PROCEEDINGS OF THE ROYAL SOCIETY A

Proc. R. Soc. A (2011) **467**, 2271–2290
doi:10.1098/rspa.2010.0543
Published online 2 March 2011

Identifying the 'inorganic gene' for high-temperature piezoelectric perovskites through statistical learning

By Prasanna V. Balachandran, Scott R. Broderick and Krishna Rajan*

Department of Materials Science and Engineering and Institute for Combinatorial Discovery, Iowa State University, Ames, IA 50011, USA

This paper develops a statistical learning approach to identify potentially new high-temperature ferroelectric piezoelectric perovskite compounds. Unlike most computational studies on crystal chemistry, where the starting point is some form of electronic structure calculation, we use a data-driven approach to initiate our search. This is accomplished by identifying patterns of behaviour between discrete scalar descriptors associated with crystal and electronic structure and the reported Curie temperature (T_C) of known compounds; extracting design rules that govern critical structure–property relationships; and discovering in a quantitative fashion the exact role of these materials descriptors. Our approach applies linear manifold methods for data dimensionality reduction to discover the dominant descriptors governing structure–property correlations (the 'genes') and Shannon entropy metrics coupled to recursive partitioning methods to quantitatively assess the specific combination of descriptors that govern the link between crystal chemistry and T_C (their 'sequencing'). We use this information to develop predictive models that can suggest new structure/chemistries and/or properties. In this manner, $BiTmO_3$–$PbTiO_3$ and $BiLuO_3$–$PbTiO_3$ are predicted to have a T_C of 730°C and 705°C, respectively. A quantitative structure–property relationship model similar to those used in biology and drug discovery not only predicts our new chemistries but also validates published reports.

Keywords: inorganic gene; high-temperature piezoelectrics; statistical learning; information theory; data-driven modelling

1. Introduction

Through many seminal papers, Alan McKay has expounded on the idea of a framework for 'Generalized Crystallography' (Mackay 1966, 1974, 1977, 1986). He has proposed that 'the crystal is a structure, the description of which is much

*Author for correspondence (krajan@iastate.edu).

Electronic supplementary material is available at http://dx.doi.org/10.1098/rspa.2010.0543 or via http://rspa.royalsocietypublishing.org.

Received 21 October 2010
Accepted 17 January 2011

smaller than the structure itself' and that this description of structure serves as a 'carrier of information' about the structure on larger length scales (MacKay 2002). He went on to suggest that these components of description of structure can help develop a 'biological approach to inorganic systems' and proposed the construction of an 'inorganic gene'. This paradigm serves as motivation underlying the present study by exploring how fundamental pieces of information, treated as discrete bits of data, can collectively characterize the stability and properties of a given crystal chemistry. We show how the use of statistical learning tools including fundamental concepts borrowed from information theory can be used to characterize a crystal structure in terms of fundamental descriptors of information (i.e. the 'genes') and how these pieces of information interact or are 'sequenced' to guide the characteristics of that crystal structure and in fact help to guide the development of new crystal chemistries and targeted physical properties.

The challenge in defining the 'gene' in inorganic crystal chemistry is to characterize the appropriate combination of discrete characteristics associated with crystal chemistry that collectively define a particular property or set of properties of the material. Normally, structure–property relationships are guided by defined functional relationships (e.g. electronic structure calculations to define energy landscapes associated with crystal chemistry). However, we propose an approach to establish such a structure–property relationship where we do not assume any specific formulation linking structure with property (Jóhannesson *et al.* 2002; Curtarolo *et al.* 2003; Woodley *et al.* 2004; Dudiy & Zunger 2006; Fischer *et al.* 2006; Sluiter 2007; Mohn & Kob 2009; Oganov & Valle 2009). Rather, we take a data-driven approach where we seek to establish structure–property relationships by identifying patterns of behaviour between known discrete scalar descriptors associated with crystal and electronic structure and observed properties of the material. From this, we extract design rules that allow us to systematically identify critical structure–property relationships, resulting in identifying in a quantitative fashion the exact role of specific combination of materials descriptors (i.e. genes) that govern a given property. This is the foundation of the concept of the quantitative structure–activity (or property) relationship (QSAR/QSPR) widely used in the field of organic chemistry and drug discovery. The mathematical underpinning of developing a QSPR-type relationship is statistical learning (a term encompassing a broad range of tools derived from statistics, data mining and machine learning). In our group, we have applied this approach to explore a variety of questions associated with crystal chemistry (Suh & Rajan 2005, 2009; Gadzuric *et al.* 2006; Rajagopalan & Rajan 2007; George *et al.* 2009; Broderick *et al.* 2010; Rajan 2010, Zenasni *et al.* 2010), and in this paper, we demonstrate that by using the QSPR concept, we can identify through the tools of statistical inference, how discrete bits of information that define a robust QSPR relationship can be sequenced to help identify new materials with new and targeted properties. The specific objective of the present study is identifying, through the sole use of statistical learning methods, new high-temperature piezoelectric ferroelectrics. However, this paper also serves as a generic template for an information science-based materials discovery and design strategy, in the spirit of Mackay's proposition of an inorganic gene.

2. Background

(*a*) *Materials chemistry of high-temperature piezoelectrics*

Historically, the design of materials chemistry for high-temperature piezoelectric behaviour has been guided by an apparent linear relationship between Goldschmidt's tolerance factor (t) and Curie temperature (T_C) at the morphotropic phase boundary (MPB) composition of the $PbTiO_3$ (PT)-based end-member solid solutions (Eitel *et al.* 2001; Duan *et al.* 2004). However, the use of the tolerance factor as a 'figure of merit' has had limited impact in developing or identifying new materials via experiment (Eitel *et al.* 2001; Duan *et al.* 2004) or computation (Baettig *et al.* 2005), owing to the fact that it captures only a very limited set of variables (i.e. ionic radii) describing a given perovskite crystal chemistry (Thomas 1997). The motivation of our work is to find alternative computational based methods that can help to refine the chemical search space and identify potentially new and promising piezoelectric materials for high-temperature applications.

The chemical search space of known and predicted perovskite-based ferroelectric compounds in $BiMeO_3$–$PbTiO_3$ solid solution is mapped in figure 1, where Me is a single cation with charge 3+ or a combination of two different cations ($Me_{1/2}Me_{1/2}$, $Me_{2/3}Me_{1/3}$ and $Me_{3/4}Me_{1/4}$) with an average charge 3+, occupying the octahedral site of the perovskite lattice (Eitel *et al.* 2001; Grinberg *et al.* 2005; Suchomel & Davies 2005; Stein *et al.* 2006; Grinberg & Rappe 2007). The solid solutions were classified based on the chemical origin of ferroelectric instability caused by Me cations. The distinction between strong (filled red circles) and weak (filled green squares) ferroelectric activity was made based on the degree of off-centring tendency of Me cations in MeO_6 octahedra. Clearly, the search space is sparse in the high-temperature region, and our goal is to explore the vast combinatorial search space and identify new high-temperature piezoelectric chemistries. In this work, we have focused primarily on identifying a new Me^{3+} cation that satisfies the following conditions:

— it must show weak ferroelectric activity;
— $BiMeO_3$ must have a stable perovskite structure at ambient or non-ambient (high-pressure/-temperature) conditions; and
— the resulting $BiMeO_3$–$PbTiO_3$ solid solution should have a high T_C.

We explore a data-driven methodology that involves applying statistical learning tools to analyse correlations between numerous scalar descriptors of electronic and crystal structure parameters of known perovskite piezoelectric compounds and using that information in turn to develop predictive models that can suggest new structure/chemistries and/or properties based purely on the formalism of statistical learning methods. This methodology is quite different from the approach that is widely reported by many groups where large numbers of high-throughput electronic structure computations are conducted to seek compound chemistries with energy minima (where data mining-related techniques are embedded in the computation to help the efficiency of the calculations);

