npj computational materials

www.nature.com/npjcompumats

ARTICLE OPEN

Check for updates

Data-augmentation for graph neural network learning of the relaxed energies of unrelaxed structures

Jason Gibson[1,2]✉, Ajinkya Hire[1,2]✉ and Richard G. Hennig[1,2]✉

Computational materials discovery has grown in utility over the past decade due to advances in computing power and crystal structure prediction algorithms (CSPA). However, the computational cost of the ab initio calculations required by CSPA limits its utility to small unit cells, reducing the compositional and structural space the algorithms can explore. Past studies have bypassed unneeded ab initio calculations by utilizing machine learning to predict the stability of a material. Specifically, graph neural networks trained on large datasets of relaxed structures display high fidelity in predicting formation energy. Unfortunately, the geometries of structures produced by CSPA deviate from the relaxed state, which leads to poor predictions, hindering the model's ability to filter unstable material. To remedy this behavior, we propose a simple, physically motivated, computationally efficient perturbation technique that augments training data, improving predictions on unrelaxed structures by 66%. Finally, we show how this error reduction can accelerate CSPA.

npj Computational Materials (2022)8:211 ; https://doi.org/10.1038/s41524-022-00891-8

INTRODUCTION

The process of discovery of functional materials, which drives innovation, has dramatically accelerated over the past decade, partially as a product of growing crystal structure databases[1–4] and improved computationally based CSPA[5], such as genetic algorithms (GA)[6], basin hoping[7], elemental substitution[8], and particle swarm techniques[9]. CSPA have played a prominent role in successfully predicting the structure and establishing the stability of high-pressure, high-temperature superconducting binary hydrides[10–12]. Many recent studies have started looking for stable ternary hydride superconductors. The addition of a third element to the binary hydrides can potentially stabilize these materials at much lower pressure[13–15]. Complex ternary and quaternary materials systems are also promising candidates for hydrogen storage applications[16].

In particular, GAs have proven their utility to identify thermodynamically stable phases efficiently; successfully identifying previously unknown materials for applications such as Li-Ge batteries[17] and solar cells[18]. Unfortunately, finding thermodynamically stable phases in ternary and quaternary systems is notoriously difficult due in part to the computationally expensive ab initio calculations required to relax and calculate the energies of GA produced structures, accounting for 99% of the algorithm's computational cost[19]. This places restrictions on the size and the composition of the unit cells, inhibiting the exploration of complex material systems.

This computational cost can be reduced by bypassing many of the costly ab initio calculations via a machine-learned filter or by implementing a more computationally efficient machine-learned surrogate potential to pre-relax structures[20]. Wu et al.[21] fitted a classical potential to the structures evaluated by density functional theory (DFT). They used the potential to pre-relax the structures in the GA and only evaluated the best structures with DFT. Jennings et al.[22] used a machine learning (ML) model to predict a structure's fitness directly and then only used DFT to evaluate structures that improved the current population. These methods are still somewhat hindered because many DFT evaluated structures are required to train a ML model to an adequate fidelity. Further, the models are specific to the materials' space the GA is searching, restricting their application to the given GA search.

Alternatively, there has been work to create universal ML models that determine a material's stability by predicting the formation energy of structures containing elements across the periodic table. Most notably, Xie et al.[23] predicted formation energy using a crystal graph convolutional neural network (CGCNN) trained on the materials project (MP) database[1]. The CGCNN represents a crystal structure as a multi-graph and builds a graph convolutional neural network on top of the multi-graph. This enables the model to learn the best features to represent the structure as opposed to the typical handcrafted feature approach[24] and achieve a formation energy validation MAE of 39 meV/atom[23]. More recently, the MAE of formation energy prediction of graph-based models continued to decrease to 21–39 meV/atom[25–29].

However, Park et al.[28] found that the model's dependence on a structure's atomic coordinates hinders the model's predictive fidelity on structures that strongly deviate from their relaxed states. Given that to obtain a structure in a relaxed state, a DFT relaxation and hence energy calculation is needed, the reported MAEs do not represent the model's ability to accurately identify unrelaxed structures that would relax to stable structures. On an unseen test set of 311 stable $ThCr_2Si_2$-type compounds, the CGCNN obtained a reasonable formation energy MAE on relaxed structures of 56 meV/atom. However, the prediction MAE was 370 meV/atom for the same test set on unrelaxed structures. This high error led to a true-positive rate (TPR) of 0.48 when filtering the unrelaxed compounds in the dataset[28].

Noh et al.[29] directly addressed the high formation energy prediction MAE of unrelaxed structures by adding two forms of regularization to the CGCNN in their CGCNN-HD approach. Replacing the softplus activation function in the convolution function with a hyperbolic tangent and adding dropout layers[30]

[1]Department of Materials Science and Engineering, University of Florida, Gainesville, FL 32611, USA. [2]Quantum Theory Project, University of Florida, Gainesville, FL 32611, USA. ✉email: jasongibson@ufl.edu; ajinkya.hire@ufl.edu; rhennig@ufl.edu

between each fully connected layer, reduced the MAE of the formation energy for a test set of unrelaxed Mg-Mn-O compounds from 518 meV/atom to 296 meV/atom.

The significant errors for unrelaxed structures are due to the limited sampling of the complex multi-dimensional configuration space of the potential energy surface (PES), with the relaxed structures only describing the minima of the surface. Since unrelaxed structures are not located at these minima, predicting a structure's formation energy at an unrelaxed configuration is a qualitatively different task[31]. To sample configurations near the minima of the PES, Smith et al.[32] applied a data augmentation technique known as normal mode sampling to a large dataset of molecular structures, resulting in a high fidelity neural network potential. However, determining the normal modes requires millions of phonon calculations which, while attainable for molecular structures, is infeasible for crystal structures.

