Supplementary Information

AI-Assisted Enhancement of Thermoelectric Properties in the $Yb_{14-x}Ca_xMn_{1-y}Al_ySb_{11}$ Zintl System

Junsu Lee,[1] Aziz Ahmed,[1] and Tae-Soo You[1*]

[1]Department of Chemistry, Chungbuk National University, Cheongju, Chungbuk 28644, Republic of Korea

CORRESPONDING AUTHOR

Name: Tae-Soo You

E-mail: tsyou@cbungbuk.ac.kr

Office: +82 (43) 261-2282

S.1. ML methodology

To accelerate the discovery of high-performance compositions within the $A_{14}MPn_{11}$-type Zintl family, a ML guided materials exploration strategy was employed. The following sections explain the details about the dataset preparation, feature generation, and model training and evaluation process.

Dataset and Preprocessing: The foundation of our models was a processed dataset derived from the publicly available "StarryData2" database.[30] Our starting point was a partially processed version of this dataset provided by Parse *et al.*,[9] which originally comprised 39,498 entries for 3,503 unique TE materials. This dataset contains the chemical composition, measured TE properties, and the corresponding measurement temperature for a wide range of important material classes, including Chalcogenides, Skutterudites, Zintls, Oxides, Clathrates, Silicides, and Half-Heuslers. Following the data cleaning procedure outlined by Parse *et al.*,[9] we refined the dataset even further. The final dataset used for model training consists of 17,844 entries, which covers 2,765 unique TE materials.

Feature Generation: The predictive accuracy of ML models depends critically on the input features. We employed a composition-based feature vector (CBFV) approach, which generates features directly from the chemical formula of each material in the dataset.[S1] Specifically, we used the MAGPIE descriptor set to generate a 154-dimensional feature vector for each composition.[31] MAGPIE-based features were chosen for their proven effectiveness in predictive modeling for materials discovery. Furthermore, the elemental properties, obtained from the MAGPIE package, are based on their Atomic Number, Atomic Weight, Periodic Table Row, Periodic Table Column, Filled Valence Orbitals (*s*, *p*, *d*, *f*), Unfilled Valence Orbitals (*s*, *p*, *d*, *f*), Mendeleev Number, Melting temperature, Covalent Radius, Electronegativity, GS Volume, GS Bandgap, GS Magnetic Moment and Space Group Number.

The final feature set for model training included these 154 MAGPIE features plus the measurement temperature as independent variables.

Model Training: The data was then partitioned into a training set (90%) and a test set (10%). This split was performed on a compositional basis, which ensures that no compositions present in the test set were included in the training set. This type of separation validates the ability of the model to generalize and predict properties for entirely new materials. The input features were then scaled and normalized, respectively, using the StandardScaler and Normalize functions from the scikit-learn library. We selected the XGBRegressor algorithm to predict the TE properties, where XGBoost is a powerful gradient-boosting framework that prevents overfitting through inherent regularization. In addition, it allows for performance tuning via numerous hyperparameters, and then provides quantitative feature importance and captures complex non-linear data trends.[9,S2] Three independent models were trained to predict σ, S, and κ_{tot}. Hyperparameter optimization for the three regression models was conducted using the RandomizedSearchCV utility in scikit-learn. The optimized values for key hyperparameters were as follows: learning_rate = 0.15, max_depth = 5, n_estimators = 1500, colsample_bytree = 0.7, subsample = 0.8, and reg_alpha = 7. These same hyperparameters were used for all subsequent analyses.

Model Evaluation: Model performance was quantified using the coefficient of determination (R^2) and the mean absolute error (MAE). An R^2 value of 1 indicates a perfect prediction, while an MAE close to 0 signifies minimal error. To ensure a robust estimate of the generalization capability of model, we performed a 10-fold cross-validation. In this procedure, the dataset was partitioned into ten folds, and each fold was iteratively held out as a test set while the remaining nine folds were used for training. In addition to these metrics, we performed a residual error analysis to further diagnose the performance of model. For each model, we analyzed the

distribution of the residuals (defined as the difference between the actual and predicted values) by plotting them as a frequency histogram. A well-behaved model should exhibit residuals that form an approximately normal distribution (a bell-shaped curve) centered at zero. A peak centered at zero indicates that the model has no systematic bias and is not consistently over- or under-predicting. After construction and rigorous evaluation, the trained ML models were used to predict the TE properties of previously unreported compositions across the target compositional space, which enables us to extrapolate property-temperature relationships and identify promising candidates for experimental synthesis.

S.2. Experimental methodology

Synthesis

All sample preparation steps were carried out within an argon-filled glovebox, where oxygen and moisture concentrations were kept below 0.1 ppm, or under vacuum. All reactants were bought from Alfa Aesar, and these included Yb (ingot, 99.9%), Ca (shot, 99.5%), Mn (pieces,99.95%), Al (pieces, 99.9%), Sb (shot, 99.9999%), and Sn (shot, 99.8%). The surfaces of Yb and Ca were cleaned by brushing with a metal brush inside an argon-filled glovebox before use. All reactant sets were introduced into alumina crucibles, accompanied by an excess quantity of Sn metal used as the metal-flux in the ratios of Yb:Ca:Mn:Al:Sb:Sn = 13:1:2.4:3.6:11:76 and 13:1:0.6:5.4:9:11:76 for $Yb_{13.30(6)}Ca_{0.70}Mn_{0.45(2)}Al_{0.55}Sb_{11}$ and $Yb_{12.61(6)}Ca_{1.39}Mn_{0.25(3)}Al_{0.75}Sb_{11}$, respectively. Notably, following previously established methods, the amounts of Mn and Al used in the loading mixtures correspond to 6 times their stoichiometric amounts in the target compositions $Yb_{13}CaMn_{0.4}Al_{0.6}Sb_{11}$ and $Yb_{13}CaMn_{0.1}Al_{0.9}Sb_{11}$, respectively. Each alumina crucible was inserted into a fused-silica tube and subsequently sealed under vacuum conditions to avoid oxidation of reactants during high-temperature processing. Then, the reactants were heated in a furnace according to the following profile: heated to 773 K over 2 h, held at 773 K for 2 h, further heated to 1273 K over 2 h, held at 1273 K for 6 h, and subsequently cooled to 973 K at a rate of 3 K/h. Afterward, the reaction ampoule was quickly taken out of the furnace and centrifuged for 3 min to separate the crystallized products from the molten Sn-metal flux.

For physical property measurements, bulk polycrystalline samples with target compositions $Yb_{14-x}Ca_xMn_{1-y}Al_ySb_{11}$ (x = 0.15, 0.65, 1; y = 0.6, 0.7, 0.8) were prepared by high-energy ball milling followed by SPS and Hot Pressing. These reactants were loaded into a stainless-steel vial along with one 0.5-inch and two 0.25-inch stainless-steel balls inside the

argon-filled glovebox. High-energy ball milling was performed using SPEX-8000M for 2 h to yield fine powders. The resulting powders were then divided and densified using two distinct techniques based on the composition. Specifically, the powders $Yb_{14-x}Ca_xMn_{1-y}Al_ySb_{11}$ (x = 0.6, 0.8) were densified via SPS process: heated to 973 K in 4 min, followed by ramping to 1023 K for over 1 min, and hold at 1023 K for 10 min under an applied pressure of 50 MPa. In contrast, samples $Yb_{14-x}Ca_xMn_{1-y}Al_ySb_{11}$ (x = 0.15, 0.65; y = 0.7, 0.8) were densified via hot pressing under the following conditions: heated at 1223 K for 2 h and held for 1.5 h under a 70 MPa uniaxial pressure.

