AI for Materials
https://www.sciltp.com/journals/aimat



Article

Check for updates

NJmatWeb: An AI-Driven Web Platform Automating Machine Learning and Symbolic Regression for Materials Science

Lei Zhang [1,2,*], Xinyi Wang [1,2] and Youle Wang [2]

[1] Key Laboratory of Spectral Imaging Technology, Xi'an Institute of Optics and Precision Mechanics of the Chinese Academy of Sciences, Xi'an 710119, China
[2] Department of Internet Engineering, School of Software, Nanjing University of Information Science & Technology, Nanjing 210044, China
* Correspondence: zhanglei1@opt.ac.cn

How To Cite: Zhang, L.; Wang, X.; Wang, Y. NJmatWeb: An AI-Driven Web Platform Automating Machine Learning and Symbolic Regression for Materials Science. *AI for Materials* **2026**, *1*(1), 14. https://doi.org/10.53941/aimat.2026.100014

Received: 24 July 2026
Revised: 4 September 2026
Accepted: 8 October 2026
Published: 26 October 2026

Abstract: Data-driven methodologies are now widely employed for predicting materials properties; however, most existing tools fail to integrate automated feature selection, interpretable modeling, and an accessible user interface into a unified pipeline. To bridge this gap, we introduce NJmatWeb, a web-based platform that automates machine learning (ML), symbolic regression and SHAP analysis for materials science applications. Designed to perform robustly on the small datasets commonly encountered in this field, NJmatWeb incorporates a symbolic regression module that yields interpretable symbolic expressions, thereby enhancing model transparency. NJmatWeb incorporates an automatic featurizer that handles both inorganic and organic materials. The platform further provides automated feature selection, feature ranking, SHAP analysis, dataset visualization, and heat map comparisons of the feature space before and after selection. Prediction accuracies of available models are reported, with results from both widely adopted materials benchmarks and diverse case studies. This study demonstrates a useful AI tool aligned with the Materials Genome Engineering, accelerating materials design by combining predictive power with interpretability in an accessible web platform.

Keywords: AI; machine learning; symbolic regression; deep learning; AI for materials

1. Introduction

The rapid advancement of data-driven methodologies has transformed scientific discovery across disciplines [1–3]. Artificial intelligence [4–6], particularly machine learning, has evolved from pattern recognition into a paradigm capable of automating complex decision-making, predictive modeling, and knowledge extraction from high-dimensional datasets [7,8]. In parallel, the emergence of embodied intelligence further expands the frontier of autonomous scientific exploration. Within this landscape, the development of accessible and integrated software platforms has become critical to democratizing AI capabilities for domain scientists who may lack specialized programming expertise [9–11]. Platforms such as scikit-learn, TensorFlow, and PyTorch provide foundational building blocks, yet they often require significant coding effort to assemble into a complete, reproducible analytical workflow tailored to a specific domain [12,13].

In materials science, the adoption of AI has accelerated significantly over the past few years [14–16]. Butler et al. [17] provided a comprehensive perspective on machine learning for molecular and materials science, outlining both opportunities and methodological challenges. Raccuglia et al. [18] demonstrated machine-learning-assisted materials discovery using data from failed experiments, highlighting the value of mining historical datasets. Tshitoyan et al. [19] showed that unsupervised word embeddings trained on materials science literature

Copyright: © 2026 by the authors. This is an open access article under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).
Publisher's Note: Scilight stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

can capture latent chemical knowledge without explicit supervision. Merchant et al. [20] scaled deep learning for materials discovery by predicting the stability of millions of novel crystal structures, achieving an order-of-magnitude increase in known stable materials. These works collectively underscore the transformative potential of AI in accelerating materials research. Beyond property prediction, machine learning has enabled a new generation of interatomic potentials that achieve near first-principles accuracy at a fraction of the computational cost. For instance, Behler and Parrinello [21] introduced the first high-dimensional neural network potential, laying the foundation for modern machine learning force fields. Thompson et al. [22] developed the Spectral Neighbor Analysis Potential framework for automated generation of quantum-accurate interatomic potentials [22]. Zhang et al. proposed the Deep Potential (DeepMD) method and its associated training protocol DP-GEN, enabling automatic generation of reliable neural network potentials for complex material systems [23]. Batzner et al. [24] introduced NequIP, an E(3)-equivariant graph neural network architecture for data-efficient and accurate interatomic potential. These methods have dramatically expanded the scope of atomistic simulations [25]. More recently, the concept of large atomic models has emerged, aiming to train universal machine learning potentials applicable across the entire periodic table [26]. The large-scale pretrained models, analogous to foundation models in natural language processing, offer the promise of transferable atomic-scale predictions that can be fine-tuned for downstream tasks with minimal additional data [27,28]. Moreover, in parallel with forward prediction, inverse design [29–31] has emerged as a critical paradigm in which AI models are used to navigate the vast compositional and structural space to identify promising candidates with targeted properties. Rather than predicting properties from a given structure, inverse design reverses the mapping: given a desired property target, the model searches for the material that satisfies it [32,33].

The integration of AI with materials research aligns closely with the goals of the Materials Genome initiative [34,35], which seeks to reduce the cost and time of materials development by combining high-throughput computation, data infrastructure, and intelligent algorithms. Despite these advances, several challenges persist. For example, many existing materials informatics [36,37] tools require users to manually curate and featurize input data, select appropriate models, and interpret results across disparate software environments. Furthermore, the black-box nature of many high-performing models limits physical interpretability [38–40]. While symbolic regression naturally produces interpretable mathematical expressions, it remains underutilized due to a lack of accessible implementations tailored to materials data [41,42].

In this study, we present NJmatWeb, a web-based infrastructure that automates the end-to-end pipeline of machine learning and symbolic regression modeling for materials property prediction. NJmatWeb integrates an automatic featurizer for both inorganic and organic materials, recursive feature elimination for dimensionality reduction, machine learning and deep learning models, and a symbolic regression module. The current version focuses on forward prediction, specifically the automated construction of accurate and interpretable models from provided datasets. While true inverse design lies beyond the scope of the present implementation, the platform provides comprehensive model interpretability via SHAP analysis, dataset visualization, and feature space comparisons before and after selection, which can help researchers identify key compositional drivers and formulate design hypotheses, thereby establishing a useful foundation for subsequent inverse design efforts. The platform is evaluated on a diverse set of case studies: steel yield strength classification, experimental band gap prediction, perovskite surface adsorption energy, perovskite photovoltage prediction, and multi-solvent stability classification. By unifying automated featurization, model training, symbolic regression, and interpretability tools within a single web interface, NJmatWeb aims to lower the barrier for materials researchers to leverage AI-driven prediction and discovery.

2. Methods

2.1. Feature Selection and Feature Ranking

A comprehensive set of candidate descriptors is initially constructed from the compositional and structural information provided in each input dataset, encompassing compositional ratios, elemental property aggregates (e.g., Pauling electronegativity, Shannon ionic radius, oxidation state averages, group number). Feature ranking is performed using two complementary strategies. Pearson correlation coefficients are computed to measure the linear relationship between each descriptor and the target property such as the adsorption energy, while impurity-based importance scores are obtained from a Random Forest model. Descriptors with very low variance are removed using a VarianceThreshold of 0.01, and highly correlated descriptors are pruned using an absolute correlation threshold of 0.98. Recursive Feature Elimination (RFE) is then applied with a Random Forest base estimator (30 trees) and a fixed random state of 42. For conventional machine-learning and DNN models, 20 descriptors are retained; for the symbolic models, six descriptors are retained to limit expression complexity and

improve interpretability. Feature selection and normalization are performed independently within each training fold to prevent information leakage from the validation fold.

2.2. Machine Learning and Deep Learning

Different machine learning models are trained on the selected feature subset. DNN is implemented with fully connected layers using an architecture of three hidden layers containing 128, 64, and 32 neurons, respectively, with ReLU activations and dropout rates of 0.30 and 0.20 after the first and second hidden layers. Training uses the Adam optimizer with a learning rate of 1×10^{-3}. A batch size of 32 was used for datasets containing no more than 5000 samples, whereas a batch size of 512 was used for larger datasets. Random Forest and Extra Trees are configured with 300 trees and a maximum depth of 10, their minimum leaf size is 3 for regression and 2 for classification. XGBoost employs 300 boosting rounds with a maximum tree depth of 6, a learning rate of 0.03 for regression and 0.05 for classification, and a subsample ratio of 0.8. Logistic Regression uses C = 0.5, class-balanced weighting, and a maximum of 2000 iterations. All models are evaluated using five-fold stratified cross-validation for classification tasks and shuffled five-fold cross-validation for regression tasks. In each cross-validation iteration, four folds (80% of the data) are used for training and the remaining fold (20%) is used for validation. This procedure is repeated five times so that each sample is used for validation exactly once. The samples are shuffled before fold assignment using a fixed random seed of 42, and class proportions are preserved across folds for classification tasks. Within each iteration, feature selection and normalization are fitted exclusively on the training folds and then applied to the held-out validation fold to prevent information leakage. Performance metrics include the coefficient of determination (R^2), mean absolute error (MAE), and root mean squared error (RMSE) for regression tasks; accuracy and F1-score for classification tasks. Reported values are the fold-wise mean ± standard deviation. The train–validation gap is calculated as the difference between the mean training and validation scores to assess potential overfitting. To interpret the predictions of the trained black-box models, SHAP (SHapley Additive exPlanations) analysis is conducted using TreeSHAP for the ensemble tree models (Random Forest, XGBoost, Extra Trees) and DeepSHAP for the DNN, both implemented via the SHAP Python library. For each model, Shapley values are computed to decompose individual predictions into additive feature contributions. The SHAP summary beeswarm plot further visualizes the distribution of Shapley values for each feature, where each dot corresponds to a single prediction instance and color encodes the feature value from low (blue) to high (red). It should be noted that the above parameters are default values and can be customized by users to optimize performance.