Recently, Honrao et al.[33] showed that GA data could be used to predict relaxed formation energies of unrelaxed structures to high accuracy. This high accuracy was achieved by augmenting the training dataset, setting the formation energy of every structure within a basin of attraction to the minima of the respective basin of attraction, essentially modeling the continuous PES as a step function.

We propose leveraging these findings to improve formation energy prediction of unrelaxed structures by augmenting our data using a simple, physically motivated perturbation technique. Figure 1 illustrates our augmentation approach, which perturbs the atomic coordinates of a relaxed structure to generate additional training points that describe the regions surrounding the minima of the PES. We then map these perturbed structures to the energy of the relaxed structure, which requires no additional ab initio calculations. We utilize the CGCNN[23] and CGCNN-HD[29] to analyze how the augmentation affects formation energy predictions. We train these models on the MP database[1] augmented by perturbed structures. The resulting CGCNN models have similar prediction errors as the original ones for relaxed structures. To show the improvement in formation energy prediction, we apply the models to a test set consisting of 623 unrelaxed Nb-Sr-H hydride structures produced from a GA structure search. We find that compared to training on only relaxed structures, training with the augmented dataset, consisting of relaxed and perturbed structures, reduced the formation energy prediction MAE from 251 meV/atom to 86 meV/atom for CGCNN and from 172 meV/atom to 82 meV/atom for CGCNN-HD, as compared to the models trained only on relaxed structure.

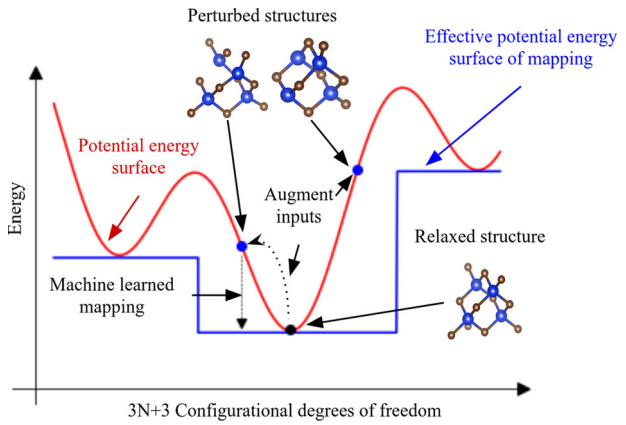

Fig. 1 Data augmentation for learning the potential energy surface (PES). The red line denotes a 2D representation of the continuous PES of materials. The blue line illustrates the effective PES, which describes the energy of a relaxed structure for a given unrelaxed input structure. Data augmentation aims to improve the machine learning of this effective PES by better sampling the configuration space. The black circle indicates the relaxed structures contained in the dataset, and the blue circles symbolize artificially generated structures for the data augmentation.

RESULTS

Model performance

Figure 2 shows the training, validation, and test (test-relaxed/test-unrelaxed) RMSE for each training epoch of the respective models. For interpretability, the trends are smoothed using an exponential moving average with a smoothing weight of 0.95. The Pearson correlation coefficients[34], with a value of 1 for perfect correlation and -1 for perfect anti-correlation, are computed between smoothed RMSE trends of Test-Relaxed/Test-Unrelaxed, Test-Relaxed/Validation, and Test-Unrelaxed/Validation.

The Pearson correlation coefficients show that the CGCNN and CGCNN-HD trained on only relaxed structures (Fig. 2(a, b)) result in anti-correlated trends between predictions on the relaxed and unrelaxed structures of the test set. These findings demonstrate that accurate predictions on relaxed structures do not lead to accurate predictions on unrelaxed structures and provide insight into the high prediction error for unrelaxed structure inputs reported in the literature. Furthermore, the CGCNN and CGCNN-HD trained on only relaxed structures display anti-correlated trends between the Test-Unrelaxed set and the validation set. This anti-correlation is detrimental to the model's predictive performance on unrelaxed structures because the validation error shows that predictions are improving and training should continue, while in actuality, unrelaxed structure predictions are getting worse. Additionally, the anti-correlation leads to an inability to correctly optimize a model's hyperparameters.

Figure 2(c) and (d) illustrate the effectiveness of training with the augmented dataset. Simply perturbing the atomic coordinates of each relaxed structure and then training on both the relaxed and perturbed structures dramatically improves the models' predictive ability on unrelaxed structures. Both the CGCNN and CGCNN-HD trained on the augmented dataset show a high Pearson correlation between the RMSE of Test-Unrelaxed/Validation. While beyond the scope of this manuscript, it is worth noting the test curves displayed in Fig. 2(c) shows the double decent behavior described in et al.[35].

Since the models trained on only relaxed structures, validation RMSE was inversely correlated to the RMSE of Test-Unrelaxed; the validation results provide no insight into the model's predictive abilities on unrelaxed structures. The model can extrapolate to unrelaxed structures only when augmented structures are included in the training dataset. Further, since the validation dataset is also augmented when the training dataset is augmented, perturbed structures that are representative of unrelaxed structures are present in the validation dataset leading to a correlation between the RMSE of Test-Unrelaxed/Validation. Hence, the validation error provides information on the models' accuracy on unrelaxed structure predictions, allowing correct hyperparameter optimization.

Figure 3 compares the models' formation energy predictions on Test-Unrelaxed to the DFT-computed formation energies. The CGCNN in Fig. 3(a) trained on only relaxed data tends to over predict E_f for the higher energy hydrides, which leads to a significant prediction MAE of 251 meV/atom. The added regularization applied to the CGCNN-HD in Fig. 3(b) improves the predictions on the higher energy hydrides. Still, the model tends to over predict E_f leading to an MAE of 172 meV/atom when training on relaxed.

