A Graph Neural Network for Predicting Energy and Stability of Known and Hypothetical Crystal Structures

Shubham Pandey,[a] Jiaxing Qu,[b] Vladan Stevanović,[a] Peter St. John,[c*] Prashun Gorai[a*]

The discovery of new inorganic materials in unexplored chemical spaces necessitates calculating total energy quickly and with sufficient accuracy. Machine learning models that provide such a capability for both ground-state (GS) and higher-energy structures would be instrumental in accelerating the screening for new materials over vast chemical spaces. Here, we develop a unique graph neural network model to accurately predict the total energy of both GS and higher-energy hypothetical structures. We use ∼16,500 density functional theory calculated total energy from the NREL Materials Database and ∼11,000 in-house generated hypothetical structures to train our model, thus making sure that the model is not biased towards either GS or higher-energy structures. We also demonstrate that our model satisfactorily ranks the structures in the correct order of their energies for a given composition. Furthermore, we present a thorough error analysis to explain several failure modes of the model, which highlights both prediction outliers and occasional inconsistencies in the training data. By peeling back layers of the neural network model, we are able to derive chemical trends by analyzing how the model represents learned structures and properties.

1 Introduction

With the advances in computing power and methodologies, computational chemistry and materials science has made great strides in accelerating discovery of molecules and materials with tailored properties.[1,2] The ability to perform large-scale *ab initio* calculations, in particular those based on density functional theory (DFT), has been instrumental in inorganic functional materials discovery.[3–7] However, computational searches have largely focused on *known* materials documented in crystallographic databases. Currently, there are ∼200,000 entries in the Inorganic Crystal Structure Database (ICSD),[8] which represents only a small part ($>10^{12}$ plausible compositions considering up to quarternary compounds[9]) of the vast chemical phase space of inorganic materials. The need for accelerated exploration of uncharted chemical spaces is shared by experimental and computational researchers.

The discovery of new inorganic compositions necessitates accurate structure prediction methods, which is a burgeoning field in itself. The general approach involves navigating the configuration space defined by the structural parameters, using a rapidly computable cost function such as total energy. The navigation of configuration space can use a variety of techniques, including simulated annealing,[10] genetic algorithms,[3,11] random structure searching,[12,13] structure prototyping,[14,15] and data mining,[16,17] etc. In these techniques, total energy is often predicted with DFT, although force field methods have also been used.[18,19] Thermodynamic phase stability, i.e., stability against decomposition, is another prerequisite in the search for new compositions. Formation enthalpy, calculated from DFT total energy, has proven immensely useful in assessing phase stability.[20–23] However, DFT total energy calculations are still computationally expensive to survey large chemical spaces with $> 10^6$ compounds. Machine learning (ML) models have emerged as a surrogate for fast prediction of total energy, formation enthalpy, and phase stability.[24–26] Here, we develop a graph neural network to predict the total energy of ground-state as well as hypothetical higher-energy structures generated for structure prediction.[16]

Crystal graph convolutional neural networks (CGCNN) have been developed to predict DFT total energy and formation enthalpy.[28–30] These deep learning models outperform traditional ML models with expert-designed feature representations. In a crystal graph, the atoms are represented by nodes and bonding interactions as edges connecting the nodes, which naturally takes into account the periodicity of crystal structures. Xie et al.[28] trained a CGCNN model on DFT-computed formation enthalpy of 46,744 crystal structures (predominantly from the ICSD) available in the Materials Project (MP) database.[20] Chen et al. proposed a generalized MatErials Graph Network (MEGNet) for molecules and materials that was trained on 60,000 crystal structures from MP.[30] Park et al. developed an improved-CGCNN (iCGCNN) by using Voronoi neighbors to represent the local environment of each node,[29] rather than connecting each node to the their first 12 nearest neighbors as is done in CGCNN. They incorporate explicit three-body interactions in their convolution functions as an improvement over only pairwise correlations in the CGCNN model. The model is trained on DFT formation enthalpy of 450,000 crystal structures in the Open Quantum Materials Database (OQMD).[22] The CGCNN and its variants exhibit similar accuracy in predicting formation enthalpy, with mean absolute error (MAE) of 0.03-0.04 eV/atom.[28–31]

[a]Colorado School of Mines, Golden, CO 80401. [b]University of Illinois at Urbana-Champaign, Urbana, IL 61801. [c]National Renewable Energy Laboratory, Golden, CO 80401. *E-mail: Peter.STJohn@nrel.gov, pgorai@mines.edu

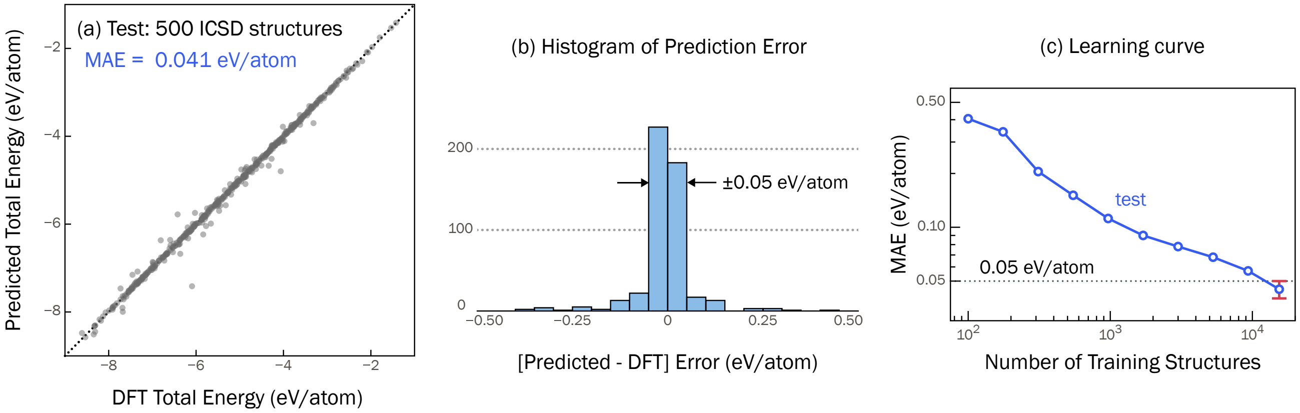

Fig. 1 CGCNN model trained on DFT total energy of ICSD structures from NREL Materials Database.[27] (a) The model predicts DFT total energy of 500 held-out crystal structures with a mean absolute error (MAE) of 0.041 eV/atom (0.95 kcal/mol). (b) Histogram of prediction errors (relative to DFT total energy) for the 500 test set structures; 82% of the structures are predicted within an error of ±0.05 eV/atom. (c) Learning curve shows that $> 10^4$ training structures are needed to achieve MAE $\leq$0.05 eV/atom.

For structure and stability predictions, it is imperative that the model is able to: (1) predict the total energy of both ground-state (GS) and higher-energy (HE) structures with similar accuracy, and (2) distinguish energetically favorable (low energy) structures from those with higher energy. The CGCNN models discussed above are trained primarily on ICSD structures, that are GS or near-GS structures. As we show in the Results and Discussion, these models are likely to be biased towards GS structures and therefore, inaccurate in predicting total energies of HE structures. While the iCGCNN model[29] is trained on both GS and HE structures, an explicit demonstration of the model performance for GS and HE structures is missing. Since the focus of that study was to improve the overall prediction accuracy, it is not clear if the resulting model can, for a given composition, correctly rank the different structures based on their total energy.

In this work, we build a hybrid CGCNN model to accurately predict the total energy of GS and HE structures by training the model on DFT total energy of ~16,500 ICSD structures from the NREL Materials Database[27] and ~11,000 hypothetical structures generated by the ionic substitution method.[32,33] The overall prediction accuracy of our hybrid model is at par with other CGCNN models (MAE = 0.04 eV/atom), with similar accuracy in predicting the total energy of GS *and* HE hypothetical structures. We demonstrate the model's capability to satisfactorily distinguish low- and higher-energy structures for a given composition. Finally, we investigate the prediction outliers and find that, in some cases, the source of the error can be traced back to the inaccuracies in the DFT total energy.

2 Results and Discussion

2.1 Model Trained on ICSD Structures

Previously reported graph neural network models for predicting total energy and formation enthalpy[28,30] were trained primarily on ICSD crystal structures with DFT total energy and formation enthalpy taken from the Materials Project.[20] For benchmarking, we train a CGCNN model on the DFT total energy of ICSD structures from the NREL Materials Database (NRELMatDB).[27] The model is trained on 15,500 crystal structures with 500 structures each withheld for validation and testing. We find that the prediction accuracy, gauged by the mean absolute error (MAE), is 0.041 eV/atom (Figure 1a). The standard deviation in the MAE is ±0.005 eV/atom, which is obtained by training 4 different models and calculating the corresponding MAE on test sets each containing 500 crystal structures, with no overlap of structures between the test sets (Figure S1). The optimized hyperparameters for the model are provided in Table S1 of the supplemental information. Hereafter, we reference this model as the "ICSD model". The learning curve is presented in Figure 1(c), which shows that at least 10^4 crystal structures are required to achieve a test MAE of <0.05 eV/atom, consistent with previous models.[28]

