

Designing Two-Dimensional Halide Perovskites Based on High-Throughput Calculations and Machine Learning

Wenguang Hu, Lei Zhang,* and Zheng Pan



Cite This: <https://doi.org/10.1021/acsami.2c00564>



Read Online

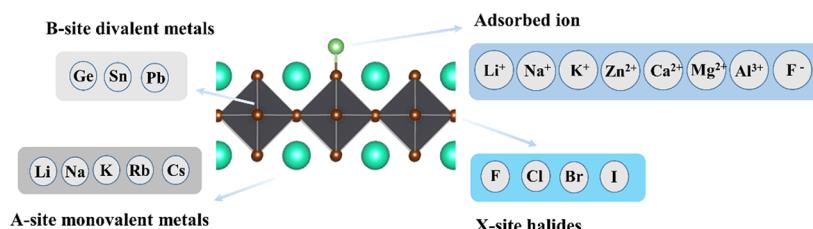
ACCESS |

Metrics & More

Article Recommendations

Supporting Information

DFT + Machine Learning → Optimize Ions + 2D A₂BX₄ Perovskites



ABSTRACT: The interactions between ions and the low-dimensional halide perovskites are critical to realizing the next-generation energy storage devices such as photorechargeable ion batteries and ion capacitors. In this study, we performed high-throughput calculations and machine-learning analysis for ion adsorption on two-dimensional A₂BX₄ halide perovskites. The first-principles calculations obtained an initial data set containing adsorption energies of 640 compositionally engineered ion/perovskite systems with diverse ions including Li⁺, Zn²⁺, K⁺, Na⁺, Al³⁺, Ca²⁺, Mg²⁺, and F⁻. The machine learning algorithms including k-nearest neighbors (KNN), Kriging, Random Forest, Rpart, SVM, and Xgboost algorithms were compared, and the Xgboost algorithm achieved the best accuracy ($r = 0.97$, $R^2 = 0.93$) and was selected to predict the virtual design space consisting of 11 976 ion/perovskite systems. The features were then analyzed and ranked according to their Pearson correlations to the output values. In particular, to better understand the features, diverse feature selection methods were employed to comprehensively evaluate the features. The machine-learning-predicted virtual design space was subsequently screened to select stable lead-free ion/perovskite systems with suitable band gaps and halogen mixing features. The present study provides a theoretical foundation to design halide perovskite materials for ion-based energy storage applications such as secondary ion batteries, ion capacitors, and solar-rechargeable batteries.

KEYWORDS: high throughput, machine learning, feature ranking, 2D halide perovskites, adsorption

INTRODUCTION

Perovskite solar cells have experienced rapid development since their first deployment in 2009 by Miyasaka et al.,¹ and have reached power conversion efficiencies as high as 25.7% in recent years owing to the excellent optoelectronic properties and defect tolerance of the halide perovskite materials. Apart from the solar cells, the halide perovskite materials have been applied for light-emitting diodes, photodetectors, scintillators, and sensors.^{2–6} However, they suffer from the notorious instability and lead contamination issues, which should be addressed for their industrial deployment. For example, the dimensional tailoring strategies have been identified to be efficient to improve the stabilities and optoelectronic properties simultaneously,^{7–11} while various passivating molecules have been proposed to optimize the perovskite surfaces and interfaces.

The interactions between the ions and the halide perovskite materials are critical for the perovskite solar cells because of the hysteresis phenomenon,^{12–14} which is associated with the ion

diffusion that poses serious challenges for the widespread deployment of the perovskite devices. The hysteresis phenomenon is mainly attributed to the facile ion movement in the perovskite bulk crystals and grain boundaries, leading to the ion accumulation at the interface; as a result, the ion movement should be preferably minimized to enhance the perovskite solar cell performance. On the other hand, the interactions between the ions and the halide perovskite materials are beneficial for the energy storage applications such as ion batteries and ion capacitors where the efficient ion movement at the electrode materials surface is constructive. In particular, the halide perovskite materials are classified as

Received: January 10, 2022

Accepted: April 15, 2022

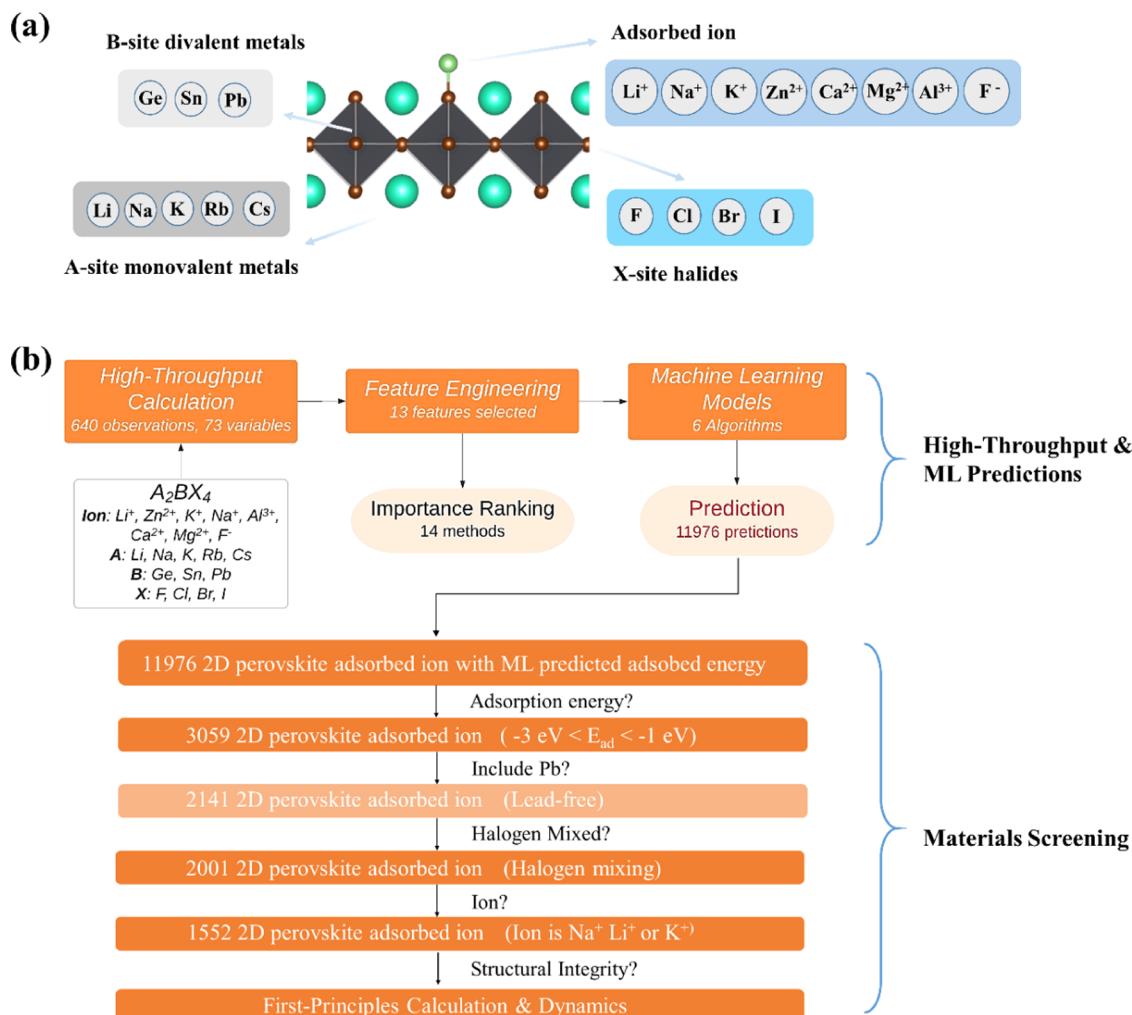


Figure 1. (a) Multidimensional virtual space displaying the ion \cdots A₂BX₄ interactions consists of ions (Li⁺, Zn²⁺, K⁺, Na⁺, Al³⁺, Ca²⁺, Mg²⁺, and F⁻), A-site species (Li, Na, K, Rb, and Cs), B-site species (Ge, Sn, Pb), and X-site species (F, Cl, Br, and I). Both train/test and prediction data sets incorporate the materials modified via the element mixing strategy. (b) Flowchart of the machine learning process for ion adsorption on A₂BX₄. First, high-throughput first-principles calculations are performed to obtain 640 observations in the train/test data set. Second, feature selection is performed to choose 13 relevant features. In addition, the features are ranked according to 14 different ranking methods to evaluate their importance. Third, machine learning models are constructed based on the comparison of six algorithms. Fourth, the Xgboost model corresponding to the best accuracy is employed to make the material predictions. Lastly, the appropriate materials are selected while additional DFT and molecular dynamics simulations are carried out to evaluate the optoelectronic properties and stability of the material candidates.

potential “optoionic materials”^{15–18} that exhibit improved ionic diffusion and charge shuttling effects at the perovskite surfaces upon the visible light irradiation for diverse energy storage and biological applications.

The lithium-ion batteries are considered the primary commercial rechargeable batteries that exhibit high power densities and play pivotal roles in electric vehicles and cell phones. Several alternatives are available to achieve higher energy density, higher specific power, or lower cost, such as those based on zinc, potassium, sodium, magnesium, aluminum, calcium, fluorine, and ammonium ions. These advanced ion batteries require efficient ion adsorption and diffusion process at the surface of the electrode materials. In addition, the metal-ion capacitor is emerging as a new electrochemical energy storage device that combines the merits of battery (high energy density) and capacitor (ultrafast charging). Interestingly, a series of recent researches have reported that optoionic materials, especially halide perovskite materials, are ideal for these applications. The halide

perovskites can be further dimensionally and compositionally engineered for energy storage applications. For example, Ahmad et al. incorporate the two-dimensional halide perovskite into a photorechargeable lithium-ion battery, achieving a capacity of 100 mAh/g, and such a device can be wirelessly charged without the necessity of a fixed charging point.¹⁹ This wireless-charging capability of the optoionic materials is attractive for use in electric vehicles and Internet-of-things (IoT). The optoionic materials are still in the embryonic stage compared with their “optoelectronic” counterparts, and it is of utmost importance to understand the interactions between the ions and the optoionic materials for the energy applications, especially the wirelessly rechargeable devices.

