npj | computational materials

www.nature.com/npjcompumats

ARTICLE OPEN

# Crystal twins: self-supervised learning for crystalline material property prediction

Rishikesh Magar[1], Yuyang Wang[1] and Amir Barati Farimani[1,2,3]✉

Machine learning (ML) models have been widely successful in the prediction of material properties. However, large labeled datasets required for training accurate ML models are elusive and computationally expensive to generate. Recent advances in Self-Supervised Learning (SSL) frameworks capable of training ML models on unlabeled data mitigate this problem and demonstrate superior performance in computer vision and natural language processing. Drawing inspiration from the developments in SSL, we introduce Crystal Twins (CT): a generic SSL method for crystalline materials property prediction that can leverage large unlabeled datasets. CT adapts a twin Graph Neural Network (GNN) and learns representations by forcing graph latent embeddings of augmented instances obtained from the same crystalline system to be similar. We implement Barlow Twins and SimSiam frameworks in CT. By sharing the pre-trained weights when fine-tuning the GNN for downstream tasks, we significantly improve the performance of GNN on 14 challenging material property prediction benchmarks.

*npj Computational Materials* (2022)8:231 ; https://doi.org/10.1038/s41524-022-00921-5

## INTRODUCTION

Machine Learning (ML) based predictive models have made rapid strides in computational chemistry due to their efficiency and performance. Characterized by their computational efficiency and accuracy, these methods are capable of faster high-throughput screening compared to classical physics models[1,2]. This capability has roots in both novel learning algorithms and improved hardware. Even though ML models can offer faster predictions, the accuracy of these models is highly correlated with the availability of clean labeled data[3]. In general, it is difficult to develop accurate and robust ML models without sufficiently large labeled data[4]. Moreover, the acquisition of labeled data is expensive as it involves performing Density Functional Theory (DFT) simulations or experiments to characterize materials[5,6]. On the other hand, gigantic databases containing structures and compositions of materials without labels (properties) are available. These databases cannot be used in supervised learning tasks due to the lack of labels. Given the availability of large unlabeled datasets, two interesting questions are raised: (1) can we develop more efficient ML models that are capable of learning the underlying structural chemistry from unlabeled data, and (2) can these models be used to make the supervised learning tasks more accurate?

In this work, we aim to address these questions by leveraging Self-Supervised Learning (SSL) for material property prediction. Unlike supervised learning which uses labels for supervision, SSL makes use of the large unlabeled data for supervision to learn robust and generalizable representations that can be used for various tasks. Recently, SSL frameworks such as SimCLR[7], Barlow Twins[8], BYOL[9], SwAV[10], MoCo[11], SimSiam[12], Albert[13], and self-supervised dialog learning[14] have been successfully applied to computer vision and natural language processing tasks. The success of these SSL methods has inspired many works in molecular ML, leading to the development of highly accurate frameworks such as MolCLR[15], dual view molecule pre-training[16], 3D Infomax[17], and numerous other popular works[18–25]. It should be noted that SSL-based methods have been developed for molecules, which have finite structures. However, the periodic crystalline materials are different from the molecules, since crystalline materials are composed of infinitely repeating unit cells of atoms, ions, or molecules. Besides, crystalline materials can have non-covalent bonds that are different from covalent bonds in molecules. Based on the differences, specialized deep learning architectures explicitly modeling crystals are required.

Most of the promising works developed for material property prediction tasks are using graph neural networks (GNN). GNNs consider non-Euclidean topology to construct a graph representation that can be learned and modified according to the task[26–28]. In general, the GNNs developed for material property prediction take input the 3D coordinates of the crystal and construct the graph by modeling atoms as the nodes and the interactions between the atoms as edges. GNNs developed for material property prediction include CGCNN[29], OGCNN[30], SchNet[31], MegNet[32], and other models[33–44]. Developments have also been made in tasks such as material structure generation and prediction[45–50] as well as identifying new materials with specific properties[51]. Despite progress being made in developing self-supervised ML architectures in the molecular ML, there is a noticeable lack of research works implementing such techniques for the periodic crystalline systems property prediction.

In this work, we introduce Crystal Twins (CT): an SSL framework for crystalline material property prediction with GNNs (Fig. 1). In pre-training, the models in the CT framework does not make use of any labeled data to learn crystalline representations, instead, it trains ML models in a self-supervised manner. In the CT framework, we use the CGCNN[29] as the encoder to learn expressive representations of crystalline system. We adapt two different SSL pretraining methods based on Barlow Twins[8] and SimSiamese[7] loss functions. In $\mathrm{CT}_{\mathrm{Barlow}}$ which uses Barlow twins loss for pre-training, the GNN encoder generates representations of two augmented instances from the same crystal and the objective of pre-training is to make the cross-correlation matrix of the two embeddings as close as possible to the identity matrix

[1]Department of Mechanical Engineering, Carnegie Mellon University, Pittsburgh, PA 15213, USA. [2]Department of Materials Science and Engineering, Carnegie Mellon University, Pittsburgh, PA 15213, USA. [3]Department of Chemical Engineering, Carnegie Mellon University, Pittsburgh, PA 15213, USA. ✉email: barati@cmu.edu

