Using SVM to Solve the Perovskite Classification Problem

F. Yang[1,*], B. Hou[1,*], E. Liu[1,*], S. Xie[1,*], and M. Thachuk[1,†]

[1] *Department of Chemistry, University of British Columbia, Vancouver BC*

Abstract

The perovskite structure is one of the most abundant structural families with wide-ranging properties and applications in solid state chemistry. The majority of the perovskites is well-described by the chemical formula ABX_3; however, due to its spacial arrangement of cations and anions, tolerance to distortion, and non-stoichiometry, the formation of the correct crystal structure depends on the combination of individual properties of A, B, and X. The relevance of each property, or feature influencing the formability of perovskites has not been well understood. The lack of research in compositional and configurational degrees of freedom makes identification and classification of the structures that have ABX_3 non-trivial. In this study, we collectively investigated the formability of perovskite structures from 181 experimental results. We demonstrate the power and utility of machine learning via training a support vector machine (SVM) classifier that uses combinations of two, three, and four features to predict the formability of a given ABX_3 formulation in the perovskite crystal structure. The model accuracy was examined by using a test set comprising of 20% the experimental data. The results indicate that the combination of the Goldschmidt tolerance factor and tetrahedral tolerance factor confidently predicted the formability of perovskite structures with the highest accuracy. We show that the algebraic combinations of the ionic radii generates the best model accuracy, suggesting that steric and geometric packing effects govern the stability of these compounds. Addition of a third or fourth feature improved the general qualities of the low-performing models; however, the addition of features did not improve the accuracy of the best performing model developed from the Goldschmidt tolerance factor and tetrahedral tolerance factor. The results shed light on the future material design under the guidance of data-driven approaches.

1 Introduction

With the deleterious consequences of the reliance on fossil fuels for energy becoming more apparent, the development of sustainable energy is becoming a more urgent concern. In pursuit of the ideal of capturing energy directly from the sun, development of photovoltaic power conversion technologies has become a major area of focus. However, as efforts are made to scale up solar energy conversion to widespread use, shortcomings of photovoltaic materials are becoming apparent. Namely, mature solar cell technologies currently used suffer from efficiencies significantly lower than theoretical limits, costly raw materials, expensive manufacturing processes, toxicity and limited long-term stability.[1] In the twenty-first century, perovskite materials have become a center of research efforts.

Historically, perovskite refers to a mineral of formula $CaTiO_3$ discovered by Gustav Rose in 1839, but now describes a class of materials with the same crystal structure, notated ABX_3. Said crystal structure was described in 1926 by Goldschmidt. [2] As shown in Fig. 1, perovskites

[0] * These authors have contributed equally to this study. For this manuscript, B. Hou wrote the description of the algorithm and part of the discussion; E. Liu wrote most of the results and discussion; S. Xie wrote the introduction and the chemical interpretation in methodology and discussion; L. Yang wrote the abstract, conclusion, part of the results and generated the figures.

are defined by a larger metal cation occupying the vertices of a cubic cell (originally Ca^{2+}, now generally A), an anion occupying the face-centres of the cubic cell (originally O^{2-}, now generally X), and a smaller metal cation occupying the octahedral cavity at the centre of the above ions (originally Ti^{4+} , now generally B). Commonly observed anions are F^-, Cl^-, Br^-, I^-, and O^{2-}.

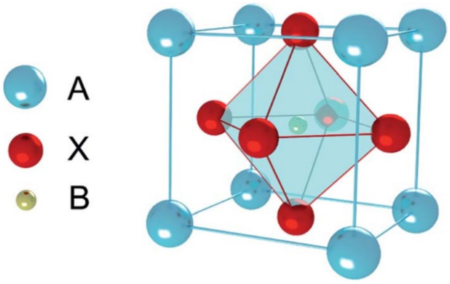


Figure 1: The crystal structure of perovskite materials. A and B correspond to metal cations and X to a halogen or oxygen anion. From [3].

Materials of the perovskite class generally have good absorptivities in the visible and near-UV ranges [4] and also support long exciton diffusion distances [5]. These properties account for the solar energy conversion ability of perovskites. Furthermore, the diversity of perovskite structures has the potential to support cheaper, less toxic materials and more diverse manufacturing techniques.[3] Thus, the characterization of perovskites of different composition is interesting in the search for a solar energy conversion material with idealized properties.

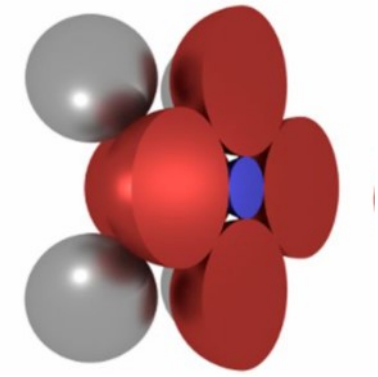


Figure 2: The cross section of the ideal cubic cell obtained by bisecting the cubic cell through the center point of the sides. Note the blue cation (B) occupying the octahedral hole formed by 6 adjacent anions (X). From [6].

The problem arises that not all compounds of formula ABX_3 crystallize in the perovskite structure; unguided attempts to develop novel perovskites may not be fruitful. Whether a particular chemical formula forms the perovskite crystal structure — referred to as formability — is dependent on the geometry of the crystal unit cell. The lowest energy lattice structure, as shown in Fig. 2, has the ions maximally packed such that spheres representing the ionic radius are contacting each other. This arrangement is energy-minimizing as the cations and anions are as close to each other as possible as dictated by electrostatic attraction. In his paper describing the perovskite crystal structure, Goldschmidt notes that this maximal-contact packing is mathematically described by

$$\sqrt{2}(r_B + r_A) = r_A + r_X, \tag{1}$$

where r_A, r_B and r_X are the ionic radii of A, B and X respectively.[2] Thus for the ideal case, where Eq. (1) holds true, dividing the right hand side of Eq. (1) by the left hand side gives a value of 1. Therefore, Goldschmidt defines the octahedral tolerance, t_o, as

$$\tau_o = \frac{r_A + r_X}{\sqrt{2}(r_B + r_A)}. \tag{2}$$

Where $\tau_{\mathbf{o}} = 1$ describes the ideal perovskite, formability has been observed to be possible within a range of $0.813 < \tau_o < 1.107$.[6] A range of different ionic radii can accommodate the perovskite structure, so a further parameter, the octahedral factor, defined as

$$o_f = \frac{r_B}{r_X}, \tag{3}$$

is used to describe the crystal structure. These two parameters have been used to predict perovskite formability using classical methods[1].[7]

It would be profitable for the discovery of new perovskite materials if the researcher could predict perovskite formability by considering several easily obtained parameters, without conducting complex calculations. In the present study we attempt to accomplish this objective through a data-driven approach in the regime of machine learning. We develop formability prediction models by applying pattern recognition to a dataset of 181 ABX_3 structures for which formability data is available. As in general perovskite studies, atom A will be selected from eight monovalent metal ions (Ag, Cs, Cu, K, Li, Na, Rb, and Tl), atom B will be select from twenty bivalent ions (Ba, Be, Ca, Cd, Co, Cr, Cu, Eu, Fe, Hg, Mg, Mn, Ni, Pb, Sn, Sr, Ti, V, Zn, and Zr) and halide X will be one of F, Cl, Br, or I. We first achieve this using literature methods [8], then attempt to further understanding by proposing several new parameters from which prediction models are built.

