Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 3 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

Machine Learning-Assisted Discovery of Advanced Materials for Next-Generation Photovoltaic Devices

Azmat Ali
Department of physics, University of chakwal
azmat.chakwal@gmail.com

ABSTRACT

The discovery of next-generation photovoltaic materials is essential for advancing sustainable energy technologies, yet conventional experimental and computational approaches face significant limitations in navigating the vast chemical and configurational spaces of multi-component systems. This review provides a comprehensive synthesis of machine learning-assisted materials discovery for emerging photovoltaic technologies, including metal halide perovskites, lead-free double perovskites, two-dimensional photovoltaics, and organic solar cells. We critically evaluate feature engineering methodologies, comparing composition-based vectorization, structural fingerprints, and graph-based atomic representations, alongside classical machine learning algorithms and advanced deep learning architectures such as graph neural networks, angle-aware equivariant networks, and hybrid frameworks. The review examines high-throughput screening pipelines that have successfully reduced candidate spaces from hundreds of thousands to experimentally viable compounds, identifies key predictive descriptors governing photovoltaic performance, and explores generative artificial intelligence models enabling inverse design of novel materials with target optoelectronic properties. We assess open data infrastructures, particularly the Perovskite Database Project, and the integration of large language models for automated literature mining, which collectively establish FAIR data ecosystems essential for data-driven discovery. Critical challenges addressed include publication bias, data degeneracy, quantum mechanical fidelity limitations, and structural discrepancies in training datasets. Finally, we discuss autonomous self-driving laboratories that close the experimental loop through Bayesian optimization and active learning, compressing discovery timelines from decades to weeks. This review establishes a framework for accelerating photovoltaic materials discovery through the synergistic integration of machine learning, high-throughput computation, and automated experimentation.

Keywords— Machine learning, photovoltaic materials, perovskites, graph neural networks, high-throughput screening, inverse design, self-driving laboratories, bandgap prediction, materials informatics, FAIR data

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 3 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

1. Introduction

The global transition toward sustainable energy infrastructure depends heavily on the rapid development and deployment of next-generation photovoltaic (PV) devices. While conventional single-junction silicon solar cells are approaching their theoretical Shockley-Queisser power conversion efficiency limit (~33.7%), emerging photoactive material classes offer opportunities to exceed these efficiency thresholds while drastically lowering manufacturing costs (Hassan et al., 2024). Among these, metal halide perovskites (ABX_3), lead-free double perovskites ($A_2BB'X_6$), two-dimensional photovoltaic (2DPV) materials, and organic solar cells (OSCs) have demonstrated unprecedented efficiency trajectories and optical tenability (Munir et al., 2023).

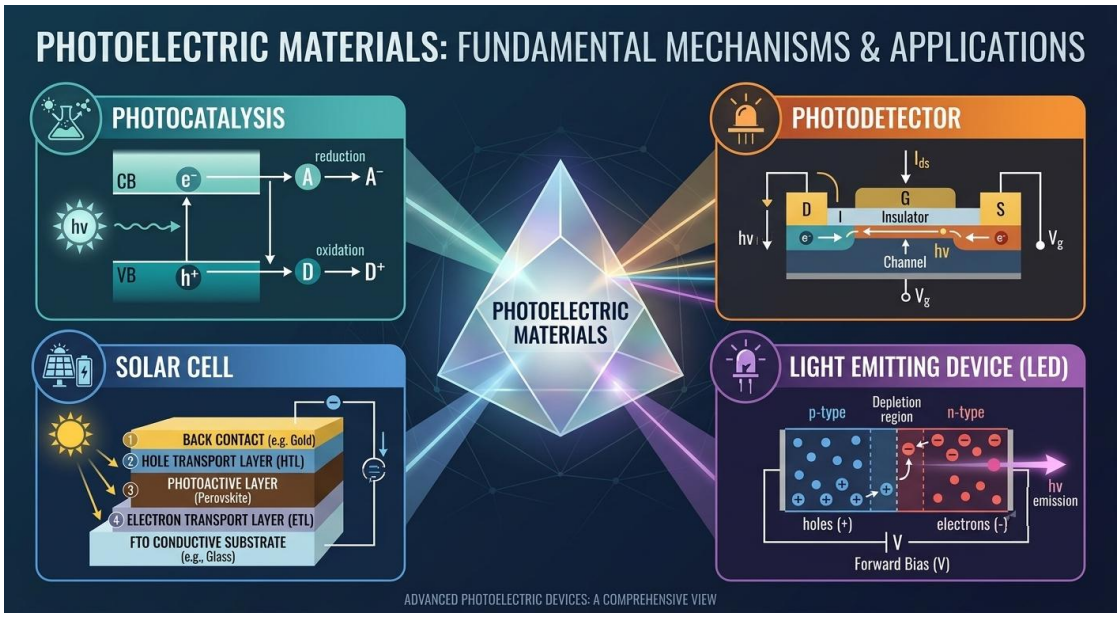
Despite these rapid efficiency gains, the vast chemical and configurational space of multi-component materials presents a formidable bottleneck to conventional experimental discovery. Traditional trial-and-error synthesis is inherently constrained by high labor costs, lengthy fabrication cycles, and difficulties in systematically mapping complex processing-microstructure-property relationships (Li et al., 2025). Concurrently, first-principles computational methods, such as Density Functional Theory (DFT), encounter severe scaling bottlenecks ($O(N^3)$ computational complexity) when deployed across broad compositional spaces involving tens of thousands of candidate compounds (Mouvet et al., 2022).