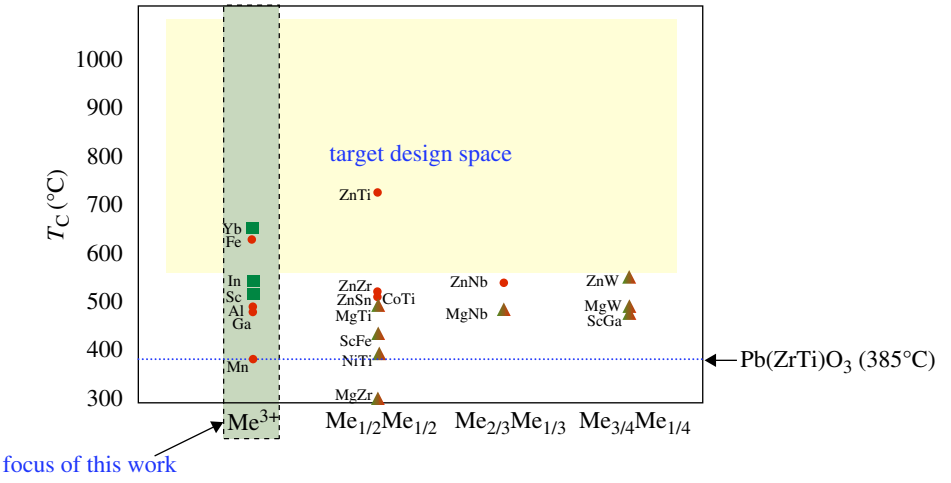

Figure 1. In this figure, we map the Curie temperature (T_C) of known and predicted perovskite-based ferroelectric compounds in the chemical space of $BiMeO_3$–$PbTiO_3$ solid solution, where Me is a single cation with charge 3+ (e.g. Al, Sc, In, etc.) or a combination of two different cations $Me_{1/2}Me_{1/2}$ (e.g. ZnTi, ZnZr, ZnSn, etc.), $Me_{2/3}Me_{1/3}$ (e.g. ZnNb, MgNb) and $Me_{3/4}Me_{1/4}$ (e.g. ZnW, MgW, ScGa) with an average charge 3+ and that occupies the octahedral site of the perovskite lattice (Eitel *et al.* 2001; Grinberg *et al.* 2005; Suchomel & Davies 2005; Stein *et al.* 2006; Grinberg & Rappe 2007). The target design space represents the high-temperature regime that is of interest to us, and, as it can be clearly seen, the chemical search space is sparse in this region with as many as only three compounds being identified. For reference, T_C of $PbZrO_3$–$PbTiO_3$ solid solution is also indicated in this figure. Our objective is to systematically explore the complex chemical search space and identify potentially new piezoelectric materials that have high T_C. In this article, we report our computational work, where we have focused particularly on identifying a suitable Me^{3+} cation (which is weakly ferroelectrically active and occupies the octahedral site of the perovskite lattice) that can significantly enhance the T_C of $BiMeO_3$–$PbTiO_3$ solid solution. The distinction between strong and weak ferroelectric activity was made based on the degree of off-centring tendency of Me cations in MeO_6 octahedra. Filled circles, Me cations that show strong ferroelectric activity; filled squares, Me cations that show weak ferroelectric activity; filled triangles, Me cations that show strong and weak ferroelectric activity. (Online version in colour.)

and then potentially new stable compounds are identified by identifying those that have energy minima but not reported in known experimental databases (Jóhannesson *et al.* 2002; Curtarolo *et al.* 2003; Woodley *et al.* 2004; Dudiy & Zunger 2006; Fischer *et al.* 2006; Sluiter 2007; Mohn & Kob 2009; Oganov & Valle 2009).

Our approach requires the need to carefully establish a dataset of descriptors on which we directly apply statistical learning tools. The number of parameters needed to predict even relatively simple structures can be large if one has to capture both geometrical and bonding characteristics of that crystal chemistry. One of the arguments we are trying to put forward in this paper is that although the potential number of variables can in fact be large, data dimensionality reduction and information theoretic techniques can help reduce

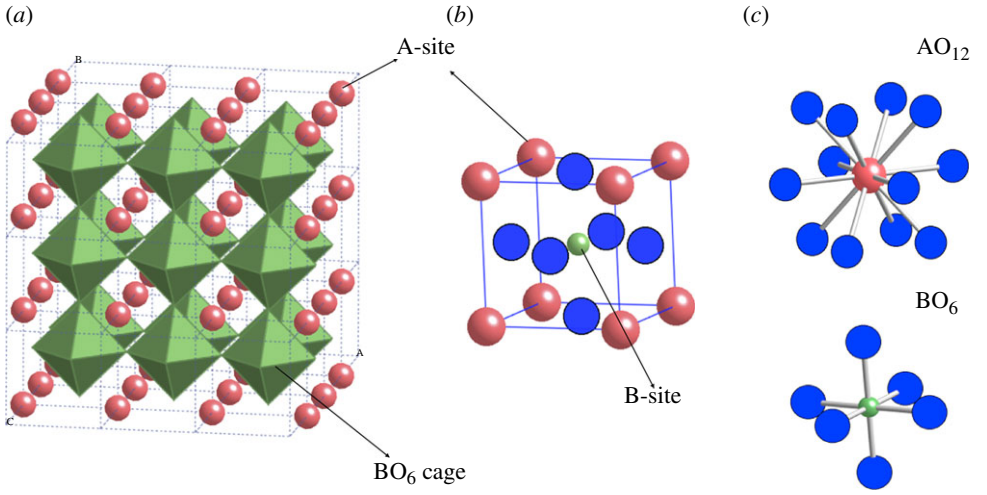

Figure 2. (*a*) A network of corner-sharing BO_6 octahedra with a large A-site cation occupying the interstitial position is shown. (*b*) The simplified unit-cell representation of cubic perovskite without showing coordination. (*c*) The geometry of the building units, AO_{12} cuboctahedra and BO_6 octahedra, with 12-coordinated A-site and 6-coordinated B-site, respectively. The description of the crystal structure in the form of structural building units presents a number of diverse choices to develop new descriptors based on the site chemistry and coordination. (Online version in colour.)

it to a manageable number. This paper describes a data mining strategy from which effective classification and predictive models can be developed using high-dimensional information.

(*b*) *Defining the chemical search space*

The search for new high-temperature piezoelectric materials by chemical modification of $PbTiO_3$ perovskite at both Pb and Ti sites has been an area of considerable interest in the last decade (Sághi-Szabó *et al.* 1998; Eitel *et al.* 2001). While there are many crystal structures that may be suitable for high-temperature piezoelectric application, such as perovskites, langasites (Damjanovic 1998) and perovskite-like layered structures (Yan *et al.* 2009), we are interested in perovskites because they have the best combination of high temperature and piezoelectric properties compared with other structures, and many perovskites are also ferroelectrics, which can be used as piezoelectric materials when poled (Cohen 2008; Rödel *et al.* 2009). The crystal structure of an ideal perovskite crystal is shown in figure 2. Following the discovery of the crucial role of Bi in enhancing the ferroelectric properties in $PbTiO_3$ (Íñiguez *et al.* 2003), numerous experimental and theoretical studies focusing on $BiMeO_3$–$PbTiO_3$ solid solutions were carried out (where Me represents a single cation with charge 3+ or a combination of cations with an average charge 3+) with the further objective of identifying a potential Me cation that can maximize both Curie temperature and ferroelectric properties of the solid solution (Suchomel & Davies 2004, 2005; Grinberg *et al.* 2005; Stein *et al.* 2006; Stringer *et al.* 2006;

Chen *et al.* 2007, 2009; Grinberg & Rappe 2007). The key findings from the earlier studies are summarized below:

— Enhancement of ferroelectric properties and Curie temperature owing to the presence of strongly ferroelectrically active Me cations (e.g. Ti^{4+}, Zn^{2+}, Fe^{3+}, etc.). These strongly ferroelectrically active Me cations cause hybridization of Me–O bonds in MeO_6 octahedra, leading to distortions resulting in significant ionic displacement from the ideal position (Cohen 1992, 2008; Rödel *et al.* 2009). The ionic displacements were responsible for enhanced polarization and ferroelectric properties. Some examples of compounds with strongly ferroelectrically active Me cations are $BiFeO_3$–$PbTiO_3$ and $Bi(ZnTi)O_3$–$PbTiO_3$.
— On the other hand, it was found that the presence of weakly ferroelectrically active Me cations (e.g. Sc^{3+}, Mg^{2+} and Yb^{3+}) can also enhance the high-temperature ferroelectric properties. In this case, the Me cations do not lead to hybridization of Me–O bonds, whereas the steric effect causes the Pb/Bi cation to avoid the larger Me/Ti cation owing to the larger wave-function overlap (therefore stronger Pauli repulsion) and move towards the smaller cation. The stronger repulsion leads to increased Pb/Bi cation displacement, which in turn results in enhanced ferroelectric behaviour (Grinberg *et al.* 2005). Some examples of compounds with weakly ferroelectrically active Me cations are $BiScO_3$–$PbTiO_3$ and $BiYbO_3$–$PbTiO_3$.

Our chemical search space is defined in electronic supplementary material, figure S1, and we have focused particularly on identifying a suitable $BiMeO_3$ perovskite end member, where Me is a single cation that is weakly ferroelectrically active with a formal charge 3+ and that can form a solid solution with $PbTiO_3$ at ambient conditions.

3. Statistical learning computational strategy

(*a*) *Introduction to tolerance factor–T_C model*

Eitel *et al.* (2001) first discovered the existence of an apparent linear relationship between tolerance factor of ABO_3 end-member compositions and Curie temperature at MPB for a large number of ABO_3–$PbTiO_3$ solid solutions, although there was some significant scatter (figure 3). Grinberg *et al.* (2005) later addressed this scatter by identifying that the data fall into two clusters, and they showed that both clusters exhibited a linear dependence of Curie temperature on the end-member tolerance factor but had different slopes. The physical reasons behind the two slopes were correlated to the differences in the ferroelectric activity of various B-site cations of the ABO_3 end-member compositions. While both models can be applied to quantitatively predict the T_C, neither predicts the perovskite phase stability of the ABO_3–$PbTiO_3$ solid solution. This is a major shortcoming because only those ABO_3–$PbTiO_3$ solid solutions that form a pure perovskite phase at ambient conditions are technologically useful (Grinberg *et al.* 2005).

