Supplementary Information

Combining Machine Learning Models with First-Principles High-Throughput Calculation to Accelerate the Search of Promising Thermoelectric Materials

Tao Fan*, Artem R. Oganov
Skolkovo Institute of Science and Technology, Bolshoy Boulevard 30, bld. 1, 121205 Moscow, Russia.
*Email: Tao.Fan@skoltech.ru

Labeling the dataset

Since our aim is to train a classification model to distinguish good and not good thermoelectric materials, the first step is labeling each sample in n type and p type datasets, respectively. Currently we have the maximum power factor (PF_{max}) value in the temperature range 300 K to 1000 K, which can represent the capability of electronic transport properties of a compound. In order to label the sample, we need to decide the boundary value of PF_{max} which separates the samples as good (label 1) and not good (label 0). Recall the definition of figure of merit ZT,

$$ZT = \frac{\alpha^2 \sigma T}{\kappa} = \frac{PF \cdot T}{\kappa}$$

thus

$$PF = \frac{ZT \cdot \kappa}{T}$$

Usually, a compound with ZT larger than 1 is thought as good thermoelectric material. The thermal conductivity κ include lattice part (κ_L) and electron part (κ_e). Here for the purpose of simplifying the problem, we assume only lattice thermal conductivity matters and omit the contribution from the electrons. But we need to emphasize that the PF_{max} is generally achieved at high carrier concentrations, where κ_e can sometimes be comparable with or even exceed κ_L. The lowest lattice thermal conductivity a compound can reach is so called amorphous limit, which is possible when the phonons have a mean free path l on the order of the interatomic spacing[1,2]. The value of the amorphous limit is usually between 0.2 $W \cdot m^{-1} \cdot K^{-1}$ and 0.5 $W \cdot m^{-1} \cdot K^{-1}$. Here we use the lower boundary 0.2 $W \cdot m^{-1} \cdot K^{-1}$. Then, we can calculate the lowest possible PF_{max} which could lead to ZT larger than 1 in the temperature range of 300 K to 1000 K,

$$\frac{1 \cdot 0.2}{1000} = 0.0002 \le PF_{max} \le \frac{1 \cdot 0.2}{300} = 0.00067 \Leftrightarrow$$
$$2\ \mu W \cdot cm^{-1} \cdot K^{-2} \le PF_{max} \le 6.7\ \mu W \cdot cm^{-1} \cdot K^{-2}$$

Choosing which value exactly as the boundary for good and not good thermoelectric materials is heuristic. If selecting a value close to 2 $\mu W \cdot cm^{-1} \cdot K^{-2}$, we actually lower the standard and expand the searching space, but at the risk of including many compounds that are actually not promising and wasting our resources. If selecting a value close to 6.7 $\mu W \cdot cm^{-1} \cdot K^{-2}$, we lift the standard and narrow the search space, but also at the risk of missing some promising compounds. Here we use the boundary value as 5 $\mu W \cdot m^{-}$

$^{1}\cdot K^{-2}$, because, practically it is rare to get such a low kappa value as 0.2 $W\cdot m^{-1}\cdot K^{-1}$, and also we are more interested in materials showing good performance near room temperature. According to this boundary value, for n type dataset, #positive:#negative = 308:444, while for p type dataset, #positive:#negative = 244:513.

Table S1. Elemental properties used to compute composition descriptors

Atomic Number	Mendeleev Number	Atomic Weight	Melting Temperature	Column
Row	Covalent Radius	Electron Negativity	# s Valence Electrons	# p Valence Electrons
# d Valence Electrons	# f Valence Electrons	Total # Valance Electrons	# Unfilled s States	# Unfilled p States
# Unfilled d States	# Unfilled f States	Total # Unfilled States	Specific Volume of 0 K Ground State	Band Gap Energy of 0 K Ground State
Electron Affinity	First Ionization Energy	Space Group Number of 0 K Ground State		

Table S2. Top 50 non-cubic n type thermoelectric materials found in this work. Entry id refers to the id number in Materials Project database. N is the band degeneracy, m_c^* the conductivity effective mass, m_d^* the density of states effective mass, E_g the band gap, and Ξ the deformation potential constant. PF_{max} was calculated within temperature range from 300 K to 1000 K

