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

Highly versatile and accurate machine learning methods for predicting perovskite properties

Ziming Chen [a], Jing Wang [b*], Canjie Li [a], Baiquan Liu [c], Dongxiang Luo [d*], Yonggang Min [b], Nianqing Fu [a*], Qifan Xue [a*]

[a] *State Key Laboratory of Luminescent Materials and Devices, Institute of Polymer Optoelectronic Materials and Devices, School of Materials Science and Engineering, South China University of Technology, Guangzhou 510640, P. R. China*
[b] *School of Materials and Energy, Guangdong University of Technology, Guangzhou 510006, China*
[c] *School of Electronics and Information Technology, Sun Yat-sen University, Guangzhou, 510275, China*
[d] Huangpu Hydrogen Innovation Center/Guangzhou Key Laboratory for Clean Energy and Materials, School of Chemistry and Chemical Engineering, Guangzhou University, Guangzhou, 510006, China

[*]Corresponding author.
Email: jingwang777@gdut.edu.cn (J. Wang); luodx@gzhu.edu.cn (D. Luo); msnqfu@scut.edu.cn (N. Fu); qfxue@scut.edu.cn （Q. Xue）.

Tabel S1. feature explanation

Feature name	explain
MagpieData avg_dev Column	Mean absolute deviation of material properties
Transition metal fraction	transition metal atomic fraction
MagpieData maximum NdUnfilled	The number of unfilled d orbital electrons of the element with the largest number of unfilled f orbital electrons
MagpieData mean NfUnfilled	Average number of electrons in unfilled f orbitals
MagpieData mean NpValence	Average number of valence electrons
MagpieData maximum Number	maximum atomic number
Minimum EN difference	The smallest electronegativity difference between elements in a compound
Volume	unit cell volume
MagpieData avg_dev NUnfilled	The average number of unfilled d and f electrons of the elements in the material

More feature details can be found on the official website of matminer.
https://hackingmaterials.lbl.gov/matminer/featurizer_summary.html

Table S2. Formation energy model training results for single and double perovskites.

□	ABX_3		$A_2B(I)B(II)X_6$	
□	R2	RMSE	R2	RMSE
Ridge	0.8838	0.3259	0.9452	0.1837
DT	0.8730	0.2947	0.9410	0.1837
RFR	0.9043	0.2968	0.9437	0.1851
XGB	0.9336	0.2546	0.9717	0.1317
SVM	0.9147	0.2784	0.9499	0.1746
MLPR	0.9149	0.2788	0.9471	0.1793

Table S3. Merge Data Formation energy Training Results.

□	Startegy1		Startegy 2	
□	R2	RMSE	R2	RMSE
Ridge	0.9172	0.2592	0.9174	0.2591
DT	0.9012	0.2508	0.9021	0.2490
RFR	0.9241	0.2466	0.9243	0.2460
XGB	0.9438	0.2073	0.9435	0.2118
SVM	0.9294	0.2373	0.9270	0.2420
MLPR	0.9319	0.2311	0.9311	0.2324

Table S4. Bandgap model training results for single and double perovskites.

□	ABX_3		$A_2B(I)B(II)X_6$	
□	R2	RMSE	R2	RMSE
Ridge	0.6070	1.0393	0.6110	0.9930
DT	0.7075	0.8184	0.7824	0.6280
RFR	0.7640	0.8153	0.8422	0.6315
XGB	0.7995	0.7419	0.8688	0.5742
SVM	0.7321	0.8617	0.8349	0.6900
MLPR	0.7626	0.8121	0.8588	0.6219

Table S5. Merge Data Bandgap Training Results.

□	Startegy 1		Startegy 2	
□	R2	RMSE	R2	RMSE
Ridge	0.5519	1.0953	0.5589	1.0867
DT	0.7229	0.7459	0.7398	0.7348
RFR	0.7960	0.7433	0.7890	0.7397
XGB	0.8310	0.6817	0.8311	0.6820
SVM	0.7431	0.7962	0.7724	0.7851

MLPR	0.7995	0.7368	0.7987	0.7376

Table S6. Bandgap model training results after merged data processed by feature engineering.

