ADVANCED ENERGY MATERIALS
www.advenergymat.de

Machine Learning for Accelerating Energy Materials Discovery: Bridging Quantum Accuracy with Computational Efficiency

Kwang S. Kim

Machine learning (ML) has revolutionized energy materials discovery through two key paradigms: ML potentials enabling quantum-accurate atomistic simulations with 2-4 orders of magnitude speedup over density functional theory, and ML-driven screening that efficiently navigates vast chemical spaces for rapid materials optimization. Advanced approaches, including graph neural networks and sparse Gaussian process regression incorporate physical symmetries and conservation laws, going beyond traditional statistical methods. Applications span battery materials, electrocatalysts, solar cells, phase change memory, and hydrogen storage systems, enabling simulations of thousands of atoms over extended timescales beyond the reach of quantum mechanical methods. Together, these complementary ML approaches enable predictive computational models spanning atomic to macroscopic scales. Current challenges include data quality, extrapolation to new chemical spaces, and physical interpretability. Emerging solutions involve equivariant architectures, active learning strategies, and physics-informed neural networks. The convergence of ML methodologies with experimental workflows can accelerate materials discovery and optimization. This addresses critical sustainable energy challenges in conversion, storage, and utilization while supporting the development of autonomous discovery platforms. In this way, ML helps overcome computational limitations in multiscale energy materials research and supports the efficient design of novel materials with tailored properties.

1. Introduction

The global transition to sustainable energy systems requires rapid innovation in materials for energy conversion, storage, and utilization. This need has catalyzed the development of computational approaches to accelerate materials discovery and optimization across batteries, catalysts, solar cells, and other energy systems.[1–4] The vast combinatorial space of elements, structural configurations, and surface compositions in the search for novel materials renders exhaustive investigation using conventional approaches impractical.[5] Because conventional approaches suffer from low efficiency, long timescales, and high computational cost, exploration and design of energy materials increasingly rely on innovative strategies. The integration of high-throughput material screening with predictive modeling has emerged as a pivotal focus area.

This perspective offers several contributions beyond prior reviews: 1) a comprehensive and contemporary integration of machine learning (ML) potentials with materials screening, showing how these paradigms can be unified; 2) a detailed comparison of sparse Gaussian process regression (SGPR) with graph neural networks (GNNs), including practical implementation guidelines; 3) introduction of the "materials digital twins (MDT)" concept with frameworks for multiscale modeling; 4) an expanded treatment of compositional ML approaches for broader chemical coverage; and 5) practical decision-making frameworks for method selection based on dataset characteristics and application requirements.

ML has emerged as a powerful approach to address these challenges, offering complementary pathways to accelerate energy materials research. The recent advancements in artificial intelligence (AI), neural network (NN), and big data techniques have significantly increased expectations that data-driven materials science would revolutionize scientific discoveries, establishing new paradigms for energy materials development.[6,7] The screening of high-performance materials alongside the development of models that link structural characteristics to functional properties has become central to the field.[8]

ML facilitates energy materials development through two primary approaches:

K. S. Kim
Center for Superfunctional Materials
Department of Chemistry
Ulsan National Institute of Science and Technology (UNIST)
50 UNIST-gil, Ulsan 44919, Republic of Korea
E-mail: kimks@unist.ac.kr

The ORCID identification number(s) for the author(s) of this article can be found under https://doi.org/10.1002/aenm.202503356

DOI: 10.1002/aenm.202503356

1) **ML Screening (MLS)** methodologies that efficiently navigate vast chemical spaces by establishing relationships between material properties and performance, reducing the computational burden associated with high-throughput screening[9–12]
2) **ML Potentials (MLPs)** methodologies bridging the gap between quantum mechanical accuracy and classical force-field efficiency, enabling simulations of complex material systems with near–first-principles fidelity and high computational speed,[13–16] and providing high-throughput screening databases spanning both well-studied materials and unexplored chemical or structural spaces.

MLPs, also referred to mostly as machine-learned interatomic potentials (MLIPs), have emerged as a critical bridge between highly accurate but computationally intensive density functional theory (DFT) and empirically derived classical potentials.[17,18] In 1995, Doren and coworkers employed ML techniques to construct interatomic potentials by mapping atomic structures to their potential energies.[17] These approaches promise to fill the gap between quantum mechanical methods and classical potentials, offering accuracy comparable to the former with computational efficiency approaching the latter.

Conventional MLS approaches relying on training with existing data encounter fundamental limitations when applied to novel materials lacking relevant data. Overcoming these limitations necessitates a new paradigm for novel materials design. Fortunately, the behavior of material systems is governed by quantum mechanical laws, and accurate solutions to these quantum mechanical equations could enable the prediction and creation of novel material phenomena.

Despite this quantum mechanical foundation, critical challenges persist in conventional computational materials science. DFT, providing essential quantum mechanical accuracy, exhibits cubic scaling with respect to system size, rendering it computationally prohibitive for systems exceeding a thousand atoms and ab initio molecular dynamics (MD) simulations beyond tens of picoseconds.[19] This constraint limits the applicability of quantum mechanical methods to many critical energy materials problems that inherently require large system sizes and extended time scales, such as ion diffusion mechanisms, interfacial dynamics, and degradation processes.[20]

This perspective examines recent advances in both MLS and MLP developments, with particular focus on their revolutionary impact on discovery and design of energy materials. We address the creation of efficient and accurate interatomic potentials using ML, and the application of diverse ML architectures for predicting critical properties. The fusion of these mutually reinforcing ML strategies establishes predictive simulations that connect atomic-level phenomena to macroscopic behavior and engineering-scale properties, creating what we term MDTs. These are real-time, bidirectionally coupled computational replicas that continuously update based on experimental feedback, providing novel insights into ion diffusion, phase transitions, and interfacial dynamics beyond what conventional multiscale modeling can achieve.

2. Machine Learning Approaches for Property Prediction and Materials Screening

While MLPs focus on accurate energy and force predictions for atomistic simulations, MLS approaches are used in direct prediction of materials properties to enable high-throughput screening.[20,21] **Figure 1** illustrates the comprehensive landscape of ML approaches for energy materials discovery, showcasing the interconnected methodologies from data preprocessing through advanced neural architectures to practical applications.

2.1. Database Construction and Data Infrastructure

High-quality databases are critical for ML in energy materials research, as data reliability significantly influences model accuracy. Community-driven efforts such as the Materials Genome Initiative, Materials Project, AFLOW, OQMD, NOMAD, and Google DeepMind's GNoME project provide large repositories of first-principles-calculated properties,[21–25] while platforms like Citrine Informatics and the Materials Data Facility standardize experimental data access.[26] However, major challenges remain. Experimental protocols vary across studies, published datasets favor successful results over failed experiments,[27,28] and data quality issues limit model generalizability.

Concrete solutions include: 1) Standardized experimental protocols through international collaboration; 2) Automated data extraction using natural language processing; 3) "Negative result" databases to counteract publication bias; 4) Uncertainty quantification (UQ) in all ML models; 5) Cross-validation protocols accounting for chemical similarity; and 6) Physics-informed models encoding conservation laws for improved robustness.

2.2. Feature Engineering and Descriptor Development

Feature engineering strongly influences materials ML performance, as it depends on appropriate descriptor selection.[29,30] Key descriptors for energy materials include adsorption energies of intermediates,[31,32] *d*-band centers determining adsorbate interactions,[33] coordination environments influencing electronic properties,[34] bond-orientational order parameters capturing local geometric arrangements,[35] and elemental properties (electronegativity, atomic radii, valence electrons).[36,37] Feature analysis uses filter, embedded, wrapper, and deep learning methods,[38] with deep neural networks (DNNs) showing particular success in nonlinear modeling.[39]

2.3. Machine Learning Architectures for Materials Property Prediction

Various ML architectures have been applied to materials property prediction, each with specific advantages for different problem types.[40–42] ML approaches are categorized into supervised learning (labeled datasets for regression/classification),[43–45] unsupervised learning (pattern discovery without labels),[46] and reinforcement learning (RL).[47] Although RL is attracting interest for materials discovery and optimization, its practical use in experimental discovery is still limited, with most success in

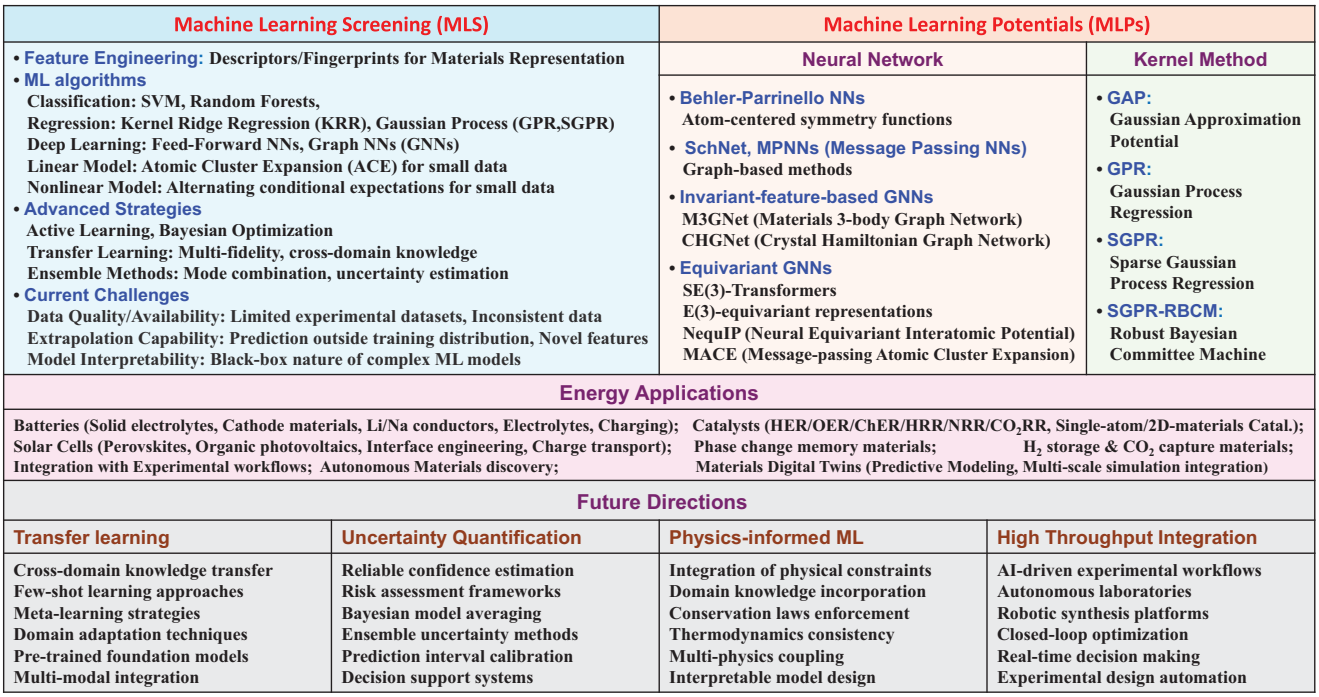

Figure 1. Machine learning approaches for energy materials discovery.

computational or simulated environments that require sequential decision-making and optimization. RL applications include autonomous experimental design, where agents learn to select optimal synthesis conditions through trial-and-error interactions with experimental systems, and inverse materials design, where RL algorithms navigate vast compositional spaces to discover materials with target properties. Recent examples include RL-driven optimization of catalytic reaction conditions, automated synthesis protocols for novel compounds, and adaptive screening strategies that learn from previous experimental outcomes to guide future material selections. In crystal structure optimization, RL agents modify atomic arrangements to achieve desired electronic or mechanical properties, receiving feedback from DFT calculations to design semiconductors with specific bandgaps. Alloy composition discovery uses RL to navigate multi-component systems, successfully identifying high-entropy alloys with superior mechanical properties. Graph-based RL frameworks represent materials as networks of atoms and bonds, enabling systematic exploration of chemical space. Multi-objective approaches balance competing properties like strength versus ductility for aerospace applications, while hierarchical RL operates across multiple length scales for composite materials design, and process optimization controls manufacturing parameters and defect engineering.

Key Algorithms:

Regression Models: Linear regression provides fast, accurate predictions for small experimental datasets.[48] Logistic regression enables rapid high-performance materials identification through classification.[49] Gaussian process regression (GPR) discovers complex relationships using unlimited parameters, particularly valuable for battery and solar cell materials.[50,51]

Neural Networks: Mathematical models imitating biological networks with fault tolerance, parallel processing, and strong nonlinear fitting capabilities.[52–54] Deep learning algorithms include convolutional NN (CNN), recurrent NN (RNN), and long short-term memory network (LSTM).[55]

Classification Algorithms: Support vector machines (SVM) divide datasets into categories with high accuracy for limited samples.[56,57] Decision trees and random forests (RFs) provide interpretable mapping relationships, with RFs sampling multiple subsets for improved performance.[58,59]

Small Dataset Techniques: The alternating conditional expectations (ACE) method finds optimal regression transformations without prior assumptions, proving valuable for predicting perovskite bandgaps with only 126 samples.[60–62]

2.4. Machine Learning Model Analysis

ML model evaluation uses cross-validation with metrics such as root mean square error (RMSE) and coefficient of determination (R^2).[63] Common validation methods include leave-one-out, *k*-fold cross-validation, and holdout validation. Effective ML models must balance prediction accuracy with practical utility, since high accuracy alone does not guarantee stability or generalization.[64]

Tables 1 and **2** provide comprehensive overviews of supervised and unsupervised/RL approaches in materials science, detailing theoretical foundations, advantages, limitations, and applications. These frameworks serve as practical guides for algorithm selection based on dataset characteristics and problem requirements.

Table 1. Supervised learning approaches in materials science.

Category/ML-algorithm[a]	Theoretical foundation	Advantages	Limitations	Applications
• *Linear Method*[b] Linear Regression, Ridge, Lasso, Elastic Net	PAC learnable, convex optimization, Tikhonov regularization, consistent learners	fast training O(nd^2) (n: data size, d: feature dimension); Reg. prevents overfitting	linear boundaries only, poor extrapolation, sensitive to outliers, assumes independence	property prediction, feature analysis, initial screening of materials database
• *Logistic Method*[c] Logistic Regression	convex log-likelihood, probabilistic framework, MLE with guarantees, generalized linear model	probabilistic output, fast inference O(d), handles small datasets, natural uncertainty	linear boundary, requires feature scaling, assumes feature indep., high-dim. overfitting	phase classification, stability prediction, synthesis feasibility, property classification
• *Support Vector Method*[d] SVM, SVR	maximum margin (γ), VC-dim.: O(R^2/γ^2) [R: radius], kernel trick, convex optimization	high-dim. efficiency, memory efficient, kernel flexibility, global optimum	critical kernel selection, O(n^3) complexity, poor scalability, no probabilistic output	non-linear prediction, limited data, high-dim. descriptors, similarity learning
• *Tree-Based Method*[e] Decision Trees, Random Forest, Extra Trees	finite VC-dimension, non-parametric, recursive splitting, information criteria	interpretable rules, handles non-linearity, robust to outliers, automatic selection	overfitting tendency, instability, biased features, exponential complexity	classification, feature ranking, catalyst screening, defect prediction
• *Ensemble: Boosting*[f] XGBoost, LightGBM, CatBoost, AdaBoost	boosting theory, weak-to-strong learning, sequential correction, margin bounds	high accuracy, handles mixed data, built-in regularization, automatic selection	hyperparameter tuning, overfitting risk, sequential training, less interpretable	high-accuracy prediction, catalyst discovery, synthesis optimization
• *Ensemble: Bagging*[g] Random Forest, Bootstrap Aggregating	bootstrap sampling, bias-variance decomp., majority voting, variance reduction	variance reduction, parallel training, robust predictions, feature importance	high memory use, Potential bias, less interpretable, limited gains	robust prediction, uncertainty estimation, feature selection, high-dim. data
• *Ensemble: Stacking*[h] Stacking, Multi-level Ensembles	meta-learning, cross-validation, generalization theory, diversity exploitation	diverse strengths, highest accuracy, bias/variance reduction, complementary algorithm	implement. complexity, computational cost, overfitting risk, interpretation	competition settings, high-stakes prediction, maximum accuracy, benchmark perform.
• *Deep Learning: Feedforward*[i] Multilayer Perceptrons, DNN	universal approx., gradient descent, backpropagation, capacity control	automatic features, universal capability, scalable, complex patterns	large datasets required, black-box nature, hyperparameter sensitivity, local minima	complex prediction, spectroscopy anal., large-scale screening, non-linear relationships
• *Deep Learning: CNN*[j] CNN, ResNet, DenseNet	translation invariance, hierarchical features, weight sharing, depth benefits	spatial patterns, parameter efficiency, translation invariant, hierarchical abstraction	large datasets, architectural complexity, computational intensity, grid-limited	image analysis, microscopy, crystal recognition, microstructure
• *Deep Learning: GNN*[k] GCN, CGCNN, MEGNet	graph theory, permutation invariance, message passing, spectral theory	structure-aware, permutation invariant, variable structures, local/global interactions	limited theory, over-smoothing, scalability: O(n), architectural complexity	crystal properties, molecular design, catalyst discovery, structure-property
• *Deep Learning: RNN*[l] LSTM, GRU, Transformers	sequence modeling, attention, memory networks, gradient flow	sequential processing, long-term dependencies, attention mechanisms, variable length	vanishing gradients, sequential bottlenecks, memory requirements, training complexity	synthesis planning, reaction prediction, time-series analysis, sequential design
• *Bayesian Method*[m] Bayesian Regression, BNN, Variational Inference	Bayes' theorem, posterior inference, prior incorporation, UQ	principled uncertainty, prior knowledge, overfitting robust, probabilistic outputs	comput. complexity, prior specification, approximation quality, scalability	decision with UQ, small data scenarios
• *Kernel Method*[n] KRR, Gaussian Process, SVM	RKHS, Mercer's theorem, representer theorem, implicit feature mapping	high-dim. mapping, non-linear modeling, theoretical foundation, global optimum	kernel selection, O(n^3) complexity, hyperparameter sensitivity, scalability	non-linear regression, similarity learning, small datasets, materials similarity
• *Gaussian Process*[o] GPR, SGPR, SGPR-RBCM	Bayesian non-parametric, RKHS, probabilistic predictions, posterior inference	natural uncertainty, prior incorporation, works with small to large datasets (SGPR-RBCM)	kernel selection, GPR: O(n^3) complexity, [SGPR: O(nm^2), RBCM: O(nm^2/p^2)]	uncertainty-aware design, active learning, Bayesian optimization

(*Continued*)

Table 1. (Continued)

Category/ML-algorithm[a)]	Theoretical foundation	Advantages	Limitations	Applications
• *Transfer Learning*[p)] Domain Adaptation, Fine-tuning, Multi-task	learning theory, domain shift, representation learning, inductive bias transfer	pre-trained models, reduced data, accelerated training, knowledge transfer	domain mismatch, negative transfer, fine-tuning complexity, source dependence	cross-domain prediction, limited data, related materials
• *Active Learning*[q)] Uncertainty Sampling, Query by Committee, Expected Improvement	information theory, optimal design, exploration-exploitation, acquisition optimization	minimizes labeling, informative samples, iterative improvement, principled selection	interaction required, acquisition design, cold start, computational overhead	experimental design, high-throughput screening, low-cost validation
• *Multi-task Learning*[r)] Multi-task Networks, Multi-output Regression	inductive transfer, shared representations, task relationships, joint optimization	concurrent related tasks, improved generalization, data efficiency, shared features	task relationships, objective balancing, increased complexity, negative transfer	multiple properties, related classes, shared features, correlated properties
• *Few-shot Learning*[s)] Meta-learning, Prototypical Network, MAML	meta-learning theory, few-shot generalization, gradient-based meta-learning	rapid adaptation, minimal data, prior experience, cross-task generalization	meta-training required, comput. complexity, limited task diversity, implementation challenge	novel materials discovery, rapid screening, minimal data scenarios
• *Online Learning*[t)] Online Gradient Descent, Perceptron	regret minimization, online optimization, mistake bounds, competitive analysis	streaming data, memory efficient, distribution adaptation, theoretical guarantees	no batch benefits, ordering sensitivity, assumption required, limited statistical analysis	real-time monitoring, streaming analysis, adaptive systems, continuous learning

[a)] n: data size, d: feature dimension, m: inducing points, p; ensemble samples. Supervised Learning: All methods require labeled training data for learning input-output mappings. PAC: Probably Approximately Correct, MLE: Maximum Likelihood Estimation, VC: Vapnik–Chervonenkis, ResNet: Residual Network, DenseNet: Densely Connected Convolutional Network, GCN: Graph Convolutional Network, GRU: Gated Recurrent Unit, RKHS: Reproducing Kernel Hilbert Space, MAML: Model-Agnostic Meta-Learning; [b)] [272–275]; [c)] [49,276]; [d)] [277–279]; [e)] [280–282]; [f)] [283–285]; [g)] [280,286]; [h)] [287,288]; [i)] [289,290]; [j)] [291–293]; [k)] [46,71,79,294]; [l)] [295–297]; [m)] [223,298]; [n)] [278,299,300]; [o)] [68,69,92,94,101,301]; [p)] [302–304]; [q)] [101,204,305]; [r)] [306–308]; [s)] [309–311]; [t)] [312,313].

3. Theoretical Foundations of Machine Learning Potentials

MLPs represent an important advance in computational materials science, bridging the accuracy of quantum mechanical methods with the efficiency of classical force fields. They learn relationships between atomic configurations and their corresponding energies and forces by training on quantum mechanical reference data. This enables prediction of material properties with near-quantum accuracy while achieving computational speeds 2–4 orders of magnitude faster than traditional DFT calculations. The key principle is to represent atomic environments with descriptors that capture local chemical information while respecting physical symmetries, then use these descriptors to predict energies and forces through various ML architectures.

