npj | computational materials

www.nature.com/npjcompumats

ARTICLE OPEN

Robust combined modeling of crystalline and amorphous silicon grain boundary conductance by machine learning

Chayaphol Lortaraprasert[1] and Junichiro Shiomi[1,2]

Knowledge in thermal and electric transport through grain boundary (GB) is crucial for designing nanostructured thermoelectric materials, where the transport greatly depends on GB atomistic structure. In this work, we employ machine learning (ML) techniques to study the relationship between silicon GB structure and its thermal and electric boundary conductance (TBC and EBC) calculated by Green's function methods. We present a robust ML prediction model of TBC covering crystalline–crystalline and crystalline–amorphous interfaces, using disorder descriptors and atomic density. We also construct high-accuracy ML models for predicting both TBC and EBC and their ratio, using only small data of crystalline GBs. We found that the variations of interatomic angles and distance at GB are the most predictive descriptors for TBC and EBC, respectively. These results demonstrate the robustness of the black-box model and open the way to decouple thermal and electrical conductance, which is a key physical problem with engineering needs.

npj Computational Materials (2022)8:219 ; https://doi.org/10.1038/s41524-022-00898-1

INTRODUCTION

Thermoelectric generators (TEGs)[1–3] are capable of directly converting heat into electricity. With low-maintenance and silent operation, TEGs are highly anticipated to reuse waste heat and power sensors and transmitters necessary to gather countless data in Internet of Things (IoT) for future society. The thermoelectric conversion efficiency is mainly determined by the figure of merit $zT = S^2T\sigma/\kappa$, where S, T, σ, and κ are Seebeck coefficient, absolute temperature, electrical and thermal conductivity, respectively. While there are materials such as Bi_2Te_3[4], PbTe[1], and SnTe[5] that exhibit high zTs, the commercialization remains challenging owing to limited abundance and toxicity of these materials. On the other hand, highly abundant and mass-producible silicon[6–10] still has a low zT at around room temperature (where most IoT sensors operate) due to its high thermal conductivity[10]. Improvement in thermoelectric performance of silicon-based materials would help larger spread of the TEG technology.

Nanocrystallization, a process of populating materials with grain boundaries (GBs), has been employed as one of the approaches to improve the zT[10–20]. The promising effect of GBs on thermoelectrics is generally attributed to phonons, primary heat carriers, being selectively scattered[10,11,14–17,19,20], carrier mobility enhancement[13,18], and energy selection of electrons[12,17]. This shows that GBs can significantly influence the thermoelectric performance of materials, and many experiments have also indicated that the thermoelectric characteristics of a single GB greatly depend on its atomistic structure[21–24]. Therefore, knowing the quantitative relationship between GB structure and its thermoelectric properties becomes advantageous when designing thermoelectric nanomaterials.

To investigate the relationship between GB structure and its physical properties, many computational efforts have been made[25–31]. For instance, excess volume of MgO at GB has been shown to correlate well with thermal conductivity obtained from molecular dynamics (MD) calculations[31]. However, the excess volume alone could not fully account for the structure–property relationships in general due to the high-dimensional nature of GB-related problems[31]. There, structural descriptors containing sufficient information to characterize different atomic structures have recently been shown to be effective in handling the intimidating dimensionality of GBs[32–36]. Several recent studies have employed these structural descriptors along with machine learning (ML) to gain more insights into the GB structure–property relationship[37–45]. For example, Fujii et al.[44] discovered promising descriptors based on suitable smooth overlap of atomic positions (SOAP)[46] that were highly correlated with MD-calculated thermal conductivity of MgO GB. They then constructed an accurate prediction model of thermal conductivity using multiple linear regression with predictors based on hierarchical clustering of these descriptors and identified the structure–property relationship from the regression model[44].

In this work, we employ the ML-descriptor approach to study the relationship between silicon GB and its thermoelectric properties, obtained by high-throughput calculations based on Green's function methods[47,48]. We present a robust ML prediction model of thermal boundary conductance (TBC) covering both crystalline–crystalline and crystalline–amorphous interfaces from input disorder descriptors and atomic density. We also construct high-accuracy ML models for predicting TBC, electrical boundary conductance (EBC), and their ratio (EBC/TBC) of crystalline GBs. Our TBC prediction model exhibits a significantly higher coefficient of determination than earlier studies[40,44]. In addition, the models reveal insights into the guiding principles to improve the thermoelectric performance that the angular and distance variation at GB are the most predictive descriptors for TBC and EBC, respectively. This suggests that populating materials with GBs that have large angular variation and small distance variation can improve the figure of merit. Moreover, the EBC/TBC prediction model also indicates that priority should be given to angular over distance variation when attempting to decouple TBC from EBC in order to increase the figure of merit.

[1]Department of Mechanical Engineering, The University of Tokyo, Tokyo, Japan. [2]Institute of Engineering Innovation, The University of Tokyo, Tokyo, Japan.
✉email: shiomi@photon.t.u-tokyo.ac.jp

Published in partnership with the Shanghai Institute of Ceramics of the Chinese Academy of Sciences

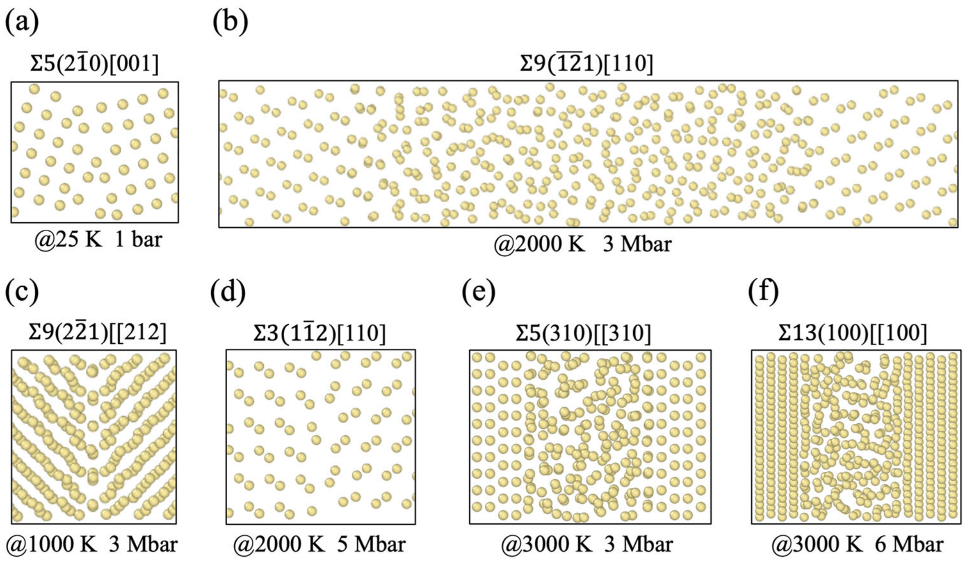

Fig. 1 Examples of grain boundaries used in this work. **a**, **c**, and **d** contain crystalline–crystalline boundaries, while **b**, **e**, and **f** contain two crystalline–amorphous boundaries.

