Supplementary information: Δ-machine learning-driven discovery of double hybrid organic-inorganic perovskites

Jialu Chen,[1] Wenjun Xu[1], Ruiqin Zhang*[1,2]

[1]Department of Physics, City University of Hong Kong, Hong Kong SAR, People's Republic of China

[2]Beijing Computational Science Research Center, Beijing 100193, People's Republic of China

*Corresponding author.

Email address: aprqz@cityu.edu.hk

Mean absolute error (MAE), mean squared error (MSE), and coefficient of determination (R^2) were calculated to evaluate model accuracy. They are defined by Equations (S1), (S2), and (S4), correspondingly. $\bar{y}$ denotes the average value for a dataset with samplings marked y_1, …, y_N, and f_1, …, f_N denote predicted values.

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N}|y_i - f_i| \quad \text{(S1)}$$

$$\mathrm{MSE} = \frac{1}{N}\sum_{i=1}^{N}(y_i - f_i)^2 \quad \text{(S2)}$$

$$\bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i \quad \text{(S3)}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - f_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2} \quad \text{(S4)}$$

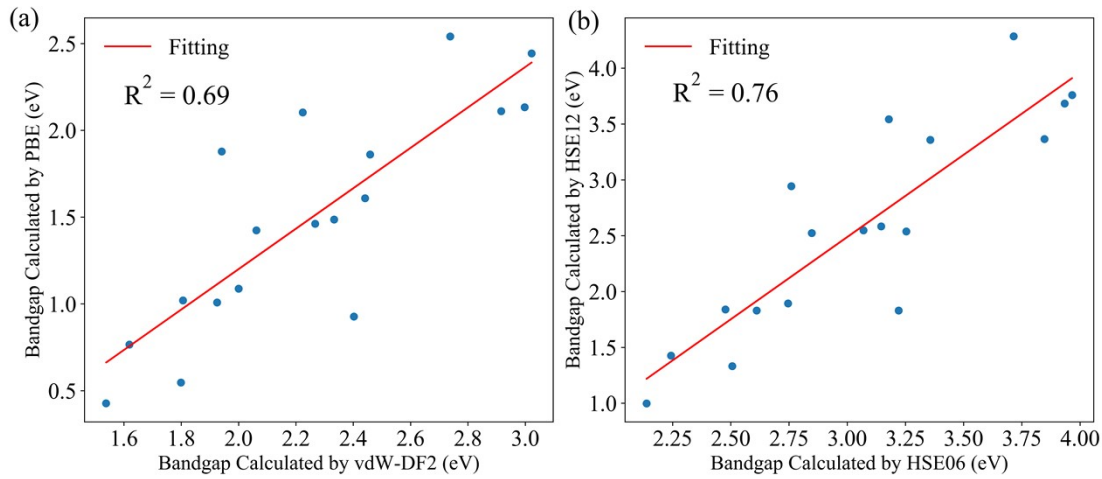

Fig. S1. Linear regression between bandgaps of vdW-DF2 and PBE (a), and between HSE06 and HSE12 (b). Bandgaps of PBE and HSE12 are from the first dataset, while vdW-DF2 and HSE06 are from the second dataset.

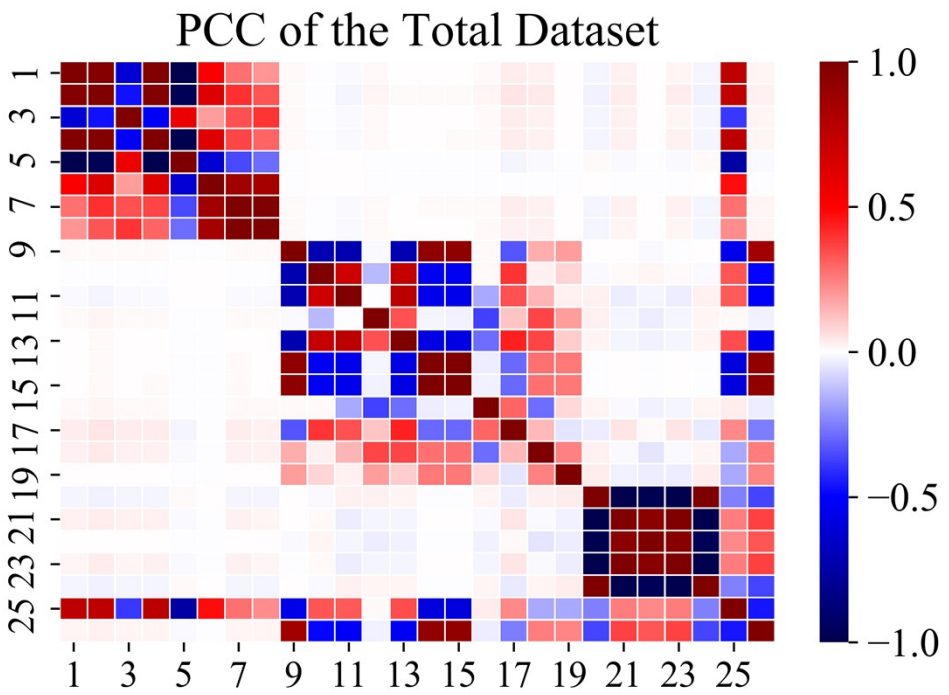

Fig. S2. The heat map of the Pearson correlation coefficient (PCC) matrix among 26 initial features of the total dataset.

