ARTICLE OPEN

Chemical hardness-driven interpretable machine learning approach for rapid search of photocatalysts

Ritesh Kumar[1] and Abhishek K. Singh[1]✉

Strategies combining high-throughput (HT) and machine learning (ML) to accelerate the discovery of promising new materials have garnered immense attention in recent years. The knowledge of new guiding principles is usually scarce in such studies, essentially due to the 'black-box' nature of the ML models. Therefore, we devised an intuitive method of interpreting such opaque ML models through SHapley Additive exPlanations (SHAP) values and coupling them with the HT approach for finding efficient 2D water-splitting photocatalysts. We developed a new database of 3099 2D materials consisting of metals connected to six ligands in an octahedral geometry, termed as 2DO (octahedral 2D materials) database. The ML models were constructed using a combination of composition and chemical hardness-based features to gain insights into the thermodynamic and overall stabilities. Most importantly, it distinguished the target properties of the isocompositional 2DO materials differing in bond connectivities by combining the advantages of both elemental and structural features. The interpretable ML regression, classification, and data analysis lead to a new hypothesis that the highly stable 2DO materials follow the HSAB principle. The most stable 2DO materials were further screened based on suitable band gaps within the visible region and band alignments with respect to standard redox potentials using the GW method, resulting in 21 potential candidates. Moreover, $HfSe_2$ and $ZrSe_2$ were found to have high solar-to-hydrogen efficiencies reaching their theoretical limits. The proposed methodology will enable materials scientists and engineers to formulate predictive models, which will be accurate, physically interpretable, transferable, and computationally tractable.

npj Computational Materials (2021)7:197 ; https://doi.org/10.1038/s41524-021-00669-4

INTRODUCTION

Hydrogen is one of the most promising fuels, which can meet the ever-increasing energy demands. Most of the currently utilized methods for hydrogen production are not eco-friendly and release pollutants in various forms. On the other hand, photocatalytic water splitting is truly considered a 'green method' for the generation of hydrogen. Traditionally, photocatalysts were based on bulk oxide materials such as TiO_2[1]. Large band gaps, low harvesting of visible light, and high tendency for charge recombination limit their wide-scale applications[2]. Recently, there has been a surge in finding various classes of two-dimensional (2D) materials for various applications including photocatalysis[3]. 2D materials offer several advantages such as increased active sites per surface area, enhanced charge separation and transport over their bulk counterparts[2]. They also mitigate the viability of charge carrier recombination by reducing the distance required for photogenerated electrons and holes for reaching the active sites. Among 2D materials, transition metal chalcogenides[4,5], and carbon nitrides[6,7] have been widely investigated. The other notable 2D materials to exhibit promising photocatalytic properties are the 2D layered double hydroxides (LDHs)[8,9]. These 2D materials belong to the octahedral symmetry group (O_h) or 1T phase, in which six hydroxyl ligands (OH^{-1}) are attached to a metal atom in the octahedral geometry. However, even after several years of extensive research into finding efficient 2D photocatalysts, none have surpassed desired limits of hydrogen generation[9]. It is primarily due to the fact that both theoretical and experimental investigations are driven by chemical intuitions. Manual search for promising materials among the vast chemical space through computational or experimental means is a very tedious, time-, and energy-intensive process.

To overcome this challenge, high-throughput (HT) methods have been developed, which screen materials based on simultaneous fulfillment of several physical and chemical properties. This approach, therefore, narrows down the search space considerably for the identification of materials with desired properties[10–12]. HT-based studies have been successful in the discovery of promising materials for various applications such as photovoltaics[13], batteries[14], and (photo)electrocatalysis[10]. Another emerging approach is to utilize machine learning (ML) to uncover the hidden pattern among existing data and predict desired material target properties at a nominal cost compared to the conventional theoretical or experimental methods. The ML methods can accelerate the identification of new materials. Coupling the HT approach with ML methods not only accelerates the process of finding desired materials, but can also unravel the physics of the underlying process[15,16]. However, such studies generally do not identify the reason for key design principles discovered, which do not advance scientific knowledge beyond simply discovering promising materials.

Herein, we coupled the interpretable ML (iML) and HT-based approach to identify stable octahedral 2D (2DO) photocatalysts for water splitting. The first tier of the HT scheme consisted of selecting 2DO materials with high thermodynamic and dynamic stability. The overall stability was decided based on strongly constrained and appropriately normed (SCAN)-calculated formation energies and convex hull distances along with elastic stiffness coefficients and Γ-point phonon calculations. The total computational time required to perform the relevant DFT calculations amounted to about three years or 450,000 CPU core hours, justifying the need for an ML-based study to quickly screen new stable 2D materials. Hence, highly accurate ML methods, including mean feature ranking and Bayesian hyperparameter optimization

[1]Materials Research Centre, Indian Institute of Science, Bangalore, Karnataka 560012, India. ✉email: abhishek@iisc.ac.in

were performed to predict formation energies and convex hull distances and to classify overall stabilities into nine classes. The chemical hardness-based features, which have hitherto been unutilized for ML applications in materials science, can achieve both the accuracy of the elemental features and distinguishability of the structural features. We found that the 2DO materials with chemically hard-soft or soft-soft interactions have optimum GW band gaps, while hard-hard interactions render them unsuitable for harvesting visible light. The simultaneous fulfillment of suitable band gaps and band alignments resulted in 21 promising 2DO photocatalysts. Upon calculating the solar-to-hydrogen efficiencies for all the 21 2DO candidates, the highest efficiencies reached up to ~18% (theoretical limit) for $HfSe_2$ and $ZrSe_2$. Our proposed approach utilizing elemental and chemical hardness features can also discover other promising materials for a range of applications.

RESULTS

Database generation

LDHs have been widely employed for electrocatalytic reactions such as water oxidation[17] and have also been extensively utilized as photoelectrocatalysts[18]. In the LDH class of materials, metals can have three oxidation states: +2 (double hydroxide), +3 (oxyhydroxide), and +4 (double oxide) (Supplementary Fig. 1). The two ligands (or functional groups) have oxidation states of (−1, −1), (−1, −2), and (−2, −2) in double hydroxides, oxyhydroxides, and double oxides, respectively. By changing the combination of metals and ligands, there is a huge possibility of 2D materials (of the order of ~10^4). We have generated a new database of 2D materials, denoted as the 2DO database, which belongs to the 2D LDH class of materials. The structural analogs of the LDH class of materials in the 2DO database have been termed as octahedrane (Fig. 1a), octahedrene (Fig. 1b), and octahedryne (Fig. 1c), respectively. We have considered the metals, which most commonly exist in these oxidation states[19], shown using the periodic table in Fig. 1d. Similarly, the choice of ligands was limited to those having an oxidation state of either −1 or −2 (Fig. 1d). A total of 15, 18, and 16 metals with +2, +3, and +4 oxidation states, respectively, are included for generating the 2DO database. Similarly, 13 and 16 ligands with −1 and −2 oxidation states are employed. Additional constraint of charge neutrality leads to a total of 3099 2DO materials. The thermodynamic, dynamic and electronic properties have been calculated for all the 2DO materials.

Feature generation

In order to establish the structure-property relationship for the 2DO materials, we have generated their features/attributes. The simplest and the most-widely used elemental features are based on the composition of a material. The elemental features utilized in our study are based on Magpie data[20], such as electronegativity, electron affinity, ionization energy. They are denoted by **{E}** (151 attributes) and were generated using the Matminer package[21]. The 2DO materials can also be represented through structural features such as packing fraction, Voronoi tessellations[22], which were generated using the Catlearn package[23]. The 335 elemental and structural attributes are jointly denoted by **{E, S}**. The elemental and structural features are described in detail in the works of Ward et al.[20] and Ward et al.[22], respectively. Further, we have also utilized another set of features based on local chemical hardness to capture the interactions between metals and ligands. Pearson and Parr defined the chemical hardness as[24]:

$$\eta_S = 1/2\left(\frac{\partial^2 E}{\partial N^2}\right)_Z \approx 1/2(I_S - A_S) \tag{1}$$

where E, N, I_S, A_S, and Z are the total energy, number of electrons, ionization energy, electron affinity, and atomic number of the chemical species, respectively. Here, species refer to the cations (e.g., Fe^{+2}, Al^{+3}) or ligands (e.g., OH^-, O^{-2}) and not the elements present in a compound (e.g., Fe, Al, O, H). Local chemical hardness values have been routinely applied to check the stability of molecules, acid-base adducts, and coordination complexes[25,26]. According to the hard and soft acids and bases (HSAB) principle, a molecule composed of hard (soft) acid and hard (soft) base should be more stable than that formed by a hard (soft) acid and soft (hard) base. Here, the phrases 'hard' and 'soft' are based on the classification by Pearson, i.e., the chemically hard acids and bases have η_S values greater than 8.5 and 4.5, respectively. The species having lower than these η_S values are soft acids and bases[27]. Analogous to the case of molecules, it is expected that the chemical hardness can play an important role in determining the stability of the 2DO materials. Therefore, we utilized the local chemical hardness values and their derived arithmetic mean, geometric mean, and standard deviation as features. The η_S values, calculated using the experimental ionization energies and electron affinities of the corresponding species (Eq. (1)), have been taken from several works by Pearson[24,27,28]. Accurate I_S and A_S can also be calculated by using DFT with large basis sets such as 6-311++G and cc-pVTZ, leading to the obtained η_S values in good agreement with the actual values[29,30]. The geometrical mean of chemical hardness (GM(η)) and mean difference in the chemical hardness of metal and ligands ($\overline{\Delta\eta}$) were evaluated through the

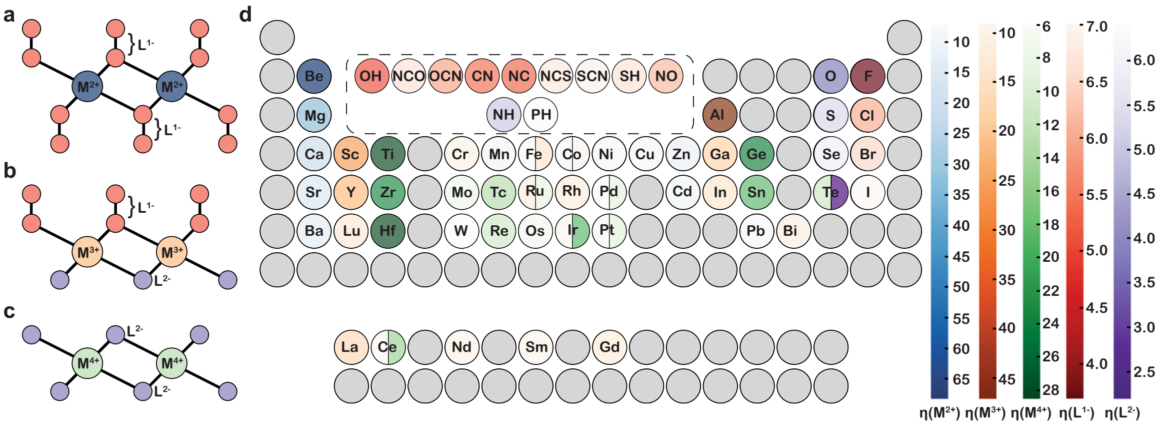

Fig. 1 The 2DO database. Structural representation of 2DO materials consisting of (**a**) octahedrane, (**b**) octahedrene, and (**c**) octahedryne. **d** Periodic table showing Pearson chemical hardness values of all metal cations and ligands (legend bars on the right side) used in the present study.

following expressions:

$$\mathrm{GM}(\eta) = \sqrt[3]{\eta_{\mathrm{M}}\eta_{\mathrm{L}_1}\eta_{\mathrm{L}_2}} \tag{2}$$

$$\overline{\Delta\eta} = \frac{(\eta_{\mathrm{M}} - \eta_{\mathrm{L}_1}) + (\eta_{\mathrm{M}} - \eta_{\mathrm{L}_2})}{2} \tag{3}$$

where η_{M}, η_{L_1}, and η_{L_2} are the Pearson chemical hardness of metal cation and the two ligands bonded to the metal cation, respectively. Combining elemental and chemical hardness features (Supplementary Table 1) resulted in 162 attributes represented by {$\mathbf{E}$, $\boldsymbol{\eta}$}. Further, the values of η_{S} for the hard cations or acids in our database span a wide range (8.5–67.8), while that of the soft cations and hard (soft) ligands or bases span very narrow ranges, i.e., 5.5–8.5 and 4.5–7.0 (2.2–4.5), respectively.

