PNAS
www.pnas.org

Supplementary Information for

Atom2Vec: learning atoms for materials discovery

Quan Zhou, Peizhe Tang, Shenxiu Liu, Jinbo Pan, Qimin Yan, and Shou-Cheng Zhang

Shou-Cheng Zhang
Email: sczhang@stanford.edu

This PDF file includes:

Supplementary text
Figs. S1 to S5
References for SI reference citations

Other supplementary materials for this manuscript include the following:

www.pnas.org/cgi/doi/10.1073/pnas.1801181115

S1. Statistics of compound data

There are 60605 inorganic compounds in total in Materials Project database[1]. Because environments with too many atom types usually depends on structures heavily, they do not help and even impair atom learning if only chemical formulas are taken into account. Hence, as mentioned in Method, only binary, ternary and quaternary compounds are used to learn atom vectors in this work. The population of each symbol (equivalently, the number of environments with this symbol as target atom) is also examined. It is found that there is a sharp population drop between common and rare symbols. A threshold, 1% of the maximal population which is located near the drop, is selected to filter out rare symbols. Only atom-environment pairs of common symbols (about 100) are used to learn atom vectors.

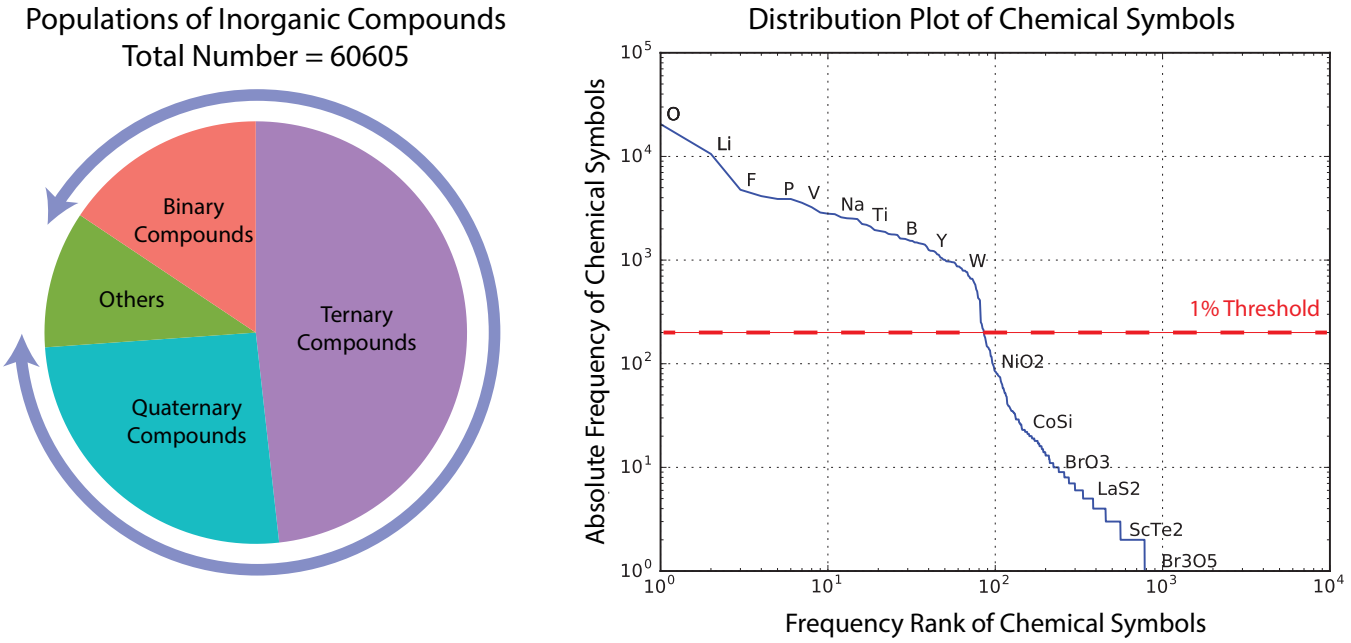

FIG. S1: (Color online) **Statistics of compound data for atom vector learning.** Population distributions of binary, ternary, quaternary and other compounds. Population plot of chemical symbols.

S2. Singular values in model-free method

Singular value decomposition[2] is applied on the atom-environment matrix in our model-free method. The singular values are shown in Fig. S2 in descending order. As seen in Fig. S2, a fat tail of non-vanishing values follows a quick decrease for the first ten, which indicates that even higher dimensions could describe meaningful aspects of atoms. In this work, atom vectors of dimension $d = 10, 20, 30$ are chosen as examples.

