Journal of Materials Chemistry C

ROYAL SOCIETY OF CHEMISTRY

PAPER

View Article Online
View Journal

Cite this: DOI: 10.1039/d6tc00296j

Prediction of aqueous stable lead-free hybrid halide perovskites for efficient solar water splitting using machine learning and molecular dynamics

Mahalaxmi Chandramohan,[a] Madhana Gopal,[b] Tumpa Sadhukhan,[b] Athira Nambiar*[c] and Meenal Deo *[a]

Received 29th January 2026,
Accepted 29th March 2026

DOI: 10.1039/d6tc00296j

rsc.li/materials-c

Hybrid organic–inorganic halide perovskites (HOIPs) have garnered significant attention in many optoelectronic applications due to their high efficiency and tunable bandgaps. However, their application in the field of solar water splitting remains largely unexplored, attributed to their instability in aqueous environments and the fact that their valence and conduction band edges fail to straddle the redox potential of water, which prevents unbiased solar water splitting. To address this, various machine learning (ML) regression models are employed in this work to predict the band gap and band edge values for 3D-HOIPs that are critical for solar water splitting applications. Application-driven screening identified 21 potential lead-free perovskites with a statistically calculated STH efficiency of ≥10%. In addition, density functional theory (DFT) and *ab initio* molecular dynamics (AIMD) simulation highlighted the structural and aqueous stability of the selected HOIP, namely $FmSnI_2Br$, with the bandgap close to the ML predicted value, making it suitable for water splitting applications.

Introduction

The depletion of fossil fuels and climate change have driven society to seek renewable energy sources, with hydrogen energy being a key focus.[1] One of the sustainable methods for producing "green hydrogen" is through photoelectrochemical (PEC) or solar water splitting, which uses solar light to split water into hydrogen (H_2) and oxygen (O_2), using a semiconducting photoelectrode.[2] This approach offers a cheaper alternative to electrolyser-based methods but faces challenges of low efficiency and therefore lacks commercialization. Recently, hybrid organic–inorganic halide perovskites (HOIPs), especially lead-based surface-protected HOIPs, have been explored as promising photoelectrodes for PEC, due to their high absorption coefficient, easy solution processed synthesis and engineerable electronic structure properties.[3] Such tunable properties based on compositional engineering in HOIPs make them well suitable for solar water splitting applications.[4,5]

Fundamentally, an efficient photoelectrode requires not only the optimal band gap but it should also exhibit a favourable band alignment, which straddles the redox potentials of water. Specifically, this means the valence band edge of the photoelectrode should lie below ∼0.815 V and the conduction band edge above ∼−0.414 V *versus* the normal hydrogen electrode (NHE, pH = 7). Beyond energetic alignment, practical deployment of HOIP photoelectrodes demands long-term stability in aqueous environments and solar to hydrogen (STH) efficiency >10% to be commercially viable.[2] Along with these criteria, toxic-free and cost-effectiveness are other crucial factors for commercialization. In this context, 3D-HOIPs are more suitable due to their larger absorption coefficients and lower band gaps over their low dimensional counterparts, allowing them to capture a broader spectrum of sunlight, which is vital for optimizing energy harvesting and conversion processes.[6] Unfortunately, the currently explored regime of 3D-HOIP photoelectrodes for PEC applications has been limited to ABX_3 with only A = MA^+ and FA^+, B = Pb^{2+} and X = I^- and Br^-, where MA^+ and FA^+ cations start to degrade rapidly when they are in contact with water.[7] Although some organic cations show native hydrophobic nature, they are less explored for the PEC application and this indeed highlights the need to explore novel aqueous stable organic cations in HOIPs for PEC applications.

In order to explore new cation based HOIPs, either experimental routes or density functional theory (DFT) calculations

[a] Department of Physics and Nanotechnology, SRM Institute of Science and Technology, Kattankulathur, Tamil Nadu, India. E-mail: meenald@srmist.edu.in
[b] Department of Chemistry, SRM Institute of Science and Technology, Kattankulathur, Tamil Nadu, India
[c] Department of Computational Intelligence, SRM Institute of Science and Technology, Kattankulathur, Tamil Nadu, India. E-mail: athiram@srmist.edu.in

are required for material property analysis. Both of these approaches are trial-and-error based, which could be both time and resource consuming. In this regard, machine learning (ML) emerges as a novel and efficient solution, capable of rapidly exploring the vast compositional and structural space of HOIPs.[8] Earlier, ML models have been trained on target properties such as band gaps or photoconversion efficiency (in the case of solar cells) and subsequently applied to novel compositions for prediction, thereby accelerating the materials designing that meets specific requirements.[9] For example, the first report on lead-free 3D-HOIPs accelerated by ML was published by Lu *et al.*, employing the gradient boosting regression (GBR) algorithm, to predict the band gap values of 5156 possible candidates for photovoltaics.[10] This breakthrough triggered a surge in predicting various properties of halide perovskites, including the prediction of their band gap through structure–property relationship,[11–14] ferroelectric properties,[15,16] formation energy,[17–20] the efficiency of perovskite solar cells[21–23] and so on. Recently, Agarwal *et al.* predicted properties such as decomposition energy, band gap and photovoltaic efficiency by integrating Wasserstein generative adversarial networks (WGANs) with random forest regression for multicomponent halide perovskite discovery.[24]

But one of the primary concerns related to HOIPs is that, for various reasons, the exploration in the 3D realm has largely been confined to 'triple-cation-mixed-halide' perovskites, $FA_xMA_yCs_{1-x-y}Pb_aSn_{1-a}Br_mCl_nI_{1-m-n}$,[25–28] both in experimental studies and through ML acceleration, with only a few reported literature studies investigating other possible alternatives.[10,29,30] Secondarily, even though, interest has been shown on several perovskite property predictions, the determination of band edge values for lead-free 3D-HOIPs, which is a crucial factor for solar water splitting applications, is yet to be explored using ML approaches. In this context, Biswas *et al.* calculated the valence band maximum (VBM) and conduction band minimum (CBM) values of perovskites for solar water splitting purely grounded on Mulliken's electronegativity based empirical formula without indulging any ML studies for band edge predictions.[31] Recently, Yang *et al.* trained ML models on empirically derived band edge values reported by Nakajima T and Sawada K based on the same Mulliken's electronegativity formula.[32,33] In the latter report, the statistically calculated band edge values were manually shifted to match the experimentally observed values based on the A-site cations (MA^+, FA^+ and Cs^+) of the HOIP system, by compromising the accuracy and reliability of the results. Therefore, as per our current knowledge, there remains no single report on the exact ML-predicted VBM and CBM values of emerging semiconductors, specifically perovskites for optoelectronic applications, marking a significant gap in the literature, and this represents a fundamental bottleneck in the available literature studies.

To overcome the above-mentioned research gaps, we have predicted the band gap and band edge values of lead-free 3D-HOIPs, using classic ML regression techniques such as GBR, random forest regression (RFR), decision tree regression (DTR), extreme gradient boosting (XGB) and genetic programming symbolic regression (GPSR). Specifically, GPSR is an interpretable ML approach that elucidates the black-box model by generating physics-inspired mathematical expressions. While symbolic regression (SR) has been explored in perovskite solar cells,[34] OER activity,[35] and PEC photovoltage stability,[36,37] its application remains largely unexplored. To the best of our knowledge, this is the first attempt to predict valence and conduction band edge values of 3D-HOIPs using relatively simple, data-driven ML techniques, such as GBR and GPSR, grounded in experimental measurements.

By leveraging this predictive power, we navigated the compositional space of 3D-HOIPs and conducted a systematic screening, ultimately identifying 21 promising compositions with mixed halides that meet key criteria for solar water splitting, especially favorable band gap, band edge values, and statistically estimated PEC efficiency. One of the compounds was selected based on the promising band gap and STH efficiency value and validated for the structural stability and band gap value through DFT calculations. Building on this validation, *ab initio* molecular dynamics (AIMD) simulations were performed to assess the aqueous stability of the selected material, confirming its suitability for practical deployment in solar water splitting applications.

