Distributed representations of atoms and materials for machine learning

Supplementary Information for the article "Distributed Representations of Atoms and Materials for Machine Learning"

Luis M. Antunes*[1], Ricardo Grau-Crespo[1], and Keith T. Butler[1,2]

[1] *Department of Chemistry, University of Reading, Whiteknights, Reading RG6 6DX, United Kingdom.*
* l.m.antunes@pgr.reading.ac.uk
[2] *SciML, Scientific Computing Department, Rutherford Appleton Laboratory, Harwell OX11 0QX, United Kingdom.*

Supplementary Notes

1. Comprehensive Results

Supplementary Tables 1 to 10 contain comprehensive results for the experiments described in the article, reporting the performance for all utilized combinations of representation type, embedding size, and pooling type. In all experiments, due to intrinsic limitations of the Atom2Vec approach, Atom2Vec vectors [1] could not be created with dimensions greater than the number of atoms being considered. Similarly, one-hot vectors are limited in dimensionality to the number of atoms being considered. Finally, pre-trained Mat2Vec vectors [2] were used, and their dimensionality was limited to 200. All tasks reported utilized the ElemNet feed-forward neural net architecture (consisting of 17 layers), with L2 regularization instead of dropout.

2. Preliminary Results with Structure-based Architectures

The experiments described in the paper were performed using the ElemNet architecture as a standard (with the exception of the Elpasolite Formation Energy task). We do not experiment with various different kinds of neural network-based architectures because the aim of the work is to introduce a new (and more accessible and effective) way of learning distributed atom representations, and not a particular combination of representation and architecture, nor to establish a new performance benchmark on a task. Nevertheless, here we report preliminary results on the use of SkipAtom embeddings with two different structure-based architectures: CGCNN [3] and MEGNet [4]. These results highlight two important points: first, that SkipAtom embeddings are effective in the context of neural network architectures in general (and not only with an ElemNet architecture), and second, that they can improve the performance of models that incorporate structure information.

The CGCNN model is a convolutional graph neural network that operates on datasets that incorporate crystal structure information. It can be used for classification and regression tasks. The paper that introduced the CGCNN model used a dataset of 27,430 compounds from the Materials Project to build a regression model for predicting band gap. The CGCNN paper reports 0.388 eV MAE. They create a 60/20/20 train/validation/test split: they train on 60% and validate on 20% after each epoch; then they pick the best model according to the validation score, and evaluate on the test set. The 0.388 eV MAE is on the test set. Here, we use the CGCNN codebase [5] to reproduce the results, and to evaluate using SkipAtom vectors as the atom representations. The CGCNN architecture requires that atom representations are provided. By default, a binary feature vector is provided (see [3] for more details). In Supplementary Table 11, we compare the results of using 200-dimensional SkipAtom embeddings to the results of using the default binary feature vectors.

The MEGNet model [4] consists of a graph neural network that can be used to predict properties of molecules and crystals. It requires that atoms are given a predefined representation. Alternatively, one-hot atom vectors can be provided, and an embedding table is learned during training, which results in learned atom representations. Here, we use the MEGNet codebase [6] to compare the performance of the MEGNet model with (and without) the SkipAtom embeddings on the Elpasolite Formation Energy task. In Supplementary Table 12, we compare the results of using 200-dimensional SkipAtom embeddings to the results of using the default one-hot vectors, in the context of the Elpasolite Formation Energy prediction task with the MEGNet model. Note that in the article we report a MAE of 0.1089 eV/atom using the original architecture with concatenated atom vectors (that does not include structure information).

3. Derivation of Materials Graphs

As described in the article, the SkipAtom approach relies on the conversion of the unit cells of materials to a graph representation. From this graph, atom pairs are derived for training. The graph representing a material can be derived using any approach desired, but in this work, an approach is used which is based on Voronoi decomposition [7], which identifies nearest neighbours using solid angle weights to determine the probability of various coordination environments. Specifically, the *CrystalNN* neighbour finding algorithm was used to construct the graphs [8, 9], as implemented in the *pymatgen* package (version 2021.2.8.1) [10].

A brief description of the *CrystalNN* algorithm is provided here for convenience, but for more details, the reader is referred to the original descriptions [8, 9]. The first step in the algorithm involves the assignment of a multi-component weight to each atom pair in the structure, such that these weights correspond to the likelihood that two atoms are neighbours. The weight consists of various components, including the solid angle obtained from a Voronoi construction based on the crystal structure, a penalty for atoms that are too far apart, and the electronegativity difference between the atoms. The next step involves projecting these multi-component weights onto a quadrant of the unit circle, ordered from largest to smallest weights, and computing the area under the circle between adjacent weights to obtain neighbour likelihoods. Finally, the coordination number with the highest probability for each site is selected.

4. Learning Representations of Atoms in their Oxidation States

As stated in the article, one limitation of the SkipAtom approach is that it does not provide representations of atoms in different oxidation states. Since it is (often) possible to unambiguously infer the oxidation states of atoms in compounds, it is, in principle, possible to construct a SkipAtom training set of pairs of atoms in different oxidation states. The number of atom types would increase by several fold, but would still be within limits that allow for efficient training. Here, we demonstrate this by incorporating two additional atom types: Fe(II) and Fe(III). We continue to learn a separate embedding for neutral Fe.

