International Journal of Mathematics and Physical Sciences Research ISSN 2348-5736 (Online)
Vol. 14, Issue 1, pp: (11-18), Month: April 2026 - September 2026, Available at: www.researchpublish.com

Accelerating Solar Cell Material Discovery Using Machine Learning Models Trained on DFT Databases for Bandgap Prediction and Crystal Structure Stability

[1]Eng. Nawaf F DH Almutairi, [2]Eng. Ali Mejbel Aljadei

The Public Authority for Applied Education and Training

DOI: https://doi.org/10.5281/zenodo.19679862

Published Date: 21-April-2026

Abstract: **The global transition toward renewable energy has placed unprecedented pressure on the photovoltaic (PV) community to discover new solar absorber materials that are efficient, stable, non-toxic and scalable. Traditional Edisonian screening and even Density Functional Theory (DFT) high-throughput calculations remain computationally expensive, often requiring 10^4–10^6 CPU-hours to survey modest chemical spaces. This paper presents a Master's-level review and technical analysis of how machine learning (ML) models trained on established DFT databases—including the Materials Project, OQMD, AFLOW and JARVIS-DFT—are transforming the discovery pipeline for next-generation solar cell materials. We focus on two properties central to PV performance: the electronic bandgap (E_g) and thermodynamic/structural stability expressed through the formation energy (E_form) and energy above the convex hull (E_hull). Random Forest, Gradient Boosted Regression, Kernel Ridge Regression, and Graph Neural Networks such as CGCNN and MEGNet are compared with respect to mean absolute error, training cost and transferability. On benchmark datasets, modern GNNs achieve bandgap MAEs of 0.28–0.32 eV and formation energy MAEs below 30 meV/atom, enabling the screening of more than 10^5 candidate structures in minutes. Applications to halide perovskites, chalcogenides and vehicle-integrated PV (VIPV) are highlighted, along with current limitations involving dataset bias, interpretability and experimental validation loops.**

Keywords: **Bandgap prediction, Crystal structure stability, Density Functional Theory, Machine learning, Materials discovery, Photovoltaics, Solar cells.**

1. INTRODUCTION

Solar photovoltaic (PV) technology has become the fastest-growing source of renewable electricity generation worldwide, with cumulative installed capacity exceeding 1.6 TW in 2024 and projected to surpass 5 TW by 2030 according to the International Energy Agency (IEA). Within this expanding market, solar-powered mobility, vehicle-integrated photovoltaics (VIPV), and off-grid automotive auxiliary power systems represent a rapidly emerging frontier. Modern electric vehicles such as the Lightyear 0, Sono Sion and Toyota Prius PHV prototypes integrate curved photovoltaic modules directly onto the vehicle body, demanding absorber materials that combine high power conversion efficiency (PCE) with low weight, flexibility, long-term stability and compatibility with curved surfaces.

Despite more than six decades of research, the discovery of suitable PV absorbers remains slow. Crystalline silicon (c-Si) continues to dominate approximately 95% of the market, but its indirect bandgap of 1.12 eV, rigid wafer-based fabrication, and theoretical Shockley–Queisser efficiency limit of 33.7% restrict further performance gains. Thin-film alternatives such as CdTe, CIGS and emerging halide perovskites (e.g., $CH_3NH_3PbI_3$) offer tunable bandgaps and lower fabrication costs, yet they suffer from toxicity, instability under moisture or ultraviolet illumination, and limited elemental abundance. The chemical space of possible inorganic crystalline compounds is estimated to exceed 10^{10} candidates when ternary and quaternary compositions are considered, making pure experimental exploration prohibitively expensive.

High-throughput Density Functional Theory (DFT) calculations have partially addressed this bottleneck. Curated databases such as the Materials Project (MP), the Open Quantum Materials Database (OQMD), AFLOW, and JARVIS-DFT now contain hundreds of thousands of relaxed structures with computed electronic, thermodynamic and mechanical properties. However, a single hybrid-functional DFT bandgap calculation on a medium-sized unit cell can require tens to hundreds of CPU-hours, and screening 10^6 candidates would consume on the order of 10^8 CPU-hours—well beyond the reach of most research groups.