ML workflow

For each target properties (ΔE_{f}, ΔE_{hull}, overall stability), we employed a common methodology for ML as outlined in Fig. 2. Since the number of features generated is >100 in each of the feature sets, only the most prominent features have to be selected to increase the speed of ML algorithms and to prevent overfitting. For this purpose, mean feature ranking[31] have been performed, for which several types of ML algorithms have been utilized for measuring scores of all the features. These ML algorithms include Random Forests (RF), linear regression (or logistic regression for classification), least absolute shrinkage and selection operator (LASSO), recursive feature elimination (RFE), and extreme gradient boosting (XGBoost)[32]. This approach samples important features in a better way as each type of ML algorithm measures the correlation between each feature and the target variable in a unique manner. The different ways in which each method ranks the features are described in Supplementary Note 1. All features could be ranked differently by each algorithm and the features with the highest mean ranking should be selected. The top 15 features receiving the highest mean scores have been chosen. In order to ensure that they are not linearly correlated, we calculated Pearson's correlation coefficients (p) between any two features from the list. The p is defined as:

$$p = \frac{\mathrm{cov}\,(x_i, x_j)}{\sigma_{x_i}\sigma_{x_j}} \tag{4}$$

where cov(x_i, x_j) and $\sigma_{x_{i/j}}$ are the covariance of features x_i and x_j and standard deviation of the feature $x_{i/j}$, respectively. From the pair of features having $|p| > 0.80$, we selected the feature having a higher mean score. The selected features have been utilized in the rest of the ML studies.

After feature engineering, the performance of various ML algorithms are compared using the PyCaret package[33]. The PyCaret evaluates different models using default hyperparameters (parameters of a model initialized before training) and 10-fold cross-validation. It sorts their performance according to the

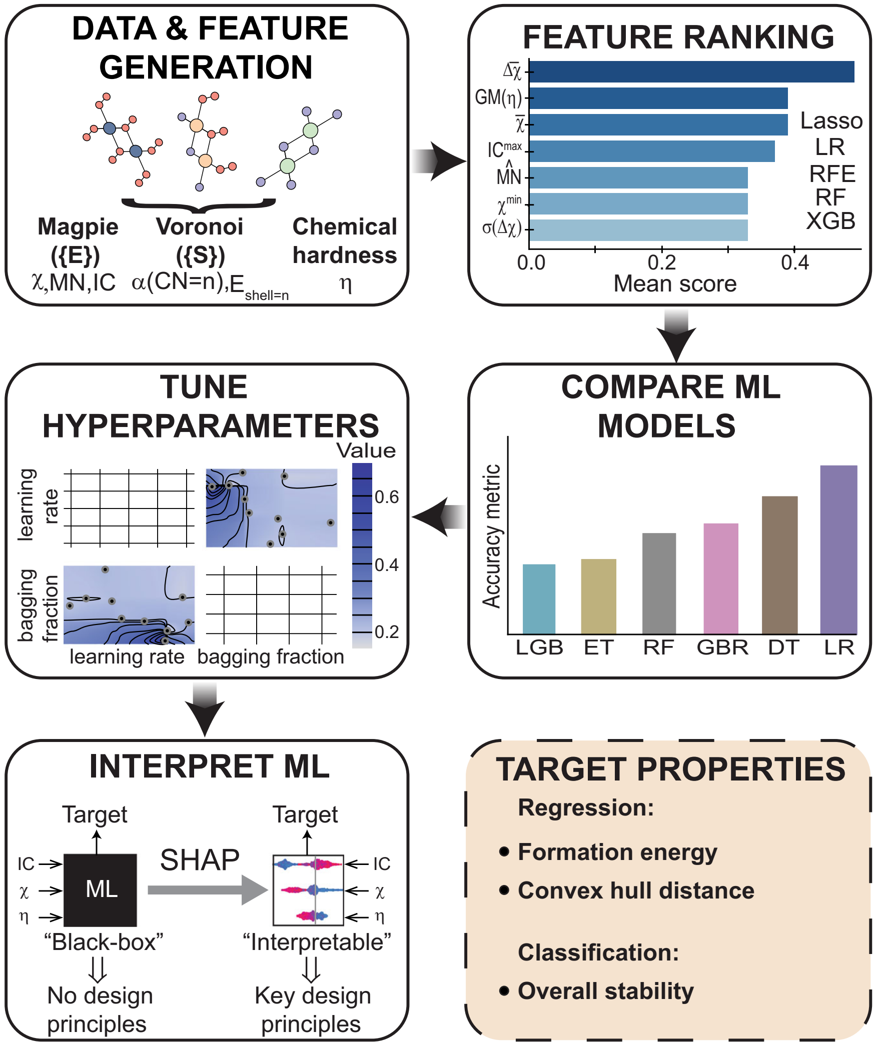

Fig. 2 ML workflow. Schematic of the workflow for machine learning applied in the current study.

desired metric such as root-mean-square error (RMSE). After the selection of the best ML algorithm, their hyperparameters need to be optimized, since using default values may not lead to optimal performance. The conventional methods such as random search or grid search methods are time-consuming and it is usually not guaranteed that the best hyperparameters have been found within the search space provided. In order to circumvent this problem, we utilized Bayesian optimization of hyperparameters using Tree-structured Parzen Estimator (TPE) algorithm as implemented in the Optuna package[34]. The process of Bayesian hyperparameter optimization has been explained in detail in Supplementary Note 2. Their optimization process can be summarized through hyperparameter importance, slice, and contour plots, which have also been explained in Supplementary Note 2.

Several accuracy metrics such as coefficient of determination (R^2), RMSE, mean absolute error (MAE), and Mathews correlation coefficient (MCC) (for classification) were evaluated for the ML model obtained after hyperparameter optimization over 2000 random trials for train-test split. The ML model corresponding to the random trial yielding minimum train/test RMSE and maximum train/test R^2 (or MCC score for classification) was selected. Furthermore, to reveal the effect of each feature on ML-predicted target values globally and locally, several types of SHAP plots such as feature importance, dependence, individual, and multioutput plots were generated using the SHAP package[35] (details of SHAP described in Supplementary Note 3).

Formation energy as a criterion for thermodynamic stability

As a first step in the HT screening, we checked the thermodynamic stability of the 2DO materials. Formation energy (ΔE_f) is the metric that is applied universally to evaluate the thermodynamic stability. The general expression for calculating ΔE_f is:

$$\Delta E_f = E_{2D} - \frac{E_M}{n} - \sum_i E^i_{ref} \tag{5}$$

where E_{2D}, E_M, n, and E^i_{ref} are the total energies of the 2D material, bulk metal, number of metal atoms in unit cell, and reference molecules or compounds, respectively. All the structures of the metals and reference molecules or compounds have been chosen in their respective standard states. For instance, E^i_{ref} for the octahedrane $Ni(OH)_2$ is the total energies of oxygen and hydrogen molecules. Generally, the PBE functional along with some corrections for reference molecules are utilized for calculating the formation energies[36,37]. They introduce several ambiguities and may not be reproducible to a different dataset[38]. Recently, the SCAN functional[39] under the meta-GGA (MGGA) approximation[40] was developed, which has been found to be very accurate for formation energies, compared to other commonly used functionals[41–43]. Therefore, the PBE-relaxed structures were further optimized using the SCAN functional.

The trends in the ΔE_f values for 3099 2DO materials are difficult to be established through simple data analysis. Hence, to gain insights into the factors governing ΔE_f and for accelerating its prediction, the ML scheme was applied for the prediction of ΔE_f as the target property. Feature ranking, including removal of highly correlated features, was performed for all the three feature sets. The best features obtained are shown in Supplementary Figs. 2a, 4a, and Fig. 3a and listed in Supplementary Table 1. Interestingly, the GM(η) values received the second best mean score, while the local chemical hardness features were not selected. GM(η) is equivalent to the global hardness value (i.e., η values corresponding to the material)[44,45]. We also checked the effect of changing the number of features on the performance of the ML models

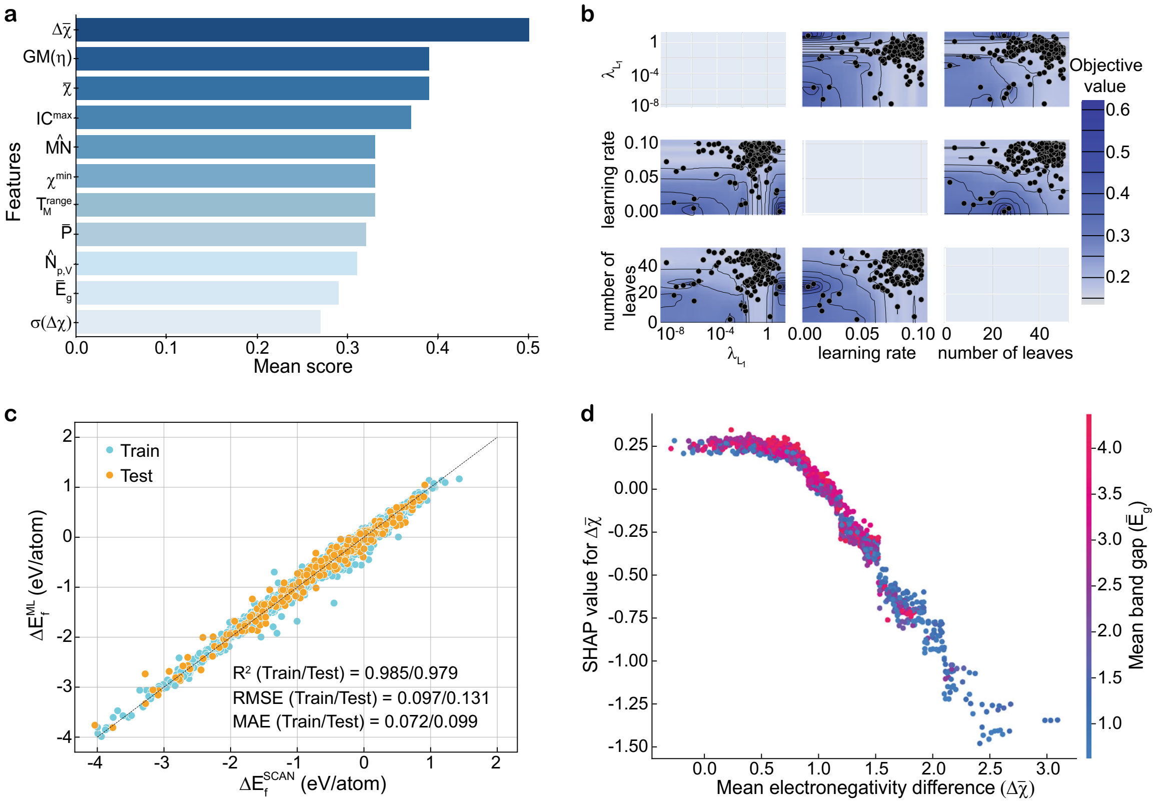

Fig. 3 **ΔE_f regression analysis. a** Mean score of best features selected from the {**E**, ***η***} feature set (highly correlated features removed) for ΔE_f^{ML}. **b** Contour plot of best three hyperparameters utilized in the ML(LGBM) model using the selected {**E**, ***η***} features for ΔE_f regression. **c** Parity plot of DFT and ML(LGBM)-predicted formation energies using the selected {**E**, ***η***} features. **d** SHAP dependence plot for $\overline{\Delta\chi}$ feature for ΔE_f^{ML} ({**E**, ***η***}).