3.1. Evolution of Machine Learning Potentials

The development of MLPs has progressed through increasingly sophisticated representations and architectures.[18] Early approaches handled mainly low-dimensional systems such as small molecules in vacuum, often relying on empirically derived force fields.[65] The Behler–Parrinello NN was a major breakthrough, transforming atomic coordinates into symmetry-adapted fingerprints before processing with fully connected NNs.[13] This ensured invariance under rotation, translation, and permutation of atoms, although it required careful feature engineering. These symmetry-preserving representations reduced the NN search space, improving accuracy and transferability. Kernel-based methods, notably Gaussian approximation potentials (GAPs) by Bartók et al.[66] offered an alternative by defining similarity measures between atomic environments using GPR. These methods achieved excellent accuracy but faced scalability challenges with large datasets.[5,67] A significant advancement was the development of the SGPR framework to address the $O(n^3)$ scaling of standard GPR (n: data size), which becomes computationally intractable for training datasets n exceeding $\approx 10^4$ samples. SGPR introduces a low-rank approximation using a reduced set of m inducing points that capture essential statistics of the full dataset, reducing the computational complexity to $O(nm^2)$, where $m << n$.[68]

Recent ensemble approaches have further addressed scaling issues. The Bayesian committee machine (BCM) potential[69,70] divides training data among multiple "local expert" SGPR models, combining predictions through principled weighting schemes. The robust BCM (RBCM) framework can improve the scaling of kernel regressors, with training complexity approaching $O(nm^2/p^2)$ where p is the number of local experts under ideal conditions of non-overlapping expert domains, while maintaining accuracy comparable to full GPR models.[70] Actual performance improvements depend on data distribution and expert assignment strategies.

Concurrently, message passing NNs (MPNNs)[71,72] made significant advances by treating molecules and materials as graphs, with atoms as nodes and interactions as edges. This representation naturally captures atomic connectivity and facilitates efficient information exchange. SchNet,[73,74] a pioneering MPNN for materials, demonstrated continuous-filter convolutional layers to model quantum interactions while respecting physical symmetries.

Table 2. Unsupervised learning and reinforcement learning approaches.

Category/ML-algorithm[a]	Theoretical foundation	Advantages	Limitations	Applications
• *Dimensionality Reduction*[b] PCA, *t*-SNE, UMAP	spectral methods, manifold learning, random projection, concentration inequalities, eigenvalue decomposition	variance preservation, non-linear embedding, visualization, noise reduction, comput. efficiency	information loss, parameter sensitivity, poor out-of-sample extension, interpretation issues	feature reduction, materials space mapping, data visualization, clustering preprocess.
• *Clustering*[c] *k*-means, DBSCAN, HDBSCAN, Hierarchical	Lloyd's algorithm, density-based clustering, graph theory, spectral theory, linkage criteria	pattern discovery, outlier detection, hierarchical structures, no labeling required, diverse algorithm	distance metric bias, initialization bias, cluster count choice, scalability, labels unavailable	materials discovery, phase identification, chemspace exploration, anomaly detection
• *Deep Learning – Generative*[d] VAE, GANs, Normalizing Flows	variational inference, adversarial training, likelihood maximization, invertible transformations, latent variable modeling	novel generation, data augmentation, latent exploration, unsupervised representation	training instability, mode collapse, evaluation challenges, computational requirements	inverse mater. design, novel structure generation, data augmentation, chem. chemspace exploration
• *Bayesian Methods*[e] Bayesian Clustering, Mixture Models, Dirichlet Processes	Bayesian non-parametrics, Dirichlet processes, variational inference, Markov Chain Monte Carlo sampling	automatic cluster number, principled uncertainty, hierarchical structures, model selection	comput. intensity, convergence issues, hyperparameter sensitivity, scalability limitations	unsupervised classification, phase autodetection, hierarchical organization
• *Online Learning*[f] Online Clustering, Streaming PCA, Online NMF	online optimization, streaming algorithms, incremental learning, forgetting factors	real-time processing, memory efficiency, concept drift adaptation, continuous learning	limited theory proofs, parameter tuning, stability issues, batch processing benefits lost	real-time monitoring, streaming data anal., adaptive clustering, continuous discovery
• *Reinforcement Learning*[g] Q-Learning, DQN, Actor-Critic, Policy Gradients	Markov decision process, Bellman equations, temporal difference, policy optimization, exploration-exploitation	seqential decision-making capability, adaptive behavior, action space explor., no explicit supervision	sample inefficiency, reward engineering, environ. requirements, stability issues, convergence limited	autonomous labs, exp. design optimiz., synthesis optimiz., process control, adaptive discovery

[a] All the above cases belong to unsupervised learning methods, except for RL which is neither supervised nor unsupervised learning methods. Unsupervised Learning methods discover patterns without labeled data. RL operates through environmental rewards/penalties rather than explicit supervision. PCA: principal component analysis, *t*-SNE: *t*-distributed stochastic neighbor embedding, UMAP: uniform manifold approximation and projection, DBSCAN: density-based spatial clustering of applications with noise, HDBSCAN: Hierarchical DBSCAN, VAE: variational autoencoder, GAN: generative adversarial network, NMF: non-negative matrix factorization, DQN: deep Q-network; [b] [314–316]; [c] [317–319]; [d] [320–322]; [e] [323–325]; [f] [326–328]; [g] [329–331].

3.2. Graph Neural Networks for Materials Representation

GNNs have become powerful tools for materials representation, explicitly capturing atomic connectivity and local environments. Two main categories are used: invariant models and equivariant models, each suited to different applications.

3.2.1. Invariant-feature-based GNN Architectures

Invariant-feature-based GNN architectures predict scalar properties, like total energy, that remain unchanged under rigid transformations (rotation, translation) and atom permutations. These models employ distance-based features and operations preserving invariance throughout the network.

Materials 3-body graph network (M3GNet)[75,76] represents a significant advancement in invariant-feature-based GNNs for materials. This universal model predicts multiple properties across diverse chemical spaces by incorporating three-body interactions through explicit angular terms, enabling accurate MD simulations and structure relaxations by simultaneously predicting energies, forces, and stress tensors.

CHGNet (crystal Hamiltonian GNN)[77] extends the invariant framework by using magnetic moments as charge proxies, enhancing the capability to describe both atomic and electronic degrees of freedom. By incorporating crystal symmetry operations, CHGNet demonstrates transferability across the periodic table; long-range electrostatics are typically handled approximately or via hybrid schemes.

Other notable invariant architectures include MEGNet (materials graph network),[78,79] which uses edge-conditioned convolutions to model bond interactions, DimeNet (directional message passing NN),[80] which incorporates directional information through spherical Bessel functions and spherical harmonics, and other GNN-based graph embedding methods.[81]

3.2.2. Equivariant GNN Architectures

Equivariant models represent a more comprehensive approach to incorporating symmetry constraints.[82–86] Unlike invariant models, equivariant architectures ensure that vector and tensor outputs transform appropriately under rotation and reflection operations. This is particularly important for predicting

directional quantities such as forces, dipole moments, and stress tensors.[87–89]

Equivariant GNNs overcome fundamental limitations of invariant models by ensuring that vector and tensor outputs transform correctly under spatial operations, enabling accurate prediction of directional quantities such as forces, dipole moments, and stress tensors. Specifically, equivariant models satisfy the condition $f(gx) = g'f(x)$, where g represents a group operation (rotation/reflection) and g' is the representation of g acting on the output space. This mathematical framework enables direct prediction of forces, dipole moments, and stress tensors without numerical differentiation, significantly improving both accuracy and computational efficiency. The key innovation lies in using spherical harmonics as basis functions and constructing message-passing operations that preserve equivariance at each layer, resulting in models that require less training data and exhibit superior generalization to new configurations.

Neural equivariant interatomic potential (NequIP)[83,84] implements $E(3)$-equivariant convolutions using spherical harmonics as basis functions. By enforcing equivariance at each layer, NequIP achieves remarkable data efficiency, requiring fewer training examples to reach high accuracy compared to invariant models.[83,87] This efficiency arises from the strong inductive bias provided by equivariance constraints, reducing the model's need to learn symmetry operations from data.

Message passing atomic cluster expansion (MACE)[90] represents another significant advance in equivariant architectures. By combining atomic cluster expansion with equivariant NNs, MACE achieves linear scaling with system size while maintaining high accuracy across diverse chemical environments.[85] The model's multipole expansion approach enables efficient capture of complex many-body interactions.

3.3. Sparse Gaussian Process Regression and Bayesian Committee Machine

While GNNs have garnered significant attention, kernel-based methods like SGPR offer distinct advantages for material systems where data efficiency is critical.[68,69,91,92] SGPR reduces $O(n^3)$ scaling of traditional GPR by using low-rank approximations with reduced sets of inducing points capturing essential training data features.

GAP[66,67] is a popular implementation of this approach that combines compact descriptors of local atomic environments[93] with GPR[94] to machine-learn potential energy surfaces. The GAP framework has been successfully applied to various systems, including elemental materials like carbon,[14] silicon,[15] phosphorus,[95] and tungsten,[96] as well as multicomponent systems.[97,98]

For atomic environments represented by descriptors $\rho = \{\rho_i\}_{i=1}^{n}$, SGPR defines potential energy as:

$$E(\rho, \chi) = \sum_i^n \sum_j^m K(\rho_i, \chi_j)\, \mu_j \tag{1}$$

where $\chi = \{\chi_j\}_{j=1}^{m}$ are the inducing descriptors, K is a covariance kernel (typically squared exponential), and $\mu = \{\mu_j\}_{j=1}^{m}$ are weight vectors, obtained through the approximation:

$$K_{nn} \approx K_{nm} K_{mm}^{-1} K_{nm}^{T} \tag{2}$$

This formulation allows SGPR to maintain the accuracy of full GPR while reducing computational complexity to $O(nm^2)$.

To further enhance the scalability and chemical coverage of SGPR, the BCM framework partitions training data into specialized "expert" models focused on specific chemical domains.[69,70] In RBCM, predictions for new structures combine outputs of multiple expert models using confidence-weighted averaging. The total energy is approximated as:

$$E \approx \hat{S} \sum_a \left(\beta_a / s_a^2\right) E_a \tag{3}$$

where E_a is energy prediction from local expert model a, $s^2{}_a$ is predictive variance, $\beta_a = \log(s^2{}_p / s^2{}_a)$ is a differential entropy term measuring the information gain of each expert relative to the combined prediction, and $\hat{S}$ is a normalization factor. This weighting scheme ensures that experts more familiar with a given chemical environment contribute more significantly to the final prediction. The BCM approach offers key advantages: i) Data efficiency with high accuracy using remarkably small datasets; ii) Built-in UQ through Bayesian framework; iii) Modular expandability for new chemical environments without retraining entire systems.

3.4. Comparison of GNNs and SGPR

GNNs and SGPR represent two distinct but complementary ML paradigms for materials science. The optimal method for a given task is not universal but depends on key factors such as dataset size, data diversity, and the need for UQ. A systematic comparison reveals their respective strengths and ideal operating regimes.

3.4.1. Architectural Distinctions

GNNs, by treating materials as graphs with atoms as nodes and interactions as edges, naturally capture the graph-like nature of molecular and crystalline systems. These architectures employ message-passing mechanisms (using models like MEGNet and DimeNet) to learn local and global information through iterative feature propagation, making them highly expressive and well-suited for modeling complex structural relationships. GNNs scale favorably with system size, typically with O(n) complexity, making them ideal for high-throughput screening of large datasets. However, this expressivity requires significant amounts of labeled data and substantial computational resources during both training and inference. They are also known to give overconfident but incorrect predictions when extrapolating to novel chemical spaces.

In contrast, SGPR is a kernel-based method that operates on descriptors of local atomic environments (such as smooth overlap of atomic potential and atom-centered symmetry functions). It addresses the computational intractability of standard GPR, which scales as $O(n^3)$, by using sparse approximation with inducing points to reduce complexity to $O(nm^2)$. With the RBCM framework, scalability is further improved to $O(nm^2/p^2)$ through dataset partitioning among multiple local experts.

The most significant distinction is their approach to uncertainty. SGPR, being a Bayesian method, intrinsically provides

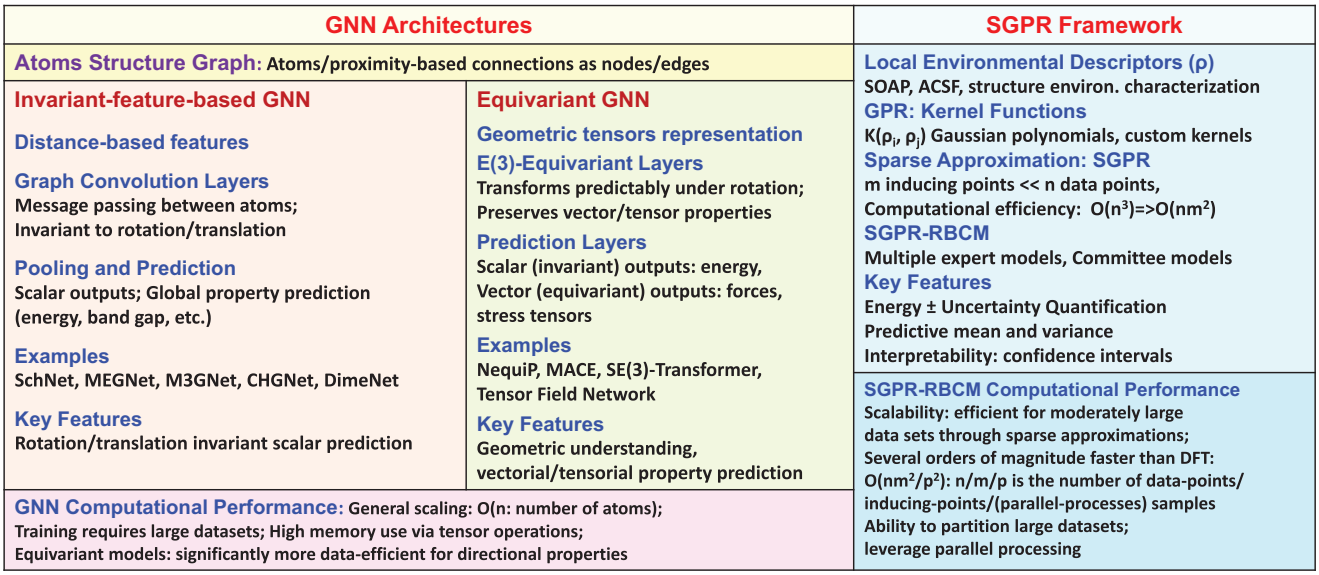

Figure 2. Comparison of graph neural network (GNN) and sparse gaussian process regression (SGPR) frameworks.

Bayesian predictive variance and natural uncertainty estimates critical for high-stakes applications and active learning workflows. This built-in UQ enables SGPR to achieve superior data efficiency—the model's awareness of uncertainty regions enables strategic, information-maximizing selection of new data points, providing a direct causal link to achieving ≈90% reduction in required experiments in some cases. GNNs require additional techniques (deep ensembles, dropout, or BNNs including Laplace approximations, Stochastic Weight Averaging Gaussian, and Improved Variant of Online Newton) for UQ, making them less suitable when reliable error bounds are critical for decision-making.

While GNNs possess high expressivity through deep architectures and non-linear activation functions, this can lead to overfitting with limited training data. SGPR provides more conservative predictions with appropriate uncertainty bounds, suitable for high-stakes decision-making.[99,100] **Figure 2** illustrates the fundamental architectural differences between these approaches.

3.4.2. Performance Across Dataset Scales

Figure 3 demonstrates how method performance varies with dataset size and model-performance/UQ, revealing distinct optimal operating regimes:

1) *Small datasets (100–1K samples)*: GPR methods achieve high performance with natural UQ, though uncertainty estimates depend on training data coverage.
2) *Medium datasets (300–30K samples)*: This represents a critical Decision Point. Around this region, the choice is highly dependent on the specific problem, the diversity of the chemical space, and the need for UQ. For < 10K samples, SGPR methods can achieve high performance with natural UQ. For > 10 *K* samples clustered into distinct, well-separated chemical domains, an SGPR-RBCM approach would be favored due to its ability to partition data and leverage expert domain knowledge.
3) *Large datasets (10K+ samples)*: GNNs demonstrate clear advantages in scalability and pattern recognition capabilities, making them the superior choice for high-throughput screening of broad chemical spaces.

3.4.3. Method Selection Framework

The choice between GNNs and SGPR is not one of universal superiority but rather a strategic decision tailored to the stage of the materials discovery pipeline. As illustrated in Figure 3, **Table 3** provides a practical decision matrix consolidating architectural differences, data requirements, UQ, and scalability characteristics into actionable guidance for method selection.

The optimal strategy often involves a multi-stage approach, where an SGPR-based method is used in exploratory phases to guide efficient data acquisition, followed by a transition to GNNs for large-scale screening once sufficient data becomes available. The choice between GNNs and SGPR depends on the following primary factors:

1) *Data Availablity & Diversity*: GNNs require substantial datasets (>10*K* samples) with diverse chemical coverage, while SGPR excels with smaller, high-quality datasets (< 10*K* samples), where expert domain knowledge can guide active sampling strategies. The chemical diversity of a dataset could often be a more important factor than its size. For medium-sized datasets (3*K*-30*K*) clustered into distinct chemical domains, SGPR-RBCM is advantageous as it partitions data and leverages domain expertise. GNNs may train such data, but often underperform without abundant, expensive samples in each domain.
2) *Uncertainty Requirements*: SGPR provides natural Bayesian uncertainty estimates critical for high-stakes applications (like

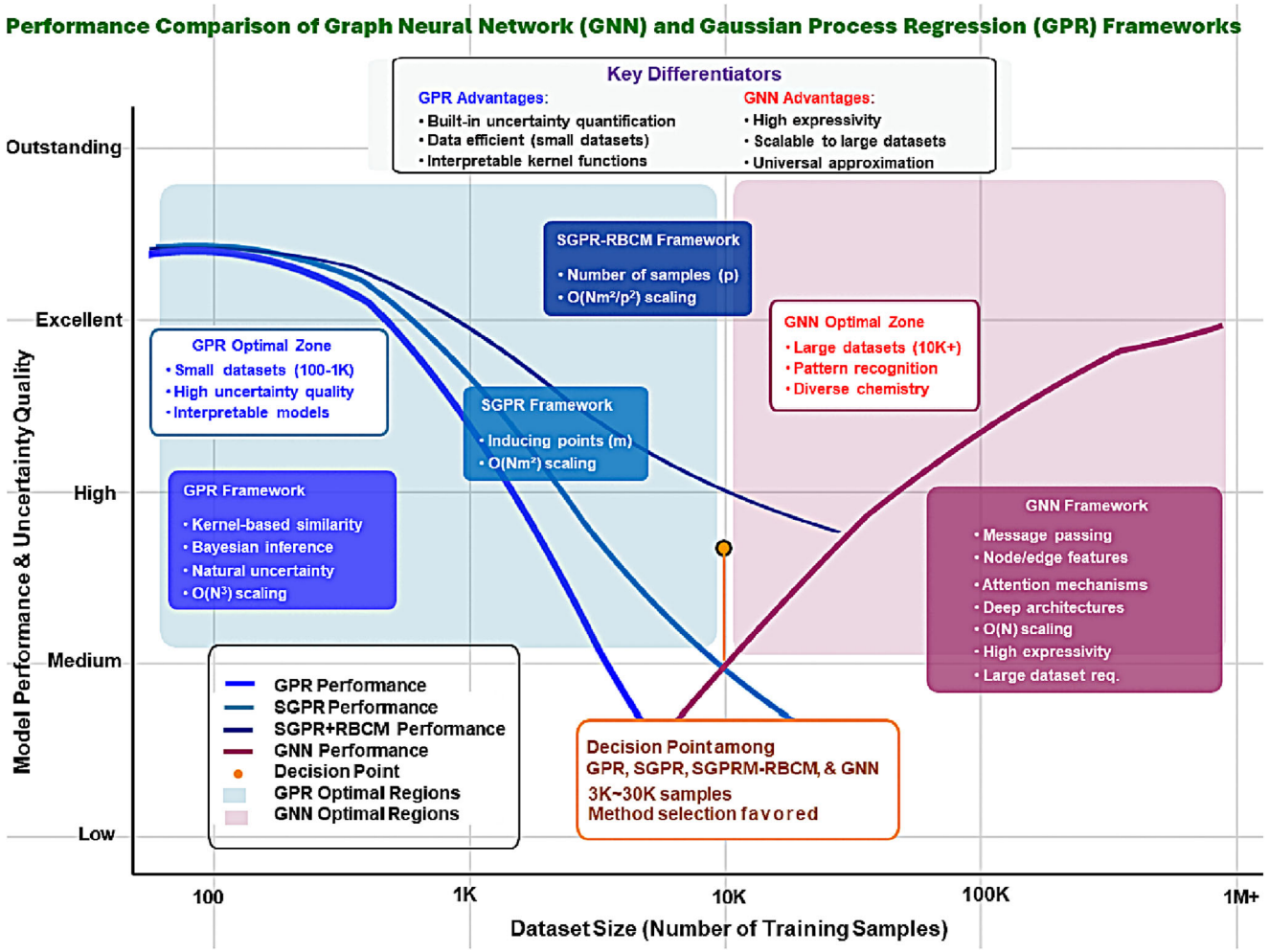

Figure 3. Illustration of performance comparison of GNN versus GPR(+SGPR, SGPR-RBCM) across data scales.

drug discovery or nuclear materials) and active learning workflows. GNNs require post-hoc methods (deep ensembles, dropout, or BNNs) for UQ, making them less suitable when reliable error bounds are critical.

3) *Computational Constraints*: GNNs offer near-O(n) scaling (for fixed local connectivity, though not general) with highly optimized GPU-accelerated batch processing for large-scale screening. Practical efficiency, however, depends on network depth, connectivity patterns, and cutoff radii. SGPR-RBCM scales as O(nm^2/p^2) under ideal conditions. This remains CPU-efficient (with moderate parallel computations for p ensemble samples) for small-to-medium datasets requiring high-accuracy on limited samples.

4) *Interpretability and Transferability*: SGPR kernel functions provide direct insight into feature importance and similarity measures between atomic environments, which is a key aspect

Table 3. GNN versus SGPR selection guide for energy materials.

Criterion	GNNs favored	SGPR-RBCM favored
Dataset size	> 10*K* samples, diverse chemistry	< 30*K* samples, quality focus
Architecture	Message passing, attention mechanisms	Kernel-based, Bayesian inference
Uncertainty	Requires post-hoc calibration	Natural Bayesian uncertainty
Interpretability	Requires specialized tools	Direct kernel interpretability
Scalability	O(n) under ideal condition, parallel batch processing	O(nm^2/p^2) under ideal condition, CPU efficient, sparse matrices
Applications	Large-scale screening, pattern recognition	Active learning, interpretable models
Computational	High throughput processing	Data-efficient predictions
Material coverage	Broad, diverse material species	Focused domains with expert knowledge

of its interpretability, yielding reliable uncertainty for predictions beyond training data. GNN interpretability requires specialized techniques and may produce overconfident predictions when extrapolating.

