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

# Supplementary information:

# A domain knowledge enhanced machine learning method to predict the properties of halide double perovskite $A_2B^+B^{3+}X_6$

Xiao Wei[a], Yunong Zhang[a], Xi Liu[a], Junjie Peng[a], Shengzhou Li[b], Renchao Che[c], Huiran Zhang[a, d, *]

[a] School of Computer Engineering and Science, Shanghai University, Shanghai, 200444, China.

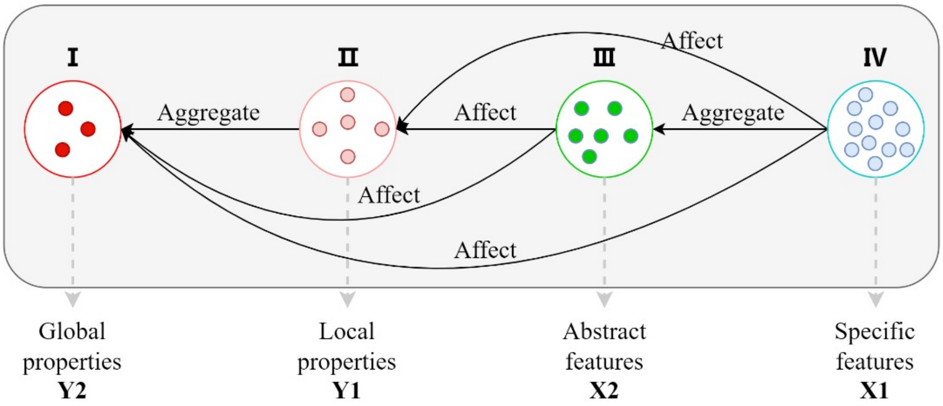
[b] Department of Computer Science, University of Tsukuba, Tsukuba, Ibaraki 305-8573, Japan.

[c] Laboratory of Advanced Materials, Shanghai Key Lab of Molecular Catalysis and Innovative Materials, Department of Materials Science, Fudan University, Shanghai 200438, China.

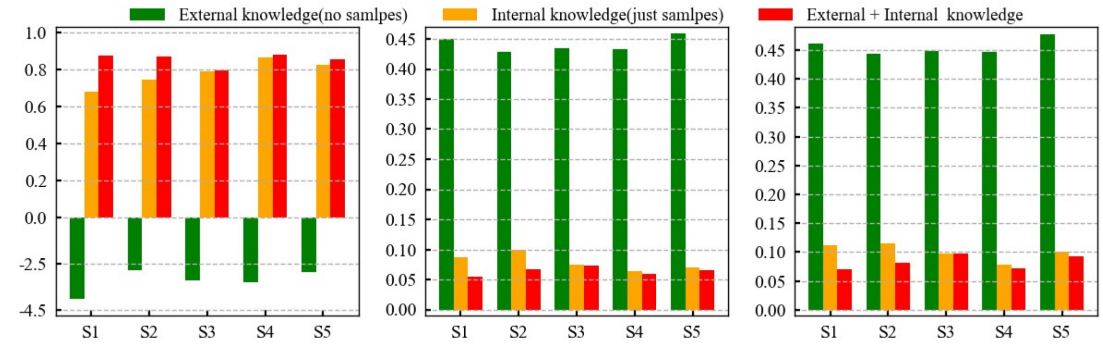
[d] Key Laboratory of Silicate Cultural Relics Conservation (Shanghai University), Ministry of Education, Shanghai University, Shanghai, 200444, China

[*] Corresponding author; E-mail: hrzhangsh@shu.edu.cn (H, Zhang).