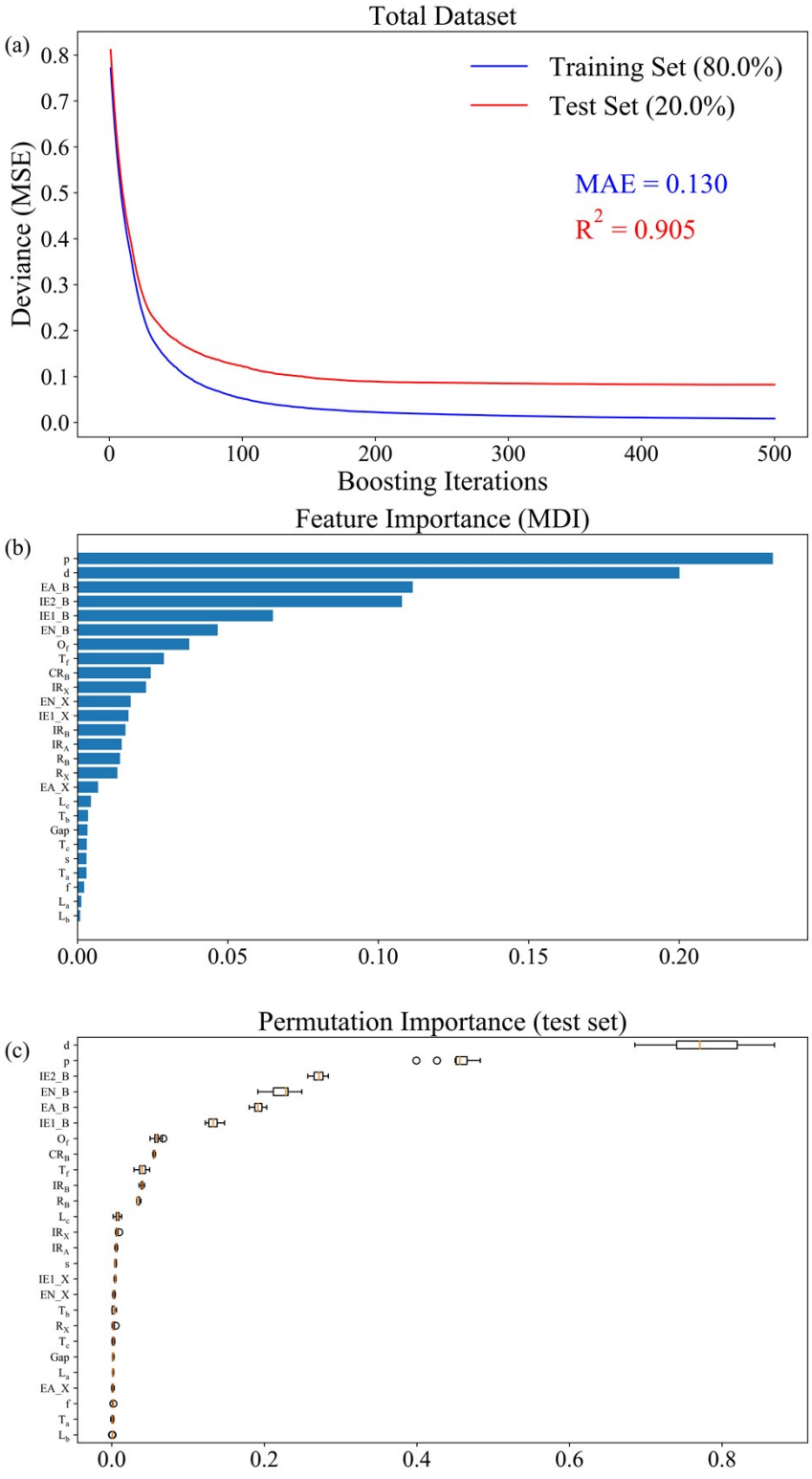

Fig. S3. The converge of model deviance, MDI, and permutation importance of the test set for the first model in Table S4.

