# Ionic species representations for materials informatics

Cite as: APL Mach. Learn. 2, 036112 (2024); doi: 10.1063/5.0227009
Submitted: 5 July 2024 • Accepted: 28 August 2024 •
Published Online: 19 September 2024

View Online Export Citation CrossMark

Anthony Onwuli,[1] Keith T. Butler,[2,a)] and Aron Walsh[1,b)]

## AFFILIATIONS

[1] Department of Materials, Imperial College London, London SW7 2AZ, United Kingdom
[2] Department of Chemistry, University College London, London WC1H 0AJ, United Kingdom

[a)] **Electronic mail:** k.t.butler@ucl.ac.uk
[b)] **Author to whom correspondence should be addressed:** a.walsh@imperial.ac.uk

## ABSTRACT

High-dimensional representations of the elements have become common within the field of materials informatics to build useful, structure-agnostic models for the chemistry of materials. However, the characteristics of elements change when they adopt a given oxidation state, with distinct structural preferences and physical properties. We explore several methods for developing embedding vectors of elements decorated with oxidation states. Graphs generated from 110 160 crystals are used to train representations of 84 elements that form 336 species. Clustering these learned representations of ionic species in low-dimensional space reproduces expected chemical heuristics, particularly the separation of cations from anions. We show that these representations have enhanced expressive power for property prediction tasks involving inorganic compounds. We expect that ionic representations, necessary for the description of mixed valence and complex magnetic systems, will support more powerful machine learning models for materials.

## I. INTRODUCTION

Materials informatics has served as a powerful field of study for the discovery and optimization of functional materials with engineered properties. For machine learning applications, the representations of materials is an important subfield, where state-of-the-art performances have been achieved through graph representations that incorporate information about both composition and structure.[1–9] This success has been made evident through the material science specific benchmarking suite Matbench, where graph neural networks (GNNs) dominate for tasks that incorporate structure.[10]

In the large data regime, neural network-based models have been able to learn representations of the elements. In addition to these learned descriptors of the elements, hand-crafted representations of elements and compositions continue to play a role in materials informatics for property prediction.[11–18] An often overlooked aspect when dealing with traditional element representations is the role of ions. Ions, and the knowledge of oxidation states, play a significant role in both the structural and electronic properties of materials such as electrical conductivity,[19,20] chemical bonding, and magnetism[21] as a result of their electronic configuration differing from the parent atom.[22,23] For example, $Fe^0$ is the building block of ferromagnetic iron metal, $Fe^{3+}$ is found in the antiferromagnetic insulator $Fe_2O_3$ (hematite), while a mixture of $Fe^{2+}$ and $Fe^{3+}$ is found in the ferrimagnetic crystals of $Fe_3O_4$ (magnetite).

For composition-only property prediction (also known as structure-agnostic learning), compositions can be represented as composition-based feature vectors (CBFVs), which are derived from element embeddings. Typically, the element embeddings are combined through pooling operations (commonly descriptive statistics) to make a CBFV, which can be used as an input for machine learning. This enables material informatics practitioners to make machine-learning property predictions in the absence of explicit structural information.

The choice of the underlying element embedding used to make the CBFV does impact the performance of the property prediction task of interest. Depending on the size of the training data, CBFVs, which lack domain knowledge (such as random representations or one-hot encoded atomic representations), can perform comparatively to CBFVs built from element properties.[24]

A common aspect of structure-agnostic approaches is that they treat element information for compositions. This raises the question of whether there is any utility for having representations or even incorporating knowledge of the oxidation states into structure-agnostic learning. When screening inorganic compositions, oxidation states are typically considered a form of chemical heuristics to achieve charge-balancing for enumerative algorithms,[25,26] to suggest why one compound may form rather than another[27] or as a measure to check the validity of compositions suggested by generative models.[28,29] Often, these oxidation states do not see much further use beyond these applications in screening studies despite their central role in the explanation of properties in inorganic chemistry.

In this study, we develop high-dimensional representations of ions and assess their utility for materials informatics tasks. The `SkipAtom`[30] approach for developing distributed representations for the chemical elements is adapted to develop distributed representations of ionic species. We call our adapted formalism `SkipSpecies`. The `SkipSpecies` representation is then benchmarked in a structure-agnostic setting on two regression tasks (formation energy and bandgap) and two classification tasks (metallic and magnetic classification) with differing dimension sizes and different pooling operations. We find that ionic representations can perform better than standard element representations on tasks that are linked to the electronic structure of a material.

## II. METHODS

### A. Dataset construction

The materials data are taken from the Materials Project database (version: 2022.10.28).[31] The Materials Project API implemented in `pymatgen` (version: 2023.7.14)[33] is used to query the oxidation states route of the database to obtain 154 718 non-deprecated structures.

We query for the following properties: *material_id*, *structure*, *formula_pretty*, *possible_species*, and *method*. From this query, we filter out materials that cannot be assigned oxidation states by removing entries where the *method* field is None. The *method* field in this dataset has three possible values: "bond valence analysis," none, and "oxidation state guess." "bond valence analysis" refers to materials where the oxidation states can be assigned using the `BVAnalyzer` class in `pymatgen`[33] using the bond valence algorithm, which uses element-based parameters[34] based on the ICSD.[35] "Oxidation State Guess" refers to materials where the oxidation states are assigned using the `oxi_state_guesses` method implemented in the `Composition` class of the `pymatgen.core.composition` submodule. This filtering returns 116 363 structures with assigned oxidation states. A further filter is applied to remove structures with non-integer oxidation states, resulting in a final dataset of 110 160 oxidation-state decorated structures. Of these remaining structures, the majority (103 687/110 160) where assigned oxidation states by bond valence analysis, with the remaining structures being assigned oxidation states through the oxidation state guess method. The distribution of the oxidation states of the elements is shown in Fig. 1.

To construct the property datasets, the Material Project IDs (material_id) are used to query the materials' summary route for the bandgap, formation energy per atom, and the metallic and magnetic classification.

### B. SkipSpecies training

Following the approach in the original paper, which introduced `SkipAtom`,[30] a Voronoi decomposition approach[36] was used to convert the dataset of 110 160 oxidation state-decorated structures into graphs and then from these graphs, another dataset of the co-occurring species pairs is developed. Species that are

