Supporting information: "Data-driven analysis of the electronic-structure factors controlling the work functions of perovskite oxides"

Yihuang Xiong,[a], Weinan Chen,[a] Wenbo Guo,[b] Hua Wei,[b] and Ismaila Dabo[a,c]

[a] Department of Materials Science and Engineering, and Materials Research Institute, The Pennsylvania State University, University Park, PA, USA
[b] College of Information Sciences and Technology, The Pennsylvania State University, University Park, PA 16802, USA
[c] Institutes of Energy and Environment, The Pennsylvania State University, University Park, PA 16802, USA

1 Regression tree

A regression tree is depicted schematically in Fig. S1.

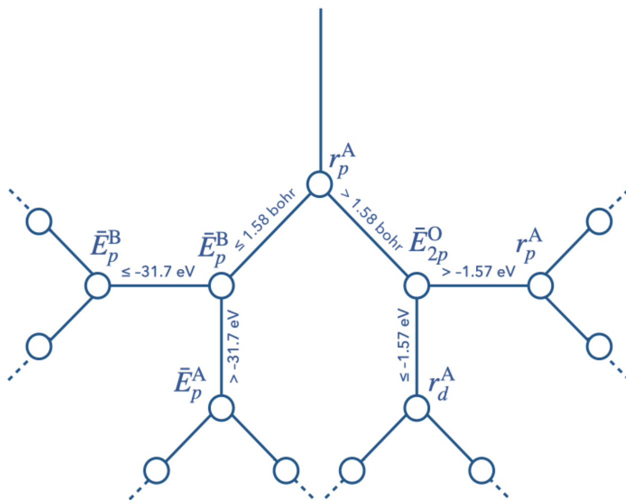

Fig. S1 Decision tree in the random forest model. Following the decision rule shown at each branching, the dataset is split into subsets.

2 Geometric mean of the electronegativity of the perovskite constituents ($\chi_{\mathrm{M}}^{\mathrm{ABO_3}}$) compared to the work function

Various methods have been used to predict the absolute band edge energies, including an empirical method based on the geometric mean of the atomic Mulliken electronegativities (χ_{M}) [1]. As previously pointed out [2,3], this method cannot distinguish between materials that have the same chemical formula but different crystal structures. We plot the work functions with respect to χ_{M} in Fig. S2. It can be seen that for either AO or BO_2 terminations, the correlation is scattered. In fact, χ_{M} is shown to be much less relevant than other features through the recursive reduction process presented in Sec. III of the main text.

3 Feature elimination based on correlation matrix

In this work, we start by using a set of 38 features. As discussed in the main text, using highly correlated features would not deteriorate the performance of random forest regressor; however, features that carry similar information would dilute the importance score, which could lead to the misidentification of the predominant features. To overcome this issue, we first perform a Pearson correlation analysis to remove strongly correlated features. The Pearson correlation coefficient measures the linear relationship between variables. The resulting coefficients lie between –1 to 1, indicating a negative or positive correlation. The Pearson correlation coefficient r_{xy} is calculated as

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}}, \tag{1}$$

where x_i, y_i represent the individual point, and $\bar{x}$, $\bar{y}$ stand for the sample mean of the variables. The computed correlation matrix is shown in Fig. S3.