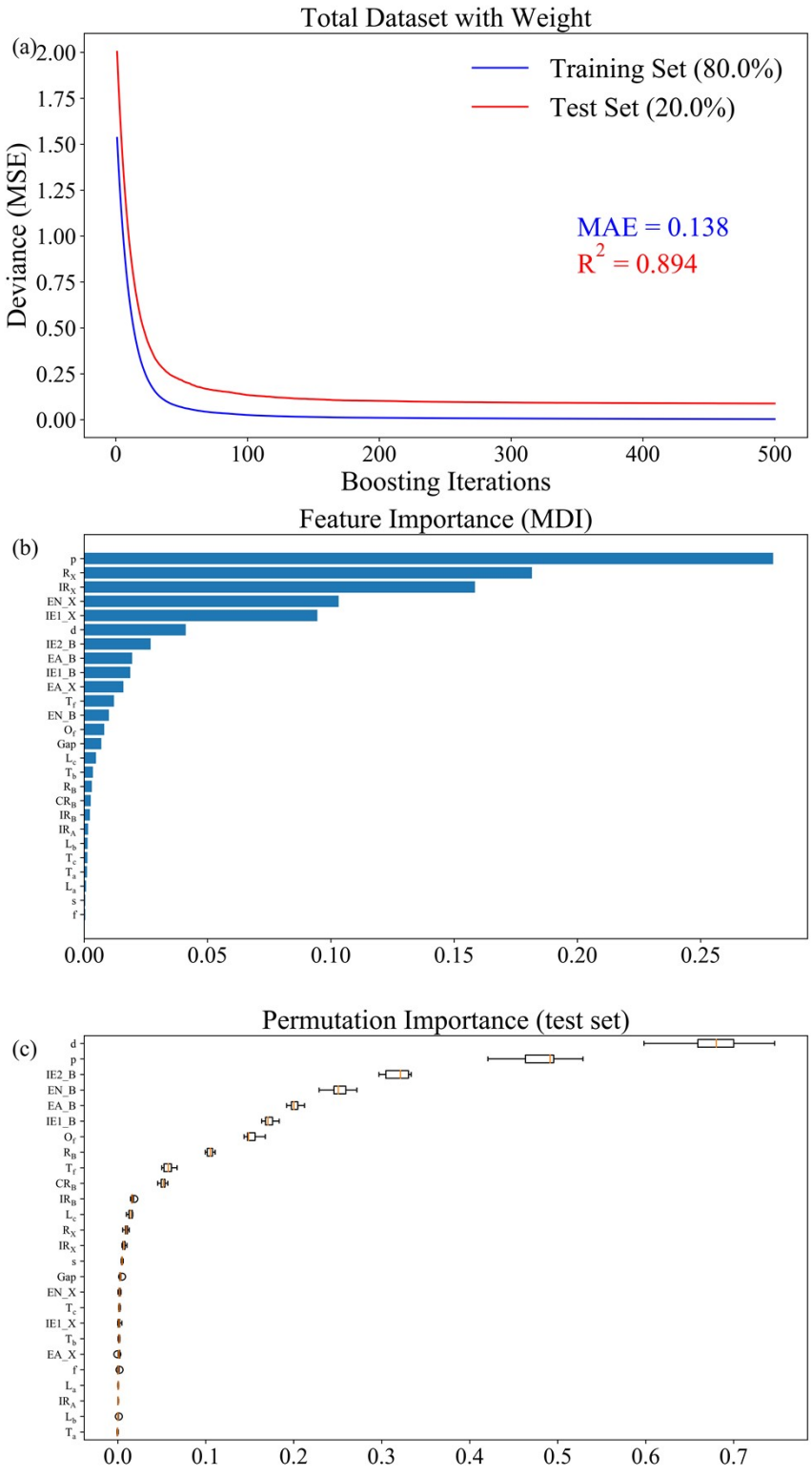

Fig. S4. The converge of model deviance, MDI, and permutation importance of the test set for the second model in Table S4, with a fitting weight of 100 for structures from the second dataset.