## Supplementary Figures

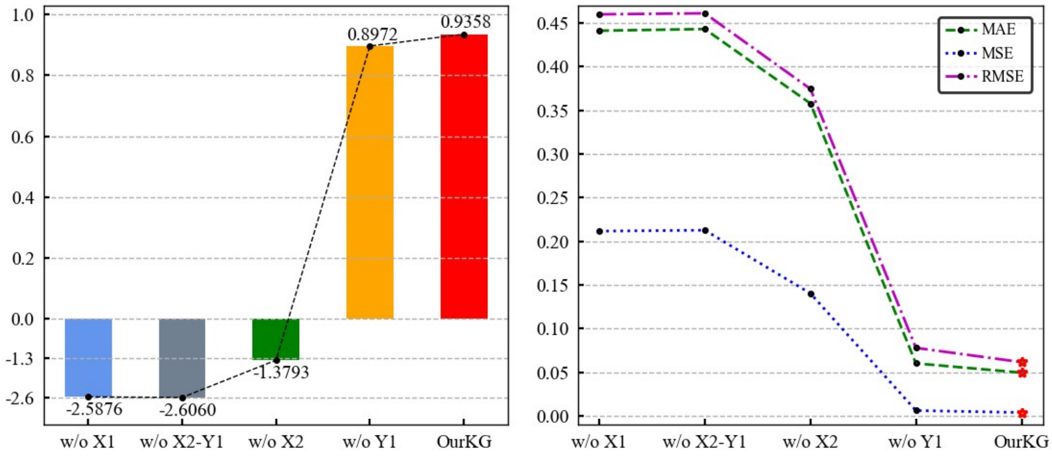


Supplementary Figure 1: Structural pattern of domain knowledge in materials Science. It mainly includes two aspects: features and properties, Specifically, it has four layers: specific features layer(IV), abstract features layer(III), local properties layer(II) and global properties layer(I).



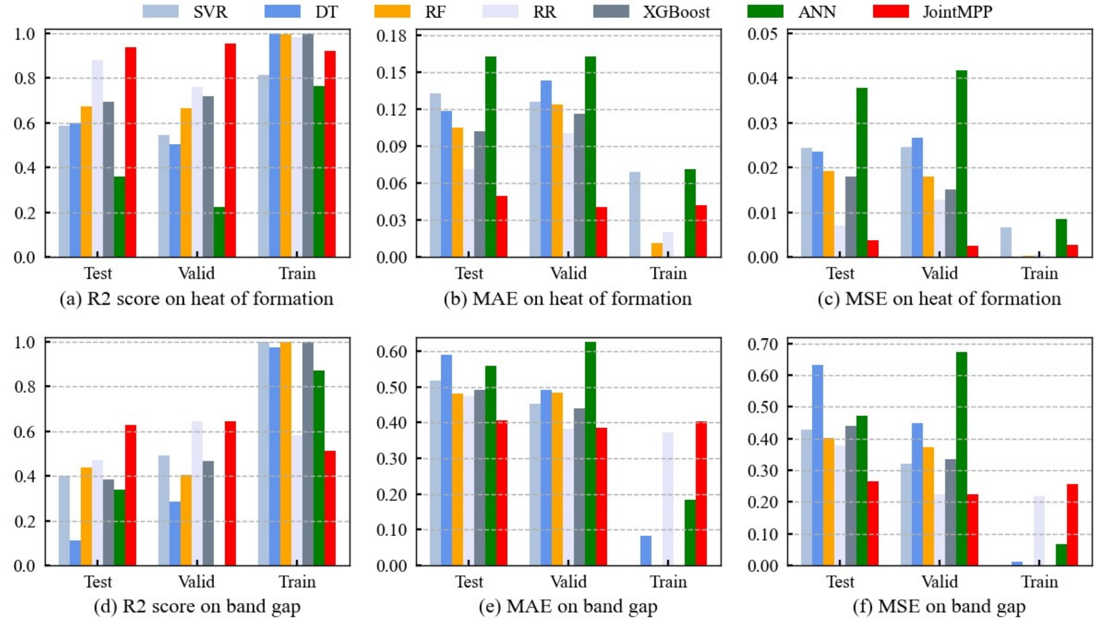
(a) R2 score (b) MAE (c) RMSE

Supplementary Figure 2: Performance testing of material domain knowledge with different types. (a) The coefficient of determinations $R^2$ on five experimental datasets. (b) Average absolute error $MAE$ on five experimental datasets. (c) Root mean square error $RMSE$ on five experimental datasets. As for S1, S2, S3, S4 and S5, they represent five groups of experimental datasets obtained from randomly sampling. It indicates that the external domain knowledge(no sample features) is difficult to fit the distribution pattern of real dataset. Compared with the knowledge graph constructed in this paper, the prediction performance of the internal knowledge graph(just sample features) is average. It proves that our hierarchical knowledge graph have an important guiding role in predicting material properties and can significantly improve model performance.



