

## MATERIALS SCIENCE

## Leverage electron properties to predict phonon properties via transfer learning for semiconductors

Zeyu Liu<sup>1</sup>, Meng Jiang<sup>2</sup>, Tengfei Luo<sup>1,3\*</sup>

Electron properties are usually easier to obtain than phonon properties. The ability to leverage electron properties to help predict phonon properties can thus greatly benefit materials by design for applications like thermoelectrics and electronics. Here, we demonstrate the ability of using transfer learning (TL), where knowledge learned from training machine learning models on electronic bandgaps of 1245 semiconductors is transferred to improve the models, trained using only 124 data, for predicting various phonon properties (phonon bandgap, group velocity, and heat capacity). Compared to directly trained models, TL reduces the mean absolute errors of prediction by 65, 14, and 54% respectively, for the three phonon properties. The TL models are further validated using several semiconductors outside of the 1245 database. Results also indicate that TL can leverage not-so-accurate proxy properties, as long as they encode composition-property relation, to improve models for target properties, a notable feature to materials informatics in general.

## INTRODUCTION

For metals, it is well known that electrical and thermal transport properties are directly connected in a linear relationship governed by the Wiedemann-Franz law (1). This is simply because free electrons are the common carriers for both electrical conduction and heat transfer in metals. However, there is no such universal relation for semiconductors since these two types of transport are respectively dominated by electrons (or holes) and phonons (2). From the first-principles theory, we know that the electron states of semiconductors are determined by the ground state charge density and that the phonon states depend on both the ground state charge density and its linear response to the atomic displacement (3). In other words, electron and phonon properties are inherently connected, but the relationship is much more complicated than that seen in metals, and no analytical formula currently exists. The ability to leverage electron properties to predict phonon properties will be enormously impactful because obtaining electron properties, either through calculation or measurement, is much easier than for phonons. For example, calculating electron band structure only needs one self-consistent field calculation of the primitive cell in the density functional theory (DFT) framework, taking merely a few seconds for materials like silicon. However, while proven to be accurate in predicting phonon properties (4, 5), first-principles calculation of the phonon band structure (i.e., dispersion relation) needs several much more time-consuming density functional perturbation theory (DFPT) calculations on discrete points in the first Brillouin zone (4) or many slow self-consistent field calculations for large supercells (5), taking at least two orders of magnitude longer than electron band structure calculations. On the experimental side, electron band structures can be measured by angle-resolved photoemission spectroscopy (6), while the much more sophisticated inelastic neutron scattering is needed to obtain phonon dispersion relations (7).

Multiobjective optimization simultaneously considering electron and phonon properties are critically important for eventually achieving

materials by design in a variety of applications, such as thermoelectrics and wide-bandgap (WBG) semiconductors. In thermoelectrics, increasing electron properties (e.g., electrical conductivity and Seebeck coefficient) and decreasing phonon thermal conductivity at the same time are needed to improve the figure of merit,  $ZT$  (8, 9). For power electronics, developed WBG semiconductors are always preferred to have superior phonon transport to ease the thermal management challenge. However, because of the substantial disparity in difficulties for quantifying electron and phonon properties, electrical design has usually taken the priority, while phonon properties come secondary. One such example is  $\beta$ -Ga<sub>2</sub>O<sub>3</sub>, which has an optimal electronic bandgap of 4.7 eV (10), ideal for the WBG applications, but it has the lowest phonon thermal conductivity among its peers (<25 W/mK) (11, 12). In contrast, GaN, the state-of-the-art WBG material, has a value of around 230 W/mK (13, 14).

Data-driven machine learning (ML) techniques have emerged as powerful tools for materials design when simple analytical principles are not established between material compositions and properties (15). ML has been used to predict fundamental properties of materials such as electronic bandgap (16–20), formation energy (21–23), and thermodynamic properties (24, 25). However, since electron properties are easier to obtain and have usually received more attention than phonon properties, the volume of data for the former is much greater than the latter in well-established databases like the Open Quantum Materials Database (26), Automatic Flow for Materials Discovery Library (27), and the Materials Project (MP) (28). Petretto *et al.* (29) recently added phonon dispersion calculated from first-principles DFPT (3) for around 1500 semiconductors to the MP database, which has already included more than 50,000 electron band structure data. In addition, if we constrain the materials design space, like the problem studied in the present work, then the amount of available phonon data will further decrease, imposing a central challenge to constructing high-fidelity ML models (30).

Transfer learning (TL) can be a useful ML technique to overcome the problem of small data (31), and it has been successfully applied in a few materials informatics studies (23, 32–34). It can leverage knowledge learned from training with a larger dataset of one or multiple proxy properties and transfer it to improve the ML model performance for a target property with a much smaller dataset.

Copyright © 2020  
The Authors, some  
rights reserved;  
exclusive licensee  
American Association  
for the Advancement  
of Science. No claim to  
original U.S. Government  
Works. Distributed  
under a Creative  
Commons Attribution  
NonCommercial  
License 4.0 (CC BY-NC).

<sup>1</sup>Department of Aerospace and Mechanical Engineering, University of Notre Dame, Notre Dame, IN 46556, USA. <sup>2</sup>Department of Computer Science and Engineering, University of Notre Dame, Notre Dame, IN 46556, USA. <sup>3</sup>Department of Chemical and Biomolecular Engineering, University of Notre Dame, Notre Dame, IN 46556, USA. \*Corresponding author. Email: tluo@nd.edu

Considering the fact that phonon modes are inherently linked to electron structure, transferring the knowledge from ML of electron structure data may help us better predict phonon properties using TL.

In this study, we demonstrate the possibility of leveraging electron properties to predict phonon properties in semiconductors using TL. We choose one of the most accessible electron properties, bandgap, as the proxy property, and transfer the knowledge learned to improve the prediction of an important phonon property, the frequency gap between the acoustic and optical phonon modes. For brevity, we call such a frequency gap the phonon bandgap. The existence of such a phonon bandgap can reduce the possibility of combining two low-frequency acoustic phonons into one higher-frequency optical phonon in the three-phonon scattering process, the main scattering mechanism for phonon scattering at room temperature, due to the conservation of energy (35). Since acoustic phonons usually dominate the thermal transport in semiconductors (36), this reduced scattering can lead to high phonon thermal conductivity, and this has been recently observed for materials like BAs (37–41) and hydrogenated silicene (42).