In addition, project-oriented recommendations are as follows:

1) *Novel electrolyte discovery with minimal experiments*: (SGPR + active learning): Natural uncertainty guides the active learning loop, reducing expensive DFT sampling.
2) *High-throughput screening of a database with 100K+ entries*: (GNNs): Highly scalable and GPU-parallelizable for rapid processing of vast material databases.
3) *Universal potential for quaternary alloys*: (SGPR-RBCM, Compositional ML): Modular, expert-guide modeling allows systematic expansion by combining models from constituent subsystems, avoiding the need for monolithic retraining.
4) *Directional properties (*e.g., *forces in MD simulation)*: (Equivariant GNNs (e.g., NequIP, MACE)): Ensure symmetry-preserving predictions for vectorial quantities.

3.4.4. Practical Applications and Hybrid Approaches

GNN Applications: Large-scale materials screening, universal property prediction (M3GNet, CHGNet), and high-throughput processing across diverse chemical spaces.

SGPR Applications: Active learning for force field development,[101] Bayesian optimization for materials discovery,[102] and small-data scenarios requiring reliable uncertainty bounds.[103]

Hybrid Methods: Recent hybrid approaches combine both methods' strengths, incorporating UQ into GNN architectures[104,105] or using GNN-learned features as kernel functions for SGPR. These developments aim to achieve GNN scalability while maintaining SGPR's uncertainty reliability, creating more robust materials models for complex discovery pipelines.

4. Characteristic Features of Machine Learning Methods for Energy Materials

Table 4 provides a comprehensive framework for ML method selection in materials science applications, integrating theoretical foundations with practical implementation guidelines. This framework guides selection based on dataset characteristics, interpretability requirements, and domain-specific constraints, providing practical decision-making support for materials scientists implementing ML methodologies. It would be useful to refer to previous literature on theoretical foundations and advanced analysis.[12]

While Table 4 presents a general framework applicable across materials domains, the following discussion demonstrates its specific implementation in energy materials applications.

Key Materials Science Considerations encompass distinct data characteristics and algorithm selections. For small datasets, novel battery chemistries benefit from Gaussian Processes with electrochemical priors for capacity prediction, while emerging photovoltaic materials leverage Bayesian optimization for bandgap tuning with minimal synthesis experiments. Transfer learning approaches successfully bridge DFT databases with experimental property prediction. High-dimensional features in energy applications include compositional screening using the least absolute shrinkage and selection operator (LASSO) for identifying key elements in multi-component electrolytes and spectroscopic data analysis through RFs for XPS/XRD pattern characterization in catalyst studies. Structured energy data applications encompass crystal structures analyzed via GNNs for predicting structures and properties of materials[106–109] and reaction networks modeled through recurrent neural networks (RNNs) for catalytic reaction pathways and intermediates.[110] Multi-fidelity energy data integration combines DFT calculations with experimental measurements through hierarchical models bridging computational and measured properties.

Physics-Informed Constraints for Energy Materials involve energy-specific symmetries, such as crystal symmetries preserved through GNNs in solid electrolyte design and electrochemical potential constraints ensuring thermodynamic consistency in battery models. Energy conservation laws are implemented through charge conservation via physics-informed NNs for electrochemical reaction modeling and through Lagrangian NNs that enforce energy balance for thermoelectric device optimization. Scale invariance in energy materials addresses size effects in nanoparticle catalysts using scale-equivariant architectures and concentration dependencies in electrolytes using dimensional analysis. Energy-specific periodicities include cyclic processes modeled with Fourier features for battery charge–discharge cycles and seasonal variations captured through periodic kernels for renewable energy resource modeling.

Energy Applications Uncertainty and Risk Assessment addresses battery safety uncertainties through heteroscedastic models for temperature-dependent thermal runaway risk and ensemble disagreement methods for cycle life prediction uncertainty. Catalyst screening uncertainties utilize BNNs for reaction rate uncertainty in catalyst optimization and Gaussian Processes for uncertainty-guided experimental design in CO_2 conversion selectivity assessment. Solar cell reliability incorporates a mixture of experts for different degradation mechanisms in photovoltaics and calibrated uncertainty for long-term efficiency forecasting. Out-of-distribution detection employs Mahalanobis distance for identifying unexplored chemical spaces in energy materials and density-based methods for extreme operating conditions.

The Model Selection Framework considers data size with Gaussian processes for battery materials screening ($n < 100$),[111–116] tree-based methods for catalyst discovery ($100 < n < 10K$),[117–123] various ML methods for perovskite solar cells ($n \approx 10K$),[124–130] and DNNs for battery signals[131] and solid electrolyte screening[132,133] ($n > 10K$). GNNs with proper regularization ($n > 10K$) often outperform other methods due to their ability to learn hierarchical feature representations directly from the graph structure of materials. Energy-specific interpretability requirements range from high interpretability using linear models for battery safety[134,135] and thermoelectric materials performance optimization[136] to medium interpretability

Table 4. Framework for ML method selection in energy materials.

Key materials science considerations	Implementation guidelines
■ *Data Characteristics and Algorithm Selection*[a)] • *Small datasets (n < 1000):* Gaussian Processes for uncertainty, Bayesian methods for prior incorporation, Transfer Learning for domain knowledge • *High-dimensional features (d >> n):* Regularized methods (LASSO, Ridge), RFs for feature selection, NNs with dropout • *Structured data:* GNNs for crystal structures, CNNs for image data, RNNs for sequential processes • *Multi-fidelity data:* Hierarchical models, Transfer Learning between fidelity levels, Multi-task learning frameworks	■ *Validation Strategies for Materials Science*[e)] • *Time series data:* Forward chaining validation to respect temporal ordering • *Materials composition:* Leave-one-composition-out to test generalization across chemical space • *Crystal systems*: Leave-one-system-out to evaluate transferability across structure types • *Property-based:* Stratified *k*-fold ensuring balanced property distributions • *Spatial data:* Spatial cross-validation for materials with location-dependent properties
■ *Physics-Informed Constraints and Inductive Biases*[b)] • *Symmetry preservation*: GNNs with equivariant layers, invariant neural architectures • *Conservation laws:* Physics-informed NNs with constraint enforcement, Lagrangian NNs • *Scale invariance:* Appropriate feature normalization, scale-equivariant architectures, dimensional analysis • *Periodicity:* Fourier-based features, periodic kernel functions, cyclic NN architectures	■ *Performance Metrics and Evaluation*[f)] • *Regression tasks:* Mean absolute error (MAE) for robust performance assess.RMSE for penalty on large errors R^2 coefficient for explained variance interpretation Mean absolute percentage error (MAPE) for relative perform. • *Classification tasks:* Accuracy for balanced datasets *F*1-score for imbalanced classes common in mater. discovery Receiver Operating Characteristic - Area Under the Curve (ROC-AUC) for probabilistic ranking capabilities Matthews Correlation Coefficient for true performance on imbalanced data • *Ranking tasks:* Spearman correlation for monotonic relationships Kendall's tau for robustness to outliers Top-*k* accuracy for practical discovery scenarios • *Uncertainty quantification:* Calibration error for probability assessment quality Prediction interval coverage for uncertainty reliability Sharpness metrics for uncertainty informativeness
■ *Uncertainty Requirements and Risk Assessment*[c)] • *Aleatoric uncertainty:* Heteroscedastic NN, distributional regression, quantile regression • *Epistemic uncertainty (also capture aleatoric*): BNNs, GPR, ensembles, mixture of experts, dropout • *Out-of-distribution detection:* Ensemble disagreement, Mahalanobis distance, density-based methods • *Calibration:* Platt scaling, temperature scaling, isotonic regression for probability calibration	
■ *Selection Practice: Model Selection Framework*[d)] • *Data size considerations:* $n < 100 \rightarrow$ Gaussian Processes, Bayesian methods, GPR $100 < n < 10K \rightarrow$ Tree-based methods or SVM, SGPR $n \approx 10K \rightarrow$ SGPR-RBCM $n > 10K \rightarrow$ NNs with proper regularization • *Interpretability requirements:* High → Linear methods or Decision Trees Medium → Random Forest with feature importance Low acceptable → NNs or ensemble methods • *Uncertainty quantification needs:* High → Gaussian Processes or BNNs Helpful → Ensemble methods or calibrated models Not required → Any method based on performance • *Computational constraints:* Low → Linear methods or simple tree models Medium → Tree-based ensembles or kernel methods High → DNNs or large ensembles	■ *Feature Engineering and Data Preprocessing*[g)] • *Chemical descriptors:* Elemental properties, stoichiometric features, structural descriptors • *Physical descriptors:* Electronic properties, mechanical properties, thermodynamic quantities • *Structural descriptors:* Coordination numbers, bond lengths, crystal symmetries • *Compositional descriptors:* Average properties, variance measures, extremal values • *Normalization strategies: Z*-score standardization, min-max scaling, robust scaling for outliers

[a)] [46,71,95,312,326,332–335]; [b)] [246–249]; [c)] [69,99–105,336]; [d)] [223,242–245,278,301,303,335–344]; [e)] [273,345–348]; [f)] [334,349,350]; [g)] [351–355].

through RF for solid electrolytes screening.[137,138] Energy materials UQ addresses safety-critical battery thermal and aging management.[139,140]

Implementation Guidelines for energy-specific applications span multiple framework components:

Energy-Specific Validation Strategies include forward chaining validation for battery degradation studies, leave-one-element-out validation for catalyst composition space, and spatial cross-validation for geospatial energy materials.

Energy Materials Performance Metrics encompass energy storage applications (MAE/RMSE for battery capacity, R^2/MAPE for power density), energy conversion metrics (F1-score/ Receiver Operating Characteristic-Area Under the Curve (ROC-AUC)) for solar cells, Spearman correlation for catalyst activity), and energy storage safety metrics (Matthews correlation coefficient for thermal stability).

Energy-Specific Feature Engineering involves battery materials descriptors (electrochemical properties, structural features),

catalyst descriptors (electronic and surface properties), solar cell material features (optical properties, electronic band structure), and energy-specific normalization approaches (electrochemical scales, temperature dependencies).

5. Applications of Machine Learning for Energy Materials

ML techniques have found diverse applications across various energy materials domains, from batteries and catalysts to solar cells and hydrogen storage, significantly accelerating materials discovery and optimization through targeted applications in specific energy systems where ML has demonstrated immense impact.

5.1. Battery Materials Simulation and Screening

DFT has been widely used to analyze battery material performance across cathodes, anodes, and electrolytes.[141,142] The anode material generally can be made of carbon/graphite/silicon anodes, while the cathode material is normally composed of lithium-containing metal oxide.[143] Recently, the applications of ML to screen high-performing lithium-ion battery materials have been extensively studied.[144–146]

5.1.1. Solid Electrolyte Materials

SGPR-based MD simulations have provided significant insights into $Li_7P_3S_{11}$, a promising solid electrolyte material.[67] Large-scale simulations ($4 \times 2 \times 2$ supercell) over 300–1200 K revealed a phase transition at ≈450 K, through P_2S_7 ditetrahedra rotations.[92] These simulations showed that material porosity and 1D channels lead to enhanced Li^+ diffusion by an order of magnitude. Calculated ionic conductivities agree well with experimental values, with activation energies matching nuclear magnetic resonance measurements.[147]

The SGPR approach was extended to systematically screen potential solid electrolytes by examining Li^+ diffusivity in hundreds of ternary crystals. For Li-X-Y ternary systems, SGPR potentials were trained using on-the-fly adaptive sampling for structures in the Materials Project database.[23] After filtering for lithium content (>10% atoms) and electronic bandgap (> 1 eV), ≈300 structures remained, among which 22 promising crystals were identified showing significant Li diffusivity.[92]

A compositional MLP approach showed particular promise for complex quaternary systems. By combining expert models trained separately on constituent ternary compounds (Li-P-S and Li-Ge-S), quaternary system properties (Li-P-Ge-S) were accurately predicted with minimal additional training.[92] For example, this approach successfully predicted Li^+ diffusion behavior in $Li_{10}GeP_2S_{12}$, achieving prediction accuracy ($R^2 = 0.961$) comparable to specialized models.[92,147] The activation barrier for lithium diffusion in $Li_{10}GeP_2S_{12}$ was significantly lower than parent ternary systems (Li_3PS_4, Li_4GeS_4), resulting in substantially higher room-temperature ionic conductivities.[92]

For solid-state lithium-ion conductors, unsupervised ML methods were applied to prioritize candidate lists from various Li-containing materials, discovering 16 new fast Li-conductors.[148] By using a recommender system coupled with RF classification algorithm, $Li_6Ge_2P_4O_{17}$ was found as a novel lithium-ion conductor for solid-state electrolyte battery.[149] To investigate the application potential of non-flammable Li-conducting ceramics as solid electrolytes, data-driven methods for screening materials using Bayesian optimization were proposed to efficiently process data and improve searching efficiency.[150] An automated simulation optimization framework was developed to design new solid polymer electrolytes, where the materials design process started from a discrete conventional design space and transferred to a continuous coarse graining design space by simulation and iterative exploration, with Bayesian optimization applied to obtain optimal materials design output.[151]

5.1.2. Cathode Materials and Doping Effects

ML approaches have been valuable for understanding dopant effects on cathode performance. The impact of Al-doping on high-capacity Li-ion battery cathode materials containing 4*d*-elements was investigated using SGPR.[152] In lithium-rich layered oxide cathode materials (LRO: Li_2RuO_3), stability was improved by Ni-doping (LRNO) and further enhanced through additional Al-doping (Al-LRNO). SGPR-based MLPs, combined with spin-polarized DFT calculations, identified optimal cation arrangements, achieving MAE < 0.1 eV for adsorption energies. This approach revealed that in LRNO, Ni tends to form clusters due to Ni-Ru immiscibility, while in Al-LRNO, Ni and Al form alloy-type clusters due to favorable Al-Ni mixing.[152,153]

Recent ML progress has improved electrode materials by predicting battery performance. ML-driven simulations provided atomic-level insights into the mechanisms of performance enhancement in battery materials.[154] Artificial NNs (ANNs) were employed to predict formation energies and volume changes of lithium intercalation compounds, identifying promising candidates with high capacity and minimal volume expansion.[155] By applying ML tools to discharge voltage data from just the first 100 cycles, Severson et al. successfully predicted the full cycle life of commercial Li-ion batteries long before significant degradation occurred, achieving a low quantitative test error of 9.1%.[156] To address the challenge of accurately predicting battery health, a physics-informed NN was used to model battery degradation, achieving a low mean absolute percentage error of 0.87% across a comprehensive dataset of 387 batteries.[157] To forecast the remaining useful life of batteries from early-cycle data, a DNN framework with memory features was introduced, which more than halved the RMSE compared to models without memory features (the optimized framework achieved an RMSE of 6.6%, and the DNN model demonstrated a prediction accuracy of 92.1% for test data).[158]

5.2. Electrocatalyst Design and Optimization

ML approaches have accelerated efficient electrocatalyst discovery for various reactions through: i) Predicting adsorption energies and free energy changes via ML models, often combined

with high-throughput screening, rapidly estimating thermodynamic properties requiring expensive DFT calculations;[123,159,160] ii) Identifying relationships between coordination environments and catalytic activity;[161–163] and iii) Optimizing ligand effects for enhanced reactant activation.[164]

5.2.1. ML-Driven Catalyst Property Prediction and Screening

ML has increasingly been recognized as a powerful asset in the study of adsorption phenomena and catalytic reaction mechanisms. For instance, support vector regression (SVR) was used to estimate hydrogen adsorption energy. Surprisingly, even simple models using a few features performed well.[165] Among various algorithms tested for predictive models utilizing molecular structures and thermodynamic parameters as input features to estimate activation energies in gas-phase reactions, tree boosting methods were found to exhibit superior predictive capability.[166] Similarly, for methane-related adsorption on copper-based alloys, the extra trees regression model outperformed other algorithms in forecasting adsorption energies with high precision.[167] In another effort, researchers applied an ANN-driven chemisorption model to assess surface reactivity trends of metal alloys across chemically diverse spaces.[38]

For water splitting, ML-based screening has been widely applied to evaluate adsorption free energies of diverse single atom catalysts on 2D materials including graphene[123] and MXenes,[168,169] gauging their efficacy for hydrogen evolution reaction (HER), oxygen evolution reaction (OER), and oxygen reduction reaction (ORR) with screening accuracy reaching over 90% for identifying compounds with optimal activity.[123,165–169]

SGPR-based potentials have proven particularly valuable for accelerating single atom catalyst (SAC) screening. The on-the-fly adaptive sampling approach allows efficient exploration of catalyst configurations while ensuring quantum-mechanical accuracy. For HER catalysts, SGPR-based ML high-throughput screening identified promising transition metal SACs embedded in graphene, achieving high prediction accuracy while drastically reducing computational costs compared to traditional DFT-based screening.[123]

For the chlorine evolution reaction (ChER), ML-driven high-throughput screening identified efficient atomic electrocatalysts.[170] This approach combined DFT calculations with MLS and MLP models to predict ChER activity across various transition metal catalysts. The study revealed that *d*-band center positions and ligand effects are critical descriptors for ChER activity. The ML models achieved remarkable prediction accuracy (mean absolute error (MAE) < 0.1 eV) for adsorption energies, allowing efficient identification of promising catalyst candidates without exhaustive DFT calculations.

For CO_2 reduction (CO_2RR) catalysts, ML approaches have focused on predicting activity and product selectivity across diverse material classes.[171,172] Active learning approaches for efficient bimetallic alloy exploration have been demonstrated, toward optimal composition search.[173] Such approaches identified promising candidates with high activity and selectivity toward valuable products, significantly reducing computational resources required for catalyst screening. An integrated approach combining ANN, quantum mechanical simulations, and multi-scale modeling was implemented to identify active surface sites on catalyst nanoparticles[174] and predict surface reactions of metal alloys, which provided actionable insights for designing efficient CO_2 conversion catalysts. A mix of quantum simulations, NNs, and advanced modeling to locate the most active spots on catalyst particles allowed for better CO_2 reduction catalyst design.[175]

For nitrogen reduction reaction (NRR), deep NN models have been applied to screen effective electrocatalysts.[176–180] Studies on doped single atom catalysts achieved highly accurate prediction of adsorption energies and free energy barriers, identifying CrB_3C_1 as a promising catalyst with a minimal overpotential of 0.13 V for NRR.[176]

5.2.2. ML's Role in Breaking Scaling Relations

Breaking scaling relations is a profound insight that deserves emphasis as an overarching theme for the materials research field. Traditional computational chemistry often relies on linear scaling laws to simplify the search space (e.g., correlating catalytic activity with a single descriptor like the *d*-band center). However, these laws also represent fundamental limitations that constrain catalytic performance, creating trade-offs where improving one property inevitably worsens another. The ability of ML models to discover and exploit complex, non-linear patterns is a core reason they are so transformative. This is not merely about achieving computational efficiency; it is about fundamentally changing the way scientists approach discovery, moving from a search within known paradigms to a principled exploration beyond them. Here, we address simple examples for catalytic reactions.

Linear scaling relationships between adsorption energies of reaction intermediates often limit the catalytic performance of materials, creating fundamental trade-offs where improving binding for one intermediate inevitably worsens binding for another. ML approaches enable the identification and design of catalysts that break scaling relationships, allowing greater control over catalytic activity and selectivity. Traditional descriptor-based approaches are inherently limited by linear scaling relations, but ML models excel at discovering complex, non-linear patterns in high-dimensional catalyst property spaces. SVR and ANN models have successfully identified catalyst configurations that deviate significantly from expected linear scaling behavior, revealing opportunities for performance enhancement beyond conventional limits.[181] Tree boosting methods have demonstrated particular superiority in capturing non-linear relationships between molecular structures and activation energies, successfully identifying catalytic systems where activation barriers deviate from scaling relation predictions.

Single atom catalysts represent a particularly promising class for scaling relation manipulation due to their unique coordination environments that differ fundamentally from extended metal surfaces. Tang et al. demonstrated that SACs sandwiched between boron nitride and graphene sheets can achieve favorable binding energies for NRR intermediates that violate conventional scaling relationships.[180] ML models helped pinpoint the key electronic factors, specifically d-orbital hybridization with substrate π-electrons, that enable this enhanced performance, providing design principles for scaling relation engineering.[181]

For CO_2 reduction, ML-guided identification of scaling relation violations has enabled enhanced selectivity control toward valuable C_{2+} products. Active learning approaches using GPR models have identified bimetallic alloy compositions where CO^* intermediates exhibit independent binding behavior, breaking traditional scaling constraints.[182] These "scaling relation hotspots" on specific surface sites and alloy compositions represent design opportunities that were previously inaccessible through conventional optimization approaches.

The development of closed-loop research paradigms incorporating ML into every stage of catalyst development, from automated synthesis and characterization to theory-driven computation and data-driven prediction, represents the future of scaling relation engineering.[182,183] These iterative frameworks establish self-refining ecosystems that continuously discover new scaling relation violations and translate them into practical catalyst designs, transforming scaling relation limitations from fundamental constraints into design opportunities for next-generation electrocatalysts.

5.3. Solar Cell Materials

For perovskites with formula ABX_3, the bandgap is a critical parameter determining light-harvesting capability and device performance. ML methods have been applied to predict bandgaps from elemental properties. The ACE method was compared with other ML methods, including decision trees, kernel ridge regression (KRR), extremely randomized trees, AdaBoost, and gradient boosting. In moderate-sized datasets, KRR and extra trees have shown excellent bandgap prediction performance. ACE provided valuable insights through non-linear relationships.[62] For very large or structured datasets, deep or graph-based models may provide greater accuracy.