Formula	Entry id	N	m_c^*	m_d^*	E_g (eV)	Ξ (eV)	PF_{max} ($\mu W\cdot cm^{-1}\cdot K^{-2}$)
Ge_5Te_4Se	mp-1224356	12	0.129	0.533	0.226	11.022	82.032
$KBiSe_2$	mp-36539	8	0.226	1.485	0.814	7.098	74.726
TbAsSe	mp-1102476	2	0.046	0.104	0.125	4.166	59.381
DyAsSe	mp-1102952	2	0.031	0.073	0.073	4.826	50.953
YAsSe	mp-1095603	2	0.037	0.087	0.096	5.167	50.362
PbS	mp-1018115	4	0.187	0.320	0.915	8.562	44.409
HoAsSe	mp-1212091	2	0.024	0.052	0.032	4.710	43.052
PbSe	mp-1063670	4	0.186	0.339	0.578	8.079	39.249
TbAsS	mp-	3	0.120	0.185	0.054	9.167	35.943

	1101828						
YSe_2	mp-1232213	2	0.362	1.323	0.145	3.000	33.333
SmAsSe	mp-1208883	2	0.060	0.166	0.133	9.067	28.309
GeTe	mp-1080459	4	0.151	0.260	0.454	7.901	27.554
Ga_2Te_5	mp-2371	6	0.211	0.478	0.944	8.554	25.120
YAsS	mp-1102959	2	0.045	0.374	0.056	10.117	24.806
$ErAsSe_4$	mp-1213048	2	0.023	0.048	0.024	7.196	23.454
Te_4Pb_5Se	mp-1217406	4	0.172	0.479	0.590	10.084	22.717
Bi_2Se_2S	mp-1227504	4	0.375	0.719	0.721	8.396	22.620
$TlGaTe_2$	mp-3785	6	0.225	0.455	0.532	8.761	21.611
Tl_9SbTe_6	mp-34292	5	0.198	0.730	0.634	8.077	20.452
$NdTl_2InTe_4$	mp-1220095	10	0.276	1.027	0.968	8.991	19.229
$RbDy_2Ag_3Te_5$	mp-1190502	2	0.343	0.466	1.007	3.706	17.909
Tl_9SbSe_6	mp-676274	5	0.305	0.990	0.728	7.443	17.111
$BaTe_2$	mp-2150	4	0.345	1.275	0.343	7.306	17.080
Ba_2HfS_4	mp-9321	1	0.264	0.904	0.872	4.502	16.683
Tl_9BiSe_6	mp-34361	5	0.287	0.844	0.688	7.403	16.403
$TlCu_7S_4$	mp-21964	6	0.557	1.398	0.641	9.973	15.062
$BiSbSe_3$	mp-1227508	4	0.305	0.555	0.638	8.233	14.871
ErPS	mp-1191596	1	0.306	0.491	0.1137	4.485	14.762
$LuCuPbSe_3$	mp-1205378	2	0.382	0.619	0.996	8.820	14.169
Dy_3CuSe_6	mp-1225646	1	0.107	0.307	0.660	8.321	13.846
Ba_2ZrS_4	mp-3813	1	0.264	1.002	0.631	5.067	13.785
Pr_2InCuS_5	mp-1220176	4	0.639	1.661	0.941	10.550	13.756
Sm_3CuSe_6	mp-1219332	1	0.121	0.337	0.733	7.700	13.733
$Ba_5Hf_4S_{13}$	mp-557032	1	0.261	0.677	0.664	5.142	13.672

$SnPbS_2$	mp-1218951	4	0.216	0.361	0.626	8.475	13.518
GeS	mp-12910	2	0.318	1.352	0.602	7.472	13.429
PtS	mp-288	5	0.464	0.901	0.407	14.723	13.411
$Sr_3Sb_4S_9$	mp-29295	2	0.381	1.574	1.014	6.183	13.409
$LiCuS_2$	mp-766486	4	0.612	1.159	0.546	7.738	12.989
$CuSbSe_2$	mp-20331	4	0.396	0.668	0.510	9.099	12.819
Bi_2Se_3	mp-23164	4	0.256	0.527	0.607	8.475	12.794
GeSe	mp-700	2	0.258	1.237	0.867	7.181	12.542
$Ba_4Hf_3S_{10}$	mp-1189858	1	0.258	0.669	0.704	5.176	12.535
TlSe	mp-1836	4	0.212	0.543	0.197	7.317	12.486
$BaCu_2Te_2$	mp-30133	4	0.384	0.766	0.602	8.109	12.420
$LuCuPbSe_3$	mp-583052	3	0.434	0.603	1.053	7.586	12.389
$ZnCu_2SiTe_4$	mp-1078498	3	0.133	0.210	0.166	12.644	12.218
NaTe	mp-28353	2	0.196	0.373	0.520	6.282	12.026
Er_2Te_3	mp-14643	2	0.306	0.376	0.601	5.392	11.793
$TmCuPbSe_3$	mp-865269	2	0.406	0.662	1.041	8.600	11.421