Methodology

ML training for bandgap prediction

Dataset preparation. In this study, the input data set containing 869 HOIP data points was collected and merged from two different datasets reported in the literature, with additional data points individually gathered from the research articles, mainly focusing on compounds with ABX_3 composition (for more details refer SI, Note S1). Out of the total data considered for training, approximately 76.5% are from the experimentally reported literature, while the remaining ~23.5% are from computational sources. It should be noted that, while removing the duplicated entries, experimentally reported bandgaps were prioritized and this led to a total of 818 unique datapoints. The dataset was then randomly split into training (80%) and testing (20%) sets, with normal distribution curves confirming appropriate partitioning (Fig. S1, SI).

Model training

Next to dataset preparation, a unique set of physical or chemical attributes that can either indirectly or directly correlate with the target property (band gap in this case) should be defined for each data point in the input dataset, known as ''features'' or ''descriptors''.[9] Although '*n*' number of such features can be used to define the target property, it is recommended that their number be significantly lower than the input dataset to mitigate the curse of dimensionality.[29] Initially, 82 highly influential elemental features such as electronegativity, ionic radii, tolerance factor, the choice of halides and their proportional combinations and crossovers (arithmetic operators of addition, subtraction, multiplication, and ratios applied over features) are considered as in Table S1, SI.[15] The constructed

dataset is now trained using four distinct supervised regression-based ML algorithms: gradient boosting regression (GBR), random forest regression (RFR), decision tree regression (DTR), and extreme gradient boosting (XGB). During the training, the performance of each regression algorithm was assessed using the evaluation metrics, namely, coefficient of determination (R^2) and root mean squared error (RMSE) (the details can be found in Note S2, SI). After the identification of a suitable algorithm based on the evaluation metrics, hyperparameter tuning was performed using exhaustive grid search, followed by a 'forward selection strategy', a feature engineering step to forecast an optimal feature set which involves starting the loop with features from scratch and iteratively adding one feature at a time that improves the model's performance, continuing until further additions no longer lead to significant improvement. Once again highly correlated features were eliminated as a secondary refinement process. With this, the trained model is used to make predictions over the prediction dataset with the band gap as its primary target property.

ML training for band edge prediction. The ML training for band edge prediction was similar to band gap, based on the RFR, DTR, GBR and XGB algorithms with a train-test split of 90 : 10 (Fig. S2, SI), for the gathered 119 experimental datapoints reported in the literature[26] and the previously constructed feature set with 82 features. From the experimental domain, it is well understood that the band gap highly influences the position of band edges to a greater extent. Therefore, the addition of band gap resulted in a total of 83 features. At this point, it is crucial to mention that the band gap (E_g) values utilized in the band edge prediction dataset were those estimated by our XGB model. Furthermore, with the help of evaluation metrics, a suitable algorithm was chosen with RMSE as a primary deciding metric. In addition, exhaustive grid search hyperparameter tuning and a 'forward selection strategy' were carried out for potential noise reduction in the model, similar to the work flow of band gap prediction in this work. To increase the prediction accuracy, a GPSR model was also trained for the same dataset in an attempt to gain insights over the black box model of ML.

Prediction dataset. To enhance the model's predictive capacity, the chemically diverse space of HOIPs was expanded to include 29 A-site cations, selected based on available ionic radius data.[29,38] The 29 monovalent A-site cations are shown in Fig. 1 and their corresponding ionic radii gathered from the literature are tabulated in Table S2, SI. For the B-site, only divalent cations, strictly satisfying the criterion of having a reported ionic radius for a +2-oxidation state and six-fold coordination in Shannon's ionic radii database, were taken into account, leading to 30 distinct elements across the periodic table. In addition, a total of 20 different permutations of halides have been considered for the X-site, as seen in Fig. 1. Finally, a database comprised of 17 400 HOIP molecules was constructed for prospective investigation as potential candidates, considering that the initial 818 training compounds are well studied.

Computational details. All first-principles calculations were performed within the framework of spin-polarized density functional theory (DFT) using the Vienna *Ab initio* Simulation Package (VASP) which inherently employs periodic boundary conditions (PBC).[39,40] The interactions between valence electrons and ionic cores were treated using the projector augmented wave

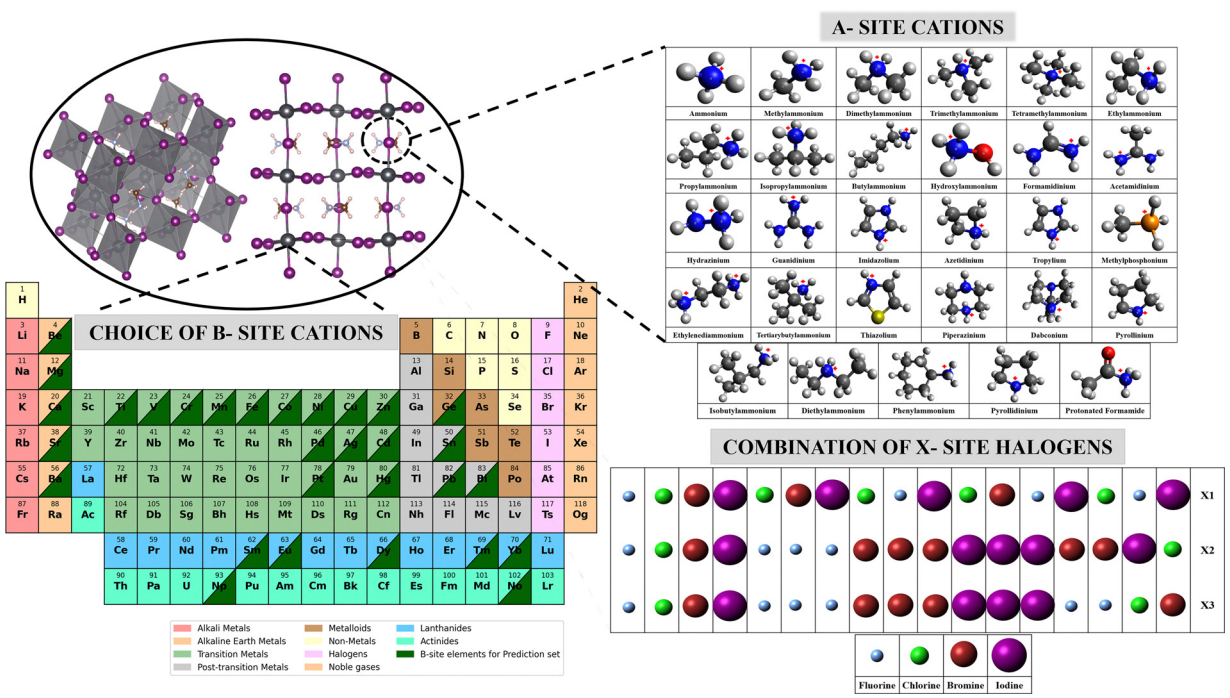

Fig. 1 Overall chemical space composition of the ABX_3 perovskite structure in the prediction dataset with 29 A-site organic cations, 30 B-site cations and 20 different combinations of halide anions.

(PAW) method. The generalized gradient approximation (GGA) with the Perdew–Burke–Ernzerhof (PBE) exchange–correlation functional was employed to describe electronic interactions.[41] An energy cutoff of 520 eV was used for the plane-wave basis set. Structural optimization was carried out using a convergence criterion of 1×10^{-8} eV for electronic self-consistency and 0.001 eV $Å^{-1}$ for ionic forces. A $2 \times 2 \times 2$ Γ-centered Monkhorst–Pack *k*-point mesh was used for geometry optimizations. The convergence threshold for total energy was set to 10^{-8} eV. Dispersion interactions were accounted for using the DFT-D3 method with Becke–Johnson damping.[42] For electronic structure analysis, including density of states (DOS) and band structure calculations, static runs were performed using a denser $9 \times 9 \times 9$ *k*-point mesh. The band structure was calculated along high-symmetry paths in the Brillouin zone generated using VASPKIT.[43]

Ab initio molecular dynamics (AIMD) simulations were performed on a slab model for the selected HOIP, cleaved along the (001) crystallographic direction.[44] The slab consisted of three inorganic B–X octahedral layers, while maintaining the stoichiometric arrangement of the organic cations within the perovskite framework. The exposed surface was terminated by a halide rich layer, representing a stable termination for halide perovskites. A vacuum region of 15 Å was introduced along the '*c*' direction (surface-normal direction) to avoid artificial interactions between periodic slab images and this vacuum space above the slab was filled with 40 water molecules, initially distributed randomly to mimic a liquid water environment. While for electronic structure calculations, PBC were applied in all three spatial directions, for the AIMD simulations of the hydrated interface, PBC were applied only in the in-plane direction. AIMD simulations were carried out within the Born–Oppenheimer approximation using the canonical (NVT) ensemble. The ionic temperature was controlled using a Nosé–Hoover thermostat at 300 K.[45,46] The equations of motion were integrated with a time step of 1 fs, and each trajectory was propagated for 8000 steps (8 ps). To reduce computational cost, the Brillouin zone was sampled only at the Γ-point during AIMD. The electronic self-consistent field convergence threshold was set to 10^{-4} eV, and Gaussian smearing with a width of 0.2 eV was applied to partial occupancies. All AIMD trajectories were recorded for subsequent structural and dynamical analyses of the water–perovskite interface.