(a) R2 score (b) Error

Supplementary Figure 3: Performance testing of material domain knowledge on different structures. (a) The $R^2$ score of domain knowledge on different hierarchies, where “w/o X1” represents the structural pattern without “specific features layer” (three layers), “w/o X2-Y1” represents the structural pattern without “abstract features layer” and “local properties layer” (two layers), “w/o X2” represents the structural pattern without “abstract features layer” (three layers), “w/o Y1” represents the structural pattern without “local properties layer” (three layers), and “OurKG” represents the structural pattern constructed in this paper (four layers). (b) Errors in performance prediction guided by different hierarchical structural patterns, including mean absolute error $MAE$, mean squared error $MSE$, and root mean square error $RMSE$. It illustrates the necessity and validity of our structural pattern with a four-layer structure, while “w/o Y1” predicts well because there are fewer local properties to be excluded.



(a) R2 score on heat of formation

(b) MAE on heat of formation

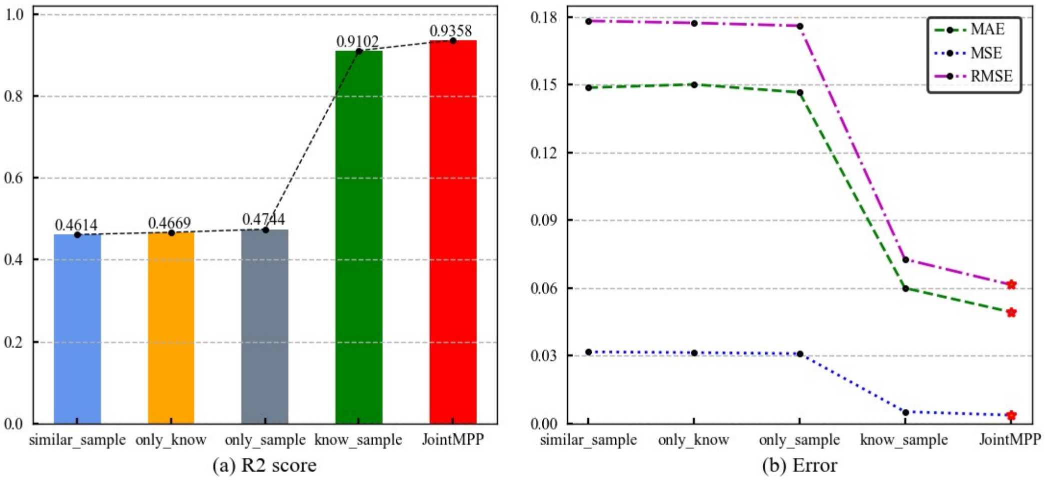
(c) MSE on heat of formation

(d) R2 score on band gap

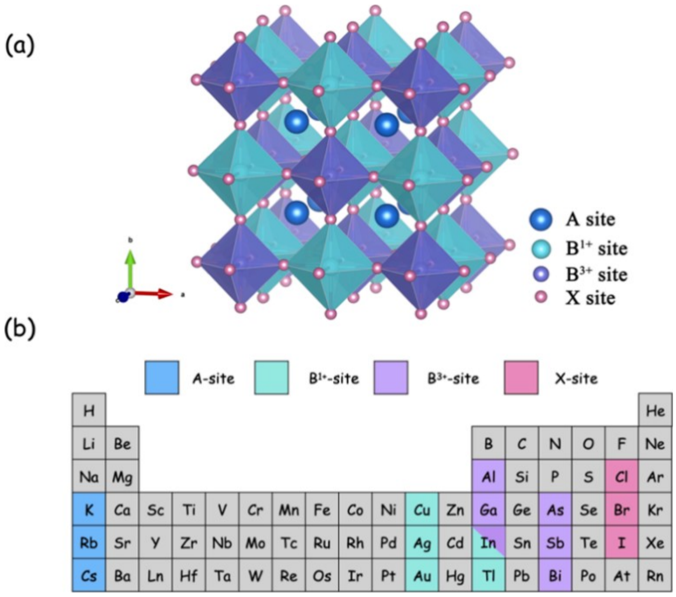
(e) MAE on band gap

(f) MSE on band gap

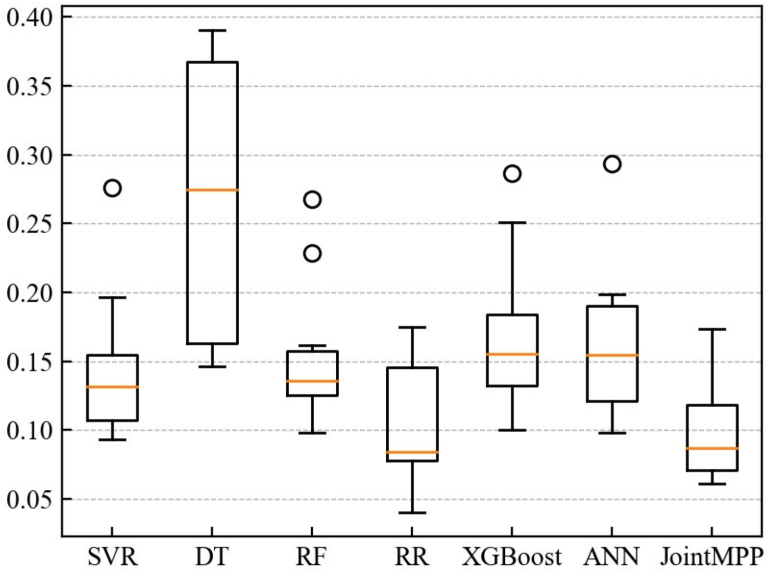
Supplementary Figure 4: Performance quantification of seven different methods. From (a) and (d), Our JointMPP achieved optimal performance on the validation and testing sets, although the prediction effect on the training set was not optimal. This shows that the method proposed in this paper has good generalization performance and robustness. From the error quantification results of MAE and MSE, it can also be seen that the superiority of the proposed JointMPP.



