Supporting Materials-I

Searching for Negative Thermal Expansion Materials with Bulk Framework Structures and their Relevant Scaling Relationships through Multi-step Machine Learning

Yu Cai[a,b], Chunyan Wang[a,b,c], Huanli Yuan[c], Yuan Guo[a,e], Jun-Hyung Cho[d], Xianran Xing[e,*], Yu Jia[*,a,b,f]

[a] Key Laboratory for Special Functional Materials of Ministry of Education, and School of Materials and Engineering, Henan University, Kaifeng 475001, China

[b] Institute of Quantum Materials and Physics, Henan Academy of Sciences, Zhengzhou 450046, China

[c] School of Physics and Telecommunication Engineering, Zhoukou Normal University, Zhoukou 466001, China

[d]Department of Physics and Research Institute for Natural Science, Hanyang University, 222 Wangsimni-ro, Seongdong-Ku, Seoul 04763, Republic of Korea

[e] Institute of Solid States Chemistry, University of Science and Technology Beijing, Beijing 100083, China

[f] Joint center for Theoretical Physics, and School of Physics and Electronics, Henan University, Kaifeng 475001, China

1. Dataset

In this study, all data on NTE property of materials, including the magnitude of coefficient and temperature range, are derived from experimental results in published literatures, and some theoretical calculation results are also included. In order to make the prediction results more reasonable, we collected NTE and PTE material data in an approximately 1:1 ratio (see Table S1), and the crystal geometry of the collected PTE materials should have the same crystal symmetry as that of the corresponding NTE ones.

Table S1 The number of NTE and PTE materials in Training Datasets

Material Type	NTE Numbers	PTE Numbers
Oxides	123	129
Cyanides	46	31
Fluorides	18	16
Others	17	54
Total	204	230

2. Chemical and Structure Features

For each collected NTE material, we have extracted the information of chemical formula (including atomic number, covalent radius, electronegativity and number of valence electrons), lattice constant, the symmetry of crystal structure, the coefficient and the temperature range of thermal expansion.

Table S2 Input features of NTE materials in machine learning

Content	Description	Example
Formula	material formula	$Sc_2Mo_3O_{12}$
Structure	crystal structure	Orthorhombic
Lattice	a, b, c, α, β, γ	9.72,9.50,13.59,90,90,90
Space group number	number representing the category of space group	60(space group symbol Pbcn)
T	range of temperature	0K-300K
CTE	coefficient of thermal expansion	-6.3
VR	Porosity	0.65
AE	average electronegativity	2.96

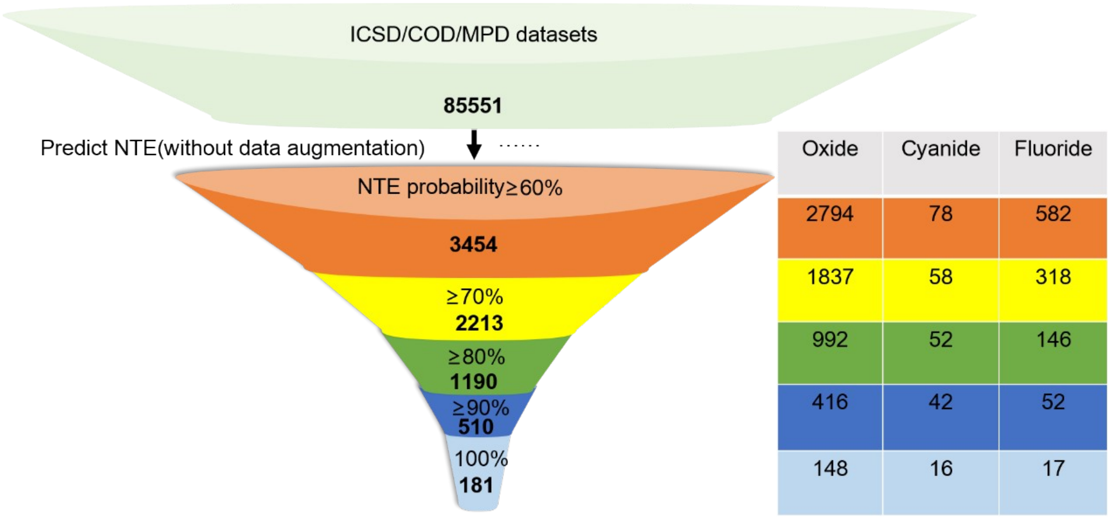
3. Data augmentation and cross-validation

By using the SMOTE algorithm and the random oversampling algorithm, the improved accuracy of dataset with data augmentation, together with the accuracy without data augmentation, is shown in Table S3. We see that data augmentation can significantly improve the accuracy of ML with smaller datasets.

Table S3 The accuracy of machine learning training with or without data augmentation

Methods	with data augmentation	without data augmentation
K-NN	97.2%	30.0%
DT	97.8%	12.0%
GBT	92.6%	52.0%
LR	53.5%	17.2%
SVR	98.8%	26.0%
RF	99.0.%	52.8%

Table S4 shows the results of using cross-validation to randomly divide the dataset into two and three parts for ML training. In the two-part cross-validation, the accuracy of the support vector machine model can reach up to 98.8%, and the accuracy of the random forest model up to 99.0%. While, in the three-part cross-validation, the accuracy also can be improved slightly, i.e., the accuracy of the support vector machine model can reach up to 99.2%, and the accuracy of the random forest model up to 99.2%.



