Electronic Supplementary Material (ESI) for Physical Chemistry Chemical Physics.

Substitution Engineering of Lead-Free Halide Perovskites for Photocatalytic Applications Assisted by Machine Learning

Tao Wang [a], Shuxin Fan [b], Hao Jin*, [a], Yunjin Yu [a], Yadong Wei *, [a]

[a] College of Physics and Optoelectronic Engineering, Shenzhen University, Shenzhen 518060, China

[b] College of Arts and Sciences, Beijing Normal University, Zhuhai 519087, China

E-mail: jh@szu.edu.cn; ywei@szu.edu.cn

Gradient boost regression. GBR is a flexible non-parametric statistical machine leaning algorithm.[1,2] The method contains a large number of decision trees that are generated sequentially. The construction of each decision tree requires the information of the previously generated decision tree. Therefore, each decision tree is based on a modified version of the original data set. The final regression algorithm is the weighted sum of these weak regression algorithms obtained by each training, as

$$F_M(x) = \sum_{m=1}^{M} T(x,\theta_m) \tag{1}$$

where m is the times of training, x is the input data, and θ_m is the distribution weight vector. The model is trained M times, and each time it produces a weak regression function T. The loss function of every weak classifier, is defined as

$$\hat{\theta} = arg_{\theta_m} min \sum_{i=1}^{N} L(y_i, F_{m-1}(x_i) + T(x_i,\theta_m)) \tag{2}$$

where $F_{m-1}(x_i)$ is the current model.

Generation of features. We generated 273 features using matminer software package, an open-source toolkit.[3] These features can be divided into the following several categories.[4,5] For details of these features, please refer to Ref 4 and 5.

Effective Coordination Number Features[5]

Effective coordination number of an atom is defined as:

$$CN_{eff} = \frac{(\sum_n A_n)^2}{\sum A_n^2} \tag{3}$$

where A_n is the area of face n in its Voronoi cell. The maximum, minimum, mean, and mean absolute deviation in coordination number are calculated as features.

Structural Heterogeneity Features[5]

These features reflect variation in the shape of local bonding environments. Bond length is defined as the Voronoi-face-area-weighted average of the distance between an atom and each neighbor:

$$\bar{l}_l = \frac{\sum A_n * \|\vec{r}_n - \vec{r}_l\|_2}{\sum A_n} \quad (4)$$

where $\vec{r}_l$ are the position vector of an atom i. A_n and $\vec{r}_n$ are the area position vector of n^{th} neighbor of atom i, respectively

The bond length variance is calculated for describing the distribution in bond lengths between each neighbor of an atom.

$$\hat{l}_l = \frac{\sum |A_n * \|\vec{r}_n - \vec{r}_l\|_2 - \bar{l}_l|}{\bar{l}_l * \sum A_n} \quad (5)$$

The maximum, minimum, mean, and mean absolute deviation of $\bar{l}_l$ and $\hat{l}_l$ are calculated as features.

The mean absolute deviation of the volume of the Voronoi cell about each atom is also used as a feature.

Chemical Ordering Features

These features are based on Warren-Cowley ordering parameters, which measure how the distribution of atoms on a lattice differs from purely-random.[6]

Maximum Packing Efficiency

The radius of the largest sphere centered on the position of the atom is equal to the distance between the center of the atom and the center of the nearest surface.

Local Environment Features[5]

These features are to describe difference in elemental properties between an atom and each neighbor. The local property difference for each atom is defined as:

$$\delta_p = \frac{\sum_n A_n * |p_n - p_i|}{\sum_n A_n} \quad (6)$$

where p_i and p_n are the elemental property of the central atom i and neighboring atom n, respectively. A_n is the area of face of atom n.

Composition-Based Features[4]

These features are only dependent on the composition of the atoms, which include: Stoichiometric Features are based on the relative fractions of elements in the structure, no matter what the elements are actually. Elemental Property Features are based on the mean, maximum, minimum, mode, range, and mean absolute deviation of 22 elemental properties. Valence Shell Features are based on the fraction of electrons in the s, p, d, and f shells of the constituent elements. Ionicity Features is about judging whether it can to form a charge-neutral ionic compound at a certain composition. Elemental properties used to compute elemental-property-based features is shown in Table S1. The selected 21 features for ML are shown in **Table S2**.