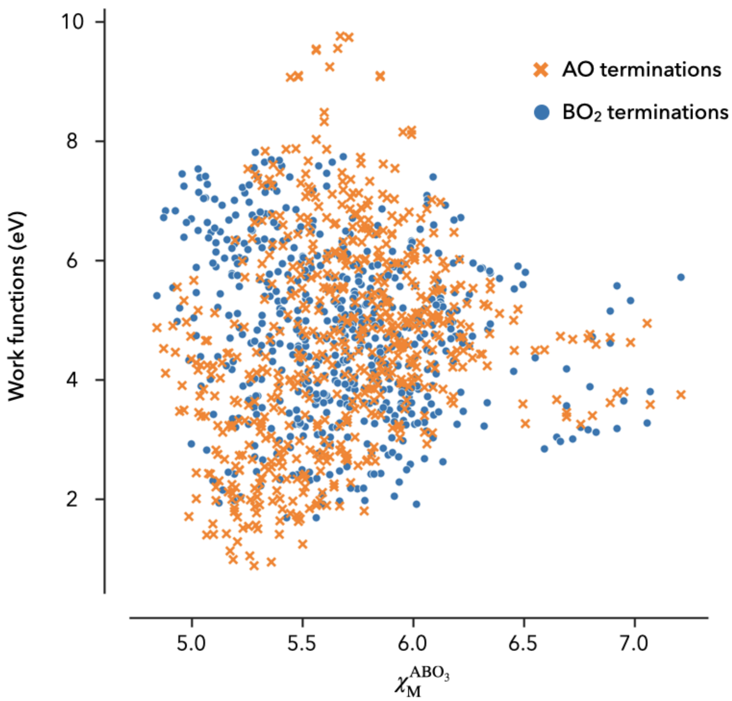

Fig. S2 Work function of AO and BO_2-terminated perovskites as a function of the geometric mean of the atomic electronegativities $\chi_M^{ABO_3}$.

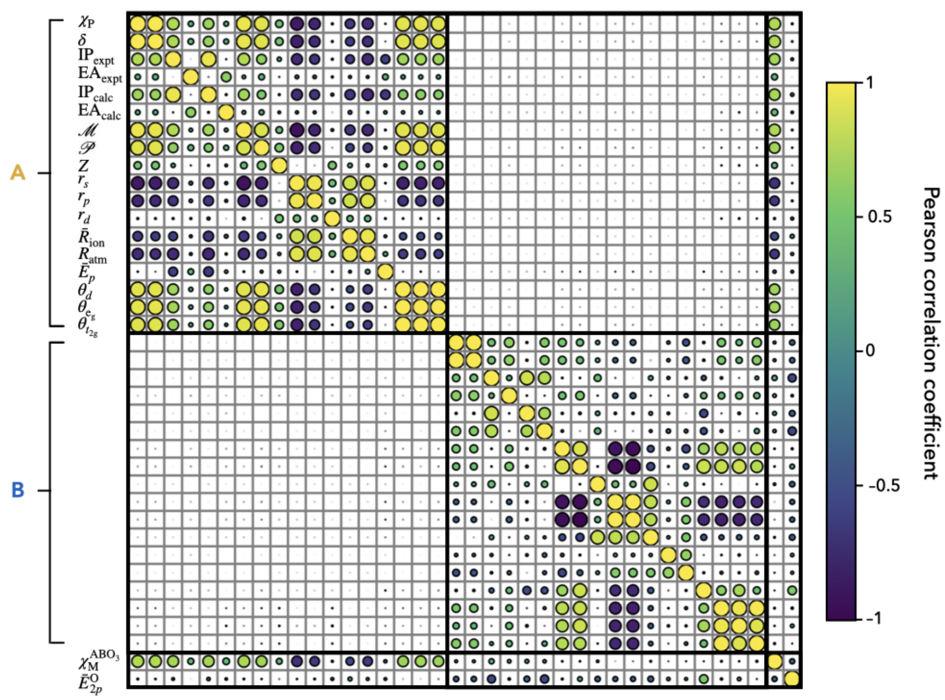

Fig. S3 Pearson matrix of the 38 features. The color indicates whether the correlation is positive or negative. The marker size represents the the magnitude of the correlation. Both the A and B elements have 18 features. The geometric mean of the atomic electronegativities ($\chi_M^{ABO_3}$) and the center of the oxygen $2p$ band ($\bar{E}_{2p}^{O}$) are also included.