Oxide	Cyanide	Fluoride
2794	78	582
1837	58	318
992	52	146
416	42	52
148	16	17

Figure S1 Machine Learning (ML) prediction results of potential NTE materials for oxides, cyanides and fluorides, respectively (without data augmentation).

In Figure S1, we also present the predicted results without data augmentation and cross-validation for comparison, the number of predicted NTE materials are almost three times that that of with data augmentation and cross-validation, showing that data augmentation and cross-validation can improved the prediction accuracy of NTE materials. With the data augmentation and the improved high accuracy of cross-validation, it is reasonable that our dataset can be used to predict new NTE materials and the prediction results should be more reliable**.**

Table S4 The training accuracy of ML using cross-validation with the dataset divide into two or three parts

ML method	two parts cross-validation	three parts cross-validation
K-NN	97.2%	97.9%
DT	97.8%	98.3%
GBT	92.6%	92.8%
LR	53.5%	54.5%
SVR	98.8%	99.2%
RF	99.0%	99.2%

4. Determination of Regression Method

In the multi-step ML method of predicting the possible NTE materials, we first use six kinds of regression algorithms to learn the training data and determine which is the best algorithm in terms of highest accuracy. These regression algorithms are K-nearest neighbour regression algorithm (K-

NN)[1], decision tree algorithm (DT)[2], gradient boosting tree algorithm (GBT)[3], support vector regression algorithm (SVR)[4], linear regression algorithm (LR)[5] and random forest regression algorithm (RF)[6], respectively.

The cross-validation tests are conducted on the six algorithms mentioned above, and the errors of the CTE between the training results and the original ones in the datasets are shown in Table S5. Again, we see from the Table S5 that the random forest method has the highest accuracy and the smallest error of 5.9 ppm/K, while the linear regression algorithm has the largest error of ~ 40 ppm/K. Therefore, we will present the predicted results of random forest algorithm in the following discussions.

Table S5 The results of cross-validation test in this work (CTE)

ML Method	Accuracy	Error (ppm/K)
K-NN	97.2%	9.6
DT	97.8%	8.5
GBT	92.6%	16.1
SVR	98.8%	6.3
LR	53.5%	40.4
RF	99.0%	5.9

Table S6 Cross-validation test results of range of temperature of NTE materials

ML Method	Accuracy	Error(K)
K-NN	94.6%	82.9
DT	88.0%	119.5
GBT	92.3%	98.8
SVR	95.1%	77.9
LR	67.7%	201.5
RF	96.9%	59.8

Same as the prediction processes of CNTE, we also conduct cross-validation tests for the prediction of range of temperature with these regression algorithms. Table S6 shows the prediction errors between the training results and the original results of upper limit of temperature range. It can be seen that the random forest method has the highest accuracy with a smallest error of ~60K. And the linear regression algorithm has the largest error of more than 200K. Therefore, the random forest method (RF) is again to be selected for prediction of upper limit of temperature range**.**