Training with the augmented dataset substantially improves the prediction MAEs for the CGCNN and CGCNN-HD, reducing the prediction MAE to 86 meV/atom and 82 meV/atom, respectively.

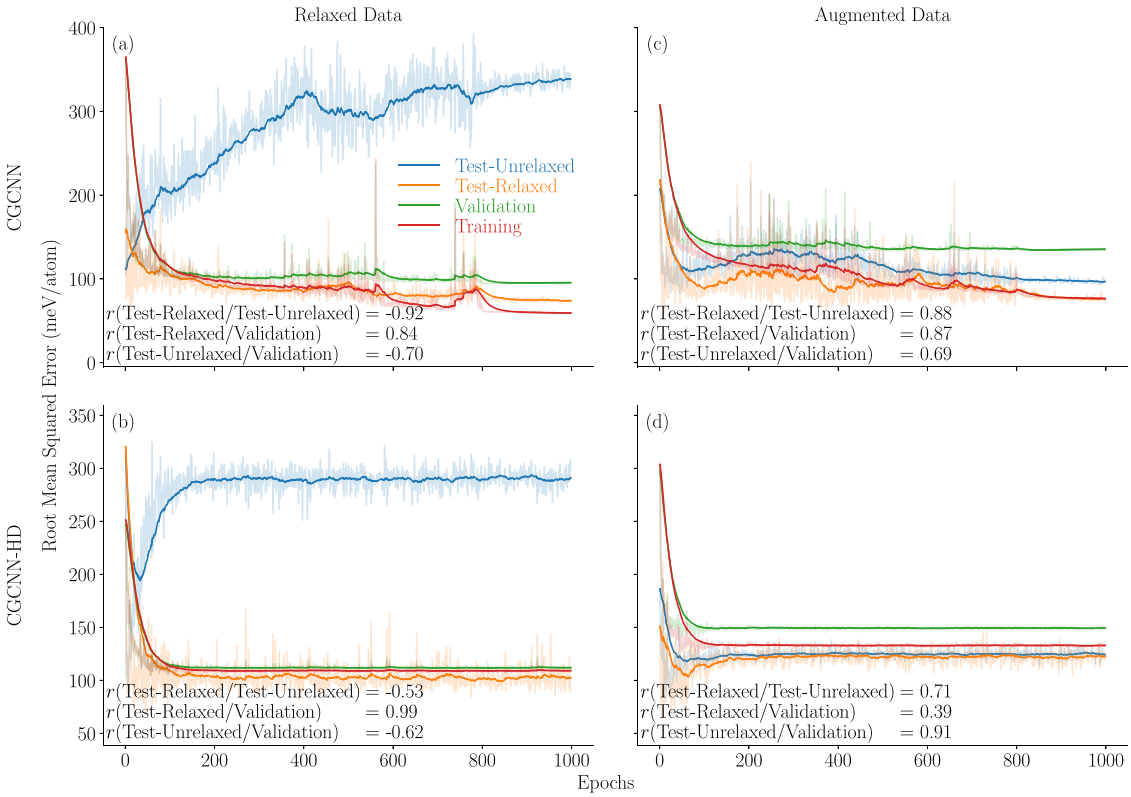

Fig. 2 Learning curves. a The CGCNN trained on relaxed data, **b** the CGCNN-HD trained on relaxed data, **c** the CGCNN trained on augmented data, **d** the CGCNN-HD trained on augmented data on the relaxed (1st column) and augmented (2nd column) data. The faded curves show the exact loss values, while the solid curves show the smoothed values. The red and green curves denote the loss on the training and validation data, respectively. The orange and blue curves display the loss for the Test-Relaxed and Test-Unrelaxed test sets, respectively. Note that the Test-Relaxed and Test-Unrelaxed datasets were not used in the training or validation of the model. The *r* values are the Pearson correlation coefficients between the stated trends.

This reduction in error improves the models ability to correctly identify low energy structures.

Interestingly, while the CGCNN-HD trained on augmented data has the lowest testing MAE, the dense region of underpredicted formation energies seen in Fig. 3(d) leads to a substantial number of misidentified unstable structures, hindering the model's ability to filter unstable structures. This will be discussed further in the proceeding section.

Owing to the large perturbations, some augmented structures may have moved to neighbor basins of attraction. Intuitively this would seem to yield substantial errors. However, as shown previously, the prediction of unrelaxed structures improved substantially. We suspect the error associated with perturbing a structure to a neighboring basin of attraction is mitigated due to the tendency of neighboring basins to cluster around similar minima on the PES[36]. Still, the large perturbations likely introduce some prediction error. A more sophisticated augmentation method could likely reduce the number of structures perturbed into neighboring basins and further improve predictions.

Interestingly, the models trained on the augmented data also display improved predictions on Test-Relaxed (Supplementary Fig. 1). However, this improvement seems specific to our test data as predictions on the relaxed validation data were better when the model was trained only on relaxed structures (Supplementary Fig. 2). The CGCNN-HD trained on only relaxed structures underpredicted the structures in Test-Relaxed likely because the training data contains relatively few transition metal hydrides and the bounded hyperbolic activation function, utilized in the CGCNN-HD's convolutional layers, has poor predictive power on unseen domains[37]. This poor predictive power on unseen domains is also the reason the CGCNN-HD models make poor predictions on the high and low formation energy structures of the MP data. Further, it is the reason the CGCNN-HD seems to make better predictions on both Test-Relaxed and Test-Unrelaxed for the model trained on the augmented dataset (Fig. 2(d)) when compared to the training and validation set. This is confirmed when using MAE instead of the RMSE as the learning curve's loss function (Supplementary Fig. 3).