To learn the representations for Fe(II) and Fe(III), we scan the materials structure database for compounds containing Fe, and determine the oxidation state of the element using a *maximum a posteriori* estimation method, as implemented in the *BVAnalyzer* class of the *pymatgen* package (version 2021.2.8.1) [10]. We then form pairs that we will add to our original training set, by keeping only the pairs where Fe(II) or Fe(III) are the target atom (i.e. the atom whose context we will learn to predict). The associated atom in the pair is represented in its neutral state. In total, there were 190,056 pairs generated in this way, and added to the original dataset.

The embeddings were then learned using the SkipAtom approach described in the article, together with this enhanced dataset. To evaluate the learned Fe(II) and Fe(III) representations, a qualitative assessment was made by comparing to Zn and Al, since Zn is generally found in its Zn(II) state, and Al is generally found in its Al(III) state. The four embeddings together, Al, Zn, Fe(II), and Fe(III), were subjected to dimensionality reduction using t-SNE, and the results are plotted in Supplementary Figure 3. It is apparent that Fe(II) resides more closely to Zn, and Fe(III) resides more closely to Al, as one might expect, at least along the first dimension.

5. Analysis of the Number of Embedding Dimensions

Across all the evaluation tasks, the performance of the SkipAtom embeddings appears to increase with the number of embedding dimensions. To better evaluate the influence of the number of embedding dimensions on the performance of the representations, a series of SkipAtom embeddings of different sizes were learned. These embeddings were then mean-pooled for the Refractive Index prediction task, and their performance is given in Supplementary Table 13. A plot of their performance on the task in given in Supplementary Figure 4. Also, these embeddings were used for the Elpasolite Formation Energy prediction task. The results are given in Supplementary Table 14 and Supplementary Figure 5

6. Analysis of Training Set Size

To analyze the influence of the training dataset size on the quality of the learned embeddings, 200-dimensional SkipAtom embeddings were learned using either all or 25% of the available training data from the Materials Project. The training dataset consisting of 25% of the available pairs was created by randomly sampling from the 15,360,652 pairs derived from the Materials Project, yielding a dataset with 3,840,163 pairs. These 200-dimensional SkipAtom embeddings were mean-pooled for the Refractive Index prediction task, and their performance is given in Supplementary Table 15.

Supplementary Tables

Supplementary Table 1: Elpasolite Formation Energy prediction results after 10-fold cross-validation. The dataset consists of 5,645 examples. The task and the model were initially described in the paper that introduced Atom2Vec (an alternative approach for learning atom vectors). [1] The target formation energies were obtained by DFT. [11] The mean best formation energy MAE on the test set after 200 epochs of training in each fold is reported. Batch size was 32, learning rate was 0.001. Note that Dim refers to the dimensionality of the atom vector; the size of the input vector is 4 × Dim. All results were generated using the same procedure on identical train/test folds.

Representation	**Dim**	**MAE (eV/atom)**
Atom2Vec	30	0.1477 ± 0.0078
SkipAtom	*30*	*0.1183* ± *0.0050*
Random	30	0.1701 ± 0.0081
Atom2Vec	86	0.1242 ± 0.0066
One-hot	86	0.1218 ± 0.0085
SkipAtom	86	0.1126 ± 0.0078
Random	86	0.1190 ± 0.0085
Mat2Vec	200	0.1126 ± 0.0058
SkipAtom	**200**	**0.1089 ± 0.0061**
Random	200	0.1158 ± 0.0050

Supplementary Table 2: OQMD Dataset Formation Energy prediction results after 10-fold cross-validation. The dataset consists of 275,424 examples. The target values were computed using DFT. [12, 13]. The mean best formation energy MAE on the test set after 100 epochs of training in each fold is reported. All results were generated using the same procedure on identical train/test folds.

Representation	**Dim**	**Pooling**	**MAE (eV/atom)**
SkipAtom	86	sum	0.0420 ± 0.0005
SkipAtom	86	mean	0.0460 ± 0.0006
SkipAtom	86	max	0.0615 ± 0.0006
Atom2Vec	86	sum	0.0396 ± 0.0004
Atom2Vec	86	mean	0.0417 ± 0.0005
Atom2Vec	86	max	0.0532 ± 0.0006
Bag-of-Atoms / One-hot	**86**	**sum**	**0.0388 ± 0.0002**
ElemNet / One-hot	86	mean	0.0427 ± 0.0007
One-hot	86	max	0.0388 ± 0.0005
Random	86	sum	0.0440 ± 0.0004
Random	86	mean	0.0468 ± 0.0006
Random	86	max	0.0572 ± 0.0007
Mat2Vec	200	sum	0.0401 ± 0.0004
Mat2Vec	200	mean	0.0444 ± 0.0007
Mat2Vec	200	max	0.0501 ± 0.0006
SkipAtom	200	sum	0.0408 ± 0.0003
SkipAtom	200	mean	0.0451 ± 0.0005
SkipAtom	200	max	0.0559 ± 0.0006
Random	200	sum	0.0417 ± 0.0004
Random	200	mean	0.0441 ± 0.0007
Random	200	max	0.0511 ± 0.0005

Supplementary Table 3: Experimental Band Gap prediction results after 2-repeated 5-fold cross-validation. The dataset consists of 4,604 examples. The target values were obtained by experiment. [14]. The mean best MAE on the test set after 100 epochs of training in each fold is reported. All results were generated using the same procedure on identical train/test folds. The reported state-of-the-art result is an MAE of 0.416 eV (Automatminer). [15] Note that the state-of-the-art result does not make use of structure, and uses composition only.

