COMPOSITIONAL AND ELEMENTAL DESCRIPTORS FOR PEROVSKITE MATERIALS

Jiří Hostaš, Hatef Shahmohamadi,
Venkataraman Thangadurai & Dennis R. Salahub
Department of Chemistry
University of Calgary
2500 University Drive NW
Calgary, AB, T2N 1N4, Canada
{jiri.hostas}@gmail.com

Maicon Pierre Lourenço
Departamento de Química e Física –
Centro de Ciências Exatas
Naturais e da Saúde – CCENS –
Universidade Federal do Espírito Santo
29500-000, Alegre, Espírito Santo, Brasil

John Garcia & Karthik Shankar
Department of Electrical and Computer Engineering
University of Alberta
9211 116 Street NW
Edmonton, AB, T6G 1H9, Canada

Alain Tchagang
Digital Technologies Research Centre
National Research Council of Canada
1200 Montréal Road
Ottawa, ON, K1A 0R6 Canada

ABSTRACT

In this extended abstract we compare the performance of different families of descriptors – *molar composition descriptor, weight composition descriptor and elemental descriptor* – for regression task (prediction of bandgap) and include examples of a classification task for perovskite oxide materials with general formulas ABO_3, $A_2BB'O_6$, and $A_xA'_{1-x}B_yB'_{1-y}O_6$. The best performance was observed for our elemental descriptor which consisted of A-site and B-site element information on: Shannon's ionic radius, ideal bond length, electronegativity, van der Waals radius, ionization energy, molar volume, atomic number, and atomic mass. The weight composition descriptor showed superior results over a simpler molar composition descriptor. The results of principal component analysis, regression models with the hyperparameters optimized using an autoML software and Wasserstein autoencoders are briefly discussed for a possible use in inverse materials design.

1 INTRODUCTION

Perovskites are one of the most versatile and studied groups of compounds with wide-ranging applications in electrocatalysis, clean energy conversion, photocatalysis and photovoltaics (Yin et al., 2019). The atoms in the original perovskite with formula CaTiO$_3$ can be substituted by many other elements which results in an enormous number of possible compositions and structures (the chemical space) with a wide range of properties. One of the most promising lines of research is the photocatalytic activity of perovskites for water splitting (Kumar et al., 2021). It is not feasible to experimentally explore all the element combinations in perovskites in the near future; therefore, theoretical predictions of geometric, electronic and catalytic properties represent a burgeoning field of ongoing research (Gvozdetskyi et al., 2023; Tao et al., 2021).

In the last 10 years, there has been a growing number of available open-access databases containing experimentally measured and synthetized materials as well as hypothetical compounds. Among them, prominent examples are the NIST Inorganic Crystal Structure Database and the Materials Project database that are on track to provide information about a million inorganic materials (Levin, 2018; Jain et al., 2013). However, the growing number of entries represents a challenge in data governance, curating data, ensuring data consistency and accuracy. This hinders efforts in applying machine learning and artificial intelligence techniques in the predictions of unknown compositions.

Many of the state-of-the-art computational studies probe the idealized pure perovskite structures and predict single, cubic, crystal geometries while experimentalists study a wide range of non-integer partial perovskites substitutions. For example, Talapatra et al. (2021) focused on perovskites with cubic symmetry and structures with less than 0.50 meV energy above the convex hull calculated using density functional theory (DFT) in a recent encouraging study of synthesizability and formability of perovskites. This allowed the authors to probe a well-defined search space exhaustively, however, this also limits the insights to the structurally ideal compounds and omits a high number of diverse compounds in the process. Pilania et al. (2015) used similar tree-based classification models and reported encouraging results on a more modest experimental dataset with 354 samples achieving about 95% average accuracy for the classification task using geometrical descriptors.

For this work, we compiled a specific dataset of single- and double-perovskite compounds with reported experimental bandgap values and promising catalytic properties for the water splitting reaction. This dataset opened a unique opportunity for a quantitative study of this narrow group of materials going beyond the predictions of formability and synthesizability. Next, we developed 3 types of descriptors and highlighted their performance for the bandgap prediction. Finally, we demonstrated how they can be visualized and transformed for future efficient inverse design applications.

2 Data collection and descriptors

While the ideal single perovskite oxides have a general formula ABO_3 and a highly symmetrical cubic structure, double perovskites have A and B sites each occupied by up to two different cations. The charge and size difference between the elements in the B site determines the arrangement in the lattice with the rock salt structure being the most prevalent. In total we have collected experimental information related to about 85 perovskite-like compounds with promising properties for the water splitting reaction with the references included in the dataset file (see Appendix A.1). The most abundant elements in our data set were iron, bismuth and barium (see Figure 2). In total, 54 compounds had a unique chemical formula and experimental bandgap value which we decided to investigate further. The bandgap values ranged between 1.0 and 4.5 eV and this important property for many applications has been used to test different descriptors for an initial regression task. The parameters and description of the PCA, Wasserstein autoencoders, cross validation, and applied regression methods is included in Appendix A.2.

2.1 Descriptors

We have compared the performance of several classes of descriptors for regression tasks which are described below. We have started with two baseline descriptors: molar and weight composition. To calculate these, we have used only the information about elements present in the A-site, B-site, chemical formula and elemental weight.

