Electronic Supplementary Information

Unveiling the Physical Mechanisms Underpinning Bandgap Variations in Chalcopyrite Crystals (ABX_2) Using Interpretable Artificial Intelligence

Xiaolan Fu,[a,#] Jiaqian Wang,[b,#] Xiaojuan Hu,[b,*] Wenwu Xu,[a,*] Sergey V. Levchenko,[c] and Zhong-Kang Han[b,*]

[a]Department of Physics, School of Physical Science and Technology, Ningbo University, Ningbo, 315211, China

[b]School of Materials Science and Engineering, Zhejiang University, Hangzhou, 310027, China

[c] Skolkovo Institute of Science and Technology, Bolshoy Boulevard 30/1, Moscow, 121205, Russia

Tel：18321635996

E-mail: hanzk@zju.edu.cn

Table of Contents

Methodology

1. Density functional theory calculations

Spin-polarized density functional theory (DFT)[1] calculations for ABX_2 chalcopyrite materials were carried out using the all-electron full-potential electronic-structure package FHI-aims[2], which employs numeric atom-centered orbital basis sets. Spin-orbit coupling was taken into account using a non-self-consistent second-variational approach [3]. Ensuring well-converged DFT energies is crucial for developing a predictive computational model. To achieve the desired convergence, we utilized the *tight* numerical settings available in FHI-aims. Each ABX_2 chalcopyrite system was modeled using a supercell tailored to the specific structure under investigation. For each configuration, we fully relaxed both the atomic positions and the lattice parameters to ensure accurate calculations. The choice of exchange-correlation (xc) functional was meticulously validated by comparing the calculated lattice parameters and band gaps with experimental values. In this study, we specifically employed the hybrid functional HSE06 [4] for these calculations. The HSE06 hybrid functional is particularly effective in accurately describing the electronic structure of these materials, especially in predicting bandgaps[5]. During our computations, we used a 6×6×3 *k*-point mesh converged to reproduce the bandgap to within 0.01 eV. This methodology enabled us to obtain high-precision band gap data, providing a solid foundation for further analysis of materials properties and aiding in the understanding and prediction of these materials' performance in various applications.

2. Model building by SISSO code

In SISSO, the feature construction was performed by creating all possible mathematical combinations (up to feature complexity equal 5) of the primary features using the operators {+, −, ·, /, log, exp, exp−, −1, 2, 3, √, 3√, |−|}[6,7]. The subspace size parameter was set to 100. A list of the primary features for model training data sets can be found in the main text. The code for SISSO and user guide are available at: https://github.com/rouyang2017/SISSO.

The importance score of descriptor components in SISSO is calculated based on RMSE for each descriptor component as follows:

$$Importance\ score = 1 - \frac{RMSE(all\ components)}{RMSE(all\ components - 1\ component)}$$

The component is removed from the descriptor, the model is refit with the remaining components, and RMSE is recalculated. Score 0 means that removing the component does not change the RMSE of the model.

3. Subgroup discovery analysis

The subgroup discovery was performed using a modified RealKD package[8]. Each feature was split to 14 subsets using 14-means clustering algorithm[9]. The obtained borders between adjacent feature value clusters (a1, a2, ...) are applied further for construction of inequalities (feature1 < a1), (feature2 ≥ a2), etc. While final result might depend on the number of considered clusters, in our previous studies we found that higher numbers of considered clusters provide essentially the same result[10]. The candidate subgroups are built as conjunctions of obtained simple inequalities. The main idea of SGD is that the subgroups are unique if the distribution of the data in them is as different as possible from the data distribution in the whole sampling. The uniqueness is evaluated with a quality function[11]. The search was done with an adapted for such

purposes Monte-Carlo algorithm[12], in which first a certain number of trial conjunctions (seeds) is generated. Afterwards, for each seed (accompanied with pruning of inequalities) the quality function is calculated. We have tested here several numbers of initial seeds: 10000, 30000, 50000, 70000 and 100000. The subgroups with the overall high quality function value were selected.

Table S1. HSE06 bandgap of 122 chalcopyrite crystals.