2.3. Symbolic Regression

NJmatWeb implements a symbolic regression framework to evolve interpretable symbolic expressions. In each cross-validation fold, six descriptors are selected from the training subset and normalized using a MinMaxScaler fitted only to that training subset. The symbolic regression models use gplearn's half-and-half initialization scheme, with initial tree depths ranging from 2 to 6 and the default tournament-selection and subtree-operation settings of the library. The terminal set comprises the selected feature variables, each transformed to the [0, 1] interval using a MinMaxScaler fitted exclusively on the training fold. The function set includes binary arithmetic operators (addition, subtraction, multiplication and division) and unary elementary functions (exponential, natural logarithm, sine, cosine, and square root). Each individual tree encodes a candidate symbolic expression mapping the feature space to a scalar prediction. For regression tasks, the evolved expression directly predicts the continuous target property, the population size is 2000 and the evolutionary search proceeds for five generations, with a parsimony coefficient of 1×10^{-4} and MAE as the fitness metric. For classification tasks, the same symbolic regression framework is used with a population size of 300 and five generations, but the symbolic expression produces a continuous discriminant score, which is subsequently converted to binary class labels via thresholding at 0.5. For conciseness, we refer to this module uniformly as symbolic regression throughout the manuscript, irrespective of task type. Selection, subtree crossover, and mutation follow the corresponding default operations implemented in gplearn; no additional crossover or mutation probabilities are imposed in the present implementation. All symbolic models use a fixed random state of 42. These settings were selected as computationally tractable defaults for interactive use. A population size of 2000 for regression tasks provides sufficient structural diversity for evolutionary search while keeping the computational cost manageable, whereas the smaller population size of 300 used for classification tasks reduces the cost of evolving a discriminant function for binary labels. Five generations was selected to provide a reproducible baseline within a reasonable runtime for web-based analysis. The parsimony coefficient of 1×10^{-4} penalizes unnecessarily complex expressions and favors compact formulas that are easier to interpret. The six-descriptor limit further reduces the search space and helps prevent overly complex or unstable expressions, particularly for the relatively small materials datasets considered

here. These values are intended as default starting points rather than universally optimal settings, and users can increase the population size or number of generations when more complex expressions or higher predictive accuracy are required. The physical interpretation of an evolved expression is based on the known chemical meaning of the descriptors and the mathematical structure of the expression. Because the descriptors are normalized and the symbolic model is optimized empirically, the resulting equation should be regarded as an interpretable, hypothesis-generating approximation rather than a unique mechanistic law.

2.4. Frontend

The analytical pipeline, encompassing feature selection, RFE, DNN training, ensemble model training, symbolic regression, and SHAP interpretation, is deployed as an interactive web application using the Gradio framework with a Python backend. The frontend interface is organized into three functional panels (Figure 1). The input panel allows researchers to upload data in CSV format and configure key hyperparameters via dropdown menus and numerical sliders, including the train-test split ratio, the number of features to retain after RFE, the population size and generation count for symbolic regression, and the choice of SHAP visualization type. The execution panel triggers the training workflow in a single click, with progress bars indicating the status of feature selection, model training, symbolic regression evolution, and SHAP computation. The results panel dynamically renders regression parity plots for all models, a comparative table of R^2, MAE and RMSE across models, feature importance rankings from both SHAP and impurity-based methods, SHAP summary beeswarm plots, and the final symbolic equation discovered by the symbolic regression displayed in LaTeX-rendered format. The entire system is containerized using Docker, enabling deployment on local machines or cloud servers for collaborative use by the materials science community.

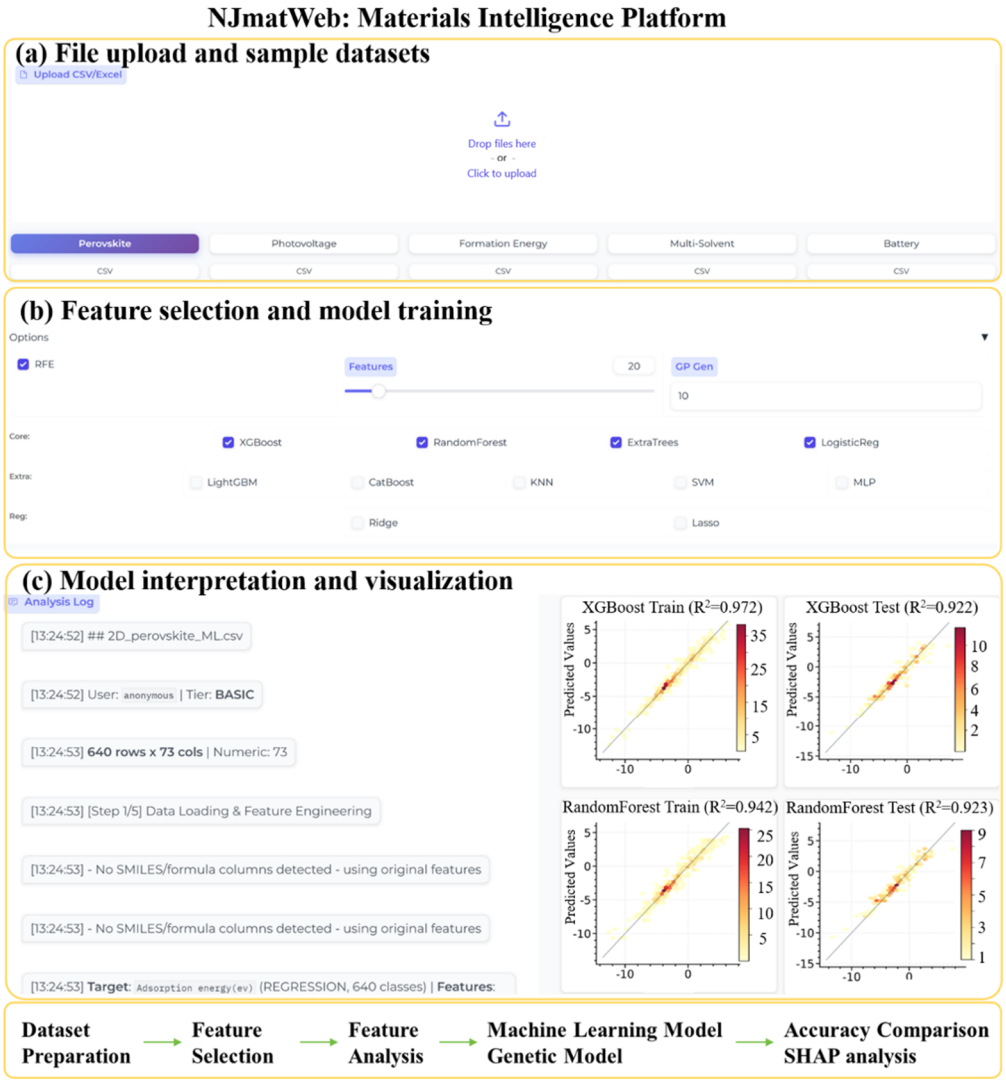


Figure 1. Screenshot of the NJmatWeb dashboard, showing the sample dataset selector, file upload interface, and the main functional modules for feature selection, symbolic regression, and machine learning.

3. Results and Discussion

3.1. Case 1: Steel Materials

The NJmatWeb platform is evaluated on the matbench_steels.csv dataset from MatBench, which contains 312 steel compositions with yield strength as the target (Table 1) [43]. Upon a single-click submission, the platform automatically detects the composition column and applies matminer Magpie-based featurization, generating 134 compositional descriptors. After variance filtering and correlation pruning, SelectFromModel reduces the feature space to 20 features. As the initial regression R^2 falls below 0.1, the system autonomously switches to binary classification by partitioning samples into high and low yield strength classes via the median threshold.

Table 1. Performance comparison of classification models for steel yield strength prediction (accuracy and F1-score).