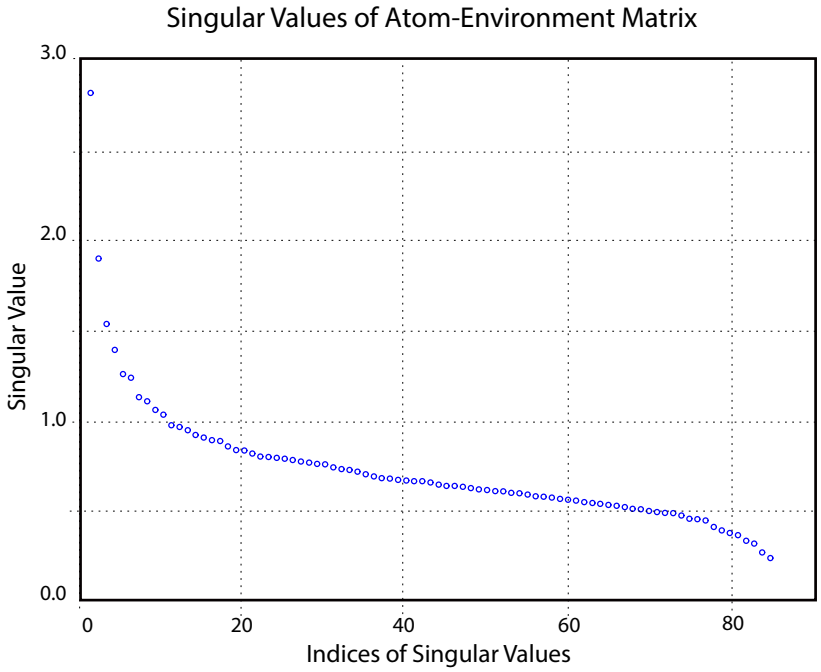

FIG. S2: (Color online) **Singular values distribution of atom-environment matrix.**

S3. Learning in model-based method

In model-based method, mini-batch stochastic gradient descent[3] is used to update atom vectors for minimization of the loss function. Fig. S3 shows learning curves for two atom vector dimensions ($d = 10$ and $d = 30$) based on two different score functions (bilinear and inverse square). When converged, the loss functions for all four cases are as high as 3.5, this means that the model itself still do not well describe the composition between atom and environment, since on average given an environment, the correctly predicted atom only gives probability of $e^{-3.5} \approx 1/30$.