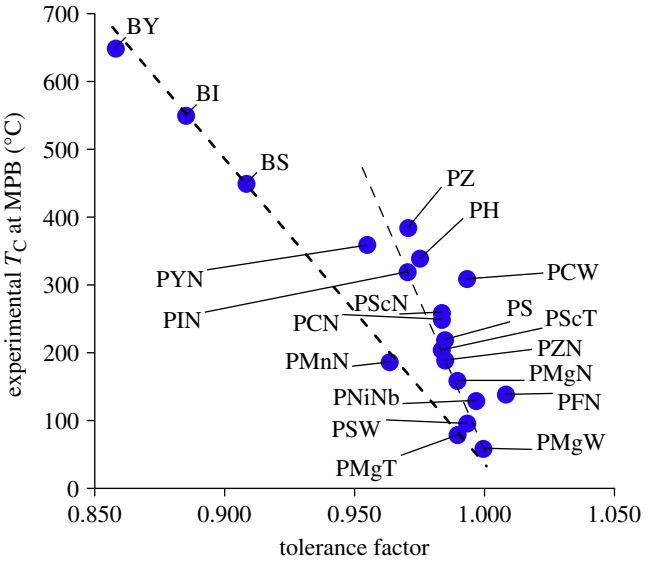

Figure 3. The univariate tolerance factor–T_C model of Eitel *et al.* (2001) is shown here. The shortcomings of the univariate tolerance factor–T_C model are clearly noticeable as the data show significant scatter owing to the presence of two clusters of compounds with different physics. This indicates that the tolerance factor is only a necessary condition and not sufficient for modelling T_C. We have addressed the shortcomings of the tolerance factor–T_C model by developing a multivariate model that considers six key crystal chemical descriptors instead of only the tolerance factor. Notation for chemical compounds and parameters are described in the electronic supplementary material. (Online version in colour.)

We have collectively addressed the above-mentioned shortcomings of the tolerance factor–T_C model in a couple of ways. Firstly, by considering additional crystal chemical descriptors, a reasonably accurate multivariate model was developed (described in §4*b*) using linear manifold methods for quantitatively predicting the T_C at MPB of ABO_3–$PbTiO_3$ solid solutions. To reduce the scatter, instead of including all ferroelectric ABO_3–$PbTiO_3$ chemistries that contain both strongly and weakly ferroelectrically active cations, we have typically considered end members that belong to $Pb(B_1B_2)O_3$ and $BiMeO_3$ perovskites, where B_1, B_2 and Me are cations that occupy the octahedral site of the perovskite lattice and Me cation is weakly ferroelectrically active. By clearly defining our chemical search space in this manner, we focus on the relevant physics that best describes our objective.

Secondly, in order to determine the perovskite phase stability of the ABO_3–$PbTiO_3$ solid solution, we have developed an independent classification model based on information theory concepts (e.g. Shannon entropy) that tracks which combination of parameters influences the perovskite structural stability by partitioning a high-dimensional dataset. As noted by Karnani *et al.* (2009), natural data structures, such as genomes, books, file systems and data servers, are repositories of information that share common characteristics. Also, they display skewed distributions and hierarchical organization, which certainly applies to

crystallographic data. The physical representation of information allows us to understand that these ubiquitous characteristics are consequences of the second law. Thus, by combining the linear manifold methods with the information theory concepts, we can identify new high-temperature piezoelectric materials.

(*b*) *Informatics-based computational strategy*

Our computational logic for designing new high-temperature piezoelectric chemistries is summarized in the form of a flow chart in the electronic supplementary material, figure S2. The logic involves three steps. (i) Identification of a relevant descriptor set that fully describes the high-temperature behaviour of ABO_3 perovskites. Thirty attributes were screened using principal component analysis (PCA) and a reduced set of six key attributes was identified that showed high correlation with the transition temperature. (ii) Development of a robust multivariate model using partial least squares (PLS) that predicts T_C at MPB of ABO_3–$PbTiO_3$ solid solutions. By applying the PLS model, new candidate chemistries were identified that are suitable for high-temperature applications. (iii) Screening for the piezoelectric behaviour in the new candidate chemistries by testing the perovskite structural stability of ABO_3 end members. For this purpose, new classification models were developed using a recursive partitioning strategy. The outcome of this analysis is important for determining whether it is possible to synthesize a pure perovskite phase in the ABO_3–$PbTiO_3$ solid solution. Only those ABO_3 end members that were classified to have a stable perovskite structure-type by recursive partitioning were chosen and identified as potential high-temperature piezoelectric materials. The mathematics of PCA, PLS and recursive partitioning in the context of our specific datasets is summarized in the electronic supplementary material.

Before elaborating on the data mining methods, we need to address the obvious concern that at first glance the statistical learning methods do not in themselves explicitly solve the energy minimization problem that the physics-based calculations do. However, this concern is addressed collectively in a couple of ways. The first is that we are searching for a high-dimensional correlation between attributes of compounds that already exist and hence are by definition stable. In fact, a corollary to this point is that mathematically we are using convex optimization methods that help to ensure we have a global minimum (Izenman 2008). Second, we test the validity of our models with respect to the target materials properties (i.e. Curie temperature in this case) by using well-established and robust methods for being able to reproduce the known data, to give us the statistical confidence of the models we develop.

4. Results and discussion

(*a*) *Identifying the relevant descriptor set: the inorganic genes*

As noted above, the tolerance factor as the sole figure of merit to design new high-temperature piezoelectric perovskite compounds appears to be insufficient. To look beyond the tolerance factor to predict new high-temperature piezoelectric materials, we have surveyed over 30 different attributes (table 1) associated with crystal geometry, bonding, thermodynamics and electronic structure of 22 simple

Table 1. Enumeration of 30 descriptors used in the principal component analysis (PCA) for identifying the relevant inorganic gene is given in this table. The underlying rationale behind choosing these different attributes associated with crystal geometry, bonding, thermodynamics and electronic structure was to fully describe the crystal chemistry of perovskite-based compounds that is relevant for modelling the ferroelectric behaviour, and the search was motivated by the past experimental and theoretical work of Abrahams *et al.* (1968), Igarashi *et al.* (1987), Singh *et al.* (1988), Ravez *et al.* (1997), Goudochnikov & Bell (2007) and Grinberg & Rappe (2007).

abbreviation	description
r_A(Å)	Shannon's (1976) ionic radii of A-site (12-coordination)
r_B(Å)	Shannon's ionic radii of B-site (6-coordination)
t	tolerance factor calculated using ionic radii
d_{A-O}(Å)	ideal A–O bond distance (Brese & O'Keeffe 1991)
d_{B-O}(Å)	ideal B–O bond distance
t_{BV}	tolerance factor calculated using d_{A-O} and d_{B-O}
A_{EA}(kJ mol^{-1})	A-site electron affinity (Hotop & Lineberger 1985)
A_{EFF-S}	A-site effective nuclear charge—Slater scale (Slater 1930)
A_{EFF-C}	A-site effective nuclear charge—Clementi scale (Clementi & Raimondi 1963)
A_{EFF-F}	A-site effective nuclear charge—Froese-Fisher scale (Froese-Fischer 1972)
B_{EFF-S}	B-site effective nuclear charge—Slater scale
B_{EFF-C}	B-site effective nuclear charge—Clementi scale
B_{EFF-F}	B-site effective nuclear charge—Froese-Fisher scale
A_{WS}(Å)	A-site Wigner–Seitz cell radius (Skriver 2004)
B_{WS}(Å)	B-site Wigner–Seitz cell radius
A_{EN-P}	A-site electronegativity—Pauling scale (Pauling 1960)
A_{EN-AR}	A-site electronegativity—Allred–Rochow scale (Allred & Rochow 1958)
A_{EN}(eV)	A-site electronegativity—absolute scale (Pearson 1988)
B_{EN-P}	B-site electronegativity—Pauling scale
B_{EN-AR}	B-site electronegativity—Allred–Rochow scale
B_{EN}(eV)	B-site electronegativity—absolute scale
D_A(Å)	ionic displacement (Grinberg & Rappe 2007) of A-site
D_B(Å)	ionic displacement of B-site
ΔH^f_{AO}(J mol^{-1})	enthalpy of formation (Saxena 1993) of A oxide
ΔH^f_{BO}(J mol^{-1})	enthalpy of formation of B oxide
ΔH^f_{ABO3}(J mol^{-1})	enthalpy of formation of ABO_3
a(Å)	lattice constant (Matsui & Nomura 1981)
b(Å)	lattice constant
c(Å)	lattice constant
V/Z(Å^3)	volume of unit cell/coordination number
T_t(K)	transition temperature

ABO_3 perovskite chemistries with known transition temperatures (Shannon 1976; Matsui & Nomura 1981; Saxena 1993; Emsley 1998; Brown 2002; Suh & Rajan 2005; Goudochnikov & Bell 2007; Grinberg & Rappe 2007; Makov *et al.* 2009; Pettersson *et al.* 2009; Rajan 2010). The transition temperature of an ABO_3 compound is defined as the temperature when the crystal structure of ABO_3 changes from low symmetry to the highest possible symmetry. While not all of the ABO_3 compounds assessed are ferroelectric, the objective of this work is unaffected, since the final goal is to suggest new perovskite-based end members