RESULTS AND DISCUSSION

Crystalline and amorphous GBs

We obtained 1228 silicon GB structures at various annealing temperatures and pressures. The resulting GB structures depended on the final temperature and pressure of the annealing process. In general, for low final temperatures, the disordered structure in the device region crystallized from both ends and formed a sharp crystalline–crystalline interface as shown in Fig. 1(a, c, d), and for high final temperatures, the structure in the device region remains being disordered resulting in amorphous slab as shown in Fig. 1(b, e, f).

In the case of amorphous slab, there are two crystalline–amorphous interfaces on the left and right side of the slab, therefore the total resistance (R_{total}) consists of the internal resistance of the slab (R_l) and the thermal boundary resistance at the interfaces (R_0). To understand their relative contribution, we performed conductance calculation with varying lengths (l) of the amorphous device region as shown in Fig. 2. The result shows that R_{total} grows linearly with l, confirming that the thermal conductivity can be approximated to be constant in the range of variation in l. This then allows us to extract $2R_0$ by extrapolating the linear profile to $l = 0$. The analysis finds that, for the range of l, $2R_0$ is significantly larger than R_l and thus the primary contributor to R_{total}. Therefore, half of the transmission through the amorphous slab case can be seen to represent transmission through a single crystalline–amorphous GB.

Consequently, we now have two types of GBs in this work, crystalline–crystalline GBs and crystalline–amorphous GBs. It becomes necessary, therefore, to classify our structures into crystalline–crystalline and crystalline–amorphous structures based on structural disorder. We employed the k-means algorithm[49] using the four disorder descriptors, i.e., σ_θ, H_ϕ, σ_l, and A_{RDF} to cluster our structures into two groups. As a result, we obtained 426 structures with crystalline–crystalline GBs and 802 structures with crystalline–amorphous GBs.

Combined predictions of TBC

Ten different supervised ML models implemented in the scikit-learn Python library[50] were examined: linear regression, k-nearest neighbors (KNN), RandomForest[51], ExtraTrees[52], GradientBoosting[53], CatBoost[54], AdaBoost[55], XGBoost[56], HistGB[57], and LightGBM[58]. The first two are classical models, while the remainings are tree-based models. Five structural descriptors, namely, the standard deviation of bond angles θ and bond lengths l (σ_θ and σ_l), the entropy of dihedral angles ϕ (H_ϕ), the RDF area (A_{RDF}), and the atomic density (ρ) were used as input predictors, and the calculated conductances of single GBs were used as prediction targets for ML models.

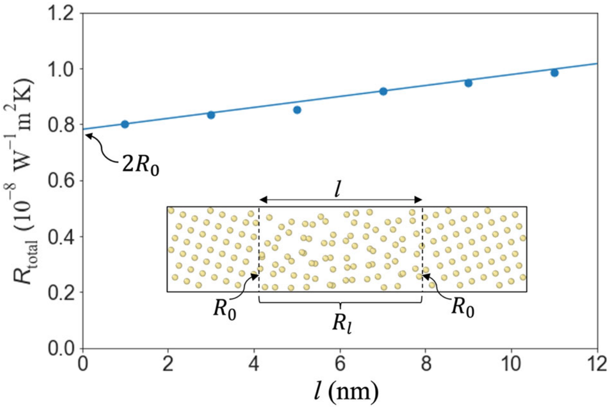

Fig. 2 Length dependence of total thermal resistance (R_{total}) of a disorderly structure. Fourier's law gives: $R_{\text{total}} = 2R_0 + R_l = 2R_0 + r \times l$, where R_0 is the crystalline–amorphous resistance, and R_l and r is the resistance and resistivity of the amorphous (device) region, respectively.

All data were divided into 85% training set and 15% test set. Six-fold cross-validation (CV) on the training set was employed to fine-tune and compare the ML predictive performance. Hyperparameters of each ML model were fine-tuned based on CV results using Tree-structured Parzen Estimator (TPE) algorithm[59–61] implemented in the Optuna Python library[62]. Root mean squared error (RMSE) was selected to evaluate the ML models since large errors are undesirable. Moreover, mean absolute percentage error (MAPE) was also calculated to assist the interpretation of results. The model with the lowest RMSE in CV was then refitted to the entire training set and finally evaluated against the test set to assess a generalization capability.

It can be seen from the CV results shown in Table 1 that linear regression had the worst performance of all the ten ML models. This indicates that the relationship between the structural descriptors in this work and thermal boundary conductance (TBC) cannot be explained by a simple linear relationship. Notably, LightGBM also performed poorly, with average error more than 20%. The other ML models performed fairly well, with average errors between 10% and 20%, indicating the robustness of tree-based models in general. On the other hand, KNN, despite being a traditional ML model, showed the second best performance, only inferior to ExtraTrees model.

Moreover, by combining the individual predictions from the five best models (ExtraTrees, KNN, GradientBoosting, CatBoost and AdaBoost), we were able to realize a powerful ensemble model with better predictions. This is attributable to the fact that combining a set of equally well performing models could balance out the individual weaknesses of each model[63].

Finally, the ensemble model was refitted to the entire training set and evaluated against the test set. As shown in Fig. 3 and Table 2, the ensemble model was able to predict TBC of both GB types with exceptional accuracy on both training and test set. Moreover, the close errors on the training and test set indicate that the ensemble model did not overfit, which commonly occurs when a ML model captures unnecessary noise and detail in the training set. The prediction results on the test set also show robustness of the ensemble model when predicting data it has never encountered before.

Table 1. Six-fold cross-validation results for thermal boundary conductance prediction of all training structures, ordered by root mean squared error (RMSE).