Based on the Pearson correlation coefficient, we removed highly-correlated features if its Pearson corelation coefficient is larger than 0.8.This process reduces the number of features from 38 to 21. The features for A are the experimental and calculated electron affinity EA_{expt}^{A}, EA_{calc}^{A}, calculated ionization potential IP_{calc}^{A}, radius of s and d orbital r_s^A and r_d^A, atomic number Z^A, band center of p orbital $\bar{E}_p^A$

and filling factor $\theta^{A}_{t_{2g}}$. The features for B are the experimental and calculated electron affinity EA^{B}_{expt} and EA^{B}_{calc}, calculated ionization potential IP^{B}_{calc}, atomic radii and averaged ionic radii R^{B}_{atm}, $\bar{R}^{B}_{ion}$, radius of d orbital r^{B}_{d}, Pauling's electronegativity χ^{B}_{P}, band center of p orbital $\bar{E}^{B}_{p}$, filling factor of e_g band $\theta^{e_g}_{d}$ and t_{2g} band $\theta^{B}_{t_{2g}}$. Pettifor's chemical scale $\mathscr{P}^{B}$. The last two are the geometric mean of the Mulliken electronegativity of the bulk perovskite $\chi^{ABO_3}_{M}$ and the $2p$ band center of oxygen $\bar{E}^{O}_{2p}$.

4 Hyperparameters tuning and recursive feature elimination

The random forest models are developed using SCIKIT-LEARN library[4]. Four primary hyperparameters are considered, including the number of trees, the maximum depth of the tree, the minimum number of samples required to split an internal node, and the number of features to consider when looking for the best split. To select the hyperparameters that optimize the performance of the random forest regressor, a grid of hyperparameter combinations were generated using GridSearch or RandomizedSearch. In addition, we also perform the recursive feature elimination to optimize the feature space. To identify the most important features, we recursively remove the least important one, as we discussed in the main text. The change of root mean square error (RMSE) with respect to the number of features is shown in Fig. S4(a). Fig. S4(b) shows the normalized importance of all selected features for AO and BO_2 termination. The aforementioned hyperparameter tuning was performed when each feature is removed. This process yields the compatible number of features and hyperparameters. To comprehensively evaluating the model's performance, fivefold cross-validation was applied at each step during the recursive feature eliminations. For AO and BO_2 interfaces, the final selected hyperparameters are shown below:

AO-termination: n_estimators: 130, max_depth = 14, max_features = "sqrt", min_samples_split = 2

BO_2-termination: n_estimators: 100, max_depth = 16, max_features = "log2", min_samples_split = 2