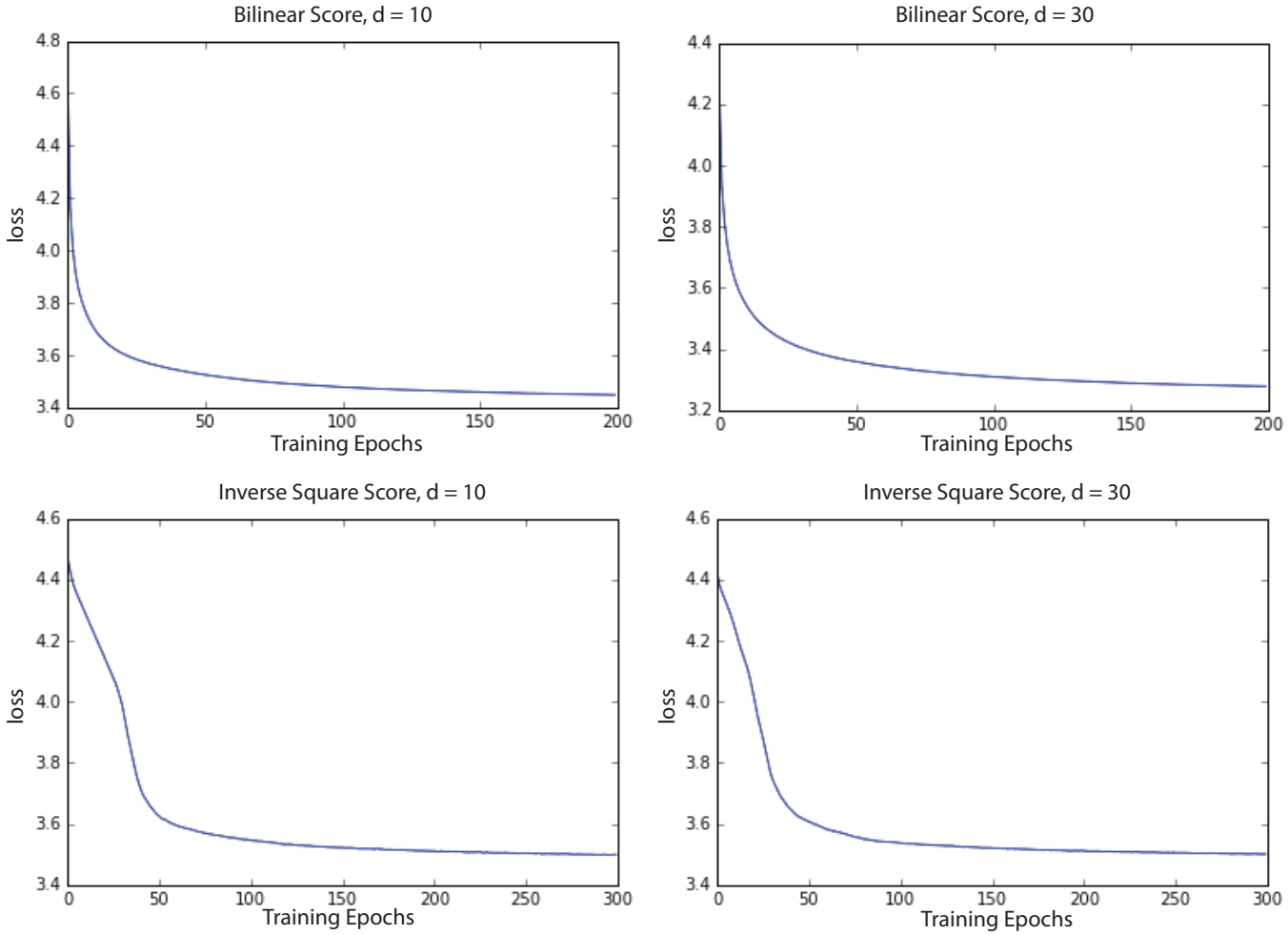

FIG. S3: (Color online) **Learning Curves for model-based methods** Two different score functions (bilinear and inverse square) and two atom vector dimension ($d = 10$ and $d = 30$) are investigated.

S4. Atom vectors from model-based method

We show atom vectors from our model-based method, in particular, from the one based on inverse square score function. The atom vectors of dimension $d = 20$ for main group elements are investigated, and projections to leading principal components are shown in Fig. S4. These vectors are learned from a small dataset including only compounds comprised of main group elements, rather than the entire dataset as in model-free method. We have also examined the case when the entire dataset is used, those vectors almost learn nothing about properties of atoms, probably due to the fact that the over-simplified model does not describe a large portion of the full dataset. This limitation actually appears in the atom vectors shown in Fig. S4 as well. Note that the first principal component of these vectors is dominant in model-based learning here, and it corresponds to the valence trend (or the columns in the periodic table) almost exactly. But projections over remaining components are extremely noisy, there appears no patterns at all beyond the first principal component in the model-based method.