Results and discussion

Bandgap prediction

Upon evaluation of the results achieved by ML algorithms for band gap training, it was observed that XGB outperformed the other models (Fig. S3 and S4, SI) by attaining an outstanding R^2 score of 0.998 and 0.989 on train and test sets with very low error metrics by capturing non-linear relationships between features and the target property (Fig. S5, S6 and Table S3, SI). With accurate hyperparameter tuning (Fig. S7, SI), along with feature reduction steps, namely, the 'forward selection strategy' (Fig. S8, SI) and 'Pearson's correlation coefficient-heatmap' (Fig. 2(a)), an optimized feature set including 17 features was achieved (Fig. 2(b)). These features are provided in Table S4 in the SI, along with their respective importance scores. The model's performance is notably higher than the R^2 scores typically reported in the literature, for regression models trained on perovskites, as can be seen from the comparison table in Table S5, SI.

Feature importance and discussion. Fig. 2(b) presents the feature importance ranking from the XGB model for band gap prediction. The atomic density of X-site halogens emerged as the most critical feature, with nearly a 50% importance score. Recalling the high correlation (0.98) between halogen density and atomic weight, where only density was retained, it becomes clear that the selection of the X-site has a significant impact on band gap values. This demonstrates that both the choice and

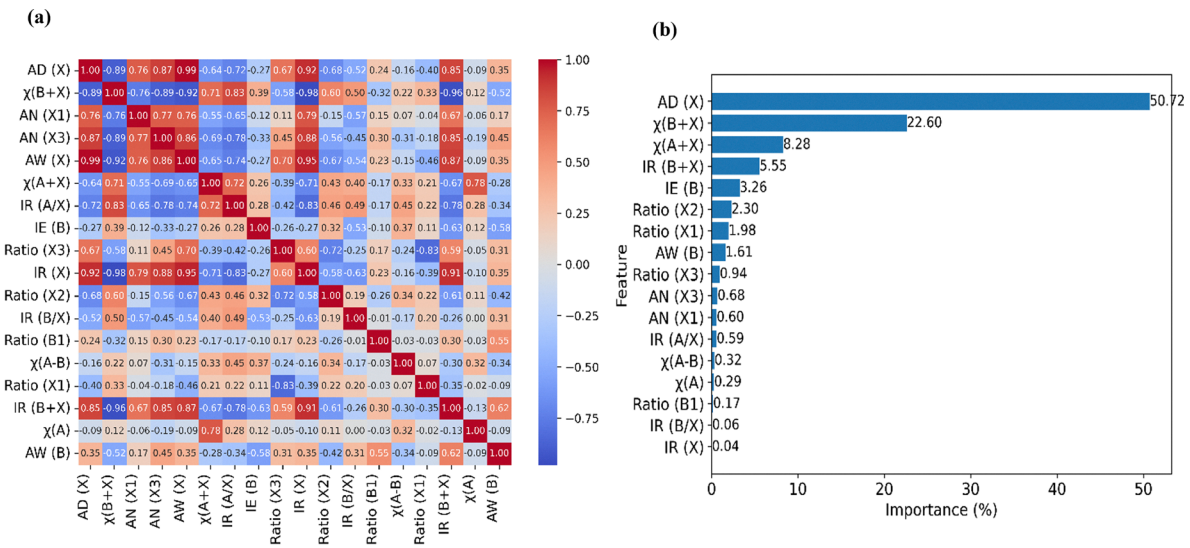

Fig. 2 (a) Pearson's correlation heatmap and (b) feature importance ranking for the XGB trained model in band gap prediction.

proportion of halides affect the band gap. The scatter plot from Fig. S9(a), SI further shows that as atomic density increases, the band gap decreases (*i.e.*, F > Cl > Br > I), consistent with experimental observations. This trend can also be linked to electronegativity, where a higher electronegativity typically results in a stronger ionic bond,[47] which in turn would elevate the band gap. For instance, fluorine (F^-) tends to exhibit higher band gaps, as shown in Fig. S10, SI, and is generally not favoured in perovskite photovoltaics. In this intricate lattice of perovskites, the B–X interaction is considered to be metavalent bonding,[48] with the valence band maximum (VBM) largely influenced by the antibonding interactions between the B-site's s-orbital and the X-site's p-orbital. Conversely, the conduction band minimum (CBM) emerges mostly from the antibonding of the B-site's p-orbital and partially from the X-site's s-orbital.[49] This orbital interplay underscores why, after the X-site, B-site features are pivotal, aligning with the ML results where χ(B + X) is ranked as the second most important feature. From Fig. S8(b), SI, it becomes evident that as the electronegativity (χ) of (B + X) increases, so does the band gap. For a specific band gap range, say 1.2–2.2 [eV], the corresponding χ(B + X) values cluster around 10.2–12.8 on the Mulliken scale. Yang *et al.* used electronegativity in Pauling's scale and the electronegativity difference between (B–X) was considered as the most important feature by the GBR algorithm which is in good agreement with our model.[27]

Intriguingly, the third-ranked feature is χ(A + X), which challenges conventional wisdom that the A-site has only an indirect influence on band gap prediction. While the A-site might not engage directly with the electronic states in the same way as the B and X sites, it plays a crucial role in shaping the perovskite's 3D framework. As seen in Fig. S8(c), SI, when χ(A + X) increases, the band gap follows it, mirroring the trend seen with χ(B + X). Also, from the correlation heatmap (Fig. 2(a)), no strong correlation values were found between χ(A + X) and other features, ruling out the possibility for it to act as proxy. In prior research conducted by Lu *et al.*, the ionic polarizability of the A-site was ranked as an even more critical feature than the B-site features, with the recommendation to use weakly polarizable molecules in the A-site.[10] As is known well, the term 'polarizability' is inversely related to electronegativity, which reflects the ease with which an electron cloud can be distorted. Also, it is well established that the A-site cation size influences the octahedral tilt, which in turn imposes subtle changes in the band gap, even if indirectly.[50] When connecting these dots together, it becomes clearly understandable that electronegativity, polarizability, and ionic size are deeply intertwined. Thus, it would be justifiable to conclude that χ(A + X) with a feature importance score of 8.28% might serve as a complex, yet crucial indicator that captures nuanced aspects of the material's behaviour, which other features alone do not fully account for, thereby asserting its strong predictive power in the model. Next is the ionic radii related feature, which is IR(B + X), (Fig. S8(d), SI), showing an inverse relation with the bandgap, which is an expected outcome. Take, for instance, $MAPbI_3$, where IR(Pb + I) equals 3.39 Å, reflecting the largest radii for both Pb and I. This directly correlates with the findings for IR(B + X), where the optimal bandgap is observed for an ionic radius, within a range of ~2.49–3.39 Å, further validating the accuracy of this feature in predicting bandgap behaviour. The features discussed so far account for 92% of the model's predictive power, while additional features (Fig. S8(e–h), SI) such as ionization energy, atomic weight of the B-site, and the octahedral factor contribute to enhancing prediction accuracy and serve as distinguishing factors among the features.