Model	Train Acc	Validation Acc	F1
Random Forest	0.989 ± 0.003	0.875 ± 0.033	0.875 ± 0.033
Extra Trees	0.978 ± 0.007	0.872 ± 0.029	0.872 ± 0.029
XGBoost	1.000 ± 0.000	0.853 ± 0.017	0.852 ± 0.017
DNN	0.876 ± 0.026	0.776 ± 0.029	0.775 ± 0.029
Symbolic Regression	0.620 ± 0.022	0.603 ± 0.037	0.599 ± 0.037
Logistic Regression	0.611 ± 0.013	0.593 ± 0.044	0.591 ± 0.044

Figure 2a presents the feature correlation matrices before and after selection. The left heatmap displays pairwise Pearson correlations among all 134 raw Magpie-derived features using a red-blue diverging color scale. A clear block-diagonal pattern emerges, indicating strongly correlated clusters arising from related elemental properties. This observation motivates the removal of features with correlation coefficients exceeding 0.98. The right heatmap shows the correlation structure among the 20 features retained after recursive feature elimination. The reduced intercorrelation among the selected features confirms that the filtering process successfully preserves a diverse and informative subset. The ROC plot (Figure 2b) overlays three tree-based classifiers. Extra Trees achieves the highest AUC of 0.928, followed by Random Forest at 0.924 and XGBoost at 0.912, all well above the diagonal baseline, demonstrating that compositional features alone provide strong discriminative power for yield strength classification. The SHAP summary beeswarm plot reveals the directional impact of each feature. The dominant Magpie descriptors represent elemental electronic structure and periodic position. For example, descriptors related to valence-electron count, periodic-table row, and atomic size can reflect differences in bonding character, atomic packing, and alloying-element effects. Their SHAP distributions indicate how these composition-derived properties increase or decrease the predicted probability of belonging to the high- or low-yield-strength class. In addition, the SHAP importance ranking and the Pearson correlation bar chart independently converge on similar dominant features. For each model, validation (left) and training (right) confusion matrices (Figure 2c) are displayed side-by-side. Random Forest achieves the highest validation accuracy of 0.875 ± 0.033 (F1 = 0.875 ± 0.033), followed by Extra Trees at 0.872 ± 0.029 (F1 = 0.872 ± 0.029) and XGBoost at 0.853 ± 0.017 (F1 = 0.852 ± 0.017). All tree-based models attain near-perfect training accuracy, but their lower validation accuracy indicates some degree of overfitting. This suggests that while tree-based models have sufficient capacity to capture composition–strength relationships, the model complexity may exceed what the dataset size (312 samples) can robustly support. Logistic Regression achieves a validation accuracy of 59.3% (F1 = 0.591) and a training accuracy of 61.1%. Per-class inspection confirms that misclassifications are roughly symmetric across high and low classes, indicating no systematic bias toward either regime. Overall, this case confirms that NJmatWeb can efficiently construct and screen multiple machine learning and deep learning algorithms on a standard steel materials benchmark, and the results demonstrate that ensemble tree methods are well suited for the steel property prediction task.

3.2. Case 2: Band Gap Prediction

The NJmatWeb platform is evaluated on the dataset with experimental band gap values (matbench_expt_gap.csv) from MatBench, comprising 4603 inorganic compounds with experimentally measured band gaps ranging from 0 to 7.5 eV (Table 2 and Figure 3). The platform automatically detects the composition column and applies matminer Magpie-based featurization, generating 134 compositional descriptors. Five regressors were evaluated using shuffled five-fold cross-validation with MinMax normalization: XGBoost (300 boosting rounds, maximum depth 6, learning rate 0.03, subsample ratio 0.8, and column-subsample ratio 0.8), Random Forest (300 trees, maximum depth 10, and minimum leaf size 3), Extra Trees (300 trees, maximum depth 10, and minimum leaf size 3), a three-layer DNN (128-64-32), and symbolic regression (population size 2000 and five generations). XGBoost achieves the

highest validation R^2 of 0.713 ± 0.023, with MAE of 0.438 ± 0.014 eV, followed by Extra Trees (R^2 = 0.671 ± 0.029) and Random Forest (R^2 = 0.650 ± 0.034). The DNN attains a validation R^2 of 0.641 ± 0.050, which is slightly lower than those of the tree-based ensemble models. This difference may arise because tree ensembles are particularly effective for medium-sized tabular datasets with nonlinear feature interactions, whereas the fixed three-layer DNN architecture may require a larger training dataset or more extensive architecture and regularization tuning to achieve its full predictive potential. Symbolic regression produces an interpretable symbolic expression with limited predictive power (validation R^2 = 0.194 ± 0.090). In the correlation heatmaps of the feature space before and after feature selection (Figure 3a), the left panel shows pairwise Pearson correlations among all 134 Magpie features, where block-diagonal structure reveals highly correlated descriptor families. Three complementary ranking methods are presented. The Pearson correlation bar chart ranks features by |Pearson r| with the target. The dominant descriptors in the SHAP analysis are related to elemental valence states, periodic-table position, and atomic size. These quantities are chemically relevant to band-gap variation because they influence orbital energies, bonding interactions, and the degree of electronic-state overlap in inorganic compounds. The SHAP results therefore provide more than a numerical ranking: they indicate how changes in these composition-derived descriptors shift the predicted band gap. Beyond the fitting curves, the symbolic regression module produces a LaTeX-rendered mathematical formula discovered through evolutionary search. While the quantitative accuracy remains poor (validation R^2 = 0.194 ± 0.090), the formula is expected to serve as an alternative hybrid feature to offer closed-form approximation of the band gap–composition relationship.

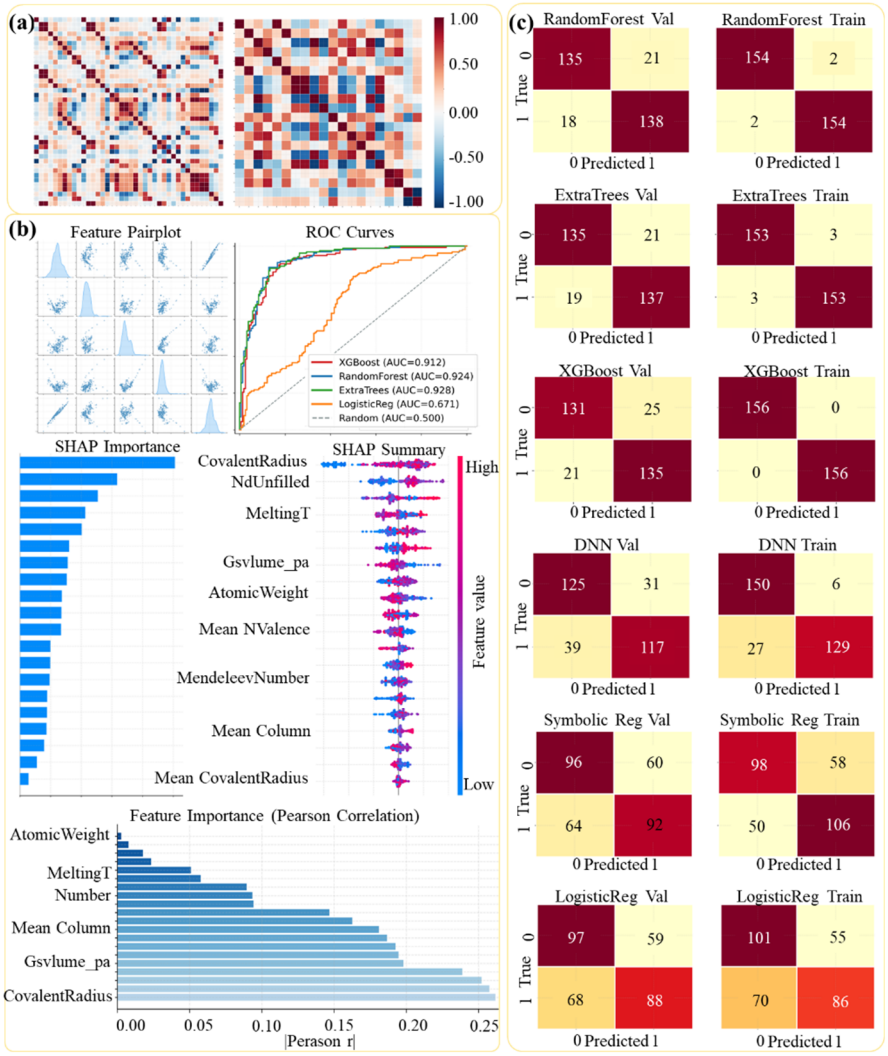


Figure 2. Automated analysis of the steel dataset in matbench following submission to NJmatWeb. (**a**) Correlation heatmaps of the feature space before and after feature selection. (**b**) Model evaluation and interpretability analysis, including ROC curves with AUC scores for XGBoost (0.912), Random Forest (0.924), and Extra Trees (0.928), along with SHAP summary plots, SHAP-based feature importance rankings, and Pearson correlation-based feature importance rankings. (**c**) Confusion matrices for six models, including XGBoost, Logistic Regression, Extra Trees, Random Forest, DNN, and symbolic regression on the validation and training sets.