Crystal	Bandgap (eV)	Crystal	Bandgap (eV)
$BeAlN_2$	0.03	$BeInN_2$	0.00
$BeTlP_2$	0.00	$BeTlAs_2$	0.03
$BeTlSe_2$	0.01	$BePbN_2$	1.05
$BePbTe_2$	0.62	$BeBiAs_2$	0.10
$MgGaN_2$	0.01	$MgTlS_2$	0.64
$MgPbSb_2$	0.41	$MgSbTe_2$	1.78
$CaTlS_2$	1.42	$CaGaN_2$	0.02
$CaBiSe_2$	1.73	$CuInN_2$	0.00
$CuGeN_2$	0.01	$CuSbAs_2$	0.03
$ZnGaN_2$	0.01	$ZnTlS_2$	0.02
$ZnPbAs_2$	0.29	$AgTlN_2$	0.01
$AgSbN_2$	0.20	$AgSbSb_2$	0.07
$CdAlN_2$	0.01	$CdInAs_2$	0.01
$PdGeN_2$	1.55	$PdGeS_2$	0.10
$PdSbTe_2$	0.06	$PdSbSb_2$	0.03
$CuAlS_2$	3.00	$CuAlSe_2$	2.07
$CuAlTe_2$	2.02	$CuGaS_2$	1.88
$CuGaSe_2$	1.05	$CuGaTe_2$	1.04

$CuInS_2$	1.02	$CuInSe_2$	0.55
$CuInTe_2$	0.76	$CuTlS_2$	0.66
$CuTlSe_2$	0.55	$CuTlTe_2$	0.28
$AgAlS_2$	3.00	$AgAlSe_2$	2.09
$AgAlTe_2$	1.89	$AgGaS_2$	1.98
$AgGaSe_2$	1.10	$AgGaTe_2$	0.92
$AgInS_2$	1.28	$AgInSe_2$	0.71
$AgInTe_2$	0.76	$AgTlS_2$	0.24
$AgTlSe_2$	0.80	$AgTlTe_2$	0.69
$BeSiN_2$	4.91	$BeSiP_2$	1.91
$BeSiAs_2$	1.71	$BeSiSb_2$	0.93
$BeGeN_2$	4.47	$BeGeP_2$	1.59
$BeGeAs_2$	1.26	$BeGeSb_2$	0.51
$BeSnN_2$	2.31	$BeSnP_2$	1.56
$BeSnAs_2$	1.23	$BeSnSb_2$	0.51
$ZnSiN_2$	4.56	$ZnSiP_2$	2.07
$ZnSiAs_2$	1.69	$ZnSiSb_2$	1.03
$ZnGeN_2$	2.73	$ZnGeP_2$	1.91
$ZnGeAs_2$	0.62	$ZnGeSb_2$	0.19
$ZnSnN_2$	0.99	$ZnSnP_2$	1.45
$ZnSnAs_2$	0.38	$ZnSnSb_2$	0.21
$CdSiN_2$	2.85	$CdSiP_2$	3.41
$CdSiAs_2$	1.10	$CdSiSb_2$	0.61
$CdGeN_2$	1.78	$CdGeP_2$	1.38
$CdGeAs_2$	0.08	$CdGeSb_2$	0.17
$CdSnN_2$	0.39	$CdSnP_2$	0.93
$CdSnAs_2$	0.07	$CdSnSb_2$	0.08

$MgSiN_2$	4.77	$MgSiP_2$	2.04
$MgSiAs_2$	1.86	$MgSiSb_2$	1.46
$MgGeN_2$	3.69	$MgGeP_2$	2.17
$MgGeAs_2$	1.20	$MgGeSb_2$	0.71
$MgSnN_2$	2.00	$MgSnP_2$	1.92
$MgSnAs_2$	1.01	$MgSnSb_2$	0.78
$PdBiP_2$	0.01	$PdBiAs_2$	0.04
$PdBiS_2$	0.04	$PdBiSe_2$	0.03
$PdBiTe_2$	0.01	$PdBiSb_2$	0.00
$BeAlN_2$	0.04	$BeAlP_2$	0.01
$BeAlAs_2$	0.06	$BeAlS_2$	1.22
$BeAlSe_2$	1.15	$BeAlTe_2$	0.03
$BeAlSb_2$	0.08	$BeGaN_2$	0.16
$BeGaP_2$	0.03	$BeGaAs_2$	0.04
$BeGaS_2$	0.00	$BeGaSe_2$	0.01
$BeGaTe_2$	0.00	$BeGaSb_2$	0.09

Table S2. Primary features utilized in AI models.