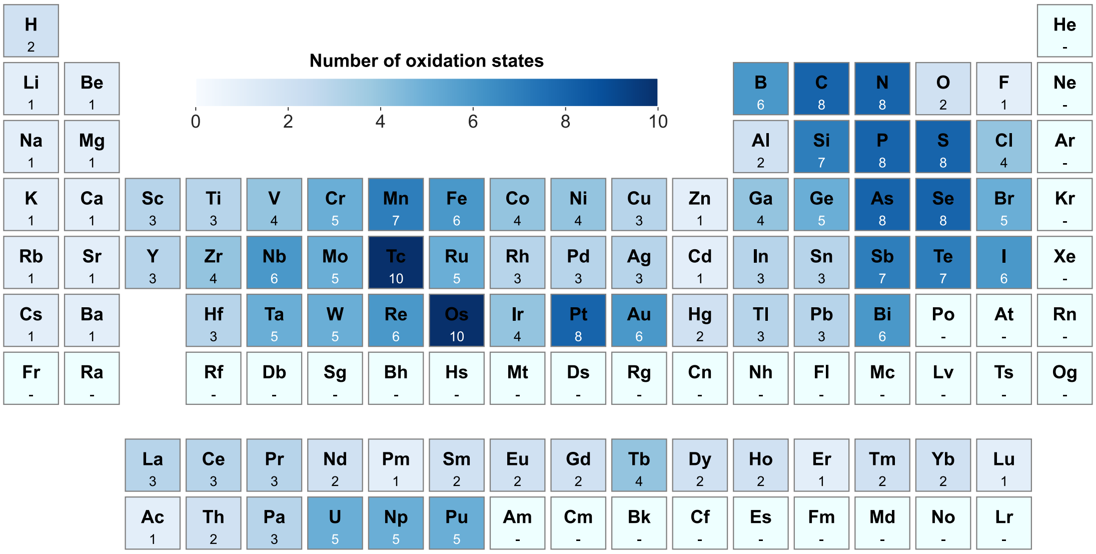

FIG. 1. Periodic table heatmap of the number of non-zero oxidation states available for each element in the dataset of 110 160 oxidation-state decorated structures obtained from the materials project.[31] This figure was created using `pymatviz`.[32]

connected in the graph representation of the structure can be considered to co-occur and make up the training pairs for learning distributed representations.

The `SkipSpecies` approach for learning representations of ionic species is that a "fake" learning task is used to predict the species that co-occur with a target species in a given structure. This task is referred to as "fake" since the aim is not to build a classifier for predicting what species will co-occur with a target species but rather to use the learned matrix of parameters as an embedding table for the species within the dataset of materials.

The general model architecture for training `SkipSpecies` models is adapted from `SkipAtom`[30] and consists of an input layer of 336 neurons, a hidden embedding layer with $d$ neurons, and an output layer with 336 neurons with softmax activation. The main modification compared to `SkipAtom` is that the input and output layers contain 336 neurons instead of 86 neurons to account for each unique species in the dataset. For the hidden dimension $d$, models with the following dimensions are chosen: 30, 86, 100, 200, 300, and 400. The input to this model is the one-hot representation of the target species, and the "fake" target variable is the one-hot representation of a species that co-occurs with the target. The loss function for this model is the cross-entropy loss between the one-hot representation of the target species and the probabilities produced by the softmax activation on the output layer of the model. The distributed representations are obtained from the embedding layer.

### *1. Induction*

Species that are under-represented in the dataset will receive fewer parameter updates. The resulting representations are likely to be of a lower quality than the representations for the more frequently represented species. This is also a problem within natural language processing (NLP). An NLP solution to this problem is to apply an optional post-processing technique called induction.[37] This involves adjusting the learned representations into a more sensible area of the representation space by using the learned vectors of their most similar species. To achieve this, a quadruple tuple of periodic group table number, row number, electronegativity, and oxidation state is used to represent each ionic species. The cosine similarity of this 4-dimensional vector is used to determine the nearest neighbors of the species. From this, the induced representation, $\hat{\boldsymbol{u}}$ can be define from the original representation $\mathbf{u}$,

$$\hat{\boldsymbol{u}} = \mathbf{u} + \frac{1}{N}\sum_{k=1}^{N} e^{-(k-1)}\mathbf{v}_k, \tag{1}$$

where $N$ is the number of the most similar species considered and $\mathbf{v}_k$ is the learned representation of the $k$th nearest neighbor. The five nearest neighbors are used in this study.

## C. Creating composition-based feature vectors

To represent an $N$-ary ionic composition, $\mathbf{X}$, which can be described as containing species $x_1, x_2, \ldots, x_N$, let us consider the quaternary case where $X = (a, b, c, d)$. We can apply a pooling operation, $f_p$, to obtain a vector that represents the composition,

$$\mathbf{V}_X = f_p(n_a, v_a, n_b, v_b, n_c, v_c, n_d, v_d), \tag{2}$$

where $n_i$ and $v_i$ are the stoichiometry and the vector representation of the species, $i$, in the composition $\mathbf{X}$, respectively. $V_X$ is the CBFV that represents the composition. Typically, the pooling operation, $f_p$, can be a combination of summary statistics, such as the mean, variance, sum, maximum, and minimum of the vectors. In this study, we create the CBFVs using the mean, maximum, and sum of the element/species vectors.

Max pooling takes the maximum value of each component in the vector representations,

$$\mathbf{V}_X = \max_{i=1}^{k} n_i \mathbf{v}_i. \tag{3}$$

The max pooling operation returns a vector of the same dimensions as the constituent vector representations, where each component of the vector is the maximum value of that component of the $m$ constituent vectors.

Mean pooling involves taking a component-wise sum of the constituent vector representations of the composition and dividing them by the total number of atoms in the composition,

$$\mathbf{V}_X = \frac{\sum_{i=1}^{k} n_i \mathbf{v}_i}{\sum_{i=1}^{k} n_i}. \tag{4}$$

The sum pooling operation is very similar to the mean pooling operation, with the exception that there is no division by the total number of atoms in the composition,

$$\mathbf{V}_X = \sum_{i=1}^{k} n_i \mathbf{v}_i. \tag{5}$$

## D. SkipSpecies evaluation

To evaluate the performance of the distributed representations of `SkipAtom` and `SkipSpecies`, the `ElemNet`[38] architecture (as chosen in the original `SkipAtom` paper[30]) with the CBFVs derived from the learned representations as inputs is applied to two classification and two regression tasks: metallic and magnetic classification, and formation energy per atom and bandgap, respectively. This work focuses on observing the difference in performance of the distributed representations for elements vs atoms as `SkipAtom` has been shown to give a superior performance to one-hot, random, and Mat2Vec[39] element representations.[30] As this dataset contained polymorphs of the same composition, it was further filtered by only keeping the composition of a polymorph with the lowest energy above the convex hull. This reduced the dataset to 71 470 materials. The property datasets are shown in Fig. 2.

The `ElemNet` model is implemented in TensorFlow[40,41] and is a 17-layer feed-forward neural network that consists of 41 024 neuron layers, 3 × 512 neuron layers, 3 × 256 neuron layers, 3 × 128 neuron layers, 2 × 64 neuron layers, and 1 × 32 neuron layer. All the layers use ReLU activation. For the classification tasks, the output layer is a single neuron layer with sigmoid activation, and the loss function is the binary cross-entropy. The regression task uses an output layer with linear activation and the loss function is the mean absolute error (MAE). The models were trained using the following hyperparameters: a maximum number of epochs of 100, a learning rate of $10^{-4}$, a batch size of 32, and an L2 lambda of $10^{-5}$.