Next, we have investigated the use of various fundamental elemental properties in the formulation of the descriptor for perovskite materials. Some of these elemental properties were studied in the past, *e.g.* ionic radii has been used for calculation of geometrical criteria such as tolerance and octahedral factors (see formulas 1–4 in Appendix A.3). To augment these geometrical criteria, we gathered 8 elemental properties: Shannon's ionic radius, ideal bond length, electronegativity, van der Waals radius, ionization energy, molar volume, atomic number, and atomic mass. We have used the pymatgen API (Ong et al., 2013) to collect the data for up to two elements in the A-site (A and A') and two elements in the B-site (B and B'). These values were scaled by the molar composition of the given element which resulted in a descriptor with $8 \times 4 = 32$ features. Next, we decided to investigate multi-site features. We calculated them as products (AB and $A'B'$) and ratios (A/B and A'/B') of the properties listed above, which resulted in 32 additional features per material.

3 Results and discussion

We decided to test the simplest molar and weight composition descriptors first. These two descriptors establish our baseline regression performance against which we can evaluate our more complex descriptors. The weight composition descriptor has an inherently higher variance than the molar composition descriptor since any two materials with the same stoichiometry containing different elements will have the same molar composition but a different weight composition. In Figure 3,

Table 1: Models and feature processing optimized for different descriptors using TPOT autoML package. XGB stands for extreme gradient boosting (XGB) regressor. Max. abs. scaling is Maximum absolute scaling. Pearson corr. coef. is Pearson correlation coefficient. Detailed information about the models used can be found in scikit-learn documentation (Pedregosa et al., 2017).

			Pearson corr. coef.	
Descriptor (# of features)	**Model**	**Feature preprocessing**	**Training**	**Testing**
Weight composition (39)	Random forest	Stacking estimator	0.99	0.41
Molar composition (40)	Adaptive boosting	Stacking estimator	0.92	0.26
Elemental descriptor (64)	XGB	One hot encoding	1.00	0.61
Elemental descriptor (48)	XGB	Max. abs. scaling + Stacking estimator	0.97	0.62
Elemental descriptor (32)	XGB	Max. abs. scaling	0.98	0.55

we can see that the features within both composition descriptors are not correlated with each other, which is because these descriptors are rather simple, sparse and lack chemical insight into material properties. The molar composition of oxygen is a constant 3/5 for all compounds and has been omitted. In the case of both composition descriptors, only 2, 3 or 4 values are non-zero. This results in a rather poor performance in the regression task with achieved 0.41 and 0.26 Pearson correlation coefficients on the testing set for weight and molar composition descriptors, respectively (Table 1). The weight composition descriptor achieved a marginally better result than the molar composition descriptor which might be a reason why it is used in certain domains of materials design, e.g. alloy discovery (Rao et al., 2022). The models and feature preprocessing methods were optimized using the Tree-Based Pipeline Optimization Tool (TPOT) autoML package and are listed in Table 1.

Next, we have investigated the correlation of the different elemental features in Figure 4. Since there are only 2 examples of alloy double perovskites which have two different elements occupying the B-site and A-site at the same time, the correlations calculated for the sparse A'/B' and $A'B'$ composition features do not bring new insights. However, we can conclude that the correlation of elemental features of the elements in either the A-site or the B-site are above 0.6 (in Figure 4 from the top left, the first four diagonal submatrices). Surprisingly, the A elemental properties are highly correlated with the elemental properties of the second element in the B-site (denoted as B' in Figure 4) and the same is true for elemental properties of the second element in the A-site (A') and the first element in the B-site (B). The high correlation between features indicates that the information encoded in the respective features could be similar and therefore redundant.

The feature generation, that is a process of combining different properties, is an active area of research and it can be used to generate new features or reduce the number of features. A successful example of this is the octahedral factor (ratio of ionic radii of B-site cations and the anion, oxygen) while a less promising one would be e.g. the ratio of atomic number and atomic weight. To study these combinations, we generated features denoted as AB and A/B which are the product and ratio of the elemental properties of the first elements in the A-site and B-site (see section 2.1).

We have optimized the model using information of the entire descriptor, [A, A', B, B', A/B, AB, A'/B', $A'B'$] first. Let us inspect the regression results optimized using the autoML software TPOT in Figure 5 and Table 1. We can see that although the model shows signs of overfitting when using the descriptor with 64 features, the performance on the test dataset is satisfactory with Pearson correlation coefficient of 0.61. The reason for overfitting here might have been that the autoML software evaluated the one hot encoding as the best performing feature preprocessing for the training set which results in a high number of features and we are dealing with few data points (or perovskites). To address this, one can use the developed descriptors and ML models here described in an iterative framework for material discovery using the uncertainty for decision making, Bayesian optimization or active learning (Lourenço et al., 2022). Through the acquisition of new perovskite examples one might converge to a better ML model. If we drop the synthetic features $A'B'$ and A'/B' (graph B in Figure 5), the overfitting is less severe while maintaining decent correlation of 0.62. However, stripping the descriptor of remaining synthetic features AB and A/B slightly decreases the performance to 0.55. For all three variants and sizes of the elemental descriptor, the extreme gradient boosting regressor performed the best while the test results were rather robust in the choice of the feature manipulations resulting in similar performance of one hot encoding and maximum absolute scaling on these data sets.

If we add also tolerance factors calculated according to equations 1–4 the regression model performance does not improve (see Figure 6) despite the fact these geometrical factors are only weakly correlated with each other as shown in Figure 7. This can be explained by the fact that the tolerance factors are calculated using the ionic radii information which is already present in the previously described descriptor and which does not show larger a contribution or importance as is indicated by the feature importance results (in Figure 8). The product of atomic numbers of A and B site is among the most important features with the importance close to 0.20 for both methods. This indicates that adding electronic information e.g. number of d-electrons might improve the models further. On average, A-site information is less important than B-site information which reflects the fact that the data set is focused on double perovskites with doped B-site.

