PCCP

ROYAL SOCIETY OF CHEMISTRY

PAPER

View Article Online
View Journal | View Issue

Cite this: *Phys. Chem. Chem. Phys.*, 2023, **25**, 32123

Machine-learning prediction of thermal expansion coefficient for perovskite oxides with experimental validation†

Kevin P. McGuinness, [a] Anton O. Oliynyk, [b] Sangjoon Lee,[c] Beatriz Molero-Sanchez [a] and Paul Kwesi Addo *[a]

Received 21st August 2023,
Accepted 14th November 2023

DOI: 10.1039/d3cp04017h

rsc.li/pccp

Perovskite oxides have been of high-interest and relatively well studied over the last 20 years due to their various applications, specifically for solid oxide fuel cells (SOFCs) and solid oxide electrolysis cells (SOECs). One of the key properties for a perovskite to perform well as a component in SOFCs, SOECs, and other high-temperature applications is its thermal expansion coefficient (TEC). The use of machine learning (ML) to predict material properties has greatly increased over the years and has proven to be a very useful tool for materials screening. The process of synthesizing and testing perovskite oxides is laborious and costly, and the use of physics-based models is often highly computationally expensive. Due to the amount of elements able to be accommodated in the ABO_3 structure and the ability for crystallographic mixing in both the A and B-sites, there are a massive amount of possible ABO_3 perovskites. In this paper, a ML model for the prediction of the TECs of $AA'BB'O_3$ perovskites is produced and applied to millions of potential compositions resulting in reliable TEC predictions for 150 451 of the compositions.

1. Introduction

The perovskite structure type has the largest number of representatives with known structure reported to date in Pearson's crystallographic data (PCD) databases, with over 20 000 experimental reports (Fig. 1).[1] (Perovskites are especially numerous, when more derivatives like High-Tc cuprates are added to the perovskite family.) Many reports on perovskites began in the late 1980s during a spark of interest in super conductive materials and since then there have been several large spikes in perovskite research due to their various applications.[2–4] The structure of perovskite is deceptively simple (Fig. 2), given that the substitution of elements in the ABX_3 structure results in various distortions, which affects the desired properties.

Perovskite-type oxides have received a considerable amount of attention over the last few decades because of their attractive physical and chemical properties. The perovskite structure is very versatile, which enables them to perform excellently in several high-interest applications, including high temperature technologies such as solid oxide fuel cells (SOFCs), and solid oxide electrolysis cells (SOECs).[5–22] In these high temperature applications one of the properties that is essential to performance is the thermal expansion coefficients (TECs) of the materials used for certain device components. Thermal expansion coefficients

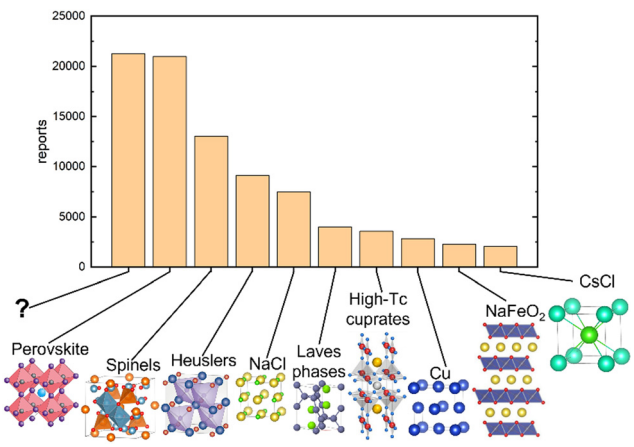

Fig. 1 Structure type statistics of top 10 most reported structure types from PCD.

[a] *SeeO2 Energy Inc., 3655 36 St NW, Calgary, AB T2L 1Y8, Canada. E-mail: founders@seeo2energy.com*
[b] *Department of Chemistry, Hunter College, City University of New York, New York, NY, 10065, USA*
[c] *Department of Applied Physics and Applied Mathematics, Columbia University, New York 10027, New York, USA*

† Electronic supplementary information (ESI) available: Machine-learning prediction of thermal expansion coefficient for perovskite oxides with experimental validation, SI.docx, TEC_predictions.xlsx and Element_Property_Table.xlsx. See DOI: https://doi.org/10.1039/d3cp04017h

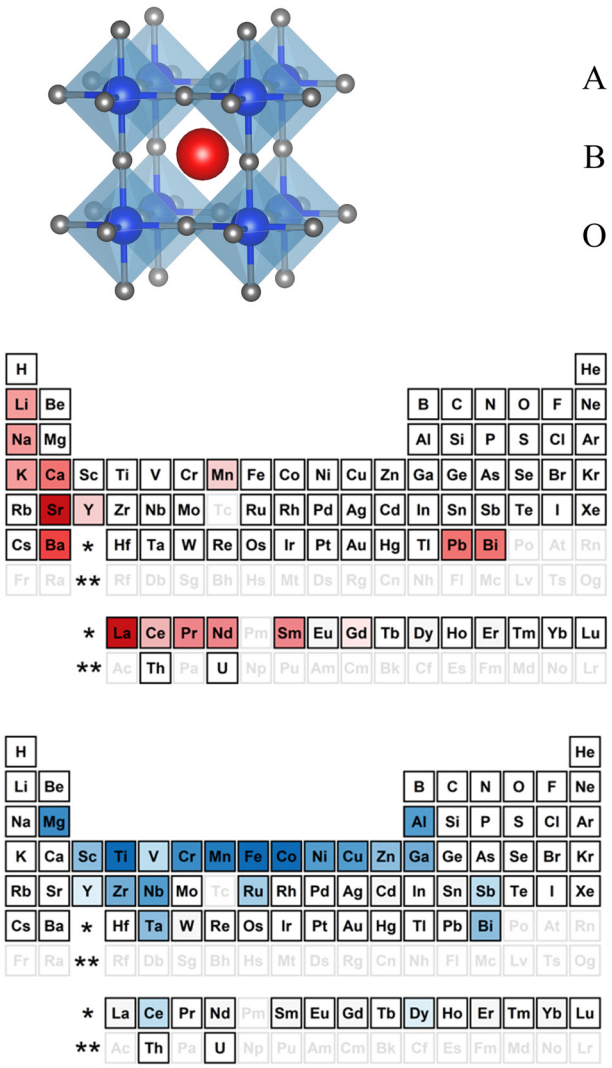

Fig. 2 Compositional map of $AA'BB'O_3$ perovskite oxides produced using compositions from the PCD, crystallographic open database, and the dataset presented in the work.

are proportionality constants of how a material expands or contracts with a change in temperature. Using materials with incompatible TECs in a device can cause strain and lead to defects, which greatly impacts the device's longevity and overall performance. For example, in SOFCs and SOECs incompatible TECs between the cell components causes delamination in the cells, which causes them to degrade rapidly.[23,24]

The ABO_3-type structure subset of perovskite oxides is able to accommodate a large range of elements in both the A and B-sites and both sites can be shared by atomic mixing, which causes there to be an immense amount of possible ABO_3-type perovskite. The overall relationship between the degree of atomic mixing in the crystallographic sites and the TEC is not very well understood as it is heavily dependent on the elements involved, and as shown later in this paper there are many element properties that effect the TEC value. Even though there has been an extensive amount of research related to perovskite oxides, only a relatively small amount of possible compositions have been studied[25–27] (Fig. 1).