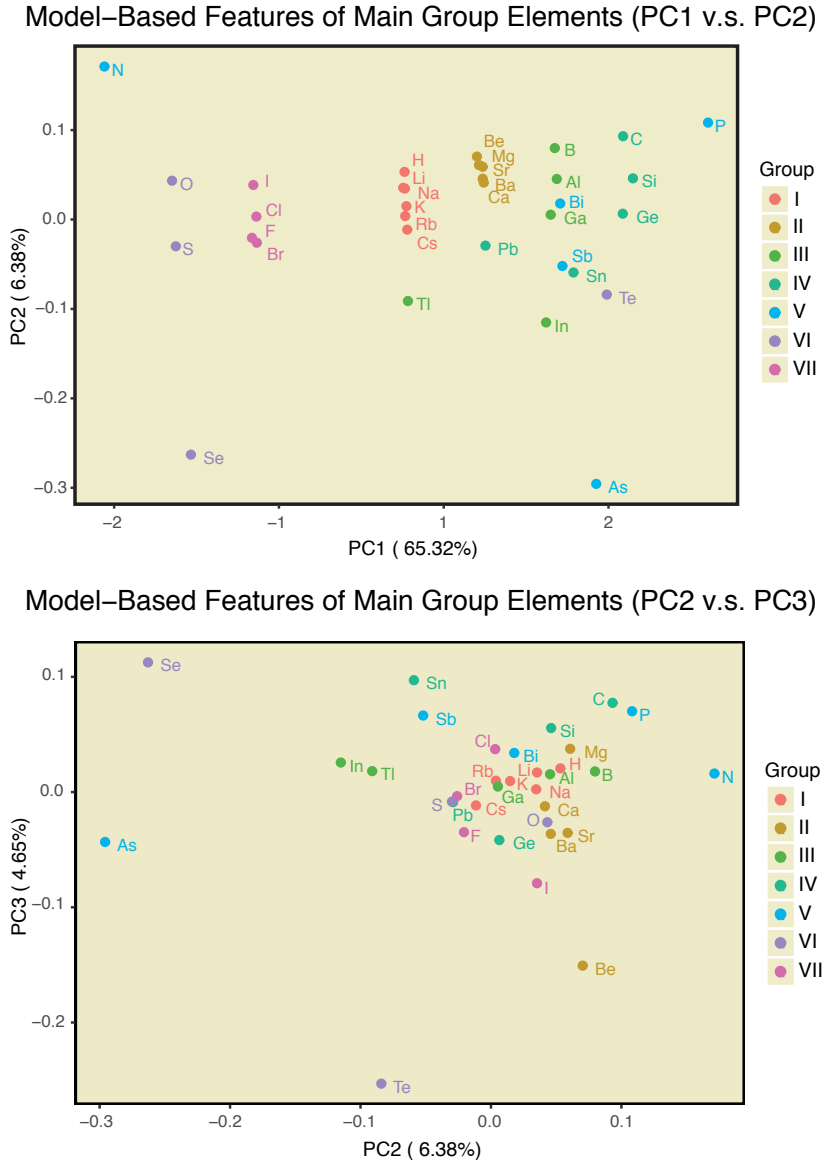

FIG. S4: (Color online) **Projection of atom vectors learned using model-based methods into leading principal components.** These atom vectors are of 20 dimensions, and inverse square score function is used in model-based learning. A small dataset which includes only compounds with atoms of main-group elements is used for learning.

S5. Elpasolite compound formation energy data

When evaluating our atom vectors quantitatively, we train neural network models to predict formation energy of elpasolites based on a dataset with formation energies computed by first principle calculation. There are nearly ten thousands of such compounds in the original dataset[4], only the compounds comprised of atoms that have our atom vector representation are used to train the model. This leaves about six thousands elpasolite samples, whose formation energy distribution is shown in Fig. S5.

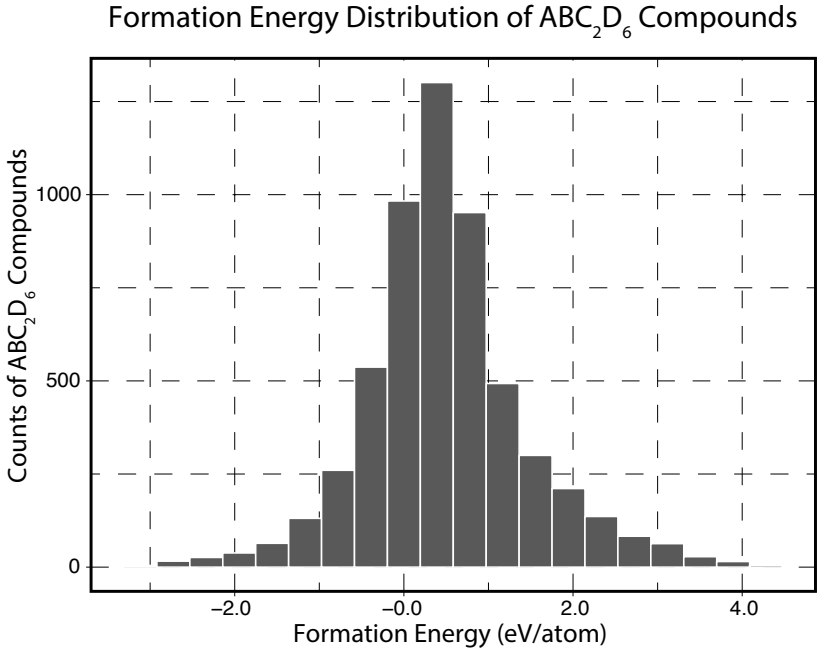

FIG. S5: (Color online) **Formation energy distribution of all ABC_2D_6 elpasolite compounds.**

References

1. Jain A, *et al*. (2013) The Materials Project: A materials genome approach to accelerating materials innovation. *APL Materials 1(1):011002*
2. Hastie T, Tibshirani R, Friedman J (2001) The Elements of Statistical Learning, Springer Series in Statistics. (Springer New York Inc., New York, NY, USA).
3. Goodfellow I, Bengio Y, Courville A (2016) Deep Learning. (MIT Press).
4. Faber FA, Lindmaa A, von Lilienfeld OA, Armiento R (2016) Machine learning energies of 2million elpasolite ABC2D6 crystals. *Phys. Rev. Lett 117(13):135502.*