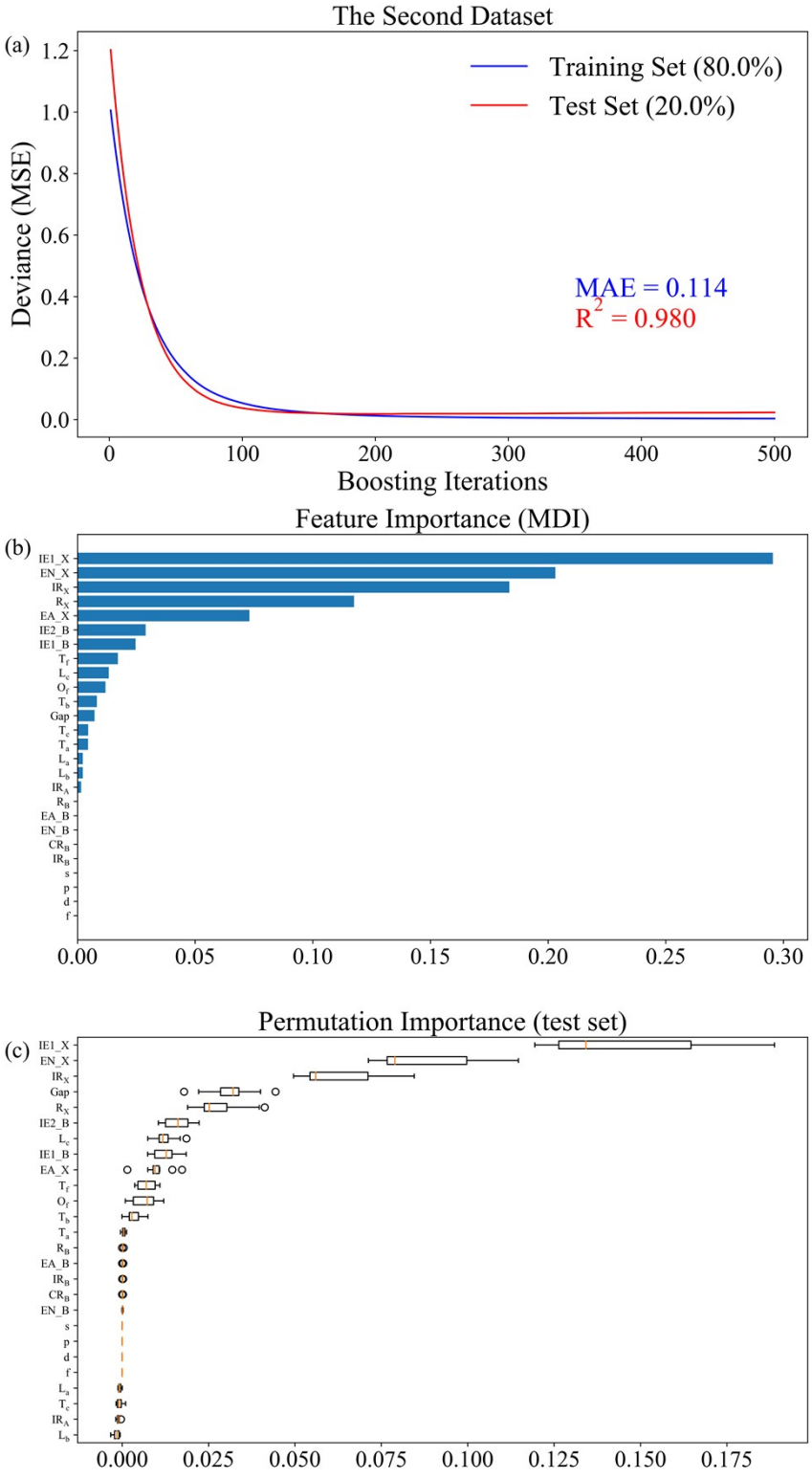

Fig. S5. The converge of model deviance, MDI, and permutation importance of test set for the third model in Table S4, by only fitting the second dataset.

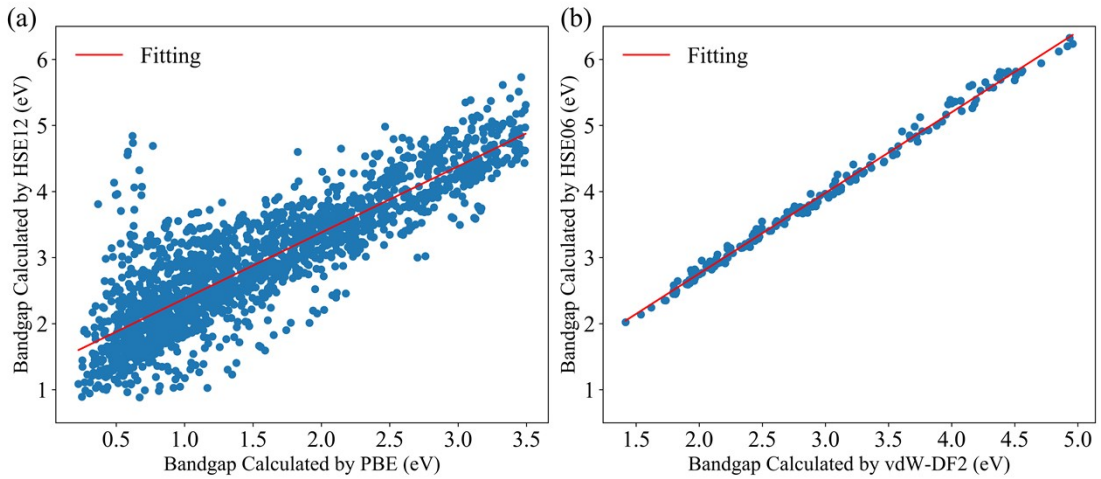

Fig. S6. Linear regression of bandgaps between PBE and HSE12 in the first dataset (a) and vdW-DF2 and HSE06 in the second dataset (b).