The formation enthalpy (ΔH_f) of a crystal structure with a chemical composition $A_xB_yC_z$ can be calculated from the DFT total energy as, $\Delta H_\mathrm{f} = E_\mathrm{total} - x\mu_A^0 - y\mu_B^0 - z\mu_C^0$, where E_total is DFT total energy of $A_xB_yC_z$ with ΔH_f and E_total expressed per formula unit and μ_i^0 are the reference chemical potentials of elements, typically under standard conditions. Since μ^0 are reference values, ΔH_f is linearly dependent on E_total. By design, the error in predicting ΔH_f is the same as in predicting E_total. The ICSD model has an MAE of 0.041 eV eV/atom for predicting DFT total energy. As such ΔH_f can be predicted with the same accuracy, which is at par with other CGCNN

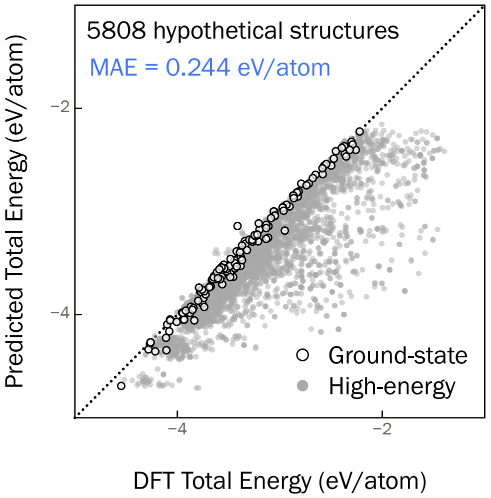

Fig. 2 Total energy of hypothetical structures (see Section 2.1 for details) predicted with the ICSD model. The total energy is systematically underpredicted for the high-energy hypothetical structures suggesting model bias toward lower-energy structures.

models reported in the literature.[28–30] Furthermore, the typical experimental error in measuring formation enthalpy is the "chemical accuracy", which on the order of 1 kcal/mol (0.043 eV/atom).[23] Assuming DFT calculated ΔH_f are reliable, the prediction error of the ICSD model is comparable to the chemical accuracy.

Figure 1(b) shows a histogram of the prediction errors relative to the DFT values, with 82% crystal structures (410 out of 500) predicted within an error of ±0.05 eV/atom. Of the remaining 90 structures lying outside the ±0.05 eV/atom error range, 51 structures are underpredicted, including PdN (space group #221) and CoMnP (space group #62) that are underpredicted by -0.733 eV/atom and -0.397 eV/atom, respectively. We find that these are higher-energy structures of those compositions reported in the ICSD, with PdN (space group #221) 0.459 eV/atom and CoMnP (space group #1) 0.400 eV/atom above the respective GS structures PdN (space group #225) and CoMnP (space group #62). Other underpredicted structures such as SiCN (space group #216) and AuN (space group #225) are highly unstable structures that lie above their respective convex hulls by 2.168 eV/atom and 1.897 eV/atom, respectively. The vast majority of ICSD structures have been determined through XRD refinement of experimentally grown crystal structures with some metastable and computationally predicted hypothetical structures. As such, ICSD is biased toward stable, GS structures; the underprediction of the high-energy/unstable structures is a testament to this inherent bias, which so far has not been acknowledged in previous studies.[28–30]

We further confirm this bias by using the ICSD model to predict the total energy of ∼5800 hypothetical structures. As described in Section 4.2 (Methods), the dataset of hypothetical structures contain, in addition to the GS structures, a number of higher-energy hypothetical structures for a given composition. The ICSD model severely underpredicts the total energy of the higher-energy hypothetical structures but accurately predicts the energy of the corresponding GS structures (Figure 2), which highlights the model bias toward GS structures. For structure and stability predictions, a model that is accurate for both GS and higher-energy structures is desired.

2.2 Model Trained on ICSD and Hypothetical Structures

To address the underestimation of the total energy of the hypothetical structures with the ICSD model, we first train a graph neural network model on the hypothetical structures separately i.e., not including the ICSD structures. The training, validation, and test sets are chosen in a way to avoid overlap of compositions across them. For instance, all the hypothetical structures associated with the composition KGeP (ABX composition) appear only in the test set (Figure 3a) but not in the training or validation set. By avoiding overlap of compositions across the sets, we can eventually test the true performance of the model in energetically ranking the different structures associated with a given composition. In addition, at least one composition type (ABX, ABX_4, ...) is present in each of the sets.

First, the overall performance of this model with MAE = 0.055 eV/atom (Figure 3a) is significantly better than the performance of the ICSD model on the same structures (Figure 2). We find that the total energy of certain composition types e.g., AX_2 (6 out of 191 compositions), that are under-represented in the hypothetical dataset are predicted with lower accuracy. In Figure 3(a), the prediction outliers are predominantly of the AX_2 composition. Nonetheless, the overall performance is comparable to the ICSD model. However, when we use this model, trained on hypothetical structures only, to predict the total energy of 1065 ICSD structures, we again find that the model performs poorly with an MAE = 0.424 eV/atom (Figure S2). As with the ICSD model (Section 2.1), this model appears to be again biased toward the hypothetical structures used in the training. To overcome this systematic bias, we find that it is practical to train a "hybrid" model simultaneously on ICSD and hypothetical structures.

A hybrid model is trained on the DFT total energy of 14,845 ICSD and 9980 hypothetical structures (in 171 compositions) and validated and tested on 800 ICSD and ∼600 hypothetical (10 compositions each) structures. An overall MAE of 0.04 is achieved across ICSD and hypothetical structures (Figure 3b), which is comparable to the prediction accuracy of the ICSD model. The standard deviation in the MAE (0.005 eV/atom) is determined by training 4 different models and calculating the corresponding MAE on test sets each containing 800 ICSD and ∼600 hypothetical structures (10 composi-

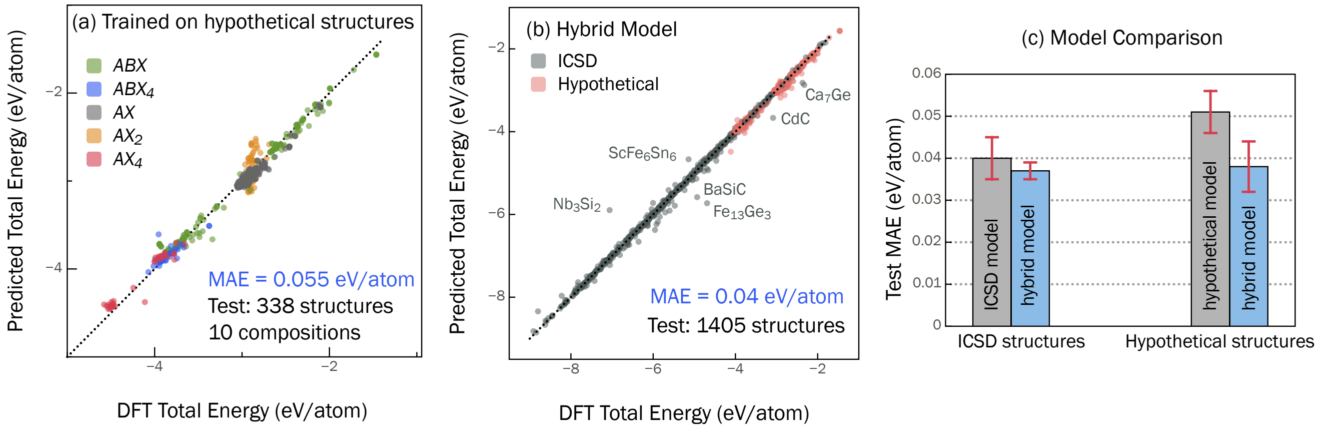

Fig. 3 (a) Predicted vs. DFT total energy of the model trained only on hypothetical structures. The data points are colored by their composition type (see Section 4.2 for details). (b) Hybrid model accurately predicts the total energy for both ICSD and hypothetical structures, with an overall MAE of 0.04 eV/atom (c) Comparison of prediction MAE of ICSD model (Figure 1a), model trained on only hypothetical structures shown in (a), and hybrid model. The standard deviation (shown as error bars) is calculated from 4 different models with non-overlapping test sets.

tions each) with no overlap in the structures (Figure S3). The learning curve is presented in Figure S4, which shows that at least 2×10^4 crystal structures (twice as required for the ICSD model) are required to achieve test MAE of <0.05 eV/atom. Figure 3(c) shows the individual MAEs for the ICSD and hypothetical structures. For comparison, the prediction MAE of the ICSD model (Section 2.1) and the model trained on the hypothetical structures alone are provided. It is evident from Figure 3(c) that the hybrid model improves the prediction accuracy for both ICSD and hypothetical structures and overcomes the model bias when each dataset is used separately to train a total energy model.

2.3 Energy Ranking of Structures

While it is crucial to have a high accuracy model for predicting total energy, it remains to be seen whether the model can rank the different structures of a given composition in the correct order of their energies. As mentioned in the Introduction, this energy ranking is desired for distinguishing energetically favorable (low energy) structures from the higher-energy unfavorable structures. Figure 4 shows the comparison between DFT and our model-predicted relative total energy ($E-E_{\mathrm{min}}$) of all the hypothetical structures for each of the 10 compositions present in the test set (Figure 3b). In general, the predicted energy rankings are in fair agreement with DFT, although there are noticeable differences depending on the composition type.

The rankings for the *ABX* type compositions e.g., KGeP, KZnSb, and NaBeAs are the most accurate i.e., the model correctly identifies the GS structure and also does not incorrectly misassign a higher-energy structure as low energy (Figure 4). The good ranking of *ABX* composition type can be attributed to the fact that *ABX* comprises the largest fraction of the training dataset of hypothetical structures (139 out of 191 compositions). The ranking for CsAs, (*AX* type composition) is satisfactory, with DFT ground-state structure predicted to be only 0.007 eV/atom higher than the GS structure predicted by the model. Moreover, none of the higher-energy structures are misassigned as low-energy structures. In the case of $KGaAs_4$, a ABX_4 type composition, the model correctly identifies the GS structure and also does not misassign any of the higher-energy structures as the ground state.