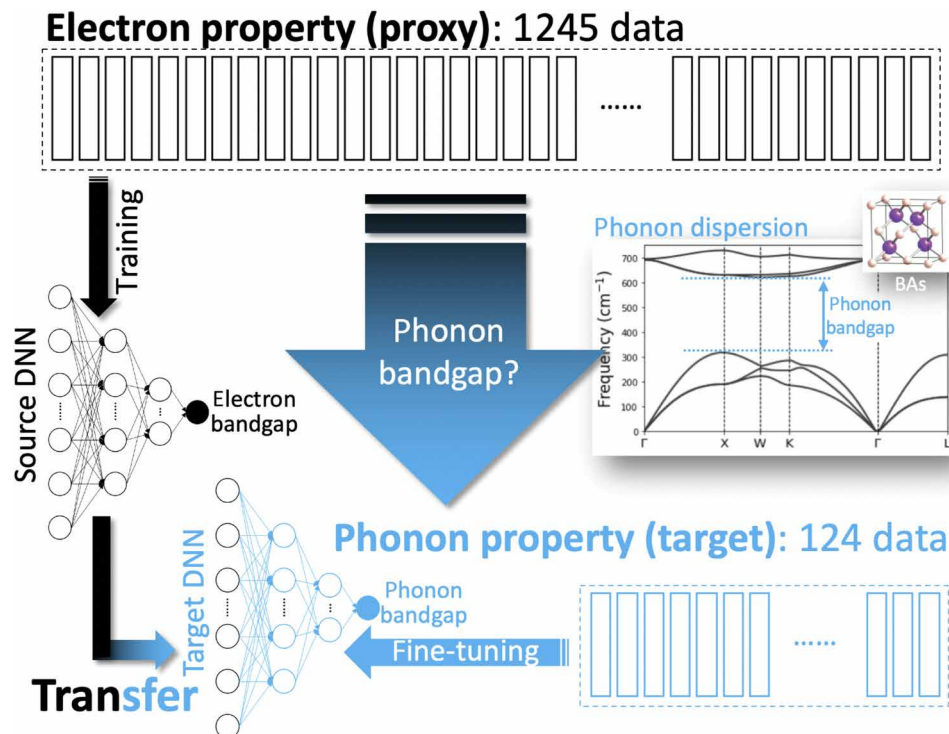
For the semiconductors that are labeled with phonon band information in the MP database (1245 data), we first down-select those semiconductors that have a phonon bandgap, reducing the number of viable phonon data down to 124. We then build a classification model to determine whether a new semiconductor is going to have a phonon bandgap. A deep neural network (DNN) linking material composition to electronic bandgap is first trained against the whole 1245 dataset, and then, TL is applied to transfer the electronic bandgap DNN to facilitate the construction of the DNN for phonon bandgap,

which is trained on the small dataset of 124 points (see Fig. 1 for complete work flow). Our results show that, even with very limited number of viable phonon data, using the knowledge learned from electronic bandgap, the phonon bandgap can be predicted with very high accuracy. The mean absolute error (MAE) of the DNN is reduced substantially by TL from 23.847 to 8.458  $\text{cm}^{-1}$  of the directly trained DNN. Last, this TL approach is extended to other phonon properties like phonon speed of sound and heat capacity. Improvement in model prediction can also be achieved in these properties with TL reducing the MAE by 15 and 54%. While the DFT-calculated electronic bandgaps are known to be underestimated because of the limitations of DFT, such systematic error in the proxy label seems to have no impact on the prediction capability on phonon properties using our TL scheme. The demonstrated success of TL from electron property to phonon property may have notable impact to materials development for a wide range of applications.

## RESULTS

### Datasets

The data used in this work are from the MP database contributed by Petretto *et al.* (29), where phonon dispersion relations for  $\sim 1500$  semiconductors were calculated by diagonalizing the dynamical matrix in the whole first Brillouin zone based on the second-order force constants calculated from the first-principles DFPT method. The DFPT-calculated phonon dispersion is of relatively high accuracy (29), which can be directly used for in silico materials design. We have removed materials with imaginary phonon frequencies in this



**Fig. 1. Schematic of TL from electron property to phonon properties.** A total of 1245 electronic bandgap data of semiconductors that have phonon information in the MP database are used as the proxy property in the source task, while the constrain on semiconductor to have a phonon bandgap [e.g., boron arsenide (BAs) dispersion in the inset] reduces the data for the target property down to 124. A DNN is pretrained on the electronic bandgap, and its architecture and parameters are transferred to the target task, where the DNN is further fine-tuned using the small phonon bandgap data.

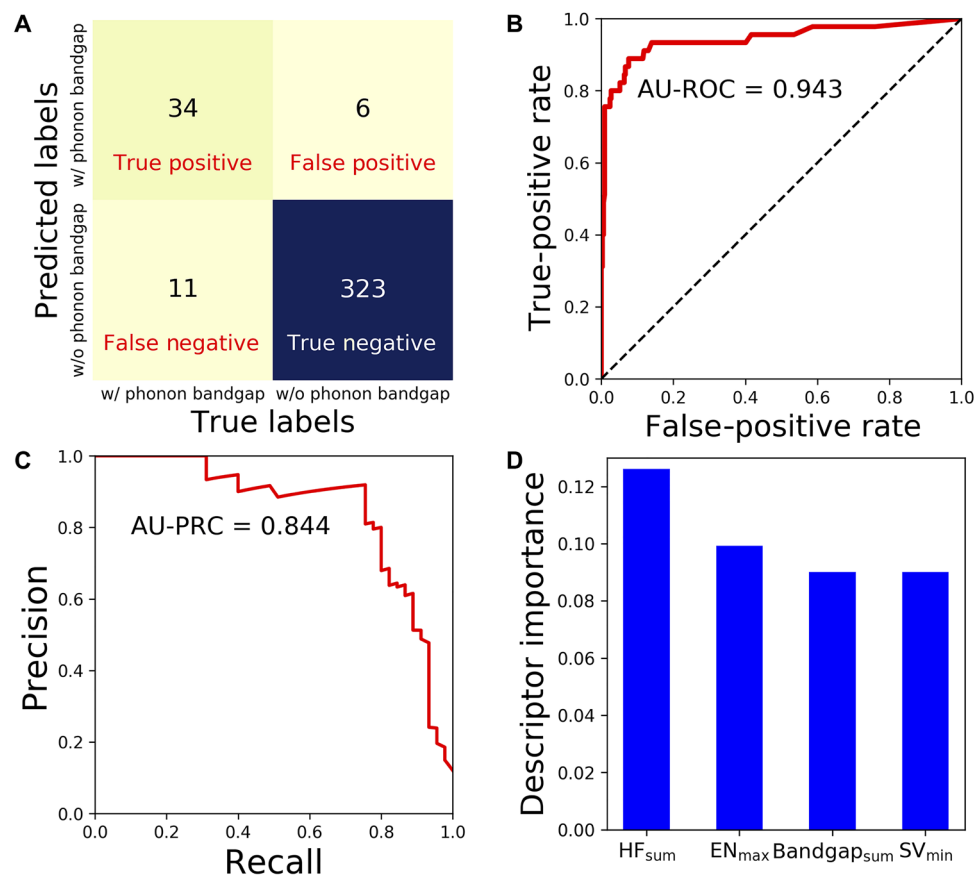
study because of their dynamical instability, leaving a total of 1245 phonon data. Among these materials, we further extract those that have a phonon bandgap (i.e., target property), reducing the amount of viable data to 124, merely 10% of the proxy data. Electronic bandgap, which is the proxy property, calculated by the DFT is also collected from the MP database (28) for all the 1245 semiconductors. The semiconductor materials are represented with fixed length vectors (see Materials and Methods). The data distributions are included in section S1, which indicate that there are no linear correlations between the proxy and target properties and they also have their own unique distributions.

### Classification model

Before predicting the phonon bandgap of a given compound, we first have to be able to tell whether it is going to have a phonon bandgap. We construct a classification model to achieve this capability. A random forest classifier (43) is trained to identify the 124 semiconductors that have phonon bandgap from a total of 1245 compounds using scikit-learn (44). The number of materials with phonon bandgap is much smaller than those without, imposing a fundamental obstacle for accurate classification. SMOTE (45), a systematic over-sampling method, for the phonon bandgap materials (positive label) implemented in imblearn (46) is used to overcome the challenge in this highly imbalanced classification problem. More details of the

model and definition of terminologies discussed below are included in Materials and Methods.

