Supplement for CrysGraphFormer: An Equivariant Graph Transformer for Prediction of Lattice Thermal Conductivity with Interpretability

Zhengyu Sun[a ‡], Weiwei Sun[*b ‡], Shaohan Li[c], Zening Yang[*a], Mutian Zhang[a], Yang Yang[a], Huayun Geng[d] and Jin Yu[*a]

a. School of Materials Science and Engineering, Southeast University, Nanjing 211189, China. E-mail: yujin@seu.edu.cn; E-mail: yangzening@seu.edu.cn

b. Key Laboratory of Quantum Materials and Devices of Ministry of Education, School of Physics, Southeast University, Nanjing 211189, China. E-mail: provels8467@gmail.com

c. Nanjing University of Science and Technology ZiJin College, Nanjing 211189, China.

d. National Key Laboratory of Shock Wave and Detonation Physics, Institute of Fluid Physics, CAEP; P.O. Box 919-102 Mianyang, Sichuan, P. R. China, 621900

1. Detailed tables

Table S1. Extra features descriptors.

Feature descriptors
mean absolute deviation in relative bond length
max relative bond length
min relative bond length
mean absolute deviation in relative cell size
mean bond length
std_dev Average bond length
mean CN_VoronoiNN
std_dev CN_VoronoiNN
density
vpa
packing fraction
a
b
c
alpha
beta
gamma
natoms
max multiplicity
min multiplicity

sum multiplicity

The calculation formulas for some features are as follows:

mean absolute deviation in relative bond length(mean_abs_dev):

$$mean\_abs\_dev = \frac{\sum_{i=1}^{n} | non\_zero\_bond\_lengths_i - mean\_bond\_len |}{n} \tag{S1}$$

max relative bond length(max_relative_len):

$$max\_relative\_len = \frac{max(non\_zero\_bond\_lengths)}{mean\_bond\_len} \tag{S2}$$

min relative bond length(min_relative_len):

$$min\_relative\_len = \frac{min(non\_zero\_bond\_lengths)}{mean\_bond\_len} \tag{S3}$$

mean absolute deviation in relative cell size(mean_abs_dev_cell_size):

$$mean\_a\_b\_c = \frac{a/b + b/c + c/a}{3} \tag{S4}$$

$$mean\_abs\_dev\_cell\_size = \frac{|a/b - mean\_a\_b\_c| + |b/c - mean\_a\_b\_c| + |c/a - mean\_a\_b\_c|}{3} \tag{S5}$$

mean bond length(mean_bond_len):

$$mean\_bond\_len = \frac{\sum_{i=1}^{n} non\_zero\_bond\_lengths_i}{n} \tag{6}$$

std_dev Average bond length(std_dev_bond_len):

$$std\_dev\_bond\_len = \sqrt{\frac{\sum_{i=1}^{n} (non\_zero\_bond\_lengths_i - mean\_bond\_len)^2}{n}} \tag{7}$$

Where n is the number of non-zero bond lengths in the structure.

mean CN_VoronoiNN(mean_CN_Voronoi):

$$mean\_CN\_Voronoi = \frac{\sum_{i=1}^{N} CN_{voronoi,i}}{N} \tag{8}$$

Where N is the number of atoms in the structure and CNvoronoi is the coordination number of the i-th atom.

std_dev CN_VoronoiNN(std_dev_CN_Voronoi):

$$std\_dev\_CN\_Voronoi = \sqrt{\frac{\sum_{i=1}^{N}(CN_{voronoi,i} - mean\_cn\_voronoi)^2}{N}} \quad (9)$$

Where N is the number of atoms in the structure and CNvoronoi is the coordination number of the i-th atom.

density: The mass per unit volume of a substance, often measured in g/cm³ for solids.

vpa (Volume per atom): The total volume of the unit cell divided by the number of atoms in the unit cell, indicating the space occupied by each atom.

packing fraction(pf): The fraction of the volume of the unit cell that is actually occupied by the atoms, representing how tightly the atoms are packed.

lattice constants: a, b, c, alpha, beta, gamma.

natoms: the number of atoms in the unit cell.

max multiplicity: The highest multiplicity of atoms in the unit cell, indicating the maximum number of equivalent positions an atom can occupy.

min multiplicity: The lowest multiplicity of atoms in the unit cell, indicating the minimum number of equivalent positions an atom can occupy.

sum multiplicity: The total sum of the multiplicities of all atoms in the unit cell, reflecting the overall symmetry and equivalent positions in the structure.