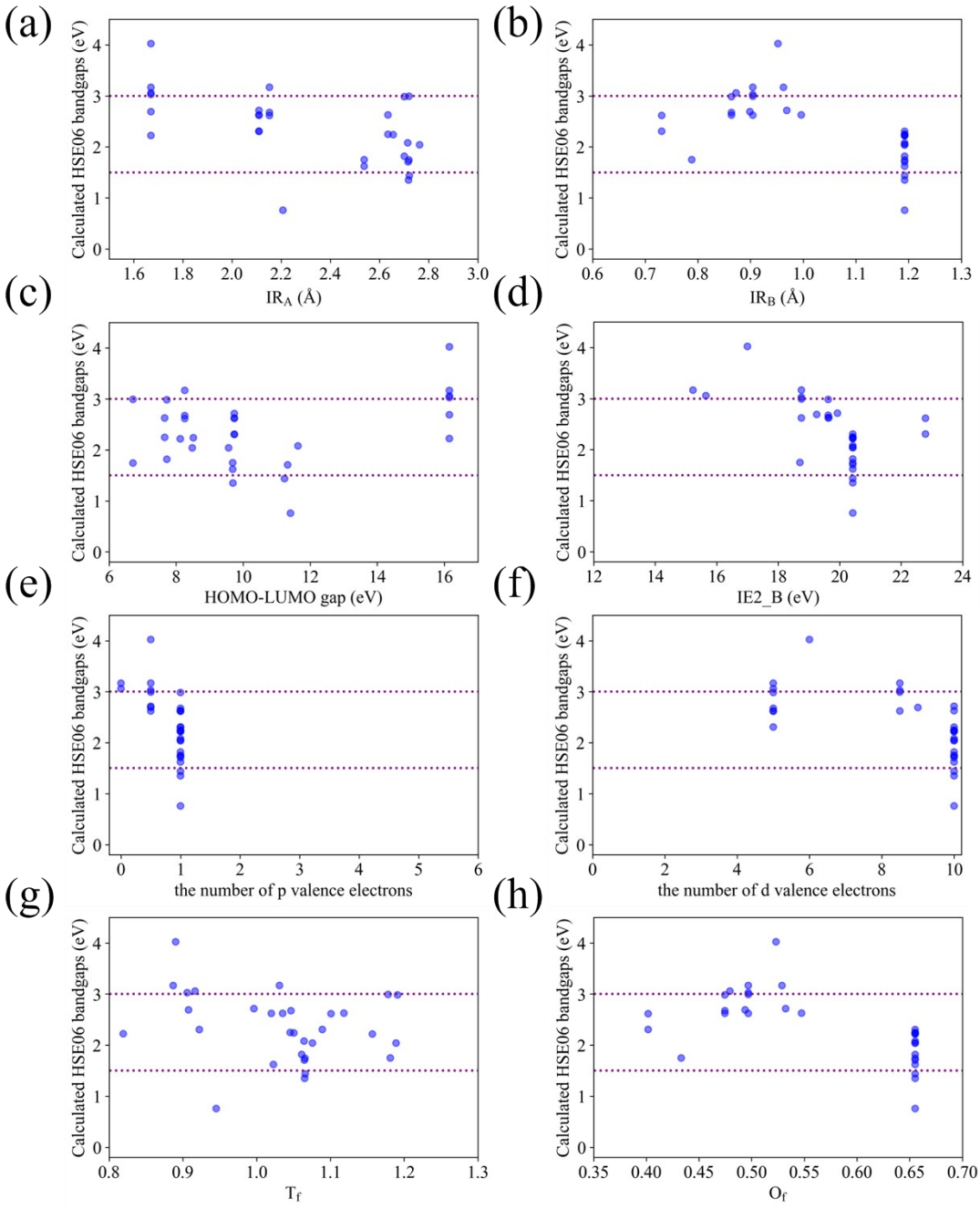

Fig. S7. Data visualization of HSE06 bandgaps of 24 screening DHOIPs with (a) effective radii of A-site cations, (b) average ionic radii of B-site atoms, (c) HOMO-LUMO gaps of A-site cations, (d) the second ionization energies of B-site atoms, (e) the number of p valence electrons, (f) the number of d valence electrons, (g) tolerance factor, and (h) octahedral factor.

Table S1. Comparison of calculated bandgaps (eV) of the same compositions in different datasets. The bandgaps of the second dataset are averaged by the same compositions.

Formula	The first dataset		The second dataset	
	PBE	HSE	vdw-DF2	HSE
$CH_3NH_3GeCl_3$	2.443	3.683	3.022	3.934
$CH_3NH_3GeBr_3$	1.486	2.583	2.333	3.146
$CH_3NH_3GeI_3$	1.008	1.830	1.925	2.611
$CH_3NH_3SnCl_3$	1.609	2.538	2.441	3.254
$CH_3NH_3SnBr_3$	1.088	1.894	2.000	2.746
$CH_3NH_3SnI_3$	0.766	1.426	1.619	2.243
$CH_3NH_3PbCl_3$	2.133	3.759	2.998	3.966
$CH_3NH_3PbBr_3$	1.861	3.358	2.459	3.357
$CH_3NH_3PbI_3$	1.424	2.524	2.062	2.848
$HC(NH_2)_2GeCl_3$	2.110	3.365	2.916	3.847
$HC(NH_2)_2GeBr_3$	1.462	2.548	2.267	3.070
$HC(NH_2)_2GeI_3$	1.020	1.840	1.806	2.477
$HC(NH_2)_2SnCl_3$	0.927	1.830	2.402	3.221
$HC(NH_2)_2SnBr_3$	0.547	1.332	1.799	2.506
$HC(NH_2)_2SnI_3$	0.427	0.998	1.538	2.138
$HC(NH_2)_2PbCl_3$	2.541	4.285	2.738	3.715
$HC(NH_2)_2PbBr_3$	2.103	3.543	2.224	3.179
$HC(NH_2)_2PbI_3$	1.878	2.943	1.941	2.760

Table S2. Properties of A-site cations including the identification N of organic cations (1~16) or atomic number (N > 16) in this work, Gap (eV), the HOMO-LUMO gap, three lengths L_a, L_b and L_c, and radius (Å), and rotational temperatures (K) in three directions T_a, T_b, and T_c.