The overall classification accuracy is found to be 95.5% on the testing data, and the F1 score for the positive label is 0.800. The confusion matrix for the testing data is shown in Fig. 2A to visualize the model classification ability. Most of the testing data are correctly classified into their own group, as characterized by the diagonal components in the confusion matrix (true positive and true negative). The scenarios of materials with phonon bandgap misclassified as without phonon bandgap (false negative) and the materials without phonon bandgap misclassified as with phonon bandgap (false positive) are shown as the off-diagonal components in Fig. 2A. The receiver operating characteristic (ROC) curve for the testing data is presented in Fig. 2B, which shows the relation between the true-positive rate and the false-positive rate. The better the classification model is, the more the ROC curve will be concentrated on the upper left corner. The area under the ROC curve (AU-ROC) is 0.5 for a random guess, and AU-ROC is 1 for a perfect classification. Our ROC curve is close to the upper left corner, and the high AU-ROC score of 0.943 indicates a relatively good classification performance. For an imbalanced classification like the one we have here, the precision-recall curve (PRC) shown in Fig. 2C is also of importance. This plots the relation between precision and recall. Similar to the ROC curve, the better the classification model is, the more the PRC will be



**Fig. 2. Classification model performance evaluated on the testing data.** (A) The confusion matrix illustrating the number of true positive, false positive, false negative, and true negative. (B) The ROC curve. (C) The precision-recall curve (PRC). (D) The top four most important descriptors identified from the random forest model. HF<sub>sum</sub>, the sum of the heat of fusion of the compound elements; EN<sub>max</sub>, the maximum value of electronegativity of the compound elements; Bandgap<sub>sum</sub>, the sum of the ground state bandgap of the compound elements; SV<sub>min</sub>, the minimum value of the sound velocity of the compound elements.

concentrated on the upper right corner. The area under the PRC (AU-PRC) is 0 for a random guess and 1 for a perfect classification model. Our AU-PRC is calculated to be 0.844, indicating that this model can separate materials with or without phonon bandgap reasonably well even for this highly imbalanced classification problem.

One advantage of using decision tree-based models like random forest is their good model explainability (47). We use this model capability to visualize the descriptor importance in Fig. 2D. The four most important descriptors for this classification model of phonon bandgap are shown, and they are the sum of the heat of fusion of the compound elements, the maximum value of elemental electronegativity, the sum of the ground state electronic bandgap of the compound elements, and the minimum value of the sound velocity of the compound elements. Although the connection between a compound property (e.g., bandgap of  $\text{Ga}_2\text{O}_3$ ) and the elemental property (e.g., bandgaps of elemental Ga and O) is not straightforward, it is still interesting to see that a classification model on a phonon property is connected to not only elemental phonon properties (e.g., sound velocity, group velocity of low-frequency phonons) but also elemental electron properties (e.g., electronic bandgap). This indicates that from this data-driven result, some connection between electron and phonon properties is suggested. It is also understandable that the sum of the heat of fusion of the compound elements and the maximum value of elemental electronegativity turn out to be the most important descriptors, as they can respectively be linked to bond strength and ionicity, both of which can directly influence phonon band structure.

### DNN model for proxy property

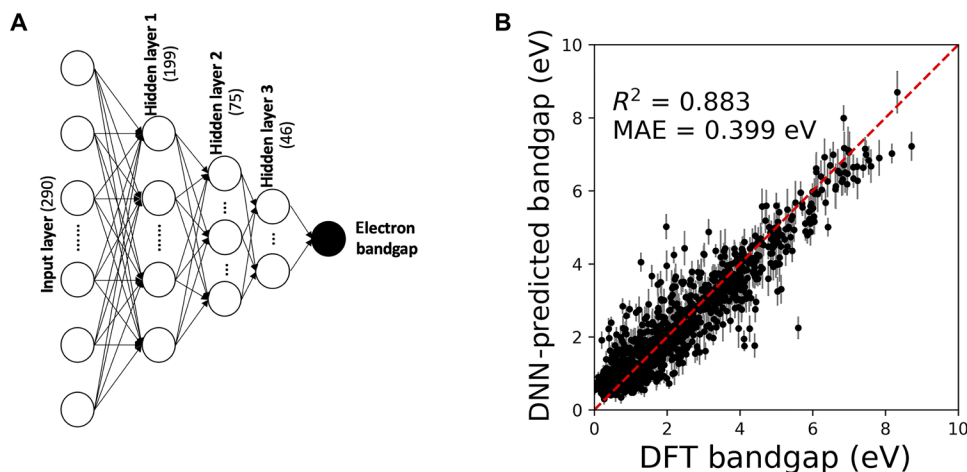
A total of 250 multilayer perceptron (MLP) DNN models with different numbers of nodes in three hidden layers are first trained on the source task of electronic bandgap (i.e., proxy property) prediction on the whole dataset for the 1245 semiconductors. The averaged MAE is 0.442 eV, and the averaged coefficient of determination ( $R^2$ ) is 0.860 for these 250 different pretrained models, which are comparable to other ML models for electronic bandgap of inorganic semiconductors (20, 22). The SD of MAE and  $R^2$  for these 250 different pretrained models are 0.012 eV and 0.007, respectively. The trained model structure with the lowest MAE is visualized in Fig. 3A, and the comparison of ML-predicted electronic bandgap against the

ground truth for this model is shown in Fig. 3B. Parameters (weight and bias for each hidden layer) in these 250 pretrained models are stored for further TL for phonon properties.