(Supplementary Table 2). Decreasing the number of features leads to an underperformed ML model, while no significant changes are observed on increasing the number of features. Hence, in order to have a balance in the speed and accuracy of ML models, the number of features selected from the feature ranking have been fixed at 15 in the rest of the ML studies.

The selected attributes were employed to check the performance of different ML algorithms separately for each feature set using the PyCaret package. LightGBM (LGBM) emerged as the best ML algorithm for all feature sets as shown in Supplementary Figs. 3, 5, and 8. LightGBM[46] is a gradient boosting framework, which performs ensemble learning of decision trees, similar to several other ML methods such as XGBoost and CatBoost. It uses leaf-wise tree growth while others use depth-wise tree growth. This renders the leaf-wise gradient boosting algorithm to converge faster than the depth-wise growth.

After selecting the best algorithm, its hyperparameters were optimized over 200 random trials with test RMSE as the objective using the Optuna package. A total of six hyperparameters were used in the LightGBM models including learning rate, number of leaves, minimum child samples, bagging fraction, feature fraction, λ_{L_1}, λ_{L_2}, and bagging frequency. Each of these hyperparameters has been explained in detail in Supplementary Note 4. The learning rate, λ_{L_1}, and the number of leaves emerged as best from the hyperparameter importance plot for the ML(LGBM) model (Supplementary Fig. 9a). The slice plots for all the six hyperparameters are shown in Supplementary Fig. 9b–i. The contour plot in Fig. 3b shows that the maximum number of trials yielding the lowest RMSEs exhibit high learning rate (~0.1) for a wide range of λ_{L_1} (10^{-4}–1) and number of leaves (25–50).

The optimized hyperparameters were used for selecting the best model in terms of lowest RMSE and highest R^2 from 2000 random train-test splits. We compared the performance of the best ML(LGBM) models built using the three feature sets. The ML (LGBM)-predicted ΔE_f using **{E}** alone are found to be slightly superior than using **{E, S}**, having comparable test RMSEs (0.139 and 0.142 eV atom^{-1}) (Supplementary Fig. 6a and b). It is in agreement with the previous reports, which show that utilizing only elemental features is sufficient for predicting ΔE_f[47,48]. However, the ML models built using only **{E}** features hit a roadblock, when applied for predicting target properties of isocompositional compounds. Such compounds are present in the 2DO database in the form of 2D linkage isomers such as Cd$(CN)_2$ and Cd$(NC)_2$, which have same composition (Cd, 2C, and 2N for both compounds), but the ligands have different connectivities to metals (through C and N in Cd$(CN)_2$ and Cd$(NC)_2$, respectively). On comparing the ML(LGBM)-predicted ΔE_f using **{E}** and **{E, S}** for the 2DO linkage isomers, it is found that their $\Delta E_\mathrm{f}^\mathrm{ML}$ (**{E}**) values are same (Supplementary Table 3). The **{E, S}** feature set can clearly distinguish the ΔE_f for such compounds. Therefore, only elemental features may not be applicable for all types of materials. However, even the structural features such as $\delta\chi^{\mathrm{min}}_{\mathrm{shell}=1}$ (minimum electronegativity difference in 1st shell neighbors; Supplementary Table 1) are not easy to interpret.

Hence, we used the {$\mathbf{E}, \boldsymbol{\eta}$} feature set for predicting ΔE_f values as an alternative for the elemental and structural features. This feature set can also distinguish the ΔE_f for the 2DO linkage isomers. For instance, in the case of linkage isomers AlO(NCS) and AlO(SCN), the ligands NCS$^-$ and SCN$^-$ are comparatively harder and softer, respectively, on account of N being smaller and more electronegative than S. Therefore, Al^{+3} (chemically hard) cation will prefer NCS$^-$ over SCN$^-$, and hence has more exothermic ΔE_f. This trend is aptly captured by the {$\mathbf{E}, \boldsymbol{\eta}$} feature set. Hence, it is more intuitive than the **{E, S}** feature set. Moreover, only connectivity information is needed in {$\mathbf{E}, \boldsymbol{\eta}$} features, while complete structural information is necessary for generating the **{E, S}** feature set. The ML(LGBM) model utilizing the selected {$\mathbf{E}, \boldsymbol{\eta}$} features also performs best when compared to the other two feature sets. The test RMSE, MAE, and R^2 were obtained to be 0.131, 0.099, and 0.979 eV atom^{-1}, respectively (Fig. 3c), which is comparable to or better than the previously reported accuracy metrics for $\Delta E_\mathrm{f}^\mathrm{ML}$ of 2D materials, e.g., Schleder et al. (RMSE = 0.205 eV atom^{-1})[49], and Siriwardane et al. (MAE = 0.083 eV atom^{-1})[50]. Hence, we utilized only {$\mathbf{E}, \boldsymbol{\eta}$} in the rest of our ML studies. Further, in order to confirm that the amount of training sample is sufficient for developing the ML models, learning curves have been plotted for the ML(LGBM) model with RMSE and R^2 as the performance metrics in Supplementary Fig. 10. It is clear that both the train and test scores (both RMSE and R^2) converge at a training size of 90%. Therefore, the training size has been fixed at 90% of the total data throughout the manuscript.

Most ML models are 'black-boxes', as the input-output processes are opaque, inhibiting the rationalization of the underlying physics[51]. Interpreting such black-box models is necessary to develop new theories and design principles for accelerating the discovery of promising materials[52]. For this purpose, we utilized SHAP values for both global (feature importance and dependence plots) and local (individual plots) interpretability. The feature importance plots for $\Delta E_\mathrm{f}^\mathrm{ML}$({$\mathbf{E}, \boldsymbol{\eta}$}) are shown in Supplementary Fig. 11a and b, where $\overline{\Delta\chi}$ (mean electronegativity difference) is found to be the most important feature. $\overline{\Delta\chi}$ has positive impact (correlation) on the model output for its lower values and negative impact for its higher values (Supplementary Fig. 11a). This trend can be confirmed by the dependence plot for the $\overline{\Delta\chi}$ feature (Fig. 3d). A SHAP feature dependence plot shows a variation of the SHAP values of a feature (e.g., $\overline{\Delta\chi}$) against the given and another interacting feature (e.g, $\overline{E}_\mathrm{g}$) values, where the color bar indicates interaction between the two features. The SHAP selects the feature that exhibits maximum interaction effects with the given feature (i.e., $\overline{\Delta\chi}$)[53]. Therefore, the dependence of the target variable can be explained in terms of both the chosen and the interacting features.

According to Fig. 3d, $\Delta E_\mathrm{f}^\mathrm{ML}$ is almost independent of $\overline{\Delta\chi}$ for values up to 1.0, subsequently decreasing monotonically with respect to the higher values of $\overline{\Delta\chi}$ and lower values of $\overline{E}_g$. It can be attributed to the ionic character (IC) of a compound varying as a function of the electronegativity difference according to the relation:

$$\mathrm{IC}\,\% = \left[1 - e^{-0.25(\chi_\mathrm{A}-\chi_\mathrm{B})^2}\right] \times 100 \qquad (6)$$

where $\chi_\mathrm{A/B}$ is the electronegativity of the element A or B. Therefore, increasing electronegativity difference leads to higher ionicity in the compounds. It is also well known that ionic compounds are mostly composed of hard acids and hard bases, thereby leading to higher stability according to the HSAB principle[28]. For lower values of $\overline{\Delta\chi}$ (<1.0), covalent character will be predominant, which could be due to either soft-soft or hard-soft type of interactions. It may result in $\Delta E_\mathrm{f}^\mathrm{ML}$ being independent of the $\overline{\Delta\chi}$ in this range. In other words, the $\overline{\Delta\chi}$ has implicit chemical hardness behavior. For the case of GM(η), $\Delta E_\mathrm{f}^\mathrm{ML}$ mostly decreases from 4 to 5, then increases from 5 to ~6, after which no clear relationship with GM(η) is observed (Supplementary Fig. 12a). Therefore, the dependence plot for GM(η) does not seem to follow any general trend. However, interestingly, most of the 2DO compounds following a decreasing trend for the GM(η) values of 4–5, have very high values (>3000 K) of the interacting feature $T_\mathrm{M}^\mathrm{range}$ (range in melting points). Upon closer inspection, such 2DO compounds are found to consist of soft ligands like NCO^{-1}, OCN^{-1}, SCN^{-1}, and NCS^{-1} and soft metal cations. Therefore, many 2DO materials with GM(η) values up to ~5 have negative SHAP values, leading to negative impact on the ML output (i.e, lower $\Delta E_\mathrm{f}^\mathrm{ML}$ values). Hence, these compounds have exothermic $\Delta E_\mathrm{f}^\mathrm{ML}$, on account of the soft-soft interactions. For higher values, the 2DO materials are not guaranteed to have only soft-soft or hard-hard type of interactions, due to the range of η_S for hard cations being largest compared to the other type of species. Hence, the negative SHAP values are observed for only few 2DO materials having GM (η) in the range of 8–14, indicating possible hard-hard interactions. The features $\overline{\Delta\chi}$ and GM(η) can identify the 2DO materials composed of hard-hard and soft-soft species, respectively, having the most exothermic ΔE_{f} values. The dependence plots for the other features $\bar{\chi}$ (mean electronegativity), and $\mathrm{IC}^{\mathrm{max}}$ (maximum ionic character) are shown in Supplementary Figs. 12b and c, respectively. $\Delta E_{\mathrm{f}}^{\mathrm{ML}}$ shows a inverse dependence on $\bar{\chi}$ up to its value of 3.0 and a direct dependence thereafter. Since elements with higher electronegativity also have higher η values, the 2DO compounds lying in the region of ~3 values of $\bar{\chi}$ are composed of hard-hard species. Hence, these compounds have the most exothermic $\Delta E_{\mathrm{f}}^{\mathrm{ML}}$ values, in agreement with the insight obtained from the discussion on $\overline{\Delta\chi}$ feature. The trends in the feature dependence plots of $\mathrm{IC}^{\mathrm{max}}$ and $\overline{\Delta\chi}$ are similar, mainly differing in the feature values at which the relationships with $\Delta E_{\mathrm{f}}^{\mathrm{ML}}$ change. $\Delta E_{\mathrm{f}}^{\mathrm{ML}}$ is independent of $\mathrm{IC}^{\mathrm{max}}$ for the 2DO compounds with lower values (<0.2) of IC^{max}, decreases up to ~0.7, and finally increases. It may be due to $\Delta\chi$ and IC being interrelated to each other (Eq. (6)).

Furthermore, the individual SHAP plots depicting local interpretability are shown for linkage isomers of the octahedrane Mg(NCS)(NCO) in Fig. 4. The features such as χ^{min} (minimum electronegativity), $\mathrm{IC}^{\mathrm{max}}$, $\bar{P}$ (mean group number), and $\bar{\chi}$ push the predicted value towards the lower (or left) side of the base value (mean of the target property over train data), while other features drive the base value towards its higher (or right) side. For example, $\mathrm{IC}^{\mathrm{max}}$ has overall negative impact on model output and its value (0.678) for the linkage isomers of Mg(SCN)(OCN) is higher than its average value over the train data (0.462) (Supplementary Fig. 11b), hence it pushes the base value towards left. The values of all features except that of GM(η) are same for all the four linkage isomers, therefore affect the $\Delta E_{\mathrm{f}}^{\mathrm{ML}}(\{\mathbf{E}, \boldsymbol{\eta}\})$ in exactly the same way. The usage of such elemental features alone cannot distinguish the target properties of the linkage isomers. Only GM (η) helps the ML(LGBM) model in differentiating the predicted ΔE_{f} values for such compounds. For instance, while GM(η) has negligible effect on $\Delta E_{\mathrm{f}}^{\mathrm{ML}}$ for Mg(NCS)(NCO) (Fig. 4b), it has more positive impact for Mg(NCS)(OCN) (Fig. 4d) and Mg(SCN)(OCN) (Fig. 4h), leading to their predicted values being more closer to the corresponding true ΔE_{f} values (Supplementary Table 3).

