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 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.

Introduction

Materials informatics has grown into a substantial field, supported by the surge in 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 are 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]

Structure would be easier to represent if we were able to apply a filter to smear atomistic properties in a mean-field manner. Although such a 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 to macroscopic properties. For example, in perovskites, slight displacement of B-site cation could induce both a local electric dipole, as well as macroscopically observable ferroelectric behaviours.[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 have 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, but the rudimentary examples include 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 quantitiative defintion.[22–25] One advanced method is to analyse the connectivity of the bonds 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]

In this paper, we take advantage of established techniques in group theory and use it to encode polyhedron shape.[32–37] In particular, we projected the distortions onto the basis vectors of the irreducible representations (irreps) to obtain 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.[38–44] 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 difference in octahedral geometry.

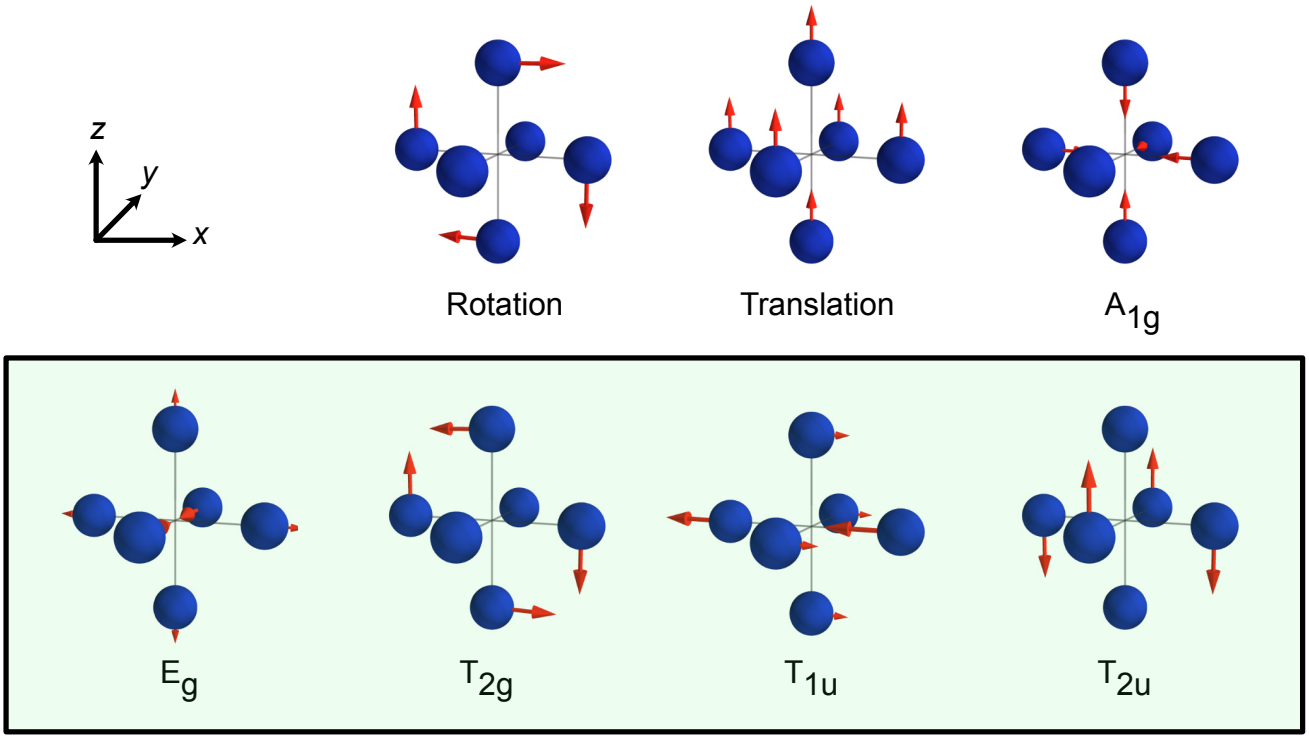

Figure 1: Basis set distortions for the irreducible representations of a six atom octahedron as found in a cubic perovskite. For multidimensional irreducible representations, only one distortion is shown. For the actual projection, we have used the four distortions presented in the bottom row. The full list is presented in Figure S1 and S2.