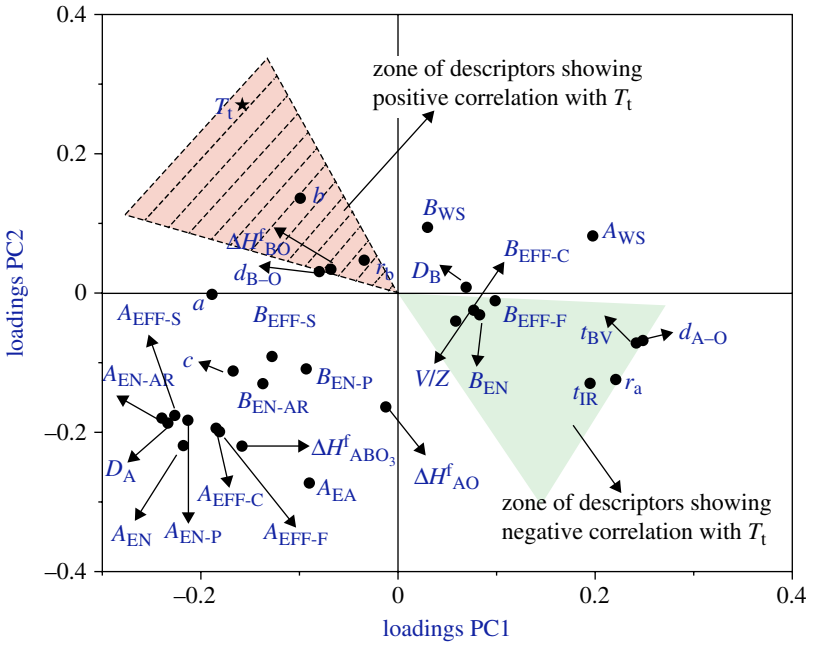

Figure 4. Loadings plot between PC1 and PC2 showing the interactions of 30 descriptors captured by PCA. Based on the angle θ, the degree of correlation between the target variable and other attributes is established. Two zones are marked in the figure that show a strong correlation with the target variable (T_t): the red zone (with stripes) signifies attributes that show positive correlation with T_t and the green zone (no stripes) signifies variables that show negative correlation with T_t. The abbreviations of the attributes are provided in table 1. (Online version in colour.)

forming solid solutions with PT. Alloying an ABO_3 perovskite compound with $PbTiO_3$ has the potential to lead to a high piezoelectric characteristic in the resulting ABO_3–$PbTiO_3$ ceramic (Grinberg & Rappe 2004).

To identify the complex relationships between physical properties and crystal chemistry and geometry from the existing knowledge base, PCA is employed (Ericksson *et al.* 2001; Rajan 2005; Ringnér 2008). The input $X = \{x_1, x_2, x_3, \ldots, x_n\} \in \mathrm{Re}^{n\times d}$ (where $n = 22$ and $d = 30$ denote the number of ABO_3 compounds and the number of physical attributes quantifying each ABO_3 compound, respectively) is initially preprocessed by mean-centring and standardization. PCA reduces the dimensionality of the data by identifying new latent variables (called principal components, PCs) that capture the largest amount of variation in the data. Each PC is a linear combination of the weighted contribution of each attribute. By comparing the magnitude and direction of the weighted contribution from each attribute, the correlation structure in the high-dimensional data is discovered).

Figure 4 (referred to as a loading plot) shows the uncovered correlations between the physical attributes for the first two PCs. The transition temperature (T_t) is the target variable against which all correlations are computed. As we are using linear manifold methods, we have employed Euclidean geometrical mapping to help interpret these plots. The degree of correlation between any attribute and

T_t is determined by the cosine of the angle (θ) between the attribute and T_t (angle between attribute origin–T_t) within the loading plot. If $\theta = 0°$, the attribute and T_t are highly positively correlated, if $\theta = 180°$, then they are highly negatively correlated and if $\theta = 90°$, there is no correlation between the attribute and T_t. In figure 4, two zones that show the strongest correlation of the attributes with T_t are explicitly marked, with the assumption that the first two PCs capture such a high percentage of the data's information that the other PCs do not need to be explicitly considered. The attributes r_B (ionic radii of B-site), $d_{B–O}$ (ideal B–O bond distance based on the bond-valence model), ΔH_{fBO} (enthalpy of formation of BO oxide) and b (lattice constant) correlate positively with T_t, while r_A (ionic radii of A-site), $d_{A–O}$ (ideal A–O bond distance based on the bond-valence model), t (tolerance factor calculated using ionic radii), t_{BV} (tolerance factor calculated using the bond-valence method), B_{EN} (B-site electronegativity—absolute scale), B_{Eff} (B-site effective nuclear charge) and V/Z (volume of unit cell/coordination number) correlate negatively with T_t. Our PCA model reproduces the well-known inverse linear relationship between tolerance factor (t) and T_t. Based on the removal of redundancy and consideration of available data, we have determined that six attributes (r_A, t, B_{EN}, $d_{A–O}$, r_B and $d_{B–O}$) are appropriate for describing T_t. By identifying these attributes, we can more fully describe the high-temperature behaviour than possible by only considering the tolerance factor (t), and the selection of only the highly correlated attributes ensures the robustness of the model.

(*b*) *Identifying new high-temperature perovskites: developing a 'QSPR'*

To test for high-T_C piezoelectric materials, we have applied PLS regression (Ericksson *et al.* 2001) to predict T_C at the MPB of the end-member $PbTiO_3$ solid solution. PLS is particularly suitable for handling sparse data with strongly correlated attributes. The piezoelectric materials database for predicting T_C as a function of six attributes (r_A, t, B_{EN}, $d_{A–O}$, r_B and $d_{B–O}$) is taken from the published work of Eitel *et al.* (2001) and Grinberg *et al.* (2005). This new QSPR formulated using PLS is given by

$$T_C = -(789.912 \times t) - (153.932 \times r_A) + (1013.981 \times r_B) + (796.5864 \times d_{B–O}) - (138.9 \times d_{A–O}) - (55.6076 \times B_{EN}) - 526.537.$$

Fifteen compounds were used for training the model and an independent set of five compounds (not used during the training) was used for testing (figure 5). Our QSPR model takes into account the physics of mismatch of bond lengths (t), ionic size (r_A and r_B), bond lengths ($d_{A–O}$ and $d_{B–O}$) and chemical bonding at the B-site (B_{EN}), thereby accounting for a far greater diversity of attributes in comparison to the previous model where only mismatch of bond lengths was considered. Some of the descriptors captured in our QSPR model are also in the original description of the tolerance factor. However, only two (r_B and r_A) of the six descriptors are explicitly used in the tolerance factor formulation,

$$\left(t = \frac{r_A + r_O}{\sqrt{2}(r_B + r_O)} \right)$$

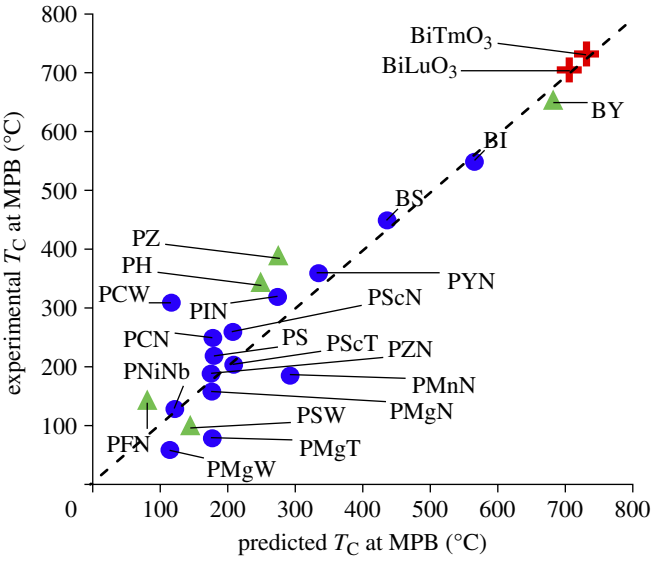

Figure 5. Multivariate predicted model (abscissa) in comparison with the measured T_C as reported in the literature (Eitel *et al.* 2001; Grinberg *et al.* 2005) is shown for the $PbTiO_3$ end members. The model was developed by using 15 chemistries and tested for five chemistries. The new figure of merit is $T_C = -(789.912 \times t) - (153.932 \times r_A) + (1013.981 \times r_B) + (796.5864 \times d_{B-O}) - (138.9 \times d_{A-O}) - (55.6076 \times B_{EN}) - 526.537$. Based on the new figure of merit, the T_C of new piezoelectric chemistries $BiTmO_3$–PT and $BiLuO_3$–PT were predicted to be 730°C and 705°C, respectively (labelled red in the figure). It should be noted that the T_C of $BiTmO_3$–PT and $BiLuO_3$–PT plotted in the figure is only the predicted value and needs to be experimentally validated. Notation for chemical compounds and parameters are described in the electronic supplementary material. Filled circles, training set; filled triangles, test set; plus symbols, new predictions. (Online version in colour.)

while the other four descriptors are not explicitly used. For end members that had more than one cation in the octahedral site, such as $Pb(B_1B_2)O_3$, we considered the arithmetic mean value of B_1 and B_2. It should be noted, although not elaborated in this paper, that the classification of Me ions into weakly and strongly ferroelectric active species can be accomplished by exploring more descriptors such as polarizability, ionic valence and ionic size.