Band edge prediction. Along with an appropriate band gap, the valence and conduction band edge positions of perovskites relative to the redox potential of water are also crucial for solar water splitting. Specifically, at pH 7, the conduction and valence band edges should be ≤ -0.414 V and ≥ 0.815 V (*vs.* NHE),[51] respectively, straddling the water redox potential scales. To determine the band edge positions ($E_{(\mathrm{VBM,CBM})}$), Castelli *et al.* used an empirical formula that primarily relies on the Mulliken electronegativity of a neutral atom (eqn (1)), provided the band gap of the molecules is known (eqn (2)).[52]

$$\chi(\mathrm{A}_a\mathrm{B}_b\mathrm{C}_c) = \left[\chi(\mathrm{A})^a \cdot \chi(\mathrm{B})^b \cdot \chi(\mathrm{C})^c\right]^{\frac{1}{(a+b+c)}} \quad (1)$$

$$E_{(\mathrm{VBM,CBM})} = \left[\chi(\mathrm{A})^a \cdot \chi(\mathrm{B})^b \cdot \chi(\mathrm{C})^c\right]^{\frac{1}{(a+b+c)}} \pm \frac{E_g}{2} \quad (2)$$

where a, b, and c correspond to the number of atoms of A, B, C elements respectively in a chemical formula.[53] This formula predicts the band edge positions of inorganic perovskites, though with minor accuracy limitations. When this formula was extended to organic–inorganic perovskites by calculating the electronegativity of A-site initially and then applying it in the empirical formula with B- and X-site values in eqn (1), the band edges for the compounds were obtained. Biswas *et al.* used a similar approach to predict the band edge placements of inorganic and mixed HOIPs.[31] However, when compared to reported VBM and CBM values for HOIPs in the literature, a wider error range emerged, making this conventional formula insufficient for screening. In the study conducted by Liu *et al.*,[26] ratios of A-, B-, and X-sites from 114 mixed perovskites were used as features, yielding RMSE values of 0.11 and 0.12 for CBM and VBM training, with predictions on 5 unseen data points. As the study focused on band gap estimation, further VBM and CBM predictions were not conducted. For the above-mentioned dataset with 119 data points, we performed ML studies to predict VBM values using the same 82 features employed earlier for bandgap prediction with E_g as an additional feature. Based on the R^2 and error values shown in Fig. 3(a) and (b), the GBR model, which outperformed for both training and test sets similarly, is chosen for further feature engineering (Table S6, SI), with an RMSE of 0.13 (on the test set).

From the forward selection plot generated during feature engineering (Fig. 3(c)), there was no significant improvement in model performance when the number of features exceeded five. Therefore, these five features: band gap (E_g), tolerance factor (t), sum of ionic radii of B- and X-sites (IR(B + X)), sum of electronegativity of B- and X-sites (χ(B + X)) and atomic density

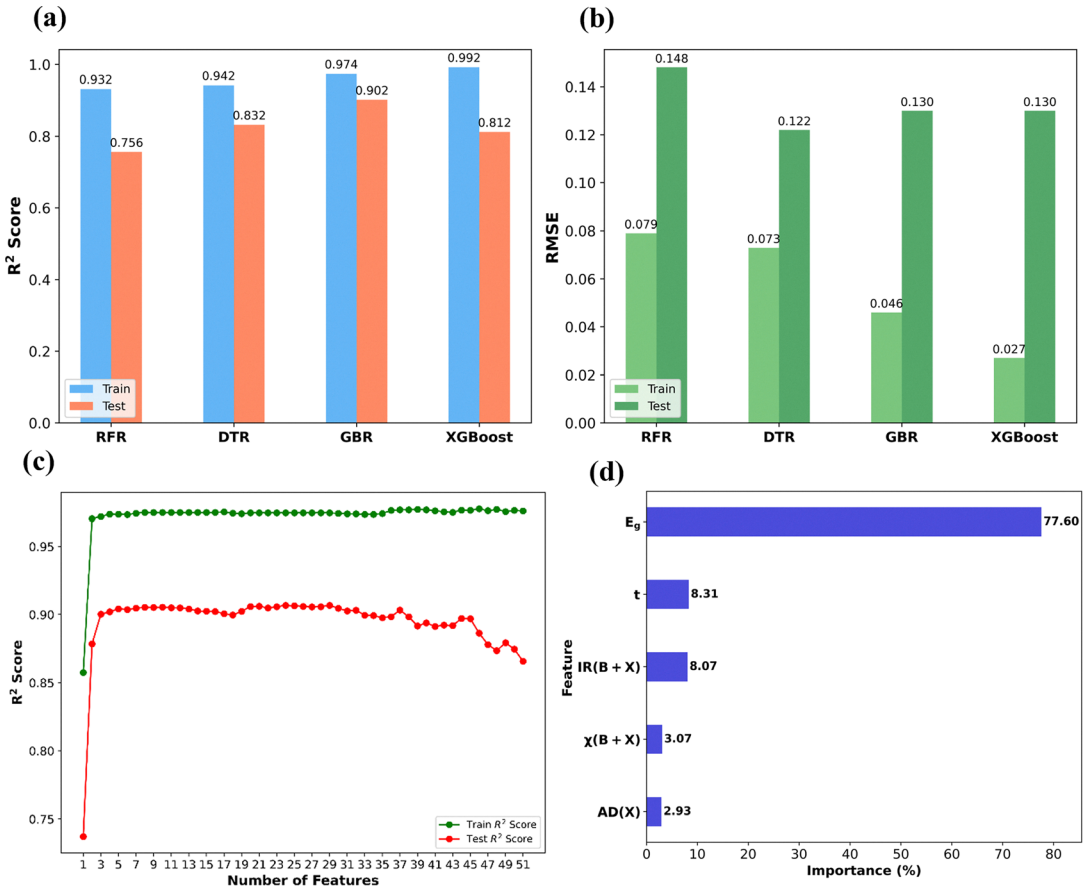

Fig. 3 (a) Comparison of R^2 scores and (b) loss functions obtained for different algorithms. (c) Visualization plot for the forward selection strategy and (d) feature importance plot for band edge prediction.

of the X-site (AD(X)), were calculated for their feature importance (Fig. 3(d)) and were used for further VBM prediction. Recalling that the VBM arises mainly from B- and X-site antibonding interactions confirms the importance of ML-extracted features. Next, to establish a clear analytical equation linking the selected descriptors to the target property, in this study, the genetic programming symbolic regressor (GPSR) from the 'gplearn' library was employed to derive a numerical relationship for predicting the VBM, based on the final features selected from the trained GBR algorithm.[54] The model achieved convergence with an RMSE of ~0.16, resulting in the formula (eqn (3)):

$$E_{VBM} = [0.451 - (E_g + t)] - IR(B + X) \quad (3)$$

where E_g = bandgap, t = old tolerance factor, and IR(B + X) = sum of ionic radii of B- and X-sites (Table S7, SI). With the derived formula, predictions on VBM values were made for 17 400 compounds and then CBM values were calculated using XGB predicted band gap values.

Table S8, SI observation reveals that GBR, GPSR and empirical formula methods for band edge predictions do not consistently perform well for every compound. The GBR model performs well for some compounds while GPSR model predicts accurately for others. To address this inconsistency, values from all three methods were averaged. These averaged band edge values were further considered for the materials screening.

Materials screening

Identifying suitable candidates for PEC water splitting requires not only meeting target properties but also adhering to structural criteria, which is essential for a stable 3D framework and efficient charge carrier dynamics.[55] Primarily, the dataset was screened out to 152 viable candidates as shown in Fig. 4, by applying key criteria, including charge neutrality, bandgap (1.23–2.2 eV),[56] and structural stability parameters, namely, the old tolerance factor (t = 0.8–1),[38,57] octahedral factor (μ = 0.414–0.732), and new tolerance factor ($\tau < 4.18$),[58] to facilitate the screening of 3D-HOIPs without significant octahedral distortion. These criteria highlighted the importance of A-site cations with ionic radii between 2.16–2.58 Å and >0.9 Å for B-site and excluded compounds with crystallographic instabilities or unsuitable ionic radii, ensuring the selection of stable 3D HOIPs.

