ACS APPLIED MATERIALS & INTERFACES

www.acsami.org

Research Article

Machine Learning Accelerated Discovery of Promising Thermal Energy Storage Materials with High Heat Capacity

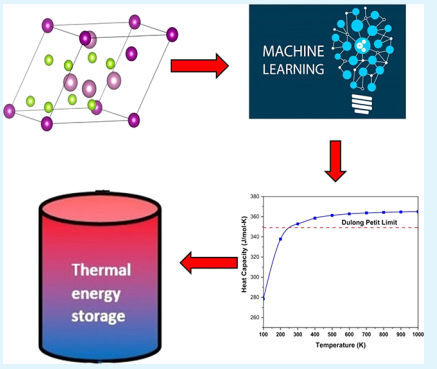
Joshua Ojih, Uche Onyekpe, Alejandro Rodriguez, Jianjun Hu, Chengxiao Peng,* and Ming Hu*

Cite This: https://doi.org/10.1021/acsami.2c11350

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Thermal energy storage offers numerous benefits by reducing energy consumption and promoting the use of renewable energy sources. Thermal energy storage materials have been investigated for many decades with the aim of improving the overall efficiency of energy systems. However, finding solid materials that meet the requirement of high heat capacity has been a grand challenge for material scientists. Herewith, by training various machine learning models on 3377 high-quality data from full density functional theory (DFT) calculations, we efficiently search for potential materials with high heat capacity. We build four traditional machine learning models and two graph neural network models. Cross-comparison of the prediction performance and model accuracy was conducted among different models. The deeperGATGNN model exhibits high prediction accuracy and is used for predicting the heat capacity of 32,026 structures screened from the open quantum material database. We gain deep insight into the correlation between heat capacity and structure descriptors such as space group, prototype, lattice volume, atomic weight, etc. Twenty-two structures were predicted to possess high heat capacity, and the results were further validated with DFT calculations. We also identified one special structure, namely, $MnIn_2Se_4$, with space group no. 227 ($Fd\bar{3}m$), that exhibits extremely high heat capacity, even higher than that of the Dulong–Petit limit at room temperature. This study paves the way for accelerating the discovery of novel thermal energy storage materials by combining machine learning with minimal DFT inquiry.



KEYWORDS: *Thermal Energy Storage, Materials Discovery, Machine Learning, Graph Neural Network, Heat Capacity, Dulong–Petit Limit*

INTRODUCTION

The rapid development of global industrialization and population has led to the increase in the demand of energy. However, the use of fossil fuels has caused severe environmental pollution by greenhouse gas emission,[1] which has led to great research interest and development of renewable energy. Thermal energy storage (TES) systems have emerged as promising solutions in solving the problem of storing the excess energy produced from renewable energy sources and made available for later use, and developing TES materials is the core element of the TES system.[2] TES is very important in many engineering applications. It has been applied mostly in solar energy systems,[3] and it can also be applied to store heat in building structures, to couple waste heat and district heating systems, and to couple heat pumps and combined heat power generators in district heating networks. TES materials can be classified into sensible heat storage (SHS), latent heat storage (LHS), and thermochemical heat storage (THS) materials. Each storage material occurs in a different physical state, storing and release energy differently. For example, SHS materials store and release energy through a change in temperature,[4] and LHS materials store the energy mainly based on phase change.[5,6] However, the THS system utilizes reversible thermochemical reactions for heat storage. In this paper, we search for materials based on SHS, and those are materials which are affected by an increase or decrease of the material temperature. Thus, it is desirable for the storage medium to have as high as possible heat capacity,[2] as the heat capacity of a material is the amount of heat required to change the temperature of the material.

All materials have a capability of absorbing and storing heat due to the fact they have mass (m) and specific heat capacity (c_p) at constant pressure. As described by the thermodynamics, for a temperature difference ΔT, the amount of energy stored in SHS material is given by

$$Q_{sensible} = mc_p\Delta T \quad (1)$$

As clearly seen from eq 1, the $Q_{sensible}$ depends on the heat capacity of the material. For solid materials, the specific heat

Received: June 26, 2022
Accepted: September 6, 2022

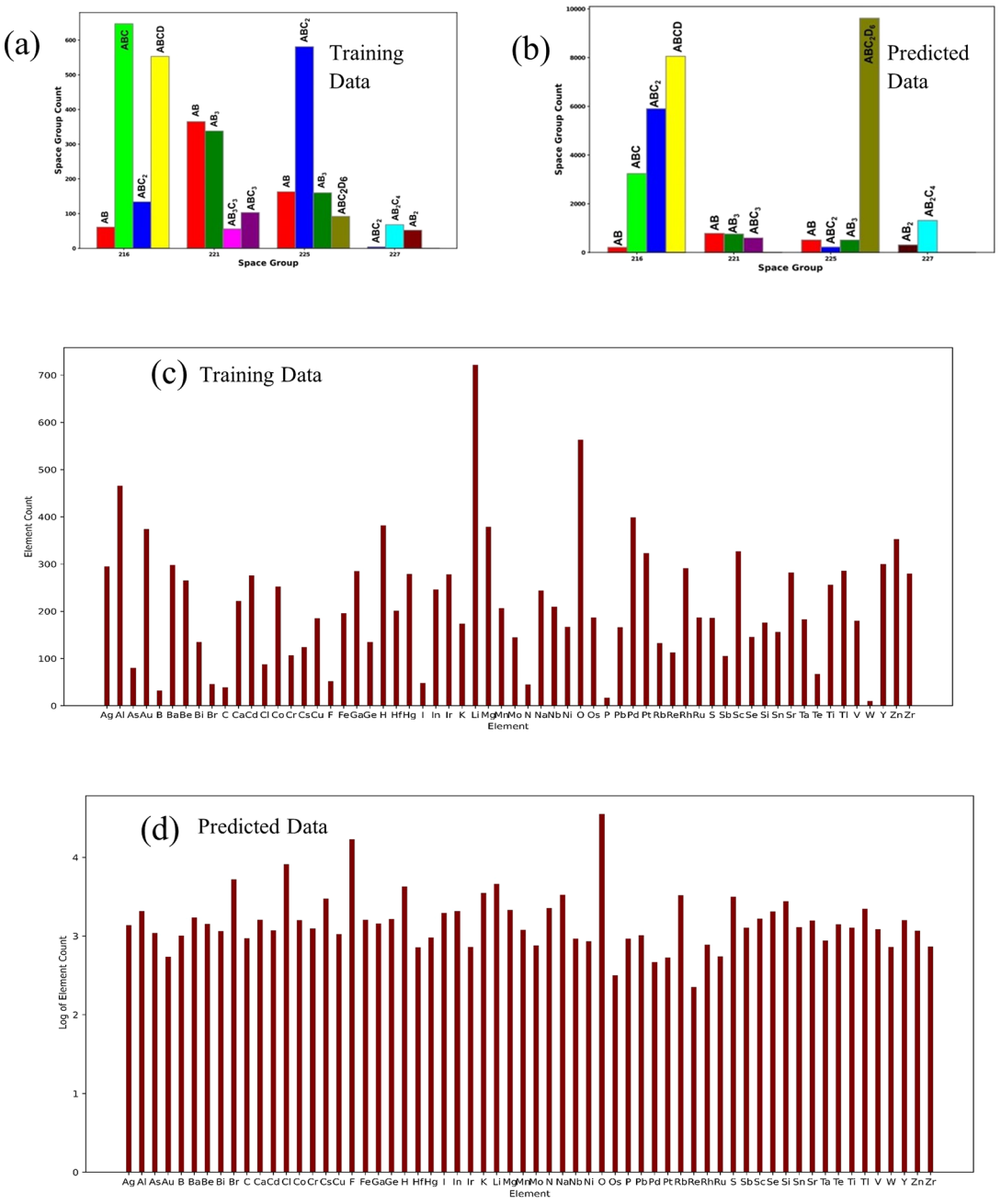


Figure 1. Statistics (structure type and element distribution) for training and prediction data. (a,b) Space group count for training and predicted data. (c,d) Element distribution for training and predicted data.

per mole of a substance has an upper limit of about 3*NR* (the so-called Dulong–Petit law), where R = 8.31441 J/mol-K being the molar gas constant and N the molar number of atoms in the material. Thus, the molar thermal energy q_{mol} stored in solids can be approximated by

$$q_{mol} \leq 3NR\Delta T \quad (2)$$

In the past decade, material scientists have been using high-throughput screening (HTS) coupled with density functional theory (DFT) for structure–property prediction with high accuracy in search of novel materials.[7,8] DFT is highly accurate but less efficient, and hence it is computationally expensive and time-consuming.[7,9] Machine learning (ML) methods offer the possibility of reducing the number of DFT calculations needed to discover new materials because ML models are based on statistical predictions rather than physical-based calculations, hence they are computationally less expensive.[10]

ML has been used for prediction of mechanical properties of metal alloys,[11,12] band gap energies of crystals,[13,14] and formation energies of crystals.[15−17] Kauwe et al.[10] used ML to predict the heat capacity of solid inorganics over a wide range of temperatures. Their work showed that the ML algorithm can be presented as an alternative method to predict heat capacity for any material at a wide range of temperatures, with better accuracy and less time than traditional DFT methods. In medicine, Odigwe et al.[18] utilized ML techniques to predict the expected magnitude of heart failure patients' left ventricular end-systolic volume, 3 months after cardiac resynchronization therapy placement, within a 17% median margin of error. Though ML is highly efficient, it has few

https://doi.org/10.1021/acsami.2c11350
ACS Appl. Mater. Interfaces XXXX, XXX, XXX−XXX

limitations which reduces its accuracy in prediction; such limitations include measurement error,[19] lack of generality and precision, reliance on high-quality data,[20,21] inability to determine high level concepts,[22] and poor extrapolation.[23,24] In this work, we implement four traditional ML algorithms and two graph neural network (GNN) algorithms. The accuracy of the traditional ML depends on the effective input representation of the crystal structures. In this study, we represent our crystal structure with many simple descriptors. We also use GNN models which do not need any input representation, as they learn directly from the crystal structures. After thoroughly screening of the 50,000 crystal structures from the OQMD, we identify 22 structures with high heat capacity and one of which has heat capacity even higher than that of the traditional Dulong−Petit limit. We also compare our results between each ML model and DFT calculations to validate our findings.