X-ray crystallographic analysis

The crystal structure and phase purity of the title compound were characterized using PXRD and SXRD. The PXRD pattern was collected at room temperature using a Bruker D8 diffractometer equipped with an area detector and monochromatic Cu Kα_1 radiation (λ = 1.54187 Å). The collection step size was set to 0.02° in the 20° ≤ 2θ ≤ 80° range with a total exposure time of 30 min as shown in Fig. 5 and Supplementary Information Fig. S10. The phase purity of the sample was determined by comparing the collected PXRD patterns with the SXRD-simulated pattern of $Yb_{13.30(6)}Ca_{0.70}Mn_{0.45(2)}Al_{0.55}Sb_{11}$. The SXRD data of $Yb_{13.30(6)}Ca_{0.70}Mn_{0.45(2)}Al_{0.55}Sb_{11}$ and $Yb_{12.61(6)}Ca_{1.39}Mn_{0.25(3)}Al_{0.75}Sb_{11}$ were collected at room temperature using a Bruker SMART APEX2 CCD-based diffractometer equipped with Mo Kα_1 radiation (λ = 0.71073 Å). Initially, several metallic lustrous cubic-shaped single crystals were chosen from the product batch. After the quick quality check, the best single crystal was carefully selected, and a full data collection was conducted using Bruker APEX2 program.[S3] Data reduction, integration, and unit cell parameter refinements were executed using the SAINT program.[S4] SADABS was used to perform semiempirical absorption corrections based

on equivalents.[S5] The entire set of reflections from the compound was in good agreement with the tetragonal crystal system, and eventually, a space group $I4_1/acd$ (No. 142) was chosen for these isotypic Zintl phases. Detailed crystal structures were solved by direct methods and refined to convergence by full matrix least-squares methods on F^2. The refined parameters include the scale factor, atomic positions including anisotropic displacement parameters (ADPs), extinction coefficients, and the occupancy factors of the Yb/Ca and Mn/Al mixed sites. In the last stage of a refinement cycle, atomic positions were standardized using STRUCTURE TIDY. Important crystallographic data, atomic positions with ADPs, and several selected interatomic distances are provided in Table 1 and Supplementary Information Tables S1 and S2. These data can be obtained free of charge from The Cambridge Crystallographic Data Centre via www.ccdc.cam.ac.uk/data_request/cif. Depository number is CCDC–2505407 for $Yb_{13.30(6)}Ca_{0.70}Mn_{0.45(2)}Al_{0.55}Sb_{11}$ and CCDC–2505408 for $Yb_{12.61(6)}Ca_{1.39}Mn_{0.25(3)}Al_{0.75}Sb_{11}$.

Electronic structure calculation

To understand the overall electronic structures of the title compounds including the correlation between chemical compositions and band gap sizes, a series of theoretical TB-LMTO-ASA were performed on the four different structural models using the Stuttgart TB-LMTO47 program.[24-27] For practical reasons, these two hypothetical models were designed to have the tetragonal subgroup $I4_1$ (No. 80) rather than the refined space group tetragonal $I4_1/acd$ (No. 142) to apply the idealized compositions of $Yb_{13}CaMnSb_{11}$ and $Yb_{13}CaMn_{0.5}Al_{0.5}Sb_{11}$. Lattice parameters and atomic sites for these models were taken from the SXRD refinement results of $Yb_{11.80(1)}Ca_{2.20}MnSb_{11}$[12] and $Yb_{13.30(6)}Ca_{0.70}Mn_{0.45(2)}Al_{0.55}Sb_{11}$, respectively. Further structural details about this hypothetical model are provided in Supplementary Information Table S3. In

the ASA method, space is filled with overlapping Wigner-Seitz (WS) atomic spheres.[28] The local density approximation was adopted to treat exchange and correlation. All relativistic effects, except spin-orbit coupling, were taken into account using a scalar relativistic approximation.[28] The symmetry of the potential inside each WS sphere was considered spherical, and a combined correction was used to take into account the overlapping part. The radii of the WS sphere were obtained by requiring the overlapping potential to be the best possible approximation to the full potential and were determined by an automatic procedure.[28] The WS radii used for the two models are as follows: Yb, 1.929−1.956 Å; Ca, 1.954 Å; Mn, 1.613 Å; Sb, 1.775−1.960 Å for $Yb_{13}CaMnSb_{11}$; Yb, 1.934−1.974 Å; Ca, 1.958 Å; Mn, 1.603 Å; Al, 1.602 Å; Sb, 1.767−1.959 Å for $Yb_{13}CaMn_{0.5}Al_{0.5}Sb_{11}$. The used basis sets were 6s, 6p and 5d orbitals for Yb, 4s, 4p and 3d orbitals for Ca, 4s, 4p and 3d orbitals for Mn, 3s, 3p and 3d orbitals for Al, 5s, 5p, 5d and 4f orbitals for Sb. The Löwdin downfolding technique[S6] was employed for the Yb 6p, Ca 4p, Al 3d, and Sb 5d and 4f orbitals. Integration in *k*-space was performed using the tetrahedron method, and the self-consistent charge density was computed across the first Brillouin zone using 320 irreducible *k*-points for both models.[29]

Electrical transport property measurement

The four samples of $Yb_{14-x}Ca_xMn_{1-y}Al_ySb_{11}$ (x = 0.15, 0.65, 1; y = 0.6, 0.7, 0.8) were prepared into a bar shape (3 mm × 3 mm × 10 mm) for the electrical transport property measurements. The densities of these samples were proven higher than 95% according to the geometric density measurement method. The longer direction of each bar-shaped sample coincided with the direction in which the properties were measured. The σ and the S were simultaneously measured between 303 and 720 K under the He atmosphere using a ULVAC-RIKO ZEM-3 instrument system.

Thermal conductivity measurement

The thermal diffusivity of the three disk-shaped title compounds was determined using the flash diffusivity method with a Netzsch LFA 467 HyperFlash instrument, under an argon atmosphere, over the temperature range of 298 to 723 K. The measurement was conducted using a flash diffusion method where a short laser burst irradiated the front surface of the disk, while an IR detector recorded and analyzed the temperature change on the rear surface. The thermal conductivity κ was determined using the equation: $\kappa = DC_p\rho$, where D = thermal diffusivity, C_p = heat capacity, and ρ = density. The Dulong−Petit value ($3R$/atom, R = gas constant) was used for C_p.[22] The total thermal conductivity κ_{tot} was calculated by the sum of the lattice κ_{latt} and the electronic κ_{elec} thermal conductivities. The electronic thermal conductivity was evaluated using the Wiedemann-Franz law: $\kappa_{elec} = L\sigma T$, where L is the temperature-dependent Lorenz number, which was derived from the single-parabolic band model using the temperature-dependent Seebeck coefficient S.[23] Subsequently, the lattice thermal conductivity was calculated using: $\kappa_{latt} = \kappa_{tot} - \kappa_{elec}$.

S.3. Dataset and Model Analysis

The dataset used for training our ML models, comprising 2765 unique compounds, was categorized into established TE material families to understand its composition Fig. S2. The distribution is heavily weighted towards a few major families, with IV-VI chalcogenides (18.5%), Skutterudites (10.7%), and V-VI chalcogenides (8.3%) being the most prevalent.

The statistical profiles of the training data for σ, S, and κ_{tot} were visualized as histograms (Fig. S3). The distribution for $\log_{10}\sigma$, which spans from 0 to 7, is slightly left-skewed with a peak centered around 4.8. In contrast, the distribution for $\log_{10}\kappa_{tot}$ is slightly right-skewed, with most data concentrated between -0.25 and 0.75 and a peak at approximately 0.35. More distinctively, the S exhibits a non-symmetric bimodal distribution, a clear sign of the combined *p*-type and *n*-type materials sourced from the StarryData2 database. The two modes are centered near +150 μV/K (p-type) and -150 μV/K (n-type), with the taller peak for the p-type materials indicating their higher prevalence in the dataset.

Fig. S4 summarizes the temperature-dependent TE properties from the experimental dataset used to train our ML model. The data generally shows degenerate semiconductor or metallic behavior, and it highlights the common trade-off where materials with high σ (e.g., Silicides, Heuslers) also suffer from high κ_{tot}. In contrast, families like Chalcogenides and Argyrodites are excellent thermal insulators. Within this landscape, the $A_{14}MPn_{11}$-type Zintl family data is particularly noteworthy. It exhibits the behavior of a *p*-type degenerate semiconductor and, at high temperatures (~1150 K), combines a strong S (>200 μV/K) with a remarkably low κ_{tot} of approximately 0.6 W/m·K. This promising profile from the training data established the $A_{14}MPn_{11}$ system as a prime candidate for our ML-guided discovery and subsequent experimental validation.

The predictive accuracy and generalization capability of the trained models are visualized

through the residual error distributions shown in Fig. S5. A key indicator of model quality is the absence of systematic bias, which is confirmed by the fact that all six histograms are centered precisely at a residual error of zero. This demonstrates that the models do not systematically overestimate or underestimate the target properties. Furthermore, the distributions of the residual error for all properties approximate a Gaussian (bell-curve) shape, which implies that the errors are random rather than indicative of underlying patterns that the models failed to capture. A comparison between the training and testing sets reveals critical insights into model generalization. For all three properties, the residual error plots for the training data exhibit narrow and tall distributions (See Fig. S5(a), (c), (e)), which signifies a very low error margin and high precision on the data which the model has already seen. In contrast, the distributions of these errors for the testing data are visibly shorter and wider as observed in Fig. S5(b), (d), (f). This is particularly pronounced for σ, which also displays a higher frequency of outliers as observed in parity plot in Fig. 1(d). This classic signature indicates a degree of overfitting, where the models' predictive power of the models is somewhat diminished when applied to new and unseen data. This performance gap suggests that σ is inherently more sensitive to subtle variations in composition and electronic structure within the Zintl phase system, which makes it intrinsically harder to predict with high precision.