Being done with crystallographic stability, the next step involved applying valence and conduction band edge criteria. VBM values were first converted from the absolute vacuum scale (AVS) to the NHE scale using eqn (4):

$$-E_{AVS} = E_{NHE} + 4.44\ V \quad (4)$$

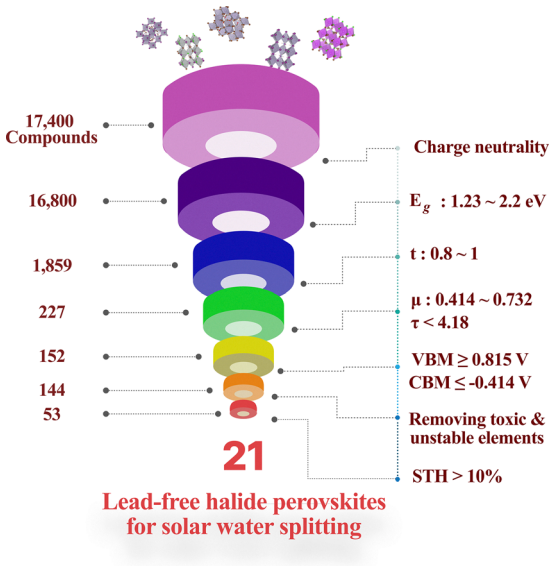

Fig. 4 Hierarchical screening of HOIPs for PEC water splitting application.

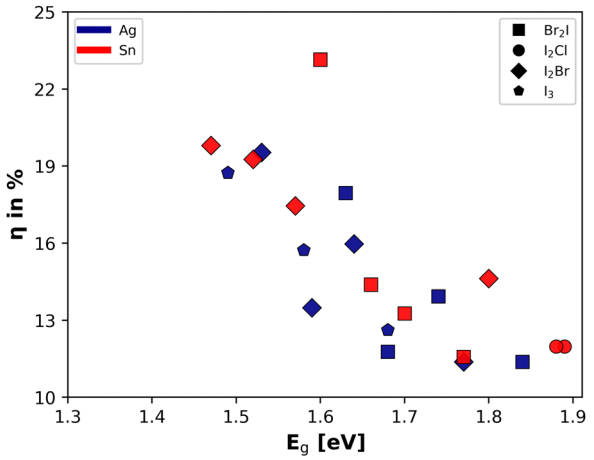

Fig. 5 Estimated STH efficiency values plotted against the predicted bandgap.

where E_{AVS} is the energy at AVS and E_{NHE} is the energy at the NHE scale.[59,60] After conversion, compounds with a VB edge $\geq$0.815 V and a CB edge ≤ -0.414 V were retained, considering their suitability for solar water splitting, narrowing the dataset to 144 compounds. Toxic elements (Pb^{2+}, Pd^{2+}, Cd^{2+}, and Hg^{2+}) were then excluded, reducing the selection to 81 viable 3D HOIPs, featuring 8 unique A-site and 3 unique B-site cations with varying halide combinations. Out of the 3 B-site cations namely Ag^{2+}, Sn^{2+} and Bi^{2+}, the practical existence of Bi in the +2 state is still experimentally difficult and hence compounds with Bi^{2+} were eliminated resulting in 53 compounds.

The solar to hydrogen (STH) conversion efficiency (η), being a key parameter for PEC, is statistically calculated for the remaining HOIPs excluding the repeatability of compounds from the training dataset (refer Note S3, SI for calculation details)[31,61] with 21 compounds exhibiting the commercial level estimated PEC efficiency of $\geq$10% as shown in Fig. 5. The band alignments for the final 21 3D-HOIPs are shown in Fig. 6, with respect to the NHE scale, with all the HOIPs exhibiting band edges that straddle the redox potential values of water, and their band edge values are also tabulated in Table S9, SI. The material with the highest efficiency, 23.14%, is $MPSnBr_2I$ (MP = methylphosphonium). Very recently, Zhang *et al.* and their group have experimentally investigated $MPSnI_3$ and $MPSnBr_3$, reporting bandgaps of 1.43 eV and 2.62 eV, respectively, which were not included in our training dataset.[62,63] Our ML model predicts corresponding bandgaps of 1.30 eV for $MPSnI_3$ and 2.00 eV for $MPSnBr_3$, confirming its effective predictive capability.

Model validation

Density functional theory. Since MP^+, MA^+ and other cations like Hz^+, HA^+, PA^+ and Az^+ have already been reported experimentally,[64–68] the next material of interest in this study is $FmSnI_2Br$ ($CHONH_3SnI_2Br$), with a calculated efficiency of 19.79% and a ML predicted band gap of 1.47 eV. To the best of our knowledge, the above selected material stands as an experimentally unexplored organic cation in the A-site chemistry of HOIPs. For the above selected material, a case study was conducted to gain more insights into its electronic, structural and aqueous stability. Initially to check for structural stability using DFT calculations, the orthorhombic phase of $MASnI_3$, serving as the structural prototype, was taken and optimized. The calculated lattice constants (a = 8.37 Å, b = 12.54 Å, c = 8.77 Å) were found to be consistent with previously reported values,[69] and the structure maintained the integrity of the corner-sharing $[SnI_6]^{2-}$ octahedral framework without any distortion (Fig. S11, SI). This optimized $MASnI_3$ structure was then used as a reference to construct the target system by substituting the A-site cation and halide atoms accordingly. A key consideration in halide substitution was the choice of position: axial *vs.* equatorial within the octahedral network. Both substitution modes were modelled and evaluated energetically. It was observed that equatorial substitution consistently resulted in lower total energies after relaxation, indicating it to be the thermodynamically favorable configuration. The system was therefore constructed with equatorial halide substitution and subsequently optimized. The optimized lattice parameters are a = 8.31 Å, b = 11.83 Å, and c = 8.66 Å for the $FmSnI_2Br$ system. The results reveal only minor variations in lattice constants compared to $MASnI_3$, indicating that the substituted systems retain the overall orthorhombic perovskite framework, without significant structural distortion as shown in Fig. 7(a). This suggests that the chemical modifications did not compromise the structural stability of the base lattice. After structural optimization, self-consistent computations were carried out to examine the band structure and electronic density of states. With a DFT-calculated band gap of 1.48 eV, which is in excellent agreement with its ML-predicted value, $FmSnI_2Br$ (shown in Fig. 7(b)) is found to be a desirable direct bandgap system for optoelectronic applications. It is important to note that although the inclusion of spin–orbit coupling (SOC) generally improves the accuracy of the calculated band-gaps, it was not included in the present work. Because in halide perovskites,

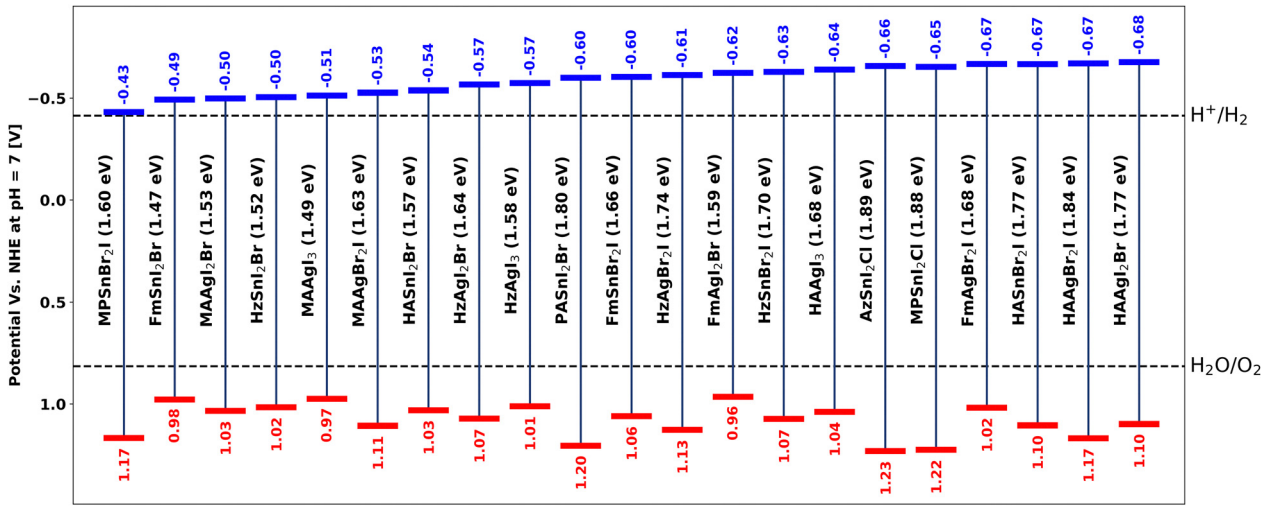

Fig. 6 Band alignments of the top 21 HOIPs with $\eta \geq 10\%$.