□	Startegy 1		Startegy 2	
□	R2	RMSE	R2	RMSE
RFR	0.8250	0.6884	0.8242	0.6900
XGB	0.8407	0.6571	0.8384	0.6604
MLPR	0.8000	0.7382	0.8110	0.7431

Table S7. Formation energy model training results after merged data processed by feature engineering

□	Startegy 1		Startegy 2	
□	R2	RMSE	R2	RMSE
RFR	0.9287	0.2388	0.9292	0.2382
XGB	0.9480	0.2052	0.9473	0.2060
MLPR	0.9362	0.2342	0.9363	0.2265

Table S8. Standard deviation of the result after 100 training sessions of the bandgap model

□	Startegy 1		Startegy 2	
□	R2	RMSE	R2	RMSE
RFR	0.0502	0.1079	0.0500	0.1081
XGB	0.0567	0.1284	0.0518	0.1166
MLPR	0.0633	0.1279	0.0552	0.1152

Table S9. Standard deviation of the result after 100 training sessions of the Formation energy model

□	Startegy 1		Startegy 2	
□	R2	RMSE	R2	RMSE
RFR	0.0378	0.0778	0.0376	0.0775
XGB	0.0276	0.0652	0.0252	0.0622
MLPR	0.0286	0.0622	0.0279	0.0623

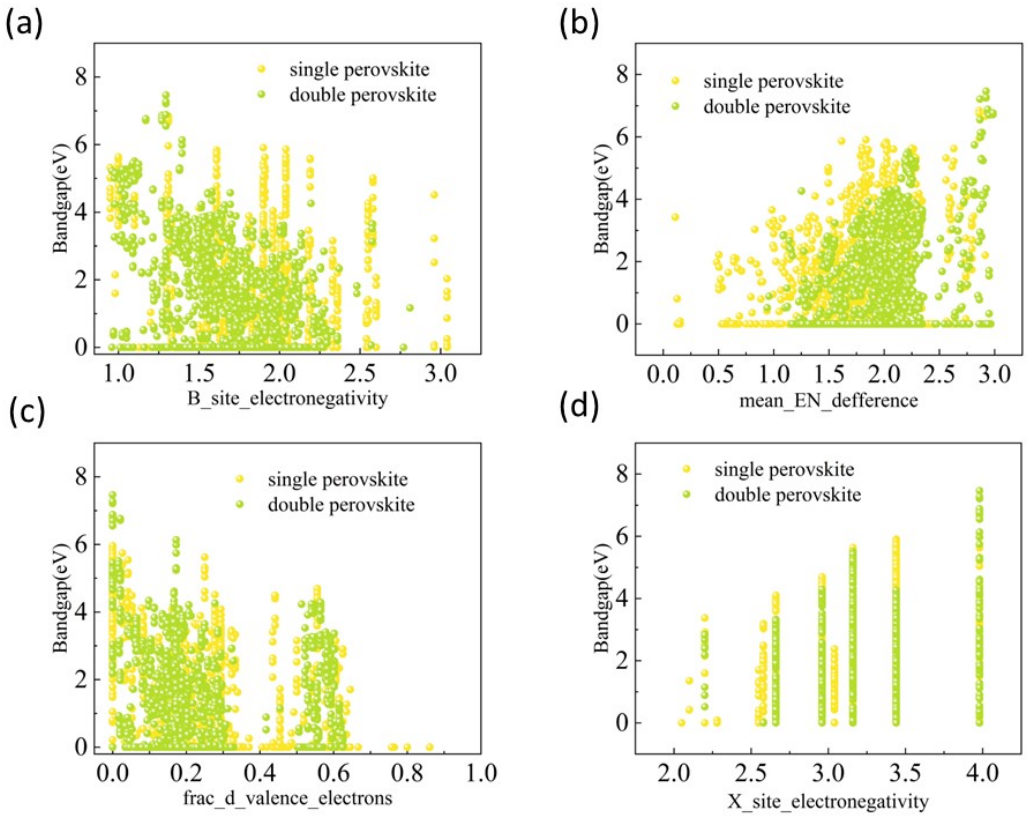

Figure S1. B_site_eletronegativity, mean EN difference, frac d valence electrns, X_site_ eletronegativity and bandgap distribution relationship