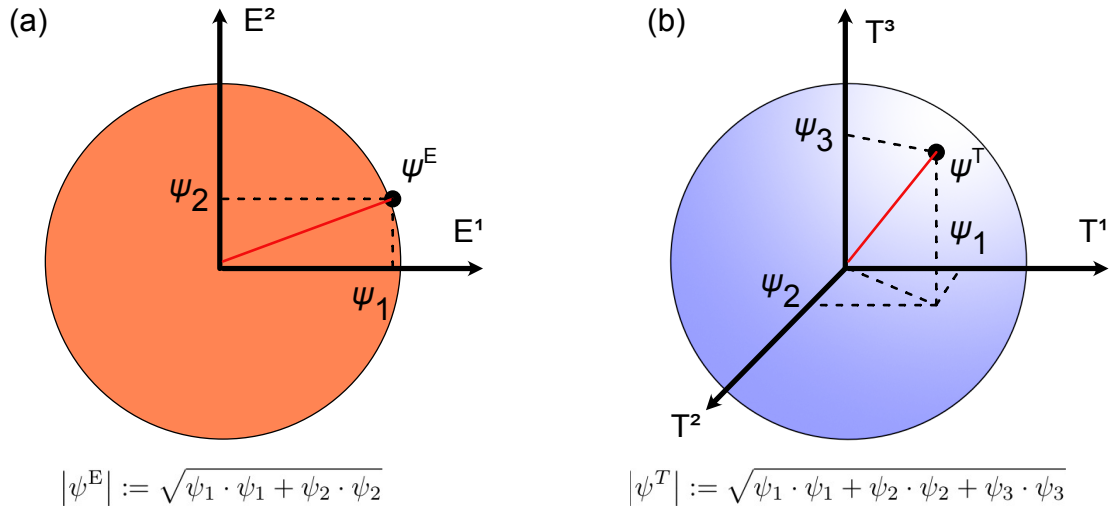

Figure 2: Illustration of how amplitudes are averaged within a multidimensional irreducible representation. (a) a two dimensional irreducible representation (E_g), (b) three dimensional irreducible representations (T_{2g}, T_{1u} and T_{2u}).

Methodology

Defining basis vectors

The goal of this section is to calculate complete and orthogonal basis distortions (basis vectors) of the irreps. The irreps fulfill the "great orthogonality theorem",[45]

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

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 $\Gamma^{(i)}$.[46] We cannot directly use this however, because the specific elements of Γ are unknown *a priori*. Therefore, throughout the section, we make 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 notations follows Ref 45.

Firstly, we need to calculate 18 dimensional reducible representations, which is a direct product between 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.[47] On the other hand, six dimensional representations $\tilde{\Gamma}^{(6)}(R)$ depends 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 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 trace of equation 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 (single 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",[45] 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 $f^{(i)}_\kappa$, the κ-th element within irrep i of the function F

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

Again a problem arises due to lack of the knowledge on $\Gamma^{(i)}(R)$. Analogically to the relation

between equation 1 and 2, there is a slightly restricted version,[48] 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 a 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 arbitrary 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 as an initial vector. Although the number of trial initial vectors is arbitrary, this choice is the minimum number required to generate all the 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 equals to 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 equation 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 1 (full list in Figure S1 and S2).

Projection to the normal 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 same shape but different size, 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 obtained the distortion vector by comparing it against ideal octahedron with bonding length of 1.0 Å. By applying this scaling, the resulting distortion amplitudes for octahedra of same shape, but different size 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.[49,50] 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 1. 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 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 (equation 3), the resulting amplitudes of the basis vectors would have had dependence on the choice of the axis. For example if the T_{1u} distortion in Figure 1 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.[51]

Our approach was to use the total length spanned by vectors within the irreps. As shown in Figure 2, we have calculated the length of the vectors in two or three dimensional space using the Euclidean distance,

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

Here the summation is over the dimension inside irrep (i). Just like the Euclidean distance of a given point from the origin remains 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 2, 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.