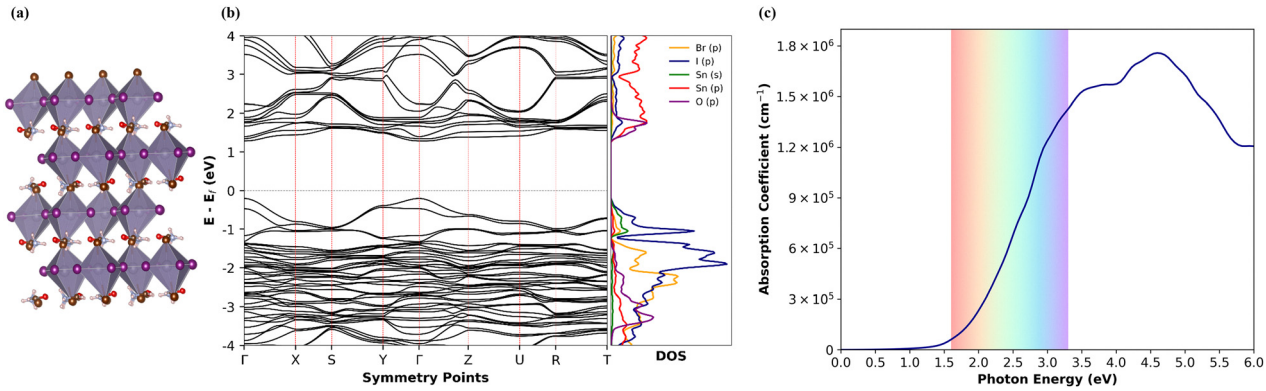

Fig. 7 (a) Structure of optimized $FmSnI_2Br$, (b) band structure of $FmSnI_2Br$, and (c) calculated absorption coefficient plot.

SOC effects are particularly significant in Pb-based HOIPs, where they can reduce the band gap by ~1 eV, whereas Sn-based systems exhibit a comparatively smaller effect (~0.3–0.4 eV).[70,71] Since the DFT calculations in this study were primarily intended to validate the trends predicted by the ML model rather than to obtain quantitatively precise band-gap values, the omission of SOC is not expected to affect the overall conclusions. Moreover, including SOC would significantly increase the computational cost.

From the DOS plot, the O-(p) states appear near the conduction band edge together with the dominant Sn-(p) contribution. This feature is particularly noteworthy because most reported HOIPs do not include oxygen within the A-site environment, and it is generally observed that the electronic states associated with organic A-site cations lie several electron volts below the VBM.[72] The presence of O-derived states near the band edge therefore reflects a distinct electronic configuration arising from this mixed-halide, oxygen-containing framework and points to an interesting direction for further investigation. In addition, Fig. 7(c) shows that the absorption onset begins approximately at 1.5 eV, which is in good agreement with the DFT calculated electronic band gap of 1.48 eV discussed earlier. The absorption coefficient near the onset is on the order of $>3.5 \times 10^4$ cm^{-1} and increases rapidly with increasing photon energy, reaching values on the order of 10^5 cm^{-1} in the visible region and 10^6 cm^{-1} in the ultraviolet region. Such absorption coefficient values are characteristic of well-established Sn-based perovskites and are consistent with the previously reported literature,[73,74] highlighting the strong light-harvesting capability of the material. Thus, with DFT calculations validating the electronic structure of $FmSnI_2Br$, it is further assessed for its behavior in an aqueous environment with the help of AIMD simulation.

AIMD simulation. The structural and dynamical response of the mixed-halide tin perovskite ($FmSnI_2Br$) under hydration was investigated using AIMD. A slab model containing the inorganic framework (SnI_2Br) and organic moieties in stoichiometric proportion was exposed to 40 water molecules in the vacuum region, and the trajectory was propagated for 8 ps (8000 steps) under near-ambient conditions. This set up is

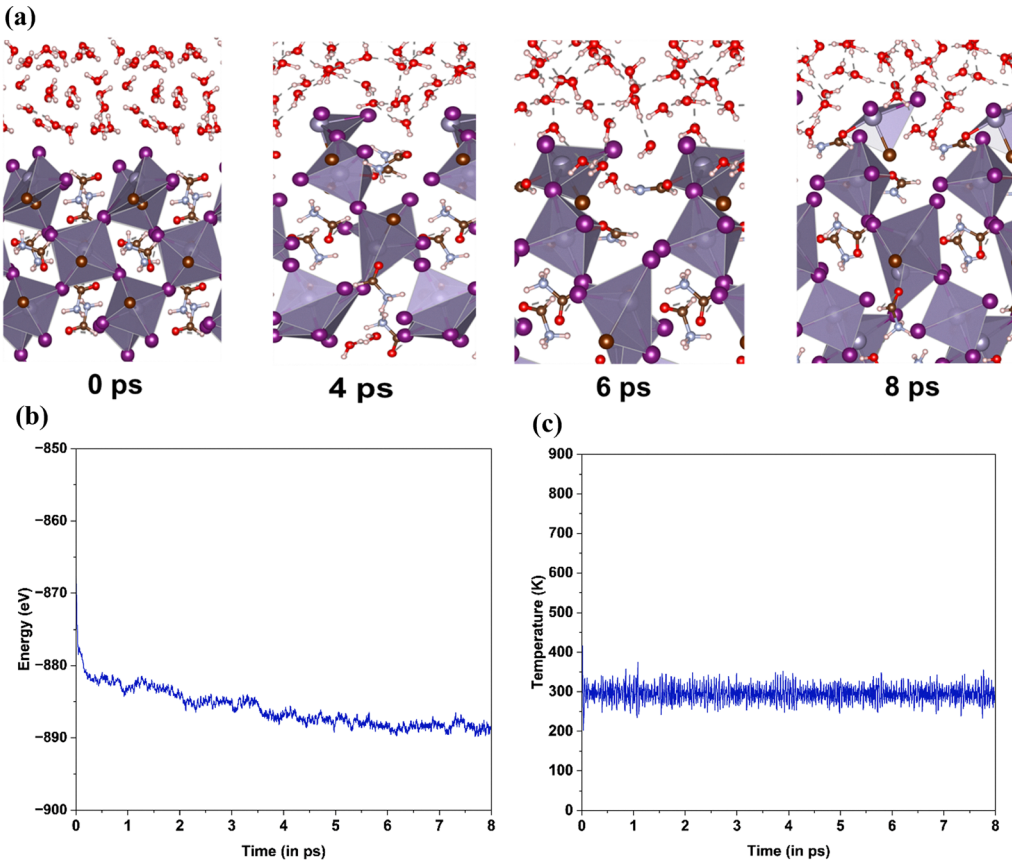

Fig. 8 (a) AIMD relaxed structures of the $FmSnI_2Br$-001 surface, (b) energy *vs.* time, and (c) temperature *vs.* time plot.

intended to probe the early-stage aqueous interaction between water molecules and the perovskite surface and to examine possible initial surface infiltration events. Because, in the context of solar water splitting, the perovskite surface is directly exposed to the aqueous environment, and it is this surface–water interface that governs the initial structural distortions, which can subsequently evolve into more extensive degradation of the perovskite. Here, the trajectories show rapid formation of hydrogen bonds between water molecules and surface halide ions, along with local distortions of the surface Sn-halide octahedra. These processes represent the initial steps of hydration-induced structural perturbation and typically occur within the first few picoseconds, as widely reported in previous AIMD studies of halide perovskite–water interfaces.[75–77]

The relaxed structures are shown in Fig. 8(a). The temporal evolution of the total energy revealed a two-stage behavior: an initial decrease during the first 3 ps as the system relaxed from its constructed geometry, followed by stabilization around a mean baseline with only thermal fluctuations which is shown in Fig. 8(b). This energetic relaxation corresponds to the redistribution of water molecules near the surface and minor adjustments of the lattice in response to hydration. The well-controlled temperature profile, maintained close to 300 K with only minor oscillations, confirmed that the simulation provided a stable configuration, which is shown in Fig. 8(c). Radial distribution functions further elucidated the local coordination environment. Sn–I and Sn–Br correlations retained sharp peaks near 2.7–2.9 Å, consistent with stable octahedral bonding and preservation of the inorganic framework (Fig. S12, SI). These results portray a hydrated $FmSnI_2Br$ system that is dynamically flexible yet structurally robust. This atomic-level picture highlights the balance between stability and interfacial dynamics in hydrated tin-based perovskites and provides a framework for developing stabilization strategies through surface passivation and compositional engineering.