2 Methodology

2.1 General Workflow of Machine Learning

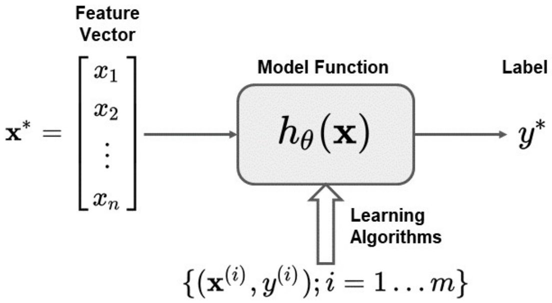
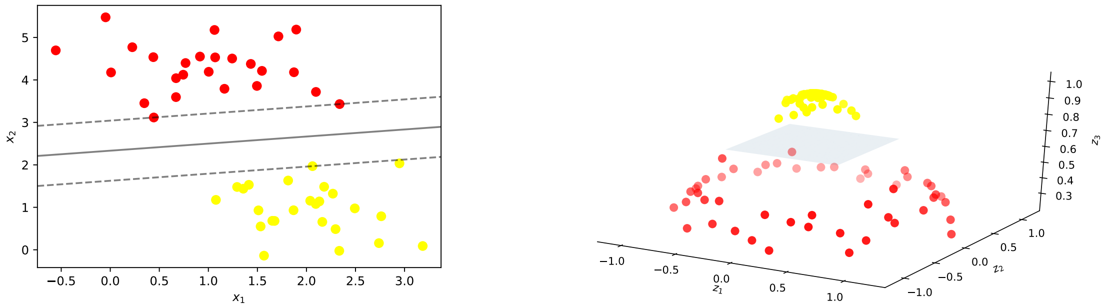

Feature Vector
$\mathbf{x}^* = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$
Model Function
$h_\theta(\mathbf{x})$
Label
y^*
Learning Algorithms
$\{(\mathbf{x}^{(i)}, y^{(i)}); i = 1 \dots m\}$

Figure 3: General workflow of supervised machine learning algorithms: The model function h_θ is trained from the training set $\{\mathbf{x}^{(i)}, y^{(i)}\}$ with m labelled points based on some learning algorithms, and predicts on the unseen data $\mathbf{x}^*$. The predicted label attached to $\mathbf{x}^*$ is shown as y^*.

The main themes in machine learning are pattern recognition and model prediction.[9] Specifically, in the area of supervised learning, the task is to construct a relation or probabilistic model between a finite number of input data and their labels. The total available data are obtained from theoretical calculations labelled by experimentally verified formability. We split the total data into two sets: one for training the model, and the other for testing the model's prediction accuracy. Shown in Fig. 3, each training point is usually encoded in a multidimensional feature vector $\mathbf{x}$, specifying the general properties of the perovskite. The complexity of the algorithm increases with the number of features x_i being considered. Two and three dimensional problems ($n = 2, 3$) are the primary focus considering the results in literature [8]. The variable y denotes the label (1 for perovskite and 0 for non-perovskite) that we are trying to attach to a given feature vector. Labels in the training set are predetermined by experimental results, and machines can 'learn' from that *a priori* knowledge. When y can only take 1 or 0, it is known as a binary classification problem. The core of our project is to capture a model function h_θ from a finite number of training data, such that the function can predict whether given feature vector $\mathbf{x}^*$ can form perovskite with high accuracy. The accuracy of a model is defined by the ratio of correctly predicted samples to the

[1] The classical method for classifying perovskite is rather empirical. Common routines involve constructing a structure map which is simply drawing a line between a two-feature class.

total samples in the test set. As the model has no previous knowledge of those test set samples, this estimation reflect the general performance of the model with broader applications.



(a) Linear SVM

(b) Kernel SVM

Figure 4: Examples of classifiers used in the SVM method. (a) Linearly-separable data classified by a simple linear SVM. The dashed lines illustrate the maximized margin and the solid line is the predicted linear boundary. (b) A kernel SVM adapting the feature mapping in higher dimension. Note that the original two-dimensional data may not be nearly separable. Feature mapping into higher dimensions may resolve this problem. Adapted from [10].

2.2 Support Vector Machine

2.2.1 Linear SVM

Under the framework of machine learning described in Section 2.1, the support vector machine provides a systematic way of determining the model function (classifier) h_θ for a classification problem. Simply speaking, SVM seeks for the optimal hyperplane (or the decision boundary) that best separates the two classes. The position of the hyperplane is tuned by the algorithms based on the marginal distance between the plane and the feature vectors in feature space. To start with, let us first consider a two dimensional feature space with two linearly separable classes. Labelled by red and yellow dots in Fig 4, the two classes, perovskite and non-perovskite, can presumably be separated by a line. If the model function is defined as $h_{\boldsymbol{\theta},b}(\mathbf{x}) = \boldsymbol{\theta}^\intercal \mathbf{x} + b$, the hyperplane is given as $\boldsymbol{\theta}^\intercal \mathbf{x} + b = 0$, where for the 2D case, $\mathbf{x} = [x_1, x_2]^\intercal$, $\boldsymbol{\theta} = [\theta_1, \theta_2]^\intercal$ and b is a constant [2]. Clearly, hyperplane $h_{\boldsymbol{\theta},b}(\mathbf{x}) = 0$ represents a line in 2D feature space. To choose the best position of the line, define the geometric margin γ as

$$\gamma = \min_{i=1\dots m} \gamma^{(i)} = \min_{i=1\dots m} \frac{y^{(i)}(\boldsymbol{\theta}^\intercal \mathbf{x}^{(i)} + b)}{\|\boldsymbol{\theta}\|}. \tag{4}$$