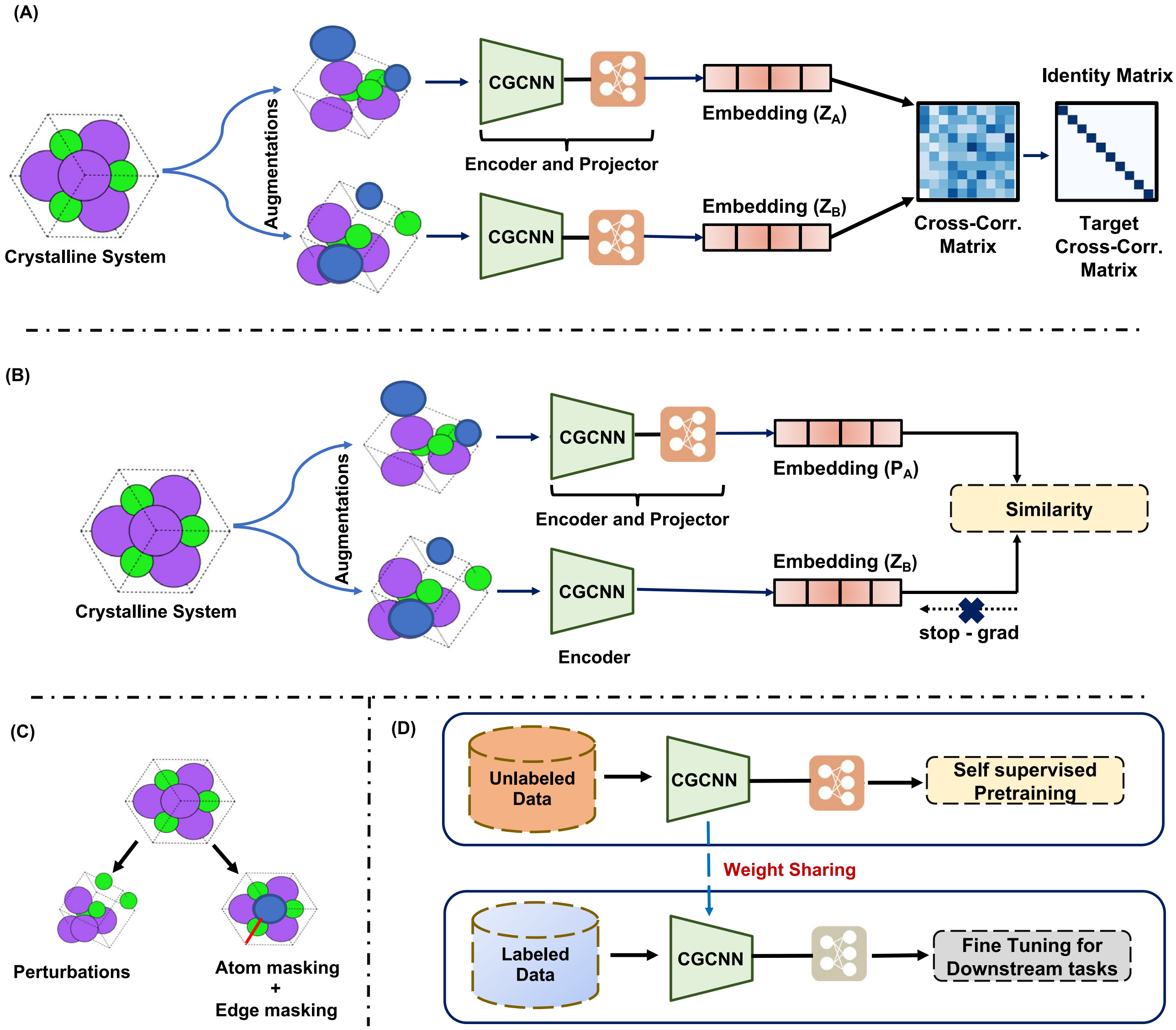

**Fig. 1 Overview of the crystal twins (CT) framework.** We propose two methodologies for SSL pre-training based on the Barlow Twins loss and SimSiamese loss function. The CT framework takes the structural file (CIF) as the input and then augments the structure to create two different augmented instances. (**A**) In $CT_{Barlow}$, each instance is passed to the CGCNN graph encoder followed by a projector to generate embedding. The pre-training objective aims to maximize the cross-correlation between the two embeddings. (**B**) The $CT_{SimSiam}$, each instance is passed through same CGCNN encoder branch to generate embeddings. One branch has an projector MLP head after the encoder and the other branch has stop-gradient operation. The pre-training objective is to maximize similarity between the embeddings. (**C**) To create augmented instances, three augmentation techniques are used in this work: random perturbations, atom masking, and edge masking. (**D**) In the pre-training stage we trained using SSL. In the fine-tuning stage, the pre-trained weights are shared with the encoder (CGCNN) which is trained to predict the material property.

(Fig. 1A). In the other model $CT_{SimSiam}$ that uses SimSiamese[7] loss function for pre-training, the objective is to maximize the cosine similarity between the embeddings generated from the graph encoder CGCNN for two augmented instances. Additionally, in $CT_{SimSiam}$, one branch has the stop gradient operation and the other has predictor head after the graph encoder (Fig. 1B). To create augmented instances, we introduce the combination of three different augmentation techniques: random perturbations, atom masking, and edge masking (Fig. 1C). The representations learned by the encoder are later used for downstream material property prediction tasks in the fine-tuning stage (Fig. 1D). In the pre-training stage, graph encoder learns representations from unlabeled data. Using the pre-trained weights to initialize the graph encoder for fine-tuning, both $CT_{Barlow}$ and $CT_{SimSiam}$ demonstrate superior prediction performances on 14 challenging datasets. We also compare the performance of the CT models with other competitive supervised learning baselines. We have successfully demonstrated the use of self-supervised learning for crystalline material property prediction.

## RESULTS

To comprehensively evaluate the performance of models using the CT framework, we test its performance on 13 challenging regression benchmark datasets and 1 classification dataset. The capabilities of the models in the CT framework are tested on a wide variety of properties including exfoliation energy, frequency of the highest frequency optical phonon mode peak, band gap, formation energy, refractive index, bulk modulus, shear modulus, Fermi energy, and metallicity. An overview of the datasets used for benchmarking the performance of the models in the CT framework is shown in Table 1. Among the total 14 datasets, we benchmark the performance of the models on 9 datasets (Table 2) from the MatBench suite and the remaining 5 datasets (Table 3) follow the datasets used in previously published works of CGCNN[29] and OGCNN[30]. More detailed descriptions of these datasets are available in the Supplementary Information.

**Table 1.** Overview of the datasets used for benchmarking the performance of the CT framework.

| Dataset | # Crystals | Property | Unit |
|---|---|---|---|
| JDFT2D(JDFT)[63] | 636 | Exfoliation Energy | meV per atom |
| Phonons[64] | 1,265 | Last Phdos Peak | 1 per cm |
| HOIP[61] | 1,345 | Band Gap | eV |
| Lanthanides[62] | 4,166 | Formation Energy | eV per atom |
| Dielectric[65] | 4,764 | Refractive Index | Unitless |
| GVRH[59,66] | 10,987 | Shear Modulus | $\log_{10} VRH$ |
| KVRH[66] | 10,987 | Bulk Modulus | $\log_{10} VRH$ |
| Perovskites[67] | 18,928 | Formation Energy | eV per atom |
| Fermi Energy[68] | 26,447 | Fermi energy | eV |
| Formation Energy (FE)[68] | 26,078 | Formation energy | eV per atom |
| Band Gap (BG)[68] | 26,709 | Band Gap | eV |
| MP-Is Metal (Is Metal)[68] | 106,113 | Metallicity | NA |
| MP-Gap (MP-BG)[68] | 106,113 | Band Gap | eV |
| MP-E-Form (MP-FE)[68] | 132,752 | Formation Energy | eV per atom |

Five of these datasets are in accordance to the previously published works of CGCNN[29] and OGCNN[30]. Additionally, we have also benchmarked on 9 datasets aggregated from the MatBench suite[42].