Table S2. shows the hyperparameter configuration of CrysGraphFormer trained on the lattice thermal conductivity dataset.

Parameters	Value
CrysGraphFormer Layers	6
Multi Heads	8
Angle SBF dim	256
Bond RBF dim	80
Hidden dim	256
Atom dim	93
Learning rate	0.001
Batch size	64
Cutoff	6
Epochs	400
Weight decay	1e-5
warmup_steps	2000

Table S3. shows the hyperparameter configuration of other model trained on the lattice thermal conductivity dataset.

Parameters	CGCNN	MEGNET	GATGNN	ALIGNN	Matformer	DeeperGATGNN	CrystalFormer
lr	0.01	0.001	0.005	0.001	0.001	0.005	0.0005

Batch size	256	64	256	64	64	100	256
Epochs	300	1000	200	300	300	500	500

Table S4. Lattice thermal conductivity of 59 thermoelectric materials as predicted by the model versus the results of DFT calculations. $LTC_{Predict\ (1)}$ (W m^{-1} K^{-1}) is the prediction result using 6 layers of CrysGraphFormerLayer, $LTC_{Predict\ (2)}$ (W m^{-1} K^{-1}) is the prediction result after removing all designed modules, and $LTC_{Predict\ (3)}$ (W m^{-1} K^{-1}) is the prediction result with only 2 layers of CrysGraphFormerLayer.

Formula	Space Group	LTC_{DFT}	$LTC_{Predict\ (1)}$	$LTC_{Predict\ (2)}$	$LTC_{Predict\ (3)}$
Ag4O2	224	2.14	0.24	1.28	0.36
Ag4O4	14	3.9	0.64	1.62	0.68
Ag4O6	43	8.77	0.88	1.60	0.88
Al4Ru2	70	14.19	9.25	2.60	10.56
As4Cd2	98	2.53	1.75	1.65	1.25
As4Te6	12	1.03	3.04	1.75	2.10
As8Na8	19	0.55	1.21	1.65	1.30
As8Na8	14	0.84	1.15	1.63	1.30
As8Rb8	19	0.33	0.51	1.42	0.59
Ba2Se4	15	1.31	2.04	1.96	2.01
Bi2Se3	166	2.32	2.59	1.92	3.56
Bi4I4	12	0.82	0.81	1.43	1.40
Bi4Se8	12	0.34	2.21	1.69	2.04
Br4Te2	136	1.25	0.94	1.33	0.92
C4Os4	198	22.16	13.58	3.31	12.06
Cl2Cu1	12	6.21	1.38	1.66	1.14
Co6O8	227	1.32	3.75	2.24	3.28
Cr3Si6	180	20.51	18.35	3.96	15.41
Cr3Si6	181	21.22	18.5	3.96	13.59
Cs4Te16	14	0.28	0.42	1.22	0.68
Cu4O2	224	4.43	0.24	1.05	0.21
F4Hg2	136	1.9	0.77	1.68	0.87
F6Mn2	55	7.64	3.87	2.52	4.79
Fe4Si4	198	22.62	10.35	3.51	14.26
Ga4Os2	70	8.4	8.26	2.31	6.40
Ge4O8	205	14.35	8.34	2.94	8.77
Ge4Se4	62	2.48	2.65	2.10	3.04
Hf1Se2	164	5.63	6.35	3.00	7.96

