, and used

it to train an SVR model. We have chosen potential energy as it only depends on atomic coordinate and serves as a good benchmark to assess the capability of capturing thermally induced lattice distortions. The featurisation procedure was the same as above, except we also make the use of A_{1g} distortions relative to the 0 K geometry optimised structure. Since there were eight octahedra in the calculation cell as shown in Figure 11(a), the dimensionality of the feature was 40. This is a good example of a case where the scaling distortion could be used to incorporate the domain knowledge of the problem. Figure 11(b) shows the SVR prediction of the test data with the zero of the energy set to be the average potential energy during the BOMD simulation (raw values shown in Figure S15). The r^2 and mean absolute error for the test (training) data was 0.956 (0.989) and 31.2 meV (18.2 meV), respectively. The high accuracy is notable considering the model contains no explicit information regarding the atomic positions of Cs and Pb. We believe that since Pb and Cs are heavier, the displacement is smaller and thus the change in the inter-atomic distances could dominantly be taken into account by considering I. Since this is finite temperature calculation, given a configuration of I, there are numerous possible positions of Cs and Pb, so such a degree of freedom is likely to be a reason for some outliers seen in Figure 11(b). We therefore expect increasing error with temperature. This result shows the efficiency of the featurisation method towards supervised learning.

Discussion on other applications

Comparing our method to other types of local structural featurisation, conventional methods that incorporate basis set expansion of the local environment will capture a wider variety of environments and may be better suited for training general purpose machine-learned force-fields that could describe solid-liquid transitions for example. In contrast to their generality, these types of expansions typically have orders of magnitude larger feature size, $\sim$1000 for SOAP,[82] which require significant data and training time. Furthermore, in such high dimensional methods, the features are not guaranteed to correspond directly with the

displacement of atoms and thus may obstruct analyses based on conventional symmetry arguments.[82] Our approach will have an advantage in encoding polyhedral distortions in cases where the dataset size is not large enough to train general featurisation techniques. A similar discussion holds when compared with graph neural networks, where our method has an advantage in smaller datasets.[83]

Like other approaches, our method is best suited for encoding local properties, but there are often cases where one wants to treat global properties. If the number of polyhedron sites is fixed throughout the dataset, the features could be used directly, like in the BOMD analysis above. A problem occurs when a dataset includes a variable number of polyhedron sites, and in cases where the analysis method only accepts fixed-size input. The simplest solution is to use sum or average pooling. The choice between these two could be made by whether or not the property of the interest is intensive or extensive nature.[8] The use of recurrent neural network based methods such as set2set could further improve the performance.[10]

Conclusion

We have shown that using a group theoric approach, distortions in polyhedra can be encoded into a small vector. As a case study, we have shown their efficacy towards representing the structures of ABO_3 stoichiometry oxides. In addition to recovering intuitively understandable trends, we presented the close relations between octahedra connectivity and their distortions, which are likely to be smeared out by some of the conventional analyses. As a co-product, we were able to find $SrCrO_3$, which contained a rich variety of ferroic behaviours. We further showed that it is capable of predicting the potential energy of $CsPbI_3$ accurately with supervised machine learning. All of these analyses were performed solely on the information of the structures and additional information such as thermodynamic stability and electronic structure will likely elucidate additional trends. We emphasise that this method is not exclusive and synergistic effects are expected when combined with other featurisation

techniques. Finally, the results of this study are based on simple dimensional analysis with the potential for further improvements using more sophisticated non-linear approaches such as deep neural networks. We expect that these developments will open a path to more accurate ML models and support further materials discoveries.

Acknowledgement

We thank funding support from Yoshida Scholarship Foundation, Japan Student Services Organization, and Centre for Doctoral Training on Theory and Simulation of Materials at Imperial College London funded by the EPSRC (EP/L015579/1). Via our membership of the UK's HEC Materials Chemistry Consortium, which is funded by EPSRC (EP/R029431), this work used the ARCHER2 UK National Supercomputing Service (http://www.archer2.ac.uk).

Supporting Information Available

Supporting Information Available: Detailed results on distortion analyses and density functional theory calculations omitted in the main text. This material is available free of charge via the Internet at http://pubs.acs.org.