Filtering unstable hydrides

To evaluate the models' ability to filter energetically unfavorable structures, we removed the MP database's correction applied to hydrogen-containing compounds and constructed a convex hull using the five known competing phases of the Nb-Sr-H system. Then, based on this constructed convex hull, we computed the hull distance $E_{\text{Hull}}^{\text{DFT}}$ of all the structures in the test set, utilizing their DFT-computed formation energy, and defined all structures with $E_{\text{Hull}}^{\text{DFT}} <= 0$ as stable. As a result, ten structures in the test set met the stability criteria.

To construct the receiver operating characteristic (ROC) curve, shown in Fig. 4 the predicted formation energies were used to compute hull distance ($E_{\text{Hull}}^{\text{ML}}$). To compute a range of true positives, false positives, true negatives, and false negatives, we varied the stability criteria ($E_{\text{Hull}}^{\text{ML}} < [-200, -199, \ldots, 699, 700]$ meV/atom) of $E_{\text{Hull}}^{\text{ML}}$ over a range of hull distances that ensure a completed ROC curve. We defined a true positive as a stable structure predicted as stable, a false positive as an unstable structure predicted as stable, a true negative as an unstable structure predicted as unstable, and a false negative as a stable structure predicted as unstable.

The ROC curve provides a graphical way to balance the models accuracy and computational cost. To demonstrate this we assume

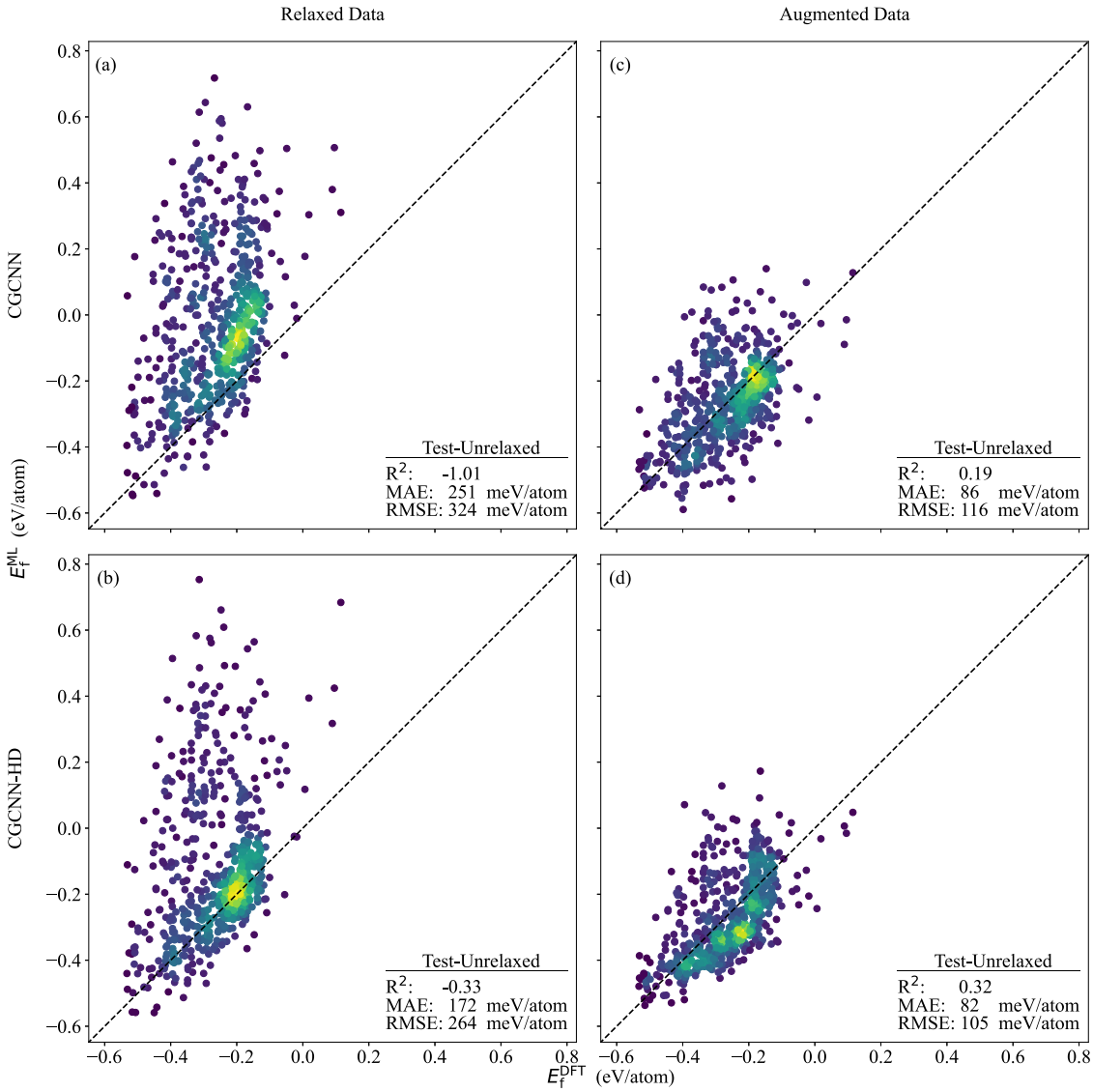

Fig. 3 Comparison of target and predicted energies of the test set of unrelaxed structures. a The CGCNN trained on relaxed data, **b** the CGCNN-HD trained on relaxed data, **c** the CGCNN trained on augmented data, **d** the CGCNN-HD trained on augmented data. The *x*-axis denotes the DFT-computed formation energies, while the *y*-axis denotes the predicted formation energies of the Test-Unrelaxed set. The values reported in the lower right are the coefficient of determination, R^2, the MAE, and the RMSE.

our test data is randomly generated and consider two hypothetical cases in which the models may be utilized. Case 1 emulates a study where identifying all stable structures is desirable (TPR = 1.0). Case 2 emulates a study where missing some stable structures is acceptable (TPR = 0.7).