Apart from the traditional experimental works and first-principles calculations that are often associated with “trial-and-error” issues, the data-driven studies combining the high-throughput calculations (or experiments) and the machine learning methods are considered the new paradigm to advance the material research and have been deployed for the rapid

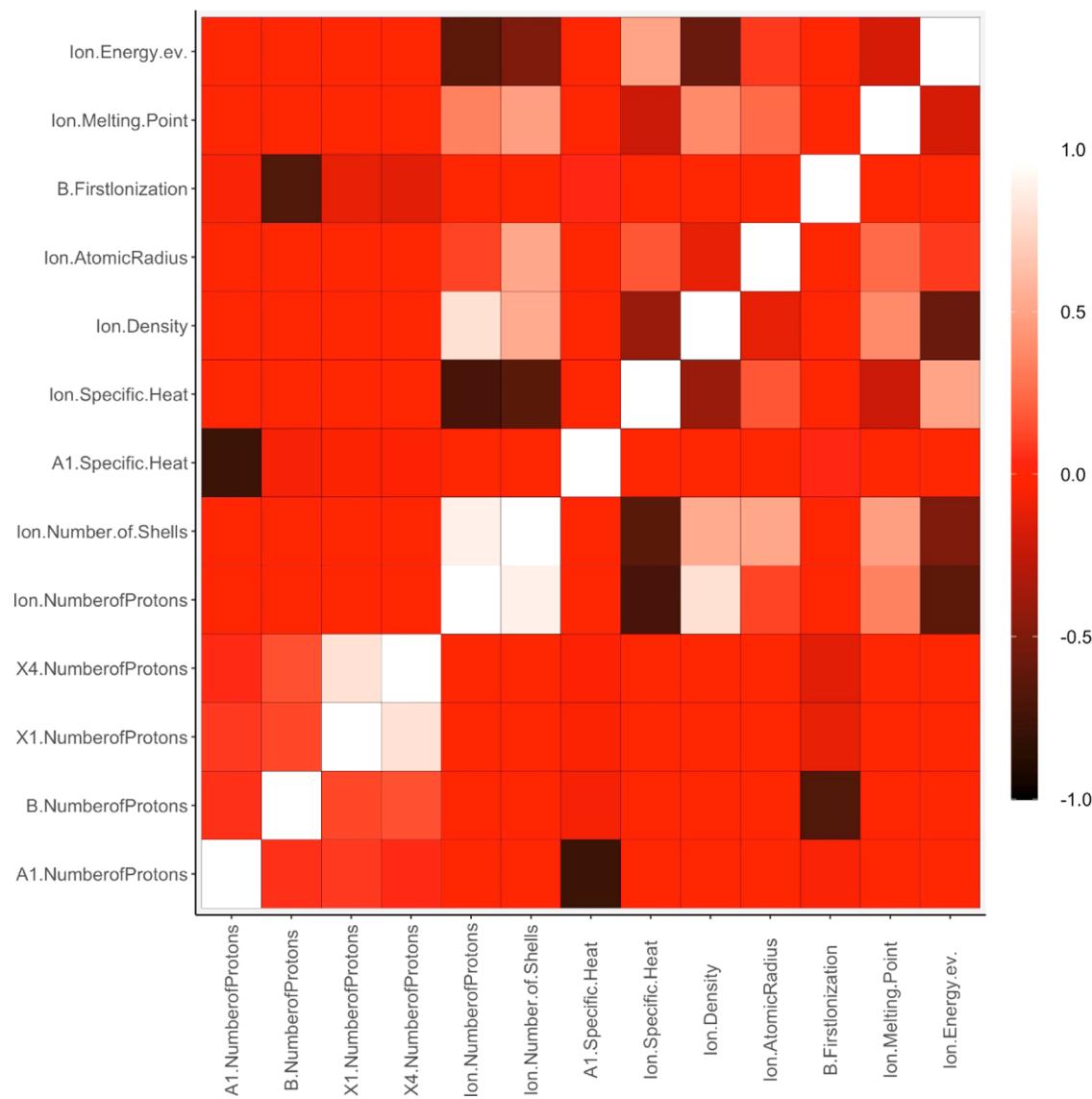


Figure 2. Heat map representing the correlations between 13 remaining features after the feature selection step.

development of new materials for solar cell and batteries. The machine learning methods are effective to predict the virtual space for the materials screening, such that the new candidates can be easily obtained and more scientific information on the ideal candidates can be extracted from the predicted data set.^{20,21}

In this study, we performed density functional theory (DFT) calculations and machine learning investigations to evaluate the interactions between the ions and the two-dimensional halide perovskites toward the energy storage applications; the focus was on adsorption energy and served as the target output. Appropriate features were selected, which were then employed for the construction of the machine learning models. In addition, the features were ranked according to their correlations with the output values using 14 disparate ranking methods that were suggested to provide a more comprehensive understanding of the importance of the features. The accuracies of six machine learning algorithms were compared, and Xgboost achieved the best accuracy and was employed to predict the virtual space. The virtual space consisted of the adsorption energies of various ion/A₂BX₄ systems, with ions

(Li⁺, Zn²⁺, K⁺, Na⁺, Al³⁺, Ca²⁺, Mg²⁺, and F⁻), A-site species (Li, Na, K, Rb, and Cs), B-site species (Ge, Sn, Pb), and X-site species (F, Cl, Br, and I). The proper lead-free perovskite candidates with mixed halogen species are screened from the predicted virtual space and are further evaluated via the first-principles calculations and molecular dynamics to understand their optoelectronic and stability performance. The element mixing strategy was included to prepare the train/test and predict data sets, since the halogen mixing strategy had been reported to be viable to improve the light-harvesting and charge transport properties of the halide perovskite materials. The flowchart of the overall machine learning process is depicted in Figure 1.

METHODS

First-Principles Calculations. The first-principles calculations are performed on the two-dimensional ion/perovskite systems with various ions (Li⁺, Zn²⁺, K⁺, Na⁺, Al³⁺, Ca²⁺, Mg²⁺, and F⁻), A-site species (Li, Na, K, Rb, and Cs), B-site species (Ge, Sn, Pb), and X-site species (F, Cl, Br, and I) to construct the database. The calculations are carried out in CASTEP,²² using Perdew–Burke–Ernzerhof (PBE) functional and 430 eV energy cutoff. The convergence criteria for the

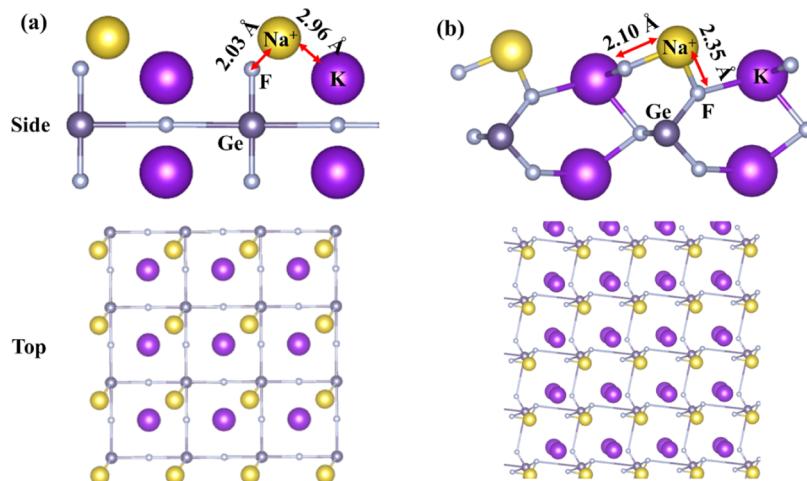


Figure 3. Representative structures of the ion/perovskite systems obtained via high-throughput first-principles calculations. Here, the $\text{Na}^+/\text{K}_2\text{GeF}_4$ system is shown as an example. (a) Initial structure before the geometrical optimization of $\text{Na}^+/\text{K}_2\text{GeF}_4$. (b) $\text{Na}^+/\text{K}_2\text{GeF}_4$ structure after the geometrical optimization, with the two ion–halogen bonds (2.10 and 2.35 Å) highlighted.

geometrical optimization are 10^{-5} eV for the energy, 0.03 eV/Å for the force, and 0.002 Å for the displacement. A vacuum layer of 10 Å is inserted into the unit cell to avoid unnecessary interactions between the neighboring two-dimensional layers. The van der Waals forces are considered using the Tkatchenko–Scheffler (TS) scheme²³ and the properties are calculated using a $4 \times 4 \times 1$ k-point set. To prepare the initial atomic structures for fair comparisons, we place the ion directly on top of the uppermost halogen atoms of the perovskite substrate since such coordination on the dangling bond is suggested to be representative of the low-dimensional perovskite adsorption studies.

The adsorption energy (E_{ad}) is expressed as

$$E_{\text{ad}} = E_{\text{ion+perovskite}} - (E_{\text{ion}} + E_{\text{perovskite}}) \quad (1)$$

where $E_{\text{ion+perovskite}}$ is the total energy of the optimized system consisting of both ion adsorbate and perovskite substrate, $E_{\text{perovskite}}$ is the energy of the A_2BX_4 perovskite structure, and E_{ion} is the energy of the ion adsorbate.

Feature Selection. Pearson correlation coefficients are calculated in the feature engineering step to eliminate the redundant features, which are depicted in the heat map for each feature pair (Figure S1); this demonstrates that a significant number of raw features are highly correlated with each other and some of them should be eliminated. The features with the Pearson correlation coefficients >0.8 or <-0.8 are removed such that only the appropriate one in the feature pair remains for the next round. Consequently, 13 features are selected from the raw features based on the Pearson correlation coefficients (Figure 2).

Hyperparameters. Six machine learning algorithms are compared to construct the models, including KNN, Kriging, Random Forest, Rpart, SVM, and Xgboost. The specific parameters values are determined according to the following procedure (Figure S2). First, $\sim 10\,000$ combinations of the hyperparameter tuples are generated. Second, the RMSE values are calculated for each hyperparameter combination. Third, the optimal hyperparameter tuple is selected according to the smallest RMSE. The following parameters are adopted for the machine learning process. SVM: cost = 2.15, γ = 6.44, degree = 2. KNN: k = 9, distance = 1. Kriging: bias.correct = TRUE, nugget.estim = TRUE, optim.method = BFGS. RPART: minsplit = 10, minbucket = 4, cp (complexity) = 0.001, maxdepth = 15. Xgboost: nrounds = 30, max_depth = 50, tree_method = hist, grow_policy = lossguide, training = TRUE.

Ranking Methods for Feature Importance. For a comprehensive evaluation of the feature importance and to avoid the bias caused by the Pearson method, 14 different ranking methods are employed, including anova, conditional mutual information maximization (cmim), Pearson correlation, diversity-induced self-

representation (disr), find_correlation, information_gain, joint mutual information (jmi), jmim, kruskal_test, mim, mrmr, normalized joint mutual information maximization (njmim), relief, and variance. The explanation of these ranking methods that have been employed in data science for statistical analysis is presented in the Supporting Information.

RESULTS AND DISCUSSION

High-Throughput Calculation. The 640 materials in the train/test data set based on the high-throughput calculation demonstrate that the adsorption energies strongly depend on the types of the ion adsorbates (Figure S3). In general, the ion adsorbates display the bidentate anchoring mode on the perovskite substrate via two ion–halogen bonds (Figure 3). The fluorine ion displays stronger binding strength (i.e., more negative adsorption energy) with the two-dimensional perovskite substrates, which is ascribed to the existence of the F–F halogen bonds and the electrostatic forces between F^- and A^+ . In contrast, Zn^{2+} and Al^{3+} exhibit more positive adsorption energies, with that of the former ranging from 0 to 5 eV and the latter from -2.5 to 7.5 eV. The positive sign indicates that the adsorption is unstable. It should be noted that $|E_{\text{ad}}|$ should not be too large to prevent the structure from disintegrating. In contrast, Li^+ , Na^+ , K^+ , and Ca^{2+} exhibit more negative adsorption energies ranging from -7 to -2 eV. In addition, the adsorption energies depend on the compositions of the halide perovskite substrate; for example, the adsorption energies increase when the B-site element in the halide perovskite substrate change from Ge to Sn to Pb. For the A-site and B-site species, when the elements move down in the periodic table, the distributions of the adsorption energies become narrower. To sum up, the adsorption energies of the 640 two-dimensional ion/perovskite systems in the train set determined by the first-principles calculations are strongly dependent on the types of the ion adsorbates and their distributions are influenced by the perovskite compositions.