Machine learning (ML) and artificial intelligence (AI) have emerged as transformative tools capable of circumventing these computational and experimental limitations. By leveraging historical computational databases and high-throughput experimental datasets, ML models capture non-linear structure-property relationships, enabling real-time property prediction, targeted candidate screening, and autonomous inverse material design (Mohamed, 2025). This review delivers a comprehensive synthesis of ML-assisted materials discovery in next-generation photovoltaics. It critically evaluates feature representation frameworks, graph neural network architectures, generative design models, open data infrastructures, and the integration of automated self-driving laboratories (Miao & Sun, 2025).

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 3 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

Figure 1: Fundamental Optoelectronic Mechanisms and Device Architectures Across Key Photovoltaic and Photoelectric Applications.



2. Machine Learning Algorithms and Computational Frameworks

2.1 Feature Engineering and Material Representations

The predictive fidelity of any machine learning model relies fundamentally on the mathematical representation of the input material, known as featurization or vectorization. In photovoltaic research, material inputs are broadly categorized into composition-based feature vectors (CBFVs), structural fingerprints, and graph-based atomic representations (Abolade, 2023).

Compositional representations utilize elemental physical properties such as ionic radii, electronegativity, valence electron counts, and atomic mass aggregated via weighted statistical operators including minimum, maximum, mean, and range. Feature libraries such as Magpie, Oliynyk, and mat2vec convert stoichiometric formulas into fixed-length numerical vectors. While CBFVs allow rapid screening of vast uncharacterized chemical spaces without requiring prior structural relaxation, they fundamentally lack the spatial resolution necessary to differentiate structural polymorphs or capture complex octahedral tilting angles in halide perovskites (Barua & Kleinke, 2025).

To capture three-dimensional spatial arrangements, graph-based representations treat crystal structures as mathematical graphs $G = (V, E)$, where nodes V represent atoms characterized by local atomic properties, and edges E correspond to interatomic bonds or interatomic distance vectors. Universal atomic embeddings, such as CrystalTransformer UAEs (ct-UAEs), dynamically learn chemical and spatial environments, yielding substantial improvements in accuracy when transferred across graph neural networks (Ullah et al., 2026).

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 3 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

2.2 Comparative Model Architectures and Performance

Machine learning algorithms deployed for photovoltaic property prediction range from classical supervised algorithms to deep, angle-aware graph neural networks. Classical machine learning (CML) algorithms such as Random Forest (RF), eXtreme Gradient Boosting (XGBoost), and Support Vector Regression (SVR) excel when trained on tabular, descriptor-engineered datasets. For example, XGBoost has established superior predictive metrics for determining the dimensionality of hybrid organic-inorganic perovskites (HOIPs) and forecasting solar cell power conversion efficiency (PCE) from process descriptors (Howlader, 2025).

Graph Neural Networks (GNNs), including the Crystal Graph Convolutional Neural Network (CGCNN), MatErials Graph Network (MEGNet), Atomistic Line Graph Neural Network (ALIGNN), and Global Attention Graph Neural Network (GATGNN), operate directly on atomic coordinates. While early architectures like CGCNN and MEGNet achieved sub-0.05 eV/atom accuracy for formation energies (E_f), recent angle-aware and equivariant architectures (such as ALIGNN, GemNet-OC, and E2GNN) have pushed predictive errors down to the ~10 meV scale (Wang et al., 2025).

To address complex crystal structures, hybrid architectures have been constructed. The RFNET framework integrates graph-extracted spatial embeddings with Multi-Layer Perceptron (MLP) structured feature vectors, achieving superior bandgap (E_g) predictive accuracy compared to standard GNN models. Furthermore, structural descriptors such as cell volume and B-X bond lengths, generated via Kolmogorov-Arnold Networks (KAN), have been integrated into SVR models to enhance bandgap prediction in lead-free perovskite systems (Xiao et al. 2023).

Table 1: Performance comparison of computational and machine learning models for photovoltaic material property prediction.

Model Architecture	Input Representation Type	Target Property	Prediction Error / Metric (MAE/RMSE/ R2)	Primary Structural / Physical Insight
CGCNN	Crystal Graph (Atoms & Bonds)	Bandgap (E_g)	MAE: 0.560 eV, R^2: 0.677	Learns intrinsic structural representations directly from periodic atomic coordinates.
MEGNet	Crystal Graph with Global State Vector	Bandgap (E_g)	MAE: 0.734 eV, R^2: 0.532	Incorporates macroscopic physical conditions including temperature, pressure, and entropy.