Convex hull as a criterion for thermodynamic stability

ΔE_{f} is not the sole criterion for determining the thermodynamic stability of materials. Convex hull distance (ΔE_{hull}), which is defined as the decomposition energy of a particular phase into most stable phases and is equal to the distance from the convex hull line[38], is another metric for evaluating thermodynamic stability. Considering them together is a more stringent criterion than by using only formation energies. For instance, a 2D material lying above the convex hull may decompose into its competing bulk or other 2D phases even if it has exothermic formation energy. The convex hull for all the 3099 2DO materials were constructed from the 1418 most stable bulk phases corresponding to their elemental compositions, extracted from the Open Quantum Materials Database (OQMD)[54]. The ΔE_{hull} values were determined relative to the energies of competing bulk and 2D materials by utilizing the Pymatgen[55] package. The energies of the bulk compounds obtained from OQMD were recalculated first using PBE and then using SCAN functional with same parameters as used for the 2D materials. The convex hull constructions have been shown for few binary and ternary compounds in Supplementary Fig. 13. All compounds lying on the convex hull line or at distances <0.2 eV atom^{-1} from the line are considered thermodynamically stable.

We applied our ML scheme to these ΔE_{hull} values as the target property using only {$\mathbf{E}, \boldsymbol{\eta}$} feature set. The best features selected

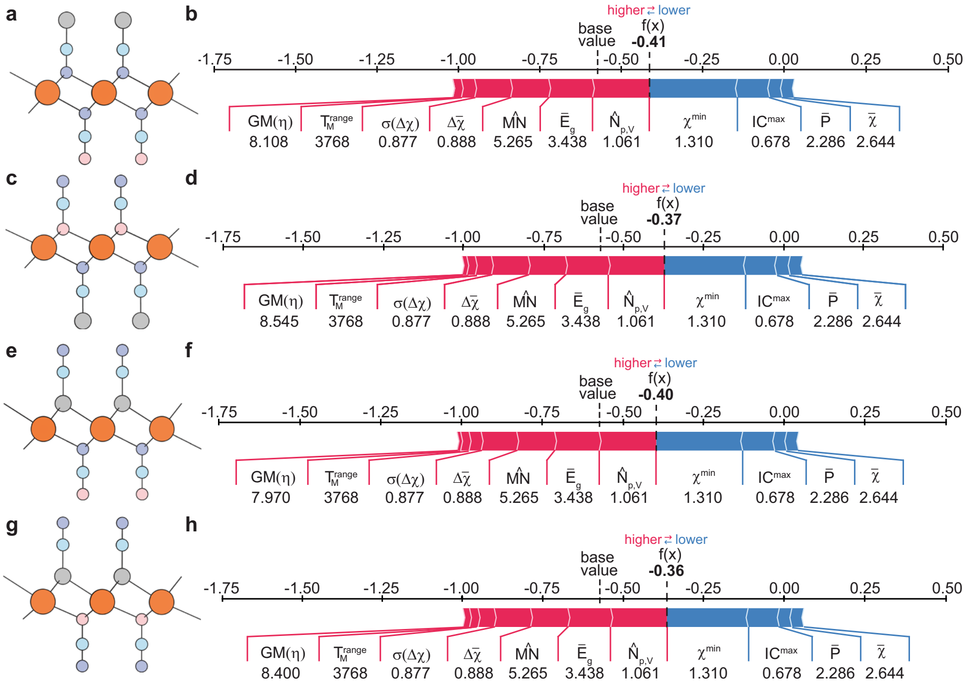

Fig. 4 Local interpretability into ΔE_{f} regression using SHAP. Optimized structures and individual SHAP plots using the $\Delta E_{\mathrm{f}}^{\mathrm{ML}}(\{\mathbf{E}, \boldsymbol{\eta}\})$ for (**a**) and (**b**) Mg(NCS)(NCO); (**c**) and (**d**) Mg(NCS)(OCN); (**e**) and (**f**) Mg(SCN)(NCO); (**g**) and (**h**) Mg(SCN)(OCN) compounds. The numbers below the features denote their values for a particular observation, predicted values are shown under `f(x)', and base value (−0.572 eV atom^{-1}) is the mean of the model output over the train data. The features that push the predicted value higher (to the right) are shown in red, and those pushing the prediction lower (to the right) are shown in blue. For a particular observation, if the value of a feature having positive (negative) impact on model output is more than its mean value, then it will push the base value towards right (left).

from mean feature ranking after removing highly correlated features are shown in Supplementary Fig. 14a. Using PyCaret, LightGBM again emerged as the best algorithm (Supplementary Fig. 15). The three best hyperparameters obtained using Optuna are minimum child samples, learning rate, and feature fraction (Supplementary Fig. 16a). The contour plot for the three hyperparameters and the slice plots for all hyperparameters are shown in Supplementary Figs. 16b and 17, respectively. The regions yielding the lowest RMSEs have become narrower for the pairs of minimum child samples and feature fraction with the learning rate. The best ML(LGBM) model was obtained by optimizing the accuracy metrics for 2000 random train-test splits, using the obtained hyperparameters. The model achieves a good accuracy as it has a test MAE, RMSE, and R^2 values of 0.066, 0.090 eV atom^{-1}, and 0.896, respectively (Supplementary Fig. 18a). To the best of our knowledge, there is only one other study reporting ML-predicted ΔE_{hull}, by Bartel et al. with a best MAE of 0.06 eV atom^{-1} for bulk materials using the composition-based features[48]. However, it is to be noted that the ML(LGBM) model does not perform very well for the data points having ΔE_{hull} values >2 eV atom^{-1}. It may be because there are only 13 such data points compared to 3085 data points existing in the range of 0–1.5 eV atom^{-1}. Nevertheless, the 2DO materials in the higher ΔE_{hull} value region are very unstable and unlikely to be synthesized. The only area of interest is between 0 and 0.2 eV atom^{-1}, where the ML(LGBM) model performs really well. We also do not find any case in which ΔE_{hull} in the region of >2 eV atom^{-1} are predicted within the 0–0.2 eV region and vice versa. It establishes the efficacy of the developed ML model for predicting ΔE_{hull}.

We calculated the SHAP values to gain insights into the best ML (LGBM) model. The SHAP feature importance plot shown in Fig. 5 depicts that $\hat{N}_{\mathrm{p,V}}$ (average deviation in the number of valence *p* electrons) is the most important feature followed by $\eta_{\mathrm{L}}^{\mathrm{max}}$ (maximum chemical hardness of ligand) and $\eta_{\mathrm{L}}^{\mathrm{min}}$ (minimum chemical hardness of ligand). The lower and higher values of $\hat{N}_{\mathrm{p,V}}$ have positive and negative impacts on the ML(LGBM) model, respectively, which is also verified by its SHAP dependence plot shown in Supplementary Fig. 19a. Its slope increases up to the $\hat{N}_{\mathrm{p,V}}$ value of 1.0 (hence, $\Delta E_{\mathrm{hull}}^{\mathrm{ML}}$ will also increase), subsequently the slope starts to decrease. For the case of $\eta_{\mathrm{L}}^{\mathrm{max}}$, most of its lower values have a significant negative impact on the ML(LGBM) model and the higher values have a positive impact. The exact opposite trend is shown by the $\eta_{\mathrm{L}}^{\mathrm{min}}$ feature, also captured by their corresponding dependence plots (Supplementary Figs. 19b and c). Further, lowest values of $\Delta E_{\mathrm{hull}}^{\mathrm{ML}}$ will occur for the points at which all the features have highest negative impact on the ML(LGBM) model. For example, few 2DO materials having $\hat{N}_{\mathrm{p,V}}$ values in the range of 1.80–2.25 lead to very low (<0.1) $\Delta E_{\mathrm{hull}}^{\mathrm{ML}}$ values. The $\hat{N}_{\mathrm{p,V}}$ values of ~2 are obtained for ligands such as OH^{-1}, SH^{-1}, O^{-2}, S^{-2}, Se^{-2}, F^{-1}, Cl^{-1}, Br^{-1}, and I^{-1}, when they are not attached to *p*-block metals. The hard ligands such as OH^{-1}, O^{-2}, F^{-1}, and Cl^{-1}, when bonded to hard metal cations, yield low $\Delta E_{\mathrm{hull}}^{\mathrm{ML}}$ values. Similarly, the combination of soft ligands such as SH^{-1}, S^{-2}, Se^{-2}, Br^{-1}, and I^{-1} and soft metal cations result in lower convex hull distances. A few 2DO materials having $\eta_{\mathrm{L}}^{\mathrm{max}}$ and $\eta_{\mathrm{L}}^{\mathrm{min}}$ values in the range of 5–7 (hard ligands) or ~4 (soft ligands), also lead to negative SHAP values, indicating higher thermodynamic stability. The corresponding interacting features ($\hat{\chi}$ and $\overline{N}_{\mathrm{U}}$) have high numerical values in these regions. These findings signify that both soft and hard ligands lead to lower ΔE_{hull} values when they interact with soft and hard metal cations, respectively. Furthermore, the individual SHAP plots for the 2DO linkage isomers of Mg (NCS)(NCO) are shown in Supplementary Fig. 20. It was again found that only the chemical hardness-based features ($\eta_{\mathrm{L}}^{\mathrm{max}}$ and $\eta_{\mathrm{L}}^{\mathrm{min}}$) enable the ML(LGBM) model in discerning the target property (ΔE_{hull}) for these compounds.

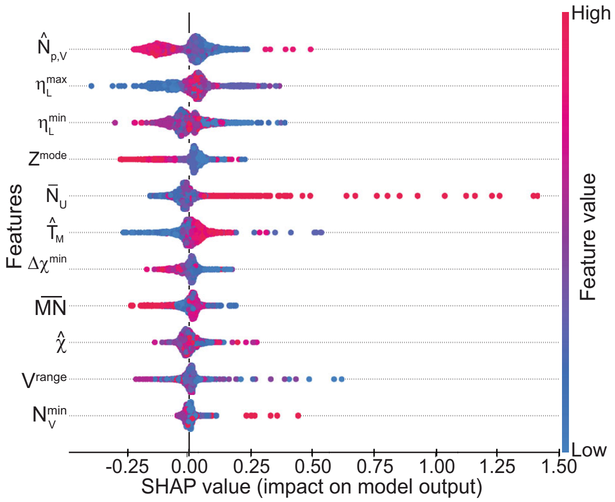

Fig. 5 Global interpretability into ΔE_{hull} regression using SHAP. SHAP feature importance plot for $\Delta E_{\mathrm{hull}}^{\mathrm{ML}}(\{\mathbf{E}, \boldsymbol{\eta}\})$ (LGBM model), arranged in the order of their decreasing importance.

Classification of overall stability

A concern with most of the HT studies for identifying photocatalysts is that they include only their thermodynamic stabilities while ignoring dynamic stabilities[10,11]. The thermodynamic stability of a 2D material only specifies its energetic preference with respect to competing bulk and 2D compounds, where all the structures are considered in their frozen ground states. In experimental conditions, the atoms do not remain stationary, so the realizability of the 2D material is affected to a reasonable extent by its dynamical stability[56]. Generally, the elastic constants (C_{ij}) and minimum eigenvalue of dynamical matrix ($|\tilde{\omega}_{\mathrm{min}}^2|$) determine the dynamic stability of the 2D material. In order to consider both types of stabilities, we also calculated the elastic constants and $|\tilde{\omega}_{\mathrm{min}}^2|$ for all the 3099 2DO materials. The elastic constants are evaluated through the stress-strain relationship or generalized Hooke's law[57]:

$$\boldsymbol{\sigma} = \mathbf{C}\boldsymbol{\epsilon} \tag{7}$$

where $\boldsymbol{\epsilon} = \{\epsilon_1, \epsilon_2, \epsilon_3, \epsilon_4, \epsilon_5, \epsilon_6\}$ are the set of applied strains, $\boldsymbol{\sigma} = \{\sigma_1, \sigma_2, \sigma_3, \sigma_4, \sigma_5, \sigma_6\}$ are the corresponding set of stresses, and $\mathbf{C}$ is the 6x6 elastic constant matrix or the stiffness tensor. The internal routine of VASP was utilized for generating six finite lattice distortions of $2 \times 2 \times 1$ supercells of the 2DO materials and the elastic constants were subsequently derived from Eq. (7). The indices 1, 2, and 3 represent normal or planar strain/stress and the indices 4, 5, and 6 represent shear strain/stress. For 2D materials, only the planar strain/stress is to be considered in the *x* and *y* directions for the strain and stress tensors, resulting in three elastic stiffness coefficients—C_{11}, C_{22}, and C_{12}[38]. These elastic constants are then multiplied by the length of the *c*-axis to represent them in the units of Nm^{-1} [58]. Further, the density functional perturbation theory (DFPT) was utilized to obtain dynamical matrices of the $2 \times 2 \times 1$ supercells of all the 2DO materials. The dynamical matrix was diagonalized at the Γ-point to get eigenvalues, which is square of the mass-weighted phonon frequencies ($\tilde{\omega}$)[38]. The negative eigenvalues correspond to imaginary frequencies and hence signify dynamical instability. The workflow for the interpretable ML regression described in the preceding sections can be easily extended for C_{ij} and $\tilde{\omega}$ as the target properties.