Recent studies have leveraged ML techniques to screen potential photovoltaic materials and predict key properties. ML models have been developed to predict perovskite structure stability. A decision tree based data screening method combined with 1D tolerance factor showed that the proposed ML framework can accurately identify 92% of compounds in 576 ABX_3 materials.[184] Based on known crystal structure information, classification models predicting new perovskite halides using SVM algorithms have been established, discovering several new ABX_3 compositions with perovskite crystal structure.[185]

Integration of ML with first-principles calculations has been particularly valuable for designing efficient interfaces in perovskite solar cells. SGPR-based MLPs have enabled large-scale simulations of complex interfaces between SnO_2 electron transfer layers and perovskite absorbers, providing atomic-level insights into interfacial stability and charge transport mechanisms.[186,187] These studies showed that when Cl-containing precursors are used, the SnO_2/perovskite interface spontaneously forms a crystalline $FASnCl_3$ interlayer bonding coherently to both layers. This interlayer facilitates efficient charge extraction while reducing interfacial recombination, explaining exceptional device performance. ML-enabled simulations also explained why TiO_2, despite being chemically similar to SnO_2, forms less stable interfaces with perovskites. The difference arises from the ability of Sn to form intermediate oxidation states that bridge the chemical environments of the oxide and perovskite layers. These insights have guided experimental strategies for interface engineering, contributing to record efficiency (25.8% at the time of the cited research) in perovskite solar cells.[187]

Recent advances in ML for perovskite materials include methods to identify promising halide perovskites for photovoltaic applications by leveraging high-quality bandgap datasets. Wang et al. developed a strategy that combines advanced computational techniques with ML to search for optimal perovskite compositions, identifying 14 new materials for solar cells.[188] A recent review discusses ML approaches for perovskite solar cells in depth.[189]

For organic photovoltaics, deep learning architectures have been used to estimate the highest occupied molecular orbital (HOMO) levels of organic semiconductors, showcasing the efficiency of transfer learning techniques in screening organic solar cells across diverse molecular datasets.[190] RF algorithms applied to conjugated polymers for polymer-fullerene solar cells achieved top-tier performance in predicting photovoltaic suitability.[191] A major step forward has been the experimental realization of ML-forecasted materials. Supervised models trained to correlate chemical structure with optoelectronic performance guided the synthesis of ten novel molecular donors, all of which demonstrated experimental results consistent with ML-based predictions.[192]

5.4. Phase Change Memory Materials

MLPs have been applied extensively to investigate phase-change memory (PCM) materials, crucial for non-volatile memory technologies. NN potentials have been created for GeTe, accurately reproducing crystalline, liquid, and amorphous properties.[97] The approach was successfully applied to explore the material's structure and phase transitions in simulations with thousands of atoms, inaccessible to direct DFT calculations.

Using the GAP framework, MLPs have been developed for the ternary PCM compound $Ge_2Sb_2Te_5$,[97] enabling creation of detailed 7200-atom models. This provided remarkable insights into material local structure and facilitated smaller models for in-depth chemical bonding studies. This approach was later expanded to describe various Sb-Te alloys,[193,194] revealing composition-dependent structural motifs and crystallization behaviors.

For exploring PCM behaviors in realistic memory device geometries and conditions, MLPs trained with quantum-mechanical data simulated all Ge-Sb-Te compositions used in PCMs.[195] These significantly enhanced the speed and accuracy of atomistic simulations, supporting simulations of multiple thermal cycles and operations crucial for neuromorphic computing. Large-scale device models containing over half a million atoms demonstrate the capability to accurately depict critical processes in PCM-based memory devices, enabling device-scale PCM simulations (10^5–10^6 atoms) with ab-initio-level local structure/energy fidelity.

5.5. Hydrogen Storage and CO_2 Capture Materials

Large-scale simulations of hydrogen adsorption and diffusion in nanoporous materials have helped identify promising storage candidates. Crystal graph convolutional neural network (CGCNN) and related architectures have been used to screen metal–organic frameworks (MOFs) and covalent organic frameworks (COFs) for hydrogen storage,[196] predicting key properties such as binding energies, volumetric capacities, and kinetic barriers with near-DFT accuracy at reduced computational cost. ML approaches can predict quantum-chemical properties of MOFs for accelerated materials discovery, such as adsorption energies and diffusion barriers in complex porous materials.[197] This has enabled rapid screening of thousands of MOF structures for hydrogen storage, identifying several promising candidates with high gravimetric and volumetric capacities.

MOFs have garnered significant interest as promising candidates for physical CO_2 adsorption, due to highly tunable architectures and porosity. A comprehensive multiscale approach combining DFT, grand canonical Monte Carlo (GCMC) simulations, and ML techniques evaluated how pore surface chemistry and structural topology contribute to CO_2 uptake.[198] ML algorithms combined with multivariate statistical analysis and first-principles calculations evaluated 2932 MOF structures for electronic conductivity.[199] A ML-based structure-property classifier capable of rapidly assessing CO_2 adsorption potential achieved a tenfold reduction in computational demand without compromising accuracy.[200]

To determine the factors most critical for CO_2 adsorption, a predictive model using the RF algorithm was developed within a quantitative structure-attribute relationship (QSAR) framework.[201] This finding emphasized pressure as a dominant factor, along with three other key structural parameters, offering practical insights for guiding material selection and optimization. An ML-accelerated high-throughput screening methodology pinpointed COFs with superior CO_2 capture capability, with the RF approach emerging as highly effective for modeling complex adsorption behavior.[202]

6. Advanced Screening Strategies and Workflows

6.1. Active Learning and Bayesian Optimization

Active learning has emerged as a powerful strategy for efficient exploration of vast materials spaces.[101,203] By iteratively selecting the most informative samples for evaluation, active learning approaches can dramatically reduce computational or experimental resources required for materials discovery.[204] **Figure 4** illustrates the typical active learning workflow that has become central to modern materials discovery, showing how initial models trained on limited data iteratively improve through strategic selection of new data points that maximize information gain.

Bayesian optimization, a specific form of active learning, has been particularly successful in materials screening applications.[102] Bayesian optimization can efficiently identify optimal compositions of complex oxides for energy applications, using acquisition functions that balance exploration and exploitation of the design space.[205] Similar approaches have been employed to optimize organic photovoltaic material performance,

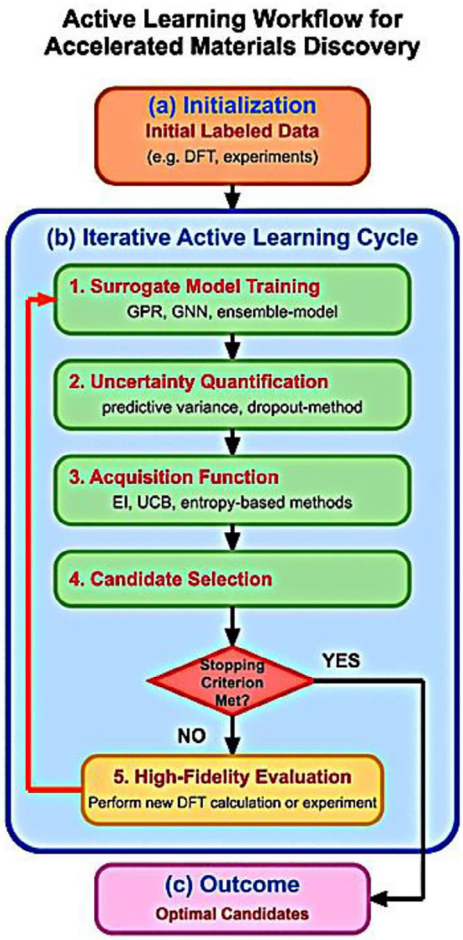

Figure 4. Active Learning Workflow for Accelerated Materials Discovery (see Table 5). The iterative cycle begins with an initial model trained on limited data, which is then used to predict properties across a large search space. The acquisition function identifies the most informative samples for evaluation, balancing exploration of unknown regions with exploitation of promising areas. New data points are selectively acquired to maximize information gain, and the model is retrained with expanded data. This process continues until convergence or satisfactory performance is achieved, dramatically reducing expensive calculations or experiments required compared to traditional screening approaches. Without a reliable way to quantify uncertainty, the acquisition function has no principled way to select the next most informative data point. The deep, causal link is that UQ directly leads to superior data efficiency in exploratory, low-data regimes. The model's awareness of where it is least certain about regions of chemical space is what enables the strategic, information-maximizing selection of new data points.

achieving significant improvements in power conversion efficiency with minimal experimental trials.[206]

Table 5 provides a detailed comparison of the three main families of acquisition functions used in active learning for materials discovery, each offering different approaches to the fundamental exploration-exploitation tradeoff that is central to efficient screening.

Integration of active learning with high-throughput computational workflows has given rise to autonomous materials discovery frameworks.[207] Autonomous materials search systems

Table 5. Comparison of acquisition functions for active learning in materials discovery.

Feature	Entropy-based methods (Uncertainty Sampling)	UCB (Upper confidence bound)	EI (Expected improvement)
•*Core Idea*	Query max $\sigma(x)$ $\sigma(x)$: prediction uncertainty (max entropy). BALD (Bayesian Active Learning by Disagreement)	Query max UCB(x)= $\mu(x) + \kappa\ \sigma(x)$. μ: mean prediction: exploitation σ: uncertainty (standard deviation) κ: exploration hyperparameter	Query max EI(x) = $\int(y\text{-}f^*)\varphi(y)dy$, the point most likely to be better than the best we've seen so far. f^*: current best value $\varphi(y)$: normal distribution
•*Main*	Classification	Classification & Regression	Regression (Bayesian Optim.)
•*Balances*	Primarily focuses on model confusion, often leading to exploiting decision boundaries	Explicitly balances exploration and exploitation via hyperparameter	Balances exploration and exploitation, aiming to improve over the current best
•*Model Needs*	A model that outputs class probabilities (e.g., Logistic Regression, Softmax Net)	A model that provides both mean and uncertainty (e.g., Gaussian Process, BNN)	A model that provides both mean and uncertainty (e.g., Gaussian Process)
•*Pros*	Intuitive, simple, refining decision boundaries.	Principled trade-off, flexible, strong theoretical guarantees.	Very sample-efficient for finding optima.
•*Cons*	May ignore unexplored regions	Requires tuning hyperparameters	Less intuitive for classification; may ignore model improvement

combine ML, DFT calculations, and Bayesian optimization for self-driving discovery of functional materials.[208] Such systems demonstrate the ability to navigate complex compositional spaces and identify promising candidates with minimal human intervention.

SGPR-based approaches have integrated active learning through built-in UQ.[68–70,91,92] The spilling factor $s(\rho) = K(\rho,\rho) - K_{\rho m}\ K_{mm}{}^{-1}\ K_{\rho m}{}^{T}$ quantifies predictive uncertainty for given local chemical environments, providing natural acquisition functions for active learning. This uncertainty-driven sampling ensures that new DFT calculations are performed precisely where they provide maximum information gain. On-the-fly adaptive sampling algorithms[68,69] have demonstrated remarkable efficiency in building accurate potentials with minimal training data, achieving ≈90% reduction in required experiments.

6.2. Multi-Fidelity and Transfer Learning

Multi-fidelity learning approaches have proven valuable for materials screening by leveraging data from multiple sources with varying accuracy and computational cost.[209,210] These methods enable efficient computational resource allocation by performing expensive high-fidelity calculations only when necessary, while using lower-fidelity approximations for initial screening.[102]

Multi-fidelity GPR has been demonstrated to accurately predict bandgaps using DFT calculations at different theory levels, achieving significant speedups compared to relying solely on high accuracy calculations.[211] Similarly, multi-fidelity NNs have been employed to effectively transfer the information from large datasets of semi-empirical calculations to improve predictions for more accurate but limited DFT data.[212] Multi-fidelity graph networks to accurately predict material properties even with limited high-fidelity data by incorporating lower-fidelity computational data, enhance the model's understanding of structural features. This approach significantly reduces the error in predicting experimental bandgaps and provides a method for modeling disorder in materials through learned elemental embeddings, a key advancement in computational materials science.[213] Transfer learning offers another approach for leveraging knowledge across different domains or property prediction tasks.[214]

6.3. Multi-Objective Optimization and Pareto Frontiers

Energy materials often require optimization across multiple competing objectives, necessitating approaches that identify optimal trade-offs rather than single best solutions.[215] Multi-objective optimization frameworks combined with ML have enabled efficient exploration of Pareto frontiers, representing solution sets where no objective can be improved without degrading another.[216]

Multi-objective Bayesian optimization has been employed to discover transition metal complexes with optimal spin-splitting energies and redox potentials for spin-crossover applications. By explicitly modeling trade-offs between these properties, such approaches identify promising candidates that would be overlooked by single-objective optimization. For battery materials, multi-objective active learning frameworks have been developed to optimize fast-charging protocols for lithium-ion batteries, balancing charging speed against capacity degradation. This approach effectively optimized fast-charging protocols while reducing experimental costs by ≈90%, demonstrating value for multi-objective optimization for both operational protocols and materials design.[217]

7. Key Advantages of ML Approaches for Energy Materials

7.1. Computational Efficiency and Quantum-Mechanical Accuracy

MLPs have emerged as critical bridges between highly accurate but computationally intensive DFT and empirically derived

classical potentials.[5,16,17,67] This computational acceleration enables: i) Simulations of energy materials systems with thousands of atoms with accuracy depending on training data similarity; ii) Extended time scales necessary for observing ion diffusion, phase transitions, and reaction phenomena, though long-timescale accuracy requires careful validation; iii) Comprehensive screening of material candidates with different compositions and structures.

GNNs can handle large systems due to favorable scaling properties. By representing atoms as nodes and interactions as edges, GNNs inherently capture the graph-like nature of molecular and crystalline systems, allowing message-passing mechanisms to efficiently handle systems with numerous atoms.[78,79] The local nature of these interactions generally results in linear scaling with system size, though GNNs typically require substantial training data to achieve optimal performance.

For kernel-based methods like SGPR, training individual expert models typically requires minutes to hours, depending on dataset size, with ensemble approaches adding negligible overhead.[69] Despite the excellent accuracy of kernel methods, their $O(nm^2)$ scaling can become a limitation when modeling diverse chemical systems. The RBCM framework effectively addresses scaling challenges by partitioning datasets and inducing points into multiple local expert models, reducing computational complexity to $O(nm^2/p^2)$.[70] Applied to organic systems and oxygen-containing compounds relevant to energy conversion and storage, this approach maintains quantum-mechanical accuracy while enabling seamless model expansion to incorporate new chemical elements and functional groups.

7.2. Chemical Transferability and Expansion of Accessible Chemical Space

A critical advantage of modern ML approaches is their ability to transfer knowledge across chemical spaces. Graph-based NNs have shown particular promise for transferability across diverse materials. CGCNNs can directly process crystal structures, capturing periodic atomic arrangements and bonding patterns with high accuracy.[45] This transferability stems from the ability to transform non-transferable atomic configurations into transferable representations, enabling predictions across broad chemical spaces.

The compositional MLP approach offers a powerful strategy for extending chemical coverage via modular expansion.[92,218] By training expert models on specific chemical subsystems and combining them through frameworks like RBCM, researchers can systematically build coverage of complex chemical spaces without requiring exhaustive training across all possible compositions.[70,218]

This modular strategy has shown particular promise for multi-component energy materials. By combining expert models trained on lithium-sulfide and germanium-sulfide systems, quaternary lithium-germanium-sulfide solid electrolyte properties were successfully predicted with minimal additional training.[92,147] Similarly, combining expert models for different hydrocarbon families has enabled accurate modeling of complex organic systems.[70,219,220]

7.3. Targeted Improvement Through Active Learning and Uncertainty Quantification

Modern ML methodologies incorporate active learning strategies that significantly enhance the efficiency of model development. These approaches identify configurations with high uncertainty for targeted sampling, incrementally improve models with minimal additional DFT calculations, and adapt to newly encountered physical processes during simulations. This approach enables efficient computational resource use by focusing new high-level calculations where they provide the most information gain, particularly valuable for exploring complex potential energy surfaces of energy materials.[101] SGPR and related kernel-based models intrinsically provide Bayesian predictive variance, though the reliability of these uncertainty estimates depends on model assumptions and training data coverage, and can be overconfident when extrapolating beyond training distributions; standard GNNs require post-hoc methods, such as ensembling or stochastic dropout, for uncertainty estimation.[221–224]

7.4. Bridging Scales Through Multi-Scale Modeling

ML approaches enable bridging across length and time scales critical for energy materials: i) Connecting atomic-scale properties to device-level performance; ii) Combining thermodynamic predictions with kinetic models for more complete performance estimates; iii) Accounting for materials evolution under operating conditions. This multi-scale modeling capability is essential for developing comprehensive models of energy devices, where performance is determined by phenomena occurring across multiple stages.[225] By combining information from different scales into unified ML frameworks, researchers can achieve more accurate and predictive models capturing hierarchical nature of materials properties.

Almost all ML potentials are local, accounting for interactions between an atom and its neighbors up to some cutoff radius.[1] While there exist some nonlocal models, for most systems, reasonable cutoff radii enable highly accurate results without requiring explicit long-range terms. For systems where very long-range interactions play a critical role, hybrid approaches that combine MLPs with classical force fields for electrostatic and dispersion effects have proven effective.[70,218]

8. Challenges and Future Directions

Despite significant progress in applying ML to energy materials, several challenges remain to be addressed for realizing the full potential of these approaches.

8.1. Data Quality and Availability

ML-driven materials screening faces significant challenges related to data quality and availability.[1] Many material properties lack standardized experimental protocols, leading to inconsistencies in reported values across studies. Published datasets are biased toward successful materials rather than failures, creating bias that can limit the predictive capability of ML models.

Training accurate MLPs requires high-quality reference data, typically from DFT calculations. Different ML architectures have varying data requirements. While kernel-based methods like GAPs can achieve reasonable accuracy with relatively small datasets,[66] GNN-based approaches often require substantially more training data to achieve comparable results. This presents a challenge for materials in complex phase spaces or in cases where high-level quantum mechanical calculations are computationally prohibitive.

Future advances will require greater integration of data from multiple sources, including experimental measurements, computational predictions, and literature text mining.[226] Development of automated data extraction tools, standardized metadata schemas, and improved UQ methods will be crucial for building more comprehensive and reliable materials databases.[227]

8.2. Extrapolation to New Chemical Spaces

One primary limitation of current MLPs is inability to extrapolate to novel chemical spaces. While these models excel at interpolating within chemical space covered by training data, they often fail when presented with substantially different environments or bonding patterns. Different ML architectures have varying capabilities—kernel-based methods like SGPR and GAP tend to provide appropriate uncertainty estimates when predicting outside training distribution, while deterministic NNs may give overconfident but incorrect predictions.[99,100]

Equivariant architectures offer promising pathways for improving extrapolation capabilities. By explicitly encoding physical symmetries, these models can generalize more effectively to new environments with limited training data.[87,90] The strong inductive bias provided by equivariance constraints reduces the model's need to learn symmetry operations from data, improving data efficiency and generalization.

Complementary to equivariant architectures, ensemble approaches provide another pathway for systematically expanding chemical coverage. The compositional MLP approach offers a promising strategy for extending chemical coverage via modular expansion.[92,218] The RBCM framework[70] combines multiple SGPR experts, each trained on specific chemical domains, through Bayesian weighting schemes, enabling straightforward extension to new elements and functional groups. The built-in UQ in such ensemble models naturally flags regions where extrapolation may be unreliable, guiding additional high-level calculations precisely where they would provide the greatest information gain.

The modular nature of compositional approaches also creates a pathway toward increasingly universal MLPs. As computational resources grow and expert model libraries expand, compositional MLPs will play an increasingly important role in accelerating materials discovery across diverse chemical spaces.

8.3. Long-Range Interactions

Since most ML potentials are local, relying on a fixed cutoff radius to capture interactions between an atom and its neighbors, they face difficulties in accurately treating long-range interactions such as electrostatics and dispersion forces. While computationally efficient with linear scaling, this approach fundamentally neglects interactions beyond the cutoff. This limitation is particularly problematic for systems where long-range forces dominate, including ionic solids, layered materials, and polar molecules.

Several approaches have been developed to address this challenge. For kernel-based methods like SGPR, strategies include adding explicit long-range terms via analytical corrections to complement ML predictions of short-range interactions, and hierarchical modeling that trains separate ML models to capture interactions at different length scales.[70] Transformer-based architectures naturally capture global dependencies through self-attention mechanisms, with recent equivariant transformers maintaining symmetry constraints required for physical accuracy.[228] As a hybrid approach for systems with critical long-range interactions, MLPs can be combined with classical force fields that accurately capture electrostatic and dispersion effects at long distances.[70,218] This method maintains quantum-mechanical accuracy for local interactions while efficiently modeling extended physical effects.

8.3.1. Quantitative Performance Metrics

Specific computational cost and accuracy benchmarks can be assessed for each approach.

Analytical Corrections: The most computationally efficient approach, adding explicit long-range terms to complement ML predictions of short-range interactions.[70] For dispersion interactions, approaches like D3 corrections[229] can expand interaction ranges with minimal computational overhead (10–20% above base MLP cost) and accuracy improvements of 2–5 meV atom^{-1} for dispersion-dominated systems. However, this method lacks transferability for chemically heterogeneous systems and fails at interfaces, working best only for neutral, homogeneous condensed-phase materials.

Hierarchical Modeling: Co-trains separate ML models to capture interactions at different length scales, with distance scaling relationships varying from r^{-1} (Coulomb) to r^{-6} (dispersion), and system-dependent r^{-2} to r^{-5} for van der Waals in nanostructures.[230] Recent work achieved significant speedup (much faster than single-scale approaches with multiple-time-step integrators) without accuracy loss on potential energy or simulation-derived quantities[231] by using small, efficient models for short-time-scale interactions with large, expressive models for remaining interactions. Training overhead is moderate, with manageable inference overhead, though implementation complexity is higher.

Transformer-Based Architectures: Self-attention mechanisms naturally capture global dependencies but suffer from quadratic scaling limitations. Recent equivariant transformers use self-attention to model interactions at any distance without predefined cutoffs while maintaining symmetry constraints required for physical accuracy.[104] These models can learn complex, orientation-dependent interactions by treating charges as equivariant objects. While providing a theoretically comprehensive solution, quadratic scaling limits application to

Table 6. Performance metrics for computational methods of long-range interactions.

Approach	Theoretical Basis	Computational Cost	Accuracy	Ideal Application
•*Analytical Corrections*	Inverse power laws	Low (10-20% overhead)	Low-to-Medium	Homogeneous, neutral condensed-phase materials
•*Hybrid ML/Classical*	ML + Classical Force Fields (FFs)	Medium (2× overhead vs classical FF)	High	Large systems (>10*K* atoms) with polar/ionic character
•*Equivariant Transformers*	Self-attention, equivariant charges	High (quadratic scaling)	Highest	Small systems (< 1*K* atoms) with critical long-range effects
•*Charge Equilibration*	Variational charge equilibration	Medium (nearly O(*N*) scaling)	High	Medium systems (< 1*K* atoms) needing high electrostatic accuracy

smaller systems (< 1*K* atoms) where global interactions are critical.