For case 1, the CGCNN trained on the augmented data performed best, successfully identifying all stable structures at a filtration criterion of $E_{Hull}^{ML} < 39$ meV/atom while classifying 130 unstable structures as stable, yielding a 5-fold reduction in the number of energy calculations needed to identify all stable structures. For case 2, again, the CGCNN trained on the augmented data performed best, obtaining a TPR of 0.7 while only classifying 54 unstable structures as stable with a filtration criterion of $E_{Hull}^{ML} < 6$ meV/atom. The performance of all the models is summarized in Table 1, models trained on only relaxed structures required a higher E_{Hull}^{ML} and had substantially more FP compared to models trained on the augmented dataset.

The naive approach of setting the stability criteria to be the same for E_{Hull}^{ML} and E_{Hull}^{DFT}, restricts the ability to select a balance of accuracy and computational cost. For example, at stability criteria of $E_{Hull}^{ML} < 0$, the CGCNN trained with the augmented dataset obtained a TPR of 0.6, classifying 47 unstable structures as stable. While this performance is acceptable, as shown previously, at stability criteria of $E_{Hull}^{ML} < 39$ meV/atom the model can correctly identify all stable structures, which may be more desirable for a given application.

DISCUSSION

We proposed a simple, physically motivated, computationally efficient perturbation technique that augmented our data to represent the PES better, dramatically improving unrelaxed structure formation energy predictions. To augment our dataset, we add a perturbed structure for every relaxed structure and map it to the same energy as the relaxed structure. Thus, representing an additional point for a given basin of attraction of the energy landscape. Compared to training on only relaxed structures, training with an augmented dataset consisting of one relaxed and one perturbed structure for every relaxed structure, prediction MAEs of the CGCNN and CGCNN-HD were reduced from 251 meV/

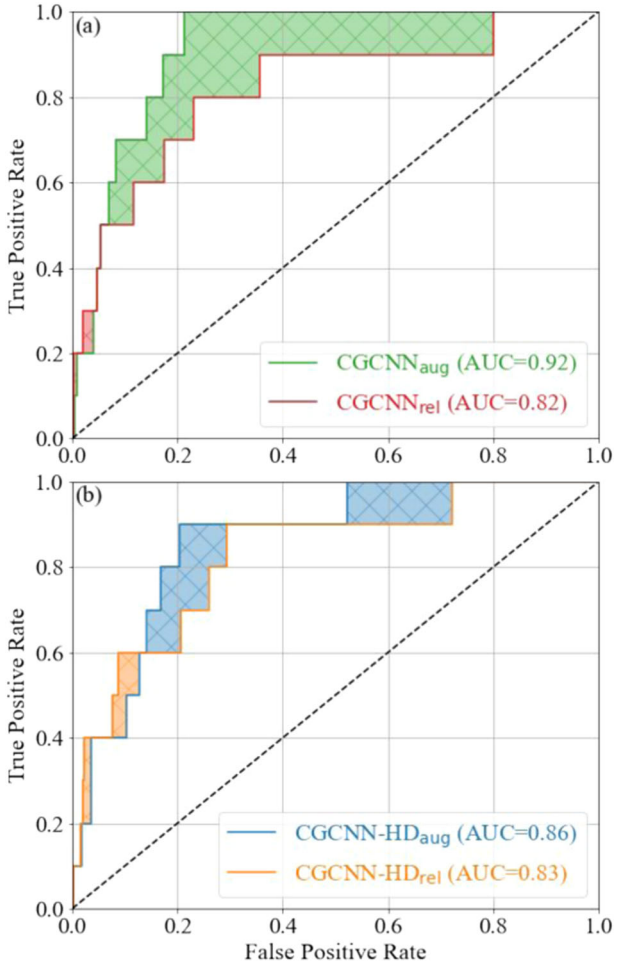

Fig. 4 Receiver operating characteristic curve. Figure displays the **a** CGCNN and **b** CGCNN-HD ability to classify stable structures when applied to unrelaxed structures. The *x*-axis is the fraction of unstable structures classified as stable. The *y*-axis is the fraction of stable structures classified as stable. The *aug* subscript represents the model was trained on the augmented data. The *rel* subscript represents the model was trained on the relaxed data. The dashed-black line represents a random classifier. The shaded regions signify where a model is performing better. The area under curve (AUC) is reported for all models.

atom and 172 meV/atom to 86 meV/atom and 82 meV/atom, respectively. Further, we showed that formation energy predictions for relaxed structures inputs were anti-correlated to predictions for unrelaxed structures inputs when training on only relaxed structures while training on the augmented data correlated relaxed and unrelaxed predictions RMSE. Finally, we utilize a ROC curve to show two cases where our method may be useful in accelerating CSPA. While more advanced augmentation techniques likely exist, this work showed the surprising effectiveness of a relatively simple method of augmentations that outperformed the current state of the art in formation energy prediction of unrelaxed structures.

METHODS

Data augmentation

For training, we use two datasets derived from the MP database[1] accessed on December 10, 2021. The 1st dataset, referred to as the relaxed dataset, consists of 126 k relaxed structures from the MP database, 20% of this data is held out for validation. The 2nd dataset, referred to as the augmented dataset, consists of the relaxed set and one perturbed structure for every relaxed structure.

Table 1. Models' filtering performance.

Rel.	Case 1		Case 2	
Aug.	TPR = 1.0		TPR = 0.7	
Metric	CGCNN	CGCNN HD	CGCNN	CGCNN HD
E_{Hull}^{ML} < (meV/atom)	576	328	159	91
	39	85	6	21
Number of FP	491	442	110	130
	130	320	54	89

Models' stability criteria and the number of false positives for case 1 and case 2. The upper right value correspond to the models trained on the relaxed data while the numbers reported in the lower left correspond to the models trained on the augmented data.