Representation	**Dim**	**Pooling**	**MAE (eV)**
SkipAtom	*86*	*sum*	*0.3495 ± 0.0020*
SkipAtom	86	mean	0.3737 ± 0.0091
SkipAtom	86	max	0.3954 ± 0.0090
Atom2Vec	86	sum	0.3922 ± 0.0087
Atom2Vec	86	mean	0.4005 ± 0.0080
Atom2Vec	86	max	0.4070 ± 0.0048
Bag-of-Atoms / One-hot	86	sum	0.3797 ± 0.0022
ElemNet / One-hot	86	mean	0.4060 ± 0.0072
One-hot	86	max	0.3823 ± 0.0046
Random	86	sum	0.4109 ± 0.0058
Random	86	mean	0.4286 ± 0.0058
Random	86	max	0.4389 ± 0.0028
Mat2Vec	200	sum	0.3529 ± 0.0007
Mat2Vec	200	mean	0.3886 ± 0.0000
Mat2Vec	200	max	0.3625 ± 0.0070
SkipAtom	**200**	**sum**	**0.3487 ± 0.0085**
SkipAtom	200	mean	0.3737 ± 0.0069
SkipAtom	200	max	0.3985 ± 0.0049
Random	200	sum	0.4058 ± 0.0004
Random	200	mean	0.4181 ± 0.0010
Random	200	max	0.4289 ± 0.0067

Supplementary Table 4: Theoretical Band Gap prediction results after 2-repeated 5-fold cross-validation. The dataset consists of 106,113 examples. The target values were obtained by DFT-GGA. [16, 17]. The mean best MAE on the test set after 100 epochs of training in each fold is reported. All results were generated using the same procedure on identical train/test folds. The reported state-of-the-art result is an MAE of 0.228 eV (CGCNN). [15] Note that the state-of-the-art result makes use of structure.

Representation	**Dim**	**Pooling**	**MAE (eV)**
SkipAtom	86	sum	0.2791 ± 0.0008
SkipAtom	86	mean	0.2807 ± 0.0003
SkipAtom	86	max	0.3512 ± 0.0017
Atom2Vec	86	sum	0.2692 ± 0.0008
Atom2Vec	86	mean	0.2712 ± 0.0025
Atom2Vec	86	max	0.3289 ± 0.0016
Bag-of-Atoms / One-hot	86	sum	0.2611 ± 0.0008
ElemNet / One-hot	**86**	**mean**	**0.2582 ± 0.0003**
One-hot	86	max	0.2603 ± 0.0004
Random	86	sum	0.3238 ± 0.0005
Random	86	mean	0.3180 ± 0.0016
Random	86	max	0.4096 ± 0.0008
Mat2Vec	200	sum	0.2741 ± 0.0002
Mat2Vec	200	mean	0.2744 ± 0.0005
Mat2Vec	200	max	0.3256 ± 0.0002
SkipAtom	*200*	*sum*	*0.2736 ± 0.0008*
SkipAtom	200	mean	0.2753 ± 0.0006
SkipAtom	200	max	0.3351 ± 0.0013
Random	200	sum	0.3083 ± 0.0021
Random	200	mean	0.3095 ± 0.0009
Random	200	max	0.3733 ± 0.0010

Supplementary Table 5: Bulk Modulus prediction results after 2-repeated 10-fold cross-validation. The dataset consists of 10,987 examples. The target values were computed using DFT-GGA. [18]. The mean best MAE on the test set after 100 epochs of training in each fold is reported. All results were generated using the same procedure on identical train/test folds. The reported state-of-the-art result is an MAE of 0.0679 log(GPa) (Automatminer). [15] Note that the state-of-the-art result makes use of structure.

Representation	**Dim**	**Pooling**	**MAE (log(GPa))**
SkipAtom	86	sum	0.0790 ± 0.0002
SkipAtom	*86*	*mean*	*0.0789 ± 0.0002*
SkipAtom	86	max	0.0867 ± 0.0000
Atom2Vec	86	sum	0.0795 ± 0.0005
Atom2Vec	86	mean	0.0810 ± 0.0004
Atom2Vec	86	max	0.0861 ± 0.0002
Bag-of-Atoms / One-hot	86	sum	0.0861 ± 0.0002
ElemNet / One-hot	86	mean	0.0853 ± 0.0001
One-hot	86	max	0.0861 ± 0.0003
Random	86	sum	0.0916 ± 0.0002
Random	86	mean	0.0908 ± 0.0004
Random	86	max	0.0997 ± 0.0001
Mat2Vec	**200**	**sum**	**0.0776 ± 0.0000**
Mat2Vec	200	mean	0.0779 ± 0.0003
Mat2Vec	200	max	0.0813 ± 0.0003
SkipAtom	200	sum	0.0786 ± 0.0003
SkipAtom	200	mean	0.0785 ± 0.0000
SkipAtom	200	max	0.0888 ± 0.0002
Random	200	sum	0.0887 ± 0.0003
Random	200	mean	0.0871 ± 0.0001
Random	200	max	0.0960 ± 0.0004

Supplementary Table 6: Shear Modulus prediction results after 2-repeated 10-fold cross-validation. The dataset consists of 10,987 examples. The target values were computed using DFT-GGA. [18]. The mean best MAE on the test set after 100 epochs of training in each fold is reported. All results were generated using the same procedure on identical train/test folds. The reported state-of-the-art result is an MAE of 0.0849 log(GPa) (Automatminer). [15] Note that the state-of-the-art result makes use of structure.