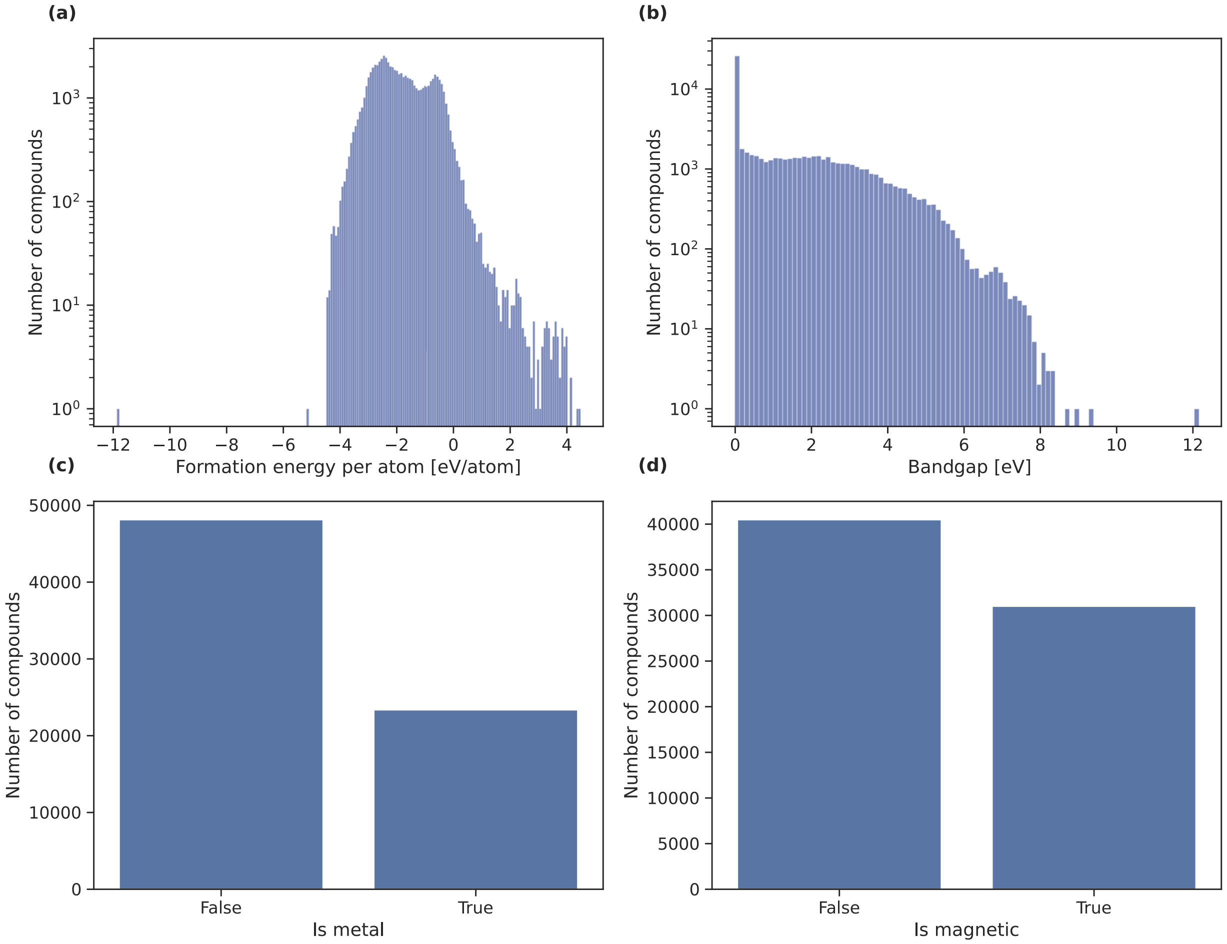

**FIG. 2.** Distributions of the target variables of the 71 470 compositions in the property prediction dataset. (a) Distribution of the formation energy per atom. (b) Distribution of bandgap. (c) Distribution of metallic classification label. (d) Distribution of magnetic classification label.

Twice-repeated fivefold cross-validation is performed to evaluate the performance of the compound representations on the property prediction tasks. The reported metric on each task is the average MAE and the average AUC (area under the receiver-operating characteristic curve) for the regression and classification tasks, respectively. The error in the metrics is the standard deviation across the twice-repeated fivefold cross-validation.

## III. RESULTS AND DISCUSSION

### A. SkipSpecies training dataset

Figure 3 shows the distribution of unique components (elements/species) per structure in the dataset that has been curated for training the `SkipSpecies` vectors, as described in Sec. II B. In Figs. 3(a) and 3(b), the one-component materials come from polymorphs of $H_2$ and $N_2$, which from the oxidation states route of the Materials Project[31] are automatically assigned oxidation states of 0+. Figure 3(a) shows that the mode number of unique elements in the dataset is three, reflecting that ternary materials are the most frequent in this dataset although this is very closely followed by quaternary materials. This is reversed in Fig. 3(b) when we consider the unique species within the structures, where the most frequent number becomes four. This distribution shift arises from the presence of many mixed-valence compounds within this dataset. Further evidence of this is shown in that the maximum number of components when we consider species becomes ten vs the case when we only consider the elements and would thus have at most nine component compounds.

Figure 3(c) further shows the breakdown between the number of components that each structure has when we consider either unique species or elements. With the exception of unary and nonary materials, mixed valency is present in all the materials. What is further illustrated is that the materials in this dataset can have one or more elements where mixed valency is observed. An example of this is the elemental quaternary material $Li_{10}GeP_2S_{12}$, a known lithium superionic conductor,[42] where depending on the crystal structure, it can be an ionic quinary material (mp-696 128) where