N	Formula	L_a	L_b	L_c	Radius	Gap	T_a	T_b	T_c
1	NH_4^+	3.943	3.940	3.687	2.037	16.152	8.586	8.586	8.586
2	$CH_3NH_3^+$	4.628	4.385	4.187	2.692	11.413	4.067	0.959	0.959
3	$(CH_3)_2NH_2^+$	6.637	4.457	4.184	3.319	11.235	1.447	0.415	0.367
4	$(CH_3)_3NH^+$	6.592	6.298	4.953	3.313	11.329	0.388	0.388	0.225
5	$(CH_3)_4N^+$	6.550	6.337	6.099	3.311	11.636	0.223	0.223	0.223
6	$CH_3CH_2NH_3^+$	6.630	4.545	4.180	3.315	9.697	1.415	0.399	0.353
7	$CH_3CH_2CH_2NH_3^+$	7.784	4.726	4.172	3.959	8.483	1.127	0.170	0.160
8	$(CH_3)_2CHNH_3^+$	6.650	6.406	5.073	3.370	9.567	0.369	0.369	0.213
9	$CH_3CH_2CH_2CH_2NH_3^+$	7.581	5.605	5.356	3.792	8.125	0.359	0.141	0.127
10	$OHNH_3^+$	4.511	4.211	4.091	2.573	9.738	6.687	1.178	1.137
11	$HC(NH_2)_2^+$	6.425	5.093	2.900	3.213	7.659	2.877	0.493	0.421
12	$CH_3C(NH_2)_2^+$	6.428	6.014	4.167	3.294	7.724	0.474	0.432	0.233
13	$NH_2NH_3^+$	4.593	4.060	3.851	2.625	8.259	5.321	1.092	1.048
14	$C(NH_2)_3^+$	6.359	6.315	2.400	3.239	8.515	0.486	0.486	0.243
15	$C_3H_8N^+$	6.111	6.103	4.550	3.095	9.689	0.528	0.496	0.297
16	$C_3H_5N_2^+$	6.451	6.403	2.400	3.317	6.714	0.449	0.439	0.222
55	Cs^+	3.620	3.620	3.620	1.810	14.566	0.000	0.000	0.000

Table S3. Effective radii (Å) by Kieslich et al., initial radius (Å) calculated by Mulitwfn, and scaled ionic radii (Å) with a factor of 0.82 for organic cations.

N	Formula	Kieslich et al.	Initial Radius	Scaled Radius
1	NH_4^+	1.46	2.037	1.670
2	$CH_3NH_3^+$	2.17	2.692	2.207
3	$(CH_3)_2NH_2^+$	2.72	3.319	2.722
4	$(CH_3)_3NH^+$		3.313	2.717
5	$(CH_3)_4N^+$	2.92	3.311	2.715
6	$CH_3CH_2NH_3^+$		3.315	2.718
7	$CH_3CH_2CH_2NH_3^+$		3.959	3.246
8	$(CH_3)_2CHNH_3^+$		3.370	2.763
9	$CH_3CH_2CH_2CH_2NH_3^+$		3.792	3.109
10	$OHNH_3^+$	2.16	2.573	2.110
11	$HC(NH_2)_2^+$	2.53	3.213	2.635
12	$CH_3C(NH_2)_2^+$		3.294	2.701
13	$NH_2NH_3^+$	2.17	2.625	2.153
14	$C(NH_2)_3^+$	2.78	3.239	2.656
15	$C_3H_8N^+$	2.50	3.095	2.538
16	$C_3H_5N_2^+$	2.58	3.317	2.720

Table S4. The data and hyperparameters of GBR models during the feature selection process.

	1	2	3
Dataset	Total	Total	Second
Weight	No	Yes	No
Boosting stages	500	500	500
Learning rate	0.05	0.05	0.02
Maximum depth of nodes	8	8	3
Minimum number to split a node	3	3	8
Training set size	80%	80%	80%

Table S5. The data and hyperparameters of GBR models after the feature selection process.

	Model1	Model2	Model3	Model4	Model5
Dataset	Second	Second	Total	Total	Total
Weight	No	No	Yes	Yes	Yes
Boosting stages	500	500	500	500	500
Learning rate	0.02	0.02	0.05	0.03	0.03
Maximum depth of nodes	3	3	8	7	7
Minimum number to split a node	8	8	3	4	4
Training set size	20%	20%	80%	80%	80%

Table S6. Common valences of cations in B-site in the range of +1 to +3.

Atomic number	Chemical symbol	Valences
4	Be	2
5	B	3
6	C	
7	N	3
12	Mg	2
13	Al	3
14	Si	
15	P	3
20	Ca	2
21	Sc	3
22	Ti	2 3
23	V	2 3
24	Cr	2 3
25	Mn	2 3
26	Fe	2 3
27	Co	2 3
28	Ni	2 3
29	Cu	1 2 3
30	Zn	2
31	Ga	3
32	Ge	2
33	As	3
38	Sr	2
39	Y	3
40	Zr	
41	Nb	3
42	Mo	3
43	Tc	
44	Ru	3
45	Rh	3
46	Pd	1 2 3
47	Ag	1 2 3
48	Cd	2
49	In	3
50	Sn	
51	Sb	3
56	Ba	2
72	Hf	
73	Ta	3
74	W	
75	Re	
76	Os	
77	Ir	3
78	Pt	2
79	Au	1 3
80	Hg	1 2
81	Tl	1 3
82	Pb	2
83	Bi	3

Table S7. Calculated bandgaps of screening perovskites by our machine learning models. The bandgaps are not available for structures with failed convergence.