The code to perform the polyhedron analysis proposed in this study is freely available from `https://github.com/KazMorita/polyhedron_distortion` (latest version) or `https://doi.org/10.5281/zenodo.5255356` (archived version).

References

(1) Alom, M. Z.; Taha, T. M.; Yakopcic, C.; Westberg, S.; Sidike, P.; Nasrin, M. S.; Esesn, B. C. V.; Awwal, A. A. S.; Asari, V. K. The History Began from Alexnet: A Comprehensive Survey on Deep Learning Approaches. 2018, arXiv:1803.01164. arXiv.org e-Print archive. `https://arxiv.org/abs/1803.01164`, (accessed December 14, 2021).

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

(3) de Pablo, J. J. et al. New Frontiers for the Materials Genome Initiative. *npj Comput. Mater.* **2019**, *5*, 41.

(4) Horton, M. K.; Dwaraknath, S.; Persson, K. A. Promises and Perils of Computational Materials Databases. *Nat. Comp. Sci.* **2021**, *1*, 3–5.

(5) Saal, J. E.; Oliynyk, A. O.; Meredig, B. Machine Learning in Materials Discovery: Confirmed Predictions and Their Underlying Approaches. *Annu. Rev. Mater. Res.* **2020**, *50*, 49–69.

(6) Musil, F.; Grisafi, A.; Bartók, A. P.; Ortner, C.; Csányi, G.; Ceriotti, M. Physics-Inspired Structural Representations for Molecules and Materials. *Chem. Rev.* **2021**, *121*, 9759–9815.

(7) George, J.; Hautier, G. Chemist versus Machine: Traditional Knowledge Versus Machine Learning Techniques. *Trends Chem.* **2021**, *3*, 86–95.

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

(9) Schütt, K. T.; Sauceda, H. E.; Kindermans, P.-J.; Tkatchenko, A.; Müller, K.-R. Schnet – A Deep Learning Architecture for Molecules and Materials. *J. Chem. Phys.* **2018**, *148*, 241722.

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

(11) Goodfellow, I.; Bengio, Y.; Courville, A. *Deep Learning*; MIT press, 2016.

(12) Morita, K.; Davies, D. W.; Butler, K. T.; Walsh, A. Modeling the Dielectric Constants of Crystals Using Machine Learning. *J. Chem. Phys.* **2020**, *153*, 024503.

(13) Davies, D.; Butler, K.; Jackson, A.; Skelton, J.; Morita, K.; Walsh, A. Smact: Semiconducting Materials by Analogy and Chemical Theory. *JOSS* **2019**, *4*, 1361.

(14) Goodall, R. E. A.; Parackal, A. S.; Faber, F. A.; Armiento, R.; Lee, A. A. Rapid Discovery of Novel Materials by Coordinate-free Coarse Graining. 2021, arXiv:2106.11132. arXiv.org e-Print archive. `https://arxiv.org/abs/2106.11132`, (accessed December 14, 2021).

(15) Martin, L. W.; Rappe, A. M. Thin-Film Ferroelectric Materials and Their Applications. *Nat. Rev. Mater.* **2016**, *2*, 16087.

(16) Smith, M. B.; Page, K.; Siegrist, T.; Redmond, P. L.; Walter, E. C.; Seshadri, R.; Brus, L. E.; Steigerwald, M. L. Crystal Structure and the Paraelectric-to-Ferroelectric Phase Transition of Nanoscale $BaTiO_3$. *J. Amer. Chem. Soc.* **2008**, *130*, 6955–6963.

(17) Zhang, X.; Liu, Q.; Luo, J.-W.; Freeman, A. J.; Zunger, A. Hidden Spin Polarization in Inversion-Symmetric Bulk Crystals. *Nat. Phys.* **2014**, *10*, 387–393.

(18) Nguyen, L. T.; Cava, R. J. Hexagonal Perovskites as Quantum Materials. *Chem. Rev.* **2020**, *121*, 2935–2965.