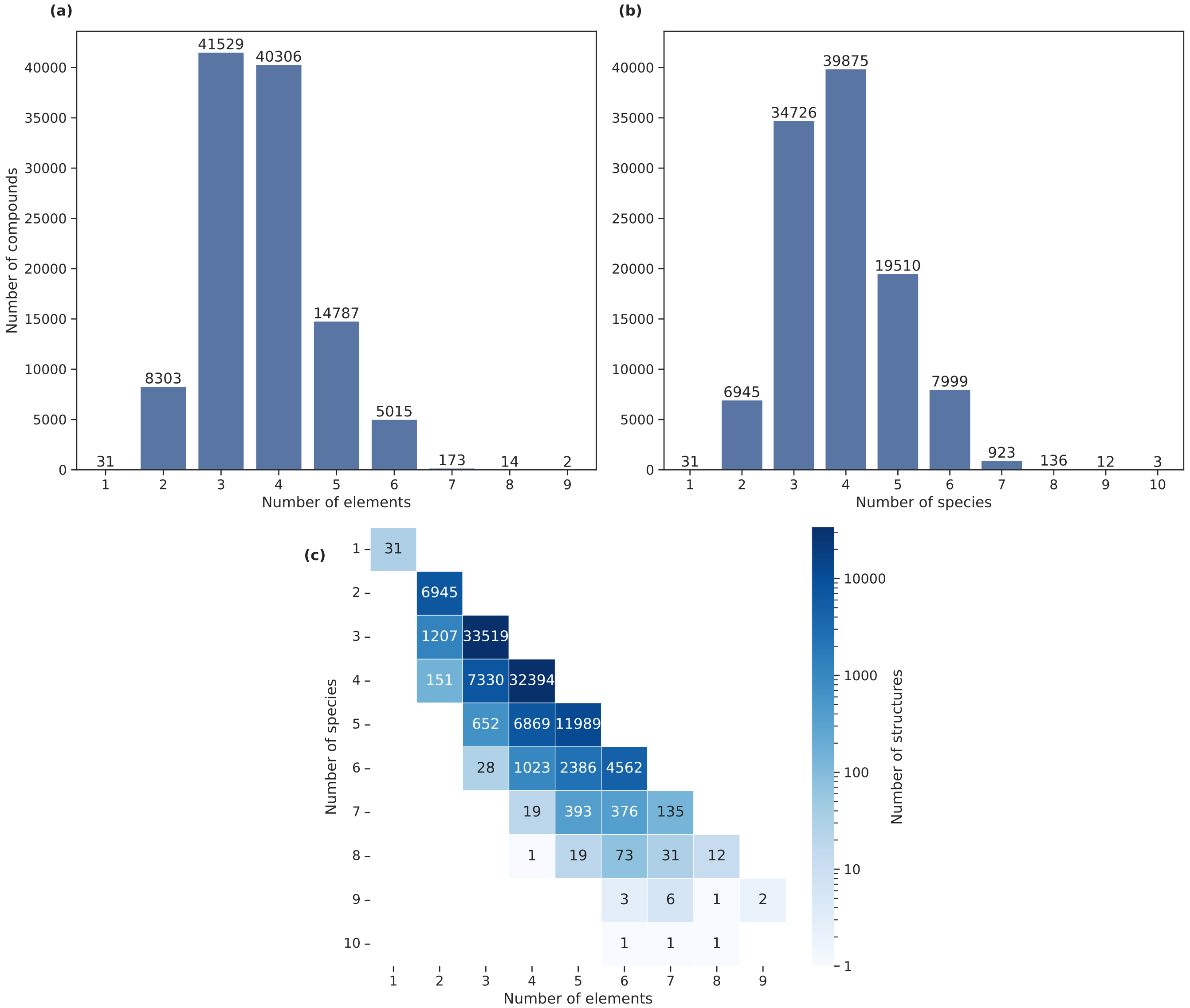

**FIG. 3.** Component distribution of the curated oxidation state decorated dataset. (a) Distribution of unique elements per structure in the dataset. (b) Distribution of unique ionic species per structure in the dataset. (c) Heatmap of the number of structures containing unique elements and unique species. The number of structures was transformed with log 10 and is displayed as the heatmap color.

S exhibits mixed valency as $S^-$ and $S^{2-}$ or an ionic senary material (mp-696 138) where P and S both exhibit mixed valency as and $P^{4+}$ and $P^{5+}$, and $S^-$ and $S^{2-}$, respectively.

## B. Learned representations

Various techniques exist to reduce high-dimensional data to make visualization easier for us to understand. In Fig. 4, we have chosen the uniform manifold approximation and projection (UMAP)[43] and the t-stochastic neighbors embedding (t-SNE).[44] Visualizing the embeddings can provide a qualitative understanding of the quality of the learned representations. It is apparent that we can recover expected chemical trends but also find patterns that we would not intuitively expect. For example, in Fig. 4(a), we find that the halides with the exception of the chloride ion appear close to each other within this space, with $Br^-$ and $I^-$ close to each other than with $F^-$. To further emphasize this, a red ellipse has been drawn around these halide ions, which further highlights how the chloride ion is missing in this cluster in the UMAP reduction of the `SkipSpecies`. All the halides cluster together when t-SNE is used for the dimension reduction, as shown in Fig. 4(b) (the iodide and bromide anions overlap each other in the reduced space). The anions can generally

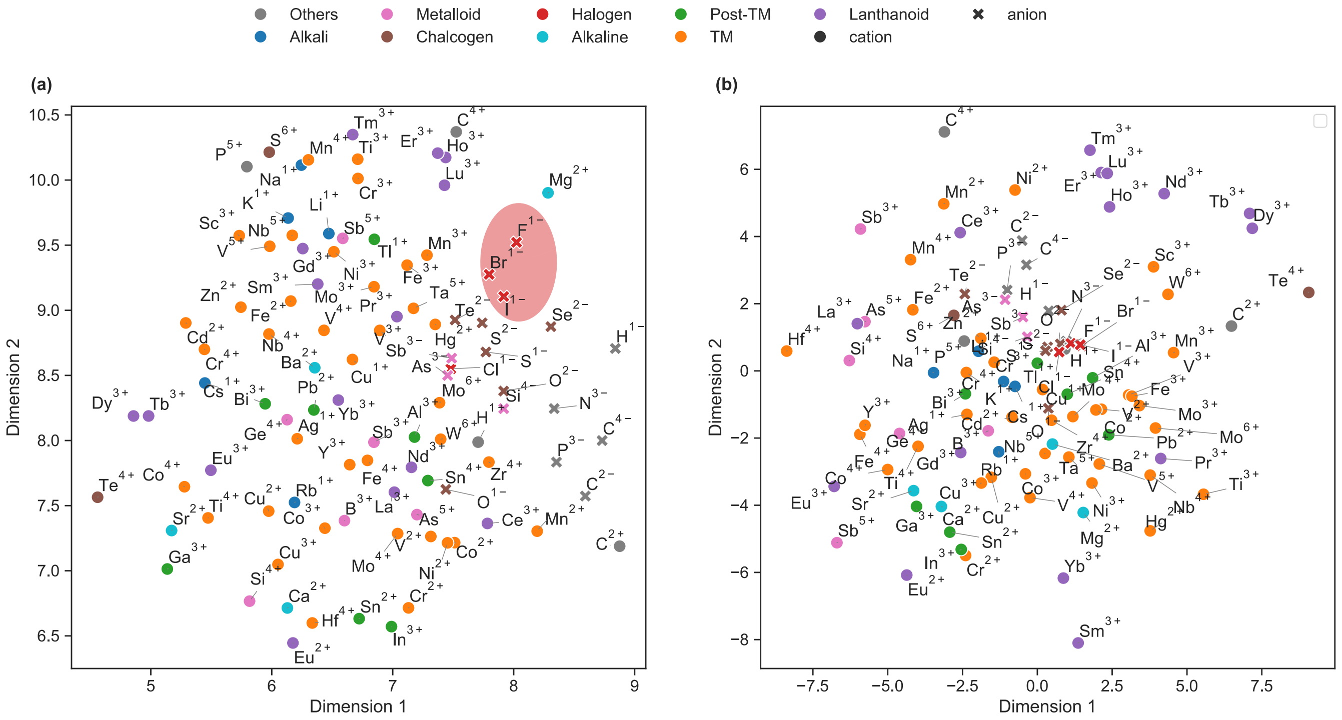

**FIG. 4.** Dimension-reduced SkipSpecies vectors. (a) UMAP was used to reduce the 200-dimension SkipSpecies representation to 2 dimensions. The red ellipse around $F^{1-}$, $Br^{1-}$, and $I^{1-}$ serves to emphasize their cluster and their separation from $Cl^{1-}$ in the plot. (b) t-SNE was used to reduce the representation to 2 dimensions. The 100 most prevalent ions in the dataset are shown for visual clarity.

be separated from the cations in the reduced space, although in the t-SNE figure, they form a cluster with a few outliers, including $Te^{2-}$, $C^{2-}$, and $C^{4-}$. For the UMAP figure, we can still observe a cluster although it is more spread out compared to the t-SNE space.

## C. Property prediction evaluation

### *1. Pooling effect*

In Fig. 5, the error metrics of the property prediction tasks using the induced SkipSpecies representation are shown over a range of dimensions of the representation, with different curves representing the different choices of pooling to make the CBFV. Independent of the task, the dimension, and the representation, it is evident that creating a CBFV using max-pooling leads to worse performance for the property prediction tasks. This result likely occurs as not all the information that is present from the constituent species vectors is used in the max pooling operation, whereas sum and mean pooling do use information from all the constituent species vectors. It is possible that the max-pooling operation can, in some cases, neglect the element or species that is most important within a particular composition for the prediction of particular properties. To further expand on this rationale, the max-pooling approach is very sensitive to outliers, as species vectors that may have anomalously high components can dominate the components of the CBFV. Mean-pooling would be less sensitive to outliers due to taking the average over each component, making it more robust to outliers. The resulting max-pooled CBFV as such may not fully describe a composition.