Intuitively, $\gamma^{(i)}$ quantifies the distance between each feature vector and the hyperplane, and the geometric margin γ is the shortest distance among all training points. In Fig. 4(a), the four points on the dash line boundaries characterize the geometric margin, and are referred to as the support vectors. The best hyperplane is the one that maximizes the geometric margin γ, in other word, the ultimate goal of a linear SVM can be written as a optimization problem as follows:

$$\operatorname*{arg\,max}_{\boldsymbol{\theta},b} \gamma \qquad \text{s.t. } \gamma^{(i)} \geq \gamma,\, i = 1\dots m. \tag{5}$$

Mathematically, the above formulation can be further transformed into an equivalent convex optimization problem [3]

$$\begin{gathered} \operatorname*{arg\,min}_{\boldsymbol{\theta},b,\xi_i} \frac{1}{2}\|\boldsymbol{\theta}\|^2 + C\sum_{i=1}^{m} \xi_i \\ \text{s.t. } y^{(i)}(\boldsymbol{\theta}^\intercal \mathbf{x}^{(i)} + b) \geq 1 - \xi_i,\, i = 1\dots m, \end{gathered} \tag{6}$$

[2] The $\boldsymbol{\theta}$ vector discussed in Section 2.1 includes parameter b as the zeroth order term. As a convention in SVM, parameter b is usually written explicitly in the expression

[3] Problem proposed in Eq. (5) cannot be solved by convex optimization algorithms.

where C is a hyperparameter[4], and ξ_i are slack variables. Algorithmically, the term $C\sum_{i=1}^{m}\xi_i$ adds a penalty on an otherwise strictly right-or-wrong classifier, resulting in a so-called soft-margin classifier. By selecting parameter C, one can control the tolerance for misclassification. As a rule of thumb for many machine learning models, this penalty term will avoid over-fitting due to the bias towards the outliers. If we further assume[5] $\boldsymbol{\theta}$ can be expressed as a linear combination of feature vectors $\mathbf{x}^{(i)}$ in the training set with proportionality constant α_i, i.e. $\boldsymbol{\theta} = \sum_{i=1}^{m}\alpha_i y^{(i)}\mathbf{x}^{(i)}$, the term $\frac{1}{2}\|\boldsymbol{\theta}\|^2$ in Eq. (6) becomes

$$\frac{1}{2}\|\boldsymbol{\theta}\|^2 = \frac{1}{2}\boldsymbol{\theta}^{\intercal}\boldsymbol{\theta} = \frac{1}{2}\sum_{i,j}^{m}\alpha_i\alpha_j y^{(i)}y^{(j)}\langle\mathbf{x}^{(i)},\mathbf{x}^{(j)}\rangle, \tag{7}$$

and the classifier $h_{\boldsymbol{\theta},b}$ becomes

$$h_{\boldsymbol{\theta},b}(\mathbf{x}) = \boldsymbol{\theta}^{\intercal}\mathbf{x} + b = \sum_{i}^{m}\alpha_i y^{(i)}\langle\mathbf{x}^{(i)},\mathbf{x}\rangle + b, \tag{8}$$

where the inner product $\langle\mathbf{x}^{(i)},\mathbf{x}^{(j)}\rangle := (\mathbf{x}^{(i)})^{\intercal}\mathbf{x}^{(j)}$. Surprisingly, all the main quantities in the algorithms can be written as an inner product between two vectors in the feature space. As we will see in the following discussion, changing this inner product into a kernel function enables us to capture more complex decision boundaries.

2.2.2 Kernel SVM

The above classification algorithm applies only to linearly separable data, while the formability of perovskite cannot be predicted by a linear classifier. If we are more ambitious and want to capture the non-linear decision boundaries, we need to expand the feature space into higher dimensions, in which the data are more easily classified. Fig.4(b) shows pictorially how a feature mapping from 2D to 3D can result a set of linear separable data. Consider a feature mapping [6] $\phi : \mathbf{x} \mapsto \phi(\mathbf{x})$, where $\mathbf{x} \in \mathbb{R}^{\mathrm{n}}$ and $\phi(\mathbf{x}) \in \mathbb{R}^{\mathrm{N}}, \mathrm{N} = 1\ldots\infty$. In order to capture more complex feature structures in high dimensions, instead of working with $\langle\mathbf{x}^{(i)},\mathbf{x}^{(j)}\rangle$ in Eq. (7) and (8), let us consider a generalized inner product

$$\kappa(\mathbf{x}^{(i)},\mathbf{x}^{(j)}) = K_{ij} := \langle\phi(\mathbf{x}^{(i)}),\phi(\mathbf{x}^{(j)})\rangle. \tag{9}$$

Here, κ is called the kernel function and $\boldsymbol{K}$ is the corresponding kernel matrix in terms of the finite training set. Compared with Eq. (8), the decision boundary for a kernel SVM is thus

$$h_{\boldsymbol{\theta},b}(\mathbf{x}) = \sum_{i}^{m}\alpha_i y^{(i)}\kappa(\mathbf{x}^{(i)},\mathbf{x}) + b. \tag{10}$$

Introducing the kernel functions will not only simplify working with features in higher dimensions, but also implicitly include our assumption on the data patterns. Many popular kernels used in SVM are:

$$\begin{aligned} \text{Linear:} \quad & \kappa(\mathbf{x},\mathbf{x}') = \mathbf{x}^{\intercal}\mathbf{x}' \\ \text{Polynomial:} \quad & \kappa(\mathbf{x},\mathbf{x}') = (1+\mathbf{x}^{\intercal}\mathbf{x}')^k \\ \text{Gaussian:} \quad & \kappa(\mathbf{x},\mathbf{x}') = \exp\left\{\frac{|\mathbf{x}-\mathbf{x}'|^2}{\sigma^2}\right\}. \end{aligned}$$

Applying the linear kernel to Eq. (10) will retrieve the linear SVM discussed in previous section. Gaussian kernel (also called Round Basis Function kernel) is so widely used that they are often set as the default in many SVM packages. It represents a feature mapping ϕ in the infinite dimensions, which cannot be captured by traditional methods. We will also use this kernel in the classifier.

[4] A hyperparameter is parameter that controls the behavior of the learning algorithm. Choosing such parameters can sometimes be empirical, but there are also ways to select such parameters to maximize the performance of the model. See cross validation.

[5] This is generally not an assumption but a theorem. See discussion on representer theorem.