(19) Eerenstein, W.; Mathur, N. D.; Scott, J. F. Multiferroic and Magnetoelectric Materials. *Nature* **2006**, *442*, 759–765.

(20) Khomskii, D. I.; Streltsov, S. V. Orbital Effects in Solids: Basics, Recent Progress, and Opportunities. *Chem. Rev.* **2020**, *121*, 2992–3030.

(21) Batra, R.; Song, L.; Ramprasad, R. Emerging Materials Intelligence Ecosystems Propelled by Machine Learning. *Nat. Rev. Mater.* **2020**, 1–24.

(22) Hoppe, R. Effective Coordination Numbers (ECoN) and Mean Fictive Ionic Radii (MEFIR). *Z. Kristallogr. - Cryst. Mater.* **1979**, *150*, 23–52.

(23) Brunner, G. O. A Definition of Coordination and its Relevance in the Structure Types AlB_2 and NiAs. *Acta Crystallogr., Sect. A* **1977**, *33*, 226–227.

(24) O'Keefe, M.; Brese, N. Atom Sizes and Bond Lengths in Molecules and Crystals. *J. Am. Chem. Soc.* **1991**, *113*, 3226–3229.

(25) Zimmermann, N. E. R.; Jain, A. Local Structure Order Parameters and Site Fingerprints for Quantification of Coordination Environment and Crystal Structure Similarity. *RSC Adv.* **2020**, *10*, 6063–6081.

(26) Isayev, O.; Oses, C.; Toher, C.; Gossett, E.; Curtarolo, S.; Tropsha, A. Universal Fragment Descriptors for Predicting Properties of Inorganic Crystals. *Nat. Commun.* **2017**, *8*, 15679.

(27) Waroquiers, D.; Gonze, X.; Rignanese, G.-M.; Welker-Nieuwoudt, C.; Rosowski, F.; Göbel, M.; Schenk, S.; Degelmann, P.; André, R.; Glaum, R.; Hautier, G. Statistical Analysis of Coordination Environments in Oxides. *Chem. Mater.* **2017**, *29*, 8346–8360.

(28) Bartók, A. P.; Kondor, R.; Csányi, G. On Representing Chemical Environments. *Phys. Rev. B* **2013**, *87*, 184115.

(29) Rupp, M.; Tkatchenko, A.; Müller, K.-R.; von Lilienfeld, O. A. Fast and Accurate Modeling of Molecular Atomization Energies with Machine Learning. *Phys. Rev. Lett.* **2012**, *108*, 058301.

(30) Huo, H.; Rupp, M. Unified Representation of Molecules and Crystals for Machine Learning. 2018, arXiv:1704.06439. arXiv.org e-Print archive. `https://arxiv.org/abs/1704.06439`, (accessed December 14, 2021).

(31) Cumby, J.; Attfield, J. P. Ellipsoidal Analysis of Coordination Polyhedra. *Nat. Commun.* **2017**, *8*, 14235.

(32) Perez-Mato, J. M.; Orobengoa, D.; Aroyo, M. I. Mode Crystallography of Distorted Structures. *Acta Crystallogr., Sect. A: Found. Adv.* **2010**, *66*, 558–590.

(33) Kerman, S.; Campbell, B. J.; Satyavarapu, K. K.; Stokes, H. T.; Perselli, F.; Evans, J. S. O. The superstructure determination of displacive distortionsviasymmetry-mode analysis. *Acta Crystallogr., Sect. A: Found. Crystallogr.* **2012**, *68*, 222–234.

(34) Islam, M. A.; Rondinelli, J. M.; Spanier, J. E. Normal Mode Determination of Perovskite Crystal Structures with Octahedral Rotations: Theory and Applications. *J. Phys.: Condens. Matter* **2013**, *25*, 175902.

(35) Schranz, W.; Rychetsky, I.; Hlinka, J. Polarity of Domain Boundaries in Nonpolar Materials Derived from Order Parameter and Layer Group Symmetry. *Phys. Rev. B* **2019**, *100*, 184105.

