Advanced Intelligent Discovery

WILEY VCH

ADVANCED INTELLIGENT DISCOVERY Open Access

REVIEW OPEN ACCESS

Feature Selection for Machine Learning-Driven Accelerated Discovery and Optimization in Emerging Photovoltaics: A Review

Jiyun Zhang[1,2] | Jiayi Tan[1,2] | Qizhen Song[2] | Tian Du[1,2] | Jens Hauch[1] | Christoph J. Brabec[1,2]

[1]Helmholtz-Institute Erlangen-Nürnberg (HI ERN), High-Throughput Methods in Photovoltaics, Forschungszentrum Jülich GmbH, Immerwahrstraße 2, Erlangen, Germany | [2]Faculty of Engineering, Department of Material Science, Institute of Materials for Electronics and Energy Technology (i-MEET), Friedrich-Alexander-University Erlangen-Nuremberg (FAU), Martensstrasse 7, Erlangen, Germany

Correspondence: Jiyun Zhang (jiyun.zhang@fau.de) | Tian Du (t.du@fz-juelich.de) | Christoph J. Brabec (christoph.brabec@fau.de)

Received: 3 February 2025 | **Revised:** 28 April 2025 | **Accepted:** 27 May 2025

Funding: Bavarian State Government, Grant/Award Number: 44-6521a/20/4; SolMAP; SolarTAP–a Technology Acceleration Platform for Emerging Photovoltaics Project by Helmholtz Association; Aufbruch Bayern initiative of the state of Bavaria; EnCN and Solar Factory of the Future; German Research Foundation, Grant/Award Number: SFB953-No. 182849149; GRK2495, Grant/Award Number: ITRG2495

Keywords: emerging photovoltaics | feature selection | gaussian process regression | high-dimensional data | high-throughput experimentation | machine learning | mRMR method

ABSTRACT

Developing reliable emerging photovoltaic (e-PV) technologies requires high-throughput material discovery, device design, and processing optimization. However, the effective process of the resulting high-dimensional, multivariate datasets remains a significant challenge. Integrating feature selection methods and machine learning (ML) provides a robust solution to reduce data dimensionality, improve predictive accuracy, and uncover material performance mechanisms. This review summarizes the advancements in synergizing feature selection methods, particularly the maximum relevance minimum redundancy (mRMR) method embedded, with Gaussian process regression (GPR) to advance e-PVs research. It highlights the importance of integrating feature selection with ML and high-throughput experimentation (HTE) frameworks to accelerate material screening, optimize manufacturing processes, and predict stability. Additionally, the review discusses key challenges such as data quality and model scalability and offers promising strategies to address these limitations. This data-driven approach offers a systematic pathway toward the accelerated discovery and optimization of e-PV technologies.

1 | Introduction

Developing next-generation photovoltaic technologies, such as organic photovoltaics (OPVs) and perovskite solar cells (PSCs), is critical to meeting the ever-growing global demand for renewable energy [1, 2]. These emerging technologies, with their high photoelectric conversion efficiency, low manufacturing costs, and excellent flexibility, offer viable solutions for achieving sustainable energy goals. However, challenges such as further improving efficiency, enhancing environmental stability, and optimizing preparation processes remain barriers to their large-scale application [3–5]. Automated experimental platforms have emerged as powerful tools for addressing these issues [6–10]. These platforms enable high-throughput experimentation (HTE), processing large samples in parallel while improving efficiency, minimizing errors, and ensuring reproducibility [11–14]. For example, researchers can use these automated platforms to explore high-dimensional material combinations and process parameter spaces to accelerate material screening and device optimization [15–18]. These platforms generate large-scale

Jiyun Zhang and Jiayi Tan contributed equally to this study.

datasets encompassing material composition, preparation conditions, device performance, and environmental stability information within a short timeframe [19–22].

However, effectively processing such high-dimensional and multivariate data to extract meaningful insights is posed with various challenges, including data redundancy, noise interference, and inefficiency in data analysis [23, 24]. These challenges limit the pace of material and process development, and add complexity to experimental workflows in current photovoltaic research [25]. Feature selection techniques offer an effective solution to tackle these challenges. By isolating the most relevant variables from a large dataset, they reduce data dimensionality and redundancy, enhance the predictive accuracy of machine learning (ML) models, and reveal key mechanisms underlying the discovered material and device performance [26–29]. Within the integrated framework of ML-driven HTE platforms, feature selection plays a pivotal role in optimizing material properties, process parameters, and device designs. By integrating automated experimentation, ML, and advanced data processing methods, the development of e-PV research is entering a new era [30]. These holistic approaches address major challenges in renewable energy and paves the way for transformative advancements in the global energy landscape [31, 32].

This review highlights the critical role of feature selection in advancing ML-driven approaches for e-PV material discovery and optimization. It discusses techniques such as the maximum relevance minimum redundancy (mRMR)-embedded Gaussian process regression (GPR) model, which streamline high-dimensional datasets, reduce computational burden, and identify key variables influencing material properties and device performance. The integration of feature selection with ML and HTE frameworks is shown to effectively uncover structure-property relationships, optimize processing parameters, and improve environmental stability analysis. The review summarizes key applications in process optimization, stability evaluation, and the discovery of low-dimensional and lead-free hybrid organic–inorganic perovskites (HOIPs) with optimal properties. Possible challenges such as data quality and model scalability are briefly discussed, alone with prospectives strategies for addressing these limitations.

2 | Feature Selection Methods

Feature selection identifies the most relevant features to reduce data dimensionality, improve model performance, and enhance data interpretability [33, 34]. Based on selection mechanism, these methods are classified into filter, wrapper, and embedded methods (Figure 1) [35, 36]. Filter methods independently evaluate feature importance using statistical metrics or information theory indicators. They select features based on intrinsic properties and relevance to the target variable, without considering feature interactions or relying on a specific ML model, and applied as preprocessing step before model training [23]. Filter methods using statistical metrics or information theory indicators to assess feature importance independently of ML models are highly efficient for high-dimensional data. Examples include variance screening, chi-square test, and the mRMR method, based on mutual information and the Laplacian Score [37, 38]. However, filter methods may fail to capture interactions between features, limiting their uses in complex datasets. Wrapper methods, such as recursive feature elimination, combine ML models to evaluate feature subsets through multiple rounds of training [39, 40]. These methods can capture complex feature interactions but are computationally intensive, especially for datasets with a multitude of features, where computational costs grow exponentially. Embedded methods select features during the training process by analyzing feature weights generated by the model. Common techniques include Lasso regression with L_1 regularization and tree-based feature selection. These methods strike a balance between computational efficiency and model accuracy, although their effectiveness can depend on the selected model and regularization parameters, potentially introducing variability in the results. For datasets with redundant or highly correlated features, dimensionality reduction methods like principal component analysis (PCA) [41–44], linear discriminant analysis (LDA) [45], and locally linear embedding (LLE) [46] simplify feature dimensions. These methods reduce noise and computational costs but can obscure the physical or chemical meaning of features and may fail to capture nonlinear interactions, especially in image-type datasets. Each method has unique characteristics and is suitable for different data structures and research needs.

In addition to traditional linear methods such as PCA and mRMR, advanced nonlinear feature selection techniques have attracted growing attention for their ability to identify complex interactions in high-dimensional materials datasets. These methods can be broadly classified into kernel-based methods, manifold learning strategies, and tree-based ensemble models. Kernel-based methods, such as Hilbert–Schmidt Independence Criterion combined with Lasso regression (HSIC-Lasso), capture nonlinear dependencies between features and targets by mapping data into a

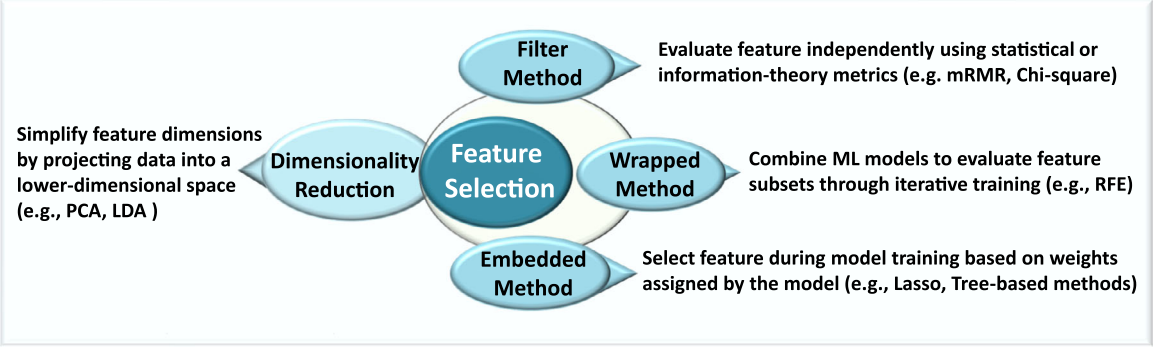

FIGURE 1 | Feature selection methods and simplified categories.

reproducing kernel Hilbert space (RKHS) [47–49]. These methods, which leverage criteria like the Hilbert–Schmidt Independence Criterion or minimizing the trace of the conditional covariance operator, identify predictive and nonredundant features beyond the scope of linear models. They are particularly effective at uncovering hidden nonlinear correlations in complex material systems. Additionally, GPR, often used for regression, can also support feature selection through automatic relevance determination (ARD) kernels. In this context, GPR assigns a length-scale parameter to each feature, optimizing them during training. Features with shorter length-scales contribute more to predictive performance, to implicit feature ranking and selection. GPR has proven effective for complex, high-dimensional photovoltaic datasets by modeling nonlinear relationships and enabling embedded feature selection [50]. Manifold learning approaches, such as Isomap and t-SNE, provide powerful tools for nonlinear dimensionality reduction. Isomap preserves global geodesic distances, while t-SNE effectively visualizes high-dimensional data by preserving local structures [51]. By projecting data onto low-dimensional intrinsic manifolds, these techniques improve visualization and can enhance the robustness and interpretability of ML models. Tree-based ensemble models, such as random forests and XGBoost, enable scalable and interpretable feature selection by leveraging feature importance scores derived from decision-tree splits [52]. These models are well-suited for high-dimensional and noisy datasets due to their ability to model nonlinear dependencies and handle data heterogeneity. Furthermore, techniques like FeatBoost, a boosting-based strategy, improve tree-based feature selection by reweighting samples and prioritizing informative features, thereby reducing redundancy and optimizing subset relevance.