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 3 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

RFNET	Hybrid (GNN Graph Embeddings + MLP Vectors)	Bandgap (E_g)	MAE: 0.486 eV, R^2: 0.758	Reduces candidate screening pools by 8–14% compared to standard baseline models.
ALIGNN / GemNet-OC	Angle-Aware Equivariant Graph	Formation Energy (E_f)	MAE : ~0.010 eV/atom	Resolves three-dimensional bond-angle configurations and interatomic forces with high fidelity.
XGBoost Ensemble	Molecular / Processing Descriptors	Perovskite Dimensionality / PCE	Accuracy: 88% (Dimensionality)	Identifies critical topological index drivers such as MaxAbsEStateIndex.
ATL-CGCNN	Adaptive Transfer Learning Graph	HSE06 Bandgap (E_g)	Verified by DFT validation	Successfully transfers knowledge from large PBE datasets to high-fidelity HSE06 targets.

2.3 Advanced Learning Paradigms: Transfer Learning and AutoML

A central bottleneck in training deep graph models for energy materials is data scarcity, specifically regarding high-fidelity bandgap values computed via expensive hybrid functional DFT (e.g., HSE06) or GW approximations. Standard Generalized Gradient Approximation (GGA-PBE) calculations exhibit systematic derivative discontinuities (Δ_{xc}), causing bandgap underestimations of 40–50% (Xinyue & Wang, 2026).

To overcome this without incurring extreme computational costs, Adaptive Transfer Learning (ATL) protocols pre-train graph neural networks on massive low-fidelity DFT datasets (such as the Materials Project) before transferring and fine-tuning weights on smaller, high-fidelity datasets. For instance, the ATL-CGCNN architecture utilizes adaptive fine-tuning on high-confidence samples to accurately predict the HSE06 bandgaps of unreported hybrid organic-inorganic perovskites (Buterez et al., 2024).

Additionally, Automated Machine Learning (AutoML) frameworks including AutoGluon, TPOT, and H2O.ai automate model selection, hyperparameter tuning, and feature extraction. These pipelines accelerate informatics workflows and establish direct links to automated robotic laboratories (LeDell et al., 2020).

3. High-Throughput Screening and Property Prediction Across PV Material Classes

3.1 Metal Halide Perovskites and Lead-Free Alternatives

Metal halide perovskites (ABX_3) possess exceptional optoelectronic attributes,

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 3 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

including tunable optical bandgaps, high absorption coefficients, defect tolerance, and long charge-carrier diffusion lengths. However, commercial adoption is hindered by thermal instability, moisture sensitivity, and the toxicity of lead (Pb^{2+}). Consequently, machine learning focus has shifted toward screening stable, non-toxic lead-free alternatives, such as double perovskites ($A_2BB'X_6$) (Soltani et al., 2025).

High-throughput ML workflows combining goal-driven screening with DFT verification have demonstrated significant efficiency in filtering vast chemical spaces. In a representative multi-condition screening campaign of 16,400 candidate double perovskites, ML models evaluating photoelectric properties and thermal stability reduced the pool to 10 promising lead-free candidates for experimental synthesis (Kumar, 2026).

Feature importance analyses across perovskite screening datasets consistently reveal a clear physical hierarchy: atomic descriptors at the B/B'-site cations dictate bandgap energy levels (E_g) through d-p orbital hybridization, whereas the X-site halide anions govern thermodynamic stability, moisture sensitivity, and overall device efficiency (PCE) (Gholamrezaei et al., 2026).

3.2 Two-Dimensional Photovoltaics (2DPV) and Layered Structures

Two-dimensional (2D) materials offer structural robustness, suppressed ion migration, and quantum confinement effects ideal for stable photoactive layers. However, screening 2D systems is complicated by non-covalent van der Waals interlayer spaces and complex organic spacer configurations (Ullah et al., 2025).

High-throughput ML protocols combined with DFT calculations have successfully identified novel 2DPV candidates from massive crystal databases. Screening 187,093 experimentally identified inorganic crystal structures yielded 26 high-efficiency 2DPV candidates, highlighting compounds such as Sb_2Se_2Te, Sb_2Te_3, and Bi_2Se_3 as exhibiting conversion efficiencies superior to conventional 2D monolayers. Model feature evaluations established the packing factor (P_f) as the single most critical structural descriptor governing photovoltaic performance in 2D inorganic networks (Jin et al., 2026).

For 2D hybrid organic-inorganic perovskites (2D HOIPs), standard GNNs often fail because they treat the inorganic octahedral layers and organic spacer molecules as an isotropic 3D network. To address this structural anisotropy, the Site-Attention-Based Graph Convolutional Neural Network (SATGNN) separates convolution operations between the organic A-site cation layers and the inorganic BX_6 octahedral layers. Incorporating a site-attention layer enables the model to resolve how organic spacer steric hindrance and inorganic octahedral distortion independently modulate the bandgap (Dudakov et al., 2026).