Table S3. Top 50 non-cubic p type thermoelectric materials found in this work. Entry id refers to the id number in Materials Project database. N is the band degeneracy, m_c^* the conductivity effective mass, m_d^* the density of states effective mass, E_g the band gap, and Ξ the deformation potential constant. PF_{max} was calculated within temperature range from 300 K to 1000 K

Formula	Entry id	N	m_c^*	m_d^*	E_g (eV)	Ξ (eV)	PF_{max} ($\mu W \cdot cm^{-1} \cdot K^{-2}$)
$KBiSe_2$	mp-36539	12	0.356	2.423	0.814	7.264	73.372
Ge_5Te_4Se	mp-1224356	4	0.034	0.112	0.226	13.397	61.018
TbAsS	mp-1101828	2	0.034	0.110	0.054	5.117	59.956
$NaBiSe_2$	mp-35015	4	0.151	0.649	0.679	9.070	34.868
Ba_2HfS_4	mp-9321	6	0.878	2.902	0.872	5.903	34.083
$NaSbS_2$	mp-1173513	4	0.053	0.417	0.089	11.117	32.156
$LiBiS_2$	mp-33526	4	0.159	0.517	0.811	10.942	27.934

Ba_2ZrS_4	mp-3813	6	0.888	2.887	0.631	6.346	27.435
$Ba_4Hf_3S_{10}$	mp-1189858	6	0.776	2.182	0.704	7.148	27.416
PbS	mp-1018115	2	0.211	0.535	0.915	10.235	26.501
DyAsSe	mp-1102952	2	0.036	0.073	0.073	9.236	25.257
YAsSe	mp-1095603	2	0.043	0.089	0.096	9.417	24.958
PbSe	mp-1063670	4	0.156	0.385	0.579	16.364	23.621
TbAsSe	mp-1102476	2	0.057	0.110	0.125	9.367	23.303
$Ba_4Zr_3S_{10}$	mp-14883	6	0.787	2.152	0.485	7.780	22.199
Te_4Pb_5Se	mp-1217406	4	0.134	0.443	0.590	12.333	21.496
$Sm_{10}Se_{19}$	mp-29832	4	0.398	2.587	0.269	9.367	21.099
$Ba_3Zr_2S_7$	mp-9179	6	0.809	2.443	0.528	8.359	19.701
Er_2Se_3	mp-1225508	4	0.517	1.317	0.230	7.780	19.586
$Ba_3Zr_2S_7$	mp-554172	6	0.809	2.443	0.528	8.359	19.396
MgS_2	mp-1185953	6	0.822	3.948	0.932	8.628	19.136
HoAsSe	mp-1212091	2	0.026	0.052	0.032	9.311	18.522
DyPS	mp-1192185	3	0.613	1.015	0.295	7.270	17.628
TbPS	mp-1190952	3	0.657	1.082	0.355	7.150	17.436
GeS	mp-12910	4	0.137	0.307	0.602	10.118	17.146
YPS	mp-1191026	3	0.634	1.061	0.307	7.450	17.118
CoAsS	mp-4627	4	0.814	1.612	0.942	13.671	16.943
Zn_4CdSe_5	mp-1215624	5	0.803	1.757	0.920	7.913	16.667
SnS	mp-559676	4	0.146	0.344	0.365	12.498	15.443
ErAsSe	mp-1213048	2	0.025	-0.048	0.024	9.111	14.926

$Tm_{10}S_{19}$	mp-1204382	4	0.968	6.831	0.552	9.850	14.523
$SnTe_4Pb_3$	mp-1218925	4	0.051	0.146	0.060	11.683	14.484
ErTeAs	mp-1212634	3	0.262	2.011	0.248	10.310	14.180
GeTe	mp-1080459	2	0.086	0.215	0.454	11.641	13.893
SmPS	mp-1191565	3	1.077	1.775	0.380	6.267	13.608
Li_5SbS	mp-767409	2	0.284	0.857	0.605	7.331	13.561
DyTeAs	mp-1212803	3	0.265	2.188	0.225	10.271	13.204
$BaAg_2GeS_4$	mp-7394	5	0.753	1.842	0.585	11.884	12.880
YbY_2Se_4	mp-1193988	2	0.485	2.337	1.122	7.667	12.633
Ga_2HgTe_4	mp-1224839	3	0.406	1.534	0.326	9.780	12.507
$SnPb_4S_5$	mp-1218954	4	0.156	0.649	0.085	10.568	12.341
K_2Se_3	mp-7670	6	1.536	8.316	0.627	4.302	12.318
$GeTe_2Pb$	mp-1224318	2	0.031	0.055	0.067	10.468	12.205
$YbTm_2Se_4$	mp-1193091	2	0.459	1.244	1.140	7.700	11.607
$YbEr_2Se_4$	mp-1192445	2	0.462	1.248	1.129	7.667	11.389
$YbHo_2Se_4$	mp-1194224	2	0.464	1.251	1.117	7.633	11.221
$CaIn_2Te_4$	mp-677072	5	0.425	1.310	0.653	10.178	10.823
NdPS	mp-1191667	3	1.562	2.775	0.310	6.144	10.738
SmTeAs	mp-1208842	3	0.265	1.219	0.167	10.056	10.475
Sm_2S_3	mp-1403	3	0.780	1.330	0.750	8.350	10.438