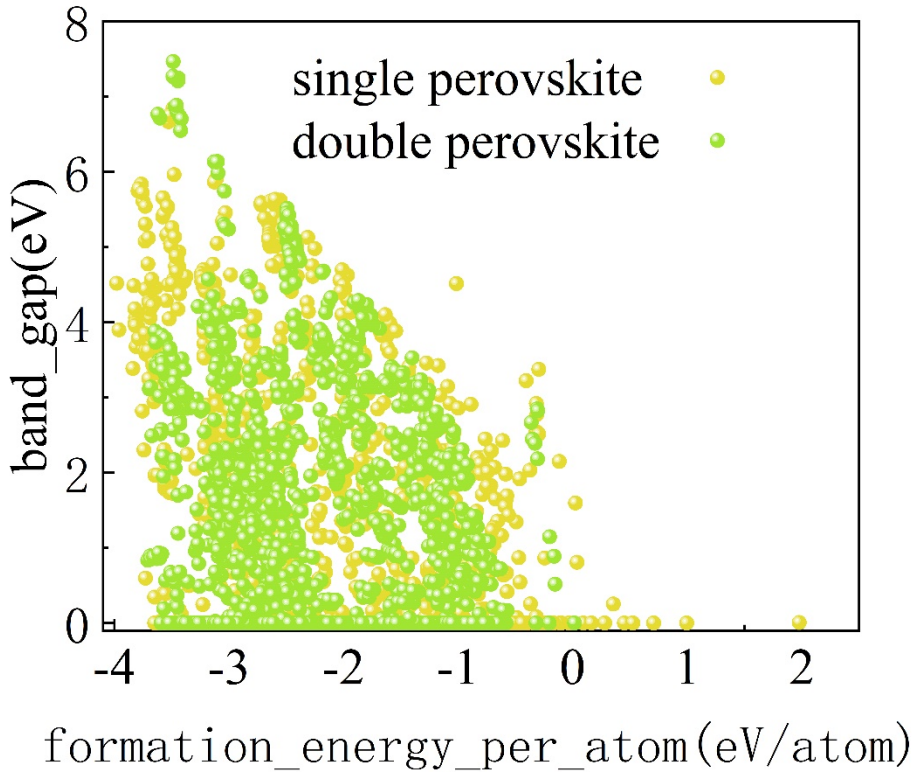

Figure S2. Formation energy and bandgap distribution relationship

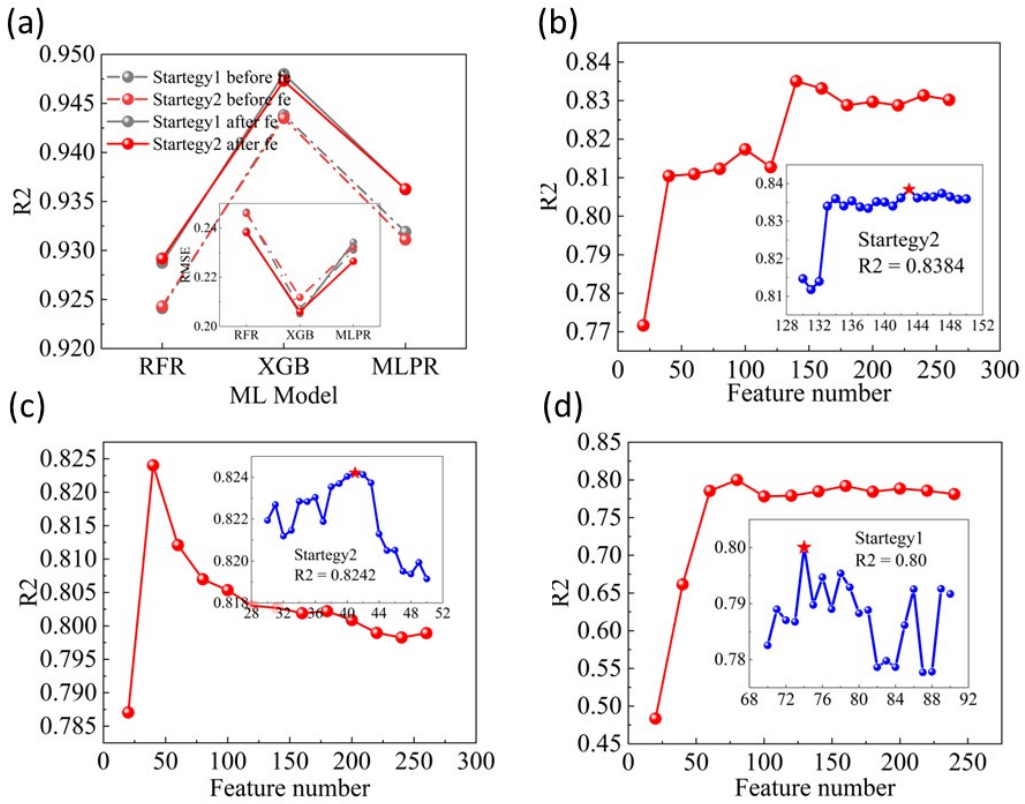

Figure S3. Feature engineering of Formation energy Model; (a) RFR, XGBoost and MLPR three models after feature engineering processing model effect; (b) XGBoost model training R2 changes with the number of features; (c) MLPR