## Benchmarking the models on the MatBench Suite

The MatBench[42] suite consists of multiple material property prediction datasets. In this work, we consider 9 datasets that have crystal structures as input for benchmarking our self-supervised learning models $CT_{Barlow}$ and $CT_{SimSiam}$. We compare the results of our framework with the previously published supervised learning baselines available on MatBench. The protocols for benchmarking the performance of $CT_{Barlow}$ and $CT_{SimSiam}$ are exactly same as introduced in MatBench. We make use of nested 5 fold cross validation to generate the results in Table 2. The detailed hyperparameters used for finetuning models are listed in the Supplementary Information (Supplementary Table 3). We observe that the models trained using SSL based approach consistently outperform the supervised learning CGCNN baseline. Improved results for models in CT framework over the CGCNN baseline are observed for 7 out of the 9 datasets. For the Is Metal dataset, the performance of the models in CT framework are within the standard deviation of the supervised model. We also compare the performance of our SSL model with AMMExpress[42] model in the MatBench suite. We observe that models in the CT framework outperform AMMExpress on 6 out of the 9 datasets. Additionally, we also benchmark our model against the state-of-the-art model for material property prediction ALIGNN[43]. It was observed that our model performs better than ALIGNN only for the classification task. It must be noted that the ALIGNN achieves this high performance by modeling three-body interactions whereas CGCNN models two-body interactions. The enhancement of explicitly modeling three-body interactions gives ALIGNN more expressive power than CGCNN making it a more accurate baseline. Since we are using CGCNN as the graph encoder model in the CT framework, the $CT_{Barlow}$ and $CT_{SimSiam}$ are essentially modeling two-body interactions and are unable to compete with ALIGNN. The improvements demonstrated in our results over supervised

**Table 2.** Mean and standard deviation of test MAE of Crystal Twins (CT) in comparison to the supervised baselines on MatBench[42] regression benchmarks.

| Dataset | JDFT[63] | Phonons[64] | Dielectric[65] | GVRH[59,66] | KVRH[66] | Perovskites[67] | MP-BG[68] | MP-FE[68] | Is Metal[68] |
|---|---|---|---|---|---|---|---|---|---|
| # Crystals | 636 | 1265 | 4764 | 10,987 | 10,987 | 18,928 | 106,113 | 132,752 | 106,113 |
| CGCNN[29] | 49.24 ± 11.58 | 57.36 ± 12.31 | 0.599 ± 0.083 | 0.089 ± 0.001 | 0.071 ± 0.002 | 0.045 ± 0.001 | 0.297 ± 0.003 | 0.033 ± 0.001 | **0.952 ± 0.007** |
| AMMExpress[42] | **39.84 ± 09.88** | 56.17 ± 06.80 | **0.315 ± 0.067** | 0.087 ± 0.002 | 0.065 ± 0.002 | 0.200 ± 0.009 | 0.282 ± 0.006 | 0.172 ± 0.208 | 0.909 ± 0.001 |
| ALIGNN[43] | 43.42 ± 08.95 | **29.53 ± 02.11** | 0.345 ± 0.087 | **0.071 ± 0.001** | **0.057 ± 0.003** | **0.029 ± 0.001** | **0.186 ± 0.003** | **0.022 ± 0.001** | 0.913 ± 0.001 |
| $CT_{Barlow}$ | 46.79 ± 19.92 | 50.33 ± 08.88 | 0.434 ± 0.100 | 0.086 ± 0.004 | 0.067 ± 0.003 | 0.042 ± 0.001 | 0.264 ± 0.011 | 0.037 ± 0.001 | 0.945 ± 0.004 |
| $CT_{SimSiam}$ | 48.38 ± 18.68 | 48.86 ± 07.69 | 0.417 ± 0.079 | 0.087 ± 0.003 | 0.067 ± 0.003 | 0.042 ± 0.001 | 0.281 ± 0.025 | 0.037 ± 0.000 | 0.947 ± 0.003 |

The results in the table follow the protocols from MatBench. The best performing result has been shown in boldface and next best performing result has been underlined. The mean and the standard deviation are calculated over 5 folds following the MatBench protocol.

**Table 3.** Mean and standard deviation of test MAE of Crystal Twins (CT) in comparison to the supervised baselines on 5 regression benchmarks.

| Dataset | HOIP[61] | Lanthanides[62] | Fermi Energy[68] | FE[68] | BG[68] |
|---|---|---|---|---|---|
| # Crystals | 1333 | 4166 | 26,447 | 26,078 | 26,709 |
| GIN[20] | 0.666 ± 0.123 | 0.197 ± 0.038 | 0.605 ± 0.015 | 0.109 ± 0.007 | 0.601 ± 0.038 |
| CGCNN[29] | 0.170 ± 0.013 | 0.080 ± 0.003 | 0.400 ± 0.003 | 0.040 ± 0.001 | 0.369 ± 0.003 |
| OGCNN[30] | 0.164 ± 0.013 | 0.072 ± 0.002 | 0.446 ± 0.018 | 0.035 ± 0.001 | 0.353 ± 0.008 |
| $GIN_{Barlow}$ | 0.395 ± 0.007 | 0.094 ± 0.000 | 0.478 ± 0.125 | 0.085 ± 0.003 | 0.337 ± 0.004 |
| $CT_{Barlow}$ | 0.153 ± 0.003 | 0.058 ± 0.001 | 0.399 ± 0.004 | **0.025 ± 0.001** | 0.328 ± 0.002 |
| $CT_{SimSiam}$ | **0.140 ± 0.004** | **0.054 ± 0.001** | **0.384 ± 0.004** | **0.024 ± 0.001** | **0.302 ± 0.001** |

The benchmark datasets have been taken from previously published OGCNN[30]. The best performing result has been shown in boldface and the second best performing result has been underlined. The mean and the standard deviation have been calculated over 3 different runs.