[6] There are many ways to expand a feature space into high dimension. For example, if $\mathbf{x} = [x_1, x_2, x_3]^{\intercal}$, mapping ϕ can map $\mathbf{x}$ in a combinatorial fashion: $\phi(\mathbf{x}) = [x_1x_1, x_1x_2, x_1x_3 \ldots x_3x_3]$. Note that $\mathbf{x} \in \mathbb{R}^3$ and $\phi(\mathbf{x}) \in \mathbb{R}^9$

2.3 Feature Selection in Chemical Space

Feature ID	Symbol	Feature Description
0	r_A	Shannon's ionic radius of cation A
1	r_B	Shannon's ionic radius of cation B
2	r_X	Shannon's ionic radius of anion X
3	τ_o	Goldschmidt tolerance, $\frac{r_A+r_X}{\sqrt{2}(r_B+r_A)}$
4	o_f	Octahedral factor, $\frac{r_B}{r_X}$
5	t_f	Tetrahedral hole factor, $\frac{r_A}{r_X}$
6	r_{A-X}	Unit valence bond length for A-X bond
7	r_{B-X}	Unit valence bond length for B-X bond
8	$\frac{r_A^{s+p}}{r_X^{s+p}}$	Where r_A^{s+p} and r_X^{s+p} are the Zunger's pseudopotential core radii sum of the A and X atoms respectively
9	$\frac{r_B^{s+p}}{r_X^{s+p}}$	Where r_B^{s+p} and r_X^{s+p} are the Zunger's pseudopotential core radii sum of the B and X atoms respectively
10	$IE_1(A)$	The first ionization energy of cation A
11	$IE_1(B)$	The first ionization energy of cation B
12	$IE_2(B)$	The second ionization energy of cation B
13	τ_{rev}	Revised tolerance factor, $\frac{r_X}{r_B} - n_A(n_A - \frac{r_A/r_B}{\ln(r_A/r_B)})$ where n_A is the oxidation state of A
14	τ_t^B	Tetrahedral tolerance with cation B, $\frac{\sqrt{3}r_X}{\sqrt{2}(r_X+r_B)}$
15	τ_t^A	Tetrahedral tolerance with cation A, $\frac{\sqrt{3}r_X}{\sqrt{2}(r_X+r_A)}$

Table 1: Definition of the features and their corresponding symbols and numbering considered in the machine learning model. Further discussion is offered in Section 3.

For convenience in application, the features considered when constructing the predictive model should have readily available data from which the formability of novel perovskites can be made. In this study we consider 16 features in the chemical space, listed in Table 1. Classical studies [7] and prior machine learning studies [8] in predicting perovskite formability have focused on geometric factors — reasonable as the geometry of ionic packing is the primary factor in the thermodynamics of crystal packing. Hence, in concordance with the literature precedent set by Pilania *et al.* [8], we consider the effective radii of the the ions occupying the A, B and X sites (features 0, 1, 2), and parameters derived from these ionic radii: the Goldschmidt tolerance (feature 3) and octahedral factor (feature 4) as discussed in Section 1. The effective ionic radius of a given ion refers to the radius of the sphere that ion occupies in a crystal cell. These values are obtained statistically from crystallographic data, and well tabulated in the literature.[11]

In addition to ionic radii, which essentially considers ions as hard spheres in the crystal structure, Pilania *et al.* consider other quantitative descriptions of the space ions occupy, which we too have incorporated into our model. Namely, these are features 6 through 9. Features 6 and 7 are unit valence bond lengths for the A-X and B-X bonds, r_{A-X} and r_{B-X}. This is the theoretical bond length, in the bond valence formalism, between two atoms where each partner in the bond is bonded only to the other.[12] These values are calculated from crystallographic data and well compiled in the concerned databases. In calculating feature 8 and feature 9, r^{s+p} refers to the

Zunger's pseudopotential core radii sum. It refers to the core distance of the wave function where the pseudopotential crosses zero for s and p orbitals for a given angular momentum.[13] Essentially, Zunger's formalism classifies electron density around a nucleus into a core region and a valence region. r^{s+p}, then, corresponds to the spherical border between these two regions.[13] Values for r^{s+p} are determined by DFT, and some values are tabulated in the literature.

Recently, Bartel et al. presented a new data-driven tolerance descriptor based on Goldschmidt's tolerance factor defined more than 90 years earlier.[14] Specifically,

$$\tau_{rev} = \frac{r_X}{r_B} - n_A(n_A - \frac{r_A/r_B}{\ln(r_A/r_B)}), \tag{11}$$

where the oxidation state of A (n_A) is considered to ionic radii. Any ABX_3 combination that obtains $\tau_{rev} < 4.18$ is likely to be a stable perovskite. Bartel *et al.* found the ionic radii processed in this fashion can better estimate interatomic bond distances of the perovskite structure [14]. They experimented on increasing the dimensionality of the descriptor through 2D structure map by another machine learning method [15] and concluded that the best performing descriptor depends only on oxidation state and ionic radii. To compare the predictive strength of this revised tolerance and this latter assertion, we included τ_{rev} as feature 13.

In an effort to further current understanding of this field we sought also to consider several new features. To enrich the feature space beyond geometric parameters and to consider electronic parameters, we included the ionization energies of A and B as features. It should be noted that for the 181 perovskites studied the A cation was always in the +1 oxidation state and B in the +2 oxidation state. As such, we consider the first ionization energy of A (feature 10) and the first and second ionization energies of B (features 11 and 12).

It is worth noting that several non-perovskite ABX_3 crystal structures exist. Of these, subclasses of the wollastonite structure are among most often observed.[16] As shown in Fig. 5, in the crystal structure the metal cation is sequestered in a tetrahedral hole [17] formed by four adjacent anions rather than an octahedral hole in perovskite. We therefore reasoned that if past papers have found success in predicting formability by looking for ion radii that support the perovskite structure, it may be beneficial to the prediction of formability if we could eliminate ABX_3 formulations that favour alternate crystal structures. Therefore, we sought to develop the equivalence of a tolerance factor for an anion situated in a tetrahedral hole which would indicate a non-perovskite structure.

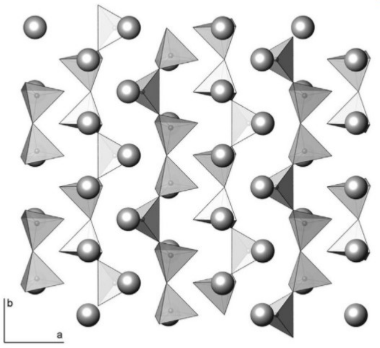


Figure 5: The wollastonite ABX_3 crystal structure. The vertices of the tetrahedron shown represent the centres of X anions and at the centre of which is trapped a metal cation. From [17].