Name	Abbreviation	Unit
Number of Valence Electrons	NVE	/
Molar Volume[13]	MV	cm^3/mol
Ionization Energy	IE	kJ/mol
Atomic Weight	AW	a.m.u.
Atomic Radius Calculated [14]	ARC	pm
Atomic Radius Empirical[15]	ARE	pm
Covalent Radius	CR	pm
Electron Affinity	EA	kJ/mol
Pauling electronegativity[16–19]	EN	/
Heat of Fusion	HF	kJ/mol
Thermal Conductivity[20–23]	TC	W/(m·K)
Heat of Vaporization	HV	kJ/mol

Table S3. RMSE (eV) of models for every iteration in CV20 (Feature complexity: 5; Dimension: from 1 to 5).

Iteration	1D	2D	3D	4D	5D
01	1.32	1.12	0.30	0.28	0.27
02	0.74	0.51	0.45	0.47	0.37
03	0.54	0.77	0.80	0.58	0.55
04	0.62	0.66	0.73	0.81	0.85
05	0.21	0.34	0.29	0.40	0.42
06	0.88	0.86	0.85	0.76	0.79
07	0.51	0.65	0.52	0.45	0.50
08	0.56	0.44	0.45	0.62	0.66
09	0.68	0.52	1.30	1.22	1.14
10	0.48	0.44	0.40	0.41	0.30
11	0.55	0.40	0.44	0.42	0.34
12	0.97	0.78	0.46	0.47	0.45
13	0.42	0.38	0.42	0.48	0.43
14	0.57	0.36	0.42	0.43	0.36
15	0.58	0.51	0.44	0.41	0.34
16	0.85	0.50	0.56	0.40	0.41
17	0.74	0.51	0.43	0.42	0.26
18	1.00	0.87	0.94	1.05	0.90
19	0.70	0.70	0.78	0.78	0.87
20	3.12	3.14	3.22	3.30	3.32

Table S4. Descriptor components, coefficients, and importance scores of the SISSO model used for predicting bandgap. The value of the intercept is 0.59.

Symbol	Descriptor	Coefficient	Importance score
d_1	(EAb × TCa × |EAa–EAb|) ÷ (MVb × ARCc)	0.001	0.72
d_2	|HVa + HVb–EAa–|IEb–IEc + HFb||	-0.002	0.37
d_3	|(MVb ÷ AWb)–(log(ENb) × (MVa ÷ AWa))|	2.24	0.22
d_4	(TCb ÷ TCc) ÷ (MVb ÷ MVc–EAc ÷ HVa)	-0.00002	0.18
d_5	(ENa × TCa + EVb × TCc) ÷ (HFa)2	-0.087	0.14

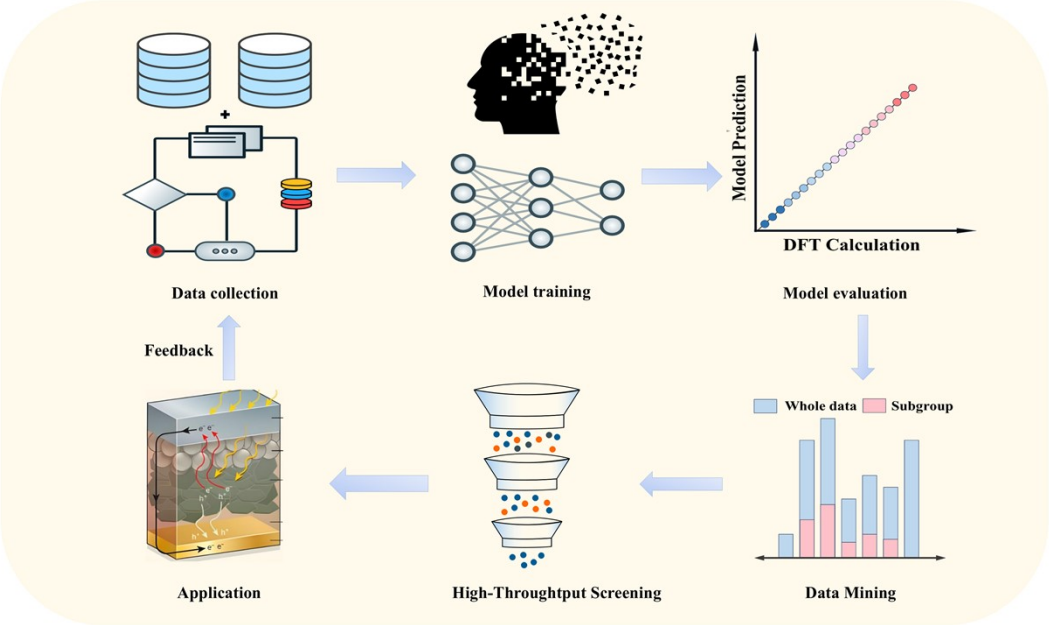
Table S5. The subgroups with highest quality obtained with SGD that minimize ("Below" prefix) or maximize ("Above" prefix) bandgap for different thresholds. Degenerate propositions are also shown. The degeneracy is determined by the condition that the size of the subgroup increases by less than 1% of the original subgroup when the proposition is removed. The units of the features are shown in Table I. Atoms selected by each proposition are shown in parenthesis. For degenerate propositions, only differences ("+" for addition and "–" for removal) in selected species are shown.