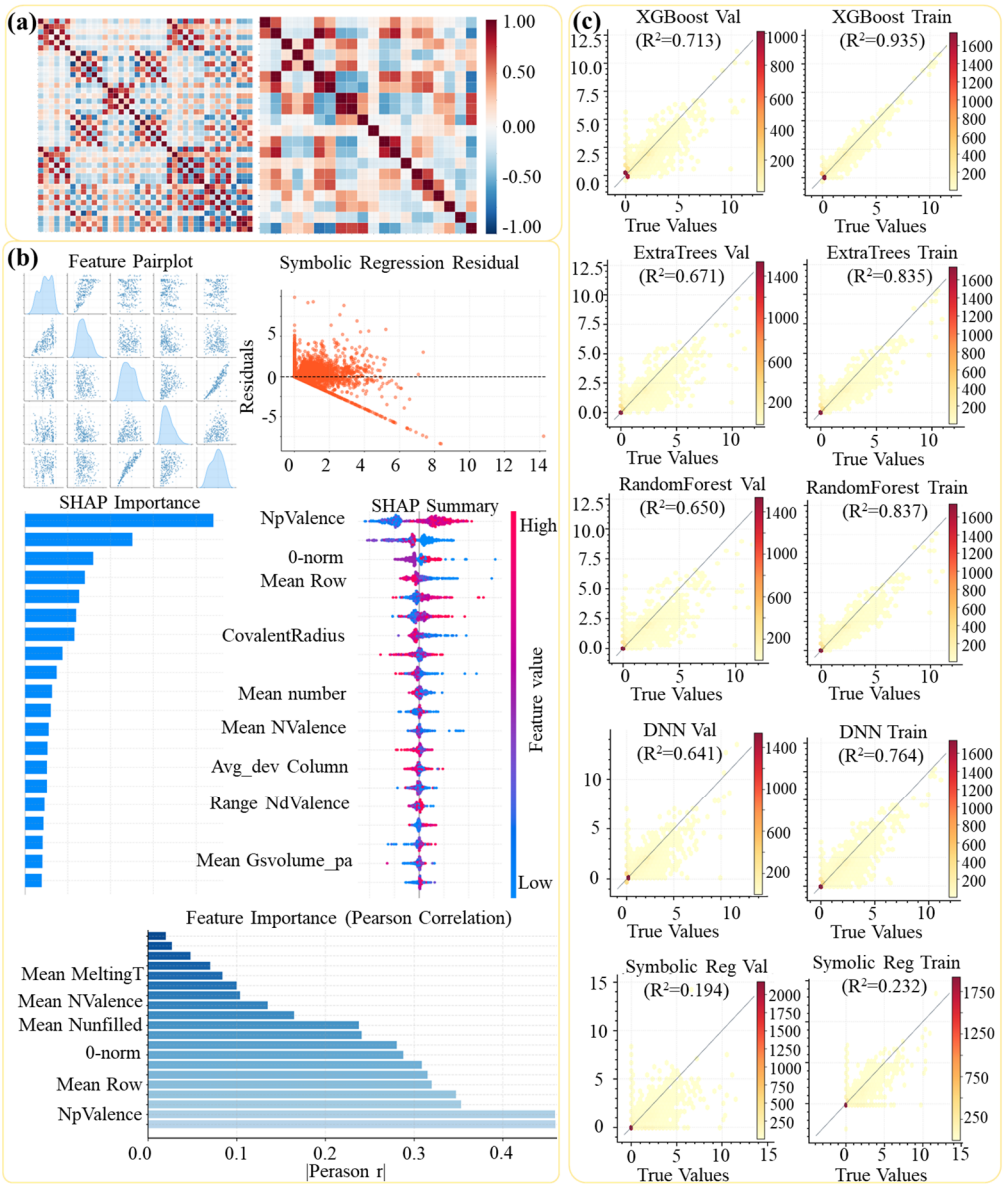


Figure 3. Experimental band gap prediction using NJmatWeb. (**a**) Correlation heatmaps of the feature space before and after feature selection. (**b**) Model evaluation and interpretability analysis, including SHAP summary plots, SHAP-based feature importance rankings, and Pearson correlation-based feature importance rankings. (**c**) Regression parity plots for machine learning and symbolic regression models on the validation and training set.

Table 2. Matbench experimental band gap prediction.

Model	Train R^2	Validation R^2	MAE	RMSE	Gap (Train-Validation)
XGBoost	0.935 ± 0.003	0.713 ± 0.023	0.438 ± 0.014	0.774 ± 0.059	0.222
Extra Trees	0.835 ± 0.016	0.671 ± 0.029	0.480 ± 0.030	0.829 ± 0.063	0.165
Random Forest	0.837 ± 0.007	0.650 ± 0.034	0.489 ± 0.023	0.854 ± 0.072	0.187
DNN	0.764 ± 0.040	0.641 ± 0.050	0.469 ± 0.040	0.865 ± 0.085	0.124
Symbolic Regression	0.232 ± 0.061	0.194 ± 0.090	0.712 ± 0.045	1.296 ± 0.115	0.038

3.3. Case 3: Adsorption Energy

For the adsorption energy prediction task [44], five representative regression models, including DNN, Random Forest, XGBoost, Extra Trees, and symbolic regression, are systematically evaluated (Table 3 and Figure 4). Among all models, XGBoost achieves the highest validation R^2 of 0.946 ± 0.016, followed by DNN (R^2 = 0.928 ± 0.014), Random Forest (R^2 = 0.922 ± 0.014), and Extra Trees (R^2 = 0.918 ± 0.015). These machine learning models provide highly accurate adsorption energy predictions, demonstrating that NJmatWeb effectively captures the nonlinear relationship between perovskite structures and surface adsorption behavior. The symbolic regression model achieves a lower validation R^2 of 0.455 ± 0.131; however, it produces a closed-form mathematical

expression that directly links material descriptors with adsorption energy, thereby providing physical interpretability that cannot be obtained from black-box models. The symbolic expression provides a qualitative descriptor-level interpretation of the adsorption-energy trends. In the evolved equation shown in Figure 4a, ion density (X_1) and ion melting point (X_5) are combined through logarithmic, square-root, ratio, and additive terms, indicating a nonlinear and non-additive descriptor–property relationship. SHAP analysis identifies the contribution and direction of individual descriptors in the predictions, whereas symbolic regression additionally provides an explicit candidate functional form describing how selected descriptors may interact mathematically. Thus, the two approaches provide complementary information: SHAP offers feature-attribution information, while symbolic regression provides a compact equation that can be used to formulate testable hypotheses. The descriptor names and their original units corresponding to the symbolic variables are provided in Table S3. Notably, the symbolic regression model exhibits the smallest train–validation R^2 gap of 0.030, compared with 0.034–0.050 for the machine-learning models, suggesting that the compact symbolic expression is less prone to fitting fold-specific fluctuations. However, the small gap should not be interpreted alone as evidence of superior generalization, because its absolute validation R^2 remains relatively low (0.455 ± 0.131). Instead, the result reflects the lower capacity and stronger complexity constraint of symbolic regression, highlighting its role as an interpretable, hypothesis-generating model rather than the most accurate predictor. To sum up, regression comparisons of Extra Trees, Random Forest, XGBoost, and symbolic regression embedded in NJmatWeb demonstrate the complementary roles of the modeling paradigms: machine learning models maximize predictive accuracy of the small materials science dataset, whereas symbolic regression uncovers explicit scientific knowledge in the form of analytical mathematical equations. Together, they enable NJmatWeb to provide both high-performance prediction and interpretable mechanistic insights for adsorption phenomena on perovskite surfaces.

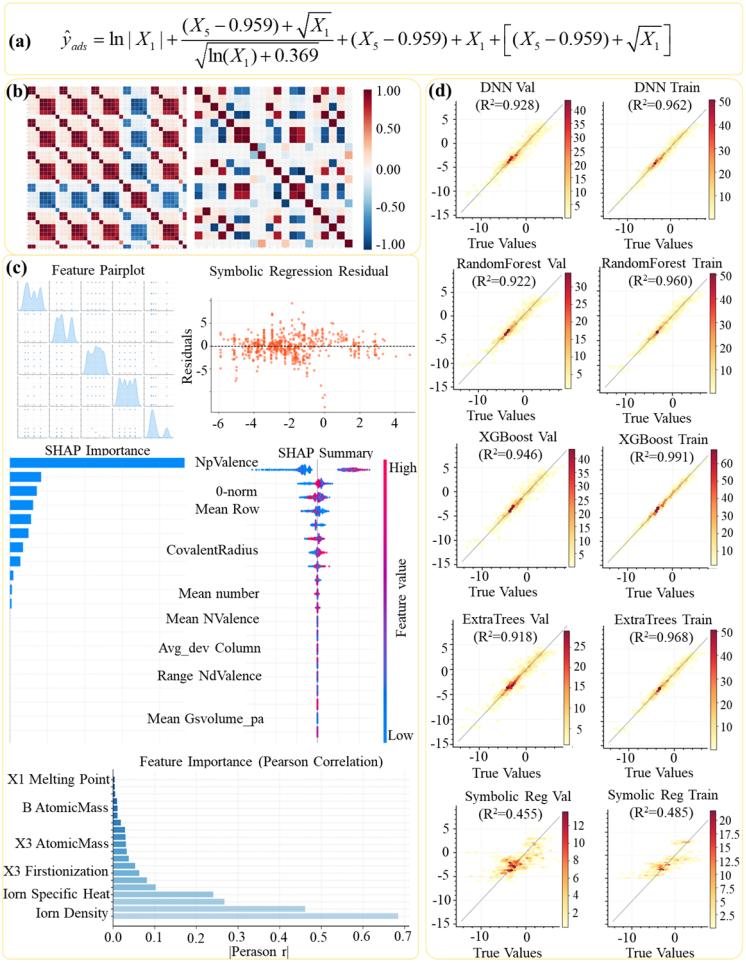


Figure 4. Analysis of perovskite adsorption energy using NJmatWeb. (**a**) Mathematical equation for adsorption energy automatically derived by the symbolic regression module. (**b**) Heat map comparison of the full descriptor set and the selected descriptor subset. (**c**) Integrated model interpretability analysis, including SHAP value distributions, feature importance rankings based on SHAP and Pearson correlation, and residual analysis of the symbolic regression model. (**d**) Regression performance of DNN, Extra Trees, Random Forest, XGBoost, and symbolic regression, showing fitting results for the training (**right**) and validation (**left**) datasets.