Representation	**Dim**	**Pooling**	**MAE (log(GPa))**
SkipAtom	*86*	*sum*	*0.1014 ± 0.0001*
SkipAtom	86	mean	0.1025 ± 0.0002
SkipAtom	86	max	0.1102 ± 0.0002
Atom2Vec	86	sum	0.1029 ± 0.0000
Atom2Vec	86	mean	0.1054 ± 0.0000
Atom2Vec	86	max	0.1089 ± 0.0005
Bag-of-Atoms / One-hot	86	sum	0.1137 ± 0.0005
ElemNet / One-hot	86	mean	0.1155 ± 0.0001
One-hot	86	max	0.1140 ± 0.0002
Random	86	sum	0.1195 ± 0.0002
Random	86	mean	0.1199 ± 0.0001
Random	86	max	0.1260 ± 0.0001
Mat2Vec	**200**	**sum**	**0.1014 ± 0.0002**
Mat2Vec	200	mean	0.1035 ± 0.0001
Mat2Vec	200	max	0.1050 ± 0.0000
SkipAtom	**200**	**sum**	**0.1014 ± 0.0000**
SkipAtom	200	mean	0.1024 ± 0.0001
SkipAtom	200	max	0.1111 ± 0.0001
Random	200	sum	0.1167 ± 0.0002
Random	200	mean	0.1163 ± 0.0002
Random	200	max	0.1223 ± 0.0000

Supplementary Table 7: Refractive Index prediction results after 2-repeated 5-fold cross-validation. The dataset consists of 4,764 examples. The target values were computed using DFPT-GGA. [19]. The mean best MAE on the test set after 100 epochs of training in each fold is reported. All results were generated using the same procedure on identical train/test folds. The reported state-of-the-art result is an MAE of 0.299 n (Automatminer). [15] Note that the state-of-the-art result makes use of structure.

Representation	**Dim**	**Pooling**	**MAE (n)**
SkipAtom	86	sum	0.3369 ± 0.0014
SkipAtom	*86*	*mean*	*0.3275 ± 0.0004*
SkipAtom	86	max	0.3561 ± 0.0013
Atom2Vec	86	sum	0.3419 ± 0.0013
Atom2Vec	86	mean	0.3308 ± 0.0016
Atom2Vec	86	max	0.3522 ± 0.0005
Bag-of-Atoms / One-hot	86	sum	0.3576 ± 0.0002
ElemNet / One-hot	86	mean	0.3409 ± 0.0016
One-hot	86	max	0.3547 ± 0.0013
Random	86	sum	0.3625 ± 0.0012
Random	86	mean	0.3593 ± 0.0006
Random	86	max	0.3891 ± 0.0021
Mat2Vec	200	sum	0.3272 ± 0.0004
Mat2Vec	**200**	**mean**	**0.3236 ± 0.0017**
Mat2Vec	200	max	0.3428 ± 0.0004
SkipAtom	200	sum	0.3340 ± 0.0012
SkipAtom	200	mean	0.3247 ± 0.0015
SkipAtom	200	max	0.3618 ± 0.0026
Random	200	sum	0.3598 ± 0.0053
Random	200	mean	0.3543 ± 0.0006
Random	200	max	0.3824 ± 0.0019

Supplementary Table 8: Bulk Metallic Glass Formation prediction results after 2-repeated 5-fold stratified cross-validation. The dataset consists of 5,680 examples. The target values were obtained from experiment. [20, 21]. The mean best ROC-AUC on the test set after 100 epochs of training in each fold is reported. All results were generated using the same procedure on identical train/test folds. The reported state-of-the-art result is an ROC-AUC of 0.858 (RF). [15] Note that the state-of-the-art result does not make use of structure, and uses composition only.

Representation	**Dim**	**Pooling**	**ROC-AUC**
SkipAtom	86	sum	0.9312 ± 0.0007
SkipAtom	*86*	*mean*	*0.9346* ± *0.0010*
SkipAtom	86	max	0.9243 ± 0.0005
Atom2Vec	86	sum	0.9306 ± 0.0026
Atom2Vec	86	mean	0.9316 ± 0.0012
Atom2Vec	86	max	0.9300 ± 0.0008
Bag-of-Atoms / One-hot	86	sum	0.9277 ± 0.0004
ElemNet / One-hot	86	mean	0.9322 ± 0.0014
One-hot	86	max	0.9289 ± 0.0016
Random	86	sum	0.9262 ± 0.0011
Random	86	mean	0.9274 ± 0.0006
Random	86	max	0.9243 ± 0.0020
Mat2Vec	200	sum	0.9280 ± 0.0004
Mat2Vec	200	mean	0.9348 ± 0.0024
Mat2Vec	200	max	0.9253 ± 0.0009
SkipAtom	200	sum	0.9327 ± 0.0022
SkipAtom	**200**	**mean**	**0.9349 ± 0.0019**
SkipAtom	200	max	0.9268 ± 0.0002
Random	200	sum	0.9274 ± 0.0019
Random	200	mean	0.9302 ± 0.0016
Random	200	max	0.9298 ± 0.0009

Supplementary Table 9: Experimental Metallicity prediction results after 2-repeated 5-fold stratified cross-validation. The dataset consists of 4,921 examples. The target values were obtained from experiment. [14]. The mean best ROC-AUC on the test set after 100 epochs of training in each fold is reported. All results were generated using the same procedure on identical train/test folds. The reported state-of-the-art result is an ROC-AUC of 0.917 (Random Forest). [15] Note that the state-of-the-art result does not make use of structure, and uses composition only.