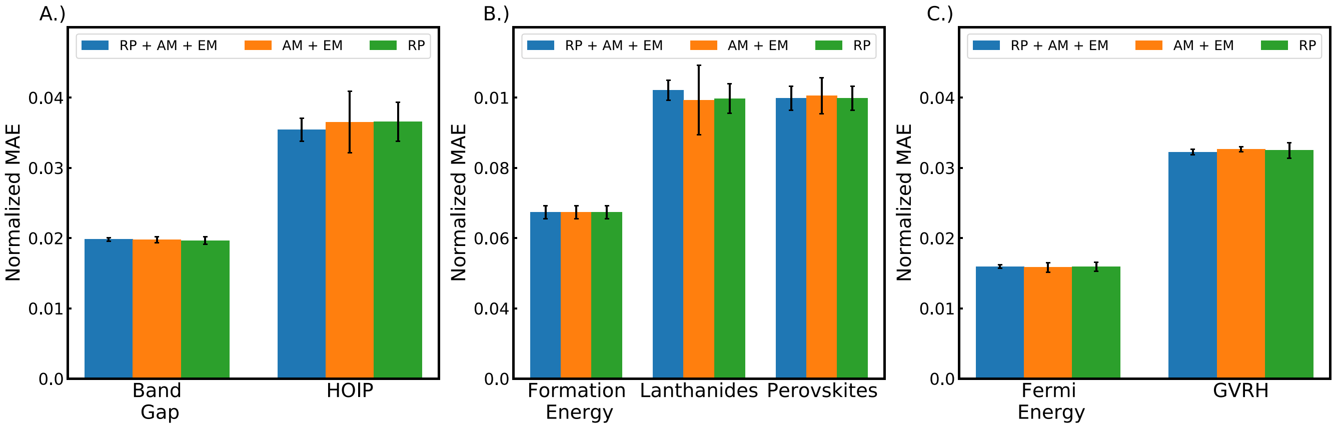

**Fig. 2 Ablation study of three augmentation techniques, random perturbation (RP), atom masking (AM), and edge masking (EM), for $CT_{Barlow}$ model.** (**A**) Evaluating the effect of different augmentation techniques in Band Gap and HOIP dataset where the label is band gap. (**B**) Evaluating the effect of augmentation techniques on the FE, Lanthanides, and Perovskites datasets for which the label is formation energy. (**C**) Evaluating the effect of different augmentation strategies on the Fermi energy and $\log_{10}$ VRH - shear modulus of the structures prediction. The error bars indicate variation in MAE over 3 different runs.

learning baselines CGCNN[29], AMMExpress[42] show the promise of using SSL for learning representation of crystalline materials.

## Benchmarking the models on additional datasets

Apart from benchmarking the performance on datasets from the MatBench suite. We also benchmarked the performance of our models on additional datasets similar to previously published works OGCNN[30] and CGCNN[29]. The datasets include properties like formation energy, band gap and Fermi energy. As we pre-trained the model with CGCNN encoder, the comparison with the CGCNN model is the most direct and fair, and it offers insights into how self-supervised learning methods can help in predicting the crystalline material properties with a high degree of accuracy. We also compare the performance of the models in the CT framework with other popular supervised GNN models, i.e., GIN[20] and OGCNN[30] for the datasets in Table 3. We would like to note that all the models used for comparison in Table 3 are trained with the same hyperparameters as suggested in their publicly available codes. The train/validation/test split for all the datasets is the same and set to 0.6/0.2/0.2 following previous standard benchmarking protocols. The data splitting is performed randomly following the protocols in the previously published works. The test Mean Absolute Errors (MAEs) for the supervised training baselines and the models in CT framework are shown in Table 3. The detailed hyperparameters used for supervised models are listed in the supplementary Table 4.

It is observed that the CT models outperform all supervised learning baselines on all the 5 regression tasks. We would like to note that the performance improvements (Supplementary Table 1) achieved by the CT models over the baseline CGCNN model are non-trivial. We observed an average improvement of 17.09% for $CT_{Barlow}$ and 21.83% for $CT_{SimSiam}$ when compared to CGCNN. The results in Table 3 clearly demonstrate the merit of using self-supervised learning frameworks for periodic crystal property prediction. In order to test the generic nature of our SSL framework, we also implement GIN[20] pre-trained via the Barlow Twins loss. We observed impressive gains in performance for the $GIN_{Barlow}$ over the supervised GIN model. The average improvement of the $GIN_{Barlow}$ model when compared to the supervised GIN model is 36.97%. The improvement in case of $GIN_{Barlow}$ indicates that CT framework can be applied with other graph encoder architectures and performance gains may be expected for those GNN models when compared to their supervised counterpart.

## Ablation study

To compare the effectiveness of the different augmentations techniques, we pre-train three $CT_{Barlow}$ models, (1) using only random perturbation augmentations (RP), (2) using only atom masking and edge masking augmentations (AM+EM), (3) using all three random perturbation, atom masking, and edge masking augmentations (RP+AM+EM). We report the MAE of the model on different fine-tuning datasets to determine the effectiveness of the augmentation techniques (Fig. 2).

The performance of AM+EM augmentation is better than RP for perovskites, BG and GVRH datasets, whereas RP augmentation has better performance than AM+EM for Fermi energy, lanthanides, and HOIP datasets. For FE dataset the performance of both RP and AM+EM augmentation techniques is the same. It must be noted that the performance of models trained with different augmentation techniques is almost identical, making it difficult to conclusively ascertain which augmentation technique is better. Moreover, we also observe that the effectiveness of the augmentation techniques is dataset dependent. We would also like to note that the standard deviation of MAE is always lower when using the pre-trained model with all augmentation techniques. Therefore, using a combination of all three augmentation techniques is most effective.

## Understanding the CT representations

To understand the CT representations, we visualize the representations from the pre-trained and fine-tuned $CT_{Barlow}$ framework in comparison to the CGCNN model in 2D using t-SNE[52]. The t-SNE representation maps the embedding based on the similarity in the 2D space. The comparison between the representations of the CGCNN model and the $CT_{Barlow}$ model for the perovskites dataset is shown in Fig. 3. Each point is colored by the formation energy of perovskites which is the label that the model is trained on in the fine-tuning stage.

We observe that the t-SNE projection from the $CT_{Barlow}$ model has a better clustering, namely, the crystalline materials with higher formation energy are clustered at the top left of the t-SNE projection plot (Fig. 3B) when compared to the CGCNN (Fig. 3A). Similarly the materials with lower formation energy are clustered at the bottom of the t-SNE plot (Fig. 3B) for the $CT_{Barlow}$ model. For example, perovskites $InOsO_3$ and $LaReO_3$ with relatively lower formation energies of −0.58 and −0.64 eV/atom, respectively, are clustered closely together in t-SNE projection from $CT_{Barlow}$ compared to CGCNN. This demonstrates the generalizability of the representations learned by the $CT_{Barlow}$ model when