Table 3. Performance comparison of different models for predicting adsorption energy on perovskite surfaces.

Model	Train R^2	Validation R^2	MAE	RMSE	Gap (Train-Validation)
DNN	0.962 ± 0.009	0.928 ± 0.014	0.566 ± 0.069	0.817 ± 0.096	0.034
Random Forest	0.960 ± 0.005	0.922 ± 0.014	0.577 ± 0.053	0.850 ± 0.114	0.037
XGBoost	0.991 ± 0.010	0.946 ± 0.016	0.453 ± 0.101	0.707 ± 0.139	0.046
Extra Trees	0.968 ± 0.006	0.918 ± 0.015	0.590 ± 0.063	0.875 ± 0.123	0.050
Symbolic Regression	0.485 ± 0.095	0.455 ± 0.131	1.663 ± 0.212	2.256 ± 0.422	0.030

3.4. Case 4: Perovskite Photovoltage

To further demonstrate the versatility of NJmatWeb, the platform is applied to the classification of perovskite photovoltage (Table 4 and Figure 5) [45]. The ensemble learning models achieve the best classification performance, with Extra Trees reaching the highest validation accuracy and F1-score of 0.908 ± 0.051, followed by Random Forest (accuracy = 0.903 ± 0.047) and XGBoost (accuracy = 0.882 ± 0.042). In comparison, the DNN achieved a validation accuracy of 0.795 ± 0.075, while Logistic Regression and the symbolic regression model achieved validation accuracies of 0.621 ± 0.049 and 0.577 ± 0.059, respectively. However, the high training accuracies (0.958–0.995) combined with lower validation accuracies indicate that these models exhibit overfitting, as evidenced by the train–validation gaps ranging from 0.050 to 0.113. Despite this overfitting, the ensemble models still achieve strong validation performance, confirming that the selected descriptors capture meaningful compositional signals for photovoltage classification: the symbolic expression generates a continuous discriminant score rather than directly predicting the photovoltage value. The equation primarily combines the descriptors X_0, X_3, and X_4, corresponding to the oxygen-related descriptor of component A, the oxygen-related descriptor of component B, and the ETA-BetaP descriptor, respectively. These terms can be viewed as proxies for the chemical environment, polarity, and electronic or topological characteristics of the material components. Their multiplicative and nested-ratio structure indicates that the classification boundary depends on nonlinear interactions among the component descriptors rather than on a single independent feature. The symbolic regression exhibited lower predictive performance than the ensemble learning models, but it provides interpretable mathematical relationships between material descriptors and photovoltage, complementing the high predictive accuracy of black-box machine learning algorithms. These results further demonstrate that NJmatWeb integrates accurate prediction with model interpretability, enabling both reliable classification and scientific insight for diverse materials science tasks.

3.5. Case 5: Multi-Solvent System

To further demonstrate the applicability of NJmatWeb to materials optimization, we investigate the classification of multi-solvent systems for enhancing the photocurrent stability of perovskite materials in aqueous environments (Table 5 and Figure 6) [46]. Identifying suitable solvent combinations is essential for improving the environmental stability and long-term optoelectronic performance of functional hybrid organic-inorganic materials. NJmatWeb enables rapid identification of promising solvent systems through data-driven classification, providing an efficient alternative to conventional trial-and-error approaches. As summarized in Table 5, ensemble learning models consistently achieved excellent classification performance. Extra Trees and Random Forest achieved the highest validation accuracy (0.884 ± 0.098 and 0.884 ± 0.071, respectively), followed by XGBoost (accuracy = 0.865 ± 0.091) and DNN (accuracy = 0.816 ± 0.070). In contrast, Logistic Regression and the symbolic regression model yielded validation accuracies of 0.690 ± 0.108 and 0.691 ± 0.151, respectively, indicating that the relationship between solvent descriptors and stability is highly nonlinear and is better captured by ensemble learning methods. Despite the overfitting (near-perfect training accuracies but lower validation accuracies), the validation accuracies of the ensemble models (0.865–0.884) are substantially higher than those of the linear baseline (0.690), confirming that the selected descriptors encode meaningful information for distinguishing effective multi-solvent systems. The corresponding interpretability analysis further identifies the key descriptors governing photocurrent stability through SHAP analysis and feature importance ranking. The SHAP analysis provides a materials interpretation of the solvent-classification results by linking the predicted stability to the molecular composition and relative solvent descriptors. These descriptors may reflect differences in polarity, hydrogen-bonding ability, molecular size, and other intermolecular-interaction characteristics that influence the aqueous environment of the perovskite material. The results suggest that solvent selection is governed by combined molecular effects rather than by a single descriptor. Importantly, the symbolic regression model achieved a validation accuracy of 0.691 ± 0.151 (F1 = 0.691 ± 0.150). Although less accurate than the ensemble learning models, it provides an explicit mathematical expression linking material descriptors to stability. These results

provide accurate prediction of solvent combinations and reveal the dominant physicochemical factors influencing optoelectronic stability.

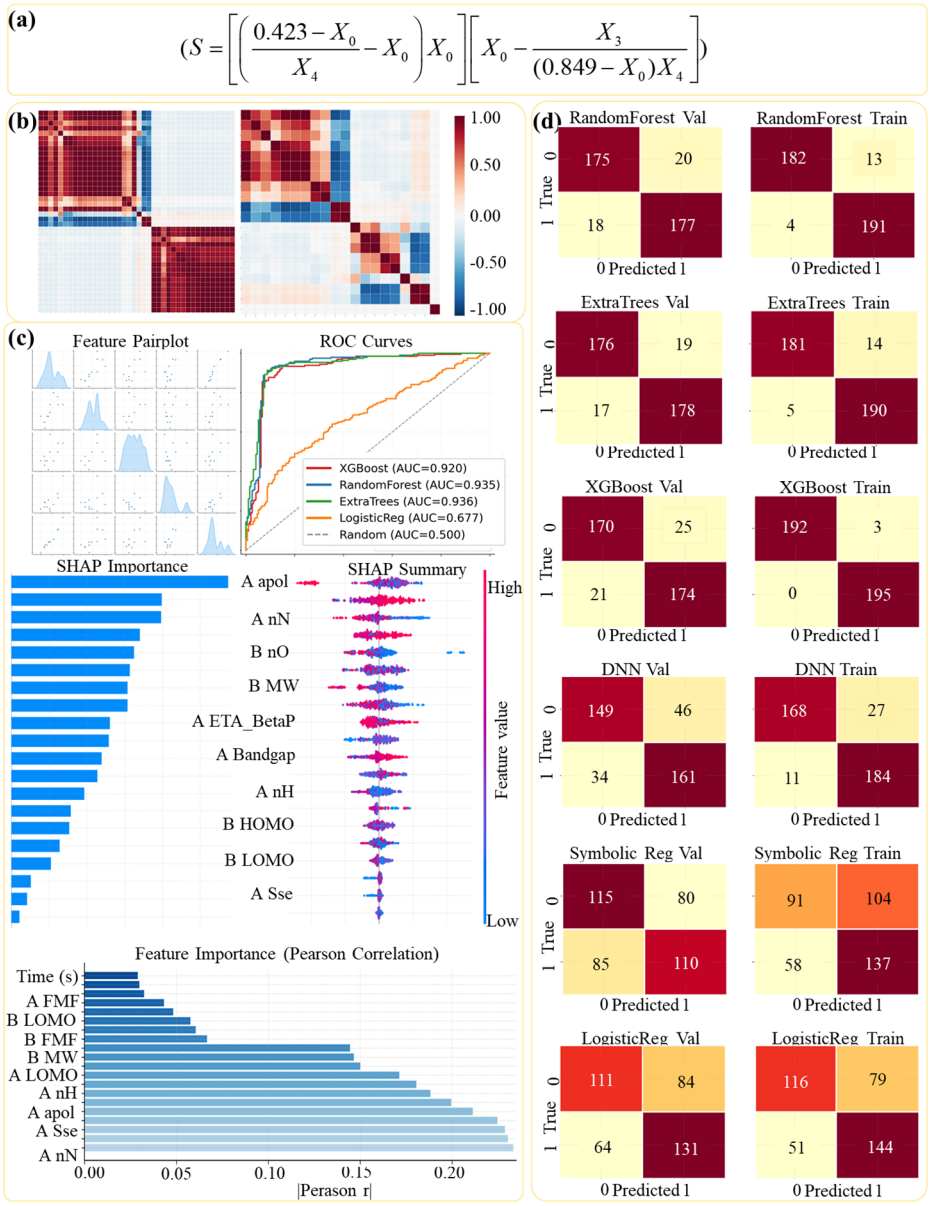


Figure 5. Analysis of perovskite photovoltage prediction using NJmatWeb. (**a**) Heat map comparison of the full descriptor set and the selected descriptor subset. (**b**) Integrated model interpretability analysis, including SHAP value distributions, feature importance rankings based on SHAP and Pearson correlation, and ROC curves. (**c**) Mathematical equation automatically generated by the symbolic regression module and classification performance of DNN, Extra Trees, Logistic Regression, Random Forest, XGBoost and symbolic regression on the training (**right**) and validation (**left**) datasets.