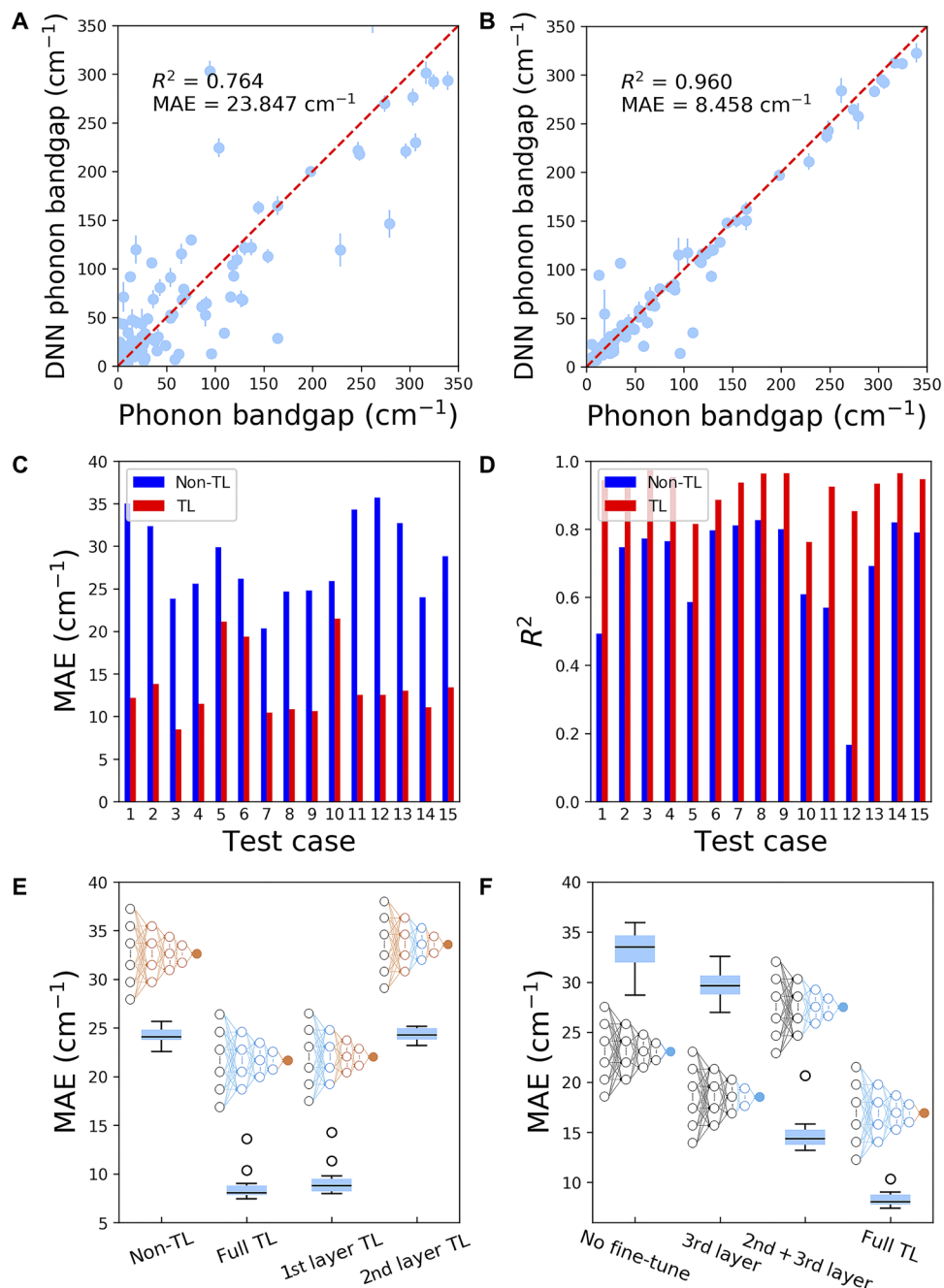
### TL model for phonon properties

We then train 250 MLP models using the TL scheme for phonon bandgap with the help of the 250 pretrained models (see Materials and Methods). Accurate prediction of phonon bandgap for the 124 compounds with phonon bandgap is achieved using this TL scheme, despite very limited available data. For the TL model with the best MAE performance, compared to the DNN model directly trained using the 124 phonon data (Fig. 4A) with the same architecture, the TL scheme significantly reduces the MAE by over 60% from 23.874 to 8.458  $\text{cm}^{-1}$ , and  $R^2$  also sees a major improvement from 0.764 to 0.960 (Fig. 4B). To ensure that the superior prediction of the TL model on phonon bandgap is not due to overfitting, we used an additional approach of performance evaluation, where we hide 40% of the data as the testing dataset and the model is only trained on the remaining 60% data. These random dataset splits are conducted 15 times, and both the TL and non-TL models are trained and evaluated against the same dataset splits to make a fair comparison. With even smaller dataset available for training, high prediction accuracy with an averaged MAE of 13.217  $\text{cm}^{-1}$  and an averaged  $R^2$  of 0.928 can still be achieved on the test set for the TL model, and it can outperform the non-TL model in every single case tested in both MAE and  $R^2$ , as illustrated in Fig. 4 (C and D). This confirms that the high accuracy of the TL model is not a result of overfitting. In addition, as can be seen from Fig. 4D, there are some cases with very low  $R^2$ , and one of them even has  $R^2 < 0.2$  from direct training (non-TL). The TL scheme is able to improve these cases markedly, so their  $R^2$  are similar to those other cases. In another words, TL can eliminate low and nonpredictive models, which is especially useful and important when the given small data cannot yield a meaningful model.

To further ensure the validity of our TL models, we conducted two more experiments. We first take different levels of TL by evaluating the cases where the pretrained parameters are only used for the first or the second hidden layer, leaving all other parameters initialized randomly, which are respectively denoted as the “1st layer TL” and “2nd layer TL” cases in Fig. 4E (see insets for illustration). It can be



**Fig. 3. DNN model for proxy property.** (A) The MLP DNN architecture of the best-performing model for electronic bandgap. In the parentheses are the numbers of neurons in each layer. (B) DNN-predicted electronic bandgap versus DFT-calculated bandgap (ground truth) for this model.



**Fig. 4. TL model for phonon bandgap.** Predicted phonon bandgap versus DFPT calculation using (A) the non-TL and (B) the TL model with the same DNN architecture. (C) MAE and (D)  $R^2$  on 15 different random testing datasets for non-TL and TL models. (E) Box plot comparison of MAE for the cases where no pretrained parameters on electronic bandgap are used for phonon bandgap model (“non-TL”), TL model takes pretrained parameters from all three hidden layers (full TL), TL model receiving only the first hidden layer parameters from pretrained model (1st layer TL), and TL model receiving pretrained parameters for the second layer only (2nd layer TL). Insets show transferred layers in blue and nontransferred in orange. (F) Box plot of MAE for different TL fine-tune experiments: “No fine-tune,” all parameters in the three hidden layers are imported from the pretrained model but are not allowed to be fine-tuned in retraining; “3rd layer,” the first two hidden layers are frozen, and only the third layer is fine-tuned; “2nd + 3rd layer,” only the second and third hidden layers are fine-tuned, while the first layer remains unchanged; and full TL, normal TL that all hidden layers are allowed to be fine-tuned. Insets show fine-tuned layers in blue and not fine-tuned layers in black.

seen that the performance of transferring information of only the first hidden layer is just slightly inferior to the case where all hidden layers are transferred (denoted as “full TL” in Fig. 4E), indicating that majority of the useful knowledge comes from the first hidden layer. Considering the fact that the information of the second hidden

layer is a complex nonlinear mixture of the first hidden layer, using only the pretrained parameters for the second hidden layer should have little overall impact and the final performance is expected to be similar to fully randomizing the initial parameters of all layers. This indeed agrees with our observation where transferring only the second

hidden layer parameters leads to similar MAE as the non-TL model (Fig. 4E).

The second experiment is conducted by different fine-tuning approaches, where we first import pretrained parameters for all three hidden layers, but some of the hidden layers are not allowed to update (i.e., frozen) in the fine-tuning process on the small phonon bandgap data. When all hidden layers are frozen during retraining, large prediction errors are obtained (Fig. 4F). We then lift the restrictions gradually for the cases where the first two hidden layers are frozen and then only the first hidden layer is frozen (insets in Fig. 4F). When all three hidden layers are free to update, then it is just our normal TL model. From Fig. 4F, we can observe that with more transferred hidden layers allowed to be fine-tuned, the TL models progressively perform better and the MAE approaches the normal TL case. The reasonable and expected results from both experiments further ensure that our implementation of TL from electronic bandgap to phonon bandgap is robust.

With the success on phonon bandgap, we further test the generalizability of TL by extending it to another two phonon-relevant properties, including the speed of sound (i.e., low-frequency phonon group velocity) and the heat capacity at 300 K, of the 124 semiconductors that have phonon bandgap. They are both of great importance for phonon thermal transport, as lattice thermal conductivity is also proportional to heat capacity and the square of phonon group velocity (2), besides phonon relaxation time, which is related to scattering and, thus, the phonon bandgap (35). The speed of sound (unit: km/s) is calculated by averaging the phonon group velocities of the three acoustic modes, which are calculated from the phonon dispersion relation at the Brillouin zone center. The heat capacity (unit:  $\text{J mol}^{-1} \text{K}^{-1}$ ) is calculated from the Bose-Einstein distribution based on the phonon density of states. The calculated properties are validated with available experimental data, and good agreement has been achieved. For instance, the calculated longitudinal and transverse speed of sound along the [100] direction in diamond are 12.735 and 17.392 km/s, respectively; the measured values are 12.82 and 17.52 km/s (48), and the calculated heat capacity of diamond is  $6.297 \text{ J mol}^{-1} \text{K}^{-1}$ , compared with the reference value of  $6.109 \text{ J mol}^{-1} \text{K}^{-1}$  (49). Some more comparison of calculated properties and reference data are included in section S2. Both TL models and non-TL models are trained and evaluated in the same way as the case for phonon bandgap using the 124 semiconductors with phonon bandgap. The model performance for the speed of sound and heat capacity on the 124 semiconductors with phonon bandgap is summarized in Table 1, and the box plots of MAE and  $R^2$  are shown in Fig. 5. By using TL, the model prediction accuracy for speed of sound has been improved from 0.501 to 0.433 km/s in MAE, a 13.6% decrease, and  $R^2$  is improved from 0.763 to 0.838. The MAE of heat capacity is decreased from 6.002 to 2.793  $\text{J mol}^{-1} \text{K}^{-1}$ , reduced by 53.4% using TL, and  $R^2$  is increased from 0.883 to 0.985. The detailed parity plot of these two properties is included in section S3. Even compared to the DNN model for these two properties trained using the complete 1245 available data, the prediction accuracy of the TL models is similar (Table 1), despite the fact that the TL models used only 10% of the data for training.

