Scholars Journal of Engineering and Technology

OPEN ACCESS

Abbreviated Key Title: Sch J Eng Tech
ISSN 2347-9523 (Print) | ISSN 2321-435X (Online)
Journal homepage: https://saspublishers.com

Artificial Intelligence and Machine Learning in the Design of Nanomaterials for Next-Generation Solar Cells

Mohammad Arsalan Aslam[1], Muhammad Rafi Ud Din Farhan[2], Ihsan Ullah[3], Syed Muhammad Abu Bakar Shah[4]*, Aqsa Nisar[5]

[1]Department of Energy for Circular Economy, The Open University of Sri Lanka
[2]Institute of Data Science, University of Engineering and Technology (UET), Lahore, Pakistan
[3]Department of Electrical Engineering, University of Engineering and Technology (UET), Peshawar, Pakistan
[4]Institute of Physics, The Islamia University of Bahawalpur, Pakistan
[5]Department of Chemistry, University of Agriculture, Faisalabad, Pakistan

DOI: https://doi.org/10.36347/sjet.2026.v14i07.002 | **Received:** 09.05.2026 | **Accepted:** 20.06.2026 | **Published:** 09.07.2026

***Corresponding author:** Syed Muhammad Abu Bakar Shah
Institute of Physics, The Islamia University of Bahawalpur, Pakistan

Abstract | **Original Research Article**

The development of third-generation photovoltaics perovskite solar cells (PSCs), organic solar cells (OSCs), and quantum dot solar cells (QDSCs) is constrained by the impracticality of exhaustively exploring multidimensional chemical and processing parameter spaces using conventional trial-and-error experimentation. The discovery of stable, non-toxic, and high-efficiency nanomaterials requires navigation of combinatorial spaces spanning elemental compositions, molecular architectures, synthesis routes, and processing conditions that far exceed experimental throughput capabilities. Artificial intelligence (AI) and machine learning (ML) offer a paradigm shift by enabling data-driven prediction, optimization, and generation of materials prior to laboratory synthesis. Through training on curated experimental datasets, high-throughput density functional theory (DFT) calculations, and multi-scale simulations, ML models can approximate complex structure–property relationships, identify optimal synthesis windows, and guide experimental efforts with quantifiable uncertainty. This review provides a systematic evaluation of ML methodologies applied to photovoltaic nanomaterial design, covering data acquisition and representation, supervised learning, deep learning, generative models, active learning, and self-driving laboratories. Key challenges including data scarcity, interpretability, domain transfer, and reproducibility are critically assessed.

Keywords: Perovskite solar cells, Machine learning, Photovoltaics, Materials discovery, High-throughput screening, Power conversion efficiency.

1. INTRODUCTION

1.1 Motivation: The Photovoltaics Landscape

Decarbonization of global energy systems requires scalable, cost-effective, and efficient photovoltaic technologies. While first-generation crystalline silicon (c-Si) and second-generation thin-film chalcogenides dominate present commercial markets, single-junction devices are fundamentally bounded by the Shockley Queisser radiative efficiency limit of approximately 33.7% for terrestrial illumination (Shockley & Queisser, 1961). Third-generation photovoltaics encompassing perovskites, organic semiconductors, and quantum dots aim to circumvent these limitations through nanostructured active layers, solution-processable fabrication, and bandgap tunability. Perovskite solar cells have achieved certified power conversion efficiencies (PCEs) exceeding 26% in single-junction configurations, rivaling established technologies after only a decade of development (de la Asunción-Nadal *et al*., 2025; Jacobsson *et al*., 2022). Organic solar cells have surpassed 19% PCE through advances in non-fullerene acceptors (NFAs) (Li *et al*., 2021; Lopez *et al*., 2017), while quantum dot solar cells offer unique advantages in infrared spectral harvesting and multi-exciton generation (Rainò *et al*., 2018).

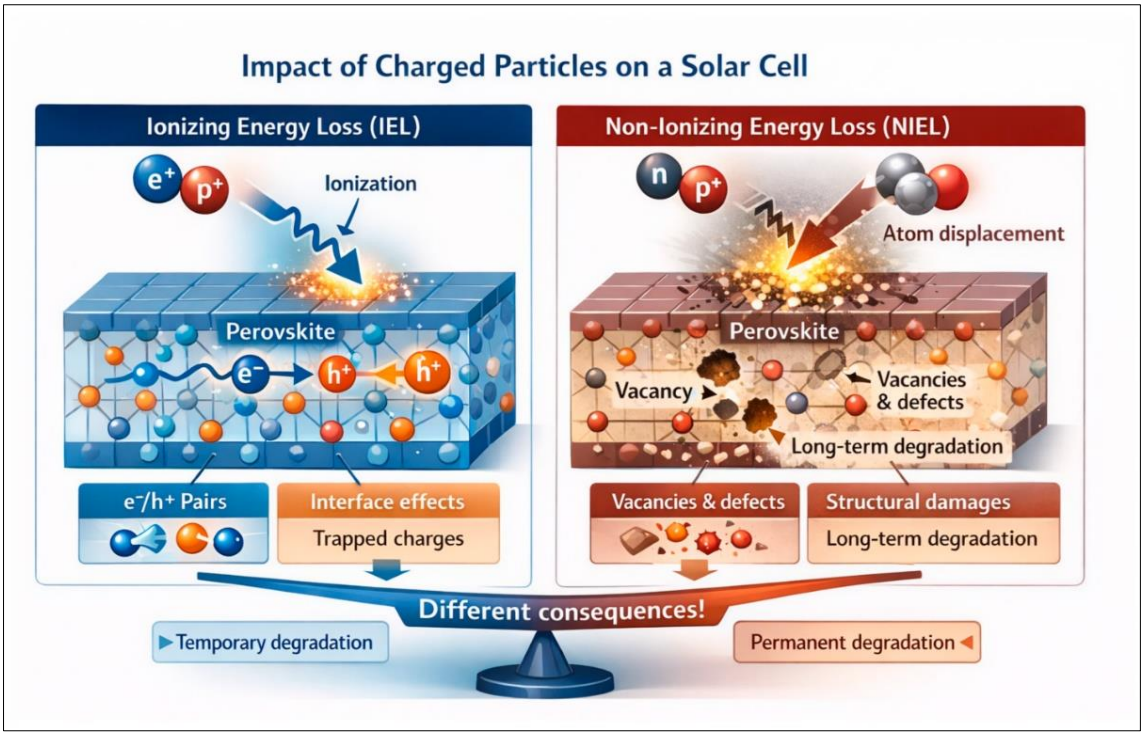

Fig. 1: Impact of charged particles on perovskite solar cell degradation. Ionizing energy loss (IEL) from electrons (e^-) and protons (p^+) generates electron-hole pairs and trapped charges at interfaces, causing temporary degradation. Non-ionizing energy loss (NIEL) from neutrons (n^+) and protons (p^+) induces atomic displacement, creating vacancies and defects that lead to permanent structural damage and long-term degradation

1.2 The Experimental Bottleneck

Despite this progress, deployment is impeded by persistent challenges: perovskite instability under moisture, thermal stress, and illumination; OSC sensitivity to nanoscale bulk heterojunction (BHJ) morphology; and QDSC limitations from surface trap states and charge transport barriers (Butler *et al*., 2018; Morgan & Jacobs, 2020). Even with coarse sampling of three compositional parameters (ten values each) in a mixed-cation mixed-halide perovskite system, the space exceeds 1,000 unique compositions. Including synthesis variables expands the space to millions of possible configurations far beyond the throughput of any single research group. Traditional trial-and-error methods are ill-suited to such high-dimensional optimization, and publication bias toward positive results creates an incomplete empirical record that distorts subsequent hypothesis generation (Morgan & Jacobs, 2020; Schmidt *et al*., 2019).