(36) Mochizuki, Y.; Sung, H.-J.; Takahashi, A.; Kumagai, Y.; Oba, F. Theoretical Exploration of Mixed-Anion Antiperovskite Semiconductors M_3XN(M=Mg,Ca,Sr,Ba; X=P,As,Sb,Bi). *Phys. Rev. Mater.* **2020**, *4*, 044601.

(37) Yang, R. X.; Skelton, J. M.; da Silva, E. L.; Frost, J. M.; Walsh, A. Assessment of Dynamic Structural Instabilities Across 24 Cubic Inorganic Halide Perovskites. *J. Chem. Phys.* **2020**, *152*, 024703.

(38) Howard, C. J.; Stokes, H. T. Group-Theoretical Analysis of Octahedral Tilting in Perovskites. *Acta Crystallogr., Sect. B: Struct. Sci* **1998**, *54*, 782–789.

(39) Wagner, N.; Puggioni, D.; Rondinelli, J. M. Learning from Correlations Based on Local Structure: Rare-Earth Nickelates Revisited. *J. Chem. Inf. Model.* **2018**, *58*, 2491–2501.

(40) Lewis, J. W.; Payne, J. L.; Evans, I. R.; Stokes, H. T.; Campbell, B. J.; Evans, J. S. O. An Exhaustive Symmetry Approach to Structure Determination: Phase Transitions in $Bi_2Sn_2O_7$. *J. Amer. Chem. Soc.* **2016**, *138*, 8031–8042.

(41) Castelli, I. E.; Olsen, T.; Datta, S.; Landis, D. D.; Dahl, S.; Thygesen, K. S.; Jacobsen, K. W. Computational Screening of Perovskite Metal Oxides for Optimal Solar Light Capture. *Energy Environ. Sci.* **2012**, *5*, 5814–5819.

(42) Fabini, D. H.; Laurita, G.; Bechtel, J. S.; Stoumpos, C. C.; Evans, H. A.; Kontos, A. G.; Raptis, Y. S.; Falaras, P.; Van der Ven, A.; Kanatzidis, M. G.; et al., Dynamic Stereochemical Activity of the Sn^{2+} Lone Pair in Perovskite $CsSnBr_3$. *J. Amer. Chem. Soc.* **2016**, *138*, 11820–11832.

(43) Correa-Baena, J.-P.; Nienhaus, L.; Kurchin, R. C.; Shin, S. S.; Wieghold, S.; Putri Hartono, N. T.; Layurova, M.; Klein, N. D.; Poindexter, J. R.; Polizzotti, A.; Sun, S.; Bawendi, M. G.; Buonassisi, T. A-Site Cation in Inorganic $A_3Sb_2I_9$ Perovskite Influences Structural Dimensionality, Exciton Binding Energy, and Solar Cell Performance. *Chem. Mater.* **2018**, *30*, 3734–3742.

(44) Filip, M. R.; Giustino, F. The Geometric Blueprint of Perovskites. *Proc. Natl. Acad. Sci.* **2018**, *115*, 5397–5402.

(45) Maughan, A. E.; Ganose, A. M.; Scanlon, D. O.; Neilson, J. R. Perspectives and Design Principles of Vacancy-Ordered Double Perovskite Halide Semiconductors. *Chem. Mater.* **2019**, *31*, 1184–1195.

(46) Tao, Q.; Xu, P.; Li, M.; Lu, W. Machine Learning for Perovskite Materials Design and Discovery. *npj Comput. Mater.* **2021**, *7*, 23.

(47) Talapatra, A.; Uberuaga, B. P.; Stanek, C. R.; Pilania, G. A Machine Learning Approach for the Prediction of Formability and Thermodynamic Stability of Single and Double Perovskite Oxides. *Chem. Mater.* **2021**, *33*, 845–858.

(48) Tinkham, M. *Group Theory and Quantum Mechanics*; Dover publications, Inc, 2003.

(49) Togo, A.; Tanaka, I. First Principles Phonon Calculations in Materials Science. *Scr. Mater.* **2015**, *108*, 1–5.

(50) Dresselhaus, M. S.; Dresselhaus, G.; Jorio, A. *Group Theory: Application to the Physics of Condensed Matter*; Springer Science & Business Media, 2007.