Hf2Se6	11	4.55	4.08	2.26	4.12
In4S4	64	1.99	1.39	1.71	1.05
I4Te2	136	2.57	0.65	1.22	0.58
Ir4N8	14	25.43	13.73	3.21	10.17
K4O2	141	2.19	0.92	1.52	0.95
K6Te6	189	0.38	0.4	1.25	0.61
K8Te12	62	0.34	0.38	1.23	0.44
Li8P8	14	2.02	3.45	2.58	5.36
Mo3Si6	180	14.96	17.21	3.94	16.24
N6W3	152	5.07	5.1	2.66	4.19
N8W4	33	7.35	3.24	2.60	3.35
Na6S6	189	1.67	1.34	1.84	1.82
Na8P8	19	0.89	2.4	1.88	2.36
O11Zr6	25	3.04	6.86	2.68	4.57
O6Pb4	31	0.89	2.52	2.12	2.50
O6Ti4	167	7.64	13.86	2.77	8.56
O12Ti8	62	5.17	8.94	2.87	6.17
Os2P4	58	30.53	19.72	3.58	16.11
Os4Si4	198	16.76	12.66	3.71	12.67
Pb2S2	186	1.76	2.77	2.07	2.56
Rb2Te5	12	0.48	0.89	1.25	0.75
Rb6Te6	189	0.39	0.28	1.12	0.41
Ru4Si4	198	12.88	12.69	3.60	12.77
Sc4Se6	15	1.76	2.69	2.22	3.34
Sc4Te6	15	1.09	1.95	1.98	1.80
Se2Sn2	63	1.27	2.33	2.02	1.78
Se2Sn2	129	2.98	1.59	2.18	1.19
Se2Zr1	164	5.57	6.63	2.92	6.47
Se4Sn4	62	2.42	1.96	1.99	1.35
Se6Zr2	11	4.48	3.9	2.06	3.54
Si6W3	180	11.83	13.48	3.80	15.34

Table S5. The comparison of the lattice thermal conductivity of 55 ternary semiconductor materials predicted by the model versus the results of DFT calculations.

Formula	Material_id	LTC_{DFT}	$LTC_{Predict}$

Ba2Cu4S4	mp-4255	1.2333	2.5017
Y2TeS2	mp-1216058	6.6667	4.9043
Ba2Cd4P4	mp-8279	5.0000	3.5302
As2Ca2Rb2	mp-9845	1.5000	1.3108
K2Li2S2	mp-8430	2.6000	1.7357
Ag2Te4Y2	mp-12903	0.6667	1.8274
K8Rb4Sb4	mp-976148	0.2000	0.6906
Bi4Cs4K8	mp-867339	0.2000	0.4534
Li8Sb4Tl4	mp-1184973	0.5000	1.2813
K4Na4S4	mp-28383	1.4333	1.0801
Li2Rb2Se2	mp-9250	1.6667	1.0785
CsP3Zn4	mp-1101093	1.8333	2.3335
Na4SeTe	mp-1221139	2.3667	1.0591
Cs2Na2Se2	mp-8658	0.8667	0.6424
K4Rb4S4	mp-28760	0.5667	0.5941
Ca2Rb2Sb2	mp-9846	1.3000	1.3537
AgIn5Se8	mp-571103	2.2000	1.0801
Cd4P4Sr2	mp-8277	3.3333	3.9464
AgIn5Te8	mp-569813	1.4667	0.7829
Ba2Ga4Te8	mp-38117	1.5000	2.0516
As3Cd4Cs	mp-1079080	0.6667	0.5657
Cs4K8Sb4	mp-581024	0.2000	0.6940
K2Mg2P2	mp-1018737	4.0000	4.0498
Bi2Ca2K2	mp-773137	1.0333	1.3456
Li2Rb2S2	mp-8751	1.6000	1.3268
Ba2Mg4P4	mp-8278	5.5333	6.4007
Na2Rb2S2	mp-8799	1.2333	1.0778
As4Ba2Mg4	mp-8280	5.0000	5.9652
Cs2Na2Te2	mp-5339	0.9333	0.4932
Pb2S4Sn2	mp-1218951	1.0333	2.1290
In2P2S8	mp-20790	3.2333	2.6868
As3RbZn4	mp-975144	1.0667	1.3071
Ag2K2Se2	mp-16236	0.3667	0.9626
In4Mg2Te8	mp-1222182	3.2000	1.6165