In recent years the use of machine learning (ML) to predict material properties and to computationally screen for potential high performance materials has substantially increased.[28–34] However, these efforts are typically limited to theoretical study and rarely result in experimental validations.[31] In comparison to using ML, the process of synthesizing and testing the properties of perovskites is very time consuming and costly. While there are physics-based simulation techniques and ML models for predicting TECs, they typically involve computationally costly calculations, knowledge of certain experimentally determined structural properties or have significant error.[8,35–37] Targeting a high amount of TEC predictions and focusing on predictability based solely on composition is a challenge and in the current study it will be seen if it can be done.

In this paper, a ML derived prediction model for the thermal expansion coefficient of quinary $A_{1-x}A'_xB_{1-y}B'_yO_3$ perovskites, focusing on binary substitution of each metal site, where no intersite mixing is expected is discussed. The model was trained and tested using a manually compiled dataset from experimental data found in the literature (Table S1, ESI†). Using a dataset only made up of quinary $AA'BB'O_3$ instead of one with all types of ABO_3 perovskites reduced the diversity of materials being used to train and test the model and requires a less robust model to make accurate predictions and gain chemical knowledge behind the simpler phenomenon. It also had the benefit of simplifying how the data was formatted as the number of elements in each material is constant. The most beneficial aspect of this model is that they only use variables related to the properties of the constituent elements in the material and thus can be applied to a given chemical formula nearly instantly, which is demonstrated in this paper. The models used in this work can be used as an effective tool for screening of perovskite oxides by ruling out material compositions that have a TEC outside of a specified range.

2. Experimental

The model was prepared based on quinary $AA'BB'O_3$ perovskites found in the literature and in-lab gathered data. The dataset consisted of 146 samples. The TEC values recorded were taken from large range tests, usually from about 25 °C to 900–1000 °C. The features were calculated based on the composition of the samples using data from the element property table included in the ESI.† All string values (labels) were changed to numerical labels. The zero values were assumed to be meaningful, and not an equivalent of N/A. The features with empty cells were excluded, which changed the x-vector block from 948 to 902 features. Depending on the methods used, the missing values were replaced with the best guess or the feature/sample were excluded from the training set. The split of training/validation was 90/10 and the validation split method was random. To find the optimal model, the k-fold method with random splits was used to train multiple train-validation sets. The methods that were considered in this study consisted of partial least squares (PLS) regression, support vector machine (SVM) regression, principal component regression (PCR), local weighted regression

(LWR), multiple linear regression (MLR), artificial neural network (ANN), and a gradient-boosted decision tree (GBDT) regression model (XGBoost). The models were built with PLS_Toolbox software by Eigenvector in a MATLAB environment.[38] Two samples were selected for blind validation of the model, their features were calculated in a similar manner to the training dataset, and the predicted values were compared to the lab-measured values as an experimental validation.

3. Results and discussion

3.1 PLS

Various methods have been tested for the dataset. The first method was partial least squares regression. While being the simplest method, it might produce a decent result, while allowing for additional visualization of the data and the model. The principle of PLS is similar to principal component analysis (PCA), where dimensions (902 features) were reduced to three latent values (LVs). The root mean square error (RMSE) of the cross-validated dataset was 2.89 ($\times 10^{-6}$ K^{-1}), while the bias remained very low: $\times 10^{-14}$. The cumulative variance captured by the model is great, 59.56% at 3 LVs, which is a good indication that the problem might not require a complex solution.

The PLS model gives information on the potential limitations of the model and the nature of the dataset. For instance, the 95% confidence level ellipse (Fig. 3a), gives us a good idea that the samples located outside it might be very different from the rest of the data points. The sample $La_{0.2}Sr_{0.8}Co_{0.9}Sb_{0.1}O_3$ is the sample with the lowest content of the rare-earth metal (La), while other samples have at least 0.3 index next to a rare-earth metal. The two other samples both contain Al and are the only samples with Al, while the majority of the data set contains transition metals as the B-site cations in the $AA'BB'O_3$ perovskites formula. The arrays of samples that are lined can also reveal more information, such as the samples measured in the same batch, or their composition being a gradual solid solution substitution. The two samples for experimental validation (purple squares) are within the confidence ellipse. Regardless, the cross-validated model fit line is far from the ideal 1 : 1 measured/predicted ratio (Fig. 3b). While the experimental validation samples were within proximity from the point where overestimation turns into underestimation, the model deviation from the ideal line is too noticeable to speculate whether this method could be applied to the experimental validation samples and for this reason the predicted samples are not plotted, but the predicted values are included in Table 1.