Identification	A	B	B'	EPBE	EHSE06
1	NH_4^+	B	Tl	3.77	/
2	NH_4^+	Mg	Ag	0.07	/
3	NH_4^+	Al	Tl	3.89	/
4	NH_4^+	Ca	Ge	3.24	/
5	NH_4^+	Ca	Pd	1.38	3.06
6	NH_4^+	Sc	Tl	3.0	/
7	NH_4^+	Ti	Tl	/	/
8	NH_4^+	V	Tl	0.0	/
9	NH_4^+	Mn	Tl	0.0	/
10	NH_4^+	Fe	Tl	0.0	/
11	NH_4^+	Co	Tl	0.92	3.03
12	NH_4^+	Ni	Tl	0.0	/
13	NH_4^+	Ga	Tl	2.86	/
14	NH_4^+	Ge	Sr	3.25	/
15	NH_4^+	Ge	Ba	3.35	/
16	NH_4^+	Sr	Pd	1.56	3.17
17	NH_4^+	Nb	Tl	0.0	/
18	NH_4^+	Mo	Tl	0.0	/
19	NH_4^+	Ru	Tl	0.0	/
20	NH_4^+	Rh	Tl	1.15	2.69
21	NH_4^+	Pd	Tl	0.0	/
22	NH_4^+	In	Tl	2.35	/
23	NH_4^+	Ta	Tl	0.0	/
24	NH_4^+	Tl	Tl	1.4	2.22
25	$CH_3NH_3^+$	B	Tl	2.33	/
26	$CH_3NH_3^+$	Al	Tl	3.71	/
27	$CH_3NH_3^+$	Ca	Ge	3.13	/
28	$CH_3NH_3^+$	Co	Tl	0.83	3.03
29	$CH_3NH_3^+$	Ga	Tl	2.96	/
30	$CH_3NH_3^+$	Ge	Sr	3.47	/
31	$CH_3NH_3^+$	Ge	Ba	3.48	/
32	$CH_3NH_3^+$	In	Tl	2.43	/
33	$CH_3NH_3^+$	Tl	Tl	0.23	0.76
34	$(CH_3)_2NH_2^+$	Al	Tl	3.17	/
35	$(CH_3)_2NH_2^+$	Ga	Tl	2.94	/
36	$(CH_3)_2NH_2^+$	Ge	Sr	3.47	/
37	$(CH_3)_2NH_2^+$	Ge	Ba	3.36	/
38	$(CH_3)_2NH_2^+$	In	Tl	2.43	/
39	$(CH_3)_2NH_2^+$	Tl	Tl	0.77	1.44
40	$(CH_3)_3NH^+$	Al	Tl	3.63	/
41	$(CH_3)_3NH^+$	Ga	Tl	2.63	/
42	$(CH_3)_3NH^+$	Ge	Sr	3.14	/
43	$(CH_3)_3NH^+$	Ge	Ba	3.27	/
44	$(CH_3)_3NH^+$	In	Tl	2.44	/
45	$(CH_3)_3NH^+$	Tl	Tl	0.94	1.71