Na8Rb4Sb4	mp-975275	0.4000	0.9537
As2K2Mg2	mp-1019089	3.7000	2.6392
Ba3Bi3Na3	mp-31235	1.8667	0.8098
As2Ba2Li2	mp-10616	3.1667	3.5054
CaPbSe2	mp-1227206	3.7333	2.8995
PbS4Sr3	mp-1218560	2.4667	2.4236
K4Na4Se4	mp-28595	1.3000	0.8906
As2Cd2K2	mp-1018736	1.3000	0.6691
Al4Ba2Te8	mp-38394	1.6667	2.5234
Bi4K4Na8	mp-863707	0.8000	0.8299
K4Na4Te4	mp-8755	0.8667	0.6923
Ba4N4Zn2	mp-9307	2.5333	2.8852
Bi2K2Mg2	mp-1019105	2.2667	0.9111
Ga4Sr2Te8	mp-33880	1.2667	2.0259
K2Mg2Sb2	mp-7089	3.6000	1.9295
Ba3Na3P3	mp-9732	1.3000	2.2953
K2Li2Te2	mp-4495	2.3667	1.0025
Cs2Na2S2	mp-6973	1.0333	0.8893
As2K2Li4	mp-28994	2.3000	1.9935
Ba2In4Te8	mp-33985	1.3000	1.3955
Ag4As4S4	mp-984714	0.7000	0.6798

2.Evaluate metrics

Calculation formula of mean absolute error (MAE):

$$MAE(x,y)=\frac{\sum_{i=1}^{n}\left|y_i-x_i\right|}{n} \tag{S10}$$

Here, x is the predicted value, y is the target value, and n is the total number of samples.

The calculation formula of R-squared (R^2):

$$R^2=1-\frac{\sum_i (y'^i-y^i)^2}{\sum_i (\bar{y}-y^i)^2} \tag{S11}$$

Here y' is the predicted value, y^i is the target value, and $\bar{y}$ is the mean of the sample.

3. Detailed figures

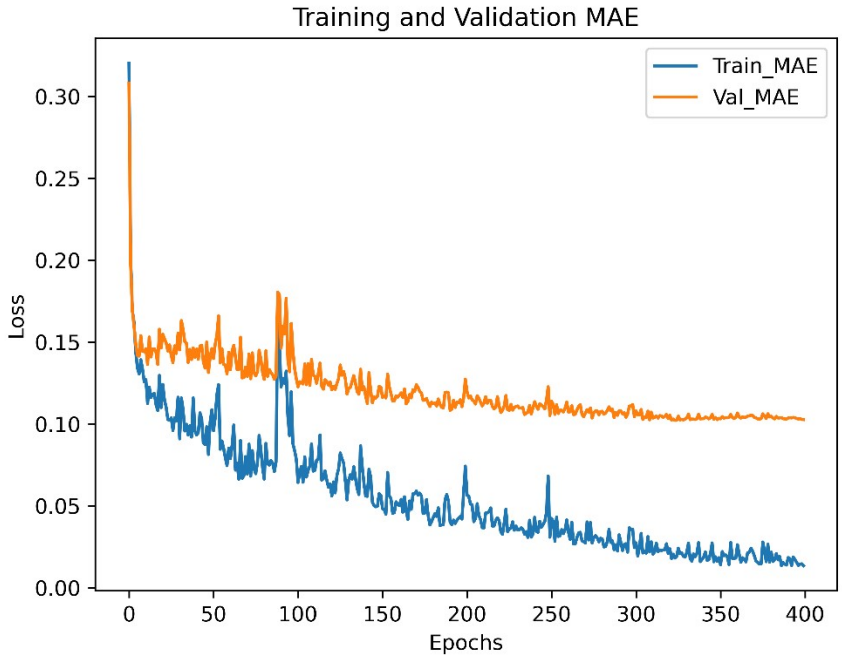

Fig. S1 shows the training loss curve of CrysGraphFormer on the lattice thermal conductivity dataset.