Thygesen et al.[38] classified thermodynamic and dynamic stabilities of 2D materials into high, medium, or low according to a criterion based on the values of the above four properties as shown in Table 1. This criterion classifies the overall stability of 2D materials into nine classes – **Hh**, **Hm**, **Hl**, **Mh**, **Mm**, **Ml**, **Lh**, **Lm**, and **Ll**, where uppercase letters denote the thermodynamic stability and lowercase letters denote dynamic stability and **H/h** ≡ high, **M/m** ≡ medium, **L/l** ≡ low. The maximum number of 2DO materials in our database belong to the **Mm** class (49.21%), while other classes constitute remaining half of the database, with the **Ll** containing only 8 (0.26%) samples. Hence, the octahedral 2D materials database is highly skewed in terms of the overall stability classes.

In the previous sections, we established chemical hardness along with few elemental features to be the key factors deciding the thermodynamic stability of the 2DO materials. To check if they also affect the overall stability of the 2DO materials, we performed ML classification for the above nine classes using the workflow described earlier. In order to remove inherent bias in the imbalanced data, the lowest populated classes (**Hl** and **Lh**) were oversampled using the SMOTE method. It leads to a change in the class population of the train data after oversampling, as shown in Supplementary Table 4. This data has been subsequently utilized for choosing the best features from mean feature ranking. The less-correlated, important features are shown in Supplementary Fig. 21a, and their Pearson correlation coefficients in Supplementary Fig. 21b. The suitability of various ML algorithms were checked for the classification task using the selected features and the Extra Trees (ET) classifier emerged as the best model (Supplementary Fig. 22). The ET algorithm performs ensemble-based bagging of decision trees and is hence similar to the popular RF algorithm. However, the ET method does not bootstrap samples (i.e., sampling is done without replacement) and nodes are split on random rather than the best splits, thus differing from the RF[59]. For hyperparameter optimization, the MCC score for test data was selected as the objective. It led to the final ML(ET) model correctly classifying all the imbalanced classes better than those built using other metrics such as accuracy, precision, and F_1 score due to the highest number of true predicted cases (Supplementary Table 5). The contour, and slice plots for the hyperparameter optimization corresponding to the MCC score as the objective are shown in Supplementary Fig. 23 and described in Supplementary Note 5. The optimized hyperparameters were then used to maximize the MCC score over 200 random train-test splits and 10 SMOTE trials, leading to the best ML(ET) model. The confusion matrix and receiver operating characteristic (ROC) curve obtained from this model are shown in Fig. 6a and Supplementary Fig. 24, respectively. Further, the individual precision, recall and F_1 scores are shown in Supplementary Table 6. Most of the values for these metrics are >0.80 for the individual classes, depicting that the final ML(ET) model is fairly accurate for the imbalanced dataset.

As a next step, the SHAP values were evaluated for the best ML (ET) model. The feature importance plot, in this case, describes not only the impact of each feature on the overall model but also on each class separately as shown in Fig. 6b. $T_{\mathrm{M}}^{\mathrm{range}}$ feature affects the ML(ET) model to the highest extent. Among individual classes, the largest impact of each feature is on the highest-populated class **Mm**. Although the chemical hardness-related features—$\overline{\Delta\eta}$ and GM(η) have the lowest overall impact on the ML(ET) model, they exhibit no bias towards any class apart from the **Mm** class. The other features are biased towards at least one other class apart from **Mm**, e.g., $\overline{\mathrm{IC}}$ (mean ionic character) is more biased towards **Ml** and **Lm** classes. We also performed local interpretability of the ML(ET) model using SHAP multioutput decision plots (explained in Supplementary Note 6) for the 2DO linkage isomers $Ba(CN)_2$ and $Ba(NC)_2$ as an illustration (Supplementary Fig. 25). The octahedranes $Ba(CN)_2$ and $Ba(NC)_2$ belong to the classes **Mm** and **Mh**, respectively, on account of the NC^{-1} ligand being slightly harder than the CN^{-1} ligand. Utilizing only elemental features would misclassify both compounds to the same class. Therefore, upon utilizing the ML(ET) model with the selected {$\mathbf{E}, \boldsymbol{\eta}$} features, it was found that all classes except **Mh** and **Mm** receive zero probabilities (initial probability for each class = 1/9 = 0.11). The probabilities of classes **Mh** and **Mm** received for $Ba(CN)_2$ are 0.48 and 0.52, respectively, and hence it is assigned to the **Mm** class. On the other hand, the probabilities of classes **Mh** and **Mm** received for $Ba(CN)_2$ are 0.54 and 0.46, respectively, and hence it is

Table 1. Stability criteria and number of 2D materials belonging to each class of stability.

Dynamic (→) Thermodynamic (↓)	$\lvert\tilde{\omega}^2_{\mathrm{min}}\rvert < 10^{-5}$, $C_{ii} > 0$ (high)	$10^{-5} < \lvert\tilde{\omega}^2_{\mathrm{min}}\rvert < 2$, $C_{ii} > 0$ (medium)	$\lvert\tilde{\omega}^2_{\mathrm{min}}\rvert > 2$, $C_{ii} < 0$ (low)
$\Delta E_{\mathrm{f}} < 0.2$, $\Delta E_{\mathrm{hull}} < 0.2$ **(High)**	114	366	55
$\Delta E_{\mathrm{f}} < 0.2$, $\Delta E_{\mathrm{hull}} > 0.2$ **(Medium)**	154	1525	456
$\Delta E_{\mathrm{f}} > 0.2$, $\Delta E_{\mathrm{hull}} > 0.2$ **(Low)**	8	229	192

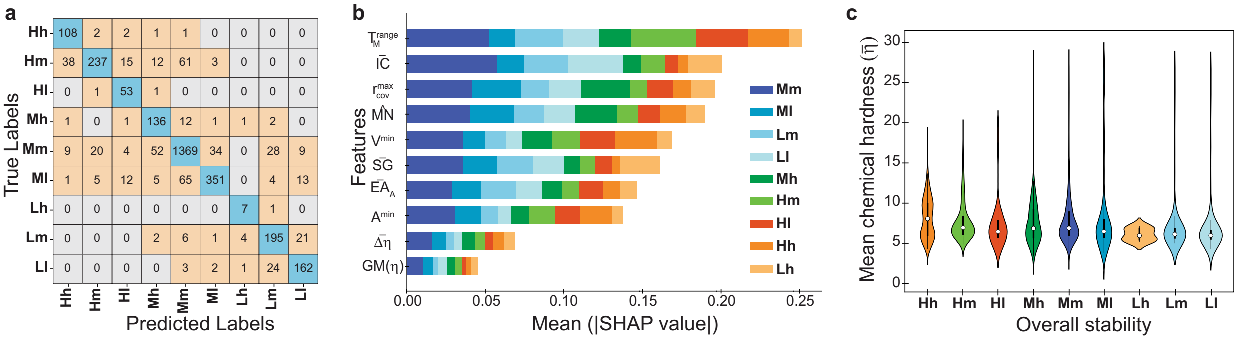

Fig. 6 Overall stability classification. **a** Confusion matrix for multiclass classification of overall stability using the ML(ET) model. The numbers in the cyan and orange-colored boxes represent the instances of correctly and incorrectly predicted samples. **b** SHAP feature importance plot for overall stability using the ML(ET) model. **c** Violin plots of $\overline{\eta}$ as a function of the overall stability classes. The outer shells, thick lines, and thin lines represent the probability densities of complete data, interquartile range (i.e., between 25% and 75% of data), and 1.5 times the interquartile range, respectively. The white circles represent the median values.

assigned to the **Mh** class. This is attributed to the contribution of the chemical hardness-based features $\overline{\Delta\eta}$ and GM(η) to the model being reversed for the two compounds (Supplementary Fig. 25a and b).

Finally, the profile of the overall stability of the 2DO materials as a function of a chemical hardness attribute has been shown using violin plots in Fig. 6c. The arithmetic mean of the local chemical hardness values ($\overline{\eta}$) are shown on the *y*-axis, and the overall stability classes on the *x*-axis. The median value for the **Hh** class is highest compared to other classes. It decreases from **Hh** to **Hl**, then increases slightly for the **Mh** class and remains constant for the medium thermodynamic stability classes, i.e., **Mh**, **Mm**, and **Ml**. Similarly, the medians of the violin plots remain constant for the low thermodynamic stability classes, but are lower than that of the medium thermodynamic classes. In particular, the violin plot for the **Hh** class shows two distributions (with high probabilities) above and below the median value, respectively, which are highest and lowest compared to that of the other classes. It depicts that these two distributions correspond to the 2DO materials composed of hard-hard and soft-soft interactions, respectively. This observation is in accordance with the HSAB principle, according to which a compound formed from hard-hard and soft-soft metal cations and ligands should be more stable than that from the hard-soft or soft-hard species. Therefore, the HSAB principle can predict the overall stability of the 2DO materials to a reasonable extent. Further, the lowering in median values observed for the classes **Hm** and **Hl** can be attributed to one of the components (metals or ligands) being softer, leading to a decrease in the dynamic stabilities.

Screening stable 2D photocatalysts

In the preceding sections, we discussed the first and essential criteria for screening a photocatalyst, i.e., its stability[60]. We next analyze the other critical factors related to the electronic structure of the 2D material, which decide whether it can enable photocatalysis. The process of photocatalysis usually occurs in three steps. The first step involves the absorption of light in the form of photons by the material. The energy of the photon to be absorbed should correspond to the band gap of the material. Hence, only semiconductors can act as photocatalysts. The electrons, upon photon absorption, get excited into the conduction band, leaving behind the oppositely charged holes in the valence band. The second step involves the migration of these photogenerated charge carriers to the active sites. Upon reaching the active sites, the electrons and holes are utilized in the reduction and oxidation reactions, respectively. However, the two reactions can proceed only if the energy of the electron (hole) is greater (lower) than the standard reduction (oxidation) potential of the reaction.

Therefore, the fundamental requirements for a photocatalyst is that it should be highly stable, should have band gap within the visible region and should satisfy the standard redox potentials of the the reaction under consideration (water splitting in our case). The energy gap (or band gap) between the highest occupied and the lowest unoccupied electronic energy levels is usually defined in two ways, i.e., optical and fundamental band gaps[61]. The optical gap (E_g^{opt}) of material corresponds to the energy of the lowest electronic transition, accessible via absorption of a single photon. The optical band gaps are obtained from the UV–visible (absorption) spectra using the Tauc method[62,63]. On the other hand, the fundamental band gap (E_g^{fund}) is the difference between ionization energy (I) and electron affinity (A) of the material. The fundamental band gaps are obtained by a combination of photoelectron spectroscopy and inverse photoemission spectroscopy[61]. Usually, the fundamental band gaps are larger than the optical band gaps, i.e., $E_g^{fund} \geq E_g^{opt}$. The band edges on the other hand are obtained through electrochemical methods such as the Mott-Schottky plot[64]. The band edge values obtained through such methods lie on the normal hydrogen electrode (NHE) scale. For obtaining experiment-level accuracy in the band gaps and band edges through computational methods, the GW approximation is the most preferred technique[65]. However, the band gaps obtained from the GW method correspond to the fundamental band gaps rather than the optical band gaps. Moreover, the positions of the band edges obtained from GW, followed by alignment with the corresponding vacuum levels, lie on the absolute potential scale. They can be rescaled to the NHE reference by adding a value of 4.44 eV[66].