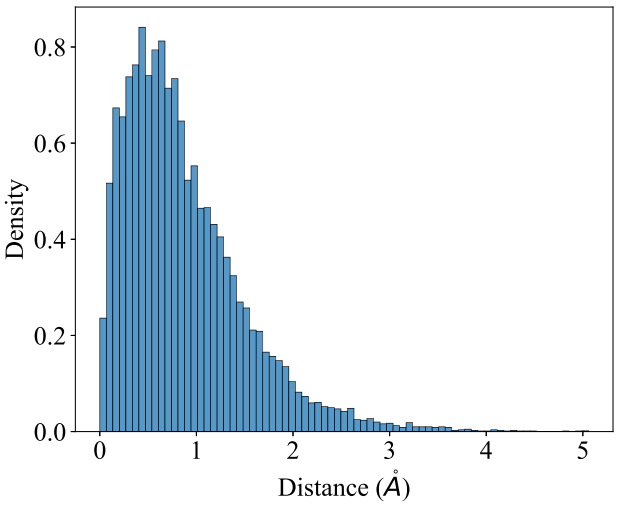

Fig. 5 Distribution sampled for perturbation. Distribution of the displacement of atoms during structural relaxation in three separate GA runs. The displacement is measured by the change in fractional coordinates multiplied by the lattice vector matrix of the relaxed structure.

We augment the data by perturbing the coordinates, $\mathbf{R}_i$ of all atoms, *i*, in each relaxed structures using a displacement vector

$$\Delta\mathbf{R}_i = (M_i^x, M_i^y, M_i^z) \quad (1)$$

where each Cartesian component is obtained by multiplying a direction vector, randomly sampled from a unit sphere with a scalar value randomly sampled from the displacement distribution. The displacement distribution was determined by analyzing the displacements of atoms during relaxation in three separate GA structure searches. The distance was determined by first taking the difference between the initial and final structure's fractional coordinates using the minimum image convention[38]. These differences were then multiplied by the lattice vector matrix of the relaxed structure to obtain the cartesian displacement vector and the euclidean norm for this vector. Figure 5 shows the resulting distribution of displacements, which was then fitted to a Gaussian mixture model (GMM) as implemented in the open python library scikit-learn[39]. The value of M_i was then selected by randomly sampling the GMM. Information about the change in lattice vectors and volume can be found in Supplementary Fig. 4.

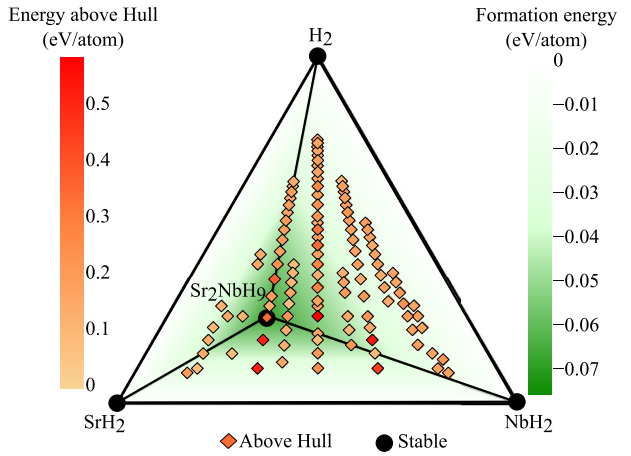

Fig. 6 The convex hull of the energies of the predicted structures for the NbH_2-SrH_2-H_2 materials system. The distance from the convex hull measures the thermodynamics stability of the various candidate compounds. The green shading indicates the formation energy of the thermodynamically stable compounds and mixtures relative to the three compounds NbH_2, SrH_2, and H_2. We identify a previously unknown ternary hydride, Sr_2NbH_9.

Training

The CGCNN and CGCNN-HD are trained on both the relaxed and augmented data. The model's architecture was determined by performing a grid search on the CGCNN trained on the augmented data. Supplementary Table 2 provides the range of parameters considered in the grid search. The architecture that minimized the validation error consists of 3-graph convolutional layers followed by 6-hidden layers with 64 neurons each. This architecture was then used for all models. Interestingly while past works[40] have found the CGCNN can scale up to 25 graph convolutional layers we found that models with more than eight hidden layers suffered from the vanishing gradient problem[41] when training on the relaxed data while training on the augmented data allowed for deeper models. we speculate the models' ability to scale with the number of graph convolutional layers and not the number of hidden layers is a product of the graph convolutional layers containing batch normalization. The remaining model hyperparameters are set to the values reported in ref. [23] for CGCNN and ref. [29] for the CGCNN-HD.

Test data

To provide test data for our models, we performed a GA search over the ternary system formed by H_2-Sr_6NbH_{16}-Nb_6SrH_{16}. We used the Genetic Algorithm for Structure and Phase Prediction (GASP) python package[42,43] for performing the GA search. Our aim with the search was to produce high hydrogens-containing structures that might show superconductivity. We remove the elemental hydrogen structures and partitioned this data into two test sets consisting of the relaxed and unrelaxed hydrides, referred to as Test-Relaxed and, Test-Unrelaxed respectively. Additionally. since the MP database contains few hydride structures, this data provides a challenging test case.