	Proposition	Elements	Degenerate propositions:	Elements
Below-1	ARCc ≥ 98	P, As, Se, Sb, Te	TCc ≥ 0.24	
	EAa ≤ 125.6	Cu, Pd, Ag, Be, Ca, Cd, Mg, Zn	AWa ≤ 112.41 ENa ≤ 2.20	
	EAa ≥ 53.7	Cu, Pd, Ag, Au, Cd		
	MVb ≥ 11.93	Si, Ge, In, Sn, Sb, Tl, Pb, Bi		
Below-0.5	AWb ≥ 72.64	Ge, In, Sn, Sb, Tl, Pb, Bi	MVb ≥ 13.65	
	TCa<120	Pd, Cd		
	IEc ≤ 999.6	S, As, Se, Sb, Te	EAc ≥ 78	
Above-2.5	ARCc<103	N, P, S	TCc < 0.38 IEc > 947	
	EAa>125.6	Au	AWa > 112.41 ENa > 2.2	

	HFb ≥ 7	Al, Si, Ge, Sn, Sb, Bi	EAb ≥ 42.5	
	MVb ≤ 18.18	Al, Si, Ga, Ge, In, Sn, Sb, Tl	TCb ≥ 16 AWb ≤ 121.76	+Pb –Tl
Above-2	AWa>112.41	Au	EAa > 125.6 ENa > 2.2	
	HFb ≥ 6.30	Al, Si, Ge, Sn, Sb, Bi	EAb ≥ 42.5	
	MVb ≤ 18.27	Al, Si, Ga, Ge, In, Sn, Sb, Tl, Pb	TCb ≥ 16 AWb ≤ 121.76	–Tl, –Pb
	IEc ≥ 941	N, P, S, As, Se	ARCc ≤ 114	

Table S6. Direct-gap materials with bandgaps between 0.6-2 eV, calculated with HSE06 hybrid functional.

Materials	bandgap	Materials	bandgap	Materials	bandgap
$CdSiSb_2$	0.61	$CuGaTe_2$	1.04	$PdGeN_2$	1.55
$ZnGeAs_2$	0.62	$BePbN_2$	1.05	$BeSnP_2$	1.56
$CuTlS_2$	0.66	$CuGaSe_2$	1.05	$BeGeP_2$	1.59
$AgInSe_2$	0.71	$CdSiAs_2$	1.10	$ZnSiAs_2$	1.69
$MgGeSb_2$	0.71	$AgGaSe_2$	1.10	$BeSiAs_2$	1.71
$AgInTe_2$	0.76	$MgGeAs_2$	1.20	$CaBiSe_2$	1.73
$CuInTe_2$	0.76	$BeAlS_2$	1.22	$CdGeN_2$	1.78
$MgSnSb_2$	0.78	$BeSnAs_2$	1.23	$MgSbTe_2$	1.78
$AgGaTe_2$	0.92	$BeGeAs_2$	1.26	$MgSiAs_2$	1.86
$BeSiSb_2$	0.93	$AgInS_2$	1.28	$CuGaS_2$	1.88
$CdSnP_2$	0.93	$CdGeP_2$	1.38	$AgAlTe_2$	1.89
$ZnSnN_2$	0.99	$CaTlS_2$	1.42	$ZnGeP_2$	1.91
$MgSnAs_2$	1.01	$ZnSnP_2$	1.45	$MgSnP_2$	1.92
$CuInS_2$	1.02	$MgSiSb_2$	1.46	$AgGaS_2$	1.98
$ZnSiSb_2$	1.03				



Schematic S1. Schematic of the active learning strategy for predicting the bandgap of ternary chalcopyrite crystals (ABX_2).