The additional diversity of the QSPR model has a clear advantage as compared with the model based solely on tolerance factor. For many compounds, the QSPR model is in reasonable agreement with the tolerance factor model. However, in some cases, the mismatch of bond length is not sufficient for modelling the physics of the system. For the systems predicted here, $BiLuO_3$–$PbTiO_3$ is predicted to have a higher T_C than any systems included in the training dataset; however, this result is not found when using the tolerance factor model. Therefore, we conclude that our developed QSPR is highly robust in predicting the T_C of unknown compounds (figure 5) and has a more broad significance when applied to new materials. Based on this QSPR model, a search of all the elements in the periodic table that best satisfy the correlation criterion involving the combination of attributes was performed. The search has resulted in generating four new

ABO_3 chemistries ($BiTmO_3$, $BiLuO_3$, $BiHoO_3$ and $BiErO_3$) as potential high-T_C materials. Having identified the new chemistries, we then tested them for their crystal structure-type.

(*c*) *Screening for piezoelectric behaviour: 'sequencing the gene'*

To test for the perovskite structural stability, a new classification model was developed using a recursive partitioning strategy (Witten & Frank 2000; Hall *et al.* 2009) on a large database (taken from the work of Zhang *et al.* 2007 and references therein) of 355 ABO_3 stoichiometric compounds (227 perovskites and 128 non-perovskites) to track which combination of parameters influences the perovskite structural stability by partitioning a high-dimensional dataset. The outcome of this analysis is important for determining whether it is feasible to synthesize a pure perovskite phase in the $BiBO_3$–$PbTiO_3$ solid solution (where B = Tm, Lu, Ho, Er). Our hypothesis is, if $BiTmO_3$, $BiLuO_3$, $BiHoO_3$ and $BiErO_3$ compounds are predicted to have a stable perovskite structure-type at ambient or non-ambient (high pressure/temperature) condition, then we propose that it is possible to experimentally obtain a pure perovskite phase in $BiBO_3$–$PbTiO_3$ solid solution (where B = Tm, Lu, Ho, Er). Here, we explain the relevance of this hypothesis using a few examples based on experimental observations.

It is well known that obtaining a pure Bi-based perovskite is difficult under conventional processing methods at ambient conditions. For example, a pure perovskite phase in $BiScO_3$ is synthesized only at 6 GPa pressure and 1140°C temperature (Belik *et al.* 2006*a*,*b*) and in $BiMnO_3$ a pure perovskite phase is obtained only at pressures greater than 4 GPa and 750°C temperature (Montanari *et al.* 2005). However, solid solutions of $BiScO_3$–$PbTiO_3$ (Zhang *et al.* 2003) and $BiMnO_3$–$PbTiO_3$ (Woodward & Reaney 2004) have been experimentally synthesized and are shown to have a pure perovskite phase. Even in the case of very low tolerance factor end members such as $BiYbO_3$ (tolerance factor = 0.857), there are experimental reports that confirm the limited solubility of $BiYbO_3$ in $PbTiO_3$. Feng *et al.* (2009) using conventional ceramic processing methods synthesized a solid solution of $0.05BiYbO_3$–$0.95PbTiO_3$ with the highest perovskite phase purity of 97.83 per cent. Obtaining a pure perovskite phase in $BiYbO_3$ when synthesized at ambient conditions is extremely difficult (Drache *et al.* 2004), and we note that there is no experimental or theoretical study on structural phase transitions in $BiYbO_3$ at high-pressure/-temperature conditions. In this work, we have identified for the first time the existence of a stable perovskite structure-type in $BiYbO_3$ via a recursive partitioning strategy at high-pressure/-temperature conditions, and this structural stability at high-pressure/-temperature conditions explains the limited solubility of $BiYbO_3$ in $PbTiO_3$ at ambient conditions. Alloying $BiYbO_3$ with $PbTiO_3$, which has a large c/a ratio, can help stabilize a perovskite phase by applying chemical pressure (Ahart *et al.* 2008).

In this work, we apply our classification model to qualitatively determine the feasibility of synthesizing a pure perovskite phase in the $BiBO_3$–$PbTiO_3$ solid solution (where B = Tm, Lu, Ho, Er). In order to capture the physics of perovskite stability at high-pressure/-temperature conditions, we have included ABO_3 perovskite compounds such as $BiScO_3$ (Belik *et al.* 2006*a*,*b*), $BiMnO_3$ (Montanari *et al.* 2005), $BiAlO_3$ (Belik *et al.* 2006*a*,*b*), $NaSbO_3$ (Mizoguchi *et al.* 2004) and

$YInO_3$ (Shannon 1967) that are experimentally known to have a stable perovskite structure-type only at extreme pressure/temperature conditions. Therefore, the design rules that we extract from our classification model are applicable to identify new perovskites at both ambient and high-pressure/-temperature conditions. Using the Shannon entropy as a selection criterion, a hierarchical set of design rules was formulated to develop classification schemes that hitherto have been approached by empirical observation (Plenio & Vitelli 2001; Shell 2008; Karnani *et al.* 2009).

The expected information required to classify an ABO_3 compound solely based on its proportion in the database D is given by the Shannon entropy $H(D)$, which is defined as

$$H(D) = -\sum_{i=1}^{m} p_i \log_2(p_i),$$

where p_i is the probability that an arbitrary tuple in 'D' belongs to perovskite crystal structure or not. A log function of base 2 is used, because the information is encoded in bits and m is an integer with distinct values defining m distinct classes (Han & Kamber 2006). We formulated our recursive partitioning as a binary classification problem. Further details on the construction and interpretation of the dendrogram are provided in the electronic supplementary material.

The aim of the classification is to track precisely which and how variables contribute to perovskite structural stability. The output from a recursive partitioning analysis is a dendrogram (or a tree diagram) with branches grown on each node (attribute) to classify whether a particular ABO_3 compound forms a perovskite crystal structure. The advantage of the recursive partitioning method is that it can efficiently model nonlinear relationships in any arbitrary form even when the attributes show strong interactions (Hawkins *et al.* 1997). Our recursive partitioning model classified 336 out of 355 compounds accurately (95% accuracy), and the model was validated by a standard 10-fold cross-validation technique used in statistics.

The dendrogram model used for predicting new perovskites is shown in figure 6. According to the dendrogram, $d_{A–O}$ (ideal A–O bond length calculated based on the bond-valence method) is the most significant attribute impacting the phase stability of perovskite compounds, followed by the tolerance factor. The leaf nodes that are labelled 'yes' and 'no' indicate compounds that may have a stable perovskite structure-type or not a perovskite, respectively. From the dendrogram, design rules were extracted for predicting new potentially stable perovskite compounds. Of the 227 perovskite compounds, 184 obeyed the following rule: if $d_{A–O} > 2.453$ and $t_{IR} \leq 1.090863$ and $r_A/r_B > 1.509872$ and $B_{EN}–O_{EN} > 1.42$ and $r_A/r_B \leq 2.5625$, then the ABO_3 compound is a perovskite, where $d_{A–O}$ is the ideal bond length based on the bond-valence model, t_{IR} is the tolerance factor calculated using ionic radii, r_A/r_B is the ionic radii ratio of A-site to B-site and $B_{EN}–O_{EN}$ is the electronegativity difference (Pauling scale) between B-cation and O-anion. A total of 11 design rules were formulated for testing the perovskite structural stability.

By applying the dendrogram to the four candidate ABO_3 compounds, only two compounds, $BiTmO_3$ and $BiLuO_3$, were identified as having a stable perovskite crystal structure at high-pressure/-temperature conditions.