1.3 The AI/ML Alternative

Machine learning offers a fundamentally different approach (Butler *et al*., 2018; Schmidt et al., 2019). Rather than exhaustively exploring parameter space, ML algorithms learn empirical mappings from material representations to target properties. Once trained, surrogate models can evaluate thousands of hypothetical candidates per second, prioritize experiments likely to yield improved performance, and quantify prediction uncertainty to guide exploration (Shahriari *et al*., 2016; Snoek *et al*., 2012). ML does not replace physical understanding but complements it: ML models identify non-linear correlations that escape human intuition (Morgan & Jacobs, 2020), while physical constraints embedded into ML architectures improve generalization and reduce data requirements (Bartók *et al*., 2013; Himanen *et al*., 2019).

2. Theoretical Foundations and Problem Formalization

2.1 Photovoltaic Device Physics: Key Figures of Merit

The power conversion efficiency (PCE) is the ratio of electrical power output to incident optical power: $\eta = Voc \cdot Jsc \cdot FF / Pin$, where Voc is open-circuit voltage, Jsc is short-circuit current density, FF is fill factor, and Pin is incident power density (100 mW/cm² under AM1.5G illumination) (Shockley & Queisser, 1961). In the radiative limit, Voc is bounded by the bandgap Eg, with non-radiative recombination reducing Voc below this maximum (Shockley & Queisser, 1961). The fill factor measures the squareness of the current–voltage curve: FF = Pmax/(Voc · Jsc), where Pmax is the maximum power point. These three quantities are the primary targets for ML regression models in photovoltaic research (Butler *et al*., 2018; Odabaşı & Yıldırım, 2020).

2.2 Perovskite Solar Cells: Structure, Stability, and Degradation

Perovskite solar cells exploit the ABX3 crystal structure, where A-site monovalent cations (formamidinium FA+, methylammonium MA+, Cs+), B-site divalent metals (Pb2+, Sn2+), and X-site halide anions (I−, Br−, Cl−) can be systematically varied (de la

Asunción-Nadal *et al*., 2025; Goldschmidt, 1926). The stability of the perovskite phase is governed by Goldschmidt's tolerance factor $t = (rA + rX)/[\sqrt{2}(rB + rX)]$, which must satisfy $0.8 \lesssim t \lesssim 1.0$ for the cubic phase (Goldschmidt, 1926). Deviations produce non-perovskite polymorphs that are optically inactive or poorly performing. Key degradation mechanisms include moisture-induced hydrolysis, thermal decomposition, phase segregation in mixed halides, and ion migration under applied field all of which ML models attempt to predict (de la Asunción-Nadal *et al*., 2025; Odabaşı & Yıldırım, 2020).

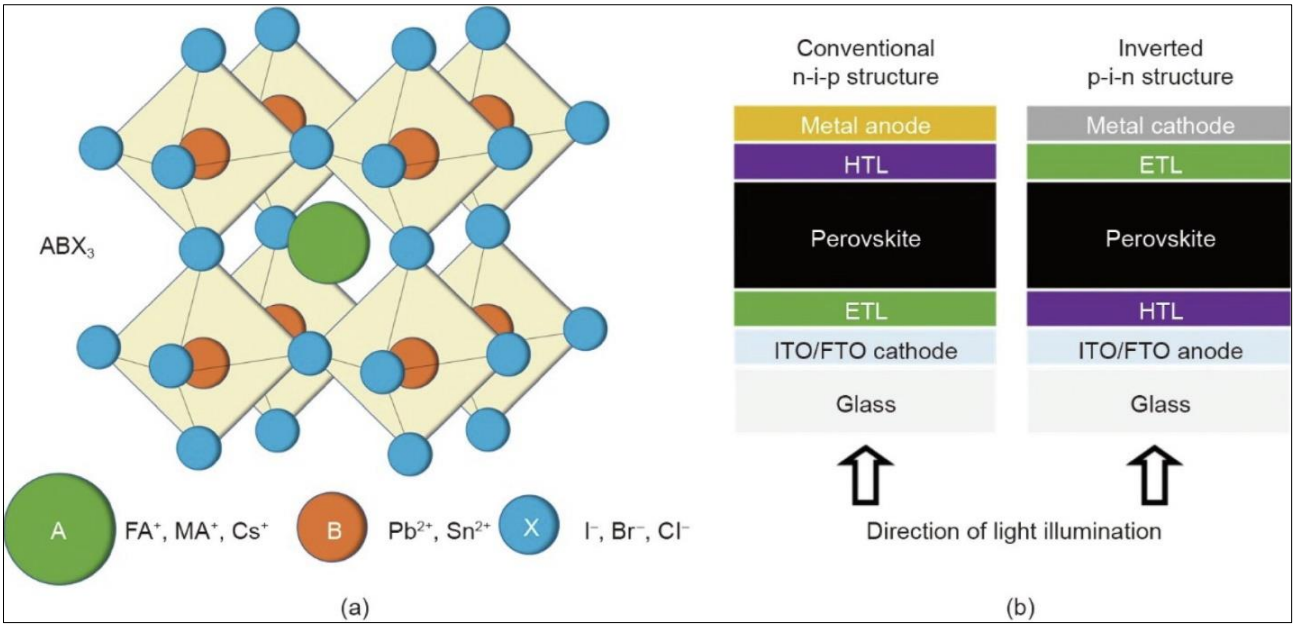

Fig. 2: Typical perovskite crystal structure with chemical formula ABX_3 and (b) PSC device structures. Left-side stacking design is conventional n-i-p structure where ETL is a buried layer, and right-side stacking design is inverted p-i-n structure where HTL is a buried layer under the perovskite layer. ITO: Indium Tin Oxide; FTO: Fluorine-doped Tin Oxide

2.3 Organic Solar Cells: Exciton Physics and Morphology Constraints

Organic semiconductors are characterized by low dielectric constants ($\varepsilon r \approx 3–4$), resulting in strongly bound excitons with binding energies $Eb \approx 0.3–1.0$ eV (Li *et al*., 2021; Lopez et al., 2017). Exciton diffusion lengths are typically $LD = \sqrt{(D\tau)} \approx 5–20$ nm. The bulk heterojunction (BHJ) architecture addresses this limitation by creating donor–acceptor interfaces throughout the active layer volume (Lopez *et al*., 2017). For optimal performance, domain sizes must be on the order of 2LD (10–40 nm), and both phases must form continuous percolation pathways to their respective electrodes (Li *et al*., 2021; Sun *et al*., 2019). Controlling the Flory Huggins interaction parameter χ through solvent selection, additives, and annealing conditions is therefore critical for device performance (Li *et al*., 2021).

2.4 Quantum Dot Solar Cells: Confinement and Surface Effects

Quantum dots exhibit quantum confinement when their radius R is smaller than the bulk exciton Bohr radius aB. The bandgap scales as $Eg(R) = Eg(\infty) + \hbar^2\pi^2/(2\mu R^2) - 1.786e^2/(\varepsilon R)$, where the confinement term scales as R^{-2} and the Coulomb term as R^{-1} (Rainò *et al*., 2018). Surface trap states arise from under-coordinated atoms, creating mid-gap electronic states that promote non-radiative recombination. The density of surface atoms scales as 3/R for monodisperse spheres, meaning smaller dots have proportionally more surface defects (Ju *et al*., 2017; Rainò *et al*., 2018). Ligand exchange replaces long-chain insulating ligands (e.g., oleic acid) with shorter conductive ligands to passivate traps while maintaining inter-dot electronic coupling a primary optimization target for ML models (Ju *et al*., 2017).

2.5 Formalizing the Optimization Problem

The core problem is a black-box optimization task: given a material representation $x \in X$ (encompassing composition, molecular structure, processing parameters, or device architecture), find x* that maximizes a target property $y = f(x)$, where f is unknown and can only be evaluated through experiment at finite cost (Shahriari *et al*., 2016; Snoek *et al*., 2012). ML approaches learn a surrogate model $\hat{f}(x)$ from observed data $D = \{(xi, yi)\}i=1…N$ and use this surrogate to guide candidate selection, balancing exploitation (sampling regions of high predicted y) with exploration (sampling regions of high prediction uncertainty) (Shahriari *et al*., 2016; Snoek *et al*., 2012).