Model	RMSE (GWm^{-2}K^{-1})	MAPE
Ensemble	0.064	9.9%
ExtraTrees	0.078	11.3%
KNN	0.081	12.4%
GradientBoosting	0.087	15.6%
CatBoost	0.091	14.4%
AdaBoost	0.094	17.5%
XGBoost	0.098	19.8%
RandomForest	0.100	18.6%
HistGB	0.107	19.3%
LightGBM	0.108	23.0%
Linear regression	0.187	28.8%

Mean absolute percentage error (MAPE) was calculated for interpretation purpose. The ensemble model was constructed by averaging individual predictions from the five best models.

Furthermore, it becomes apparent that the model performed slightly better on CV than on the test set. This is expected since we fine-tuned the hyperparameters of the models based on CV results, thus the ensemble model was adjusted to give an optimal performance on CV. Consequently, the test set in fact evaluates the generalization ability of the ensemble model. Moreover, as the performance on the test set was comparable to that in cross-validation, which is the average result from different splits, it can be said that it is not by chance that the randomly selected test set in this work yielded remarkable prediction results.

In addition, it can be observed that, while the ensemble model approximately had an equivalent performance on both GB types for the training set, it performed better on crystalline–crystalline GBs than on crystalline–amorphous GBs for the test set. As shown in Table 2, the ensemble model had smaller RMSE on crystalline–crystalline GBs than crystalline–amorphous GBs despite the former type having higher overall TBCs for both training and test set.

The somewhat superior performance of the ensemble model on crystalline–crystalline GBs may be attributed to the fact that the calculated inverse of R_{total} in the crystalline–amorphous case is not exactly the TBC at the interface with minor contribution from the thin amorphous slab (Fig. 2). Another possible explanation is the origin of the disorder features (σ_θ, H_ϕ, σ_l, and A_{RDF}). While the disorder features were obtained from the entire structure in the simulation domain, for crystalline–crystalline GBs, they reflect the disorder of the atoms near the GB, i.e., the local interface disorder since there is virtually no disorder in the internal crystal region away from the interface. On the other hand, in the case of crystalline–amorphous GB, the disorder features reflect that of the amorphous slab in addition to the interface. Therefore, the act of approximating the resistance and features at the crystalline–amorphous interface could potentially have confused the prediction model, hence the inferior performance. Nevertheless, the interesting point here is that despite the approximation, the prediction of conductance was quite successful even for crystalline–amorphous interfaces.

Figure 5 shows the feature importances of the ensemble model, which are the average relative importances of features over the five best models (except KNN) fitted on the entire training set. In general, tree-based models attempt to minimize loss, which is the

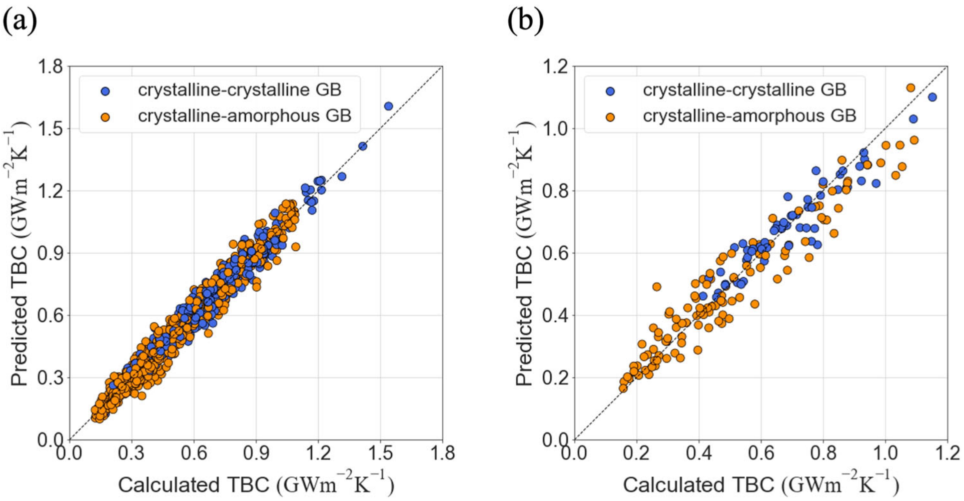

Fig. 3 Parity plot between calculated and predicted thermal boundary conductance (TBC) from the ensemble model. Parity plots of **a** training set and **b** test set. For the test set, the model performed better on crystalline–crystalline GBs than on crystalline–amorphous GBs.

Table 2. Summary of thermal boundary conductance (TBC) prediction results on training and test set.

	Training set		Test set	
	RMSE ($GWm^{-2}K^{-1}$)	MAPE	RMSE ($GWm^{-2}K^{-1}$)	MAPE
Crystalline–crystalline (426 GBs)	0.044	5.8%	0.056	6.2%
Crystalline–amorphous (802 GBs)	0.047	8.1%	0.073	12.5%
Total (1228 GBs)	0.046	7.3%	0.068	10.4%

Root mean squared error (RMSE) and mean absolute percentage error (MAPE) were calculated as evaluation metrics.

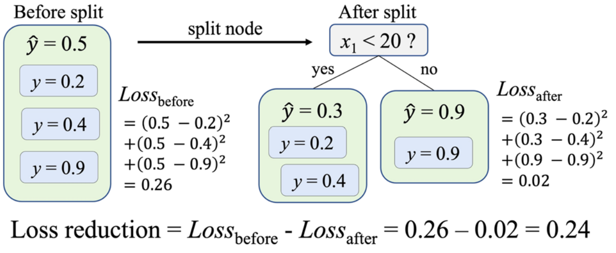

Fig. 4 Demonstration of computing loss reduction resulting from one split in a decision tree. This reduction value for a given feature is summed all over a single tree, then gets averaged over all trees. y, $\hat{y}$, and x_1 denote the target variable, predicted value, and selected feature for splitting, respectively.

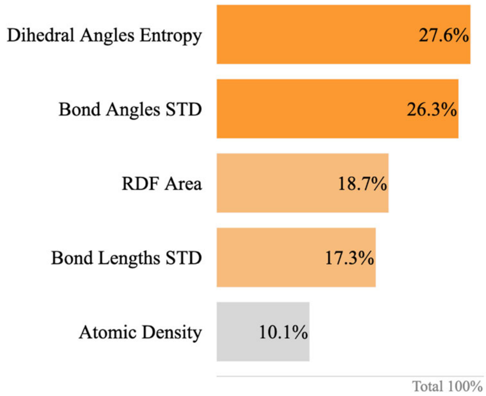

Fig. 5 Feature importances for TBC prediction. Angles-related features are the most predictive. Atomic density is the weakest predictor but still relevant.

variance of a target variable within each leaf node in case of regression, as illustrated in Fig. 4. Feature importance of tree-based models is the total reduction in loss resulting from splitting each node within each tree by that feature, averaged over all trees. The values of feature importances are normalized so that the total importance of all features is unity, i.e., the relative importances are computed. In essence, the importance of a feature represents the contribution of that feature to accurately predicting the target variable, which is the calculated conductance in this work.