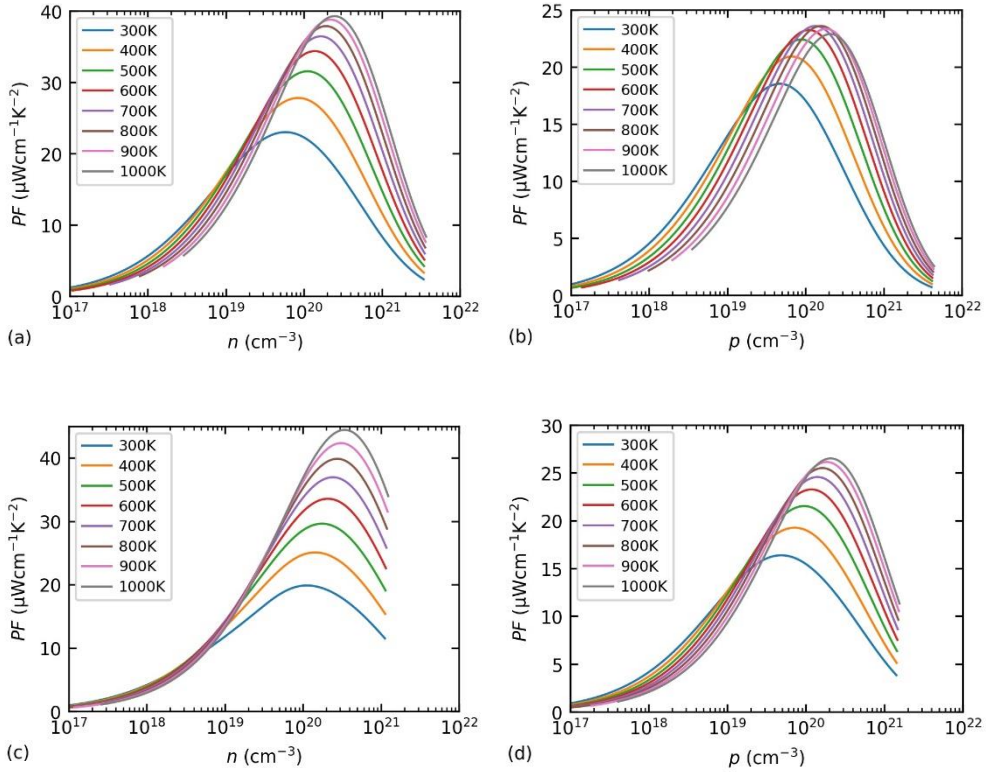


Figure S1. Power factor at varying temperatures and carrier concentrations for (a, b) PbSe and (c, d) PbS.

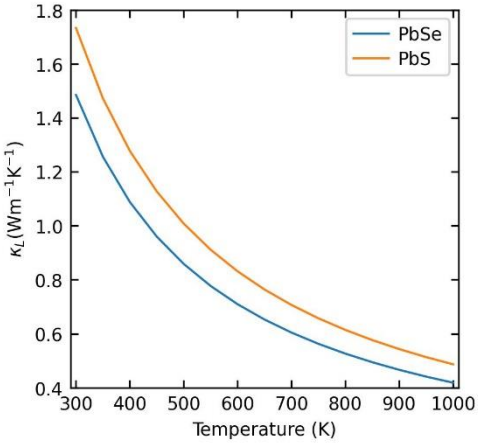


Figure S2. Lattice thermal conductivity of PbSe and PbS.