Representation	**Dim**	**Pooling**	**ROC-AUC**
SkipAtom	*86*	*sum*	*0.9645 ± 0.0012*
SkipAtom	86	mean	0.9575 ± 0.0003
SkipAtom	86	max	0.9561 ± 0.0020
Atom2Vec	86	sum	0.9582 ± 0.0008
Atom2Vec	86	mean	0.9541 ± 0.0005
Atom2Vec	86	max	0.9548 ± 0.0006
Bag-of-Atoms / One-hot	86	sum	0.9600 ± 0.0012
ElemNet / One-hot	86	mean	0.9485 ± 0.0007
One-hot	86	max	0.9599 ± 0.0014
Random	86	sum	0.9559 ± 0.0021
Random	86	mean	0.9460 ± 0.0008
Random	86	max	0.9426 ± 0.0037
Mat2Vec	**200**	**sum**	**0.9655 ± 0.0014**
Mat2Vec	200	mean	0.9570 ± 0.0008
Mat2Vec	200	max	0.9634 ± 0.0013
SkipAtom	200	sum	0.9645 ± 0.0008
SkipAtom	200	mean	0.9572 ± 0.0008
SkipAtom	200	max	0.9589 ± 0.0010
Random	200	sum	0.9541 ± 0.0002
Random	200	mean	0.9454 ± 0.0001
Random	200	max	0.9508 ± 0.0011

Supplementary Table 10: Theoretical Metallicity prediction results after 2-repeated 5-fold stratified cross-validation. The dataset consists of 106,113 examples. The target values were obtained by DFT-GGA. [16, 17]. The mean best ROC-AUC on the test set after 100 epochs of training in each fold is reported. All results were generated using the same procedure on identical train/test folds. The reported state-of-the-art result is an ROC-AUC of 0.977 (MEGNet). [15] Note that the state-of-the-art result makes use of structure. Notable is that the CGCNN model in the same study achieves an ROC-AUC of 0.954, also using structure, which is comparable to the performance of the Mat2Vec representation, which uses only composition.

Representation	**Dim**	**Pooling**	**ROC-AUC**
SkipAtom	86	sum	0.9520 ± 0.0002
SkipAtom	86	mean	0.9506 ± 0.0000
SkipAtom	86	max	0.9440 ± 0.0000
Atom2Vec	*86*	*sum*	*0.9526 ± 0.0001*
Atom2Vec	86	mean	0.9506 ± 0.0001
Atom2Vec	86	max	0.9450 ± 0.0003
Bag-of-Atoms / One-hot	86	sum	0.9490 ± 0.0002
ElemNet / One-hot	86	mean	0.9477 ± 0.0001
One-hot	86	max	0.9487 ± 0.0003
Random	86	sum	0.9444 ± 0.0000
Random	86	mean	0.9433 ± 0.0002
Random	86	max	0.9330 ± 0.0001
Mat2Vec	**200**	**sum**	**0.9528 ± 0.0002**
Mat2Vec	200	mean	0.9517 ± 0.0001
Mat2Vec	200	max	0.9469 ± 0.0005
SkipAtom	200	sum	0.9524 ± 0.0001
SkipAtom	200	mean	0.9507 ± 0.0001
SkipAtom	200	max	0.9454 ± 0.0001
Random	200	sum	0.9453 ± 0.0002
Random	200	mean	0.9441 ± 0.0001
Random	200	max	0.9380 ± 0.0000

Supplementary Table 11: Band gap prediction results on the test set of 27,430 compounds from the Materials Project, split 60/20/20, using the CGCNN model. Training was performed for 100 epochs, a learning rate of 0.01 was used, along with a batch size of 256. The default settings provided by library were used for the other hyperparameters.

Input Representation	**MAE (eV)**
CGCNN binary feature vector	0.381
SkipAtom 200-dim	**0.371**

Supplementary Table 12: Elpasolite formation energy prediction results with the MEGNet architecture. This model incorporates crystal structure.

Input Representation	**MAE (eV/atom)**
one-hot atom vectors + embedding table	0.0685
SkipAtom 200-dim	**0.0568**

Supplementary Table 13: Refractive Index prediction results after 2-repeated 5-fold cross-validation using mean-pooled SkipAtom embeddings of various dimensions. The mean best MAE on the test set after 100 epochs of training in each fold is reported. All results were generated using the same procedure on identical train/test folds.