To relax and evaluate the energies of the candidate structures generated by GASP, we use VASP[44–47] with the projector augmented wave method[48] and the Perdew-Burke-Ernzerhof (PBE) generalized gradient approximation for the exchange-correlation functional[49]. We use a *k*-point density of 40 per inverse Å with the Methfessel-Paxton scheme and a smearing of 100 meV for the Brillouin zone integration, and a cutoff energy of 250 eV for the plane-wave basis set. The GA search was terminated after 771 DFT relaxations. We recomputed all energies using the VASP inputs generated by the MPRelaxset class of pymatgen to ensure consistency between the training and test sets and computed the formation energies[50]. Figure 6 shows the ternary convex hull of the Nb-Sr-H system produced using the GA-generated structure and the known competing phases from the MP database. Noteworthy, our structure search found previously unreported, thermodynamically stable ternary hydride, Sr_2NbH_9. Preliminary analysis of Sr_2NbH_9 suggest that the band gap closes at around 100 GPa.

DATA AVAILABILITY

Data will be made available upon reasonable requests.

CODE AVAILABILITY

Code for implementing the model on both cpus and gpus, training the models, augmenting training data are available at https://github.com/JasonGibsonUfl/Augmented_CGCNN.

Received: 27 April 2022; Accepted: 7 September 2022;
Published online: 30 September 2022

REFERENCES

1. Jain, A. et al. The materials project: a materials genome approach to accelerating materials innovation. *APL Mater.* **1**, 011002 (2013).
2. Kirklin, S. et al. The open quantum materials database (oqmd): assessing the accuracy of dft formation energies. *NPJ Comput. Mater.* **1**, 15010 (2015).
3. Draxl, C. & Scheffler, M. Nomad: The fair concept for big data-driven materials science. *MRS Bull.* **43**, 676–682 (2018).
4. Curtarolo, S. et al. Aflow: an automatic framework for high-throughput materials discovery. *Comput. Mater. Sci.* **58**, 218–226 (2012).
5. Oganov, A. R., Pickard, C. J., Zhu, Q. & Needs, R. J. Structure prediction drives materials discovery. *Nat. Rev. Mater.* **4**, 331–348 (2019).
6. Revard, B., Tipton, W. & Hennig, R. Genetic algorithm for structure and phase prediction. https://github.com/henniggroup/GASP-python (2018).
7. Wales, D. J. & Doye, J. P. K. Global optimization by basin-hopping and the lowest energy structures of lennard-jones clusters containing up to 110 atoms. *J. Phys. Chem. A* **101**, 5111–5116 (1997).
8. Noh, J. et al. Unveiling new stable manganese based photoanode materials via theoretical high-throughput screening and experiments. *Chem. Commun.* **55**, 13418–13421 (2019).
9. Chen, B. et al. Phase stability and superconductivity of lead hydrides at high pressure. *Phys. Rev. B* **103**, 035131 (2021).
10. Duan, D. et al. Pressure-induced metallization of dense (h2s)2h2 with high-tc superconductivity. *Sci. Rep.* **4**. https://doi.org/10.1038/srep06968 (2014).
11. Liu, H., Naumov, I. I., Hoffmann, R., Ashcroft, N. W. & Hemley, R. J. Potential high-tc superconducting lanthanum and yttrium hydrides at high pressure. *Proc. Natl Acad. Sci. USA* **114**, 6990–6995 (2017).
12. Peng, F. et al. Hydrogen clathrate structures in rare earth hydrides at high pressures: possible route to room-temperature superconductivity. *Phys. Rev. Lett.* **119** https://doi.org/10.1103/physrevlett.119.107001 (2017).
13. Sun, Y., Lv, J., Xie, Y., Liu, H. & Ma, Y. Route to a superconducting phase above room temperature in electron-doped hydride compounds under high pressure. *Phys. Rev. Lett.* **123**. https://doi.org/10.1103/physrevlett.123.097001 (2019).
14. Cataldo, S. D., Heil, C., von der Linden, W. & Boeri, L. LaBH8 : towards high- tc low-pressure superconductivity in ternary superhydrides. *Phys. Rev. B* **104** https://doi.org/10.1103/physrevb.104.l020511 (2021).
15. Hilleke, K. P. & Zurek, E. Tuning chemical precompression: Theoretical design and crystal chemistry of novel hydrides in the quest for warm and light superconductivity at ambient pressures. *J. Appl. Phys.* **131**, 070901 (2022).
16. Huang, Y., Cheng, Y. & Zhang, J. A review of high density solid hydrogen storage materials by pyrolysis for promising mobile applications. *Ind. Eng. Chem. Res.* **60**, 2737–2771 (2021).
17. Tipton, W. W., Matulis, C. A. & Hennig, R. G. Ab initio prediction of the li5ge2 zintl compound. *Comput. Mater. Sci.* **93**, 133–136 (2014).
18. Nguyen, M. C. et al. New layered structures of cuprous chalcogenides as thin film solar cell materials: cu_2Te and cu_2Se. *Phys. Rev. Lett.* **111**, 165502 (2013).
19. Heiles, S. & Johnston, R. L. Global optimization of clusters using electronic structure methods. *Int J. Quantum Chem.* **113**, 2091–2109 (2013).