## DISCUSSION

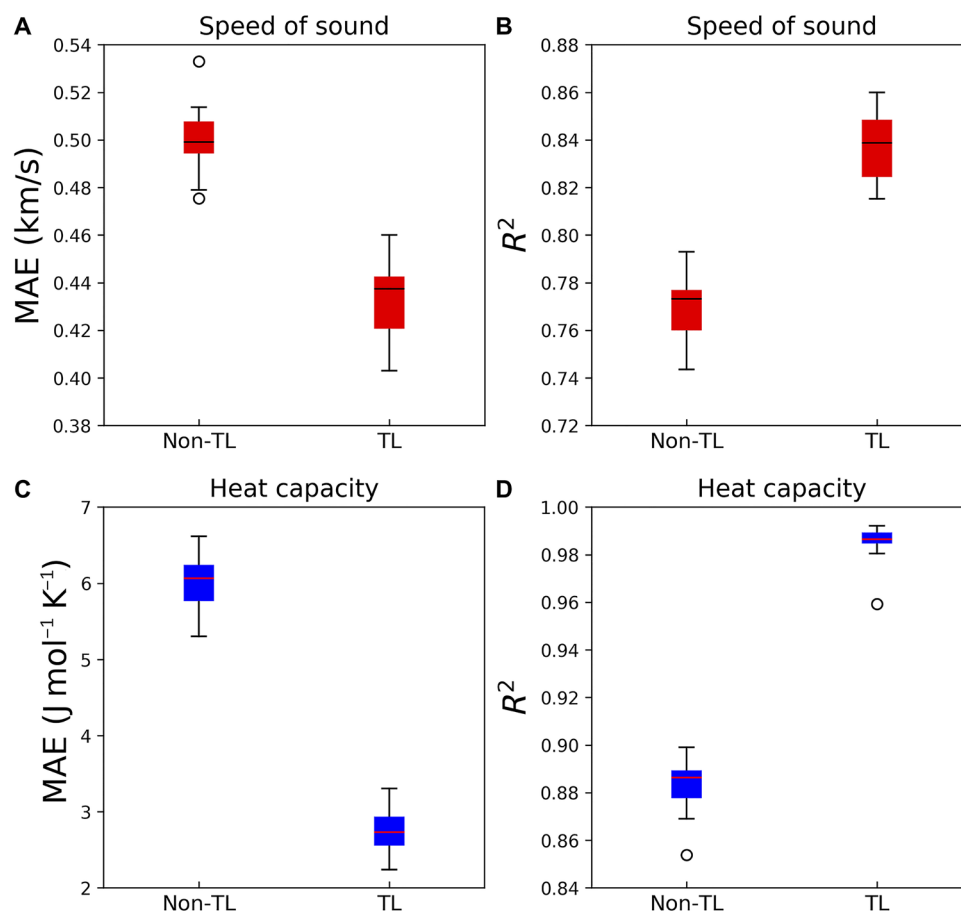
With the robust TL model for phonon properties, we further test it with several III-V semiconductors, which have been the focus of materials research for electronics applications where phonon properties

**Table 1. Model performance for phonon bandgap, speed of sound, and heat capacity.**

	Phonon bandgap ( $\text{cm}^{-1}$ )	Speed of sound (km/s)	Heat capacity ( $\text{J mol}^{-1} \text{K}^{-1}$ )
Non-TL model	MAE: 23.847	MAE: 0.501	MAE: 6.002
	$R^2$ : 0.764	$R^2$ : 0.763	$R^2$ : 0.883
TL model	MAE: 8.458	MAE: 0.433	MAE: 2.739
	$R^2$ : 0.960	$R^2$ : 0.838	$R^2$ : 0.985
Learning from full data		MAE: 0.455	MAE: 2.390
		$R^2$ : 0.870	$R^2$ : 0.989

are also important. There are four important III-V semiconductors—InN, GaSb, InAs, and InSb—that are absent from the 1245 semiconductors database since their DFT-calculated electronic bandgaps in the MP database are zero (i.e., metal) because of the systematic underestimation of electronic bandgap by the DFT method (50). We note that the classification and the TL models for phonon properties take the elemental descriptors of compounds as input with no explicit constraints on the electronic bandgap of the compounds. For instance, for InN, an important semiconductor with promising applications in high-speed electronics and solar cells (51), the experimental bandgap is around 0.7 eV (52), but the DFT data in the MP database indicates it as a metal (53). Our classification model, on the other hand, can successfully classify it to have a phonon bandgap, and the phonon bandgap is predicted to be  $208.382 \text{ cm}^{-1}$  using the TL model, which compares favorably with  $215.0 \text{ cm}^{-1}$  from DFPT (54). The TL-predicted speed of sound and heat capacity are  $37.483 \text{ J mol}^{-1} \text{K}^{-1}$  and 4.029 km/s, respectively, which are also close to the reference data of  $41.73 \text{ J mol}^{-1} \text{K}^{-1}$  (55) and 3.80 km/s (56). The other three semiconductors—GaSb, InAs, and InSb—miscalculated by the DFT to be metals also have wide applications (57–59). Results from our classification and TL models are again very reasonable for these compounds, as summarized in Table 2. Note that this TL model relies on the phonon properties of the semiconductors with phonon bandgap, making the model applicability inevitably biased toward those materials with phonon bandgap, which likely only make up a small portion of the large chemical space. It is advised to first identify whether the semiconductor is with or without a phonon bandgap and then apply the TL model to predict the phonon properties when studying an unseen material.

These results not only demonstrate the generalizability of our models but also imply a very important feature of TL. Although DFT underestimates electronic bandgap (50), the DFT theory still captures some true relationship between the material composition and their electronic bandgap, and the calculated data inherently encode such information. Thus, even if the proxy property (i.e., electronic bandgap) contains errors, the knowledge that connects the material composition to the proxy property can still be transferred to benefit the model for the target property, and the TL model can have high fidelity as long as the training data for the target property are accurate. In the present study, DFPT-calculated phonon properties are known to be accurate (29). This feature can be important to materials informatics in general since one may use simple calculations to massively produce proxy labels, although they might not be very



**Fig. 5. TL model performance for speed of sound and heat capacity.** Box plots of MAE and  $R^2$  for (A and B) speed of sound and (C and D) heat capacity with and without TL.