Dim	**MAE (n)**
30	0.3278 ± 0.0008
86	0.3262 ± 0.0002
200	0.3248 ± 0.0015
300	0.3252 ± 0.0005
400	0.3267 ± 0.0017
800	0.3263 ± 0.0000

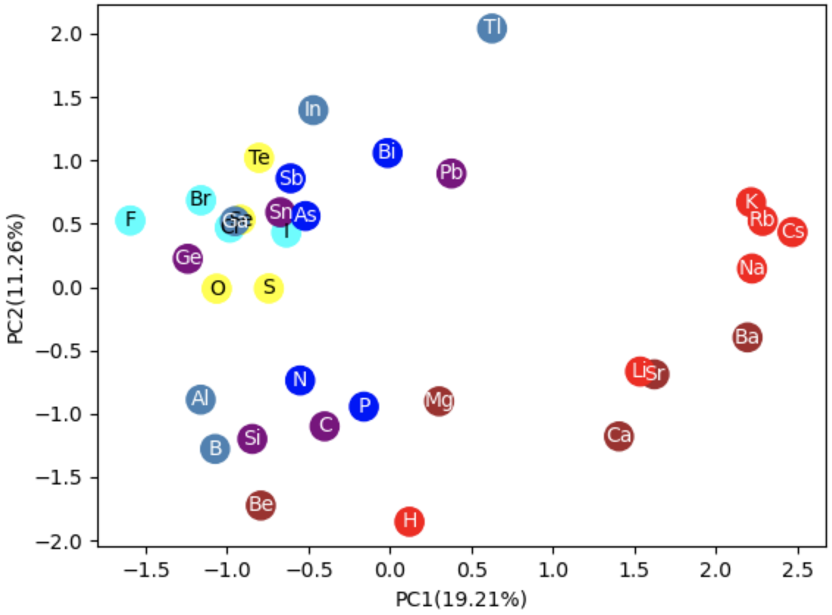
Supplementary Table 14: Elpasolite Formation Energy prediction results after 10-fold cross-validation using SkipAtom embeddings of various dimensions. The mean best MAE on the test set after 200 epochs of training in each fold is reported. All results were generated using the same procedure on identical train/test folds.

Dim	**MAE (eV/atom)**
30	0.1183 ± 0.0050
86	0.1126 ± 0.0078
200	0.1089 ± 0.0061
300	0.1082 ± 0.0053
400	0.1085 ± 0.0029
800	0.1056 ± 0.0034

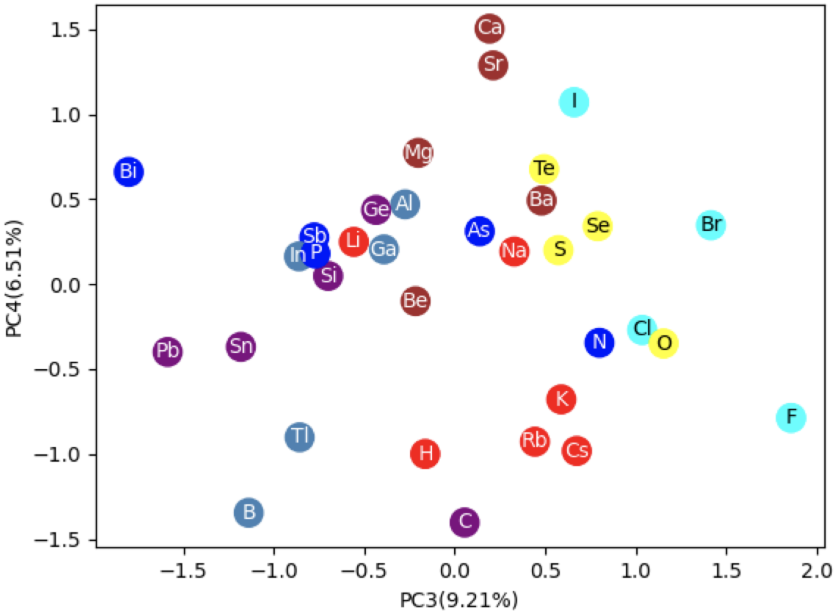
Supplementary Table 15: Refractive Index prediction results after 2-repeated 5-fold cross-validation using 200-dim mean-pooled SkipAtom embeddings learned with different amounts of training data. The mean best MAE on the test set after 100 epochs of training in each fold is reported. All results were generated using the same procedure on identical train/test folds.

Dim	**% of training data**	**MAE (n)**
200	25	0.3256 ± 0.0003
200	100	0.3248 ± 0.0015