model training R2 changes with the number of features (d) R2 of RFR model training changes with the number of features;

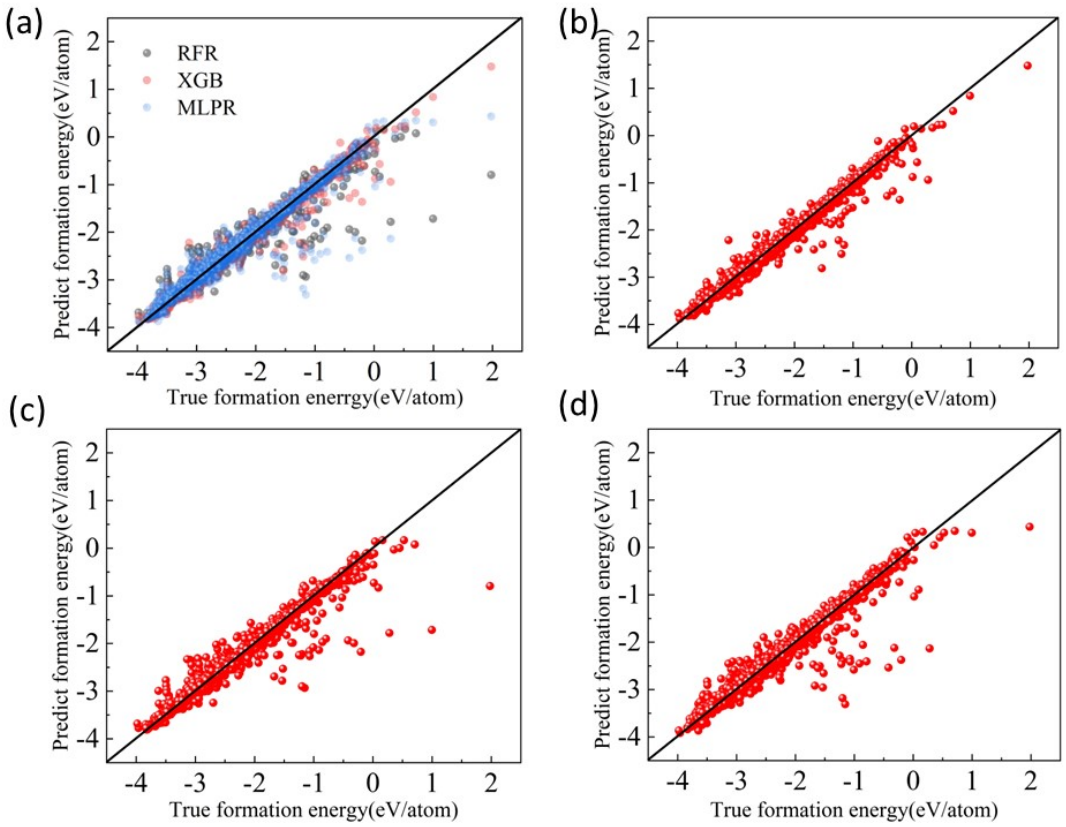

Figure S4. Fitting trend chart of the formation energy model；(a) Fitting trend chart of XGB&RFR&MLPR;(b) Fitting trend chart of XGB;(c) Fitting trend chart of RFR;(d) Fitting trend chart of MLPR;