Feature statistics can help us understand the nature of the TEC phenomenon within the perovskite dataset. Table 2 lists the top ten features according to the variable importance in projection (VIP) and selectivity ratio feature statistics. While feature statistics show only the potential of each variable to influence the model, it is important to understand that it is not the individual variables that govern the model, rather the combination of variables that is important. It is necessary to see that out of 902 features the best-performing features from different statistic methods appear in the top 10 of the feature lists. In this case, the most common high-performance features are various

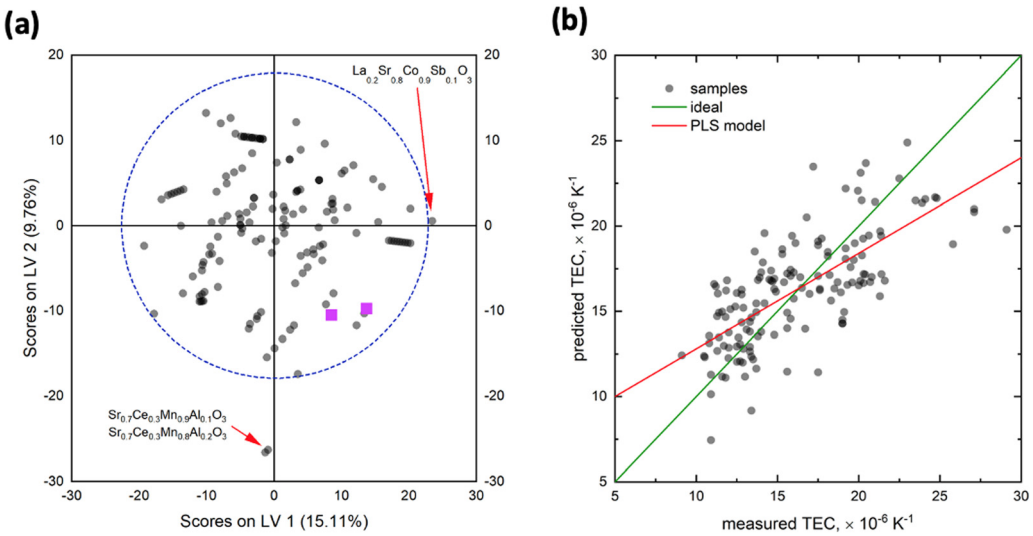

Fig. 3 (a) PLS model confidence ellipse (b) plot of predicted TEC *vs.* measured TEC for the PLS model TEC predictions. R^2: 0.63.

Table 1 Experimentally validated samples (values $\times 10^{-6}$ K^{-1})

				Other methods			
Sample	Experimental TEC	PLS predictions	SVM prediction	PCR	LWR	ANN	XGBoost
$La_{0.3}Ca_{0.7}Fe_{0.7}Cr_{0.3}O_3$	11.83	16.13	12.70	18.69	13.76	12.94	18.70
$La_{0.3}Sr_{0.7}Fe_{0.7}Cr_{0.3}O_3$	16.30	18.75	15.83	20.30	17.77	17.39	20.68

Table 2 Variable statistics in PLS model

Top 10 features (VIP)	Top 10 features (selectivity ratio)
B crystal radius	**A Pauling EN**
B gamma	**A Allred EN**
A Pauling EN	**A Allred–Rockow EN**
A Allred EN	**A density**
A density	**A Nagle EN**
A Allred–Rockow EN	**A Mulliken EN**
B bulk modulus	A ScRLDA Exc potential
A Mulliken EN	A RLDA Exc potential
A Nagle EN	A LSD Exc potential
B′ ionization energy	A LDA Exc potential

electronegativity scales. The data redundancy is not an issue given that PLS method is based on dimensionality reduction. From the list, the factors important in perovskites are similar to other works that study the formation of perovskite phases. For example, electronegativity is crucial to differentiate A and B-site cations in perovskite, along with the size factors, that typically attribute large size to an A-site cations.[39–41] Interestingly, bulk modulus is also in the top list of feature statistics. Bulk modulus is a compressibility factor, directly related to expansion/compression process, the very phenomenon TEC is about.

3.2 SVM

Next, the support vector machine method is implemented, which has demonstrated numerous successes in the solid-state chemistry field, given that the method is exceptionally successful with limited datasets, common in materials research.[42] Epsilon-SVR with radial basis function type kernel was used with cost parameter = 3.1623, epsilon = 0.01, and gamma = 0.001. While the parameters are reasonable for an SVM model (Fig. 4), the number of support vectors is 137, which is too large for a 146 sample dataset, which is reflected in the gamma parameter and might be considered as overfitting. Cross-validated RMSE = 1.54 ($\times 10^{-6}$ K^{-1}), reduced RMSE value of 0.38, and bias being in the range on $\times 10^{-1}$ is good.

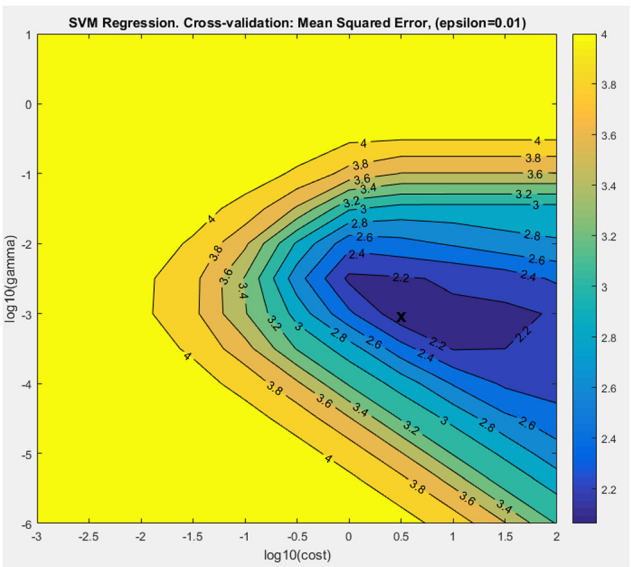

Fig. 4 SVM parameters optimization.

The model's fit (Fig. 5) shows the promising potential for extrapolation or for entire database screening. It should be noted that the samples that show the most overestimation or underestimation are Ba-containing samples, which are rare in the database and contain Sb in the composition. Similarly to Al-containing samples highlighted in the PLS model, the SVM-highlighted Sb-containing samples are outside of a typical compositional range for the B substitution element, which typically is a transition metal. It is worth pointing out the most underestimated sample also features a small atomic percent of the substitution, which is generally not ideal for machine learning models.

3.3 Other methods

As a survey run, other ML methods have been used to estimate their potential: principal component regression, local weighted regression with 10 local points, and multiple linear regression. Out of these three methods, the MLR fit was decent, however, the model predicts a negative TEC value for some samples, making the model physically unreasonable. Feature scores were extracted at the same time, with a significant overlap of the most important variables to the ones listed in Table 2. Among interesting additions to the important feature list is B' weighted Young modulus (Young modulus value multiplied by the atomic fraction of B′), which is physically related to the previously mentioned bulk modulus. This tells us that the perovskite TEC problem is not purely chemical, but also depends on the mechanical properties of the constituent elements.