The mean and sum-pooled CBFVs have comparable performances with each other. This can be rationalized as both pooling operations aggregate information from the constituent vectors that make up the CBFV. This aggregation of info preserves information about the species that make up the composition. Mathematically, the sum and mean-pooled CBFVs of a composition $\mathbf{X}$ are related as follows:

$$\mathbf{V}_{X,mean} = \frac{1}{\sum_{i=1}^{k} n_i} \times \mathbf{V}_{X,sum}. \tag{6}$$

The division by the total number of atoms to create the mean-pooled vector is a constant and hence the sum-pooled and mean-pooled CBFV have a linear relationship to each other.

### *2. Dimension effect*

For creating SkipSpecies representations of the chemical species, there is a choice of dimensions for the resulting distributed representations. The effect of the choice in the dimensionality is shown in Fig. 5. It can be observed that, generally, as the number of dimensions increases, the performance of the models also increases. This trend is more dramatic for the max-pooled CBFVs. For both the

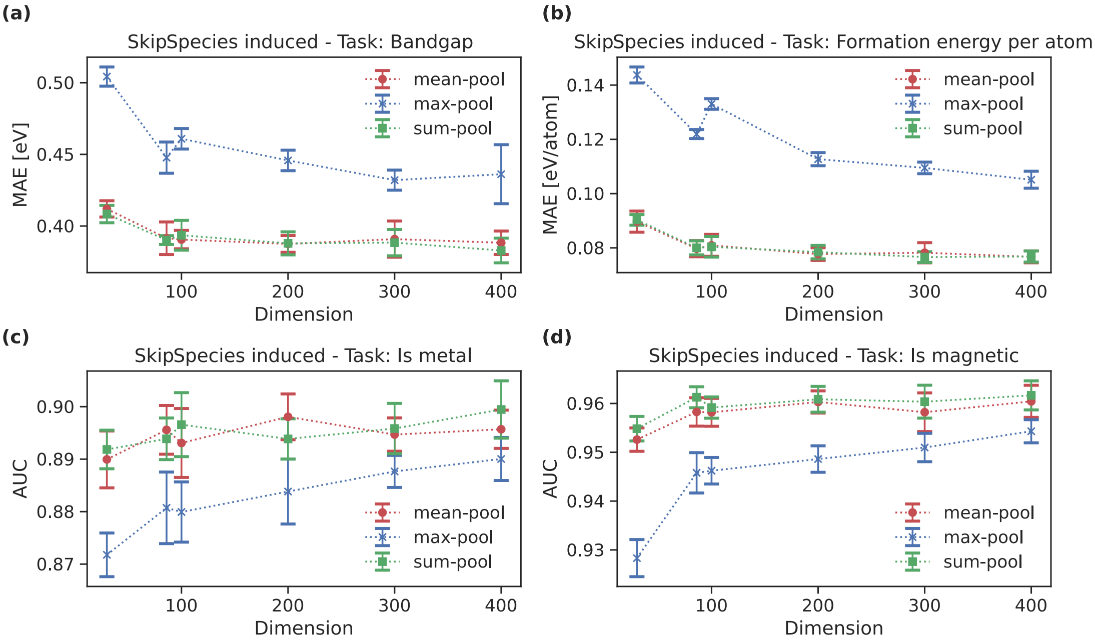

**FIG. 5.** Error metrics for the SkipSpecies-induced representation for the four property prediction tasks. (a) Mean absolute error (MAE) for the bandgap prediction task. (b) MAE for the formation energy per atom task. (c) Area under the curve (AUC) for the metallic classification task. (d) AUC for the magnetic classification task. The different colored lines in each inset refer to the pooling operation applied to make the CBFV. All the figures depict the average error metric over the twice repeated fivefold cross-validation. The error bars represent the standard deviation.

sum-pooled and mean-pooled CBFVs, there are marginal increases in performance beyond 200 dimensions.

Within the NLP field for training word embeddings, an arbitrary dimension is often chosen. For the word embeddings, setting a higher number of dimensions usually results in higher-quality embeddings up until a saturation point.[45] We have observed this from our property prediction tasks. The performance tends to improve with dimension size because a higher number of dimensions can capture more complex relationships between the co-occurring pairs. However, the effectiveness of increasing dimensions is ultimately constrained by the available data size, which limits the ability to learn meaningful patterns.

### 3. Representation effect

To better visualize the effect of the choice of representation on the performance of the four property prediction tasks, heatmaps of the sum-pooled representations have been shown in Fig. 6. For the bandgap prediction task, the SkipSpecies representations perform better than the induced SkipAtom representation across all dimensions. While the choice to apply induction to the SkipSpecies representation does appear to offer a slight improvement to the MAE, the error in these values makes it hard to discern if the application of induction is significant in upgrading the performance of the SkipSpecies representation. For the formation energy per atom task, the induced SkipAtom representation outperforms both SkipSpecies representations across all the dimensions. For each of the representations, there is a marginal improvement in the MAE beyond 200 dimensions, if any. The induced SkipSpecies representation performs the best on both the metallic and magnetic classification tasks, with the SkipAtom representation performing the worst.

To further highlight the effect of the representation on the model performance, we have shown a plot of how the validation metric (MAE for the regression tasks and AUC for the classification tasks) changes during training in Fig. 7. Except for the formation energy task shown in Fig. 7(d), the element representation SkipAtom performs the worst compared to the ionic SkipSpecies representations. For the other three tasks, the SkipSpecies representation achieves better results from the start of the training process, and this is maintained throughout the 100 epochs.

For the bandgap task, an ionic representation may offer better performance than a representation based on the neutral atom due to the knowledge of the oxidation states. This can be rationalized by considering that the oxidation state of an ion allows the model to distinguish between the properties of different materials containing the same element. The loss or gain of electrons affects both the effective radius of a species, impacting its local structural environment and electronic configuration, both of which can alter the bandgap. As the embeddings are learned such that species which occur within similar environments should be similar, ionic representations may provide the flexibility to describe different types