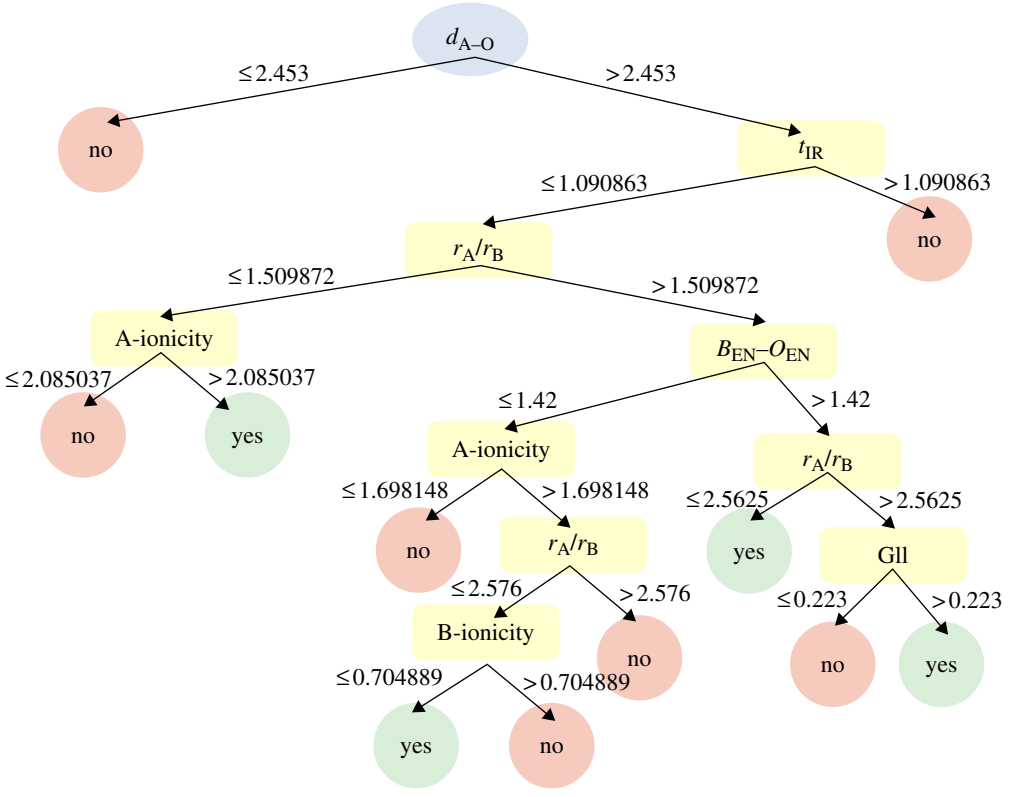

Figure 6. The dendrogram (or tree diagram) classification model developed based on the recursive partitioning method for identifying new potentially stable perovskite compounds is shown. We used the Shannon entropy as a selection criterion to identify key descriptors, and a hierarchical set of design rules were formulated to develop classification schemes that have been approached by empirical observation. The leaf nodes that are labelled 'yes' or 'no' indicate compounds that may have a stable perovskite structure-type or not a perovskite, respectively. From the dendrogram, 11 design rules were formulated for testing the perovskite structural stability. By applying the dendrogram to the four candidate high-temperature materials $BiErO_3$, $BiHoO_3$, $BiTmO_3$ and $BiLuO_3$, only two compounds, $BiTmO_3$ and $BiLuO_3$, were identified as having the stable perovskite crystal structure at high-pressure/-temperature conditions. As a result, $BiTmO_3$–$PbTiO_3$ and $BiLuO_3$–$PbTiO_3$ solid solutions were identified as new perovskite compounds with a significantly high T_C while having piezoelectric behaviour. The dendrogram application of other Bi-based systems $BiMEO_3$, where ME = Cr, Co, Ga and Ni, also identifies them as having the perovskite crystal structure in agreement with the literature (Ishiwata *et al.* 2002; Baettig *et al.* 2005; Goujon *et al.* 2008; Oka *et al.* 2010). In the dendrogram, d_{A-O} is the ideal A–O bond length calculated based on the bond-valence method, t_{IR} is the tolerance factor from ionic radii data, r_A is ionic radii (Shannon's scale) of the A-site cation with coordination number 12, r_B is the ionic radii (Shannon's scale) of the B-site cation with coordination number 6, B_{EN}–O_{EN} is the electronegativity difference (Pauling's scale) between B-site and O-site, A-ionicity is the product of r_A/r_O and A_{EN}–O_{EN}, B-ionicity is the product of r_B/r_O and B_{EN}–O_{EN} and GII is the global stability index (Zhang *et al.* 2007). (Online version in colour.)

Experimental synthesis of $BiTmO_3$ and $BiLuO_3$ compounds at ambient pressure has been attempted in the past but was unsuccessful in synthesizing a pure perovskite phase (Drache *et al.* 2005); however, there are no data available on

synthesizing $BiTmO_3$ and $BiLuO_3$ compounds at high-pressure/-temperature conditions. Therefore, we predict for the first time the existence of a stable perovskite phase in $BiTmO_3$ and $BiLuO_3$ compounds at high-pressure/-temperature conditions. This result indicates that Tm^{3+} (thulium) is the largest cation (with an ionic radius of 0.88 Å in sixfold coordination) that can occupy the octahedral site of a $BiMeO_3$ perovskite lattice without impacting its phase stability. The dendrogram also predicts the existence of a stable perovskite phase in $BiYbO_3$ at high-pressure/-temperature conditions. $BiYbO_3$–$PbTiO_3$ is known as a potential high-temperature piezoelectric material (Eitel *et al.* 2001; Feng *et al.* 2009), and there are experimental reports that confirm the limited solubility of $BiYbO_3$ in $PbTiO_3$, thereby forming a solid solution (Feng *et al.* 2009). Thus, we conclude that it is possible to experimentally obtain a pure perovskite phase in $BiLuO_3$–$PbTiO_3$ and $BiTmO_3$–$PbTiO_3$ solid solutions. Based on the QSPR and the recursive partitioning model, two new perovskite end members were identified ($BiTmO_3$–$PbTiO_3$ and $BiLuO_3$–$PbTiO_3$) and predicted to have a high T_C of 730°C and 705°C at the MPB, respectively, while having piezoelectric behaviour.

The focus of this report has been solely on identifying new $BiMeO_3$–$PbTiO_3$ materials chemistries with higher Curie temperatures, where Me is a weakly ferroelectrically active cation with a formal charge 3+. We fully realize that other electronic structure parameters such as polarizability and other microstructural parameters play a critical role in defining a useful high-temperature piezoelectric material. This involves exploring a larger and more diverse chemical space that includes more than one Me cation that is strongly ferroelectrically active, which is presently being done, as well as experimental verification of our results, which will be reported in upcoming publications.

5. Summary

We have identified two new perovskite-based piezoelectric crystal chemistries, $BiTmO_3$–$PbTiO_3$ and $BiLuO_3$–$PbTiO_3$, with significantly higher Curie temperature using a highly efficient and robust computational strategy based on statistical learning and information theory concepts. The data mining strategy we have developed also permits us to identify key physical attributes that appear to govern the properties of a given crystal chemistry (e.g. piezoelectrics with a high Curie temperature), providing a mechanistic-based discovery process and not just a heuristic strategy. Finally, this paper helps to establish the efficacy of informatics as an approach to refine the chemical search space for materials discovery and to hence serve as a broader template for materials design in other applications.

The authors acknowledge support from the Air Force Office of Scientific Research, grant nos FA9550-06-10501 and FA9550-08-1-0316; the National Science Foundation: NSF-IMI programme grant no. DMR-08-33853, NSF-ARI Program: CMMI 09-389018; NSF-CDI Type II program: grant no. PHY 09-41576 and NSF-AF grant no. CCF09-17202, DARPA N/MEMS Science & Technology Fundamentals program, grant no. HR0011-06-0049, Dr D. L. Polla, Program Manager, and Army Research Office grant no. W911NF-10-0397. KR would also like to acknowledge support from Iowa State University through the Wilkinson Professorship in Interdisciplinary Engineering.

References

Abrahams, S. C., Kurtz, S. K. & Jamieson, P. B. 1968 Atomic displacement relationship to Curie temperature and spontaneous polarization in displacive ferroelectrics. *Phys. Rev.* **172**, 551–553. (doi:10.1103/PhysRev.172.551)

Ahart, M. *et al.* 2008 Origin of morphotropic phase boundaries in ferroelectrics. *Nature* **451**, 545–548. (doi:10.1038/nature06459)

Allred, A. L. & Rochow, E. G. 1958 A scale of electronegativity based on electrostatic force. *J. Inorg. Nucl. Chem.* **5**, 264–268. (doi:10.1016/0022-1902(58)80003-2)

Baettig, P., Schelle, C. F., LeSar, R., Waghmare, U. V. & Spaldin, N. 2005 Theoretical prediction of new high-performance lead-free piezoelectrics. *Chem. Mater.* **17**, 1376–1380. (doi:10.1021/cm0480418)

Belik, A. A. *et al.* 2006*a* $BiScO_3$: centrosymmetric $BiMnO_3$-type oxide. *J. Am. Chem. Soc.* **128**, 706–707. (doi:10.1021/ja057574u)

Belik, A. A., Wuernisha, T., Kamiyama, T., Mori, K., Maie, M., Nagai, T., Matsui, Y. T. & Takayama-Muromachi, E. 2006*b* High-pressure synthesis, crystal structures, and properties of perovskite-like $BiAlO_3$ and pyroxene-like $BiGaO_3$. *Chem. Mater.* **18**, 133–139. (doi:10.1021/cm052020b)

Brese, N. E. & O'Keeffe, M. 1991 Bond-valence parameters for solids. *Acta Cryst. B* **47**, 192–197. (doi:10.1107/S0108768190011041)

Broderick, S. R., Nowers, J. R., Narasimhan, B. & Rajan, K. 2010 Tracking chemical processing pathways in combinatorial polymer libraries via data mining. *J. Comb. Chem.* **12**, 270–277. (doi:10.1021/cc900145d)

Brown, I. D. 2002 *The chemical bond in inorganic chemistry: the bond valence model.* Oxford, UK: Oxford University Press.