Finally, we have used two unsupervised techniques, Wasserstein autoencoders (WAE) and principal component analysis (PCA), to visualize our dataset when the elemental descriptor with 64 features is used. The preliminary results show that autoencoders can cluster our materials according to the cationic content rather well (see Figure 1, left). We have investigated how the size of the descriptor changes the principal components next. When we reduce the number of features which are used as an input for PCA, the representation or latent space rotates but remains qualitatively similar (see Figure 9). Surprisingly, a visually larger change can be observed when sparse features $A'B'$ and A'/B' are neglected. However, the results discussed above are not sufficient to comment on which method provides a more efficient representation of the latent space and a careful evaluation and hyperparameter optimization of WAE for inverse design applications(Ren et al., 2022) is planned in the near future.

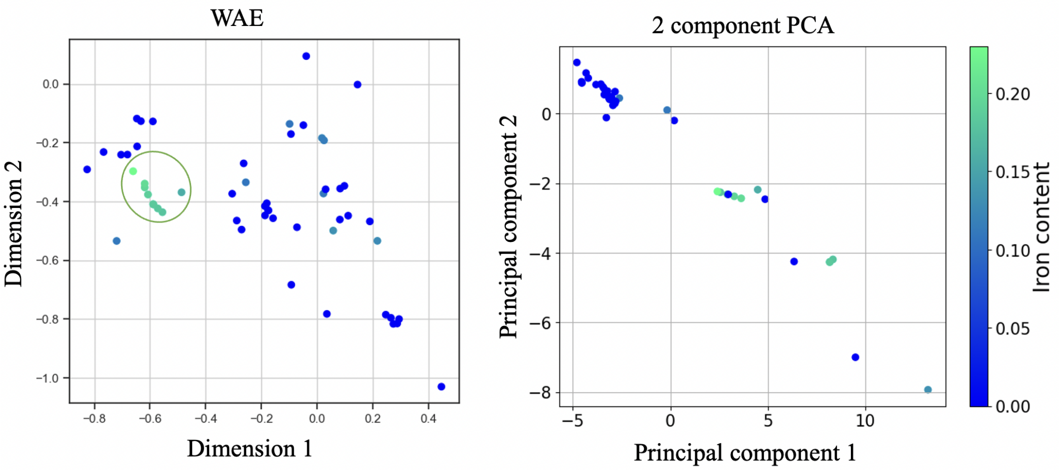

Figure 1: Wasserstein autoencoder results (WAE) on the left and Principal Component Analysis (PCA) on the right, the standard scaling was applied on 64 elemental features.

4 Conclusion

We compared different families of descriptors for the regression and classification tasks for alloy perovskite oxide materials with a general formula $A_xA'_{1-x}B_yB'_{1-y}O_6$. The weight composition descriptor showed superior results over a simpler molar composition descriptor. The best performance for regression tasks was observed for our elemental descriptor which consisted of A-site and B-site values of: Shannon's ionic radius, ideal bond length, electronegativity, van der Waals radius, ionization energy, molar volume, atomic number, and atomic mass. Wasserstein autoencoders showed promising results for the use in inverse design.

REFERENCES

Leo Breiman. Random forests. *Machine Learning*, 45:5–32, 2001.

Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pp. 785–794, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450342322. doi: 10.1145/2939672.2939785.

Volodymyr Gvozdetskyi, Balaranjan Selvaratnam, Anton O. Oliynyk, and Arthur Mar. Revealing hidden patterns through chemical intuition and interpretable machine learning: A case study of binary rare-earth intermetallics rx. *Chemistry of Materials*, ASAP, 2023. doi: 10.1021/acs.chemmater.2c02425.

Lauri Himanen, Marc O.J. Jäger, Eiaki V. Morooka, Filippo Federici Canova, Yashasvi S. Ranawat, David Z. Gao, Patrick Rinke, and Adam S. Foster. Dscribe: Library of descriptors for machine learning in materials science. *Computer Physics Communications*, 247:106949, 2020.

Jiri Hostas, Maicon Pierre Lourenço, John Garcia, Hatef Shahmohamadi, Alain Tchagang, Karthik Shankar, Venkataraman Thangadurai, and Dennis R. Salahub. Compositional and elemental descriptors for perovskite materials. In *Workshop on "Machine Learning for Materials" ICLR 2023*, 2023. URL `https://openreview.net/forum?id=Q5KVnwFUrkH`.

Anubhav Jain, Shyue Ping Ong, Geoffroy Hautier, Wei Chen, William Davidson Richards, Stephen Dacek, Shreyas Cholia, Dan Gunter, David Skinner, Gerbrand Ceder, and Kristin A. Persson. Commentary: The materials project: A materials genome approach to accelerating materials innovation. *APL Materials*, 1(1):011002, 2013.

Pawan Kumar, Suresh Mulmi, Devika Laishram, Kazi M. Alam, Ujwal K Thakur, Venkataraman Thangadurai, and Karthik Shankar. Water-splitting photoelectrodes consisting of heterojunctions of carbon nitride with a p-type low bandgap double perovskite oxide. *Nanotechnology*, 32(48): 485407, 2021.

Trang T Le, Weixuan Fu, and Jason H Moore. Scaling tree-based automated machine learning to biomedical big data with a feature set selector. *Bioinformatics*, 36(1):250–256, 2020.

Igor Levin. NIST inorganic crystal structure database, NIST standard reference database number 3, national institute of standards and technology, gaithersburg md, 20899. doi: 10.18434/M32147, 2018. Accessed: 2023-02-08.