(51) Kuhn, H. W. The Hungarian Method for the Assignment Problem. *Naval research logistics quarterly* **1955**, *2*, 83–97.

(52) Ong, S. P.; Richards, W. D.; Jain, A.; Hautier, G.; Kocher, M.; Cholia, S.; Gunter, D.; Chevrier, V. L.; Persson, K. A.; Ceder, G. Python Materials Genomics (Pymatgen): A Robust, Open-Source Python Library for Materials Analysis. *Comput. Mater. Sci.* **2013**, *68*, 314–319.

(53) Goodall, R. E. A.; Lee, A. A. Order Matters: Sequence to Sequence for Sets. 2016, arXiv:1511.06391. arXiv.org e-Print archive. `https://arxiv.org/abs/1511.06391`, (accessed December 14, 2021).

(54) Pedregosa, F. et al. Scikit-learn: Machine Learning in Python. *J. Mach. Learn Res.* **2011**, *12*, 2825–2830.

(55) Jain, A.; Ong, S. P.; Hautier, G.; Chen, W.; Richards, W. D.; Dacek, S.; Cholia, S.; Gunter, D.; Skinner, D.; Ceder, G.; Persson, K. A. Commentary: The Materials Project: A Materials Genome Approach to Accelerating Materials Innovation. *APL Mater.* **2013**, *1*, 011002.

(56) Ong, S. P.; Cholia, S.; Jain, A.; Brafman, M.; Gunter, D.; Ceder, G.; Persson, K. A. The Materials Application Programming Interface (API): A Simple, Flexible and Efficient API for Materials Data Based On Representational State Transfer (Rest) Principles. *Comput. Mater. Sci.* **2015**, *97*, 209–215.

(57) Togo, A.; Tanaka, I. `Spglib`: A Software Library for Crystal Symmetry Search. 2018, arXiv:1808.01590. arXiv.org e-Print archive. `https://arxiv.org/abs/1808.01590`, (accessed December 14, 2021).

(58) Blöchl, P. E. Projector Augmented-Wave Method. *Phys. Rev. B* **1994**, *50*, 17953.

(59) Kresse, G.; Furthmüller, J. Efficiency of Ab-Initio Total Energy Calculations for Metals and Semiconductors Using a Plane-Wave Basis Set. *Comput. Mater. Sci.* **1996**, *6*, 15–50.

(60) Kresse, G.; Furthmüller, J. Efficient Iterative Schemes for Ab Initio Total-Energy Calculations Using a Plane-Wave Basis Set. *Phys. Rev. B* **1996**, *54*, 11169.

(61) VISE, v0.1.13; GitHub repository: `https://github.com/kumagai-group/vise`, (accessed December 14, 2021).

(62) Perdew, J. P.; Burke, K.; Ernzerhof, M. Generalized gradient approximation made simple. *Phys. Rev. Lett.* **1996**, *77*, 3865.

(63) Heyd, J.; Scuseria, G. E.; Ernzerhof, M. Hybrid Functionals Based on a Screened Coulomb Potential. *J. Chem. Phys.* **2003**, *118*, 8207–8215.

(64) Momma, K.; Izumi, F. Vesta 3 for Three-Dimensional Visualization of Crystal, Volumetric and Morphology Data. *J. Appl. Crystallogr.* **2011**, *44*, 1272–1276.

(65) Frost, J. M.; Butler, K. T.; Walsh, A. Molecular ferroelectric contributions to anomalous hysteresis in hybrid perovskite solar cells. *APL Mater.* **2014**, *2*, 081506.

(66) Mattoni, A.; Filippetti, A.; Caddeo, C. Modeling hybrid perovskites by molecular dynamics. *J. Phys.: Condens. Matter* **2016**, *29*, 043001.

(67) Csonka, G. I.; Perdew, J. P.; Ruzsinszky, A.; Philipsen, P. H. T.; Lebègue, S.; Paier, J.; Vydrov, O. A.; Ángyán, J. G. Assessing the performance of recent density functionals for bulk solids. *Phys. Rev. B* **2009**, *79*, 155107.

