Crystal-Structure Descriptor for Binary Materials Based on Coordination Numbers

Aryan Agal, Joakim Kattelus, Nikita Tepliakov
Ecole Polytechnique Fédérale de Lausanne
Course: CSS-433 Machine Learning

Abstract— **In this Project, we developed an original crystal-structure descriptor for binary compounds that describes chemical bonding in a given material through the coordination numbers of atoms. This descriptor is rotationally invariant, independent of the choice of a unit cell, and has a fixed length, which makes it attractive for applications in machine learning. We used the developed descriptor to train models based on the neural network and gradient boosted trees method, using approximately 19,000 binary materials from the Materials Project database. Our model performs classification of binary materials as conductors or insulators with the accuracy of 90.2% and predicts their formation energies within 0.83 eV/atom.**

I. Introduction

Despite the development of computational facilities, the precise modelling of crystal materials still takes a large amount of time and cannot be used for rapidly covering a wide range of structures. This is an area where *machine learning* can significantly outperform the conventional techniques. In particular, machine learning can be used to predict formation energies of crystals, guide chemical experiments, or search the materials with nontrivial topological properties [1].

Implementation of machine learning algorithms in materials science requires *crystal-structure descriptors*, which would transform crystal structures into readable input for machine learning models. One of the simplest descriptors is based on the Coulomb matrix [2], whereas a more sophisticated one explicitly takes into account angles between chemical bonds [3]. A characteristic feature of these descriptors is that their length depends on the number of atoms per unit cell, which only allows to train models for a specific class of materials.

In this Project, we aimed at developing a simple crystal-structure descriptor for binary materials that would contain all the essential information on the atomic environments and be applicable to the structures of an arbitrary number of distinguishable atoms. We proposed a descriptor based on coordination numbers and used it to predict conduction properties and formation energies of binary materials from the Materials Project database [4]. Our results show that a descriptor based on coordination numbers is a fast and reliable way of describing crystal structures for machine learning applications.

II. Crystal-Structure Descriptor

A given crystal-structure descriptor must always satisfy a number of general symmetry requirements. First of all, it should be *rotationally invariant*, that is, it must not depend on the orientation of a unit cell with respect to the coordinate axes. This requirement stems from a more general difficulty for neural networks to deal with the rotated data [5]. Another requirement is that the descriptor must not depend on the choice of a unit cell, which is often made arbitrary and bears no physical meaning. Finally, it is also desirable that the descriptor is independent of the number of atoms per unit cell, which varies between different materials.

In order to satisfy the latter condition, we focus on *binary compounds* $\mathrm{A}_x\mathrm{B}_y$, which are a giant class of materials comprising approximately $19,000$ structures in the Materials Project database. First and foremost, we have to choose a way of encoding species of A and B atoms in the form of variables. The most straightforward approach is to characterise an element with its number in the periodic table. A disadvantage of this approach is that it does not provide much physical sense and

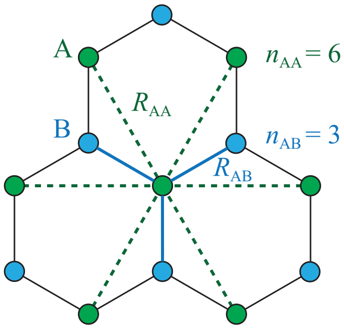

Fig. 1. Example of NNs choice in a two-dimensional honeycomb lattice. Each A atom has three NNs of type B at distance R_{AB} and six NNs of type A at distance R_{AA}. The situation is symmetrical for atoms of sort B.

may lead to the wrong fitting by machine learning models.

Instead, we shall use the period P and group G of an element in the periodic table. This is a natural way of taking into account the periodicity of atomic chemical properties, which are determined by the number of valence and core electrons. This approach might fall behind in case of lanthanides and actinides, which belong to the same period and group but differ in the number of f-electrons. However, these electrons do not contribute significantly to the formation energies [6], which allows us to disregard them in our analysis.