Maicon Pierre Lourenço, Lizandra Barrios Herrera, Jiří Hostaš, Patrizia Calaminici, Andreas M. Köster, Alain Tchagang, and Dennis R. Salahub. Automatic structural elucidation of vacancies in materials by active learning. *Phys. Chem. Chem. Phys.*, 24:25227–25239, 2022.

Shyue Ping Ong, William Davidson Richards, Anubhav Jain, Geoffroy Hautier, Michael Kocher, Shreyas Cholia, Dan Gunter, Vincent Chevrier, Kristin A. Persson, and Gerbrand Ceder. Python materials genomics (pymatgen) : A robust, open-source python library for materials analysis. *Computational Materials Science*, 68:314–319, 2013. doi: doi:10.1016/j.commatsci.2012.10.028.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Fabian Pedregosa, Gael Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Edouard Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12: 2825–2830, 2017.

Gubernatis Pilania, Prasanna Venkataraman Balachandran, James E. Gubernatis, and Turab Lookman. Classification of abo3 perovskite solids: a machine learning study. *Acta Crystallographica Section B*, 71(5):507–513, 2015.

Ziyuan Rao, Po-Yen Tung, Ruiwen Xie, Ye Wei, Hongbin Zhang, Alberto Ferrari, T.P.C. Klaver, Fritz Körmann, Prithiv Thoudden Sukumar, Alisson Kwiatkowski da Silva, Yao Chen, Zhiming Li, Dirk Ponge, Jörg Neugebauer, Oliver Gutfleisch, Stefan Bauer, and Dierk Raabe. Machine learning enabled high-entropy alloy discovery. *Science*, 378(6615):78–85, 2022.

Zekun Ren, Juhwan Tian, Siyu Isaac Parkerand Noh, Felipe Oviedo, Guangzong Xing, Jiali Li, Qiaohao Liang, Ruiming Zhu, Armin G. Aberle, Shijing Sun, Xiaonan Wang, Yi Liu, Qianxiao Li, Senthilnath Jayavelu, Kedar Hippalgaonkar, Yousung Jung, and Tonio Buonassisi. An invertible crystallographic representation for general inverse design of inorganic crystals with targeted properties. *Matter*, 5:314–335, 2022.

Anjana Talapatra, Blas P. Uberuaga, Christopher R. Stanek, and Ghanshyam Pilania. A machine learning approach for the prediction of formability and thermodynamic stability of single and double perovskite oxides. *Chemistry of Materials*, 33(3):845–858, 2021. doi: 10.1021/acs.chemmater.0c03402.

Qiuling Tao, Pengcheng Xu, Minjie Li, and Wencong Lu. Machine learning for perovskite materials design and discovery. *npj Comput Mater*, 7:23, 2021.

Ilya Tolstikhin, Olivier Bousquet, Sylvain Gelly, and Bernhard Schoelkopf. Wasserstein autoencoders, 2017. URL `https://arxiv.org/abs/1711.01558`.

Guido Van Rossum and Fred L. Drake. *Python 3 Reference Manual*. CreateSpace, Scotts Valley, CA, 2009. ISBN 1441412697.

Logan Ward, Ankit Agrawal, Alok Choudhary, and Christopher Wolverton. A general-purpose machine learning framework for predicting properties of inorganic materials. *npj Computational Materials*, 2:16028, 2016.

Wes McKinney. Data Structures for Statistical Computing in Python. In *Proceedings of the 9th Python in Science Conference*, pp. 56 – 61, 2010. doi: 10.25080/Majora-92bf1922-00a.

Wan-Jian Yin, Baicheng Weng, Jie Ge, Qingde Sun, Zhenzhu Li, and Yanfa Yan. Oxide perovskites, double perovskites and derivatives for electrocatalysis, photocatalysis, and photovoltaics. *Energy Environ. Sci.*, 12:442–462, 2019.

Quan Zhoua, Peizhe Tanga, Shenxiu Liua, Jinbo Panb, Qimin Yanb, and Shou-Cheng Zhang. A general-purpose machine learning framework for predicting properties of inorganic materials. *PNAS*, 115(28):E6411, 2018.

A Appendix

Acknowledgements

We are grateful to Compute Canada/WestGrid (now the Digital Research Alliance of Canada) for computational resources. Work supported by the National Research Council of Canada, Artificial Intelligence for Design program and by the Natural Sciences and Engineering Research Council of Canada, Discovery Grant (RGPIN-2019-03976).

A.1 Database

Database file which includes references to the experimental works can be downloaded at:

`https://github.com/jiri-hostas/EDA-and-ML-for-Perovskites`

A.2 Methods: Exploratory data analysis, regression models and classification models

The initial data collection and data management were done using python version 3.9, pandas, scikit-learn and chemparse (Van Rossum & Drake, 2009; Wes McKinney, 2010; Pedregosa et al., 2017). We explored the performance of popular regression methods using an automated Tree-Based Pipeline Optimization Tool (TPOT) written in python Le et al. (2020). This software optimizes machine learning pipelines using genetic programming. We have used default settings, leave one out cross validation and 10 generations with the size of population 50 which resulted in the evaluation of 550 ML pipeline configurations for each descriptor.

Since our data set size was limited to 54 compounds, we were not able to afford to use a separate validation which will be addressed in the future by collecting more data by combining theoretical and experimental data. We will focus on the discussion of results for two tree-based methods, extreme gradient boosting and random forest (Chen & Guestrin, 2016; Breiman, 2001).