5. First principles calculations and results

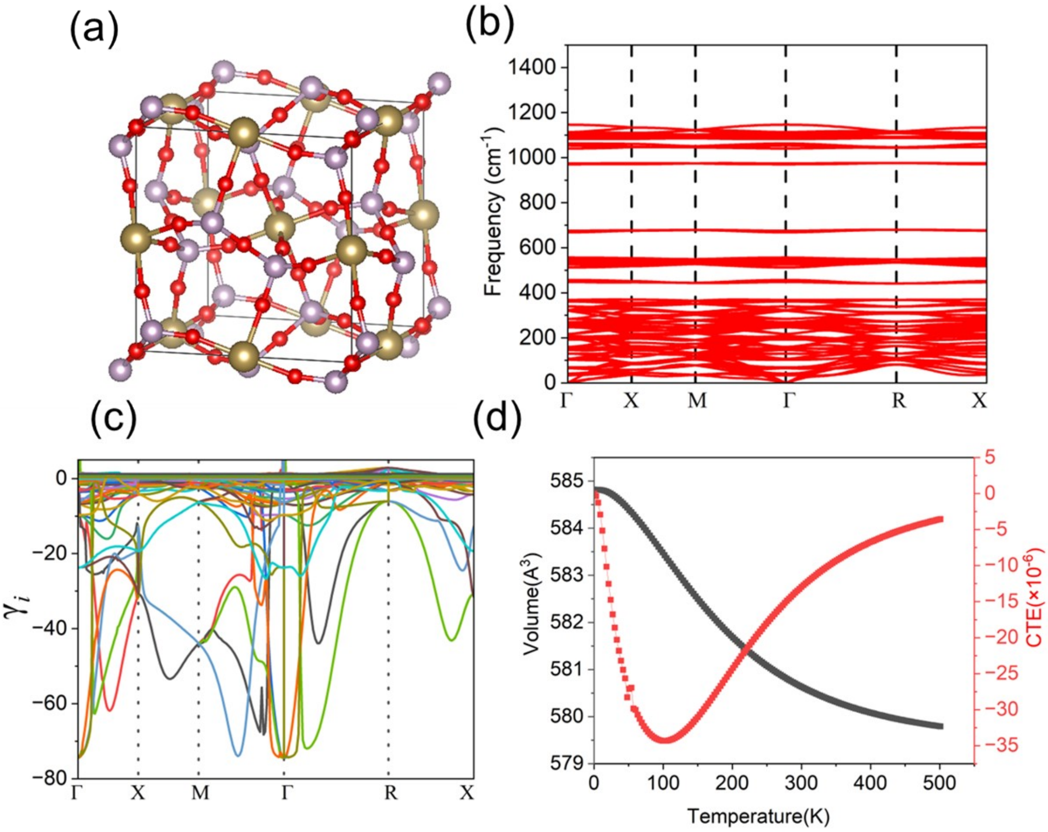


Figure S2 The calculation of coefficient of NTE of TaP_2O_7 using the DFT with the quasi harmonic approximatiom. (a) Crystal Structure, (b) Phonon Spectrum, (c) Gruneisen parameters and (d) Lattice volume and CTE with changes of temperature, respectively.

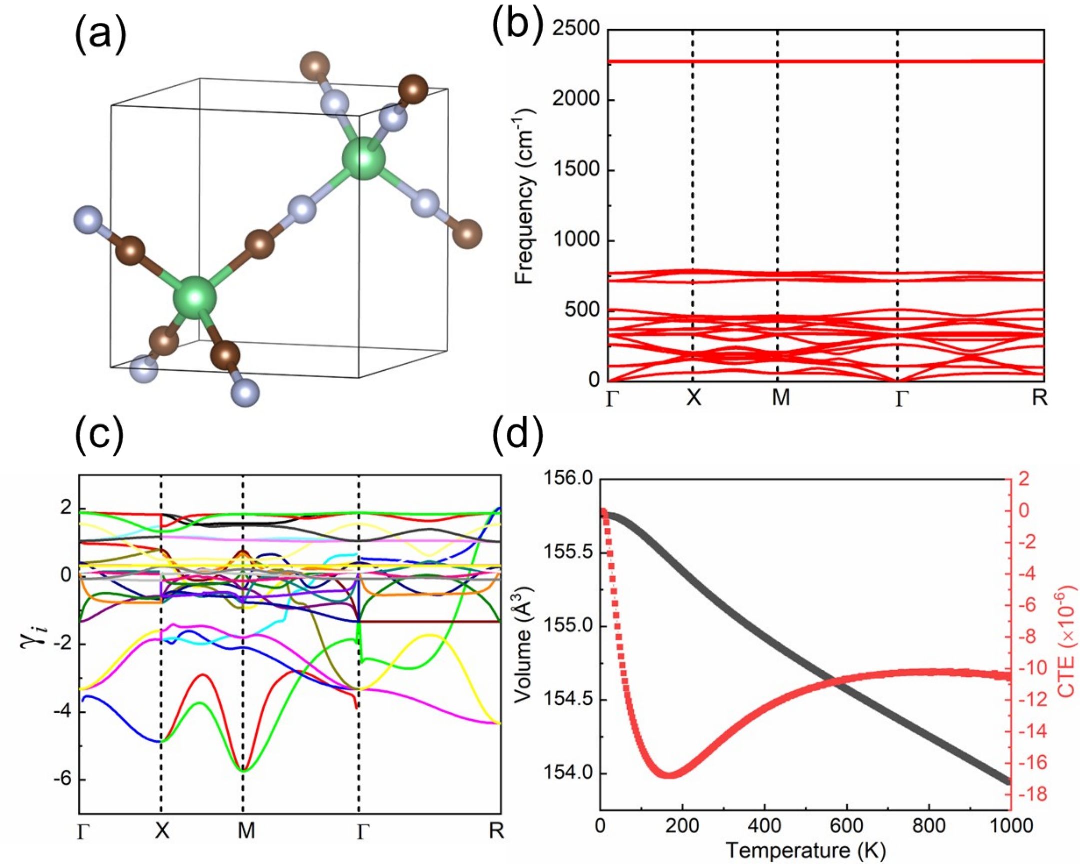


Figure S3 The calculation of coefficient of NTE of Be (CN)$_2$ using the DFT with the quasi harmonic approximatiom. (a) Crystal Structure, (b) Phonon Spectrum, (c) Gruneisen parameters and (d) Lattice volume and CTE with changes of temperature, respectively.