METHODOLOGY

Heat Capacity Calculation by DFT. Optimal performance of ML models requires high-quality training data either from high-throughput calculations or from experiment. The initial data used in this paper to train and build the heat capacity predictive models were obtained from our first-principles calculations of 3377 cubic crystal structures from four different space groups (specifically space group numbers 216, 221, 225, and 227), eight different atom types, and 66 elements, as shown in Figure 1. The initial data downloaded from the OQMD database were 50,000. We then screened the data by removing any structures with a band gap equal to zero, as we are only interested in semiconductors. Finally, we obtained 32,026 structures after screening. The initial crystal structures (atomic positions and lattice parameters) were downloaded from the OQMD database[25,26] and then were reoptimized by our own computational parameters. We did not explore the prediction of space groups that were not contained within the training data. This is because we only successfully reoptimized ∼50,000 cubic structures with the same space groups as the training data. Our predicted data have many structures with few atoms (e.g., less than 5) in the primitive cells, while large primitive cell structures (such as above 20 atoms) are few. We also did not explore disordered supercells; all predicted structures are ordered crystalline phases. The DFT calculations were performed using the plane-wave basis projector-augmented wave (PAW) method,[27] within the Perdew−Burke−Ernzerhof exchange-correlation functional,[28] as implemented in the Vienna ab initio simulation package (VASP).[29−31] The cutoff energy was set to 520 eV for all crystal structures. The energy and force criteria for the DFT calculation of structure optimization were set at 10^{-6} eV and 10^{-4} eV/Å, respectively. The phonon band structures were determined using the supercell approach implemented in the PHONOPY package.[32] The second- and third-order interatomic force constants (IFCs) required for phonon band structure calculation were calculated using a compressive sensing lattice dynamics (CSLD) method,[33] which extracts the IFCs from the Taylor-expanded interatomic forces in terms of atomic displacements via advanced compressive sensing techniques. All atoms in the supercells were randomly displaced with a magnitude of 0.03 Å by the PHONOPY package. The advantage of this method is the significantly lowered requirement of supercell numbers for IFCs, reducing the needed DFT for converged IFCs. For example, for quaternary (ABCD, space group number 225) Heusler structures, a traditional finite displacement method would require at magnitude $\sim 10^2$ structures to get the third-order interatomic force constants even if only the third nearest neighbor is considered, whereas only 12−16 randomly displaced supercell structures are needed from CSLD. With IFCs obtained by DFT, the phonon dispersions were calculated by the PHONOPY package, and the heat capacity was further calculated by the ShengBTE package.[34] This technique of calculating heat capacity by the ShengBTE package using second-order harmonic interatomic force constants has been widely used for various materials in the past few years.[35,36]

Data Generation and Analysis for Machine Learning Model Training. Table 1 shows some of the DFT-calculated heat capacities compared to experimental data. From Table 1, we can observe that our DFT result is in very good agreement with the experimental result. Both training and predicted data contain four different space groups, eight atom types, and 66 elements in the periodic table except the La and Ac elements. In Figure 1, we present the statistics of our training and predicted data. Figure 1a,b shows the number of structures and atom types in each space group, respectively. This statistic gives us insight on what space group and atom type we can explore to find novel TES materials. We observe that the space group 227 in our training data is relatively small compared to other space group, but our model was able to predict all high heat capacity from the 227 space group and AB_2C_4 atom type (see details below). Figure 1c,d shows the distribution of elements in our training and predicted data, respectively. All elements showing up in the predicted data occurred in the training structures, which makes the prediction of the trained models more reasonable.

Table 1. Experimental and DFT-Calculated Heat Capacity of Some Materials from OQMD

material ID	formula	experiment (J/g-K)	DFT (J/g-K)
3683	KCl	0.695[58]	0.653
110592	NaCl	0.85[59]	0.822
1224082	ZnS	0.469[60]	0.465
1222438	LiF	1.5617[61]	1.5134
1223683	ZnO	0.495[62]	0.501
12376	AlAs	0.49[63]	0.435
1104282	AlN	0.819[64]	0.802
5686	NaF	1.088[65]	1.084
11589	PbS	0.285[66]	0.2028
1104590	KI	0.313[67]	0.297
2577	MgO	1.0[68]	0.937

Figure 2 illustrates the relation between each attribute and the target. The somewhat monotonic relationships between the variables justify the use of the Spearman's rank correlation in evaluating the monotonic relationship between the attributes and the target, as is presented Figure 3. High multicollinearity was observed among the independent variables "Average Weight" and "Total Weight", "Number Density" and "Bond Length", with values of 0.73 and −0.87, respectively, where zero is ideal (representing no correlation), and 1 and −1 represent a perfect correlation (unideal for the independent variables). In this study, we define correlation values greater than 0.7 as high collinearity. Nevertheless, multicollinearity is undesirable as it enables less reliable statistical inferences. The statistical relevance of the observed correlations is measured

https://doi.org/10.1021/acsami.2c11350
ACS Appl. Mater. Interfaces XXXX, XXX, XXX−XXX

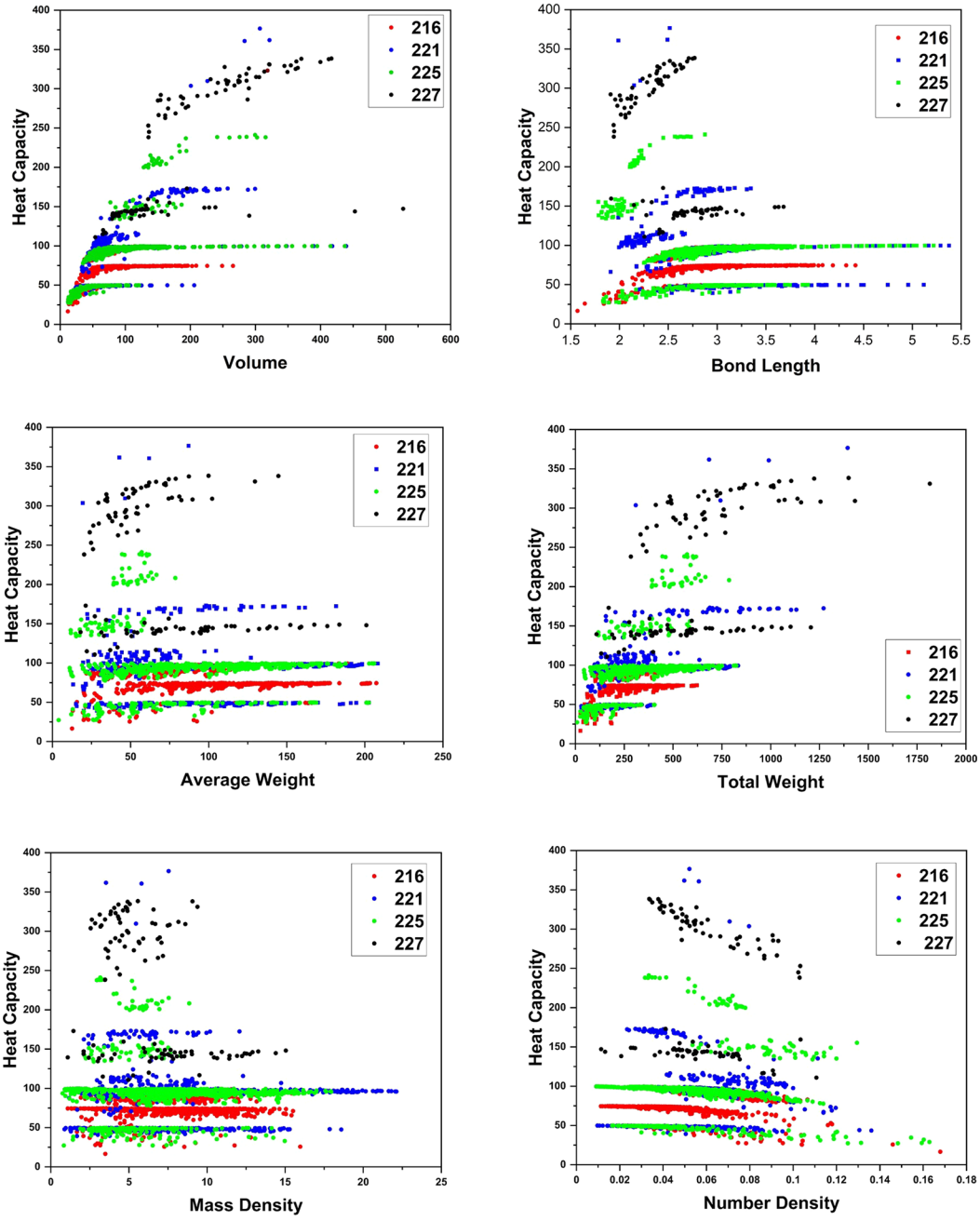


Figure 2. Scatter plot between attributes and target variables with different colors representing different space groups. This justifies the use of Spearman's rank correlation in evaluating the monotonic relationship between the attributes and target.

using the p value, with the assumption of the null hypothesis being true. The p values for the observed correlations between the input features are below 5% or 0.05, thus rejecting the null hypothesis and supporting the unlikelihood of the correlations happening by chance.