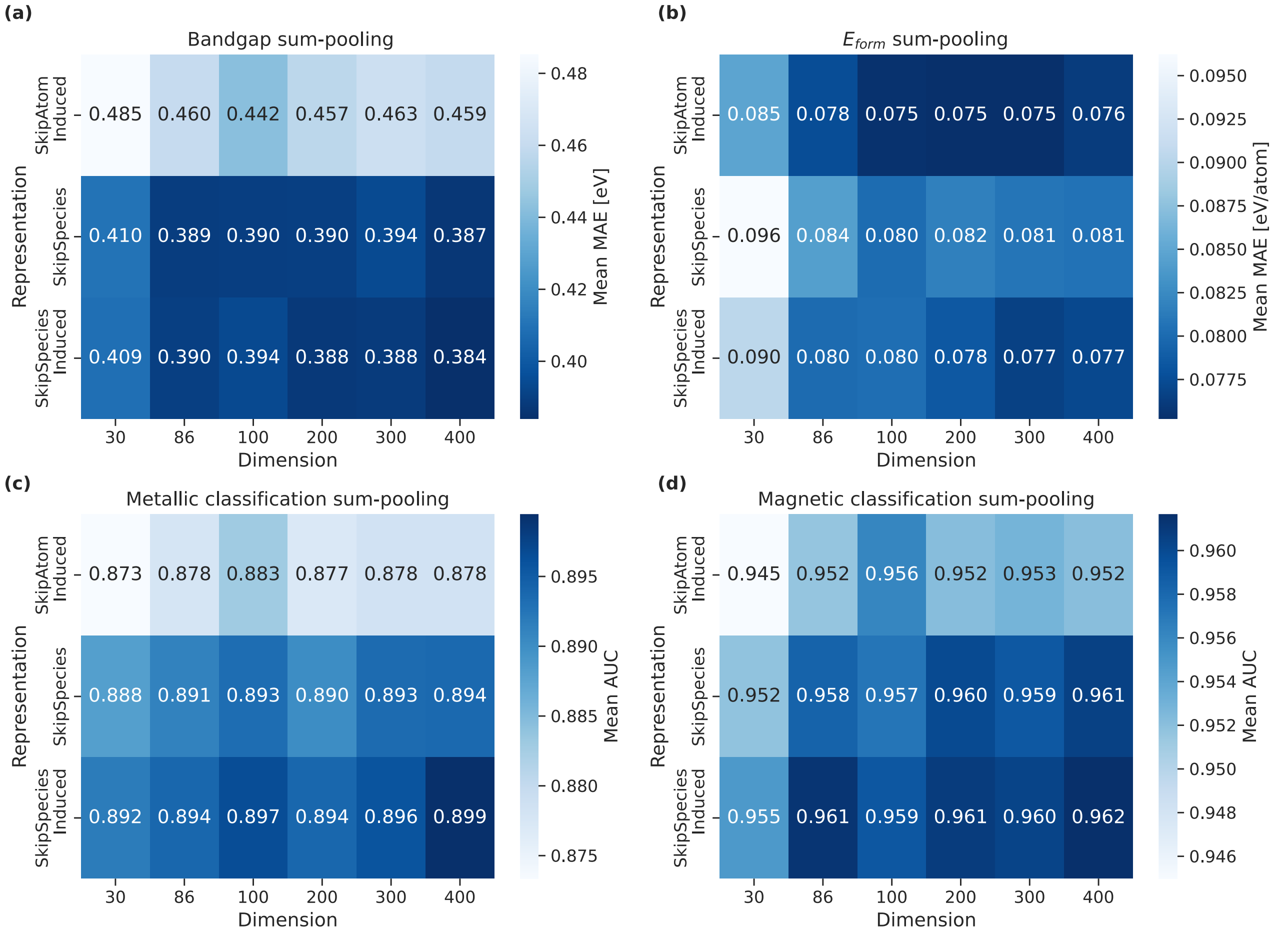

**FIG. 6.** Heatmaps of the performance of the three representation schemes where the constituent vectors were sum-pooled. (a) Heatmap for the bandgap prediction task. (b) Heatmap for the formation energy per atom prediction task. (c) Heatmap for the metallic classification task. (d) Heatmap for the magnetic classification task. The darker colors correspond to better performance in each task. All the figures depict the average error metric over the twice repeated fivefold cross-validation.

of compounds containing the same element. For example, $TiO_2$ containing $Ti^{4+}$ has a wide bandgap above 3 eV, while $Ti_2O_3$ containing $Ti^{3+}$ has a small bandgap closer to 0 eV. This flexibility may not be captured by atom-only representations.

The metallic classification task can be rationalized by an important caveat that this task is based on Materials Project[31] data. This classification is based on the bandgap calculated with semi-local density functional theory, so it is possible that some of the compounds in this dataset that are labeled metals could, in fact, be semiconductors due to a bandgap underestimation at this level of theory.[46,47] In addition, it is likely that many known metallic compounds are likely to have been excluded from the construction of the dataset. This occurred due to the need to assign oxidation states to materials, which typically leads to the exclusion of many intermetallic compounds.

The magnetic classification task is less common for property prediction tasks, as noted by its absence from MatBench.[10] The boost in performance from considering ionic representations possibly arises from how magnetism arises from the electronic structure of a material through the spins of unpaired electrons centered on atomic sites. The oxidation states of the ions implicitly encode whether particular ions may possess paired or unpaired electrons depending on the crystal environments and species that they co-occurred within during the original training of the distributed representations, which could explain the difference in performance between the `SkipSpecies` and `SkipAtom` representations.

It is important to note that the performances shown will be influenced both by the quality of the representation and the chosen model architecture. Factors that can affect the quality of the representations include how often a particular species is represented in the dataset. In this work, we applied induction[37] to the `SkipSpecies` representation to compensate for the under-represented species. The induction appears to offer a small boost in performance based on the error metrics.

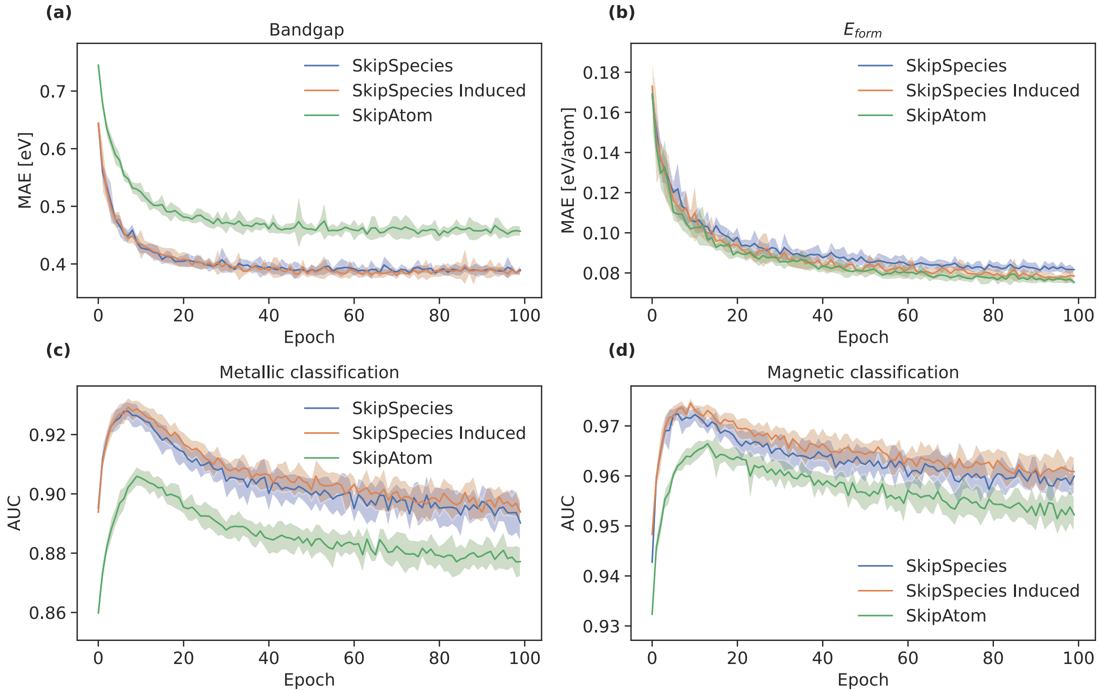

**FIG. 7.** Validation during training for the property prediction tasks using the 200 dimension atomic/ionic representations. (a) Validation MAE for the bandgap prediction task. (b) Validation MAE for the formation energy per atom task. (c) Validation AUC for the metallic classification task. (d) Validation MAE for the magnetic classification. The performance metrics are averaged over the twice-repeated fivefold cross-validation. The error is the standard deviation across the twice-repeated fivefold cross-validation.