Machine Learning Models. Six machine learning algorithms are compared, including KNN, Kriging, Random Forest, Rpart, SVM, and Xgboost (Figure S4). The Xgboost algorithm ($r = 0.968$, $R^2 = 0.93$) achieves the best accuracy (Figure 4) and is consequently selected for material prediction via the preparation of the virtual design space. The Kriging

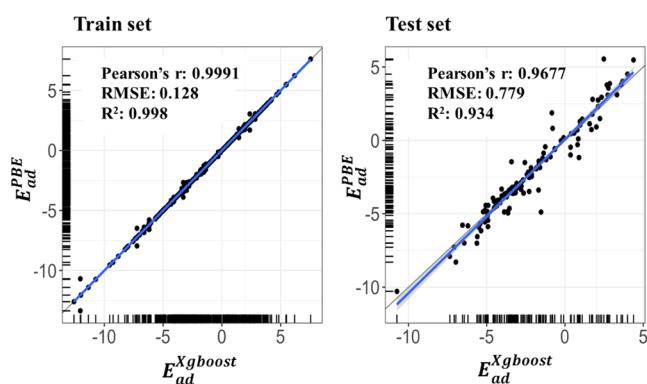


Figure 4. Machine learning model constructed by Xgboost, which offers the best accuracy among the six machine learning models and is selected for the virtual space prediction.

algorithm also exhibits decent performance and achieves a Pearson coefficient r above 0.9.

Feature Ranking. The features are ranked according to their Pearson correlation coefficients with the target variable (adsorption energy) (Figure 5). The adsorption energy has positive correlations with the density of the ion, the melting point of the ion, the number of protons of the ion, the number of shells of the ion, the atomic radius of the ion, and the number of protons of the B-site element. In contrast, it has negative correlations with the first ionization energy of B, the

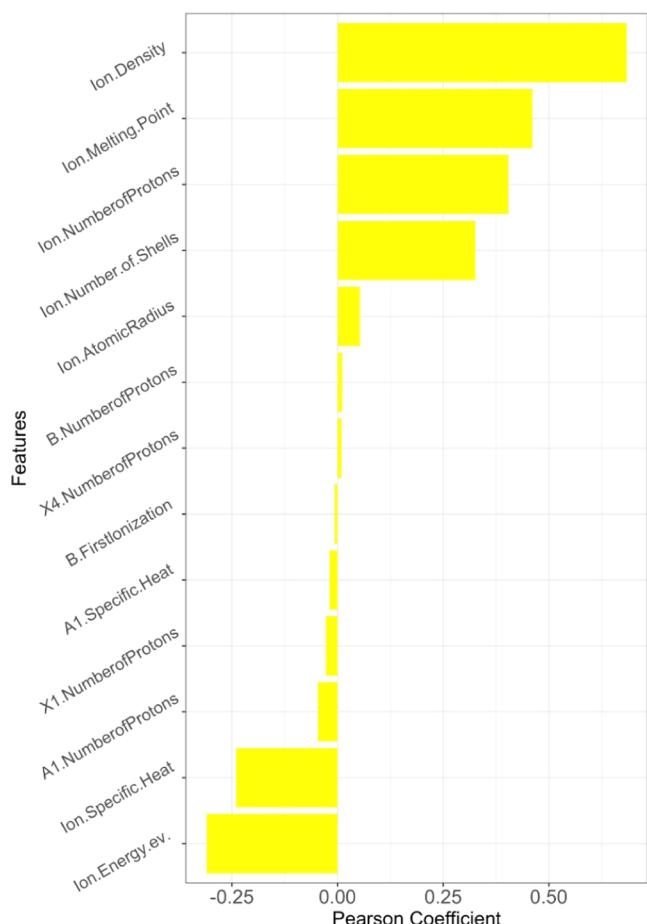


Figure 5. Ranking of the features according to their Pearson correlations with the output values.

specific heat of A, the specific heat of the ion, and the energy of the ion. The densities, melting points, and the number of shells of the ion adsorbates exhibit high rankings, which highlight the predominant contributions by the types of the ion adsorbates. In addition, the statistical ranking methods highlight the importance of ion density in adsorption energy. This is attributed to the high impacts of the mass and volume of the ion adsorbates that influence the adsorbate–adsorbent interactions. Moreover, the ion density is positively correlated with the adsorption energy (larger ion densities lead to higher adsorption energy and equivalently weaker binding); this is ascribed to the fact that the larger ion density signifies the smaller element radius and the stronger localization of the electrons, which weakens the interactions between the ion adsorbates and the perovskite substrate. This corroborates with the observation in Figure S3, where the metals possessing larger ion densities (such as Zn^{2+} , Al^{3+} , and Mg^{2+}) than lithium exhibit larger adsorption energy and weaker binding strength. The importance of the 13 features follows the order: Ion.Density > Ion.Melting.Point > Ion.NumberofProtons > Ion.Number.of.Shells. > Ion.Energy > Ion.Specific.Heat > Ion.AtomicRadius > A1.NumberofProtons > X1.NumberofProtons > A1.Specific.Heat > B.NumberofProtons > X4.NumberofProtons > B.FirstIonization.

Effects of Ranking Methods. In the literature, the features are commonly ranked according to their Pearson correlation coefficients with the outputs; however, this ranking method can sometimes introduce bias as the Pearson coefficients neglect several statistical attributes such as context, redundancy, and complementariness.^{24,25} As a result, in this study, the importance of the features is ranked according to various statistical ranking methods, including anova, cmim, correlation, disr, find_correlation, information_gain, jmi, jmim, kruskal_test, mim, mrmr, njmim, relief, and variance (Figure 6 and Table S1). These methods are proposed to comprehensively evaluate the feature importance, while the Pearson correlation coefficient alone is insufficient. The ion density ranks first for most ranking methods such as anova, cmim, Pearson correlation, and njmi despite the low ranking for the krusk method. The krusk method is inconsistent with most ranking methods and is thus discouraged in the feature ranking process. In addition, the atomic radius of the ion ranks equally high for the majority of the ranking methods, especially for cmim, disr, find_correlation, jmi, jmim, mim, mrmr, and njmi where the atomic radius of the ion ranks the first. In contrast, the first ionization of the B species ranks low for almost all of the ranking methods. As a result, the ranking methods demonstrate slight deviations in the feature importance, which should be addressed in future material feature studies to systematically evaluate the feature importance. To sum up, the different ranking methods consistently suggest the importance of the density of the ion adsorbate for the adsorption energies of the hybrid systems, while the importance of the atomic radius is insufficiently addressed by the Pearson method.

Visualization of the Virtual Design Space. The Xgboost algorithm is employed to prepare the prediction data set consisting of 11976 predicted ion/perovskite systems, which are visualized based on the features (Figures S5 and S6). The new materials in the virtual space should exhibit appropriate adsorption energies to simultaneously satisfy the structural integrity and the strong binding criteria. For example, the absolute values of the adsorption energy should be neither too

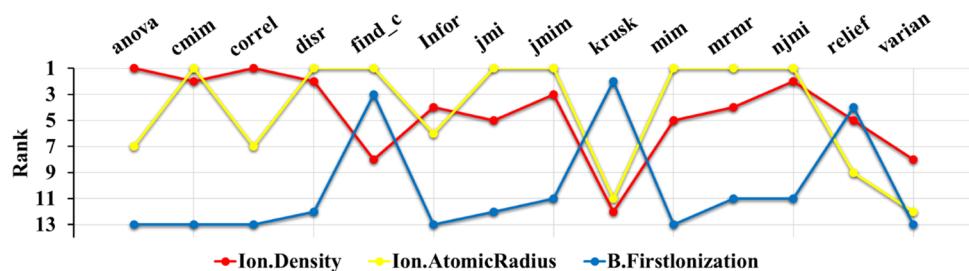


Figure 6. Importance ranking of three selected features (Ion.Density, Ion.AtomicRadius, and B.FirstIonization) according to 14 different ranking methods including anova, cmim, correlation, disr, find_correlation, information_gain, jmi, jmim, kruskal_test, mim, mrmr, njmim, relief, and variance.

Table 1. Screened Ion/Perovskite Candidates ($\text{Li}^+/\text{K}_2\text{GeClBr}_3$, $\text{K}^+/\text{Rb}_2\text{SnFBr}_3$, $\text{K}^+/\text{Rb}_2\text{SnBr}_3\text{I}$, $\text{Na}^+/\text{Cs}_2\text{SnFCl}_2\text{I}$, and $\text{Na}^+/\text{LiCsSnFBr}_2\text{I}$) with Their Feature Inputs^a

type	NP _{A1}	NP _B	NP _{X1}	NP _{X4}	NP _I	NS _I	SH _{A1}	SH _I	D _I	AR _I	FI _B	MP _I	E _I
$\text{K}_2\text{GeClBr}_3$ and Li^+	19	32	18	35	3	2	0.76	3.58	0.534	2.1	7.90	453.9	-200.05
$\text{Rb}_2\text{SnFBr}_3$ and K^+	37	50	9	35	19	4	0.36	0.76	0.862	2.8	7.34	336.5	-790.35
$\text{Rb}_2\text{SnBr}_3\text{I}$ and K^+	37	50	35	53	19	4	0.36	0.76	0.862	2.8	7.34	336.5	-790.35
$\text{Cs}_2\text{SnFCl}_2\text{I}$ and Na^+	55	50	9	53	11	3	0.24	1.23	0.971	2.2	7.34	371.2	-1304.18
$\text{LiCsSnFBr}_2\text{I}$ and Na^+	3	82	9	35	11	3	3.58	1.23	0.971	2.2	7.42	371.2	-1304.18

^aNP_{A1} is the A1 number of protons. NP_B is the B number of protons. NP_{X1} is the X1 number of protons. NP_{X4} is the X4 number of protons. NP_I is the ion number of protons. NS_I is the ion number of shells. SH_{A1} is the A1 specific heat. SH_I is the ion specific heat. D_I is the ion density. AR_I is the ion atomic radius. FI_B is the B first ionization. MP_I is the ion melting point. E_I is the ion energy.

large nor too small; the appropriate range of the adsorption energies is proposed to be $-5 \text{ eV} < E_{\text{ad}} < 0 \text{ eV}$. In general, the adsorption energies strongly depend on the ion density, the ion melting point, and the ion atomic radius, while the adsorption energies are positively correlated with ion density.

Materials Screening and First-Principles Evaluation.

The 11976 predicted ion/perovskite systems are further screened to obtain the potential two-dimensional perovskite candidates for ion-based semiconductor devices. First, the proposed appropriate energy range is proposed to be $-5 \text{ eV} < E_{\text{ad}} < 0 \text{ eV}$, which rules out the unstable hybrid systems. Second, the lead-containing halide perovskite systems are eliminated, which addresses the lead contamination issue. Third, the samples with halogen mixing are selected for the next round, which is suggested to enhance the optoelectronic properties and the interfacial interactions.^{26,27} In addition, the perovskite candidates containing the lithium, sodium, and potassium ions are temporarily focused since these types of ions are predominantly employed for ion battery applications. The structural integrities are visually checked to eliminate the ion/perovskite structures with unacceptable atomic drifting that signifies disintegration. As a result, five exemplar ion/perovskite systems in the prediction set, $\text{Li}^+/\text{K}_2\text{GeClBr}_3$, $\text{K}^+/\text{Rb}_2\text{SnFBr}_3$, $\text{K}^+/\text{Rb}_2\text{SnBr}_3\text{I}$, $\text{Na}^+/\text{Cs}_2\text{SnFCl}_2\text{I}$, and $\text{Na}^+/\text{LiCsSnFBr}_2\text{I}$ (Table 1), are proposed as the potential candidates for ion capacitors and are further evaluated via first-principles calculations to predict their optoelectronic and stability performances. The adsorption energies calculated by the PBE functional and predicted by the Xgboost model are consistent with each other (Figure 7): $\text{K}^+/\text{Rb}_2\text{Sn}_3\text{I}$ exhibits the PBE-based adsorption energy of -2.93 eV and the machine-learning-based adsorption energy of -2.98 eV ; $\text{Na}^+/\text{Cs}_2\text{SnFCl}_2\text{I}$ exhibits the PBE-based adsorption energy of -1.37 eV and the machine-learning-based adsorption energy of -1.08 eV ; $\text{Na}^+/\text{LiCsSnF}_2\text{I}$ exhibits the PBE-based adsorption energy of -1.97 eV and the machine-learning-based adsorption energy of -1.79 eV .