Hybrid ML-Classical Methods: Combines MLPs for local interactions with classical force fields for long-range effects. Hybrid potentials using optimized multi-timestep integrators reduced computational overhead from 40× to only 2× relative to classical force fields.[232] The deep potential long-range approach shows excellent energy conservation with small drift (≈0.4 meV/H_2O (≈1 *K*) per 100 ps)[233] and demonstrated effectiveness on systems >10^6 atoms using exascale computing, representing a highly practical solution for large systems with significant polar or ionic character.

Charge Equilibration Methods: Fourth-generation MLPs incorporating charge equilibration (Q_{eq}) provide accurate electrostatic descriptions but traditionally require calculation of dense Coulomb matrices, resulting in quadratic scaling with respect to the number of atoms *N*.[234] Recent advances using particle mesh methods reduce complexity from O(N^3) to nearly O(*N*) while avoiding explicit computation of Coulomb matrix elements.[234] Variational charge equilibration approaches demonstrate equivalent performance while being less computationally expensive than traditional fourth-generation methods.[235] Memory requirements are reduced by half through sparse matrix storage, making them best suited for smaller systems (< 1*K* atoms) where superior electrostatic accuracy is critical.

A streamlined comparison of these approaches is presented in **Table 6**.

8.3.2. Critical Evaluation and System-Size Recommendations

Explicit recommendations can be made based on system size and computational constraints.

1) *Small Systems (< 1K atoms)*: Transformer-based or full charge equilibration methods optimal[236]
2) *Medium Systems (1K–10K atoms)*: Hybrid approaches show best accuracy-efficiency trade-offs, with particular success for polar materials such as electrolytes.[237] Modern implementations enable million-atom polarizable force field (PFF) simulations.[232]
3) *Large Systems (>10K atoms)*: Analytical corrections remain most practical, with careful validation required.[230]

8.3.3. Fundamental Limitations

We now explicitly discuss the inherent trade-offs:

1) *Accuracy* versus *Efficiency Dilemma*: While rapid expansion of MLPs is beginning to overcome the classically pervasive accuracy-efficiency trade-off, fundamental limitations persist.
2) *Scalability Constraints*: Attention mechanisms, while theoretically capable of capturing all long-range interactions, suffer from quadratic complexity that makes them computationally prohibitive for realistic system sizes.
3) *Transferability Issues*: Simple analytical corrections work well for homogeneous systems but will ultimately lead to incorrect predictions for systems where dispersion contributions beyond the local atomic environment vectors cannot be neglected.

8.3.4. Future Directions

Several promising research directions are emerging:

1) *Emerging Architectures*: Sum-of-GNNs (SOG-Net) approach showing promise for maintaining "close-to-linear computational complexity during training and simulation" while adaptively capturing "diverse long-range decay behaviors"[238]
2) *Multi-Scale Integration*: Co-training approaches that separate fast and slow degrees of freedom, showing significant computational advantages
3) *Hardware-Optimized Methods*: Adaptation of long-range methods for modern GPU architectures and distributed computing environments

This field continues evolving toward more sophisticated approaches that balance the fundamental accuracy-efficiency trade-off, with method selection requiring careful consideration of system characteristics, computational resources, and accuracy requirements for specific applications.

8.4. Interpretability and Explainability

As ML models for materials screening become increasingly complex, ensuring interpretability becomes both more challenging and important.[227] Black-box models may achieve high predictive accuracy but provide limited scientific insight into the underlying physical principles governing materials behavior.[239]

8.4.1. Architecture-Dependent Approaches

Different ML architectures offer varying degrees of interpretability. Kernel-based methods like GAP and SGPR provide direct connections between training data and predictions through kernel functions,[94,240] making it straightforward to identify which reference configurations influence particular predictions.[3] DNNs typically offer less transparent relationships, though recent advances in explainable AI are addressing this limitation.[241]

8.4.2. Attention Mechanisms in GNNs

Recent efforts have focused on developing more interpretable ML approaches for materials screening.[73] Attention mechanisms in GNNs can reveal which atomic interactions contribute most significantly to predicted properties.[242] Recent studies demonstrate that attention weights identify important substructures critical for materials to achieve desired properties.[243] Crystal graph attention networks have shown particular promise in highlighting specific atomic environments and bonding patterns that correlate with experimental property trends.[244]

8.4.3. SHAP Analysis for Feature Importance

Shapley additive explanations (SHAP) values provide quantitative measures of feature importance that reveal structure-property relationships.[245] Applications in energy materials include catalyst screening, where SHAP analysis reveals the relative contributions of electronic and geometric descriptors to predicted activity, and battery materials research for identifying features influencing capacity and cycling stability. In perovskite solar cells, SHAP analysis identified dominant molecular descriptors affecting degradation, leading to successful experimental validation.

8.4.4. Gradient-Based Attribution Methods

Gradient-based techniques like integrated gradients quantify how material property changes influence model predictions, providing optimization guidance.[246] These methods identify which material features most strongly influence predicted properties, helping researchers understand which material characteristics drive performance, for example, in battery materials and electrocatalyst discovery,[199] though they face challenges with discontinuous changes and high-dimensional descriptor spaces common in materials systems.

8.4.5. Model-Agnostic Methods and Future Directions

Local interpretable model-agnostic explanations (LIME) and other model-agnostic methods provide local explanations applicable to any ML architecture.[245] Key challenges remain including high-dimensional materials data, complex feature interactions, and distinguishing correlation from causation. Promising directions include counterfactual explanation methods and integration of physics-based knowledge with interpretability techniques. The aim is to develop interpretable ML approaches that not only achieve high predictive accuracy but also yield insights to guide experimental design and improve understanding of materials behavior.

8.4.6. Physical Insights

Perhaps most promising is integration of physics-based knowledge with data-driven approaches, creating hybrid models that combine interpretability of first-principles methods with ML efficiency.[247] Physics-informed NNs, which encode known physical laws and constraints directly into model architectures, represent one approach to achieving this balance. Similarly, symbolic regression techniques aim to discover analytical expressions that describe material properties in terms of physically meaningful descriptors.[248]

8.5. Physics-Informed Machine Learning

Incorporating physical knowledge into ML frameworks improves prediction accuracy and interpretability, addressing limitations of purely data-driven models that struggle with out-of-distribution inputs or generating physically realistic predictions when data is scarce. Primary approaches for physics-informed ML are as follows: i) *Physics-constrained NNs*: Architectures that explicitly enforce physical laws and constraints can provide more reliable predictions;[247] ii) *Hybrid models*: Combining ML with physics-based simulations to leverage the strengths of both approaches; iii) *Theory-guided data science*: Using theoretical insights to guide feature selection and model architecture for improved performance and interpretability.[249] By encoding conservation laws and thermodynamic consistency, these approaches reduce reliance on large datasets while improving generalization and interpretability. Physics-informed ML ensures predictions remain physically meaningful and provides actionable insights for materials design. By balancing data-driven learning with physical constraints, these methods guide experimental efforts more effectively while providing a fundamental understanding of structure-property relationships.

8.6. Integration with Experimental Workflows

The integration of computational ML approaches with experimental materials synthesis and characterization represents a frontier with enormous potential. Self-driving laboratories for materials discovery combine ML with automated synthesis and characterization to accelerate the identification of promising materials.[250] These closed-loop systems iterate between prediction, synthesis, testing, and model refinement to rapidly optimize material properties.[251]

Key developments in this area include: i) Real-time ML guidance to adapt experimental parameters on-the-fly based on incoming data;[252] ii) Automated high-throughput characterization techniques coupled with ML analysis to provide rapid feedback;[253] iii) Multimodal ML approaches that integrate information from multiple experimental techniques, including spectroscopy, diffraction, and imaging, to provide more comprehensive understanding of materials. These approaches promise to

dramatically reduce the time and resources required for materials discovery by intelligently navigating the vast design space of possible compositions, structures, and processing conditions.

New strategies are accelerating materials discovery by efficiently exploring large design spaces of compositions and synthesis methods. Despite the value of experimental data, its collection remains slow and often yields sparse datasets. With advances in automation and 3D printing, autonomous laboratories are becoming feasible. Recently, smart robots have been developed that perform high-throughput experiments, enabling rapid search for improved material compositions.[254]

It is worthwhile to address some promising cases for ML-based materials discovery. Cost-benefit would show a drastic reduction in development cycles with 10-100× acceleration compared to traditional approaches. In practice, ML-driven workflows, coupled with automated synthesis and high-throughput testing, have demonstrably shortened the discovery-to-prototype cycle for certain classes of materials. For example, the Microsoft-PNNL collaboration demonstrated ML-guided solid electrolyte discovery that identified 18 promising candidates from 32 million materials in just 80 h-a process that would traditionally require 20 years-with final material synthesis and validation completed within 9 months.[255] In heterogeneous catalysis, closed-loop optimization platforms using Bayesian optimization have identified optimal catalyst compositions for CO_2-to-methanol conversion within 6 weeks across 144 catalysts, representing order-of-magnitude reductions in experimental campaigns compared to traditional decade-long research cycles.[256] Automated catalyst optimization has achieved experimental time cost reductions of ≈60-fold compared to manual operations, while active learning workflows for CO_2 reduction catalysts have identified optimal Cu-Al compositions through iterative cycles that achieved over 80% Faraday efficiency compared to ≈66% for pure Cu.[257] For battery electrode materials, ML approaches have accelerated the screening and identification of thousands of potential electrode materials for Na/K-ion batteries with voltages rivaling their Li-ion counterparts, reducing traditional trial-and-error approaches from multi-year timescales to months.[258] These examples illustrate the tangible potential of digital and autonomous R&D pipelines to accelerate materials innovation in practice, even though the degree of acceleration remains highly dependent on the specific material system and integration level of automation.

Step-by-step protocols for ML-guided synthesis include: 1) Initial screening using ML models to identify promising candidates; 2) Automated synthesis parameter optimization using Bayesian optimization; 3) Real-time characterization with ML-guided analysis; 4) Feedback loops for iterative refinement; 5) Integration with existing laboratory infrastructure through modular design approaches.

8.6.1. A-Lab's Autonomous Materials Discovery Platform

Ceder and coworkers reported the A-Lab,[259] an autonomous laboratory for the solid-state synthesis of inorganic powders. While the platform demonstrates a high success rate in synthesizing targeted materials, the initial selection of these targets is guided by large-scale ab initio phase-stability data from databases like the Materials Project. The platform integrates robotics with computational tools, literature data, and ML to plan and execute experiments.

1) *Target Identification*: The A-Lab identifies synthesis targets using large-scale ab initio phase-stability data from databases like Materials Project and Google DeepMind. These calculations identified 58 targets that were predicted to be stable or near-stable.
2) *Recipe Generation*: The A-Lab generates initial synthesis recipes for each target using ML models trained on synthesis data from literature. For each compound, up to five initial recipes are proposed based on chemical similarity.
3) *Performance*: Over 17 days of continuous operation, the A-Lab successfully synthesized 41 of the 58 targets, representing a 71% success rate for solid-state inorganic synthesis. Success rates vary significantly across different material classes and synthesis methods. The A-Lab's active-learning cycle, which proposes improved recipes when initial ones fail, was able to optimize synthesis for nine targets, six of which had a zero yield initially. However, only 37% of the 355 synthesis recipes tested by the A-Lab produced the desired targets.

8.6.2. Actionable Implementation Pathways

Level 1-Computational Pre-screening Integration: Experimentalists can immediately implement ML-guided target identification using existing computational resources. The A-Lab methodology demonstrates systematic selection of synthesis targets using DFT-calculated phase stability data from databases like Materials Project.[259] This approach identified 58 viable targets from millions of possible compositions, focusing experimental efforts on thermodynamically favored compounds.

Level 2-Semi-Automated Synthesis Optimization: Integration of Bayesian optimization with existing laboratory equipment enables systematic parameter optimization without full automation. Demonstrated protocols include: 1) DoE-guided screening of temperature, time, and precursor ratios; 2) Real-time adjustment of synthesis parameters based on intermediate characterization; 3) Active learning algorithms that minimize required experiments while maximizing information gain.

Level 3-Closed-Loop Autonomous Systems: Full autonomy requires integration of robotic synthesis, automated characterization, and ML decision-making. Successful implementations like the A-Lab achieved 71 (41 of 58 targets) over 17 continuous days of operation.[259]

8.6.3. ML-Guided Synthesis Protocols

Protocol 1: Literature-Informed Recipe Generation: target selection, recipe generation, temperature prediction, initial synthesis, yield assessment

Protocol 2: Active Learning Optimization: thermodynamic analysis, pathway prediction, adaptive experimentation, and iterative refinement for failed attempts

Protocol 3: Multi-Modal Characterization Integration: real-time X-ray diffraction analysis, spectroscopic feedback, morphological assessment, property validation

8.6.4. Scalability Analysis: Throughput Scalability

Manual synthesis: 1-5 samples per day

Level 1 (Computational Integration): Standard laboratory computing infrastructure: 5–20 samples per day

Level 2 (Semi-Automated): Automated synthesis equipment (heating, mixing, dosing): Timeline: 20–100 samples per day

Level 3 (Full Autonomy): Comprehensive robotic laboratory infrastructure: 50–200 samples per day (demonstrated by A-Lab)[259]

Quantified Benefits from Deployed Systems:

1) *A-Lab Performance*: 71% synthesis success rate versus 20–30% for traditional approaches in solid-state inorganic synthesis[259]
2) *Time Acceleration*: 10–100× faster than specific manual synthesis-characterization workflows,[260] (actual speedup varies significantly by synthesis complexity and characterization requirements)
3) *Performance*: A mobile robot to autonomously search for improved photocatalysts for hydrogen production, with 1000 times faster autonomous workflow than manual methods and at least ten times faster than semi-automated robotic workflows. The robot also identified a photocatalyst mixture that was six times more active than the initial formulations.[252]
4) *Labor Optimization*: 90% reduction in routine experimental tasks, freeing researchers for strategic planning[261]

8.6.5. Open-Source ML Tools and Databases

In this section, open-source ML tools and databases for materials are listed in **Table 7**.

8.6.6. Future Directions and Emerging Technologies

Large Language Model (LLM) Integration: Recent developments in LLM-based reaction development frameworks demonstrate automated experimental design and natural language interaction with laboratory systems.[257,262] These systems eliminate coding requirements for experimentalists while maintaining sophisticated optimization capabilities. LLMs can be fine-tuned on synthesis protocol datasets to generate procedures for diverse materials, significantly streamlining discovery by reducing trial-and-error approaches. They also enable natural language interfaces for laboratory automation, democratizing access to advanced instrumentation and freeing researchers to focus on strategic planning. However, LLMs face significant challenges when applied to scientific problems. Models trained solely on text can struggle with physical laws governing chemical reactions, lacking explicit constraints to conserve atoms or electrons, potentially leading to nonsensical "alchemical" results. This highlights the necessity of integrating physical knowledge with data-driven approaches. Physics-informed LLMs could explicitly track electrons in reactions to ensure that generated outputs are both linguistically plausible and physically meaningful.

Multi-Laboratory Networks: Emerging frameworks enable distributed autonomous experimentation across multiple institutions, sharing synthesis strategies and expanding accessible chemical space through collaborative autonomous research networks.[263]

The convergence of ML algorithms, robotic automation, and systematic data management represents a fundamental transformation of materials discovery methodology. By implementing these tiered approaches, experimentalists can immediately begin realizing benefits while building toward fully autonomous discovery platforms that promise to accelerate materials innovation by orders of magnitude.

9. Materials Digital Twins for Energy Applications

The integration of MLPs with multiscale modeling enables the development of MDTs -dynamic computational models that replicate materials behavior across scales through real-time, bidirectional coupling with physical systems. Unlike conventional static multiscale models, MDTs incorporate continuous learning and are updated with operational data, providing adaptive, data-driven intelligence. Traditional multiscale modeling remains invaluable for hierarchical, physics-based understanding, but MDTs extend this paradigm by embedding continuous feedback loops between experiment and computation. The key distinction lies in their bidirectional data flow and adaptive refinement, where experimental feedback perpetually improves model fidelity, in contrast to the fixed predictive hierarchies of classical approaches.

The future lies in combining both paradigms to create intelligent, adaptive materials systems that leverage deep physical understanding with real-world operational data. **Figure 5** illustrates the conceptual framework for such systems, showing how atomic-scale phenomena connect to device-level performance through hierarchical modeling approaches that span quantum, mesoscale, and macroscale domains.

These MDTs enable comprehensive visualization and prediction of material behavior under various operating conditions, facilitating materials property optimization before experimental synthesis. Several categories can be defined: imaginary, monitoring, predictive, prescriptive, autonomous, and recollection MDTs, enabling i) structure generation and property prediction prior to experimental synthesis; ii) visualization of dynamic processes such as ion diffusion, phase transitions, and reactions; iii) design of high-performance materials for batteries, catalysts, solar cells, and other energy applications.

Industrial implementations of battery MDTs have demonstrated significant practical impact. Recent pilot studies show that MDT-enabled battery management systems can achieve real-time monitoring capabilities, enabling predictive maintenance strategies.[264,265] For example, comprehensive battery MDTs incorporating electrochemical-thermal-mechanical coupling have been successfully deployed for grid-scale energy storage systems, enabling asset-specific optimal decisions and degradation analysis.[266] MDT-driven specific capacity predictions showed a low average deviation at different C-rates.[265] The MDT aggregated data such as paste viscosity, coating thickness, and material homogeneity, generated a live prediction of the battery cell quality[267] and provided quality indicators and metrics based on the entire production process.[268] MDTs show a promising

Table 7. Open-source ML tools and databases for materials.

Tool/Database[a)]	Type	Prime-Focus	Key features	URL/repository
Materials Project	Database	Inorganic materials	140K+ materials, DFT calculations, ML-predicted properties	next-gen.materialsproject.org
AFLOW	Database	Crystal structures	Automated materials discovery, 4M compounds, 8B+ properties REST API	aflowlib.org
NOMAD	Repository	Materials data	Open materials data repository, 19M+ million entries and 113+ TB of data, FAIR data principles	NOMAD-lab.eu
JARVIS-DFT	Database	Materials properties	AI-ready dataset, 40K+ materials, 3D crystal structures	jarvis.nist.gov
Pymatgen	Library	Materials analysis	Python library for materials analysis, structure manipulation	pymatgen.org
ASE	Library	Atomistic simulations	Atomic Simulation Environment, interfaces to many codes	ase-lib.org
SchNet	Framework	Neural networks	Deep learning for atomistic systems, continuous-filter CNNs	github.com/atomistic-machine-learning/schnetpack
CGCNN	Framework	Graph neural networks	Crystal Graph Convolutional Neural Networks	github.com/txie-93/cgcnn
MEGNet	Framework	Property prediction	Graph networks for materials structure & property prediction	github.com/materialsvirtuallab/megnet
ALIGNN	Framework	Graph neural networks	Atomistic Line Graph Neural Network	github.com/usnistgov/alignn
DeepMD-kit	Framework	Molecular dynamics	Deep learning potential energy surfaces	github.com/deepmodeling/deepmd-kit
Materials Cloud	Platform	Computational tools	Web platform for computational materials science	materialscloud.org
QM9	Dataset	Quantum chemistry	133K small, stable organic molecules with quantum properties	pytorch-geometric.readthedocs.io
MP-20	Dataset	Materials discovery	20K materials for benchmarking ML algorithms	https://next-gen.materialsproject.org/materials/mp-20
Open Catalyst Project	Dataset	Catalysis	1.3M+ catalyst-adsorbate systems for renewable energy	opencatalystproject.org
Catalysis-Hub	Database	Catalysis	Surface reactions, reaction energies, barriers	catalysis-hub.org
MatBench	Benchmark	Materials properties	Standardized benchmark suite for ML in materials science, 13 prediction tasks	matbench.materialsproject.org
OMDB	Database	Organic materials	Computational database of organic photovoltaic materials	omdb.mathub.io
NOMAD Laboratory CoE	Toolkit	High-throughput	Tools for automated materials discovery workflows	NOMAD-coe.eu/NOMAD-coe/
GNoME	Database	Compositions & raw energies	520K quasi-stable materials (>380K stable) from DFT & measurements	github.com/google-deepmind/materials_discovery
Omol25	Dataset	Quantum chemistry	100M+ quantum chemical calculations, a family of Universal Models for Atoms	www.medvolt.ai/blog/meta-omol25-uma-models-molecular-simulation
PyMatGen	Toolkit Database	Python library Mater. analysis	Geometric/electronic structure, reaction, phase/Pourbaix diagrams, diffusion	github.com/materialsproject/pymatgen

[a)] Databases & Repositories: Materials Project: Largest inorganic materials database, AFLOW: High-throughput crystal structure database. NOMAD: FAIR data repository for materials science, JARVIS: NIST's materials property database. ML Frameworks & Libraries: Pymatgen: Essential Python toolkit for materials, ASE: Interface to atomistic simulation codes, SchNet: Continuous-filter NNs, MEGNET: materials graph network, ALIGNN: Advanced GNNs. Benchmarks & Standards: MatBench: Standardized ML benchmarks, MP-20: Materials discovery benchmark dataset, Open Catalyst: Catalysis-specific benchmarks. Specialized Applications: Energy Storage: Materials Project (battery materials), OMDB (organics), Catalysis: Open Catalyst Project, Catalysis-Hub, Photovoltaics: OMDB, Materials Project (bandgaps), Fuel Cells: JARVIS (ionic conductivity), Materials Project.