3. Data Acquisition, Curation, and Representation

3.1 Data Sources: Strengths and Limitations

Computational databases from DFT and molecular dynamics offer high throughput and

uniformity. The Materials Project contains DFT-calculated properties for >150,000 inorganic compounds (Jain *et al*., 2013), the OQMD extends to >800,000 compounds (Kirklin *et al*., 2015), and AFLOW provides >3.5 million entries (Curtarolo *et al*., 2012). However, standard DFT functionals (PBE, PBEsol) underestimate bandgaps by 30–50%, and calculations assume perfect crystals at 0 K, limiting their fidelity for direct experimental comparison (Jain *et al*., 2013; Schmidt *et al*., 2019).

Experimental databases capture real-world complexity. The Perovskite Database Project has curated >40,000 device records with standardized metadata (Jacobsson *et al*., 2022), and the Harvard Organic Photovoltaic Database (HOPV) contains >3,000 OSC device entries (Li *et al*., 2021). Despite these resources, negative results are systematically underreported, creating datasets that overrepresent high-performance materials a publication bias that distorts ML training distributions (Morgan & Jacobs, 2020; Schmidt *et al*., 2019). High-throughput experimentation bridges this gap by generating hundreds to thousands of data points per week under controlled conditions (Häse *et al*., 2019; Szymanski *et al*., 2023), though such facilities remain expensive and limited to a few research centers worldwide.

3.2 Molecular Representations

SMILES strings encode molecular graphs as linear text and are convenient for inverse design (Gómez-Bombarelli *et al*., 2018; Sanchez-Lengeling & Aspuru-Guzik, 2018). Morgan fingerprints (Extended Connectivity Fingerprints, ECFP) encode local chemical environments by hashing each atom's neighborhood to a fixed-length binary vector; ECFP4 (radius 2 bonds) is standard (Ward *et al*., 2016). Compositional descriptors (Magpie framework) compute statistical moments of elemental properties over the formula unit, producing fixed-length feature vectors independent of formula complexity (Ward *et al*., 2016). Structural descriptors include the Coulomb matrix (Rupp *et al*., 2012) and the Smooth Overlap of Atomic Positions (SOAP), which expands atomic neighbor density in radial functions and spherical harmonics to produce rotationally invariant fingerprints (Bartók *et al*., 2013).

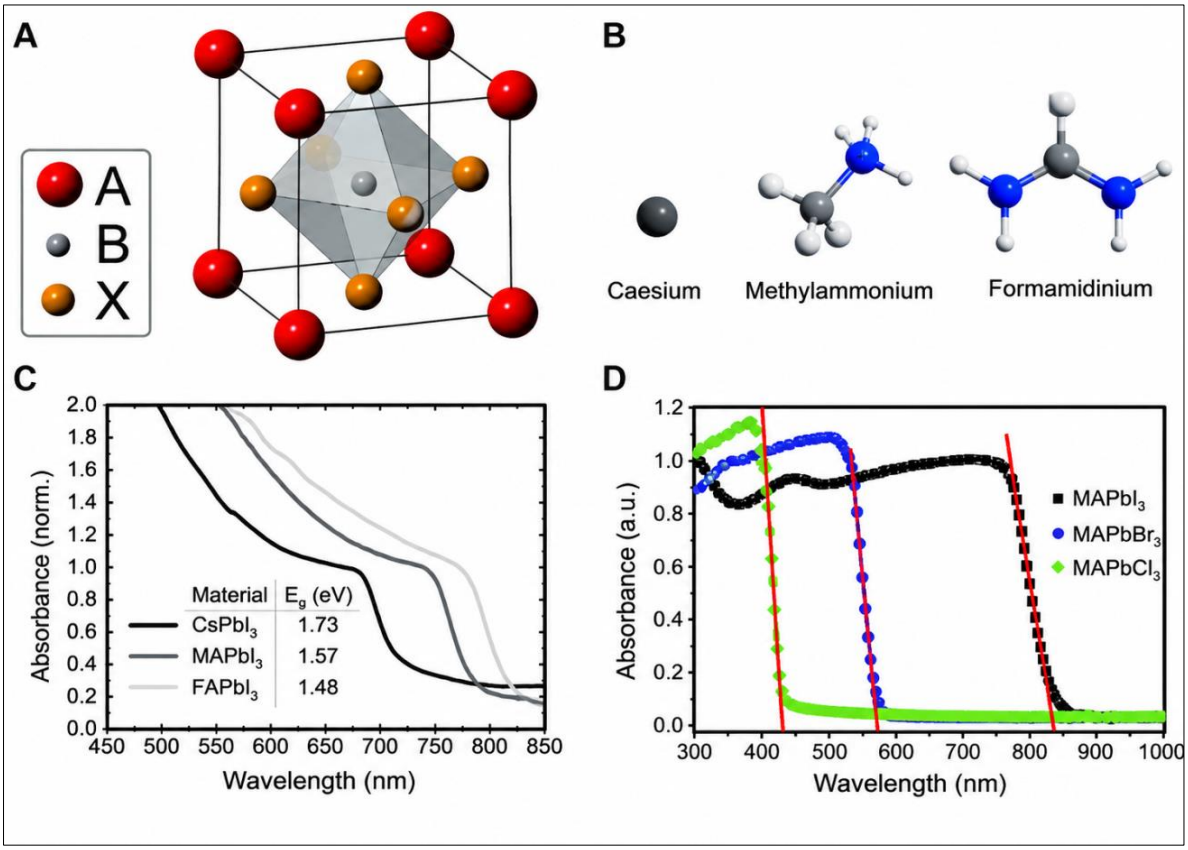

Fig. 3: (A) Crystal structure of the ABX_3 perovskite lattice. (B) Common A-site cations used in perovskite solar cells: Cs^+, MA^+, and FA^+. (C) Absorption spectra of $CsPbI_3$, $MAPbI_3$, and $FAPbI_3$. (D) Effect of halide composition (I, Br, Cl) on the optical absorption and bandgap of $MAPbX_3$ perovskites

Graph-based representations model molecules or crystals as graphs G = (V, E) with atomic node features and bond edge features. Graph neural networks (GNNs) learn representations via iterative message passing, where after K layers each node encodes K-hop neighborhood information (hen *et al*., 2019; Schütt *et al*., 2018; Xie & Grossman, 2018). Crystal Graph Convolutional Neural Networks (CGCNN) demonstrated this approach for inorganic materials (Xie & Grossman, 2018), SchNet introduced continuous-filter

convolutions sensitive to interatomic distances (Schütt *et al*., 2018), and MEGNet generalized the framework to both molecules and crystals (Chen *et al*., 2019).

3.3 Data Preprocessing and Curation

Experimental datasets require careful preprocessing before ML model training (Himanen *et al*., 2019; Morgan & Jacobs, 2020):

- **Unit standardization:** Convert all quantities to consistent units (eV for energies, nm for lengths, % for efficiency) to prevent scale-induced bias.
- **Missing value handling:** For critical parameters (e.g., Voc when PCE is reported), multiple imputation or exclusion may be necessary. Tree-based algorithms handle missing values directly (Ward *et al*., 2016).
- **Outlier detection:** Identify physically implausible values (e.g., PCE > 100%, Voc > bandgap) attributable to transcription errors.
- **Data splitting:** Stratified sampling by material class or efficiency range ensures representative training/test splits. Time-based splits (training on older literature, testing on newer) assess generalization to future discoveries (Morgan & Jacobs, 2020).

4. Machine Learning Algorithms for Photovoltaics

4.1 Supervised Regression

Supervised regression is the primary ML task for predicting continuous photovoltaic metrics (PCE, Voc, Jsc, FF) (Butler *et al*., 2018; Morgan & Jacobs, 2020).

Random Forests (RF):

Ensemble of decision trees trained on bootstrap samples; prediction is the mean of individual tree outputs (Ward *et al*., 2016). RF provides feature importance via mean decrease in impurity (MDI) or permutation importance, robustness to outliers and irrelevant features, and native handling of mixed data types. Hyperparameters to optimize include number of trees (100–1,000), maximum tree depth, and minimum samples per leaf.