The Spearman's rank correlation between the attributes and the target is shown in Figure 3. It can be observed from Figure 3 that the "Number Density" attribute is negatively correlated with the heat capacity variable, while the "Total Weight", "Average Weight", "Mass Density", "Volume", and "Bond Length" are positively correlated with the heat capacity. It can be further observed that the "Total Weight" and "Volume" attributes have a relatively stronger correlation with the target with correlation values of 0.59 and 0.66, respectively, where 1 and −1 represent a perfect correlation (an ideal correlation) and zero represents no correlation. The p value results suggest the unlikelihood of the correlations for the "Number Density", "Mass Density", and "Average Weight" attributes with the "Heat Capacity" target.

Outliers were found to be present within the data, as presented in Figure 4. The outliers are defined in this study as observations outside the upper boundary and lower boundary for each attribute as defined in eqs 3−5. Nevertheless, the outliers in this data set were left unaddressed as there was no valid reason for their exclusion; they are more representative of reality. Furthermore, all attributes were normalized to a 0−1 bound to reduce learning bias.

$$IQR = Q3 - Q1 \tag{3}$$

https://doi.org/10.1021/acsami.2c11350
ACS Appl. Mater. Interfaces XXXX, XXX, XXX−XXX

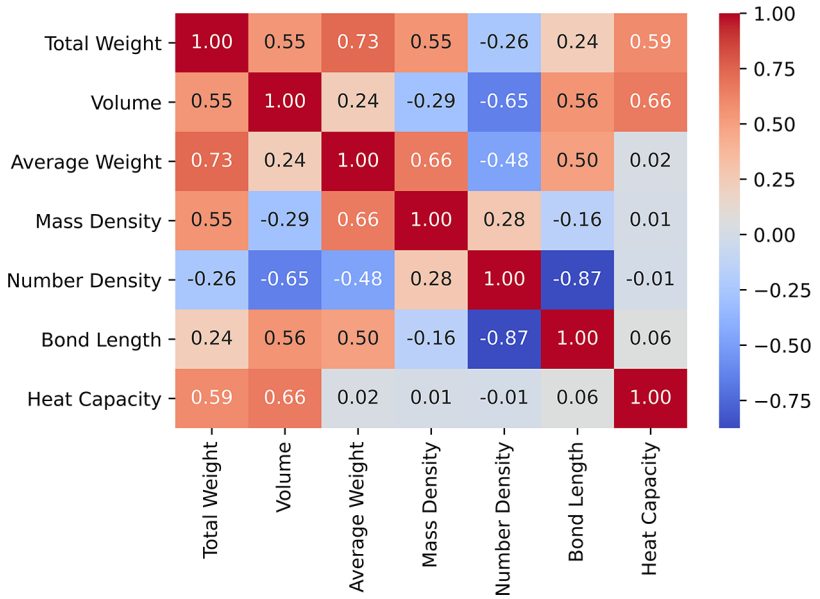


Figure 3. Spearman's rank multicollinearity study between the independent variables and the dependent variable.

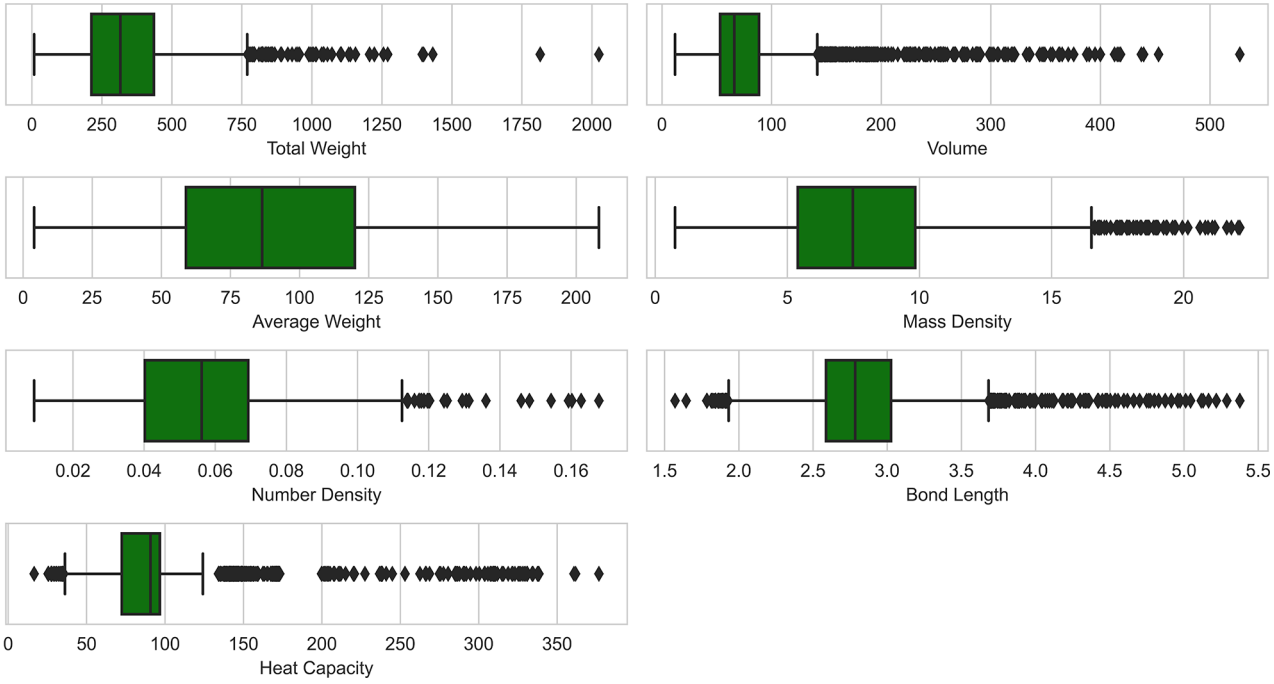


Figure 4. Outliers within the independent and dependent variables.

$$\text{upper boundary} = Q3 + 1.5 \times \text{inter-quartile range} \quad (4)$$

$$\text{lower boundary} = Q1 - 1.5 \times \text{inter-quartile range} \quad (5)$$

where inter-quartile range = $Q3 - Q1$, $Q1$ is the 25th quartile, and $Q3$ is the 75th quartile.

Machine Learning Model Training. Four traditional ML models and two graph neural networks are investigated in this study, namely, the linear regression, random forest, extreme gradient boosting, Catboost, crystal graph convolution neural network (CGCNN), and global attention graph neural network (deeperGATGNN). We discuss in this subsection, the learning scheme of the machine learning approach, the summary of the theoretical underpinnings of these models, and a description of the implementation, optimization, and validation of the investigated machine learning approaches. The learning task is formulated to map the relationship between the input features and the target ("Heat Capacity"). The objective of the learning process is defined in eq 6, where the learning objective is defined to be the minimization of the mean absolute error (MAE) between the ML-predicted heat capacity values and the DFT-calculated heat capacity.

$$\text{minimize: } J(\theta_0, \theta_1) = \frac{1}{2m} \sum_{i=1}^{m} |h_\theta(x^{(i)}) - y^{(i)}| \quad (6)$$

Linear Regression. Linear regression is used to find the best line of fit which describes the statistical relation between a defined set of predictors and target variables.[37] It is used for the value of the target (dependent variable) based on the

https://doi.org/10.1021/acsami.2c11350
ACS Appl. Mater. Interfaces XXXX, XXX, XXX−XXX