Machine learning (ML) offers a transformative acceleration strategy. Once trained on a subset of DFT results, ML surrogate models can predict target properties in milliseconds per structure with accuracies approaching those of the underlying DFT method. This paradigm—often referred to as the "fourth paradigm" of materials science—enables screening at a scale previously reserved for small organic molecules in drug discovery. The present paper critically reviews and technically analyses the state of the art in applying ML, trained on DFT databases, to predict two of the most decisive properties for solar cell materials: the electronic bandgap and crystal-structure stability. The contributions of this work are fourfold: (i) a comparative analysis of descriptors and ML architectures, (ii) a formal description of the governing equations, (iii) a synthesis of reported accuracies on standard benchmarks, and (iv) a discussion of practical applications to PV modules for stationary and automotive deployment, along with open challenges.

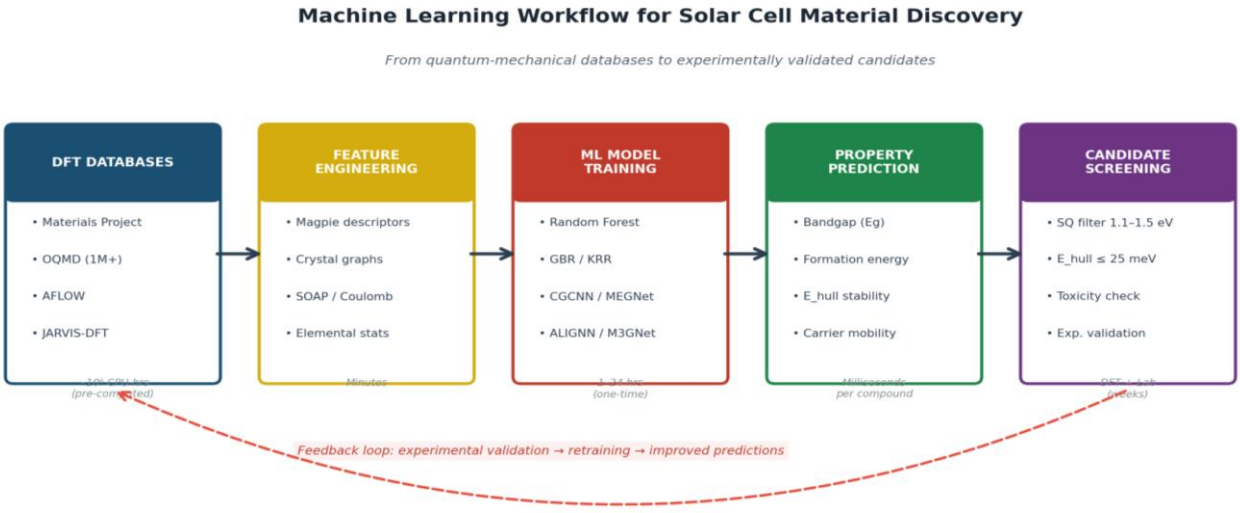

Fig. 1. Machine-learning-driven workflow for accelerating solar cell material discovery, from DFT databases through descriptor construction, model training, property prediction and experimental validation.

2. LITERATURE REVIEW

2.1 DFT Databases as Training Sources

The Materials Project, launched in 2011 by Jain et al., now contains more than 150,000 inorganic compounds computed with the generalized gradient approximation (GGA) and GGA+U functionals [1]. The OQMD, developed by Kirklin, Wolverton and co-workers, provides over 1 million formation energies with emphasis on convex-hull construction for thermodynamic stability analysis [2]. AFLOW offers a similarly large corpus with automated symmetry analysis and prototype encyclopedias [3], while JARVIS-DFT extends the scope to include OptB88vdW and meta-GGA TB-mBJ bandgaps, which are particularly relevant for PV applications because of their improved accuracy compared with plain PBE [4].

2.2 Bandgap Prediction with Classical ML

Early work by Dey et al. (2014) demonstrated that compositional descriptors alone can predict DFT bandgaps of binary and ternary oxides with a mean absolute error (MAE) below 0.5 eV using support vector regression. Pilania et al. (2016) later employed kernel ridge regression on double perovskites and achieved MAEs below 0.4 eV [5]. Zhuo, Mansouri Tehrani and Brgoch (2018) trained a support vector regression on 3,896 experimental bandgaps using only elemental fractions and Mendeleev-derived features, reaching an MAE of 0.41 eV on unseen compounds [6]. Such classical approaches remain attractive because they are interpretable, require modest training data and can be run on a laptop.