It is obvious from Fig. 5 that all of the disorder features (descriptors), i.e., σ_θ, H_ϕ, σ_l, and A_{RDF} are important to predict TBC. Particularly, the angles-related features (σ_θ and H_ϕ) were the strongest predictors of TBC, i.e., they yielded the highest reduction in TBC variance in a tree node. These two features accounted for more than 50% of the total importance for TBC prediction. The distance-related features (σ_l and A_{RDF}) are also crucial for the prediction, with 36% total importance. While atomic density (ρ) was the weakest predictor, it is still a relevant feature since excluding it from the input features greatly exacerbated the predictive performance of the ensemble model. The fact that this feature is still important is attributed to its accounting for varying pressure during the annealing process of GB structure.

As above, the ML-descriptor approach employed in this work is capable of capturing the complex relationship between the atomistic details of GB structure and TBC. The obtained insight cannot be discovered by the traditional methods such as acoustic mismatch model (AMM)[64] and diffusive mismatch model (DMM)[65] as these models do not account for the atomic structure at GB.

Thermoelectric properties predictions of crystalline–crystalline GBs

Regarding electrical boundary conductance (EBC), we discovered that a large number of GBs had zero or extremely small values of maximum EBCs within a realistic range of chemical potential ($-1\ \text{eV} \leq \mu \leq 1\ \text{eV}$). Excluding these GBs left us with 75 crystalline–crystalline and 211 crystalline–amorphous GBs that are electronically realistic.

We repeated the ML-training process on all 286 electronically realistic GBs and the prediction results were mediocre (30.9% and 37.7% average error for TBC and EBC cross-validation, respectively). These inferior results indicate that a large amount of data is necessary in order to predict the conductances of both GB types simultaneously. However, using only 75 crystalline–crystalline GBs (with 20% randomly selected test data) yielded extremely accurate prediction results, as shown in Fig. 6 and Table 3. Six-fold cross-validation (CV) on the training set was also employed to fine-tune ten ML models based on RMSE. An ensemble model, however, was constructed by averaging the individual predictions from the three best models of each property instead of five since this delivered better results.

In addition to TBC and EBC, we also trained the models to directly predict EBC/TBC, which is proportional to the thermoelectric figure of merit (zT). By directly predicting EBC/TBC, we can identify the important descriptors that need to be prioritized when attempting to decouple one conductance from the other in order to increase zT of silicon materials. The Seebeck coefficient was excluded in this work since it is a bulk property in general, in contrast to the boundary conductance which is always an interfacial property (see Supplementary Discussion).

It can be observed from Table 3 that TBC was predicted with merely around 2% average error. Moreover, the training and test results were extremely close, indicating that the ensemble model barely overfitted the training data, thus should have correctly captured the underlying structure–property relationship. In addition, our model also has a significantly higher coefficient of determination ($R^2 = 0.99$) on the test set than previous works predicting the thermal transport property of grain boundary[40,44] despite using less data.

As is the case with combined predictions of both GB types, the angles-related structural descriptors, particularly the entropy of dihedral angles (H_ϕ), were the strongest predictors for TBCs of crystalline–crystalline GBs, as illustrated in Fig. 6b. This result

