Electronic Supplementary Material (ESI) for Journal of Materials Chemistry A.

Supplementary Information

Inverse design of experimentally synthesizable crystal structures by leveraging computational and experimental data

S1. ALIGNN

ALIGNN comprises two types of graphs: atomic graphs and line graphs. An atomic graph is denoted as G_A = (atom_v_i, bond_e_{ij}), where atomic sites serve as graph nodes (v_i) and bonds between sites as graph edges (e_{ij}). In the context of crystals, a bond is defined as the connection between atoms with a distance less than a certain cutoff value R. A line graph is represented as G_L = (bond_v_{i_j}, angle_e_{ij}), using bonds as graph nodes and the angles formed between two bonds as graph edges. The nodes in the line graph correspond one-to-one with the edges in the atomic graph, sharing the same representation. For crystal structures with fewer than three atoms, a 2·2·2 expansion is performed in the x, y, and z directions before converting to a graph structure. This expansion ensures an adequate number of edges and bond angles to ensure the successful use of ALIGNN.

ALIGNN can be generalized to the message passing neural network (MPNN) architecture[33] whose core concept involves establishing relationships and feature representations between nodes through message passing. The operations of MPNN typically encompass several key steps, including:

(1) Initializing node and edge representations. The representation of atom_v_i is initialized through a one-hot encoding of the atomic number and the degree. Each bond_e_{ij} or bond_v_i is embedded with a vector containing distance information (r_{ij}) between neighboring atoms i and j. The angle_e_{ij} is embedded with a vector that includes angle information ($\theta_{ij,jk}$) between neighboring bonds ij and jk. Specifically, the initiation of the feature vector is accomplished by employing a Gaussian Basis Function (GBF).

$$atom\_v_i^0 = Onehot\left(atomic\ number\right) \oplus Onehot\left(\text{degree}\right) \text{* MERGEFORMAT (2)}$$

$$bond\_e_{ij}^0 = bond\_v_{i\_j}^0 = \exp\left[-\eta\left(r_{ij} - \mu\right)^2\right] \text{* MERGEFORMAT (3)}$$

$$angle\_e_{ij,jk}^0 = \exp\left[-\eta\left(\theta_{ij,jk} - \mu\right)^2\right] \quad \text{* MERGEFORMAT (4)}$$

(2) Message passing. The essence of MPNN lies in updating node representations through message passing. In each message passing stage, each node receives messages from its adjacent nodes as well as the edges connecting them. The update formula for node attributes and edge attributes can be expressed as follows:

$$v_i^{t+1} = v_i^t + SiLU\left(LayerNorm\left(W_{src}^t v_i^t + \sum \hat{e}_{ij}^t W_{dst}^t v_i^t\right)\right) \quad \text{* MERGEFORMAT (5)}$$

$$\hat{e}_{ij}^t = \sigma\left(e_{ij}^t\right) \Big/ \left(\sum_{k \in N_i} \sigma\left(e_{ik}^t\right) + \varepsilon\right) \quad \text{* MERGEFORMAT (6)}$$

$$e_{ij}^t = e_{ij}^{t-1} + SiLU\left(LayerNorm\left(W_{gate}^t Z_{ij}^{t-1}\right)\right) \quad \text{* MERGEFORMAT (7)}$$

$$z_{ij} = h_i \oplus h_j \oplus e_{ij} \quad \text{* MERGEFORMAT (8)}$$

Here, z_{ij} represents the concatenation of information from connected nodes and edges in the graph; The edge gate vector is denoted as $_{ij}$; W_{src}, W_{dst}, and W_{gate} are weight matrices. The activation function SiLU (Sigmoid Linear Unit) is applied, and σ represents a sigmoid activation. Additionally, LayerNorm signifies the Layer Normalization operation.

(3) Message aggregation. After nodes receive messages from neighboring nodes, these messages are aggregated, typically through pooling operations such as summation or averaging. This process results in a graph representation (v_g) that captures information from all nodes.

$$v_g = Pool\left(v_0^T, v_1^T, ..., v_n^T\right) \quad \text{* MERGEFORMAT (9)}$$

In this work, mean pooling is select.