Conclusion

In conclusion, we have applied simple yet effective ML methods for predicting suitable organic inorganic halide perovskites for solar water splitting applications. We found that the XGB model performed well for band gap prediction, while the GBR model achieved superior performance in predicting band edge values, reflected by higher R^2 and lower RMSE, compared to other models. Also, the calculated STH efficiency for the ML-identified HOIPs was a maximum of 23.14%. We have screened 21 suitable HOIPs with STH $>10\%$ and having suitable band alignment with water redox potentials. In addition, we have successfully identified an unexplored HOIP, namely, $FmSnI_2Br$, exhibiting favorable bandgap, band alignment and aqueous stability for solar water splitting, which was predicted by leveraging the

power of ML integrated with DFT and AIMD results. Furthermore, unravelling the chemical and elemental patterns that govern the aqueous stability of perovskites through deep learning could lead to groundbreaking advances in the field of solar water splitting.

Author contributions

Mahalaxmi Chandramohan: data curation, software, methodology, visualization, writing – original draft; Madhana Gopal: software, validation, writing – original draft (DFT and AIMD); Tumpa Sadhukhan: formal analysis, investigation, resources, writing – review and editing (DFT and AIMD); Athira Nambiar: methodology, supervision (ML), writing – review and editing; Meenal Deo: conceptualization, project administration, supervision, writing – review and editing.

Conflicts of interest

There are no conflicts to declare.

Data availability

The data supporting this article have been included as part of the supplementary information (SI). Supplementary information: Details on dataset preparation, ML model training, Computational and AIMD simulation details, evaluation metric plots, hyperparameter tuning, feature engineering plots, scatter plots between band gap and important features, details of VBM and CBM predictions, STH efficiency calculations, ML-identified 21 HOIPs for solar water splitting application and other technical details. See DOI: https://doi.org/10.1039/d6tc00296j.

The dataset used for training can also be accessed through GitHub: https://github.com/Ahamm00/Band-gap-and-band-edge-prediction-of-HOIPs.

Acknowledgements

The computational resources provided by the High-Performance Computing Center (HPCC) of SRM Institute of Science and Technology, Kattankulathur, India are greatly acknowledged. MMC and MD would like to acknowledge Dr. Rudra Banerjee for the initial discussions on the density functional theory. The authors acknowledge open access funding provided by SRM Institute of Science and Technology, Kattankulathur, through an institutional agreement with the Royal Society of Chemistry.

References

1 F. Alasali, M. I. Abuashour, W. Hammad, D. Almomani, A. M. Obeidat and W. Holderbaum, *Energy Sci. Eng.*, 2024, **12**, 1934–1968.
2 P. Nandigana, S. Pari, D. Sujatha, M. Shidhin, C. Jeyabharathi and S. K. Panda, *ChemistrySelect*, 2023, **8**, e202204731.
3 Y. Liu, Y. Liu and Y. Guo, *Mater. Chem. Front.*, 2023, **7**, 5215–5246.
4 Y. Zhang, Q. Chen, H.-S. Yang, D. Kim, I. Shin, B. R. Lee, J. H. Kim, D. K. Moon, K. H. Kim and S. H. Park, *ACS Appl. Mater. Interfaces*, 2021, **13**, 33172–33181.
5 G. Grancini and M. K. Nazeeruddin, *Nat. Rev. Mater.*, 2019, **4**, 4–22.
6 X. Chen, H. Zhou and H. Wang, *Front. Chem.*, 2021, **9**, 715157.
7 J. Hidalgo, W. Kaiser, Y. An, R. Li, Z. Oh, A. F. Castro-Méndez, D. K. LaFollette, S. Kim, B. Lai, J. Breternitz, S. Schorr, C. A. R. Perini, E. Mosconi, F. De Angelis and J. P. Correa-Baena, *J. Am. Chem. Soc.*, 2023, **145**, 24549–24557.
8 C. W. Myung, A. Hajibabaei, J. H. Cha, M. Ha, J. Kim and K. S. Kim, *Adv. Energy Mater.*, 2022, **12**, 2202279.
9 Y. Liu, X. Tan, J. Liang, H. Han, P. Xiang and W. Yan, *Adv. Funct. Mater.*, 2023, **33**, 2214271.
10 S. Lu, Q. Zhou, Y. Ouyang, Y. Guo, Q. Li and J. Wang, *Nat. Commun.*, 2018, **9**(1), 3405.
11 V. Gladkikh, D. Y. Kim, A. Hajibabaei, A. Jana, C. W. Myung and K. S. Kim, *J. Phys. Chem. C*, 2020, **124**, 8905–8918.
12 D. O. Obada, E. Okafor, S. A. Abolade, A. M. Ukpong, D. Dodoo-Arhin and A. Akande, *Mater. Sci. Semicond. Process.*, 2023, **161**, 107427.
13 A. Talapatra, B. P. Uberuaga, C. R. Stanek and G. Pilania, *Commun. Mater.*, 2023, **4**, 1–14.
14 J. Lan, S. Yuan, H. Zheng, W. Yang, W. Li and J. Fan, *J. Phys. Chem. C*, 2023, **127**, 23412–23419.
15 S. Lu, Q. Zhou, L. Ma, Y. Guo and J. Wang, *Small Methods*, 2019, **3**(11), 1900360.
16 Y. Sun, X. Wang, C. Hou and J. Ni, *J. Phys. Chem. C*, 2023, **127**, 23897–23905.
17 E. T. Chenebuah, M. Nganbe and A. B. Tchagang, *Mater. Today Commun.*, 2021, **27**, 102462.
18 H. Park, R. Mall, F. H. Alharbi, S. Sanvito, N. Tabet, H. Bensmail and F. El-Mellouhi, *Phys. Chem. Chem. Phys.*, 2019, **21**, 1078–1088.
19 Z. Chen, J. Wang, C. Li, B. Liu, D. Luo, Y. Min, N. Fu and Q. Xue, *J. Mater. Chem. C*, 2024, **12**(38), 15444–15453.
20 W. Hu and L. Zhang, *Mater. Today Commun.*, 2023, **35**, 105841.
21 L. Zhang, Y. Huang, L. Yan, J. Ge, X. Ma, Z. Liu, J. You, A. K. Y. Jen and S. Frank Liu, *Adv. Intell. Syst.*, 2024, **6**, 2300678.
22 M. M. Salah, Z. Ismail and S. Abdellatif, *Mater. Renew. Sustain. Energy*, 2023, **12**, 187–198.
23 W. Hussain, S. Sawar and M. Sultan, *RSC Adv.*, 2023, **13**, 22529–22537.
24 I. Agrawal, M. Biswas and A. Mannodi-Kanakkithodi, *Comput. Mater. Sci.*, 2025, **258**, 113989.
25 J. Li, B. Pradhan, S. Gaur and J. Thomas, *Adv. Energy Mater.*, 2019, **9**(46), 1901891.
26 Y. Liu, W. Yan, H. Zhu, Y. Tu, L. Guan and X. Tan, *Org. Electron.*, 2022, **101**, 106426.
27 C. Yang, X. Chong, M. Hu, W. Yu, J. He, Y. Zhang, J. Feng, Y. Zhou and L. W. Wang, *ACS Appl. Mater. Interfaces*, 2023, **15**, 40419–40427.