The XGBRegressor algorithm provided a feature importance score, as displayed in Fig. S6. First, the most important feature in predicting $\log_{10}\sigma$ is mode_GSbandgap, which refers to the mode number of the DFT bandgap of elemental solid among elements in the composition. The bandgap of a solid directly impacts its σ, where a smaller band gap allows electrons to easily jump to a higher energy level (conduction band) with minimal energy input, enabling current flow and high σ. Conversely, a large bandgap restricts electron movement, resulting in poor σ and making the material an insulator.[S7-S9] Second, the most important feature in

predicting S is max_NdUnfilled, which refers to the maximum number of unfilled valence d orbitals among elements in composition. The number of unfilled valence d orbitals in a material's composition significantly impacts its S, primarily through its influence on the electronic band structure and carrier behavior. Specifically, the presence of unfilled d orbitals can lead to a higher density of states near the Fermi level, which can even enhance the S. Additionally, the specific arrangement and filling of these d orbitals can affect the effective mass of charge carriers and their scattering mechanisms, which further influences the S.[S10-S14] Last, the max_Mendeleev number which refers to the Maximum Mendeleev number among elements in composition is the most important feature that affects $\log_{10}\kappa_{tot}$ prediction. The Mendeleev number, a measure of an position of elements in the periodic table, influences κ_{tot} indirectly through its correlation with atomic mass, electron configuration, and bonding characteristics. These factors, in turn, affect how heat is transferred through the material.[S15-S17] Other important features that affect their respective ML models can be explained in a similar way. We have also performed the feature engineering test in which the least important features were removed gradually from the full list of 154, and the performance of each model was evaluated on the remaining features in terms of R^2 values. The results of the feature engineering test on the training dataset (see Fig. S7) shows that the R^2 values for the $\log_{10}\sigma$, S, and $\log_{10}\kappa_{tot}$ models remained high (above 0.9) even as a large number of features were removed. Performance did not degrade significantly until the feature count was reduced to 7, 3, and 7 features for each model, respectively. Therefore, given that predictive performance was maintained across a wide range of feature counts, we have used the full set of 154 MAGPIE features for model evaluation to avoid prematurely discarding potentially useful information.

S.4. Benchmarking Model Predictions Against Experimental Data

To test the practical utility of our ML models, we validated their predictions against experimental results for our title system at 720 K, as shown in Fig. S12. This case study serves not only to assess the numerical accuracy of the predictions but, more importantly, to validate the ability of the model to guide a successful materials discovery strategy. As established in our broader statistical analysis (Fig. 1, 2, and S5), the models for $\log_{10}\kappa_{tot}$ and S are highly robust, while the model for $\log_{10}\sigma$ is the most challenging. The results of this six-compound study align perfectly with these established characteristics, which showcase both the strengths of the model in trend prediction as well as its quantitative limitations.

A key success of the ML model was its ability to correctly forecast the compositional trends across the material system. In our experiments, we practically observed a sequential decrease in σ with the introduction of Ca and then Al, a trend that was successfully captured by our ML model. Similarly, for the S, the model correctly forecasted the positive sign, the linear temperature dependence, and the significant enhancement of S values upon co-substitution. For κ_{tot}, the model and experiment were again in agreement with both showing that co-substitution effectively lowered κ_{tot} through enhanced alloy scattering. While the qualitative trends were predicted with high fidelity, the quantitative accuracy varied by property. In line with its higher statistical robustness, the S prediction for $Yb_{13}CaMn_{0.1}Al_{0.9}Sb_{11}$ was exceptionally accurate (0.23% error). However, the model struggled with the magnitude of change for σ in the Al-substituted compounds, which leads to a quantitative overestimation (e.g., 103.46% for $Yb_{13}CaMn_{0.1}Al_{0.9}Sb_{11}$). This specific inaccuracy is a practical manifestation of the known statistical variance of the model for σ, which is inherently more difficult to predict due to complex scattering effects.

The ultimate test of the model was its ability to identify a path toward improved TE

performance, and our current model proved to be an outstanding success. As shown in Fig. S12(d), the experimental *ZT* was sequentially improved by the co-substitution strategy that the ML model originally proposed. The final quinary compounds significantly outperformed their ternary and quaternary predecessors, which confirms that the model successfully identified a valid route to superior *ZT* values. While the final *ZT* values for the Al-substituted materials were quantitatively overestimated—a direct and expected consequence of propagating the error from the σ prediction—the strategic guidance was fundamentally correct. For instance, the model achieved a brilliant prediction for the non-substituted $Yb_{13}CaMnSb_{11}$ (0.62% error) and, more importantly, correctly identified that co-substitution was the key to unlocking higher performance. This case study, therefore, validates our ML approach not as a perfect numerical solution, but as a powerful and effective tool that helps navigate complex compositional spaces and accelerates the discovery of high-performance TE materials.

Table S1 Atomic coordinates and equivalent isotropic displacement parameters (U_{eq}[a]) from the SXRD refinements for $Yb_{13.30(6)}Ca_{0.70}Mn_{0.45(2)}Al_{0.55}Sb_{11}$ and $Yb_{12.61(6)}Ca_{1.39}Mn_{0.25(3)}Al_{0.75}Sb_{11}$

Atom	Wyckoff site	Occupation	*x*	*y*	*z*	U_{eq}[a](Å^2)
			$Yb_{13.30(6)}Ca_{0.70}Mn_{0.45(2)}Al_{0.55}Sb_{11}$			
M1	32*g*	0.948(4)/0.052	0.02219(2)	0.37637(2)	0.00394(2)	0.01517(15)
M2	32*g*	0.954(4)/0.046	0.04252(2)	0.07355(2)	0.17145(2)	0.01152(14)
M3	32*g*	0.949(4)/0.051	0.34095(3)	0.06991(2)	0.09236(2)	0.01588(15)
M4	16*e*	0.950(4)/0.050	0.35496(3)	0	1/4	0.01088(19)
Mn/Al	8*a*	0.45(2)/0.55	0	1/4	3/8	0.0081(10)
Sb1	32*g*	1	0.36050(3)	0.25443(3)	0.05968(3)	0.01186(16)
Sb2	32*g*	1	0.13050(3)	0.02613(3)	0.04736(3)	0.01179(16)
Sb3	16*f*	1	0.13593(3)	0.38593(3)	1/8	0.00989(18)
Sb4	8*b*	1	0	1/4	1/8	0.0122(2)
			$Yb_{12.61(6)}Ca_{1.39}Mn_{0.25(3)}Al_{0.75}Sb_{11}$			
M1	32*g*	0.898(4)/0.102	0.02213(3)	0.37660(3)	0.00457(2)	0.0129(2)
M2	32*g*	0.912(4)/0.088	0.04281(3)	0.07350(3)	0.17126(2)	0.0097(2)
M3	32*g*	0.890(4)/0.110	0.34082(3)	0.06955(3)	0.09219(2)	0.0137(2)
M4	16*e*	0.905(5)/0.095	0.35504(4)	0	1/4	0.0091(3)
Mn/Al	8*a*	0.25(3)/0.75	0	1/4	3/8	0.0087(17)
Sb1	32*g*		0.36153(4)	0.25418(4)	0.05966(3)	0.0097(2)
Sb2	32*g*		0.13044(4)	0.02578(4)	0.04731(3)	0.0097(2)
Sb3	16*f*		0.13582(4)	0.38582(4)	1/8	0.0083(3)
Sb4	8*b*	1	0	1/4	1/8	0.0101(3)

[a]M = Ca/Yb mixed-site.