(4) Output layer. The readout phase calculates output for the entire graph using fully connected layers.

$$output = v_g W + b \quad \text{* MERGEFORMAT (10)}$$

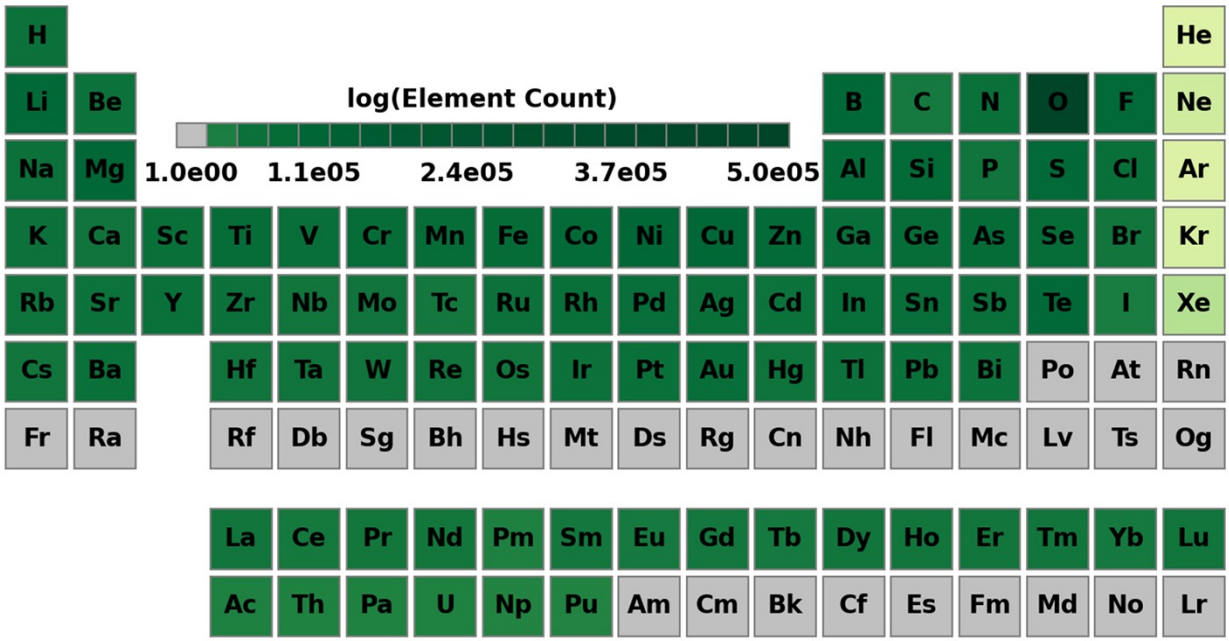

Fig. S1. Visualization of element prevalence in OQMD, shown as a heatmap on a periodic table.