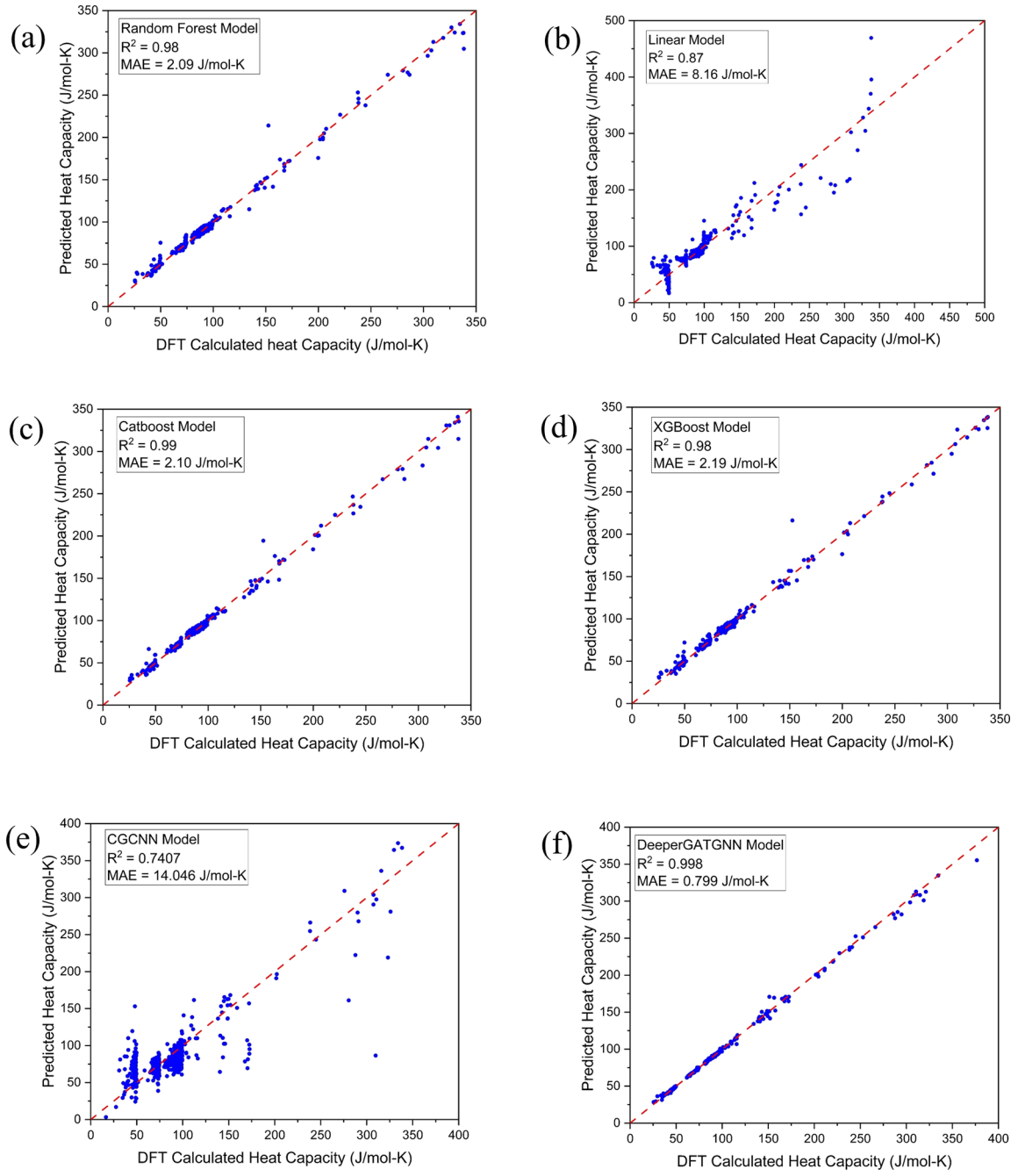


Figure 5. Testing results of the ML and GNN model. (a–d) Testing results for the traditional ML model, random forest, linear, Catboost and XGBoost models. (e,f) CGCNN and deeperGATGNN model.

information provided on the predictors (independent variables). The operation of the linear regression model is generally governed by

$$\text{target} = w \times \text{predictor} + b \tag{7}$$

where w refers to the slope of the line and b represents the intercept of the line.

Random Forest. The random forest (RF) machine learning algorithm was introduced by Breiman in 2001.[38] It is characterized by a collection of many individual decision trees, thus operating as an ensemble. The output of the ensemble is selected as the average of the predictions of each individual tree. The RF is created by the random feature selections and bagging as introduced by Breiman,[39] which helps to reduce overfitting. The performance of the RF algorithm on regression problems has been demonstrated in several application areas such as predicting atomic local environment[40] and lattice thermal conductivity of crystalline material.[41] See ref 38 for more information on random forest.

Extreme Gradient Boosting (XGBoost). Boosting proposed in 1999 by Schapire[42] is an ensemble method that aggregates and integrates multiple base (tree) learners to produce better predictions of classification and regression problems. Boosting outperforms bagging when the data are characterized by less noise.[43,44] The XGBoost specifically as proposed by Chen and Guestrin in 2016[45] is a scalable and efficient implementation of the gradient boosting technique proposed by Friedman et al. in 2001.[46,47] It is characterized by

https://doi.org/10.1021/acsami.2c11350
ACS Appl. Mater. Interfaces XXXX, XXX, XXX−XXX

an efficient linear model solver and tree learning algorithm. The XGBoost creates new decision trees to fit the residuals of previous decision trees through a process of continuous iteration with the aim of improving the prediction accuracy with each passing iteration. The XGBoost model makes use of the average of the predictions of the individual number of trees within a given sample.[48] More details on XGBoost can be found in Chen et al.[45]

Catboost. Catboost has found successful applications to several problems.[49,50] Initially proposed by Prokhorenkova et al. in 2017,[51] Catboost is based on the gradient boosting decision tree which uses a complex ensemble learning approach. During the learning stage, decision trees are sequentially constructed to produce subsequent trees with decrease loss. The decision tree learns from the preceding trees and influences the creation of the next tree, thus boosting the performance of the model. One of the ways the Catboost model differs from other gradient boost techniques is in its ordered boosting mechanism which addresses the target leakage problem of traditional gradient boosting methods caused by gradient bias.

Graph Neural Network. CGCNN was development by Xie et al. in 2018,[52] and deeperGATGNN[53] was further developed by Omee et al. in 2021. Both GNN codes have found success in the material discovery for accurate and efficient prediction of material properties. Both CGCNN and deeperGATGNN extract features from the crystal structure, which are then used for training the model. They combine the descriptors and learning model into one inseparable step. The model learns material properties directly from the connection of atoms in the crystal.

ML Model Optimization, Validation, and Prediction. For the effective application of the machine learning models to the SHC prediction problem, the models were optimized using the sklearns randomized search library.[54] The machine learning models are trained on 80% of the randomly sampled data set using the Scikit-learn library.[55] The results describing the performance of the model on the training data set are presented in Figure 5. Figure 5 highlights the results describing the performance validation of the trained model on the outstanding 20% of the data set. The average of the results obtained via a cross-validation with five folds demonstrates the superiority of the Catboost model compared to the other traditional ML models, but the graph attention neural network (deeperGATGNN) has the best performance over all models used.

RESULTS AND DISCUSSION

Figure 5a–f shows the testing results for all trained models. The MAE for the deeperGATGNN model, as shown in Figure 5f, was relatively small, and the R^2 value is much higher compared to that of other ML models, which justified the use of the deeperGATGNN model for the further screening of potential TES materials. We determined that the Catboost model also has a good R^2 value compared to that of the deeperGATGNN model, but the corresponding MAE is much higher. According to the testing results from Figure 5, the deeperGATGNN model was finally chosen for the prediction of 32,026 structures from the OQMD database in order to search for potential TES materials with high heat capacity. Another advantage of the deeperGATGNN model is that there is no manual feature extraction needed. The only input for the model is the 3D atomic structures with basic information on elements, atomic positions, etc. After using the deeperGATGNN model to predict the heat capacity of 32,026 structures, we finally chose 22 highest predictions of heat capacity to be validated by DFT calculations. Addition of more data will improve the performance of our model, as shown in Figure S2, where the 22 recommended structures (calculated by DFT) were added to the previous training data. There is a slight improvement in our result, as seen in the MAE, which is due to the small amount of new data added (22), compared with the previous large amounts of training data (3377).

We would like to emphasize that it is possible to use active learning to further improve the deeperGATGNN model, provided that a significantly large amount of new data will be added. We have tried to add the 22 recommended structures (calculated by DFT) to the previous training data and retrain the deeperGATGNN model. There is only a slight improvement in our result in terms of MAE (results not shown for brevity), which is due to the small amount of new data added (22), compared with a previous large amount of training data (3377). However, performing high precision DFT calculations on very large number of new structures is very time- and resource-consuming. Due to the limited time and resources, we cannot perform large-scale active learning approaches to 32,026 structures. Another reason for not using the active learning approach is that we find the MAE of our trained deeperGATGNN model is already low, i.e., ∼0.8 J/mol-K, compared to the heat capacity range of 350 J/mol-K, meaning the MAE is within 0.23% of the heat capacity range. The R^2 score of the deeperGATGNN model is already 0.998, leaving very little room for further improvement by active learning.

Figure 6 shows the comparison of heat capacity between the prediction by the deeperGATGNN model and DFT calculations for our 22 highest heat capacity structures. We can see that the DFT calculations confirmed that our 22 structures have relatively high heat capacity (above 300 J/mol-K), which makes them potential TES materials. The agreement between the deeperGATGNN model prediction and DFT results is good, meaning that the deeperGATGNN model is indeed well trained and it captures the inherent nature and properties of the atomic structures in terms of vibrational frequency and phonon density of states.

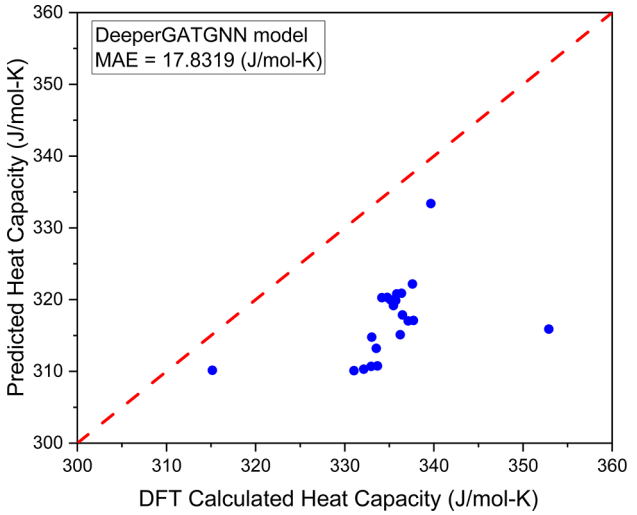


Figure 6. DFT result and deeperGATGNN prediction for the 22 selected structures with high heat capacity.

https://doi.org/10.1021/acsami.2c11350
ACS Appl. Mater. Interfaces XXXX, XXX, XXX−XXX