Overall, these nonlinear feature selection techniques significantly enhance the analytical capabilities available for high-throughput, data-intensive photovoltaic research, facilitating the discovery of complex structure–property relationships.

Given the diversity of feature selection methods, it is important to assess their relative advantages and limitations when applied to complex, high-dimensional datasets in e-PV research. Table 1 presents a comparative summary of different feature selection methods, including their respective strengths, limitations, and recommended applications scenarios.

Feature selection methods are widely used in PV research to address challenges related to high-dimensional data processing [37, 53, 54]. While both linear and nonlinear feature selection methods provide valuable tools for dimensionality reduction and insight extraction, each approach has limitations when applied individually to process complex, high-dimensional datasets in e-PV research. Linear methods often fail to capture higher-order or nonlinear dependencies critical for accurate modeling, while sole nonlinear techniques, though powerful, can become computationally intensive or prone to overfitting without proper dimensionality control.

To overcome these challenges, hybrid frameworks combining the strengths of both paradigms have gained increasing attention. A promising strategy is the hierarchical integration of mRMR method and GPR equipped with ARD kernels. This mRMR-GPR framework combines efficient, information-theoretic feature filtering with nonlinear, kernel-based modeling to offer robust, interpretable, and scalable feature selection for high-throughput photovoltaic material discovery and optimization. In the first stage, mRMR hierarchically reduces the feature space by preserving variables with high relevance to the output while minimizing redundancy. This two-tier filtering process minimizes the risk of excluding weakly correlated but functionally critical features. In the second stage, the reduced feature set is evaluated using GPR with an ARD kernel, which quantifies the nonlinear contribution of each feature through optimized length-scale parameters. Features with higher predictive significance—indicated by shorter length scales—are prioritized for final model development. By combining the computational efficiency of filter-based selection with the representational power and interpretability of kernel-based modeling, this hybrid approach provides a robust and scalable solution for feature selection in complex scientific domains. Information theory-based approaches are particularly effective in capturing nonlinear relationships in e-PV research, with key applications in optimizing material properties, refining process conditions, and analyzing environmental stability [55]. When incorporated into an ML-driven autonomous framework, feature selection accelerates material discovery and performance optimization. The following section summarizes the applications of these feature selection-embedded frameworks in data-driven optimization strategies for e-PV technologies.

TABLE 1 | Summary of the advantages, limitations, and application suitability of different feature selection methods.

Method type	Advantages	Limitations	Applications
Filter Method	Fast; scalable; model-independent; simple implementation	Ignores feature interactions; mostly linear; may overlook synergistic effects	Initial descriptor screening in high-dimensional datasets
Wrapper Method	Captures complex interactions; can significantly improve model accuracy	Computationally intensive; prone to overfitting; limited scalability	Small-to-medium datasets where predictive accuracy is a priority
Embedded Method	Integrated with model training; balances efficiency and interpretability	Model-dependent; may overlook nonlinear or higher-order interactions	General tasks requiring interpretable models and moderate computational cost
Nonlinear and advanced Method	Capture nonlinear and high-order feature dependencies; scalable with parallel computing	Interpretation may require auxiliary tools; increased computational cost	Complex datasets with unknown or nonlinear structure–property relationships

3 | ML-Driven Discovery and Optimization of New Photovoltaic Materials

Figure 2 illustrates a typical data-driven workflow that integrates HTE, ML, and feature engineering for e-PV material discovery and process optimization. The process begins with formulating a research plan, wherein specific targets, hypotheses, or concepts are defined to guide the subsequent experiments. Based on these objectives, high-throughput materials and device processing are performed to generate a broad range of experimental samples and datasets. Experimental data, supplemented by curated databases and literature, are then systematically collected to form a high-dimensional dataset. A critical step in the workflow is feature selection, where statistical and ML-based techniques, such as mRMR and PCA, are used to extract the most informative descriptors. This step reduces data dimensionality, enhances model interpretability, and improves computational efficiency [56]. Subsequently, ML models such as GPR, gradient boosting methods, and deep learning frameworks are trained using the refined feature sets to predict material properties and device performances. Model outputs are iteratively refined through feedback loops, wherein explainable AI methods such as SHapley Additive exPlanations (SHAP) and Local Interpretable Model-agnostic Explanations (LIME) enable the interpretation of feature contributions and physicochemical insights [17, 57]. This iterative workflow enables the identification of promising photovoltaic materials and the optimization of fabrication parameters, offering a systematic pathway for accelerating material innovation.

4 | Integrating Feature Selection into ML-Driven HTE Frameworks for Photovoltaic Research

As the efficiency of e-PV technologies improves, stability issues have become a bottleneck hindering their commercial application. Investigating the effects of chemical structures, material components, processing conditions, and device parameters on long-term operational stability are challenged by the nonlinear and often interdependent relationships among these variables. ML-enhanced autonomous frameworks, particularly feature selection techniques like mRMR combined with GPR, emerged as powerful tools to tackle the complexity [58]. The following section highlights representative studies that demonstrate the application of such frameworks in advancing e-PV research, with a particular focus on how feature selection contributes to uncovering structure–property–performance relationships.

4.1 | Process Optimization: Linking Process Parameters to Performance

Understanding and predicting how processing parameters influence microstructural features is critical to the large-scale production of OPVs [59]. However, traditional methods are often time-consuming, complex, and lack the statistical ability needed for comprehensive evaluation. An ML-driven high-throughput platform was developed to address these limitations. Automated UV-Vis spectral modeling was employed to extract key morphological features, which modeled the spectra as superpositions of donor and acceptor contributions from ordered and amorphous phases, characterized by parameters such as vibronic replica strength and bandwidth. Using a refined mRMR-embedded GPR model, Liu et al. systematically selected features that were both highly relevant and nonredundant (Figure 3A) [60]. This feature selection process involved iteratively adding features to a predictive model, guided by GPR to minimize RMSE on test datasets [62].

The process retained only features that provided new and nonredundant information. GPR surrogate models were validated, and uncertainty was assessed using bootstrap-like methods with multiple train/test splits and a reliable shuffle criterion. The effects of solvents, such as chloroform and eco-friendly o-xylene, and PCBM additives on layer thickness, microstructural order, and stability, were analyzed. PCBM was found to improve fill factor (FF) stability by reducing energetic

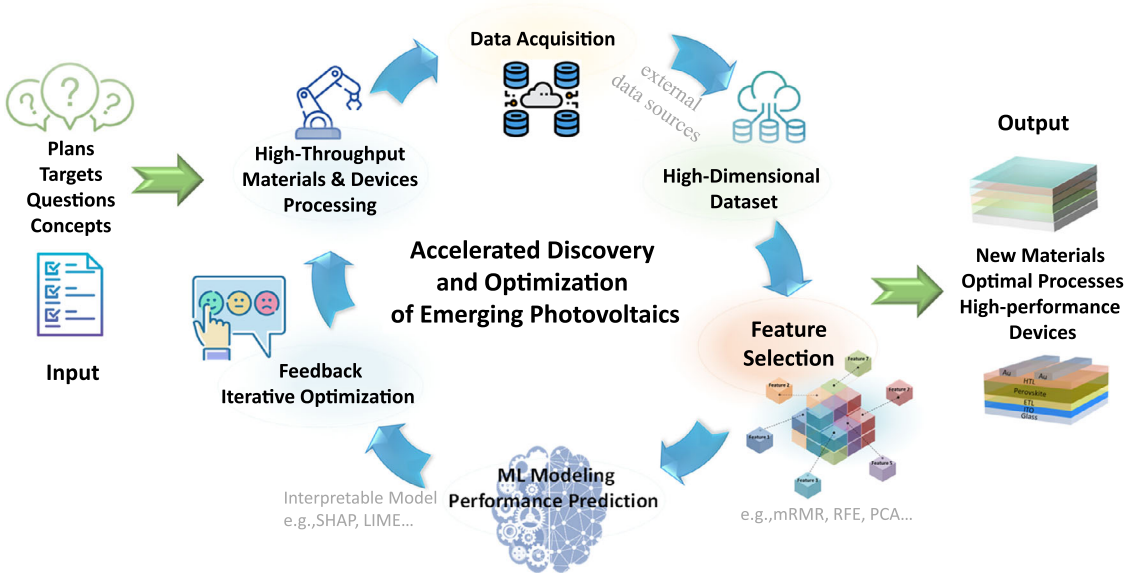

FIGURE 2 | A typical data-driven optimization workflow through integrating feature selection and ML-driven automation for the discovery of new photovoltaic materials.