We thus get four input parameters for a binary material: P_{A}, P_{B}, G_{A}, and G_{B}. It is convenient to replace them with their linear combinations $A_{\pm} = P_{\mathrm{A}} \pm G_{\mathrm{A}}$ and $B_{\pm} = P_{\mathrm{B}} \pm G_{\mathrm{B}}$. One can see that A_{+} and B_{+} are always different for distinct elements, which is useful for symmetrising the descriptor, as will be shown further.

In order to obtain a crystal-structure descriptor of a fixed length, we suggest a formalism based on averaged coordination numbers. In high-symmetry crystal structures each lattice site is surrounded by a number of symmetrically-equivalent atoms, the nearest neighbours (NNs), located at the same distance from the reference point. What is important is that the main properties of crystal structures can often be calculated assuming that each atom only interacts with its NNs.

The number of NNs is called the coordination number n of a given atom. We argue that this number carries enough information on the spatial environment of each atom and can be used instead of explicitly considering angles between chemical bonds. In order to capture all the interactions in binary materials, we use four different coordination number: n_{AA}, n_{AB}, n_{BA}, and n_{BB}. For example, n_{AB} is defined as the number of NNs of type B surrounding atom A. The corresponding distances to the NNs are given by R_{AA}, R_{AB}, R_{BA}, and R_{BB}. The choice of these parameters is illustrated in Fig. 1.

Finally, we introduce parameters U_{A} and U_{B} showing the numbers of given atoms per unit cell. Due to the symmetry breaking in distorted structures, different atoms of the same sort may have different environment, which yields different coordination numbers and distances. We average this parameters for each atom type, by summing the values for the distinct atoms within the unit cell and dividing it by U_{A} or U_{B}.

In total, we have 14 features in our crystal-structure descriptor: seven parameters for the A atom, $\mathbf{x}_{\mathrm{A}} = \{A_{+}, A_{-}, U_{\mathrm{A}}, n_{\mathrm{AA}}, n_{\mathrm{AB}}, R_{\mathrm{AA}}, R_{\mathrm{AB}}\}$, and similar seven features $\mathbf{x}_{\mathrm{B}}$ for the atom of sort B. Using these parameters straight away may lead to inaccurate predictions, because this descriptor changes upon swapping atoms A and B in a database entry. We thus use a symmetrised descriptor, which is defined as

$$\mathbf{x}_{+} = \mathbf{x}_{\mathrm{A}} + \mathbf{x}_{\mathrm{B}}, \tag{1}$$

$$\mathbf{x}_{-} = (\mathbf{x}_{\mathrm{A}} - \mathbf{x}_{\mathrm{B}}) \, \mathrm{sign}(A_{+} - B_{+}). \tag{2}$$

Here we used the previously mentioned fact that A_{+} and B_{+} are always different for binary materials. The resulting descriptor $\mathbf{x} = \{\mathbf{x}_{+}, \mathbf{x}_{-}\}$ is symmetric with respect to replacement $\mathrm{A} \leftrightarrow \mathrm{B}$ and does not lose any information upon the linear transformations.

III. Machine Learning Models

In this part, we use the developed crystal-structure descriptor to predict various physical properties of binary materials in the Materials Project database [4], which contains 18,962 binary materials. In the first part, we aim at binary classification of materials as conductors or insulators according to their bandgap. In the second part we want to predict a continuous value of the formation energy per atom E_{at}.

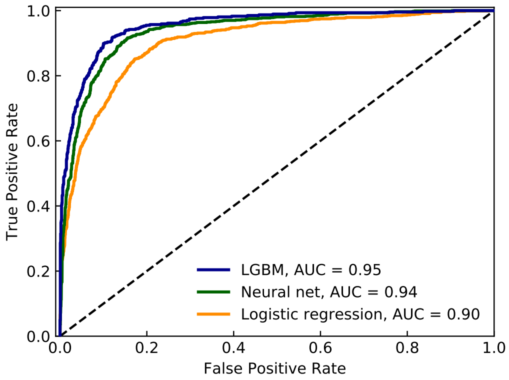

Fig. 2. Receiver operating characteristic (ROC) curve for the logistic regression, neural network, and gradient boosted trees method with the corresponding area under curve (AUC) parameters.