Chen, J., Hu, P., Sun, X., Sun, C. & Xing, X. 2007 High spontaneous polarization in $PbTiO_3$–$BiMeO_3$ systems with enhanced tetragonality. *Appl. Phys. Lett.* **91**, 171–907. (doi:10.1063/1.2794742)

Chen, J., Tan, X., Jo, X. & Rödel, J. 2009 Temperature dependence of piezoelectric properties of high-T_C $Bi(Mg_{1/2}Ti_{1/2})O_3$–$PbTiO_3$. *J. Appl. Phys.* **106**, 034109. (doi:10.1063/1.3191666)

Clementi, E. & Raimondi, D. L. 1963 Atomic screening constants from SCF functions. *J. Chem. Phys.* **38**, 2686. (doi:10.1063/1.1733573)

Cohen, R. E. 1992 Origin of ferroelectricity in perovskite oxides. *Nature* **358**, 136–138. (doi:10.1038/358136a0)

Cohen, R. E. 2008 First-principles theories of piezoelectric materials. In *Springer series in materials science on piezoelectricity*, vol. 114 (eds W. Heywang, K. Lubitz & W. Wersing), pp. 471–492. Berlin, Germany: Springer.

Curtarolo, S., Morgan, D., Persson, K., Rodgers, J. & Ceder, G. 2003 Predicting crystal structures with data mining of quantum calculations. *Phys. Rev. Lett.* **91**, 135–503. (doi:10.1103/PhysRevLett.91.135503)

Damjanovic, D. 1998 Materials for high temperature piezoelectric transducers. *Curr. Opin. Mater. Sci.* **3**, 469–473. (doi:10.1016/S1359-0286(98)80009-0)

Drache, M., Roussel, P., Wignacourt, J.-P. & Conflant, P. 2004 $Bi_{17}Yb_7O_{36}$ and $BiYbO_3$: two new compounds from the Bi_2O_3–Yb_2O_3 equilibrium phase diagram determination. *Mater. Res. Bull.* **39**, 1393–1405. (doi:10.1016/j.materresbull.2004.04.034)

Drache, M., Roussel, P., Conflant, P. & Wignacourt, J-P. 2005 $Bi_{17}Yb_7O_{36}$ and $BiYbO_3$ crystal structures of thulium and lutetium homologous compounds. *Solid State Sci.* **7**, 269–276.

Duan, R., Speyer, R. F., Alberta, E. & Shrout, T. R. 2004 High Curie temperature perovskite $BiInO_3$-$PbTiO_3$ ceramics. *J. Mater. Res.* **19**, 2185–2193. (doi:10.1557/JMR.2004.0282)

Dudiy, S. V. & Zunger, A. 2006 Searching for alloy configurations with target physical properties: impurity design via a genetic algorithm inverse band structure approach. *Phys. Rev. Lett.* **97**, 046401. (doi:10.1103/PhysRevLett.97.046401)

Eitel, R. E., Randall, C. A., Shrout, T. R., Rehrig, P. W., Hackenberger, W. & Park, S. E. 2001 New high temperature morphotropic phase boundary piezoelectrics based on $Bi(Me)O_3$–$PbTiO_3$ ceramics. *Jpn J. Appl. Phys.* **40**, 5999–6002. (doi:10.1143/JJAP.40.5999)

Emsley, J. 1998 *The elements*, 3rd edn. Oxford, UK: Clarendon Press.

Ericksson, L., Johansson, E., Kettaneh-Wold, N. & Wold, S. 2001 *Multi- and megavariate data analysis: principles, applications*. Umea, Sweden: UmetricsAb.

Feng, G., Rongzi, H., Jiaji, L., Zhen, L., Lihong, C. & Changsheng, T. 2009 Phase formation and characterization of high temperature $xBiYbO_3$-$(1-x)PbTiO_3$ piezoelectric ceramics. *J. Eur. Ceram. Soc.* **29**, 1687–1693. (doi:10.1016/j.jeurceramsoc.2008.09.024)

Fischer, C. C., Tibbetts, K. J., Morgan, D. & Ceder, G. 2006 Predicting crystal structure by merging data mining with quantum mechanics. *Nat. Mater.* **5**, 641–646. (doi:10.1038/nmat1691)

Froese-Fischer, C. 1972 Average-energy-of-configuration Hartree–Fock results for the atoms helium and radon. *Atom. Data Nucl. Data Tables* **4**, 301–399. (doi:10.1016/S0092-640X(72)80008-1)

Gadzuric, S., Suh, C., Gaune-Escard, M. & Rajan, K. 2006 Extracting information from molten salt database. *Met. Trans. A* **37**, 3411–3414. (doi:10.1007/S11661-006-1034-6)

George, L., Hrubiak, R., Rajan, K. & Saxena, S. 2009 Principal component analysis on properties of binary and ternary hydrides and a comparison of metal versus metal hydride properties. *J. Alloy. Compd* **478**, 731–735. (doi:10.1016/j.jallcom.2008.11.137)

Goudochnikov, P. & Bell, A. J. 2007 Correlations between transition temperature, tolerance factor and cohesive energy in 2+ : 4+ perovskites. *J. Phys. Condens. Matter* **19**, 176–201. (doi:10.1088/0953-8984/19/17/176201)

Goujon, C., Darie, C., Bacia, M., Klein, H., Ortega, L. & Bordet, P. 2008 High pressure synthesis of $BiCrO_3$, a candidate for multiferroism. *J. Phys. Conf. Ser.* **121**, 022009. (doi:10.1088/1742-6596/121/2/022009)

Grinberg, I. & Rappe, A. M. 2004 Silver solid solution peizoelectrics. *Appl. Phys. Lett.* **85**, 1760. (doi:10.1063/1.1787946)

Grinberg, I. & Rappe, A. M. 2007 First-principles calculations, crystal chemistry and properties of ferroelectric perovskites. *Phase Trans.* **80**, 351–368. (doi:10.1080/01411590701228505)

Grinberg, I., Suchomel, M. R., Davies, P. M. & Rappe, A. M. 2005 Predicting morphotropic phase boundary locations and transition temperatures in Pb- and Bi-based perovskite solid solutions from crystal chemical and first-principles calculations. *J. Appl. Phys.* **98**, 094111. (doi:10.1063/1.2128049)

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P. & Witten, I. H. 2009 The WEKA Data Mining software: an update. *SIGKDD Explorations*, **11**, 10–18. (doi:10.1145/1656274.1656278)

Han, J. & Kamber, M. 2006 *Data mining: concepts and techniques*, 2nd edn, pp. 291–310. Amsterdam, The Netherlands: Elsevier.

Hawkins, D. M., Young, S. S. & Ruskino III, A. 1997 A. Analysis of a large structure-activity data set using recursive partitioning. *Quant. Struct. Act. Rel.* **16**, 296–302. (doi:10.1002/qsar.19970160404)

Hotop, H. & Lineberger, W. C. 1985 Binding energies in atomic negative ions: II. *J. Phys. Chem. Ref. Data* **14**, 731–750. (doi:10.1063/1.555735)

Igarashi, K., Koumoto, K. & Yanagida, H. 1987 Ferroelectric Curie points at perovskite-type oxides. *J. Mater. Sci.* **22**, 2828–2832. (doi:10.1007/BF01086478)

Íñiguez, J., Vanderbilt, D. & Bellaiche, L. 2003 First-principles study of $(BiScO_3)_{1-x}$–$(PbTiO_3)_x$ piezoelectric alloys. *Phys. Rev. B* **67**, 224–107. (doi:10.1103/PhysRevB.67.224107)

Ishiwata, S., Azuma, M., Takano, M., Nishibori, E., Takata, M. & Kato, K. 2002 High pressure synthesis, crystal structure and physical properties of a new Ni(II) perovskite $BiNiO_3$. *J. Mater. Chem.* **12**, 3733–3737. (doi:10.1039/b206022a)

Izenman, A. J. 2008 Modern multivariate statistical techniques. In *Regression, classification and manifold learning* (eds G. Casella, S. Fienberg & I. Olkin). Springer texts in statistics. New York, NY: Springer.

Jóhannesson, G. H., Bligaard, T., Ruban, A. V., Skriver, H. L., Jacobsen, K. W. & Nørskov, J. K. 2002 Combined electronic structure and evolutionary search approach to materials design. *Phys. Rev. Lett.* **88**, 255–506. (doi:10.1103/PhysRevLett.88.255506)

Karnani, M., Pääkkönen, K. & Annila, A. 2009 The physical character of information. *Proc. R. Soc. A* 465, 2155–2175. (doi:10.1098/rspa.2009.0063)