Consider now the tetrahedral arrangement shown in Fig. 6 where anions, X, are at certain diagonal vertices of a cube and a cation, B, is at the center. The B-X bond length, l, can be determined by applying the Pythagorean theorem to the right triangle formed by half the face diagonal, $\frac{d\sqrt{2}}{2}$, and half the side length, $\frac{d}{2}$:

$$l = \sqrt{(\frac{d\sqrt{2}}{2})^2 + (\frac{d}{2})^2} = \frac{\sqrt{3}}{2}d. \tag{12}$$

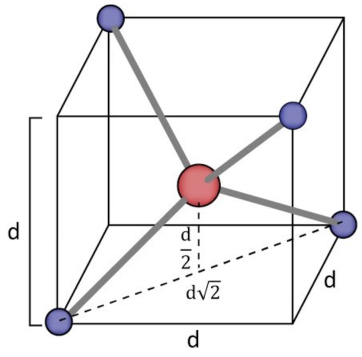


Figure 6: A representation of a cation (red) sequestered in a cavity formed from four anions (blue). Ionic radii are not drawn to scale for clarity.

Now, in the ideal, maximally packed crystal cell, the anions are such that they each contact the adjacent anions tangentially and the cation maximally fills the cavity such that it also tangentially contacts all four anions. In this configuration, the bond length is equal to the sum of the radii of X and B, and the diagonal is equal to twice the radius of X. Then, taking the fraction of l over the diagonal, $d\sqrt{2}$, using the expression for l from Eq. (12), gives:

$$\frac{l}{d\sqrt{2}} = \frac{\frac{\sqrt{3}}{2}d}{d\sqrt{2}} = \frac{r_X + r_B}{2r_X}. \tag{13}$$

Simplifying to the relationship:

$$\frac{\sqrt{3}}{\sqrt{2}} = \frac{r_X + r_B}{r_X}, \tag{14}$$

for the ideal tetrahedral cavity geometry. Therefore, we can assign a tetrahedral tolerance factor:

$$\tau_t^B = \frac{\sqrt{3}r_X}{\sqrt{2}(r_X + r_B)}. \tag{15}$$

Now Eq. (15) corresponds to cation B sequestered in the the tetrahedral cavity. However, since these non-perovskite structures are more diverse, cation A can also be sequestered in the tetrahedral cavity. Therefore, we additionally define:

$$\tau_t^A = \frac{\sqrt{3}r_X}{\sqrt{2}(r_X + r_A)}. \tag{16}$$

These last tetrahedral tolerance factors complete the feature space we consider in this study as features 14 and 15.

2.4 Model Design and Implementation

The total data set include a 181 pre-labelled points with 16 features proposed in the previous section. [15][18][14] 80% of the data form the training set and 20% form the test set. This split ratio will be applied to all feature combinations. As shown in Section 2.2.1, SVM is sensitive to the distance between feature vectors and the hyperplane. Considering that different features may take values in a wide range, we standardize each feature vector before training the model by

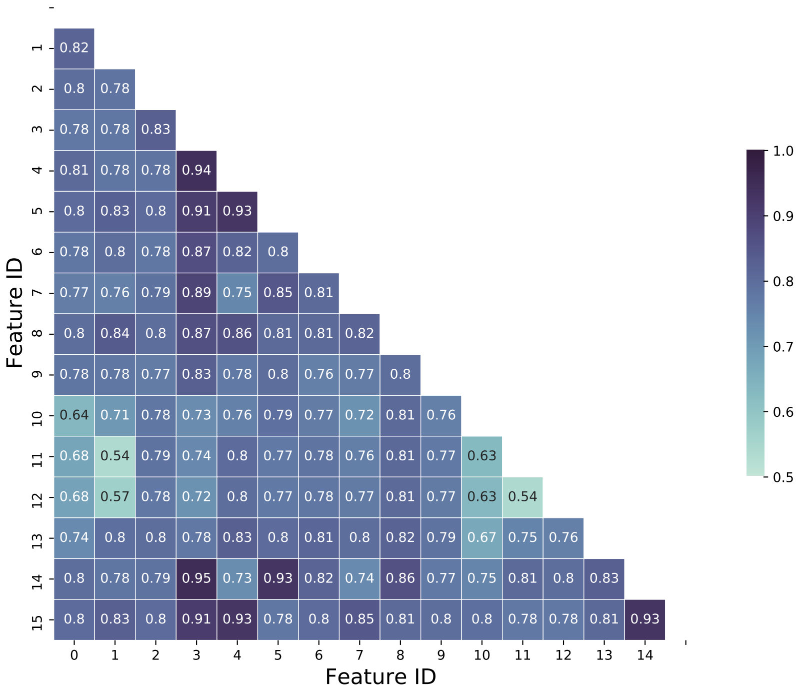
$$\tilde{\mathbf{x}} = \frac{\mathbf{x} - \mu}{\sigma},$$

where μ and σ are the mean and standard deviation of all the training data. In this study, we used the SVM module available in the scikit-learn [19] *Python* package *version 0.23.2* to generate classifier models described above. There are two hyperparameters as discussed in previous sections controls the performance of the algorithm: the soft-margin C and the kernel parameter σ. Determining those parameters can be empirical and sometimes non-trivial. As discussed in [20], the two parameters are selected as $C = 1$ and $\sigma^2 = 1/n$, where n is the dimension of the feature vector. This result is based on a grid-search cross validation[7]. With those parameters fixed, the model combines two and three features. The performance of those models are discussed in the next section.

3 Results and Discussions

In this study, we have built SVM models to take combinations of two features (120 groups), three features (560 groups) and four features (1820 groups). Each combination was repeated 100 times with randomly selected 80% training set and 20% test set compositions. The accuracy of different combinations was estimated from the averages of those 100 runs on the test set. Results show that features 3,4,14,15 i.e. the octahedral factor, the Goldschmidt tolerance factor, and the tetrahedral tolerance factor of cation A and B are the most influential features to model performance. In the following sections, we will first discuss how different numbers of features can affect the model performance with the focus on the two-feature combinations in Section 3.1 and then explain the possible chemical interpretations behind those best-performing models in Section 3.2. Section 3.3 will focus on some possible explanations of the misclassified compounds. As we will see, those points failing to be captured by the model arise mainly because of their unique chemical properties rather than a flaw of the algorithm. We will end with the limitations of our model in 3.4 and some further improvements for future study.

3.1 Implication of Feature Space Dimensions



Feature ID	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1	0.82														
2	0.8	0.78													
3	0.78	0.78	0.83												
4	0.81	0.78	0.78	0.94											
5	0.8	0.83	0.8	0.91	0.93										
6	0.78	0.8	0.78	0.87	0.82	0.8									
7	0.77	0.76	0.79	0.89	0.75	0.85	0.81								
8	0.8	0.84	0.8	0.87	0.86	0.81	0.81	0.82							
9	0.78	0.78	0.77	0.83	0.78	0.8	0.76	0.77	0.8						
10	0.64	0.71	0.78	0.73	0.76	0.79	0.77	0.72	0.81	0.76					
11	0.68	0.54	0.79	0.74	0.8	0.77	0.78	0.76	0.81	0.77	0.63				
12	0.68	0.57	0.78	0.72	0.8	0.77	0.78	0.77	0.81	0.77	0.63	0.54			
13	0.74	0.8	0.8	0.78	0.83	0.8	0.81	0.8	0.82	0.79	0.67	0.75	0.76		
14	0.8	0.78	0.79	0.95	0.73	0.93	0.82	0.74	0.86	0.77	0.75	0.81	0.8	0.83	
15	0.8	0.83	0.8	0.91	0.93	0.78	0.8	0.85	0.81	0.8	0.8	0.78	0.78	0.81	0.93

Figure 7: Accuracy of two-feature combinations, with each combination averaged over 100 runs. The corresponding feature ID can be found in Table 1.