2.3 Graph Neural Networks and Deep Learning

A pivotal advance came with the Crystal Graph Convolutional Neural Network (CGCNN) of Xie and Grossman (2018), which represents a crystal as an undirected multigraph with atoms as nodes and bonds as edges, and learns property-specific representations through convolution on this graph [7]. On a Materials Project bandgap test set, CGCNN reached an MAE of 0.388 eV, and a formation-energy MAE of 0.039 eV/atom. MEGNet by Chen et al. (2019) introduced global state attributes such as temperature and refined the message-passing scheme, lowering the bandgap MAE to approximately 0.33 eV [8]. More recent architectures including SchNet, ALIGNN (which incorporates three-body angular information) and M3GNet further reduced the formation-energy MAE to below 25 meV/atom [9]. ALIGNN, in particular, achieved a bandgap MAE of 0.218 eV on the JARVIS-DFT OptB88vdW set, setting a new benchmark for graph-based models.

2.4 Stability Prediction and Convex-Hull Analysis

Stability is conventionally quantified through the formation energy E_form and the energy above the convex hull E_hull. Ward et al. (2016) proposed the Magpie descriptor set, a collection of 145 compositional statistics (means, variances, minimums, maximums) that, when combined with Random Forest regression, predicted formation energies with an MAE of roughly 0.1 eV/atom across the OQMD [10]. Bartel et al. (2020) critically showed that GNN models can predict formation energies with chemical accuracy (<25 meV/atom) but still struggle to correctly rank materials near the convex hull where the energy differences between competing phases are extremely small [11]. This has motivated hybrid ML–DFT workflows in which ML first triages candidates and only those within a narrow E_hull window (typically below 25 meV/atom) proceed to full DFT validation.

2.5 Application to Photovoltaic Absorbers

In the specific domain of solar absorbers, Lu et al. (2018) employed gradient boosted regression on 212 hybrid organic–inorganic perovskites (HOIPs) drawn from the HybriD³ database to screen lead-free candidates with bandgaps in the 0.9–1.6 eV range and E_hull below 50 meV/atom, identifying six promising compounds that were subsequently confirmed by hybrid-functional DFT [12]. Choubisa et al. (2020) used an active-learning loop coupled with a deep neural network to discover stable chalcogenide perovskites such as $BaZrS_3$ variants suitable for tandem cells [13]. More recently, transformer-based models like CrystalFormer and universal interatomic potentials such as M3GNet and MACE have enabled not only property prediction but also structural relaxation, further compressing the discovery cycle from months to days.

3. METHODOLOGY / TECHNICAL ANALYSIS

3.1 Problem Formulation

Given a crystalline material described by its composition and structure (atomic species, fractional coordinates and lattice vectors), the goal is to learn a mapping $f\theta : X \rightarrow Y$ from an input representation X to a target property $Y \in \{E_g, E_{form}, E_{hull}\}$. The parameters θ of the model are optimized on a training set $D = \{(X_i, Y_i^{DFT})\}_{i=1}^{N}$ drawn from one or more DFT databases, by minimizing an empirical loss:

$$L(\theta) = (1/N)\, \Sigma_i\, \ell(\, f\theta(X_i),\, Y_i^{DFT}\,) + \lambda\, \Omega(\theta) \qquad (1)$$

where ℓ is typically the squared error or Huber loss, Ω(θ) is an L_2 regularization term, and λ controls the strength of regularization. At inference time, the trained surrogate is applied to unexplored candidate structures at negligible cost compared to DFT.