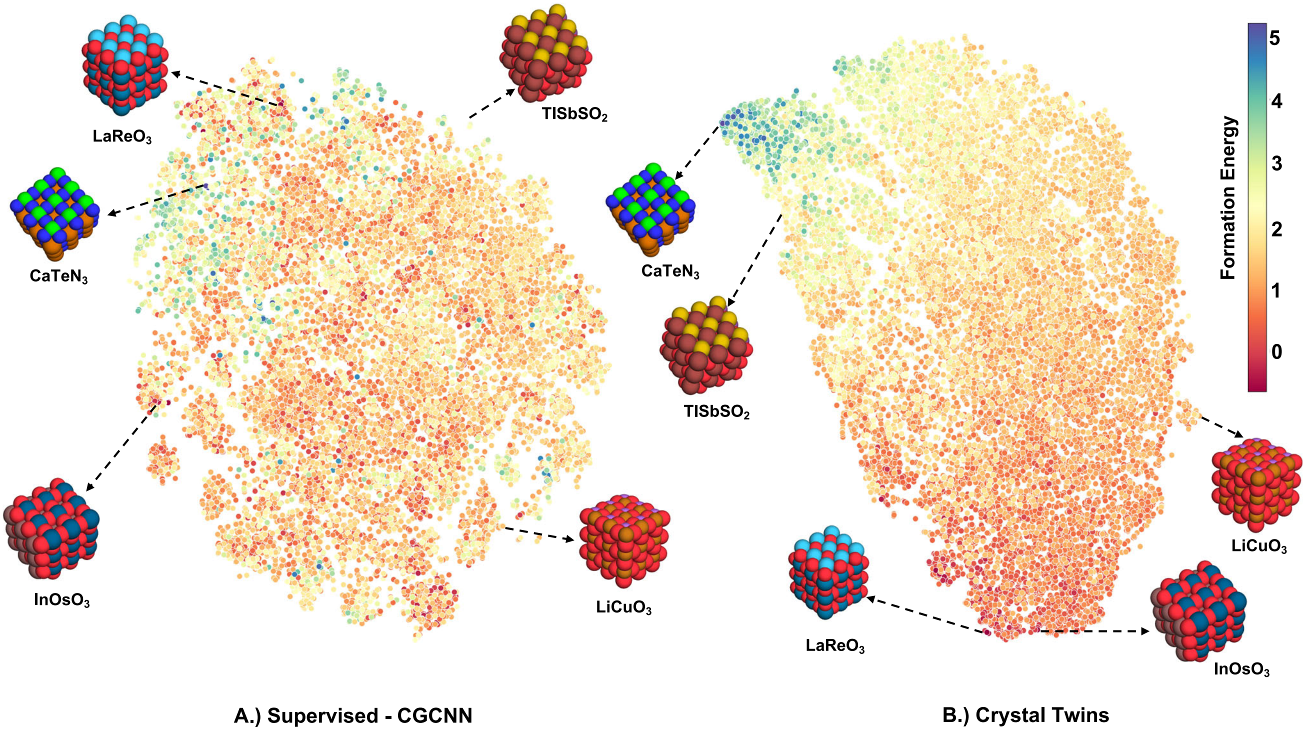

**Fig. 3 Visualizing the embeddings space for the perovskites dataset using t-SNE.** Every point on the t-SNE plot is colored corresponding to the formation energy of the crystalline system. (**A**) The t-SNE plot for the embedding was generated from the CGCNN model. (**B**) The t-SNE plot for the embedding was generated from the graph encoder of CT model after fine-tuning.

compared to supervised learning. Such representation learnt from the CT framework can also be used to characterize and understand the large chemical space of materials.

## DISCUSSION

In this work, we develop Crystal Twins (CT), a generic SSL framework for crystalline material property prediction. In this framework, we propose two SSL strategies using the twin graph neural networks to learn representations by leveraging the Barlow Twins loss and SimSiamese loss during pre-training. The models in CT framework ($CT_{Barlow}$ and $CT_{SimSiam}$) achieve superior performance compared to other competitive supervised learning baselines. The models in CT framework demonstrate high generalizability and robustness by learning representations that can be used to predict a variety of properties like formation energy, band gap, Fermi energy, shear modulus, bulk modulus, and refractive index of different crystalline materials. The pre-training of models in the CT framework has been performed on significantly less amount of data compared to SSL models in other domains like molecular machine learning, computer vision, and natural language processing. In general, SSL models are known to demonstrate better performance with larger unlabeled data as it allows them to learn more generalizable representations. We expect the models in the CT framework to demonstrate a superior performance with larger training data when compared to our current results. The representations learned by the models in the CT framework are of great promise and can open up avenues for exciting research in understanding the chemical space and designing materials with desired properties.

## METHODS

In this section, we describe the components of the CT framework (Fig. 1). In general, SSL frameworks employ correlations in the input itself to learn robust and generalizable representations from unlabeled data.[53]. As a part of CT framework we propose two different SSL pretraining models namely $CT_{Barlow}$ and $CT_{SimSiam}$. For the $CT_{Barlow}$, the goal during pre-training is to force the empirical cross-correlation matrix created from the encoder embeddings of two different augmentations generated by the same crystal towards the identity matrix. All the elements in the cross-correlation matrix lie between $-1$ and 1, with 1 representing maximum correlation. Intuitively, since the embeddings are generated from augmentations of the same crystalline system, the cross-correlation matrix must be close to the identity matrix. For the $CT_{SimSiam}$, the pre-training objective is to maximize the cosine similarities between encoder embeddings of augmented instances generated from the same crystalline system. To avoid model collapse, $CT_{SimSiam}$ implements an extra projection head on one side of the twin networks and applies the stop-gradient technique on the other side in training. Using such objectives during pre-training allows the graph encoder to learn robust representations. To create the augmented instances, we use augmentation techniques, including atom masking, edge masking, and random perturbation (refer Supporting Information). The embeddings for the augmented instances of the crystalline system are generated via the CGCNN graph encoder. We pre-train the CGCNN model with two SSL strategies using Barlow Twins and SimSiamese loss function. The weights of the pre-trained self-supervised model are used to initialize the graph encoder model during the fine-tuning stage for material property prediction.

### Graph neural network encoder

Most recent successful deep learning approaches for crystalline material property prediction are based on GNNs because of their ability to capture structural geometry and chemistry. In a crystal graph ($G$), we consider the atoms as the nodes ($V$), and interactions between them are modeled via edges ($E$). In general, GNNs aggregate information from the neighborhood of the node to construct embeddings that are updated iteratively. The update for the GNN can be described as in Eq. (1).

$$\boldsymbol{h}_v^{(k)} = \mathrm{COMBINE}^{(k)}\left(\boldsymbol{h}_v^{(k-1)}, \mathrm{AGGREGATE}^{(k)}\left(\{\boldsymbol{h}_u^{(k-1)} | u \in \mathcal{N}(v)\}\right)\right), \tag{1}$$

where $\boldsymbol{h}_v^{(k)}$ is the feature of the node $v$ at the $k$-th layer and $\boldsymbol{h}_v^{(0)}$ is initialized by node feature $\boldsymbol{x}_v$. $\mathcal{N}(v)$ denotes the set of all the neighbors of node $v$. $\boldsymbol{a}_v^{(k)}$ is the output from the aggregation operation at the $k^{\mathrm{th}}$ layer. The aggregation operation collects the features of neighboring nodes and the combination operation combines the original node feature with the aggregated features. To extract the feature of the entire crystal system, $\boldsymbol{h}_G$, readout operation integrates all the node features among the graph $G$ as given in Eq. (2).

$$\boldsymbol{h}_G = \mathrm{READOUT}\left(\{\boldsymbol{h}_v^{(k)} | v \in G\}\right). \tag{2}$$

The readout operations such as summation, averaging, and max pooling are most commonly used[54].