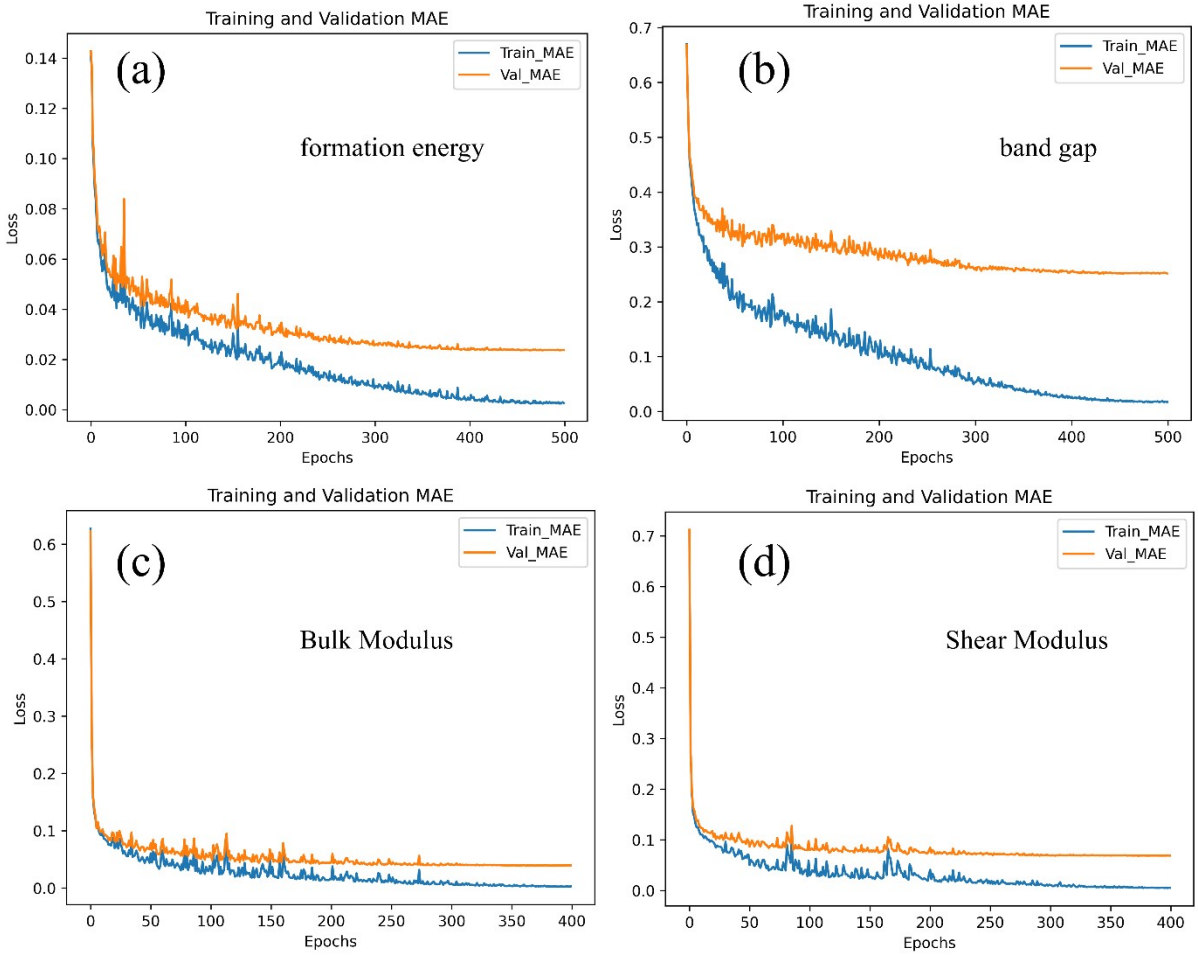

Fig. S2 Training loss curve of CrysGraphFormer on the mp-2018.6.1 dataset. (a) Training loss curve on the energy dataset. (b) Training loss curve on the band gap dataset. (c) Training loss curve on bulk modulus dataset. (d) Training loss curve on shear modulus dataset.