Table S2 Selected interatomic distances (Å) for $Yb_{13.30(6)}Ca_{0.70}Mn_{0.45(2)}Al_{0.55}Sb_{11}$ and $Yb_{12.61(6)}Ca_{1.39}Mn_{0.25(3)}Al_{0.75}Sb_{11}$

Atomic pair	Bond distance (Å)	
	$Yb_{13.30(6)}Ca_{0.70}Mn_{0.45(2)}Al_{0.55}Sb_{11}$	$Yb_{12.61(6)}Ca_{1.39}Mn_{0.25(3)}Al_{0.75}Sb_{11}$
M1[a]-Sb1	3.1490(6)	3.1422(9)
M1[a]-Sb1	3.7373(6)	3.7192(9)
M1[a]-Sb2	3.1651(6)	3.1511(9)
M1[a]-Sb2	3.2794(6)	3.2675(9)
M1[a]-Sb3	3.2847(5)	3.2674(7)
M1[a]-Sb4	3.4265(4)	3.4120(6)
M2[a]-Sb1	3.1893(6)	3.1962(9)
M2[a]-Sb1	3.2171(8)	3.2208(10)
M2[a]-Sb2	3.1897(6)	3.1830(9)
M2[a]-Sb2	3.2111(8)	3.2024(9)
M2[a]-Sb3	3.2149(6)	3.2079(9)
M2[a]-Sb4	3.1914(4)	3.1830(6)
M3[a]-Sb1	3.1723(6)	3.1661(9)
M3[a]-Sb1	3.4325(7)	3.4371(9)
M3[a]-Sb2	3.2125(8)	3.205(1)
M3[a]-Sb2	3.2307(8)	3.2256(9)
M3[a]-Sb2	3.7142(7)	3.7018(9)
M3[a]-Sb3	3.1700(6)	3.1582(9)
M4[a]-Sb1 (× 2)	3.2568(6)	3.2597(9)
M4[a]-Sb2 (× 2)	3.1333(6)	3.1217(8)
M4[a]-Sb3 (× 2)	3.3590(3)	3.3548(6)
Mn/Al-Sb1 (× 4)	2.7364(6)	2.7154(8)
Sb3-Sb4 (× 2)	3.1999(5)	3.1877(7)

[a]M = Ca/Yb mixed-site.

Table S3 Detailed structural information of the structure model of $Yb_{13}CaMnSb_{11}$ and $Yb_{13}CaMn_{0.5}Al_{0.5}Sb_{11}$

Chemical formula	$Yb_{13}CaMnSb_{11}$		$Yb_{13}CaMn_{0.5}Al_{0.5}Sb_{11}$
Space group		$I4_1$ (No.80)	
Unit cell dimensions (Å)	a = 16.5885		a = 16.646
	c = 21.976		c = 22.1332
Volume (Å^3)	6047.319		6132.873

Atom	Wyckoff site	Atomic coordinates		
		x	y	z
		$Yb_{13}CaMnSb_{11}$		
Yb1	8*b*	0.0224	0.1296	0.378
Yb2	8*b*	0.4776	0.1296	0.372
Yb3	8*b*	0.9776	0.3704	0.372
Yb4	8*b*	0.5224	0.3704	0.378
Yb5	8*b*	0.0422	0.8237	0.5469
Yb6	8*b*	0.4578	0.8237	0.2031
Yb7	8*b*	0.9578	0.6763	0.2031
Yb8	8*b*	0.5422	0.6763	0.5469
Yb9	8*b*	0.3413	0.8205	0.4676
Yb10	8*b*	0.1587	0.8205	0.2824
Yb11	8*b*	0.6587	0.6795	0.2824
Yb12	8*b*	0.8413	0.6795	0.4676
Yb13	8*b*	0.3548	0.25	0.125
Yb14	8*b*	0.8548	0.25	0.125
Mn1	4*a*	0	0	0.25
Mn2	4*a*	0	0	0.75
Sb1	8*b*	0.3587	0.0048	0.4347
Sb2	8*b*	0.1413	0.0048	0.3153
Sb3	8*b*	0.6413	0.4952	0.3153
Sb4	8*b*	0.8587	0.4952	0.4347
Sb5	8*b*	0.8697	0.7231	0.3275
Sb6	8*b*	0.6303	0.7231	0.4225
Sb7	8*b*	0.1303	0.7769	0.4226
Sb8	8*b*	0.3697	0.7769	0.3274
Sb9	8*b*	0.6357	0.6357	0

Sb10	8*b*	0.8643	0.1357	0
Sb11	4*a*	0	0	0
Sb12	4*a*	0	0	0.5
		$Yb_{13}CaMn_{0.5}Al_{0.5}Sb_{11}$		
Yb1	8*b*	0.0222	0.1264	0.3789
Yb2	8*b*	0.4778	0.1264	0.3711
Yb3	8*b*	0.9778	0.3736	0.3711
Yb4	8*b*	0.5222	0.3736	0.3789
Yb5	8*b*	0.0425	0.8235	0.5465
Yb6	8*b*	0.4575	0.8235	0.2035
Yb7	8*b*	0.9575	0.6765	0.2035
Yb8	8*b*	0.5425	0.6765	0.5465
Yb9	8*b*	0.159	0.8199	0.2826
Yb10	8*b*	0.659	0.6801	0.2826
Yb11	8*b*	0.841	0.6801	0.4674
Yb12	8*b*	0.355	0.25	0.125
Yb13	8*b*	0.855	0.25	0.125
Ca	8*b*	0.341	0.8199	0.4674
Mn	4*a*	0	0	0.75
Al	4*a*	0	0	0.25
Sb1	8*b*	0.3605	0.0044	0.4347
Sb2	8*b*	0.1395	0.0044	0.3153
Sb3	8*b*	0.6395	0.4956	0.3153
Sb4	8*b*	0.8605	0.4956	0.4347
Sb5	8*b*	0.1305	0.7761	0.4224
Sb6	8*b*	0.3695	0.7761	0.3276
Sb7	8*b*	0.8695	0.7239	0.3276
Sb8	8*b*	0.6305	0.7239	0.4224
Sb9	8*b*	0.6359	0.6359	0
Sb10	8*b*	0.8641	0.1359	0
Sb11	4*a*	0	0	0
Sb12	4*a*	0	0	0.5

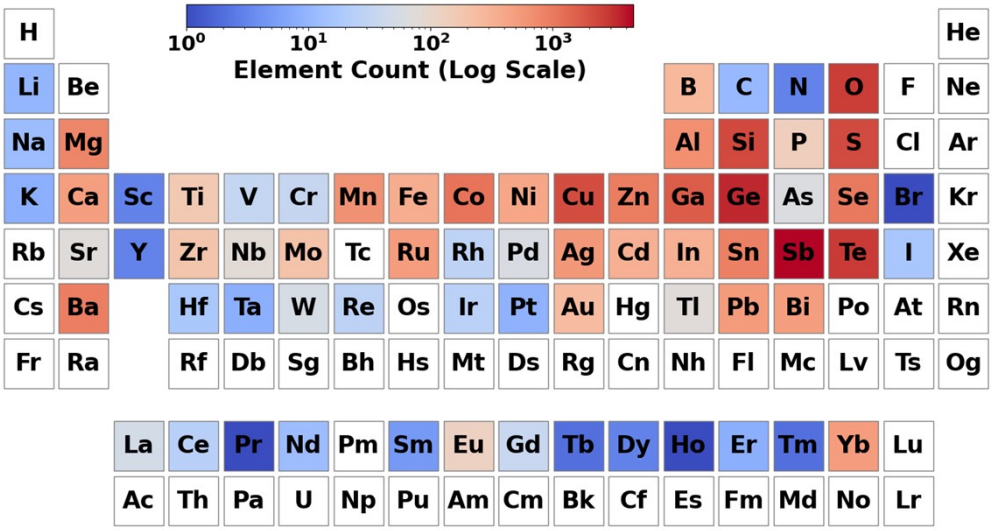

Fig. S1 A periodic table heatmap that illustrates the frequency of each element within the dataset. The color intensity, on a logarithmic scale, corresponds to the total count of each element across all materials.