On the other hand, for the AX_2 type compositions (e.g. $ZnAs_2$, $CdSb_2$, $CdBi_2$), the energy ranking of the structures requires a more detailed examination. The model correctly identifies the GS structure of $ZnAs_2$; however, a few high-energy structures are also identified as low energy. This energy ranking can be considered satisfactory because in practical structure prediction implementations one would consider a few lowest energy structures as candidates for the GS structure. Similarly, the DFT GS structure of $CdSb_2$ is predicted to be only 0.009 eV/atom above the model-predicted GS, which will qualify the true GS structure as one of the lowest energy structures. The model-predicted GS structure has a DFT relative energy ($E-E_{\mathrm{min}}$) of 0.007 eV/atom. Finally, the energy ranking for $CdBi_2$ is inaccurate since the DFT ground-state structure is predicted to be 0.171 eV/atom above the model predicted GS structure. It is evident from Figure 4 that the relative energies of all the $CdBi_2$ structures lie in a limited window of ~0.25 eV/atom, unlike the *ABX*, *AX*, and ABX_4 type compositions. It is a more challenging to rank the structure in the correct order of their energies when all or a large fraction of the structures have similar energies i.e., the energy differences cannot be sufficiently resolved.

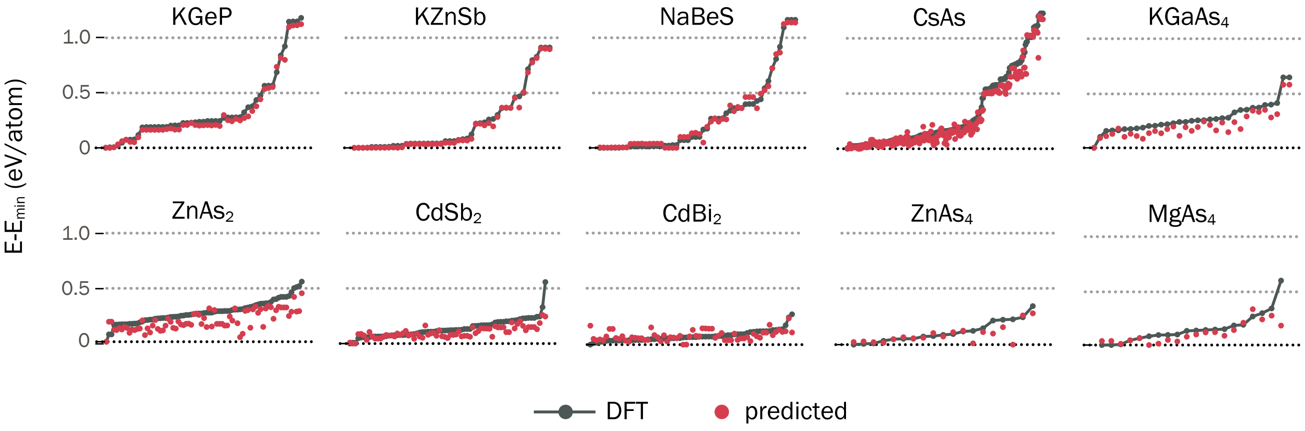

Fig. 4 Predicted relative energy ($E - E_{\min}$) of hypothetical structures of 10 different compositions from the test set in Figure 3(b) compared with DFT. For 9 out of 10 compositions, the predicted GS either matches or is within 0.025 eV/atom of the DFT ground-state structure.

For AX_4 type compositions, the energy rankings are similar to $ZnAs_2$ and $CdSb_2$, wherein the GS structures of $ZnAs_4$ and $MgAs_4$ are among the lowest energy structures predicted by the model, with their DFT relative energies 0.023 eV/atom and 0.036 eV/atom, respectively. At the same time, a few high-energy structures are also identified as low energy.

While the model satisfactorily ranks the energies of hypothetical structures, we also inspect the rankings of known structures to establish the robustness of the model. We chose the known polymorphs of MgO and ZnO from the ICSD database as representative examples. Figure S5 shows the comparison between DFT and hybrid model predicted energy rankings. Out of the 9 reported polymorphs of MgO, the model correctly labels the GS rocksalt structure and also does not misassign the higher-energy structures as low energy. Similarly, out of the 5 reported polymorphs of ZnO, the model correctly labels the GS wurtzite structure and accurately ranks the higher-energy structures. In summary, the model satisfactorily ranks the energy of the structures for most of composition types. For 9 out of 10 hypothetical compositions (Figure 4), the predicted GS structure either exactly matches or is within 0.025 eV/atom of the the DFT ground-state structure.

2.4 Analysis of Prediction Errors

We perform a thorough analysis of the large prediction errors in Figure 3(b). Such an analysis is useful in attributing the error to either prediction outlier or inconsistency in the training data. The hybrid model, presented in Figure 3(b), predicts the total energy of ~79% (1105 out of 1405) structures with <0.05 eV/atom error. However, 7 crystal structures (labelled in the figure) are either over- or under-predicted by 0.500 eV/atom, which are, interestingly, all ICSD structures. We analyze each of these structures on a case by case basis to understand the source of the error.

$Fe_{13}Ge_3$ (space group #221, ICSD id 150584) is severely underpredicted by 1.039 eV/atom relative to the DFT total energy. In this case, our analysis reveals that the DFT total energy is inaccurate. In magnetic compounds containing transition metals, the total energy is sensitive to the configuration of the magnetic moments.[34] $Fe_{13}Ge_3$ has a ferromagnetic ground state; however, the DFT total energy in NREL-MatDB is for the non-magnetic configuration. Upon recalculating the DFT total energy with ferromagnetic configuration, the prediction error is reduced to +0.08 eV/atom. This example highlights that DFT materials databases may contain occasional inconsistencies that can be flagged through machine learning regression.

The total energy of BaSiC (space group #107, ICSD id 168413) and CdC (space group #225, ICSD id 183177) are underpredicted by 0.651 eV/atom and 0.582 eV/atom, respectively. We find that both are hypothetical structures that were proposed in computational studies but not experimentally realized (ICSD contains a small fraction of hypothetical structures). These specific structures of BaSiC and CdC lie 0.795 eV/atom and 1.706 eV/atom above their respective convex hulls, which indicates that these high-energy structures are likely unstable. While the hybrid model is trained to predict the total energy of both GS and higher-energy structures, the training dataset of hypothetical structures span 24 elements (see Methods), including Ba, Cd, and Si but not C. The underprediction in the case of BaSiC and CdC is indicative of the remnant bias in the model towards lower-energy structures for compounds containing elements that are not in the hypothetical structure dataset.

The total energy for Ca_7Ge (space group #225, ICSD id 43321) is underpredicted by 0.545 eV/atom. Upon analyzing the crystal structure of this intermetallic compound, we find that the Ca-Ge bond lengths associated with the Ca(*4b*) Wyck-

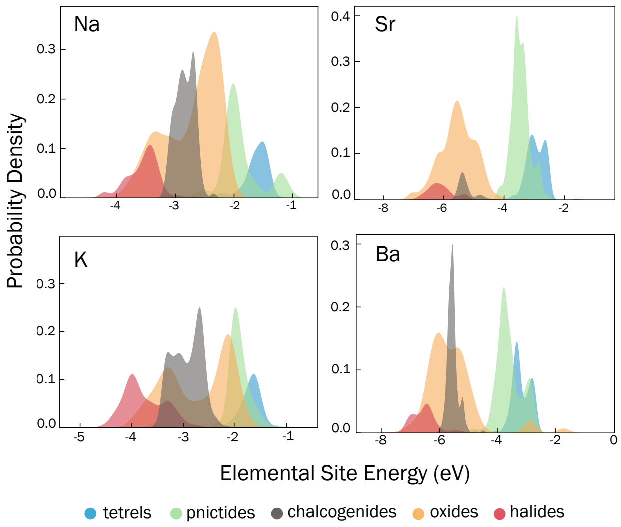

Fig. 5 Probability density of elemental site energies of Na, K, Sr, and Ba. The energy distribution provides chemical trends learned by the model. Generally, the site energies are more negative when the electropositive elements such as Na, K, Sr, and Ba are bonded to more electronegative anions (halides) than when bonded to less electronegative anions (pnictides, tetrels).