[7] When doing the cross validation, the total data set will be split into three sets: training, development, and test set. The development is specifically for determining the hyperparameters. The parameter that minimize the development set error will be selected

We start with considering different combinations of two features. Fig. 7 shows a table of accuracy of two feature combination with 100 runs. The prediction accuracy is defined as the ratio of the correctly predicted data to the total data in the test set. Notice that a high prediction accuracy of 95 % is obtained from the model constructed by feature pair (3,14), which corresponds to the Goldschmidt octahedral tolerance τ_o and tetrahedral tolerance τ_s^B with cation B. Then, we observe that (3,4), (4,5), (4,15), and (14,15) also construct highly accurate models. The results indicate that the Goldschmidt tolerance factor (3), the octahedral factor (4), the tetrahedral hole factor (5), tetrahedral tolerance with cation B (14) and with cation A (15) are more relevant to predicting the formability of the perovskite structure. This statement is also supported in later discovery of models combining three and four features. Also we notice that features coming from either element A or B alone are not correlated with the formability of perovskite structures. For example, the ionization energies of B, and its combination with ionic radius of B (1) yield nearly random predictions (54%). One thing to point out is that the ionic energy was standardized to optimize the performance of the model. The values of those energies are in the order of 10^2, which is significantly greater than other features. As SVM is sensitive to the distance between feature vectors, the standardization is necessary to keep each features in a similar order of magnitude.

With the discovery from the results of the two-feature combinations, we applied the same calculation to construct models using groups of three and four features. The results of this calculation show that the general results of prediction performance has increased (see supporting materials), with a significant number of models having 90% and above accuracy. For higher dimensions, most of the feature groups demonstrate a relatively high accuracy above 80% and no model's accuracy is below 66%.

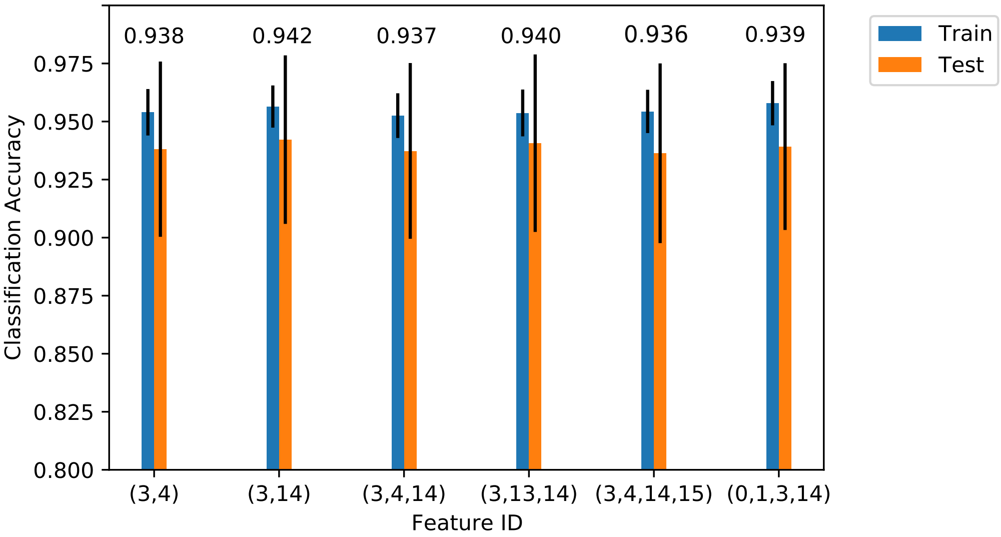
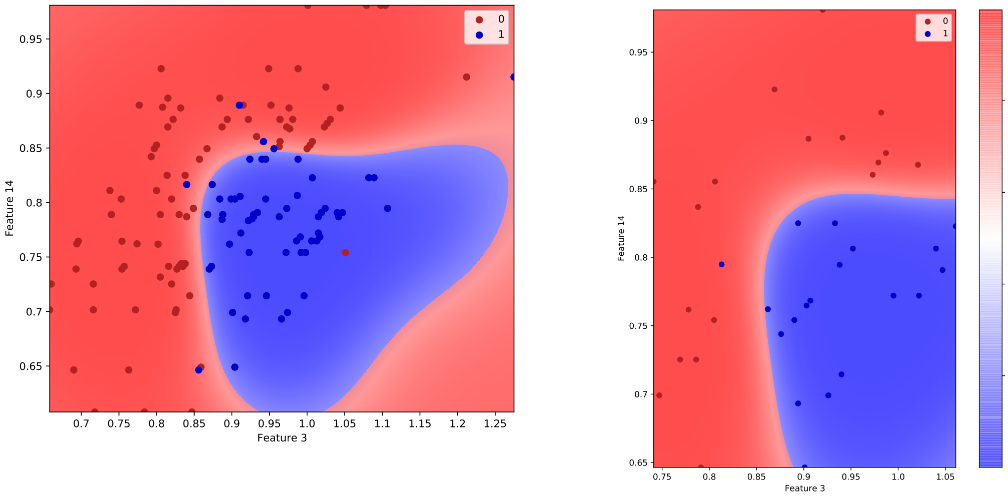


Figure 8: Best performing feature combinations from two, three, and four-membered group with 5000 runs ranked by their prediction accuracy on the randomly selected 80% of the training set and 20% of the test set. Error bars represent the standard deviation of accuracy within those runs.

Fig. 8, shows the comparison of the best performing models from two, three, and four-membered groups. Interestingly, adding more features into the model does not necessarily increase the maximum performance. The model with the highest averaged accuracy is still constructed by (3,14). Addition of a third or fourth feature does not influence the result significantly. This trend in groups of three and four features is expected. We observe that models combining two of feature 3, 4, 5, 14, and 15 generally give highly accurate prediction. Thus, these features might be determining to the performance of a model, and models combining three or four of these features might also yield high accuracy. Here we notice that the best performing three and four feature groups, combinations involving feature 3, 4, 14, 15, construct the best performing models.This supports that the Goldschmidt tolerance factor (3), the octahedral factor (4), the tetrahedral hole factor (5), tetrahedral tolerance factor with cation B (14) are important factors to determine the formability of perovskite structures. In addition, we observe that the ionic radii, r_A (feature 0) and r_B (feature 1) construct a model with accuracy of 93.9%, combining Goldschmidt tolerance factor (3) and tetrahedral tolerance factor (14). This suggests that in formability prediction, the steric effects of A and B element combined with the geometry packing effects are paramount.