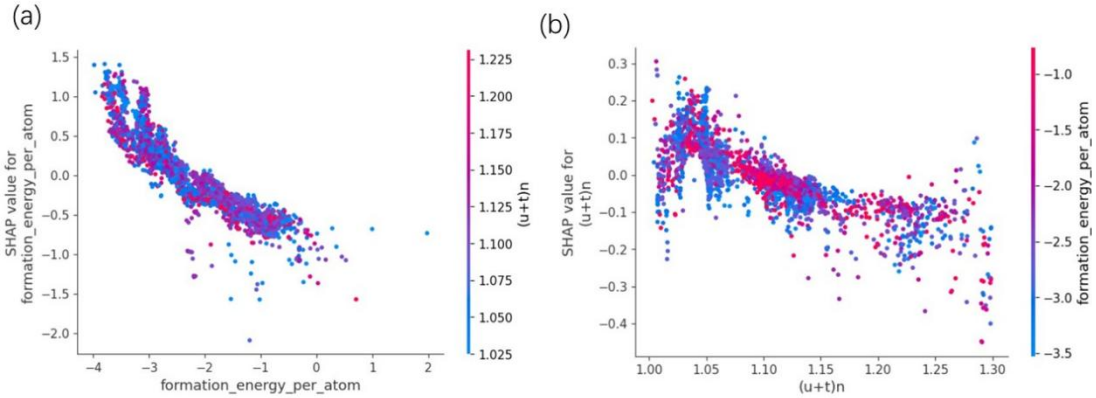

Figure S5. shap dependency graph of formation energy/$(\mu+t)^{\eta}$;

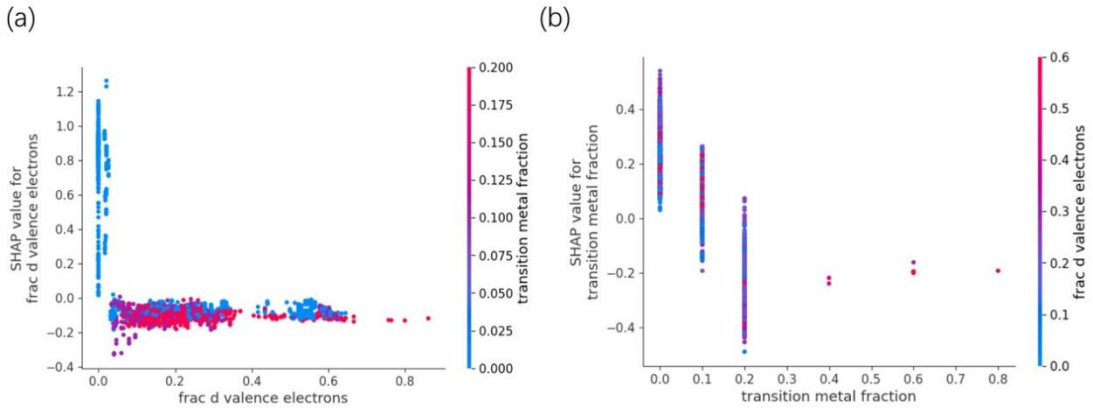

Figure S6. shap dependency graph of frac d valence electrons/transition metal fraction;

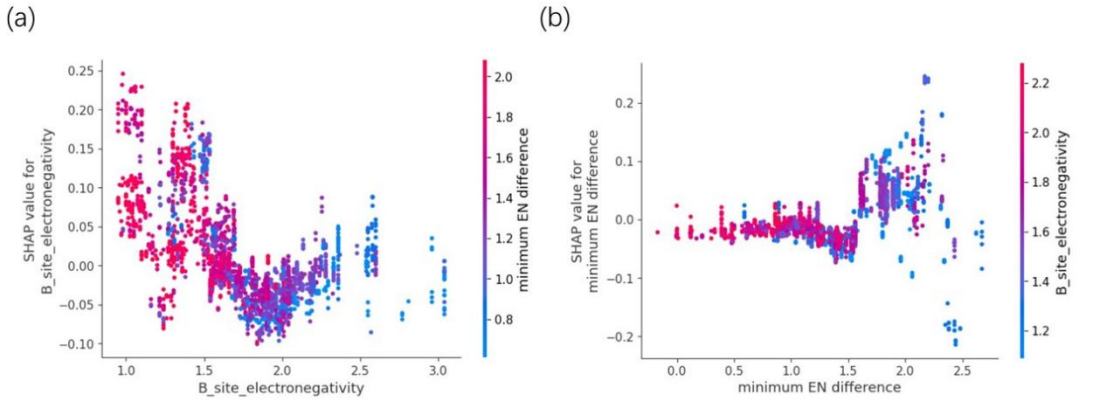

Figure S7. shap dependency graph of B site electronegativity/minimum EN difference;

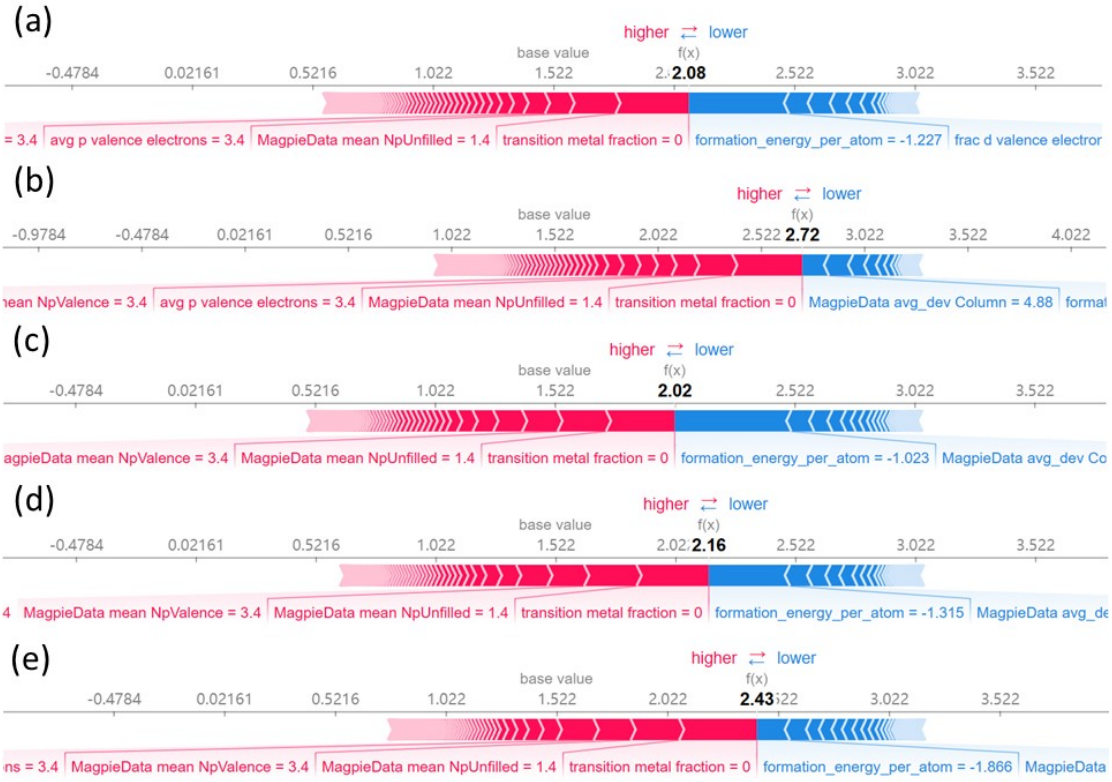

Figure S8. SHAP value of a single sample, where (a) is $CsSnBr_3$; (b) is $CsSnCl_3$; (c) is $CsPbI_3$; (d) is $CsPbBr_3$; (e) is $CsPbCl_3$;

Model hyperparameters

Bandgap Model

Ridge

ABX_3: {'alpha': 10, 'solver': 'saga'}

$A_2B(I)B(II)X_6$: {'alpha': 0.1, 'solver': 'lsqr'}

Startegy 1: {'alpha': 10, 'solver': 'lsqr'}

Startegy 2: {'alpha': 10, 'solver': 'lsqr'}

DT

ABX_3:{‘max_depth’:72,
’max_features’:183,’max_leaf_nodes’:172’,’min_samples_leaf’:5,’min_samples_split’:4}

$A_2B(I)B(II)X_6$:{‘max_depth’:24,’max_features’:70,’max_leaf_nodes’:293’,’min_samples_leaf’:3,
’min_samples_split’:3}

Startegy1:{‘max_depth’:40,’max_features’:270,’max_leaf_nodes’:108,’min_samples_leaf’:4,’min_samples_split’:2}

Startegy2:{‘max_depth’:60,’max_features’:199,’max_leaf_nodes’:144,’min_samples_leaf’:4,’min_samples_split’:2}

#RFR

ABX_3:{‘max_depth’:95,‘max_features’:’sqrt’,’min_samples_leaf’:1,’min_samples_split’:3,’n_estimators’:855}

$A_2B(I)B(II)X_6$:{‘max_depth’:95,‘max_features’:’sqrt’,
’min_samples_leaf’:1,’min_samples_split’:3,’n_estimators’:855}

Startegy1:{‘max_depth’:17,‘max_features’:’sqrt’,’min_sample_leaf’:1,’min_samples_split’3,’n_estimators’:944}

Startegy2:{‘max_depth’:72,‘max_features’:’sqrt’,’min_sample_leaf’:1,’min_samples_split’5,’n_estimators’:1756}

#SVR

ABX_3:{‘kernel’:rbf,’C’:70 ,’gamma’:0.0006，‘epsilon’:0.3}

$A_2B(I)B(II)X_6$: {‘kernel’:rbf,’C’:129 ,’gamma’:0.006631，‘epsilon’:0.1935}

Startegy 1: {‘kernel’:rbf,’C’:32,’gamma’:0.004135，‘epsilon’:0.128}

Startegy 2: {‘kernel’:rbf,’C’:22,’gamma’:0.003007，‘epsilon’:0.2721}

#XGBoost
ABX_3:{‘n_estimtors’:437,’eta’=0.0444,’reg_alpha’:4.182,’reg_lambda’:0.8283,’gamma’:0,’max_depth’:8,’colsample_bytree’:0.694,’colsample_bylevel’:0.6674,’colsample_bynode’:0.7819,’min_child_weight’:3.771}
$A_2B(I)B(II)X_6$:{‘n_estimtors’:211,’eta’=0.06295,’reg_alpha’:2.845,’reg_lambda’:0.02552,’gamma’:0,’max_depth’:9,’colsample_bytree’:0.8108,’colsample_bylevel’:0.7524,’colsample_bynode’:0.5698,’min_child_weight’:9.998}
Startegy1:{‘n_estimtors’:382,’eta’=0.0263,’reg_alpha’:0,’reg_lambda’:0,’gamma’:0,’max_depth’:10,’colsample_bytree’:1,’colsample_bylevel’:0.5,’colsample_bynode’:1,’min_child_weight’:9.998}
Startegy2:{‘n_estimtors’:343,’eta’=0.03052,’reg_alpha’:0.5829,’reg_lambda’:1.032,’gamma’:0,’max_depth’:9,’colsample_bytree’:0.5653,’colsample_bylevel’:0.5576,’colsample_bynode’:0.6966,’min_child_weight’:9.416}

#MLPR
ABX_3:{'activation':'logistic', 'alpha': 0.0630936388609197, 'hidden_layer_sizes': (136,34), 'learning_rate_init': 0.00010425699084203143, 'solver': 'adam',’learning_rate‘:’constant’}
$A_2B(I)B(II)X_6$:{'activation':'logistic','alpha':0.1,'hidden_layer_sizes':(200,40，10) , 'learning_rate_init': 0.00010425699084203143, 'solver': 'adam',’learning_rate‘:’constant’}
Startegy1:{'activation': 'logistic', 'alpha': 0.1, 'hidden_layer_sizes_1': (78,57,69), 'learning_rate_init': 0.01,'solver': 'adam'，‘learning_rate’:’constant’}
Startegy2:{'activation': 'logistic', 'alpha': 0.1, 'hidden_layer_sizes_1': (195,178,69), 'learning_rate_init': 0.01,'solver': 'adam',‘learning_rate’:’constant’}

Formation energy Model

#DT
ABX_3:{‘max_depth’:18,
’max_features’:70,’max_leaf_nodes’:298’,’min_samples_leaf’:3,’min_samples_split’:2}
$A_2B(I)B(II)X_6$:{‘max_depth’:20,’max_features’:158,’max_leaf_nodes’:168’,’min_samples_leaf’:6,’min_samples_split’:3}
Startegy1:{‘max_depth’:31,’max_features’:150,’max_leaf_nodes’:165,’min_samples_leaf’:7,’min_samples_split’:3}
Startegy2:{‘max_depth’:25,’max_features’:171,’max_leaf_nodes’:169,’min_samples_leaf’:6,’min_samples_split’:2}