We have compared the results of principal component analysis (PCA) and Wasserstein autoencoders (WAE) (Tolstikhin et al., 2017). While the former method is used primarily for visualization and data dimension reduction, the latter is a new class of algorithms for building generative models. We have used both methods to visualize the datasets. PCA is readily available in scikit-learn and we followed the WAE implementation by Rao et al. (2022) written using the python package pytorch (Pedregosa et al., 2017; Paszke et al., 2019)

A.3 Geometrical criteria for perovskite formability

In the literature, there exist different geometrical criteria for perovskite formability and those listed below were tested: the tolerance factor calculated as

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

and octahedral factor

$$r = \frac{\frac{r_A}{r_X} + 1}{\sqrt{2(r+1)^2 + \Delta r^2)}} \tag{2}$$

, where r$_A$ and r$_B$ are ionic radii of the cations and r$_X$ is the ionic radii of the anion, oxygen. These factors can be generalized for double-perovskites resulting in the generalized tolerance factor Pilania et al. (2015):

$$r = \frac{r_B + r_{B'}}{2r_X} \tag{3}$$

and the octahedral mismatch:

$$\Delta r = \frac{r_B - r_{B'}}{2r_X}. \tag{4}$$

These aforementioned geometrical criteria were used as descriptors in this work. They were calculated as combinations of ionic radii of the anions and cations present in the perovskite lattice. To explore descriptors which could be derived using similar feature combinations, we decided to augment the A, A', B and B' elemental descriptors with the product and ratios of these properties (AB, A'B', A/B and A'/B'). This approach is similar to the other one-dimensional vector descriptors such as Magpie Ward et al. (2016), where one encodes physical, chemical, electronic, ionic and basic properties of the material, with the exception that we generate such a vector for each A or B site separately.

There is a large number of descriptors that can be used to encode 3D arrangement in the crystal lattice (such as Orbital Field Matrix (Zhoua et al., 2018), Sine Matrix, Many-body Tensor Representation and others (Himanen et al., 2020)). However, it was not possible to apply them here without making serious assumptions about the structure of the materials since there is only limited structural information about the majority of the materials studied in this work.

A.4 Figures

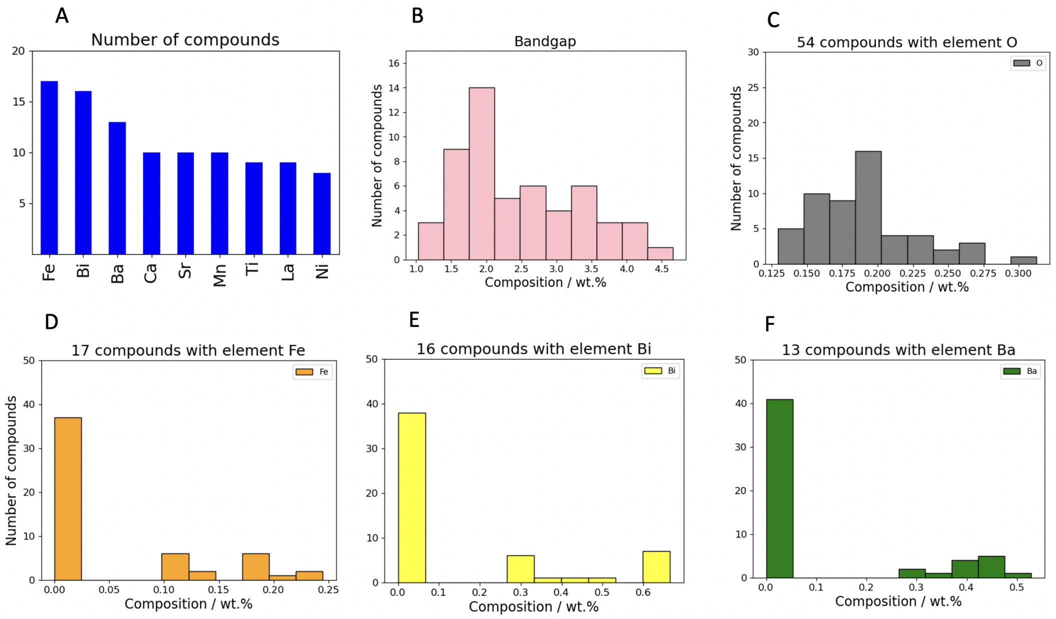

Figure 2: Dataset characteristics: A) Number of perovskites containing the given element; B) Distribution of experimental band gap values in [eV]; and distribution of weight composition for oxygen, iron, bismuth and barium elements are shown in C, D, E and F, respectively.

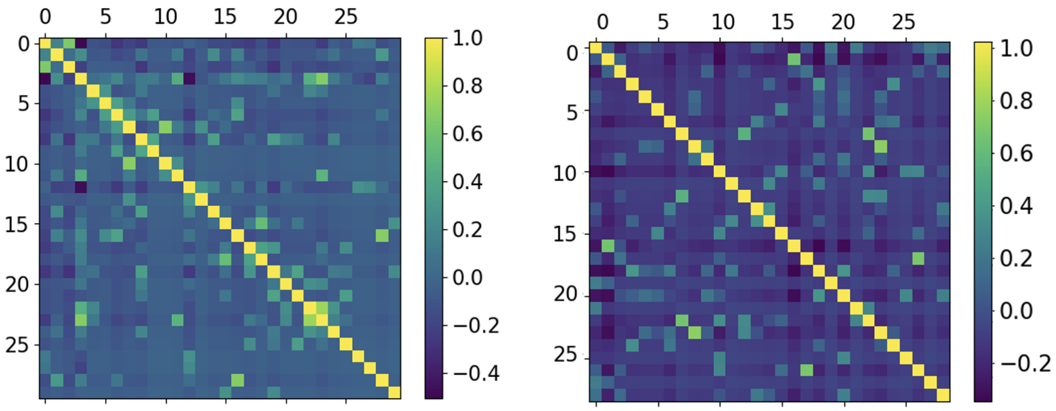

Figure 3: Correlation matrices for weight (left) and molar (right) composition descriptors with 30 and 29 features, respectively. Oxygen molar composition feature is a constant 3/5 for all perovskites, so it has been omitted.