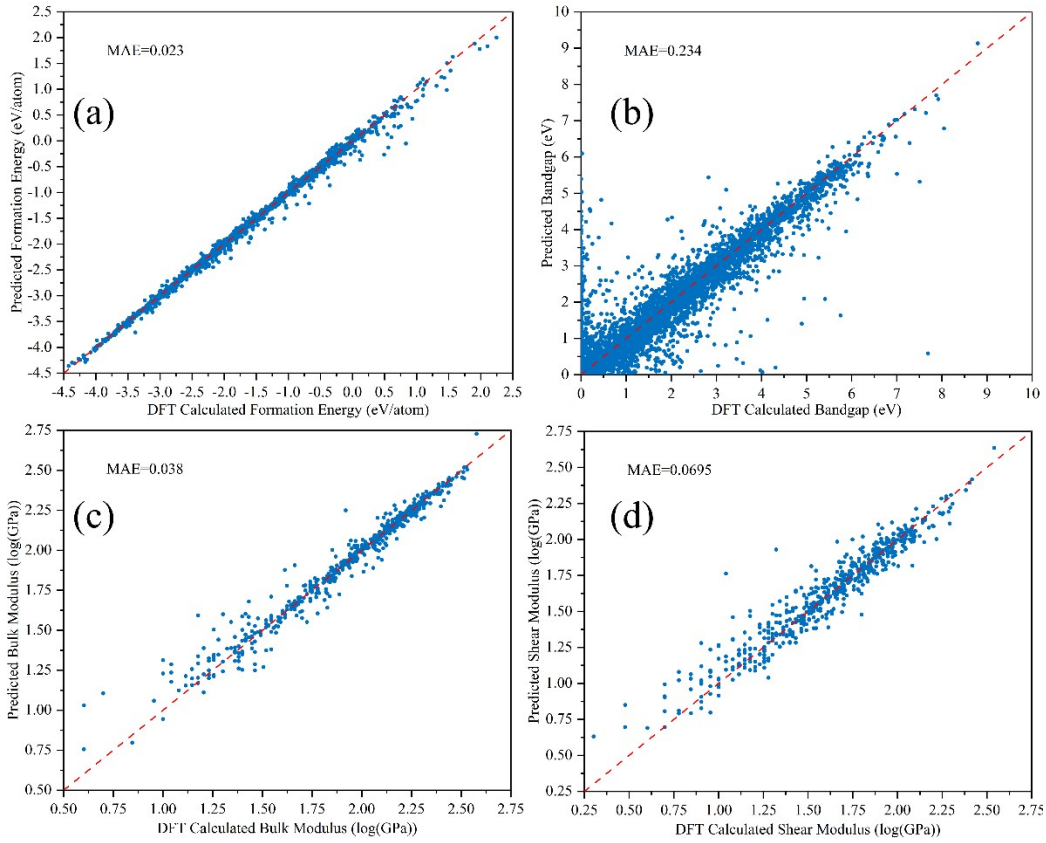

Fig. S3 Comparison of the error between predicted value and target value of CrysGraphFormer on the mp-2018.6.1 test set. (a) Form a prediction error graph on the test set. (b) Prediction error plot on the bandgap test set. (c) Prediction error plot on the bulk modulus test set. (d) Prediction error plot on the shear modulus test set.

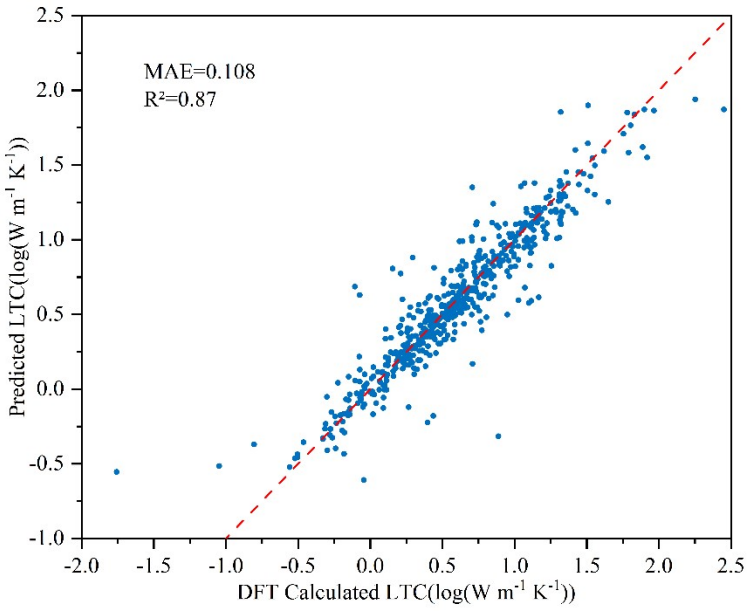

Fig. S4 Comparison of the error between the predicted value and the target value of CrysGraphFormer on the lattice thermal conductivity test set.