3.2 Best performed model and improvements



(a) **(b)**

Figure 9: The prediction contour for both (a) training set and (b) test set with feature pairs (τ_o, τ_t^B). The color bar labels the probability of forming the perovskite, with 1 (blue) to be the most likely and 0 (red) the least likely. The axes has the original units for both parameters, as we inversely transformed the standardized data to the actual feature space after constructing the model.

As discussed Section 3.1, the Goldschmidt octahedral tolerance τ_o and tetrahedral tolerance τ_s^B with cation B stood out with a prediction accuracy of 95%. Figure 9 shows the prediction contour generated by the SVM model along with all the training and test data. Red (non-perovskite) and blue (perovskite), and the saturation of color relates to the probability of prediction. The darker the color is, the more confident the prediction is. The non-linear decision boundary is perfectly captured by the Gaussian kernel. The model allows some misclassification in either the perovskite or non-perovskite regions due to the soft-margin constant C. Note that the boundary curve is a closed loop whose shape is constraint by the support vectors. This prediction encodes physical meanings as both factors should lie in a finite instead of an open region.

Even from a cursory examination of Fig. 7, it becomes apparent that feature combinations including feature 3, τ_o, are generally noticeably better than other combinations. Examining Fig. 9, and imagining the data points collapsed onto the 'Feature 3' axis alone, it becomes apparent why. As well documented by past studies, τ_o values proximal to 1 predict that a perovskite will form.[21][7][22] The vast majority of perovskites occur in the range $0.85 < \tau_o < 1.10$. However, a significant number of non-perovskites also occur in this range. Using τ_o alone to predict formability would generate a large number a false positives. This necessitates some second feature to be considered to resolve these false-negatives out.

In past studies, this second feature considered has been the octahedral factor, o_f. Through Fig. 9 we demonstrate that the tetrahedral tolerance factor with cation B, τ_t^B, that we define in Section 2.3, can effectively carry out the elimination of false positives that occur when only Goldschmidt's τ_o is used. Specifically, these false positives occur when $\tau_t^B > 0.85$. This confirms that τ_t^B is predicting formability in a manner consistent with our expectations discussed in Section 2.3: structures with τ_t^B proximal to 1 are non-perovskite structures, since $\tau_t^B = 1$ corresponds to the ideal tetrahedral cavity geometry. It is also interesting to note that the range of τ_o values which predict a perovskite structure, $0.85 < \tau_o < 1.10$, is near identical to the range where τ_t^B eliminates false positives by identifying tetrahedral cavity formation, $\tau_t^B > 0.85$. Interestingly though, returning to Fig. 7, τ_t^B is not necessarily much better as a predicting feature when combined with most other features, further suggesting that τ_t^B alone is not notable, but functions as a discriminator of false positives for features like τ_o. Another feature of interest is that those ABX_3 structures that fall within the range of values for both ideal perovskite formation and tetrahedral cavity formation are preferentially non-perovskite and presumably tetrahedral. This seems to indicate that when conditions allow both, formation of the tetrahedral cavity structure is more favourable than the desired perovskite structure. Further examination of the crystal structures that fall into this description may prove interesting. In sum, though pairing τ_t^B with Goldschmidt's τ_o yields marginal improvements to model accuracy, the success of this model demonstrates that perovskite formability can be predicted both positively, by identifying conditions that form the ideal perovskite, and negatively, by identifying conditions that favour the formation of an alternative geometry.

3.3 Interpretation of the Outliers

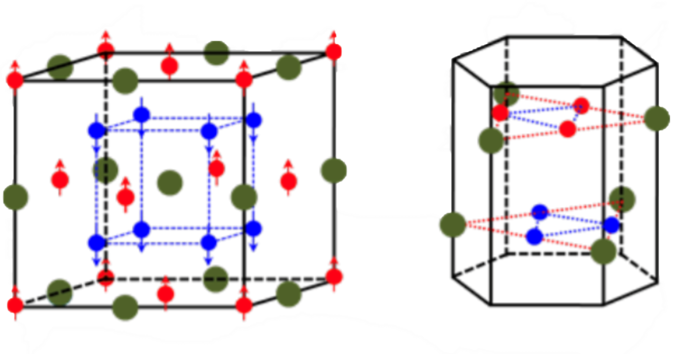


Figure 10: Cubic crystal structure (Left), Hexagonal crystal structure (Right). Hexagonal structure can be transformed to cubic structure under high pressure and vice versa. From [23].

.

We looked into a few misclassified compounds in our model. There are in total six perovskite structures (RbF–PbF_2, CsF–BeF_2, KCl–$FeCl_2$, TlI–MnI_2, RbI–SnI_2, TlI–PbI_2) classified as non-perovskite and only one non-perovskite system (CsF–MnF_2) is misclassified as a perovskite. Most of these discrepancies occur near the decision boundary, which means that their predicted probability is close to 50%. As for the false negative predictions shown in Fig. 9 (perovskite classified as non-perovskite), two compounds have lower octahedral tolerance, three have larger tetrahedral factor and one has both factors larger than the prediction boundary. Additionally, some of the outliers have special chemical properties which data driven methods struggle to capture. For example, compound TlI–MnI_2, while being a perovskite, has slightly higher tetrahedral tolerance τ_s^B (0.88) than most of the typical structures. This may arise from the large polarizability of the I^- ions. As the I^- is attracted by the two cations, the electron cloud will be distorted into an ellipsoid shape, causing an increase of the radius r_X. From the definition of the tetrahedral tolerance, we see that the distortion can increase the tetrahedral factor to some extent.