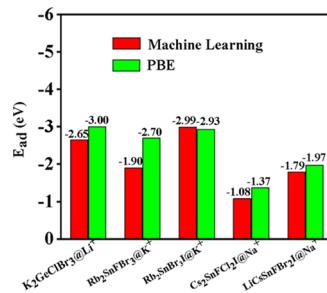


Figure 7. Comparison of adsorption energies of $\text{Li}^+/\text{K}_2\text{GeClBr}_3$, $\text{K}^+/\text{Rb}_2\text{SnFBr}_3$, $\text{K}^+/\text{Rb}_2\text{SnBr}_3\text{I}$, $\text{Na}^+/\text{Cs}_2\text{SnFCl}_2\text{I}$, and $\text{Na}^+/\text{LiCsSnFBr}_2\text{I}$ between the machine-learning-predicted results and the first-principles calculation results.

The electronic, optical, and dynamic properties of $\text{K}^+/\text{Rb}_2\text{SnBr}_3\text{I}$, $\text{Na}^+/\text{Cs}_2\text{SnFCl}_2\text{I}$, and $\text{Na}^+/\text{LiCsSnF}_2\text{I}$ are obtained at the PBE level to evaluate their optoelectronic and stability performances. The band gaps of $\text{K}^+/\text{Rb}_2\text{SnBr}_3\text{I}$, $\text{Na}^+/\text{Cs}_2\text{SnFCl}_2\text{I}$, and $\text{Na}^+/\text{LiCsSnF}_2\text{I}$ are 1.69, 1.88, and 1.39 eV, respectively. The ion adsorbates exhibit distinctive coordination in nearby halogen species; for example, the sodium ion in $\text{Na}^+/\text{Cs}_2\text{SnFCl}_2\text{I}$ exhibits a $\text{Na}\cdots\text{F}$ distance of 2.09 Å and the potassium ion in $\text{K}^+/\text{Rb}_2\text{SnBr}_3\text{I}$ exhibits a $\text{K}\cdots\text{I}$ distance of 3.49 Å (Figure 8). The projected density of states (PDOS) spectra suggest that the ion adsorption introduces shallow defect states in these systems since no additional states are present in the original band gap when the ion adsorbates are available; this indicates minimized charge recombination during the charge transport stage in a photorechargeable energy storage device. The valence bands are mainly contributed by the p-orbitals of the halogen atoms while the conduction bands are mainly from the B-site species and the ion adsorbates, which are consistent with the findings in the perovskite literature. The molecular dynamic calculations (Figure 9) demonstrate that $\text{K}^+/\text{Rb}_2\text{SnBr}_3\text{I}$, $\text{Na}^+/\text{Cs}_2\text{SnFCl}_2\text{I}$,

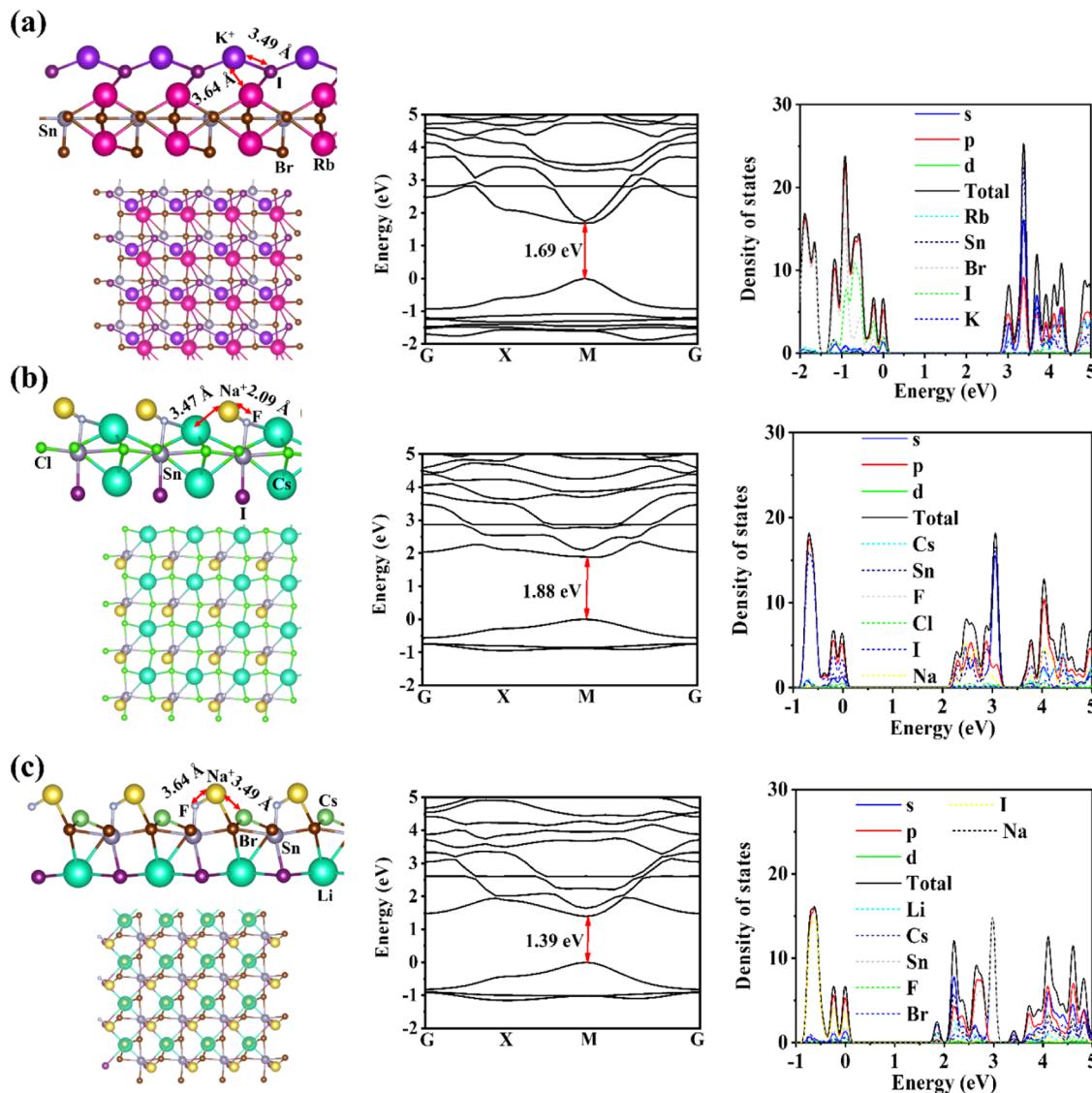


Figure 8. Optimized atomic structures (top and side view), band structures, and PDOS spectra of the exemplar screened perovskite candidates: (a) $\text{K}^+/\text{Rb}_2\text{SnBr}_3\text{I}$, (b) $\text{Na}^+/\text{Cs}_2\text{SnFCl}_2\text{I}$, and (c) $\text{Na}^+/\text{LiCsSnF}_2\text{I}$.

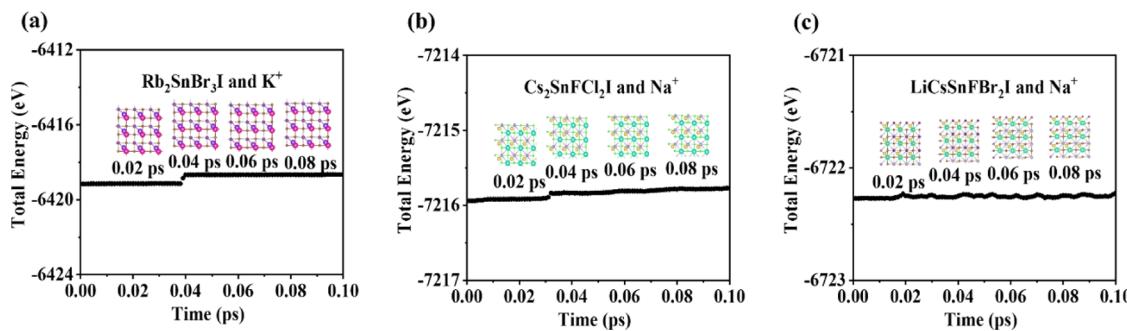


Figure 9. Molecular dynamics of the selected ion/perovskite systems with NVT ensemble at 298 K: $\text{K}^+/\text{Rb}_2\text{SnBr}_3\text{I}$ (a), $\text{Na}^+/\text{Cs}_2\text{SnFCl}_2\text{I}$ (b), and $\text{Na}^+/\text{LiCsSnF}_2\text{I}$ (c).

and $\text{Na}^+/\text{LiCsSnF}_2\text{I}$ exhibit stable structures at the room temperature. $\text{K}^+/\text{Rb}_2\text{SnBr}_3\text{I}$ and $\text{Na}^+/\text{Cs}_2\text{SnFCl}_2\text{I}$ display slight energy increase at 0.04 and 0.03 ps, while the energy of $\text{Na}^+/\text{LiCsSnF}_2\text{I}$ slightly fluctuates during the simulation time. Nevertheless, the fluctuation is small and the ion/perovskite structures are suggested to be thermodynamically stable. The

typical ion migration pathways and the corresponding diffusion barriers on the two-dimensional halide perovskite surface determined by the DFT calculations are provided to further examine the machine-learning-predicted materials (Figure S7). For the $\text{K}^+/\text{Rb}_2\text{SnBr}_3\text{I}$ system, the first-principles calculation demonstrates that the diffusion barrier depends on the ion

diffusion direction; in this case, the energy barrier maximum is 0.211 eV, while the diffusion barrier is smaller in the reverse direction. As a result, the solar energy is suggested to be sufficient to activate the efficient diffusion of the charge carriers that is photo-excited from the valence band to the conduction band in the two-dimensional halide perovskite system. Similarly, the diffusion of the Na^+ in the $\text{Na}^+/\text{Cs}_2\text{SnFCl}_2\text{I}$ system is efficient and the diffusion barriers are 0.151 and 0.027 eV in the representative pathways; the diffusion of the Na^+ in the $\text{Na}^+/\text{LiCsSnF}_2\text{I}$ system is also efficient and exhibits diffusion barriers of 0.129 and 0.014 eV, respectively. In conclusion, the ion diffusion on the halide perovskite materials is efficient, and these machine-learning-predicted samples are suggested to be applicable for ion-based energy storage systems.

CONCLUSIONS

The high-throughput calculations, machine learning, and feature ranking were performed to explore the adsorption of different ions on two-dimensional halide perovskite systems for their applications in ion-based energy devices. The adsorption energies strongly depended on the types of the ion adsorbates and the constituent species in the perovskite substrate. To comprehensively understand the feature importance, 14 different ranking methods were presented to help the feature evaluation, which was more statistically robust than the Pearson correlation ranking method. A database consisting of 11 976 two-dimensional ion/perovskite systems was predicted via the Xgboost algorithm, after which the materials were screened according to the lead-free and halogen mixing criteria. The selected materials candidates were further evaluated via first-principles and molecular dynamic calculations to explain their optoelectronic properties and stabilities. The present machine-learning-based study provides a theoretical platform for the development of the two-dimensional halide perovskite materials for solar-rechargeable ion capacitors and perovskite-based ion batteries.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at <https://pubs.acs.org/doi/10.1021/acsami.2c00564>.