Dataset processing

In order to apply the projection, we have obtained 46,048 materials from the *Materials Project* database accessed through the API in *pymatgen* package.[50,52,53] We then searched for materials containing octahedra with stoichiometry of ABO_3 resulting in 492 materials. The octahedra was constructed by selecting all the B-site cations with distinct Wyckoff positions using the *spglib* and applying the *CrystalNN* algorithm to detect the surrounding anions.[25,54] Since this method extracts multiple octahedra for a single material, we obtained 552 distinct octahedron in total, in which we have treated them as independent data points. It is worth adding that 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 density functional theory (DFT) potential energy landscapes.

Density functional theory calculations

Although we have largely applied the method to openly available from the *Materials Project* database,[50,52,53] for validation we performed some calculations with stricter condition. The plane-wave DFT within projector-augmented wave scheme in calculations were performed using the *VASP*.[55–57] The input file was automatically generated via *VISE* package,[58] 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 functional.[59] The visualisation of the structures were done using the *VESTA* software.[60]

Results and discussion

Projection onto normal distortions

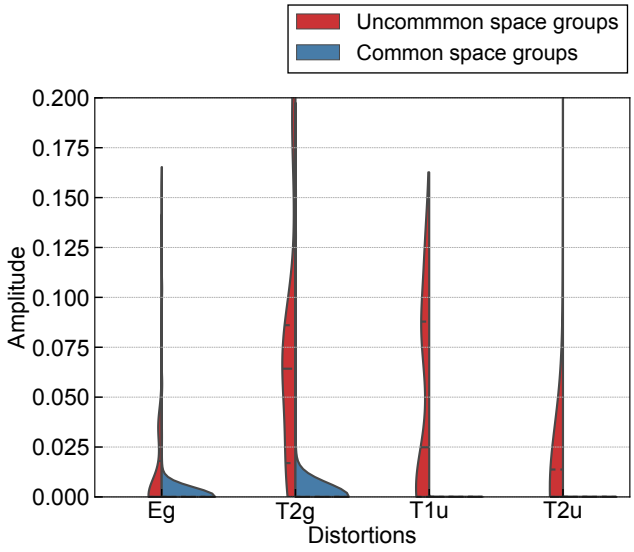

Figure 3: Distortion amplitudes for 552 ABO_3 materials. The blue and red shading refers to materials that belong or do not belong to the common space groups for distorted perovskites, respectively.

The distribution of distortion amplitudes for all 552 materials are presented in Figure 3. The materials are categorised by whether or not they belong to the common point shared perovskite space group (cubic $Pm\bar{3}m$, tetragonal P4mm, tetragonal P4/mmm, tetragonal P4/mbm, tetragonal I4/mcm, orthorhombic Pnma, orthorhombic Amm2, orthorhombic Cmcm, monoclinic $P2_1/m$, rhombohedral R3m, rhombohedral R3c, and rhombohedral $R\bar{3}c$).[32] The number of materials in common and uncommon space groups were 443 and 109, respectively. From Figure 3, 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 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 the fact that there was no clear

chemical trends (Figure S4∼S7), this result suggests a strong relation between the crystal structure and the local distortions of the octahedra.