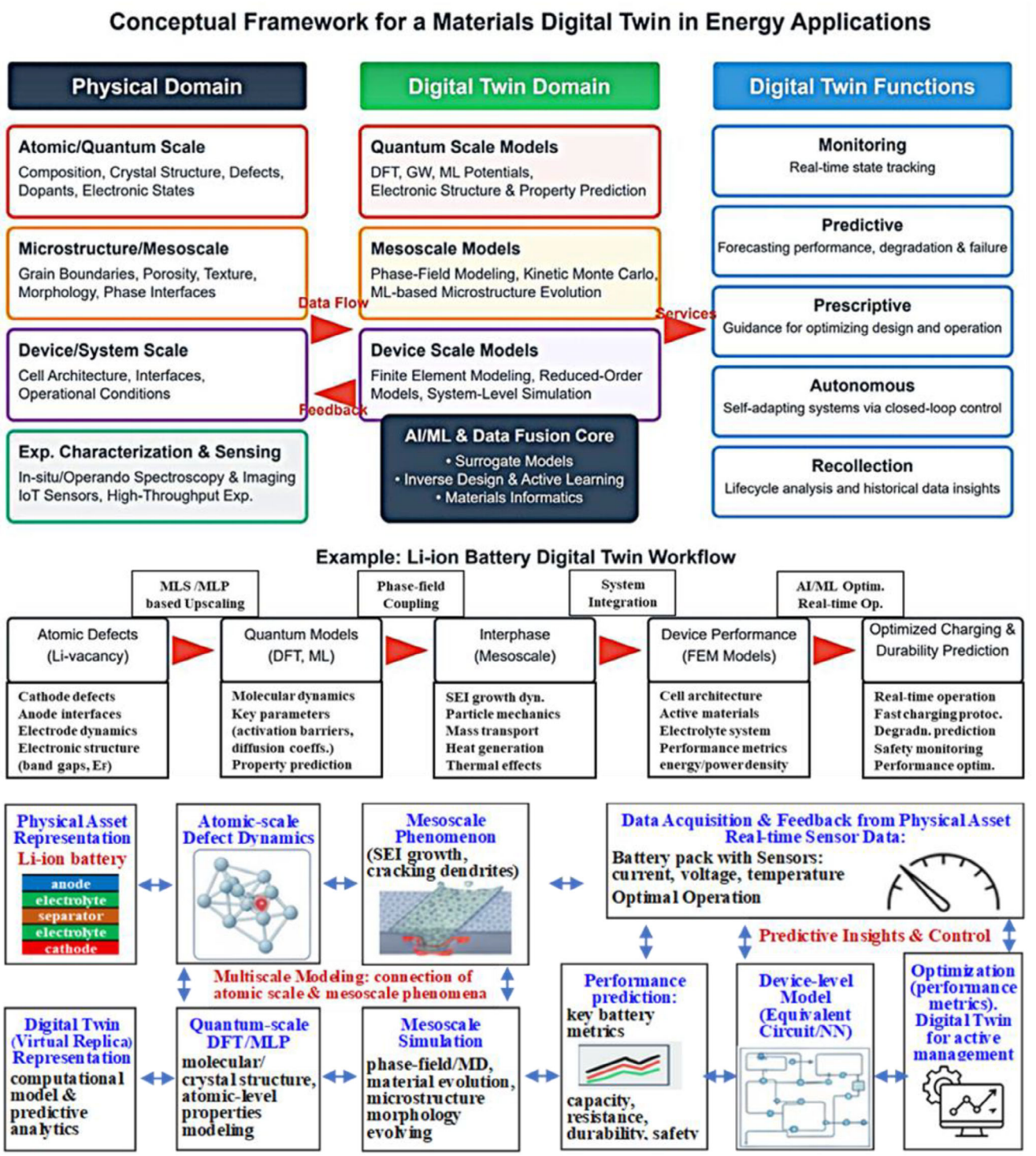

Figure 5. Conceptual Framework for a Materials Digital Twin in Energy Applications, with Lithium-Ion Batteries as an Example. (*top*) framework illustrating a three-domain architecture: Physical Domain encompassing atomic/quantum scale phenomena, microstructure/mesoscale features, and device/system scale characteristics; MDT Domain containing corresponding computational models across quantum, mesoscale, and device scales; MDT Functions providing operational capabilities. (*middle*) Li-ion battery MDT workflow showing the progression from atomic defects and cathode interfaces through quantum models and MD to system integration and real-time optimization. (*bottom*) Bidirectional data flow between the physical Li-ion battery asset and its digital replica, where atomic-level defects (Li-vacancy) are captured through quantum-scale DFT/MLP modeling, mesoscale phenomena such as solid electrolyte interphase (SEI) growth are simulated via phase-field/MD methods, and device-level performance predictions are generated through equivalent circuit models and NNs. Real-time sensor data enables continuous model updating and predictive control optimization, facilitating active management of complex energy storage systems across temporal and spatial scales to enhance performance, reliability, and operational efficiency.

potential to achieve several benefits, such as significantly enhancing battery lifetime and safety, optimizing lifetime and value through asset-specific control, enabling rapid and reliable root-cause analysis of battery failure, and increasing the predictive capabilities of a single asset and reducing uncertainty by pooling data.[266]

Digital manufacturing approaches have been successfully implemented for perovskite solar cell optimization. High-throughput robotic synthesis platforms combined with ML-guided optimization have achieved reproducible fabrication of perovskite devices with power conversion efficiencies exceeding 20%.[269] These systems demonstrate rapid screening of compositional space, with automated synthesis-characterization loops reducing development time from months to weeks.[270] Large-scale perovskite solar panels (4.5 m^2 total area) have been successfully integrated into stand-alone solar farm infrastructure with peak power exceeding 250 W, demonstrating the scalability of MDT-guided optimization.[271] These installations provide continuous monitoring data that validates and refines MDT models under real operating conditions.

The ultimate vision extends beyond improved materials simulation to autonomous materials discovery platforms. Future developments should focus on closed-loop systems that can formulate hypotheses about materials with target properties, plan and execute synthesis experiments, and interpret results through Bayesian active learning frameworks. Through self-learning processes with continuously expanding training data, these systems will build "universal potential libraries" applicable to diverse materials challenges. This represents a paradigm shift from traditional trial-and-error materials discovery to autonomous, quantum-accurate materials innovation platforms that can accelerate the development of next-generation energy technologies.

10. Ethical Considerations and Responsible AI

The implementation of ML in materials science raises several ethical considerations that require careful attention. Data ownership and sharing protocols must be established to ensure fair access to computational resources and findings. Bias in automated discovery systems can perpetuate existing inequalities in materials research if not properly addressed through diverse training datasets and validation protocols. Research shows that data bias, inherent in the data collection process, significantly influences the error and reliability of ML model predictions, limiting the material space that can be reliably discovered by a given model. This is particularly problematic as published datasets often favor successful experiments, creating a strong positive bias that can reinforce existing perspectives and limit the discovery of novel materials outside conventional research domains. Responsible AI practices in materials science include transparent reporting of model limitations, UQ, and validation protocols to ensure reproducible and reliable results.

11. Conclusion and Outlook

ML approaches have revolutionized energy materials discovery and design, enabling remarkable computational efficiency while maintaining quantum-mechanical accuracy. The application of GNNs and SGPR in developing MLPs has facilitated atomistic simulations of complex materials systems with thousands of atoms and extended time scales, providing crucial insights into phenomena such as ion diffusion, phase transitions, and interfacial dynamics. ML-driven screening methodologies enable high-throughput screening of vast chemical spaces, significantly reducing computational burden while accelerating prediction of critical properties for energy materials applications.

The realization of comprehensive materials development has been significantly advanced by the MLP development that can accurately predict material behavior across multiple scales.[68,69,92] The "AI+X" framework, which combines SGPR-based simulations with first-principles potentials, has shown particular promise for creating MDTs of energy materials systems. The "X" component encompasses diverse material elements relevant to energy applications. As an example, one application could envision designing efficient, durable, non-toxic, non-flammable Li-ion batteries using materials composed of selected elements. The MDT would provide a comprehensive visualization of the battery's materials structure, reaction phenomena, performance during charging/discharging, and changes in stability and durability. Through this approach, complex phenomena such as ion diffusion, phase transitions, and interfacial dynamics can be simulated with high accuracy and efficiency.

Future directions extend beyond improved materials simulation toward autonomous discovery platforms, marking a shift from traditional trial-and-error approaches. Through self-learning processes with continuously expanding training data, it is possible to build "built-in universal potential libraries" that can eventually be applied to all "Things+Phenomena." As the data accumulate, the system can engage in self-learning, building universal potential libraries through cumulative learning.

However, several fundamental limitations continue to constrain ML's full potential in energy materials discovery. *Data quality and availability* remain major challenges, with high-quality experimental data scarce for many systems, particularly novel compositions and extreme operating conditions. *Model interpretability* limitations hinder scientific insight generation, as "black box" approaches provide limited understanding of why certain materials exhibit superior performance. Significant *theory-experiment gaps* persist, with models trained on idealized computational data often failing to capture real-world complexities, including defects, grain boundaries, and environmental factors. *Scalability and transferability* issues limit model applicability across different material classes, while *integration barriers*, including technical complexity and economic constraints impede adoption in experimental and industrial environments.

To address these challenges, critical research directions include: development of standardized, community-wide databases with comprehensive UQ; advancement of physics-informed ML approaches that integrate domain knowledge while maintaining interpretability; creation of multi-scale frameworks that seamlessly couple atomistic to device-level phenomena; implementation of robust UQ methods that account for model limitations and extrapolation risks; development of cost-effective autonomous laboratory systems accessible to diverse research environments; and establishment of collaborative frameworks for data sharing and model validation across institutions.

The integration of these complementary approaches is enabling the realization of MDT that simulates material behavior across multiple scales and provides actionable insights for designing next-generation energy technologies. As these methods continue to evolve and become more deeply integrated with experimental workflows, systematic addressing of current limitations through coordinated research efforts will determine whether ML-driven materials discovery can fulfill its promise of revolutionizing energy technology development. Success in overcoming these challenges will enable ML to play an increasingly central role in accelerating the transition to sustainable energy systems, transforming materials discovery from empirical trial-and-error to predictive, physics-informed design.

Abbreviations

ACE, alternating conditional expectation; ANN, artificial neural network; BCM, Bayesian committee machine; BNN, Bayesian neural network; CHGNet, crystal Hamiltonian graph neural network; CGCNN, crystal graph convolutional neural network; CNN, convolutional neural network; DFT, density functional theory; DimeNet, directional message passing neural net; DNN, deep neural network; GAP, Gaussian approximation potential; GNN, graph neural network; GPR, Gaussian process regression; KRR, kernel ridge regression; LASSO, least absolute shrinkage and selection operator; LLM, large language model; LSTM, long short-term memory network; M3GNet, materials 3-body graph network; MACE, message passing atomic cluster expansion; MAE, mean absolute error; MAPE, mean absolute percentage error; MEGNet, materials graph network; MD, molecular dynamics; MDT, materials digital twin; ML, machine learning; MLS, machine learning screening; MLIP, machine learning interatomic potential; MLP, machine learning potential; MPNN, message passing neural network; NequIP, neural equivariant interatomic potential; NN, neural network; RBCM, robust Bayesian committee machine; RF, random forest; RNN, recurrent neural network; RL, reinforcement learning; RMSE, root mean square error; SGPR, sparse Gaussian process regression; SHAP, Shapley additive explanations; SVM, support vector machine; SVR, support vector regression; UQ, uncertainty quantification

Acknowledgements

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT: Ministry of Science and ICT) (No. RS-2024-00346451).

Conflict of Interest

The author declares no conflict of interest.

Keywords

energy materials, Gaussian processes, machine learning potentials, machine learning screening, neural networks

Received: June 18, 2025
Revised: September 18, 2025
Published online: October 25, 2025

[1] K. T. Butler, D. W. Davies, H. Cartwright, O. Isayev, A. Walsh, *Nature* **2018**, *559*, 547.
[2] G. R. Schleder, A. C. M. Padilha, C. M. Acosta, M. Costa, A. Fazzio, *J. Phys.: Mater.* **2019**, *2*, 032001.
[3] C. Chen, Y. Zuo, W. Ye, X. Li, Z. Deng, S. P. Ong, *Adv. Energy Mater.* **2020**, *10*, 1903242.
[4] E. Kocer, T. W. Ko, J. Behler, *Annu. Rev. Phys. Chem.* **2022**, *73*, 163.
[5] V. L. Deringer, M. A. Caro, G. Csányi, *Adv. Mater.* **2019**, *31*, 1902765.
[6] P. De Luna, J. Wei, Y. Bengio, A. Aspuru-Guzik, E. Sargent, *Nature* **2017**, *552*, 23.
[7] A. Chen, X. Zhang, Z. Zhou, *InfoMat* **2020**, *2*, 553.
[8] Y. Liu, O. C. Esan, Z. Pan, L. An, *Energy AI* **2021**, *3*, 100049.
[9] Z. W. Ulissi, A. J. Medford, T. Bligaard, J. K. Nørskov, *Nat. Commun.* **2017**, *8*, 14621.
[10] K. Tran, Z. W. Ulissi, *Nat. Catal.* **2018**, *1*, 696.
[11] J. F. Rodrigues Jr, L. Florea, M. C. F. de Oliveira, D. Diamond, O. N. Oliveira Jr, *Discover Materials* **2021**, *1*, 12.
[12] S. Shalev-Shwartz, S. Ben-David, *Understanding Machine Learning: From Theory to Algorithms*, Cambridge University Press, Cambridge UK **2014**.
[13] J. Behler, M. Parrinello, *Phys. Rev. Lett.* **2007**, *98*, 146401.
[14] P. Rowe, V. L. Deringer, P. Gasparotto, G. Csányi, A. Michaelides, *J. Chem. Phys.* **2020**, *153*, 034702.
[15] A. P. Bartók, J. Kermode, N. Bernstein, G. Csányi, *Phys. Rev. X* **2018**, *8*, 041048.
[16] R. Jinnouchi, F. Karsai, G. Kresse, *Phys. Rev. B* **2019**, *100*, 014105.
[17] T. B. Blank, S. D. Brown, A. W. Calhoun, D. J. Doren, *J. Chem. Phys.* **1995**, *103*, 4129.
[18] V. L. Deringer, A. P. Bartók, N. Bernstein, D. M. Wilkins, M. Ceriotti, G. Csányi, *Chem. Rev.* **2021**, *121*, 10073.
[19] J. Neugebauer, T. Hickel, *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2013**, *3*, 438.
[20] L. Himanen, A. Geurts, A. S. Foster, P. Rinke, *Adv. Sci.* **2019**, *6*, 1900808.
[21] A. Jain, G. Hautier, S. P. Ong, K. Persson, *J. Mater. Res.* **2016**, *31*, 977.
[22] N. Nosengo, *Nature* **2016**, *533*, 22.
[23] A. Jain, S. P. Ong, G. Hautier, W. Chen, W. D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder, K A. Persson, *APL Mater.* **2013**, *1*, 011002.
[24] C. Toher, C. Oses, D. Hicks, E. Gossett, F. Rose, P. Nath, D. Usanmaz, D. C. Ford, E. Perim, C. E. Calderon, J. J. Plata, Y. Lederer, M. Jahntek, W. Setyawan, S. Wang, J. Xue, K. Rasch, R. V. Chepulskii, R. H. Taylor, G. Gomez, H. Shi, A. R. Supka, R. A. R. A. Orabi, P. Gopal, F. T. Cerasoli, L. Liyanage, H. Wang, I. Siloi, L. A. Agapito, C. Nyshadham, et al., *Handbook of Materials Modeling*, Springer Cham, Cham Switzerland **2018**, p. 1.
[25] M. Peplow, *Nature* **2023**, *10*, 1.
[26] S K. Suram, Y. Xue, J. Bai, R. Le Bras, B. Rappazzo, R. Bernstein, J. Bjorck, L. Zhou, R. B van Dover, C P. Gomes, J M. Gregoire, *ACS Comb. Sci.* **2017**, *19*, 37.
[27] Y. Zhang, C. Ling, *npj Comput. Mater.* **2018**, *4*, 25.
[28] P. Raccuglia, K C. Elbert, P D. F. Adler, C. Falk, M B. Wenny, A. Mollo, M. Zeller, S A. Friedler, J. Schrier, A J. Norquist, *Nature* **2016**, *533*, 73.
[29] Q. Tong, P. Gao, H. Liu, Y. Xie, J. Lv, Y. Wang, J. Zhao, *J. Phys. Chem. Lett.* **2020**, *11*, 8710.
[30] S. Khalid, T. Khalil, S. Nasreen, in *Proceedings of Science, Information (SAI) Conference*, IEEE, New York NY USA **2014**, 372.
[31] J. K. Nørskov, T. Bligaard, J. Rossmeisl, C. H. Christensen, *Nat. Chem.* **2009**, *1*, 37.
[32] S. Back, J. Yoon, N. Tian, W. Zhong, K. Tran, Z. W. Ulissi, *J. Phys. Chem. Lett.* **2019**, *10*, 4401.
[33] B. Hammer, J. K. Nørskov, *Nature* **1995**, *376*, 238.
[34] F. Calle-Vallejo, J. I. Martínez, J. M. García-Lastra, P. Sautet, D. Loffreda, *Angew. Chem., Int. Ed.* **2014**, *53*, 8316.

[35] X. Ma, Z. Li, L. E. K. Achenie, H. Xin, *J. Phys. Chem. Lett.* **2015**, *6*, 3528.
[36] F. Calle-Vallejo, N G. Inoglu, H.-Y. Su, J I. Martínez, I C. Man, M T. M. Koper, J R. Kitchin, J. Rossmeisl, *Chem. Sci.* **2013**, *4*, 1245.
[37] O. Isayev, C. Oses, C. Toher, E. Gossett, S. Curtarolo, A. Tropsha, *Nat. Commun.* **2017**, *8*, 15679.
[38] A. Mangal, E. A. Holm, *Integrating Materials, Manufacturing Innovation* **2018**, *7*, 87.
[39] J. Schmidhuber, *Neural Networks* **2015**, *61*, 85.
[40] W. Sha, Y. Guo, Q. Yuan, S. Tang, X. Zhang, S. Lu, X. Guo, Y.-C. Cao, S. Cheng, *Advanced Intelligent Systems* **2020**, *2*, 1900143.
[41] J. A. Warren, *MRS Bull.* **2018**, *43*, 452.
[42] A. Jain, *Current Opinion in Solid State, Materials Science* **2024**, *33*, 101189.
[43] A. Agrawal, A. Choudhary, *APL Mater.* **2016**, *4*, 053208.
[44] J. Schmidt, M. R. G. Marques, S. Botti, M. A. L. Marques, *npj Comput. Mater.* **2019**, *5*, 83.
[45] S. P. Ong, *Comput. Mater. Sci.* **2019**, *161*, 143.
[46] T. Xie, J. C. Grossman, *Phys. Rev. Lett.* **2018**, *120*, 145301.
[47] S. L. Brunton, J. N. Kutz, *Data-Driven Science, Engineering: Machine Learning, Dynamical Systems, Control*, Cambridge University Press, Cambridge UK **2019**.
[48] K. Sodeyama, Y. Igarashi, T. Nakayama, Y. Tateyama, M. Okada, *Phys. Chem. Chem. Phys.* **2018**, *20*, 22585.
[49] A. D. Sendek, Q. Yang, E. D. Cubuk, K.-A. N. Duerloo, Y. Cui, E. J. Reed, *Energy Environ. Sci.* **2017**, *10*, 306.
[50] E. Schulz, M. Speekenbrink, A. Krause, *Journal of Mathematical Psychology* **2018**, *85*, 1.
[51] Y. Okamoto, Y. Kubo, *ACS Omega* **2018**, *3*, 7868.
[52] M. Hellström, J. Behler, *Handbook of Materials Modeling*, Springer Cham, Cham Switzerland **2018**, p. 1.
[53] B. Kim, S. Lee, J. Kim, *Sci. Adv.* **2020**, *6*, aax9324.
[54] H. K. D. H. Bhadeshia, *ISIJ Int.* **1999**, *39*, 966.
[55] T. N. Sainath, O. Vinyals, A. Senior, H. Sak, in *Proceedings of IEEE International Conference on Acoustics, Speech, Signal Processing (ICASSP)*, IEEE, New York NY USA **2015**, 4580.
[56] W. C. Lu, X. B. Ji, M. J. Li, L. Liu, B. H. Yue, L. M. Zhang, *Advanced Manufacturing* **2013**, *1*, 151.
[57] N. Kireeva, V. S. Pervov, *Phys. Chem. Chem. Phys.* **2017**, *19*, 20904.
[58] P. V. Balachandran, J. Theiler, J. M. Rondinelli, T. Lookman, *Sci. Rep.* **2015**, *5*, 13285.
[59] Y. Li, *Materials Science, Engineering: A* **2006**, *433*, 261.
[60] L. Breiman, J. H. Friedman, *J. Am. Stat. Assoc.* **1985**, *80*, 580.
[61] L. Breiman, J. H. Friedman, *J. Am. Stat. Assoc.* **1985**, *80*, 614.
[62] V. Gladkikh, D. Y. Kim, A. Hajibabaei, A. Jana, C. W. Myung, K. S. Kim, *J. Phys. Chem. C* **2020**, *124*, 8905.
[63] F. Maleki, N. Muthukrishnan, K. Ovens, C. Reinhold, R. Forghani, *Neuroimaging Clinics of North America* **2020**, *30*, 433.
[64] C. Sutton, M. Boley, L. M. Ghiringhelli, M. Rupp, J. Vreeken, M. Scheffler, *Nat. Commun.* **2020**, *11*, 4428.
[65] J. Cai, X. Chu, K. Xu, H. Li, J. Wei, *Nanoscale Advances* **2020**, *2*, 3115.
[66] A. P. Bartók, M. C. Payne, R. Kondor, G. Csányi, *Phys. Rev. Lett.* **2010**, *104*, 136403.
[67] A. P. Bartók, G. Csányi, *Int. J. Quantum Chem.* **2015**, *115*, 1051.
[68] A. Hajibabaei, C. W. Myung, K. S. Kim, *Phys. Rev. B* **2021**, *103*, 214102.
[69] S. Y. Willow, A. Hajibabaei, M. Ha, D. C. Yang, C. W. Myung, S. K. Min, G. Lee, K. S. Kim, *Chemical Physics Reviews* **2024**, *5*, 041307.
[70] S. Y. Willow, S. Kim, D. C Yang, M. Ha, A. Hajibabaei, J. W. Yang, K S. Kim, C. W. Myung, *Chemical Physics Reviews* **2025**, *6*, 021401.
[71] J. Gilmer, S. S. Schoenholz, P. F. Riley, O. Vinyals, G. E. Dahl, *Proceedings of the 34th International Conference on Machine Learning* **2017**, *70*, 1263.
[72] D. K. Duvenaud, D. Maclaurin, J. Iparraguirre, R. Bombarell, T. Hirzel, A. Aspuru-Guzik, R. P. Adams, *Advances in Neural Information Processing Systems* **2015**, *28*, 2224.
[73] K. T. Schütt, F. Arbabzadah, S. Chmiela, K. R. Müller, A. Tkatchenko, *Nat. Commun.* **2017**, *8*, 13890.
[74] K. T. Schütt, P.-J. Kindermans, H. E. Sauceda Felix, S. Chmiela, A. Tkatchenko, K.-R. Müller, *Advances in Neural Information Processing Systems* **2017**, *30*, 991.
[75] C. Chen, S. P. Ong, *Nature Computational Science* **2022**, *2*, 718.
[76] S. D. Griesemer, Y. Xia, C. Wolverton, *Nature Computational Science* **2023**, *3*, 934.
[77] B. Deng, P. Zhong, K. Jun, J. Riebesell, K. Han, C. J. Bartel, G. Ceder, *Nature Machine Intelligence* **2023**, *5*, 1031.
[78] C. Chen, W. Ye, Y. Zuo, C. Zheng, S. P. Ong, *Chem. Mater.* **2019**, *31*, 3564.
[79] A S. Rosen, V. Fung, P. Huck, C T. O'Donnell, M K. Horton, D G. Truhlar, K A. Persson, J M. Notestein, R Q. Snurr, *npj Comput. Mater.* **2022**, *8*, 112.
[80] J. Gasteiger, J. Groß, S. Günnemann, arXiv 2020, arXiv:2003.03123.
[81] S. Khoshraftar, A. An, *ACM Transactions on Intelligent Systems, Technology* **2024**, *15*, 1.
[82] V. G. Satorras, E. Hoogeboom, M. Welling, *Proceedings of the 38th International Conference on Machine Learning* **2021**, *139*, 9323.
[83] I. Batatia, S. Batzner, D. P. Kovács, A. Musaelian, G N. C. Simm, R. Drautz, C. Ortner, B. Kozinsky, G. Csányi, *Nature Machine Intelligence* **2025**, *7*, 56.
[84] S. Batzner, A. Musaelian, L. Sun, M. Geiger, J P. Mailoa, M. Kornbluth, N. Molinari, T E. Smidt, B. Kozinsky, *Nat. Commun.* **2022**, *13*, 2453.
[85] F. Drautz, *Phys. Rev. B* **2019**, *99*, 014104.
[86] X. Gong, H. Li, N. Zou, R. Xu, W. Duan, Y. Xu, *Nat. Commun.* **2023**, *14*, 2848.
[87] B. Kozinsky, A. Musaelian, A. Johansson, S. Batzner, in *SC '23: Proceedings of the International Conference for High Performance Computing, Networking, Storage, Analysis*, ACM, New York NY USA **2023**, 1.
[88] F. B. Fuchs, D. E. Worrall, V. Fischer, M. Welling, *Advances in Neural Information Processing Systems* **2020**, *33*, 1970.
[89] A. Musaelian, S. Batzner, A. Johansson, L. Sun, C J. Owen, M. Kornbluth, B. Kozinsky, *Nat. Commun.* **2023**, *14*, 579.
[90] I. Batatia, D. P. Kovács, G. N. Simm, C. Ortner, G. Csányi, *Advances in Neural Information Processing Systems* **2022**, *35*, 11423.
[91] A. Hajibabaei, C. W. Myung, K. S. Kim, *arXiv* **2020**, arXiv:2009.13179.
[92] A. Hajibabaei, K. S. Kim, *J. Phys. Chem. Lett.* **2021**, *12*, 8115.
[93] S. De, A. P. Bartók, G. Csányi, M. Ceriotti, *Phys. Chem. Chem. Phys.* **2016**, *18*, 13754.
[94] C. E. Rasmussen, C. K. I. Williams, *Gaussian Processes for Machine Learning*, MIT Press, Cambridge MA USA **2006**.
[95] V. L. Deringer, M. A. Caro, G. Csányi, *Nat. Commun.* **2020**, *11*, 5461.
[96] M. A. Wood, M. A. Cusentino, B. D. Wirth, A. P. Thompson, *Phys. Rev. B* **2019**, *99*, 184305.
[97] F C. Mocanu, K. Konstantinou, T. H. Lee, N. Bernstein, V L. Deringer, G. Csányi, S R. Elliott, *J. Phys. Chem. B* **2018**, *122*, 8998.
[98] L. Shenoy, C D. Woodgate, J B. Staunton, A P. Bartók, C S. Becquart, C. Domain, J R. Kermode, *Phys. Rev. Mater.* **2024**, *8*, 033804.
[99] J. P. Janet, C. Duan, T. Yang, A. Nandy, H. J. Kulik, *Chem. Sci.* **2019**, *10*, 7913.
[100] R. Chalapathy, S. Chawla, *arXiv* **2019**, arXiv:1901.03407.
[101] T. Lookman, P. V. Balachandran, D. Xue, R. Yuan, *npj Comput. Mater.* **2019**, *5*, 21.
[102] B. Shahriari, K. Swersky, Z. Wang, R. P. Adams, N. de Freitas, *Proc. IEEE* **2015**, *104*, 148.
[103] F. Musil, A. Grisafi, A. P. Bartók, C. Ortner, G. Csányi, M. Ceriotti, *Chem. Rev.* **2021**, *121*, 9759.