Descriptions of ranking methods; heap map of the Pearson correlation coefficients; process to generate hyperparameter tuples and comparison of performances of different machine learning models; visualization of the adsorption energy distributions; fitting results of six machine learning models; a complete list of the ranking of 13 features; visualization of the prediction data set; visualization of the adsorption energy distributions versus the features; typical surface diffusion path and energy barrier for the ion diffusion of the screened candidates ([PDF](#))

AUTHOR INFORMATION

Corresponding Author

Lei Zhang – Institute of Advanced Materials and Flexible Electronics (IAMFE), School of Chemistry and Materials Science, Nanjing University of Information Science & Technology, Nanjing 210044, China; Department of Applied Physics, School of Physics and Optoelectronic Engineering, Nanjing University of Information Science & Technology,

Nanjing 210044, China;  orcid.org/0000-0001-6873-7314; Email: 002699@nuist.edu.cn

Authors

Wenguang Hu – Institute of Advanced Materials and Flexible Electronics (IAMFE), School of Chemistry and Materials Science, Nanjing University of Information Science & Technology, Nanjing 210044, China; Department of Applied Physics, School of Physics and Optoelectronic Engineering, Nanjing University of Information Science & Technology, Nanjing 210044, China

Zheng Pan – Institute of Advanced Materials and Flexible Electronics (IAMFE), School of Chemistry and Materials Science, Nanjing University of Information Science & Technology, Nanjing 210044, China; Department of Applied Physics, School of Physics and Optoelectronic Engineering, Nanjing University of Information Science & Technology, Nanjing 210044, China

Complete contact information is available at:
<https://pubs.acs.org/10.1021/acsami.2c00564>

Author Contributions

The author has given approval to the final version of the manuscript.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

This work was supported by the Jiangsu Provincial Science and Technology Project Basic Research Program (Natural Science Foundation of Jiangsu Province) (No. BK20211283). The authors acknowledge computational support from the NSCCSZ, Shenzhen, China.

REFERENCES

- (1) Kojima, A.; Teshima, K.; Shirai, Y.; Miyasaka, T. Organometal Halide Perovskites as Visible-Light Sensitizers for Photovoltaic Cells. *J. Am. Chem. Soc.* **2009**, *131*, 6050–6051.
- (2) Xie, C.; Zhao, X.; Ong, E. W. Y.; Tan, Z.-K. Transparent Near-Infrared Perovskite Light-Emitting Diodes. *Nat. Commun.* **2020**, *11*, No. 4213.
- (3) Kim, Y.-H.; Kim, S.; Kakekhani, A.; Park, J.; Park, J.; Lee, Y.-H.; Xu, H.; Nagane, S.; Wexler, R. B.; Kim, D.-H.; et al. Comprehensive Defect Suppression in Perovskite Nanocrystals for High-Efficiency Light-Emitting Diodes. *Nat. Photonics* **2021**, *15*, 148–155.
- (4) Gandini, M.; Villa, I.; Beretta, M.; Gotti, C.; Imran, M.; Carulli, F.; Fantuzzi, E.; Sassi, M.; Zaffalon, M.; Brofferio, C.; et al. Efficient, Fast and Reabsorption-Free Perovskite Nanocrystal-Based Sensitized Plastic Scintillators. *Nat. Nanotechnol.* **2020**, *15*, 462–468.
- (5) Xie, A.; Hettiarachchi, C.; Maddalena, F.; Witkowski, M. E.; Makowski, M.; Drozdowski, W.; Arramel, A.; Wee, A. T. S.; Springham, S. V.; Vuong, P. Q.; et al. Lithium-Doped Two-Dimensional Perovskite Scintillator for Wide-Range Radiation Detection. *Commun. Mater.* **2020**, *1*, No. 37.
- (6) Yang, S.; Xu, Z.; Xue, S.; Kandlakunta, P.; Cao, L.; Huang, J. Organohalide Lead Perovskites: More Stable than Glass under Gamma-Ray Radiation. *Adv. Mater.* **2019**, *31*, No. 1805547.
- (7) Gao, Y.; Wei, Z.; Hsu, S.-N.; Boudouris, B. W.; Dou, L. Two-Dimensional Halide Perovskites Featuring Semiconducting Organic Building Blocks. *Mater. Chem. Front.* **2020**, *4*, 3400–3418.
- (8) Yu, Y.; Zhang, D.; Yang, P. Ruddlesden-Popper Phase in Two-Dimensional Inorganic Halide Perovskites: A Plausible Model and the Supporting Observations. *Nano Lett.* **2017**, *17*, 5489–5494.

- (9) Yin, J.; Brédas, J.-L.; Bakr, O. M.; Mohammed, O. F. Boosting Self-Trapped Emissions in Zero-Dimensional Perovskite Heterostructures. *Chem. Mater.* **2020**, *32*, 5036–5043.
- (10) Romani, L.; Bala, A.; Kumar, V.; Speltini, A.; Milella, A.; Fracassi, F.; Listorti, A.; Profumo, A.; Malavasi, L. PEA₂SnBr₄: A Water-Stable Lead-Free Two-Dimensional Perovskite and Demonstration of Its Use as a Co-Catalyst in Hydrogen Photogeneration and Organic-Dye Degradation. *J. Mater. Chem. C* **2020**, *8*, 9189–9194.
- (11) Shi, R.; Zhang, Z.; Fang, W.-H.; Long, R. Charge Localization Control of Electron–Hole Recombination in Multilayer Two-Dimensional Dion–Jacobson Hybrid Perovskites. *J. Mater. Chem. A* **2020**, *8*, 9168–9176.
- (12) Zhong, Y.; Hufnagel, M.; Thelakkat, M.; Li, C.; Huettner, S. Role of PCBM in the Suppression of Hysteresis in Perovskite Solar Cells. *Adv. Funct. Mater.* **2020**, *30*, No. 1908920.
- (13) Senanayak, S. P.; Abdi-Jalebi, M.; Kamboj, V. S.; Carey, R.; Shivanna, R.; Tian, T.; Schweicher, G.; Wang, J.; Giesbrecht, N.; Di Nuzzo, D.; et al. A General Approach for Hysteresis-Free, Operationally Stable Metal Halide Perovskite Field-Effect Transistors. *Sci. Adv.* **2020**, *6*, No. eaaz4948.
- (14) Jiang, J.; Zou, X.; Lv, Y.; Liu, Y.; Xu, W.; Tao, Q.; Chai, Y.; Liao, L. Rational Design of Al₂O₃/2D Perovskite Heterostructure Dielectric for High Performance MoS₂ Phototransistors. *Nat. Commun.* **2020**, *11*, No. 4266.
- (15) Cai, J.; Ma, W.; Hao, C.; Sun, M.; Guo, J.; Xu, L.; Xu, C.; Kuang, H. Artificial Light-Triggered Smart Nanochannels Relying on Optoionic Effects. *Chem* **2021**, *7*, 1802–1826.
- (16) Podjaski, F.; Lotsch, B. V. Optoelectronics Meets Optoionics: Light Storing Carbon Nitrides and Beyond. *Adv. Energy Mater.* **2021**, *11*, No. 2003049.
- (17) Senocrate, A.; Kotomin, E.; Maier, J. On the Way to Optoionics. *Helv. Chim. Acta* **2020**, *103*, No. e2000073.
- (18) He, M.; Zhang, L. Prediction of Solar-chargeable Battery Materials: A Text-mining and First-principles Investigation. *Int. J. Energy Res.* **2021**, *45*, 15521–15533.
- (19) Ahmad, S.; George, C.; Beesley, D. J.; Baumberg, J. J.; De Volder, M. Photo-Rechargeable Organo-Halide Perovskite Batteries. *Nano Lett.* **2018**, *18*, 1856–1862.
- (20) Liu, Y.; Guo, B.; Zou, X.; Li, Y.; Shi, S. Machine Learning Assisted Materials Design and Discovery for Rechargeable Batteries. *Energy Storage Mater.* **2020**, *31*, 434–450.
- (21) Liu, Y.; Wu, J. M.; Avdeev, M.; Shi, S. Q. Multi-Layer Feature Selection Incorporating Weighted Score-Based Expert Knowledge toward Modeling Materials with Targeted Properties. *Adv. Theory Simul.* **2020**, *3*, No. 1900215.
- (22) Segall, M. D.; Lindan, P. J. D.; Probert, M. J.; Pickard, C. J.; Hasnip, P. J.; Clark, S. J.; Payne, M. C. First-Principles Simulation: Ideas, Illustrations and the CASTEP Code. *J. Phys.: Condens. Matter* **2002**, *14*, 2717–2744.
- (23) Tkatchenko, A.; Scheffler, M. Accurate Molecular Van Der Waals Interactions from Ground-State Electron Density and Free-Atom Reference Data. *Phys. Rev. Lett.* **2009**, *102*, No. 073005.
- (24) Ahlgren, P.; Jarneving, B.; Rousseau, R. Requirements for a Cocitation Similarity Measure, with Special Reference to Pearson's Correlation Coefficient. *J. Am. Soc. Inf. Sci. Technol.* **2003**, *54*, 550–560.
- (25) Paruolo, P.; Saisana, M.; Saltelli, A. Ratings and Rankings: Voodoo or Science? *J. R. Stat. Soc. Ser. A* **2013**, *176*, 609–634.
- (26) Yin, W.-J.; Yan, Y.; Wei, S. Anomalous Alloy Properties in Mixed Halide Perovskites. *J. Phys. Chem. Lett.* **2014**, *5*, 3625–3631.
- (27) Jesper Jacobsson, T.; Correa-Baena, J.-P.; Pazoki, M.; Saliba, M.; Schenk, K.; Grätzel, M.; Hagfeldt, A. Exploration of the Compositional Space for Mixed Lead Halogen Perovskites for High Efficiency Solar Cells. *Energy Environ. Sci.* **2016**, *9*, 1706–1724.

Design of two-dimensional halide perovskites based on high-throughput calculations and machine learning

*Wenguang Hu,^{a**ab**} Lei Zhang^{**ab***} and Zheng Pan^{**ab**}*

^a *Institute of Advanced Materials and Flexible Electronics (IAMFE), School of Chemistry and Materials Science, Nanjing University of Information Science & Technology, 210044, Nanjing, China.*

Email: 002699@nuist.edu.cn

^b *Department of Applied Physics, School of Physics and Optoelectronic Engineering, Nanjing University of Information Science & Technology, 210044, Nanjing, China.*

KEYWORDS High-Throughput; Machine Learning; Feature Ranking; 2D Halide Perovskites; Adsorption

Supporting Information

Descriptions of Ranking Methods:

The anova method is a type of hypothesis testing which is used to find out the experimental results by analyzing the variance of the different survey groups. The cmim method iteratively selects features by maximizing mutual information with a target variable conditioned on previously selected features. the disr approach considers both the representativeness and diversity properties of features.

Find_correlation analyzes a correlation matrix of the data attributes that report on attributes that can be removed. Information gain evaluates the gain of each variable in the context of the target variable. Jmi is adopted to measure the joint redundant and complementary effect of features in feature selection methods. Jmim is a feature selection technique using joint mutual information along with the maximum of minimum approach.¹ Kruskal_test is a nonparametric test that is referred to as one-way anova on ranks and is an extension of the Man-Whitney test to situations where more than two levels/populations are involved. The mim method adopts mutual information to measure each feature's

relevancy to the class label and does not consider redundancy and complementariness among features.