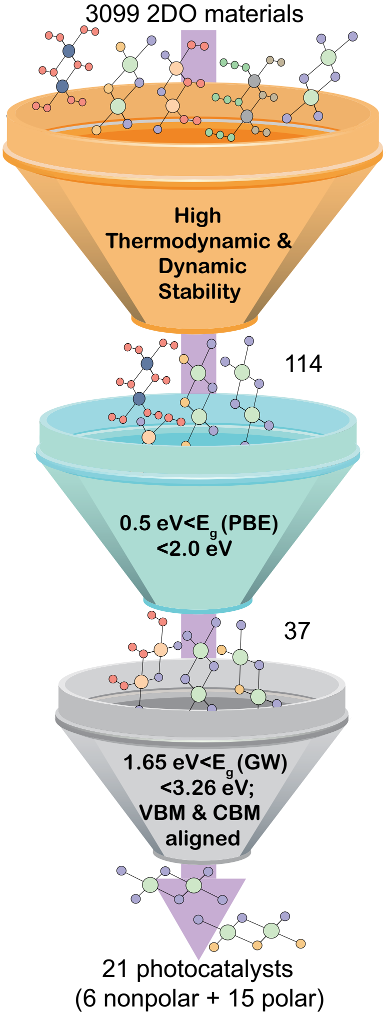
3099 2DO materials
High Thermodynamic & Dynamic Stability
114
0.5 eV<E$_g$(PBE) <2.0 eV
37
1.65 eV<E$_g$(GW) <3.26 eV; VBM & CBM aligned
21 photocatalysts (6 nonpolar + 15 polar)

Fig. 7 HT workflow. The HT scheme utilized in the current study for screening stable 2DO photocatalysts. The first tier screened 3099 2DO materials based on their overall stability. The second tier selected semiconductors from the **Hh** class having PBE band gaps between 0.5 and 2.0 eV. The selected 37 2DO materials were finally screened based on suitable GW band gaps and band alignments.

On the basis of the aforementioned conditions, we proceed towards second tier of HT screening of the 2DO materials as water-splitting photocatalysts. Among all the 2DO materials, only 114 compounds belonging to the **Hh** class were selected for further studies, as shown in Fig. 7. The PBE band structure calculations were performed for these compounds, and 37 2DO materials with PBE band gaps (E_g^{PBE}) in the range of 0.5–2.0 eV were selected. This criterion was chosen to ensure the 2DO materials to be optically active in the visible light region, consistent with previous HT studies on photocatalysis[10]. Further, it is expected that the chemical hardness may also have an effect

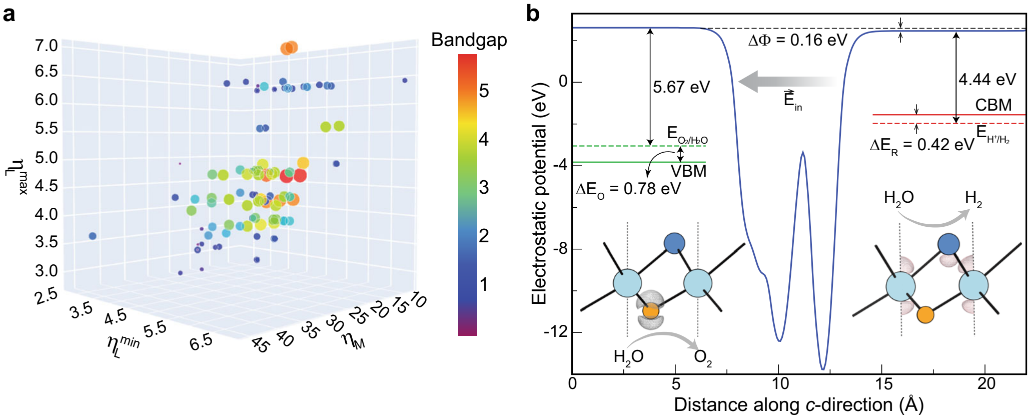

Fig. 8 Analysis of promising 2DO photocatalysts. **a** Variation of band gaps of 2DO semiconductors belonging to the **Hh** class as a function of η_L^{max}, η_L^{min}, and η_M. Both the color bar and circle-sizes are plotted according to their E_g^{PBE} values. **b** Electrostatic potential for the 2D intrinsically polarized material BiSeI. Band-decomposed charge densities corresponding to the VBM (gray isosurface) and the CBM (purple isosurface) of BiSeI are shown in left and right insets, respectively. The values for potentials of the photogenerated holes (ΔE_O) and electrons (ΔE_R) are also shown. The cyan, blue, and orange circles denote Bi, I, and Se atoms, respectively. The isosurface values was taken to be 0.02 eÅ^{-3}.

on the band gap of the 2DO materials. To verify this, the E_g^{PBE} values of the 2DO semiconductors belonging to the **Hh** class have been plotted against the η_S values of all three components (metal cation and ligands) in Fig. 8a. It is observed that the 2DO materials composed of both hard cations ($\eta_M > 8.5$) and hard ligands ($\eta_L > 4.5$) have large band gaps, those with hard-soft interactions have moderate band gaps, and those with soft-soft interactions have low band gaps. This finding has been further verified for all the semiconductors in our database (Supplementary Fig. 26). The observed trend in band gaps can be attributed to the compounds formed from hard-hard interactions being predominantly ionic in nature[28], leading to large band gaps. Therefore, the 2DO materials with soft-soft interactions or hard-soft interactions (where at least one component should be soft) will invariably result in lower band gaps, making them suitable candidates for photocatalysis. In general, the PBE-calculated band gaps are underestimated with respect to the experimental band gaps. In order to obtain highly accurate band gaps, the GW calculations were performed. The 2DO materials with the GW band gaps (E_g^{GW}) lying outside the visible region (1.65–3.26 eV)[67] were left out from this subset, resulting in 29 potential 2DO photocatalysts. All of these 29 2DO materials are either the octahedrenes or octahedrynes but not the octahedranes. Again, the reason can be attributed to the octahedrane materials in the **Hh** class being composed of either very hard or very soft metal cations leading to their band gaps lying outside the visible range.

Apart from the band gap, the other essential requirement for a material to host photocatalytic properties is that it should also satisfy the oxidation and reduction potentials of the water-splitting reaction. In order to find possible water-splitting photocatalysts among the 29 2DO materials, the positions of valence band maximas (VBMs) and conduction band minimas (CBMs) were compared with respect to the standard water oxidation (E_{H^+/H_2}) and water reduction (E_{O_2/H_2O}) potentials. Here, the GW-obtained VBM and CBM values were aligned with respect to the vacuum levels of the 2DO materials, calculated using the electrostatic potential method. The values of E_{H^+/H_2} and E_{O_2/H_2O} at pH = 0 with respect to the vacuum level are –4.44 and –5.67 eV[68], respectively. We found six nonpolar 2DO materials—HfS_2, ZrS_2, PtS_2, $HfSe_2$, $ZrSe_2$, and $PtSe_2$, on which both hydrogen evolution reaction (HER) and oxygen evolution reaction (OER) can take place (Supplementary Table 7). There are also several 2DO materials, in which the metal is attached to two different ligands such as BiSeI. For such 2D materials, vacuum levels corresponding to the top (001) and bottom (00$\bar{1}$) surfaces will be inequivalent, leading to an intrinsic electric field. They are called 2D intrinsically polarized materials[69,70]. Their E_{H^+/H_2} and E_{O_2/H_2O} levels have to be aligned against the vacuum levels of surfaces on which the conduction and valence band edges are localized. For instance, the reduction and oxidation levels for the octahedrene BiSeI will be aligned with respect to the vacuum levels of (001) and (00$\bar{1}$) surfaces due to the localization of CBM and VBM on the respective surfaces (Fig. 8b). Among the intrinsically polarized 2DO materials, further 15 candidates were found to be suitable for the complete photocatalytic water-splitting reaction (Supplementary Table 7). Such materials facilitate the HER and OER on different surfaces (e.g., HER on (001) and OER on (00$\bar{1}$) for BiSeI as shown in Fig. 8b). It leads to suppressed backward reaction of the evolved H_2 and O_2 gases[60] and charge recombination[70].

The photocatalytic activity of the 21 promising candidates can be compared based on solar efficiency metrics such as applied bias photon-to-current efficiency (ABPE), incident photon-to-current efficiency (IPCE), absorbed photon-to-current efficiency (APCE), and solar-to-hydrogen efficiency (η_{STH})[71]. Among all the four metrics, η_{STH} is the most vital metric to characterize a photocatalyst device, while the other metrics represent diagnostic efficiency measurements. It describes the overall efficiency of the photocatalyst when no external potential is applied[71]. It is defined as the ratio of the generated chemical energy to the input solar energy. The chemical energy produced equals the H_2 production rate multiplied by the standard Gibbs free energy change per mole of H_2. The input solar energy is the incident illumination power density multiplied by the electrode surface area, where the illumination source should closely resemble the shape and intensity of the Air Mass 1.5 Global (AM1.5G) standard[72]. The original definition of η_{STH} can be modified to depend on the band gap and band edges of the photocatalyst, according to the following expression[73]:

$$\eta_{STH} = \frac{\Delta G \int_E^\infty \frac{P(\hbar\omega)}{\hbar\omega} d(\hbar\omega)}{\int_0^\infty P(\hbar\omega) d(\hbar\omega) + \Delta\Phi \int_{E_g}^\infty \frac{P(\hbar\omega)}{\hbar\omega} d(\hbar\omega)} \tag{8}$$

where $P(\hbar\omega)$ is the AM1.5G spectral irradiance of the solar spectrum as a function of the photon energy $\hbar\omega$, ΔG is the standard Gibbs free energy change corresponding to the water splitting (1.23 eV) reaction, $\Delta\Phi$ is the difference in vacuum levels of the two surfaces, and E_g is the band gap of the 2D material. The term E determines the energy required by photons to drive the

HER and OER, and its evaluation is described in Supplementary Note 7. The η_{STH} values were calculated for all the 21 photocatalysts screened from the HT scheme using equation (8) to compare their overall efficiencies. The highest η_{STH} values were obtained for $HfSe_2$ and $ZrSe_2$ (both having η_{STH} of 17.14%), which is comparable to the theoretical limit (~18%) for the photocatalytic water-splitting reaction[73]. Furthermore, among the 21 selected 2DO materials, 1T-ZrS_2 ($\eta_{STH} = 2.03\%$) and 1T-$PtSe_2$ ($\eta_{STH} = 11.09\%$) have already been synthesized experimentally and proven to be suitable photocatalysts[74–76]. This validates our proposed approach for finding efficient 2D photocatalysts and establishes $ZrSe_2$ and $HfSe_2$ as the two most promising candidates for photocatalysis. Creating heterojunctions from the set of 21 2DO photocatalysts to form either type-II[7] or Z-scheme photocatalysts[77] would further increase their intrinsic photocatalytic efficiencies. Such type of photocatalysts mimic the behavior of 2D intrinsically polarized materials on different monolayers.