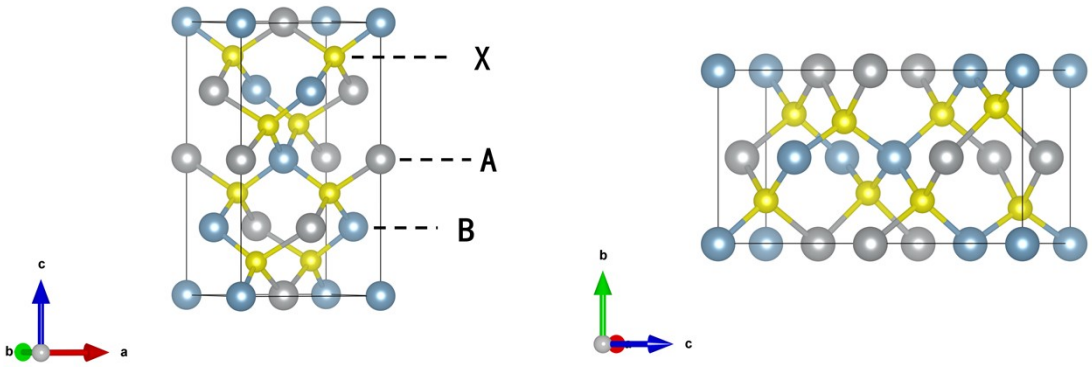


Figure S1. Geometric structure of the Chalcopyrite crystal (ABX_2).