Table 4. Performance comparison of machine learning and symbolic regression models for perovskite photovoltage prediction.

Model	Train Acc	Validation Acc	F1
Random Forest	0.958 ± 0.006	0.903 ± 0.047	0.903 ± 0.047
Extra Trees	0.958 ± 0.006	0.908 ± 0.051	0.908 ± 0.051
XGBoost	0.995 ± 0.002	0.882 ± 0.042	0.882 ± 0.042
DNN	0.923 ± 0.017	0.795 ± 0.075	0.794 ± 0.075
Symbolic Regression	0.606 ± 0.069	0.577 ± 0.059	0.549 ± 0.099
Logistic Regression	0.674 ± 0.020	0.621 ± 0.049	0.618 ± 0.050

Table 5. Performance comparison of machine learning and symbolic regression models for multi-solvent classification.

Model	Train Acc	Validation Acc	F1
Extra Trees	0.959 ± 0.011	0.884 ± 0.098	0.884 ± 0.099
XGBoost	0.995 ± 0.007	0.865 ± 0.091	0.864 ± 0.092
Random Forest	0.959 ± 0.011	0.884 ± 0.071	0.884 ± 0.071
DNN	0.983 ± 0.024	0.816 ± 0.070	0.815 ± 0.070
Symbolic Regression	0.718 ± 0.074	0.691 ± 0.151	0.691 ± 0.150
Logistic Regression	0.811 ± 0.076	0.690 ± 0.108	0.684 ± 0.111

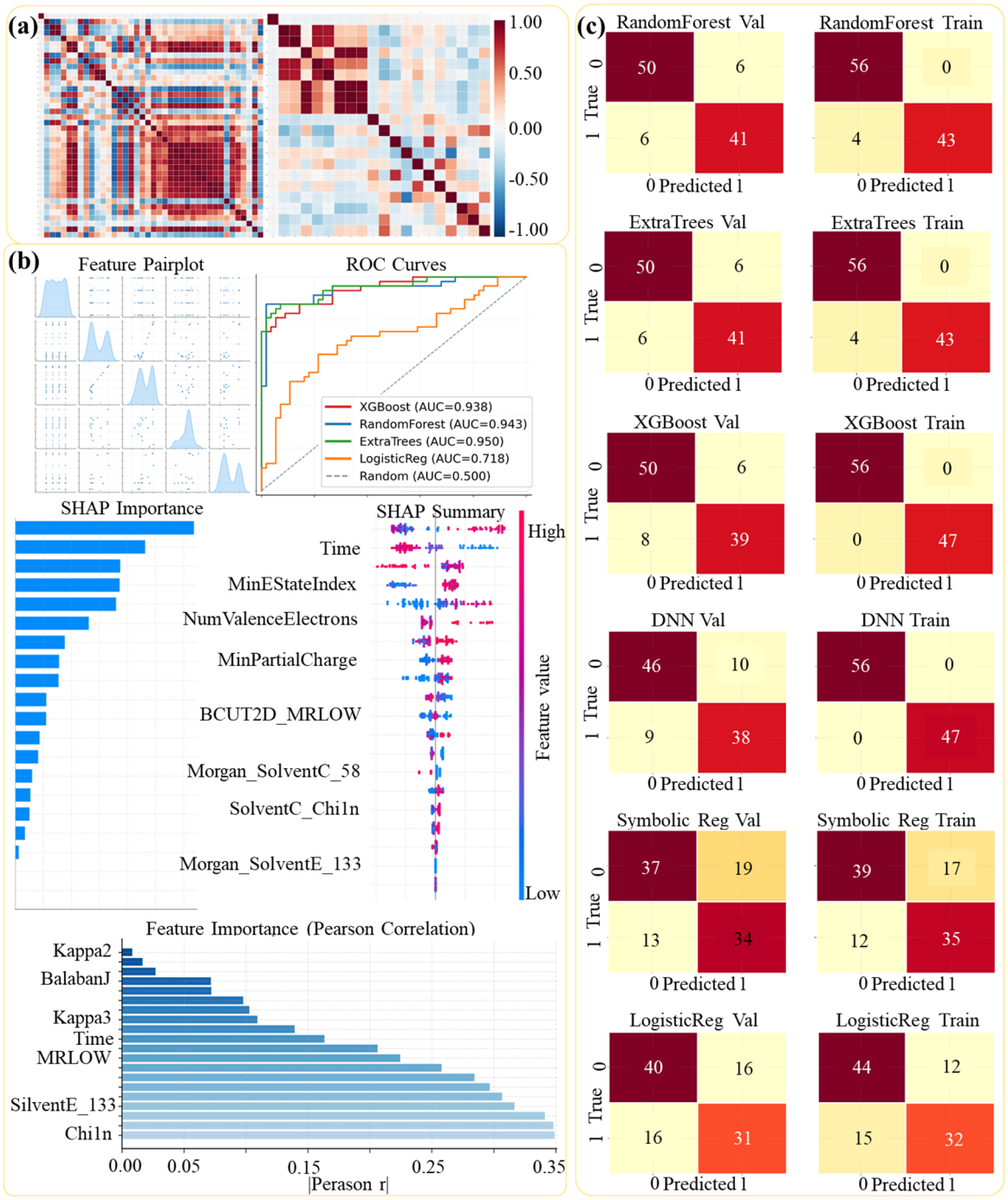


Figure 6. Analysis of multi-solvent prediction for perovskite materials using NJmatWeb. (**a**) Heat map comparison of the full descriptor set and the selected descriptor subset. (**b**) Integrated model interpretability analysis, including SHAP value distributions, feature importance rankings based on SHAP and Pearson correlation, and ROC curves. (**c**) Classification performance of machine learning and symbolic regression models on the training (**right**) and validation (**left**) datasets.

3.6. Cross Platform Comparison and Scope of NJmatWeb

To clarify the positioning of NJmatWeb, we compared its workflow with representative materials-informatics tools and AI platforms, including Matminer, Automatminer, MLMD, JARVIS-Leaderboard, NJmat 2.0, and a manually assembled Matminer + SHAP workflow (Table 6).

Table 6. Compact capability comparison of NJmatWeb with representative materials-informatics tools and AI platforms. "√" indicates an integrated or explicitly reported capability; "—" indicates that the capability is not a primary integrated function in the cited workflow. "External" indicates that the function can be added through separate packages or user-written code.

Platform/Tool	Featurization	Auto ML/CV	SHAP	Symbolic Regression	Design/Benchmarking	Web
NJmatWeb (this work)	√	√	√	√	Forward prediction	√
Matminer [47]	√	—	External	External	—	—
Manual Matminer + SHAP [47]	√	—	√	—	—	—
Automatminer [43]	√	√	—	—	Benchmarking	—
MLMD [12]	√	√	—	—	Active learning/design	√
JARVIS-Leaderboard [13]	Partial	Benchmarking	—	—	√	√
NJmat 2.0 [8]	√	Partial	Partial	Partial	Broad design scope	√

Matminer provides extensive materials featurization, data-access, and visualization utilities, whereas feature selection, model training, validation, and interpretation generally require additional user-assembled code. Automatminer emphasizes automated materials-property prediction and model selection. MLMD has a broader design-oriented scope involving automated descriptor engineering, model optimization, surrogate optimization, and active learning. JARVIS-Leaderboard is primarily a community-driven infrastructure for standardized benchmarking across multiple materials-design task categories. NJmat 2.0 provides a broader collection of materials-AI and data-analysis functions, including machine learning, NLP/LLM, machine-learning potentials, graph neural networks, and data exploration.

In comparison, NJmatWeb focuses on a compact forward-prediction workflow that combines automatic featurization for inorganic and organic inputs, recursive feature elimination, five-fold cross-validation, multiple machine-learning and deep-learning models, SHAP analysis, symbolic regression, and interactive visualization in a single web interface. Relative to a manually assembled Matminer + SHAP workflow, this integration reduces the programming and coordination required to proceed from raw composition or molecular input to feature selection, model evaluation, descriptor interpretation, and symbolic-expression generation. The main contribution of NJmatWeb is therefore workflow integration and accessibility rather than the introduction of a new machine-learning algorithm or a claim of universally superior predictive accuracy.