DISCUSSION

In summary, an iML-HT approach is developed towards establishing structure-stability relationships for finding efficient 2D photocatalysts. A new database called the 2DO database containing around 3000 octahedral 2D materials is developed based on 2D LDHs. The four physical properties— ΔE_f, ΔE_{hull}, C_{ij}, and $|\tilde{\omega}^2_{min}|$, which ultimately decide the thermodynamic and dynamic stability, are calculated for all the 2DO materials using the first-principles calculations. The factors governing the two fundamental physical properties (ΔE_f and ΔE_{hull}) are revealed through state-of-the-art ML techniques. The elemental and chemical hardness-based features help the ML models discerning the prediction of target properties for the 2DO linkage isomers, establishing their superiority over the structural features. We also performed ML multiclass (nine classes) classification of overall stability for the 2DO materials and found it to be reasonably accurate for the highly imbalanced dataset. Interestingly, we found that the maximum 2DO materials belonging to the **Hh** class comprised of hard-hard and soft-soft interactions. Hence, the stability of the 2DO materials is mainly governed by the HSAB principle, which is widely applied for molecules and transition metal complexes. Furthermore, all the highly stable 2DO materials are screened for their potential applications in photocatalytic water splitting, from which 21 potential candidates are selected, with the efficiencies of $HfSe_2$ and $ZrSe_2$ reaching the theoretical limit. The predicted efficiencies for these compounds are experimentally realizable. It is due to the existence of few photocatalysts having solar-to-hydrogen efficiencies close to the theoretical limit of 18%. For example, Fe_2O_3, Ta_3N_5, and TiO_2 nanotubes have solar-to-hydrogen efficiencies of 12.9%[72], 15.9%[72], and 16.0%[78], respectively.

The predicted PBE and GW band gaps for few 2DO materials are also compared with the experimentally observed band gaps (E_g^{expt}) in Supplementary Table 8. Except for $PtSe_2$, all other E_g^{expt} values are optical band gaps, while the GW method measures the fundamental band gaps. As discussed, the E_g^{fund} values are always greater than the E_g^{opt} values. This effect is more pronounced in the 2D materials, leading to very large exciton binding energies (difference in E_g^{fund} and E_g^{opt})[38,79]. Moreover, the calculated GW band gaps are in excellent agreement with previously reported values for few known 2DO compounds[38,79]. Apart from the basic requirements of suitable band gaps and band edges, a photocatalyst should also possess other desirable physical and chemical properties, such as high charge carrier mobilities for fast charge migration, large visible light-harvesting for practical applications, low charge recombination for efficient charge utilization, low tendency for back reaction and low overpotentials to have high catalytic activities. These properties can be easily evaluated through first-principles to verify the efficacy of the proposed 2D photocatalysts[7,60]. However, most photocatalysts suffer from high charge recombination in their bulk form, leading to reduced photocatalytic activity. The intrinsically polarized 2DO materials in our database reduces the charge recombination propensity in two ways—by decreasing the distance required for the photogenerated charge carrier to migrate to the active sites due to nanostructuring, and by helping the electrons and holes to accumulate at the opposite sides of the materials, leading to spatial separation. Hence, these 2DO materials are also expected to have efficient photocatalytic activities. In addition, our HT scheme can also help in the identification of photocatalysts suitable for other reactions such as carbon dioxide reduction reaction (CO_2RR)[80], nitrogen reduction reaction (NRR)[81], and pollutant degradation[2], which require further investigations. There is also a possibility for the 2DO materials belonging to the **Hm** (e.g., $Ni(OH)_2$[82]) and **Mh** (e.g., $Co(OH)_2$[83]) classes to act as potential candidates for photo(electro)catalysis. Apart from catalysis, the 2DO materials from our database can also be promising for applications in spintronics, 2D ferromagnets, quantum computers, and topological insulators, to name a few. This can be attributed to the wide range of band gaps from 0 to >5 eV exhibited by these compounds and the presence of magnetic and heavy atoms.

METHODS

Density functional theory calculations

All the first-principles calculations were performed using density functional theory (DFT) as implemented in the Vienna ab initio simulation package (VASP version 5.4.4)[84]. Electron-ion interactions were described by all electron projector augmented wave (PAW) pseudopotentials[85]. A vacuum of 20 Å was included along the *c*-direction to prevent interactions among the periodic images. Both ionic positions and cell shapes of the 2D structures were optimized using the Perdew-Burke-Ernzerhof (PBE) functional under generalized gradient approximation (GGA)[86]. The Brillouin zone was sampled by a 15 × 15 × 1 Monkhorst-Pack k-point grid. The plane waves with a kinetic energy cutoff of 500 eV were used in all the calculations. The relaxation was performed using a conjugate gradient scheme until the energies and each component of forces were $<10^{-6}$ eV and 0.005 eV $Å^{-1}$, respectively.

GW calculations

The GW calculations were carried out within many-body perturbation theory using non-self-consistent GW approximation (G0W0)[87] as implemented in the VASP. G0 is the Green's function of the electrons, and W0 denotes the screened Coulomb interactions. The input parameters utilized for the G0W0 calculations included 60 frequency grids, 100 empty bands per atom, 500 eV energy cutoff, and a 11 × 11 × 1 Monkhorst-Pack k-grid.

ML training and post-processing

The ML models were developed to learn pattern among the existing data ({**X**, *y*}) by mapping the input attributes ($\{\mathbf{X}\} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$) to the target property (*y*) through the hypothesis function (*h*): $y_i = h(\mathbf{x}_i)$. All the codes related to ML were built using Scikit-learn (version 0.23) package[88] of the Python (version 3.6) programming language. The complete data was divided into 90% and 10% for training and testing of the ML models, respectively. PyCaret (version 2.0) package was used for comparing the performance of various ML algorithms. For hyperparameter optimization of the selected ML algorithm, the Optuna (version 2.3.0) package was used. SHAP (version 0.36.0) package was utilized for interpreting ML models. For multiclass classification, resampling of imbalanced data was performed using the Synthetic Minority Oversampling Technique (SMOTE)[89] method as implemented in the imblearn package[90].

DATA AVAILABILITY

The data used in developing the ML models are freely available at https://github.com/ritesh001/HT-iML_Photocatalysis in the form of spreadsheets. All the electronic, structural, and mechanical properties of 2DO materials utilized in the present study will be soon uploaded on the aNaNt database[91].

CODE AVAILABILITY

The Jupyter notebooks for reproducing main results in the manuscript is also freely available at https://github.com/ritesh001/HT-iML_Photocatalysis.

Received: 7 June 2021; Accepted: 6 November 2021;
Published online: 09 December 2021

REFERENCES

1. Jafari, T. et al. Photocatalytic water splitting – the untamed dream: a review of recent advances. *Molecules* **21**, 900 (2016).
2. Luo, B., Liu, G. & Wang, L. Recent advances in 2D materials for photocatalysis. *Nanoscale* **8**, 6904–6920 (2016).
3. Zhang, H. Introduction: 2D materials chemistry. *Chem. Rev.* **118**, 6089–6090 (2018).
4. Lu, Q., Yu, Y., Ma, Q., Chen, B. & Zhang, H. 2D transition-metal-dichalcogenide-nanosheet-based composites for photocatalytic and electrocatalytic hydrogen evolution reactions. *Adv. Mater.* **28**, 1917–1933 (2016).
5. Maity, N., Srivastava, P., Mishra, H., Shinde, R. & Singh, A. K. Anisotropic interlayer exciton in gese/sns van der Waals heterostructure. *J. Phys. Chem. Lett.* **12**, 1765–1771 (2021).
6. Zhang, J., Zhang, M., Sun, R.-Q. & Wang, X. A facile band alignment of polymeric carbon nitride semiconductors to construct isotype heterojunctions. *Angew. Chem. Int. Ed.* **51**, 10145–10149 (2012).
7. Kumar, R., Das, D. & Singh, A. K. C_2N/WS_2 van der waals type-II heterostructure as a promising water splitting photocatalyst. *J. Catal.* **359**, 143–150 (2018).
8. Gunjakar, J. L., Kim, I. Y., Lee, J. M., Lee, N.-S. & Hwang, S.-J. Self-assembly of layered double hydroxide 2D nanoplates with graphene nanosheets: An effective way to improve the photocatalytic activity of 2D nanostructured materials for visible light-induced O_2 generation. *Energy Environ. Sci.* **6**, 1008–1017 (2013).
9. Zhao, Y. et al. Two-dimensional photocatalyst design: a critical review of recent experimental and computational advances. *Mater. Today* **34**, 78–91 (2020).
10. Kuhar, K. et al. Sulfide perovskites for solar energy conversion applications: computational screening and synthesis of the selected compound $LaYS_3$. *Energy Environ. Sci.* **10**, 2579–2593 (2017).
11. Castelli, I. E. et al. New cubic perovskites for one-and two-photon water splitting using the computational materials repository. *Energy Environ. Sci.* **5**, 9034–9043 (2012).
12. Shinde, A. et al. Discovery of manganese-based solar fuel photoanodes via integration of electronic structure calculations, Pourbaix stability modeling, and high-throughput experiments. *ACS Energy Lett.* **2**, 2307–2312 (2017).
13. Castelli, I. E. et al. New light-harvesting materials using accurate and efficient bandgap calculations. *Adv. Energy Mater.* **5**, 1400915 (2015).
14. Kahle, L., Marcolongo, A. & Marzari, N. High-throughput computational screening for solid-state Li-ion conductors. *Energy Environ. Sci.* **13**, 928–948 (2020).
15. Juneja, R., Yumnam, G., Satsangi, S. & Singh, A. K. Coupling the high-throughput property map to machine learning for predicting lattice thermal conductivity. *Chem. Mater.* **31**, 5145–5151 (2019).
16. Rajan, A. C. et al. Machine-learning-assisted accurate band gap predictions of functionalized MXene. *Chem. Mater.* **30**, 4031–4038 (2018).
17. Wu, L. et al. Recent advances in self-supported layered double hydroxides for oxygen evolution reaction. *Research* **2020**, 1–17 (2020).
18. Lu, X. et al. 2D layered double hydroxide nanosheets and their derivatives toward efficient oxygen evolution reaction. *Nano-Micro Lett.* **12**, 1–32 (2020).
19. Greenwood, N. N. & Earnshaw, A. *Chemistry of the Elements* (Butterworth-Heinemann, 1997).
20. Ward, L., Agrawal, A., Choudhary, A. & Wolverton, C. A general-purpose machine learning framework for predicting properties of inorganic materials. *npj Comput. Mater.* **2**, 1–7 (2016).
21. Ward, L. et al. Matminer: an open source toolkit for materials data mining. *Comput. Mater. Sci.* **152**, 60–69 (2018).
22. Ward, L. et al. Including crystal structure attributes in machine learning models of formation energies via Voronoi tessellations. *Phys. Rev. B* **96**, 024104 (2017).
23. Hansen, M. H. et al. An atomistic machine learning package for surface science and catalysis. Preprint at https://arxiv.org/abs/1904.00904 (2019).
24. Parr, R. G. & Pearson, R. G. Absolute hardness: companion parameter to absolute electronegativity. *J. Am. Chem. Soc.* **105**, 7512–7516 (1983).
25. Pearson, R. G. Hard and soft acids and bases, HSAB, part 1: Fundamental principles. *J. Chem. Educ.* **45**, 581–586 (1968).
26. Pearson, R. G. Hard and soft acids and bases, HSAB, part II: Underlying theories. *J. Chem. Educ.* **45**, 643–648 (1968).
27. Pearson, R. G. Absolute electronegativity and hardness: application to inorganic chemistry. *Inorg. Chem.* **27**, 734–740 (1988).
28. Pearson, R. G. Hard and soft acids and bases – the evolution of a chemical concept. *Coord. Chem. Rev.* **100**, 403–425 (1990).
29. Shankar, R., Senthilkumar, K. & Kolandaivel, P. Calculation of ionization potential and chemical hardness: a comparative study of different methods. *Int. J. Quantum Chem.* **109**, 764–771 (2009).
30. De Proft, F. & Geerlings, P. Calculation of ionization energies, electron affinities, electronegativities, and hardnesses using density functional methods. *J. Chem. Phys.* **106**, 3270–3279 (1997).
31. Mukherjee, M., Satsangi, S. & Singh, A. K. A statistical approach for the rapid prediction of electron relaxation time using elemental representatives. *Chem. Mater.* **32**, 6507–6514 (2020).
32. Chen, T. & Guestrin, C. Xgboost: A scalable tree boosting system. In *Proc. 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794 (2016).
33. PyCaret: An open source, low-code machine learning library in python. https://pycaret.org.
34. Optuna: A hyperparameter optimization framework. https://optuna.readthedocs.io/en/stable/index.html.
35. SHAP (shapley additive explanations): A game theoretic approach to explain the output of any machine learning model. https://shap.readthedocs.io/en/latest.
36. Pandey, M. & Jacobsen, K. W. Heats of formation of solids with error estimation: the mBEEF functional with and without fitted reference energies. *Phys. Rev. B* **91**, 235201 (2015).
37. Stevanović, V., Lany, S., Zhang, X. & Zunger, A. Correcting density functional theory for accurate predictions of compound enthalpies of formation: fitted elemental-phase reference energies. *Phys. Rev. B* **85**, 115104 (2012).
38. Haastrup, S. et al. The computational 2D materials database: High-throughput modeling and discovery of atomically thin crystals. *2D Mater.* **5**, 042002 (2018).
39. Sun, J. et al. Accurate first-principles structures and energies of diversely bonded systems from an efficient density functional. *Nat. Chem.* **8**, 831 (2016).
40. Becke, A. D. & Roussel, M. R. Exchange holes in inhomogeneous systems: a coordinate-space model. *Phys. Rev. A* **39**, 3761 (1989).
41. Wang, Z., Guo, X., Montoya, J. & Nørskov, J. K. Predicting aqueous stability of solid with computed Pourbaix diagram using SCAN functional. *npj Comput. Mater.* **6**, 1–7 (2020).
42. Yang, J. H., Kitchaev, D. A. & Ceder, G. Rationalizing accurate structure prediction in the meta-GGA SCAN functional. *Phys. Rev. B* **100**, 035132 (2019).
43. Friedrich, R. et al. Coordination corrected ab initio formation enthalpies. *npj Comput. Mater.* **5**, 1–12 (2019).
44. Datta, D. Geometric mean principle for hardness eualization: a corollary of Sanderson's geometric mean principle of electronegativity equalization. *J. Phys. Chem.* **90**, 4216–4217 (1986).
45. Kaya, S. & Kaya, C. A new equation for calculation of chemical hardness of groups and molecules. *Mol. Phys.* **113**, 1311–1319 (2015).
46. Ke, G. et al. Lightgbm: a highly efficient gradient boosting decision tree. *Adv. Neural Inform. Proc. Syst.* **30**, 3146–3154 (2017).
47. Murdock, R. J., Kauwe, S. K., Wang, A. Y.-T. & Sparks, T. D. Is domain knowledge necessary for machine learning materials properties? *Integr. Mater. Manuf. Innov.* **9**, 221–227 (2020).
48. Bartel, C. J. et al. A critical examination of compound stability predictions from machine-learned formation energies. *npj Comput. Mater.* **6**, 1–11 (2020).
49. Schleder, G. R., Acosta, C. M. & Fazzio, A. Exploring two-dimensional materials thermodynamic stability via machine learning. *ACS Appl. Mater. Interfaces* **12**, 20149–20157 (2019).
50. Siriwardane, E. M., Joshi, R. P., Kumar, N. & Çakır, D. Revealing the formation energy–exfoliation energy–structure correlation of MAB phases using machine learning and DFT. *ACS Appl. Mater. Interfaces* **12**, 29424–29431 (2020).
51. Esterhuizen, J. A., Goldsmith, B. R. & Linic, S. Theory-guided machine learning finds geometric structure-property relationships for chemisorption on subsurface alloys. *Chem* **6**, 3100–3117 (2020).
52. Wagner, N. & Rondinelli, J. M. Theory-guided machine learning in materials science. *Front. Mater.* **3**, 28 (2016).
53. Lundberg, S. M. et al. From local explanations to global understanding with explainable AI for trees. *Nat. Mach. Intell.* **2**, 56–67 (2020).
54. Saal, J. E., Kirklin, S., Aykol, M., Meredig, B. & Wolverton, C. Materials design and discovery with high-throughput density functional theory: the open quantum materials database (OQMD). *JOM* **65**, 1501–1509 (2013).
55. Ong, S. P. et al. Python materials genomics (Pymatgen): a robust, open-source python library for materials analysis. *Comput. Mater. Sci.* **68**, 314–319 (2013).
56. Malyi, O. I., Sopiha, K. V. & Persson, C. Energy, phonon, and dynamic stability criteria of two-dimensional materials. *ACS Appl. Mater. Interfaces* **11**, 24876–24884 (2019).
57. Shang, S., Wang, Y. & Liu, Z.-K. First-principles elastic constants of α- and θ-Al_2O_3. *Appl. Phys. Lett.* **90**, 101909 (2007).