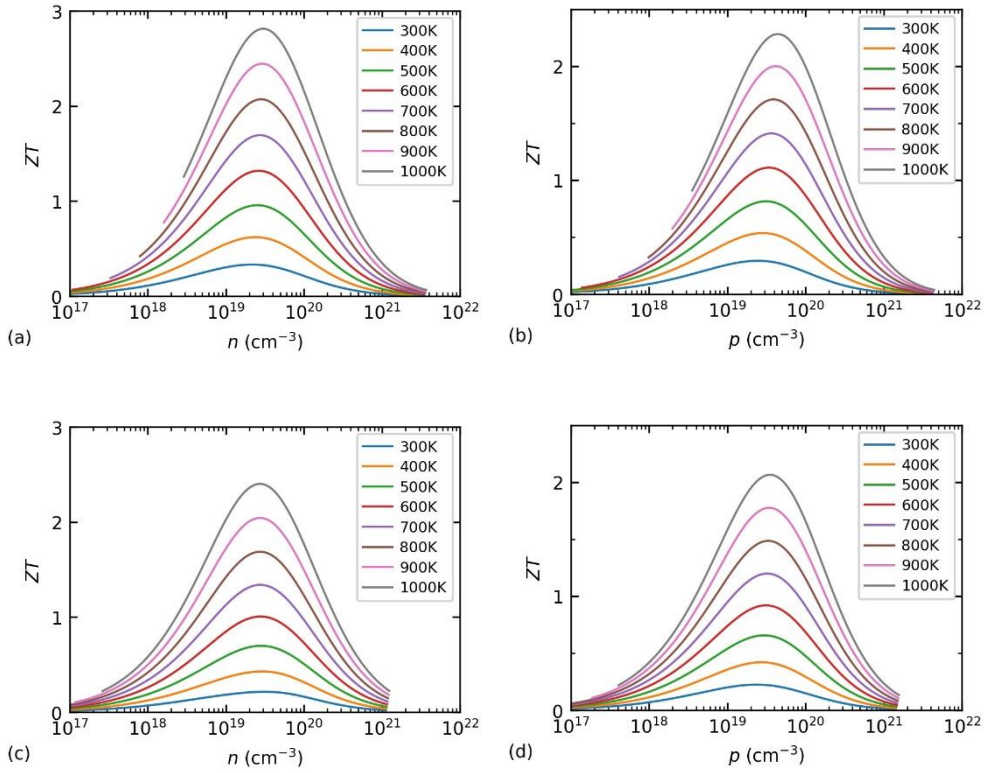


Figure S3. Figure of merit at varying temperatures and carrier concentrations for (a, b) PbSe and (c, d) PbS.

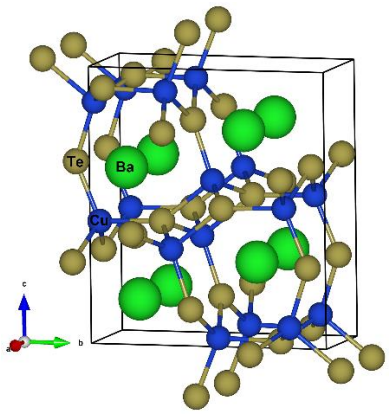
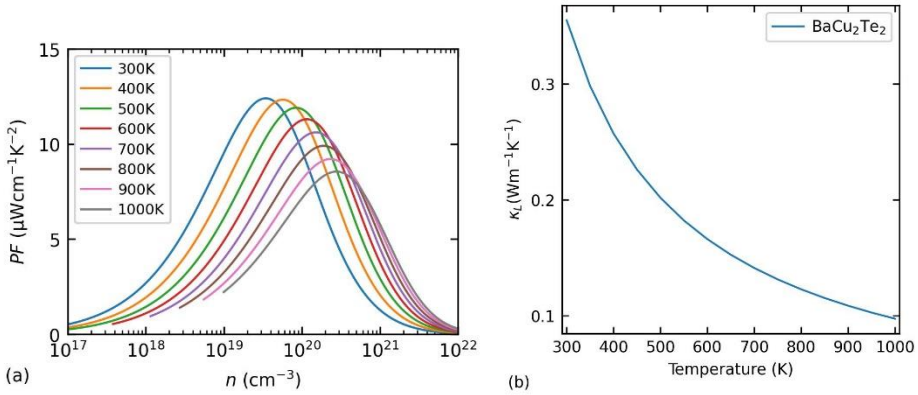


Figure S4. Crystal structures of $BaCu_2Te_2$.



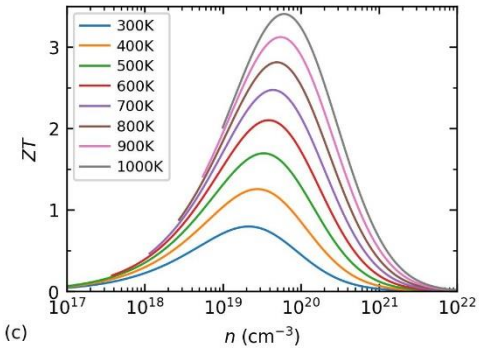


Figure S5. Thermoelectric properties of $BaCu_2Te_2$ for n type transport. (a) power factor, (b) lattice thermal conductivity, and (c) figure of merit.