The broader development of materials AI is moving toward heterogeneous workflows that combine machine learning, high-performance computing, robotics, automated experimentation, and iterative feedback [48,49]. NJmatWeb currently addresses the data-processing, prediction, and interpretation stages of this cycle. It does not yet provide active learning, automated candidate generation, uncertainty-aware inverse design, or closed-loop experimental validation. Therefore, NJmatWeb should be regarded as an accessible and interpretable forward-prediction platform that can provide a foundation for future inverse-design and autonomous-discovery workflows.

4. Limitations

Despite the promising performance demonstrated across multiple materials science datasets (e.g., Tables S1 and S2), NJmatWeb has several limitations that provide opportunities for future development. First, the reliability of the models depends strongly on the size, quality, and chemical diversity of the input dataset. For small or highly imbalanced datasets, the reported performance may be sensitive to the train–validation partition and may not fully represent generalization to chemically distinct materials. The relatively large differences between training and test performance observed for some tree-based models indicate a risk of overfitting, which is common for ensemble methods on small materials datasets. Therefore, model performance should be interpreted together with train–test gaps and, where possible, assessed using repeated cross-validation and external datasets.

Second, the current implementation primarily relies on composition-, structure-, and molecular-descriptor-based representations via Matminer and RDKit. These descriptors do not fully capture microstructure, defects, surface termination, crystallographic disorder, processing history, adsorption geometry, or dynamic environmental effects. Consequently, the learned relationships may provide useful data-driven correlations without constituting a complete physical description of the material system.

Third, the SHAP results explain the contribution of descriptors to a fitted predictive model but do not establish causal mechanisms. Similarly, the symbolic regression equations are empirical approximations whose form and coefficients depend on the selected descriptors, data distribution, and optimization settings. They should therefore be regarded as interpretable, hypothesis-generating relationships rather than universal physical laws.

Fourth, NJmatWeb currently performs forward prediction and model interpretation rather than autonomous inverse materials discovery. It does not yet include active learning, closed-loop experimental feedback, or automated candidate synthesis and validation.

These limitations point to several directions for future development. First, future versions will incorporate a prediction interface that enables users to upload unseen compositions or descriptors and directly obtain predicted material properties, allowing rapid virtual screening prior to experimental validation. Second, we plan to integrate LLM-based assistants capable of automatically interpreting datasets, recommending suitable machine learning algorithms, explaining symbolic regression equations in natural language, and generating publication-ready reports from the prediction results. Such capabilities will lower the barrier for researchers without expertise in artificial intelligence. Third, we aim to combine NJmatWeb with high-throughput calculations and automated experimental fabrication and characterization to establish an AI-driven self-driving laboratory. By continuously integrating computational and experimental feedback into model retraining, NJmatWeb could evolve from a predictive platform into an autonomous materials discovery system capable of accelerating the design and optimization of functional materials.

5. Conclusions

This study introduced NJmatWeb, a web-based infrastructure that automates machine learning and symbolic regression modeling for materials property prediction within a single accessible interface. The platform is validated across multiple case studies spanning diverse materials systems and tasks, including steel yield strength classification, experimental band gap prediction, perovskite surface adsorption energy analysis, perovskite photovoltage prediction, and multi-solvent stability classification for perovskite optoelectronic materials. NJmatWeb integrates recursive feature elimination for dimensionality reduction, SHAP analysis for model interpretability, and symbolic regression, producing explicit mathematical expressions that complement black-box models. The automatic featurizer supports both inorganic compositions and organic SMILES inputs without requiring manual descriptor engineering. Together, these results demonstrate that NJmatWeb provides a unified, interpretable, and user-friendly platform to accelerate the design and discovery of advanced functional materials. The data-driven materials science website is provided in http://njmat.cn.

Supplementary Materials

The additional data and information can be downloaded at: https://media.sciltp.com/articles/others/26102 61404592615/AIMat-26070234-SM.pdf. Table S1. Performance comparison of machine learning and symbolic regression models for perovskite property prediction using the matbench_perovskites.csv dataset. Table S2. Performance comparison of machine learning and symbolic regression models for perovskite formation energy prediction [1]. Table S3. Mapping between symbolic-regression variables and the corresponding descriptors. Reference [50] is cited in supplementary materials.

Author Contributions

L.Z.: conceptualization, supervision, writing—original draft preparation. X.W.: software, visualization, writing—reviewing and editing. Y.W.: supervision. All authors have read and agreed to the published version of the manuscript.

Funding

This research was funded by the Chinese Academy of Sciences (CAS Leading Action Plan), grant number E65543Z101. The APC was funded by the Chinese Academy of Sciences (CAS Leading Action Plan), grant number E65543Z101.

Institutional Review Board Statement

Not applicable.

Informed Consent Statement

Not applicable.

Data Availability Statement

The data that supports the findings of this study are available within the article. The code is provided in https://github.com/Zhang-NJ-Lab/NJmatWeb. The website can be accessed via http://njmat.cn.

Conflicts of Interest

The authors declare no conflict of interest.

Use of AI and AI-Assisted Technologies

During the preparation of this work, the authors used DeepSeek to improve the language and readability of the manuscript. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

References

1. Ahlawat, D.; Mishra, V.; Singh, S.; et al. A Family of Large Language Models for Materials Research with Insights into Model Adaptability in Continued Pretraining. *Nat. Mach. Intell.* **2026**, *8*, 435–448. https://doi.org/10.1038/s42256-026-01199-8.
2. Ge, H.; Liu, J.; Sebek, M.; et al. AI-Assisted Wafer-Scale Exfoliation and Transfer of 2D Materials: Status, Challenges and Perspectives. *AI Sci.* **2025**, *1*, 013002. https://doi.org/10.1088/3050-287x/ae0353.
3. Yin, R.A.; Ren, Z.; Yin, Z.; et al. Collaborative AI Enhances Image Understanding in Materials Science. *AI Mater.* **2026**, *1*, 6. https://doi.org/10.53941/aimat.2026.100006.
4. Zang, Y.; Zhang, Z.; Li, S.; et al. Fiber Neural Networks for the Intelligent Optical Fiber Communication Signal Processing. *iOptics* **2025**, *1*, 100009. https://doi.org/10.1016/j.iopt.2025.100009.
5. Liu, Z.K.; Li, B. Perspectives on Thermodynamics for AI and AI Safety. *AI Sci.* **2026**, *2*, 022001. https://doi.org/10.1088/3050-287x/ae7e9b.
6. Herbst, R.; Stille, D.; Gwilliam, G.F.; et al. Unlocking the Past: The Potential of Large Language Models to Revolutionize Transcription of Natural History Collections. *Data Intell.* **2025**, *7*, 237–264. https://doi.org/10.3724/2096-7004.di.2025.0011.
7. Lin, Z.; Yang, C. Artificial Intelligence Design of Sustainable Aluminum Alloys: A Review. *Comput. Mater. Contin.* **2026**, *86*, 1–33. https://doi.org/10.32604/cmc.2025.070735.
8. Zhang, L.; Deng, H. NJmat 2.0: User Instructions of Data-Driven Machine Learning Interface for Materials Science. *Comput. Mater. Contin.* **2025**, *83*, 1–11. https://doi.org/10.32604/cmc.2025.062666.
9. Gu, C.; Wang, L.; Leng, J.; et al. MIS-VIO: Deep Multi-Modal Visual-Inertial Localization with Cross-Modality Interaction and Selection. *iOptics* **2025**, *1*, 100019. https://doi.org/10.1016/j.iopt.2025.100019.
10. Lin, J.; He, Y.; Du, Q.; et al. AI-Powered Retinal Projection for Immersive Autostereoscopic Display. *iOptics* **2026**, *2*, 100030. https://doi.org/10.1016/j.iopt.2026.100030.
11. Li, H.; Wang, J.; Jiang, Z.; et al. EVCharging-GPT: Predicting Electric Vehicle User Charging Behavior Using Large Language Models. *Data Intell.* **2025**, *7*, 336–357. https://doi.org/10.3724/2096-7004.di.2025.0002.
12. Ma, J.; Cao, B.; Dong, S.; et al. MLMD: A Programming-Free AI Platform to Predict and Design Materials. *npj Comput. Mater.* **2024**, *10*, 59. https://doi.org/10.1038/s41524-024-01243-4.
13. Choudhary, K.; Wines, D.; Li, K.; et al. JARVIS-Leaderboard: A Large Scale Benchmark of Materials Design Methods. *npj Comput. Mater.* **2024**, *10*, 93. https://doi.org/10.1038/s41524-024-01259-w.
14. Lu, T.; Li, M.; Lu, W.; et al. Recent Progress in the Data-Driven Discovery of Novel Photovoltaic Materials. *J. Mater. Inform.* **2022**, *2*, 7. https://doi.org/10.20517/jmi.2022.07.
15. Zhang, S.; Wang, W.Y.; Wang, X.; et al. Large Language Models Enabled Intelligent Microstructure Optimization and Defects Classification of Welded Titanium Alloys. *J. Mater. Inform.* **2024**, *4*, 34. https://doi.org/10.20517/jmi.2024.64.
16. Zhang, L.; Wang, Y.; Zhou, Y.; et al. AI-Guided Design of Optoelectronic Molecularly Modified $CH_3NH_3PbI_3$ Perovskite Thin Films with Improved Aqueous Stability. *AI Sci.* **2026**, *2*, 025006. https://doi.org/10.1088/3050-287x/ae6b24.
17. Butler, K.T.; Davies, D.W.; Cartwright, H.; et al. Machine Learning for Molecular and Materials Science. *Nature* **2018**, *559*, 547–555. https://doi.org/10.1038/s41586-018-0337-2.
18. Raccuglia, P.; Elbert, K.C.; Adler, P.D.F.; et al. Machine-Learning-Assisted Materials Discovery Using Failed Experiments. *Nature* **2016**, *533*, 73–76. https://doi.org/10.1038/nature17439.