28 C. Ren, Y. Wu, J. Zou and B. Cai, *Materials*, 2024, **17**(11), 2686.
29 T. Wu and J. Wang, *Nano Energy*, 2019, **66**, 104070.
30 T. Wu and J. Wang, *ACS Appl. Mater. Interfaces*, 2020, **12**, 57821–57831.
31 M. Biswas, R. Desai and A. Mannodi-Kanakkithodi, *Phys. Chem. Chem. Phys.*, 2024, **26**(35), 23177–23188.
32 Y. Yang, H. Li, Q. Hua, J. Chen, S. Ren, Z. Chen, W. Sun, Y. Li, C. Sun, Y. Ye, R. Li, B. Qu, H. Wang, Y. Liu, Z. Chen, J. Zhang and L. Xiao, *Mater. Futures*, 2025, **4**(3), 035601.
33 T. Nakajima and K. Sawada, *J. Phys. Chem. Lett.*, 2017, **8**, 4826–4831.
34 Y. Hu, X. Hu, L. Zhang, T. Zheng, J. You, B. Jia, Y. Ma, X. Du, L. Zhang, J. Wang, B. Che, T. Chen and S. Liu, *Adv. Energy Mater.*, 2022, **12**(41), 2201463.
35 B. Weng, Z. Song, R. Zhu, Q. Yan, Q. Sun, C. G. Grice, Y. Yan and W. J. Yin, *Nat. Commun.*, 2020, **11**(1), 3513.
36 L. Zhang, W. Hu, M. He and S. Li, *ACS Appl. Energy Mater.*, 2023, **6**, 5177–5187.
37 Z. Pan, Y. Zhou and L. Zhang, *ACS Appl. Mater. Interfaces*, 2022, **14**, 9933–9943.
38 S. Alidoust, F. Jamalinabijan and A. Tekin, *ACS Appl. Energy Mater.*, 2024, **7**, 785–798.
39 J. Hafner, *J. Comput. Chem.*, 2008, **29**, 2044–2078.
40 G. Sun, J. Kürti, P. Rajczy, M. Kertesz, J. Hafner and G. Kresse, *J. Mol. Struct. THEOCHEM*, 2003, **624**, 37–45.
41 Z. Wu and R. E. Cohen, *Phys. Rev. B: Condens. Matter Mater. Phys.*, 2006, **73**, 235116.
42 S. Grimme, J. Antony, S. Ehrlich and H. Krieg, *J. Chem. Phys.*, 2010, **132**(15), 154104.
43 V. Wang, N. Xu, J. C. Liu, G. Tang and W. T. Geng, *Comput. Phys. Commun.*, 2021, **267**, 108033.
44 D. Marx and J. Ürg Hutter, *NIC Series*, 2000, vol. 1, pp. 301–449.
45 S. Nosé, *J. Chem. Phys.*, 1984, **81**, 511–519.
46 W. G. Hoover, *Phys. Rev. A*, 1986, **34**, 2499.
47 A. M. Ganose, D. O. Scanlon, A. Walsh and R. L. Z. Hoye, *Nat. Commun.*, 2022, **13**, 1–4.
48 M. Wuttig, C.-F. Schön, M. Schumacher, J. Robertson, P. Golub, E. Bousquet, C. Gatti, J.-Y. Raty, M. Wuttig, C.-F. Schön, M. Schumacher, J. Robertson, P. Golub, E. Bousquet and J.-Y. Raty, *Adv. Funct. Mater.*, 2022, **32**, 2110166.
49 H. Dong, C. Zhang, X. Liu, J. Yao and Y. S. Zhao, *Chem. Soc. Rev.*, 2020, **49**, 951–982.
50 Q. Lin, D. J. Kubicki, M. K. Omrani, F. Alam and M. Abdi-Jalebi, *J. Mater. Chem. C*, 2023, **11**, 2449–2468.
51 W. Li, H. Tian, L. Ma, Y. Wang, X. Liu and X. Gao, *Mater. Adv.*, 2022, **3**, 5598–5644.
52 I. E. Castelli, D. D. Landis, K. S. Thygesen, S. Dahl, I. Chorkendorff, T. F. Jaramillo and K. W. Jacobsen, *Energy Environ. Sci.*, 2012, **5**, 9034–9043.
53 R. Long, B. Li and Q. Mi, *AIP Adv.*, 2020, **10**(6), 065224.
54 Welcome to gplearn's documentation! — gplearn 0.4.2 documentation, https://gplearn.readthedocs.io/en/stable.
55 K. Bienkowski, R. Solarska, L. Trinh, J. Widera-Kalinowska, B. Al-Anesi, M. Liu, G. K. Grandhi, P. Vivo, B. Oral, B. Yılmaz and R. Yıldırım, *ACS Catal.*, 2024, **14**, 6603–6622.
56 S. S. Khamgaonkar, A. Leudjo Taka and V. Maheshwari, *Adv. Funct. Mater.*, 2024, 2405414.
57 I. E. Castelli, J. M. García-Lastra, K. S. Thygesen and K. W. Jacobsen, *APL Mater.*, 2014, **2**(8), 081514.
58 C. J. Bartel, C. Sutton, B. R. Goldsmith, R. Ouyang, C. B. Musgrave, L. M. Ghiringhelli and M. Scheffler, *Sci. Adv.*, 2019, **5**(2), eaav0693.
59 H. Zhang, Y. Zhang, Y. Zhang, H. Li, M. Ou, Y. Yu, F. Zhang, H. Yin, Z. Mao and L. Mei, *Nat. Commun.*, 2024, **15**(1), 6783.
60 T. E. A. Frizon, A. A. Vieira, F. N. da Silva, S. Saba, G. Farias, B. de Souza, E. Zapp, M. N. Lôpo, H. de, C. Braga, F. Grillo, S. F. Curcio, T. Cazati and J. Rafique, *Front. Chem.*, 2020, **8**, 360.
61 C. F. Fu, J. Sun, Q. Luo, X. Li, W. Hu and J. Yang, *Nano Lett.*, 2018, **18**, 6312–6317.
62 H. Y. Zhang, X. G. Chen, Z. X. Zhang, X. J. Song, T. Zhang, Q. Pan, Y. Zhang and R. G. Xiong, *Adv. Mater.*, 2020, **32**, 2005213.
63 H. Y. Zhang and R. G. Xiong, *Chem. Commun.*, 2023, **59**, 920–923.
64 A. D'Annibale, R. Panetta, O. Tarquini, M. Colapietro, S. Quaranta, A. Cassetta, L. Barba, G. Chita and A. Latini, *Dalton Trans.*, 2019, **48**, 5397–5407.
65 E. V. Campbell, B. Dick, A. L. Rheingold, C. Zhang, X. Liu, Z. V. Vardeny and J. S. Miller, *Chem. – Eur. J.*, 2018, **24**, 222–229.
66 R. Panetta, G. Righini, M. Colapietro, L. Barba, D. Tedeschi, A. Polimeni, A. Ciccioli and A. Latini, *J. Mater. Chem. A*, 2018, **6**, 10135–10148.
67 J. Tian, D. B. Cordes, A. M. Z. Slawin, E. Zysman-Colman and F. D. Morrison, *Inorg. Chem.*, 2021, **60**, 12247–12254.
68 D. Ma, Z. Xu, F. Wang and X. Deng, *CrystEngComm*, 2019, **21**, 1458–1465.
69 A. Dendane, B. Rerbal, T. Ouahrani, A. Molina-Sanchez, A. Muñoz and D. Errandonea, *RSC Adv.*, 2024, **14**, 19880–19890.
70 T. Das, G. Di Liberto and G. Pacchioni, *J. Phys. Chem. C*, 2022, **126**, 2184–2198.
71 J. Kang, *Phys. Rev. Mater.*, 2020, **4**, 085405.
72 D. A. Egger, A. M. Rappe and L. Kronik, *Acc. Chem. Res.*, 2016, **49**, 573–581.
73 S. Sen, S. Gopalan, R. Sellappan, A. N. Grace and P. Sonar, *Adv. Energy Sustainability Res.*, 2023, **4**, 2300110.
74 M. M. Byranvand, W. Zuo, R. Imani, M. Pazoki and M. Saliba, *Chem. Sci.*, 2022, **13**, 6766–6781.
75 W. Kaiser, D. Ricciarelli, E. Mosconi, A. A. Alothman, F. Ambrosio and F. De Angelis, *J. Phys. Chem. Lett.*, 2022, **13**, 2321–2329.
76 Y. Gao, D. Lin, P. Liu, T. Shi and W. Xie, *Mater. Chem. Front.*, 2024, **8**, 785–799.
77 S. Cheng and H. Zhong, *J. Phys. Chem. Lett.*, 2022, **13**, 2281–2290.