Connectivity analysis

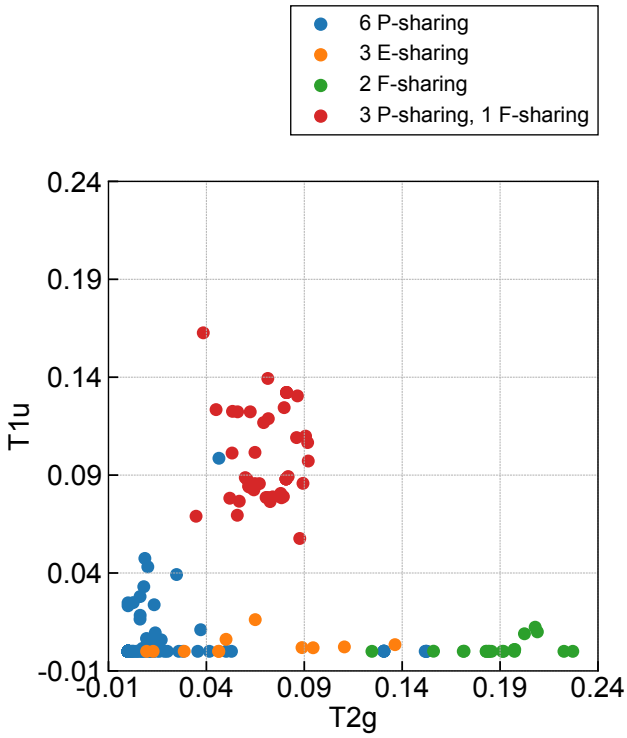

Figure 4: Relation between T_{1u} distortion against the T_{2g} distortion. Each points represent an octahedron site and are coloured according to their connectivity with other octahedra. Blue points are connected with via six point sharing (6 P-sharing), orange points are connected with via three edge sharing (3 E-sharing), red points are connected with via three point sharing and one face sharing (3 P-sharing and 1 F-sharing), and green points are connected with via two face sharing (2 F-sharing).

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 4). The four connectivities in Figure 4 are: six point sharing (6 P-sharing, A and C in Figure 5), three edge sharing (3 E-sharing, B in Figure 5), two face sharing (2 F-sharing, E in Figure 5), and three point sharing and one face sharing (3 P-sharing and 1 F-sharing, D in Figure 5).

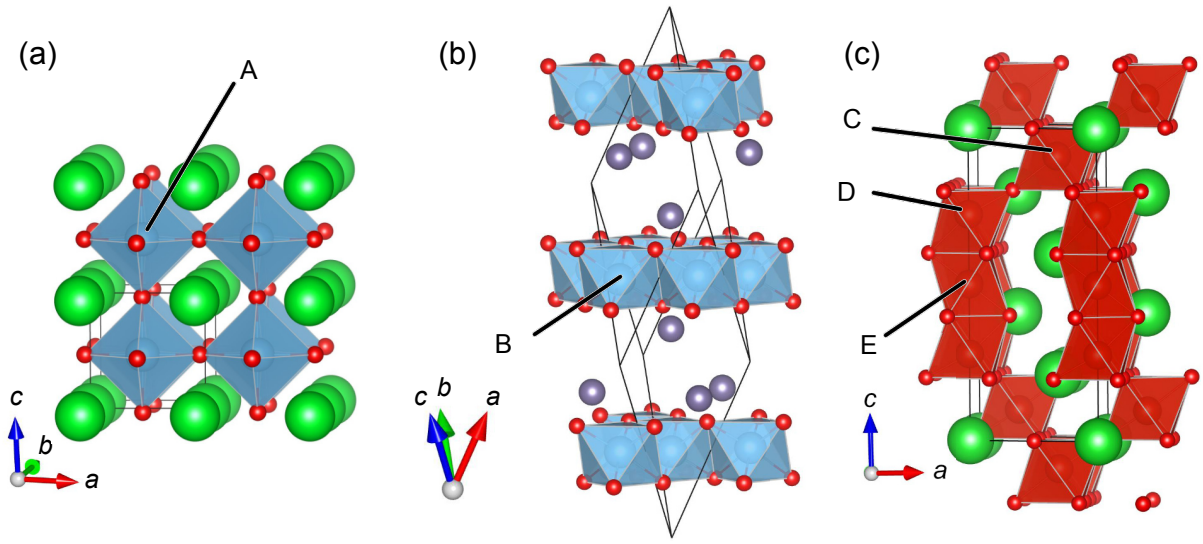

Figure 5: Structures of (a) cubic $SrTiO_3$, (b) rhombohedral $GeTiO_3$, and (c) hexagonal $BaVO_3$. Octahedra are composed of TiO_6, TiO_6, and VO_6, respectively. Distinct Wycoff positions are labeled by A to E.

Clearly, a cluster of distortion amplitudes are distinguishable about $(T_{2g},T_{1u}) = (0.075,0.100)$. Two interesting observation 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. This is analogous to the situation in improper ferroelectrics where coexistence of two distortions create a ferroelectric distortion.[61–63] Secondly, this cluster is composed mostly of three point 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 5(c). Accounting the fact that three point sharing and one face sharing octahedra were not seen outside of this cluster, this result indicates that hexagonal phases could support distortions much larger than that seen in point shared perovskites. The one fully-point-shared outlier in the cluster was $BiFeO_3$, which exhibited unusually large distortion. The possible origin of this distortion is an interplay between Bi lone pairs and Fe Jahn-Teller distortion.[64,65]

Outside of this cluster, the T_{1u} distortion was generally small. Most of the fully point shared octahedra and fully edge sharing octahedra had an ideal octahedron structure, which made the data points to be scattered around the zero amplitude point. Two face sharing

octahedra interestingly, had 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 5, the uniaxial strain due to being sandwiched by neighbouring octahedra likely to have caused the compression of the octahedron.

Clustering analysis

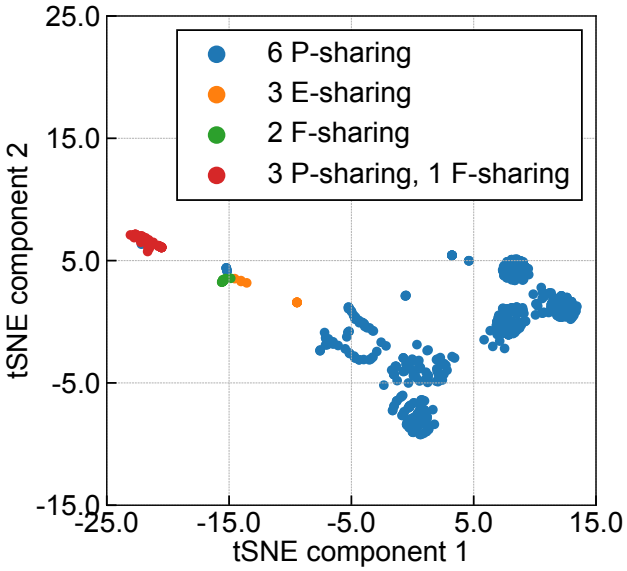

Figure 6: The clustering of different octahedron connectivity plotted on the dimensionally reduced axis obtained through t-distributed stochastic neighbor embedding (t-SNE).

Up to here, we have made discussions based on the trends in Figure 4, however such a discussion may be overlooking trends in higher dimensions. Therefore, we have performed dimensional 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.[66,67] The result is shown in Figure 6. Different octahedral connectivities are clearly separated. This result is fortuitous since it indicates that the shape of octahedra are largely determined by their connectivities with neighbouring octahedra. In other words, geometrical network of bonds dominantly determine the shape of the octahedra rather than the chemical property of individual bonds.