(68) Indenbom, V. Phase Transitions Without Change of the Atom Number in the Crystal Unit Cell. *Kristallografiya* **1960**, *5*, 115–125.

(69) Levanyuk, A. P.; Sannikov, D. G. Improper Ferroelectrics. *Phys.-Usp.* **1974**, *17*, 199–214.

(70) Benedek, N. A.; Fennie, C. J. Why Are There So Few Perovskite Ferroelectrics? *J. Phys. Chem. C* **2013**, *117*, 13339–13349.

(71) Wang, J.; Neaton, J. B.; Zheng, H.; Nagarajan, V.; Ogale, S. B.; Liu, B.; Viehland, D.; Vaithyanathan, V.; Schlom, D. G.; Waghmare, U. V.; ; Spaldin, N. A.; Rabe, K.; Wuttig, M.; Ramesh, R. Epitaxial $BiFeO_3$ Multiferroic Thin Film Heterostructures. *Science* **2003**, *299*, 1719–1722.

(72) Neaton, J. B.; Ederer, C.; Waghmare, U. V.; Spaldin, N. A.; Rabe, K. First-principles study of spontaneous polarization in multiferroic $BiFeO_3$. *Phys. Rev. B* **2005**, *71*, 014113.

(73) Walsh, A.; Payne, D. J.; Egdell, R. G.; Watson, G. W. Stereochemistry of Post-Transition Metal Oxides: Revision of the Classical Lone Pair Model. *Chem. Soc. Rev.* **2011**, *40*, 4455–4463.

(74) Van der Maaten, L.; Hinton, G. Visualizing data using t-SNE. *J. Mach. Learn Res.* **2008**, *9*, 2579–2605.

(75) Pedregosa, F. et al. Scikit-Learn: Machine Learning in PYthon. *J. Mach. Learn Res.* **2011**, *12*, 2825–2830.

(76) Iasir, A. R. M.; Lombardi, T.; Lu, Q.; Mofrad, A. M.; Vaninger, M.; Zhang, X.; Singh, D. J. Electronic and magnetic properties of perovskite selenite and tellurite compounds: $CoSeO_3$, $NiSeO_3$, $CoTeO_3$, and $NiTeO_3$. *Phys. Rev. B* **2020**, *101*, 045107.

(77) Sawaguchi, E.; Akishige, Y.; Yamamoto, T.; Nakahara, J. Phase Transition in Hexagonal Type $BaTiO_3$. *Ferroelectrics* **1989**, *95*, 29–36.

(78) Hashemizadeh, S.; Biancoli, A.; Damjanovic, D. Symmetry Breaking in Hexagonal and Cubic Polymorphs of $BaTiO_3$. *J. Appl. Phys.* **2016**, *119*, 094105.

(79) Ogawa, N.; Ogimoto, Y.; Ida, Y.; Nomura, Y.; Arita, R.; Miyano, K. Polar Antiferromagnets Produced with Orbital Order. *Phys. Rev. Lett.* **2012**, *108*, 157603.

(80) Gupta, K.; Mahadevan, P.; Mavropoulos, P.; Ležaić, M. Orbital-Ordering-Induced Ferroelectricity in $SrCrO_3$. *Phys. Rev. Lett.* **2013**, *111*, 077601.

(81) Ding, Y.; Cao, L.; Wang, W.; Jing, B.; Shen, X.; Yao, Y.; Xu, L.; Li, J.; Jin, C.; Yu, R. Bond Length Fluctuation in Perovskite Chromate $SrCrO_3$. *J. Appl. Phys.* **2020**, *127*, 075106.

(82) Parsaeifard, B.; De, D. S.; Christensen, A. S.; Faber, F. A.; Kocer, E.; De, S.; Behler, J.; von Lilienfeld, O. A.; Goedecker, S. An assessment of the structural resolution of various fingerprints commonly used in machine learning. *Mach. Learn.: Sci. Technol.* **2021**, *2*, 015018.

(83) Fung, V.; Zhang, J.; Juarez, E.; Sumpter, B. G. Benchmarking graph neural networks for materials chemistry. *npj Comp. Mater.* **2021**, *7*, 84.