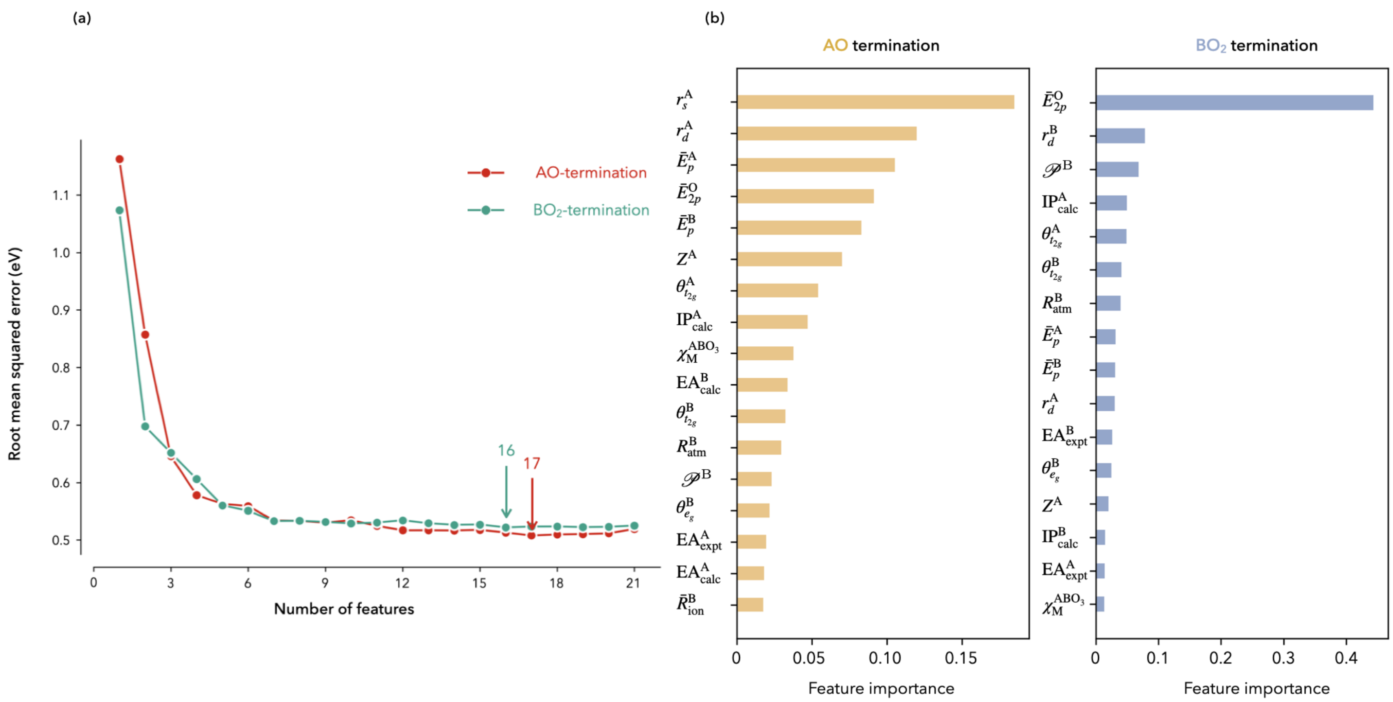

Fig. S4 (a) Averaged RMSE with respect to the number of features for the AO and BO_2 terminations. The results show that the best performances are achieved when 17 and 16 number of features are included in the random forest, respectively. (a) The overall importance ranking of AO and BO_2 terminations.

5 Projected density of states of selected perovskites

We examined the partial dependence of the work functions of perovskites with respect to the bulk oxygen $2p$ band center ($\bar{E}^{O}_{2p}$). We found that the correlation between $\bar{E}^{O}_{2p}$ and the work function is almost linear, except for the ones with deep $\bar{E}^{O}_{2p}$ (< -4 eV). We pay specific attention to the compounds that are close to the plateau of the PDP of $\bar{E}^{O}_{2p}$ and the work functions. Those perovskites turn out to have relatively low work functions, but the decrease of the work function is limited, especially for AO terminations, as shown in Fig. 7(a) in the main text. To understand this trend, we plot the projected density of states of selected perovskites with AO terminations in Fig. S5. We found that for the perovskites with $3d$ transition metals, the $\bar{E}^{O}_{2p}$ energy level is almost constant with respect to vacuum