Mackay, A. L. 1966 Generalized crystallography. *Comp. Maths. Appl. B* **12**, 21–37.
Mackay, A. L. 1974 Generalized structural geometry. *Acta Cryst. A* **30**, 440–447.
Mackay, A. L. 1977 The generalized inverse and inverse structure. *Acta Cryst. A* **33**, 212–215.
Mackay, A. L. 1986 Generalized crystallography science on form. In *Proc. of the 1st Int. Symposium on Form, Tsukuba, Japan, 26–30 November 1986* (eds Y. Kato, R. Takaki & J. Toriwaki), pp. 615–620. Tokyo, Japan: KTK Scientific Publishers.
Mackay, A. L. 2002 Generalized crystallography. *Struct. Chem.* **13**, 215–220. (doi:10.1023/A:1015838303255)
Makov, G., Gattinoni, C. & De Vita, A. 2009 Ab initio based multi-scale modeling for materials science. *Model. Simul. Mater. Sci. Eng.* **17**, 084008. (doi:10.1088/0965-0393/17/8/084008)
Matsui, T. & Nomura, S. 1981 *Landolt–Börnstein: numerical data in functional relationships in science and technology*, vol. 16a (ed. K. H. Hellwege). New York, NY: Springer-Verlag.
Mizoguchi, H., Woodward, P. M., Byeon, S-H. & Parise, J. B. 2004 Polymorphism in $NaSbO_3$: structure and bonding in metal oxides. *J. Am. Chem. Soc.* **126**, 3175–3184. (doi:10.1021/ja038365h)
Mohn, C. E. & Kob, W. 2009 A genetic algorithm for the atomistic design and global optimization of substitutionally disordered materials. *Comp. Mater. Sci.* **45**, 111–117. (doi:10.1016/j.commatsci.2008.03.046)
Montanari, E., Righi, L., Claestani, G., Migliori, A., Gilioli, E. & Bolzoni, F. 2005 Room temperature polymorphism in metastable $BiMnO_3$ prepared by high-pressure synthesis. *Chem. Mater.* **17**, 1765–1773. (doi:10.1021/cm048250s)
Oganov, A. R. & Valle, M. 2009 How to quantify energy landscapes of solids. *J. Chem. Phys.* **130**, 104–504. (doi:10.1063/1.3079326)
Oka, K. *et al.* 2010 Pressure-induced spin-state transition in $BiCoO_3$. *J. Am. Chem. Soc.* **132**, 9438–9443. (doi:10.1021/ja102987d)
Pauling, L. 1960 *The nature of the chemical bond*, 3rd edn. Ithaca, NY: Cornell University Press.
Pearson, R. G. 1988 Absolute electronegativity and hardness: application to inorganic chemistry. *Inorg. Chem.* **27**, 734–740. (doi:10.1021/ic00277a030)
Pettersson, F., Suh, C., Saxén, H., Rajan, K. & Chakraborti, N. 2009 Analyzing sparse data for nitride spinels using data mining, neural networks and multiobjective genetic algorithm. *Mater. Manuf. Process* **24**, 2–9. (doi:10.1080/10426910802539762)
Plenio, M. B. & Vitelli, V. 2001 The physics of forgetting: Landauer's erasure principle and information theory. *Contemp. Phys.* **42**, 25–60. (doi:10.1080/00107510010018916)
Rajagopalan, A. & Rajan, K. 2007 Informatics based optimization of crystallographic descriptors for framework structures. In *Combinatorial and high-throughput discovery and optimization of catalysts and materials* (eds W. Maier & R. A. Potyrailo). Boca Raton, FL: CRC Press.
Rajan, K. 2005 Materials informatics. *Mater. Today* **8**, 38–45. (doi:10.1016/S1369-7021(05)71123-8)
Rajan, K. 2010 Data mining and inorganic crystallography. In *Data mining in crystallography—structure and bonding series*, vol. 134 (eds D. W. M. Kuleshova & N. Liudmila), pp. 59–87. Berlin, Germany: Springer-Verlag.
Ravez, J., Pouchard, M. & Hagenmuller, P. 1997 Chemical bonding, a relevant tool for designing new perovskite-type ferroelectric materials. *Ferroelectrics* **197**, 161–173. (doi:10.1080/00150199708008406)
Ringnér, M. 2008 What is principal component analysis? *Nat. Biotechnol.* **26**, 303–305. (doi:10.1038/nbt0308-303)
Rödel, J., Klaus, W. K., Seifert, T. P., Anton, E.-M., Granzow, T. & Damjanovic, D. 2009 Perspective on the development of lead-free piezoceramics. *J. Am. Ceram. Soc.* **92**, 1153–1177. (doi:10.1111/j.1551-2916.2009.03061.x)
Sághi-Szabó, G., Cohen, R. E. & Krakauer, H. 1998 First-principles study of piezoelectricity in $PbTiO_3$. *Phys. Rev. Lett.* **80**, 4321–4324. (doi:10.1103/PhysRevLett.80.4321)
Saxena, S. K. 1993 *Thermodynamic data on oxides and silicates: an assessed dataset based on thermochemistry and high pressure phase equilibrium.* Berlin, Germany: Springer-Verlag.
Shannon, R. D. 1967 Synthesis of some new perovskites containing indium and thallium. *Inorg. Chem.* **6**, 1474–1478. (doi:10.1021/ic50054a009)

Shannon, R. D. 1976 Revised effective ionic radii and systematic studies of interatomic distances in halides and chalcogenides. *Acta Cryst. A* **32**, 751–767.

Shell, M. S. 2008 The relative entropy is fundamental to multiscale and inverse thermodynamic problems. *J. Chem. Phys.* **129**, 144108. (doi:10.1063/1.2992060)

Singh, K., Bopardikar, D. K. & Atkare, D. V. 1988 A compendium of $T_C - U_s$ and $P_s - \Delta z$ data for displacive ferroelectrics. *Ferroelectrics* **82**, 55–67. (doi:10.1080/00150198808201337)

Skriver, H. L. 2004 Materials science databases, CAMP, DTU. See http://guglen.dk/databases/hlsPT/ptable.php.

Slater, J. C. 1930 Atomic shielding constants. *Phys. Rev.* **36**, 57–64. (doi:10.1103/PhysRev.36.57)

Sluiter, M. H. F. 2007 Lattice stability prediction of elemental tetrahedrally closed packed structure. *Acta Mater.* **55**, 3707–3718. (doi:10.1016/j.actamat.2007.02.016)

Stein, D. M., Suchomel, M. R. & Davies, P. K. 2006 Enhanced tetragonality in (x)$PbTiO_3$-(1-x)$Bi(B'B'')O_3$ systems: $Bi(Zn_{3/4}W_{1/4})O_3$. *Appl. Phys. Lett.* **89**, 132–907. (doi:10.1063/1.2357871)

Stringer, C. J., Shrout, T. R., Randall, C. A. & Reaney, I. M. 2006 Classification of transition temperature behavior in ferroelectric $PbTiO_3$–$Bi(Me'Me'')O_3$ solid solutions. *J. Appl. Phys.* **99**, 024–106. (doi:10.1063/1.2163986)

Suchomel, M. R. & Davies, P. K. 2004 Predicting the position of the morphotropic phase boundary in high temperature $PbTiO_3$–$Bi(B'B'')O_3$ based dielectric ceramics. *J. Appl. Phys.* **96**, 4405–4410. (doi:10.1063/1.1789267)

Suchomel, M. R. & Davies, P. K. 2005 Enhanced tetragonality in $(x)PbTiO_3$-$(1-x)Bi(Zn_{1/2}Ti_{1/2})O_3$ and related solid solution systems. *Appl. Phys. Lett.* **86**, 262–905. (doi:10.1063/1.1978980)

Suh, C. & Rajan, K. 2005 Virtual screening and QSAR formulations for crystal chemistry. *QSAR Comb. Sci.* **24**, 114–119. (doi:10.1002/qsar.200420057)

Suh, C. & Rajan, K. 2009 Data mining and informatics for crystal chemistry: establishing measurement techniques for mapping structure–property relationships. *Mater. Sci. Tech.* **25**, 466–471. (doi:10.1179/174328409X430483)

Thomas, N. W. 1997 Beyond the tolerance factor: harnessing X-ray and neutron diffraction data for the compositional design of perovskite ceramics. *Br. Ceram. T* **96**, 7–15.

Witten, I. H. & Frank, E. 2000 *Data mining, practical machine learning tools and techniques with Java implementations*. San Diego, CA: Morgan Kaufman.

Woodley, S. M., Sokol, A. A. & Catlow, C. R. A. 2004 Structure prediction of inorganic nanoparticles with predefined architecture using a genetic algorithm. *Z. Anorg. Allg. Chem.* **630**, 2343–2353. (doi:10.1002/zaac.200400338)

Woodward, D. I. & Reaney, I. M. 2004 A structural study of ceramics in the $(BiMnO_3)_x$–$(PbTiO_3)_{1-x}$ solid solution series. *J. Phys. Condens. Matter* **16**, 8823–8834. (doi:10.1088/0953-8984/16/49/002)

Yan, H., Ning, H., Kan, Y., Wang, P. & Reece, M. J. 2009 Piezoelectric ceramics with super-high Curie points. *J. Am. Ceram. Soc.* **92**, 2270–2275. (doi:10.1111/j.1551-2916.2009.03209.x)

Zhang, S., Randall, C. A. & Shrout, T. R. 2003 High Curie temperature piezocrystals in $BiScO_3$–$PbTiO_3$ perovskite system. *Appl. Phys. Lett.* **83**, 3150–3152. (doi:10.1063/1.1619207)

Zhang, H., Li, N., Li, K. & Xue, D. 2007 Structural stability and formability of ABO_3-type perovskite compounds. *Acta Cryst. B* **63**, 812–818. (doi:10.1107/S0108768107046174)

Zenasni, H., Aourag, H., Broderick, S. R. & Rajan, K. 2010 Electronic structure prediction via data-mining the empirical pseudopotential method. *Phys. Status. Solidi. B* **247**, 115–121. (doi:10.1002/pssb.200945268)