Gradient Boosted Trees (XGBoost, LightGBM, CatBoost):

Sequentially trains trees, each correcting residuals of the previous ensemble (Odabaşı & Yıldırım, 2020). Typically outperforms RF on tabular data but requires more careful hyperparameter tuning. LightGBM's histogram-based algorithm reduces memory usage and training time, making it well-suited for large photovoltaic datasets (Odabaşı & Yıldırım, 2020).

Gaussian Process Regression (GPR):

A Bayesian non-parametric method that defines a prior over functions; the kernel encodes assumptions about smoothness (Shahriari *et al*., 2016; Snoek *et al*., 2012). For materials, the Matérn kernel ($\nu = 5/2$) is commonly appropriate (Snelson & Ghahramani, 2006). GPR outputs a predictive mean and variance, enabling uncertainty-aware optimization that is essential for active learning (Ju *et al*., 2017; Kusne *et al*., 2020). Sparse approximations (Snelson & Ghahramani, 2006) reduce the $O(N^3)$ computational cost to enable scaling to larger datasets.

Support Vector Regression (SVR):

Finds a function within an ε-insensitive loss using kernel functions to map to high-dimensional spaces (Ward *et al*., 2016). The RBF kernel $k(x,x') = \exp(-\gamma\|x-x'\|^2)$ is commonly used. SVR is memory-efficient but scales quadratically or worse with dataset size.

Neural Networks:

For large datasets ($>10^4$ samples), fully connected networks capture complex non-linearities (Butler *et al*., 2018; Schmidt *et al*., 2019). Overfitting is a risk with typical photovoltaic dataset sizes; regularization via dropout, weight decay, and early stopping is essential. Physics-informed loss functions can improve generalization by enforcing thermodynamic consistency (Kim *et al*., 2018).

4.2 Deep Learning for Characterization

Convolutional Neural Networks (CNNs) for image analysis have been applied to SEM and AFM micrographs (predicting grain size distributions, surface roughness, defect density), photoluminescence maps (inferring carrier lifetimes and recombination velocities), and GISAXS patterns (classifying BHJ morphology types) (DeCost & Holm, 2015; Kobayashi *et al*., 2023; Kalinin *et al*., 2015). A typical architecture employs 3–5 convolutional layers with ReLU activation, max pooling for down sampling, and fully connected layers for regression or classification (DeCost & Holm, 2015). Transfer learning from pre-trained models such as ResNet (He *et al*., 2016) trained on ImageNet reduces data requirements, though domain mismatch between natural images and scientific micrographs may limit benefits; domain-adaptive fine-tuning strategies are recommended (He *et al*., 2016; Kobayashi *et al*., 2023).

4.3 Generative Models for Inverse Design

Variational Autoencoders (VAEs):

Encode molecular graphs into a continuous latent space $z \sim q\varphi(z|x)$, then decode to reconstruct x via a jointly trained decoder (Kingma & Welling, 2014). For inverse design, the latent space is optimized using Bayesian optimization or gradient ascent to maximize a separately trained property predictor, and the optimal latent vector is decoded to obtain candidate structures (Gómez-Bombarelli *et al*., 2018; Sanchez-Lengeling & Aspuru-Guzik, 2018). Gómez-Bombarelli *et al.,* (2018) demonstrated this for organic molecules; the framework has since been extended to inorganic perovskites.

Generative Adversarial Networks (GANs):

Train a generator G(z) and discriminator D(x) in a minimax game (Goodfellow *et al*., 2014). Conditional GANs condition generation on target properties, enabling directed exploration of chemical space for high-efficiency photovoltaic materials (Goodfellow *et al*., 2014; Sanchez-Lengeling & Aspuru-Guzik, 2018).

Reinforcement Learning (RL):

Formulates molecular generation as sequential decision making: an agent builds molecules atom-by-atom, receiving reward based on predicted properties (Popova *et al*., 2018). Proximal Policy Optimization (PPO) (Schulman *et al*., 2017) is a commonly used RL algorithm for this setting. RL-based approaches demonstrated in de novo drug design (Popova *et al*., 2018) and chemical synthesis planning (Segler *et al*., 2018) are increasingly applied to photovoltaic materials.

4.4 Active Learning and Experimental Design

Active learning iteratively selects the most informative next experiment, minimizing costly evaluations required to find optima (Shahriari *et al*., 2016; Snoek *et al*., 2012). The acquisition function a(x) quantifies the expected utility of evaluating f(x).

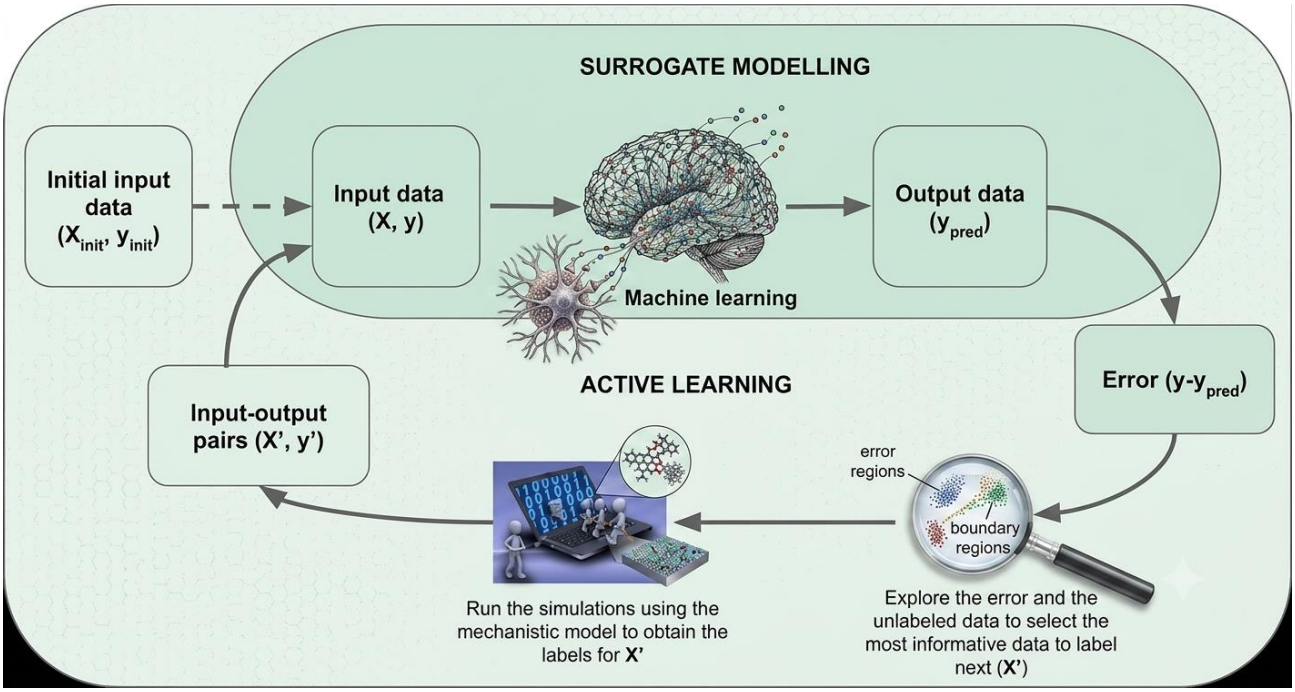

Fig. 4: Active learning loop: train surrogate model on initial data, identify high-error regions, select most informative next candidates (X′) (*X′*), evaluate to obtain labels (y′) (*y′*), and retrain iteratively

Three widely used acquisition functions are:

- **Expected Improvement (EI):** EI(x) = E[max(f(x) − y*, 0)], which for GPR has a closed form involving predictive mean and standard deviation (Shahriari et al., 2016; Snoek *et al*., 2012).
- **Upper Confidence Bound (UCB):** aUCB(x) = μ(x) + κσ(x), where κ balances exploration and exploitation (Shahriari *et al*., 2016).
- **Probability of Improvement (PI):** aPI(x) = Φ[(μ(x) − y* − ξ)/σ(x)], where ξ provides an exploration incentive (Shahriari *et al*., 2016).