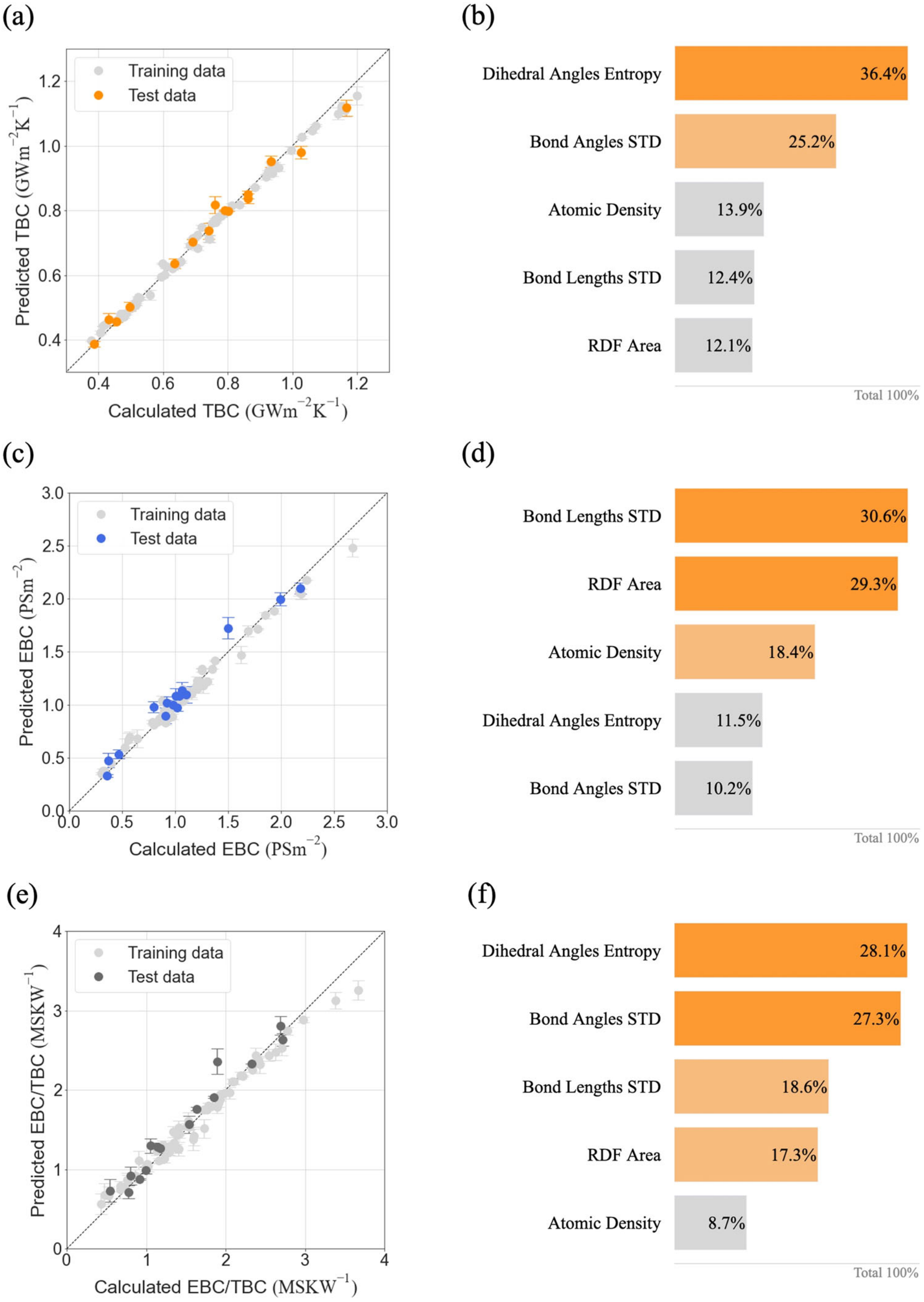

Fig. 6 Prediction results of 75 crystalline–crystalline grain boundaries that are electroincally realistic. Panels **a**, **c**, and **e** demonstrate parity plots between calculated and ensemble-predicted values with error bars indicating standard deviations of the ensemble predictions, while **b**, **d**, and **f** show feature importance of thermal boundary conductance (TBC), electrical boundary conductance (EBC), and EBC/TBC respectively.

further reinforces the previous insight that the angles variation at GB is highly predictive of TBC.

For EBC, although the errors were higher than those of TBC, the ensemble model still enabled good predictions with approximately 9% average error on the test set. In contrast to TBC, distance-related features, i.e., σ_l and A_{RDF} were the most predictive for EBC, as shown in Fig. 6d. This shows potential decoupling of EBC from TBC by decoupling distance-related features from angles-related features. Notably, atomic density was also a strong predictor of EBC.

Finally, predicting EBC/TBC directly was the most difficult task, with ~10% average error on the test set. This result is somewhat intuitive since the ML models should be able to predict each individual property more easily. However, by directly predicting EBC/TBC, we were able to discover the most important features to prioritize when attempting to decouple TBC from EBC, which are the angles-related features, as shown in Fig. 6f.

Furthermore, more data than 75 crystalline–crystalline GBs did not significantly improve the ML predictive performance (see Supplementary Fig. 1). This indicates that, given predictive input

Table 3. Summary of prediction results for thermal and electrical boundary conductance (TBC and EBC) of 75 electronically realistic crystalline–crystalline grain boundaries.

	Training set		Test set	
Property	RMSE	MAPE	RMSE	MAPE
TBC ($GWm^{-2}K^{-1}$)	0.018	1.9%	0.026	2.3%
EBC (PSm^{-2})	0.066	5.4%	0.095	8.8%
EBC/TBC ($MSKW^{-1}$)	0.114	6.7%	0.164	10.1%

Root mean squared error (RMSE) and mean absolute percentage error (MAPE) were calculated as evaluation metrics.

descriptors and high-quality data, even a small amount of data can be sufficient for a ML model to capture the structure–property relationship and thus accurately predict the property. On the other hand, a large amount of data was necessary for predicting the conductances of crystalline–amorphous GBs. This suggests that while the quantity of data is important for a large search space, the quality of data is more important for a small search space.

Moreover, since both TBC and EBC had negative correlations with the degree of structural disorder represented by four descriptors in this work, one possible solution for thermoelectrics is to increase the population of GBs with large angular variation and small distance variation.

In summary, we employed the ML-descriptor approach identify the relationship between silicon GB structure and its thermoelectric properties. Using structural descriptors representing angular and radial variation as well as atomic density, together with high-throughput conductance calculations based on Green's function method, we were able to build a robust ML prediction model of thermal conductance covering both crystalline and amorphous GBs. Moreover, we presented high-accuracy ML models for predicting EBC and TBC and their ratio which is directly proportionate to thermoelectric figure of merit, using only small data of 75 crystalline GBs. We also discovered that the variations of interatomic angles and distance at GB are the most predictive descriptors of TBC and EBC respectively, and that angular variation is more important to EBC/TBC prediction than distance variation. This suggests that constructing GBs with large angular and small distance variation in materials can improve thermoelectric performance of silicon. The insights from this work also open the way to reveal the underlying mechanisms governing GB physical properties. In addition, the extreme robustness of the ML model predicting a large range of interfaces expands the applicability of ML methodology to explore wider search space of material structures and phenomena.

METHODS

Grain boundary (GB) structure construction

Firstly, fifteen Si GBs were generated by tilting or twisting two crystals about their [001], [110], [310], [212] and [100] axes for various misorientation angles, implemented in a Python library called Aimsgb[66]. These structures were relaxed by an annealing method using molecular dynamics (MD) simulation performed in Large-scale Atomic/Molecular Massively Parallel Simulator (LAMMPS) software[67]. The two tilted or twisted Si crystals were placed as the contact regions in Fig. 7 and a melt slab of Si, obtained by heating a single-crystal Si up to 7000 K, was inserted into the device region in between to form an initial structure for further annealing.

Annealing of the initial structure started at 4000 K with a stepwise temperature decrease to 25 K over 1275 ps. The pressure during the process was set to 1 bar. The GB structure was finally optimized with a conjugate gradient algorithm. The optimized Stillinger-Weber interatomic potential[68] was used and one unit layer on each side was fixed throughout the whole process. We repeated the process 10 times for each GB using different initial velocities and selected the most energetically stable structure.

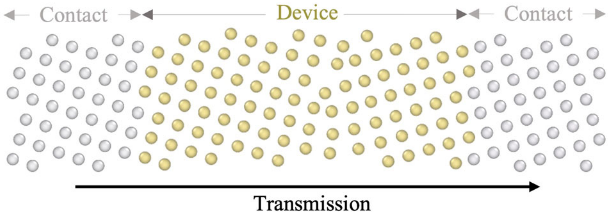

Fig. 7 The model set up in Green's function calculations. The contact regions are the unit layers fixed during the annealing process.

To obtain more structures for training ML models, we varied the values of the final temperature and pressure of the annealing process. Particularly, in addition to the standard condition of 25 K and 1 bar, the annealing calculations with the final temperature of 3000, 2000, and 1000 K, and the constant pressure from 0.1 to 0.6 Mbar with an increment of 0.1 Mbar were performed[69]. All structures were also energetically optimized after the annealing. Furthermore, we took four additional structures with different initial velocity distributions for each amorphous condition (2000 and 3000 K), giving us the total of 84 structures for each of 15 base GBs obtained initially. After excluding 32 unstable structures, we overall obtained 1228 stable or metastable GB structures.