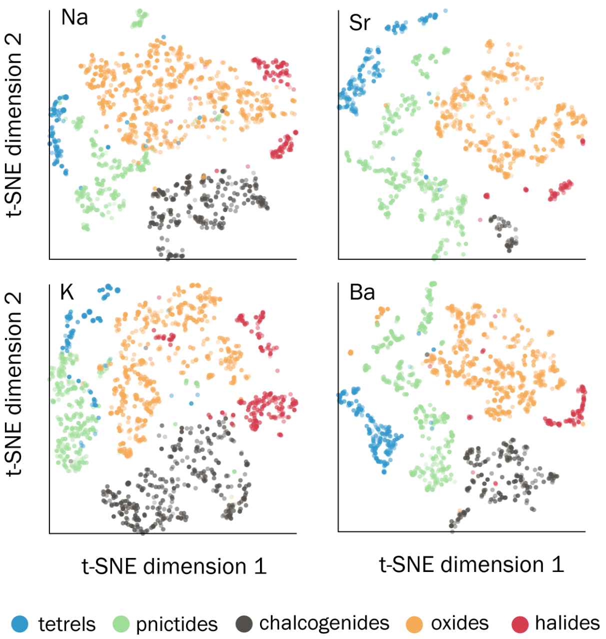

Fig. 6 t-SNE visualizations of the PCA-reduced elemental embeddings of Na, K, Sr, and Ba, shown as representative examples. The training set extracted embeddings are analyzed to draw chemical trends learned by the model. The embeddings lie in four major clusters, depending on the local environment (oxides, chalcogenides, halides, pnictides, tetrels) of the element of interest.

off site is 3.4 Å (Figure S6), which is significantly longer than typical Ca-Ge bond length (3 Å) in other Ca-Ge compounds, e.g. CaGe, Ca_2Ge, and Ca_5Ge_3. We perform a k-Nearest Neighbor (kNN) analysis on the penultimate site embeddings (see Methods) to identify other structures in the training set with embeddings that resemble Ca_7Ge. The purpose of the kNN is to find a number of training samples closest in distance to a point in the test set. Principal component analysis (PCA) is first used to reduce the embedding space to 10 dimensions, and the ten nearest neighbors for each site in Ca_7Ge is found from embeddings for sites in the training dataset. There are two unique Wyckoff sites of Ca (*4b*, *24d*) in Ca_7Ge; their 10 nearest neighbors are shown in Figure S6, which suggests that the *4b* site more resembles Sr and Ba (larger ionic radius than Ca), consistent with the long Ca-Ge bond lengths. This could also explain why Ca_7Ge is furthest from the convex hull (0.093 eV/atom) compared to other Ca-Ge structures.

Another outlier, Nb_3Si_2 (space group #127, ICSD id 645431), is overpredicted by 1.163 eV/atom. In training the model, we directly use ICSD structures rather than DFT-relaxed structures. While in most cases, the DFT-relaxed structures are not far from the ICSD structures, there are exceptions where this is not the case, such as for Nb_3Si_2. Using the DFT-relaxed structure instead of the ICSD structure reduces the error to 0.054 eV/atom.

Not all prediction errors are easily explainable as arising from the underlying DFT database. The source of error for $ScFe_6Sn_6$ (space group #191), which is overpredicted by 0.470 eV/atom, could not be identified and we believe that it is a case of prediction outlier. We thus have identified several causes of prediction errors, ranging from inconsistency in DFT data to simply model inaccuracy.

2.5 Chemical Trends

Interpretability of predictive neural network models remains intrinsically challenging. While direct physical interpretation of the CGCNN model in this work may not be possible, we compare trends in the model predictions with general chemical principles. Specifically, we identify trends in the learned elemental site energies (see Methods) through dimensionality reduction techniques such as PCA and t-distributed stochastic neighbor embedding (t-SNE). In conjunction, we also analyze the probability density of the elemental site energies.

We chose electropositive elements from group 1 (Na, K) and group 2 (Sr, Ba) as representative examples to identify trends in the learned elemental site energies. Figure 5 shows the probability density as a function of the elemental site energy for these elements. Figure 6 presents the corresponding two-dimensional t-SNE projections performed on the elemental embeddings. The site energy distributions in Figure 5 are calculated for all the sites in training set crystal structures for a given element. Only ICSD structures are considered in this analysis to avoid any unphysical effects arising from the hypothetical high-energy structures. For example, there are 1095 unique Na-containing structures, with 7056 unique Na Wyckoff sites. We find that when the element of interest is bonded to more electronegative anions – halogens (F, Cl, Br, I), oxygen (O), or chalcogens (Se, Se, Te), the resulting elemental site energies are more negative than when bonded only to less electronegative anions – tetrels (C, Si, Ge, Sn, Pb) or pnictogens (N, P, As, Sb, Bi). For example, out of 7056 sites for Na, 3458 sites bonded only to either halogens, oxygen, or chalcogens span an energy range of [-4.36, -1.97] eV whereas, the 1134 sites bonded only to tetrels or pnictogens span a lower energy range of [-3.17, -1.06] eV.

Notably the energy distribution for oxides span a wider energy range overlapping with other anion types, which can be attributed to the large variety of oxide compositions and structures and the different cation coordinations. Generally, Na, K, Sr, and Ba prefer octahedral coordination (6-fold coordination) when bonded to oxygen (e.g. rocksalt Na_2O, BaO) but there can be a departure from this typical behavior depending on the presence of other cations. For instance, Na sites in $Na_{17}Al_5O_{16}$ (space group #8) and $Na_{14}Al_4O_{13}$ (space group #14) are 3-fold, 4-fold, and 5-fold coordinated with some of the elemental site energies lying in the "tail" of the oxides (near the peak of pnictides) energy distribution (Figure S7). As such, some of the Na sites in these compounds behave as if they are bonded to pnictogens rather than oxygen. The presence of Al, which generally prefers tetrahedral coordination, causes this departure from the typical behavior.

The t-SNE projections in Figure 6 offer an additional dimension (compared to the 1-D site energy distribution in Figure 5) to visualize the learned elemental distributions. The t-SNE projections reveal distinct clusters depending on the anion type consistent with the observation of peaks in the probability density energy distributions (Figure 5). The separation into different clusters suggests that the chemical identity of the cation-anions bonds, at least for the 4 representative elements considered here, governs the learned elemental embedding. Consistent with the elemental site energy distribution, some Na sites in $Na_{17}Al_5O_{16}$ (space group #8) and $Na_{14}Al_4O_{13}$ (space group #14) lie in the cluster of pnictide embeddings (Figure S7).

2.6 Assessment of Thermodynamic Stability

Thermodynamic phase stability against decomposition into competing phases is a prerequisite for searching new materials and can be assessed through a convex hull construction.[23] Materials that lie on the convex hull are considered stable, i.e., the energy above the hull (ΔE_{hull}) is zero. Materials lying above the hull ($\Delta E_{\text{hull}} > 0$) are either unstable or metastable. The convex hull is defined as a convex envelope connecting the GS structures in a given chemical space and can be computed from DFT total energy by calculating formation enthalpy. For instance, in the binary Li-P chemical space, the convex hull connects elemental Li and P, and stable phases Li_3P, LiP, LiP_7, LiP_5, and Li_3P_7.

Compound	N	Space Group #	$\Delta E_{\text{hull,DFT}}$	$\Delta E_{\text{hull,pred}}$
LiF	1	225	0	0
PbTe	1	225	0	0
CdTe	1	186	0	0
Mg_3Sb_2	1	206	0	0
Li_2S	2	62	0	0
Li_3P	5	194	0	0
$Ca_2Ta_2FO_6$	6	227	0	0
Li_5OCl_3	8	140	0	0
$NaGaSb_4$	8	62	0.01	0.012
Na_3SbS_4	10	217	0.006	0.006
$La_9RbIr_4O_{24}$	13	12	0	0
Na_3PS_4	19	114	0	0
Na_3WFO_4	21	62	0	0
$MgSr_2Si_2O_7$	25	113	0	0
$Na_2Ba_3N_8C_4$	27	14	0	0
$K_2Mo_3Mn_2O_{12}$	40	1	0	0
$FeKNaSi_4O_{10}$	70	2	0	0
CdTe–ZnTe	3	186:216	stable	stable
CdTe–CdO	10	186:225	stable	stable
CdTe–$CuGaSe_2$	22	186:122	stable	stable

Table 1 Comparison of bulk thermodynamic phase stability calculated with DFT and with total energies predicted by the CGCNN model. N is the number of competing phases in the chemical space, not including the elemental phases. The predicted energy above the convex hull ($\Delta E_{\text{hull,pred}}$) for the unstable/metastable structures are consistent with DFT ($\Delta E_{\text{hull,DFT}}$), which are both expressed in eV/atom. Prediction of interface chemical stability is also compared for three representative examples (CdTe–ZnTe, CdTe–$CuGaSe_2$, and CdTe–CdO).

To demonstrate the accuracy of the hybrid model in predicting bulk thermodynamic phase stability, we perform convex hull analysis on a set of well-known materials (Table 1) by using the model predicted total energy of all the competing phases. Here, we consider all the competing phases documented in the ICSD. The number of competing phases (N) in each chemical space is shown in Table 1. We also perform the convex hull analysis to determine the accuracy of the model in predicting the interfacial chemical stability between two different materials; stable solid-solid interfaces are desired in various applications including solid-state batteries (stable electrode-electrolyte interface)[35,36] and solar cells (stable absorber-contact layer interface).[37]

In Table 1, we find that the predicted bulk stability (with the hybrid model) is consistent with the DFT stability. Moreover, ΔE_{hull} for Na_3SbS_4 and $NaGaSb_4$, the two compounds that lie above the hull, the predicted and DFT values are in excellent agreement. CdTe–ZnTe, CdTe–CdO, and CdTe–$CuGaSe_2$ interfaces are predicted to be stable, again consistent with DFT. Given the good agreement in bulk and interface stability predictions, we believe that the hybrid CGCNN model can be used to reliably and rapidly assess the phase stability of new phases.

3 Conclusions

In summary, we have developed a graph neural network model capable of reliably predicting DFT total energy of both ground-state and higher-energy structures. A hybrid model trained on both GS and higher-energy hypothetical structures achieves a lower error than models trained on either GS or hypothetical structures alone. The accuracy of the resulting model is sufficient to rank the small differences in energy typically encountered between structures with the same composition. The model can, therefore, serve the purpose of rapidly screening the energetics of different configurations for a given composition, a critical step in elucidating the structure and stability of new chemistries.