19. Tshitoyan, V.; Dagdelen, J.; Weston, L.; et al. Unsupervised Word Embeddings Capture Latent Knowledge from Materials Science Literature. *Nature* **2019**, *571*, 95–98. https://doi.org/10.1038/s41586-019-1335-8.
20. Merchant, A.; Batzner, S.; Schoenholz, S.S.; et al. Scaling Deep Learning for Materials Discovery. *Nature* **2023**, *624*, 80–85. https://doi.org/10.1038/s41586-023-06735-9.
21. Behler, J.; Parrinello, M. Generalized Neural-Network Representation of High-Dimensional Potential-Energy Surfaces. *Phys. Rev. Lett.* **2007**, *98*, 146401. https://doi.org/10.1103/physrevlett.98.146401.
22. Thompson, A.P.; Swiler, L.P.; Trott, C.R.; et al. Spectral Neighbor Analysis Method for Automated Generation of Quantum-Accurate Interatomic Potentials. *J. Comput. Phys.* **2015**, *285*, 316–330. https://doi.org/10.1016/j.jcp.2014.12.018.
23. Zhang, L.; Han, J.; Wang, H.; et al. Deep Potential Molecular Dynamics: A Scalable Model with the Accuracy of Quantum Mechanics. *Phys. Rev. Lett.* **2018**, *120*, 143001. https://doi.org/10.1103/physrevlett.120.143001.
24. Batzner, S.; Musaelian, A.; Sun, L.; et al. E(3)-Equivariant Graph Neural Networks for Data-Efficient and Accurate Interatomic Potentials. *Nat. Commun.* **2022**, *13*, 2453. https://doi.org/10.1038/s41467-022-29939-5.
25. Muuga, K.; Knijff, L.; Zhang, C. Molecular Electrostatic Potentials from Machine Learning Models for Dipole and Quadrupole Predictions. *AI Sci.* **2026**, *2*, 025002. https://doi.org/10.1088/3050-287x/ae531a.
26. Chen, C.; Ong, S.P. A Universal Graph Deep Learning Interatomic Potential for the Periodic Table. *Nat. Comput. Sci.* **2022**, *2*, 718–728. https://doi.org/10.1038/s43588-022-00349-3.
27. Wood, B.M.; Dzamba, M.; Fu, X.; et al. UMA: A Family of Universal Models for Atoms. *arXiv* **2025**, arXiv:2506.23971.
28. Zhang, D.; Liu, X.; Zhang, X.; et al. DPA-2: A Large Atomic Model as a Multi-Task Learner. *npj Comput. Mater.* **2024**, *10*, 293. https://doi.org/10.1038/s41524-024-01493-2.
29. Ansari, M.; Watchorn, J.; Brown, C.E.; et al. dZiner: Rational Inverse Design of Materials with AI Agents. *arXiv* **2024**, arXiv:2410.03963.
30. Adachi, Y.; Chen, T.T.; Sun, F.; et al. A Review on Inverse Analysis Models in Steel Material Design. *Mater. Genome Eng. Adv.* **2024**, *2*, e71. https://doi.org/10.1002/mgea.71.
31. Sanchez-Lengeling, B.; Aspuru-Guzik, A. Inverse Molecular Design Using Machine Learning: Generative Models for Matter Engineering. *Science* **2018**, *361*, 360–365. https://doi.org/10.1126/science.aat2663.
32. Noh, J.; Kim, J.; Stein, H.S.; et al. Inverse Design of Solid-State Materials via a Continuous Representation. *Matter* **2019**, *1*, 1370–1384. https://doi.org/10.1016/j.matt.2019.08.017.
33. Boussinot, G.; Apel, M. Phase Field and Analytical Study of Mushy Zone Solidification in a Static Thermal Gradient: From Dendrites to Planar Front. *Acta Mater.* **2017**, *122*, 310–321. https://doi.org/10.1016/j.actamat.2016.09.053.
34. Xie, J. Prospects of Materials Genome Engineering Frontiers. *Mater. Genome Eng. Adv.* **2023**, *1*, e17. https://doi.org/10.1002/mgea.17.
35. Xu, K.; Lu, Y.; Liu, L.; et al. MatterMind: A User-Friendly Material Design Platform Assisted by Large Language Model. *Mater. Genome Eng. Adv.* **2026**, *4*, e70075. https://doi.org/10.1002/mgea.70075.
36. Hu, Y.C. Data-Driven Prediction of the Glass-Forming Ability of Modeled Alloys by Supervised Machine Learning. *J. Mater. Inform.* **2023**, *3*, 1. https://doi.org/10.20517/jmi.2022.28.
37. Zhang, L.; Zhou, J.; Chen, X. Data-Driven Exploration and First-Principles Analysis of Perovskite Material. *J. Mater. Inform.* **2024**, *4*, 13. https://doi.org/10.20517/jmi.2024.20.
38. He, F.; Lai, H.; Liu, J.; et al. Solving Mathematical Problems Using Large Language Models: A Survey. *Data Intell.* **2025**, *7*, 907–946. https://doi.org/10.3724/2096-7004.di.2025.0064.
39. Zeng, Y.; Jin, X. AI-Empowered Optical Metasurfaces: From Unit-Cell Optimization to System-Level Integration. *iOptics* **2025**, *1*, 100004. https://doi.org/10.1016/j.iopt.2025.100004.
40. Ullah, H.; Zhou, S.; Du, K.; et al. Intrinsic Dark-Field Fourier Ptychographic Diffraction Tomography under Non-Matched Illumination. *iOptics* **2025**, *1*, 100006. https://doi.org/10.1016/j.iopt.2025.100006.
41. Pan, Z.; Zhou, Y.; Zhang, L. Photoelectrochemical Properties, Machine Learning, and Symbolic Regression for Molecularly Engineered Halide Perovskite Materials in Water. *ACS Appl. Mater. Interfaces* **2022**, *14*, 9933–9943. https://doi.org/10.1021/acsami.2c00568.
42. Xie, J.; Zhang, L. Machine Learning and Symbolic Regression for Adsorption of Atmospheric Molecules on Low-Dimensional TiO_2. *Appl. Surf. Sci.* **2022**, *597*, 153728. https://doi.org/10.1016/j.apsusc.2022.153728.
43. Dunn, A.; Wang, Q.; Ganose, A.; et al. Benchmarking Materials Property Prediction Methods: The Matbench Test Set and Automatminer Reference Algorithm. *npj Comput. Mater.* **2020**, *6*, 138. https://doi.org/10.1038/s41524-020-00406-3.
44. Hu, W.; Zhang, L.; Pan, Z. Designing Two-Dimensional Halide Perovskites Based on High-Throughput Calculations and Machine Learning. *ACS Appl. Mater. Interfaces* **2022**, *14*, 21596–21604. https://doi.org/10.1021/acsami.2c00564.
45. Zhang, L.; Hu, W.; He, M.; et al. Optimizing Photoelectrochemical Photovoltage and Stability of Molecular Interlayer-Modified Halide Perovskite in Water: Insights from Interpretable Machine Learning and Symbolic Regression. *ACS Appl. Energy Mater.* **2023**, *6*, 5177–5187. https://doi.org/10.1021/acsaem.2c04066.

46. Huang, Y.; Li, S.; Zhang, L. Accelerated Multisolvent Prediction for Aqueous Stable Halide Perovskite Materials. *ACS Appl. Mater. Interfaces* **2023**, *15*, 48771–48784. https://doi.org/10.1021/acsami.3c09507.
47. Ward, L.; Dunn, A.; Faghaninia, A.; et al. Matminer: An Open Source Toolkit for Materials Data Mining. *Comput. Mater. Sci.* **2018**, *152*, 60–69. https://doi.org/10.1016/j.commatsci.2018.05.018.
48. Nematov, D.; Hojamberdiev, M. Machine Learning-Driven Materials Discovery: Unlocking Next-Generation Functional Materials—A Review. *Comput. Condens. Matter* **2025**, *45*, e01139. https://doi.org/10.1016/j.cocom.2025.e01139.
49. Pyzer-Knapp, E.O.; Pitera, J.W.; Staar, P.W.J.; et al. Accelerating Materials Discovery Using Artificial Intelligence, High Performance Computing and Robotics. *npj Comput. Mater.* **2022**, *8*, 84. https://doi.org/10.1038/s41524-022-00765-z.
50. Hu, W.; Zhang, L. High-Throughput Calculation and Machine Learning of Two-Dimensional Halide Perovskite Materials: Formation Energy and Band Gap. *Mater. Today Commun.* **2023**, *35*, 105841. https://doi.org/10.1016/j.mtcomm.2023.105841.