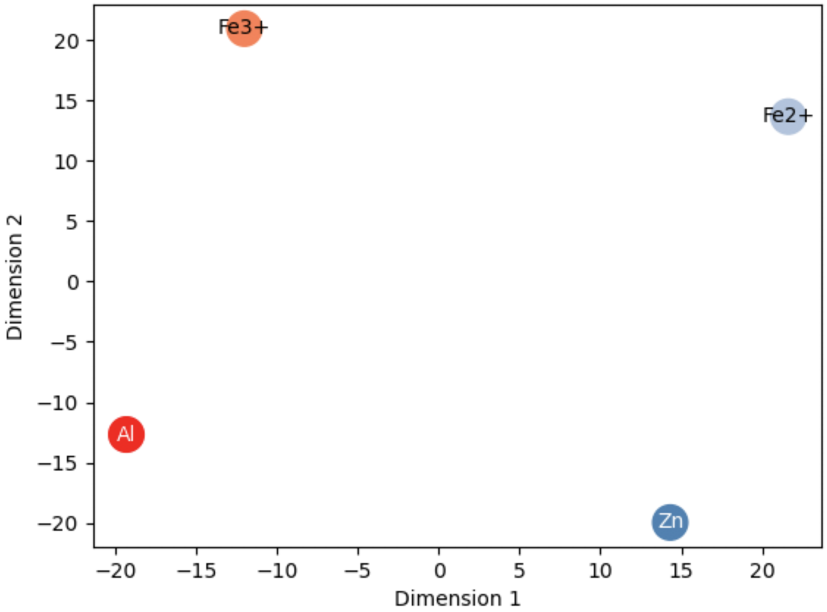
Supplementary Figures



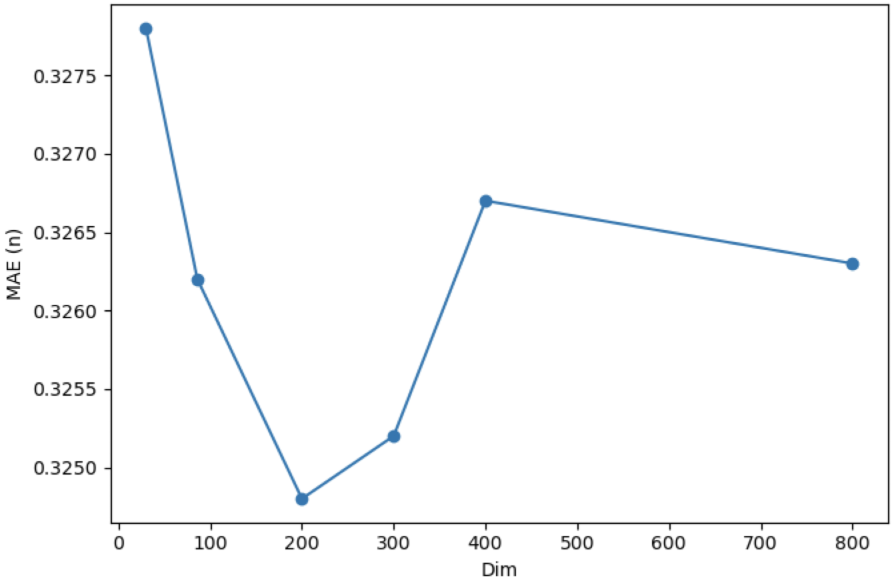
Supplementary Figure 1: Principal Component Analysis of SkipAtom Representations. The first two principal components of the SkipAtom 200-dim vectors for 34 atoms are depicted.



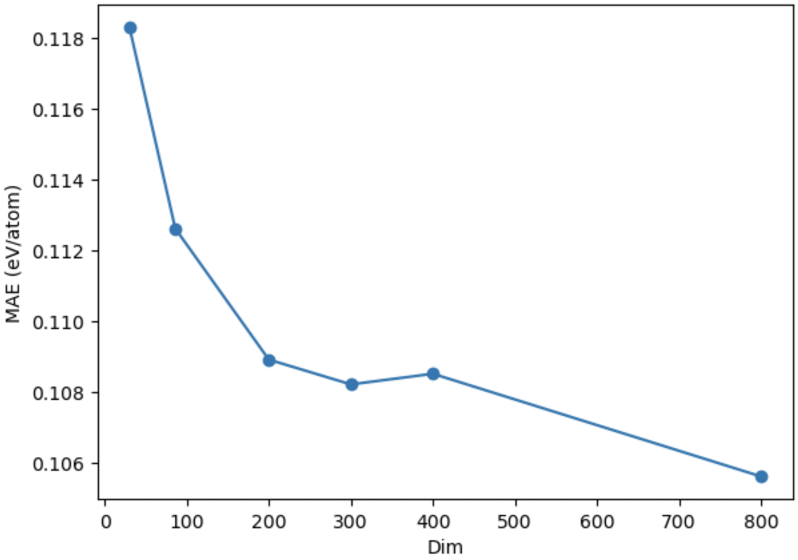
Supplementary Figure 2: Principal Component Analysis of SkipAtom Representations. The third and fourth principal components of the SkipAtom 200-dim vectors for 34 atoms are depicted.



Supplementary Figure 3: Dimensionally reduced SkipAtom vectors for Al and Zn, and for Fe(II) and Fe(III). The vectors were reduced from 200 dimensions to 2 dimensions using t-SNE.



Supplementary Figure 4: A plot of MAE results for the Refractive Index prediction task, obtained using 2-repeated 5-fold cross-validation, for a number of different embedding sizes. The SkipAtom embeddings were mean-pooled.



Supplementary Figure 5: A plot of MAE results for the Elpasolite Formation Energy prediction task, obtained using 10-fold cross-validation, for a number of different embedding sizes. The SkipAtom embeddings were concatenated.

Supplementary References

[1] Quan Zhou, Peizhe Tang, Shenxiu Liu, Jinbo Pan, Qimin Yan, and Shou-Cheng Zhang. Learning atoms for materials discovery. *P. Natl. Acad. Sci.*, 115(28):E6411–E6417, 2018.

[2] Vahe Tshitoyan, John Dagdelen, Leigh Weston, Alexander Dunn, Ziqin Rong, Olga Kononova, Kristin A Persson, Gerbrand Ceder, and Anubhav Jain. Unsupervised word embeddings capture latent knowledge from materials science literature. *Nature*, 571(7763):95–98, 2019.