## IV. CONCLUSIONS

We have explored the development of ionic species representations for crystals from chemical data using machine learning. This work builds upon SkipAtom[30] and some simple ion featurizers in Matminer[17] based on oxidation states and electronegativity.

The SkipSpecies ionic representations can be used to develop property prediction models with lower errors than comparable atomic representations for predicting properties such as the bandgap or classifying compositions as metals and non-metals, and magnetic or non-magnetic, suggesting that there may be some utility for ionic representations for predicting the properties of compositions. One caveat is that ionic representations are more restrictive compared to element representations, as the oxidation states in the composition have to be known to use them for property prediction in a structure agnostic setting. In addition, the quality of the trained SkipSpecies is dependent on the correct assignment of charges used to build the dataset of pairs. While tools such as pymatgen[33] and BERTOS[48] can be used to assign oxidation states to compositions, this does introduce an additional step into a workflow to predict properties and may fail to be assigned physical charges for certain compounds.

These ionic representations may find use for property prediction in approaches that create or generate compositions alongside knowledge of the oxidation states of the constituent elements, rather than having to decorate an already existing set of compositions where this information is not already known. One example where these ionic representations could be used is within SMACT-based workflows, as the chemical filters used to generate compositional spaces also return both the constituent elements and the oxidation states of the compositions. These representations can be used for both property prediction on these spaces, as well as providing an alternative means to visualize the compositional space as opposed to using elemental representations to visualize this space, since compositions of the same formula but with elements in different oxidation states would be different points in the ionic space instead of the same point in the elemental space.

Distributed species representations may have applications for crystal structure assignment by analogy through ionic substitutions, as pairwise similarity values can be derived from the vectors using distance or similarity measures. Alternatively, similarity measures can be applied to compositional feature vectors derived from these representations to suggest what known materials are similar to hypothetical compositions as part of synthesizability models.[49] Finally, we note that such representations are not limited to compositional (structure free) models, but could be used, for example, to initialize node vectors on graph-based models of materials structure and properties.

## ACKNOWLEDGMENTS

We thank Luis M. Antunes and Ricardo Grau-Crespo for the insightful discussions. A.O. acknowledges EPSRC for a Ph.D.

studentship (Grant No. EP/T51780X/1). We acknowledge the UK Materials and Molecular Modeling Hub for computational resources, which is partially funded by EPSRC (Grant Nos. EP/P020194/1 and EP/T022213/1). We also acknowledge the Imperial College Research Computing Service for its computational resources.

## AUTHOR DECLARATIONS

### Conflict of Interest

The authors have no conflicts to disclose.

### Author Contributions

**Anthony Onwuli**: Conceptualization (equal); Data curation (equal); Formal analysis (equal); Investigation (equal); Methodology (equal); Writing – original draft (equal). **Keith T. Butler**: Conceptualization (equal); Project administration (equal); Supervision (equal); Writing – review & editing (equal). **Aron Walsh**: Conceptualization (equal); Project administration (equal); Supervision (equal); Writing – review & editing (equal).

## DATA AVAILABILITY

The data that support the findings of this study are openly available as an archive at https://doi.org/10.5281/zenodo.12733915 and are maintained at https://github.com/WMD-group/skipspecies. The ionic representation schemes have also been including in the `ElementEmbeddings` package available from https://github.com/WMD-group/ElementEmbeddings.

## REFERENCES

[1] A. Ihalage and Y. Hao, "Formula graph self-attention network for representation-domain independent materials discovery," Adv. Sci. **9**, 2200164 (2022).

[2] C. Chen, W. Ye, Y. Zuo, C. Zheng, and S. P. Ong, "Graph networks as a universal machine learning framework for molecules and crystals," Chem. Mater. **31**, 3564–3572 (2019).

[3] T. Xie and J. C. Grossman, "Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties," Phys. Rev. Lett. **120**, 145301 (2018).

[4] J. Schmidt, L. Pettersson, C. Verdozzi, S. Botti, and M. A. L. Marques, "Crystal graph attention networks for the prediction of stable materials," Sci. Adv. **7**, eabi7948 (2021).

[5] S.-Y. Louis, Y. Zhao, A. Nasiri, X. Wang, Y. Song, F. Liu, and J. Hu, "Graph convolutional neural networks with global attention for improved materials property prediction," Phys. Chem. Chem. Phys. **22**, 18141–18148 (2020).

[6] K. Choudhary and B. DeCost, "Atomistic line graph neural network for improved materials property predictions," npj Comput. Mater. **7**, 185–188 (2021).

[7] R. Ruff, P. Reiser, J. Stühmer, and P. Friederich, "Connectivity optimized nested line graph networks for crystal structures," Digit. Discov. **3**, 594–601 (2024).

[8] J. Gasteiger, J. Groß, and S. Günnemann, "Directional message passing for molecular graphs," in International Conference on Learning Representations (ICLR), 2020.

[9] J. Gasteiger, S. Giri, J. T. Margraf, and S. Günnemann, "Fast and uncertainty-aware directional message passing for non-equilibrium molecules," in Machine Learning for Molecules Workshop, NeurIPS, 2020.

[10] A. Dunn, Q. Wang, A. Ganose, D. Dopp, and A. Jain, "Benchmarking materials property prediction methods: The Matbench test set and Automatminer reference algorithm," npj Comput. Mater. **6**, 138–210 (2020).

[11] E. D. Cubuk, A. D. Sendek, and E. J. Reed, "Screening billions of candidates for solid lithium-ion conductors: A transfer learning approach for small data," J. Chem. Phys. **150**, 214701 (2019).

[12] A. D. Sendek, Q. Yang, E. D. Cubuk, K.-A. N. Duerloo, Y. Cui, and E. J. Reed,"Holistic computational structure screening of more than 12000 candidates for solid lithium-ion conductor materials," Energy Environ. Sci. **10**, 306–320 (2017).

[13] A. O. Oliynyk, E. Antono, T. D. Sparks, L. Ghadbeigi, M. W. Gaultois, B. Meredig, and A. Mar, "High-throughput machine-learning-driven synthesis of full-Heusler compounds," Chem. Mater. **28**, 7324–7331 (2016).

[14] B. Meredig, A. Agrawal, S. Kirklin, J. E. Saal, J. W. Doak, A. Thompson, K. Zhang, A. Choudhary, and C. Wolverton, "Combinatorial screening for new materials in unconstrained composition space with machine learning," Phys. Rev. B **89**, 094104 (2014).

[15] L. Ward, A. Agrawal, A. Choudhary, and C. Wolverton, "A general-purpose machine learning framework for predicting properties of inorganic materials," npj Comput. Mater. **2**, 16028 (2016).

[16] L. M. Ghiringhelli, J. Vybiral, S. V. Levchenko, C. Draxl, and M. Scheffler, "Big data of materials science: Critical role of the descriptor," Phys. Rev. Lett. **114**, 105503 (2015).

