Supporting Information

Coupling the High-Throughput Property Map to Machine Learning for Predicting Lattice Thermal Conductivity

Rinkle Juneja†‡, George Yumnam†‡, Swanti Satsangi†, and Abhishek K. Singh†

†Equal contribution
‡Materials Research Centre, Indian Institute of Science, Bangalore 560012, India

Lattice thermal conductivity, phonon dispersion, and Grüneisen parameter of 116 compounds

For all the screened compounds, lattice thermal conductivity, phonon dispersion, and Grüneisen parameter are calculated. The 2 ultralow and 2 ultrahigh $\kappa_l$ compounds are discussed in the main manuscript. The remaining 116 are presented here. Figures S1 - S5 show the lattice thermal conductivity, Figures S6 - S10 show the phonon dispersion, and Figures S11 - S15 show the Grüneisen parameter, respectively.

![Graphs showing lattice thermal conductivity, phonon dispersion, and Grüneisen parameter](image)

Figure S1: Calculated DFT lattice thermal conductivity as a function of temperature for the nonmetallic and dynamically stable compounds.
Figure S2: Calculated DFT lattice thermal conductivity as a function of temperature for the nonmetallic and dynamically stable compounds.

Figure S3: Calculated DFT lattice thermal conductivity as a function of temperature for the nonmetallic and dynamically stable compounds.
Figure S4: Calculated DFT lattice thermal conductivity as a function of temperature for the nonmetallic and dynamically stable compounds.

Figure S5: Calculated DFT lattice thermal conductivity as a function of temperature for the nonmetallic and dynamically stable compounds.
Figure S6: Phonon dispersion for the nonmetallic and dynamically stable compounds.

Figure S7: Phonon dispersion for the nonmetallic and dynamically stable compounds.
Figure S8: Phonon dispersion of nonmetallic and dynamically stable compounds.

Figure S9: Phonon dispersion for the nonmetallic and dynamically stable compounds.
Figure S10: Phonon dispersion for the nonmetallic and dynamically stable compounds.

Figure S11: Grüneisen parameter as a function of frequency for the nonmetallic and dynamically stable compounds.
Figure S12: Grüneisen parameter as a function of frequency for the nonmetallic and dynamically stable compounds.

Figure S13: Grüneisen parameter as a function of frequency for the nonmetallic and dynamically stable compounds.
Figure S14: Grüneisen parameter as a function of frequency for the nonmetallic and dynamically stable compounds.

Figure S15: Grüneisen parameter as a function of frequency for the nonmetallic and dynamically stable compounds.
Learning curve

Figure S16 shows the learning curve for the developed model. As a function of increasing training data, the train and test rmse converge around 0.2, showing no overfitting for 90 percent of training data.

![Learning curve graph]

Figure S16: Learning curve for the developed GPR model.

Table S1: Comparison of DFT calculated and experimental lattice thermal conductivity $T = 300$ K.

<table>
<thead>
<tr>
<th>Compound</th>
<th>$\kappa_l^{DFT}$ (W/mK)</th>
<th>$\kappa_l^{exp}$ (W/mK)</th>
</tr>
</thead>
<tbody>
<tr>
<td>PbF$_2$ (Fm-3m)</td>
<td>1.33</td>
<td>1.4 [1]</td>
</tr>
<tr>
<td>PbSe (Fm-3m)</td>
<td>1.47</td>
<td>2 [2]</td>
</tr>
<tr>
<td>PbS (Fm-3m)</td>
<td>1.00</td>
<td>2.9 [2]</td>
</tr>
<tr>
<td>InP (F-43m)</td>
<td>79.01</td>
<td>68 [3]</td>
</tr>
<tr>
<td>AlAs (F-43m)</td>
<td>76.30</td>
<td>80 [3]</td>
</tr>
<tr>
<td>AlP (F-43m)</td>
<td>76.20</td>
<td>90 [3]</td>
</tr>
<tr>
<td>CdTe (F-43m)</td>
<td>9.18</td>
<td>7.5 [2]</td>
</tr>
<tr>
<td>GaAs (F-43m)</td>
<td>35.42</td>
<td>44 [3]</td>
</tr>
<tr>
<td>GaP (F-43m)</td>
<td>124.19</td>
<td>100 [2]</td>
</tr>
<tr>
<td>CdS (P6$_3$mc)</td>
<td>13.87</td>
<td>16 [2]</td>
</tr>
</tbody>
</table>

Table S2: Comparison of DFT and model predicted lattice thermal conductivity of 8 independent compounds at $T = 300$ K.

<table>
<thead>
<tr>
<th>Compound</th>
<th>$\ln(\kappa_l)^{DFT}$ (W/mK)</th>
<th>$\ln(\kappa_l)^{ML}$ (W/mK)</th>
</tr>
</thead>
<tbody>
<tr>
<td>BaO (Fm-3m)</td>
<td>0.92</td>
<td>0.97</td>
</tr>
<tr>
<td>LiF (Fm-3m)</td>
<td>2.54</td>
<td>3.70</td>
</tr>
<tr>
<td>LiCl (Fm-3m)</td>
<td>1.20</td>
<td>2.73</td>
</tr>
<tr>
<td>ZnO (F-43m)</td>
<td>4.06</td>
<td>4.89</td>
</tr>
<tr>
<td>CuCl (F-43m)</td>
<td>0.46</td>
<td>1.51</td>
</tr>
<tr>
<td>CdO (Fm-3m)</td>
<td>2.01</td>
<td>1.89</td>
</tr>
<tr>
<td>KH (Fm-3m)</td>
<td>2.14</td>
<td>2.38</td>
</tr>
<tr>
<td>InAs (F-43m)</td>
<td>3.00</td>
<td>2.06</td>
</tr>
</tbody>
</table>
**GPR model using Slack descriptors**

A machine learning prediction model is developed using the Slack descriptors in order to have the comparison with the high-throughput proposed descriptors. Figure S17(a) shows the scatter plot for the DFT calculated versus the ML predicted $\kappa_l$ using Slack descriptors and (b) shows the corresponding residual. The rmse for this model is $\sim 0.30$, which is higher than the prediction model using high-throughput proposed descriptors.