58. Choudhary, K., Cheon, G., Reed, E. & Tavazza, F. Elastic properties of bulk and low-dimensional materials using van der Waals density functional. *Phys. Rev. B* **98**, 014107 (2018).
59. Geurts, P., Ernst, D. & Wehenkel, L. Extremely randomized trees. *Mach. Learn.* **63**, 3–42 (2006).
60. Singh, A. K., Mathew, K., Zhuang, H. L. & Hennig, R. G. Computational screening of 2D materials for photocatalysis. *J. Phys. Chem. Lett.* **6**, 1087–1098 (2015).
61. Bredas, J.-L. Mind the gap! *Mater. Horiz.* **1**, 17–19 (2014).
62. Makuła, P., Pacia, M. & Macyk, W. How to correctly determine the band gap energy of modified semiconductor photocatalysts based on UV-Vis spectra. *J. Phys. Chem. Lett.* **9**, 6814–6817 (2018).
63. Zhang, Y. & Xu, X. Machine learning band gaps of doped-TiO_2 photocatalysts from structural and morphological parameters. *ACS omega* **5**, 15344–15352 (2020).
64. Beranek, R. (Photo) electrochemical methods for the determination of the band edge positions of TiO_2-based nanomaterials. *Adv. Phys. Chem.* **2011**, 1–20 (2011).
65. van Schilfgaarde, M., Kotani, T. & Faleev, S. Quasiparticle self-consistent GW theory. *Phys. Rev. Lett.* **96**, 226402 (2006).
66. Trasatti, S. The absolute electrode potential: an explanatory note (Recommendations 1986). *Pure Appl. Chem* **58**, 955–966 (1986).
67. Starr, C. *Biology: Concepts and Applications* (Thomson Brooks/Cole, 2005).
68. Weast, R. C., Astle, M. J. & Beyer, W. H. *Handbook of Physics and Chemistry* (CRC Press, Boca Raton, 1986).
69. Fan, Y., Song, X., Qi, S., Ma, X. & Zhao, M. Li-III-VI bilayers for efficient photocatalytic overall water splitting: The role of intrinsic electric field. *J. Mater. Chem. A* **7**, 26123–26130 (2019).
70. Li, X., Li, Z. & Yang, J. Proposed photosynthesis method for producing hydrogen from dissociated water molecules using incident near-infrared light. *Phys. Rev. Lett.* **112**, 018301 (2014).
71. Chen, Z. et al. Accelerating materials development for photoelectrochemical hydrogen production: standards for methods, definitions, and reporting protocols. *J. Mater. Res.* **25**, 3–16 (2010).
72. Murphy, A. et al. Efficiency of solar water splitting using semiconductor electrodes. *Int. J. Hydrog. Energy* **31**, 1999–2017 (2006).
73. Fu, C.-F. et al. Intrinsic electric fields in two-dimensional materials boost the solar-to-hydrogen efficiency for photocatalytic water splitting. *Nano Lett.* **18**, 6312–6317 (2018).
74. Wang, Y. et al. Monolayer $PtSe_2$, a new semiconducting transition-metal-dichalcogenide, epitaxially grown by direct selenization of Pt. *Nano Lett.* **15**, 4013–4018 (2015).
75. Sun, X. et al. An efficient and extremely stable photocatalytic $PtSe_2$/FTO thin film for water splitting. *Energy Technol.* **8**, 1900903 (2020).
76. Wen, Y., Zhu, Y. & Zhang, S. Low temperature synthesis of ZrS_2 nanoflakes and their catalytic activity. *RSC Adv.* **5**, 66082–66085 (2015).
77. Zhang, R. et al. Direct Z-scheme water splitting photocatalyst based on two-dimensional van der waals heterostructures. *J. Phys. Chem. Lett.* **9**, 5419–5424 (2018).
78. Lin, Y. et al. Semiconductor nanostructure-based photoelectrochemical water splitting: a brief review. *Chem. Phys. Lett.* **507**, 209–215 (2011).
79. Rasmussen, F. A. & Thygesen, K. S. Computational 2D materials database: Electronic structure of transition-metal dichalcogenides and oxides. *J. Phys. Chem. C* **119**, 13169–13183 (2015).
80. Chen, Y. et al. Two-dimensional nanomaterials for photocatalytic CO_2 reduction to solar fuels. *Sustainable Energy Fuels* **1**, 1875–1898 (2017).
81. Kumar, R. & Singh, A. K. Electronic structure based intuitive design principle of single-atom catalysts for efficient electrolytic nitrogen reduction. *ChemCatChem* **12**, 5456–5464 (2020).
82. Yu, J., Pan, S., Zhang, Y., Liu, Q. & Li, B. Facile synthesis of monodispersed α-Ni$(OH)_2$ microspheres assembled by ultrathin nanosheets and its performance for oxygen evolution reduction. *Front. Mater.* **6**, 124 (2019).
83. Balasubramanian, P., He, S.-B., Deng, H.-H., Peng, H.-P. & Chen, W. Defects engineered 2D ultrathin cobalt hydroxide nanosheets as highly efficient electrocatalyst for non-enzymatic electrochemical sensing of glucose and L-cysteine. *Sens. Actuators B Chem.* **320**, 128374 (2020).
84. Kresse, G. & Hafner, J. Ab initio molecular dynamics for liquid metals. *Phys. Rev. B* **47**, 558–561 (1993).
85. Kresse, G. & Joubert, D. From ultrasoft pseudopotentials to the projector augmented-wave method. *Phys. Rev. B* **59**, 1758–1775 (1999).
86. Perdew, J. P., Burke, K. & Ernzerhof, M. Generalized gradient approximation made simple. *Phys. Rev. Lett.* **77**, 3865–3868 (1996).
87. Shishkin, M. & Kresse, G. Implementation and performance of the frequency-dependent GW method within the paw framework. *Phys. Rev. B* **74**, 035101 (2006).
88. scikit-learn: Machine learning in python. https://scikit-learn.org/stable.
89. Chawla, N. V., Bowyer, K. W., Hall, L. O. & Kegelmeyer, W. P. SMOTE: Synthetic minority over-sampling technique. *J. Artif. Intell. Res.* **16**, 321–357 (2002).
90. Lemaître, G., Nogueira, F. & Aridas, C. K. Imbalanced-learn: a python toolbox to tackle the curse of imbalanced datasets in machine learning. *J. Mach. Learn. Res.* **18**, 1–5 (2017).
91. aNANt: A functional materials database. http://anant.mrc.iisc.ac.in.

ACKNOWLEDGEMENTS

The authors also acknowledge supports from Institute of Eminence (IoE) grant from MHRD, DST Indo-Korea grant, and DST Nanomission. The authors thank Materials Research Centre (MRC) and Supercomputer Education and Research Centre (SERC), and Solid State and Structural Chemistry Unit (SSCU), Indian Institute of Science, Bangalore for providing the required computational facilities.

AUTHOR CONTRIBUTIONS

A.K.S. conceived the project. R.K. performed the calculations and analyzed the results. R.K. and A.K.S. contributed to the discussion of results and the writing of the manuscript.

COMPETING INTERESTS

The authors declare no competing interests.

ADDITIONAL INFORMATION

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41524-021-00669-4.

Correspondence and requests for materials should be addressed to Abhishek K. Singh.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.