In both of these task, we rely on two methods: neural network implemented in TensorFlow [7] and gradient boosted trees method with LightGBM (LGBM) [8]. Each neural network has ReLU activations, batch normalisation, and a dropout with probability of 50%, following each hidden layer. For both neural nets and LGBM, cross validation was done by splitting the data with a 80-20 train-test split. For LGBM, the test set was further split into a 50-50 test and holdout set, as the test set was used for hyperparameter optimisation. In the beginning we also experimented with normalising the input from the descriptor, but this did not improve the results.

A. Conductor-Insulator Classification

As a baseline for classifying materials as either conducting or insulating, we implement the logistic regression. We trained it with 5-Fold Cross Validation on 80% of data, achieving the accuracy of 86% on the test set and area under curve (AUC) parameter of 0.90. In Fig. 2 we plot the receiver operating characteristic (ROC) for the logistic regression, as well as the other two methods.

We then trained a neural network with 3 hidden layers, each containing 50 nodes, using the Adam optimiser [9], cross entropy loss, and learning rate of 1.5×10^{-4} over 200 epochs. Weights were initialised using Kaiming initialisation [10] and biases as zeros. We achieved a classification accuracy of 88.8% and AUC of 0.94 on the test set. Training with different learning rates for over 500 epochs did not lead to any notable improvements in the prediction accuracy, while significantly slowing down the training process.

Next we trained a LGBM model, which minimises the cross entropy log loss. We used a learning rate of 0.005, 85% sub-sampling of the training data and 75% of the features for each tree. The trees had a maximum depth of 10, maximum 100 leaves, and minimum 50 observations per leaf. The number of trees was optimised using 5-fold cross validation. LGBM achieved the best accuracy of 90.2% and AUC of 0.95 on the holdout set.

B. Prediction of Formation Energies

Next we focus on predicting energies per atom $E_{\rm at}$ in the binary compounds. We start with a linear regression model as a baseline. The values of $E_{\rm at}$ in the input data are characterised by the mean value of $\bar{E}_{\rm at} = -5.3$ eV/atom and the standard deviation of $\Delta E_{\rm at} = 2.4$ eV/atom. Using the linear regression, we achieved the root mean square error of 1.9 eV/atom. Since this accuracy is comparable to the deviation of entries in the initial data array, the linear model does not provide sufficient accuracy in prediction energies per atom.

However, the linear model is instructive for analysing the dependencies of atomic energies on the crystal-structure parameters. Making an inverse transformation from Eq. (1) and assuming that element A is heavier than element B, we get the following list of coefficient for each crystal-structure parameter:

P_A	P_B	G_A	G_B	U_A	U_B	$n_{\rm AA}$	$n_{\rm BB}$	$n_{\rm AB}$	$n_{\rm BA}$	$R_{\rm AA}$	$R_{\rm BB}$	$R_{\rm AB}$	$R_{\rm BA}$
0.4	-2.3	0.3	0.4	0.1	-0.5	-0.1	-0.2	0.1	0.1	0.5	0.6	0.7	0.8

One can see that the atomic energies (that are negative) reduce strongly with increasing the period $P_{\rm B}$ of the lighter atom, which is related to the rapid increase in the total number of electrons in the system. The formation energies also decrease in absolute values with the interatomic distances R, as moving atoms away from each other weakens the chemical bonds. Surprisingly, the coordination numbers have little impact on the energies, indicating that total numbers $U_{\rm A}$ and $U_{\rm B}$ may carry enough information for approximating the atomic energies.