One of the most common machine-learning techniques, artificial neural network was also applied to test it. It is worth mentioning that even before starting with ANN, there were not high expectations, since ANN works wonderfully well with large amounts of data, which is not the case in materials science.[43]

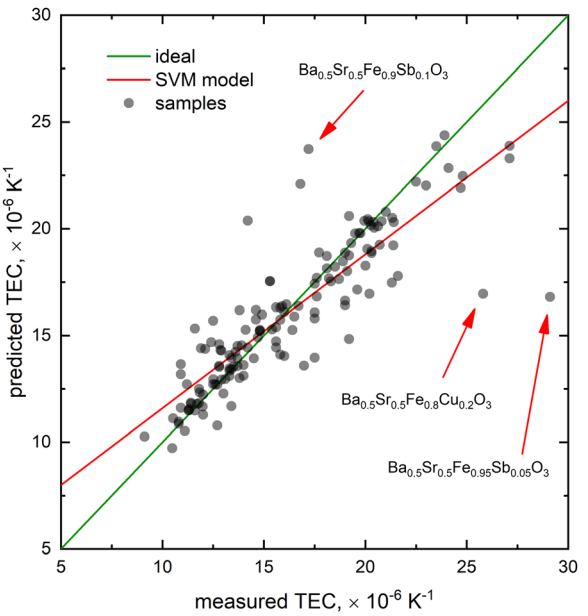

Fig. 5 Plot of predicted TEC *vs.* measured TEC for the SVM model TEC predictions. R^2: 0.88.

ANN resulted in RMSE = 1.52 ($\times 10^{-6}$ K^{-1}), with bias being in the range of $\times 10^{-2}$, however, the best results were achieved with only 2 layer-1 nodes, definitely not an expected or typical parameter for such a complex method. For a comparison, the model fit is available in the ESI† (Fig. S1).

XGBoost (gradient-boosted decision tree) regression is another method that has been applied. While it produces a fit and overall results that are similar to the PLS model, this method has a great advantage, as it allows us to score the features. The set of the top most important features was different from the list in Table 2. From the newer insights (XGBoost) the most important variable was A-site cation atomic weight, which is expected given that most of the A-site cations are heavy alkali metals or rare-earth elements.

Summarizing the trial runs, it can be concluded that the factors that typically influence perovskite formation, such as A *vs.* B-site cation electronegativity and size echo well with similar parameters from ML methods, and second the crystallographic patterns proposed by Pauling, Villars, and Pettifor. Additional insight that has been obtained from ML is that mechanical factors play a significant role, especially the compressibility factor. This is expected, given that the property being studied (TEC) is related to changes that manifest themselves in mechanical form, such as expansion. While the other methods provided additional insight, they did not perform as well as the SVM model.

3.4 Feature selection

Feature selection is a typical approach to make the model less bulky and more sensitive to the variable changes. Feature selection is used to reduce the number of variables, while improving model statistics, inherently it means that a significant part of the *x*-block carries useless information and produces noise. Feature selection is an iterative process of removing/adding features until a better model than the one started with is produced. The most common feature selection approaches (besides simple variable statistics) are genetic algorithm (GA), which is a combination of mutation, crossover, and selection steps, iPLS and rPLS, which use scored features (top of this list is given in Table 2) and then apply an interval or recursive weighted method to this list. Given that the features in Table 2 already show chemically meaningful features, there is a reasonable expectation that these methods might be successful (Table 3).

GA took the most time, which is expected, however produced insignificant improvement, by lowering the RMSE statistics by less than 0.001 ($\times 10^{-6}$ K^{-1}) with a model that uses 259 out of 902 features. While the improvement is less than what would be considered useful, it is important to see how the most successful methods (PLS and SVM) from the trial run respond to these newly selected features. The PLS method was improved, resulting in a better fit with just two LVs, that capture in total 33.20% of all variance, which is an extremely successful result, considering that there were an initial 902 variables. This resulted in all samples being within the confidence ellipse, except for $La_{0.2}Sr_{0.8}Co_{0.9}Sb_{0.1}O_3$ which also showed problems as described above. The same selected features were used for the SVM model which resulted in improved predictions (Fig. 6). The improved SVM model give fairly accurate TEC predictions for the two external validation samples, with an error of 7.23% and 2.67%. The prediction accuracies of ML models are can be assessed by their usefulness in real world situations. In SOFCs and SOECs a TEC difference between adjacent cell components of 7.23% or 2.67% is considered to be good, meaning predictions with errors this low give useful and actionable results.[44–46]

Table 3 Feature selection model results (values $\times 10^{-6}$ K^{-1})

Sample	Default PLS	iPLS-FS PLS	GA PLS	Default SVM	iPLS-FS SVM	GA SVM
$La_{0.3}Ca_{0.7}Fe_{0.7}Cr_{0.3}O_3$	16.13	17.87	16.28	**12.70**	13.19	**12.65**
$La_{0.3}Sr_{0.7}Fe_{0.7}Cr_{0.3}O_3$	18.75	18.92	18.25	**15.83**	16.14	**15.59**

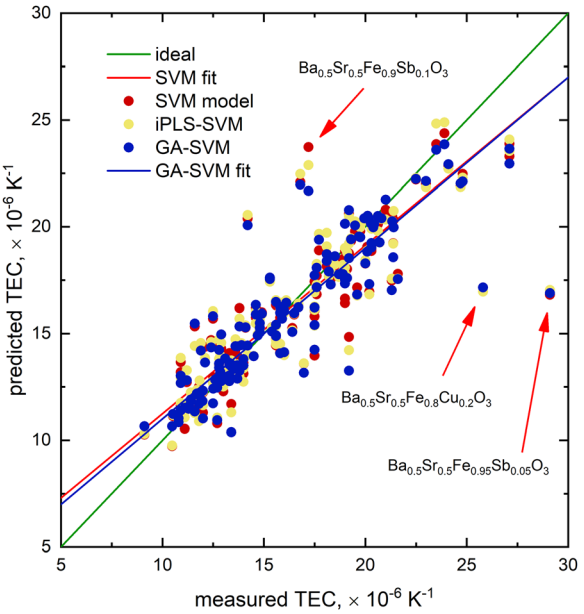

Fig. 6 Comparison of the SVM models produced using different feature selection methods.