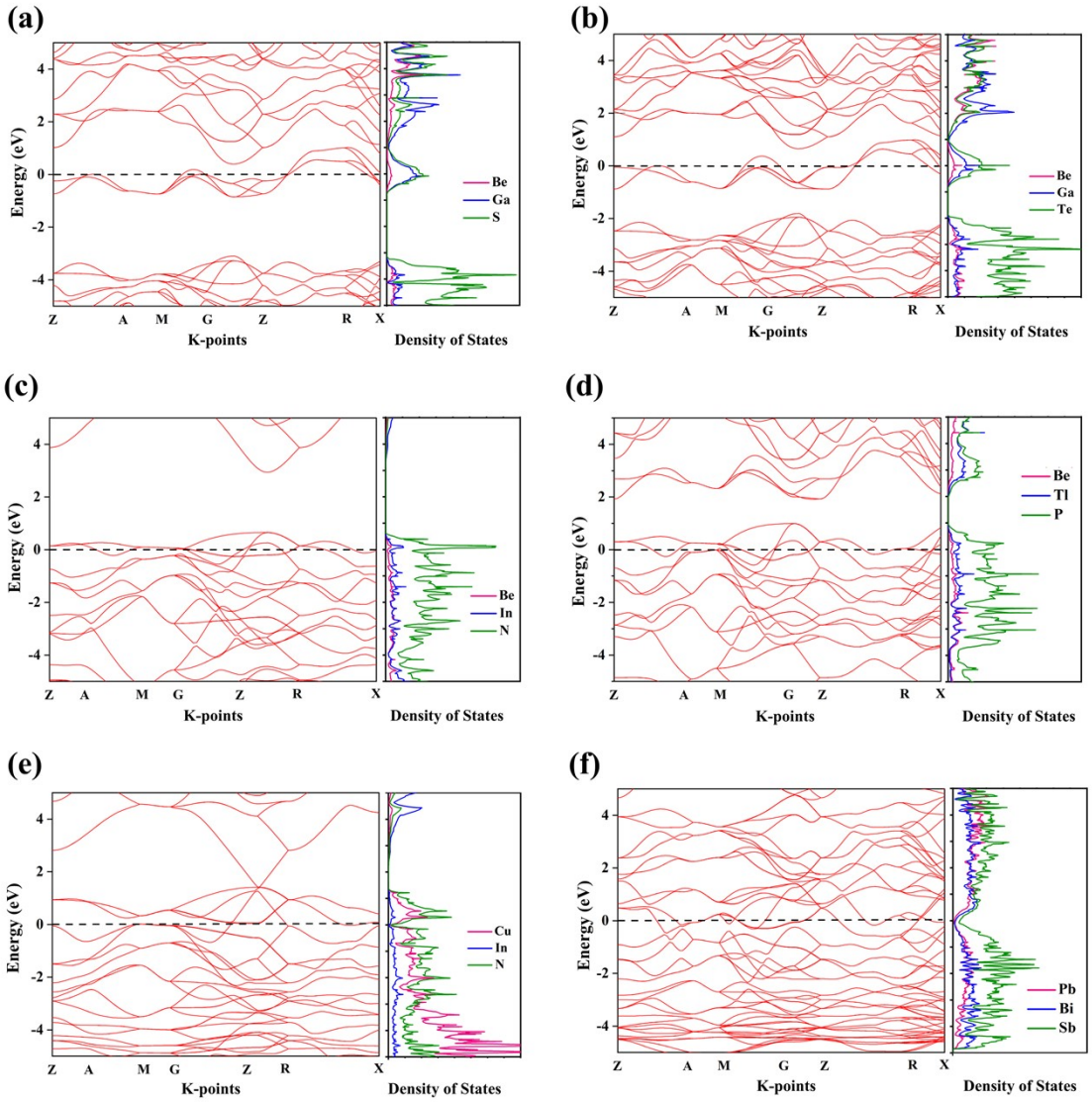


Figure S2. The bandstructure and density of states (DOS) for the chalcopyrite crystalline (ABX_2) materials with a zero bandgap.