**Table 2. TL-predicted phonon properties and reference values for some III-V semiconductors not included in the original database.**

	Does it have phonon bandgap?		Phonon bandgap ( $\text{cm}^{-1}$ )		Heat capacity ( $\text{J mol}^{-1} \text{K}^{-1}$ )		Speed of sound ( $\text{km/s}$ )	
	Predicted	Reference	Predicted	Reference	Predicted	Reference	Predicted	Reference
InN (mp-22205)	Yes	Yes (54)	$208.382 \pm 3.524$	215.0 (54)	$37.483 \pm 1.733$	41.73 (55)	$4.029 \pm 0.155$	3.80 (55)
GaSb (mp-1156)	Yes	Yes (61)	$9.129 \pm 4.262$	23.3 (61)	$45.67 \pm 2.775$	47.87 (62)	$3.141 \pm 0.146$	3.17 (62)
InAs (mp-20305)	Yes	Yes (67)	$12.405 \pm 4.509$	15.8 (61)	$45.886 \pm 2.208$	47.43 (55)	$3.127 \pm 0.102$	3.03 (55)
InSb (mp-20012)	No	No (61)			$45.663 \pm 3.090$	47.32 (62)	$2.667 \pm 0.131$	2.66 (62)

accurate, for further TL applications on small available target labels. For example, for developing materials for a new application where experimental data are sparse and difficult to collect, one may perceivably use computer simulations (e.g., coarse-grain molecular simulations) to quickly generate large volumes of data as the proxy labels and then leverage the TL scheme to build a predictive model using both the small experimental data and large available proxy labels.

In summary, we have demonstrated the ability to leverage the more readily accessible electron properties to help predict phonon properties using a TL strategy. By only using element-level compositional descriptors, a classification model using the SMOTE scheme

can accurately classify whether a compound has a phonon bandgap. We then train DNN MLP models on the proxy property of electronic bandgap (1245 data) and use the same model architectures and parameters as initial values in training the target property of phonon bandgap, which only has 10% of the proxy property in data volume. The obtained TL model is found to have high accuracy and notably outperforms the directly trained model (i.e., non-TL model) on the small data. The TL scheme is also extended to construct models for other phonon properties including low-frequency phonon group velocity (i.e., speed of sound) and heat capacity, and improvements over non-TL models are also achieved. Our work indicates a strong underlying connection between the electron and phonon

properties, and this connection can be leveraged by the data-driven TL approach without the need for complex analytical expressions. For semiconductors like InN, which are mistakenly calculated to be metals in the proxy property database, they can still be correctly classified, and their phonon properties are predicted using TL models with accuracy. This suggests that TL can leverage not-so-accurate proxy labels, as long as they have encoded some true composition-property relation, for improving models for target labels, a feature that can have notable implications to materials informatics in general.

## MATERIALS AND METHODS

### Materials representation

The phonon bandgap is extracted from the phonon dispersion relation in the database studied, where materials where the frequency of the lowest-energy optical phonon mode is higher than that of the highest-energy acoustic phonon mode are considered as semiconductors with a phonon bandgap and the frequency difference is defined as the phonon bandgap. To represent these materials in a fixed-length vector for ML algorithms, element-level compositional descriptors implemented by the XenonPy project (32) are used, where a total of 290 elemental property descriptors, such as the atomic number, atomic radius, etc., are included. For example, given a compound Ga<sub>2</sub>O<sub>3</sub>, the elemental properties like the atomic number and radius for elements Ga and O are extracted from the XenonPy element property database, and then, operations like summation, weighted average, weighted variance (e.g., weights of 0.4 for Ga and 0.6 for O), and the maximum or minimum of the elemental properties are performed to produce the 290-dimension descriptors for Ga<sub>2</sub>O<sub>3</sub>. Besides elemental descriptors, we have also tested descriptors with some crystal structural information but found no improvement in model accuracy, and the details are included in section S4. We thus do not include crystal information as part of the descriptors, which speeds up the featurization process.

### Classification model

Seventy percent of the whole dataset is randomly chosen as the training data, and the remaining 30% is for testing for constructing the random forest model. We perform a fivefold cross-validation grid search with F1 score as the criterion for the optimized number of trees and the maximum depth of the tree in the random forest classifier on the SMOTE oversampled training data. Here, the F1 score is defined as the harmonic mean of the precision and recall. The precision is defined as the ratio between correctly classified phonon bandgap materials (true positive) and all materials classified to have phonon bandgap (true positive + false positive), describing how precise our model is given a classification result. The recall is the ratio between correctly classified phonon bandgap materials (true positive) and all materials that actually have a phonon bandgap (true positive + false negative), describing how sensitive our model can capture the real positive cases. The model with the best cross-validation F1 score is chosen for the testing dataset. For ROC, the true-positive rate is the ratio between correctly classified phonon bandgap materials and all materials that actually have phonon bandgap, and the false-positive rate is the ratio between materials without phonon bandgap incorrectly classified as materials with phonon bandgap and all materials with no phonon bandgap.

### DNN model for proxy property

The MLP model is trained using PyTorch (60). For the electronic bandgap (proxy property), there are three hidden layers in our MLP models, and the number of nodes in each hidden layer is randomly selected. A total of 250 models with different numbers of nodes in each hidden layer are pretrained, and the model performance metrics are calculated using the ground truth and the properties averaged by 15 independent fivefold cross-validation predictions. The MAE and  $R^2$  of the 250 models are then averaged.

### Transfer learning

In TL, we build MLP models with exactly the same architecture as the pretrained models for the proxy property, and the parameters from these pretrained models, except those for the output layer, are used as the initial parameters. These transferred models are then retrained using the small phonon bandgap data, the fine-tuning process. Different TL experiments by only transferring the parameters of some hidden layers or fine-tuning the parameters of selected layers are also performed to test the robustness of the TL scheme. The model performance is evaluated using the same method as that used for the proxy electron property.

### SUPPLEMENTARY MATERIALS

Supplementary material for this article is available at <http://advances.sciencemag.org/cgi/content/full/6/45/eabd1356/DC1>