The mRMR approach selects features with a high correlation with the class (output) and a low correlation between themselves. The njmim approach is based on the maximum of the minimum criterion. Relief takes a filter-method approach to feature selection that is notably sensitive to feature interactions.² The variance threshold is a simple baseline approach to feature selection that removes all zero-variance features, i.e., the features that have the same value in all samples.

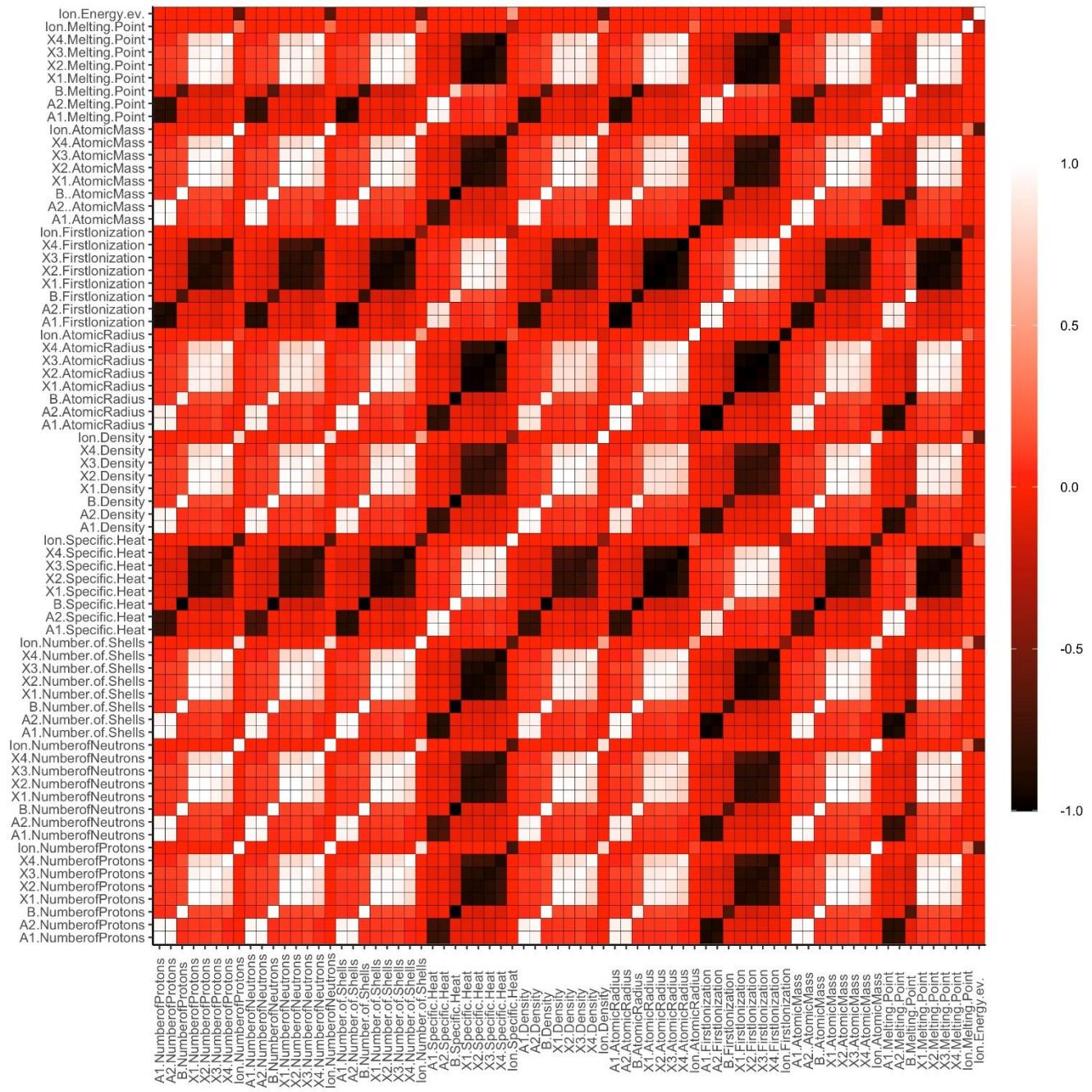


Figure S1. Heat map of the Pearson correlation coefficients between the raw features.

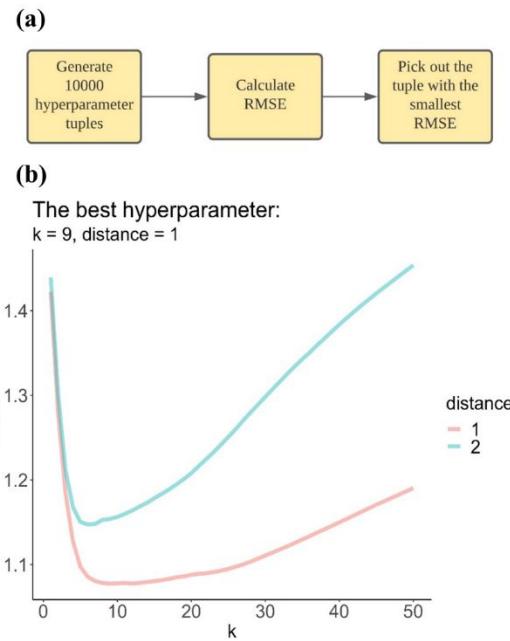


Figure S2. (a) Process to generate hyperparameter tuples. (b) Comparison of performances of different machine learning models using different hyperparameter tuples for the KNN model; the minkowski distance and k are shown as an example.

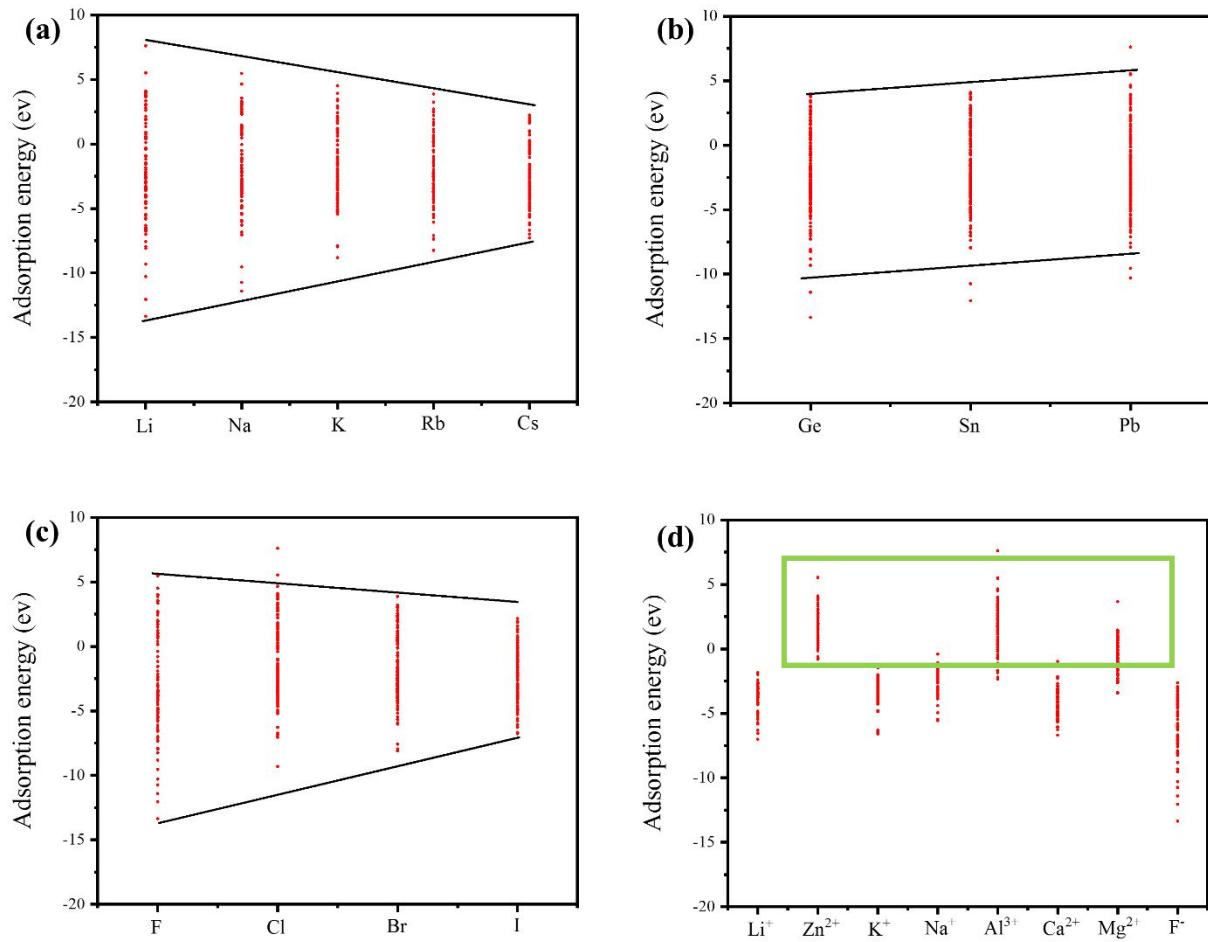


Figure S3. Visualization of the adsorption energy distributions versus the elemental types in the training/test dataset. (a) Adsorption energy versus A-site species (Li, Na, K, Rb and Cs). (b) Adsorption energy versus B-site species (Ge, Sn and Pb). (c) Adsorption energy versus X-site species (F, Cl, Br, I). (d) Adsorption energy versus the ion species (Li^{2+} , Zn^{2+} , K^+ , Na^+ , Al^{3+} , Ca^{2+} , Mg^{2+} and F^-). The metals possessing larger ion densities (such as Zn^{2+} , Al^{3+} and Mg^{2+}) than lithium that exhibit larger adsorption energy and weaker binding strength are highlighted.