3.3 Organic Solar Cells (OSCs)

In organic photovoltaics, efficiency depends on the complex molecular interaction, energy level alignment, and bulk heterojunction (BHJ) morphology between electron donors and acceptors. Machine learning models applied to OSCs focus on forecasting key performance metrics including open-circuit voltage (V_{oc}), short-circuit current density (J_{sc}), highest occupied molecular orbital (HOMO), lowest unoccupied molecular orbital (LUMO) energy levels, and power conversion efficiency (PCE)

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 3 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

(Shafian et al., 2026).

By training supervised models (e.g., Random Forest, XGBoost, and Neural Networks) on molecular descriptors derived from quantum chemical simulations (DFT/TD-DFT), researchers can rapidly evaluate millions of donor-acceptor chemical combinations. Feature selection routines demonstrate that frontier orbital energy offsets ($\Delta E_{\text{HOMO-LUMO}}$) and optical absorption spectra profiles are primary drivers of predicted device outputs (Jiang et al., 2021).

Figure 2: Four-Pillar Machine Learning Discovery Workflow for Organic Solar Cells Integrating Database Curation, Model Training, Combinatorial Screening, and Experimental Feedback.

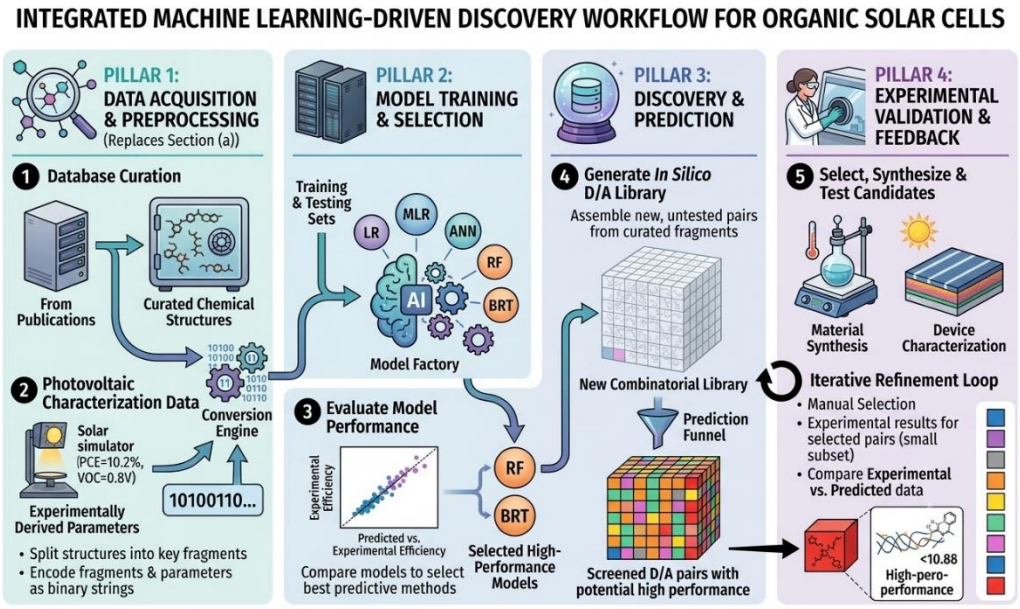


Table 2: Representative high-throughput machine learning discovery pipelines for emerging photovoltaic materials.

Photovoltaic Material System	Initial Candidate Space Size	Computational & Machine Learning Pipeline	Final Validated High-Performance Candidates	Key Predictive Physical Descriptors
Lead-Free Double Perovskites ($A_2BB'X_6$)	16,400 compounds	ML multi-condition property models + DFT thermal/bandgap filtering	10 stable lead-free double perovskites	B-site atomic radius, valence state, and electronegativity.
Inorganic 2D	187,093	Classification	26	Packing factor

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 3 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

Photovoltaics (2DPV)	inorganic crystal structures	ML + High-throughput DFT efficiency calculation	candidates (Sb_2Se_2Te , Sb_2Te_3 , Bi_2Se_3)	(P_f), coordination geometry.
Hybrid Organic-Inorganic Perovskites (2D HOIPs)	Synthetic molecular space	Dimensionality XGBoost + Layer-specific SATGNN	7 DFT-validated optimal bandgap HOIPs	Steric hindrance index, octahedral distortion angle.
Organic Bulk Heterojunctions (OSCs)	Combinatorial donor-acceptor space	Molecular descriptor ML + TD-DFT quantum chemistry	Highly efficient polymer donor/non-fullerene acceptor pairs	$\Delta E_{\text{HOMO-LUMO}}$, optical absorption spectra overlap.

4. Generative AI Models and Inverse Design Protocols

4.1 De Novo Molecular and Crystal Generation

While high-throughput screening relies on filtering pre-existing crystal structures or combinatorial libraries, inverse design utilizes generative AI models to construct novel chemical structures optimized for target photovoltaic metrics directly from latent spaces (Orry et al., 2006).

Generative architectures deployed in materials science include Variational Autoencoders (VAEs), Adversarial Autoencoders (AAEs), Objective-Reinforced Generative Adversarial Networks (ORGAN), Character-level Recurrent Neural Networks (CharRNN), REINVENT, and GraphINVENT. In de novo polymer donor design for organic photovoltaics, these generative models explore chemical space by modifying molecular graph strings (SMILES/SELFIES) or 3D atomic coordinates, continuously guided by property-predictive reward functions. This approach allows researchers to generate previously unconsidered chemical structures that satisfy specific optoelectronic constraints (Menon & Ranganathan, 2022).

4.2 Multi-Objective Bayesian Optimization

Photovoltaic performance requires optimizing multiple conflicting physical properties simultaneously: a material must maximize power conversion efficiency (PCE) and optical absorption while maintaining low thermodynamic formation energy (E_f) and high operational stability (Cetinkaya et al., 2025).

Bayesian Optimization (BO) coupled with multi-objective Pareto selection algorithms guides inverse design workflows toward optimal efficiency-stability trade-offs. By constructing probabilistic surrogate models (typically Gaussian Processes or ensemble neural networks), BO quantifies predictive uncertainty through acquisition functions such as Expected Improvement (EI) or Upper Confidence Bound (UCB). This active learning loop selectively identifies candidate materials that either maximize target performance or minimize structural uncertainty, drastically lowering the total number of expensive DFT or experimental validations required (Jung et al., 2023).

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 3 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

5. Open Data Infrastructures, Process-Property Informatics, and Literature Mining

5.1 The Perovskite Database Infrastructure

Data-driven materials discovery relies heavily on open-access, standardized computational and experimental repositories. A notable advancement in photovoltaic informatics is the Perovskite Database Project (perovskitedatabase.com), an open-access repository adhering strictly to FAIR (Findable, Accessible, Interoperability, and Reusable) data principles (Himanen et al., 2019).

The database aggregates manually and automatically extracted records for over 42,400 individual solar cell devices collected from more than 15,000 peer-reviewed publications, covering over 400 distinct parameters per device. To maintain structural consistency across multi-layer devices, the schema enforces a substrate-first, left-to-right stacking convention that meticulously catalogues every functional layer. This includes the substrate, electron transport layer (ETL), photoactive perovskite absorber, hole transport layer (HTL), and back contact (Jacobsson et al., 2022). Precise formatting delimiters govern data entry: vertical bars (|) partition distinct physical layers, semicolons (;) delimit co-deposited chemical reagents, and double forward brackets (>>) indicate multi-step sequential reactions. In addition to complete stack sequences and synthesis parameters, the database records critical photovoltaic outputs (V_{oc} , J_{sc} , FF , PCE), quantum efficiency profiles, and operational stability measurements such as T_{80} (the time required for a device to lose 20% of its initial power conversion efficiency) (Martelli et al., 2017).

5.2 Descriptor Importance and Processing Parameter Analytics

Machine learning models trained on the Perovskite Database provide insight into the processing parameters that dictate solar cell performance. Using non-linear statistical association metrics such as Mutual Information (MI) and SHAP (SHapley Additive exPlanations), researchers can rank feature importance directly from empirical device records (Su et al., 2026).

MI analyses reveal that the choice of X-site halide anion exerts the strongest statistical influence over overall photovoltaic output (PCE, V_{oc}, J_{sc}, FF), surpassing the influence of A-site organic cations. Film thickness, optical bandgap, and B-site cation choices represent the next most significant parameter group (Jena et al., 2019).

Regarding fabrication conditions, ML analyses uncover strong non-linear relationships between ambient humidity during perovskite crystallization and ultimate device performance. Specifically, relative humidity levels maintained strictly between 30% and 40% yield peak PCE values, whereas deviations above or below this window trigger accelerated defect formation or incomplete precursor conversion (Hernandez-Torres et al., 2024).

5.3 Automated Literature Mining via Large Language Models

To prevent data infrastructures from becoming obsolete as literature expands, natural language processing (NLP) and Large Language Models (LLMs) are used to automate data extraction. Fine-tuning models such as LLaMA on the Perovskite Database schema enables automated parsing of unstructured text, tables, and experimental sections in

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 4, Issue 3 (2026)

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

newly published research (Han et al., 2024).
Fine-tuned LLMs achieve high extractions scores (e.g., F1 scores exceeding 87%) in mapping complex synthesis protocols, precursor concentrations, and performance metrics into structured FAIR database entries. This automated knowledge capture bridges the gap between manual experimental synthesis and dynamic, real-time machine learning training pipelines (Lee et al., 2026).

6. Challenges, Bottlenecks, and Future Perspectives

6.1 Data Degeneracy and Publication Bias

A persistent challenge in training machine learning models on empirical photovoltaic data is publication bias the selective reporting of exceptional device efficiencies and the systemic suppression of unsuccessful experiments ("negative data") (Alhazmi et al., 2025).
This bias creates severe data degeneracy and imbalanced distributions in public databases, causing predictive models to overestimate the feasibility of unvalidated synthesis recipes. Overcoming this bottleneck requires the institutional adoption of standardized laboratory data management systems (LIMS) that record failed synthetic routes, alongside reporting negative results directly to FAIR data infrastructures (Branco et al., 2016).

6.2 Quantum Mechanical Fidelity and Structural Discrepancies

The accuracy of high-throughput machine learning pipelines remains limited by the quality of its training data. Most large crystal repositories rely on semi-local DFT functionals (PBE) that ignore long-range van der Waals forces and miscalculate electronic bandgaps due to derivative discontinuities (Δ_{xc}) (Sampaio et al., 2024).
Deploying multi-fidelity learning frameworks which bridge fast semi-local PBE calculations, high-level HSE06/GW calculations, and empirical experimental measurement data is essential for improving predictive accuracy across non-standard chemical spaces (Mannodi-Kanakkithodi et al., 2026).

6.3 Autonomous Self-Driving Laboratories

The future of machine learning in photovoltaics lies in closing the loop between computational prediction and experimental validation. Autonomous self-driving laboratories (SDLs) merge active learning software with automated robotics, liquid handlers, thin-film deposition setups, and inline characterization suites (Yao et al., 2023).
In a fully integrated autonomous discovery loop, Bayesian optimization algorithms or generative models first propose high-confidence chemical formulations and precise processing parameters. Automated liquid-handling workstations then prepare precursor solutions, deposit films via robotic spin-coating or inkjet printing, and thermally anneal the photoactive layers. Inline testing modules immediately record absorbance spectra, J-V curves, and operational decay rates (Stanton et al., 2022). These experimental results are streamed directly back into the surrogate machine learning model, dynamically updating the probability distribution and refining subsequent search directions without human intervention. By eliminating manual experimental

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

bottlenecks and suppressing human variability, autonomous self-driving laboratories hold the potential to compress material discovery timelines from decades down to a matter of weeks (Desimpel et al., 2026).

Conclusion

The integration of machine learning into photovoltaic materials discovery has fundamentally transformed the paradigm of materials research, enabling rapid navigation of vast chemical spaces that were previously inaccessible through conventional experimental or computational approaches alone. This review has systematically demonstrated that ML-assisted methodologies, spanning classical algorithms, graph neural networks, and generative models, have achieved remarkable predictive accuracy in forecasting critical optoelectronic properties including bandgaps, formation energies, and power conversion efficiencies, often surpassing traditional DFT calculations in throughput while maintaining competitive fidelity. The successful application of high-throughput screening pipelines, such as the identification of lead-free double perovskites from 16,400 candidates and 26 novel 2DPV materials from over 187,000 crystal structures, validates the efficacy of data-driven strategies in accelerating discovery cycles. However, significant challenges persist that must be addressed to realize the full potential of ML in photovoltaics research. Publication bias and the systemic exclusion of negative experimental results create imbalanced datasets that lead models to overestimate synthesis feasibility and device performance. The reliance on semi-local DFT functionals introduces systematic bandgap underestimations of 40–50%, necessitating multi-fidelity learning frameworks that bridge computational approximations with experimental measurements. Furthermore, the translation of ML predictions into experimentally realized devices remains limited by the gap between computational descriptors and real-world processing conditions. The emergence of autonomous self-driving laboratories represents the most promising pathway forward, closing the loop between computational prediction, robotic synthesis, and inline characterization. By combining Bayesian optimization with automated experimentation, these platforms can dynamically refine search strategies without human intervention, compressing discovery timelines from decades to weeks. Future research must prioritize the institutional adoption of standardized data management systems that capture both successful and failed synthesis attempts, the development of physically informed neural architectures that respect crystallographic symmetries, and the integration of multi-modal data streams encompassing synthesis parameters, processing conditions, and device performance metrics. Ultimately, the continued convergence of machine learning, high-throughput computation, and automated experimentation will establish a self-accelerating discovery ecosystem capable of sustainably powering the global transition toward renewable energy technologies.

References

Abolade, Y. A. (2023). Bridging Mathematical Foundations and Intelligent Systems: A Statistical and Machine Learning Approach. *Communication In Physical Sciences*, *9*(4), 774-784.

Alhazmi, A., Kholoud, M., & Eke, C. I. (2025). A Systematic Review of Advances in

Online ISSN 3007-3197
Print ISSN 3007-3189

Deep Learning Architectures for Efficient and Sustainable Photovoltaic Solar Tracking: Research Challenges and Future Directions. *Sustainability*, *17*(21), 9625.

Barua, N. K., & Kleinke, H. (2025). Machine learning predictions of thermopower for thermoelectric material screening. *ACS Applied Energy Materials*, *8*(21), 16110-16121.