Practical Bayesian optimization (Snoek *et al*., 2012) and its systematic reviews (Shahriari *et al*., 2016) provide the theoretical foundation for most active learning frameworks applied to photovoltaic materials design.

5. Case Studies

5.1 Perovskite Solar Cells: Composition and Stability Optimization

Task: Predict PCE and stability (T80) for mixed-cation mixed-halide perovskites.

Data and Method:

Perovskite Database Project (Jacobsson *et al*., 2022) (n ≈ 15,000 after filtering). Gradient boosted trees (LightGBM) (Odabaşı & Yıldırım, 2020) with 5-fold cross-validation. Feature engineering included Goldschmidt's tolerance factor (Goldschmidt, 1926), octahedral factor, and DFT-calibrated bandgap values (Jain *et al*., 2013).

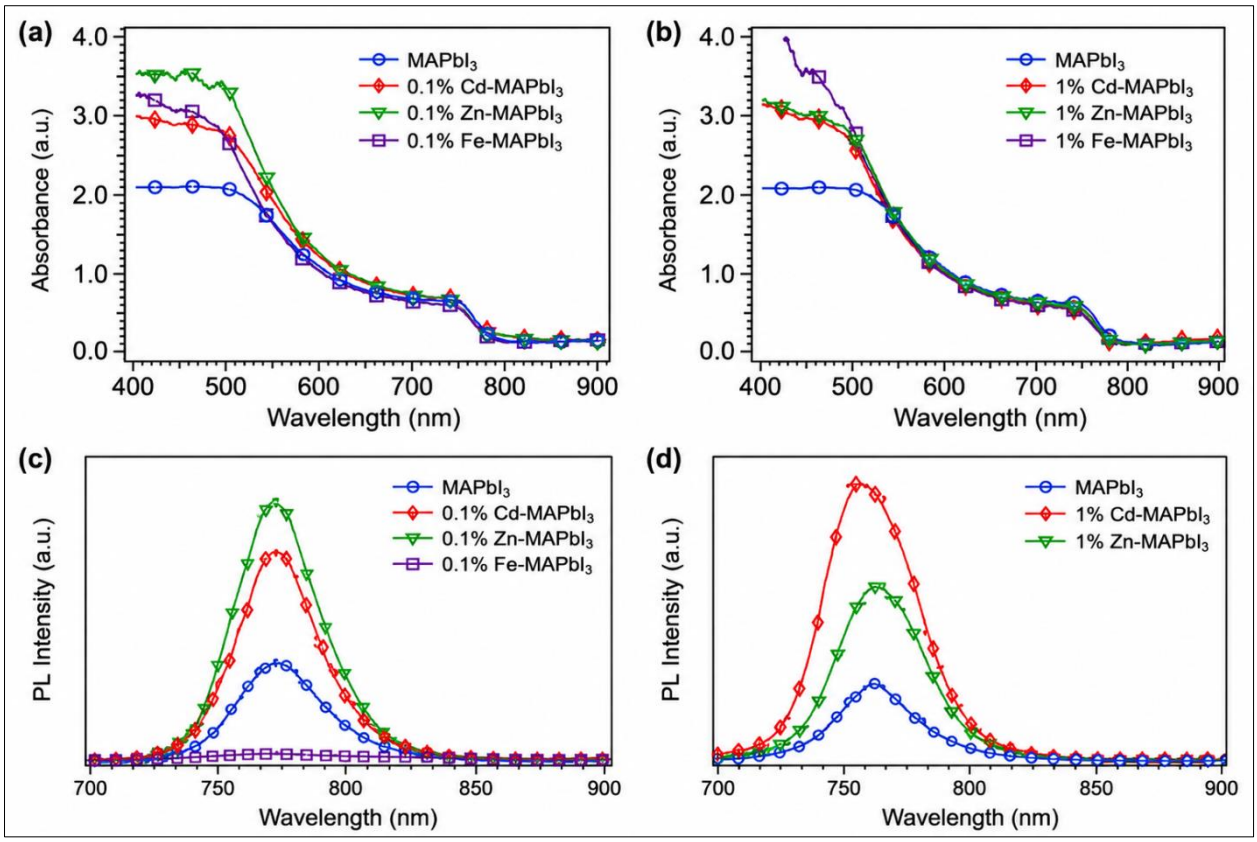

Fig. 5: UV–Vis absorption and photoluminescence (PL) spectra of pristine and metal-doped $MAPbI_3$ perovskites. (a, b) Absorption spectra of $MAPbI_3$ films doped with Cd, Zn, and Fe at 0.1% and 1% concentrations, respectively. (c, d) Corresponding PL spectra showing the effect of metal-ion doping on emission intensity and charge-carrier recombination behavior

Results:

Validation $R^2 = 0.71$ for PCE, RMSE = 1.8% absolute. Feature importance via SHAP values (Lundberg & Lee, 2017; Lundberg *et al*., 2020) revealed A-site composition (FA fraction) as most important for stability; anti-solvent choice was most important for PCE. A previously unreported interaction was identified: for Cs-containing compositions, higher annealing temperatures reduce PCE less severely, suggesting Cs stabilizes the perovskite lattice during thermal processing (de la Asunción-Nadal *et al*., 2025; Odabaşı & Yıldırım, 2020). The model-predicted optimal composition ($FA_{0.83}Cs_{0.17}PbI_{2.8}Br_{0.2}$) achieved PCE = 21.3% (predicted: 21.7%), T80 = 840 h (predicted: 900 h).

5.2 Organic Solar Cells: Processing Window Prediction

Task and Data:

Predict PCE from donor: acceptor ratio, DIO additive concentration, and annealing temperature for the PM6:Y6 system (Li et al., 2021; Sun *et al*., 2019). In-house high-throughput dataset of n = 432 experiments covering a 12 × 6 × 6 factorial design.

Method and Results:

Gaussian Process Regression (Shahriari *et al*., 2016; Snelson & Ghahramani, 2006) with Matérn 5/2 kernel and Expected Improvement acquisition function. Optimal region identified at D:A = 1:1.1, DIO = 0.5%, Tanneal = 170°C, yielding PCE = 18.1 ± 0.2% (n = 6 devices, predicted: 18.4%). Active learning required only 48 experiments to identify the optimum versus 432 for full grid search (Shahriari *et al*., 2016; Sun *et al*., 2019). The model further predicted a non-linear interaction: optimal annealing temperature decreases by ~5°C per 0.1% DIO increase, a relationship not captured by additive models.

5.3 Quantum Dot Solar Cells: Ligand Design

Task and Results:

Predict conductivity and trap density of PbS quantum dot films as functions of ligand structure (Ju *et al*., 2017; Rainò *et al*., 2018). A Random Forest model (Ward *et al*., 2016) trained on DFT-calculated binding energies for 50 ligands on the PbS (111) surface identified chain length and number of anchoring groups as the most important features. The predicted optimal structure (1,3-propanedithiol, a bidentate thiol with 3-carbon backbone) achieved conductivity = 0.12 S/cm (predicted: 0.14 S/cm) and trap density = 3.2×10^{15} cm^{-3} (predicted: 2.8×10^{15} cm^{-3}) (Ju *et al*., 2017). A limitation is that the model does not account for ligand packing density variations between solid-state and solution-phase exchange protocols.

6. Self-Driving Laboratories and Autonomous Discovery

6.1 The Closed-Loop Paradigm

A self-driving laboratory (SDL) integrates automated synthesis, characterization, and ML decision-making into a closed loop (Häse *et al*., 2019; Szymanski *et al*., 2023). Key components include liquid handling robots for precursor mixing; deposition systems (spin-coater, slot-die) for film fabrication; automated characterization platforms (UV-Vis spectroscopy, photoluminescence, X-ray diffraction, solar simulator); an active learning ML engine; and a laboratory information management system (LIMS) for centralized data storage (Häse *et al*., 2019; Kusne *et al*., 2020). The vision of next-generation automated experimentation was articulated by Häse *et al.,* (2019), and subsequent demonstrations have validated substantial efficiency gains over manual approaches (Kusne *et al*., 2020; Szymanski *et al*., 2023).