In this work, we implement the CGCNN[29] architecture as the GNN encoder. We choose CGCNN because of its competitive performance and computational efficiency when compared to other GNN baselines. Moreover, CGCNN is one of the most widely benchmarked baseline models for material property prediction allowing us to compare the performance of our SSL framework with CGCNN and other baselines. To encode crystal features and obtain an embedding, we use mean pooling to generate a latent representation with the dimension of 64. After the GNN encoder, the projection head with 2 MLP layers is attached to generate the final embedding on which the SSL loss functions are applied for pre-training. Additionally, we also implement a general purpose graph neural network GIN[20] to test the fidelity of SSL methods on another architecture apart from CGCNN.

To generate self-supervised learning representations, we need to construct different augmentations of the crystalline system. Inspired by AugLiChem,[55] we devise three different augmentation techniques (Fig. 1C), namely, random perturbation, atom masking and edge masking. The random perturbation augmentation perturbs each atom by a distance drawn from the uniform distribution between 0 Å and 0.05 Å. For atom masking, we randomly mask 10% of the atoms in the crystal, similarly for edge masking we randomly mask 10% of the edge features between two neighboring atoms. More details on atom masking and edge masking are provided in the Supplementary Information (Supplementary Fig. 1). These augmentations are applied to the crystalline systems and two augmented instances are generated randomly on the fly at each epoch during pre-training. These augmented instances are fed into the GNN encoder to generate embeddings on which SSL loss functions are applied.

### Barlow Twins loss

In the pre-training stage for $\mathrm{CT}_{\mathrm{Barlow}}$, we use the Barlow Twins loss function to learn graph representations from crystals. This loss is based on the redundancy reduction principle proposed by neuroscientist H. Barlow[56,57] and was introduced to SSL by Zbontar et al.[8]. We use the Barlow Twins loss function in CT because of its high performance and ease of implementation. Moreover, the Barlow Twins loss, unlike other contrastive loss functions, does not explicitly require positive and negative pairs for pre-training. The Barlow Twins loss function is applied to the cross-correlation matrix created from encoder-generated embeddings of the two different augmentations generated from the same crystalline system. The Barlow Twins loss function is represented by Eq. (3),

$$L_{\mathrm{BT}} \triangleq \sum_i (1 - C_{ii})^2 + \lambda \sum_i \sum_{j \neq i} C_{ij}^2, \tag{3}$$

where $\boldsymbol{C}$ is the cross-correlation matrix of embeddings from two augmented instances, the cross correlation matrix is given by Eq. (4). The $\lambda$ used in this work is 0.0051 same as the original paper.

$$C_{ij} \triangleq \frac{\sum_b \boldsymbol{Z}_{b,i}^A \boldsymbol{Z}_{b,j}^B}{\sqrt{\left(\boldsymbol{Z}_{b,i}^A\right)^2}\sqrt{\left(\boldsymbol{Z}_{b,j}^B\right)^2}} \tag{4}$$

where $b$ is the index of the data in batch and $i, j$ index the vector dimensions of the projector output ($Z^A$ and $Z^B$), for both the augmented instances $A$ and $B$ from the same crystalline material.

### SimSiam loss

We developed another variant of CT that uses SimSiam[12] denoted as $\mathrm{CT}_{\mathrm{SimSiam}}$. In this case, an extra MLP head $f(\cdot)$ is added to the GNN backbone to map the latent vector $Z$ to $P$, namely $\boldsymbol{P} = f(\boldsymbol{Z})$. The distance between two vectors is defined as Eq. (5).

$$\mathcal{D}(\boldsymbol{P}_b^A, \boldsymbol{Z}_b^B) = -\frac{\boldsymbol{P}_b^A}{\|\boldsymbol{P}_b^A\|_2} \cdot \frac{\boldsymbol{Z}_b^B}{\|\boldsymbol{Z}_b^B\|_2}, \tag{5}$$

where $b$ denotes the index of the data in a batch and $A$, $B$ denote two augmented instances from the same crystalline material. The objective to minimize given a batch is further shown in Eq. (6).

$$L_{\mathrm{SimSiam}} \triangleq \frac{1}{2}\sum_b \left(\mathcal{D}(\boldsymbol{P}_b^A, \mathtt{stopgrad}(\boldsymbol{Z}_b^B)) + \mathcal{D}(\boldsymbol{P}_b^B, \mathtt{stopgrad}(\boldsymbol{Z}_b^A))\right), \tag{6}$$

where $\mathtt{stopgrad}(\cdot)$ means the gradient is not back propagated on this branch of the $\mathrm{CT}_{\mathrm{SimSiam}}$ model. Such an asymmetric architecture and the stop-gradient operation avoid the collapse of learned representations.

### Training details

In the pre-training stage, the embedding dimension of the CGCNN encoder is set to 128 for the $\mathrm{CT}_{\mathrm{Barlow}}$ and 256 for the $\mathrm{CT}_{\mathrm{SimSiam}}$ model. We use the Adam optimizer[58] with a learning rate of 0.00001 and a batch size of 64 and pre-train the model for 15 epochs. The other hyperparameters for the CGCNN (graph encoder) model are kept the same as in the original paper. The train/validation ratio for pre-training data is 95%/5%. For pre-training, we combine the datasets from the Matminer database[59] and the hypothetical Metal-Organic Framework dataset[60], aggregating a total of 428,275 samples. The labels in the datasets are not used during CT pre-training. Additional details about the hyperparameters during the pretraining stage are available in Supplementary Table 2. In the fine-tuning stage, we add a randomly initialized MLP head with two fully connected layers to generate the final property prediction. The $\mathrm{CT}_{\mathrm{Barlow}}$ and $\mathrm{CT}_{\mathrm{SimSiam}}$ are tested on a variety of datasets including some datasets from previously published work OGCNN[30] and matbench suite[42]. Additional details about the hyperparameters used during the finetuning stage are mentioned in Supplementary Table 3.