Branco, P., Torgo, L., & Ribeiro, R. P. (2016). A survey of predictive modeling on imbalanced domains. *ACM computing surveys (CSUR)*, *49*(2), 1-50.

Buterez

Cetinkaya, C., Cokduygulular, E., Aykut, M. Y., Erkal, O., Aydogmus, F., & Kinaci, B. (2025). Artificial intelligence-empowered functional design of semi-transparent optoelectronic and photonic devices via deep Q-learning. *Scientific Reports*, *15*(1), 13508.

Desimpel, S., Dorbec, M., Van Geem, K. M., & Stevens, C. V. (2026). Bayesian optimization for chemical reactions. *Chemical Society Reviews*, *55*(5), 2731-2775.

Dudakov, I. V., Rubtsov, I., Marchenko, E. I., Eliseev, A. A., Kvashnin, D. G., Tarasov, A. B., ... & Korolev, V. V. (2026). CoRE 2D-HOIP DB: Computation-ready, experimental database of two-dimensional hybrid organic–inorganic perovskites. *Journal of Chemical Information and Modeling*.

Gholamrezaei Sarvelat, S., Amoopour, E., & Abdolahzadeh Ziabari, A. (2026). Machine Learning-Based Optimization of Perovskite Solar Cell Architecture Using Real-World Database and Physical-Driven Features. *Advanced Theory and Simulations*, *9*(3), e01547.

Han, S., Wang, M., Zhang, J., Li, D., & Duan, J. (2024). A review of large language models: Fundamental architectures, key technological evolutions, interdisciplinary technologies integration, optimization and compression techniques, applications, and challenges. *Electronics*, *13*(24), 5040.

Hassan, Q., Viktor, P., Al-Musawi, T. J., Ali, B. M., Algburi, S., Alzoubi, H. M., ... & Jaszczur, M. (2024). The renewable energy role in the global energy Transformations. *Renewable Energy Focus*, *48*, 100545.

Hernandez-Torres, J. A., Torreglosa, J. P., Sanchez-Herrera, R., Bischi, A., & Baccioli, A. (2024). Development of an optimized non-linear model for precise dew point estimation in variable environmental conditions. *Applied Sciences*, *14*(22), 10508.

Himanen, L., Geurts, A., Foster, A. S., & Rinke, P. (2019). Data-driven materials science: status, challenges, and perspectives. *Advanced Science*, *6*(21), 1900808.

Howlader, A. S. (2025). Machine learning applications in renewable energy: Predictive analytics for solar cell performance optimization and energy yield forecasting. *American Journal of Scholarly Research and Innovation*, *4*(01), 392-427.

Jacobsson, T. J., Hultqvist, A., García-Fernández, A., Anand, A., Al-Ashouri, A., Hagfeldt, A., ... & Unger, E. (2022). An open-access database and analysis tool for perovskite solar cells based on the FAIR data principles. *Nature Energy*, *7*(1), 107-115.

Online ISSN 3007-3197
Print ISSN 3007-3189

Jena, A. K., Kulkarni, A., & Miyasaka, T. (2019). Halide perovskite photovoltaics: background, status, and future prospects. *Chemical reviews*, *119*(5), 3036-3103.

Jiang, D., Wu, Z., Hsieh, C. Y., Chen, G., Liao, B., Wang, Z., ... & Hou, T. (2021). Could graph neural networks learn better molecular representation for drug discovery? A comparison study of descriptor-based and graph-based models. *Journal of cheminformatics*, *13*(1), 12.

Jin, H., Zhang, H., Li, J., Wang, T., Wan, L., Guo, H., & Wei, Y. (2020). Discovery of novel two-dimensional photovoltaic materials accelerated by machine learning. *The journal of physical chemistry letters*, *11*(8), 3075-3081.

Jung, J., Park, K., Cho, B., Park, J., & Ryu, S. (2023). Optimization of injection molding process using multi-objective bayesian optimization and constrained generative inverse design networks. *Journal of Intelligent Manufacturing*, *34*(8), 3623-3636.

Kumar Yadav, A. (2026). Artificial intelligence in accelerating materials discovery: opportunities and challenges. *ChemistrySelect*, *11*(7), e07191.

LeDell, E., & Poirier, S. (2020, July). H2o automl: Scalable automatic machine learning. In *Proceedings of the AutoML Workshop at ICML* (Vol. 2020, p. 24)., D., Janet, J. P., Kiddle, S. J., Oglic, D., & Lió, P. (2024). Transfer learning with graph neural networks for improved molecular property prediction in the multi-fidelity setting. *Nature communications*, *15*(1), 1517.

Lee, J., Noh, J., Lee, H., Lee, Y. L., Im, J., & Seo, J. (2026). Recent advances in data-driven and artificial intelligence-integrated perovskite solar cells: From design to self-driving laboratories. *InfoMat*, *8*(4), e70124.