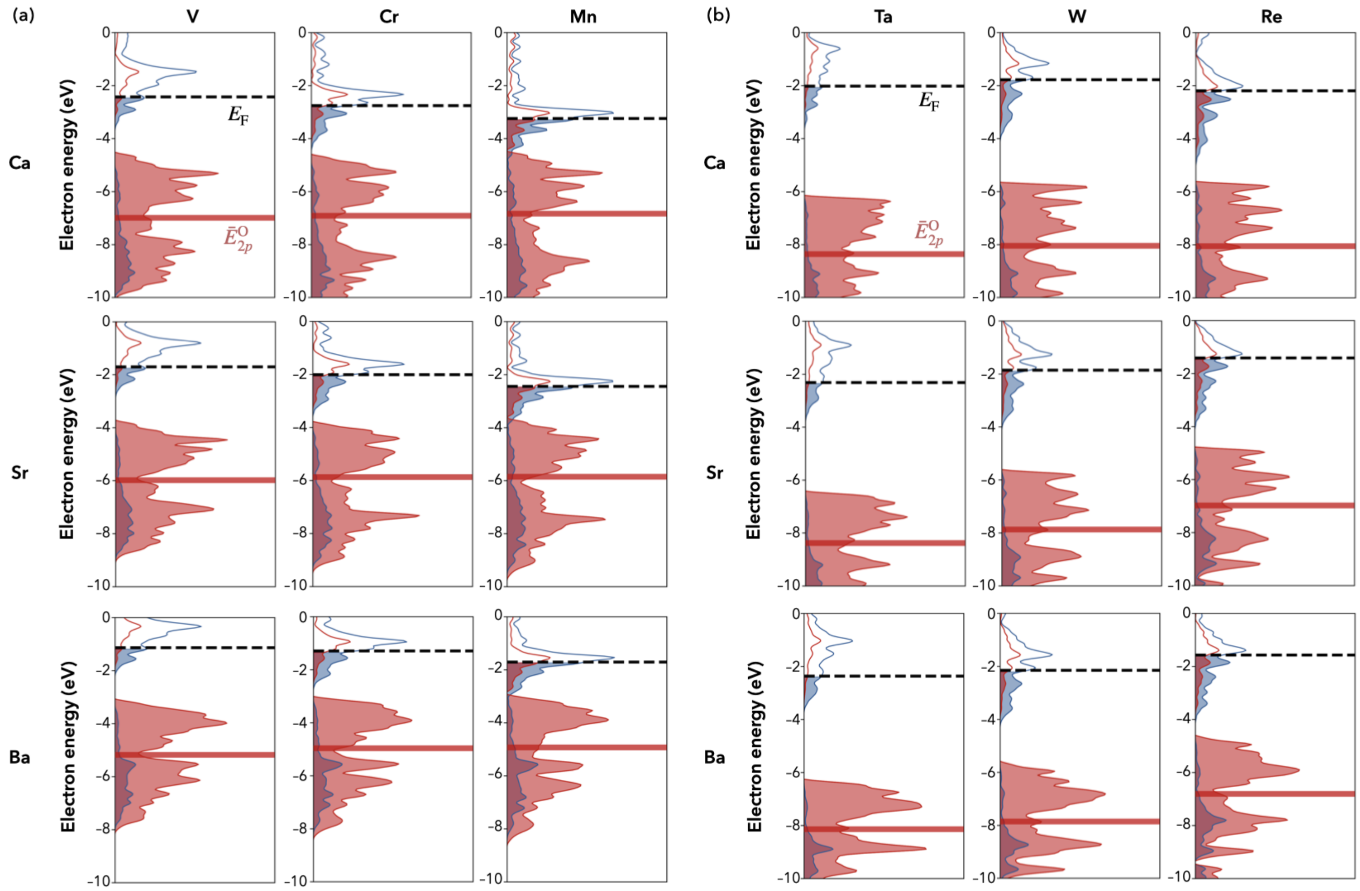

Fig. S5 Projected density of states for perovskite slabs with AO terminations. The representative compositions contains alkaline earth metals (Ca, Sr, Ba) for A site, and (a) 3d (V, Cr, Mn) and 5d (Ta, W, Re) transition metals for B sites. The vacuum energy is referenced to 0, the Fermi energy (E_F) and oxygen 2p band center $\bar{E}^{\mathrm{O}}_{2p}$ are shown with dashed line and red solid line.

(the vacuum energy is referenced to 0 eV). On the other hand, for Ta, W, Re, the hybridization between the d bands and the oxygen 2p bands involves a noticeable shift of the oxygen 2p bands with respect to vacuum. This shift largely cancels out the decrease in $\bar{E}^{\mathrm{O}}_{2p}$ and ultimately alters the correlation between $\bar{E}^{\mathrm{O}}_{2p}$ and the work function.

Notes and references

[1] I. E. Castelli, D. D. Landis, K. S. Thygesen, S. Dahl, I. Chorkendorff, T. F. Jaramillo, and K. W. Jacobsen, Energy Environ. Sci. **5**, 9034 (2012).

[2] V. Stevanović, S. Lany, D. S. Ginley, W. Tumas, and A. Zunger, Phys. Chem. Chem. Phys. **16**, 3706 (2014).

[3] Q. Yan, J. Yu, S. K. Suram, L. Zhou, A. Shinde, P. F. Newhouse, W. Chen, G. Li, K. A. Persson, J. M. Gregoire, and J. B. Neaton, PNAS **114**, 3040 (2017).

[4] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, J. Mach. Learn. Res. **12**, 2825 (2011).