Table S1. (Reproduced from Ref.4) Elemental properties used to compute elemental-property-based attributes. s

Atomic Number	Mendeleev Number	Atomic Weight	Melting Temperature	Column
Row	Covalent Radius	Electronegativity	s Valence Electrons	p Valence Electrons
d Valence Electrons	f Valence Electrons	Total Valence Electrons	Unfilled s States†	Unfilled p States†
Unfilled d States	Unfilled f States	Total Unfilled States	Specific Volume of 0 K Ground State	Band Gap Energy of 0 K Ground State

Magnetic Moment (per atom) of 0 K ground state	Space Group Number of 0 K Ground State	

Table S2. 21 feature names

No	Feature	No	Feature
1	average deviation of column in the periodic table (local difference)	12	mean of number of s valence electrons (local difference)
2	compound possible	13	average deviation of s valence electrons
3	range of Mendeleev Number (local difference)	14	average deviation of melting temperature
4	average deviation of space group number (local difference)	15	range local of specific volume of ground state (local difference)
5	average deviation of Mendeleev Number (local difference)	16	mean of space group number
6	range of the number of unfilled electrons	17	average deviation of melting temperature (local difference)
7	range of the number of valence electrons (local difference)	18	minimum of Mendeleev Number (local difference)
8	minimum of the number of unfilled electrons	19	maximum of melting temperature
9	fraction of d valence electrons	20	maximum of the number of unfilled f electrons
10	maximum of the number of unfilled electrons (local difference)	21	minimum of column in the periodic table (local difference)
11	mean of atomic weight		

In order to analyze the hidden trends within the data, we visualize the relationship between the four important features and the bandgaps, as shown in **Fig. S1**. In **Fig. S1 (b)**, results show the bandgap of the materials that can be ionically bonded vary greatly. As seen from **Fig. S1 (a), (c)**, and **(d)**, no obvious trend is observed. It is worth noting that the three features are related to the local environment of the atoms, which means that different substituted sites have great influences on the bandgaps of the materials.

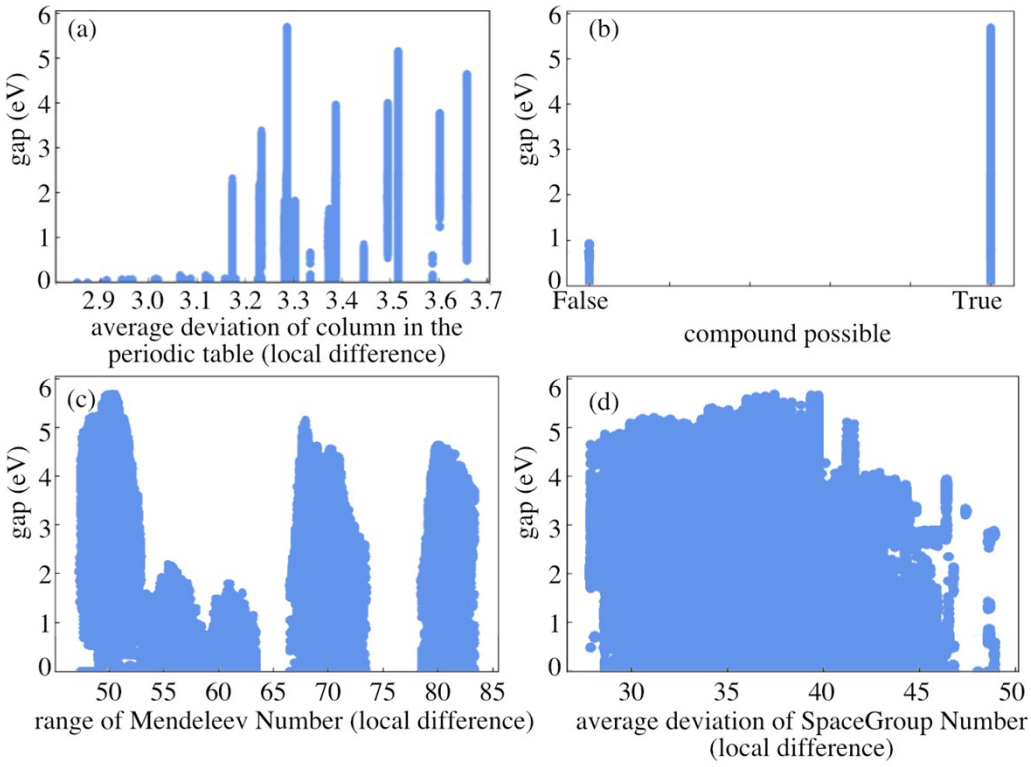

Figure S1. Data visualization of predicted bandgaps with (a) average deviation of the columns of the constituent atoms, (b) determining whether a material is ionically bonded, (c) range of the Mendeleev Numbers of the constituent atoms (d) average deviation of the space group number of the elementary substance formed by the constituent atoms.

We select a set of systems with the same concentration as an example and looked at the structures with the smallest (η_{STH}=13%), median (η_{STH}=14%), and largest (η_{STH}=15%) STH efficiency, as shown in **Fig. S2 (a)**, **(b)**, and **(c)** respectively. In this case, the value of η_{STH} is different due to different substituted sites.