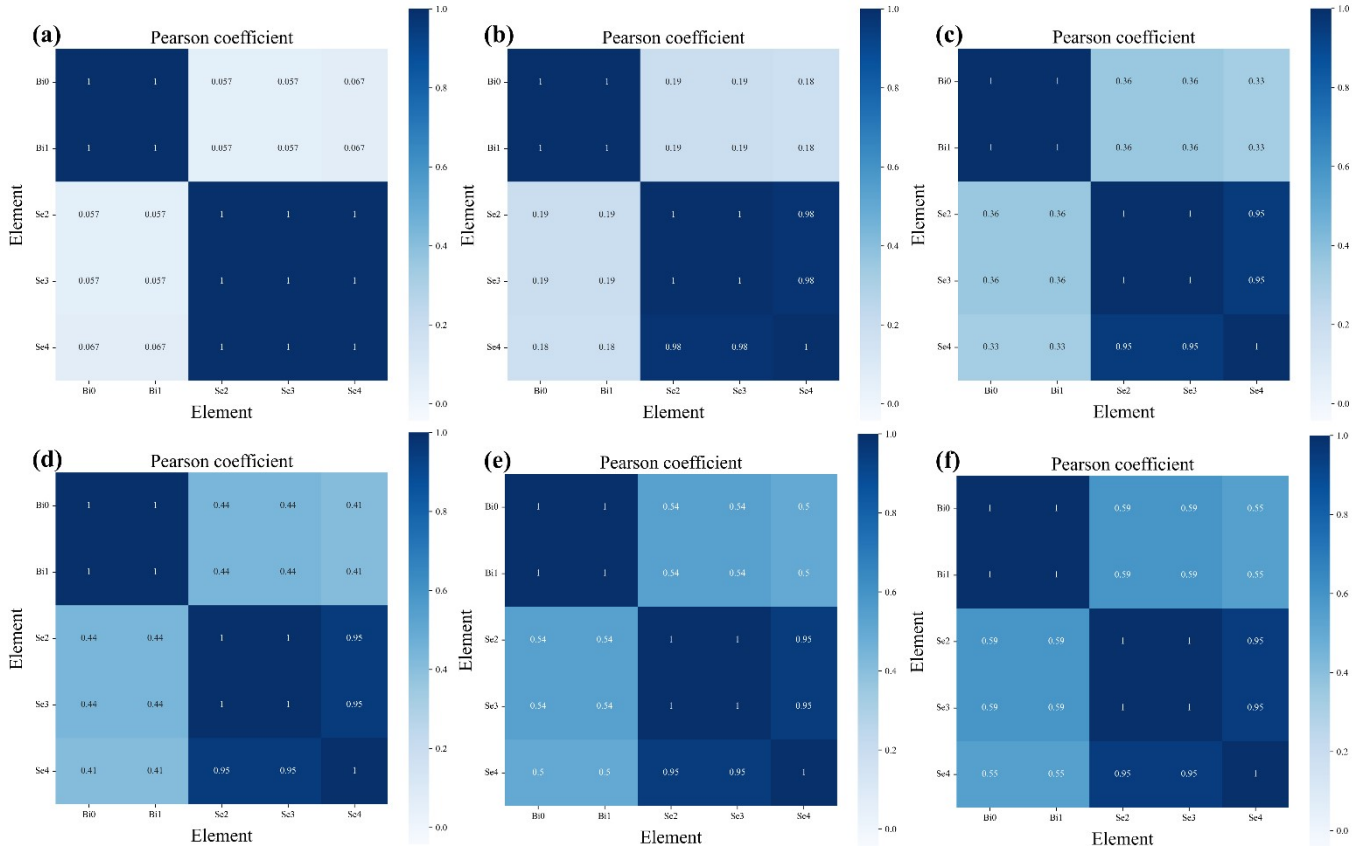

Fig. S5 Correlation analysis of atomic features in Bi_2Se_3. Panels (a) to (f) represent the features extracted from the 1st to the 6th layers of the CrysGraphFormerLayer, respectively.