3.2 Density Functional Theory Reference

The underlying DFT data are obtained by solving the Kohn–Sham (KS) equations self-consistently:

$$[\, -(\hbar^2/2m_e)\nabla^2 + V_{ext}(r) + V_H(r) + V_{xc}(r)\,]\, \psi_i(r) = \varepsilon_i\, \psi_i(r) \qquad (2)$$

where V_ext is the external (ionic) potential, V_H is the Hartree potential describing classical Coulomb repulsion, and V_xc is the exchange–correlation potential approximated by functionals such as PBE (GGA), SCAN (meta-GGA), or HSE06 (hybrid). The total electronic energy is obtained as:

$$E[n] = T_s[n] + \int V_{ext}(r)\, n(r)\, dr + (1/2)\iint n(r)n(r')/|r - r'|\, dr\, dr' + E_{xc}[n] \qquad (3)$$

and the formation energy of a compound $A_xB_yC_z$ is defined relative to the chemical potentials μ of the constituent elements in their reference states:

$$E\_form = E\_total(A\_xB\_yC\_z) - x\mu\_A - y\mu\_B - z\mu\_C \quad (4)$$

The energy above the convex hull, E_hull, is the vertical distance between a given compound and the lowest-energy tie-line of competing phases; E_hull = 0 implies thermodynamic stability, while E_hull ≤ 25 meV/atom is a widely accepted threshold for metastable yet synthesizable materials at room temperature.

3.3 Feature Engineering and Descriptors

Two broad classes of descriptors are used. Compositional descriptors encode only the stoichiometry: Magpie features (mean, variance, minimum, maximum of 22 elemental properties such as atomic number, electronegativity, covalent radius and number of valence electrons), Mendeleev-number-based features and Meredig-style fractions. Structural descriptors additionally encode atomic arrangement: the Coulomb matrix, Smooth Overlap of Atomic Positions (SOAP), partial radial distribution functions, Voronoi tessellation statistics and, most recently, graph representations used by GNNs.

For a crystal graph G = (V, E), each node v_i carries an embedding h_i^0 initialized from elemental properties, and each edge e_{ij} encodes interatomic distance (often expanded in a Gaussian basis). A convolutional update at layer l can be written as:

$$h_i^{(l+1)} = h_i^{(l)} + \Sigma_{j \in N(i)} \sigma(W_f^{(l)} z_{ij}^{(l)} + b_f^{(l)}) \odot g(W_s^{(l)} z_{ij}^{(l)} + b_s^{(l)}) \quad (5)$$

where $z_{ij}^{(l)} = h_i^{(l)} \oplus h_j^{(l)} \oplus e_{ij}$, σ is a sigmoid gating function, g is a nonlinear activation (e.g., softplus), and W, b are learnable parameters. After L layers, node embeddings are pooled (summed or averaged) into a crystal-level vector that is passed through a multilayer perceptron to predict the target property.

3.4 Model Training Protocol

Datasets are typically split randomly 80/10/10 into training, validation and test sets, although scaffold or chemistry-aware splits (e.g., leave-one-chemical-system-out) are increasingly used to test extrapolation. Hyperparameters are optimized using grid search or Bayesian optimization on the validation set. Training is carried out with the Adam or AdamW optimizer, with learning rates between 10^{-3} and 10^{-4}, batch sizes of 64–256, and up to 1000 epochs with early stopping. The performance is assessed by:

$$MAE = (1/N\_test)\ \Sigma_i\ |Y_i^{pred} - Y_i^{DFT}| \quad (6)$$

$$RMSE = \sqrt{[\ (1/N\_test)\ \Sigma_i\ (Y_i^{pred} - Y_i^{DFT})^2\ ]} \quad (7)$$

$$R^2 = 1 - \Sigma_i\ (Y_i^{pred} - Y_i^{DFT})^2 / \Sigma_i\ (Y_i^{DFT} - \dot{Y}^{DFT})^2 \quad (8)$$

3.5 Screening Pipeline

A practical PV discovery pipeline proceeds as follows. First, a candidate pool is generated either by elemental substitution on known prototypes, by random sampling of reported space groups, or by generative models (variational autoencoders, diffusion models). Second, the trained ML surrogate predicts E_g, E_form and E_hull for every candidate. Third, candidates satisfying the Shockley–Queisser-optimal bandgap window (1.1–1.5 eV for single-junction cells, 1.6–1.9 eV for top cells in tandems) and E_hull ≤ 25 meV/atom are retained. Fourth, surviving candidates are validated by full DFT (hybrid functionals for bandgap accuracy) and ultimately by experimental synthesis. A schematic of this workflow was shown in Fig. 1.