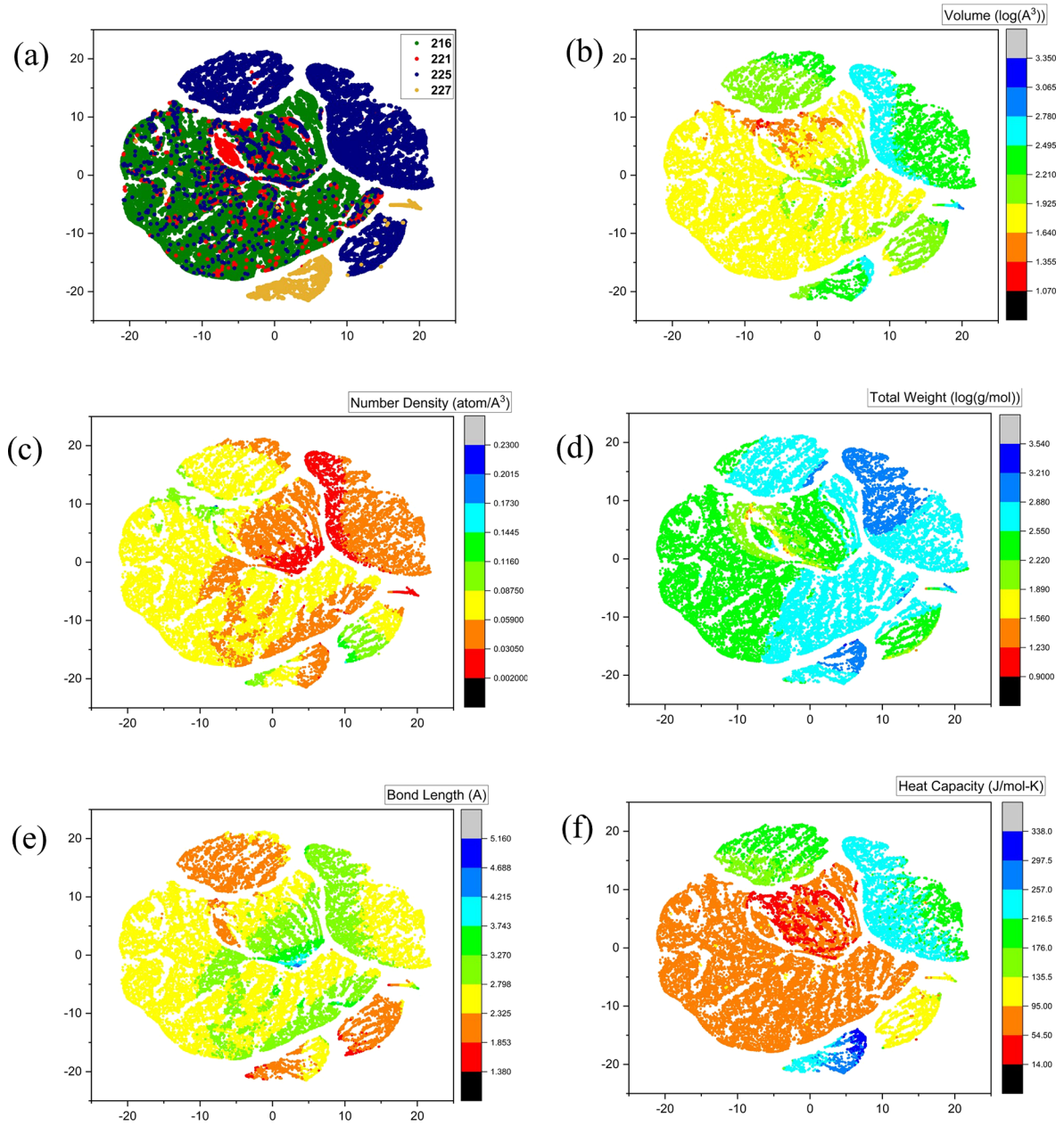


Figure 7. t-SNE plot with perplexity of 50 showing analysis and insight into the different magnitude of heat capacity among all 32,026 predicted structures. (a) Distribution of the different space group in the predicted structures. (b–e) Volume, number density, total weight, and bond length distribution. (f) Heat capacity distribution for the predicted structures.

In order to gain deep insight into the model training and structure–property relationship, Figure 7 presents the t-distributed stochastic neighbor embedding (t-SNE) plot for exploring high-dimensional data. The t-SNE method was introduced by Van der Maaten and Hilton in 2008.[56] It is a nonlinear dimensional reduction suited for embedding high-dimensional data for visualization in a low-dimensional space.[56,57] This helps us to visualize our high-dimensional data in a 2D plot and get a correlation between target properties and material descriptors and insight of potential structures as TES materials. Figure 7a shows the distribution of our predicted structures based on space group number. We can see that the structures with space group 227 are few compared to other space group structures and are for the most part isolated from the other space groups as seen by the yellow islands. The isolation of space group 227 structures from other structures can be understood in terms of the much higher structural symmetry of space group 227 compared to other space groups. In particular, space groups 225 and 216 are very similar except that different elements usually occupy the same sites for space group 216, and thus the structural symmetry is reduced. Figure 7b–e shows the color distribution of some of the atomic properties (volume, number density, total weight, and bond length) on the same t-SNE plot as Figure 7a. It is seen that those distributions have different color patterns, meaning that the structures have very diverse atomic features. Nevertheless, the dominant yellow and green color in those plots indicate that the atomic features are not uniformly distributed, which is quite normal considering our screening structure pool is huge (32,026 structures), and diverse

https://doi.org/10.1021/acsami.2c11350
ACS Appl. Mater. Interfaces XXXX, XXX, XXX–XXX

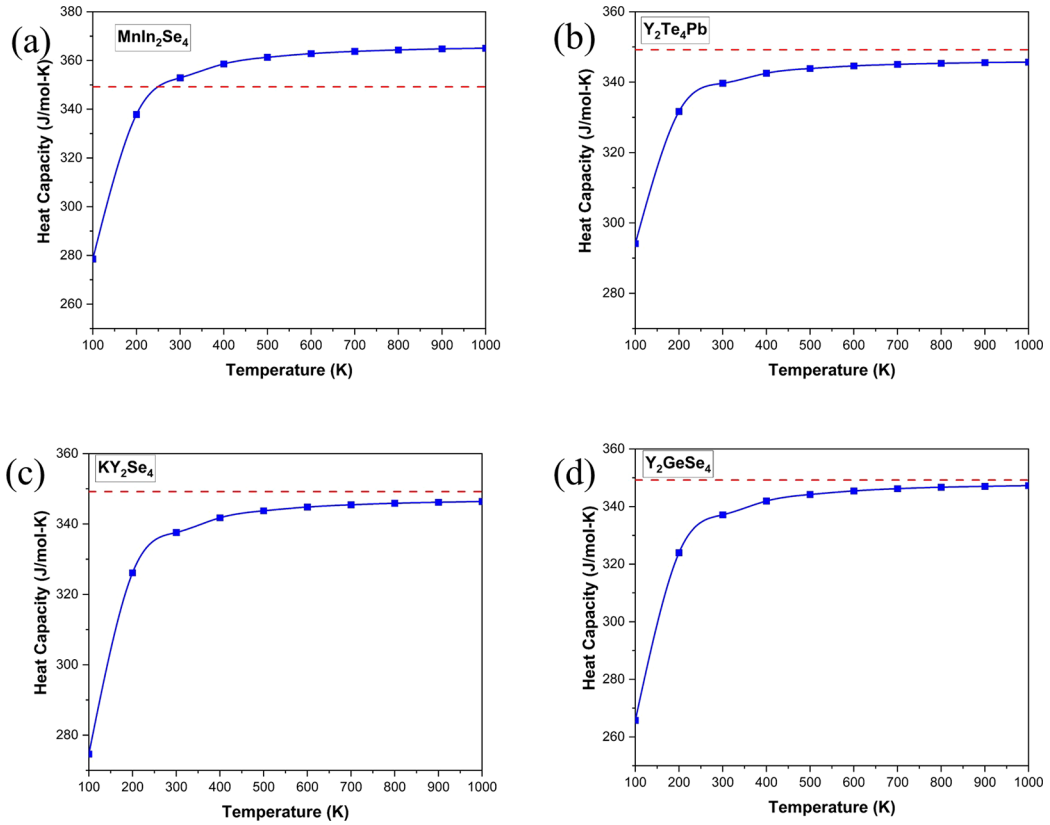


Figure 8. Temperature-dependent four top highest heat capacity materials: (a) $MnIn_2Se_4$, (b) Y_2Te_4Pb, (c) KY_2Se_4, and (d) Y_2GeSe_4. The dashed line denotes the Dulong–Petit limit. The heat capacity of $MnIn_2Se_4$ exceeds the Dulong–Petit limit near and above room temperature, whereas the other three materials' heat capacity only approaches the Dulong–Petit limit at elevated temperatures.

structures should have been included to the largest extent. Figure 7f shows the distribution of the molar heat capacity. Observably, the 227 space group shown in Figure 7a has some of the highest heat capacity values ranging from 250 to 330 J/mol-K in Figure 7f, followed by space groups 225, 216, and 221 in general. As determined, the closest correlation with the heat capacity is the volume in Figure 7b, whereby the volume scales with the heat capacity for most structures. This is also seen by the Spearman's rank of 0.66 for the volume in Figure 3 owing the highest correlation with the heat capacity out of all the other physical properties listed there. In a sense, the t-SNE plots serve as alternative visual representations of the Spearman's rank. As such, the total weight with rank value of 0.59 is also closely tied to the space group and heat capacity, as observed by the higher total weight in space group 227 structures and 221 toward the lower end. Although the number density and bond length have low Spearman's ranks with the heat capacity, some conclusions could be made from the evaluated structures. For instance, many of the color contours for Figure 7c,d are matching those in the volume figure where notably the number density follows the inverse relationship of the volume, and the bond length follows a direct relationship with the volume and is physically intuitive. Overall, properties such as volume and total weight are easily calculated from structure files and in turn could serve as quick indicators for thermal energy storage materials in which the heat capacity is the dominant characteristic for the performance.