[104] T. Xie, A. France-Lanord, Y. Wang, Y. Shao-Horn, J. C. Grossman, *Nat. Commun.* **2019**, *10*, 2667.
[105] S. Ryu, Y. Kwon, W. Y. Kim, *Chem. Sci.* **2019**, *10*, 8438.
[106] P. Reiser, M. Neubert, A. Eberhard, L. Torresi, C. Zhou, C. Shao, H. Metni, C. van Hoesel, H. Schopmans, T. Sommer, P. Friederich, *Commun. Mater.* **2022**, *3*, 93.
[107] S. Gong, C. Duan, J. Zhu, T. Yang, A. Nandy, H. J. Kulik, *Sci. Adv.* **2023**, *9*, adi3245.
[108] C. Chen, W. Ye, Y. Zuo, C. Zheng, S. P. Ong, *Commun. Mater.* **2021**, *2*, 92.
[109] A. Merchant, S. Batzner, S. S. Schoenholz, M. Aykol, G. Cheon, E. D. Cubuk, *Nature* **2023**, *624*, 80.
[110] D. Fooshee, A. Mood, E. Gutman, M. Tavakoli, G. Urban, F. Liu, *Mol. Syst. Des. Eng.* **2018**, *3*, 442.
[111] R. Zhang, C. H. Ji, X. Zhou, T. Liu, G. Jin, Z. Pan, Y. Liu, *Energy* **2024**, *285*, 129277.
[112] O. P. D, P. S. Babu, I. V, A. B, V. S, K. C, *Sci. Rep.* **2024**, *14*, 16036.
[113] H. Valladares, T. Li, L. Zhu, H. El-Mounayri, A. M. Hashem, A. E. Abdel-Ghany, A. Tovar, *J. Power Sources* **2022**, *528*, 231026.
[114] P. Tagade, K. S. Hariharan, S. Ramachandran, A. Khandelwal, A. Naha, S. M. Kolake, S H. Han, *J. Power Sources* **2020**, *445*, 227281.
[115] Y. Tang, S. Zhong, P. Wang, Y. Zhang, Y. Wang, *Sci. Rep.* **2024**, *14*, 23524.
[116] S. Wang, J. Liu, X. Song, H. Xu, Y. Gu, J. Fan, B. Sun, L. Yu, *Nano-Micro Lett.* **2025**, *17*, 287.
[117] Z. W Chen, Z. Lu, L. X. Chen, M. Jiang, D. Chen, C. V. Singh, *Chemical Catalysis* **2021**, *1*, 183.
[118] C. Wang, B. Wang, C. Wang, A. Li, Z. Chang, R. Wang, *npj Comput. Mater.* **2025**, *11*, 111.
[119] D. H. Mok, H. Li, G. Zhang, C. Lee, K. Jiang, S. Back, *Nat. Commun.* **2023**, *14*, 7303.
[120] J. Park, I. Chung, H. Jeong, D. Lee, Y. Yun, *Applied Catalysis B: Environmental, Energy* **2025**, *361*, 124622.
[121] B. M. Abraham, M. V. Jyothirmai, P. Sinha, F. Viñes, J. K. Singh, F. Illas, *WIREs Computational Molecular Science* **2024**, *14*, 1730.
[122] H. Liu, K. Liu, H. Zhu, W. Guo, Y. Li, *RSC Adv.* **2024**, *14*, 7276.
[123] M. Umer, S. Umer, M. Zafari, M. Ha, R. Anand, A. Hajibabaei, A. Abbas, G. Lee, K. S. Kim, *J. Mater. Chem. A* **2022**, *10*, 6679.
[124] Q. Tao, P. Xu, M. Li, W. Lu, *npj Comput. Mater.* **2021**, *7*, 23.
[125] S. Valsalakumar, S. Bhandari, A. Roy, T. K. Mallick, J. Hinshelwood, S. Sundaram, *npj Comput. Mater.* **2024**, *10*, 212.
[126] A. Khan, J. Kandel, H. Tayara, K. T. Chong, *Mol. Inf.* **2024**, *43*, 202300217.
[127] Y. Liu, X. Tan, J. Liang, H. Han, P. Xiang, W. Yan, *Adv. Funct. Mater.* **2023**, *23*, 2214271.
[128] H. Rafique, G. Abbas, M. J. Mendes, P. Barquinha, R. Martins, E. Fortunato, H. Águas, S. Jana, *Nano-Micro Lett.* **2026**, *18*, 44.
[129] C. Ren, Y. Wu, J. Zou, B. Cai, *Materials* **2024**, *17*, 2686.
[130] S. Ji, Y. Zhang, Y. Huang, Z. Yu, Y. Zhou, X. Lin, *Materials* **2024**, *17*, 3741.
[131] E. Chemali, P. J. Kollmeyer, M. Preindl, A. Emadi, *J. Power Sources* **2018**, *400*, 242.
[132] L.-T. Wu, B. J. Hwang, J.-C. Jiang, *Chem. Eng. J.* **2025**, *515*, 163801.
[133] J. Li, M. Zhou, H.-H. Wu, L. Wang, J. Zhang, N. Wu, K. Pan, G. Liu, Y. Zhang, J. Han, X. Liu, X. Chen, J. Wan, Q. Zhang, *Adv. Energy Mater.* **2024**, *14*, 2304480.
[134] H. Liu, S. Ma, J. Wu, Y. Wang, X. Wang, *Front. Energy Res.* **2021**, *9*, 639741.
[135] K. Hatakeyama-Sato, T. Tezuka, M. Umeki, K. Oyaizu, *J. Am. Chem. Soc.* **2020**, *142*, 3301.
[136] Y. Wang, C. Zhong, J. Zhang, J. Liu, K. Hu, J. Chen, X. Lin, *Journal of Materials Informatics* **2025**, *5*, 41.
[137] A. K. Mishra, S. Rajput, M. Karamta, I. Mukhopadhyay, *ACS Omega* **2023**, *8*, 16419.
[138] K. Li, J. Wang, Y. Song, Y. Wan, *Nat. Commun.* **2023**, *14*, 2789.
[139] Z. Ahmad, T. Xie, C. Maheshwari, J. C. Grossman, V. Viswanathan, *ACS Cent. Sci.* **2018**, *4*, 996.
[140] M. Alipour, L. Yin, S. S. Tavallaey, D. Brandell, *J. Power Sources* **2023**, *579*, 233273.
[141] G. Ceder, *MRS Bull.* **2010**, *35*, 693.
[142] A. Jain, Y. Shin, K. A. Persson, *Nat. Rev. Mater.* **2016**, *1*, 15004.
[143] R. Marom, S. F. Amalraj, N. Leifer, D. Jacob, D. Aurbach, *J. Mater. Chem.* **2011**, *21*, 9938.
[144] V. L. Deringer, *Journal of Physics: Energy* **2020**, *2*, 041003.
[145] K. Liu, Z. Wei, Z. Yang, K. Li, *J. Cleaner Prod.* **2021**, *289*, 125159.
[146] A. Van Der Ven, Z. Deng, S. Banerjee, S. P. Ong, *Chem. Rev.* **2020**, *120*, 6977.
[147] S. Wenzel, D. A. Weber, T. Leichtweiss, M. R. Busche, J. Sann, J. Janek, *Solid State Ionics* **2016**, *286*, 24.
[148] Y. Zhang, X. He, Z. Chen, Q. Bai, A. M. Nolan, C. A. Roberts, D. Banerjee, T. Matsunaga, Y. Mo, C. Ling, *Nat. Commun.* **2019**, *10*, 5260.
[149] K. Suzuki, K. Ohura, A. Seko, Y. Iwamizu, G. Zhao, M. Hirayama, I. Tanaka, R. Kanno, *J. Mater. Chem. A* **2020**, *8*, 11582.
[150] M. Nakayama, K. Kanamori, K. Nakano, R. Jalem, I. Takeuchi, H. Yamasaki, *Chem. Rec.* **2019**, *19*, 771.
[151] Y. Wang, T. Xie, A. France-Lanord, A. Berkley, J A. Johnson, Y. Shao-Horn, J. C. Grossman, *Chem. Mater.* **2020**, *32*, 4144.
[152] M. Ha, A. Hajibabaei, D. Y. Kim, A. N. Singh, J. Yun, C. W. Myung, K. S. Kim, *Adv. Energy Mater.* **2022**, *12*, 2201497.
[153] D. Kim, H. Nam, Y.-H. Cho, B. C. Yeo, So-H Cho, J.-P. Ahn, K.-Y. Lee, S. Y. Lee, S. S. Han, *ACS Catal.* **2019**, *9*, 8702.
[154] G. Houchins, V. Viswanathan, *J. Chem. Phys.* **2020**, *153*, 054124.
[155] O. Allam, B. W. Cho, K. C. Kim, S. S. Jang, *RSC Adv.* **2018**, *8*, 39414.
[156] K A. Severson, P M. Attia, N. Jin, N. Perkins, B. Jiang, Zi Yang, M H. Chen, M. Aykol, P K. Herring, D. Fraggedakis, M Z. Bazant, S J. Harris, W C. Chueh, R D. Braatz, *Nat. Energy* **2019**, *4*, 383.
[157] F. Wang, Z. Zhai, Z. Zhao, Y. Di, X. Chen, *Nat. Commun.* **2024**, *15*, 4332.
[158] S. M. Wickramaarachchi, S. A. D. Suraweera, D. M. P. Akalanka, V. Logeeshan, C. Wanigasekara, *Computers* **2025**, *13*, 147.
[159] M. Zafari, D. Kumar, M. Umer, K. S. Kim, *J. Mater. Chem. A* **2020**, *8*, 5209.
[160] J. Sun, A. Chen, J. Guan, Y. Han, Y. Liu, X. Niu, M. He, L. Shi, J. Wang, X. Zhang, *Energy & Environmental Materials* **2024**, *7*, 12693.
[161] T. Toyao, Z. Maeno, S. Takakusagi, T. Kamachi, I. Takigawa, K. I. Shimizu, *ACS Catal.* **2020**, *10*, 2260.
[162] A. Khorshidi, J. Violet, J. Hashemi, A. A. Peterson, *Nat. Catal.* **2018**, *1*, 263.
[163] R. Jinnouchi, R. Asahi, *J. Phys. Chem. Lett.* **2017**, *8*, 4279.
[164] M. Ha, D. Y. Kim, M. Umer, V. Gladkikh, C. W. Myung, K. S. Kim, *Energy Environ. Sci.* **2021**, *14*, 3455.
[165] M. O. J. Jäger, E. V. Morooka, F. F. Canova, L. Himanen, A. S. Foster, *npj Comput. Mater.* **2018**, *4*, 37.
[166] S. Choi, Y. Kim, J. W. Kim, Z. Kim, W. Y. Kim, *Chem.-Eur. J.* **2018**, *24*, 12354.
[167] T. Toyao, K. Suzuki, S. Kikuchi, S. Takakusagi, K. I. Shimizu, I. Takigawa, *J. Phys. Chem. C* **2018**, *122*, 8315.
[168] R. Anand, A. S. Nissimagoudar, M. Umer, M. Ha, M. Zafari, S. Umer, G. Lee, K. S. Kim, *Adv. Energy Mater.* **2021**, *11*, 2102388.
[169] R. Anand, A. S. Nissimagoudar, J. Park, J. Mun, G. Lee, K. S. Kim, *J. Mater. Chem. A* **2022**, *10*, 22500.
[170] M. Ha, P. Thangavel, N. K. Dang, D. Y. Kim, S. Sultan, J. S. Lee, K. S. Kim, *Small* **2023**, *19*, 2300240.