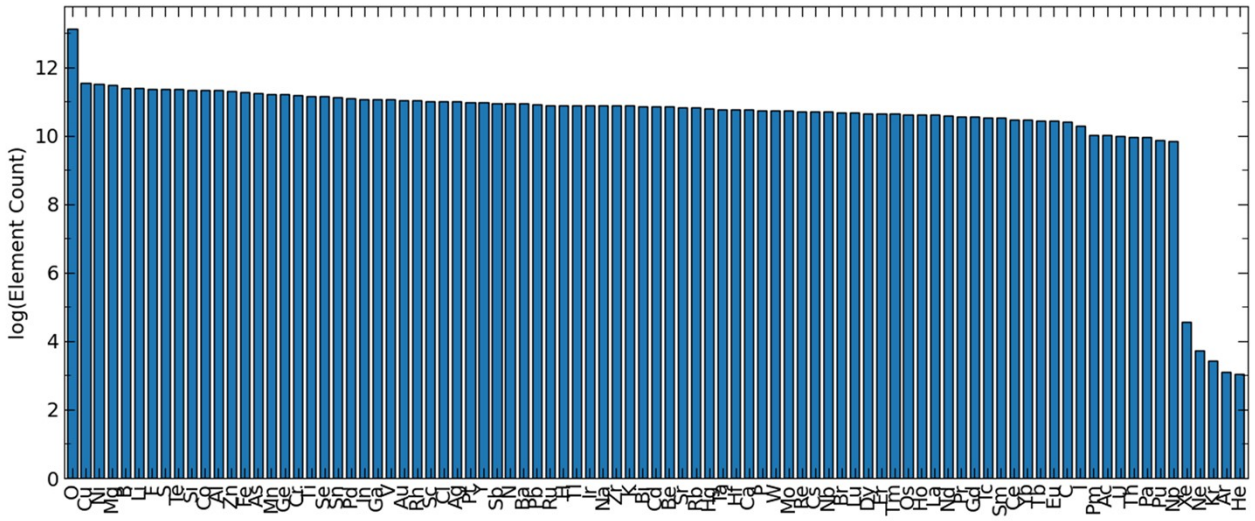

Fig. S2. Visualization of element prevalence in OQMD, shown as a histogram.

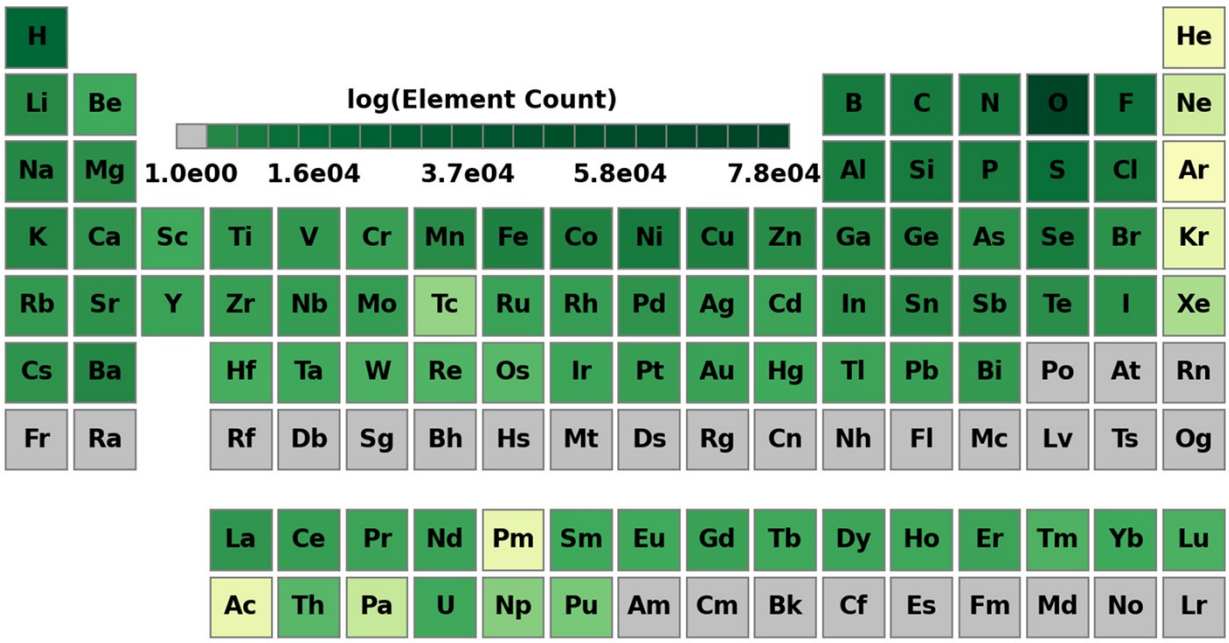

Fig. S3. Visualization of element prevalence in OQMD with ICSD label, shown as a heatmap on a periodic table.