Li, J., Feng, J., Luo, J., Jiang, B., Zheng, X., Song, Q., ... & Ma, Y. (2025). Self-optimizing machine learning potential assisted automated workflow for highly efficient complex systems material design. *arXiv preprint arXiv:2505.08159*.

Mannodi-Kanakkithodi, A., Huang, M., Gorai, P., & Kavanagh, S. R. (2026). Accelerating point-defect simulations using data-driven and machine learning approaches.

Martelli, A., Ravenscroft, A., & Holden, S. (2017). *Python in a Nutshell: A desktop quick reference*. " O'Reilly Media, Inc.".

Menon, D., & Ranganathan, R. (2022). A generative approach to materials discovery, design, and optimization. *ACS omega*, *7*(30), 25958-25973.

Miao, N., & Sun, Z. (2025). Machine learning-enabled optoelectronic material discovery: a comprehensive review. *Journal of Materials Informatics*, *5*(3), N-A.

Mohamed, N. (2025). Artificial intelligence and machine learning in cybersecurity: a deep dive into state-of-the-art techniques and future paradigms. *Knowledge and Information Systems*, *67*(8), 6969-7055.

Mouvet, F., Villard, J., Bolnykh, V., & Rothlisberger, U. (2022). Recent advances in first-principles based molecular dynamics. *Accounts of Chemical Research*, *55*(3), 221-230.

Munir, J., Jamil, M., Jbara, A. S., Ullah, H., Murtaza, H., Ali, H. E., & Ain, Q. (2023). Electronic structure and optical and thermoelectric response of lead-free double perovskite BaMgLaBiO6: a first-principles study. *Journal of Computational*

Online ISSN 3007-3197
Print ISSN 3007-3189

Electronics, *22*(5), 1482-1494.

Orry, A. J., Abagyan, R. A., & Cavasotto, C. N. (2006). Structure-based development of target-specific compound libraries. *Drug discovery today*, *11*(5-6), 261-266.

Sampaio da Silva, C., Boos, J. A., Goldowsky, J., Blache, M., Schmid, N., Heinemann, T., ... & Revol, V. (2024). High-throughput platform for label-free sorting of 3D spheroids using deep learning. *Frontiers in bioengineering and biotechnology*, *12*, 1432737.

Shafian, S., Mohd Salehin, F. N., Lee, S., Ismail, A., Mohamed Shuhidan, S., Xie, L., & Kim, K. (2025). Development of organic semiconductor materials for organic solar cells via the integration of computational quantum chemistry and AI-powered machine learning. *ACS Applied Energy Materials*, *8*(2), 699-722.

Soltani, S., Hjiri, M., Ahmed, N. I. A., Jbeli, A., Aldukhayel, A. M., & Althumairi, N. A. (2025). Metal halide perovskites for energy applications: recent advances, challenges, and future perspectives. *RSC advances*, *15*(27), 21811.

Stanton, S., Maddox, W., Gruver, N., Maffettone, P., Delaney, E., Greenside, P., & Wilson, A. G. (2022, June). Accelerating Bayesian optimization for biological sequence design with denoising autoencoders. In *International conference on machine learning* (pp. 20459-20478). PMLR.

Su, Y., Wang, K., Guan, X., Wu, Y., Zhang, H., Xie, F., & Chu, J. (2026). Machine Learning Descriptors for Mapping Structure-Property-Performance Relationships of Perovskite Solar Cells. *Advanced Energy Materials*, *16*(7), e05294.

Ullah, K., Hussain, A., Munsif, M., Sohn, K. S., & Baik, S. W. (2026). Differential attention transformer-enhanced graph neural network for accurate material property prediction. *Engineering Applications of Artificial Intelligence*, *176*, 114749.

Ullah, K., Ul Haq, A., Golovynskyi, S., Hidouri, T., Qu, J., & Golovynska, I. (2025). Perovskite Nanocrystals, Quantum Dots, and Two-Dimensional Structures: Synthesis, Optoelectronics, Quantum Technologies, and Biomedical Imaging. *Nanomaterials*, *16*(1), 30.

Wang, Z., Li, W., Wang, S., & Wang, X. (2025). The future of catalysis: applying graph neural networks for intelligent catalyst design. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, *15*(2), e70010.

Xiao, H., Li, R., Shi, X., Chen, Y., Zhu, L., Chen, X., & Wang, L. (2023). An invertible, invariant crystal representation for inverse design of solid-state materials using generative deep learning. *Nature Communications*, *14*(1), 7027.

Xinyue, Z., & Wang, S. (2026). Graph Representation Learning for Battery Energy Systems in Few-Shot Scenarios: Methods, Challenges and Outlook. *Batteries*, *12*(1), 11.

Yao, Z., Lum, Y., Johnston, A., Mejia-Mendoza, L. M., Zhou, X., Wen, Y., ... & Seh, Z. W. (2023). Machine learning for a sustainable energy future. *Nature Reviews Materials*, *8*(3), 202-215.