3.5 Predictions

Once the model was finalized, the list of $AA'BB'O_3$ perovskite compositions to predict TEC for was generated and the model was applied. Given that the number of potential candidates could be quite large, some limitations to possible A and B-site cations and mixings (A–A′, B–B′, and A–B) were applied. Element statistics were gathered from the crystallography open database (COD), the PCD database and the dataset used in this work (Tables S2 and S3, ESI†).[1] The index range of oxygen in the perovskites was set to 2.55–3 in order to maximize the number of compounds for statistical purpose. To differentiate A and B-site cations in ABO_3 formula, a simple size criterion is sufficient, since A-site cation is larger than B in most cases (only 3.6% exceptions), when ionic radii are compared. The exceptions were typically attributed to the presence of Bi and Nb elements in the composition. From 302

Table 4 Composition of A and B site in perovskites

A-Site elements	B-Site elements	A or B-site elements
Ba	Al	Bi
Ca	Cd	Ce
Eu	Co	Dy
K	Cr	Er
La	Cu	Gd
Li	Fe	Mn
Na	Ga	Sm
Nd	Mg	Y
Pb	Nb	
Pr	Nd	
Sr	Ni	
	Ru	
	Sb	
	Sc	
	Sn	
	Ta	
	Ti	
	V	
	W	
	Yb	
	Zn	
	Zr	

$AA'BB'O_3$ perovskite reports, A–A′, B–B′, and A–B combinations were summarized in Tables S4–S6 (ESI†). The list of elements for perovskite candidate generation is available in Table 4. Taking into account the size factor, some limitations were additionally applied. For example, Yb, Dy, Er, Gd, La, Mn, Sm, Y were only considered as a B-site element with Ba and K in the A-site.

For the test run, 10 000 compositions were generated, then were processed through the routine of generating features and their TEC values were predicted (Fig. 7). As expected, most of the candidate compositions returned a prediction of the same value (TEC = 14.06×10^{-6} K^{-1}), which is an indication of the sample compositions being outside the confidence region of the ML model. This is not an indication of a bad model, rather a way to warn the researcher that the model had to deal with values (or elements) it has not seen before, which results in a fixed value prediction, better than a diverse random guess. On the other hand, some series showed a great value distribution, like the $La_xSr_{1-x}B_yB'_{1-y}O_3$ series, which had TEC values predicted from 10×10^{-6} K^{-1} to 24×10^{-6} K^{-1}. The series with Ba also had successful predictions, with values from 13×10^{-6} K^{-1} to 16×10^{-6} K^{-1}.

Next, 3 593 726 $A_xA'_{1-x}B_yB'_{1-y}O_3$ compositions were generated with x and y ranging from 0.2 to 0.8, with a 0.1 composition step, and additional points at 0.25 and 0.75. In these compositions, some of the elements did not have Allred electronegativity values, which were needed for the model. One option was to remove the features that were composed of Allred values, which could potentially negatively affect the model quality, and the other option was to get the missing Allred values for the elements that lack them, and proceed with the prediction. The latter option was selected. In order to get the missing Allred values for some of the elements, a separate ML model that can successfully predict the values was tested and created. To predict the values of the elements supervised learning algorithms were used to generate

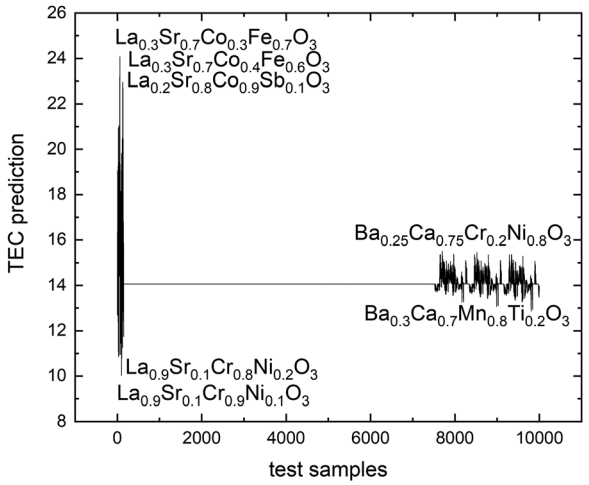

Fig. 7 TEC predictions for the 10 000 composition test run.

regression models. The algorithms include (1) support-vector machines, (2) decision tree, (3) multi-layer perception neural network (MLP-NN), (4) cubic spline interpolation, and (5) Gaussian process regression (GPR). The RMSE values between the actual and predicted Allred values were compared (Table S7, ESI†). SVMs, MLP-NN, and cubic spline interpolation resulted in high RMSE values. The poor performance can be attributed to the smoothness of the regression curves. The Decision Tree algorithm, in contrast, produced a RMSE value of 0, demonstrating a perfect alignment between actual and predicted Allred values. However, each of the predicted values was identical to the value of the adjacent element with 1 lower atomic number. Out of the 5 algorithms, GPR with the Radial Basis Function (RBF) kernel produced the best regression model. With the RMSE value of 6.78×10^{-10}, the unknown Allred values were successfully interpolated with the GPR method. As a result, Predicted Allred values (Table 5), were used for feature calculation and then applied to predict TEC values.

In total 150 451 compositions had a reliably predicted TEC, with the list available in ESI.† The histogram distribution of predicted values is shown in Fig. 8. From the deviation of the model line from the ideal line, a correction equation could be extracted, and then applied to predicted values. In principle, this should lead to the corrected values being in closer

Table 5 Predicted Allred EN values

		Allred EN predicted
Niobium	Nb	1.83
Technetium	Tc	2.12
Ruthenium	Ru	2.15
Tellurium	Te	2.52
Promethium	Pm	1.16
Europium	Eu	1.18
Terbium	Tb	1.21
Ytterbium	Yb	1.26
Hafnium	Hf	1.39
Tantalum	Ta	1.79
Rhenium	Re	2.83
Osmium	Os	2.78

agreement with the experimentally measured values. The correction used the following equation (eqn (1)):

$$\text{Corrected value} = \text{value} + (\text{value} - 15.7) \times 0.72 \quad (1)$$

All the elements that appeared in the training/testing dataset also appeared in compositions inside the confidence region of the model. The A-site elements that did not appear in compositions included in the training set but appeared in the model confidence region were Eu, K, Na, and Pb, whereas compositions including Li, Mn, Y, and Yb in the A-site were also not in the training set but weren't in the model confidence region. The main cause of this is likely due to the differences in size between the grouping of Li, Mn, Y, and Yb and the grouping of Eu, K, Na, and Pb. The only elements that were in the B-site for generated compositions but were not in the B-site for any of the compositions in the model confidence region were La and Y, which were only used as potential B-site elements when Ba and K were the A-site elements. This is not very surprising as the training set did not have any compositions that included La or Y in the B-site and they both have larger ionic radii then all the B-site elements included in the training set.