20. Xie, S. R., Rupp, M. & Hennig, R. G. Ultra-fast interpretable machine-learning potentials. https://arxiv.org/abs/2110.00624 (2021).
21. Wu, S. Q. et al. Adaptive genetic algorithm for crystal structure prediction. *J. Phys. Condens. Matter* **26** http://arxiv.org/abs/1309.4742https://doi.org/10.1088/0953-8984/26/3/035402 (2013).
22. Jennings, P. C., Lysgaard, S., Hummelshøj, J. S., Vegge, T. & Bligaard, T. Genetic algorithms for computational materials discovery accelerated by machine learning. *NPJ Comput. Mater.* **5**, 1–6 (2019).
23. Xie, T. & Grossman, J. C. Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. *Phys. Rev. Lett*. **120**, 145301 (2018).
24. Ahmad, Z., Xie, T., Maheshwari, C., Grossman, J. C. & Viswanathan, V. Machine learning enabled computational screening of inorganic solid electrolytes for suppression of dendrite formation in lithium metal anodes. *ACS Cent. Sci.* **4**, 996–1006 (2018).
25. Choudhary, K. & DeCost, B. Atomistic line graph neural network for improved materials property predictions. *NPJ Comput. Mater*. **7**, 185 (2021).
26. Cheng, J., Zhang, C. & Dong, L. A geometric-information-enhanced crystal graph network for predicting properties of materials. *Commun. Mater*. **2**, 92 (2021).
27. Chen, C., Ye, W., Zuo, Y., Zheng, C. & Ong, S. P. Graph networks as a universal machine learning framework for molecules and crystals. *Chem. Mater.* **31**, 3564–3572 (2019).
28. Park, C. W. & Wolverton, C. Developing an improved crystal graph convolutional neural network framework for accelerated materials discovery. *Phys. Rev. Mater.* **4**, 063801 (2020).
29. Noh, J., Gu, G. H., Kim, S. & Jung, Y. Uncertainty-quantified hybrid machine learning/density functional theory high throughput screening method for crystals. *J. Chem. Inf. Model* **60**, 1996–2003 (2020).
30. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. & Salakhutdinov, R. Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn Res* **15**, 1929–1958 (2014).
31. Goodall, R. E. A., Parackal, A. S., Faber, F. A., Armiento, R. & Lee, A. A. Rapid discovery of stable materials by coordinate-free coarse graining. *Sci. Adv.* **8**, eabn4117 (2022).
32. Smith, J. S., Isayev, O. & Roitberg, A. E. Ani-1: an extensible neural network potential with dft accuracy at force field computational cost. *Chem. Sci.* **8**, 3192–3203 (2017).
33. Honrao, S. J., Xie, S. R. & Hennig, R. G. Augmenting machine learning of energy landscapes with local structural information. *J. Appl. Phys.* **128**, 085101 (2020).
34. Freedman, D., Pisani, R. & Purves, R. *Statistics*. 4th edn (WW Norton & Company, 2007).
35. Nakkiran, P. et al. Deep double descent: where bigger models and more data hurt. *J. Stat. Mech. Theory Exp.* **2021**, 124003 (2021).
36. Revard, B. C., Tipton, W. W. & Hennig, R. G. in *Prediction and Calculation of Crystal Structures* (eds Atahan-Evrenk, S. & Aspuru-Guzik, A.) 181–222 (Springer International Publishing, 2014).
37. Kim, Y. et al. Deep learning framework for material design space exploration using active transfer learning and data augmentation. *NPJ Comput. Mater.* **7**, 140 (2021).
38. Deiters, U. K. Efficient coding of the minimum image convention. *Z. Phys. Chem. (N. F.)* **227**, 345–352 (2013).
39. Pedregosa, F. et al. Scikit-learn: Machine learning in python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
40. Omee, S. S. et al. Scalable deeper graph neural networks for high-performance materials property prediction. *Patterns* **3**, 100491 (2022).
41. Goh, G. B., Hodas, N. O. & Vishnu, A. Deep learning for computational chemistry. *J. Comput Chem.* **38**, 1291–1307 (2017).
42. Tipton, W. W. & Hennig, R. G. A grand canonical genetic algorithm for the prediction of multi-component phase diagrams and testing of empirical potentials. *J. Phys. Condens. Matter* **25**, 495401 (2013).
43. Revard, B. C., Tipton, W. W., Yesypenko, A. & Hennig, R. G. Grand-canonical evolutionary algorithm for the prediction of two-dimensional materials. *Phys. Rev. B* **93**, 054117 (2016).
44. Kresse, G. & Hafner, J. Ab initio molecular dynamics for liquid metals. *Phys. Rev. B* **47**, 558–561 (1993).
45. Kresse, G. & Hafner, J. Ab initio molecular-dynamics simulation of the liquid-metal–amorphous-semiconductor transition in germanium. *Phys. Rev. B* **49**, 14251–14269 (1994).
46. Kresse, G. & Furthmüller, J. Efficiency of ab-initio total energy calculations for metals and semiconductors using a plane-wave basis set. *Comput. Mater. Sci.* **6**, 15 – 50 (1996).
47. Kresse, G. & Furthmüller, J. Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set. *Phys. Rev. B* **54**, 11169–11186 (1996).
48. Blöchl, P. E. Projector augmented-wave method. *Phys. Rev. B* **50**, 17953–17979 (1994).
49. Perdew, J. P., Burke, K. & Ernzerhof, M. Generalized gradient approximation made simple. *Phys. Rev. Lett.* **77**, 3865–3868 (1996).
50. Wang, A. et al. A framework for quantifying uncertainty in DFT energy corrections. *Sci. Rep.* **11**, 15496 (2021).

ACKNOWLEDGEMENTS

This work was supported by the National Science Foundation under grants Nos. PHY-1549132, the Center for Bright Beams, and the software fellowship awarded to J.B.G. by the Molecular Sciences Software Institute funded by the National Science Foundation (Grant No. ACI-1547580). Computational resources were provided by the University of Florida Research Computing Center.

AUTHOR CONTRIBUTIONS

J.B.G. and R.G.H. conceived the Augmentation technique and training strategy. J.B.G. implemented the Augmentation technique and performed the model training and analysis. ACH performed the genetic algorithm structure search to provide testing data. J.B.G., A.C.H., and R.G.H. contributed to the writing of the manuscript.

COMPETING INTERESTS

The authors declare no competing interests.

ADDITIONAL INFORMATION

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41524-022-00891-8.

Correspondence and requests for materials should be addressed to Jason Gibson, Ajinkya Hire or Richard G. Hennig.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.