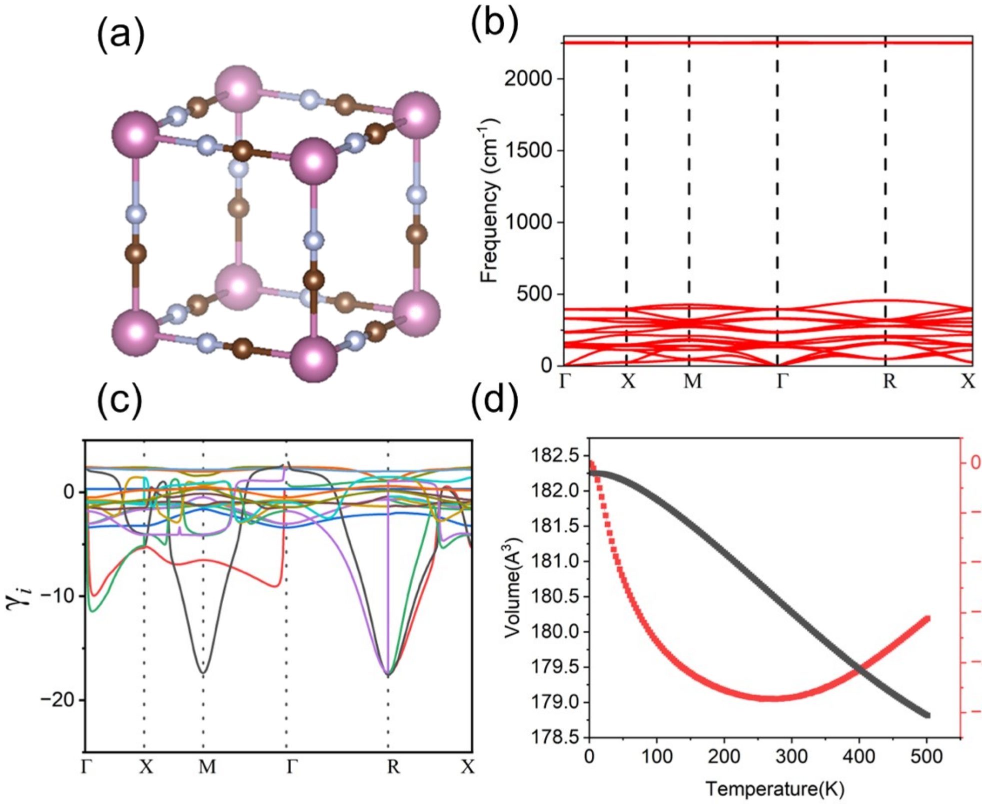


Figure S4 The calculation of coefficient of NTE of In $(CN)_3$ using the DFT with the quasi harmonic approximatiom. (a) Crystal Structure, (b) Phonon Spectrum, (c) Gruneisen parameters and (d) Lattice volume and CTE with changes of temperature, respectively.

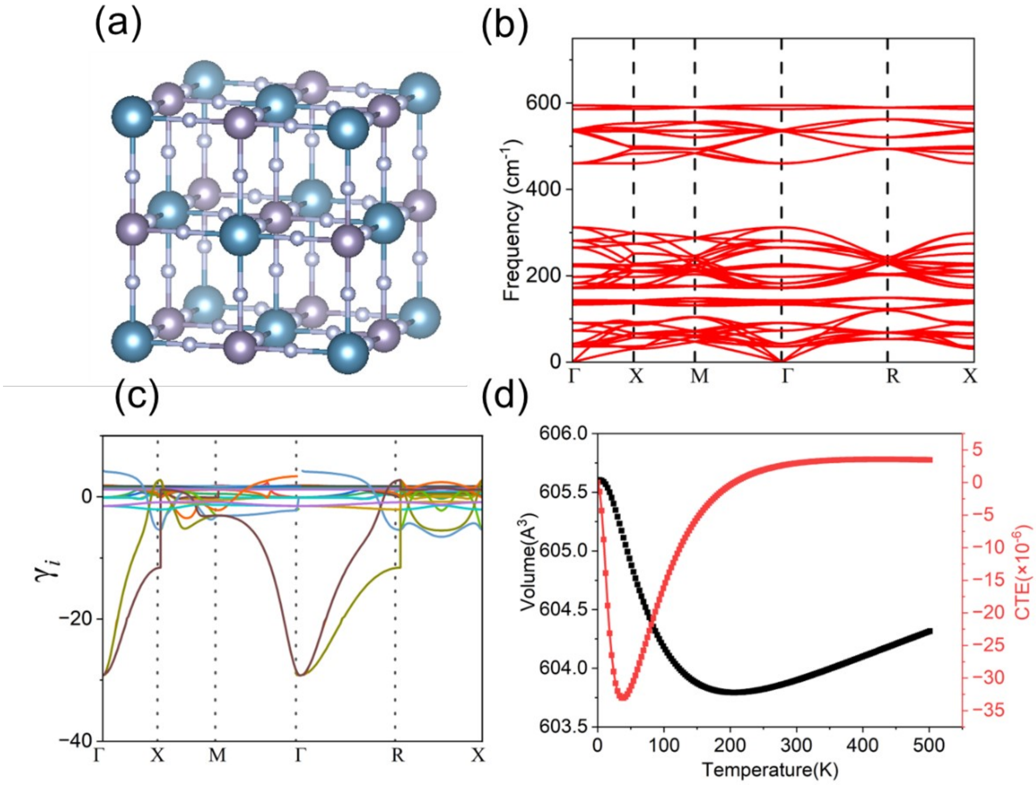


Figure S5 The calculation of coefficient of NTE of $CaSnF_6$ using the DFT with the quasi harmonic approximatiom. (a) Crystal Structure, (b) Phonon Spectrum, (c) Gruneisen parameters and (d) Lattice volume and CTE with changes of temperature, respectively.