Supplementary Figure 5: Predictive performance of JointMPP with respect to different modules. Based on the model structure of JointMPP, the ablation experimental method is divided into four types: "similar_sample" means that the properties prediction is made by using similar samples and query samples. "only_know" indicates that predictions are made only from domain knowledge, and the node representation is initialized with sample features. "only_sample" represents training and testing using only the features of query samples. "know_sample" indicates that it employs material domain knowledge to enhance presentation learning for limited samples. Our JointMPP achieves the best fitted performance ($R^2$ is 0.9358) and the lowest error ($MAE$, $MSE$ and $RMSE$ are all minimal). It illustrates that domain knowledge is critical to improving the performance of a limited sample learning approach in materials science.


(a)
A site
B¹⁺ site
B³⁺ site
X site
b
a
(b)
A-site
B¹⁺-site
B³⁺-site
X-site
H He
Li Be B C N O F Ne
Na Mg Al Si P S Cl Ar
K Ca Sc Ti V Cr Mn Fe Co Ni Cu Zn Ga Ge As Se Br Kr
Rb Sr Y Zr Nb Mo Tc Ru Rh Pd Ag Cd In Sn Sb Te I Xe
Cs Ba Ln Hf Ta W Re Os Ir Pt Au Hg Tl Pb Bi Po At Rn

Supplementary Figure 6: The crystal structures [1] of halide double perovskite with A, $B^{1+}$, $B^{3+}$ and X positions are respectively shown in deepskyblue, light green, purple as well as pink (a). Element selection at each point of halide double perovskite (b).



Supplementary Figure 7: Stability analysis of seven different methods (the evaluation metric is $RMSE$).

## Supplementary Tables

Supplementary Table 1: Descriptors of halide double perovskites[1]. It includes 540 samples, 33 features and two property (heat of formation and band gap).

| No. | Features name | Meanings |
|---|---|---|
| 1 | distance_a | Distance between cation at $A^+$ site |
| 2 | distance_b1 | Distance between cation at $B^{1+}$ site |
| 3 | distance_b2 | Distance between cation at $B^{3+}$ site |
| 4 | cubic | Space group of crystal 1 |
| 5 | ortho | Space group of crystal 2 |
| 6 | eleneg_a | Electronegativity of $A^+$site |
| 7 | eleneg_b1 | Electronegativity of $B^{1+}$ site |
| 8 | eleneg_b2 | Electronegativity of $B^{3+}$ site |
| 9 | eleneg_x | Electronegativity of $X^-$ site |
| 10 | hoe_a | Highest occupied energy level of $A^+$ site |
| 11 | hoe_b1 | Highest occupied energy level of $B^{1+}$ site |
| 12 | hoe_b2 | Highest occupied energy level of $B^{3+}$ site |
| 13 | hoe_x | Highest occupied energy level of $X^-$ site |
| 14 | ionenergy_a | Ionization energy of $A^+$ site |
| 15 | ionenergy_b1 | Ionization energy of $B^{1+}$ site |
| 16 | ionenergy_b2 | Ionization energy of $B^{3+}$ site |
| 17 | ionenergy_x | Ionization energy of $X^-$ site |
| 18 | luep_a | Lowest unoccupied energy level of $A^+$ site |
| 19 | luep_b1 | Lowest unoccupied energy level of $B^{1+}$ site |
| 20 | luep_b2 | Lowest unoccupied energy level of $B^{3+}$ site |
| 21 | luep_x | Lowest unoccupied energy level of $X^-$ site |
| 22 | rs_a | Radius of s-orbital of $A^+$ site |
| 23 | rs_b1 | Radius of s-orbital of $B^{1+}$ site |
| 24 | rs_b2 | Radius of s-orbital of $B^{3+}$ site |
| 25 | rs_x | Radius of s-orbital of $X^-$ site |
| 26 | rp_a | Radius of p-orbital of $A^+$ site |
| 27 | rp_b1 | Radius of p-orbital of $B^{1+}$ site |
| 28 | rp_b2 | Radius of p-orbital of $B^{3+}$ site |
| 29 | rp_x | Radius of p-orbital of $X^-$ site |
| 30 | rd_a | Radius of d-orbital of $A^+$ site |
| 31 | rd_b1 | Radius of d-orbital of $B^{1+}$ site |
| 32 | rd_b2 | Radius of d-orbital of $B^{3+}$ site |
| 33 | rd_x | Radius of d-orbital of $X^-$ site |

Supplementary Table 2: Quantitative prediction results of data quality improvement methods on heat of formation and band gap, where K1 stands for “domain knowledge for sample enhancement”, K2 represents “domain knowledge for feature selection”, and K3 is “domain knowledge for feature extraction”. (train:test = 2:8)

| **Methods** | **heat of formation** | | | | **Band gap** | | | |
|---|---|---|---|---|---|---|---|---|
| | **R2** | **MAE** | **MSE** | **RMSE** | **R2** | **MAE** | **MSE** | **RMSE** |
| SVR | 0.5874 | 0.1323 | 0.0243 | 0.1560 | 0.3983 | 0.5178 | 0.4292 | 0.6551 |
| DT | 0.5988 | 0.1186 | 0.0237 | 0.1538 | 0.1121 | 0.5902 | 0.6334 | 0.7958 |
| RF | 0.6738 | 0.1052 | 0.0192 | 0.1387 | 0.4373 | 0.4805 | 0.4014 | 0.6336 |
| RR | 0.8793 | 0.0710 | 0.0071 | 0.0844 | 0.4713 | 0.4725 | 0.3772 | 0.6141 |
| XGBoost | 0.6938 | 0.1017 | 0.0181 | 0.1344 | 0.3850 | 0.4927 | 0.4387 | 0.6624 |
| ANN | 0.3585 | 0.1626 | 0.0378 | 0.1945 | 0.3387 | 0.5588 | 0.4718 | 0.6869 |
| SVR-K1 | 0.6482 | 0.1213 | 0.0208 | 0.1441 | 0.4124 | 0.5008 | 0.4192 | 0.6474 |
| DT-K1 | 0.6881 | 0.1059 | 0.0184 | 0.1357 | 0.2326 | 0.5446 | 0.5474 | 0.7399 |
| RF-K1 | 0.6808 | 0.1041 | 0.0188 | 0.1372 | 0.4443 | 0.4782 | 0.3964 | 0.6296 |
| RR-K1 | 0.8924 | 0.0654 | 0.0064 | 0.0797 | 0.5468 | 0.4546 | 0.3233 | 0.5686 |
| XGBoost-K1 | 0.7374 | 0.0900 | 0.0155 | 0.1245 | 0.4762 | 0.4521 | 0.3737 | 0.6113 |
| ANN-K1 | 0.8409 | 0.0804 | 0.0094 | 0.0969 | 0.6585 | 0.3909 | 0.2436 | 0.4936 |
| SVR-K2 | 0.6053 | 0.1288 | 0.0233 | 0.1526 | 0.4178 | 0.5107 | 0.4153 | 0.6444 |
| DT-K2 | 0.6268 | 0.1108 | 0.0220 | 0.1484 | 0.3142 | 0.5233 | 0.4892 | 0.6994 |
| RF-K2 | 0.6823 | 0.1020 | 0.0187 | 0.1369 | 0.4725 | 0.4631 | 0.3763 | 0.6134 |
| RR-K2 | 0.9556 | 0.0412 | 0.0026 | 0.0512 | 0.5990 | 0.4270 | 0.2861 | 0.5349 |
| XGBoost-K2 | 0.6725 | 0.1043 | 0.0193 | 0.1390 | 0.4241 | 0.4705 | 0.4108 | 0.6409 |
| ANN-K2 | 0.5110 | 0.1352 | 0.0289 | 0.1698 | 0.7101 | 0.3568 | 0.2068 | 0.4547 |
| SVR-K3 | 0.9118 | 0.0563 | 0.0052 | 0.0721 | 0.5163 | 0.4710 | 0.3450 | 0.5874 |
| DT-K3 | 0.8458 | 0.0727 | 0.0091 | 0.0954 | 0.3652 | 0.4883 | 0.4528 | 0.6729 |
| RF-K3 | 0.8860 | 0.0619 | 0.0067 | 0.0820 | 0.5185 | 0.4617 | 0.3435 | 0.5861 |
| RR-K3 | 0.9253 | 0.0544 | 0.0044 | 0.0664 | 0.4309 | 0.5025 | 0.4060 | 0.6371 |
| XGBoost-K3 | 0.8848 | 0.0615 | 0.0068 | 0.0824 | 0.4560 | 0.4786 | 0.3881 | 0.6229 |
| ANN-K3 | 0.9099 | 0.0595 | 0.0053 | 0.0729 | 0.5630 | 0.4447 | 0.3118 | 0.5583 |
| JointMPP | 0.9358 | 0.0494 | 0.0038 | 0.0615 | 0.6265 | 0.4073 | 0.2664 | 0.5162 |

Supplementary Table 3: Prediction results of six classical machine learning methods[1] in massive training samples. It mainly includes Decision Trees (DT), Artificial Neural Network (ANN), Random Forest (RF), Ridge Regression (RR), Support Vector Regression (SVR) and XGBoost. (train: test = 8: 2)

| **Methods** | **Heat of formation** | | | | **Band gap** | | | |
|---|---|---|---|---|---|---|---|---|
| | **R2** | **MAE** | **MSE** | **RMSE** | **R2** | **MAE** | **MSE** | **RMSE** |
| SVR | 0.9692 | 0.0373 | 0.0021 | 0.0459 | 0.6389 | 0.3771 | 0.265 | 0.5148 |
| DT | 0.9835 | 0.0258 | 0.0011 | 0.0336 | 0.7634 | 0.2349 | 0.1737 | 0.4167 |
| RF | 0.9923 | 0.018 | 0.0005 | 0.0229 | 0.8945 | 0.1906 | 0.0774 | 0.2782 |
| RR | 0.9888 | 0.019 | 0.0008 | 0.0277 | 0.7017 | 0.3532 | 0.219 | 0.4679 |
| XGBoost | 0.9986 | 0.0075 | 0.0001 | 0.0099 | 0.9046 | 0.1732 | 0.07 | 0.2647 |
| ANN | 0.9535 | 0.046 | 0.0032 | 0.0564 | 0.7562 | 0.327 | 0.179 | 0.423 |
| JointMPP | 0.9876 | 0.0226 | 0.0008 | 0.0291 | 0.7672 | 0.3239 | 0.1708 | 0.4133 |

Supplementary Table 4: Stability quantization result of seven different methods, and the evaluation metric is $R^2$ and $RMSE$. ($R^2$ is the top table and $RMSE$ is the bottom table).

| Methods | DT | ANN | RF | RR | SVR | XGBoost | Ours |
|---|---|---|---|---|---|---|---|
| Sample 1 | 0.5988 | 0.3585 | 0.6738 | 0.8793 | 0.5874 | 0.6938 | 0.9358 |
| Sample 2 | -2.5046 | 0.3451 | 0.5802 | 0.8398 | 0.4963 | 0.5266 | 0.8756 |
| Sample 3 | 0.8256 | 0.5560 | 0.8060 | 0.8992 | 0.7290 | 0.6522 | 0.4322 |
| Sample 4 | 0.5725 | 0.6754 | 0.6516 | 0.9683 | 0.8288 | 0.8023 | 0.8700 |
| Sample 5 | 0.8539 | 0.4546 | 0.6853 | 0.6267 | 0.6072 | 0.4341 | 0.4721 |
| Sample 6 | 0.1579 | 0.7890 | 0.7856 | 0.4677 | 0.7601 | 0.5191 | 0.7951 |
| Sample 7 | 0.5489 | 0.8354 | 0.6994 | 0.4636 | 0.7299 | 0.1474 | 0.7242 |
| Sample 8 | -2.4395 | 0.1113 | 0.6500 | 0.5366 | 0.1297 | 0.2276 | 0.8818 |
| Sample 9 | 0.6295 | 0.7860 | 0.6357 | 0.8956 | 0.7399 | 0.7335 | 0.8534 |
| Sample 10 | 0.1894 | 0.7251 | 0.4910 | 0.8630 | 0.8252 | 0.4864 | 0.9113 |
| Sample 1 | 0.1560 | 0.1538 | 0.1387 | 0.0844 | 0.1344 | 0.1945 | 0.0615 |
| Sample 2 | 0.1408 | 0.3713 | 0.1285 | 0.0794 | 0.2509 | 0.1605 | 0.0700 |
| Sample 3 | 0.1158 | 0.1904 | 0.0980 | 0.0706 | 0.1312 | 0.1482 | 0.1676 |
| Sample 4 | 0.0935 | 0.1477 | 0.1333 | 0.0403 | 0.1004 | 0.1287 | 0.0815 |
| Sample 5 | 0.1499 | 0.3567 | 0.2283 | 0.1462 | 0.1800 | 0.1767 | 0.1738 |
| Sample 6 | 0.1048 | 0.3449 | 0.0991 | 0.1561 | 0.1484 | 0.0983 | 0.0969 |
| Sample 7 | 0.2762 | 0.3869 | 0.2679 | 0.1752 | 0.2862 | 0.2935 | 0.1256 |
| Sample 8 | 0.1964 | 0.3904 | 0.1245 | 0.1433 | 0.1850 | 0.1984 | 0.0724 |
| Sample 9 | 0.1228 | 0.1465 | 0.1453 | 0.0778 | 0.1243 | 0.1114 | 0.0922 |
| Sample 10 | 0.0949 | 0.2043 | 0.1619 | 0.0840 | 0.1626 | 0.1190 | 0.0676 |

Supplementary Table 5: Experimental data[2] for classifying perovskite materials.

| **Number** | **Compound ABX** | **Compound AB1B2X** |
|---|---|---|
| train sets | 460 | 734 |
| test sets | 116 | 184 |
| Positive sample | 313 | 868 |
| Negative sample | 263 | 50 |
| Element A | 49 | 35 |
| Element B | 67 | --- |
| Element B1 | --- | 52 |
| Element B2 | --- | 48 |
| Element X | 5 | 5 |

Data Resources: https://www.science.org/doi/10.1126/sciadv.aav0693#supplementary-materials.

Supplementary Table 6: Perovskite classification result of ten different methods. Where, "LR1" and "LR2" stand for Logistic Regression with a L1 penalty term and a L2 penalty term respectively. "SVM" is Support Vector Machines, "KNN" is K Nearest Neighbors algorithm, "DT" is Decision Trees, "RF" is Random Forests, "ANN" is Artificial Neural Network, "GBDT" is Gradient Boosted Decision Trees and "ABC" is Ada Boost. (train: test = 8: 2)

| **Methods** | **Perovskite ABX** | | | | **Perovskite ABBX** | | | |
|---|---|---|---|---|---|---|---|---|
| | **Acc** | **MP** | **MR** | **F1** | **Acc** | **MP** | **MR** | **F1** |
| LR1 | 88.79 | 88.88 | 88.21 | 88.48 | 95.65 | 85.56 | 64.71 | 70.30 |
| LR2 | 87.07 | 86.89 | 86.70 | 86.79 | 95.65 | 85.56 | 64.71 | 70.30 |
| SVM | 87.07 | 86.77 | 86.94 | 86.85 | 95.65 | 85.56 | 64.71 | 70.30 |
| KNN | 89.66 | 89.94 | 88.97 | 89.34 | 96.20 | 82.19 | 79.14 | 80.58 |
| DT | 90.52 | 90.67 | 89.97 | 90.26 | 96.74 | 86.36 | 79.43 | 82.48 |
| RF | 91.38 | 91.42 | 90.97 | 91.17 | 97.28 | 91.73 | 79.71 | 84.58 |
| ANN | 89.66 | 89.94 | 88.97 | 89.34 | 95.65 | 85.56 | 64.71 | 70.30 |
| GBDT | 91.38 | 92.73 | 90.24 | 90.99 | 96.20 | 84.30 | 74.43 | 78.41 |
| ABC | 88.79 | 88.65 | 88.45 | 88.55 | 96.20 | 88.32 | 69.71 | 75.68 |
| JointMPP | 86.21 | 86.09 | 85.70 | 85.87 | 95.65 | 85.56 | 64.71 | 70.30 |

**Supplementary Methods**

In order to verify the improvement effect of domain knowledge fusion on training data quality, this paper proposes three kinds of data preprocessing methods guided by domain knowledge. It mainly includes:

①The Hadamard product of the correlation matrix (dimension is 90×33) of features and properties in the material domain knowledge and the original training sample matrix (dimension is 90×33) is used to obtain the training sample based on the correlation (dimension is 90×33). And then the two training samples are combined to increase the number of training dataset (dimension is 180×33);

②The sparse feature correlation (dimension is 28) is obtained by assigning the feature correlation below a certain threshold in the domain knowledge as 0, and then the feature of the original training sample (dimension is 33) is screened by this correlation to get the key sample feature (dimension 28);

③By using the dot product of the hierarchical correlation matrix (dimension is 33×9) and the original training sample matrix (dimension is 90×33), the domain knowledge can achieve the aggregate extraction of the original sample features, and the final training data dimension is 90×9.

## Supplementary References

1 Z.-t. Guo and B. Lin, *Solar Energy*, 2021, **228**, 689-699.

2 C. J. Bartel, C. Sutton, B. R. Goldsmith, R. Ouyang, C. B. Musgrave, L. M. Ghiringhelli and M. Scheffler, *Science Advances*, 2018, **5**.