Figure 8 shows the four highest recommended high heat capacities verified by DFT: $MnIn_2Se_4$, Y_2Te_4Pb, KY_2Se_4, and Y_2GeSe_4. We also calculated the heat capacity as a function of temperature. We first notice that all four highest heat capacity materials share the same crystalline structure prototype and symmetry, i.e., space group 227 ($Fd\bar{3}m$) and prototype AB_2C_4. Generally, the heat capacity of all four structures increases when temperature increases, which is well-known by the Debye and Einstein models. More interestingly, we compare the high temperature limit of heat capacity with the historically famous Dulong–Petit limit, which is specified as $3NR$, where N is molar number of atoms in the structure and R is the gas constant (R = 8.31 J/mol-K). For space group 227 and prototype AB_2C_4, n = 14, yielding the Dulong–Petit limit for heat capacity to be 349.2 J/mol-K. For structures Y_2Te_4Pb, KY_2Se_4, and Y_2GeSe_4, their heat capacity approaches the Dulong–Petit limit at high temperatures well above the Debye temperature. This phenomenon is quite understandable from the previous Debye and Einstein models, as well. However, we find an exception that, in Figure 8a, the structure $MnIn_2Se_4$ has an unexpectedly high heat capacity, which is above the Dulong–Petit limit even at room temperature. As temperature continues to increase, the heat capacity of $MnIn_2Se_4$ goes well beyond the Dulong–Petit limit. Such phenomenon is not common in solid crystalline materials.

To gain insight into the mechanism for high heat capacities, we further compare the full phonon dispersions among $MnIn_2Se_4$, Y_2Te_4Pb, KY_2Se_4, and Y_2GeSe_4 in Figure 9. First, no negative or imaginary frequencies were found in any of the phonon dispersions of the four structures. The absence of negative frequencies in the full phonon dispersions in the first Brillouin zone indicates the thermodynamical stability of those structures, which means that these structures could be

https://doi.org/10.1021/acsami.2c11350
ACS Appl. Mater. Interfaces XXXX, XXX, XXX−XXX

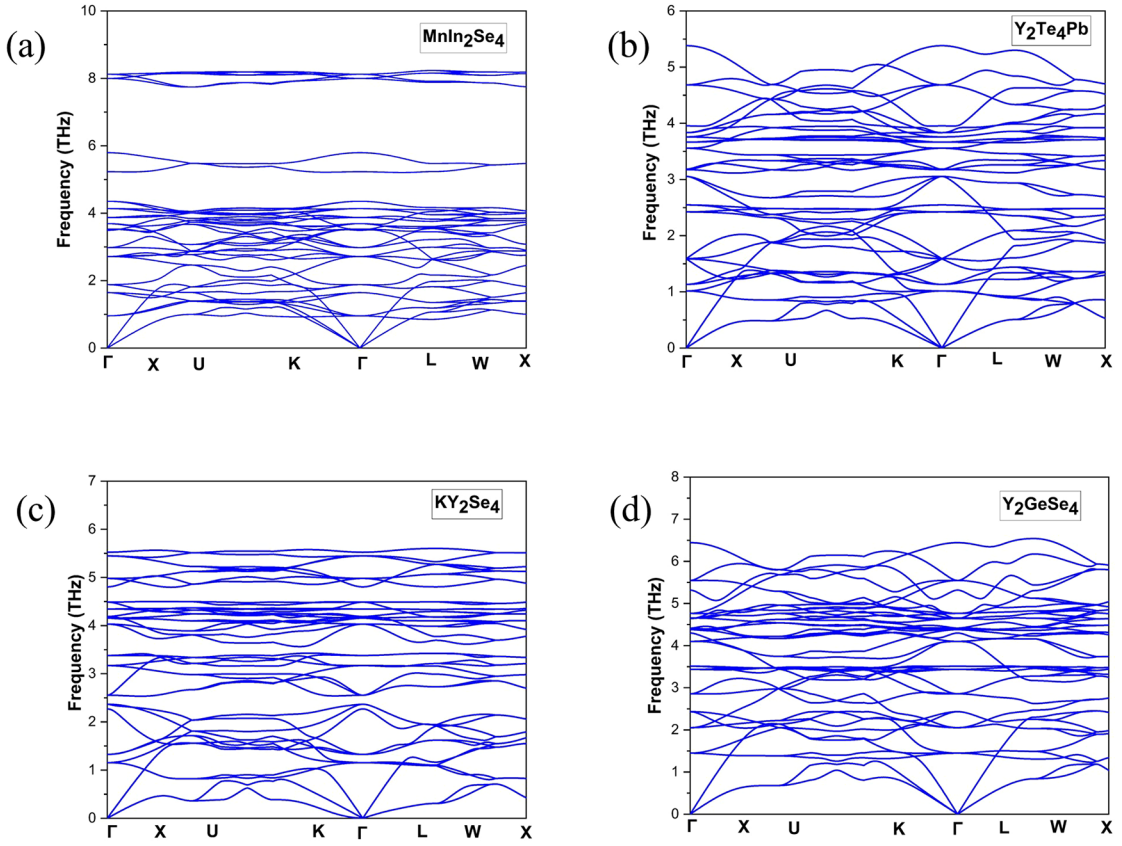


Figure 9. Phonon dispersions of (a) $MnIn_2Se_4$, (b) Y_2Te_4Pb, (c) KY_2Se_4, and (d) Y_2GeSe_4 along high symmetry paths. The non-negative phonon dispersions prove the thermodynamic stability of the structures.

synthesized experimentally provided that the formation energy has also been calculated to be negative for all four materials. Second, the phonon band structures of the four materials look very similar, which is understandable considering that all four materials share the same space group (227) and prototype (AB_2C_4). Third, it is interesting to find that all four structures have relatively low cutoff frequencies in the range of ~5.5−8 THz, which is attributed to the medium to heavy elements in those materials. According to the Bose−Einstein distribution function of phonons, $\langle n \rangle = \frac{1}{e^{\hbar\omega/k_BT} - 1}$, where $\hbar$ is Planck's constant, k_B is the Boltzmann constant, ω is the phonon frequency, and T is the absolute temperature, the corresponding characteristic frequency for the phonon energy comparable to k_BT at room temperature is $\omega^* = \frac{k_BT}{\hbar} = \frac{1.381 \times 10^{-23} \times 300}{6.626 \times 10^{-34}} = 6.25 \times 10^{12}$ s^{-1} = 6.25 THz. This means the phonon modes with a frequency less than 6.25 THz can be excited near room temperature. From the phonon dispersions in Figure 9, we know that almost all phonon modes for all four structures are populated near room temperature. This is the fundamental reason for these four materials having high heat capacities. To study the uniquely high heat capacity of $MnIn_2Se_4$, we plot the frequency-dependent accumulative heat capacity in Figure 10, from which we can see how the heat capacity increases with frequency of phonon modes. As frequency increases to the cutoff frequency (~5−6.5 THz) of the three materials (Y_2Te_4Pb, KY_2Se_4, and Y_2GeSe_4), the heat capacity of $MnIn_2Se_4$ is slightly lower than that for the other three materials. However, since $MnIn_2Se_4$ has a large phonon band gap between 6 and 8 THz (see Figure 9a), the heat capacity of

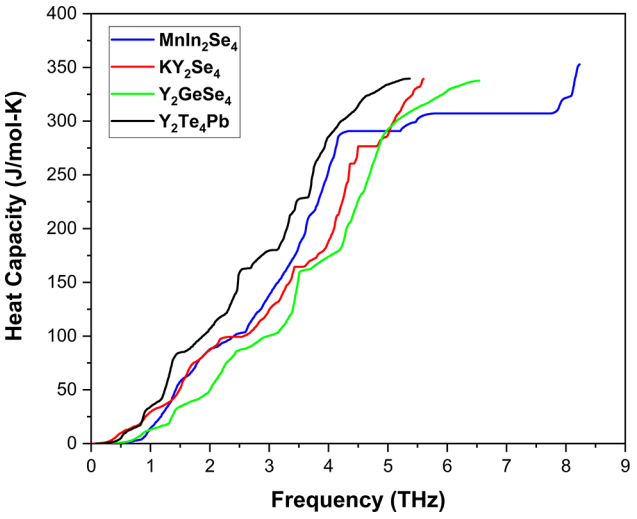


Figure 10. Comparison of frequency-dependent accumulative heat capacity among the four materials shown in Figures 8 and 9.

$MnIn_2Se_4$ increases again (see the last steep increase in Figure 10) as phonon frequencies around 8 THz contribute to the heat capacity. As we analyzed above, those phonon frequencies around 8 THz can still be populated near room temperature, leading to the total heat capacity higher than the Dulong−Petit limit.

https://doi.org/10.1021/acsami.2c11350
ACS Appl. Mater. Interfaces XXXX, XXX, XXX−XXX

CONCLUSIONS

In summary, we trained various machine learning models for screening large-scale structures in search of thermal energy storage materials with the target being high heat capacity. We built four traditional machine learning models and two graph neural network models. Cross-comparison of the prediction performance and model accuracy was conducted among different models. We finally chose the deeperGATGNN model to make a prediction on heat capacity of 32,026 structures taken from the OQMD database, due to its high prediction accuracy as manifested by the low MAE scores. Insight into the correlation between heat capacity and atomic properties such as space group, prototype, lattice volume, etc. were gained by means of the t-SNE plot. The deeperGATGNN model predicted 22 structures that potentially have high heat capacity, and the prediction was further verified by high precision DFT calculations. More interestingly, we found one structure ($MnIn_2Se_4$) exhibits extremely high heat capacity that is even higher than the Dulong–Petit limit at room temperature. Considering the total number of DFT calculations is only 0.07% of the number of all 32,026 structures that have been screened, the approach of combining machine learning and DFT is very promising for accelerating the discovery of novel materials with high efficiency and accuracy.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acsami.2c11350.