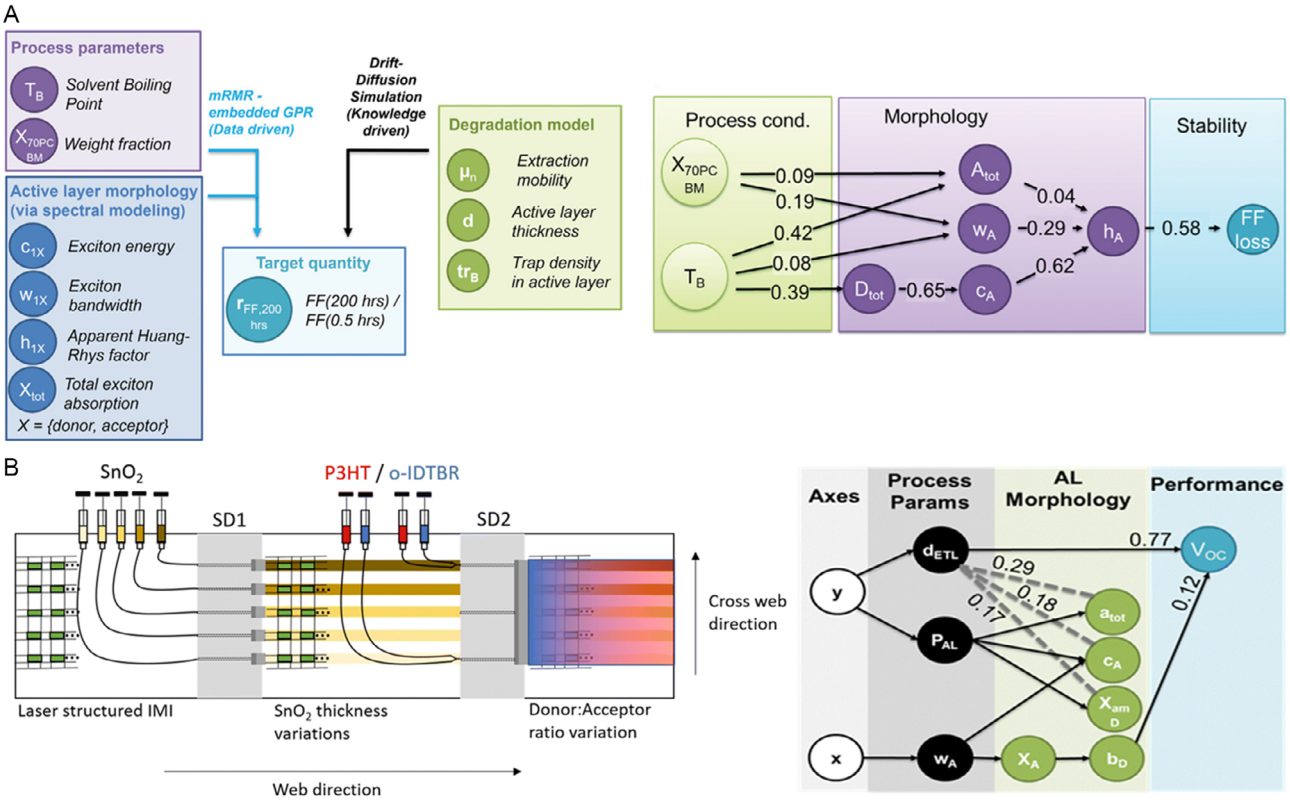

FIGURE 3 | (A) Machine-learning workflow integrating HTE, characterization, and drift-diffusion simulations to establish quantitative structure-property relationships, with mRMR-GPR analysis revealing processing–morphology–stability interdependencies in PM6:DTY6:[70]PCBM devices. Adapted with permission [60]. Copyright 2023, John Wiley and Sons. (B) Schematic of the high-throughput experiment and knowledge graph from mRMR-GPR runs, illustrating direct causations (solid lines), correlations (dashed lines), and the variance explained by each parameter. Adapted with permission [61]. Copyright 2023, Royal Society of Chemistry.

dispersion in the ordered DTY6 phase and improving chain ordering. By adjusting the additive concentration and active layer thickness, both microstructure optimization and device stability improvement were achieved. This study overcomes the limitations of large-scale statistical approaches, clarifies the causal links between processing conditions, microstructure, and stability, and provides a cost-effective framework for understanding OPV degradation and improving device performance and lifetime.

To address the challenge of scaling printed photovoltaics from laboratory optimization to industrial roll-to-roll (R2R) production, a combinatorial approach was developed to simultaneously vary the electron transport layer (ETL) thickness and donor-to-acceptor(D:A) ratio within a single slot-die coating run [61]. Using a multi-nozzle slot-die setup, 3750 OSCs were fabricated with controlled parameter variations, enabling high-throughput data collection (Figure 3B). To analyze the large multivariate dataset, a mRMR feature selection scheme embedded into GPR was applied. This approach ensured only predictors contributing significant additional variance explanations were included in the model, avoiding redundancy. The mRMR-GPR workflow incrementally selected features based on nonlinear correlation with performance metrics such as open-circuit voltage (V_{OC}), FF, and short-circuit current (J_{SC}), as demonstrated by the analysis of UV-Vis spectra-derived features. The addition of multiple predictors revealed nonlinear relationships that were obscured in single-feature analysis. The GPR model quantified these relationships and estimated prediction uncertainty through cross-validation and bootstrap-like resampling. The study identified optimal parameters of an ETL thickness of 17 ± 4 nm and a D:A ratio of 1:1.44, with voltage losses in donor-rich blends linked to incomplete ETL coverage and increased interfacial recombination. Spatial gradients in processing conditions, including temperature and ink mixing, were also uncovered and linked to variations in layer morphology and device performance. By correlating UV-Vis-derived morphological features with photovoltaic metrics, the study highlights the critical role of process-induced interfacial changes on device behavior. This integration of mRMR-GPR-based statistical modeling with HTE not only accelerats process optimization but also provids insights into subtle process fluctuations. These results demonstrate the potential of combining data-driven modeling with physics-informed analysis to enhance process control, improve production yields, and drive photovoltaic commercialization, with applicability to other e-PV technologies such as PSCs.

4.2 | Stability Analysis: Identifying Critical Degradation Mechanisms

The stability of OPV devices under ambient air and light exposure is influenced by multiple factors, including chemical structure, energy level, and morphology parameters [63, 64]. However, the complex interdependence among these

characteristics limits the effectiveness of analyzing individual parameters in isolation. Using an automated device acceleration platform (DAP), Amanda Line One, combined with mRMR-embedded GPR method, the most relevant features associated with air and light stability were identified from a dataset comprising over 40 donor–acceptor combinations (Figure 4A) [65]. Compared to traditional methods, the mRMR method significantly reduces feature redundancy and improve model performance. While standalone mRMR, implemented via the "mRMR-selection" package, ranks features based on linear correlations, mRMR-embedded GPR further enhances performance by capturing nonlinear, multidimensional trends (Figure 4B). This approach consistently outperformed linear feature selection across all feature groups and parameters. Moreover, standalone mRMR combined with GPR demonstrated scalability by handling 3200 descriptors generated using the Mordred algorithm. This model identified "ATS8s" of the acceptor as the top predictor for r_{Jsc}, explaining 70% of its variance. ATS8s represents the centered Moreau–Broto autocorrelation of order 8, reflecting electron configurations between atoms separated by eight bonds, and suggests that oxygen, rather than superoxide anions, is the primary degradation factor under the studied conditions. Guided by this automated selection strategy, accurate predictions were made based solely on pre-experimental structural features, such as the number of fluorine atoms in the acceptor (NAF) and the number of spiro-bridged structures. The study identified effective gap ($E_{g,\ eff}$) as the strongest predictor of air/light resilience. These findings demonstrate the capacity of the integrated framework to reveal causal relationships between molecular structure and stability for the rapid material screening and optimization. The study presents a versatile data-driven approach for analyzing complex material systems and paves the way for molecular inverse design by identifying structural motifs that optimize device performance while addressing stability requirements.

The development of e-PV cells for outer space applications faces significant challenges due to complex UVC degradation mechanisms, limited material datasets, and a lack of universal design principles for radiation resistance [67, 68]. To address these, a high-throughput platform was utilized to degrade and characterize hole transport materials (HTMs), generating a comprehensive dataset of stability metrics, such as photon dose at 90% absorbance (Figure 4C) [66]. A two-stage mRMR feature selection process was applied to analyze the ≈1700 molecular descriptors extracted from the dataset. In the first step, upstream mRMR reduced the descriptors to 10 based on linear correlations with stability target. In the second stage, mRMR-embedded GPR further narrowed them to three key predictors using a greedy algorithm that iteratively selected descriptors with the best predictive power while minimizing redundancy. The final GPR model revealed nonlinear relationships between predictors—such as fused aromatic ring clusters and vinyl ratios—and stability, achieving an R^2 of 0.631. The results show that fused aromatic ring clusters enhance stability, while vinyl, thiophene, and methoxy groups lead to degradation. This integrated framework of high-throughput experiments and ML not only uncovers key structure–stability relationships but also provides a platform for designing radiation-resistant materials. By using predictive modeling and feature selection, this approach accelerates the discovery and optimization of durable photovoltaics for extreme conditions application.