Of the 150 451 compositions in the model confidence region 84% of them contained Sr, which isn't surprising as nearly all (98%) of the compositions in the training set included Sr. The other 16% of compositions in the model confidence region were significantly made up of compositions containing Ce, Ca, Ba, Eu, Dy, Bi, or Pr in the A-site. It is not surprising that Ca and Ba are in this list as they are the alkaline earth metals in nearest proximity to Sr. Bi was the fourth highest occurring A-site element in the training set so it is not surprising to see a significant amount of compositions containing Bi in the confidence region of the model. It is not very clear as to why Ce, Eu, Dy, and Pr appear in much more non-Sr compositions in the confidence region than Gd, La, Nd, Sm, and Er, and due to there being many features used in the model (259) it is difficult to determine the exact cause.

Most compositions with a TEC value at the low and high ends of the predicted TEC range have both an A and B-site element with an index of 0.8 or 0.75, which is expected as a smaller amount of doping typically leads to a smaller change in TEC from what the value of the most similar undoped ABO_3. The default SVM predictions ranged from TEC values of 8.49×10^{-6} K^{-1} to 25.65×10^{-6} K^{-1}, which is close to the training set range of 9.11×10^{-6} K^{-1} to 27.1×10^{-6} K^{-1}. It was expected that there would be few default SVM predictions that are outside the training set range and due to the limited range of the training set and lower representation of compositions with low or high TECs, the default predictions at the low and high ends may be a bit conservative.

Looking at various $A_aA'_bB_{1-x}B'_xO_3$ and $A_{1-x}A'_xB_aB'_bO_3$ series (where x = 0.2–0.8), the predictions seem to follow the three typical trends seen in literature. The first one being where the TEC values continuously increase as x increases (cases where $A_aA'_bB'O_3$ or $A'_xB_aB'_bO_3$ has a larger TEC values than $A_aA'_bBO_3$ or $AB_aB'_bO_3$). The second trend seen is the opposite where the TEC values continuously decrease as x increases. The third trend observed is where the highest or lowest TEC value in the series is around x = 0.5 (sometimes x = 0.4 or x = 0.6), which seems to occur in cases where $A_aA'_bB'O_3$ or $A'_xB_aB'_bO_3$ and $A_aA'_bBO_3$ or $A_aA'_bO_3$ have similar TEC values.

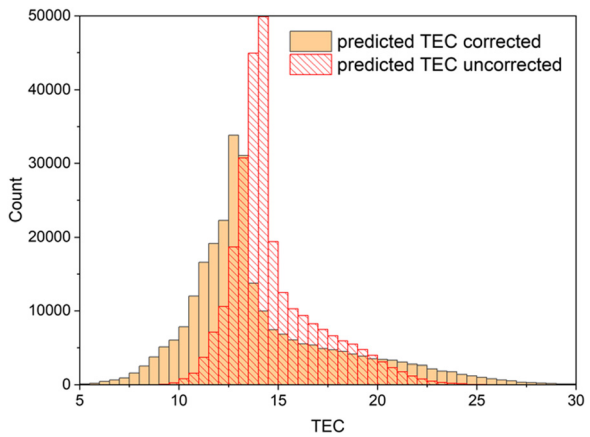

Fig. 8 Distribution of TEC predictions for the 150 451 compositions in the model confidence region.

4. Conclusion

In summary, the use of a machine learning model for the prediction of thermal expansion coefficients of $AA'BB'O_3$-type perovskites that uses features based solely on composition and was experimentally validated using two blind validation samples was successfully demonstrated. Using a manually compiled dataset consisting of 146 samples a variety of ML methods including PLS, SVM, PCR, LWR, MLR, ANN, and GBDT, were used to produce the TEC prediction models. The method that produced the highest performing model was SVM, and after a feature selection process reducing the number of features from an initial 902 to 259, the model gave a cross-validated RMSE value of 1.54×10^{-6} K^{-1}. After finalizing the model, 3 593 726 $A_xA'_{1-x}B_yB'_{1-y}O_3$ compositions were generated with x and y ranging from 0.2 to 0.8, with a 0.1 composition step, and additional points at 0.25 and 0.75. The model was applied to the generated compositions which resulted in 150 451 of the predictions being in the confidence region of the model and deemed to be reliable predictions of the TECs.

Data availability

The training dataset and the predictions have been shared in the ESI,† section.

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

SeeO2 Energy Inc. and KPM acknowledge support from the ECO CANADA Science Horizons Youth Internship program to

support this work. SeeO2 Energy Inc. would like to acknowledge Haris Masood Ansari for helpful discussions related to producing the set of potential compositions.