Additional details for training data and results for machine learning model testing (ZIP, PDF)

AUTHOR INFORMATION

Corresponding Authors

Chengxiao Peng – *Institute for Computational Materials Science, School of Physics and electronics, Henan University, Kaifeng 475004, People's Republic of China*; Email: pengcx@vip.henu.edu.cn

Ming Hu – *Department of Mechanical Engineering, University of South Carolina, Columbia, South Carolina 29208, United States*; orcid.org/0000-0002-8209-0139; Email: hu@sc.edu

Authors

Joshua Ojih – *Department of Mechanical Engineering, University of South Carolina, Columbia, South Carolina 29208, United States*; orcid.org/0000-0003-0108-8318

Uche Onyekpe – *Department of Computer and Data Science, School of Science, Technology and Health, York St. John University, York YO31 7EX, United Kingdom; Centre for Computational Sciences and Mathematical Modelling, Coventry University, Coventry CV1 5FB, United Kingdom*

Alejandro Rodriguez – *Department of Mechanical Engineering, University of South Carolina, Columbia, South Carolina 29208, United States*

Jianjun Hu – *Department of Computer Science and Engineering, University of South Carolina, Columbia, South Carolina 29208, United States*; orcid.org/0000-0002-8725-6660

Complete contact information is available at: https://pubs.acs.org/10.1021/acsami.2c11350

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

This work was supported in part by the NSF (Award Numbers 1905775, 2030128, and 2110033), NASA SC Space Grant Consortium REAP Program (Award Number 521383-RP-SC004), SC EPSCoR/IDeA Program under NSF OIA-1655740, and an ASPIRE grant from the Office of the Vice President for Research at the University of South Carolina (Project 80005046).

REFERENCES

(1) Liu, J.; Chang, Z.; Wang, L.; Xu, J.; Kuang, R.; Wu, Z. Exploration of Basalt Glasses as High-Temperature Sensible Heat Storage Materials. *ACS Omega* **2020**, *5* (30), 19236–19246.

(2) Hasnain, S. M. Review on Sustainable Thermal Energy Storage Technologies, Part I: Heat Storage Materials and Techniques. *Energy Convers. Manag.* **1998**, *39* (11), 1127–1138.

(3) Kuravi, S.; Trahan, J.; Goswami, D. Y.; Rahman, M. M.; Stefanakos, E. K. Thermal Energy Storage Technologies and Systems for Concentrating Solar Power Plants. *Prog. Energy Combust. Sci.* **2013**, *39* (4), 285–319.

(4) Samala, S.; Brahma, G. S.; Swain, T. Synthesis and Characterization of Sensible Thermal Heat Storage Mixture Containing Phosphate Compound of Cobalt and Sodium. *Sol. Energy* **2019**, *177*, 612–619.

(5) Konuklu, Y.; Ostry, M.; Paksoy, H. O.; Charvat, P. Review on Using Microencapsulated Phase Change Materials (PCM) in Building Applications. *Energy Build.* **2015**, *106*, 134–155.

(6) Nomura, T.; Sheng, N.; Zhu, C.; Saito, G.; Hanzaki, D.; Hiraki, T.; Akiyama, T. Microencapsulated Phase Change Materials with High Heat Capacity and High Cyclic Durability for High-Temperature Thermal Energy Storage and Transportation. *Appl. Energy* **2017**, *188*, 9–18.

(7) Chibani, S.; Coudert, F. X. Machine Learning Approaches for the Prediction of Materials Properties. *APL Mater.* **2020**, *8* (8), 080701.

(8) Ojih, J.; Al-fahdi, M.; Rodriguez, A. D.; Choudhary, K. Efficiently Searching Extreme Mechanical Properties via Boundless Objective-Free Exploration and Minimal First-Principles Calculations. *npg Comput. Mater.* **2022**, 1–12.

(9) 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.* **2020**, *60* (4), 1996–2003.

(10) Kauwe, S. K.; Graser, J.; Vazquez, A.; Sparks, T. D. Machine Learning Prediction of Heat Capacity for Solid Inorganics. *Integr. Mater. Manuf. Innov.* **2018**, *7* (2), 43–51.

(11) Chatterjee, S.; Murugananth, M.; Bhadeshia, H. K. D. H. *δ* TRIP Steel. *Mater. Sci. Technol.* **2007**, *23* (7), 819–827.

(12) Bhadeshia, H. K. D. H.; Dimitriu, R. C.; Forsik, S.; Pak, J. H.; Ryu, J. H. Performance of Neural Networks in Materials Science. *Mater. Sci. Technol.* **2009**, *25* (4), 504–510.

(13) Pilania, G.; Mannodi-Kanakkithodi, A.; Uberuaga, B. P.; Ramprasad, R.; Gubernatis, J. E.; Lookman, T. Machine Learning Bandgaps of Double Perovskites. *Sci. Rep.* **2016**, *6*, 19375.

(14) Dey, P.; Bible, J.; Datta, S.; Broderick, S.; Jasinski, J.; Sunkara, M.; Menon, M.; Rajan, K. Informatics-Aided Bandgap Engineering for Solar Materials. *Comput. Mater. Sci.* **2014**, *83*, 185–195.

(15) Ghiringhelli, L. M.; Vybiral, J.; Levchenko, S. V.; Draxl, C.; Scheffler, M. Big Data of Materials Science: Critical Role of the Descriptor. *Phys. Rev. Lett.* **2015**, *114* (10), 1–5.

(16) Meredig, B.; Agrawal, A.; Kirklin, S.; Saal, J. E.; Doak, J. W.; Thompson, A.; Zhang, K.; Choudhary, A.; Wolverton, C. Combinatorial Screening for New Materials in Unconstrained Composition Space with Machine Learning. *Phys. Rev. B - Condens. Matter Mater. Phys.* **2014**, *89* (9), 1–7.