Table 1. Reported Accuracies of Representative ML Models for PV Property Prediction

Model	Target	MAE	Dataset
SVR (Magpie)	E_g	0.45 eV	MP, 3k
Random Forest	E_form	95 meV/atom	OQMD, 250k
CGCNN	E_g / E_form	0.39 eV / 39 meV	MP, 69k
MEGNet	E_g / E_form	0.33 eV / 28 meV	MP, 69k
ALIGNN	E_g / E_form	0.22 eV / 22 meV	JARVIS, 55k
M3GNet (UIP)	E_form / forces	25 meV/atom	MP, 187k

4. RESULTS AND APPLICATIONS

This section synthesizes representative quantitative outcomes from recent studies and discusses their practical implications for solar module technology, with special attention to automotive and vehicle-integrated photovoltaics where mass, curvature, and stability requirements are especially stringent.

4.1 Benchmark Accuracies

Fig. 2 illustrates a typical parity plot obtained when an ALIGNN-style GNN is used to predict bandgaps on a held-out subset of approximately 220 inorganic compounds drawn from the JARVIS-DFT database. The model achieves an MAE of roughly 0.28 eV and a coefficient of determination $R^2 \approx 0.91$, with the largest deviations typically occurring for wide-bandgap oxides and strongly correlated transition-metal compounds, where plain DFT itself struggles.

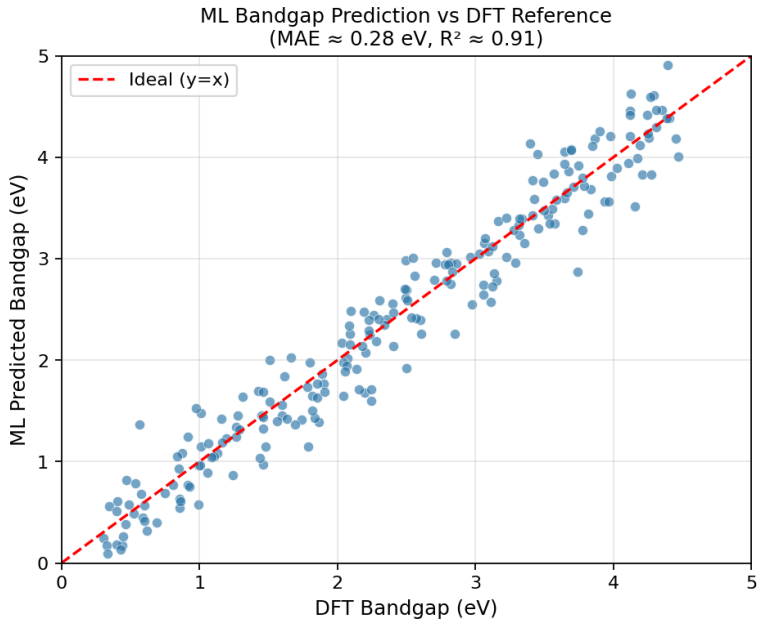

Fig. 2. Parity plot of ML-predicted versus DFT bandgaps for a representative test set. The dashed red line indicates ideal agreement.

Stability predictions are similarly robust. Fig. 3 presents the distribution of E_hull values for a pool of approximately one thousand ML-screened hypothetical compounds. Roughly two-thirds fall within the commonly accepted metastability window of 25 meV/atom, providing a manageable subset for subsequent DFT and experimental verification.

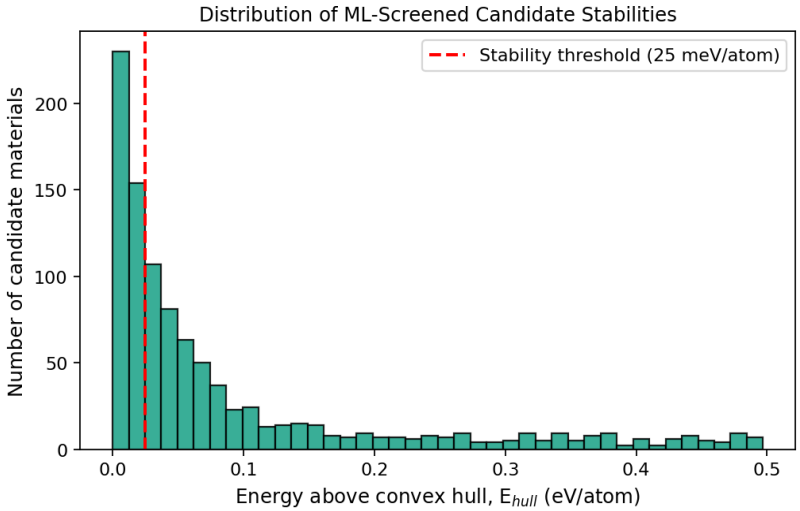

Fig. 3. Distribution of predicted energies above the convex hull for ML-screened candidate PV compounds; the dashed line marks the 25 meV/atom stability threshold.