![Figure S17](image)

Figure S17: (a) The scatter plot of DFT calculated log-scaled $\kappa_l$ versus ML model predicted $\kappa_l$ using Slack descriptors, (b) shows the difference between DFT and Slack ML $\kappa_l$.

**SISSO model using proposed descriptors and Slack descriptors**

For the identification of suitable formula using different combination of proposed descriptors ($\omega_{\text{max}}, \gamma_{c=3}, M, \text{and} V$) and Slack descriptors (n, $\theta_D$, $\gamma$, $\gamma$, $\text{and} V$), we used the sure independence screening and sparsifying operator (SISSO) [4] method. Starting from these primary features $\phi_0$ for both the cases, different mathematical operations were applied iteratively to generate feature spaces ($\phi_n$) of increasing complexity. The feature spaces $\phi_1$, $\phi_2$, and $\phi_3$ in the case of proposed descriptors consist of 70, 10860, and 350767452 features, respectively. For the Slack descriptors, $\phi_1$, $\phi_2$, and $\phi_3$ have 101, 21406, and 1364986543 features, respectively. For the case of proposed descriptors, the 1D, 2D, 3D, and 4D descriptor SISSO model for log-scaled $\kappa_l$, respectively, are

1. $\ln(\kappa_l)_{1D} = -0.207 \left( (\gamma_{c=3} + (V/M)) - (V^\frac{1}{2} \omega_{\text{max}}) \right) + 2.260$ (S1)

2. $\ln(\kappa_l)_{2D} = -0.198 \left( (\gamma_{c=3} + (V/M)) - (\ln(V) \sqrt{\omega_{\text{max}}}) \right) - 0.004 \left( \frac{M \mid (M - V) \mid}{(M + \omega_{\text{max}}) + (\omega_{\text{max}} - V)} \right) + 2.208$ (S2)

3. $\ln(\kappa_l)_{3D} = -0.209 \left( (\gamma_{c=3} + (V/M)) - (\ln(V) \sqrt{\omega_{\text{max}}}) \right) - 0.004 \left( \frac{\sqrt{\gamma_{c=3}} \mid (M - V) \mid}{(M + \omega_{\text{max}}) + (\omega_{\text{max}} - V)} \right) - 0.0003 \left( \frac{(MV) \mid (M - V) \mid}{(M - V) - \sqrt{\gamma_{c=3}}} \right) + 2.317$ (S3)

4. $\ln(\kappa_l)_{4D} = 0.488 \left( (\ln(M) \omega_{\text{max}}^\frac{1}{2}) - (\gamma_{c=3}^\frac{1}{2}) \right) - 0.419 \left( \frac{\mid (M - V) \mid}{\ln(M)} \right) - 0.023 \left( \frac{\omega_{\text{max}} \mid (M - V) \mid}{(M - V) - \sqrt{\gamma_{c=3}}} \right) - 0.013 \left( \frac{(V + \omega_{\text{max}})}{\ln(M)} \right) + 1.614$ (S4)
The accuracy of best solution obtained using SISSO increases with increase in dimensionality. For example, the rmse for 1D, 2D, 3D, and 4D is 0.94, 0.80, 0.72, and 0.66, respectively. Similar analysis was carried out for the Slack descriptors. For the case of Slack descriptors, the 1D, 2D, 3D, and 4D descriptor model for log-scaled $\kappa_l$, respectively, are

$$
\ln(\kappa_l)_{1D} = -0.062 \left( \left( M - n \right) / \sqrt{n} \right) + 6.50 \tag{S5}
$$

$$
\ln(\kappa_l)_{2D} = -0.062 \left( \left( M - n \right) / \sqrt{n} \right) + 0.0038 \left( \frac{V - \theta_D}{\ln(\gamma)} \right) + 6.50 \tag{S6}
$$

$$
\ln(\kappa_l)_{3D} = -0.867 \left( \sqrt{\frac{n}{\ln (n)}} - \frac{n^\gamma}{\ln (n)} \right) + 8.822 \left( \frac{V^{-1}/\ln(\gamma)}{(M/n) - (n/\gamma)} \right) - 2.344 \left( \frac{1}{V/\gamma - (V + \theta_D)} \right) + 10.98 \tag{S7}
$$

$$
\ln(\kappa_l)_{4D} = -0.841 \left( \gamma n^4 + \frac{\sqrt{V}}{\ln (n)} \right) - 2178.01 \left( \frac{\exp(n)/V^3}{V \gamma - (n + V)} \right) - 0.193 \left( \frac{1}{V/\gamma - (V + \theta_D)} \right) - 0.00005 \left( \frac{V^2/\ln(\gamma)}{(M/\gamma) - (M - V)} \right) + 11.29 \tag{S8}
$$

In this case also, the accuracy improves as we increase the dimensionality and complexity. The rmse for 1D, 2D, 3D, and 4D in this case is 1.01, 0.84, 0.72, and 0.63, respectively. For the less complex low dimensional Slack descriptor models, the rmse is higher than the models from proposed descriptors.

References