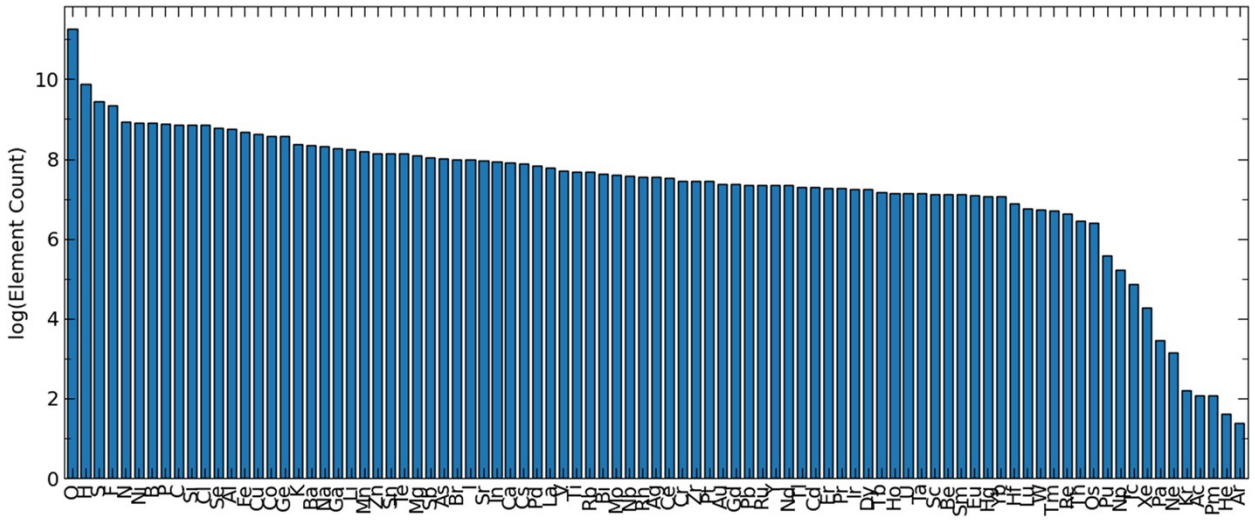

Fig. S4. Visualization of element prevalence in OQMD with ICSD label, shown as a histogram.

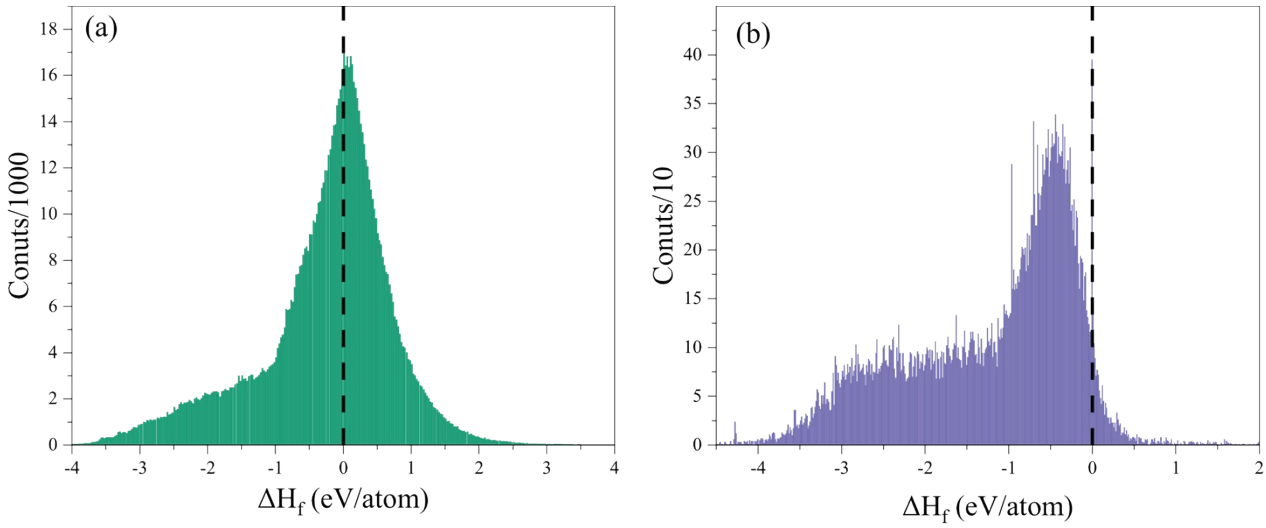

Fig. S5. Histogram of the distribution of the formation enthalpies for the entire dataset (a) and subset (b) with ICSD label.