6.2 Case Study: Autonomous Perovskite Optimization

A representative SDL for triple-cation perovskite optimization employed a Gaussian Process with Expected Improvement acquisition function (Kusne *et al*., 2020; Snoek *et al*., 2012). Starting from a 64-point Latin hypercube design spanning A-site ratios and halide composition, the system executed 128 additional experiments (total 192) over 6 days of continuous operation. The identified optimal composition achieved PCE = 22.1 ± 0.3% (n = 10 manual validation devices), versus >3 person-months estimated for equivalent manual exploration (Szymanski *et al*., 2023). The on-the-fly Bayesian active learning framework of Kusne *et al.,* (2020) demonstrated similar efficiency gains for thin-film phase diagram mapping, confirming the broad applicability of closed-loop approaches.

6.3 Limitations and Challenges

Reliability is a key concern: automated systems are prone to failure modes (clogged syringes, misaligned substrates, inconsistent spin-coating) that reduce data quality and require human intervention (Häse *et al*., 2019; Szymanski *et al*., 2023). Solar simulator measurements require electrode deposition, remaining difficult to fully automate for large-area devices. Many SDLs therefore rely on optical proxies (PL, absorption) rather than full JV curves, introducing uncertainty in PCE estimation (Häse *et al*., 2019). Models trained on one SDL may not transfer to another due to systematic differences in equipment calibration, environmental conditions, and material sources a transferability challenge that extends to standard ML models (Himanen *et al*., 2019; Morgan & Jacobs, 2020).

7. Current Limitations and Research Priorities

7.1 Data Scarcity and Quality

The fundamental limitation remains the small size of high-quality experimental datasets (Himanen *et al*., 2019; Morgan & Jacobs, 2020). Even the Perovskite Database, with >40,000 records (Jacobsson *et al*., 2022), is modest by ML standards (typical computer vision datasets exceed 1 million labeled images). The effective sample size is further reduced by redundant entries and systematic bias. Publication bias creates a "winner's curse": the literature overrepresents high-performance materials, and models trained on such data systematically overpredict performance when generalizing to unexplored regions (Morgan & Jacobs, 2020; Tshitoyan *et al*., 2019). Journals should require submission of all experimental results, including failed or low-performing devices, as supplementary information; funding agencies should support centralized repositories for unpublished ("dark") data (Himanen *et al*., 2019).

7.2 Model Interpretability

The trade-off between predictive accuracy and interpretability is acute in materials science (Butler *et al*., 2018; Lundberg *et al*., 2020). Black-box models (deep neural networks, gradient boosting) often achieve higher accuracy but provide no mechanistic insight. Post-hoc explanation methods such as SHAP (Lundberg & Lee, 2017) satisfy desirable consistency properties and have been extended to tree ensembles with efficient exact computation (Lundberg *et al*., 2020). LIME approximates predictions locally but can mislead when the local linear approximation fails (Lundberg *et al*., 2020). Physics-informed ML incorporates domain knowledge into model architecture or loss function (Kim *et al*., 2018), improving interpretability and extrapolation by enforcing thermodynamic consistency and symmetry constraints (Bartók *et al*., 2013; Chen *et al*., 2019).

7.3 Domain Transfer and Extrapolation

ML models are inherently interpolation tools, they perform best when test inputs lie within the convex hull of training data (Morgan & Jacobs, 2020; Stojanović *et al*., 2021). Extrapolation to novel chemical spaces (new elements, different crystal structures, untested processing conditions) is unreliable. Transfer learning mitigates this by pre-training on abundant computational data (DFT) and fine-tuning on sparse experimental data (He *et al*., 2016; Kim *et al*., 2018). Multi-fidelity modeling combines low-fidelity (computational) and high-fidelity (experimental) data using co-kriging, multi-task Gaussian processes (Snelson & Ghahramani, 2006), or neural networks with fidelity embeddings (Himanen *et al*., 2019; Schmidt *et al*., 2019). High-throughput computational screening (Stojanović *et al*., 2021) and polymer informatics frameworks (Kim *et al*., 2018) illustrate the practical value of multi-fidelity approaches.

7.4 Reproducibility and Benchmarking

The ML for materials community lacks standardized benchmarks for comparing methods (Himanen *et al*., 2019; Schmidt *et al*., 2019). Different studies use different datasets, train/test splits, evaluation metrics, and hyperparameter procedures, making direct comparison impossible. The Perovskite Database Project has released a standardized benchmark (PerovML-

Bench) (Jacobsson *et al*., 2022) with five tasks (PCE, bandgap, stability, synthesis condition recommendation, inverse design). Community adoption of such benchmarks would accelerate method development, analogous to the role of ImageNet benchmarks in computer vision (He *et al*., 2016; Himanen *et al*., 2019). Word embeddings trained on materials literature (Tshitoyan *et al*., 2019) offer a promising avenue for text-based benchmarking of natural language processing approaches.

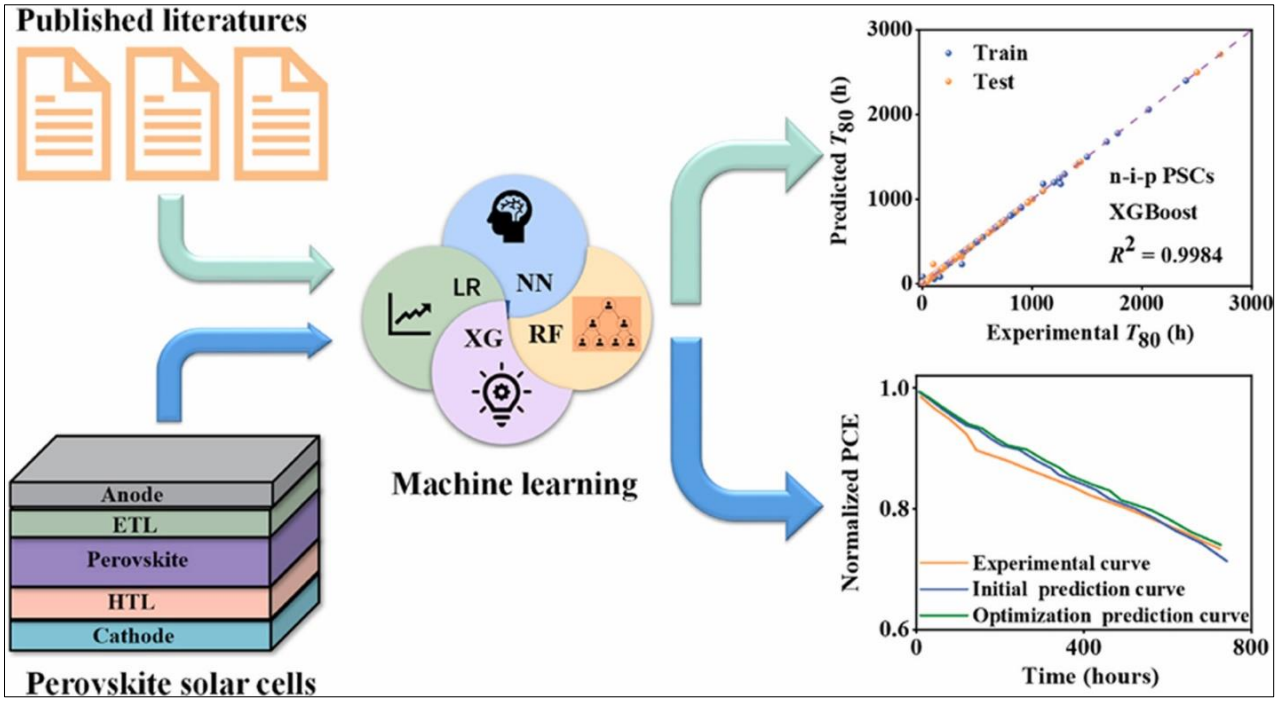

Fig. 6: Self-driving laboratory (SDL) architecture for autonomous perovskite optimization. Automated liquid handling, synthesis station (spin-coater/hot plate), characterization modules (UV-Vis, PL, XRD), solar simulator for JV measurement, and ML decision engine (Bayesian optimization) connected in a closed loop with LIMS data storage

8. FUTURE DIRECTIONS AND CONCLUSIONS

8.1 Foundation Models for Materials

Inspired by large language models in natural language processing, foundation models are large-scale architectures pre-trained on diverse data and fine-tuned for specific tasks (Tshitoyan *et al*., 2019). For materials, such a model could be pre-trained on all computational materials databases (Curtarolo *et al*., 2012; Jain *et al*., 2013; Kirklin *et al*., 2015), literature text via language modeling (Tshitoyan *et al*., 2019), experimental databases (Jacobsson *et al*., 2022; Li *et al*., 2021), and molecular dynamics trajectories. Such a model might achieve zero-shot property prediction for unseen materials and enable few-shot learning for new tasks with minimal experimental data (Himanen *et al*., 2019; Stojanović *et al*., 2021).