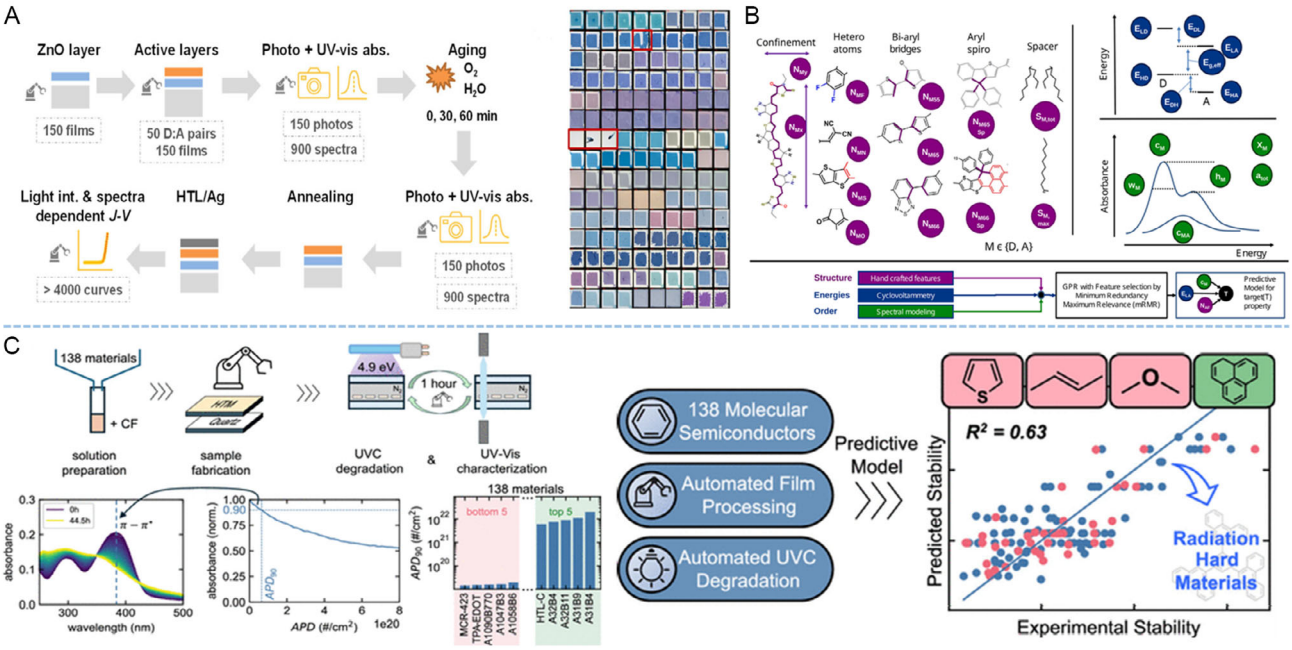

FIGURE 4 | (A) Workflow for automated organic layer and device fabrication and characterization. (B) Structural, energetic, and order-related features used to predict r_{Jsc}, with energetic values from cyclic voltammetry and order-related values from UV–vis spectral modeling before electrode deposition. ML workflow for degradation prediction incorporating features selected based on mRMR. Adapted with permission [65]. Copyright 2024, John Wiley and Sons. (C) High-throughput stability testing of 138 HTMs involved automated UVC degradation cycles and UV–vis characterization to monitor changes in absorbance over time. The decrease in the π–π^* band was used to calculate the photon dose at 90% initial absorbance (APD90) for material stability ranking. Adapted with permission [66]. Copyright 2025, American Chemical Society.

4.3 | Microstructure-Performance Relationships

Interface disorder significantly impacts the V_{OC} and bulk performance of high-efficiency OPVs, but its relationship to microstructure and device performance is not fully understood [69]. To address this, a novel hierarchical mRMR-embedded GPR (h-mRMR-GPR) model was introduced to establish comprehensive relationship links between processing parameters, microstructure, and device performance (Figure 5A) [70]. Using a high-throughput workflow, a large dataset was generated by systematically varying solid additive concentrations (c_{SA}) and annealing temperatures (T_a). The h-mRMR-GPR applied recursive feature selection to identify key nonredundant predictors of target performance metrics, such as V_{OC}, FF, and exciton efficiency. The algorithm first determined structure-property relationships by isolating predictors with high explanatory variance for each target property (Figure 5B). These first-generation predictors were then used as inputs in further mRMR-GPR runs to explore structure–structure relationships for distinguishing electronic from electrostatic interactions and linking processing conditions to microstructure. Finally, second-generation structural predictors were assessed for controllability through processing parameters, quantifying the effects of c_{SA} and T_a on interface and bulk disorder. The study revealed that interface disorder is the primary cause of voltage loss, outweighing the impact of driving force, while bulk disorder and dynamic disorder significantly influence FF and exciton dissociation efficiency. Optimizing the additive concentration (0.5:1) was shown to minimize interface disorder, reduce nonradiative losses and enhance device performance. This integrated data-driven approach not only elucidates hidden parameters influencing OPV performance but also provides a scalable framework for material optimization and autonomous high-throughput workflows in OPV development.

4.4 | Target-Driven Discovery for Novel Photovoltaics Materials

Despite advancements, the vast compositional space of HOIPs, alone with challenges such as lead toxicity and poor stability, necessitates the development of efficient discovery frameworks [13, 71–73]. To address, a target-driven high-throughput method combining ML with density functional theory (DFT) calculations was proposed to accelerate the discovery of HOIPs for photovoltaic applications [74]. A key aspect of this framework was the selection and ranking of material descriptors for effective ML models training. The framework generated a database of 230,808 HOIP candidates by varying combinations of 21 organic cations, 50 metallic cations, and 10 anions (Figure 6A). Initial screening reduced the pool to 38,086 candidates, with ML models, including Gradient Boosted Regression (GBR), Support Vector Regression (SVR), and Kernel Ridge Regression (KRR), identifying 686 candidates with suitable bandgaps for photovoltaics. Key features, such as atomic packing factor, tolerance factor, and octahedral factor, were ranked, with atomic packing factor emerging as the most critical predictor of bandgap. To ensure accuracy, 32 features were retained for modeling, even if some ranked lower. High-throughput DFT validation confirmed 132 stable, nontoxic HOIPs with ideal bandgaps

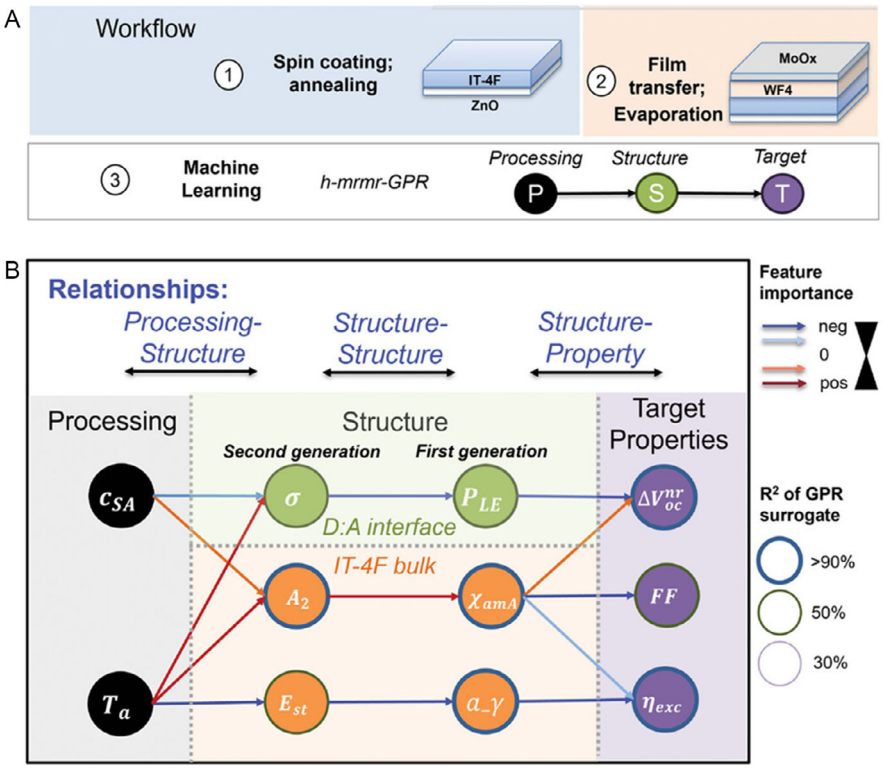

FIGURE 5 | (A) The workflow from device prepare to data analysis using hierarchical mRMR-embedded GPR model. (B) Knowledge graph illustrating the prediction of target properties from structural features and processing conditions. Adapted with permission [70]. Copyright 2024, John Wiley and Sons.

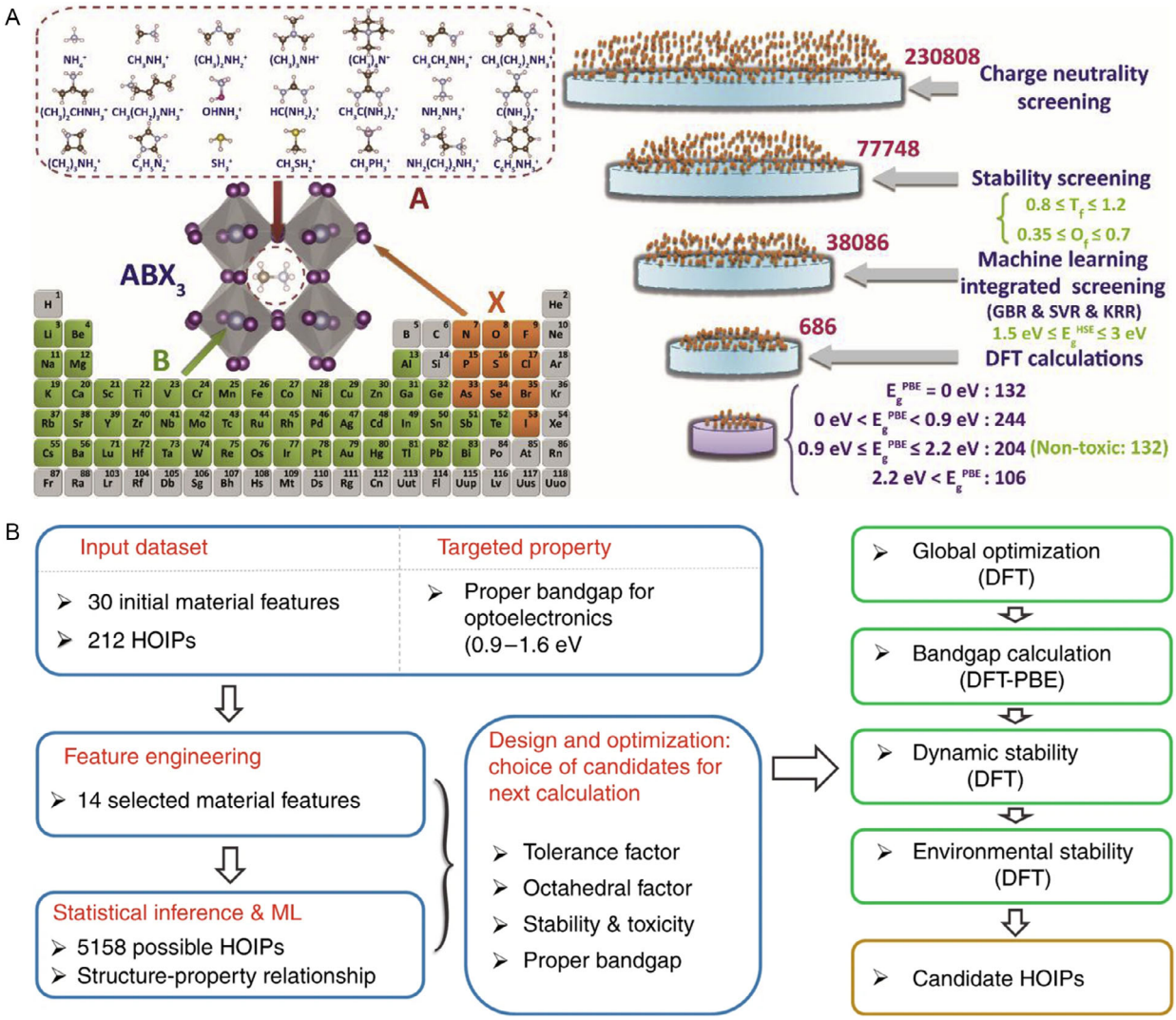

FIGURE 6 | (A) Schematic of a ML and DFT-based framework developed to screen 230 808 hypothetical candidates, using charge neutrality, stability conditions, and ML predictions, narrowing them down to 686 suitable HOIPs for solar cell applications. Further DFT verification identified 204 ideal HOIPs (including 132 nontoxic ones) with optimal bandgaps for photovoltaic use. Reproduced with permission [74]. Copyright 2019, Elsevier. (B) A lead-free HOIPs design framework integrates ML and DFT to efficiently identify stable materials with suitable bandgaps. Reproduced with permission [75]. Copyright 2018, Springer Nature.

(0.9–2.2 eV), including novel orthorhombic-like structures such as $ABSeI_2$ and $ABBrI_2$. By integrating feature selection technique, ML models, and high-throughput DFT validation, this framework accelerates material discovery, reduces computational costs, and offers a scalable approach for designing sustainable, efficient materials for photovoltaic and other advanced applications. Similarly, to address the inefficiencies of traditional trial-and-error methods, Wang's group developed a target-driven approach combining ML and DFT to screen materials for photovoltaic applications (Figure 6B) [75]. Using a stepwise elimination method, the feature selection process began with 30 initial features ranked by GBR improtance [76]. Iteratively removing the least important feature and evaluating model performance led to selection of 14 optimal features, as additional features had minimal impact on prediction accuracy. This method rapidly identified six lead-free orthorhombic HOIPs with suitable bandgaps and room-temperature stability from 5,158 candidates. A structure-property relationship for bandgap prediction was established. These efforts complement a broader framework that integrates ML and DFT to accelerate the discovery of stable, nontoxic HOIPs for photovoltaics.

4.5 | ML-Assisted Discovery of Low-Dimensional Perovskites

2D HOIPs have garnered considerable interest due to their exceptional photoelectronic properties [77–80]. However, designing HOIPs with desired bandgaps remains a significant challenge. To address this, a collaborative ML model was developed to efficiently screen materials with tailored bandgap characteristics (Figure 7A). This study employed a high-throughput screening platform capable of analyzing over one million virtual samples, alongside a three-step feature selection process for dimensionality reduction and rapid screening [81]. First, two constant features were removed, followed by correlation analysis using a cluster heat map to identify highly correlated variables. Features with Pearson's correlation coefficient ($r > 0.95$) were

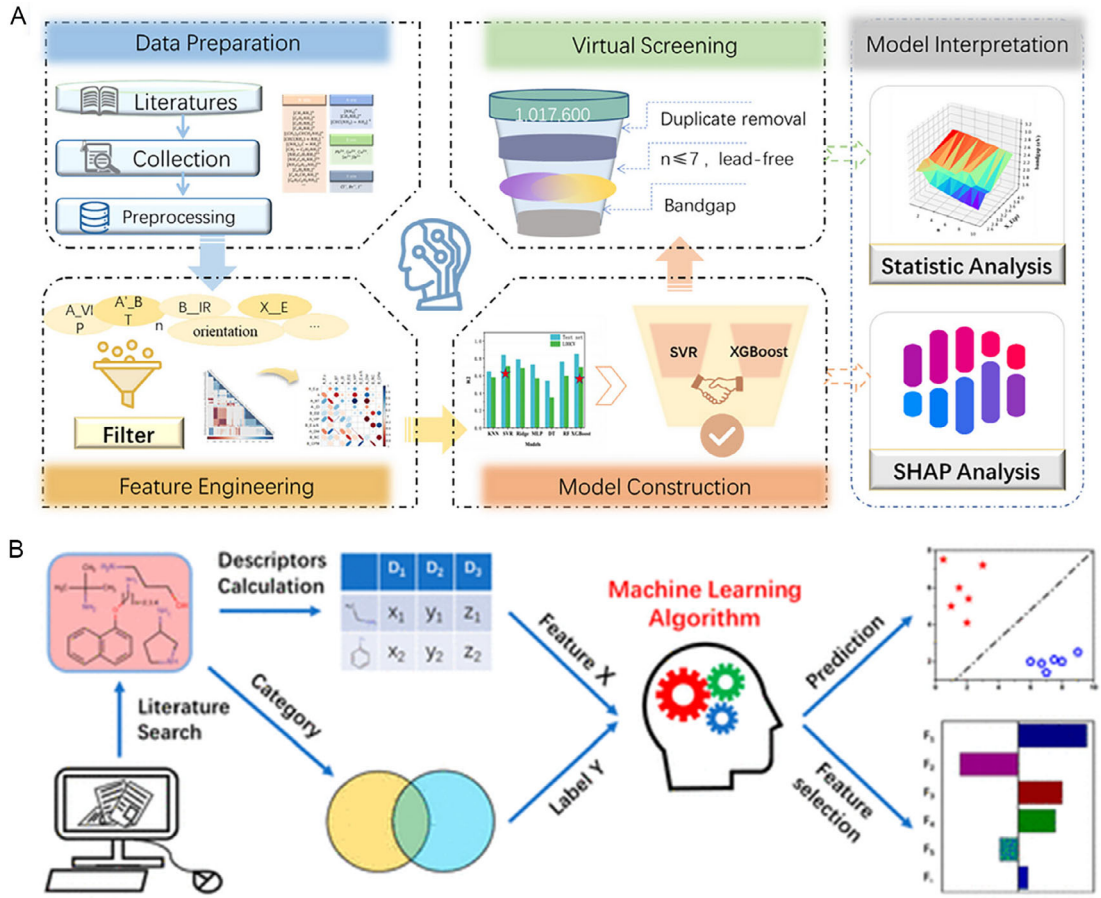

FIGURE 7 | (A) The ML-assisted flowchart for discovering 2D HOIPs with tailored bandgap. Reproduced with permission [81]. Copyright 2023, John Wiley and Sons. (B) Workflow of machine-learning-assisted exploration of low-dimensional perovskites. Reproduced with permission [82]. Copyright 2021, American Chemical Society.

filtered based on mRMR rankings to minimize redundancy and collinearity, retaining 29 features. Finally, embedded feature selection, customized for the predictive algorithms, was applied to refine the feature set. Support Vector Regression (SVR) and XGBoost models were used for bandgap prediction through forming a collaborative analysis. This approach led to the identification of 18 lead-free, stable materials with optimal bandgaps for high-efficiency solar cells, as well as 30 materials suitable for low-light applications. The framework accelerates material discovery, guides experimental design, and highlights the potential of combining ML with high-throughput platforms for functional material development.

Lyu et al. developed a supervised ML framework to predict the dimensionality of low-dimensional HOIPs (Figure 7B) [82]. The process began with initial screening to remove highly correlated features, followed by logistic regression with l1 regularization to select key descriptors. Four critical structural features were identified: steric effect index, eccentricity, largest ring size, and hydrogen-donor count. These features reflect steric, topological, and bonding properties crucial for 2D structure formation. SHAP analysis confirmed that the significance of these features, with refinements addressing molecular conformational changes due to intramolecular interactions. This approach achieved 82% prediction accuracy, providing a robust framework for accelerating the discovery and design of low-dimensional HOIPs.

5 | Challenges and Strategies

5.1 | Challenges in Feature Selection for e-PVs Research

Despite its advantages, feature selection in ML faces several challenges. The success of those method heavily depends on the acquisition of high-quality, diverse datasets. Insufficient or biased data often leads to overfitting, reducing model reliability and limiting the generalizability of predictions across new or complex material systems [35, 83]. Furthermore, many ML models operate as "black boxes," which produce accurate predictions that lack interpretability [84]. This hampers the extraction of meaningful scientific insights, hindering theory development and deeper understanding in e-PV research. Another critical limitation is the scalability of ML approaches. Although ML models can process large datasets, their computational demands, especially for advanced algorithms such as GPR and deep learning, make them resource-intensive and less accessible to smaller research teams [85, 86]. Additionally, transferability of ML models remains a challenge. Models trained on specific datasets frequently fail to generalize to new materials, environments, or varied experimental conditions, restricting their broader applicability [87]. Biases present in training data further exacerbates these issues, potentially narrowing the exploration space and leading to noninnovative solutions [83]. Moreover, reproducibility in ML-driven experiments can be undermined by

random errors in automated platforms or overlooked experimental variables, which creates challenges for consistent validation [87].

5.2 | Strategies for Overcoming Current Limitations

Addressing these limitations requires the integration of robust data protocols and hybrid modelling approaches. Combining physics-based insights with data-driven methods improves model interpretability and ensures that predictions remain grounded in scientific principles. Improving data quality begins with automated outlier detection, imputation of missing values, and domain-informed feature normalization to minimize noise and maintain consistency across heterogeneous datasets. Harmonizing features across experimental and computational domains further supports model generalization and transferability. For managing large and complex datasets, nonlinear dimensionality reduction techniques, such as autoencoders, Isomap and LLE, effectively compress feature spaces while preserving essential correlations. In parallel, the application of approximate or sparse GPR variants and model distillation strategies helps to mitigate computational overhead without compromising predictive performance.

In this context, the h-mRMR-GPR framework provides a robust and scalable solution. This framework combines information-theoretic filtering (e.g., mRMR) with nonlinear kernel-based modeling utilizing GPR equipped with ARD kernels. Such an integration effectively minimizes feature redundancy and identifies features with high nonlinear relevance. The hierarchical architecture allows for recursive feature selection and systematic feature grouping, which enables a structured resolution of complex process–structure–property relationships within e-PV systems. The use of ARD kernels within GPR introduces intrinsic, model-based measures of feature importance, thereby enhancing model interpretability and supporting the extraction of physically meaningful insights. The framework implements pretraining feature reduction to lower input dimensionality before model construction, which significantly improves computational scalability and allows GPR to function efficiently in otherwise intractable domains. The hierarchical structure aligns naturally with the multi-modal and multi-scale nature of datasets encountered in e-PV research, which facilitates cross-domain knowledge transfer for applications such as long-term stability prediction, microstructure–performance mapping, and accelerated material discovery. In addition, the framework incorporates uncertainty quantification through either the confidence intervals inherent to GPR or ensemble-based variance estimation methods. This capability strengthens the framework's effectiveness in guiding experimental prioritization under resource-limited conditions. Collectively, the h-mRMR-GPR framework establishes a scalable, interpretable, and data-efficient paradigm for ML–assisted discovery and optimization in emerging photovoltaic research.

6 | Conclusion

Feature selection is a key part of ML-driven research in e-PVs. This review summarizes its role in the management of complex datasets, the improvement of model accuracy, and the enhancement of interpretability. The review also discusses how the combination of feature selection methods with ML-driven high-throughput experimentation accelerate the discovery and optimization of new photovoltaic materials. These holistic methods help establish causal relationships among material properties, process parameters, and device performance. They support rational molecular design, scalable manufacturing, and data-driven inverse design of functional solar materials. Nonlinear and hybrid feature selection methods further improve the scalability, generalization, and transferability of ML models. This progress makes it possible to develop more efficient and stable photovoltaic technologies. Future work will need hybrid modeling frameworks that combine physics-based understanding with data-driven methods. It will also require standardized and high-quality data protocols. These improvements will make material discovery faster, reduce experimental costs, and help identify key design principles for further innovation in the field of optoelectronic materials.

Acknowledgments

The authors gratefully acknowledge the grants "ELF-PVDesign and development of solution processed functional materials for the next generations of PV technologies" (No. 44-6521a/20/4) by the Bavarian State Government, the SolMAP, and SolarTAP–a Technology Acceleration Platform for emerging Photovoltaics project by Helmholtz Association. C.J.B. gratefully acknowledges financial support through the "Aufbruch Bayern" initiative of the state of Bavaria (EnCN and "Solar Factory of the Future") and the German Research Foundation (DFG) SFB953-No.182849149, and GRK2495 (ITRG2495).

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

The data that support the findings of this study are available from the corresponding author upon reasonable request.

References

1. G. Zhang, F. R. Lin, F. Qi, et al., "Renewed Prospects for Organic Photovoltaics," *Chemical Reviews* 122, no. 18 (2022): 14180.

2. M. A. Green, A. Ho-Baillie, and H. J. Snaith, "The Emergence of Perovskite Solar Cells," *Nature Photonics* 8, no. 7 (2014): 506.

3. N. Li, X. Niu, Q. Chen, and H. Zhou, "Towards Commercialization: The Operational Stability of Perovskite Solar Cells," *Chemical Society Reviews* 49, no. 22 (2020): 8235.

4. L. J. Sutherland, H. C. Weerasinghe, and G. P. Simon, "A Review on Emerging Barrier Materials and Encapsulation Strategies for Flexible Perovskite and Organic Photovoltaics," *Advanced Energy Materials* 11, no. 34 (2021): 2101383.

5. S. Park, T. Kim, S. Yoon, C. W. Koh, H. Y. Woo, and H. J. Son, "Progress in Materials, Solution Processes, and Long-Term Stability for Large-Area Organic Photovoltaics," *Advanced Materials* 32, no. 51 (2020): 2002217.

6. J. Zhang, J. Wu, A. Barabash, et al., "Precise Control of Process Parameters for >23% Efficiency Perovskite Solar Cells in Ambient Air Using an Automated Device Acceleration Platform," *Energy & Environmental Science* 17, no. 15 (2024): 5490.

7. D. P. Tabor, L. M. Roch, S. K. Saikin, et al., "Accelerating the Discovery of Materials for Clean Energy in the Era of Smart Automation," *Nature Reviews Materials* 3, no. 5 (2018): 5.

8. J. Zhang, J. Wu, O. Stroyuk, et al., "Self-Driving AMADAP Laboratory: Accelerating the Discovery and Optimization of Emerging Perovskite Photovoltaics," *MRS Bulletin* 49, no. 12 (2024): 1284.

9. J. Yang, B. J. Lawrie, S. V. Kalinin, and M. Ahmadi, "High-Throughput Automated Exploration of Phase Growth Behaviors in Quasi-2D Formamidinium Metal Halide Perovskites," *Advanced Energy Materials* 13, no. 43 (2023): 2302337.

10. B. P. MacLeod, F. G. L. Parlane, T. D. Morrissey, et al., "Self--Driving Laboratory for Accelerated Discovery of Thin-Film Materials," *Science Advances* 6, no. 20 (2020): eaaz8867.

11. J. Zhang, B. Liu, Z. Liu, et al., "Optimizing Perovskite Thin-Film Parameter Spaces with Machine Learning-Guided Robotic Platform for High-Performance Perovskite Solar Cells," *Advanced Energy Materials* 13, no. 48 (2023): 2302594.