Some of the large errors in energy predictions are explained by identifying their source of error as inconsistencies in the underlying training data. In small-scale DFT studies, each calculation can be carefully examined by the researcher to ensure convergence. In high-throughput DFT databases, however, manual analysis must be replaced with automatic convergence criteria that can occasionally miss peculiar cases. Therefore, the training and analysis of ML models is one way that the consistency of high-throughput DFT databases can be rapidly verified. ML predictions fail where the data is poorly explained by neighboring trends, either because insufficient similar examples exist, inconsistencies in the data, or extreme sensitivity of the regressed variable with respect to structure. In addition to highlighting data inconsistencies and where additional data should be collected, prediction outliers can highlight interesting and unique chemical functionality that might otherwise go unnoticed in large databases.

There are a few limitations to the model, which remain to be addressed. The hypothetical structures used for training the hybrid model span only 24 elements, and their total energy are confined to a small range, in contrast to the wide range in the total energy of ICSD structures. To overcome this limitation, generation of additional DFT data for hypothetical structures will be done in a future work. Additionally, the current model was trained on hypothetical structures after DFT relaxations, which limits its usefulness in the forward screening of new hypothetical structures, where relaxed coordinates are not available. Generating accurate predictions with unrelaxed structures remains an unresolved problem in the field of structure prediction.

Acknowledgements

Funding for this work is provided by the U. S. Department of Energy (DOE) Advanced Research Projects Agency – Energy (ARPA-E) program. S. P., V. S., and P. G. acknowledge funding from ARPA-E under Award Number DE-AR0001205. P. S. J. acknowledges funding from ARPA-E under Award Number 19/CJ000/07/02. P.G. also acknowledges support from NSF through award DMR-2102409. JQ is funded by the NSF DIGI-MAT program, grant no. 1922758. This work was authored in part by the National Renewable Energy Laboratory (NREL), operated by the Alliance for Sustainable Energy, LLC, for the DOE under Contract No. DE-AC36-08GO28308. The views expressed in the article do not necessarily represent the views of the DOE or the U.S. Government. The research was performed using computational resources sponsored by the DOE's Office of Energy Efficiency and Renewable Energy (EERE) and located at NREL.

Author Contributions

P.G. and P.S.J. conceived and designed this research project; S.P. and P.S.J. built and trained the neural network models; J.Q and P.G. curated the training data; S.P., J.Q. and P.G. performed the first-principles DFT calculations; S.P., V.S., P.S.J. and P.G. analyzed the results; all authors participated in preparing and editing the manuscript.

4 Methods

4.1 Graph Neural Network Architecture

A crystal graph convolutional neural network (CGCNN) was constructed as depicted in Figure 7. Crystal structures are first converted to a graph using pymatgen,[38] using atomic sites as the graph nodes and distances between sites as the graph edges. Each node in the graph has exactly 12 edges, corresponding to the 12 nearest neighbor sites in the cystal while accounting for periodic boundaries. Node features include only the identity of the element at the atomic site, and edge features only included the raw distances (in Å) between the two sites. This is in contrast to other CGCNN models[28–30] that use several additional node and edge features e.g., group and period number, electronegativity etc. An embedding layer is used to convert the discrete element type of each atomic site into a 256 parameter vector, functioning similarly to a one-hot encoding of the atom type followed by a dense layer of dimension 256. Edge features are initialized from the raw distances through a radial basis function expansion, $r_i(d) = \exp\left[-\eta\,(d - c_i)\right]$ for $i \in [1, \ldots, 10]$, where d is the edge distance and η, c_i are learned parameters initialized to 7 and $[0, 0.7, 1.4, 2.1, \ldots, 6.3]$, respectively. In the CGCNN, the node and edge features are updated by passing them through a series of message layers, in which the nodes exchange information with their neighboring edges.

The structure of the message passing layers is adapted from Jørgensen et al.[39] First, for each edge, the source and target site features are concatenated with the edge's features, passed through a series of dense layers, and added to the input edge features in a residual fashion. Next, node features are updated using the features of the neighboring sites and those of the connecting edges. For each of the 12 edges pointing into a given site, the feature vector of the source site are multiplied by features of the corresponding edge before all 12 vectors are summed together. The resulting feature vector is then passed through a series of dense layers before being added to the original site feature vector in a residual fashion. Outputs from each message block are then fed as inputs into a subsequent message block for a total of 6 message layers. Final total energy predictions are produced by feeding the final site features into a 1-d output layer, producing a single energy prediction for each site. These predictions are added to a learnable mean energy for each element before being averaged over all sites in the crystal to produce a mean energy prediction.[40] Site-level contributions to the total predicted energy can therefore be extracted from this penultimate layer.

CGCNNs are trained for 500 epochs over the training data with a batch size of 64 crystals using the Adam optimizer with weight decay. The learning rate was decayed starting from an initial value of $1e^{-3}$, according to $1e^{-3}/(1+\text{epoch}/50)$, and the weight decay was similarly decayed according to $1e^{-5}/(1+\text{epoch}/50)$. The loss function minimized was the mean absolute error between predicted and DFT total energy.

4.2 Data and Preparation

Three distinct datasets of DFT-computed total energy are used in training the CGCNN models. First, we use DFT total energy of ∼14,000 ordered and stoichiometric crystal structures from the Inorganic Crystal Structure Database (ICSD)[8] that are available in the NREL Materials Database (NRELMatDB).[27] The DFT calculations are performed with VASP;[41] details of the calculations are available in Ref. 23.

During data cleanup, we identified that the DFT calculations for 1,677 structures containing fluorine were insufficiently convereged. We recalculated the DFT total energy of 874 (out of the 1677) structures with a recommended larger plane-wave energy cutoff of 540 eV. The remaining 803 structures contain transition elements that require an exhaustive search of the different magnetic configurations to determine the ground-state (GS) structure. Given the high computational cost associated with the search for the magnetic GS configurations, these 803 structures were not recalculated, and also, not included in the training data. With future applications of our CGCNN model in mind, we expanded the dataset to including DFT total energy of ∼3,900 ICSD structures containing mixed anions e.g., ZrOS, which are not currently in NRELMatDB. The DFT methodology (GGA-PBE[42]) and calculation parameters for the mixed anion compounds are consistent with those used in NRELMatDB. Combined, we use DFT total energy of ∼16,500 ICSD structures to train, validate, and test the CGCNN models. The ICSD collection IDs along with their total energy are made available through a public GitHub repository.[43] This dataset of ICSD structures span 60 elements and 12,760 unique compositions, with 2113 compositions existing in more than one one structure.

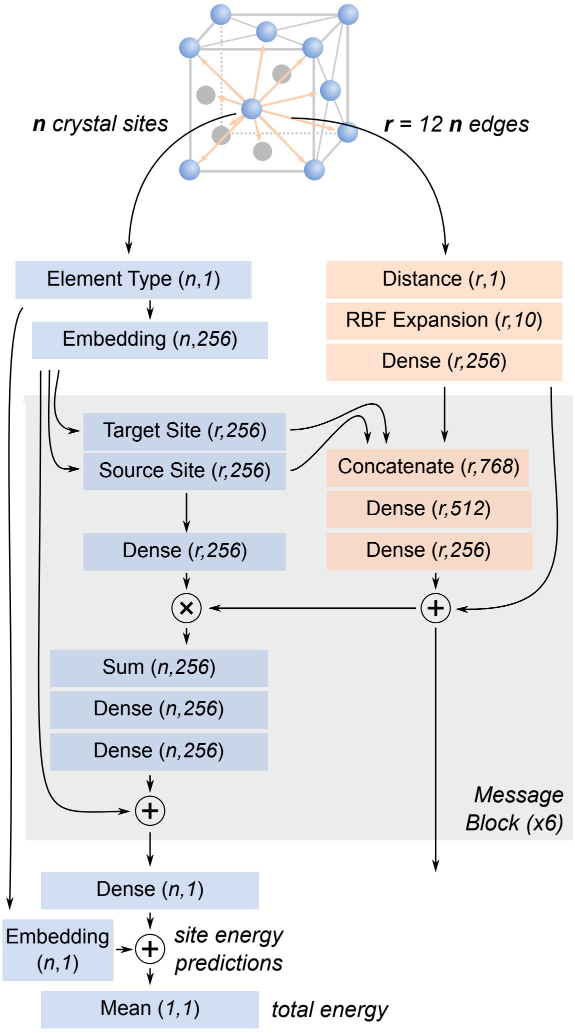

Fig. 7 Schematic of the neural network architecture. Node (atomic sites) and edge features (interatomic distances) output from each message block are fed as inputs into the subsequent block. The model predicts energy per site for all the sites in a given structure, which are averaged to get the total energy.

We also leverage a dataset of ∼11,000 hypothetical structures that were created by ionic substitutions in known prototype structures from the ICSD.[32,33] Upon ionic substitution, the decorated structures are relaxed and their total energy are calculated with DFT. The relaxed structures (as VASP POSCAR files) and the total energy are available through the GitHub repository.[43] The dataset is created for the purpose of discovering new Zintl phases.[32,33] As such, it spans 24 elements in 191 unique compositions of the type ABX (139, 6087), AX_4 (18, 318), AX (15, 3775), ABX_4 (13, 410), and AX_2 (6, 444), where the first number in paranthesis is the number of compositions and the second number is the number of structures. Here, element *A* includes Li, Na, K, Rb, Cs, Ba, Mg, Sr, Zn, Cd, element *B* are Si, Ge, Sn, Pb, Zn, Cd, Be, and *X* are group 15 elements (pnictogens) such as P, As, Sb, and Bi. KSnSb, $MgAs_4$, CdSb, $KGaSb_4$, and $ZnAs_2$ are representative compositions from this hypothetical structure dataset.