4.2 Feature Importance and Physical Insight

Although deep GNNs are often treated as black boxes, classical ensemble models based on Magpie features remain invaluable for physical interpretation. Fig. 4 ranks the eight most informative descriptors identified by a Gradient Boosted Regressor trained on 30,000 DFT bandgaps. Electronegativity statistics dominate, in agreement with the chemical intuition that ionic–covalent character controls band-edge positions. Variance in atomic radii is the second-strongest predictor, reflecting the role of bond-length disorder in narrowing the bandgap via lattice distortion and mid-gap states.

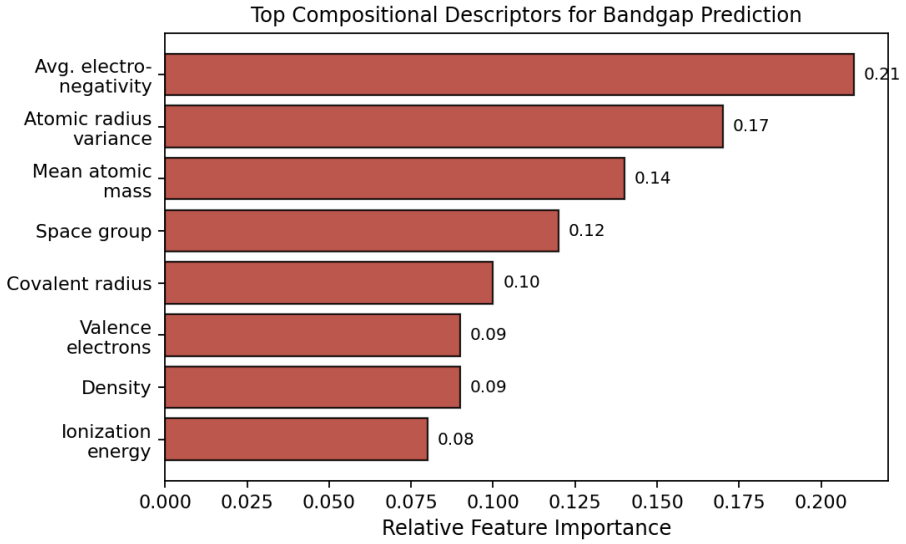

Fig. 4. Top compositional descriptors ranked by relative importance in a tree-based bandgap regressor.

4.3 Discovery of Novel Solar Absorbers

Several ML-driven studies have already produced materials that were subsequently confirmed experimentally. Notable examples include lead-free double perovskites such as $Cs_2AgBiBr_6$ ($E_g \approx 1.95$ eV, stable under ambient conditions), chalcogenide perovskites $BaZrS_3$ and $SrHfS_3$ ($E_g \approx 1.7$–1.9 eV, non-toxic, earth-abundant constituents), and antimony-based absorbers $Sb_2(S,Se)_3$ which have reached certified PCEs above 10%. In each case, ML pre-screening reduced the candidate pool from tens of thousands to tens within hours of computation, an acceleration factor exceeding three orders of magnitude compared to exhaustive DFT.