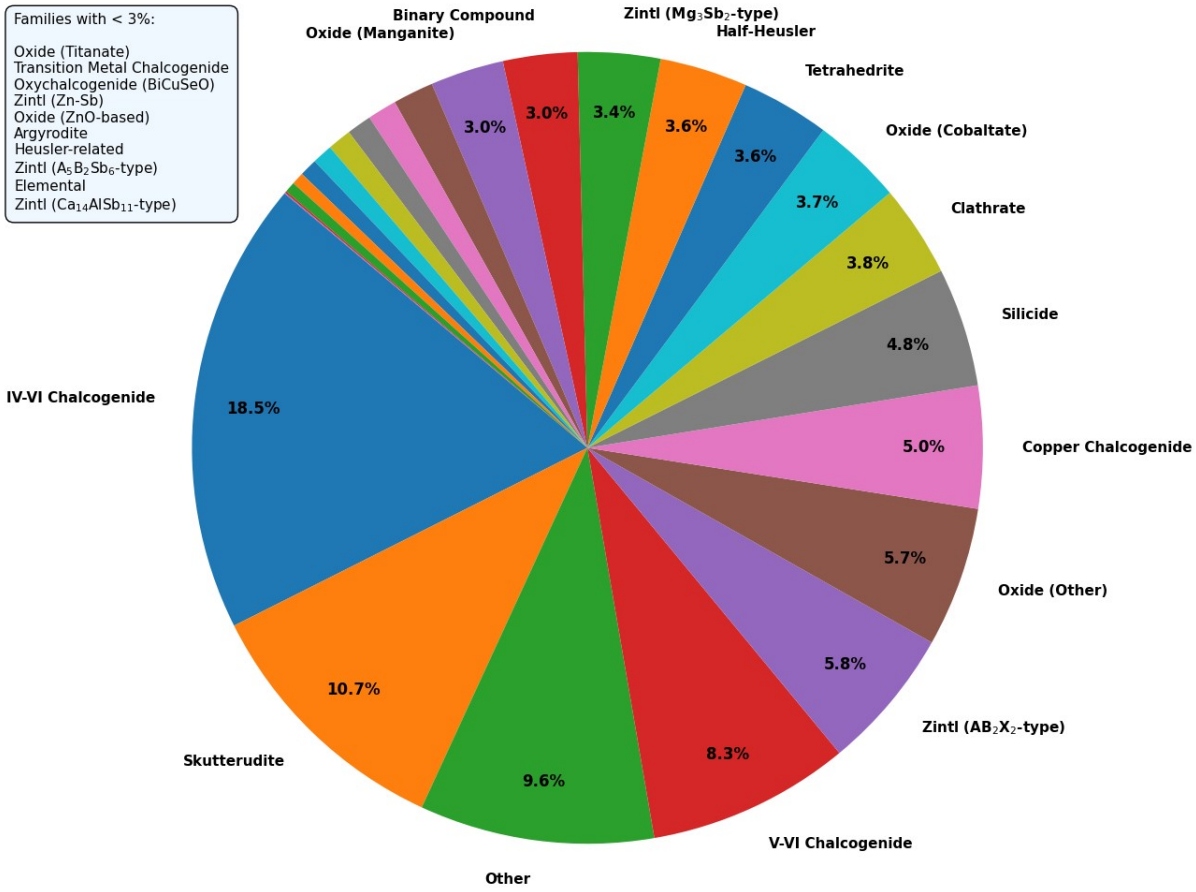

Fig. S2 A pie chart illustrating the distribution of various material families within the dataset of 2,765 unique TE compounds. A top left box shows the compositions < 3%.

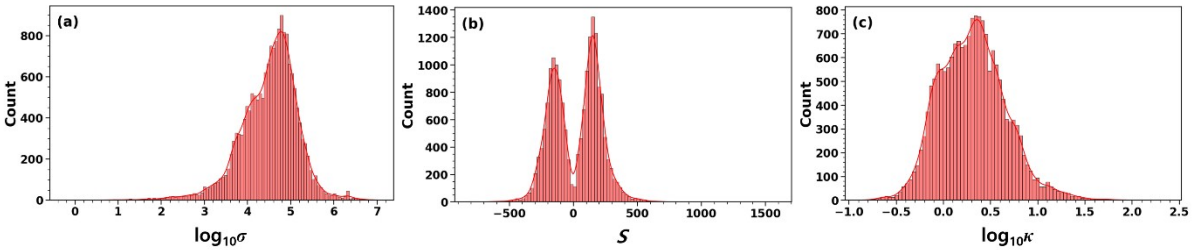

Fig. S3 Histograms visualizing the distribution of the training datasets for (a) the $\log_{10}\sigma$, (b) the S, and (c) the $\log_{10}\kappa_{\mathrm{tot}}$.

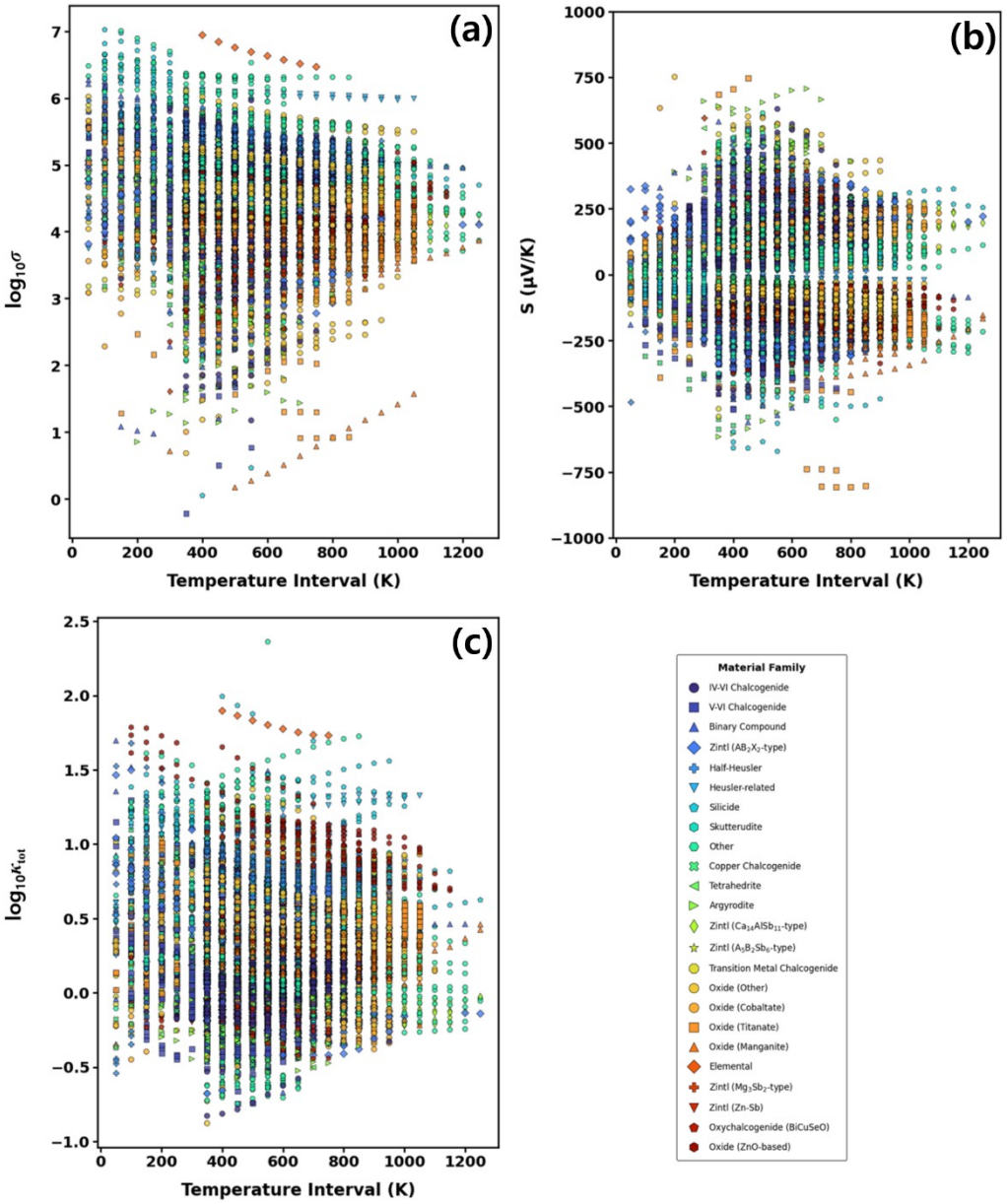

Fig. S4 TE properties of various material families from the experimental training dataset, plotted as a function of temperature for (a) $\log_{10}\sigma$, (b) S, and (c) $\log_{10}\kappa_{tot}$.

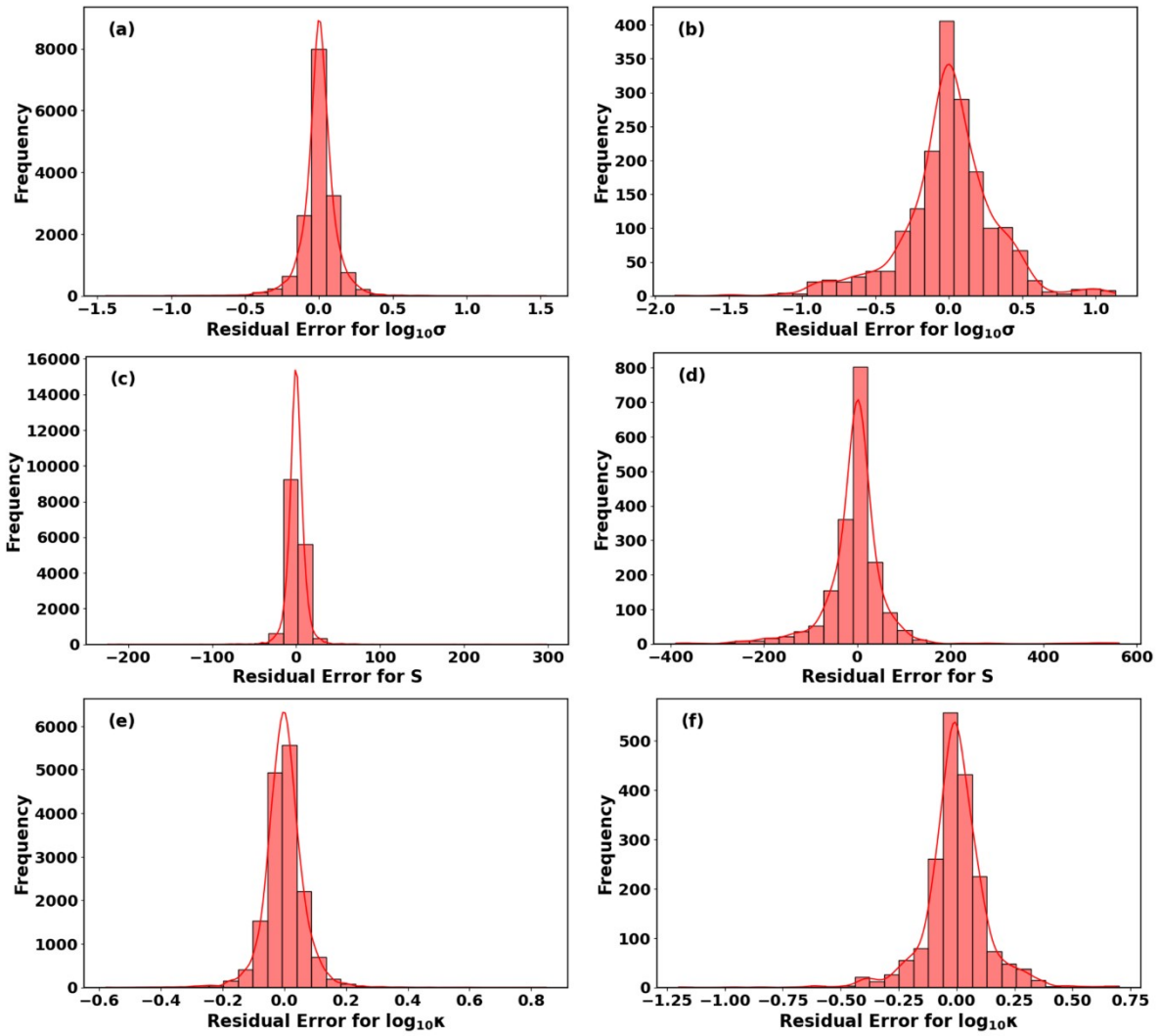

Fig. S5 Histograms of the residual errors (predicted value – true value) for the three independent ML models on both the training and testing datasets for (a, b) $\log_{10}\sigma$, (c, d) S, and (e, f) $\log_{10}\kappa_{tot}$.

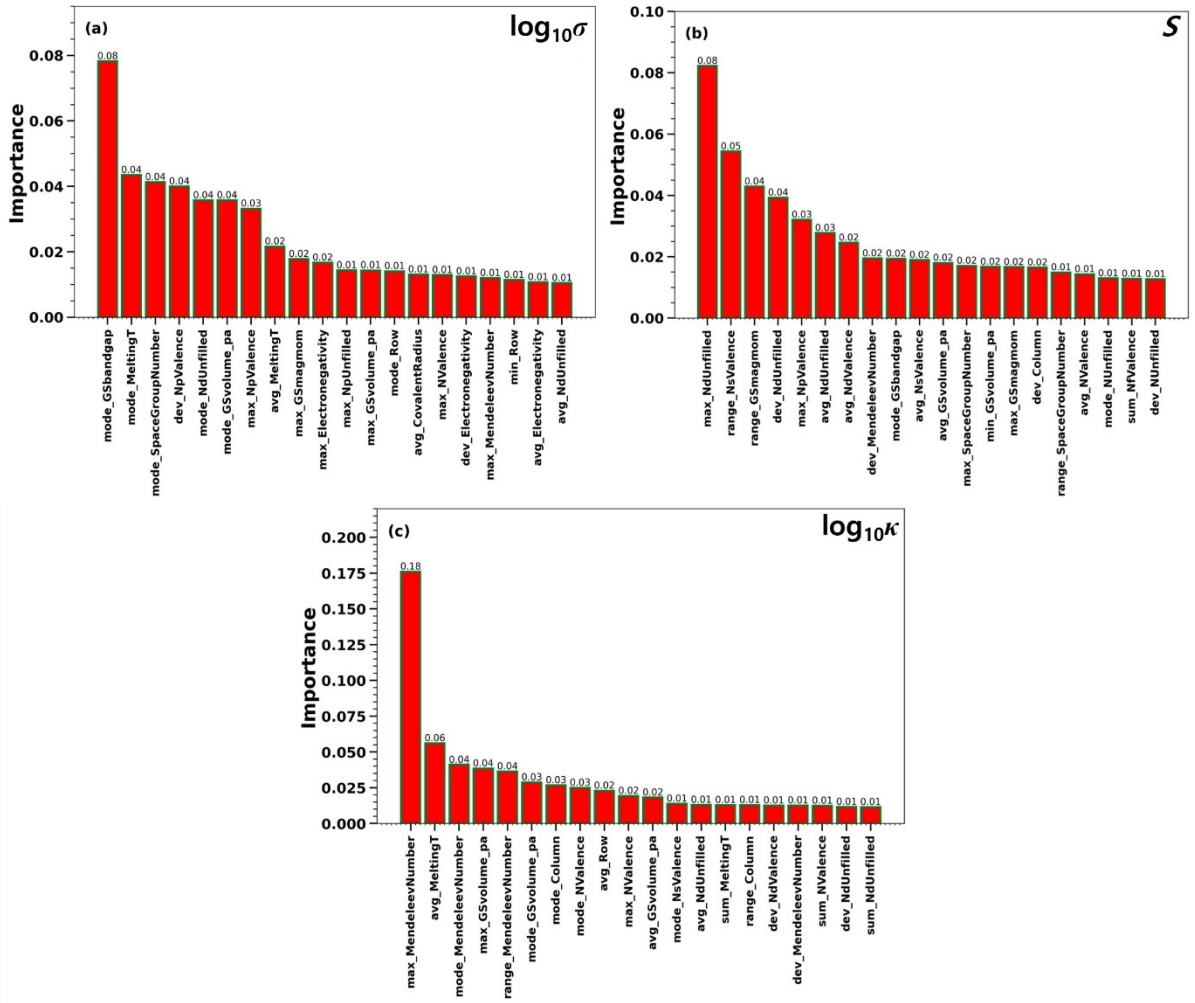

Fig. S6 Top 20 features and corresponding scores of TE properties (a) $\log_{10}\sigma$, (b) S, and (c) $\log_{10}\kappa_{tot}$ returned by the XGBRegressor model.

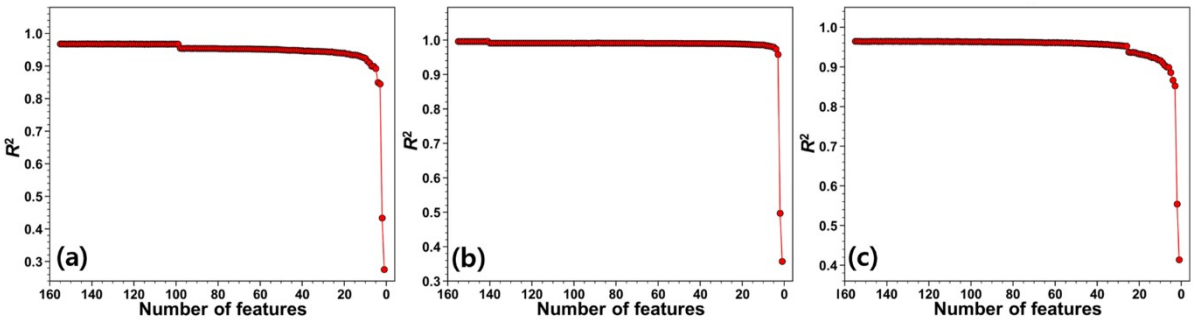

Fig. S7 Prediction performance of (a) $\log_{10}\sigma$ model, (b) S model, and (c) $\log_{10}\kappa_{tot}$ model as a function of reducing the number of features.

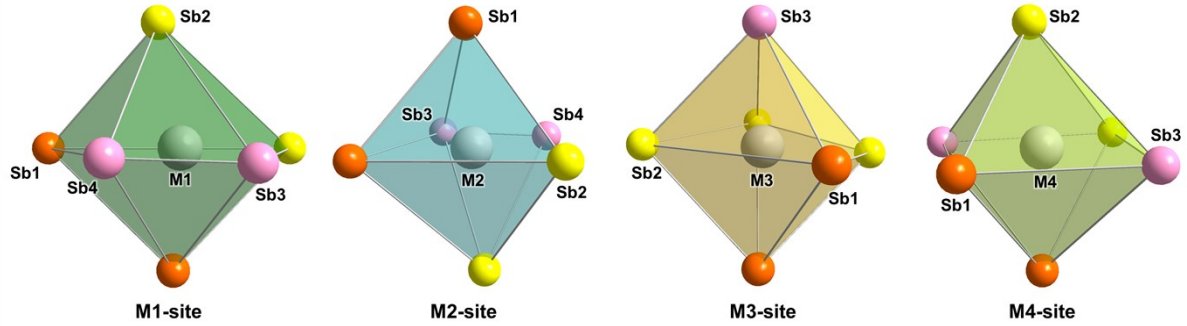

Fig. S8 Four cationic sites in $Yb_{13.30(6)}Ca_{0.70}Mn_{0.45(2)}Al_{0.55}Sb_{11}$ are illustrated as the octahedral coordination polyhedra formed by six Sb atoms. Color codes: M (Yb/Ca mixed-site), light gray; Sb1, orange; Sb2, yellow; and, Sb3 and Sb4, pink.

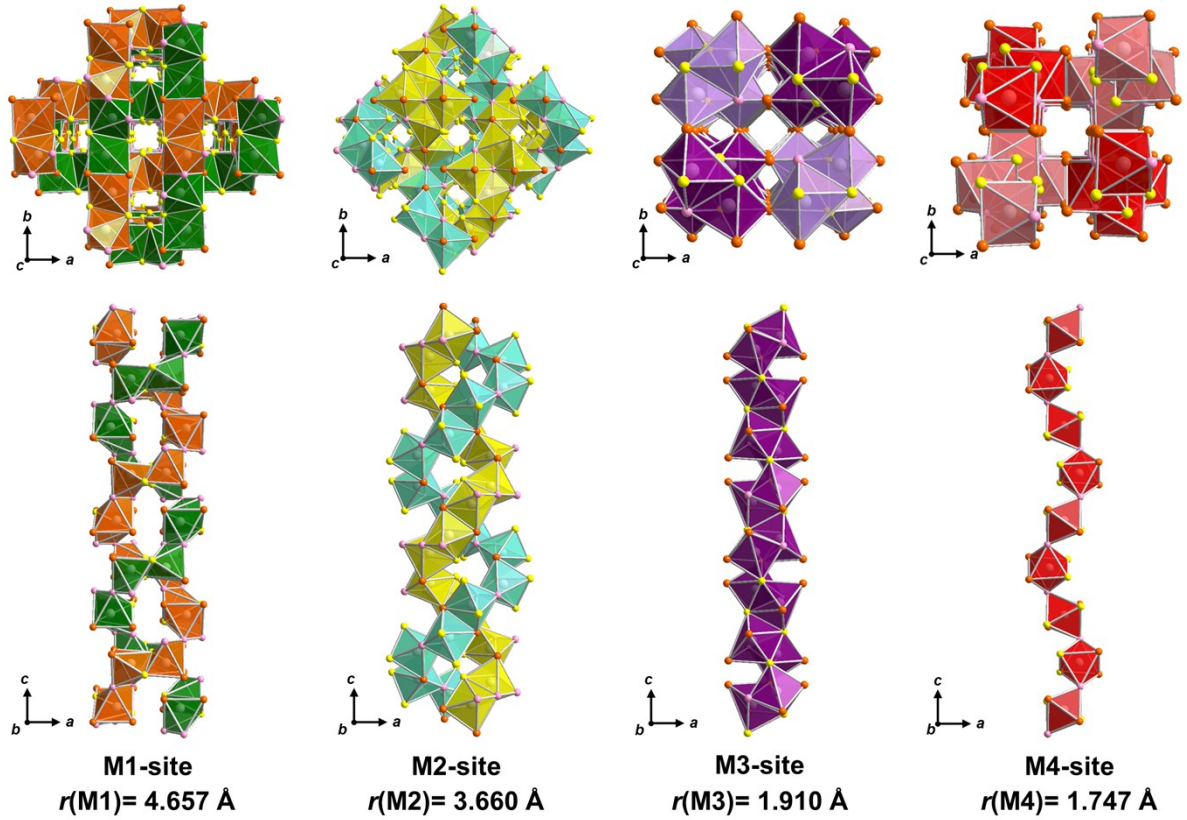

Fig. S9 Four distinct types of 1D spiral-shaped chains propagating along the c-axis direction in the unit cell of $Yb_{13.30(6)}Ca_{0.70}Mn_{0.45(2)}Al_{0.55}Sb_{11}$.

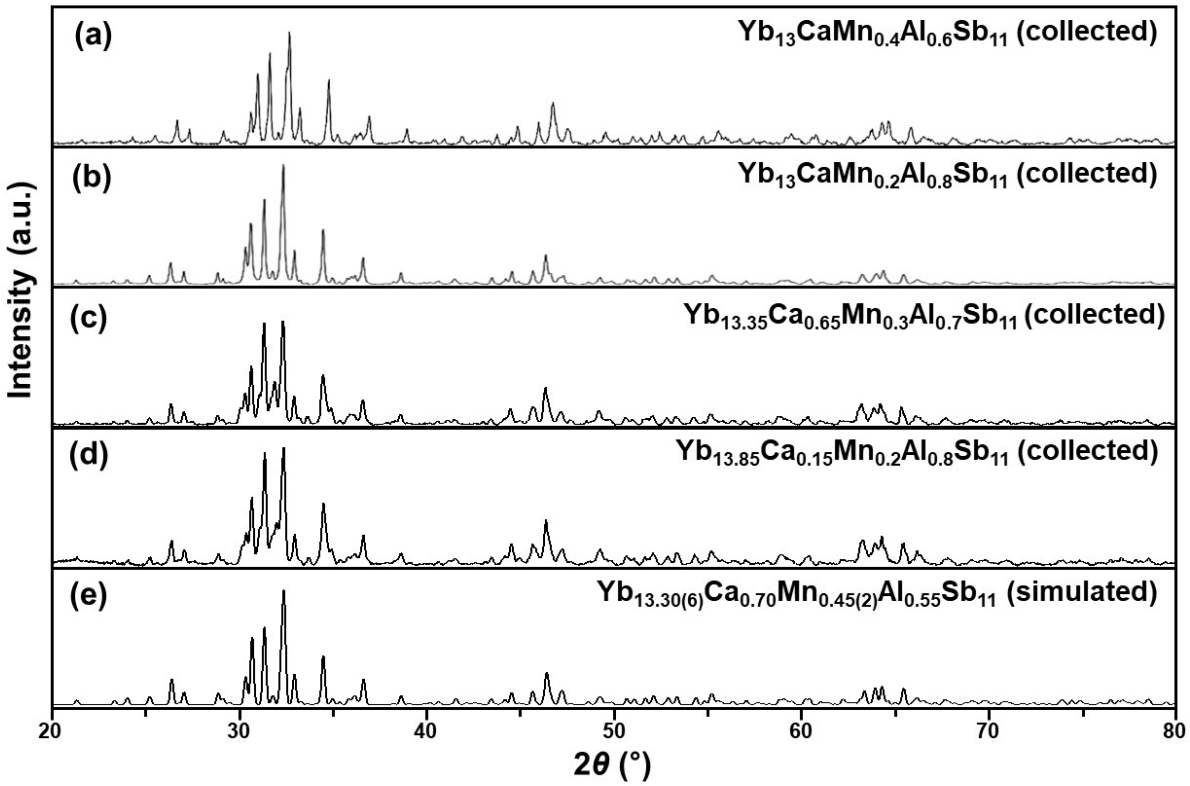

Fig. S10 Collected PXRD patterns of the three title compounds (a) $Yb_{13}CaMn_{0.4}Al_{0.6}Sb_{11}$, (b) $Yb_{13}CaMn_{0.2}Al_{0.8}Sb_{11}$, (c) $Yb_{13.35}Ca_{0.65}Mn_{0.3}Al_{0.7}Sb_{11}$, and (d) $Yb_{13.85}Ca_{0.15}Mn_{0.2}Al_{0.8}Sb_{11}$synthesized using ball milling and SPS method. A simulated PXRD pattern is also provided as a reference.