### REFERENCES AND NOTES

1. N. W. Ashcroft, N. D. Mermin, *Solid State Physics* (Saunders College, 1976).
2. G. Chen, *Nanoscale Energy Transport and Conversion: A parallel Treatment of Electrons, Molecules, Phonons, and Photons* (Oxford Univ. Press, 2005).
3. S. Baroni, S. De Gironcoli, A. Dal Corso, P. Giannozzi, Phonons and related crystal properties from density-functional perturbation theory. *Rev. Mod. Phys.* **73**, 515–562 (2001).
4. P. Giannozzi, S. Baroni, N. Bonini, M. Calandra, R. Car, C. Cavazzoni, D. Ceresoli, G. L. Chiarotti, M. Cococcioni, I. Dabo, A. D. Corso, S. de Gironcoli, S. Fabris, G. Fratesi, R. Gebauer, U. Gerstmann, C. Gougoussis, A. Kokalj, M. Lazzeri, L. Martin-Samos, N. Marzari, F. Mauri, R. Mazzarello, S. Paolini, A. Pasquarello, L. Paulatto, C. Sbraccia, S. Scandolo, G. Sclauzero, A. P. Seitsonen, A. Smogunov, P. Umari, R. M. Wentzcovitch, QUANTUM ESPRESSO: A modular and open-source software project for quantum simulations of materials. *J. Phys. Condens. Matter* **21**, 395502 (2009).
5. A. Togo, I. Tanaka, First principles phonon calculations in materials science. *Scr. Mater.* **108**, 1–5 (2015).
6. B. Lv, T. Qian, H. Ding, Angle-resolved photoemission spectroscopy and its application to topological materials. *Nat. Rev. Phys.* **1**, 609–626 (2019).
7. H. Bilz, W. Kress, *Phonon Dispersion Relations in Insulators* (Springer Science, 2012), vol. 10.
8. A. J. Minnich, M. S. Dresselhaus, Z. F. Ren, G. Chen, Bulk nanostructured thermoelectric materials: Current research and future prospects. *Energ. Environ. Sci.* **2**, 466–479 (2009).
9. J. Yan, P. Gorai, B. Ortiz, S. Miller, S. A. Barnett, T. Mason, V. Stevanović, E. S. Toberer, Material descriptors for predicting thermoelectric performance. *Energ. Environ. Sci.* **8**, 983–994 (2015).
10. H. H. Tippins, Optical absorption and photoconductivity in the band edge of  $\beta$ -Ga<sub>2</sub>O<sub>3</sub>. *Phys. Rev.* **140**, A316–A319 (1965).
11. Z. Guo, A. Verma, X. Wu, F. Sun, A. Hickman, T. Masui, A. Kuramata, M. Higashiwaki, D. Jena, T. Luo, Anisotropic thermal conductivity in single crystal  $\beta$ -gallium oxide. *Appl. Phys. Lett.* **106**, 111909 (2015).
12. M. D. Santia, N. Tandon, J. D. Albrecht, Lattice thermal conductivity in  $\beta$ -Ga<sub>2</sub>O<sub>3</sub> from first principles. *Appl. Phys. Lett.* **107**, 041907 (2015).
13. C. Mion, J. F. Muth, E. A. Preble, D. Hanser, Accurate dependence of gallium nitride thermal conductivity on dislocation density. *Appl. Phys. Lett.* **89**, 092123 (2006).
14. L. Lindsay, A. D. Broido, T. L. Reinecke, Thermal conductivity and large isotope effect in GaN from first principles. *Phys. Rev. Lett.* **109**, 095901 (2012).
15. K. Alberi, M. B. Nardelli, A. Zakutayev, L. Mitas, S. Curtarolo, A. Jain, M. Fornari, N. Marzari, I. Takeuchi, M. L. Green, M. Kanatzidis, M. F. Toney, S. Butenko, B. Meredig, S. Lany, U. Kattner, A. Davydov, E. S. Toberer, V. Stevanovic, A. Walsh, N.-G. Park, A. Aspuru-Guzik, D. P. Tabor, J. Nelson, J. Murphy, A. Setlur, J. Gregoire, H. Li, R. Xiao, A. Ludwig, L. W. Martin, A. M. Rappe, S.-H. Wei, J. Perkins, The 2019 materials by design roadmap. *J. Phys. D Appl. Phys.* **52**, 013001 (2018).
16. G. Pilania, J. E. Gubernatis, T. Lookman, Multi-fidelity machine learning models for accurate bandgap predictions of solids. *Comput. Mater. Sci.* **129**, 156–163 (2017).