## DATA AVAILABILITY

All data used in this work are publicly available. The authors used datasets from the matbench suite and Materials project, which are public data repositories. Matbench benchmark datasets are available at the website (https://matbench.materialsproject.org/). The datasets can also be found on the Materials Project website (https://materialsproject.org/). The codes and documentation for matbench can be found at the website (https://github.com/materialsproject/matbench). For the HOIP[61] (Dryad Digital Repository: https://doi.org/10.5061/dryad.gq3rg) and Lanthanides datasets[62], relevant publications have been cited from which data were obtained. These datasets are also made available at https://github.com/RishikeshMagar/Crystal-Twins.

## CODE AVAILABILITY

The code developed for this work is available at https://github.com/RishikeshMagar/Crystal-Twins.

Received: 24 May 2022; Accepted: 23 October 2022;
Published online: 10 November 2022

## REFERENCES

1. Schmidt, J., Marques, M. R., Botti, S. & Marques, M. A. Recent advances and applications of machine learning in solid-state materials science. *npj Comput. Mater.* **5**, 1–36 (2019).
2. Keith, J. A. et al. Combining machine learning and computational chemistry for predictive insights into chemical systems. *Chem. Rev.* **121**, 9816–9872 (2021).
3. Najafabadi, M. M. et al. Deep learning applications and challenges in big data analytics. *J. Big Data* **2**, 1–21 (2015).
4. Bengio, Y., Lecun, Y. & Hinton, G. Deep learning for ai. *Commun. ACM* **64**, 58–65 (2021).
5. Schleder, G. R., Padilha, A. C., Acosta, C. M., Costa, M. & Fazzio, A. From dft to machine learning: recent approaches to materials science–a review. *J. Phys. Mater.* **2**, 032001 (2019).
6. Chen, A., Zhang, X. & Zhou, Z. Machine learning: accelerating materials development for energy storage and conversion. *InfoMat* **2**, 553–576 (2020).
7. Chen, T., Kornblith, S., Norouzi, M. & Hinton, G. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning* 1597–1607 (PMLR, 2020).
8. Zbontar, J., Jing, L., Misra, I., LeCun, Y. & Deny, S. Barlow twins: self-supervised learning via redundancy reduction. In *International Conference on Machine Learning* 12310–12320 (PMLR, 2021).
9. Grill, J.-B. et al. Bootstrap your own latent-a new approach to self-supervised learning. *Adv. Neural Inf. Process. Syst.* **33**, 21271–21284 (2020).
10. Caron, M. et al. Unsupervised learning of visual features by contrasting cluster assignments. *Adv. Neural Inf. Process. Syst.* **33**, 9912–9924 (2020).
11. He, K., Fan, H., Wu, Y., Xie, S. & Girshick, R. Momentum contrast for unsupervised visual representation learning. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition* 9729–9738 (2020).
12. Chen, X. & He, K. Exploring simple siamese representation learning. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition* 15750–15758 (2021).
13. Lan, Z. et al. Albert: A lite BERT for self-supervised learning of language representations. In *International Conference on Learning Representations* (ICLR, 2019).
14. Wu, J., Wang, X. & Wang, W. Y. Self-supervised dialogue learning. In *Proc. 57th Annual Meeting of the Association for Computational Linguistics* 3857–3867 (ACL, 2019).
15. Wang, Y., Wang, J., Cao, Z. & Barati Farimani, A. Molecular contrastive learning of representations via graph neural networks. *Nat. Mach. Intell.* **4**, 279–287 (2022).
16. Zhu, J. et al. Dual-view molecule pre-training. Preprint at *arXiv* https://arxiv.org/abs/2106.10234 (2021).
17. Stärk, H. et al. 3D infomax improves gnns for molecular property prediction. In *International Conference on Machine Learning* 20479–20502 (PMLR, 2022).
18. Liu, S., Demirel, M. F. & Liang, Y. N-gram graph: simple unsupervised representation for graphs, with applications to molecules. *Adv. Neural. Inf. Process. Syst.* **32**, (2019).
19. Rong, Y. et al. Self-supervised graph transformer on large-scale molecular data. *Adv. Neural Inf. Process. Syst.* **33**, 12559–12571 (2020).
20. Hu, W. et al. Strategies for pre-training graph neural networks. In *International Conference on Learning Representations (ICLR)* (ICLR, 2020).
21. Li, P. et al. An effective self-supervised framework for learning expressive molecular global representations to drug discovery. *Brief. Bioinform.* **22**, bbab109 (2021).
22. Chithrananda, S., Grand, G. & Ramsundar, B. ChemBERTa: large-scale self-supervised pretraining for molecular property prediction. Preprint at *arXiv* https://arxiv.org/abs/2010.09885 (2020).
23. Rong, Y. et al. Grover: self-supervised message passing transformer on large-scale molecular data. In *Proc. 34th International Conference on Neural Information Processing Systems* 12559–12571 (NIPS, 2020).
24. Zhang, Z., Liu, Q., Wang, H., Lu, C. & Lee, C.-K. Motif-based graph self-supervised learning for molecular property prediction. Preprint at *arXiv* https://arxiv.org/abs/2110.00987 (2021).
25. Wang, Y., Magar, R., Liang, C. & Barati Farimani, A. Improving molecular contrastive learning via faulty negative mitigation and decomposed fragment contrast. *J. Chem. Inf. Modeling* **62**, 2714–2725 (2022).
26. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444 (2015).
27. Wu, Z. et al. A comprehensive survey on graph neural networks. *IEEE Trans. Neural Netw. Learn. Syst.* **32**, 4–24 (2020).
28. Welling, M. & Kipf, T. N. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations (ICLR 2017)* (ICLR, 2016).
29. 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).
30. Karamad, M. et al. Orbital graph convolutional neural network for material property prediction. *Phys. Rev. Mater.* **4**, 093801 (2020).
31. Schütt, K. T., Sauceda, H. E., Kindermans, P.-J., Tkatchenko, A. & Müller, K.-R. Schnet–a deep learning architecture for molecules and materials. *J. Chem. Phys.* **148**, 241722 (2018).
32. 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).
33. Louis, S.-Y. et al. Graph convolutional neural networks with global attention for improved materials property prediction. *Phys. Chem. Chem. Phys.* **22**, 18141–18148 (2020).
34. Gasteiger, J., Groß, J. & Günnemann, S. Directional message passing for molecular graphs. In *International Conference on Learning Representations* (ICLR, 2019).
35. Gasteiger, J., Giri, S., Margraf, J. T. & Günnemann, S. Fast and uncertainty-aware directional message passing for non-equilibrium molecules. Preprint at *arXiv* https://arxiv.org/abs/2011.14115 (2020).
36. Palizhati, A., Zhong, W., Tran, K., Back, S. & Ulissi, Z. W. Toward predicting intermetallics surface properties with high-throughput DFT and convolutional neural networks. *J. Chem. Inf. Modeling* **59**, 4742–4749 (2019).
37. Back, S. et al. Convolutional neural network of atomic surface structures to predict binding energies for high-throughput screening of catalysts. *J. Phys. Chem. Lett.* **10**, 4401–4408 (2019).
38. Gilmer, J., Schoenholz, S. S., Riley, P. F., Vinyals, O. & Dahl, G. E. Neural message passing for quantum chemistry. In *International Conference on Machine Learning* 1263–1272 (PMLR, 2017).
39. Unke, O. T. & Meuwly, M. Physnet: a neural network for predicting energies, forces, dipole moments, and partial charges. *J. Chem. Theory Comput.* **15**, 3678–3693 (2019).
40. Gu, G. H. et al. Practical deep-learning representation for fast heterogeneous catalyst screening. *J. Phys. Chem. Lett.* **11**, 3185–3191 (2020).
41. Jha, D. et al. Elemnet: deep learning the chemistry of materials from only elemental composition. *Sci. Rep.* **8**, 1–13 (2018).
42. Dunn, A., Wang, Q., Ganose, A., Dopp, D. & Jain, A. Benchmarking materials property prediction methods: the matbench test set and automatminer reference algorithm. *npj Comput. Mater.* **6**, 1–10 (2020).
43. Choudhary, K. & DeCost, B. Atomistic line graph neural network for improved materials property predictions. *npj Comput. Mater.* **7**, 1–8 (2021).
44. Ihalage, A. & Hao, Y. Formula graph self-attention network for representation-domain independent materials discovery. *Adv. Sci.* **9**, 2200164 (2022).
45. Moosavi, S. M., Jablonka, K. M. & Smit, B. The role of machine learning in the understanding and design of materials. *J. Am. Chem. Soc.* **142**, 20273–20287 (2020).
46. Ryan, K., Lengyel, J. & Shatruk, M. Crystal structure prediction via deep learning. *J. Am. Chem. Soc.* **140**, 10158–10168 (2018).
47. Liang, H., Stanev, V., Kusne, A. G. & Takeuchi, I. Cryspnet: crystal structure predictions via neural networks. *Phys. Rev. Mater.* **4**, 123802 (2020).
48. Long, T. et al. Constrained crystals deep convolutional generative adversarial network for the inverse design of crystal structures. *npj Comput. Mater.* **7**, 1–7 (2021).
49. Kim, S., Noh, J., Gu, G. H., Aspuru-Guzik, A. & Jung, Y. Generative adversarial networks for crystal structure prediction. *ACS Cent. Sci.* **6**, 1412–1420 (2020).
50. Xie, T., Fu, X., Ganea, O.-E., Barzilay, R. & Jaakkola, T. S. Crystal diffusion variational autoencoder for periodic material generation. In *International Conference on Learning Representations* (2021).
51. Yao, Z. et al. Inverse design of nanoporous crystalline reticular materials with deep generative models. *Nat. Mach. Intell.* **3**, 76–86 (2021).
52. van der Maaten, L. & Hinton, G. Visualizing data using t-sne. *J. Mach. Learn. Res.* **9**, 2579–2605 (2008).
53. Hadsell, R., Chopra, S. & LeCun, Y. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)* 1735–1742 (IEEE, 2006).
54. Xu, K., Hu, W., Leskovec, J. & Jegelka, S. How powerful are graph neural networks? In *International Conference on Learning Representations* (ICLR, 2018).
55. Magar, R. et al. AugLiChem: data augmentation library of chemical structures for machine learning. *Mach. learn.: sci. technol.* (IOP Publishing) (2022).
56. Barlow, H. Redundancy reduction revisited. *Network* **12**, 241–253 (2001).
57. Barlow, H. B. & Rosenblith, W. A. *Possible Principles Underlying the Transformations of Sensory Messages* 217–234 (MIT Press, 1961).
58. Kingma, D. P. & Ba, J. Adam: a method for stochastic optimization. In *International Conference on Learning Representations* (ICLR, 2015).
59. Ward, L. et al. Matminer: An open source toolkit for materials data mining. *Comput. Mater. Sci.* **152**, 60–69 (2018).

60. Wilmer, C. E. et al. Large-scale screening of hypothetical metal–organic frameworks. *Nat. Chem.* **4**, 83–89 (2012).
61. Kim, C., Huan, T. D., Krishnan, S. & Ramprasad, R. A hybrid organic-inorganic perovskite dataset. *Sci. Data* **4**, 1–11 (2017).
62. Pham, T. L. et al. Machine learning reveals orbital interaction in materials. *Sci. Technol. Adv. Mater.* **18**, 756–765 (2017).
63. Choudhary, K., Kalish, I., Beams, R. & Tavazza, F. High-throughput identification and characterization of two-dimensional materials using density functional theory. *Sci. Rep.* **7**, 1–16 (2017).
64. Petretto, G. et al. High-throughput density-functional perturbation theory phonons for inorganic materials. *Sci. Data* **5**, 1–12 (2018).
65. Petousis, I. et al. High-throughput screening of inorganic compounds for the discovery of novel dielectric and optical materials. *Sci. Data* **4**, 1–12 (2017).
66. de Jong, M. et al. Charting the complete elastic properties of inorganic crystalline compounds. *Sci. Data* **2**, 150009 (2015).
67. Castelli, I. E. et al. Computational screening of perovskite metal oxides for optimal solar light capture. *Energy Environ. Sci.* **5**, 5814–5819 (2012).
68. Jain, A. et al. Commentary: The materials project: a materials genome approach to accelerating materials innovation. *APL Mater.* **1**, 011002 (2013).

## ACKNOWLEDGEMENTS

The work has been supported by the start-up fund provided by Department of Mechanical Engineering at Carnegie Mellon University. The authors would like to thank Prakarsh Yadav, Alison Bartsch, and Cooper Lorsung for their comments on the manuscript.

## AUTHOR CONTRIBUTIONS

R.M. and A.B.F. ideated the concept. R.M. wrote the code and benchmarked the performance of the model. Y.W. assisted in the benchmarking and writing the manuscript. A.B.F. supervised the work.

## 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/s41524-022-00921-5.

**Correspondence** and requests for materials should be addressed to Amir Barati Farimani.

**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.