Table S4. The optimized hyperparameters of the models listed in Table 2

	Model	Hyperparameters
N	GBDT	subsample=0.5, min_weight_fraction_leaf=0.003, learning_rate=0.1, max_depth=7, n_estimators=650
	XGB	tree_method='exact', importance_type='gain', scale_pos_weight=1.4404332, n_jobs=4, colsample_bynode=0.7, subsample=0.9, learning_rate=0.08, max_depth=3, n_estimators=100
	AdaB	base_estimator__max_depth=3, learning_rate=0.5, n_estimators=500
	RF	min_samples_split=2, class_weight='balanced_subsample', max_depth=10, max_features=0.2, n_estimators=100
P	GBDT	subsample=0.5, min_weight_fraction_leaf=0.0043, learning_rate=0.1, max_depth=4, n_estimators=600
	XGB	tree_method='exact', importance_type='gain', scale_pos_weight= 2.0954545, n_jobs=4, colsample_bynode=1.0, subsample=0.7, learning_rate=0.09, max_depth=2, n_estimators=500
	AdaB	base_estimator__max_depth=3, learning_rate=0.3, n_estimators=500
	RF	min_samples_split=2, class_weight='balanced_subsample', max_depth=9, max_features=0.4, n_estimators=300

Tabel S5. Performance measure on test sets of GBDT algorithm with different input feature vectors. The bold ones are the best for each doping type

	features	Acc.	Prec.	Recall	F1	AUC
N	PRDF10	0.86	0.79	0.87	0.83	0.93
	PRDF16	0.84	0.76	0.90	0.82	0.94
	PRDF20	0.86	0.78	0.90	0.84	0.94
	PRDF25	0.84	0.76	0.90	0.82	0.94
	GRDF_BOP	**0.86**	**0.78**	**0.90**	**0.84**	**0.97**
	VORONOI	0.86	0.76	0.94	0.84	0.96

P	PRDF10	0.89	0.90	0.75	0.82	0.96
	PRDF16	**0.92**	**0.95**	**0.79**	**0.86**	**0.96**
	PRDF20	0.89	0.9	0.75	0.82	0.95
	PRDF25	0.91	0.95	0.75	0.84	0.96
	GRDF_BOP	0.91	0.90	0.79	0.84	0.96
	VORONOI	0.91	0.87	0.83	0.85	0.97

Table S6. Performance measure on test sets of XGB algorithm with different input feature vectors. The bold ones are the best for each doping type

	features	Acc.	Prec.	Recall	F1	AUC
N	PRDF10	0.84	0.77	0.87	0.82	0.92
	PRDF16	0.83	0.74	0.90	0.81	0.93
	PRDF20	0.84	0.76	0.90	0.82	0.94
	PRDF25	0.83	0.75	0.87	0.81	0.95
	GRDF_BOP	0.84	0.76	0.90	0.82	0.96
	VORONOI	**0.88**	**0.81**	**0.94**	**0.87**	**0.95**
P	PRDF10	0.88	0.83	0.79	0.81	0.95
	PRDF16	0.86	0.84	0.67	0.74	0.94
	PRDF20	0.87	0.85	0.71	0.77	0.91
	PRDF25	0.89	0.86	0.79	0.83	0.96
	GRDF_BOP	**0.92**	**0.95**	**0.79**	**0.86**	**0.96**
	VORONOI	0.92	0.88	0.88	0.88	0.97

Table S7. Performance measure on test sets of AdaB algorithm with different input feature vectors. The bold ones are the best for each doping type

	features	Acc.	Prec.	Recall	F1	AUC
N	PRDF10	0.86	0.79	0.87	0.83	0.92
	PRDF16	**0.89**	**0.85**	**0.90**	**0.88**	**0.95**
	PRDF20	0.83	0.76	0.84	0.80	0.91
	PRDF25	0.86	0.79	0.87	0.83	0.93
	GRDF_BOP	0.89	0.83	0.94	0.88	0.96
	VORONOI	0.87	0.77	0.97	0.86	0.97
P	PRDF10	0.86	0.88	0.62	0.73	0.96
	PRDF16	0.89	0.86	0.79	0.83	0.93
	PRDF20	0.88	0.89	0.71	0.79	0.94
	PRDF25	**0.89**	**0.9**	**0.75**	**0.83**	**0.93**
	GRDF_BOP	0.88	0.89	0.71	0.79	0.94
	VORONOI	0.86	0.81	0.71	0.76	0.93