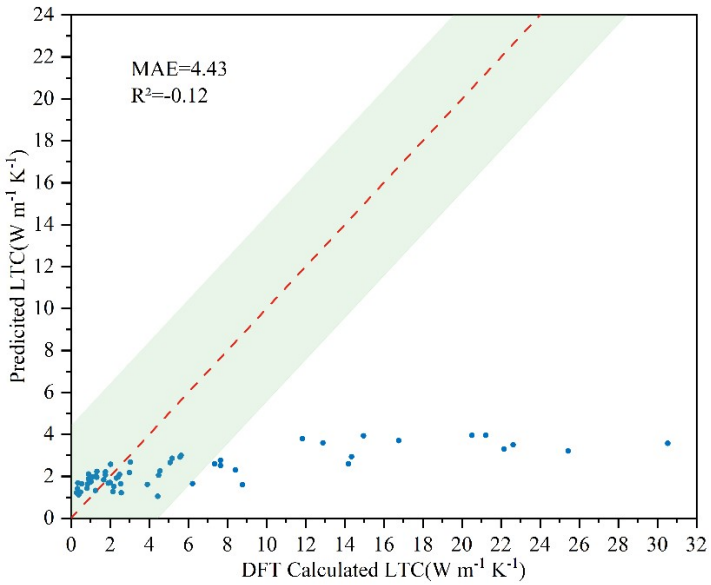

Fig. S6 The prediction results of 59 thermoelectric materials using CrysGraphFormer model with only a message-passing mechanism, and a comparison of the error between the predicted and target values on the lattice thermal conductivity test set for these 59 thermoelectric materials.

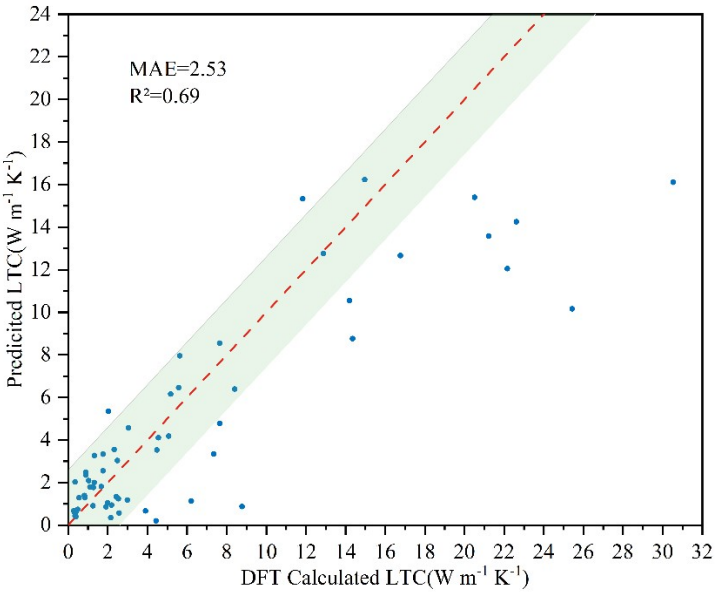

Fig. S7 The prediction results of 59 thermoelectric materials using the CrysGraphFormer model with only two layers of CrysGraphFormerLayer, and a comparison of the error between the predicted and target values of CrysGraphFormer on the lattice thermal conductivity test set for these 59 thermoelectric materials.

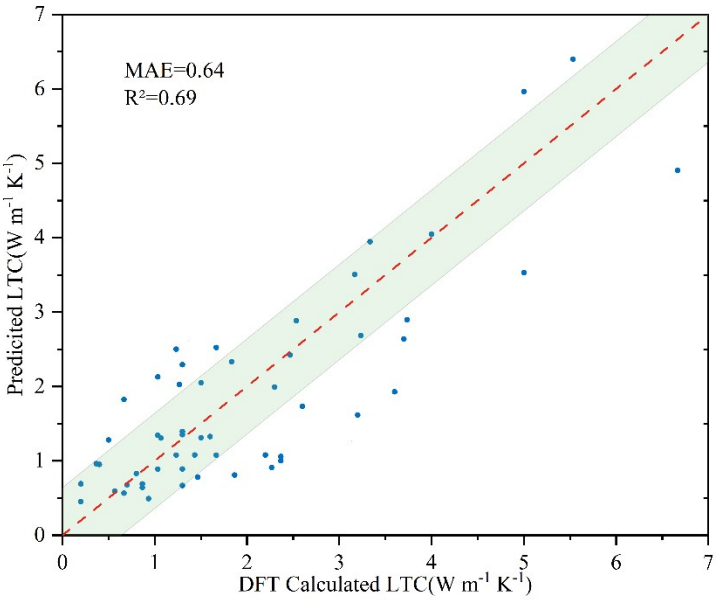

Fig. S8 Error plot of LTC prediction results for 55 ternary semiconductor materials using the CrysGraphFormer model.