17. G. Pilania, A. Mannodi-Kanakkithodi, B. P. Uberuaga, R. Ramprasad, J. E. Gubernatis, T. Lookman, Machine learning bandgaps of double perovskites. *Sci. Rep.* **6**, 19375 (2016).
18. J. Lee, A. Seko, K. Shitara, K. Nakayama, I. Tanaka, Prediction model of band gap for inorganic compounds by combination of density functional theory calculations and machine learning techniques. *Phys. Rev. B* **93**, 115104 (2016).
19. A. C. Rajan, A. Mishra, S. Satsangi, R. Vaish, H. Mizuseki, K.-R. Lee, A. K. Singh, Machine-learning-assisted accurate band gap predictions of functionalized MXene. *Chem. Mater.* **30**, 4031–4038 (2018).
20. Y. Zhuo, A. Mansouri Tehrani, J. Brگوچ, Predicting the band gaps of inorganic solids by machine learning. *J. Phys. Chem. Lett.* **9**, 1668–1673 (2018).
21. C. Sutton, L. M. Ghiringhelli, T. Yamamoto, Y. Lysoogorskiy, L. Blumenthal, T. Hammerschmidt, J. R. Golebiowski, X. Liu, A. Ziletti, M. Scheffler, Crowd-sourcing materials-science challenges with the NOMAD 2018 Kaggle competition. *npj Comput. Mater.* **5**, 111 (2019).
22. T. Xie, J. C. Grossman, Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. *Phys. Rev. Lett.* **120**, 145301 (2018).
23. D. Jha, K. Choudhary, F. Tavazza, W.-K. Liao, A. Choudhary, C. Campbell, A. Agrawal, Enhancing materials property prediction by leveraging computational and experimental data using deep transfer learning. *Nat. Commun.* **10**, 5316 (2019).
24. S. K. Kauwe, J. Graser, A. Vazquez, T. D. Sparks, Machine learning prediction of heat capacity for solid inorganics. *Integr. Mater. Manuf. Innov.* **7**, 43–51 (2018).
25. Z. Gong, Y. Wu, L. Wu, H. Sun, Predicting thermodynamic properties of alkanes by high-throughput force field simulation and machine learning. *J. Chem. Inf. Model.* **58**, 2502–2516 (2018).
26. S. Kirklın, J. E. Saal, B. Meredig, A. Thompson, J. W. Doak, M. Aykol, S. Rühl, C. Wolverton, The Open Quantum Materials Database (OQMD): Assessing the accuracy of DFT formation energies. *npj Comput. Mater.* **1**, 15010 (2015).
27. S. Curtarolo, W. Setyawan, G. L. W. Hart, M. Jahnatek, R. V. Chepulskii, R. H. Taylor, S. Wang, J. Xue, K. Yang, O. Levy, M. J. Mehl, H. T. Stokes, D. O. Demchenko, D. Morgan, AFLOW: An automatic framework for high-throughput materials discovery. *Comput. Mater. Sci.* **58**, 218–226 (2012).
28. A. Jain, S. P. Ong, G. Hautier, W. Chen, W. D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder, K. A. Persson, Commentary: The Materials Project: A materials genome approach to accelerating materials innovation. *APL Mater.* **1**, 11002 (2013).
29. G. Petretto, S. Dwaraknath, H. P. C. Miranda, D. Winston, M. Giantomasi, M. J. van Setten, X. Gonze, K. A. Persson, G. Hautier, G.-M. Rignanese, High-throughput density-functional perturbation theory phonons for inorganic materials. *Sci. Data* **5**, 180065 (2018).
30. C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)* (Springer, 2006).
31. S. J. Pan, Q. Yang, A survey on transfer learning. *IEEE Trans. Knowl. Data Eng.* **22**, 1345–1359 (2010).
32. H. Yamada, C. Liu, S. Wu, Y. Koyama, S. Ju, J. Shiomi, J. Morikawa, R. Yoshida, Predicting materials properties with little data using shotgun transfer learning. *ACS Cent. Sci.* **5**, 1717–1730 (2019).
33. X. Li, Y. Dan, R. Dong, Z. Cao, C. Niu, Y. Song, S. Li, J. Hu, Computational screening of new perovskite materials using transfer learning and deep learning. *Appl. Sci.* **9**, 5510 (2019).
34. R. MA, Y. J. Colón, T. Luo, A transfer learning study of gas adsorption in metal-organic frameworks. *ACS Appl. Mater. Interfaces* **12**, 34041–34048 (2020).
35. L. Lindsay, D. A. Broido, Three-phonon phase space and lattice thermal conductivity in semiconductors. *J. Phys. Condens. Matter* **20**, 165209 (2008).
36. X. Gu, Y. Wei, X. Yin, B. Li, R. Yang, *Colloquium*: Phononic thermal properties of two-dimensional materials. *Rev. Mod. Phys.* **90**, 041002 (2018).
37. L. Lindsay, D. A. Broido, T. L. Reinecke, First-principles determination of ultrahigh thermal conductivity of boron arsenide: A competitor for diamond? *Phys. Rev. Lett.* **111**, 025901 (2013).
38. J. S. Kang, M. Li, H. Wu, H. Nguyen, Y. Hu, Experimental observation of high thermal conductivity in boron arsenide. *Science* **361**, 575–578 (2018).
39. S. Li, Q. Zheng, Y. Lv, X. Liu, X. Wang, P. Y. Huang, D. G. Cahill, B. Lv, High thermal conductivity in cubic boron arsenide crystals. *Science* **361**, 579–581 (2018).
40. F. Tian, B. Song, X. Chen, N. K. Ravichandran, Y. Lv, K. Chen, S. Sullivan, J. Kim, Y. Zhou, T.-H. Liu, M. Goni, Z. Ding, J. Sun, G. A. G. U. Gamage, H. Sun, H. Ziyaae, S. Huyan, L. Deng, J. Zhou, A. J. Schmidt, S. Chen, C.-W. Chu, P. Y. Huang, D. Broido, L. Shi, G. Chen, Z. Ren, Unusual high thermal conductivity in boron arsenide bulk crystals. *Science* **361**, 582–585 (2018).
41. H. Ma, C. Li, S. Tang, J. Yan, A. Alatas, L. Lindsay, B. C. Sales, Z. Tian, Boron arsenide phonon dispersion from inelastic x-ray scattering: Potential for ultrahigh thermal conductivity. *Phys. Rev. B* **94**, 220303 (2016).
42. Z. Liu, X. Wu, T. Luo, The impact of hydrogenation on the thermal transport of silicene. *2D Mater.* **4**, 025002 (2017).
43. L. Breiman, Random forests. *Mach. Learn.* **45**, 5–32 (2001).
44. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, É. Duchesnay, Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
45. N. V. Chawla, K. W. Bowyer, L. O. Hall, W. P. Kegelmeyer, SMOTE: Synthetic minority over-sampling technique. *J. Artif. Intell. Res.* **16**, 321–357 (2002).
46. G. Lemaître, F. Nogueira, C. K. Aridas, Imbalanced-learn: A Python toolbox to tackle the curse of imbalanced datasets in machine learning. *J. Mach. Learn. Res.* **18**, 1–5 (2017).
47. A. Adadi, M. Berrada, Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access* **6**, 52138–52160 (2018).
48. H. J. McSkimin, P. Andreatch Jr., Elastic moduli of diamond as a function of pressure and temperature. *J. Appl. Phys.* **43**, 2944–2948 (1972).
49. R. Hultgren, P. D. Desai, D. T. Hawkins, M. Gleiser, K. K. Kelley, *Selected Values of the Thermodynamic Properties of the Elements* (American Society for Metals, 1973).
50. J. P. Perdew, Density functional theory and the band gap problem. *Int. J. Quantum Chem.* **28**, 497–523 (1985).
51. S. K. O’Leary, B. E. Foutz, M. S. Shur, L. F. Eastman, Potential performance of indium-nitride-based devices. *Appl. Phys. Lett.* **88**, 152113 (2006).
52. V. Y. Davydov, A. A. Klochikhin, R. P. Seisyan, V. V. Emtsev, S. V. Ivanov, F. Bechstedt, J. Furthmüller, H. Harima, A. V. Mudryi, J. Aderhold, O. Semchinova, J. Graul, Absorption and emission of hexagonal InN: evidence of narrow fundamental band gap. *Phys. Status Solidi* **229**, r1–r3 (2002).
53. K. Persson, Materials Data on InN (SG:186) by Materials Project (2014); doi:10.17188/1183678.
54. C. Bungaro, K. Rapcewicz, J. Bernholc, *Ab initio* phonon dispersions of wurtzite AlN, GaN, and InN. *Phys. Rev. B* **61**, 6720–6725 (2000).
55. I. Barin, O. Knacke, O. Kubaschewski, *Thermochemical Properties of Inorganic Substances: Supplement* (Springer Science & Business Media, 1977).
56. M. E. Levinshtein, S. L. Rumyantsev, M. S. Shur, *Properties of Advanced Semiconductor Materials: GaN, AlN, InN, BN, SiC, SiGe* (John Wiley & Sons, 2001).
57. P. S. Dutta, H. L. Bhat, V. Kumar, The physics and technology of gallium antimonide: An emerging optoelectronic material. *J. Appl. Phys.* **81**, 5821–5870 (1997).
58. U. Banin, Y. Cao, D. Katz, O. Millo, Identification of atomic-like electronic states in indium arsenide nanocrystal quantum dots. *Nature* **400**, 542–544 (1999).
59. S. Nadj-Perge, V. S. Pribiag, J. W. G. van den Berg, K. Zuo, S. R. Plissard, E. P. A. M. Bakkers, S. M. Frolov, L. P. Kouwenhoven, Spectroscopy of spin-orbit quantum bits in indium antimonide nanowires. *Phys. Rev. Lett.* **108**, 166801 (2012).
60. A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimselshin, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, S. Chintala, in *Advances in Neural Information Processing Systems 32* (Curran Associates Inc., 2019), pp. 8024–8035.
61. L. Lindsay, D. A. Broido, T. L. Reinecke, *Ab initio* thermal transport in compound semiconductors. *Phys. Rev. B* **87**, 165201 (2013).
62. M. Levinshtein, *Handbook Series on Semiconductor Parameters* (World Scientific, 1997), vol. 1.
63. K. Suto, J.-i. Nishizawa, Radiative recombination mechanisms in stoichiometry-controlled GaP crystals. *J. Appl. Phys.* **67**, 459–464 (1990).
64. V. Y. Davydov, Y. E. Kitaev, I. N. Goncharuk, A. N. Smirnov, J. Graul, O. Semchinova, D. Uffmann, M. B. Smirnov, A. P. Mirgorodsky, R. A. Evarestov, Phonon dispersion and Raman scattering in hexagonal GaN and AlN. *Phys. Rev. B* **58**, 12899–12907 (1998).
65. J. F. Green, T. K. Bolland, J. W. Bolland, Theoretical elastic behavior for hexagonal boron nitride. *J. Chem. Phys.* **64**, 656–662 (1976).
66. A. Inaba, A. Yoshiasa, Low-temperature heat capacity of wurtzite-type boron nitride. *Jpn. J. Appl. Phys.* **36**, 5644–5645 (1997).
67. A. Polian, M. Grimsditch, I. Grzegory, Elastic constants of gallium nitride. *J. Appl. Phys.* **79**, 3343–3344 (1996).
68. J. C. Nipko, C.-K. Loong, C. M. Balkas, R. F. Davis, Phonon density of states of bulk gallium nitride. *Appl. Phys. Lett.* **73**, 34–36 (1998).

#### Acknowledgments

**Funding:** T.L. would like to thank ONR MURI (N00014-18-1-2429) for the financial support. The computation is supported, in part, by the University of Notre Dame, Center for Researching Computing and NSF through XSEDE resources provided by TACC Stampede2 under grant number TG-CTS100078. **Author contributions:** Z.L. and T.L. conceived the idea and initiated this project. Z.L. collected the data and trained the model. Z.L., M.J., and T.L. discussed the results and wrote the manuscript. **Competing interests:** The authors declare that they have no competing interests. **Data and materials availability:** All data needed to evaluate the conclusions in the paper are present in the paper and/or the Supplementary Materials. Data are available in a GitHub repository ([https://github.com/liuzyju/phonon\\_TL](https://github.com/liuzyju/phonon_TL)). Additional data related to this paper may be requested from the authors.

Submitted 3 June 2020

Accepted 17 September 2020

Published 4 November 2020

10.1126/sciadv.abd1356

**Citation:** Z. Liu, M. Jiang, T. Luo, Leverage electron properties to predict phonon properties via transfer learning for semiconductors. *Sci. Adv.* **6**, eabd1356 (2020).