References

1 P. Villars and K. Cenzual, *Pearson's Crystal Data: Crystal Structure Database for Inorganic Compounds (on DVD). Published online*, 2021.
2 M. A. Alario-Franco, C. Chaillout, J. J. Capponi and J. Chenavas, Oxygen vacancy ordering and non stoichiometry in the $Ba_2YCu_3O_{7-x}$ superconductors, *Mater. Res. Bull.*, 1987, **22**(12), 1685–1693, DOI: **10.1016/0025-5408(87)90012-2**.
3 M. A. Alario-Franco, E. Morán-Miguélez and R. Sáez-Puche, *et al.*, The rare-earth H.T.S.C. family $Ba_2(RE)Cu_3O_7$; structural, electrical and magnetic studies (RE = Y,Nd,Sm,Eu,Gd,-Dy,Ho,Er,Tm), *Mater. Res. Bull.*, 1988, **23**(3), 313–321, DOI: **10.1016/0025-5408(88)90003-7**.
4 L. Ortega-San-Martin, *Introduction to Perovskites: A Historical Perspective*, 2020, pp. 1–41, DOI: **10.1007/978-981-15-1267-4_1**.
5 R. J. H. Voorhoeve, D. W. Johnsoi, J. P. Remeika and P. K. Gall, Perovskite Oxides, Long Known in Solid-State Chem and Physics, *Find New Appl. Catal.*, 1977, **195**, 827–833, DOI: **10.1126/science.195.4281.827**.
6 C. Sun, J. A. Alonso and J. Bian, Recent Advances in Perovskite-Type Oxides for Energy Conversion and Storage Applications, *Adv. Energy Mater.*, 2021, **11**(2), 2000459, DOI: **10.1002/aenm.202000459**.
7 A. S. Bhalla, R. Guo and R. Roy, The perovskite structure – A review of its role in ceramic science and technology, *Mater. Res. Innovations*, 2000, **4**(1), 3–26, DOI: **10.1007/s100190000062**.
8 Q. Tao, P. Xu, M. Li and W. Lu, Machine learning for perovskite materials design and discovery, *npj Comput. Mater.*, 2021, 7(1), 23, DOI: **10.1038/s41524-021-00495-8**.
9 P. K. Addo, B. Molero-Sanchez, M. Chen, S. Paulson and V. Birss, CO/CO_2 Study of High Performance $La_{0.3}Sr_{0.7}Fe_{0.7}Cr_{0.3}O_{3-\delta}$ Reversible SOFC Electrodes, *Fuel Cells*, 2015, **15**(5), 689–696, DOI: **10.1002/fuce.201400196**.
10 B. Molero-Sánchez, J. Prado-Gonjal, D. Ávila-Brande, V. Birss and E. Morán, Microwave-assisted synthesis and characterization of new cathodic material for solid oxide fuel cells: $La_{0.3}Ca_{0.7}Fe_{0.7}Cr_{0.3}O_{3-\delta}$, *Ceram. Int.*, 2015, **41**(7), 8411–8416, DOI: **10.1016/j.ceramint.2015.03.041**.
11 B. Molero-Sánchez, P. K. Addo, A. Buyukaksoy and V. Birss, GDC-Infiltrated $La_{0.3}Ca_{0.7}Fe_{0.7}Cr_{0.3}O_{3-\delta}$ Symmetrical Oxygen Electrodes for Reversible SOFCs, *ECS Trans.*, 2015, **66**(2), 185–193, DOI: **10.1149/06602.0185ecst**.
12 P. K. Addo, B. Molero-Sanchez, A. Buyukaksoy, S. Paulson and V. Birss, Sulfur Tolerance of $La_{0.3}Ca_{0.7}Fe_{0.7}Cr_{0.3}O_{3-\delta}$ (M = Sr, Ca) Solid Oxide Fuel Cell Anodes, *ECS Trans.*, 2015, **66**(2), 219–228, DOI: **10.1149/06602.0219ecst**.
13 B. Molero-Sánchez, J. Prado-Gonjal, D. Ávila-Brande, M. Chen, E. Morán and V. Birss, High performance $La_{0.3}Ca_{0.7}Cr_{0.3}Fe_{0.7}O_{3-\delta}$ air electrode for reversible solid oxide fuel cell applications, *Int. J. Hydrogen Energy*, 2015, **40**(4), 1902–1910, DOI: **10.1016/j.ijhydene.2014.11.127**.
14 B. Molero-Sánchez, P. Addo, A. Buyukaksoy, S. Paulson and V. Birss, Electrochemistry of $La_{0.3}Sr_{0.7}Fe_{0.7}Cr_{0.3}O_{3-\delta}$ as an oxygen and fuel electrode for RSOFCs, *Faraday Discuss.*, 2015, **182**, 159–175, DOI: **10.1039/C5FD00029G**.
15 J. Prado-Gonjal, M. M. González-Barrios, M. T. Fernández-Díaz, P. K. Addo and B. Molero-Sánchez, Crystal structure and electrical properties of $LaNi_{0.6}Fe_{0.2}Cu_{0.2}O_{3-\delta}$ and $LaNi_{0.6}Fe_{0.3}Cr_{0.1}O_{3-\delta}$ perovskites: Contact materials for reversible solid oxide fuel cell electrodes, *J. Solid State Chem.*, 2022, **316**, 123526, DOI: **10.1016/j.jssc.2022.123526**.
16 H. M. Ansari, D. Avila-Brande, S. Kelly, P. K. Addo and B. Molero-Sánchez, Structural, Interfacial, and Electrochemical Stability of $La_{0.3}Ca_{0.7}Fe_{0.7}Cr_{0.3}O_{3-\delta}$ Electrode Material for Application as the Oxygen Electrode in Reversible Solid Oxide Cells, *Crystals*, 2022, **12**(6), 847, DOI: **10.3390/cryst12060847**.
17 K. Singh, P. K. Addo, V. Thangadurai, J. Prado-Gonjal and B. Molero-Sánchez, $LaNi_{0.6}Co_{0.4-x}Fe_xO_{3-\delta}$ as Air-Side Contact Material for $La_{0.3}Ca_{0.7}Fe_{0.7}Cr_{0.3}O_{3-\delta}$ Reversible Solid Oxide Fuel Cell Electrodes, *Crystals*, 2022, **12**(1), 73, DOI: **10.3390/cryst12010073**.
18 B. Molero-Sánchez, P. Addo, A. Buyukaksoy and V. Birss, Performance Enhancement of $La_{0.3}Ca_{0.7}Fe_{0.7}Cr_{0.3}O_{3-\delta}$ Air Electrodes by Infiltration Methods, *J. Electrochem. Soc.*, 2017, **164**(10), F3123–F3130, DOI: **10.1149/2.0151710jes**.
19 B. Molero-Sánchez, P. Addo, E. Morán and V. Birss, *Microwave Synthesis and Sintering Methods for Reversible Solid Oxide Fuel Cell Fabrication*, *ECS Meeting Abstracts*, 2016, **MA2016-02**(40):3042, DOI: **10.1149/MA2016-02/40/3042**.
20 P. Addo, A. Ahsen, A. Buyukaksoy, B. Molero-Sánchez, O. Ozturk and V. Birss, *Understanding the Effect of Temperature on the Sulfur Tolerance of a Ca Rich Ferrite SOFC Electrode*, *ECS Meeting Abstracts*, 2016, **MA2016-02**(39):2926, DOI: **10.1149/MA2016-02/39/2926**.
21 P. Addo, S. Mulmi, B. Molero-Sánchez, P. Keyvanfar, V. Thangadurai and V. Birss, Performance Enhancement of $La_{0.3}Sr_{0.7}Fe_{0.7}Cr_{0.3}O_3$ (LSFCr) Electrodes in CO_2/CO Atmosphere, *ECS Meeting Abstracts*, 2016, **MA2016-02**(40), 3041, DOI: **10.1149/MA2016-02/40/3041**.
22 E. Sánchez-Ahijón, R. Schmidt and X. Martínez de Irujo-Labalde, *et al.*, Structural and dielectric properties of ultra-fast microwave-processed $La_{0.3}Ca_{0.7}Fe_{0.7}Cr_{0.3}O_{3-\delta}$ ceramics, *J. Solid State Chem.*, 2022, **314**, 123426, DOI: **10.1016/j.jssc.2022.123426**.
23 A. V. Nikonov, K. A. Kuterbekov, K. Z. Bekmyrza and N. B. Pavzderin, A brief review of conductivity and thermal expansion of perovskite-related oxides for SOFC cathode, *Eurasian J. Phys. Funct. Mater.*, 2018, **2**(3), 274–292, DOI: **10.29317/EJPFM.2018020309**.
24 F. Tietz, *Thermal Expansion of SOFC Materials*, 1999, vol. 5.
25 E. A. R. Assirey, Perovskite synthesis, properties and their related biochemical and industrial application, *Saudi Pharm. J.*, 2019, **27**(6), 817–829, DOI: **10.1016/j.jsps.2019.05.003**.
26 M. A. Peña and J. L. G. Fierro, Chemical structures and performance of perovskite oxides, *Chem. Rev.*, 2001, **101**(7), 1981–2017, DOI: **10.1021/cr980129f**.
27 M. Kubicek, A. H. Bork and J. L. M. Rupp, Perovskite oxides-a review on a versatile material class for solar-to-fuel