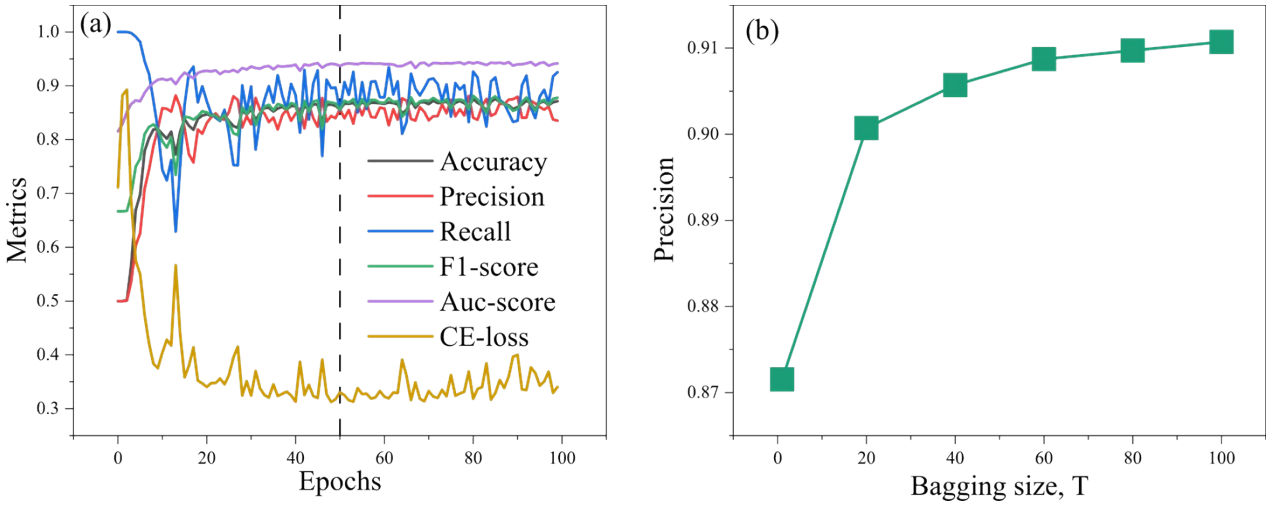

Fig. S6. The PUML performance as a function of training number (epochs) (a), and bagging size (T) (b).

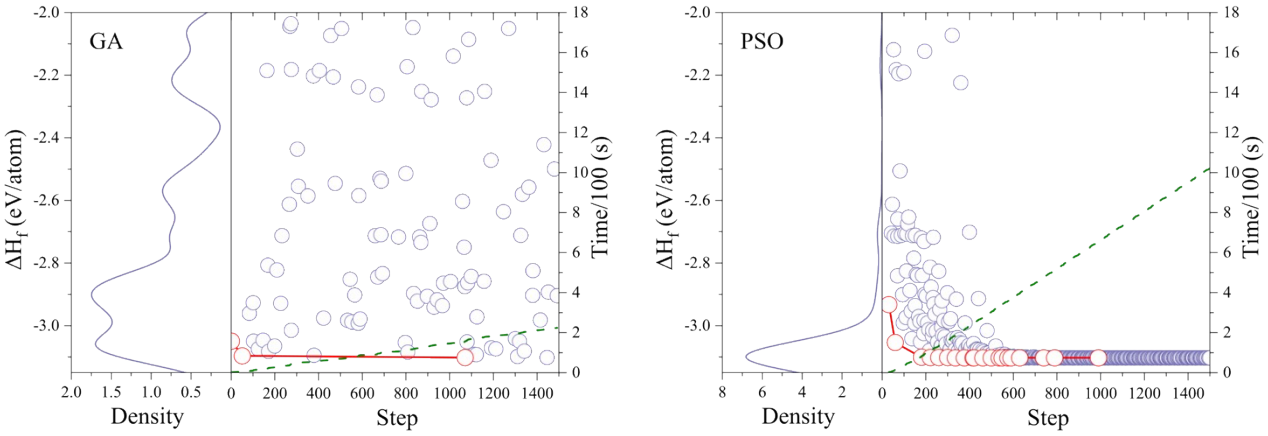

Fig. S7. The process and performance of DFT-ALIGNN-GA (P_m = 0.1) and DFT-ALIGNN-PSO (w = 0.1). The green dashed line represents the cumulative time required by each optimization algorithm, while the red solid line depicts the variation of the most stable structure over time. The left panel displays the density of states at the energy level, providing insights into the distribution of structures at different energy levels during the optimization process.