(17) Curtarolo, S.; Morgan, D.; Persson, K.; Rodgers, J.; Ceder, G. Predicting Crystal Structures with Data Mining of Quantum Calculations. *Phys. Rev. Lett.* **2003**, *91* (13), 1−4.
(18) Odigwe, B. E.; Carolina, U. S.; Carolina, U. S.; Odigwe, C. I.; Spinale, F. G. Application of Machine Learning for Patient Response Prediction to Cardiac Resynchronization Therapy. *Proceedings of the 13th ACM Int. Conf.* **2022**, 47.
(19) Fan, J.; Han, F.; Liu, H. Challenges of Big Data Analysis. *Natl. Sci. Rev.* **2014**, *1* (2), 293−314.
(20) Keith, J. A.; Vassilev-Galindo, V.; Cheng, B.; Chmiela, S.; Gastegger, M.; Müller, K. R.; Tkatchenko, A. Combining Machine Learning and Computational Chemistry for Predictive Insights into Chemical Systems. *Chem. Rev.* **2021**, *121* (16), 9816−9872.
(21) Odigwe, B. E.; Spinale, F. G.; Valafar, H. Application of Machine Learning in Early Recommendation of Cardiac Resynchronization Therapy. *arXiv* **2021**, DOI: 10.48550/arXiv.2109.06139.
(22) Bietti, A.; Mairal, J. On the Inductive Bias of Neural Tangent Kernels. *Adv. Neural Inf. Process. Syst.* 2019; https://proceedings.neurips.cc/paper/2019/file/c4ef9c39b300931b69a36fb3dbb8d60e-Paper.pdf (accessed 2022-09-12).
(23) Pun, G. P. P.; Batra, R.; Ramprasad, R.; Mishin, Y. Physically Informed Artificial Neural Networks for Atomistic Modeling of Materials. *Nat. Commun.* **2019**, *10* (1), 1−10.
(24) Seko, A.; Maekawa, T.; Tsuda, K.; Tanaka, I. Machine Learning with Systematic Density-Functional Theory Calculations: Application to Melting Temperatures of Single- and Binary-Component Solids. *Phys. Rev. B - Condens. Matter Mater. Phys.* **2014**, *89* (5), 1−9.
(25) Kirklin, S.; Saal, J. E.; Meredig, B.; Thompson, A.; Doak, J. W.; Aykol, M.; Rühl, S.; Wolverton, C. The Open Quantum Materials Database (OQMD): Assessing the Accuracy of DFT Formation Energies. *npj Comput. Mater.* **2015**, *1*, 15010.
(26) Saal, J. E.; Kirklin, S.; Aykol, M.; Meredig, B.; Wolverton, C. Materials Design and Discovery with High-Throughput Density Functional Theory: The Open Quantum Materials Database (OQMD). *Jom* **2013**, *65* (11), 1501−1509.
(27) Blöchl, P. E. Projector Augmented-Wave Method. *Phys. Rev. B* **1994**, *50* (24), 17953−17979.
(28) Perdew, J. P.; Burke, K.; Ernzerhof, M. Generalized Gradient Approximation Made Simple. *Phys. Rev. Lett.* **1996**, *77* (18), 3865−3868.
(29) Kresse, G.; Hafner, J. Ab Initio Molecular Dynamics for Liquid Metals. *Phys. Rev. B* **1993**, *47* (1), 558−561.
(30) Kresse, G.; Joubert, D. From Ultrasoft Pseudopotentials to the Projector Augmented-Wave Method. *Phys. Rev. B - Condens. Matter Mater. Phys.* **1999**, *59* (3), 1758−1775.
(31) Vargas-Hernández, R. A. Bayesian Optimization for Calibrating and Selecting Hybrid-Density Functional Models. *J. Phys. Chem. A* **2020**, *124* (20), 4053−4061.
(32) Togo, A.; Tanaka, I. First Principles Phonon Calculations in Materials Science. *Scr. Mater.* **2015**, *108*, 1−5.
(33) Zhou, F.; Nielson, W.; Xia, Y.; Ozoliņš, V. Compressive Sensing Lattice Dynamics. I. General Formalism. *Phys. Rev. B* **2019**, *100* (18), 1−15.
(34) Li, W.; Carrete, J.; Katcho, N. A.; Mingo, N. ShengBTE: A Solver of the Boltzmann Transport Equation for Phonons. *Comput. Phys. Commun.* **2014**, *185* (6), 1747−1758.
(35) Yue, S. Y.; Qin, G.; Zhang, X.; Sheng, X.; Su, G.; Hu, M. Thermal Transport in Novel Carbon Allotropes with Sp2 or Sp3 Hybridization: An Ab Initio Study. *Phys. Rev. B* **2017**, *95* (8), 1−11.
(36) Liu, T. Z.; Gall, D.; Khare, S. V. Electronic and Bonding Analysis of Hardness in Pyrite-Type Transition-Metal Pernitrides. *Phys. Rev. B* **2014**, 134102.
(37) Kalipe, G.; Gautham, V.; Behera, R. K. Predicting Malarial Outbreak Using Machine Learning and Deep Learning Approach: A Review and Analysis. *Proc. - 2018 Int. Conf. Inf. Technol. ICIT* **2018**, 33−38.
(38) Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45* (1), 5−32.
(39) Breiman, L. Bagging Predictors. *Mach. Learn.* **1996**, *24* (2), 123−140.
(40) Zheng, C.; Chen, C.; Chen, Y.; Ong, S. P. Random Forest Models for Accurate Identification of Coordination Environments from X-Ray Absorption Near-Edge Structure. *Patterns* **2020**, *1* (2), 100013.
(41) Jaafreh, R.; Kang, Y. S.; Hamad, K. Lattice Thermal Conductivity: An Accelerated Discovery Guided by Machine Learning. *ACS Appl. Mater. Interfaces* **2021**, *13* (48), 57204−57213.
(42) Schapire, R. E. A Brief Introduction to Boosting. *IJCAI Int. Jt. Conf. Artif. Intell.* **1999**, *2*, 1401−1406.
(43) Opitz, D.; Maclin, R.; Bauer, E.; Chan, P.; Stolfo, S.; Wolpert, D.; Hussain, S.; Mustafa, M. W.; Jumani, T. A.; Baloch, S. K.; Alotaibi, H.; Khan, I.; Khan, A.; Zhang, Y.; Zhao, Z.; Zheng, J.; Prokhorenkova, L.; Gusev, G.; Vorobev, A.; Dorogush, A. V.; Gulin, A.; Nobre, J.; Neves, R. F.; Schapire, R. E.; Friedman, J. H.; Hastie, T.; Tibshirani, R.; Chen, T.; Guestrin, C.; Breiman, L.; Kalipe, G.; Gautham, V.; Behera, R. K.; Scikit, L.; Pedregosa, F.; Varoquaux, G.; G, A.; M, V.; T, B.; G, O.; B, M.; P, P.; W, R.; D, V.; V, J.; P, A.; C, D.; B, M.; P, M.; D, E. Predicting Malarial Outbreak Using Machine Learning and Deep Learning Approach: A Review and Analysis. *Proc. - 2018 Int. Conf. Inf. Technol. ICIT* **1999**, *24* (2), 6638−6648.
(44) Opitz, D.; Maclin, R. Popular Ensemble Methods: An Empirical Study. *J. Artif. Intell. Res.* **1999**, *11*, 169−198.
(45) Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Min.* **2016**, 785−794.
(46) Friedman, J.; Hastie, T.; Tibshirani, R. Additive Logistic Regression: A Statistical View of Boosting (With Discussion and a Rejoinder by the Authors). *Ann. Statist.* **2000**, *28* (2), 337−407.
(47) Friedman, J. H. Greedy Function Approximation: A Gradient Boosting Machine. *Ann. Statist.* **2001**, *29* (5), 1189−1232.
(48) Nobre, J.; Neves, R. F. Combining Principal Component Analysis, Discrete Wavelet Transform and XGBoost to Trade in the Financial Markets. *Expert Syst. Appl.* **2019**, *125*, 181−194.
(49) Hussain, S.; Mustafa, M. W.; Jumani, T. A.; Baloch, S. K.; Alotaibi, H.; Khan, I.; Khan, A. A Novel Feature Engineered-CatBoost-Based Supervised Machine Learning Framework for Electricity Theft Detection. *Energy Reports* **2021**, *7*, 4425−4436.
(50) Zhang, Y.; Zhao, Z.; Zheng, J. CatBoost: A New Approach for Estimating Daily Reference Crop Evapotranspiration in Arid and Semi-Arid Regions of Northern China. *J. Hydrol.* **2020**, *588*, 125087.
(51) Prokhorenkova, L.; Gusev, G.; Vorobev, A.; Dorogush, A. V.; Gulin, A. CatBoost: Unbiased Boosting with Categorical Features. *Adv. Neural Inf. Process. Syst.* **2017**, 6638−6648.
(52) Xie, T.; Grossman, J. C. Crystal Graph Convolutional Neural Networks for an Accurate and Interpretable Prediction of Material Properties. *Phys. Rev. Lett.* **2018**, *120* (14), 145301.
(53) Omee, S. S.; Louis, S.-Y.; Fu, N.; Wei, L.; Dey, S.; Dong, R.; Li, Q.; Hu, J. Scalable Deeper Graph Neural Networks for High-Performance Materials Property Prediction. *Patterns* **2022**, *3* (5), 100491.
(54) Buitinck, L.; Louppe, G.; Blondel, M.; Pedregosa, F.; Mueller, A.; Grisel, O.; Niculae, V.; Prettenhofer, P.; Gramfort, A.; Grobler, J.; Layton, R.; Vanderplas, J.; Joly, A.; Holt, B.; Varoquaux, G. API Design for Machine Learning Software: Experiences from the Scikit-Learn Project. *arXiv* **2013**, 1−15.
(55) Pedregosa, F.; Varoquaux, G.; G, A.; M, V.; T, B.; G, O.; B, M.; P, P.; W, R.; D, V.; V, J.; P, A.; C, D.; B, M.; P, M.; D, E. Scikit-Learn: Machine Learning in {P}ython. *J. Mach. Learn. Res.* **2011**, *12*, 2825−2830.
(56) van der Maaten, L.; Hinton, G. Visualizing Data using t-SNE. *Journal of Machine Learning Research* **2008**, *9* (86), 2579−2605.
(57) Rodriguez, A.; Liu, Y.; Hu, M. Spatial Density Neural Network Force Fields with First-Principles Level Accuracy and Application to Thermal Transport. *Phys. Rev. B* **2020**, *102* (3), 35203.
(58) Palik, E. D. Potassium Chloride (KCl). *Handb. Opt. Constants Solids* **1985**, *1*, 703−718.
(59) Sarbu, I.; Sebarchievici, C. A Comprehensive Review of Thermal Energy Storage. *Sustainability* **2018**, *10* (1), 191.

https://doi.org/10.1021/acsami.2c11350
ACS Appl. Mater. Interfaces XXXX, XXX, XXX−XXX

(60) Crystran Ltd. Zinc Sulphide FLIR (ZnS); https://www.crystran.co.uk/optical-materials/zinc-sulphide-flir-zinc-sulfide-zns (accessed 2022-03-10).
(61) Mateck, Lithium Fluoride (LiF); https://mateck.com/info/lithium-fluoride-lif.html (accessed 2022-03-10).
(62) Matweb, Zinc Oxide (ZnO), Cubic Categories; https://www.matweb.com/search/datasheet.aspx?matguid=173a8f1e7cec4ce7af5dc3b90d10f756&ckck=1 (accessed 2022-03-10).
(63) Specific heat; https://www.iue.tuwien.ac.at/phd/palankovski/node35.html (accessed 2022-03-10).
(64) Material : Aluminum Nitride (AlN), Bulk. 24; https://www.memsnet.org/material/aluminumnitridealnbulk/ (accessed 2022-03-10).
(65) Crystran Ltd. Sodium Fluoride (NaF); https://www.crystran.co.uk/optical-materials/sodium-fluoride-naf (accessed 2022-03-10).
(66) El-Sharkawy, A. A.; Abou El-Azm, A. M.; Kenawy, M. I.; Hillal, A. S.; Abu-Basha, H. M. Thermophysical Properties of Polycrystalline PbS, PbSe, and PbTe in the Temperature Range 300−700 K. *Int. J. Thermophys.* **1983**, *4* (3), 261−269.
(67) Potassium Iodide (KI) Potassium Iodide (KI) Mm. *Pubchem* 1980, *84*, 78−79; https://pubchem.ncbi.nlm.nih.gov/compound/Potassium-iodide (accessed 2022-09-12).
(68) Felderhoff, M.; Bogdanović, B. High Temperature Metal Hydrides as Heat Storage Materials for Solar and Related Applications. *Int. J. Mol. Sci.* **2009**, *10* (1), 325−344.

https://doi.org/10.1021/acsami.2c11350
ACS Appl. Mater. Interfaces XXXX, XXX, XXX−XXX