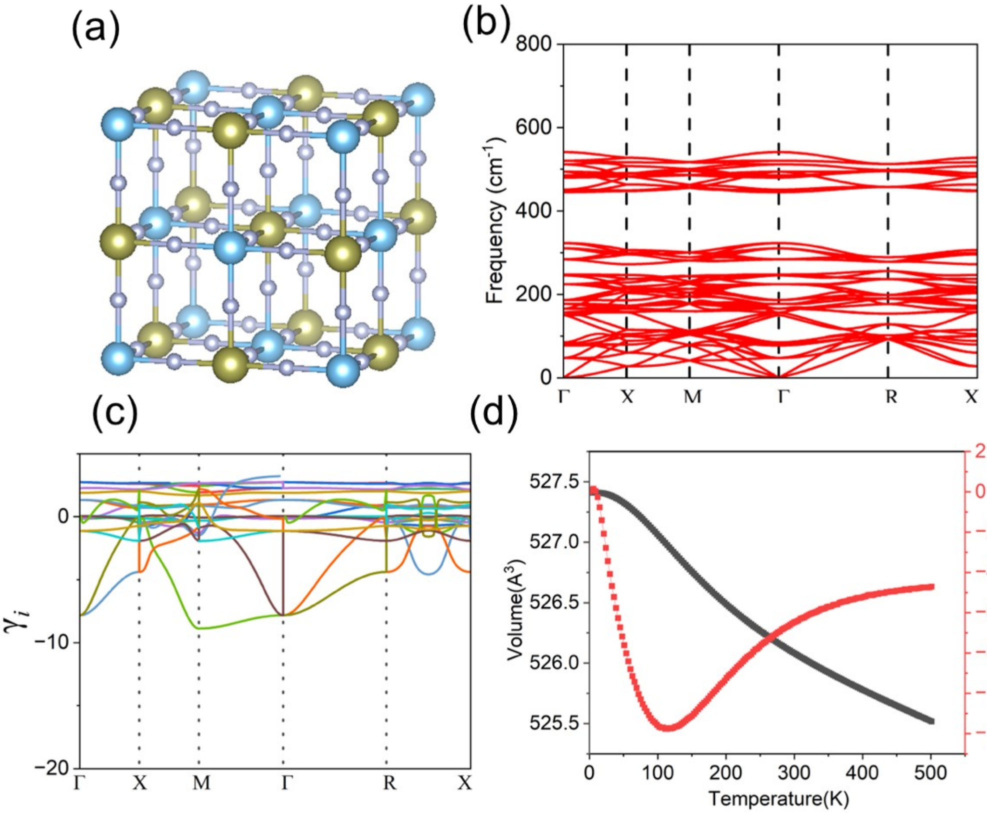


Figure S6 The calculation of coefficient of NTE of $HfTiF_6$ using the DFT with the quasi harmonic approximatiom. (a) Crystal Structure, (b) Phonon Spectrum, (c) Gruneisen parameters and (d) Lattice volume and CTE with changes of temperature, respectively.