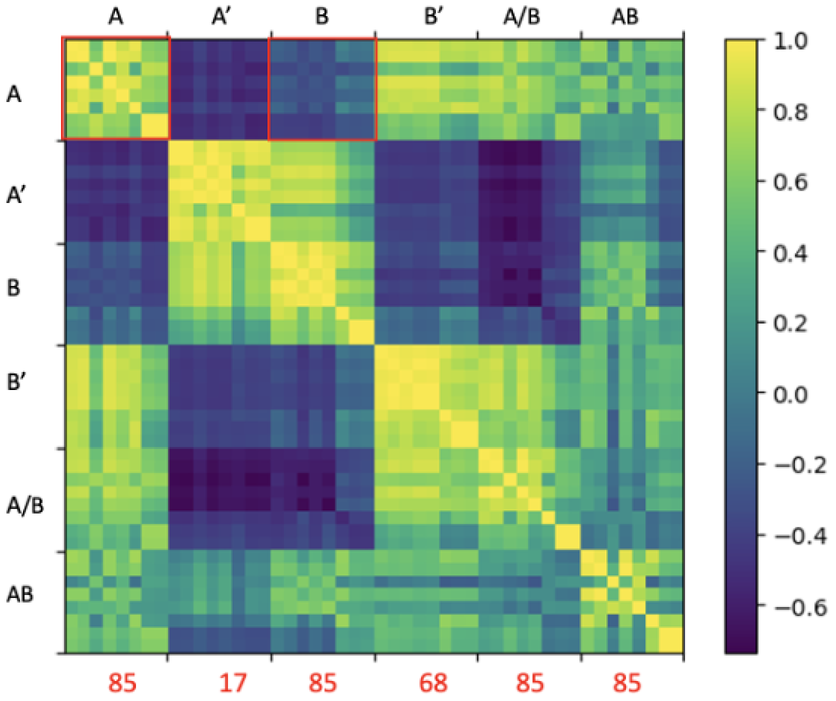

Figure 4: Correlation matrix for features in the elemental descriptors in the order: A, A', B, B', A/B, AB. The values in red under the matrix are the numbers of non-zero values of the given feature type for 85 perovskites.

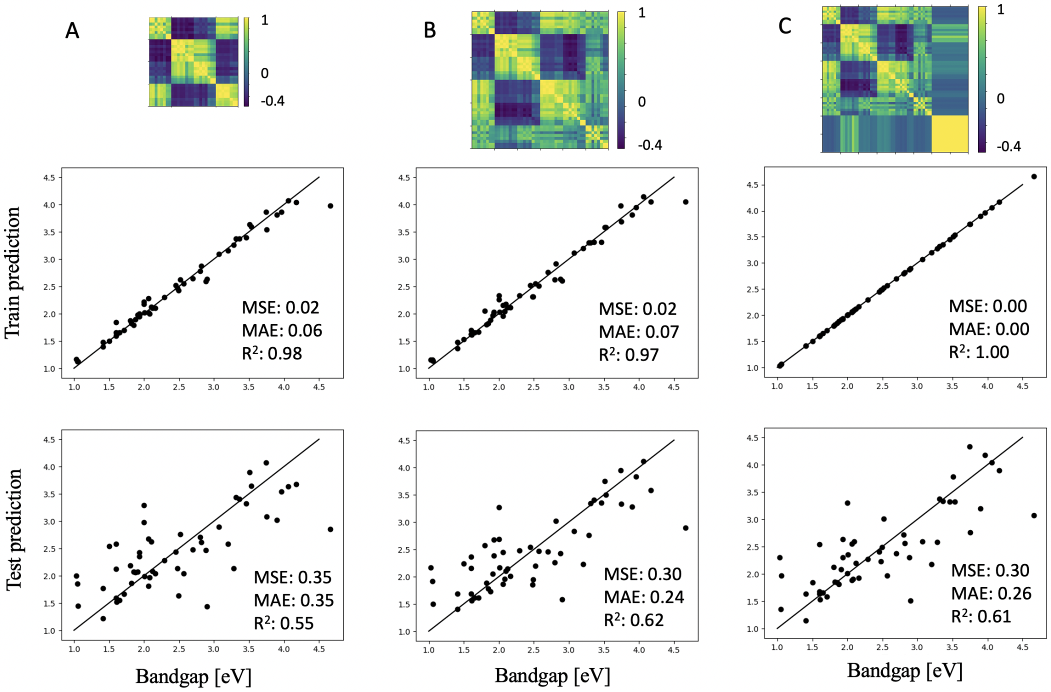

Figure 5: Regression model results for experimental bandgap prediction. We used different number of elemental descriptor features, from the left to the right: 32, 48 and 64 (for more information about the choice of elemental descriptor see section 2.1)

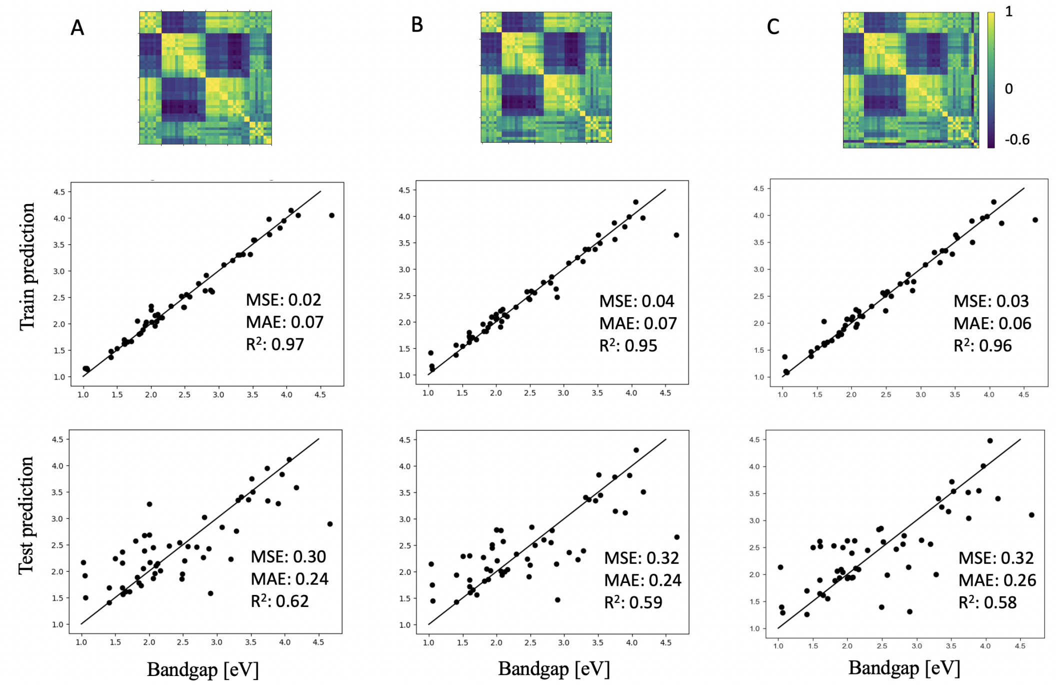

Figure 6: Regression models results for experimental bandgap prediction. Elemental descriptor with 48 features on the left and the effect of including tolerance+octahedral factor (in the centre) and also general tolerance factor and octahedral mismatch. See sections 2.1 and Appendix A.3 for more information.

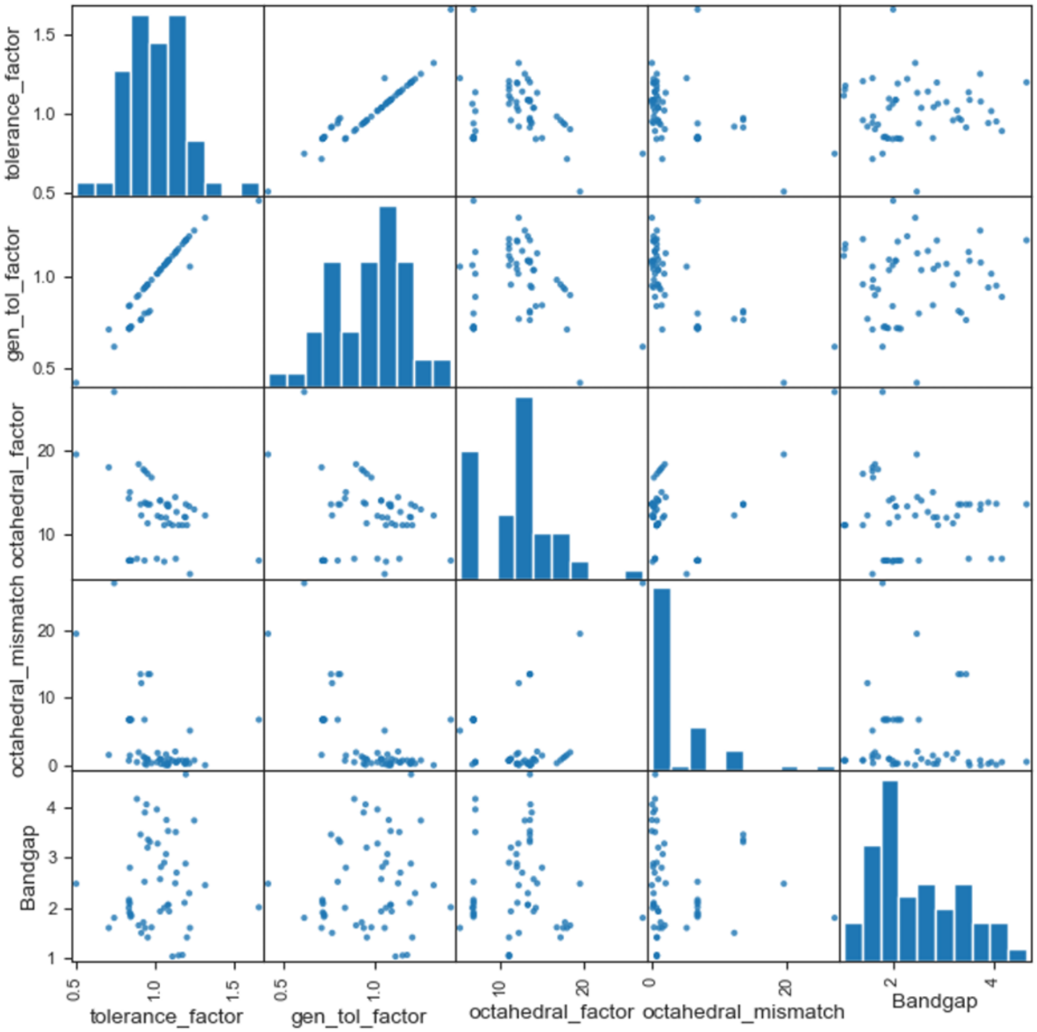

Figure 7: Correlations between tolerance factor, generalized tolerance factor, octahedral factor, octahedral mismatch and the experimental bandgap (from the left to the right and the top to the botom).

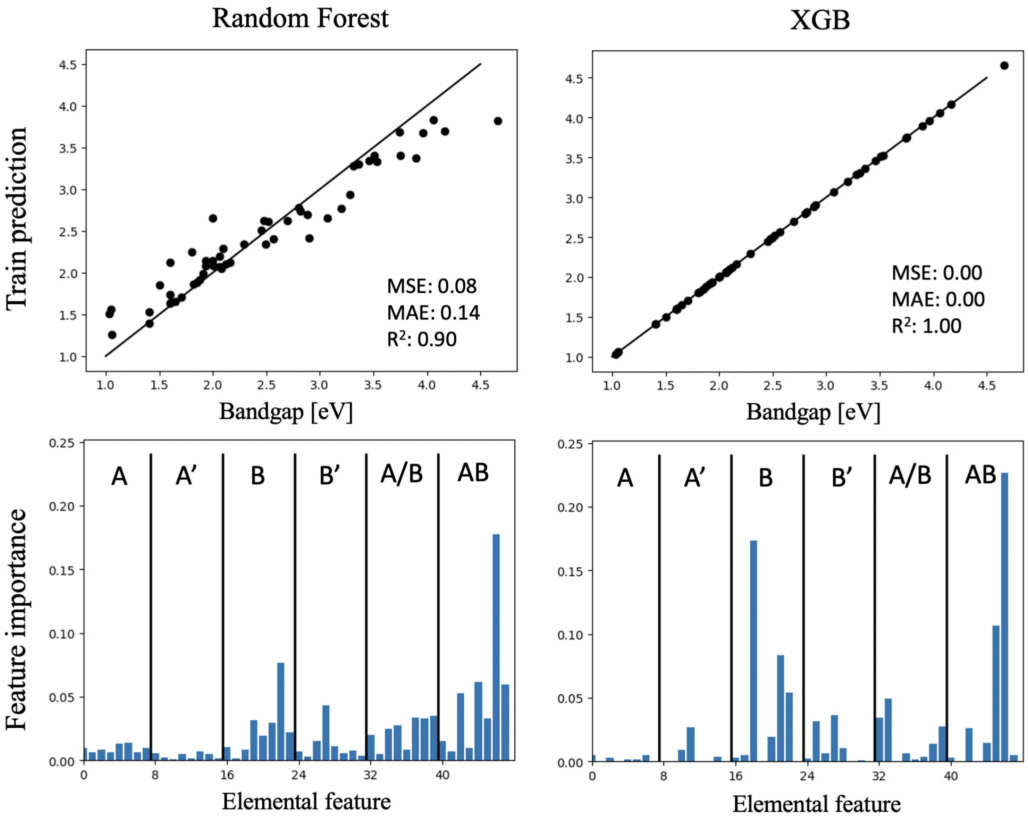

Figure 8: Feature importance results for random forest and extreme gradient boosting methods (XGB). The product of atomic numbers of A and B sites is among the most important features with the importance close to 0.20 for both methods. On average, A-site information is less important than B-site information. Due to the small data set size, overfitting is the main challenge here.

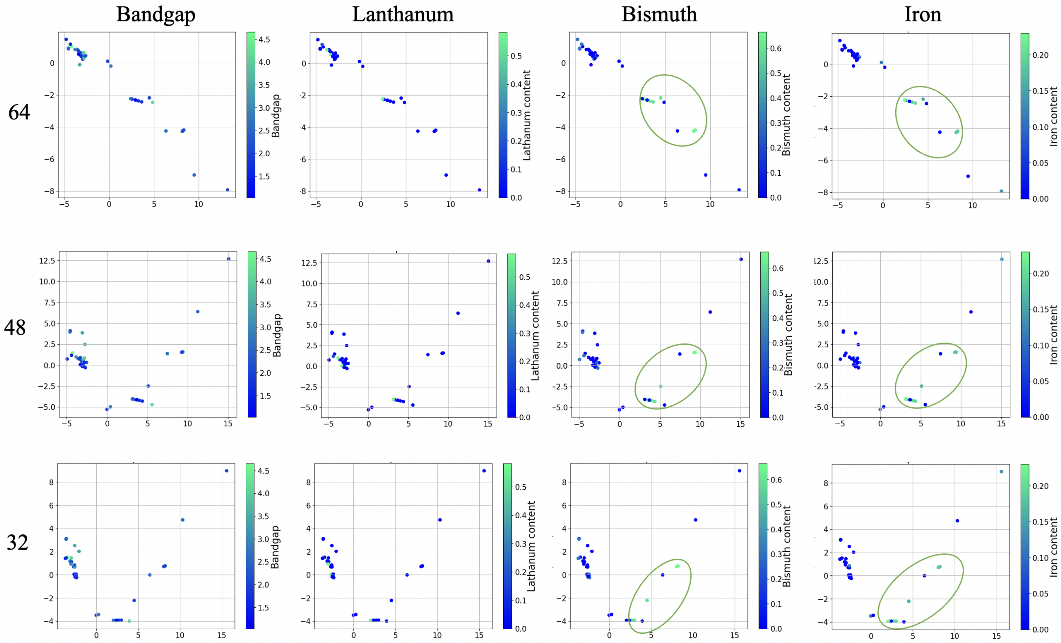

Figure 9: PCA results for elemental descriptors with 64, 48 and 32 elemental features (described in section 2.1) in the rows from the top to bottom. In the columns, the data points are colored according to content of the various elements. Principal components 1 and 2 are on the horizontal and vertical axis, respectively.