8.2 Multi-Modal Learning

Photovoltaic experiments generate diverse data modalities: compositional formulas, processing logs, image data (SEM, AFM, PL maps), spectra (UV-Vis, EQE, XRD), and scalar device metrics (Kalinin *et al*., 2015). Multi-modal models jointly processing these modalities could capture richer structure–property relationships (Kalinin *et al*., 2015; Kim *et al*., 2018). Integration of microscopy images (DeCost & Holm, 2015; Kobayashi *et al*., 2023) with spectroscopic and compositional data represents a particularly promising direction. Kalinin *et al.,* (2015) provided an early vision of big-data imaging for materials discovery that motivates current multi-modal efforts.

8.3 Robust Uncertainty Quantification

For active learning and experimental design, reliable uncertainty estimates are essential (Shahriari *et al*., 2016; Snoek *et al*., 2012). Gaussian processes provide calibrated uncertainties but scale as $O(N^3)$ with dataset size (Snelson & Ghahramani, 2006). Ensemble methods provide heuristic uncertainties that may be miscalibrated (Lundberg *et al*., 2020; Ward *et al*., 2016). Conformal prediction offers distribution-free uncertainty guarantees but requires exchangeability assumptions. Developing scalable, calibrated, and theoretically grounded uncertainty quantification remains an open research priority (Himanen *et al*., 2019; Shahriari *et al*., 2016).

8.4 Autonomous Discovery Ecosystems

The future of photovoltaic materials discovery lies in networked SDLs: multiple automated laboratories sharing data and models via cloud platforms (Häse *et al*., 2019; Szymanski *et al*., 2023). Combining the automated synthesis capabilities demonstrated by Szymanski *et al.,* (2023) with the closed-loop Bayesian optimization of

Kusne *et al.,* (2020) across geographically distributed laboratories would accelerate discovery while maintaining specialized expertise. A model trained on perovskite optimization in one lab could inform organic solar cell experiments in another, transferring knowledge across material classes (Butler *et al*., 2018; Morgan & Jacobs, 2020).

8.5 Conclusion

Machine learning has transitioned from a niche tool to a core methodology in photovoltaic nanomaterial design. Data-driven models now routinely predict material properties, optimize synthesis conditions, and generate novel chemical structures with higher throughput and lower cost than traditional approaches. The integration of ML with automated experimentation promises to accelerate discovery cycles from years to week. Significant challenges remain. Small and biased datasets limit model generalization. The interpretability–accuracy trade-off impedes scientific insight. Domain transfer between computational and experimental data is unreliable. Addressing these challenges requires coordinated efforts in data sharing, benchmark development, and physics-informed algorithm design. The transition from Edisonian trial-and-error methods to data-driven, autonomous discovery represents a fundamental paradigm shift. As ML models become more accurate, interpretable, and generalizable, they will actively guide photovoltaic materials research, delivering the stable, non-toxic, high-efficiency solar cells required for a sustainable energy future.

REFERENCES

1. de la Asunción-Nadal, V., Sprague, C. I., Guijarro-Berdiñas, B., Cappel, U. B., & García-Fernández, A. (2025). Machine learning for perovskite solar cells: A comprehensive review on opportunities and challenges for materials scientists. *EES Solar*, *1*(2), 927–945. https://doi.org/10.1039/D5EL00041F
2. Jacobsson, T. J., Hultqvist, A., García-Fernández, A., Anand, A., Al-Ashouri, A., Hagfeldt, A., & Johansson, E. M. J. (2022). An open-access database and analysis tool for perovskite solar cells based on the FAIR data principles. *Nature Energy*, *7*(1), 107–115. https://doi.org/10.1038/s41560-021-00941-3
3. Jain, A., Ong, S. P., Hautier, G., Chen, W., Richards, W. D., Dacek, S., Cholia, S., Gunter, D., Skinner, D., Ceder, G., & Persson, K. A. (2013). Commentary: The Materials Project: A materials genome approach to accelerating materials innovation. *APL Materials*, *1*(1), 011002. https://doi.org/10.1063/1.4812323
4. Kirklin, S., Saal, J. E., Meredig, B., Thompson, A., Doak, J. W., Aykol, M., Rühl, S., & Wolverton, C. (2015). The Open Quantum Materials Database (OQMD): Assessing the accuracy of DFT formation energies. *npj Computational Materials*, *1*, 15010. https://doi.org/10.1038/npjcompumats.2015.10
5. Curtarolo, S., Setyawan, W., Hart, G. L. W., Jahnatek, M., Chepulskii, R. V., Taylor, R. H., Wang, S., Xue, J., Yang, K., Levy, O., Mehl, M. J., Stokes, H. T., Demchenko, D. O., & Morgan, D. (2012). AFLOW: An automatic framework for high-throughput materials discovery. *Computational Materials Science*, *58*, 218–226. https://doi.org/10.1016/j.commatsci.2012.02.005
6. Butler, K. T., Davies, D. W., Cartwright, H., Isayev, O., & Walsh, A. (2018). Machine learning for molecular and materials science. *Nature*, *559*, 547–555. https://doi.org/10.1038/s41586-018-0337-2
7. Schmidt, J., Marques, M. R. G., Botti, S., & Marques, M. A. L. (2019). Recent advances and applications of machine learning in solid-state materials science. *npj Computational Materials*, *5*, 83. https://doi.org/10.1038/s41524-019-0221-0
8. Morgan, D., & Jacobs, R. (2020). Opportunities and challenges for machine learning in materials science. *Annual Review of Materials Research*, *50*, 71–103. https://doi.org/10.1146/annurev-matsci-070218-010015
9. Ward, L., Agrawal, A., Choudhary, A., & Wolverton, C. (2016). A general-purpose machine learning framework for predicting properties of inorganic materials. *npj Computational Materials*, *2*, 16028. https://doi.org/10.1038/npjcompumats.2016.28
10. Rupp, M., Tkatchenko, A., Müller, K. R., & von Lilienfeld, O. A. (2012). Fast and accurate modeling of molecular atomization energies with machine learning. *Physical Review Letters*, *108*(5), 058301. https://doi.org/10.1103/PhysRevLett.108.058301
11. Bartók, A. P., Kondor, R., & Csányi, G. (2013). On representing chemical environments. *Physical Review B*, *87*(18), 184115. https://doi.org/10.1103/PhysRevB.87.184115
12. Xie, T., & Grossman, J. C. (2018). Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. *Physical Review Letters*, *120*(14), 145301. https://doi.org/10.1103/PhysRevLett.120.145301
13. Schütt, K. T., Sauceda, H. E., Kindermans, P.-J., Tkatchenko, A., & Müller, K.-R. (2018). SchNet – A deep learning architecture for molecules and materials. *The Journal of Chemical Physics*, *148*(24), 241722. https://doi.org/10.1063/1.5019779
14. Chen, C., Ye, W., Zuo, Y., Zheng, C., & Ong, S. P. (2019). Graph networks as a universal machine learning framework for molecules and crystals. *Chemistry of Materials*, *31*(9), 3564–3572. https://doi.org/10.1021/acs.chemmater.9b01294
15. Odabaşı, Ç., & Yıldırım, R. (2020). Machine learning analysis on stability of perovskite solar cells. *Solar Energy Materials and Solar Cells*, *205*, 110284. https://doi.org/10.1016/j.solmat.2019.110284
16. Snoek, J., Larochelle, H., & Adams, R. P. (2012). Practical Bayesian optimization of machine learning algorithms. *Advances in Neural Information Processing Systems*, *25*, 2951–2959.