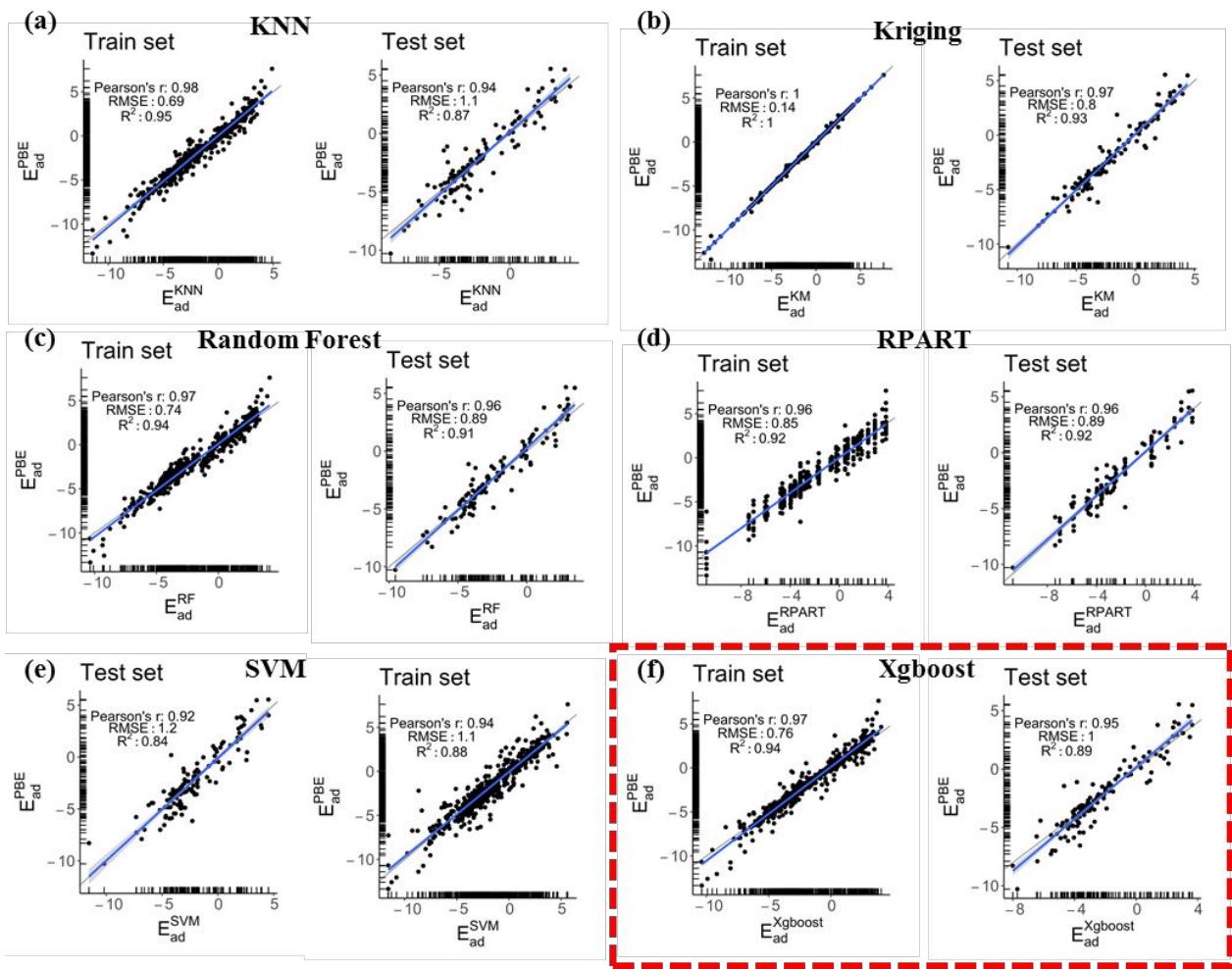


Figure S4. Machine learning models are constructed via KNN (a), Kriging (b), Random Forest (c),

Rpart (d), SVM (e) and Xgboost (f). Xgboost offers the best accuracy among the six machine learning

models and is selected for the virtual space prediction.

Table S1. A complete list of the ranking of 13 features (Ion.Density, Ion.AtomicRadius and B.FirstIonization) according to various ranking methods: anova, cmim, correlation, disr, find_correlation, information_gain, jmi, jmim, kruskal_test, mim, mrmr, njmim, relief and variance.

	anova	cmim	correlation	disr	find_correlation	information_gain	jmi
1	Ion.Density	Ion.AtomicRadius	Ion.Density	Ion.AtomicRadius	Ion.AtomicRadius	Ion.Specific.Heat	Ion.AtomicRadius
2	Ion.Melting.Point	Ion.Density	Ion.Melting.Point	Ion.Density	A1.Specific.Heat	Ion.Melting.Point	X1.NumberofProtons
3	Ion.NumberofProtons	Ion.Energy.ev.	Ion.NumberofProtons	Ion.Energy.ev.	B.Firstlonization	Ion.NumberofProtons	Ion.Energy.ev.
4	Ion.Number.of.Shells	X1.NumberofProtons	Ion.Number.of.Shells	Ion.Number.of.Shells	X1.NumberofProtons	Ion.Density	Ion.Melting.Point
5	Ion.Energy.ev.	Ion.Melting.Point	Ion.Energy.ev.	Ion.Melting.Point	Ion.Energy.ev.	Ion.Energy.ev.	Ion.Density
6	Ion.Specific.Heat	Ion.Number.of.Shells	Ion.Specific.Heat	Ion.NumberofProtons	Ion.Melting.Point	Ion.AtomicRadius	Ion.NumberofProtons
7	Ion.AtomicRadius	Ion.NumberofProtons	Ion.AtomicRadius	Ion.Specific.Heat	Ion.Specific.Heat	Ion.Number.of.Shells	X4.NumberofProtons
8	A1.NumberofProtons	Ion.Specific.Heat	A1.NumberofProtons	X1.NumberofProtons	Ion.Density	X4.NumberofProtons	Ion.Number.of.Shells
9	X1.NumberofProtons	A1.NumberofProtons	X1.NumberofProtons	X4.NumberofProtons	B.NumberofProtons	X1.NumberofProtons	Ion.Specific.Heat
10	A1.Specific.Heat	X4.NumberofProtons	A1.Specific.Heat	A1.Specific.Heat	A1.NumberofProtons	A1.Specific.Heat	A1.NumberofProtons
11	B.NumberofProtons	B.NumberofProtons	B.NumberofProtons	A1.NumberofProtons	Ion.NumberofProtons	B.NumberofProtons	A1.Specific.Heat
12	X4.NumberofProtons	A1.Specific.Heat	X4.NumberofProtons	B.Firstlonization	X4.NumberofProtons	A1.NumberofProtons	B.Firstlonization
13	B.Firstlonization	B.Firstlonization	B.Firstlonization	B.NumberofProtons	Ion.Number.of.Shells	B.Firstlonization	B.NumberofProtons

	jmim	kruskal_test	mim	mrmr	njmim	relief	variance
1	Ion.AtomicRadius	A1.NumberofProtons	Ion.AtomicRadius	Ion.AtomicRadius	Ion.AtomicRadius	X4.NumberofProtons	Ion.Energy.ev.
2	X1.NumberofProtons	B.Firstlonization	Ion.Melting.Point	X1.NumberofProtons	Ion.Density	A1.Specific.Heat	Ion.Melting.Point
3	Ion.Density	A1.Specific.Heat	Ion.NumberofProtons	A1.NumberofProtons	Ion.Energy.ev.	B.NumberofProtons	B.NumberofProtons
4	Ion.Energy.ev.	B.NumberofProtons	Ion.Energy.ev.	Ion.Density	Ion.Melting.Point	B.Firstlonization	A1.NumberofProtons
5	Ion.Melting.Point	X1.NumberofProtons	Ion.Density	B.NumberofProtons	Ion.Number.of.Shells	Ion.Density	X4.NumberofProtons
6	Ion.NumberofProtons	Ion.NumberofProtons	Ion.Number.of.Shells	Ion.Melting.Point	Ion.NumberofProtons	Ion.Melting.Point	X1.NumberofProtons
7	Ion.Number.of.Shells	Ion.Energy.ev.	Ion.Specific.Heat	X4.NumberofProtons	Ion.Specific.Heat	Ion.Specific.Heat	Ion.NumberofProtons
8	Ion.Specific.Heat	Ion.Number.of.Shells	X1.NumberofProtons	Ion.NumberofProtons	X1.NumberofProtons	Ion.Energy.ev.	Ion.Density
9	A1.NumberofProtons	Ion.Melting.Point	X4.NumberofProtons	Ion.Number.of.Shells	A1.NumberofProtons	Ion.AtomicRadius	A1.Specific.Heat
10	X4.NumberofProtons	Ion.Specific.Heat	A1.NumberofProtons	A1.Specific.Heat	X4.NumberofProtons	Ion.NumberofProtons	Ion.Specific.Heat
11	B.Firstlonization	Ion.AtomicRadius	A1.Specific.Heat	B.Firstlonization	B.Firstlonization	Ion.Number.of.Shells	Ion.Number.of.Shells
12	A1.Specific.Heat	Ion.Density	B.NumberofProtons	Ion.Energy.ev.	A1.Specific.Heat	X1.NumberofProtons	Ion.AtomicRadius
13	B.NumberofProtons	X4.NumberofProtons	B.Firstlonization	Ion.Specific.Heat	B.NumberofProtons	A1.NumberofProtons	B.Firstlonization

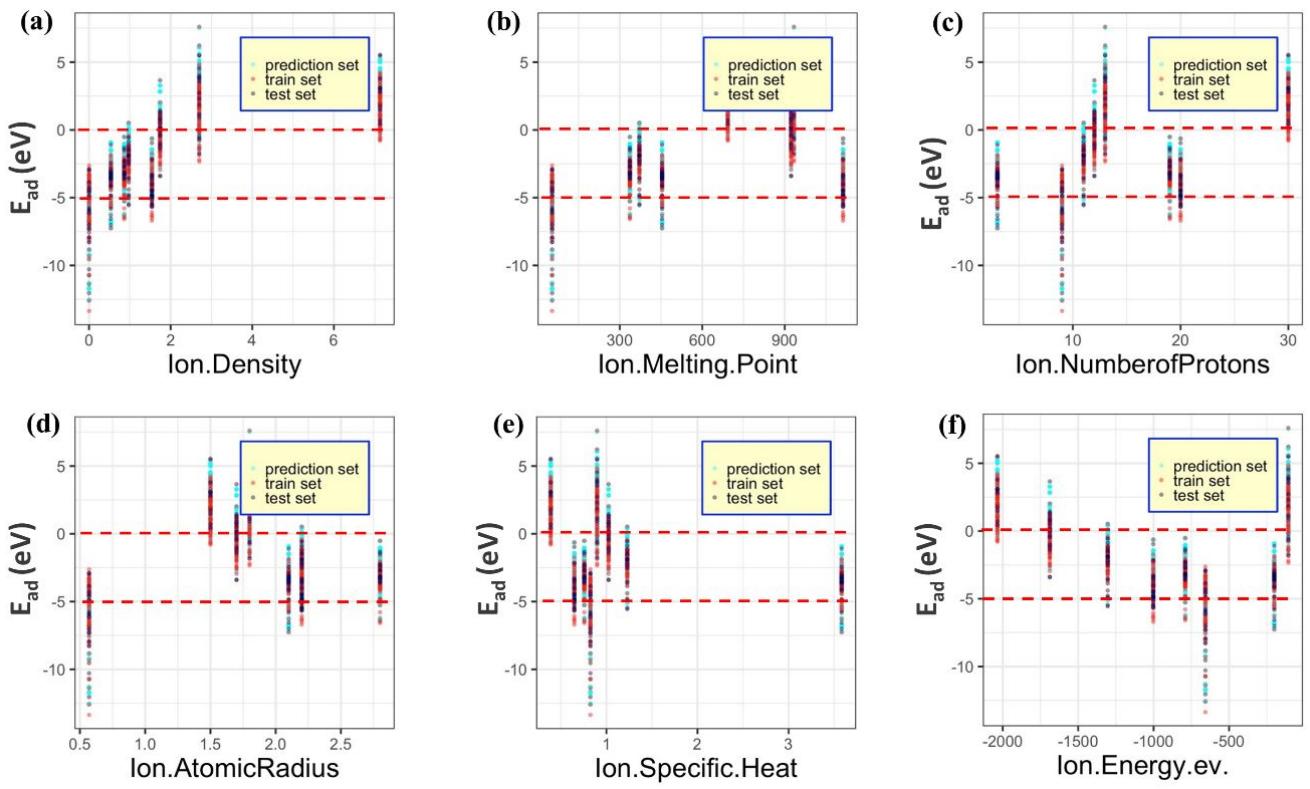


Figure S5. Visualization of the prediction dataset: the adsorption energy distributions versus the features of the ions in the prediction dataset. The proposed appropriate adsorption energy range ($-5 \text{ eV} < E_{ad} < 0 \text{ eV}$) is highlighted.