Green's function methods

Green's function methods[47,48] were selected as a means to calculate boundary conductance since they account for the quantum distribution. Moreover, these deterministic methods are also significantly faster than other common alternatives such as non-equilibrium classical molecular dynamics (NEMD) which requires the ensemble averaging of statistical computation. Thus, Green's function methods are exceptional for the ML tasks in this work where a large amount of data is advantageous.

The transmission function of phonons $\Xi(\omega)$ across a GB plane can be calculated as follows[48]:

$$\Xi_{ph}(\omega) = \mathrm{Tr}[\Gamma_L(\omega)G_{ph}(\omega)\Gamma_R(\omega)G^{\dagger}_{ph}(\omega)] \tag{1}$$

where ω is the phonon frequency, G_{ph} is the retarded Green's function of the device region, $\Gamma_L = i(\Sigma_L - \Sigma_L^{\dagger})$ and $\Gamma_R = i(\Sigma_R - \Sigma_R^{\dagger})$ represent the flow rate from the left contact into the device and from the device into the right contact, and Σ_L and Σ_R are the self-energy matrices of the left and right contact. Using the Landauer formula, we can obtain the following thermal boundary conductance (TBC)[70]:

$$\mathrm{TBC} = \frac{1}{2\pi A}\int_0^{\infty} \Xi_{ph}(\omega)\hbar\omega\left(\frac{\partial f_{BE}}{\partial T}\right)d\omega \tag{2}$$

where $\hbar$ is the reduced Planck constant, f_{BE} is the Bose-Einstein distribution, T is the temperature, and A is the GB cross-sectional area.

The transmission function of el ectrons $\Xi(E)$ at energy E can be calculated similarly[47]:

$$\Xi_{el}(E) = \mathrm{Tr}[\Gamma_L(E)G_{el}(E)\Gamma_R(E)G^{\dagger}_{el}(E)] \tag{3}$$

Electrical boundary conductance (EBC) can also be derived in the same manner as TBC:

$$\mathrm{EBC} = \frac{2e^2}{hA}\int_{-\infty}^{\infty} \Xi_{el}(E)\left(-\frac{\partial f_{FD}}{\partial E}\right)dE \tag{4}$$

where h is the Planck constant and f_{FD} is the Fermi-Dirac distribution. Since the value of EBC varies with the chemical potential (μ), we calculated the maximum EBC over a realistic range of chemical potential, $-1\,\text{eV} \le \mu \le 1\,\text{eV}$. All conductance calculations in this work were implemented in an Atomistix ToolKit package[71] with the optimized Stillinger-Weber interatomic potential[68] and the pbc-0-3 Slater Koster file[72] for phonon and electron calculation, respectively. The temperature was set to 300 K in all cases.

Structural descriptors

To represent GB quantitatively, we computed five global structural descriptors. Here, the descriptors were calculated for the entire region of a structure. The first two are the standard deviation of bond angles θ and bond lengths l, i.e., σ_θ and σ_l. Next is the entropy of dihedral angles ϕ, which is calculated as follows:

$$H_\phi = \sum_\phi p_\phi \log\left(\frac{1}{p_\phi}\right) \tag{5}$$

We also computed a descriptor based on radial distribution function (RDF) $g(r)$, which is a measure of the probability of finding an atom at a distance r away from a given reference atom. Motivated by the fact that the value of $g(r)$ converges to unity at smaller r as a structure becomes more disordered[73], we calculated the area between $g(r)$ and $g(r) = 1$ as a descriptor:

$$A_{\mathrm{RDF}} = \int_0^{r_{\mathrm{cutoff}}} |g(r) - 1|\, dr \tag{6}$$

where the cutoff radius r_{cutoff} was set to 10 Å in this work. These four descriptors (σ_θ, H_ϕ, σ_l, and A_{RDF}) quantify the degree of global disorder in a structure. σ_θ and H_ϕ represent angular variation, while σ_l and A_{RDF} represent radial variation. Finally, the global atomic density or number density ρ, which is total number of atoms divided by total volume, was computed as the last structural descriptor. All of the descriptors were calculated using the Atomic Simulation Environment (ASE) python library[74].

DATA AVAILABILITY

Relevant data that support the findings of this study are available from the corresponding author upon reasonable request.

CODE AVAILABILITY

We used Aimsgb[66], LAMMPS[67], Atomistix ToolKit package[71], ASE[74], scikit-learn[50], and Optuna[62]. Atomistix ToolKit package is a commercial software package. All other packages are freely available. Relevant custom code are available from the corresponding author upon reasonable request.

Received: 29 April 2022; Accepted: 27 September 2022;
Published online: 25 October 2022

REFERENCES

1. Biswas, K. et al. High-performance bulk thermoelectrics with all-scale hierarchical architectures. *Nature* **489**, 414–418 (2012).
2. Kim, S. I. et al. Dense dislocation arrays embedded in grain boundaries for high-performance bulk thermoelectrics. *Science* **348**, 109–114 (2015).
3. He, J. & Tritt, T. M. Advances in thermoelectric materials research: Looking back and moving forward. *Science* **357**, eaak9997 (2017).
4. Poudel, B. et al. High-thermoelectric performance of nanostructured bismuth antimony telluride bulk alloys. *Science* **320**, 634–638 (2008).
5. Zhao, L.-D. et al. Ultralow thermal conductivity and high thermoelectric figure of merit in snse crystals. *Nature* **508**, 373–377 (2014).
6. Boukai, A. I. et al. Silicon nanowires as efficient thermoelectric materials. *Nature* **451**, 168–171 (2008).
7. Hochbaum, A. I. et al. Enhanced thermoelectric performance of rough silicon nanowires. *Nature* **451**, 163–167 (2008).
8. Schierning, G. Silicon nanostructures for thermoelectric devices: a review of the current state of the art. *Phys. Status Solidi A* **211**, 1235–1249 (2014).
9. Shiomi, J. Research update: Phonon engineering of nanocrystalline silicon thermoelectrics. *APL Mater.* **4**, 104504 (2016).
10. Kashiwagi, M. et al. Scalable multi-nanostructured silicon for room-temperature thermoelectrics. *ACS Appl. Energy Mater.* **2**, 7083–7091 (2019).
11. Bux, S. K. et al. Nanostructured bulk silicon as an effective thermoelectric material. *Adv. Funct. Mater.* **19**, 2445–2452 (2009).
12. Neophytou, N. et al. Simultaneous increase in electrical conductivity and seebeck coefficient in highly boron-doped nanocrystalline si. *Nanotechnology* **24**, 205402 (2013).
13. Mehdizadeh Dehkordi, A. et al. Large thermoelectric power factor in pr-doped srtio3- δ ceramics via grain-boundary-induced mobility enhancement. *Chem. Mater.* **26**, 2478–2485 (2014).
14. Kim, S. I. et al. Dense dislocation arrays embedded in grain boundaries for high-performance bulk thermoelectrics. *Science* **348**, 109–114 (2015).
15. Miura, A., Zhou, S., Nozaki, T. & Shiomi, J. Crystalline–amorphous silicon nanocomposites with reduced thermal conductivity for bulk thermoelectrics. *ACS Appl. Mater. Interfaces* **7**, 13484–13489 (2015).
16. Zong, P.-a. et al. Skutterudite with graphene-modified grain-boundary complexion enhances zt enabling high-efficiency thermoelectric device. *Energy Environ. Sci.* **10**, 183–191 (2017).
17. Meng, X. et al. Grain boundary engineering for achieving high thermoelectric performance in n-type skutterudites. *Adv. Energy Mater.* **7**, 1602582 (2017).
18. Shi, X. et al. Extraordinary n-type mg3sbbi thermoelectrics enabled by yttrium doping. *Adv. Mater.* **31**, 1903387 (2019).
19. Tsuji, M., Murata, M., Yamamoto, A., Suemasu, T. & Toko, K. Thin-film thermoelectric generator based on polycrystalline sige formed by ag-induced layer exchange. *Appl. Phys. Lett.* **117**, 162103 (2020).
20. Zheng, Z.-h. et al. Enhanced thermoelectric performance in n-type bi2o2se by an exquisite grain boundary engineering approach. *ACS Appl. Energy Mater.* **4**, 10290–10297 (2021).
21. Tai, K., Lawrence, A., Harmer, M. P. & Dillon, S. J. Misorientation dependence of al2o3 grain boundary thermal resistance. *Appl. Phys. Lett.* **102**, 034101 (2013).
22. Furushima, Y. et al. Dislocation structures and electrical conduction properties of low angle tilt grain boundaries in linbo3. *J. Appl. Phys.* **120**, 142107 (2016).
23. Meng, X. et al. Grain boundary engineering for achieving high thermoelectric performance in n-type skutterudites. *Adv. Energy Mater.* **7**, 1602582 (2017).
24. Xu, D. et al. Thermal boundary resistance correlated with strain energy in individual si film-wafer twist boundaries. *Mater. Today Phys.* **6**, 53–59 (2018).
25. Schelling, P., Phillpot, S. & Keblinski, P. Kapitza conductance and phonon scattering at grain boundaries by simulation. *J. Appl. Phys.* **95**, 6082–6091 (2004).
26. Watanabe, T., Ni, B., Phillpot, S. R., Schelling, P. K. & Keblinski, P. Thermal conductance across grain boundaries in diamond from molecular dynamics simulation. *J. Appl. Phys.* **102**, 063503 (2007).
27. Bagri, A., Kim, S.-P., Ruoff, R. S. & Shenoy, V. B. Thermal transport across twin grain boundaries in polycrystalline graphene from nonequilibrium molecular dynamics simulations. *Nano Lett.* **11**, 3917–3921 (2011).
28. Chernatynskiy, A., Bai, X.-M. & Gan, J. Systematic investigation of the misorientation-and temperature-dependent kapitza resistance in ceo2. *Int. J. Heat. Mass Transf.* **99**, 461–469 (2016).
29. Sadasivam, S. et al. Thermal transport across metal silicide-silicon interfaces: First-principles calculations and green's function transport simulations. *Phys. Rev. B* **95**, 085310 (2017).
30. Yeandel, S. R., Molinari, M. & Parker, S. C. The impact of tilt grain boundaries on the thermal transport in perovskite srtio 3 layered nanostructures. a computational study. *Nanoscale* **10**, 15010–15022 (2018).
31. Fujii, S., Yokoi, T. & Yoshiya, M. Atomistic mechanisms of thermal transport across symmetric tilt grain boundaries in mgo. *Acta Mater.* **171**, 154–162 (2019).
32. Schoenholz, S. S., Cubuk, E. D., Sussman, D. M., Kaxiras, E. & Liu, A. J. A structural approach to relaxation in glassy liquids. *Nat. Phys.* **12**, 469–471 (2016).
33. Ramprasad, R., Batra, R., Pilania, G., Mannodi-Kanakkithodi, A. & Kim, C. Machine learning in materials informatics: recent applications and prospects. *Npj Comput. Mater.* **3**, 1–13 (2017).
34. Jäger, M. O., Morooka, E. V., Federici Canova, F., Himanen, L. & Foster, A. S. Machine learning hydrogen adsorption on nanoclusters through structural descriptors. *Npj Comput. Mater.* **4**, 1–8 (2018).
35. Patala, S. Understanding grain boundaries–the role of crystallography, structural descriptors and machine learning. *Comput. Mater. Sci.* **162**, 281–294 (2019).
36. Konstantinou, K., Mocanu, F. C., Lee, T.-H. & Elliott, S. R. Revealing the intrinsic nature of the mid-gap defects in amorphous ge2sb2te5. *Nat. Commun.* **10**, 1–10 (2019).
37. Rosenbrock, C. W., Homer, E. R., Csányi, G. & Hart, G. L. Discovering the building blocks of atomic systems using machine learning: application to grain boundaries. *Npj Comput. Mater.* **3**, 1–7 (2017).

38. Tamura, T. et al. Fast and scalable prediction of local energy at grain boundaries: machine-learning based modeling of first-principles calculations. *Model. Simul. Mater. Sci. Eng.* **25**, 075003 (2017).
39. Sharp, T. A. et al. Machine learning determination of atomic dynamics at grain boundaries. *Proc. Natl Acad. Sci.* **115**, 10943–10947 (2018).
40. Zhan, T., Fang, L. & Xu, Y. Prediction of thermal boundary resistance by the machine learning method. *Sci. Rep.* **7**, 1–9 (2017).
41. Zhu, Q., Samanta, A., Li, B., Rudd, R. E. & Frolov, T. Predicting phase behavior of grain boundaries with evolutionary search and machine learning. *Nat. Commun.* **9**, 1–9 (2018).
42. Snow, B. D., Doty, D. D. & Johnson, O. K. A simple approach to atomic structure characterization for machine learning of grain boundary structure-property models. *Front. Mater.* **6**, 120 (2019).
43. Wu, X. et al. Application of machine learning to predict grain boundary embrittlement in metals by combining bonding-breaking and atomic size effects. *Mater. (Basel)* **13**, 179 (2020).
44. Fujii, S., Yokoi, T., Fisher, C. A., Moriwake, H. & Yoshiya, M. Quantitative prediction of grain boundary thermal conductivities from local atomic environments. *Nat. Commun*. **11**, 1–10 (2020).
45. Zhang, S. et al. Predicting grain boundary damage by machine learning. *Int. J. Plast*. **150**, 103186 (2022).
46. Bartók, A. P., Kondor, R. & Csányi, G. On representing chemical environments. *Phys. Rev. B* **87**, 184115 (2013).
47. Datta, S. Nanoscale device modeling: the green's function method. *Superlattices Microstruct*. **28**, 253–278 (2000).
48. Zhang, W., Fisher, T. & Mingo, N. The atomistic green's function method: An efficient simulation approach for nanoscale phonon transport. *Numer. Heat. Transf. B: Fundam.* **51**, 333–349 (2007).
49. Lloyd, S. Least squares quantization in pcm. *IEEE Trans. Inf. Theory* **28**, 129–137 (1982).
50. Pedregosa, F. et al. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
51. Breiman, L. Random forests. *Mach. Learn.* **45**, 5–32 (2001).
52. Geurts, P., Ernst, D. & Wehenkel, L. Extremely randomized trees. *Mach. Learn.* **63**, 3–42 (2006).
53. Friedman, J. H. Stochastic gradient boosting. *Comput. Stat. Data Anal.* **38**, 367–378 (2002).
54. Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V. & Gulin, A. Catboost: unbiased boosting with categorical features. *Adv. Neural Inf. Process. Syst*. **31**, 6637–6647 (2018).
55. Schapire, R. E. *Explaining adaboost*. In *Empirical inference*, 37–52 (Springer, 2013).
56. Chen, T. & Guestrin, C. Xgboost: a scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 785–794 (ACM, 2016).
57. Guryanov, A. Histogram-based algorithm for building gradient boosting ensembles of piecewise linear decision trees. In *International Conference on Analysis of Images, Social Networks and Texts*, 39–50 (Springer, 2019).
58. Ke, G. et al. Lightgbm: a highly efficient gradient boosting decision tree. *Adv. Neural Inf. Process. Syst.* **30**, 3149–3157 (2017).
59. Bergstra, J., Bardenet, R., Bengio, Y. & Kégl, B. Algorithms for hyper-parameter optimization. *Adv. Neural Inf. Process. Syst*. **24**, 2546–2554 (2011).
60. Bergstra, J., Yamins, D. & Cox, D. Making a science of model search: Hyperparameter optimization in hundreds of dimensions for vision architectures. In *International conference on machine learning*, 115–123 (PMLR, 2013).
61. Ozaki, Y., Tanigaki, Y., Watanabe, S. & Onishi, M. Multiobjective tree-structured parzen estimator for computationally expensive optimization problems. In *Proceedings of the 2020 Genetic and Evolutionary Computation Conference*, 533–541 (ACM, 2020).
62. Akiba, T., Sano, S., Yanase, T., Ohta, T. & Koyama, M. Optuna: a next-generation hyperparameter optimization framework. In *Proceedings of the 25rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2623–2631 (ACM, 2019).
63. An, K. & Meng, J. Voting-averaged combination method for regressor ensemble. In *International Conference on Intelligent Computing*, 540–546 (Springer, 2010).
64. Little, W. The transport of heat between dissimilar solids at low temperatures. *Can. J. Phys.* **37**, 334–349 (1959).
65. Swartz, E. T. & Pohl, R. O. Thermal boundary resistance. *Rev. Mod. Phys.* **61**, 605 (1989).
66. Cheng, J., Luo, J. & Yang, K. Aimsgb: an algorithm and open-source python library to generate periodic grain boundary structures. *Comput. Mater. Sci.* **155**, 92–103 (2018).
67. Thompson, A. P. et al. LAMMPS—a flexible simulation tool for particle-based materials modeling at the atomic, meso, and continuum scales. *Comp. Phys. Comm.* **271**, 108171 (2022).
68. Lee, Y. & Hwang, G. S. Force-matching-based parameterization of the stillinger-weber potential for thermal conduction in silicon. *Phys. Rev. B* **85**, 125204 (2012).
69. Yokoi, T. & He, Y. M. Atomistic simulations of grain boundary transformation under high pressures in mgo. *Phys. B* **532**, 2–8 (2017).
70. Sivan, U. & Imry, Y. Multichannel landauer formula for thermoelectric transport with application to thermopower near the mobility edge. *Phys. Rev. B* **33**, 551 (1986).
71. Smidstrup, S. et al. Quantumatk: An integrated platform of electronic and atomic-scale modelling tools. *J. Phys. Condens. Matter* **32**, 015901 (2020).
72. Sieck, A., Frauenheim, T. & Jackson, K. Shape transition of medium-sized neutral silicon clusters. *Phys. Status Solidi B* **240**, 537–548 (2003).
73. Kirkwood, J. G. & Boggs, E. M. The radial distribution function in liquids. *J. Chem. Phys.* **10**, 394–402 (1942).
74. Larsen, A. H. et al. The atomic simulation environment—a python library for working with atoms. *J. Phys. Condens. Matter* **29**, 273002 (2017).

ACKNOWLEDGEMENTS

We would like to thank Masato Ohnishi for technical support and fruitful discussions regarding the dataset creation part. We would also like to thank Takuma Shiga for advice and support throughout the project, and Tom Hamakawa and Naoki Chiaba for discussions regarding the dataset creation part. C.L. acknowledges financial support from MEXT scholarship. This work was partially supported by JST-CREST (Grant No. JPMJCR21O2).

AUTHOR CONTRIBUTIONS

J.S. conceived the idea and supervised the entire work. C.L. constructed the grain boundary structures, calculated the high-throughput boundary conductances, conceptualized the structural descriptors, performed the machine learning predictions, and wrote the first drafts of paper as well as Supplementary Information. Both authors discussed the results and wrote the paper.

COMPETING INTERESTS

The authors declare no competing interests.

ADDITIONAL INFORMATION

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41524-022-00898-1.

Correspondence and requests for materials should be addressed to Junichiro Shiomi.

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

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