46	$(CH_3)_4N^+$	Al	Tl	3.13	/
47	$(CH_3)_4N^+$	Ga	Tl	2.54	/
48	$(CH_3)_4N^+$	Ge	Sr	3.17	/
49	$(CH_3)_4N^+$	Ge	Ba	3.01	/
50	$(CH_3)_4N^+$	In	Tl	2.0	/
51	$(CH_3)_4N^+$	Tl	Tl	1.17	2.08
52	$CH_3CH_2NH_3^+$	Al	Tl	3.67	/
53	$CH_3CH_2NH_3^+$	Ga	Tl	2.88	/
54	$CH_3CH_2NH_3^+$	Ge	Sr	3.36	/
55	$CH_3CH_2NH_3^+$	Ge	Ba	3.42	/
56	$CH_3CH_2NH_3^+$	In	Tl	2.34	/
57	$CH_3CH_2NH_3^+$	Tl	Tl	/	/
58	$CH_3CH_2NH_3^+$	Tl	Tl	1.2	2.04
59	$(CH_3)_2CHNH_3^+$	Ga	Tl	2.96	/
60	$(CH_3)_2CHNH_3^+$	Ge	Sr	3.47	/
61	$(CH_3)_2CHNH_3^+$	Ge	Ba	3.43	/
62	$(CH_3)_2CHNH_3^+$	In	Tl	2.51	/
63	$(CH_3)_2CHNH_3^+$	Tl	Tl	1.19	2.04
64	$CH_3CH_2CH_2CH_2NH_3^+$	Tl	Tl	1.33	2.22
65	$OHNH_3^+$	B	Tl	/	/
66	$OHNH_3^+$	Al	Tl	/	/
67	$OHNH_3^+$	Ca	Ge	3.36	/
68	$OHNH_3^+$	Co	Tl	/	/
69	$OHNH_3^+$	Ga	Tl	3.24	/
70	$OHNH_3^+$	Ge	Sr	3.57	/
71	$OHNH_3^+$	Ge	Ba	3.25	/
72	$OHNH_3^+$	Pd	Tl	/	/
73	$OHNH_3^+$	In	Tl	2.54	/
74	$OHNH_3^+$	Tl	Tl	1.36	2.3
75	$HC(NH_2)_2^+$	Al	Tl	3.61	/
76	$HC(NH_2)_2^+$	Ca	Ge	3.21	/
77	$HC(NH_2)_2^+$	Ga	Tl	2.91	/
78	$HC(NH_2)_2^+$	Ge	Sr	3.35	/
79	$HC(NH_2)_2^+$	Ge	Ba	3.11	/
80	$HC(NH_2)_2^+$	In	Tl	1.72	2.63
81	$HC(NH_2)_2^+$	Tl	Tl	1.41	2.25
82	$CH_3C(NH_2)_2^+$	Al	Tl	/	/
83	$CH_3C(NH_2)_2^+$	Ga	Tl	2.99	/
84	$CH_3C(NH_2)_2^+$	Ge	Sr	3.54	/
85	$CH_3C(NH_2)_2^+$	Ge	Ba	3.48	/
86	$CH_3C(NH_2)_2^+$	In	Tl	2.46	/
87	$CH_3C(NH_2)_2^+$	Tl	Tl	1.07	1.82
88	$NH_2NH_3^+$	B	Tl	/	/
89	$NH_2NH_3^+$	Al	Tl	1.32	2.67
90	$NH_2NH_3^+$	Ca	Ge	3.22	/
91	$NH_2NH_3^+$	Co	Tl	0.95	3.17
92	$NH_2NH_3^+$	Ga	Tl	2.85	/
93	$NH_2NH_3^+$	Ge	Sr	2.99	/
94	$NH_2NH_3^+$	Ge	Ba	3.17	/

95	$NH_2NH_3^+$	In	Tl	2.47	/
96	$NH_2NH_3^+$	Tl	Tl	2.47	/
97	$C(NH_2)_3^+$	Al	Tl	2.82	/
98	$C(NH_2)_3^+$	Ga	Tl	2.92	/
99	$C(NH_2)_3^+$	Ge	Sr	3.43	/
100	$C(NH_2)_3^+$	Ge	Ba	3.54	/
101	$C(NH_2)_3^+$	In	Tl	2.54	/
102	$C(NH_2)_3^+$	Tl	Tl	1.4	2.24
103	$C_3H_8N^+$	Al	Tl	3.37	/
104	$C_3H_8N^+$	Ca	Ge	3.2	/
105	$C_3H_8N^+$	Ga	Tl	2.49	/
106	$C_3H_8N^+$	Ge	Sr	3.39	/
107	$C_3H_8N^+$	Ge	Ag	0.86	1.75
108	$C_3H_8N^+$	Ge	Ba	3.25	/
109	$C_3H_8N^+$	In	Tl	2.02	/
110	$C_3H_8N^+$	Tl	Tl	0.88	1.62
111	$C_3H_5N_2^+$	Al	Tl	2.78	/
112	$C_3H_5N_2^+$	Co	Tl	0.86	2.99
113	$C_3H_5N_2^+$	Ga	Tl	2.9	/
114	$C_3H_5N_2^+$	Ge	Sr	3.19	/
115	$C_3H_5N_2^+$	Ge	Ba	3.2	/
116	$C_3H_5N_2^+$	In	Tl	2.17	/
117	$C_3H_5N_2^+$	Tl	Tl	0.9	1.74

Table S8. The results of 5-fold Cross-validation for GBR models during the feature selection. The scores were evaluated by the Scikit-learn package.

	1	2	3
Score of first split	0.897	0.914	0.951
Score of second split	0.899	0.912	0.960
Score of third split	0.894	0.913	0.963
Score of forth split	0.902	0.903	0.979
Score of fifth split	0.907	0.874	0.979
Mean score	0.900	0.903	0.967
Standard Deviation	0.004	0.015	0.011

Table S9. The results of 5-fold Cross-validation for GBR models after choosing 24 features. The scores were evaluated by the Scikit-learn package.

	Model1	Model2	Model3	Model4	Model5
Score of first split	0.965	0.963	0.903	0.793	0.712
Score of second split	0.975	0.955	0.914	0.790	0.833
Score of third split	0.985	0.945	0.903	0.730	0.750
Score of forth split	0.945	0.972	0.890	0.764	0.720
Score of fifth split	0.983	0.972	0.909	0.755	0.747
Mean score	0.970	0.961	0.904	0.766	0.752
Standard Deviation	0.014	0.010	0.008	0.024	0.043

Table S10. Comparison of PBE bandgaps calculated by CPU and GPU versions of VASP.

Identification	A	B	B'	CPU	GPU
1	NH_4^+	B	Tl	3.77	3.77
50	$(CH_3)_4N^+$	In	Tl	2.00	2.00
100	$C(NH_2)_3^+$	Ge	Ba	3.54	3.54