4.3 Analysis of Atomic Site Energy

The learned elemental site energies (Figure 7), which are the site-level contributions to the total energy, are analyzed to identify chemical trends. For specific elements, we calculate the probability density of the atomic site energies from all the ICSD structures in the dataset. We do not include the hypothetical high-energy structures in the analysis of the site energies to avoid biasing the chemical trends toward unstable structures. The distribution of pairwise distances between the learned elemental embeddings (Figure 7) will encode the relation between materials. We utilize common dimensionality reduction techniques such as principal component analysis[44] and t-distributed stochastic neighbor embedding (t-SNE),[45] as implemented in scikit-learn,[46] to analyze the multi-dimensional elemental embeddings.

References

1 J. P. Janet, F. Liu, A. Nandy, C. Duan, T. Yang, S. Lin and H. J. Kulik, *Inorg. Chem.*, 2019, **58**, 10592.

2 K. Alberi et al., *J. Phys. D: Appl. Phys.*, 2018, **52**, 013001.

3 *Computational Materials Discovery*, ed. A. R. Oganov, G. Saleh and A. G. Kvashnin, The Royal Society of Chemistry, 2019, pp. P001–455.

4 S. Curtarolo, G. L. Hart, M. B. Nardelli, N. Mingo, S. Sanvito and O. Levy, *Nat. Mater.*, 2013, **12**, 191.

5 G. Hautier, A. Jain and S. P. Ong, *J. Mater. Sci.*, 2012, **47**, 7317.

6 A. Jain, Y. Shin and K. A. Persson, *Nat. Rev. Mater.*, 2016, **1**, 1.

7 P. Gorai, V. Stevanovic and E. S. Toberer, *Nat. Rev. Mater.*, 2017, **2**, 1.

8 A. Belsky, M. Hellenbrandt, V. L. Karen and P. Luksch, *Acta Crystallographica Section B*, 2002, **58**, 364.

9 D. W. Davies, K. T. Butler, A. J. Jackson, A. Morris, J. M. Frost, J. M. Skelton and A. Walsh, *Chem*, 2016, **1**, 617.

10 K. Doll, J. C. Schön and M. Jansen, *Phys. Rev. B*, 2008, **78**, 144110.

11 S. M. Woodley, P. D. Battle, J. D. Gale and C. Richard A. Catlow, *Phys. Chem. Chem. Phys.*, 1999, **1**, 2535–2542.

12 C. J. Pickard and R. J. Needs, *J. Phys. Condens. Matter*, 2011, **23**, 053201.

13 V. Stevanovic, *Phys. Rev. Lett.*, 2016, **116**, 075503.

14 R. Gautier, X. Zhang, L. Hu, L. Yu, Y. Lin, T. O. Sunde, D. Chon, K. R. Poeppelmeier and A. Zunger, *Nat. Chem.*, 2015, **7**, 308.

15 X. Zhang, L. Yu, A. Zakutayev and A. Zunger, *Adv. Func. Mater.*, 2012, **22**, 1425.

16 G. Hautier, C. Fischer, V. Ehrlacher, A. Jain and G. Ceder, *Inorg. Chem.*, 2011, **50**, 656–663.

17 P. V. Balachandran, J. Young, T. Lookman and J. M. Rondinelli, *Nat. Comm.*, 2017, **8**, 1.

18 S. Chmiela, A. Tkatchenko, H. E. Sauceda, I. Poltavsky, K. T. Schütt and K.-R. Müller, *Sci. Adv.*, 2017, **3**, 1.

19 H. E. Sauceda, M. Gastegger, S. Chmiela, K.-R. MÃ¼ller and A. Tkatchenko, *J. Chem. Phys.*, 2020, **153**, 124109.

20 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, *APL Materials*, 2013, **1**, 011002.

21 S. Curtarolo, W. Setyawan, G. L. Hart, M. Jahnatek, R. V. Chepulskii, R. H. Taylor, S. Wang, J. Xue, K. Yang, O. Levy, M. J. Mehl, H. T. Stokes, D. O. Demchenko and D. Morgan, *Comp. Mater. Sci.*, 2012, **58**, 218.

22 S. Kirklin, J. E. Saal, B. Meredig, A. Thompson, J. W. Doak, A. Muratahan, S. Rühl and C. Wolverton, *npj Comput. Mater.*, 2015, **1**, 15010.

23 V. Stevanović, S. Lany, X. Zhang and A. Zunger, *Phys. Rev. B*, 2012, **85**, 115104.

24 C. Bartel, Q. Trewartha, A. Wang, A. Dunn, A. Jain and G. Ceder, *npj Comput. Mater.*, 2020, **6**, 97.

25 J. Schmidt, M. Marques, S. Botti and M. A. L. Marques, *npj Comput. Mater.*, 2019, **5**, 83.

26 A. M. Deml, R. O'Hayre, C. Wolverton and V. Stevanović, *Phys. Rev. B*, 2016, **93**, 085142.

27 *NREL Materials Database*, materials.nrel.gov.

28 T. Xie and J. C. Grossman, *Phys. Rev. Lett.*, 2018, **120**, 145301.

29 C. W. Park and C. Wolverton, *Phys. Rev. Materials*, 2020, **4**, 063801.

30 C. Chen, W. Ye, Y. Zuo, C. Zheng and S. P. Ong, *Chem. Mater.*, 2019, **31**, 3564.

31 C. Chen, Y. Zuo, W. Ye, X. Li and S. P. Ong, *Nat. Comp. Sci.*, 2021, **1**, 46.

32 J. Qu, V. Stevanovic, E. Ertekin and P. Gorai, *J. Mater. Chem. A*, 2020, **8**, 25306.

33 P. Gorai, A. M. Ganose, A. Faghaninia, A. Jain and V. Stevanovic, *Mater. Horiz.*, 2020, **7**, 1809.

34 P. Gorai, E. S. Toberer and V. Stevanovic, *Phys. Chem. Chem. Phys.*, 2016, **18**, 31777.

35 A. Banerjee, X. Wang, C. Fang, E. A. Wu and Y. S. Meng, *Chem. Rev.*, 2020, **120**, 6878.

36 W. D. Richards, L. J. Miara, Y. Wang, J. C. Kim and G. Ceder, *Chem. Mater.*, 2016, **28**, 266.

37 T. G. Allen, J. Bullock, X. Yang, A. Javey and S. De Wolf, *Nat. Energy*, 2019, **4**, 914.

38 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, *Computational Materials Science*, 2013, **68**, 314.

39 P. B. Jørgensen, K. W. Jacobsen and M. N. Schmidt, *arXiv:1806.03146*, 2018.

40 K. T. Schütt, H. E. Sauceda, P.-J. Kindermans, A. Tkatchenko and K.-R. Müller, *J. Chem. Phys.*, 2018, **148**, 241722.

41 G. Kresse and J. Furthmüller, *Phys. Rev. B*, 1996, **54**, 11169.

42 J. P. Perdew, K. Burke and M. Ernzerhof, *Phys. Rev. Lett.*, 1996, **77**, 3865.

43 github.com/prashungorai/papers/tree/main/2021/hybridgnn.

44 H. Hotelling, *J. Educ. Psychology*, 1993, **6**, 417–441.

45 L. van der Maaten and G. Hinton, *J. Machine Learning Research*, 2008, **9**, 2579.

46 F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot and E. Duchesnay, *J. Mach. Learn. Res.*, 2011, **12**, 2825.

— Supplemental Information —

A Graph Neural Network for Predicting Energy and Stability of Known and Hypothetical Crystal Structures

Shubham Pandey,[†] Jiaxing Qu,[‡] Vladan Stevanović,[†] Peter St. John,[*,¶] and Prashun Gorai[*,†]

[†]*Colorado School of Mines, Golden, CO 80401*

[‡]*University of Illinois at Urbana-Champaign, Urbana, IL 61801*

[¶]*National Renewable Energy Laboratory, Golden, CO 80401*

E-mail: peter.stjohn@nrel.gov; pgorai@mines.edu

1. Optimized Hyperparameters

The hyperparameters include parameters used to generate the crystal graphs, parameters of the neural networks, and parameters that control the training process. The hyperparameters are optimized through a train-validation process, on a fixed validation set. The following ranges of hyperparameters are searched: (1) batch size: 32–64, (2) embedding dimensions: 64–256, (3) number of message blocks: 4-8, and (4) learning rate: 1e^{-n}, n = 3-5. The mean absolute error of total energy prediction is reduced by 0.005 eV/atom by using a weight decay compared to when not using it.

Table S1: List of optimized hyperparameters in this work

Hyperparameter	Optimized value
Batch size	64
Embedding dimension	256
Number of message blocks	6
Learning rate	$1\text{e}1^{-3}$
Weight decay	1e^{-5}
Number of epochs	500

2. Performance of the Models Trained on Total Energy of ICSD Structures

To estimate the uncertainty in the mean absolute error (MAE) of total energy prediction, four different models are trained on the DFT total energy of ICSD structures. The uncertainty in the MAE is the standard deviation across the four models, each tested on a different hold-out test set.