[17] L. Ward, A. Dunn, A. Faghaninia, N. E. Zimmermann, S. Bajaj, Q. Wang, J. Montoya, J. Chen, K. Bystrom, M. Dylla, K. Chard, M. Asta, K. A. Persson, G. J. Snyder, I. Foster, and A. Jain, "Matminer: An open source toolkit for materials data mining," Comput. Mater. Sci. **152**, 60–69 (2018).

[18] C. J. Hargreaves, M. S. Dyer, M. W. Gaultois, V. A. Kurlin, and M. J. Rosseinsky, "The earth mover's distance as a metric for the space of inorganic compositions," Chem. Mater. **32**, 10610–10620 (2020).

[19] T. Plett, T. Gamble, E. Gillette, S. B. Lee, and Z. S. Siwy, "Ionic conductivity of a single porous $MnO_2$ mesorod at controlled oxidation states," J. Mater. Chem. A **3**, 12858–12863 (2015).

[20] J. M. Perkins, S. Fearn, S. N. Cook, R. Srinivasan, C. M. Rouleau, H. M. Christen, G. D. West, R. J. Morris, H. L. Fraser, S. J. Skinner *et al.*, "Anomalous oxidation states in multilayers for fuel cell applications," Adv. Funct. Mater. **20**, 2664–2674 (2010).

[21] N. S. Garnet, V. Ghodsi, L. N. Hutfluss, P. Yin, M. Hegde, and P. V. Radovanovic, "Probing the role of dopant oxidation state in the magnetism of diluted magnetic oxides using Fe-doped $In_2O_3$ and $SnO_2$ nanocrystals," J. Phys. Chem. C **121**, 1918–1927 (2017).

[22] I. Langmuir, "The arrangement of electrons in atoms and molecules," J. Am. Chem. Soc. **41**, 868–934 (1919).

[23] A. Walsh, A. A. Sokol, J. Buckeridge, D. O. Scanlon, and C. R. A. Catlow, "Oxidation states and ionicity," Nat. Mater. **17**, 958–964 (2018).

[24] R. J. Murdock, S. K. Kauwe, A. Y.-T. Wang, and T. D. Sparks, "Is domain knowledge necessary for machine learning materials properties?," Integr. Mater. Manuf. Innovation **9**, 221–227 (2020).

[25] D. Davies, K. Butler, A. Jackson, A. Morris, J. Frost, J. Skelton, and A. Walsh, "Computational screening of all stoichiometric inorganic materials," Chem **1**, 617–627 (2016).

[26] D. W. Davies, K. T. Butler, A. J. Jackson, J. M. Skelton, K. Morita, and A. Walsh, "SMACT: Semiconducting materials by analogy and chemical theory," J. Open Source Software **4**, 1361 (2019).

[27] D. Davies, K. T. Butler, O. Isayev, and A. Walsh, "Materials discovery by chemical analogy: Role of oxidation states in structure prediction," Faraday Discuss. **211**, 553–568 (2018).

[28] T. Xie, X. Fu, O.-E. Ganea, R. Barzilay, and T. Jaakkola, "Crystal diffusion variational autoencoder for periodic material generation," arXiv:2110.06197 [cond-mat, physics:physics] (2022).

[29] Y. Dan, Y. Zhao, X. Li, S. Li, M. Hu, and J. Hu, "Generative adversarial networks (GAN) based efficient sampling of chemical composition space for inverse design of inorganic materials," npj Comput. Mater. **6**, 184 (2020).

[30] L. M. Antunes, R. Grau-Crespo, and K. T. Butler, "Distributed representations of atoms and materials for machine learning," npj Comput. Mater. **8**(1), 44 (2022).

[31] A. Jain, S. P. Ong, G. Hautier, W. Chen, W. D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder, and K. A. Persson, "Commentary: The materials project: A materials genome approach to accelerating materials innovation," APL Mater. **1**, 011002 (2013).

[32] J. Riebesell, H. Yang, R. Goodall, and S. G. Baird (2022). "Pymatviz: Visualization toolkit for materials informatics," GitHub. https://github.com/janosh/pymatviz.

[33] S. P. Ong, W. D. Richards, A. Jain, G. Hautier, M. Kocher, S. Cholia, D. Gunter, V. L. Chevrier, K. A. Persson, and G. Ceder, "Python materials genomics (pymatgen): A robust, open-source python library for materials analysis," Comput. Mater. Sci. **68**, 314–319 (2013).

[34] M. O'Keefe and N. E. Brese, "Atom sizes and bond lengths in molecules and crystals," J. Am. Chem. Soc. **113**, 3226–3229 (1991).

[35] G. Bergerhoff, R. Hundt, R. Sievers, and I. Brown, "The inorganic crystal structure data base," J. Chem. Inf. Comput. Sci. **23**, 66–69 (1983).

[36] G. 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.

[37] M. T. Pilehvar and N. Collier, *Inducing Embeddings for Rare and Unseen Words by Leveraging Lexical Resources* (Association for Computational Linguistics, 2017).

[38] D. Jha, L. Ward, A. Paul, W.-K. Liao, A. Choudhary, C. Wolverton, and A. Agrawal, "*ElemNet*: Deep learning the chemistry of materials from only elemental composition," Sci. Rep. **8**, 17593 (2018).

[39] V. Tshitoyan, J. Dagdelen, L. Weston, A. Dunn, Z. Rong, O. Kononova, K. A. Persson, G. Ceder, and A. Jain, "Unsupervised word embeddings capture latent knowledge from materials science literature," Nature **571**, 95–98 (2019).

[40] T. Developers, Tensorflow (2023).

[41] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng, "TensorFlow: Large-scale machine learning on heterogeneous systems" (2015), https://www.tensorflow.org/about/bib; software available from tensorflow.org.

[42] N. Kamaya, K. Homma, Y. Yamakawa, M. Hirayama, R. Kanno, M. Yonemura, T. Kamiyama, Y. Kato, S. Hama, K. Kawamoto, and A. Mitsui, "A lithium superionic conductor," Nat. Mater. **10**, 682–686 (2011).

[43] L. McInnes, J. Healy, and J. Melville, "UMAP: Uniform manifold approximation and projection for dimension reduction," arXiv:1802.03426 [stat.ML] (2020).

[44] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne," J. Mach. Learn. Res. **9**, 2579 (2008).

[45] J. Pennington, R. Socher, and C. D. Manning, "GloVe: Global vectors for word representation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (IEEE, 2014) pp. 1532–1543.

[46] P. Mori-Sánchez and A. J. Cohen, "The derivative discontinuity of the exchange–correlation functional," Phys. Chem. Chem. Phys. **16**, 14378–14387 (2014).

[47] L. J. Sham and M. Schlüter, "Density-functional theory of the energy gap," Phys. Rev. Lett. **51**, 1888 (1983).

[48] N. Fu, J. Hu, Y. Feng, G. Morrison, H.-C. Z. Loye, and J. Hu, "Composition based oxidation state prediction of materials using deep learning language models," Adv. Sci. **10**, 2301011 (2023).

[49] E. R. Antoniuk, G. Cheon, G. Wang, D. Bernstein, W. Cai, and E. J. Reed, "Predicting the synthesizability of crystalline inorganic materials from the data of known material compositions," npj Comput. Mater. **9**, 155 (2023).