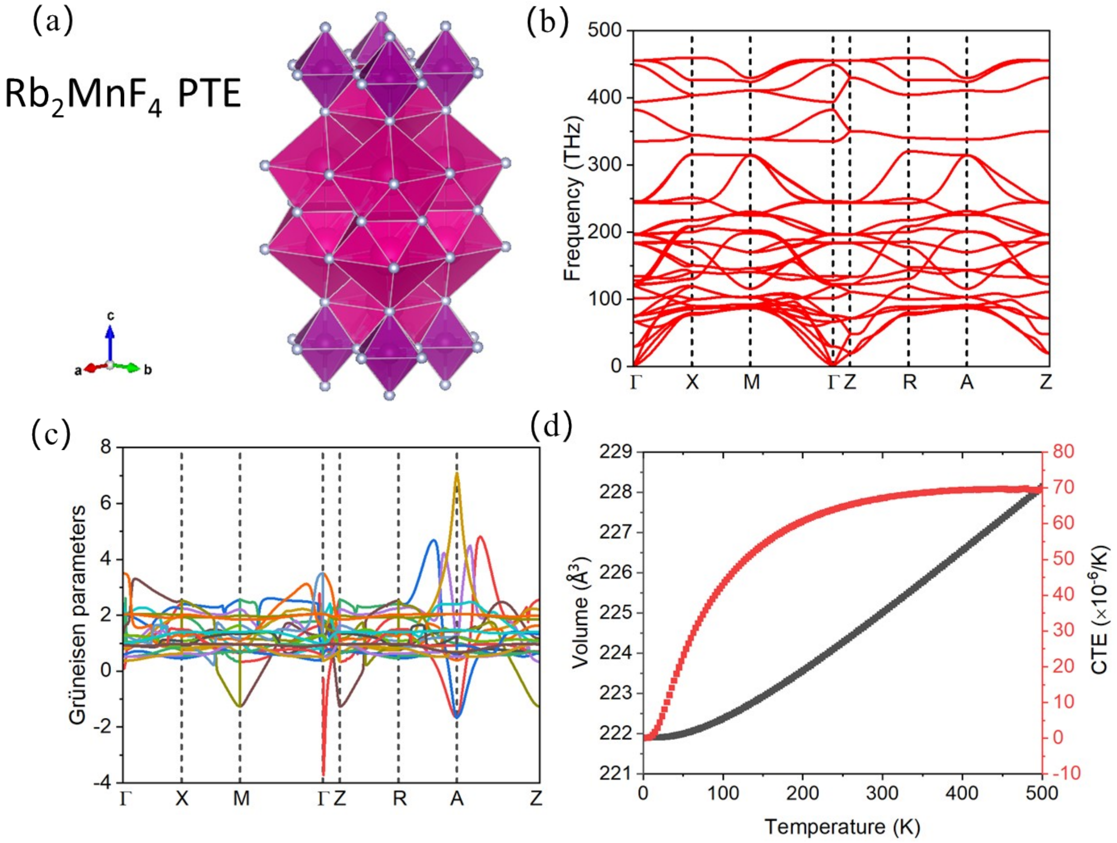


Figure S7 The calculation of coefficient of PTE of Rb$_2$MnF$_4$ using the DFT with the quasi harmonic approximatiom. (a) Crystal Structure, (b) Phonon Spectrum, (c) Gruneisen parameters and (d) Lattice volume and CTE with changes of temperature, respectively.

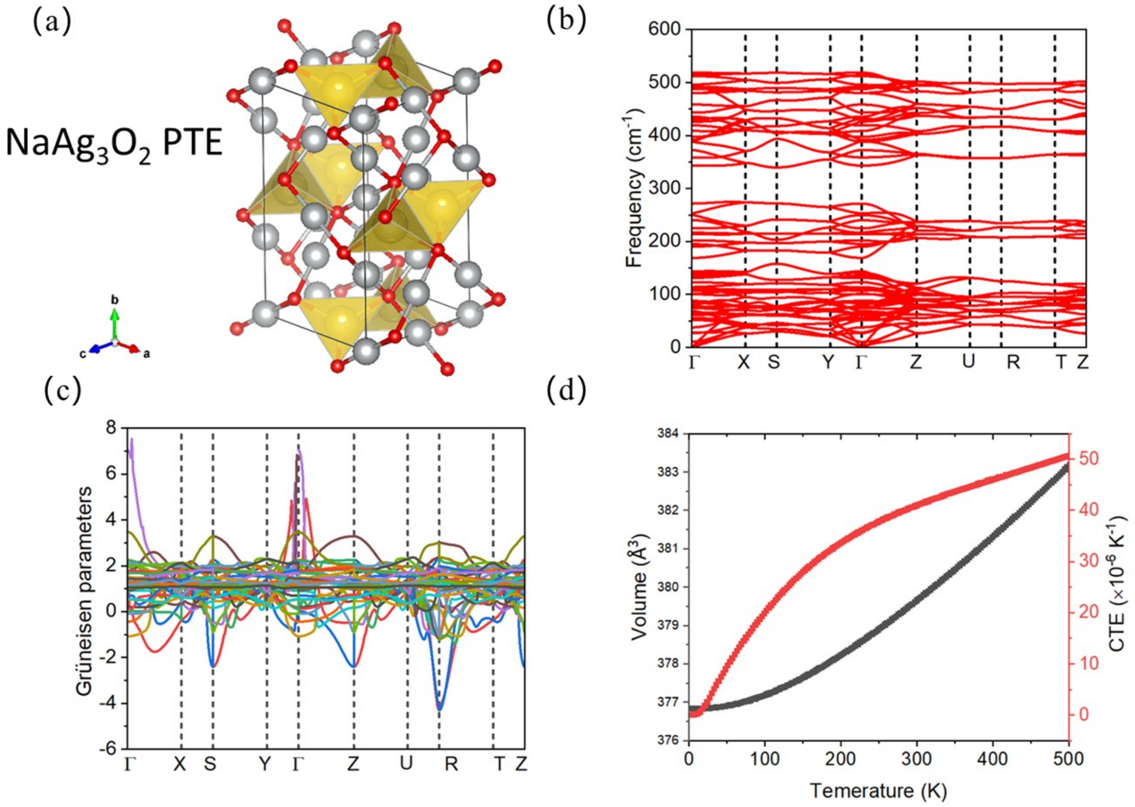


Figure S8 The calculation of coefficient of PTE of $NaAg_3O_2$ using the DFT with the quasi harmonic approximatiom. (a) Crystal Structure, (b) Phonon Spectrum, (c) Gruneisen parameters and (d) Lattice volume and CTE with changes of temperature, respectively.

6. Discussion on the Relationship between Temperature Range, Porosity, and Average Electronegativity

The material structures mentioned in the main text that compare the magnitude of CNTE are shown in the following figures S9, S10 and S11.