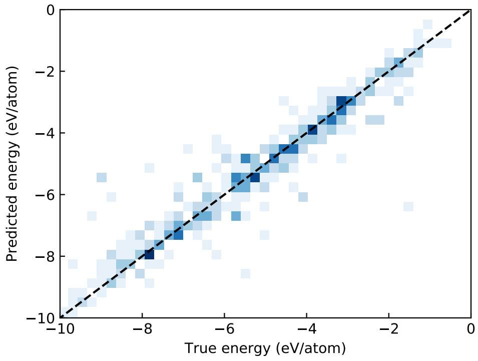

Fig. 3. Predicted energies per atom using LightGBM vs. true values taken from the database. The presented data corresponds to the holdout set of 1888 (10%) entries.

In order to achieve a better accuracy, we trained the same neural network to predict the atomic energies. However, the initialisation was slightly different, the final output layer was initialised with a weight of 0 and a bias equal to the mean of the atomic energies. After training for 500 epochs with the learning rate of 1.5×10^{-4}, we achieved the root mean square error of 0.89 eV/atom on the test set. This is several times smaller than the standard deviation in the input data.

As a final step, we use LGBM with the same parameters as for the binary classification, except that the learning rate is 0.03. LGBM achieved an accuracy of 0.83 eV/atom on the holdout set, which makes it the best model for energy predictions. In Fig. 3 we plot the predicted vs. input energies for this method.

IV. Discussion

We would like to stress that while our predictions are somewhat less accurate than those presented in previous works, this is not a flaw of our machine learning models, but a limitation imposed by the proposed crystal-structure descriptor. The latter has an advantage over more complicated approaches is that it is applicable to any binary compound and uses very simple concepts to convert crystal structures into readable arrays. This allows using this descriptor for the rapid screening of large numbers of materials, identifying classes with desirable physical properties. This quick search can be then followed by more specialised crystal-structure descriptors, which will provide better accuracy within a narrow class of materials.

V. Conclusion

In this Project, we have developed a crystal-structure descriptor for binary materials, which can be applied to materials with an arbitrary number of atoms per unit cell. The code is accessible via the following link. This descriptor encodes the crystal structure with 14 values, which include the periods and groups of elements in the periodic table, coordination numbers, distances to the nearest neighbours, and numbers of atoms per unit cell. Using this descriptor, we trained a model for classifying materials as conductors/insulators, achieving the accuracy of 90.2% with the gradient boosted trees method. In the same framework we predicted formation energies of binary compounds with the final accuracy of 0.83 eV/atom. To conclude, our descriptor provides a simple and reliable way to describe binary materials in a uniform manner, irregardless of their unit cell composition and lattice symmetry.

References

[1] K. T. Butler, *et al.* Machine learning for molecular and materials science. *Nature*, **559**, 547-555 (2018).

[2] K. T. Schütt, *et al.* How to represent crystal structures for machine learning: Towards fast prediction of electronic properties. *Phys. Rev. B* **89**, 205118 (2014).

[3] H. Wang, *et al.* DeePMD-kit: A deep learning package for many-body potential energy representation and molecular dynamics. *Comput. Phys. Commun.* **228**, 178–184 (2018).

[4] A. Jain, *et al.* The Materials Project: A materials genome approach to accelerating materials innovation. *APL Mater.* **1**, 011002 (2013).

[5] G. Cheng, *et al.* Learning rotation-invariant convolutional neural networks for object detection in VHR optical remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **54**, 7405–7415 (2016).

[6] K. A. Gschneidner Jr, *et al.* On the interrelationships of the physical properties of lanthanide compounds: The melting point, heat of formation and lattice parameter(s). *J. Less Common Met.* **17**, 1–12 (1969).

[7] M. Abadi, *et al.* TensorFlow: Large-scale machine learning on heterogeneous systems. *Software* available from tensorflow.org (2015).

[8] G. Ke, *et al.* LightGBM: A highly efficient gradient boosting decision tree. *Adv. Neural Inf. Process. Syst.* **30**, 3146–3154 (2017).

[9] D. Kingma, *et al.* Adam: A method for stochastic optimization. *International Conference on Learning Representations* (2014).

[10] H. Kaiming, *et al.* Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification. *Arxiv* 1502.01852 (2015).