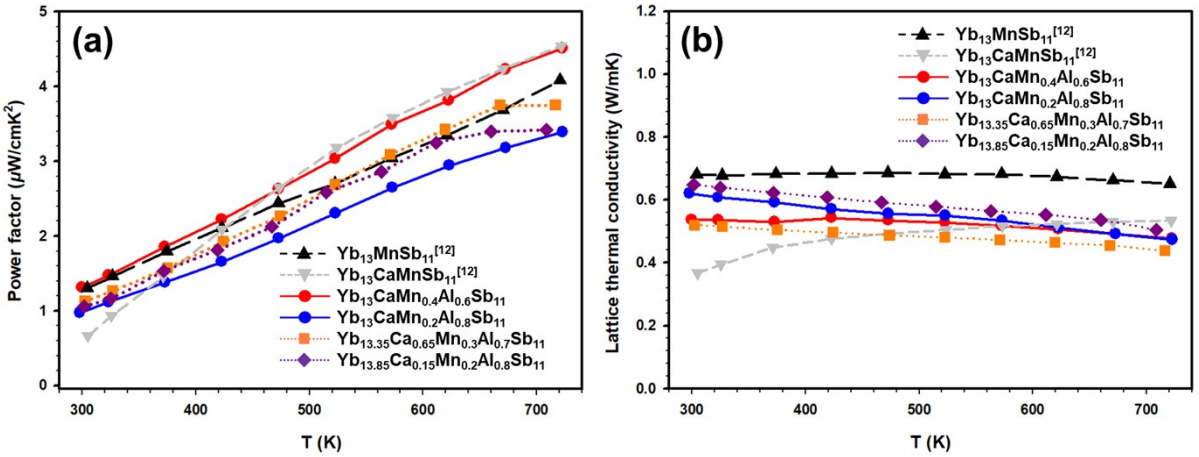

Fig. S11 Temperature-dependent (a) *PF* and (b) κ_{latt} of four title compounds $Yb_{13}CaMn_{0.4}Al_{0.6}Sb_{11}$, $Yb_{13}CaMn_{0.2}Al_{0.8}Sb_{11}$, $Yb_{13.35}Ca_{0.65}Mn_{0.3}Al_{0.7}Sb_{11}$ and $Yb_{13.85}Ca_{0.15}Mn_{0.2}Al_{0.7}Sb_{11}$ measured over the temperature range of 298-723 K. The reported TE properties of two reference compounds, $Yb_{14}MnSb_{11}$ and $Yb_{13}CaMnSb_{11}$, are also plotted for comparison purposes.

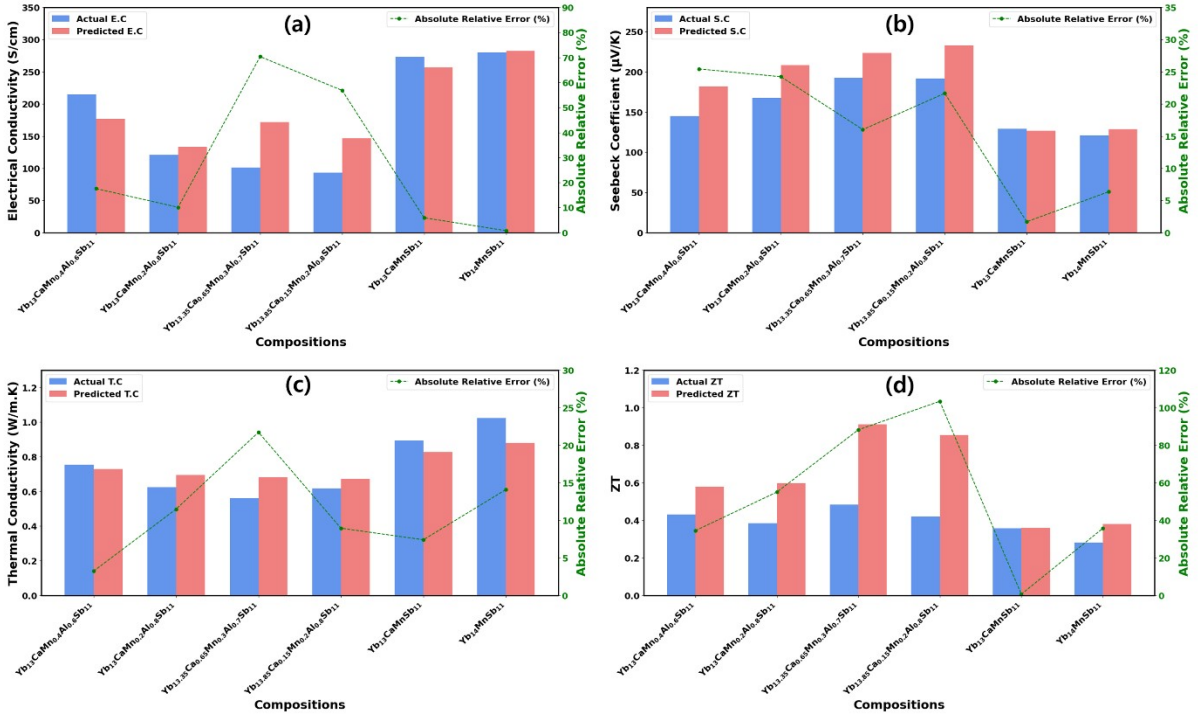

Fig. S12 Validation of the predictive model against experimental data for the title system at 720 K. The panels compare the actual (experimental, blue bars) and model-predicted (red bars) values for (a) σ, (b) S, (c) κ_{tot}, and (d) ZT. The green dashed line in each panel plots the absolute relative error percentage between the actual and predicted values, corresponding to the right-hand y-axis.

Reference

S1 S. K. Kauwe, J. Graser, A. Vazquez and T. D. Sparks, *Integr. Mater. Manuf. Innov.*, 2018, **7**, 43–51.

S2 K. E. Vipin and P. Padhan, *J. Mater. Chem. C*, 2024, **12**, 7415-7425.

S3 APEX2 (ver. 2013.6-2), Bruker AXS Inc., Madison, WI, 2013.

S4 SAINT (ver. 8.32B), Bruker AXS Inc., Madison, WI, 2013.

S5 G. M. Sheldrick, SADABS (ver. 2016/2), University of Göttingen, Göttingen, 2016.

S6 O. Jepsen and O. K. Andersen, *Z. Phys. B*, 1995, **97**, 35–47.

S7 A. Ali, A. W. Anwar, M. Moin, M. Moin, S. Nabi, R. B. Ahmed, S. Ahmad, A. Ali and S. Shaheen, *Physica B: Condensed Matter*, 2025, **712**, 417211.

S8 F. Ottomano, J. Y. Goulermas, V. Gusev, R. Savani, M. W. Gaultois, T. D. Manning, H. Lin, T. P. Manzanera, E. G. Poole, M. S. Dyer, J. B. Claridge, J. Alaria, L. M. Daniels, S. Varma, D. Rimmer, K. Sanderson and M. J. Rosseinsky, *Digit. Discov.*, 2025, **4**, 1794–1811.

S9 J. M. Crowley, J. Tahir-Kheli and W. A. Goddard III, *J. Phys. Chem. Lett.*, 2016, **7**, 1198–1203.

S10 G. Grissonnanche, G. A. Pan, H. LaBollita, D. F. Segedin, Q. Song, H. Paik, C. M. Brooks, E. Beauchesne-Blanchet, J. L. S. González, A. S. Botana, J. A. Mundy and B. J. Ramshaw, *Phys. Rev. X*, 2024, **14**, 041021.

S11 W. Dou, K. B. Spooner, S. R. Kavanagh, M. Zhou and D. O. Scanlon, *J. Am. Chem. Soc.*, 2024, **146**, 17679–17690.

S12 Y. Dai, W. Zhou, H.-J. Kim, Q. Song, X. Qian, T.-H. Liu and R. Yang, *npj Comput. Mater.*, 2022, **8**, 234.

S13 F. Garmroudi, M. Parzer, A. Riss, C. Bourgès, S. Khmelevskyi, T. Mori, E. Bauer and A. Pustogow, *Sci. Adv.*, 2023, **9**, eadj1611.

S14 M. Rafique, B. A. Kalwar, T. Xinyu, S. Yong, B. Wang, L. Li and Y. Yang, *Mater. Sci. Semicond. Process.*, 2024, **182**, 108706.

S15 A. B. Georgescu, P. Ren, A. R. Toland, S. Zhang, K. D. Miller, D. W. Apley, E. A. Olivetti, N. Wagner and J. M. Rondinelli, *Chem. Mater.*, 2021, **33**, 5591–5605.

S16 A. Zhang and Y. Li, *J. Mater. Res.*, 2023, **38**, 2049–2058.

S17 M. T. Agne, T. T. Bernges and W. G. Zeier, *PRX Energy*, 2022, **1**, 03100.