12. D. O. Baumann, F. Laufer, J. Roger, R. Singh, M. Gholipoor, and U. W. Paetzold, "Repeatable Perovskite Solar Cells through Fully Automated Spin-Coating and Quenching," *ACS Applied Materials & Interfaces* 16, no. 40 (2024): 54007.

13. Y. Zhao, T. Heumueller, J. Zhang, et al., "A Bilayer Conducting Polymer Structure for Planar Perovskite Solar Cells with over 1,400 h Operational Stability at Elevated Temperatures," *Nature Energy* 7, no. 2 (2022): 144.

14. J. Wagner, C. G. Berger, X. Du, T. Stubhan, J. A. Hauch, and C. J. Brabec, "The Evolution of Materials Acceleration Platforms: Toward the Laboratory of the Future with AMANDA," *Journal of Materials Science* 56, no. 29 (2021): 16422.

15. A. K. Y. Low, F. Mekki-Berrada, A. Gupta, et al., "Evolution-Guided Bayesian Optimization for Constrained Multi-Objective Optimization in Self-Driving Labs," *npj Computational Materials* 10, no. 1 (2024): 104.

16. J. Zhang, V. M. Le Corre, J. Wu, et al., "Autonomous Optimization of Air-Processed Perovskite Solar Cell in a Multidimensional Parameter Space," *Advanced Energy Materials* 15, no. 19 (2025): 2404957, https://doi.org/10.1002/aenm.202404957.

17. Y. Zhao, J. Zhang, Z. Xu, et al., "Discovery of Temperature-Induced Stability Reversal in Perovskites Using High-Throughput Robotic Learning," *Nature Communications* 12, no. 1 (2021): 2191.

18. T. Osterrieder, F. Schmitt, L. Lüer, et al., "Autonomous Optimization of an Organic Solar Cell in a 4-Dimensional Parameter Space," *Energy & Environmental Science* 16, no. 9 (2023): 3984.

19. M. Seifrid, R. Pollice, A. Aguilar-Granda, et al., "Autonomous Chemical Experiments: Challenges and Perspectives on Establishing a Self-Driving Lab," *Accounts of Chemical Research* 55, no. 17 (2022): 2454.

20. N. J. Szymanski, B. Rendy, Y. Fei, et al., "An Autonomous Laboratory for the Accelerated Synthesis of Novel Materials," *Nature* 624, no. 7990 (2023): 86.

21. J. Zhang, J. A. Hauch, and C. J. Brabec, "Toward Self-Driven Autonomous Material and Device Acceleration Platforms (AMADAP) for Emerging Photovoltaics Technologies," *Accounts of Chemical Research* 57, no. 9 (2024): 1434.

22. M. M. Flores-Leonar, L. M. Mejía-Mendoza, A. Aguilar-Granda, et al., "Materials Acceleration Platforms: On the Way to Autonomous Experimentation," *Current Opinion in Green and Sustainable Chemistry* 25 (2020): 100370.

23. K. M. Jablonka, D. Ongari, S. M. Moosavi, and B. Smit, "Big-Data Science in Porous Materials: Materials Genomics and Machine Learning," *Chemical Reviews* 120, no. 16 (2020): 8066.

24. R. Batra, L. Song, and R. Ramprasad, "Emerging Materials Intelligence Ecosystems Propelled by Machine Learning," *Nature Reviews Materials* 6, no. 8 (2021): 655.

25. L. Lüer, I. M. Peters, A. S. Smith, et al., "A Digital Twin to Overcome Long-Time Challenges in Photovoltaics," *Joule* 8, no. 2 (2024): 295.

26. O. Isayev, C. Oses, C. Toher, E. Gossett, S. Curtarolo, and A. Tropsha, "Universal Fragment Descriptors for Predicting Properties of Inorganic Crystals," *Nature Communications* 8, no. 1 (2017): 15679.

27. S. Wang, Y. Huang, W. Hu, and L. Zhang, "Data-Driven Optimization and Machine Learning Analysis of Compatible Molecules for Halide Perovskite Material," *npj Computational Materials* 10, no. 1 (2024): 114.

28. I. Guyon and A. Elisseeff, "An Introduction to Variable and Feature Selection," *Journal of Machine Learning Research* 3 (March 2003): 1157.

29. C. Chen, Y. Zuo, W. Ye, X. Li, Z. Deng, and S. P. Ong, "A Critical Review of Machine Learning of Energy Materials," *Advanced Energy Materials* 10, no. 8 (2020): 1903242.

30. J. Zhang, J. Wu, V. M. Le Corre, J. A. Hauch, Y. Zhao, and C. J. Brabec, "Advancing Perovskite Photovoltaic Technology through Machine Learning-Driven Automation," *InfoMat* 7, no. 5 (2025): e70005, https://doi.org/10.1002/inf2.70005.

31. B. P. MacLeod, F. G. L. Parlane, A. K. Brown, J. E. Hein, and C. P. Berlinguette, "Flexible Automation Accelerates Materials Discovery," *Nature Materials* 21, no. 7 (2022): 722.

32. M. Abolhasani and E. Kumacheva, "The Rise of Self-fDriving Labs in Chemical and Materials Sciences," *Nature Synthesis* 2, no. 6 (2023): 483.

33. J. P. Janet and H. J. Kulik, "Resolving Transition Metal Chemical Space: Feature Selection for Machine Learning and Structure-Property Relationships," *Journal of Physical Chemistry A* 121, no. 46 (2017): 8939.

34. Y. Saeys, I. Inza, and P. Larrañaga, "A Review of Feature Selection Techniques in Bioinformatics," *Bioinformatics* 23, no. 19 (2007): 2507.

35. U. M. Khaire and R. Dhanalakshmi, "Stability of Feature Selection Algorithm: A Review," *Journal of King Saud University, Computer and Information Sciences* 34, no. 4 (2022): 1060.

36. G. Chandrashekar and F. Sahin, "A Survey on Feature Selection Methods," *Computers & Electrical Engineering* 40, no. 1 (2014): 16.

37. H. Peng, F. Long, and C. Ding, "Feature Selection Based on Mutual Information Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy," *IEEE Transactions on Pattern Analysis and Machine Intelligence* 27, no. 8 (2005): 1226.

38. M. Cherrington, F. Thabtah, J. Lu and Q. Xu, "Feature Selection: Filter Methods Performance Challenges," in 2019 International Conference on Computer and Information Sciences (ICCIS), (IEEE, Sakaka, Saudi Arabia, 2019): 1.

39. R. Kohavi and G. H. John, "Wrappers for Feature Subset Selection," *Artificial Intelligence* 97, no. 1-2 (1997): 273.

40. P. Dhal and C. Azad, "A Comprehensive Survey on Feature Selection in the Various Fields of Machine Learning," *Applied Intelligence* 52, no. 4 (2022): 4543.

41. A. J. Chowdhury, W. Yang, E. Walker, O. Mamun, A. Heyden, and G. A. Terejanu, "Prediction of Adsorption Energies for Chemical Species on Metal Catalyst Surfaces Using Machine Learning," *Journal of Physical Chemistry C* 122, no. 49 (2018): 28142.

42. F. Saidi, S. Khetari, I. S. Yahia, H. Y. Zahran, T. Hidouri, and N. Ameur, "The use of Principal Component Analysis (PCA) and Partial Least Square (PLS) for Designing New Hard Inverse Perovskites Materials," *Computational Condensed Matter* 31 (2022): e00667.

43. S. Cacovich, G. Divitini, C. Ireland, F. Matteocci, A. Di Carlo, and C. Ducati, "Elemental Mapping of Perovskite Solar Cells by Using Multivariate Analysis: An Insight into Degradation Processes," *ChemSusChem* 9, no. 18 (2016): 2673.

44. G. Xia, B. Huang, Y. Zhang, et al., "Nanoscale Insights into Photovoltaic Hysteresis in Triple-Cation Mixed-Halide Perovskite:

Resolving the Role of Polarization and Ionic Migration," *Advanced Materials* 31, no. 36 (2019): 1902870.

45. P. Xanthopoulos, P. M. Pardalos, and T. B. Trafalis, "Linear discriminant analysis," in Robust Data Mining chap. 4. SpringerBriefs in Optimization.(Springer, NY: 2013): 27–33.

46. S. T. Roweis and L. K. Saul, "Nonlinear Dimensionality Reduction by Locally Linear Embedding," *Science* 290, no. 5500 (2000): 2323.

47. M. Yamada, W. Jitkrittum, L. Sigal, E. P. Xing, and M. Sugiyama, "High-Dimensional Feature Selection by Feature-Wise Kernelized Lasso," *Neural Computation* 26, no. 1 (2014): 185.

48. J. Chen, M. Stern, M. J. Wainwright, and M. I. Jordan, "Kernel Feature Selection via Conditional Covariance Minimization," *Advances in Neural Information Processing Systems* 30 (2017): 6946.

49. E. Adeli, G. Wu, B. Saghafi, L. An, F. Shi, and D. Shen, "Kernel-Based Joint Feature Selection and Max-Margin Classification for Early Diagnosis of Parkinson's Disease," *Scientific Reports* 7, no. 1 (2017): 41069.

50. K. Liu, Y. Li, X. Hu, M. Lucu, and W. D. Widanage, "Gaussian Process Regression with Automatic Relevance Determination Kernel for Calendar Aging Prediction of Lithium-Ion Batteries," *IEEE Transactions on Industrial Informatics* 16, no. 6 (2019): 3767.