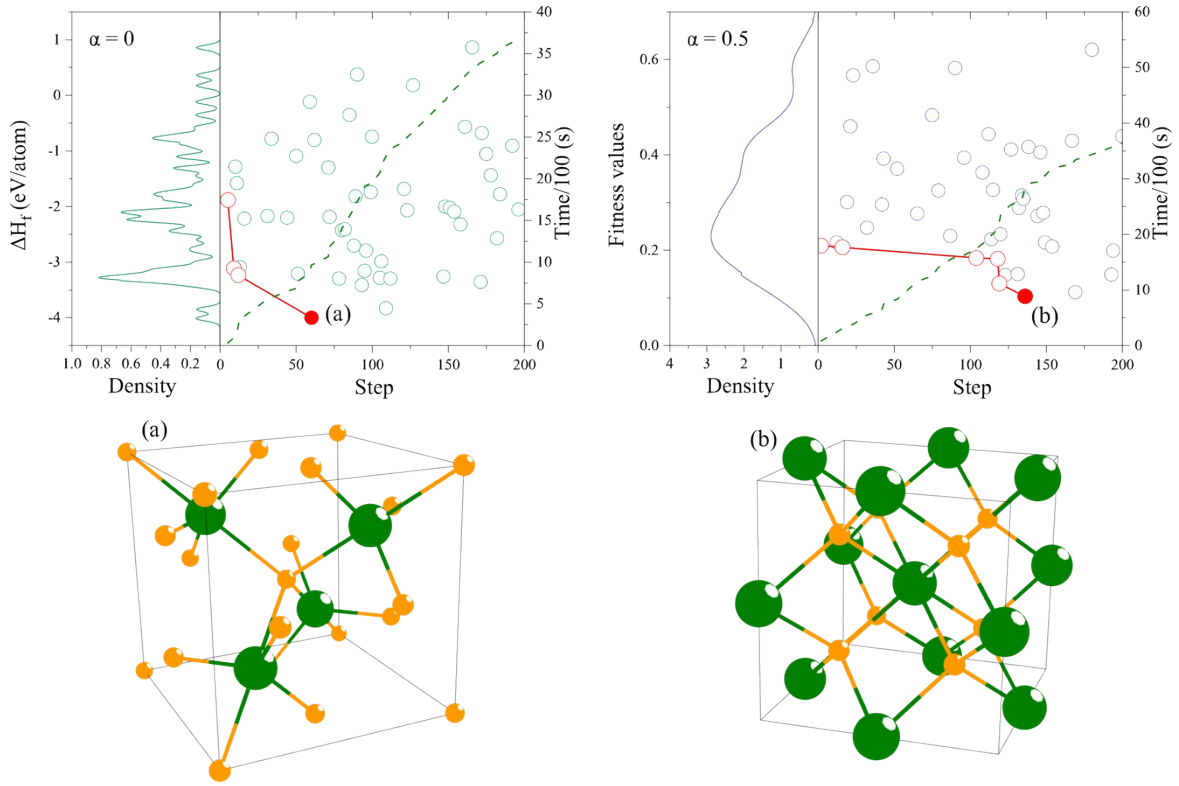

Fig. S8. The process and performance of ALIGNN-BO (α = 0) and TL-ALIGNN-BO (α = 0.5) for ThO_2. The green dashed line represents the cumulative time required by each optimization algorithm, while the red solid line depicts the variation of the most optimal structure over time. The left panel displays the density of state at the energy level, providing insights into the distribution of structures at different energy levels during the optimization process. Figure (a) displays the predicted ground-state structures obtained with α=0, while Figure (b) illustrates the optimal structures predicted with α=0.5, which have been experimentally confirmed.