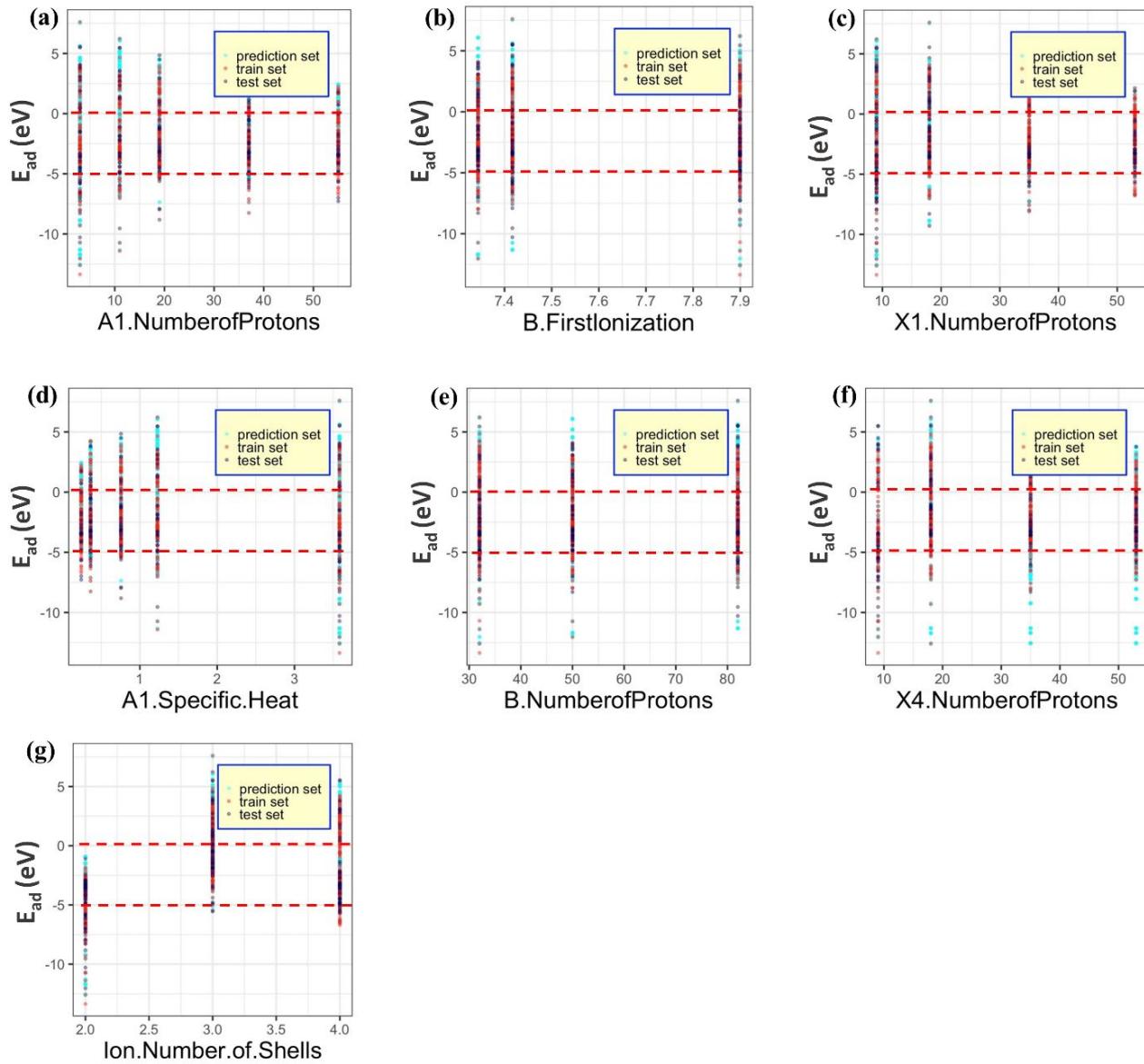


Figure S6. Visualization of the adsorption energy distributions versus the features of the halide

perovskites in the prediction dataset. The proposed appropriate energy ranges are highlighted.

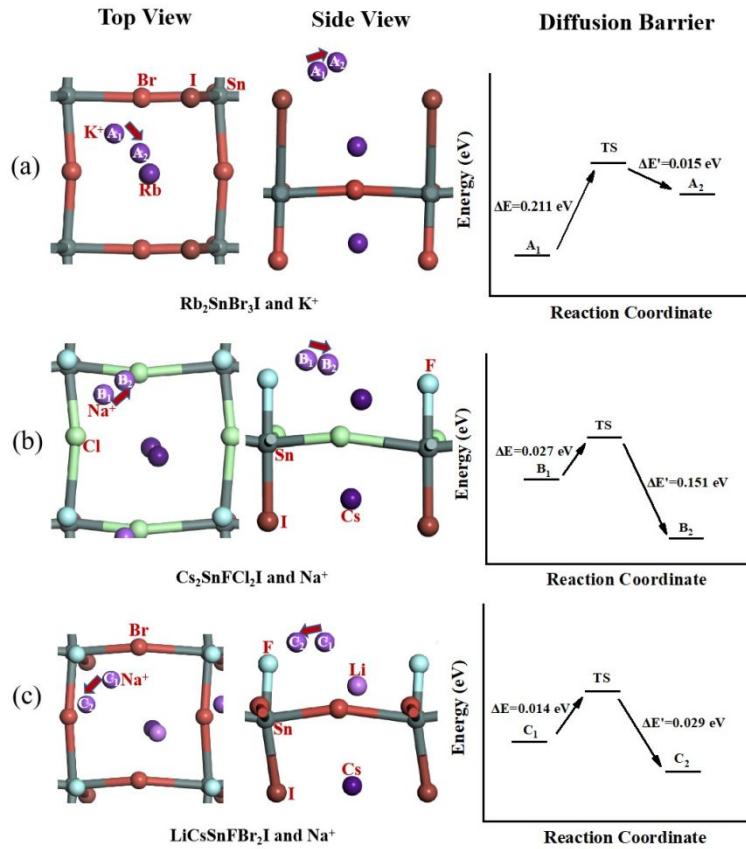


Figure S7. Typical surface diffusion path and energy barrier for the ion diffusion of the screened candidates. (a) $\text{K}^+/\text{Rb}_2\text{SnBr}_3\text{I}$, where A1 and A2 represent the K^+ positions before and after diffusion. (b) $\text{Na}^+/\text{Cs}_2\text{SnF}_2\text{I}$, where B1 and B2 represent the Na^+ positions before and after diffusion. (c) $\text{Na}^+/\text{LiCsSnF}_2\text{I}$, where C1 and C2 represent the Na^+ positions before and after diffusion.

Three typical contradictions in machine learning studies for materials science:

Contradictions between high dimension and small sample data: The data for the materials science is multi-dimensional while the sample size is much smaller compared those for other domains such as

machine vision and facial recognition; as a result, there is contradiction between high dimension and small sample data. In order to alleviate the issue, proper feature selection, feature extraction, sample augmentation, active learning, and ensemble learning methods can be employed. In this study, the database is prepared via the high-throughput DFT calculation while the features incorporate the fundamental structural and property descriptors for the constituting atoms and ions. The DFT+ML method is popular these days for materials predictions while the careful selection of features and the comprehensive feature ranking procedure offer a decent balance between the high dimension and the small sample data. It could be helpful if more data including the literature data and experimental data can be included for a more appropriate machine learning model.

Contradictions between model complexity and ease of use: The interpretability is critical for the machine learning data-driven studies for materials science. In this study, the interpretability is facilitated by the inclusion of the interpretable features, the justification of the feature importance and the chemical analysis of the prediction dataset. For example, the statistical ranking methods highlight the importance of the ion density for the adsorption energy, which highlights the high impacts of the mass and volume of the ion adsorbates to influence the adsorbate...adsorbent interactions. In addition, the Xgboost algorithm is identified to be the most suitable one for the machine learning model

construction and achieves the best accuracy ($r = 0.97$, $R^2 = 0.93$), which is thus selected to predict the virtual design space consisting of 11976 ion/perovskite systems.

Contradictions between learning results and domain knowledge: For data-driven machine learning studies for materials prediction, the domain knowledge of materials science is critical. As a result, we spent much time analyzing the domain-specific machine learning results in addition to the feature analysis procedure. The 11976 predicted ion/perovskite systems are further screened and five exemplar ion/perovskite system in the prediction set, $\text{Li}^+/\text{K}_2\text{GeClBr}_3$, $\text{K}^+/\text{Rb}_2\text{SnFBr}_3$, $\text{K}^+/\text{Rb}_2\text{Br}_3\text{I}$, $\text{Na}^+/\text{Cs}_2\text{SnF}_2\text{I}$ and $\text{Na}^+/\text{LiCsSnBr}_2\text{I}$ (Table 1), are proposed as the candidates; these potential candidates are comprehensively evaluated via the first-principles calculations to predict their optoelectronic and stability performance. For example, the optimized structures, band structures and projected density of states of the exemplar screened candidates $\text{K}^+/\text{Rb}_2\text{SnBr}_3\text{I}$, $\text{Na}^+/\text{Cs}_2\text{SnF}_2\text{I}$ and $\text{Na}^+/\text{LiCsSnF}_2\text{I}$ are discussed, while the molecular dynamics of the selected ion/perovskite systems $\text{K}^+/\text{Rb}_2\text{SnBr}_3\text{I}$, $\text{Na}^+/\text{Cs}_2\text{SnF}_2\text{I}$ and $\text{Na}^+/\text{LiCsSnFBr}_2\text{I}$ with NVT ensemble at 298K are performed to evaluate the thermodynamic stability. These highlight the importance of the domain knowledge for the machine learning studies.

Importance of feature selection step

The feature selection step is essential for the predictive ability and interpretability of the machine learning model in materials science studies. The three processing layers including sparsity evaluation, correlation evaluation, and redundancy evaluation are identified to be important for the machine learning studies,² which analyze the importance of features from different perspectives. In this study, the features representing typical structural and property descriptors of the ion/perovskite systems are selected, and the feature importance is analyzed via different ranking methods; these help interpret the machine learning data. Moreover, the relationships between features and the target outputs as well as the physical meaning are discussed to demonstrate the domain knowledge. We suggest the importance of the features and the chemical insights offered by comparing the features represent a typical feature evaluation process to help the interpretability. In addition, in this study, the machine learning model is accurate enough and the physical insights are provided via the feature selection and analysis, chemical information extraction from the virtual prediction space, while the mechanisms and scientific insights are further clarified via the DFT calculations on the predicted results; as a result, we suggest the current machine learning process is sufficient to demonstrate a valid feature selection process with domain expert knowledge. However, there are several issues that should be addressed: importantly, the predicted results should be experimentally validated, while this study only verifies the results via the DFT-based materials simulation. In fact, experimentally verifying the data-driven predicted results is

highly desired yet seldom effectively realized. In addition, more user-friendly machine learning tools and infrastructures should be established to help the experimentalists; currently there is a communication barrier between the coders carrying out the simulations and machine learning process and the experimentalists that fabricate and characterize the materials.

References:

- (1) Bennasar, M.; Hicks, Y.; Setchi, R. Feature Selection Using Joint Mutual Information Maximisation. *Expert Syst. Appl.* **2015**, *42* (22), 8520–8532. <https://doi.org/10.1016/j.eswa.2015.07.007>.
- (2) Urbanowicz, R. J.; Meeker, M.; La Cava, W.; Olson, R. S.; Moore, J. H. Relief-Based Feature Selection: Introduction and Review. *J. Biomed. Inform.* **2018**, *85*, 189–203. <https://doi.org/10.1016/j.jbi.2018.07.014>.