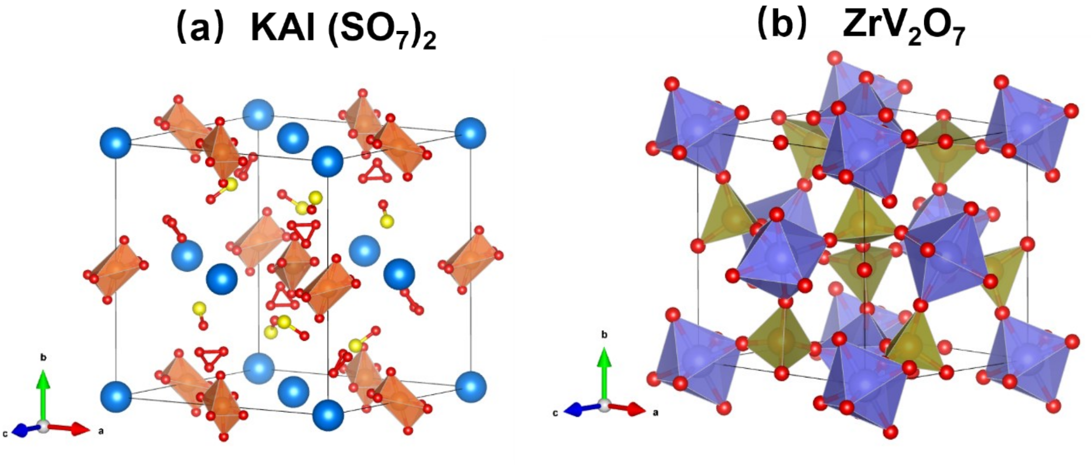


Figure S9 Comparison of porosity and CNTE between (a) KAl (SO_7) $_2$ and (b) ZrV_2O_7. Both materials belong to a cubic structure, with KAl (SO7) $_2$ porosity ~ 0.83 and CNTE=-27.8ppm/K; ZrV_2O_7 porosity~ 0.57, CNTE=-10.7ppm/K

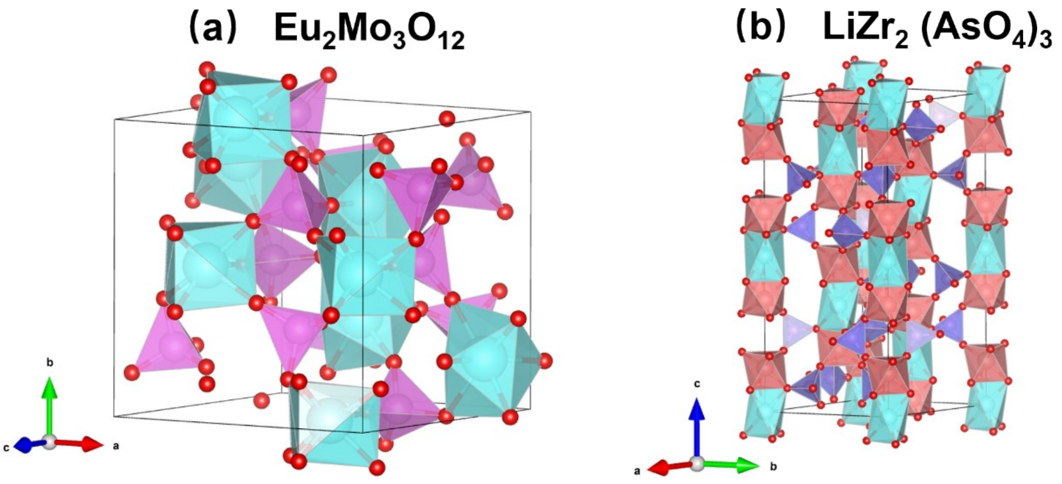


Figure S10 Comparison of the crystal structure and CNTE of (a) $Eu_2Mo_3O_{12}$ and (b) $LiZr_2(AsO_4)_3$. The porosity of both materials is ~0.6, but $Eu_2Mo_3O_{12}$ has an orthogonal structure with CNTE = -18.3ppm/K; $LiZr_2$ (AsO_4) $_3$ has a trigonal structure, CNTE = -0.36ppm/K.

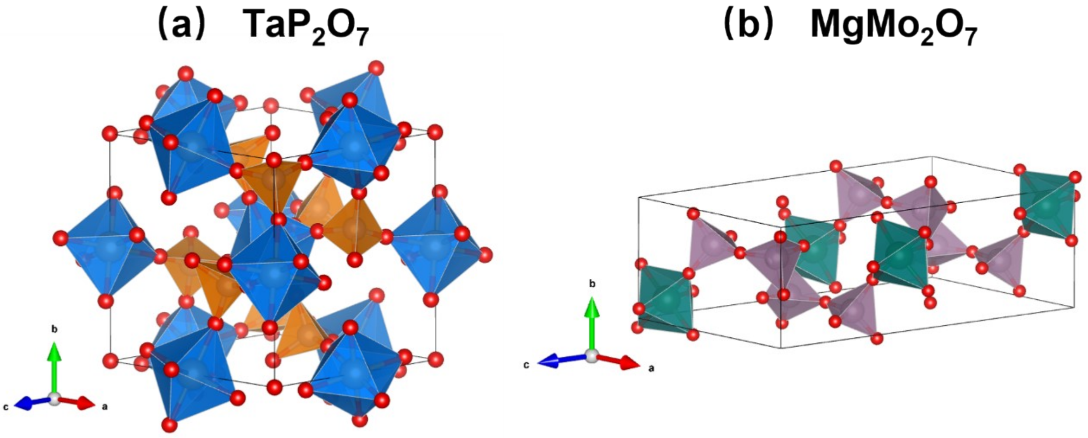


Figure S11 Comparison of the crystal structure and CNTE of (a) TaP_2O_7 and (b) $MgMo_2O_7$. The average electronegativity of both materials is ~3.0. But TaP_2O_7 has a cubic structure, with CNTE = -10.7ppm/K; $MgMo_2O_7$ has a monoclinic structure, CNTE = -5.0ppm/K.