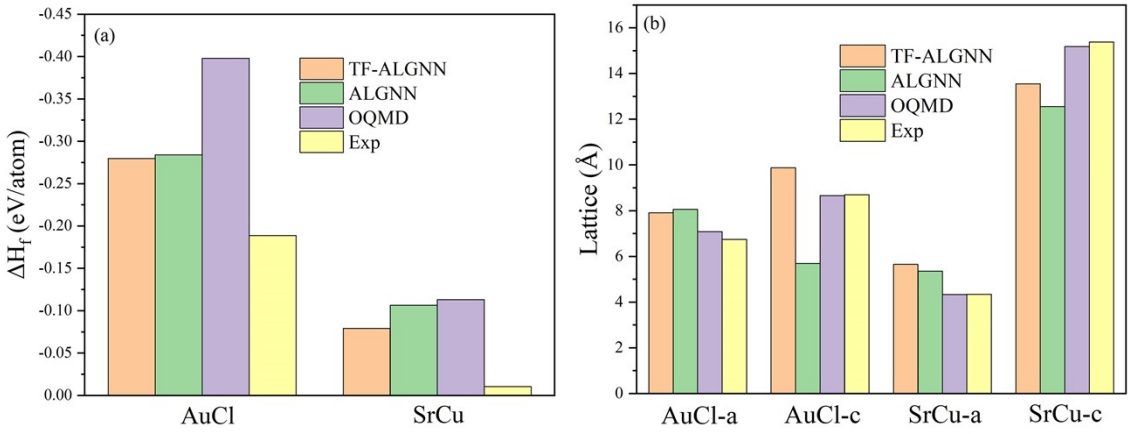

Fig. S9. Histogram of the formation enthalpies (a) and lattice constant (b) for AuCl and SrCu.

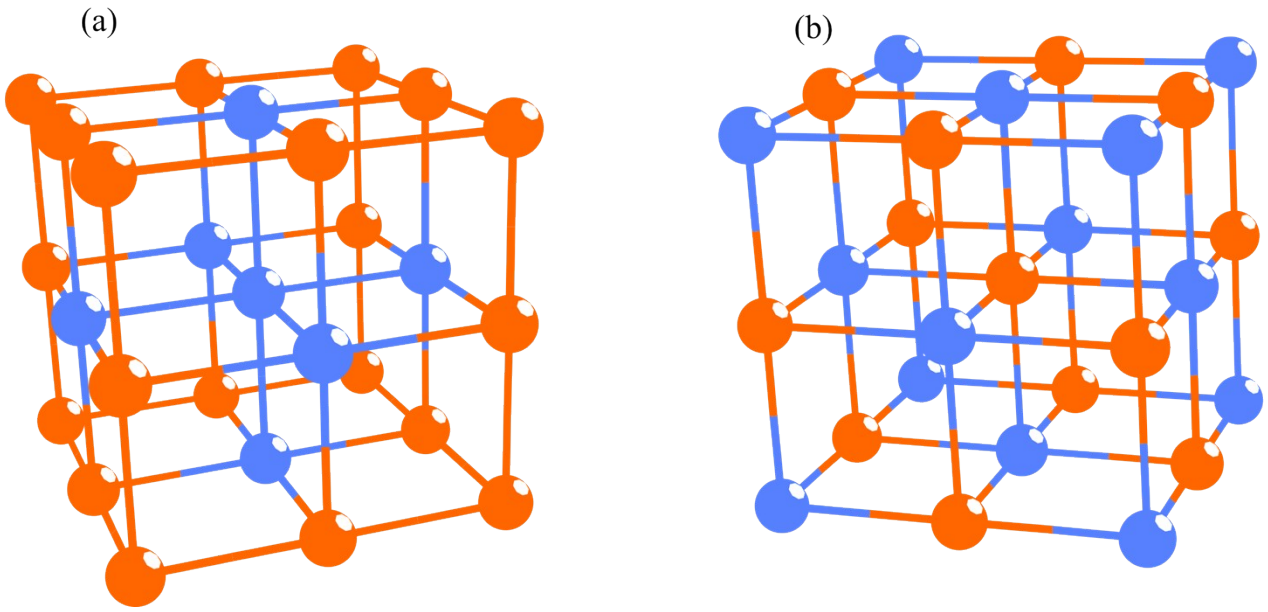

Fig. S10. Two highly similar structures. In comparison to the structure in Figure (a), the structure in Figure (b) is energetically more favorable.