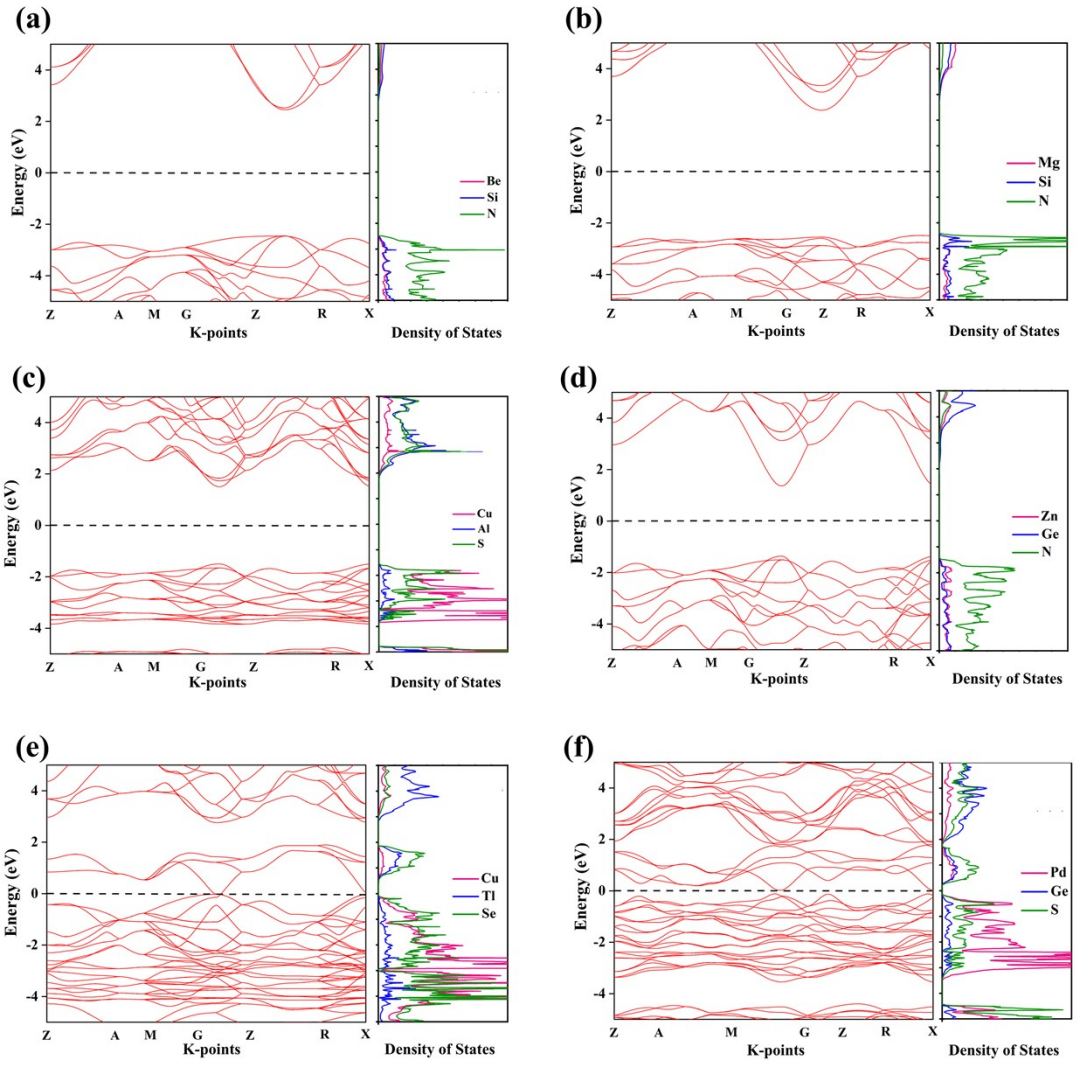


Figure S3. The bandstructures and density of states and of the Chalcopyrite crystal (ABX_2): large bandgap, moderate bandgap, and small bandgap, respectively.

References:

1 P. Hohenberg and W. Kohn, *Phys. Rev.*, 1964, 136, B864–B871.
2 W. Kohn and L. J. Sham, *Phys. Rev.*, 1965, 140, A1133–A1138.
3 W. P. Huhn and V. Blum, *Phys. Rev. Mater.*, 2017, 1, 033803.
4 V. Blum, R. Gehrke, F. Hanke, P. Havu, V. Havu, X. Ren, K. Reuter and M. Scheffler, *Comput. Phys. Commun.*, 2009, 180, 2175–2196.
5 M. Li, Y. Luo, X. Hu, G. Cai, Z. Han, Z. Du and J. Cui, *Adv. Electron. Mater.*, 2020, 6, 1901141.
6 R. Ouyang, S. Curtarolo, E. Ahmetcik, M. Scheffler and L. M. Ghiringhelli, *Phys. Rev. Mater.*, 2018, 2, 083802.
7 A. Mazheika, S. V. Levchenko and L. M. Ghiringhelli, *ArXiv Prepr. ArXiv240315816*.
8 C. Sutton, M. Boley, L. M. Ghiringhelli, M. Rupp, J. Vreeken and M. Scheffler, *Nat. Commun.*, 2020, 11, 4428.
9 T. Kanungo, D. M. Mount, N. S. Netanyahu, C. D. Piatko, R. Silverman and A. Y. Wu, *IEEE Trans. Pattern Anal. Mach. Intell.*, 2002, 24, 881–892.
10 Z.-K. Han, D. Sarker, R. Ouyang, A. Mazheika, Y. Gao and S. V. Levchenko, *Nat. Commun.*, 2021, 12, 1833.
11 A. Mazheika, Y.-G. Wang, R. Valero, F. Viñes, F. Illas, L. M. Ghiringhelli, S. V. Levchenko and M. Scheffler, *Nat. Commun.*, 2022, 13, 419.
12 B. R. Goldsmith, M. Boley, J. Vreeken, M. Scheffler and L. M. Ghiringhelli, *New J. Phys.*, 2017, 19, 013031.
13 C. N. Singman, *J. Chem. Educ.*, 1984, 61, 137.
14 Dulal C Ghosh and R. Biswas, *Int. J. Mol. Sci.*, 2002, 3, 87–113.
15 J. C. Slater, *J. Chem. Phys.*, 1964, 41, 3199–3204.
16 J. E. Huheey, E. A. Keiter, R. L. Keiter and O. K. Medhi, *Inorganic chemistry: principles of structure and reactivity*, Pearson Education India, 2006.
17 W. Porterfield, *Read. MA*. *Inorganic Chemistry. A Unified Approach, Addison-wealey,*1984.
18 A. M. James and M. P. Lord, *Macmillan's chemical and physical data*, Macmillan London, 1992.
19 L. Pauling, *Ithaca N. Y. The nature of the chemical bond, Cornell University Press,*1960.
20 D. R. Lide, *Fla. USA*. Chemical Rubber Company handbook of chemistry and physics, Fla. USA ,1998.
21 J. A. Dean, Lange's Handbook of Chemistry, 1999.
22 A. M. James and M. P. Lord, *Macmillan's chemical and physical data*, Macmillan London, 1992.
23 G. W. C. Kaye and T. H. Laby, *Tables of physical and chemical constants and some mathematical functions*, Longmans, Green and Company Limited, 1928.