Table S8. Performance measure on test sets of RF algorithm with different input feature vectors. The bold ones are the best for each doping type

	features	Acc.	Prec.	Recall	F1	AUC
N	PRDF10	0.84	0.77	0.87	0.82	0.92

	PRDF16	0.84	0.76	0.90	0.82	0.91
	PRDF20	0.87	0.80	0.90	0.85	0.93
	PRDF25	0.82	0.73	0.87	0.79	0.92
	GRDF_BOP	**0.88**	**0.81**	**0.94**	**0.87**	**0.96**
	VORONOI	0.88	0.81	0.94	0.87	0.95
P	PRDF10	0.86	0.84	0.67	0.74	0.91
	PRDF16	0.86	0.84	0.67	0.74	0.91
	PRDF20	0.88	0.86	0.75	0.80	0.93
	PRDF25	0.88	0.86	0.75	0.80	0.94
	GRDF_BOP	**0.89**	**0.90**	**0.75**	**0.82**	**0.95**
	VORONOI	0.87	0.82	0.75	0.78	0.96

Table S9. Comparison of the performance on test sets of GBDT algorithm with input feature vector including band structure descriptors and not (woB)

	features	Acc.	Prec.	Recall	F1	AUC
N	GRDF_BOP	0.86	0.78	0.90	0.84	0.97
	GRDF_BOP_woB	0.80	0.77	0.74	0.75	0.88
	VORONOI	0.86	0.76	0.94	0.84	0.96
	VORONOI_woB	0.75	0.70	0.68	0.69	0.84
	PRDF16	0.84	0.76	0.90	0.82	0.94
	PRDF16_woB	0.78	0.79	0.61	0.69	0.83
P	GRDF_BOP	0.91	0.90	0.79	0.84	0.96
	GRDF_BOP_woB	0.80	0.70	0.67	0.68	0.82
	VORONOI	0.91	0.87	0.83	0.85	0.97
	VORONOI_woB	0.80	0.68	0.71	0.69	0.81
	PRDF16	0.92	0.95	0.79	0.86	0.96
	PRDF16_woB	0.83	0.74	0.71	0.72	0.80

Table S10. Comparison of the performance on test sets of XGB algorithm with input feature vector including band structure descriptors and not (woB)

	features	Acc.	Prec.	Recall	F1	AUC
N	GRDF_BOP	0.84	0.76	0.90	0.82	0.96
	GRDF_BOP_woB	0.80	0.77	0.74	0.75	0.88
	VORONOI	0.88	0.81	0.94	0.87	0.95
	VORONOI_woB	0.75	0.70	0.68	0.69	0.84
	PRDF16	0.83	0.74	0.90	0.81	0.93
	PRDF16_woB	0.78	0.79	0.61	0.69	0.83
P	GRDF_BOP	0.92	0.95	0.79	0.86	0.96
	GRDF_BOP_woB	0.80	0.70	0.67	0.68	0.82
	VORONOI	0.92	0.88	0.88	0.88	0.97
	VORONOI_woB	0.80	0.68	0.71	0.69	0.81
	PRDF16	0.86	0.84	0.67	0.74	0.94
	PRDF16_woB	0.83	0.74	0.71	0.72	0.80