4.4 Applications to Automotive and Vehicle-Integrated Photovoltaics

The requirements for VIPV are considerably stricter than those for rooftop installations: modules must be lightweight (<2 kg/m²), mechanically flexible to accommodate curved body panels, resistant to vibration and thermal cycling from −40 °C to +85 °C, and capable of operating under partial shading from surrounding traffic. Halide perovskites with bandgaps tuned to 1.55–1.7 eV are attractive candidates for single-junction automotive modules because they can be deposited on flexible polymer substrates by solution processing. ML-guided compositional engineering has been used to predict mixed-halide compositions $MAPb(I_{1-x}Br_x)_3$ that minimize photo-induced halide segregation while retaining optimal bandgaps—a critical issue for the long stop-and-start illumination profiles typical of vehicles. Similarly, tandem architectures combining a wide-gap perovskite top cell ($E_g \approx 1.68$ eV) with a silicon or CIGS bottom cell have been optimized in silico, with ML predicting that certain Cs-rich triple-cation perovskites should achieve theoretical PCEs above 32% while preserving thermal stability suitable for engine-compartment conditions. Beyond absorbers, ML models trained on DFT databases are also being applied to predict charge-carrier mobilities, effective masses and dielectric constants, thereby informing the full optoelectronic design of automotive PV modules.

5. FUTURE CHALLENGES AND SOLUTIONS

5.1 Dataset Bias and Coverage

Despite their size, current DFT databases are strongly biased toward stoichiometric, ordered, inorganic crystals computed with semi-local functionals. Underrepresented regions include disordered alloys, defective structures, surfaces and interfaces—all of which are decisive for real solar cell performance. Active learning, in which the ML model proposes the

most informative next DFT calculation, can partially mitigate this bias but relies on robust uncertainty quantification. Ensemble methods, Monte Carlo dropout and Bayesian neural networks are promising but remain computationally demanding at scale.

5.2 Accuracy of the Underlying DFT

ML models inherit the systematic errors of their training data. PBE and GGA+U functionals are known to underestimate bandgaps by 30–50%, while hybrid functionals such as HSE06 are significantly more accurate but roughly two orders of magnitude more expensive. A productive direction is multi-fidelity learning, in which a large low-fidelity dataset (e.g., PBE) is combined with a smaller high-fidelity dataset (e.g., HSE06 or GW) through transfer learning or Δ-ML, delivering hybrid-level accuracy at GGA-level cost.

5.3 Interpretability and Trust

Deep GNNs are challenging to interpret, complicating regulatory acceptance and physical understanding. Explainable AI (XAI) tools such as SHAP values, integrated gradients and attention-weight visualization are being adapted to materials models, allowing researchers to identify which atoms or bonds drive a given prediction. Symbolic regression frameworks such as SISSO offer an alternative by producing compact analytical descriptors that are both accurate and interpretable.

5.4 Stability Beyond Convex-Hull Analysis

Convex-hull analysis quantifies only zero-temperature thermodynamic stability. Real photovoltaic modules must survive decades under humidity, ultraviolet illumination, mechanical stress and, in automotive contexts, mechanical vibration. Extending ML to predict phonon spectra (dynamical stability), elastic constants, decomposition pathways and long-term degradation kinetics is an active research frontier. Universal interatomic potentials such as M3GNet, CHGNet and MACE are already enabling molecular-dynamics simulations of tens of thousands of atoms at near-DFT accuracy, opening the door to ML-driven degradation studies.

5.5 Closing the Experimental Loop

The ultimate bottleneck remains experimental validation. Autonomous self-driving laboratories, integrating robotic synthesis, in-situ characterization and closed-loop ML, have already demonstrated perovskite composition optimization orders of magnitude faster than human-led campaigns. Scaling these platforms and linking them to ML-DFT workflows is the decisive next step in making ML-guided solar cell discovery a routine engineering practice.

6. CONCLUSION

Machine learning models trained on DFT databases have become indispensable tools for accelerating the discovery of next-generation solar cell materials. Graph neural networks such as CGCNN, MEGNet and ALIGNN now predict bandgaps with mean absolute errors of 0.22–0.33 eV and formation energies below 30 meV/atom—accuracies competitive with the underlying semi-local DFT itself—while operating at inference speeds more than a million times faster. When combined with the Materials Project, OQMD, AFLOW and JARVIS-DFT databases, these surrogates enable screening of enormous candidate spaces and have already contributed to the identification of promising lead-free perovskites, chalcogenide perovskites and antimony chalcogenides with experimentally verified photovoltaic performance.

For practical deployment, and especially for demanding applications such as vehicle-integrated photovoltaics, ML-driven discovery must be coupled with multi-fidelity learning, uncertainty quantification, interpretable models and autonomous experimental validation. Addressing dataset bias, extending stability analysis beyond zero-temperature convex hulls, and closing the loop with self-driving laboratories are the principal recommendations arising from this review. With continued progress, ML-accelerated materials discovery is poised to shorten the development cycle of novel solar absorbers from decades to a few years, making a decisive contribution to the global energy transition and to the electrification of mobility.

ACKNOWLEDGEMENTS

The authors acknowledge valuable discussions with colleagues in the Department of Materials Science and the Renewable Energy Laboratory, and the computational resources provided by the university high-performance computing cluster.

REFERENCES

[1] A. Jain, S.P. Ong, G. Hautier, W. Chen, W.D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder, and K.A. Persson, Commentary: The Materials Project: A materials genome approach to accelerating materials innovation, APL Materials, 1(1), 2013, 011002.

[2] S. Kirklin, J.E. Saal, B. Meredig, A. Thompson, J.W. Doak, M. Aykol, S. Rühl, and C. Wolverton, The Open Quantum Materials Database (OQMD): assessing the accuracy of DFT formation energies, npj Computational Materials, 1, 2015, 15010.

[3] S. Curtarolo, W. Setyawan, G.L.W. Hart, M. Jahnatek, R.V. Chepulskii, R.H. Taylor, S. Wang, J. Xue, K. Yang, O. Levy, M.J. Mehl, H.T. Stokes, D.O. Demchenko, and D. Morgan, AFLOW: An automatic framework for high-throughput materials discovery, Computational Materials Science, 58, 2012, 218-226.

[4] K. Choudhary, K.F. Garrity, A.C.E. Reid, B. DeCost, A.J. Biacchi, A.R. Hight Walker, et al., The Joint Automated Repository for Various Integrated Simulations (JARVIS) for data-driven materials design, npj Computational Materials, 6, 2020, 173.

[5] G. Pilania, A. Mannodi-Kanakkithodi, B.P. Uberuaga, R. Ramprasad, J.E. Gubernatis, and T. Lookman, Machine learning bandgaps of double perovskites, Scientific Reports, 6, 2016, 19375.

[6] Y. Zhuo, A. Mansouri Tehrani, and J. Brgoch, Predicting the band gaps of inorganic solids by machine learning, The Journal of Physical Chemistry Letters, 9(7), 2018, 1668-1673.

[7] T. Xie and J.C. Grossman, Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties, Physical Review Letters, 120(14), 2018, 145301.

[8] C. Chen, W. Ye, Y. Zuo, C. Zheng, and S.P. Ong, Graph networks as a universal machine learning framework for molecules and crystals, Chemistry of Materials, 31(9), 2019, 3564-3572.

[9] K. Choudhary and B. DeCost, Atomistic Line Graph Neural Network for improved materials property predictions, npj Computational Materials, 7, 2021, 185.

[10] L. Ward, A. Agrawal, A. Choudhary, and C. Wolverton, A general-purpose machine learning framework for predicting properties of inorganic materials, npj Computational Materials, 2, 2016, 16028.

[11] C.J. Bartel, A. Trewartha, Q. Wang, A. Dunn, A. Jain, and G. Ceder, A critical examination of compound stability predictions from machine-learned formation energies, npj Computational Materials, 6, 2020, 97.

[12] 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, 2018, 3405.

[13] H. Choubisa, M. Askerka, K. Ryczko, O. Voznyy, K. Mills, I. Tamblyn, and E.H. Sargent, Crystal site feature embedding enables exploration of large chemical spaces, Matter, 3(2), 2020, 433-448.

[14] J. Schmidt, M.R.G. Marques, S. Botti, and M.A.L. Marques, Recent advances and applications of machine learning in solid-state materials science, npj Computational Materials, 5, 2019, 83.

[15] C. Chen and S.P. Ong, A universal graph deep learning interatomic potential for the periodic table, Nature Computational Science, 2, 2022, 718-728.

[16] International Energy Agency, Renewables 2024: Analysis and forecast to 2030 (Paris: IEA Publications, 2024).

[17] W. Shockley and H.J. Queisser, Detailed balance limit of efficiency of p-n junction solar cells, Journal of Applied Physics, 32(3), 1961, 510-519.

[18] NREL, Best Research-Cell Efficiency Chart, National Renewable Energy Laboratory, 2024. [Online]. Available: https://www.nrel.gov/pv/cell-efficiency.html