51. F. Anowar, S. Sadaoui, and B. Selim, "Conceptual and Empirical Comparison of Dimensionality Reduction Algorithms (PCA, KPCA, LDA, MDS, SVD, LLE, ISOMAP, LE, ICA, t-SNE)," *Computer Science Review* 40 (2021): 100378.

52. A. Alsahaf, N. Petkov, V. Shenoy, and G. Azzopardi, "A Framework for Feature Selection through Boosting," *Expert Systems with Applications* 187 (2022): 115895.

53. S. Ramírez-Gallego, I. Lastra, D. Martínez-Rego, et al., "Fast Minimum Redundancy Maximum Relevance Algorithm for High-Dimensional Big Data," *International Journal of Intelligent Systems* 32, no. 2 (2017): 134.

54. N. T. P. Hartono, M. Ani Najeeb, Z. Li, et al., "Principled Exploration of Bipyridine and Terpyridine Additives to Promote Methylammonium Lead Iodide Perovskite Crystallization," *Crystal Growth & Design* 22, no. 9 (2022): 5424.

55. T. Lu, M. Li, W. Lu, and Zhang, T.-Y. Recent Progress in the Data-Driven Discovery of Novel Photovoltaic Materials," *Journal of Materials Informatics* 2, no. 2 (2022): 7.

56. L. Lüer, R. Wang, C. Liu, et al., "Maximizing Performance and Stability of Organic Solar Cells at Low Driving Force for Charge Separation," *Advanced Science* 11, no. 6 (2024): 2305948.

57. N. T. P. Hartono, J. Thapa, A. Tiihonen, et al., "How Machine Learning Can Help Select Capping Layers to Suppress Perovskite Degradation," *Nature Communications* 11, no. 1 (2020): 1.

58. Z. Zhao, R. Anand and M. Wang, "Maximum Relevance and Minimum Redundancy Feature Selection Methods for a Marketing Machine Learning Platform," in 2019 IEEE International Conference on Data Science and Advanced Analytics (DSAA), (IEEE, Washington, DC, 2019): 442.

59. D. Padula, J. D. Simpson, and A. Troisi, "Combining Electronic and Structural Features in Machine Learning Models to Predict Organic Solar Cells Properties," *Materials Horizons* 6, no. 2 (2019): 343.

60. C. Liu, L. Lüer, V. M. L. Corre, et al., "Understanding Causalities in Organic Photovoltaics Device Degradation in a Machine-Learning-Driven High-Throughput Platform," *Advanced Materials* 36, no. 20 (2023): 2300259.

61. M. Wagner, A. Distler, V. M. Le Corre, et al., "Cutting, Lab-to-Fab, Short: High Throughput Optimization and Process Assessment in Roll-to-Roll Slot Die Coating of Printed Photovoltaics," *Energy & Environmental Science* 16, no. 11 (2023): 5454.

62. D. Bash, Y. Cai, V. Chellappan, et al., "Multi-Fidelity High-Throughput Optimization of Electrical Conductivity in P3HT-CNT Composites," *Advanced Functional Materials* 31, no. 36 (2021): 2102606.

63. Y. Li, X. Huang, K. Ding, et al., "Non-Fullerene Acceptor Organic Photovoltaics with Intrinsic Operational Lifetimes over 30 years," *Nature Communications* 12, no. 1 (2021): 5419.

64. J. Luke, E. M. Speller, A. Wadsworth, et al., "Twist and Degrade—impact of Molecular Structure on the Photostability of Nonfullerene Acceptors and Their Photovoltaic Blends," *Advanced Energy Materials* 9, no. 15 (2019): 1803755.

65. X. Du, L. Lüer, T. Heumueller, et al., "Revealing Processing Stability Landscape of Organic Solar Cells with Automated Research Platforms and Machine Learning," *InfoMat* 6, no. 7 (2024): e12554.

66. A. J. Bornschlegl, P. Duchstein, J. Wu, et al., "An Automated Workflow to Discover the Structure-Stability Relations for Radiation Hard Molecular Semiconductors," *Journal of the American Chemical Society* 147 (2025): 1957–1967, https://doi.org/10.1021/jacs.4c14824.

67. R. Verduci, V. Romano, G. Brunetti, et al., "Energy in Space Applications: Review and Technology Perspectives," *Advanced Energy Materials* 12, no. 29 (2022): 2200125.

68. Y. Tu, J. Wu, G. Xu, et al., "Perovskite Solar Cells for Space Applications: Progress and Challenges," *Advanced Materials* 33, no. 21 (2021): 2006545.

69. H. Zhang, Y. Li, X. Zhang, Y. Zhang, and H. Zhou, "Role of Interface Properties in Organic Solar Cells: From Substrate Engineering to Bulk-Heterojunction Interfacial Morphology," *Materials Chemistry Frontiers* 4, no. 10 (2020): 2863.

70. R. Wang, L. Han, N. Li, et al., "Reducing Voltage Losses in Organic Photovoltaics Requires Interfacial Disorder Management," *Advanced Energy Materials* 14, no. 26 (2024): 2400609.

71. W. Ke and M. G. Kanatzidis, "Prospects for Low-Toxicity Lead-Free Perovskite Solar Cells," *Nature Communications* 10, no. 1 (2019): 965.

72. J. Luo, B. Liu, H. Yin, et al., "Polymer-Acid-Metal Quasi-Ohmic Contact for Stable Perovskite Solar Cells beyond a 20,000-Hour Extrapolated Lifetime," *Nature Communications* 15, no. 1 (2024): 2002.

73. J. Qin, Z. Che, Y. Kang, et al., "Towards Operation-Stabilizing Perovskite Solar Cells: Fundamental Materials, Device Designs, and Commercial Applications," *InfoMat* 6, no. 4 (2024): e12522.

74. T. Wu and J. Wang, "Global Discovery of Stable and Non-Toxic Hybrid Organic-Inorganic Perovskites for Photovoltaic Systems by Combining Machine Learning Method with First Principle Calculations," *Nano Energy* 66 (2019): 104070.

75. S. Lu, Q. Zhou, Y. Ouyang, Y. Guo, Q. Li, and J. Wang, "Accelerated Discovery of Stable Lead-Free Hybrid Organic-Inorganic Perovskites via Machine Learning," *Nature Communications* 9, no. 1 (2018): 3405.

76. J. H. Friedman, "Greedy Function Approximation: A Gradient Boosting Machine," *Annals of Statistics* 29 (2001): 1189.

77. D. H. Cao, C. C. Stoumpos, O. K. Farha, J. T. Hupp, and M. G. Kanatzidis, "2D Homologous Perovskites as Light-Absorbing Materials for Solar Cell Applications," *Journal of the American Chemical Society* 137, no. 24 (2015): 7843.

78. J. Zhang, J. Wu, S. Langner, et al., "Exploring the Steric Hindrance of Alkylammonium Cations in the Structural Reconfiguration of Quasi-2D Perovskite Materials Using a High-Throughput Experimental Platform," *Advanced Functional Materials* 32, no. 43 (2022): 2207101.

79. J. Liang, Z. Zhang, Q. Xue, et al., "C.-C. A Finely Regulated Quantum Well Structure in Quasi-2D Ruddlesden-Popper Perovskite Solar Cells with Efficiency Exceeding 20%," *Energy & Environmental Science* 15, no. 1 (2022): 296.

80. J. Zhang, S. Langner, J. Wu, et al., "Intercalating-Organic-Cation-Induced Stability Bowing in Quasi-2D Metal-Halide Perovskites," *ACS Energy Letter* 7, no. 1 (2022): 70.

81. Y. Shen, J. Wang, X. Ji, and W. Lu, "Machine Learning-Assisted Discovery of 2D Perovskites with Tailored Bandgap for Solar Cells," *Advanced Theory and Simulations* 6, no. 6 (2023): 2200922.

82. R. Lyu, C. E. Moore, T. Liu, Y. Yu, and Y. Wu, "Predictive Design Model for Low-Dimensional Organic-Inorganic Halide Perovskites Assisted by Machine Learning," *Journal of the American Chemical Society* 143, no. 32 (2021): 12766.

83. J. Schrier, A. J. Norquist, T. Buonassisi, and J. Brgoch, "In Pursuit of the Exceptional: Research Directions for Machine Learning in Chemical and Materials Science," *Journal of the American Chemical Society* 145, no. 40 (2023): 21699.

84. K. Terayama, M. Sumita, R. Tamura, and K. Tsuda, "Black-Box Optimization for Automated Discovery," *Accounts of Chemical Research* 54, no. 6 (2021): 1334.

85. H. Wang, T. Fu, Y. Du, et al., "Scientific Discovery in the Age of Artificial Intelligence," *Nature* 620 (2023): 47.

86. Z. Ren, Z. Ren, Z. Zhang, T. Buonassisi, and J. Li, "Autonomous Experiments Using Active Learning and AI," *Nature Reviews Materials* 8, no. 9 (2023): 563.

87. M. Krenn, R. Pollice, S. Y. Guo, et al., "On Scientific Understanding with Artificial Intelligence," *Nature Reviews Physics* 4, no. 12 (2022): 761.