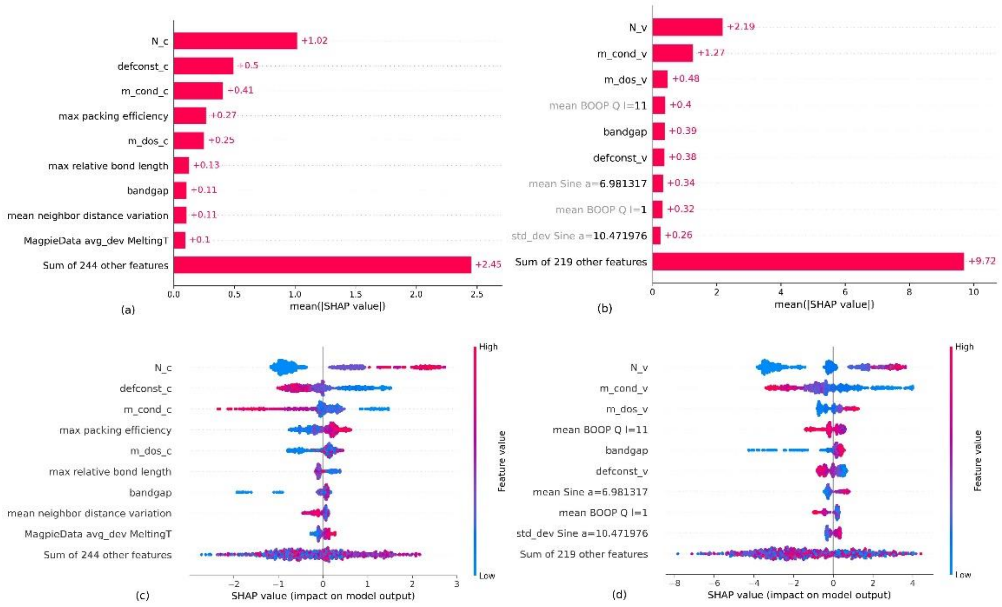


Figure S6. SHAP analysis based on XGB models in Table 2. (a)(c) for n type data, (b)(d) for p type data. (a)(b) bar chart of the average SHAP value magnitude for each input feature. (c)(d) a set of beeswarm plots, where each dot corresponds to a sample. The dot's position on the *x* axis shows the impact that feature has on the model's prediction for that sample. When multiple dots locate at the same x position, they pile up to show density.

Table S11. The optimized hyperparameters of the MEGNet models listed in Table 3

Hyperparameters	N	P
Initial learning rate	0.001	
Final learning rate	0.0001	
Weight decay	0.0001	0.0056
nblocks	3	
hidden_layer_sizes_input	(8,4)	
hidden_layer_sizes_conv	(8,8,4)	
nlayers_set2set	1	
niters_set2set	2	
hidden_layer_sizes_output	(8,8)	
cutoff	5.0	

Table S12. The optimized hyperparameters of the M3GNet models listed in Table 3

Hyperparameters	N	P
Initial learning rate	0.001	
Final learning rate	0.0001	
Weight decay	0.0026	0.0048
nblocks	3	
dim_node_embedding	32	8
dim_edge_embedding	32	8

nlayers_set2set	1	
niters_set2set	3	
units	32	8
threebody_cutoff	5.0	

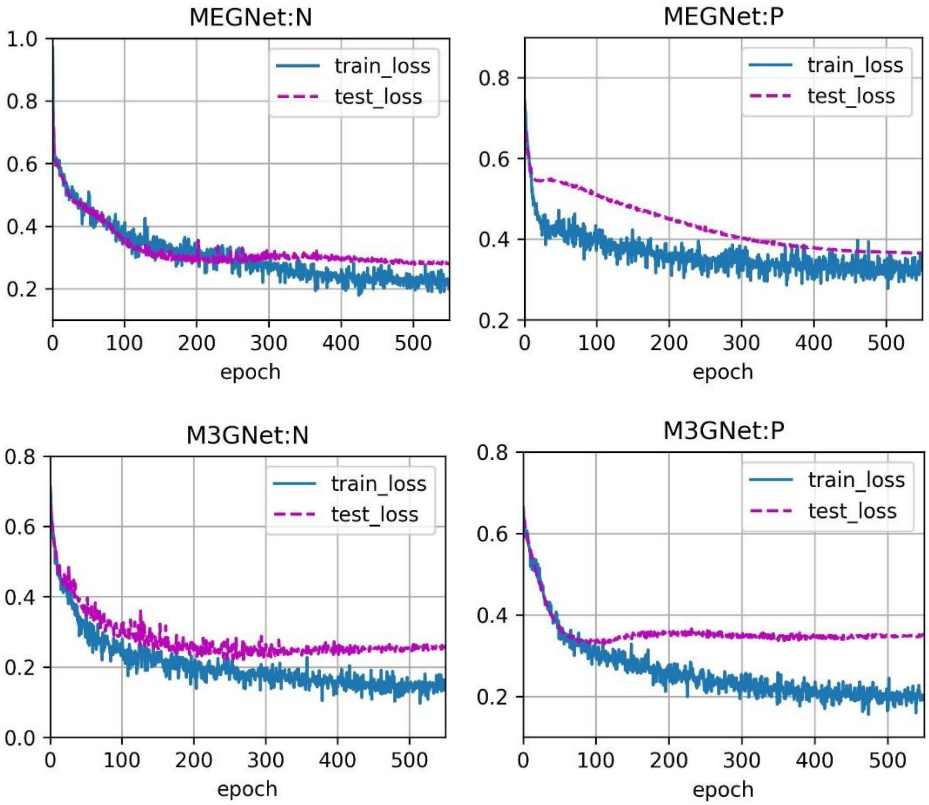


Figure S7. The final training process of the MEGNet and M3GNet models listed in Table 3.

Reference

[1] Nielsen, M. D., Ozolins, V., & Heremans, J. P. (2013). Lone pair electrons minimize lattice thermal conductivity. *Energy & Environmental Science*, *6*(2), 570-578.

[2] Zeier, W. G., Zevalkink, A., Gibbs, Z. M., Hautier, G., Kanatzidis, M. G., & Snyder, G. J. (2016). Thinking like a chemist: intuition in thermoelectric materials. *Angewandte Chemie International Edition*, *55*(24), 6826-6841.