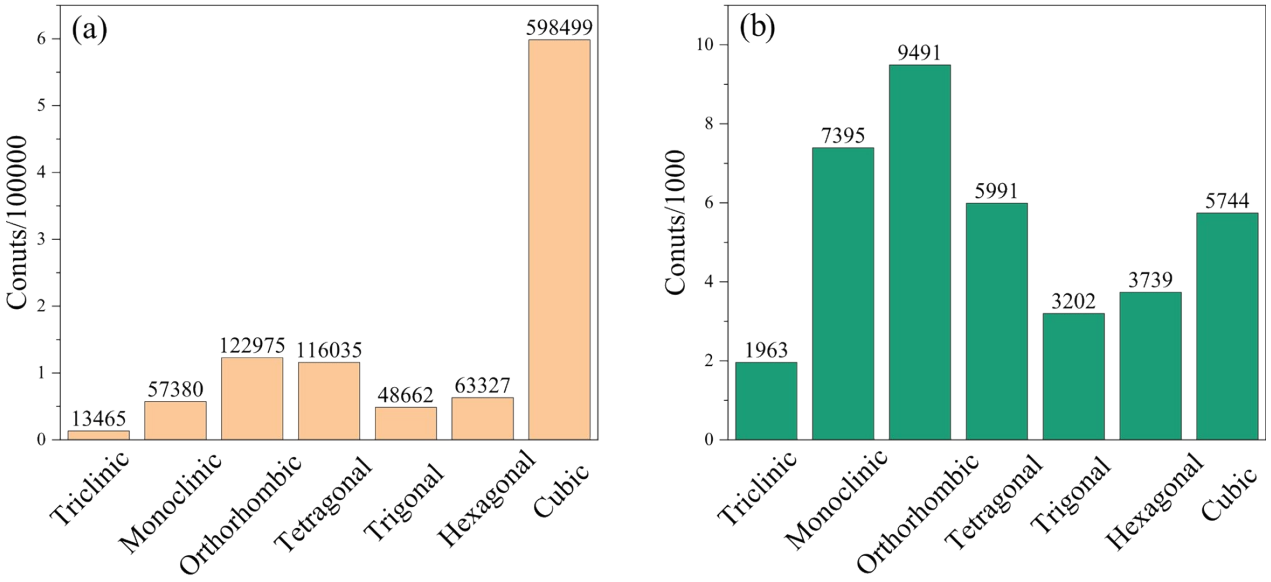

Fig. S11. Histogram of the ctystal systems for the entire dataset (a) and subset (b) with ICSD label.

Table. S1. The setup of model hyperparameters used in ALIGNN for the prediction of formation enthalpy and prediction of synthesis.

Parameters	Prediction of formation enthalpy	Prediction of synthesis
graph_max_radius	5	5
graph_max_neighbors	12	15
graph_edge_length	50	50
pre_fc_num	4	2
pre_out_channel	120	100
conv_num	4	2
post_fc_num	4	2
post_out_channel	60	50
pool	global_mean_pool	global_mean_pool
dropout_rate	0	0
batch_norm	True	True
aggr	mean	mean
act	relu	relu
epsilon	0.00001	0.00001

Table. S2. The setup of training hyperparameters used in ALIGNN for the prediction of formation enthalpy and prediction of synthesis.

Parameters	Prediction of formation enthalpy	Prediction of synthesis
train_ratio	0.8	0.8
val_ratio	0.1	0.1
loss	l1_loss	cross_entropy
epochs	200	50
lr	0.001	0.001
batch_size	4096	1024
optimizer	Adam	Adam
optimizer_args	{}	{}
scheduler	ReduceLROnPlateau	ReduceLROnPlateau
scheduler_args	{"mode":"min", "factor":0.8, "patience":10, "min_lr":0.00001, "threshold":0.0002}	{"mode":"min", "factor":0.8, "patience":10, "min_lr":0.00001, "threshold":0.0002}
Bagging_size	\	100