[3] Tian Xie and Jeffrey C Grossman. Crystal Graph Convolutional Neural Networks for an Accurate and Interpretable Prediction of Material Properties. *Phys. Rev. Lett.*, 120(14):145301, 2018.

[4] Chi Chen, Weike Ye, Yunxing Zuo, Chen Zheng, and Shyue Ping Ong. Graph Networks as a Universal Machine Learning Framework for Molecules and Crystals. *Chem. Mater.*, 31(9):3564–3572, 2019.

[5] CGCNN Library. `https://github.com/txie-93/cgcnn`.

[6] MEGNet Library. `https://github.com/materialsvirtuallab/megnet`.

[7] Georges Voronoi. Nouvelles applications des paramètres continus à la théorie des formes quadratiques. Premier mémoire. Sur quelques propriétés des formes quadratiques positives parfaites. *J. Reine Angew. Math.*, 1908:97 – 102, 1908.

[8] Nils ER Zimmermann and Anubhav Jain. Local structure order parameters and site fingerprints for quantification of coordination environment and crystal structure similarity. *RSC Adv.*, 10(10):6063–6081, 2020.

[9] Hillary Pan, Alex M Ganose, Matthew Horton, Muratahan Aykol, Kristin A Persson, Nils ER Zimmermann, and Anubhav Jain. Benchmarking Coordination Number Prediction Algorithms on Inorganic Crystal Structures. *Inorg. Chem.*, 60(3):1590–1603, 2021.

[10] Shyue Ping Ong, William Davidson Richards, Anubhav Jain, Geoffroy Hautier, Michael Kocher, Shreyas Cholia, Dan Gunter, Vincent L Chevrier, Kristin A Persson, and Gerbrand Ceder. Python Materials Genomics (pymatgen): A robust, open-source python library for materials analysis. *Comput. Mater. Sci.*, 68:314–319, 2013.

[11] Felix A Faber, Alexander Lindmaa, O Anatole Von Lilienfeld, and Rickard Armiento. Machine Learning Energies of 2 Million Elpasolite (ABC2D6) Crystals. *Phys. Rev. Lett.*, 117(13):135502, 2016.

[12] Dipendra Jha, Logan Ward, Arindam Paul, Wei-keng Liao, Alok Choudhary, Chris Wolverton, and Ankit Agrawal. ElemNet: Deep Learning the Chemistry of Materials From Only Elemental Composition. *Sci. Rep.*, 8(1):1–13, 2018.

[13] James E Saal, Scott Kirklin, Muratahan Aykol, Bryce Meredig, and Christopher Wolverton. Materials design and discovery with high-throughput density functional theory: the open quantum materials database (OQMD). *JOM*, 65(11):1501–1509, 2013.

[14] Ya Zhuo, Aria Mansouri Tehrani, and Jakoah Brgoch. Predicting the Band Gaps of Inorganic Solids by Machine Learning. *J. Phys. Chem. Lett.*, 9(7):1668–1673, 2018.

[15] Alexander Dunn, Qi Wang, Alex Ganose, Daniel Dopp, and Anubhav Jain. Benchmarking materials property prediction methods: the matbench test set and automatminer reference algorithm. *npj Comput. Mater.*, 6(1):1–10, 2020.

[16] Anubhav Jain, Shyue Ping Ong, Geoffroy Hautier, Wei Chen, William Davidson Richards, Stephen Dacek, Shreyas Cholia, Dan Gunter, David Skinner, Gerbrand Ceder, et al. Commentary: The Materials Project: A materials genome approach to accelerating materials innovation. *APL Mater.*, 1(1):011002, 2013.

[17] Shyue Ping Ong, Shreyas Cholia, Anubhav Jain, Miriam Brafman, Dan Gunter, Gerbrand Ceder, and Kristin A Persson. The Materials Application Programming Interface (API): A simple, flexible and efficient API for materials data based on REpresentational State Transfer (REST) principles. *Comput. Mater. Sci.*, 97:209–215, 2015.

[18] Maarten De Jong, Wei Chen, Thomas Angsten, Anubhav Jain, Randy Notestine, Anthony Gamst, Marcel Sluiter, Chaitanya Krishna Ande, Sybrand Van Der Zwaag, Jose J Plata, et al. Charting the complete elastic properties of inorganic crystalline compounds. *Sci. Data*, 2(1):1–13, 2015.

[19] Ioannis Petousis, David Mrdjenovich, Eric Ballouz, Miao Liu, Donald Winston, Wei Chen, Tanja Graf, Thomas D Schladt, Kristin A Persson, and Fritz B Prinz. High-throughput screening of inorganic compounds for the discovery of novel dielectric and optical materials. *Sci. Data*, 4(1):1–12, 2017.

[20] Logan Ward, Ankit Agrawal, Alok Choudhary, and Christopher Wolverton. A general-purpose machine learning framework for predicting properties of inorganic materials. *npj Comput. Mater.*, 2(1):1–7, 2016.

[21] Yoshiyuki Kawazoe. Nonequilibrium Phase Diagrams of Ternary Amorphous Alloys. *LB: New Ser., Group III: Condensed*, 37:1–295, 1997.