conversion processes, *J. Mater. Chem. A*, 2017, **5**(24), 11983–12000, DOI: **10.1039/c7ta00987a**.

28 J. Schmidt, M. R. G. Marques, S. Botti and M. A. L. Marques, Recent advances and applications of machine learning in solid-state materials science, *npj Comput. Mater.*, 2019, **5**(1), 83, DOI: **10.1038/s41524-019-0221-0**.

29 Y. Juan, Y. Dai, Y. Yang and J. Zhang, Accelerating materials discovery using machine learning, *J. Mater. Sci. Technol.*, 2021, **79**, 178–190, DOI: **10.1016/j.jmst.2020.12.010**.

30 J. Kim, D. Kang, S. Kim and H. W. Jang, Catalyze Materials Science with Machine Learning, *ACS Mater. Lett.*, 2021, **3**(8), 1151–1171, DOI: **10.1021/acsmaterialslett.1c00204**.

31 D. Morgan and R. Jacobs, Opportunities and Challenges for Machine Learning in Materials Science, *Annu. Rev. Mater. Res.*, 2020, **50**(1), 71–102, DOI: **10.1146/annurev-matsci-07021832**.

32 R. Ramprasad, R. Batra, G. Pilania, A. Mannodi-Kanakkithdi and C. Kim, Machine learning in materials informatics: recent applications and prospects, *npj Comput. Mater.*, 2017, **3**, 54, DOI: **10.1038/s41524-017-0056-5**.

33 A. Agrawal and A. Choudhary, Perspective: Materials informatics and big data: Realization of the ''fourth paradigm'' of science in materials science, *APL Mater.*, 2016, **4**(5), 053208, DOI: **10.1063/1.4946894**.

34 R. K. Vasudevan, K. Choudhary and A. Mehta, *et al.*, Materials science in the artificial intelligence age: high-throughput library generation, machine learning, and a pathway from correlations to the underpinning physics, *MRS Commun.*, 2019, **9**(3), 821–838, DOI: **10.1557/mrc.2019.95**.

35 F. Heydari, A. Maghsoudipour, M. Alizadeh, Z. Khakpour and M. Javaheri, Modeling of thermal expansion coefficient of perovskite oxide for solid oxide fuel cell cathode, *Appl. Phys. A: Mater. Sci. Process.*, 2015, **120**(4), 1625–1633, DOI: **10.1007/s00339-015-9374-y**.

36 C. Li, H. Hao and B. Xu, *et al.*, Improved physics-based structural descriptors of perovskite materials enable higher accuracy of machine learning, *Comput. Mater. Sci.*, 2021, 198, DOI: **10.1016/j.commatsci.2021.110714**.

37 J. Peng, N. S. Harsha Gunda, C. A. Bridges, S. Lee, J. Allen Haynes and D. Shin, A machine learning approach to predict thermal expansion of complex oxides, *Comput. Mater. Sci.*, 2021, 111034, DOI: **10.1016/j.commatsci.2021.111034**Published online.

38 PLS_Toolbox. Published online 2021.

39 L. Li, Q. Tao, P. Xu, X. Yang, W. Lu and M. J. Li, Studies on the regularity of perovskite formation via machine learning, *Comput. Mater. Sci.*, 2021, 199, DOI: **10.1016/j.commatsci.2021.110712**.

40 V. Sharma, P. Kumar, P. Dev and G. Pilania, Machine learning substitutional defect formation energies in ABO3perovskites, *J. Appl. Phys.*, 2020, **128**(3), 034902, DOI: **10.1063/5.0015538**.

41 L. Li, Q. Tao, P. Xu, X. Yang, W. Lu and M. J. Li, Studies on the regularity of perovskite formation via machine learning, *Comput. Mater. Sci.*, 2021, 199, DOI: **10.1016/j.commatsci.2021.110712**.

42 W. C. Lu, X. B. Ji, M. J. Li, L. Liu, B. H. Yue and L. M. Zhang, Using support vector machine for materials design, *Adv. Manuf.*, 2013, **1**(2), 151–159, DOI: **10.1007/s40436-013-0025-2**.

43 A. Y. T. Wang, R. J. Murdock and S. K. Kauwe, *et al.*, Machine Learning for Materials Scientists: An Introductory Guide toward Best Practices, *Chem. Mater.*, 2020, **32**(12), 4954–4965, DOI: **10.1021/acs.chemmater.0c01907**.

44 L. Zhang, M. Liu and J. Huang, *et al.*, Improved thermal expansion and electrochemical performances of $Ba_{0.6}Sr_{0.4}Co_{0.9}$-$Nb_{0.1}O_{3-\delta}$–$Gd_{0.1}Ce_{0.9}O_{1.95}$ composite cathodes for IT-SOFCs, *Int. J. Hydrogen Energy*, 2014, **39**(15), 7972–7979, DOI: **10.1016/j.ijhydene.2014.03.055**.

45 J. Wu and X. Liu, Recent Development of SOFC Metallic Interconnect, *J. Mater. Sci. Technol.*, 2010, **26**(4), 293–305, DOI: **10.1016/S1005-0302(10)60049-7**.

46 Y. Zhang, B. Chen and D. Guan, *et al.*, Thermal-expansion offset for high-performance fuel cell cathodes, *Nature*, 2021, **591**(7849), 246–251, DOI: **10.1038/s41586-021-03264-1**.