Beyond the three rules of the main text, we also plotted the relationships of temperature range with the porosity, as well as the average electronegativity for three kinds of compounds, see the Figure S12 (a)-(c) and Figure S13 (a)-(c), respectively. It can be seen that there is no direct correlation between temperature range with the porosity, as well as the average electronegativity, indicating that the factors affecting the temperature range of NTE materials are relatively complex, rather than the existence of a single factor correlation.

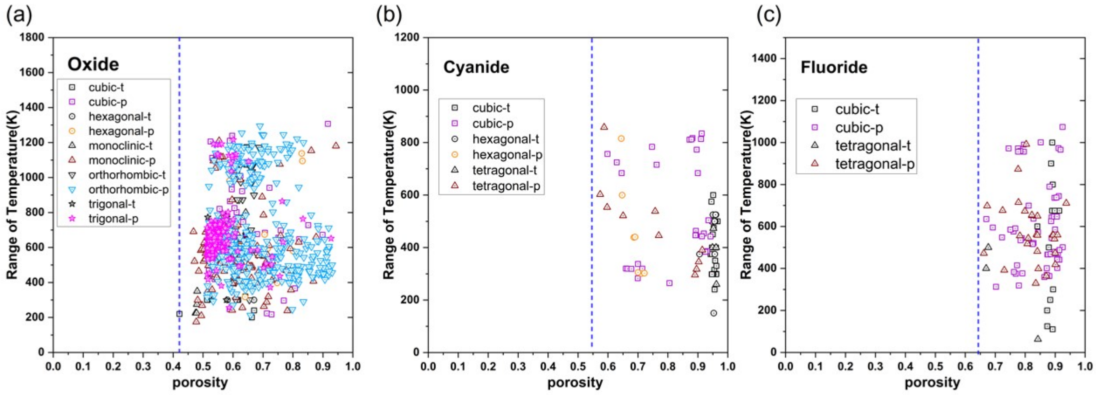


Figure S12 Relationship between temperature range and the porosity of the three kinds of NTE materials. (a)oxides, (b) cyanides and (c)fluorides, respectively.

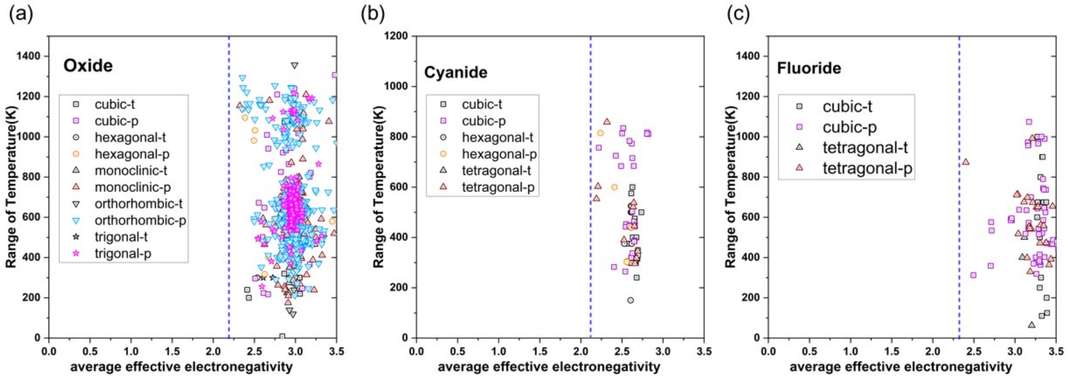


Figure S13 Relationship between temperature range and average electronegativity of the three kinds of NTE materials. (a)oxides, (b) cyanides and (c)fluorides, respectively.

References:

1. W. Yang, K. Q. Wang and W. M. Zuo, *Neurocomputing*, 2012, **83**, 31-37.
2. A. J. Myles, R. N. Feudale, Y. Liu, N. A. Woody and S. D. Brown, *Journal of Chemometrics*, 2004, **18**, 275-285.
3. G. Biau, B. Cadre and L. Rouviere, *Machine Learning*, 2019, **108**, 971-992.

4. A. J. Smola and B. Scholkopf, *Statistics and Computing*, 2004, **14**, 199-222.
5. J. A. Nelder and R. W. M. Wedderburn, *Journal of the Royal Statistical Society*, 1972, **135**, 370-384.
6. B. Talekar and S. Agrawal, *Bioscience Biotechnology Research Communications*, 2020, **13**, 245-248.
7. Kresse and Furthmuller, *Physical review. B, Condensed matter*, 1996, **54**, 11169-11186.
8. G. Kresse and D. Joubert, *Physical Review B*, 1999, **59**, 1758-1775.
9. J. P. Perdew, K. Burke and M. Ernzerhof, *Physical Review Letters*, 1996, **77**, 3865-3868.
10. A. Togo, F. Oba and I. Tanaka, *Physical Review B*, 2008, **78**.