#SVR
ABX_3:{‘kernel’:rbf,’C’:81,’gamma’:0.002175，‘epsilon’:0.08703}
$A_2B(I)B(II)X_6$: {‘kernel’:rbf,’C’:82,’gamma’:0.0001432，‘epsilon’:0.03054}
Startegy 1: {‘kernel’:rbf,’C’:32,’gamma’:0.001582，‘epsilon’:0.01925}
Startegy 2: {‘kernel’:rbf,’C’:140,’gamma’:0.0005126，‘epsilon’:0.01985}

#Ridge
ABX_3: {'alpha': 10, 'solver': 'lsqr'}
$A_2B(I)B(II)X_6$: {'alpha': 10, 'solver': 'lsqr'}
Startegy 1: {'alpha': 1, 'solver': 'svd'}
Startegy 2: {'alpha': 1, 'solver': 'svd'}

#RFR
ABX_3:{‘max_depth’:20,‘max_features’:’sqrt’,’min_samples_leaf’:1,’min_samples_split’:2,’n_estimators’:896}
$A_2B(I)B(II)X_6$:{‘max_depth’:44,‘max_features’:’sqrt’,’min_samples_leaf’:3,’min_samples_split’:3,’n_estimators’:872}
Startegy1:{‘max_depth’:16,‘max_features’:’sqrt’,’min_sample_leaf’:2,’min_samples_split’2,’n_estimators’:894}
Startegy2:{‘max_depth’:19,‘max_features’:’sqrt’,’min_sample_leaf’:2,’min_samples_split’4,’n_estimators’:905}

#XGBoost
ABX_3:{‘n_estimtors’:482,’eta’=0.09802,’reg_alpha’:0.1728,’reg_lambda’:0.8499,’gamma’:0,’max_depth’:4,’colsample_bytree’:0.8897,’colsample_bylevel’:0.8691,’colsample_bynode’:0.7906,’min_child_weight’:1.502}
$A_2B(I)B(II)X_6$:{‘n_estimtors’:478,’eta’=0.1752,’reg_alpha’:0.5263,’reg_lambda’:0.2829,’gamma’:0,’max_depth’:3,’colsample_bytree’:0.9645,’colsample_bylevel’:0.7488,’colsample_bynode’:0.9907,’min_child_weight’:1.051}
Startegy1:{‘n_estimtors’:497,’eta’=0.1557,’reg_alpha’:1.176,’reg_lambda’:0.487,’gamma’:0,’max_depth’:3,’colsample_bytree’:0.7591,’colsample_bylevel’:0.5681,’colsample_bynode’:0.7721,’min_child_weight’:2.762}
Startegy2:{‘n_estimtors’:499,’eta’=0.1201,’reg_alpha’:0.4421,’reg_lambda’:1.218,’gamma’:0,’max_depth’:4,’colsample_bytree’:0.9786,’colsample_bylevel’:0.6738,’colsample_bynode’:0.672,’min_child_weight’:9.896}

#MLPR
ABX_3:{'activation':'logistic', 'alpha': 0.025631686291210192, 'hidden_layer_sizes': (110,63), 'learning_rate_init': 0.005226903632157869, 'solver': 'adam','learning_rate‘:’constant’}
$A_2B(I)B(II)X_6$:{'activation':'logistic', 'alpha': 0.09967573041687423, 'hidden_layer_sizes': (79), , 'learning_rate_init': 0.0007669614243494098, 'solver': 'adam','learning_rate‘:’constant’}
Startegy1:{'activation': 'logistic', 'alpha':0.03467271741957278, 'hidden_layer_sizes': (198), 'learning_rate_init': 0.0031155210604508106,'solver': 'adam'，‘learning_rate’:’constant’}
Startegy2:{'activation':'logistic','alpha': 0.028072119868422, 'hidden_layer_sizes_1': (152), 'learning_rate_init': 0.001691193658192399,'solver': 'adam',‘learning_rate’:’constant’}
Feature Engineering（features number）

Band gap model

#RFR:
Startegy 1:30 features
Startegy 2:41 features

#XGBoost:
Startegy 1:135 features
Startegy 2:143 features

#MLPR:
Startegy 1:74 features
Startegy 2:103 features

Formation energy model
#RFR:
Startegy 1:52features
Startegy 2:43features

#XGBoost:
Startegy 1:218features
Startegy 2:111features

#MLPR:
Startegy 1:234features
Startegy 2:125features

Code and Data: https://github.com/czmscut/ML_pvk_bandgap_predict