[171] Y. Wang, J. Liu, Y. Wang, A. M. Al-Enizi, G. Zheng, *Small* **2017**, *13*, 1701809.
[172] R. Anand, M. Zafari, V. Gupta, G. Lee, K. S. Kim, *J. Mater. Chem. A* **2025**, *13*, 5045.
[173] M. Zhong, K. Tran, Y. Min, C. Wang, Z. Wang, C.-T. Dinh, P. De Luna, Z. Yu, A. S. Rasouli, P. Brodersen, S. Sun, O. Voznyy, C.-S. Tan, M. Askerka, F. Che, M. Liu, A. Seifitokaldani, Y. Pang, S.-C. Lo, A. Ip, Z. Ulissi, E H. Sargent, *Nature* **2020**, *581*, 178.
[174] Y. Huang, Y. Chen, T. Cheng, L.-W. Wang, W. A. Goddard, *ACS Energy Lett.* **2018**, *3*, 2983.
[175] Y. Chen, Y. Huang, T. Cheng, W. A. Goddard, *J. Am. Chem. Soc.* **2019**, *141*, 11651.
[176] M. Zafari, A. S. Nissimagoudar, M. Umer, G. Lee, K. S. Kim, *J. Mater. Chem. A* **2021**, *9*, 9203.
[177] A. N. Singh, R. Anand, M. Zafari, M. Ha, K. S. Kim, *Adv. Energy Mater.* **2024**, *14*, 2304106.
[178] M. Zafari, R. Anand, A. S. Nissimagoudar, M. Ha, G. Lee, K. S. Kim, *Nanoscale* **2024**, *16*, 555.
[179] M. Zafari, M. Umer, A. S. Nissimagoudar, R. Anand, M. Ha, S. Umer, G. Lee, K. S. Kim, *J. Phys. Chem. Lett.* **2022**, *13*, 4530.
[180] S. Tang, Q. Dang, T. Liu, S. Zhang, Z. Zhou, X. Li, X. Wang, E. Sharman, Yi Luo, J. Jiang, *J. Am. Chem. Soc.* **2020**, *142*, 19308.
[181] R. Ding, J. Chen, Y. Chen, J. Liu, Y. Bando, X. Wang, *Chem. Soc. Rev.* **2024**, *53*, 11390.
[182] D. Roy, S. C. Mandal, A. Das, B. Pathak, *Chem.-Eur. J.* **2024**, *30*, 202302679.
[183] P. M. Attia, A. Grover, N. Jin, K. A. Severson, T. M. Markov, Y.-H. Liao, M. H. Chen, B. Cheong, N. Perkins, Z. Yang, P. K. Herring, M. Aykol, S. J. Harris, R. D. Braatz, S. Ermon, W. C. Chueh, *Nature* **2020**, *578*, 397.
[184] C. J. Bartel, C. Sutton, B. R. Goldsmith, R. Ouyang, C. B. Musgrave, L. M. Ghiringhelli, M. Scheffler, *Sci. Adv.* **2019**, *5*, aav0693.
[185] G. Pilania, P. V. Balachandran, C. Kim, T. Lookman, *Frontiers in Materials* **2016**, *3*, 19.
[186] H. Min, D. Y. Lee, J. Kim, G. Kim, K. S. Lee, J. Kim, M. J. Paik, Y. K. Kim, K. S. Kim, M. G. Kim, T. J. Shin, S. Il Seok, *Nature* **2021**, *598*, 444.
[187] J. Kim, K. S. Kim, C. W. Myung, *npj Comput. Mater.* **2020**, *6*, 100.
[188] H. Wang, R. Ouyang, W. Chen, A. Pasquarello, *J. Am. Chem. Soc.* **2024**, *146*, 17636.
[189] C. W. Myung, A. Hajibabaei, J. Cha, M. Ha, J. Kim, K. S. Kim, *Adv. Energy Mater.* **2022**, *12*, 2202279.
[190] A. Paul, D. Jha, R. Al-Bahrani, W. K. Liao, A. Choudhary, A. Agrawal, presented at *2019 International Joint Conference on Neural Networks (IJCNN)*, Budapest Hungary, July **2019**, 1.
[191] S. Nagasawa, E. Al-Naamani, A. Saeki, *J. Phys. Chem. Lett.* **2018**, *9*, 2639.
[192] W. Sun, Y. Zheng, K. Yang, Q. Zhang, A. A. Shah, Z. Wu, Y. Sun, L. Feng, D. Chen, Z. Xiao, S. Lu, Y. Li, K. Sun, *Sci. Adv.* **2019**, *5*, aay4275.
[193] S. Ahmed, X. Wang, H. Li, Y. Zhou, Y. Chen, L. Sun, W. Zhang, R. Mazzarello, *Physica Status Solidi Rapid Research Letters* **2021**, *15*, 2100064.
[194] F. C. Mocanu, K. Konstantinou, J. Mavracic, S. R. Elliott, *Physica Status Solidi Rapid Research Letters* **2021**, *15*, 2000485.
[195] Y. Zhou, W. Zhang, E. Ma, V. L. Deringer, *Nat. Electron.* **2023**, *6*, 746.
[196] A. S. Rosen, S. M. Iyer, D. Ray, Z. Yao, A. Aspuru-Guzik, L. Gagliardi, J. M. Notestein, R. Q. Snurr, *Matter* **2021**, *4*, 1578.
[197] S. Callaghan, *Patterns* **2021**, *2*, 1.
[198] R. Anderson, J. Rodgers, E. Argueta, A. Biong, D. A. Gómez-Gualdrón, *Chem. Mater.* **2018**, *30*, 6325.
[199] Y. He, E. D. Cubuk, M. D. Allendorf, E. J. Reed, *J. Phys. Chem. Lett.* **2018**, *9*, 4562.
[200] M. Fernandez, P. G. Boyd, T. D. Daff, M. Z. Aghaji, T. K. Woo, *J. Phys. Chem. Lett.* **2014**, *5*, 3056.
[201] V. Kuz'min, P. G. Polishchuk, A. G. Artemenko, S. A. Andronati, *Mol. Inf.* **2011**, *30*, 593.
[202] J S. De Vos, S. Ravichandran, S. Borgmans, L. Vanduyfhuys, P. Van Der Voort, S. M. J. Rogge, V. Van Speybroeck, *Chem. Mater.* **2024**, *36*, 4315.
[203] F. Ren, L. Ward, T. Williams, K. J. Laws, C. Wolverton, J. Hattrick-Simpers, A. Mehta, *Sci. Adv.* **2018**, *4*, aaq1566.
[204] B. Settles, *Computer Sciences Technical Report 1648*, University of Wisconsin-Madison, Madison WI USA **2009**.
[205] M. L. Hutchinson, E. Antono, B. M. Gibbons, S. Paradiso, J. Ling, B. Meredig, *arXiv* **2017**, arXiv:1711.05099.
[206] X. Meng, G. E. Karniadakis, *J. Comput. Phys.* **2020**, *401*, 109020.
[207] F. Häse, L. M. Roch, A. Aspuru-Guzik, *Trends in Chemistry* **2019**, *1*, 282.
[208] Y. Iwasaki, H. Jaekyun, Y. Sakuraba, M. Kotsugi, Y. Igarashi, *Science, Technology of Advanced Materials: Methods* **2022**, *2*, 365.
[209] J. L. Gardner, H. Schulz, J. Helie, L. Sun, G. N. Simm, arXiv **2025**, arXiv:2506.14963.
[210] P. Perdikaris, M. Raissi, A. Damianou, N. D. Lawrence, G. E. Karniadakis, *Proceedings of the Royal Society A* **2017**, *473*, 20160751.
[211] G. Pilania, J. E. Gubernatis, T. Lookman, *Comput. Mater. Sci.* **2017**, *129*, 156.
[212] M. Messerly, S. Matin, A. E. A. Allen, B. Nebgen, K. Barros, J. S. Smith, N. Lubbers, R. Messerly, *arXiv* **2025**, arXiv:2505.01590.
[213] C. Chen, Y. Zuo, W. Ye, X. Li, S. P. Ong, *Nature Computational Science* **2021**, *1*, 46.
[214] S. Ju, R. Yoshida, C. Liu, S. Wu, K. Hongo, T. Tadano, J. Shiomi, *Phys. Rev. Mater.* **2021**, *5*, 053801.
[215] A. M. Gopakumar, P. V. Balachandran, D. Xue, J. E. Gubernatis, T. Lookman, *Sci. Rep.* **2018**, *8*, 3738.
[216] A. M. Schweidtmann, A. D. Clayton, N. Holmes, E. Bradford, R. A. Bourne, A. A. Lapkin, *Chem. Eng. J.* **2018**, *352*, 277.
[217] R. Zhu, W. Peng, F. Yang, M. Xie, *IEEE Transactions on Transportation Electrification* **2025**, *11*, 8327.
[218] S. Y. Willow, D. G. Kim, R. Sundheep, A. Hajibabaei, K. S. Kim, C. W. Myung, *Phys. Chem. Chem. Phys.* **2024**, *26*, 22073.
[219] M. Ha, A. Hajibabaei, S. Pourasad, K. S. Kim, *ACS Physical Chemistry Au* **2022**, *2*, 260.
[220] A. Hajibabaei, M. Ha, S. Pourasad, J. Kim, K. S. Kim, *J. Phys. Chem. A* **2021**, *125*, 9414.
[221] L. Hirschfeld, K. Swanson, K. Yang, R. Barzilay, C. W. Coley, *Journal of Chemical Information, Modeling* **2020**, *60*, 3770.
[222] A. R. Tan, S. Urata, S. Goldman, J. C. B. Dietschreit, R. Gómez-Bombarelli, *npj Comput. Mater.* **2023**, *9*, 225.
[223] Y. Gal, Z. Ghahramani, *Proceedings of Machine Learning Research* **2016**, *48*, 1050.
[224] L. V. Jospin, H. Laga, F. Boussaid, W. Buntine, M. Bennamoun, *IEEE Computational Intelligence Magazine* **2022**, *17*, 29.
[225] Z. Deng, Y. Mo, S. P. Ong, *NPG Asia Mater.* **2016**, *8*, 254.
[226] V. Tshitoyan, J. Dagdelen, L. Weston, A. Dunn, Z. Rong, O. Kononova, K A. Persson, G. Ceder, A. Jain, *Nature* **2019**, *571*, 95.
[227] L. N. Anderson, C. T. Hoyt, J. D. Zucker, A. D. McNaughton, J. R. Teuton, K. Karis, N. N. Arokium-Christian, J. T. Warley, Z. R. Stromberg, B. M. Gyori, N. Kumar, *Frontiers in Immunology* **2025**, *16*, 1502484.
[228] Y. Li, X. Zhang, M. Liu, L. Shen, *Journal of Materials Informatics* **2025**, *5*, 43.
[229] S. Grimme, J. Antony, S. Ehrlich, H. Krieg, *J. Chem. Phys.* **2010**, *132*, 154104.
[230] D. M. Anstine, O. Isayev, *J. Phys. Chem. A* **2023**, *127*, 2417.
[231] X. Fu, A. Musaelian, A. Johansson, T. Jaakkola, B. Kozinsky, *arXiv* **2023**, arXiv:2310.13756.
[232] T. Jaffrelot Inizan, T. Plé, O. Adjoua, P. Ren, H. Gökcan, O. Isayev, L. Lagardère, J.-P. Piquemal, *Chem. Sci.* **2023**, *14*, 5438.

[233] L. Zhang, H. Wang, M. C. Muniz, A. Z. Panagiotopoulos, R. Car, W. E, *J. Chem. Phys.* **2022**, *156*, 124107.
[234] M. Gubler, J. A. Finkler, M. R. Schafer, J. Behler, S. Goedecker, *Journal of Chemical Theory, Computation* **2024**, *20*, 7264.
[235] Y. Shaidu, F. Pellegrini, E. Küçükbenli, R. Lot, S. de Gironcoli, *npj Comput. Mater.* **2024**, *10*, 47.
[236] A. Gao, R. C. Remsing, *Nat. Commun.* **2022**, *13*, 1572.
[237] C. G. Staacke, H. H. Heenen, C. Scheurer, G. Csányi, K. Reuter, J. T. Margraf, *ACS Appl. Energy Mater.* **2021**, *4*, 12562.
[238] Y. Ji, J. Liang, Z. Xu, *arXiv* **2025**, arXiv:2502.04668.
[239] E. J. Braham, J. M. Ruddock, J. O. Hardin, *Patterns* **2025**, 101340.
[240] K. P. Murphy, *Machine Learning: A Probabilistic Perspective*, MIT Press, Cambridge MA USA **2012**.
[241] H. I. Aysel, X. Cai, A. Prugel-Bennett, *Appl. Sci.* **2025**, *15*, 7261.
[242] A. Lavecchia, *Drug Discovery Today* **2024**, *29*, 104067.
[243] F. Fan, A. R. N. Aouichaoui, G. Sin, *Computer Aided Chemical Engineering* **2022**, *51*, 1393.
[244] J. Schmidt, L. Pettersson, C. Verdozzi, S. Botti, M. A. L. Marques, *Sci. Adv.* **2021**, *7*, abi7948.
[245] S. M. Lundberg, S.-I. Lee, *Advances in Neural Information Processing Systems* **2017**, *30*, 4765.
[246] Y. Wang, T. Zhang, X. Guo, Z. Shen, Gradient based Feature Attribution in Explainable AI: A Technical Review, *arXiv 2024, arXiv:2403.10415.*
[247] M. Raissi, P. Perdikaris, G. E. Karniadakis, *J. Comput. Phys.* **2019**, *378*, 686.
[248] S. L. Brunton, J. L. Proctor, J. N. Kutz, *Proc. Natl. Acad. Sci. USA* **2016**, *113*, 3932.
[249] E. S. Isbrandt, R. J. Sullivan, S. G. Newman, *Angew. Chem., Int. Ed.* **2019**, *58*, 7180.
[250] B. P. MacLeod, F. G. L. Parlane, T. D. Morrissey, F. Häse, L. M. Roch, K. E. Dettelbach, R. Moreira, L. P. E. Yunker, M. B. Rooney, J. R. Deeth, V. Lai, G. J. Ng, H. Situ, R. H. Zhang, M. S. Elliott, T. H. Haley, D. J. Dvorak, A. Aspuru-Guzik, J. E. Hein, C. P. Berlinguette, *Sci. Adv.* **2020**, *6*, aaz8867.
[251] F. Häse, L. M. Roch, P. Friederich, A. Aspuru-Guzik, *Nat. Commun.* **2020**, *11*, 4587.
[252] B. Burger, P M. Maffettone, V V. Gusev, C M. Aitchison, Y. Bai, X. Wang, X. Li, B M. Alston, B. Li, R. Clowes, N. Rankin, B. Harris, R. S. Sprick, A I. Cooper, *Nature* **2020**, *583*, 237.
[253] T. Kirchdoerfer, M. Ortiz, *Computer Methods in Applied Mechanics, Engineering* **2016**, *304*, 81.
[254] M. W. Libbrecht, W. S. Noble, *Nat. Rev. Genet.* **2015**, *16*, 321.
[255] N. Baker, Microsoft Azure Quantum Blog, https://azure.microsoft.com/en-us/blog/quantum/2024/01/09/unlocking-a-new-era-for-scientific-discovery-with-ai-how-microsofts-ai-screened-over-32-million-candidates-to-find-a-better-battery/ (accessed: January 2024).
[256] A. Ramirez, E. Lam, D. P. Gutierrez, Y. Hou, H. Tribukait, L. M. Roch, C. Copéret, P. Laveille, *Chemical Catalysis* **2024**, *4*, 100888.
[257] Y. Su, X. Wang, Y. Ye, Y. Xie, Y. Xu, Y. Jiang, C. Wang, *Chem. Sci.* **2024**, *15*, 12200.
[258] C. Ling, *npj Comput. Mater.* **2022**, *8*, 33.
[259] N. J. Szymanski, B. Rendy, Y. Fei, R E. Kumar, T. He, D. Milsted, M. J. McDermott, M. Gallant, E. D. Cubuk, A. Merchant, H. Kim, A. Jain, C. J. Bartel, K. Persson, Y. Zeng, G. Ceder, *Nature* **2023**, *624*, 86.
[260] F. Delgado-Licona, M. Abolhasani, *Advanced Intelligent Systems* **2022**, *5*, 2200331.
[261] P. Nikolaev, D. Hooper, N. Perea-López, M. Terrones, B. Maruyama, *npj Comput. Mater.* **2016**, *2*, 16031.
[262] Y. Ruan, C. Lu, N. Xu, Y. He, Y. Chen, J. Zhang, J. Xuan, J. Pan, Q. Fang, H. Gao, X. Shen, N. Ye, Q. Zhang, Y. Mo, *Nat. Commun.* **2024**, *15*, 10160.
[263] F. Häse, L. M. Roch, A. Aspuru-Guzik, *Chem. Sci.* **2018**, *9*, 7642.
[264] J. M. Reniers, D. A. Howey, *Appl. Energy* **2023**, *336*, 120774.
[265] D. Yang, Y. Cui, Q. Xia, *Materials* **2022**, *15*, 3331.
[266] M. Dubarry, D. Howey, B. Wu, *Joule* **2023**, *7*, 1134.
[267] A. D. Kies, J. Krauß, A. Schmetz, R. H. Schmitt, C. Brecher, *Proc. CIRP* **2022**, *107*, 1216.
[268] H. Tang, Y. Wu, Y. Cai, F. Wang, Z. Lin, Y. Pei, *J. Energy Storage* **2022**, *47*, 103679.
[269] Z. Wang, Z. Chen, B. Wang, C. Wu, C. Zhou, Y. Peng, X. Zhang, Z. Ni, C.-Y. Chung, C.-C. Chan, J. Yang, H. Zhao, *Appl. Energy* **2025**, *377*, 124120.
[270] Z. Hui, M. Wang, X. Yin, Y. N. Wang, Y. Yue, *Comput. Mater. Sci.* **2023**, *226*, 112215.
[271] S. Pescetelli, A. Agresti, G. Viskadouros, S. Razza, K. Rogdakis, I. Kalogerakis, E. Spiliarotis, E. Leonardi, P. Mariani, L. Sorbello, M. Pierro, C. Cornaro, S. Bellani, L. Najafi, B. Martín-García, A. E. Del Rio Castillo, R. Oropesa-Nuñez, M. Prato, S. Maranghi, M. L. Parisi, A. Sinicropi, R. Basosi, F. Bonaccorso, E. Kymakis, A. Di Carlo, *Nat. Energy* **2022**, *7*, 597.
[272] R. Tibshirani, *Journal of the Royal Statistical Society: Series B* **1996**, *58*, 267.
[273] L. Ward, A. Agrawal, A. Choudhary, C. Wolverton, *npj Comput. Mater.* **2016**, *2*, 16028.
[274] A. E. Hoerl, R. W. Kennard, *Technometrics* **1970**, *12*, 55.
[275] H. Zou, T. Hastie, *Journal of the Royal Statistical Society: Series B* **2005**, *67*, 301.
[276] D. W. Hosmer Jr, S. Lemeshow, R. X. Sturdivant, *Applied Logistic Regression*, 3rd ed., Wiley, Hoboken NJ USA **2013**.
[277] C. Cortes, V. Vapnik, *Machine Learning* **1995**, *20*, 273.
[278] B. Schölkopf, A. J. Smola, *Learning with Kernels*, MIT Press, Cambridge MA USA **2002**.
[279] G. Pilania, A. Mannodi-Kanakkithodi, B. P. Uberuaga, R. Ramprasad, J. E. Gubernatis, T. Lookman, *Sci. Rep.* **2016**, *6*, 19375.
[280] L. Breiman, *Machine Learning* **2001**, *45*, 5.
[281] P. Geurts, D. Ernst, L. Wehenkel, *Machine Learning* **2006**, *63*, 3.
[282] A. Seko, A. Togo, H. Hayashi, K. Tsuda, L. Chaput, I. Tanaka, *Phys. Rev. Lett.* **2015**, *115*, 205901.
[283] Y. Freund, R. E. Schapire, *Journal of Computer, System Sciences* **1997**, *55*, 119.
[284] T. Chen, C. Guestrin, in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery, Data Mining*, ACM, New York NY USA **2016**, 785.
[285] Y. Zhuo, A. M Tehrani, J. Brgoch, *J. Phys. Chem. Lett.* **2018**, *9*, 1668.
[286] L. Breiman, *Machine Learning* **1996**, *24*, 123.
[287] D. H. Wolpert, *Neural Networks* **1992**, *5*, 241.
[288] L. Breiman, *Machine Learning* **1996**, *24*, 49.
[289] K. Hornik, M. Stinchcombe, H. White, *Neural Networks* **1989**, *2*, 359.
[290] Y. LeCun, Y. Bengio, G. Hinton, *Nature* **2015**, *521*, 436.
[291] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, *Proceedings of the IEEE* **1998**, *86*, 2278.
[292] K. He, X. Zhang, S. Ren, J. Sun, in *Proceedings of the IEEE Conference on Computer Vision, Pattern Recognition*, IEEE, New York NY USA **2016**, 770.
[293] A. Ziletti, D. Kumar, M. Scheffler, L. M. Ghiringhelli, *Nat. Commun.* **2018**, *9*, 2775.
[294] T. N. Kipf, M. Welling, in *Proceedings of the 5th International Conference on Learning Representations*, Curran Associates, Inc, New York NY USA **2017**.
[295] S. Hochreiter, J. Schmidhuber, *Neural Computation* **1997**, *9*, 1735.
[296] A. Vaswani, N. Shazeer, N. Parmar, et al., in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, Curran Associates, Inc, New York NY USA **2017**, 5998.
[297] P. Schwaller, T. Laino, T. Gaudin, P. Bolgar, C. A. Hunter, C. Bekas, A. A. Lee, *ACS Cent. Sci.* **2019**, *5*, 1572.

[298] C. M. Bishop, *Pattern Recognition, Machine Learning*, Springer, New York NY USA **2006**.
[299] J. Shawe-Taylor, N. Cristianini, *Kernel Methods for Pattern Analysis*, Cambridge University Press, Cambridge UK **2004**.
[300] M. de Jong, W. Chen, R. Notestine, K. Persson, G. Ceder, A. Jain, M. Asta, A. Gamst, *Sci. Rep.* **2016**, *6*, 34256.
[301] M. Titsias, in *Proceedings of the 12th International Conference on Artificial Intelligence, Statistics* **2009**, *5*, 567.
[302] S. J. Pan, Q. Yang, *IEEE Transactions on Knowledge, Data Engineering* **2010**, *22*, 1345.
[303] M. Long, Y. Cao, Z. Cao, J. Wang, M. I. Jordan, *IEEE Transactions on Pattern Analysis, Machine Intelligence* **2018**, *41*, 3071.
[304] Y. Kim, Y. Kim, C. Yang, K. Park, G. X. Gu, S. Ryu, *npj Comput. Mater.* **2021**, *7*, 140.
[305] D. Cohn, L. Atlas, R. Ladner, *Machine Learning* **1994**, *15*, 201.
[306] R. Caruana, *Machine Learning* **1997**, *28*, 41.
[307] S. Ruder, *arXiv* **2017**, arXiv:1706.05098.
[308] R. Ramprasad, R. Batra, G. Pilania, A. Mannodi-Kanakkithodi, C. Kim, *npj Comput. Mater.* **2017**, *3*, 54.
[309] C. Finn, P. Abbeel, S. Levine, in *Proceedings of the 34th International Conference on Machine Learning* **2017**, 1126. PMLR.
[310] J. Snell, K. Swersky, R. Zemel, in *Proceedings of the 30th International Conference on Neural Information Processing Systems*, Curran Associates, Inc, New York NY USA **2017**, 4077.
[311] D. Jha, L. Ward, A. Paul, W. K. Liao, A. Choudhary, C. Wolverton, A. Agrawal, *npj Comput. Mater.* **2019**, *5*, 108.
[312] N. Cesa-Bianchi, G. Lugosi, *Prediction, Learning, Games*, Cambridge University Press, Cambridge UK **2006**.
[313] E. Hazan, *Introduction to Online Convex Optimization*, MIT Press, Cambridge MA USA **2016**.
[314] I. T. Jolliffe, *Principal Component Analysis*, 2nd ed., Springer, New York NY USA **2002**.
[315] L. van der Maaten, G. Hinton, *Journal of Machine Learning Research* **2008**, *9*, 2579.
[316] L. M. Ghiringhelli, J. Vybiral, S. V. Levchenko, C. Draxl, M. Scheffler, *Phys. Rev. Lett.* **2015**, *114*, 105503.
[317] S. Lloyd, *IEEE Trans. Inf. Theory* **1982**, *28*, 129.
[318] M. Ester, H. P. Kriegel, J. Sander, X. Xu, in *Proceedings of the 2nd International Conference on Knowledge Discovery, Data Mining*, ACM, New York NY USA **1996**, 226.
[319] Q. Zhu, A. Samanta, B. Li, R. E. Rudd, T. Frolov, *Nat. Commun.* **2018**, *9*, 467.
[320] D. P. Kingma, M. Welling, in *Proceedings of the 2nd International Conference on Learning Representations*, Curran Associates, Inc, New York NY USA **2014**.
[321] I. Goodfellow, J. Pouget-Abadie, M. Mirza, et al., in *Proceedings of the 27th International Conference on Neural Information Processing Systems*, Curran Associates, Inc, New York, NY, USA **2014**, 2672.
[322] B. Sanchez-Lengeling, A. Aspuru-Guzik, *Science* **2018**, *361*, 360.
[323] T. S. Ferguson, *Ann. Stat.* **1973**, *1*, 209.
[324] C. E. Antoniak, *Ann. Stat.* **1974**, *2*, 1152.
[325] D. M. Blei, M. I. Jordan, *Bayesian Anal.* **2006**, *1*, 121.
[326] S. Dasgupta, in *Proceedings of the 16th Conference on Uncertainty in Artificial Intelligence*, Morgan Kaufmann Publishers Inc, Waltham MA USA **2000**, p. 143.
[327] M. Brand, in *Proceedings of the 7th European Conference on Computer Vision*, Springer, Berlin Heidelberg **2002**, p. 707.
[328] J. Mairal, F. Bach, J. Ponce, G. Sapiro, *Journal of Machine Learning Research* **2010**, *11*, 19.
[329] R. S. Sutton, A. G. Barto, *Reinforcement Learning: An Introduction*, 2nd ed., MIT Press, Cambridge MA USA **2018**.
[330] V. Mnih, K. Kavukcuoglu, D. Silver, A A. Rusu, J. Veness, M G. Bellemare, A. Graves, M. Riedmiller, A K. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg, D. Hassabis, *Nature* **2015**, *518*, 529.
[331] Z. Zhou, S. Kearnes, L. Li, R. N. Zare, P. Riley, *Sci. Rep.* **2019**, *9*, 10752.
[332] M. Mohri, A. Rostamizadeh, A. Talwalkar, *Foundations of Machine Learning*, 2nd ed., MIT Press, Cambridge, MA, USA **2018**.
[333] F. A. Faber, A. Lindmaa, O. A. von Lilienfeld, R. Armiento, *Int. J. Quantum Chem.* **2015**, *115*, 1094.
[334] T. Hastie, R. Tibshirani, J. Friedman, *The Elements of Statistical Learning*, 2nd ed., Springer, New York, NY, USA **2009**.
[335] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov, *Journal of Machine Learning Research* **2014**, *15*, 1929.
[336] P. L. Bartlett, S. Mendelson, *Journal of Machine Learning Research* **2002**, *3*, 463.
[337] A. G. Wilson, R. P. Adams, in *Proceedings of ICML* **2013**, 1067, PMLR.
[338] J. Hensman, N. Fusi, N. D. Lawrence, arXiv 2013 arXiv:1309.6835.
[339] I. Goodfellow, Y. Bengio, A. