Electronic Supplementary Material (ESI) for Materials Horizons.

Discovery of All-Inorganic Lead-Free Perovskites with High Photovoltaic Performance via Ensemble Machine Learning

Xia Cai,[1,*] Yan Li,[1] Jianfei Liu,[1] Hao Zhang,[2,†] Jianguo Pan,[1,‡] and Yiqiang Zhan[2]

[1]*College of Information, Mechanical and Electrical Engineering, Shanghai Normal University, Shanghai 200234, China*

[2]*School of Information Science and Technology, Fudan University, Shanghai 200433, China*

The distribution of collected dataset

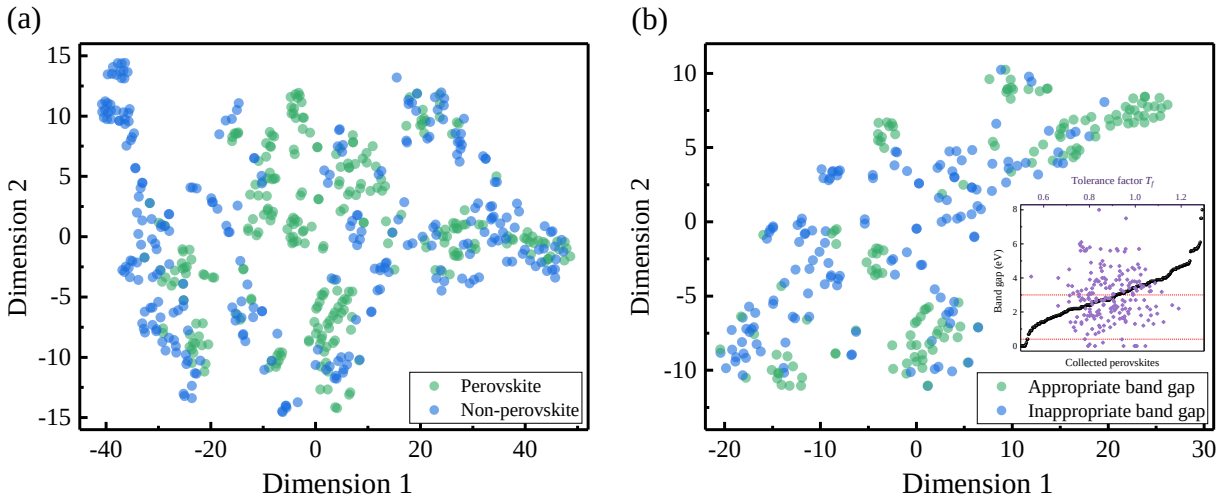

FIG. S1. Feature visualization of the dataset using t-SNE to maintain the original distributions of the high-dimensional data by preserving pairwise similarities: 68-dimensional features of the entire samples are compressed into a two-dimensional space for the tasks of (a) perovskite structure identification and (b) band gap classification, respectively. The inset is the detailed band gap distribution used for physics-inspired multi-component neural network and the visualization of tolerance factor and collected band gap.

The data guarantee for model generalization

Principal component analysis (PCA) method as a linear dimensionality reduction technique is employed on the input datasets of perovskite structure identification and band gap classification, and the prediction set with unknown materials, respectively, which can reduce the dimensionality of data while preserving its most important variations for information retention and allows us to capture the diversity of the data to some extent. The first three crucial features are utilized for distribution analysis and visualization, and the results are presented in FIG. S2, where it is evident that the statistical distributions of the three datasets remain consistent across Dimension 1 to 3. And the ranges of feature values contained in the two input datasets are substantially comparable to those in the prediction set, with disparities primarily lying in the corresponding quantities. This consistency ensures the generalization ability of the constructed model. Moreover, in the prediction set, the candidates containing magnetic elements such as Co, Ni and so on are excluded, and after the third mandatory screening about toxicity, the types of elements contained by the candidates in the final prediction set are almost similar to those in the training set.

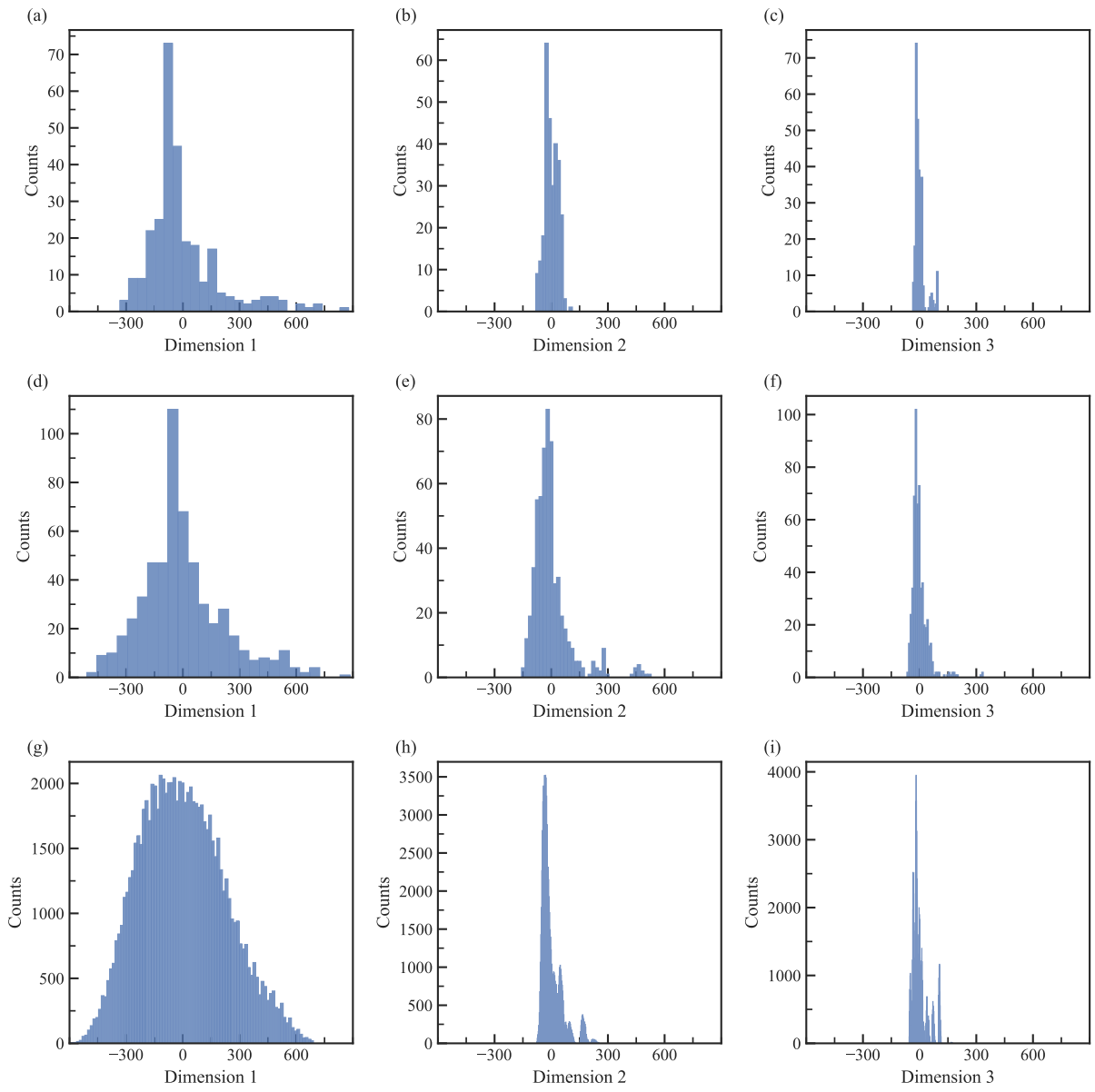

FIG. S2. The statistics on the number of data for three important feature dimensions by PCA to reduce the dimensionality of data while preserving much of the information on three dataset: (a-c) perovskite structure identification, (d-f) band gap classification and (g-i) prediction set with unknown materials, respectively.

* xcai17@fudan.edu.cn

† zhangh@fudan.edu.cn

‡ panjg@shnu.edu.cn