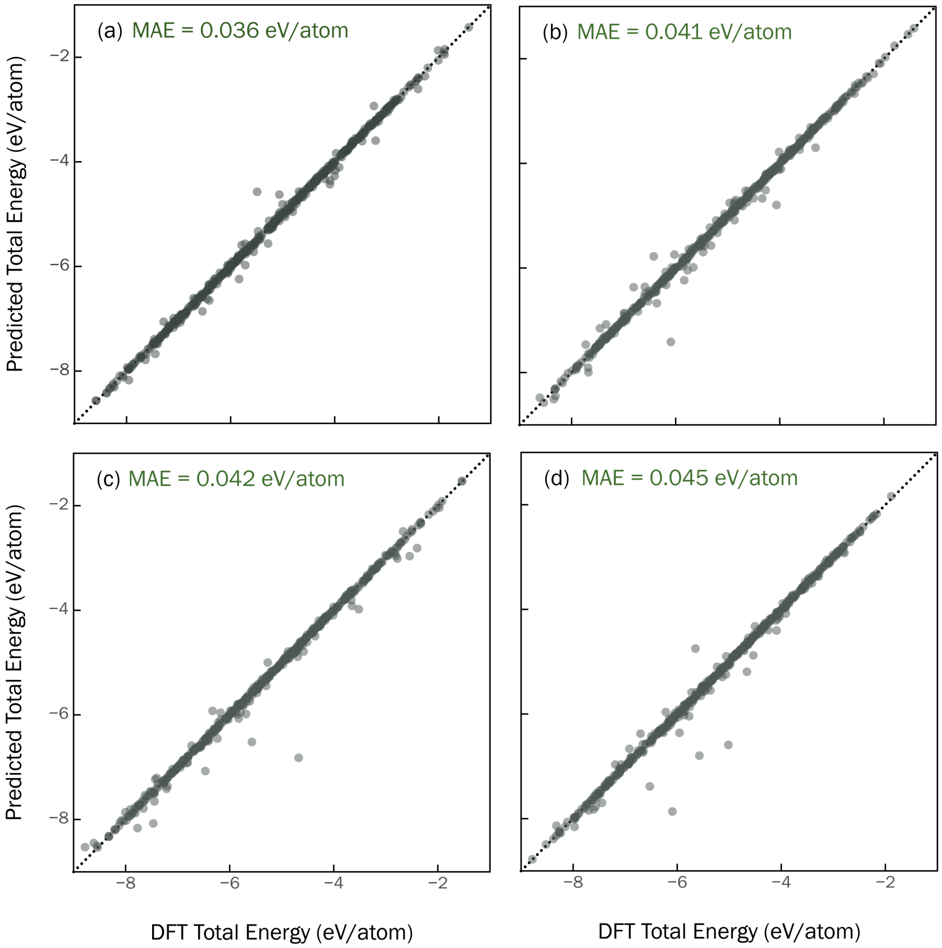

Figure S1: Convolutional neural networks trained on DFT total energy of ICSD structures from the NREL Materials Database. (a)-(d) Performance of the models trained and tested on four different sets of crystal structures. The mean absolute errors (MAEs) for the four different test sets are (a) 0.036 eV/atom, (b) 0.041 eV/atom, (c) 0.042 eV/atom, and (d) 0.045 eV/atom. The overall MAE across the four models is 0.041±0.005 eV/atom.

3. Performance of Model Trained on Total Energy of Hypothetical Structures

The model trained exclusively on the hypothetical structures is used to predict the total energy of the ICSD structures. Only a subset of ICSD structures, which contain the same 24 elements present in the hypothetical structure dataset, are chosen. The model poorly predicts the total energy of the ICSD structures with a large mean absolute error.

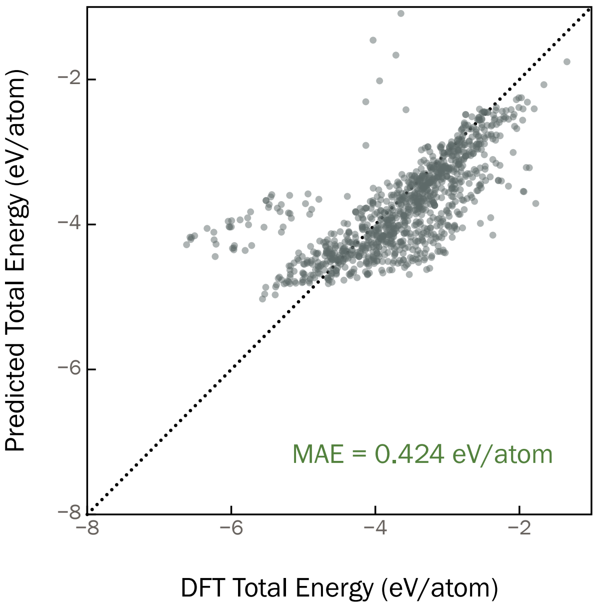

Figure S2: The total energy of 1065 ICSD structures is predicted with the model trained on the hypothetical structures alone (see Section 2.2 in the main text). The predicted total energy has large errors compared to the DFT values. The mean absolute error (MAE) of the test set prediction is 0.424 eV/atom, suggesting the model is biased towards hypothetical structures.

4. Performance of the Hybrid Model Trained on Total Energy of ICSD and Hypothetical Structures

To estimate the uncertainty in the mean absolute error (MAE) of total energy prediction, four different models are trained on the DFT total energy of ICSD and hypothetical structures. The uncertainty in the MAE is the standard deviation across the four models, each tested on a different hold-out test set. The training, validation and test sets are chosen with no overlap of compositions for the hypothetical structures.

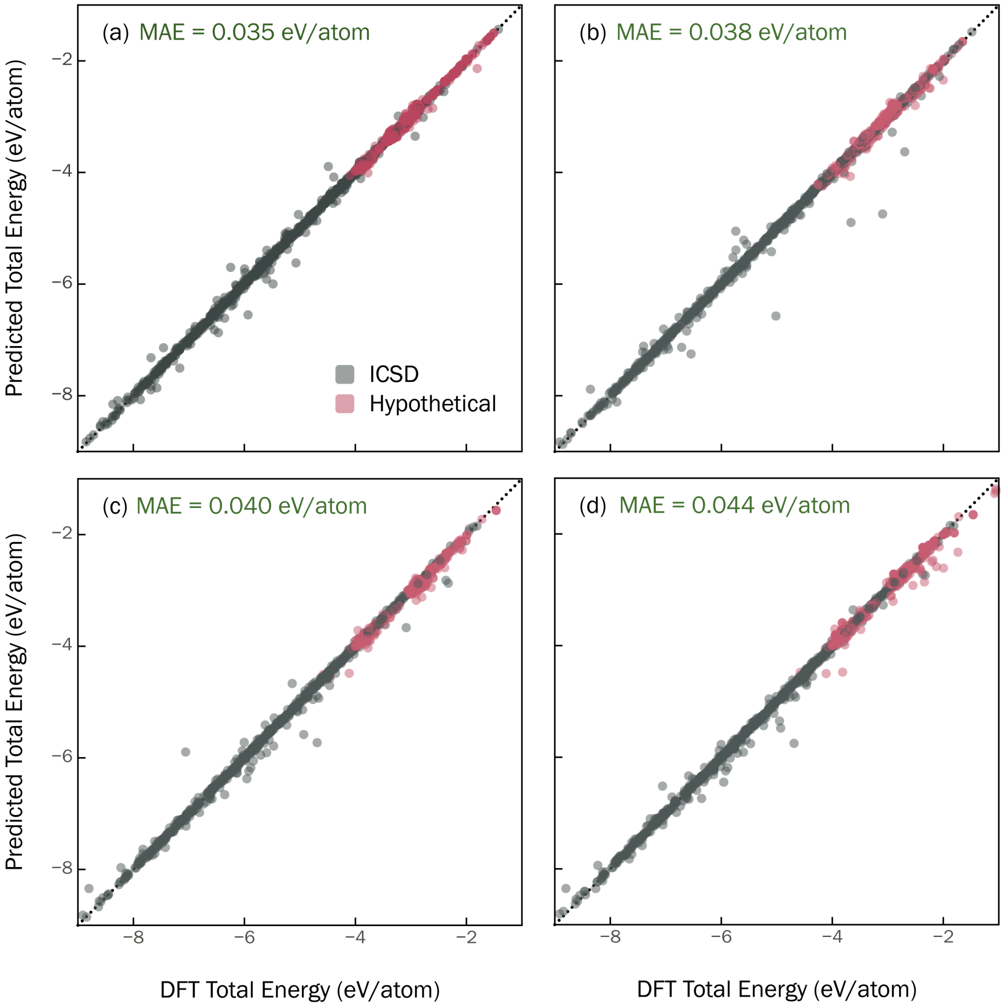

Figure S3: Convolutional neural networks trained on DFT total energy of ICSD and hypothetical structures. (a)-(d) Performance of the models trained and tested on four different sets of crystal structures. The mean absolute errors (MAEs) for the four different test sets are (a) 0.035 eV/atom, (b) 0.038 eV/atom, (c) 0.040 eV/atom, and (d) 0.044 eV/atom. Gray(red) datapoints correspond to ICSD(hypothetical) structures The overall MAE across the four models is 0.040±0.005 eV/atom.

5. Learning Curve of the Hybrid Model

A learning curve compares the performance of a model on a test set for varying number of training instances and therefore, can provide insights into whether a model is overfitted. The learning curve for the hybrid model shows that: (1) there is a systematic improvement in the model performance with the number of training crystal structures, and (2) the minimum number of training structures to achieve an MAE<0.05 eV/atom is $\sim 2 \times 10^4$.