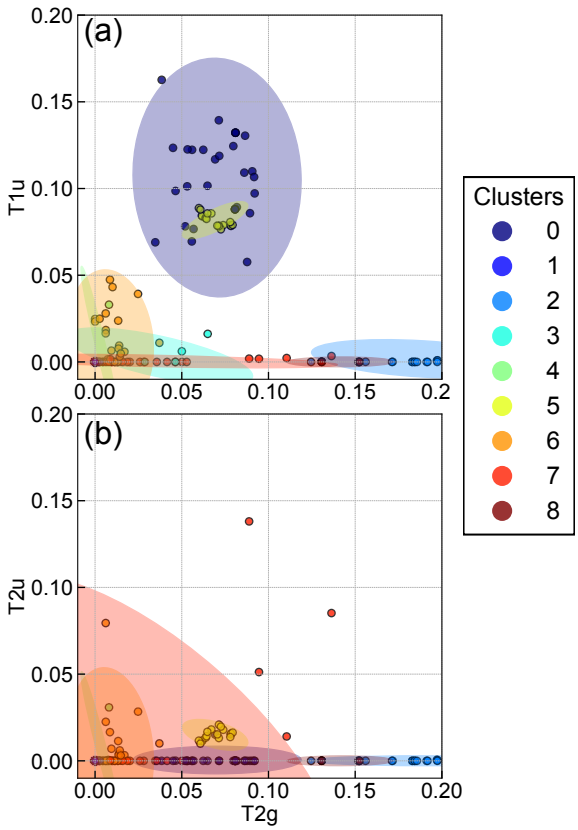

Figure 7: Clusters obtained by a Gaussian mixture model (GMM) shown in the axis of (a) T_{2g} and T_{1u}, and (b) T_{2g} and T_{2u}. The dots are coloured differently depending which of seven different cluster the point belongs to. The shading shows the extent of the multivariate Gaussian distribution defined for each clusters.

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) as implemented in the *Scikit-learn* package.[67] GMM requires 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 7 (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 7(a), a clear ellipsoid of cluster 0 can be distinguished. This cluster corresponds to the three point sharing and one face sharing in Figure 4 at (T_{2g},T_{1u}) = (0.075,0.100). Taking a closer look reveals that there are subset of materials within the ellipsoid that belong to cluster 5. Their difference is not distinguishable from Figure 7(a), but plotting against the T_{2u} distortion axis in Figure 7(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 0.02 T_{2u} distortion. This separation is not trivial from Figure 4 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

Cluster 5 in Figure 7 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 8 (a) and (b), respectively). Experimentally, the $C222_1$ is stable in the range of about 70$\sim$220 K, where it transforms in to the $P6_3/mmc$ phase at 220K.[68,69] 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.

To confirm whether the absence of the T_{2u} distortions in other ABO_3 is due to the lack of data or due to different phase stability, we have compared the energies of $P6_3/mmc$ and

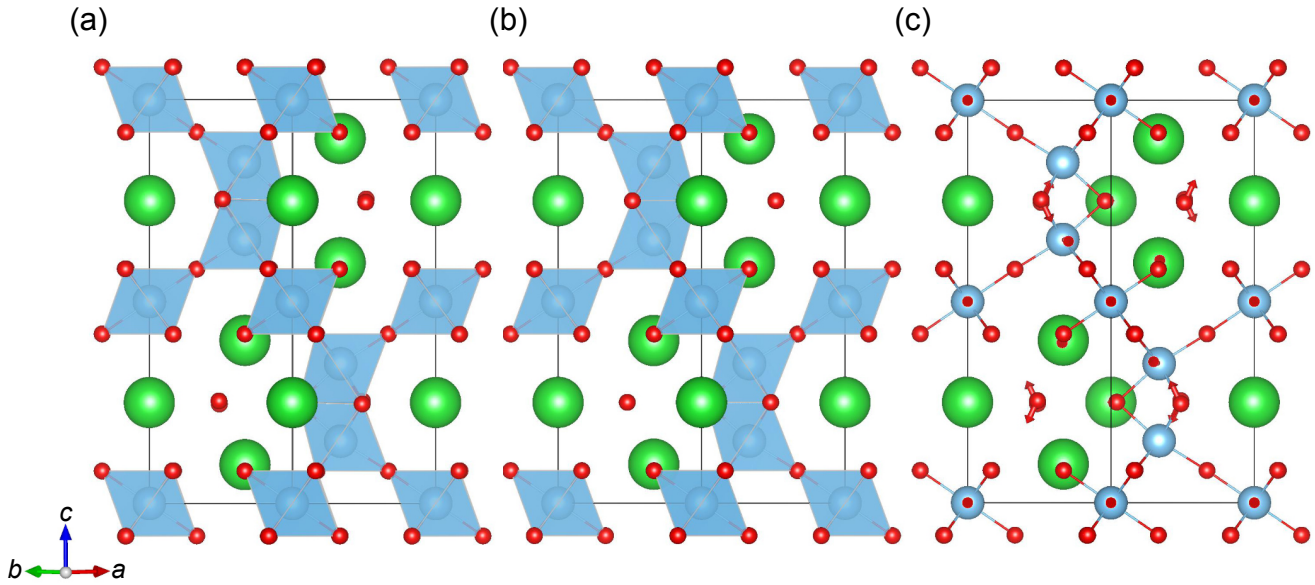

Figure 8: Structure of $BaTiO_3$ in (a) $C222_1$ and (b) $P6_3/mmc$ phase. (c) The atomic displacement to transform from $C222_1$ to $P6_3/mmc$. The blue, green, and red spheres are Ba, Ti, and O atoms respectively. The blue shading represents the TiO_6 octahedron.

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

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

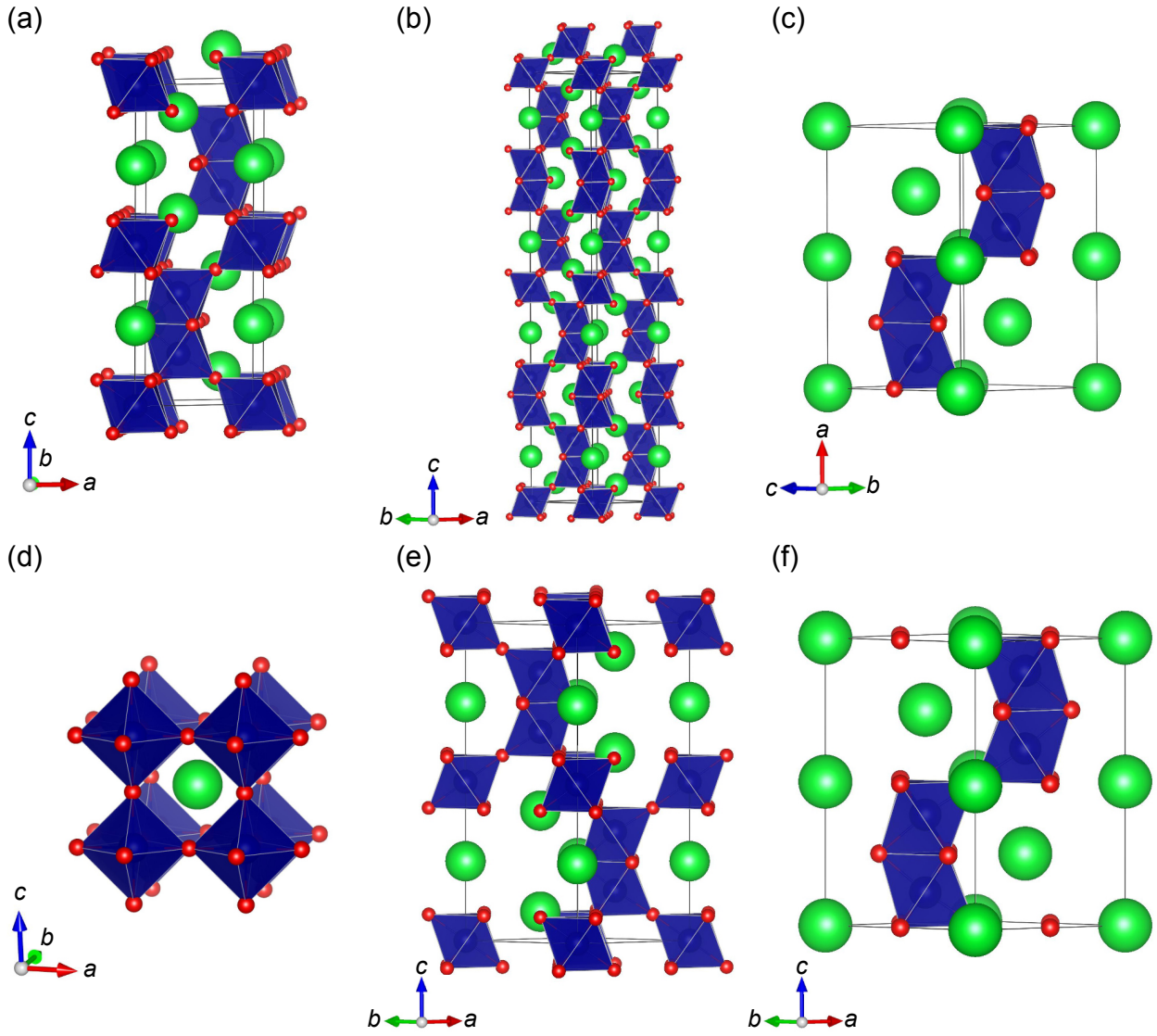

Figure 9: Structures of different $SrCrO_3$ polymorphs obtained from the *Materials Project*.[50,52,53] The details are summarised in Table S2.

the $C222_1$ phases in 11 additional compounds (Table 2). We found that in most compounds $C222_1$ phase were 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 interplay of metallicity, ferroelectricity and magnetic order. In cubic $SrCrO_3$ (Figure 9(d)), there has been reports on multiferroicity, which are induced by orbital ordering.[20,70,71] Since this material has been suggested to be internally strained,[72] we believe this is the reason for the distinct distortion behaviour of this material. For hexagonal polytypes of $SrCrO_3$ (Figure 9), as far as we are aware, there has not been previous reports but we note that the formation energy predicted by DFT is smaller than the known cubic phase (Table S2), so they should be accessible. Interestingly, within the hexagonal phases, the Ama2 phases (Figure 9(a), (b), and (c)) were calculated to be metallic, whereas $P6_3/mmc$ phases (Figure 9(e) and (f)) were insulators (Table S2).[73] Since, the ratio of point 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.

Conclusion

We have shown that using group theory, distortions in polyhedra could 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 distor-

tions, 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 rich variety of ferroic behaviours. 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 to elucidate additional trends. We would like to emphasise that this method is not exclusive, and synergistic effects are expected when used with other means of material featurisation techniques. Finally, the rich result of this study is only a result from very elementally dimensional analyses techniques, which suggest that usage of more sophisticated approaches that are suitable for higher non-linearity, such as deep neural networks, are expected to open a path towards further material discoveries.

Acknowledgement

The authors 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. 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 result 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.

(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**,

(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*.

(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.; Lee, A. A. Machine Learnt Approximations to the Bridge Function Yield Improved Closures for the Ornstein-Zernike Equation. 2021.

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

(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*.

(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*.

(30) Huo, H.; Rupp, M. Unified Representation of Molecules and Crystals for Machine Learning. 2018.

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

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

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

(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*.

(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*.

(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) 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.

(39) 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 Stereo-

chemical Activity of the Sn^{2+} Lone Pair in Perovskite $CsSnBr_3$. *J. Amer. Chem. Soc.* **2016**, *138*, 11820–11832.

(40) 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.

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

(42) 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.

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

(44) 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.

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

(46) Note that we assumed that Γ is real, which is sufficient for the analyses in this is paper. More general definition is to use complex conjugate $\Gamma^{(i)*}$ instead.

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

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

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

(50) 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.

(51) Vinyals, O.; Bengio, S.; Kudlur, M. Order Matters: Sequence to Sequence for Sets. 2016.

(52) 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.

(53) 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.

(54) Togo, A.; Tanaka, I. `Spglib`: A Software Library for Crystal Symmetry Search. 2018.

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

(56) 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.

(57) 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.

(58) VISE. `https://github.com/kumagai-group/vise`, 2021.

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

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

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

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

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

(64) 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.

(65) 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.

(66) Hinton, G. E.; Roweis, S. Stochastic Neighbor Embedding. Advances in Neural Information Processing Systems. 2003.

(67) Pedregosa, F. et al. Scikit-Learn: Machine Learning in PYthon. *JMLR* **2011**, *12*, 2825–2830.

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

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

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

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

(72) 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.

(73) The formation energies and the band gaps were obtained from the *Materials Project*.[50,52,53] For polymorphs without a band gap, we have confirmed their metallicity with our DFT calculations with HSE06 functional.