The only false positive classification which occurred inside the domain of perovskite is CsF–MnF_2. CsF–MnF_2 has a hexagonal crystal structure.[18] If the relative size of the cation at A-site is too large, then the ABX_3 crystal cell form a hexagonal system that can only transforms to the perovskite structure at elevated pressures. [24] In other words, there is possibility for this hexagonal structure to be transformed to a cubic perovskite structure under high pressure and high temperature conditions as shown in Fig. 10 ABX_3 usually crystallizes in a series of closely related structures, if A cation has a radius range from 1.0 to 1.9 and the B cation has a radius range between 0.5 and 1.2. [24] Particularly for CsF–MnF_2, the cesium cation ion at the A site has a radius of 1.67 and the manganese cation at the B site has a radius of 0.83. The cation of CsF–MnF_2 has an optimal size for hexagonal-cubic structure transformation. There may exist a possibility for hexagonal CsF–MnF_2 to be transformed to cubic perovskite structure under high pressure and temperature. Collectively looking at the misclassifications, there may be some new routes for synthesizing the special perovskites alternatively. [25]

3.4 Limitations and Future Studies

One of the disadvantages of data-driven methods like SVM is that the small set of data makes the prediction rather biased and eventually loose its prediction ability under some special conditions. Like many other machine learning methods applied in physical science, the model itself is hard to explain by theories, despite having a high prediction accuracy. However, it does allow predicting expected properties with a small subset of available data, thus reducing the workload of synthesis processes. Considering many experiment are expensive and time-consuming, using machine learning as a general guide for directing synthetic process will be beneficial to much of chemistry research. For example, double perovskites are another emerging class of frontier material that can be potentially discovered by data-driven approaches. Compared with the traditional perovskite, the B-site cation in the double perovskite has two different ions and forms a structure of $A_2BB'X_6$. As one might expected, this more complex structure results in more possible combinations of material designs, making simple trial-and-error experimental routes impossible. Unlike a total of 640 combinations of the single perovskite halide, there are over $500,000$ potential double perovskite halides. [14] As the double perovskite enlarges the chemical space of the perovskite materials, it may be promising to discover novel semiconductor alternatives.

4 Conclusion

We constructed machine learning models using support vector machine classifiers to determine the formability of a given ABX_3 formulation. The models were trained and tested using two, three, or four theoretically calculated features of chemical composition, as well as their experimental data of formability. The results reveal that the ionic radii of A and B, the Goldschmidt tolerance factor, the octahedral factor, tetrahedral tolerance factor with cation A and cation B are determining features, implying a significance of packing geometry and ionic steric effects on determining the formability of a perovskite structure. Analysis of the best forming model shows that using the formation of the counter geometry as a false signal equivalently contributes to the accuracy of the model. With the constructed models, we can efficiently determine the formability of perovskite structures based on theoretically calculated electronic and spacial parameters, improve the understanding of the configurational and compositional effects and therefore accelerate the discovery of solar energy conversion materials with desired properties.

References

[1] S. Almosni, A. Delamarre, Z. Jehl, D. Suchet, L. Cojocaru, M. Giteau, M. Behaghel, A. Julian, C. Ibrahim, L. Tatry, H. Wang, T. Kubo, S. Uchida, H. Segawa, N. Miyashita, R. Tamaki, Y. Shoji, K. Yoshida, H. Ahsan, T. Hamamura, T. Toupance, C. Olivier, S. Chambon, L. Vignau, C. Geffroy, E. Cloutet, G. Hadziioannou, N. Cavassilas, P. Rale, A. Cattoni, S. Collin, F. Gibelli, M. Paire, L. Lombez, D. Aureau, M. Bouttemy, A. Etcheberry, Y. Okada, J.-F. Guillemoles, *Sci. Technol. Adv. Mater.* **2018**, *19*, 336–369.

[2] V. M. Goldschmidt, *Naturwissenschaften* **1926**, *14*, 477–485.

[3] Y. Chen, L. Zhang, Y. Zhang, H. Gao, H. Yan, *RSC Adv.* **2018**, *8*, 10489–10508.

[4] S. De Wolf, J. Holovsky, S.-J. Moon, P. Löper, B. Niesen, M. Ledinsky, F.-J. Haug, J.-H. Yum, C. Ballif, *J. Phys. Chem. Lett.* **2014**, *5*, 1035–1039.

[5] S. D. Stranks, G. E. Eperon, G. Grancini, C. Menelaou, M. J. P. Alcocer, T. Leijten, *Science* **2013**, *342*, 341–344.

[6] M. R. Filip, F. Giustio, *Proc. Natl. Acad. Sci. U. S. A.* **2018**, *115*, 5397–5402.

[7] L. Feng, L. Jiang, M. Zhu, H. Liu, X. Zhou, C. Li, *Journal of Physics and Chemistry of Solids* **2008**, *69*, 967–974.

[8] G. Pilania, P. V. Balachandran, C. Kim, T. Lookman, *Frontiers in Materials* **2016**, *3*, 19.

[9] J. Shawe-Taylor, N. Cristianini, et al., *Kernel methods for pattern analysis*, Cambridge university press, **2004**.

[10] J. VanderPlas, *Python data science handbook: Essential tools for working with data*, " O'Reilly Media, Inc.", **2016**.

[11] R. D. Shannon, *Acta Cryst.* **1976**, *32*, 751–767.

[12] I. D. Brown, *Chem. Soc. Rev.* **1978**, *7*, 359–376.

[13] A. Zunger, M. L. Cohen, *Physical Review B* **1979**, *20*, 4082.

[14] C. J. Bartel, C. Sutton, B. R. Goldsmith, R. Ouyang, C. B. Musgrave, L. M. Ghiringhelli, M. Scheffler, *Science Advances* **2019**, *5*.

[15] R. Ouyang, S. Curtarolo, E. Ahmetcik, M. Scheffler, L. M. Ghiringhelli, *Phys. Rev. Materials* **2018**, *2*, 083802.

[16] R. S. Roth, *J. REs. NNatl. Inst. Stannd. Technol.* **1956**, *58*, 75–88.

[17] S. J. Edrees, M. M. Shukur, M. M. Obeid, *Comput. Condens. Matter* **2018**, *14*, 20–26.

[18] G. S. Rohrer, *Structure and bonding in crystalline materials*, Cambridge University Press, **2001**.

[19] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al., *the Journal of machine Learning research* **2011**, *12*, 2825–2830.

[20] G. Pilania, P. V. Balachandran, C. Kim, T. Lookman, *Frontiers in Materials* **2016**, *3*, 19.

[21] H. Zhang, N. Li, K. Li, D. Xue, *Acta Crystallographica Section B* **2007**, *63*, 812–818.

[22] A. Kumar, A. Verma, S. Bhardwaj, *The Open Applied Physics Journal* **2008**, *1*, 11–19.

[23] D. Zhang, B. Yan, S.-C. Wu, J. Kübler, G. Kreiner, S. S. P. Parkin, C. Felser, *Journal of Physics: Condensed Matter* **2013**, *25*, 206006.

[24] J. Kafalas, J. Longo, *Journal of Solid State Chemistry* **1972**, *4*, 55–59.

[25] P. V. Balachandran, J. Theiler, J. M. Rondinelli, T. Lookman, *Scientific reports* **2015**, *5*, 13285.