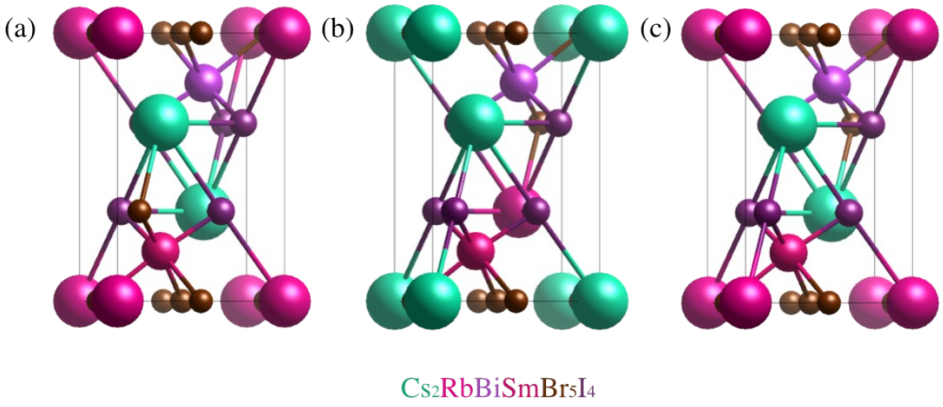

Figure S2 Schematic diagram of the configurations with (a) η_{STH} of 13%, (b) 14% and (c)15%

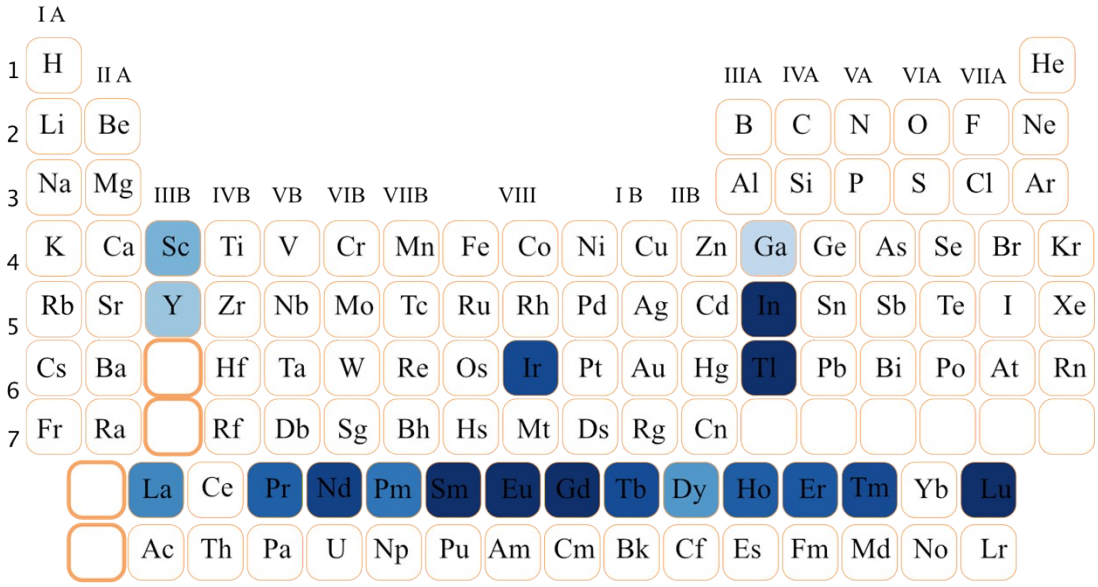

Figure R3 Periodic Table. These good B-site atoms are marked based on Fig.4(b).

References

(1) Friedman, J. H. Stochastic Gradient Boosting. *Comput. Stat. Data Anal.* **2002**, *38* (4), 367–378.

(2) Friedman, J. H. Greedy Function Approximation: A Gradient Boosting Machine. *Ann. Stat.* **2001**, *29* (5), 1189–1232.

(3) Ward, L.; Dunn, A.; Faghaninia, A.; Zimmermann, N. E. R.; Bajaj, S.; Wang, Q.; Montoya, J.; Chen, J.; Bystrom, K.; Dylla, M.; Chard, K.; Asta, M.; Persson, K. A.; Snyder, G. J.; Foster, I.; Jain, A. Matminer: An Open-Source Toolkit for Materials Data Mining. *Computational Materials Science* **2018**, *152*, 60–69.

(4) Ward, L.; Agrawal, A.; Choudhary, A.; Wolverton, C. A General-Purpose Machine Learning Framework for Predicting Properties of Inorganic Materials. *npj Comput Mater* **2016**, *2* (1), 16028.

(5) Ward, L.; Liu, R.; Krishna, A.; Hegde, V. I.; Agrawal, A.; Choudhary, A.; Wolverton, C. Including Crystal Structure Features in Machine Learning Models of Formation Energies via Voronoi Tessellations. *Phys. Rev. B* **2017**, *96* (2), 024104.

(6) Cowley, J. M. An Approximate Theory of Order in Alloys. *Physical Review* **1950**, *77* (5), 669–675.