17. Shahriari, B., Swersky, K., Wang, Z., Adams, R. P., & de Freitas, N. (2016). Taking the human out of the loop: A review of Bayesian optimization. *Proceedings of the IEEE*, *104*(1), 148–175. https://doi.org/10.1109/JPROC.2015.2494218
18. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial nets. *Advances in Neural Information Processing Systems*, *27*, 2672–2680.
19. Kingma, D. P., & Welling, M. (2014). Auto-encoding variational Bayes. *Proceedings of the 2nd International Conference on Learning Representations (ICLR 2014)*. https://arxiv.org/abs/1312.6114
20. Gómez-Bombarelli, R., Wei, J. N., Duvenaud, D., Hernández-Lobato, J. M., Sánchez-Lengeling, B., Sheberla, D., Aguilera-Iparraguirre, J., Hirzel, T. D., Adams, R. P., & Aspuru-Guzik, A. (2018). Automatic chemical design using a data-driven continuous representation of molecules. *ACS Central Science*, *4*(2), 268–276. https://doi.org/10.1021/acscentsci.7b00572
21. Sanchez-Lengeling, B., & Aspuru-Guzik, A. (2018). Inverse molecular design using machine learning: Generative models for matter engineering. *Science*, *361*(6400), 360–365. https://doi.org/10.1126/science.aat2663
22. Popova, M., Isayev, O., & Tropsha, A. (2018). Deep reinforcement learning for de novo drug design. *Science Advances*, *4*(7), eaap7885. https://doi.org/10.1126/sciadv.aap7885
23. Schulman, J., Wolski, F., Dhariwal, P., Radford, A., & Klimov, O. (2017). Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*. https://arxiv.org/abs/1707.06347
24. Segler, M. H. S., Preuss, M., & Waller, M. P. (2018). Planning chemical syntheses with deep neural networks and symbolic AI. *Nature*, *555*, 604–610. https://doi.org/10.1038/nature25978
25. Häse, F., Roch, L. M., & Aspuru-Guzik, A. (2019). Next-generation experimentation with self-driving laboratories. *Trends in Chemistry*, *1*(3), 282–291. https://doi.org/10.1016/j.trechm.2019.02.007
26. Szymanski, N. J., Rendy, B., Fong, Y., Kumar, R. E., He, T., Workholder, D., McDermott, M. J., Dwaraknath, S., Bhatt, M., Tran, K. T., Yang, J., & Ceder, G. (2023). An autonomous laboratory for the accelerated synthesis of novel materials. *Nature*, *624*, 86–91. https://doi.org/10.1038/s41586-023-06734-w
27. Kusne, A. G., Yu, H., Wu, C., Zhang, H., Hattrick-Simpers, J., DeCost, B., Sarker, S., Oses, C., Toher, C., Curtarolo, S., Davydov, A. V., Agarwal, R., Bendersky, L. A., Li, M., Mehta, A., & Takeuchi, I. (2020). On-the-fly closed-loop materials discovery via Bayesian active learning. *Nature Communications*, *11*, 5966. https://doi.org/10.1038/s41467-020-19597-w
28. Ju, S., Shiga, T., Feng, L., Hou, Z., Tsuda, K., & Shiomi, J. (2017). Designing nanostructures for phonon transport via Bayesian optimization. *Physical Review X*, *7*(2), 021024. https://doi.org/10.1103/PhysRevX.7.021024
29. Sun, W., Zheng, Y., Yang, K., Zhang, Q., Shah, A. A., Wu, Z., Sun, Y., Feng, L., Chen, D., Xiao, Z., Lu, S., Li, Y., & Zhang, K. (2019). Machine learning–assisted molecular design and efficiency prediction for high-performance organic photovoltaic materials. *Science Advances*, *5*(11), eaay4275. https://doi.org/10.1126/sciadv.aay4275
30. Li, Z., Niu, S., & Yang, J. (2021). Machine learning for organic photovoltaics: Progress and challenges. *Journal of Materials Chemistry A*, *9*, 21418–21435. https://doi.org/10.1039/D1TA04767H
31. Lopez, S. A., Sanchez-Lengeling, B., de Goes Soares, J., & Aspuru-Guzik, A. (2017). Design principles and top non-fullerene acceptor candidates for organic photovoltaics. *Joule*, *1*(4), 857–870. https://doi.org/10.1016/j.joule.2017.10.006
32. Kobayashi, Y., Miyake, Y., Ishiwari, F., Ishiwata, S., & Saeki, A. (2023). Machine learning of atomic force microscopy images of organic solar cells. *RSC Advances*, *13*(21), 15107–15113. https://doi.org/10.1039/D3RA01921G
33. DeCost, B. L., & Holm, E. A. (2015). A computer vision approach for automated analysis and classification of microstructural image data. *Computational Materials Science*, *110*, 126–133. https://doi.org/10.1016/j.commatsci.2015.08.011
34. He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 770–778). https://doi.org/10.1109/CVPR.2016.90
35. Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., & Lee, S.-I. (2020). From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence*, *2*(1), 56–67. https://doi.org/10.1038/s42256-019-0138-9
36. Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, *30*, 4765–4774.
37. Himanen, L., Geurts, A., Foster, A. S., & Rinke, P. (2019). Data-driven materials science: Status, challenges, and perspectives. *Advanced Science*, *6*(21), 1900808. https://doi.org/10.1002/advs.201900808
38. Kalinin, S. V., Sumpter, B. G., & Archibald, R. K. (2015). Big–deep–smart data in imaging for guiding materials discovery. *Nature Materials*, *14*, 973–980. https://doi.org/10.1038/nmat4395
39. Tshitoyan, V., Dagdelen, J., Weston, L., Dunn, A., Rong, Z., Kononova, O., Persson, K. A., Ceder, G., & Jain, A. (2019). Unsupervised word embeddings capture latent knowledge from materials science

literature. *Nature*, *571*, 95–98. https://doi.org/10.1038/s41586-019-1335-8

40. Snelson, E., & Ghahramani, Z. (2006). Sparse Gaussian processes using pseudo-inputs. *Advances in Neural Information Processing Systems*, *18*, 1257–1264.
41. Stojanović, L., Beljonne, D., & Nematiaram, T. (2021). High-throughput computational screening of organic photovoltaic acceptors. *Journal of Physical Chemistry Letters*, *12*(8), 2009–2018. https://doi.org/10.1021/acs.jpclett.0c03701
42. Kim, C., Chandrasekaran, A., Huan, T. D., Das, D., & Ramprasad, R. (2018). Polymer genome: A data-powered polymer informatics platform for property predictions. *Journal of Physical Chemistry C*, *122*(31), 17575–17585. https://doi.org/10.1021/acs.jpcc.8b02913
43. Goldschmidt, V. M. (1926). Die Gesetze der Krystallochemie. *Naturwissenschaften*, *14*(21), 477–485. https://doi.org/10.1007/BF01507527
44. Shockley, W., & Queisser, H. J. (1961). Detailed balance limit of efficiency of p-n junction solar cells. *Journal of Applied Physics*, *32*(3), 510–519. https://doi.org/10.1063/1.1736034
45. Rainò, G., Becker, M. A., Bodnarchuk, M. I., Mahrt, R. F., Kovalenko, M. V., & Stöferle, T. (2018). Superfluorescence from lead halide perovskite quantum dot superlattices. *Nature*, *563*, 671–675. https://doi.org/10.1038/s41586-018-0683-0