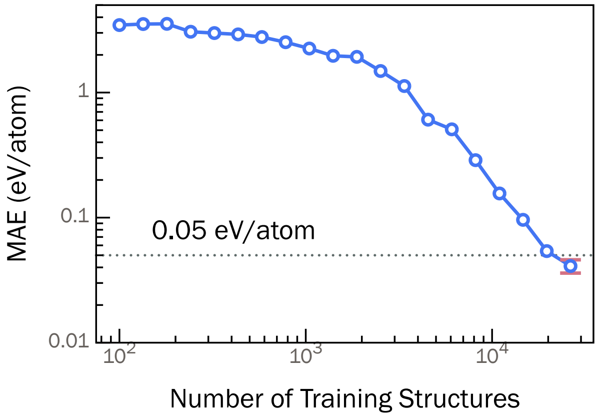

Figure S4: Learning curve for the hybrid model, showing that at least $2{\times}10^4$ crystal structures are required to achieve an MAE of <0.05 eV/atom.

6. Predicted Energy Rankings of MgO and ZnO Polymorphs

Figure 4 in the manuscript presents the energy rankings for different compositions in the hypothetical structures dataset. Here, we examine the energy rankings of two well-known binary compounds, MgO and ZnO, for which several experimentally realized and computationally proposed polymorphs are documented in the ICSD. There are 9 and 5 unique polymorphic structures reported for MgO and ZnO, respectively. The hybrid model correctly identifies the ground-state structures (rocksalt MgO, wurtzite ZnO) and also, satisfactorily ranks the other polymorphic structures.

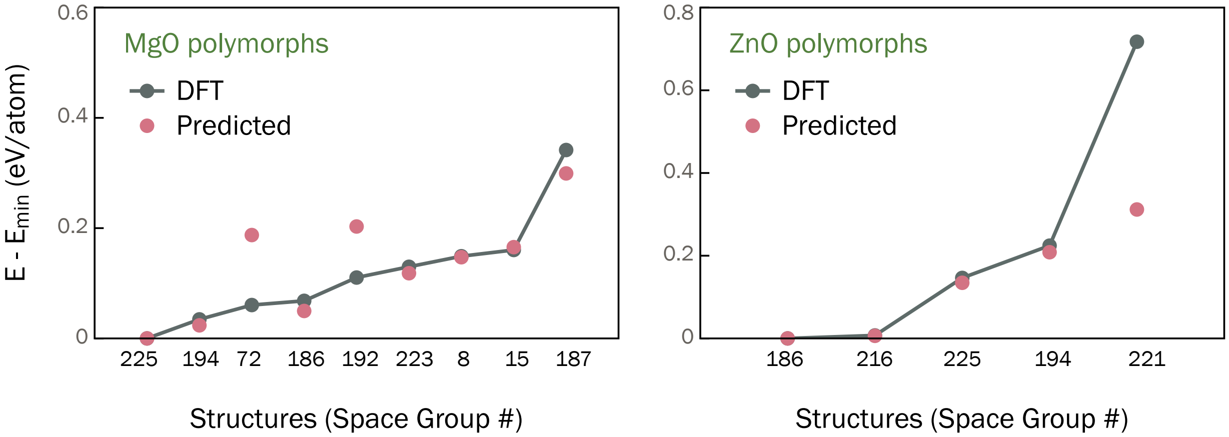

Figure S5: Predicted relative energy ($E - E_{\mathrm{min}}$) of MgO and ZnO polymorphs reported in the ICSD is compared with DFT values. The model correctly identifies the known ground-state structures of both MgO (rocksalt, space group #225) and ZnO (wurtzite, space group #186).

7. k-Nearest Neighbor Analysis of Ca_7Ge

The total energy of Ca_7Ge is severely underestimated (-0.545 eV/atom relative to the DFT value) by the hybrid model (Figure 3b). The intermetallic compound Ca_7Ge lies above the convex hull (see manuscript for details). To understand the source of the error, we perform a k-Nearest Neighbor (kNN) analysis of the elemental embeddings for all Ca and Ge sites in Ca_7Ge. From this analysis, we identify the first 10 NNs and their elemental identities. The Ca(*4b*) Wyckoff site has 9 NNs that are Ba atoms, while 1 NN is Sr. In contrast, the Ca(*24d*) Wyckoff site has 3 Ca NNs, 2 Sr, and 5 Ba. Moreover, the Ca-Ge bond lengths associated with the Ca(*4b*) site are larger compared to the Ca(*24d*) site. The Ge(*4a*) site has all 10 NNs that are Ge.

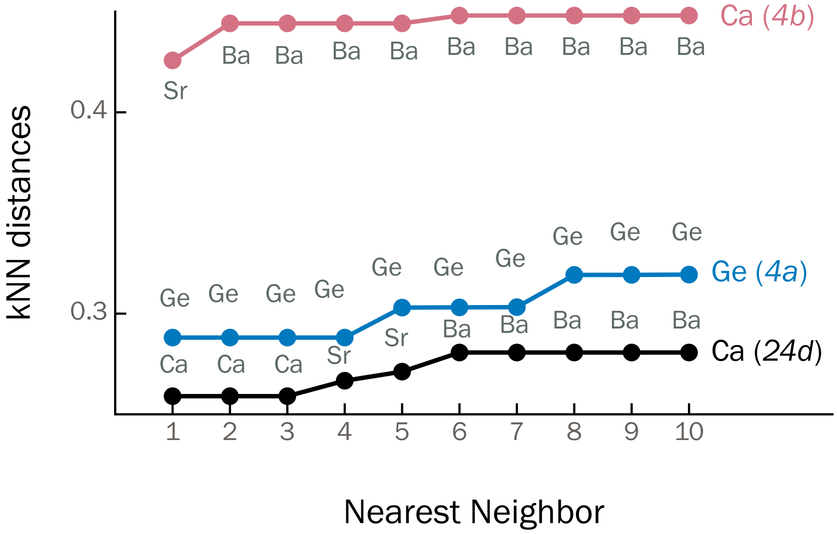

Figure S6: k-Nearest Neigbor (kNN) distances of the first 10 nearest neighbors for each Ca (*4b*, *24d* and Ge (*4a*) Wyckoff sites in Ca_7Ge. The elemental identities of the 10 nearest neighbors for each Wyckoff site are labelled.

8. Site Energies and t-SNE Projections: $Na_{17}Al_5O_{16}$, $Na_{14}Al_4O_{13}$

Chemical trends are identified by analyzing the probability density distribution of the elemental site energies (Figure 5) and t-SNE analyis of the elemental embeddings (Figure 6). In some cases, there can be a departure from the general trends. For example, some of the Na sites in $Na_{17}Al_5O_{16}$ (space group #8) and $Na_{14}Al_4O_{13}$ (space group #14) are 3-fold and 4-fold coordinated with elemental site energies in the "tail" of the oxides (near the peak of pnictides) energy distribution. The elemental embeddings for those same Na sites lie in the pnictogen cluster in the t-SNE projection. The other Na sites that lie closer to the peak of the oxides energy distribution (in the oxides cluster in t-SNE projection) are 5-fold coordinated.

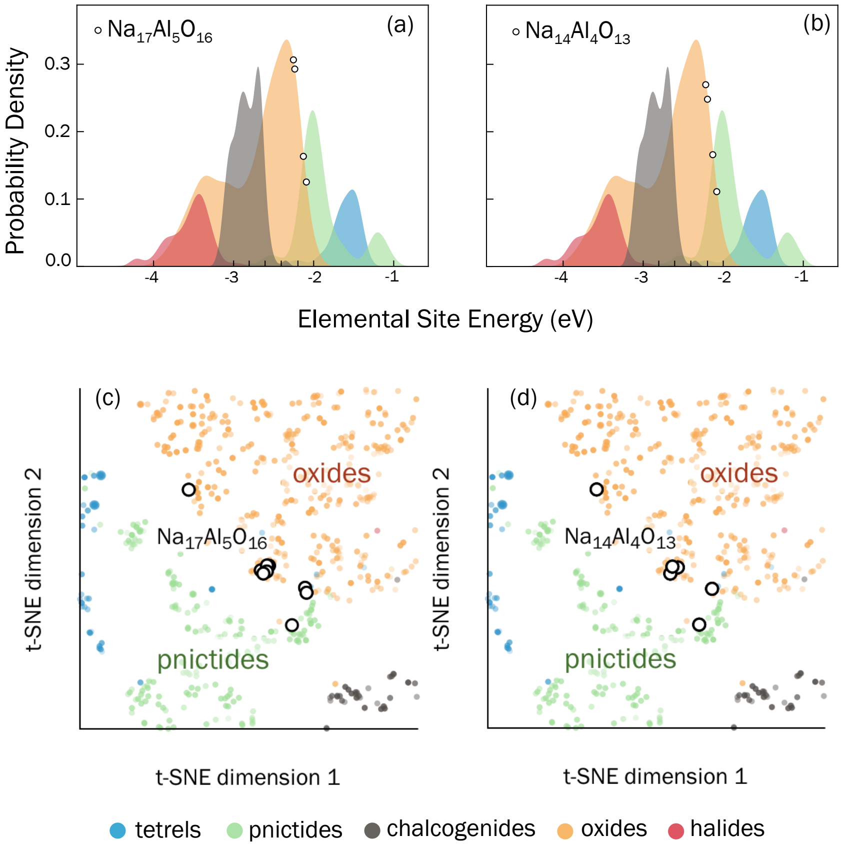

Figure S7: Elemental site energy and t-SNE projection of elemental embedding of Na sites in (a, c) $Na_{17}Al_5O_{16}$ and (b, d) $Na_{14}Al_4O_{13}$ are marked with open circles.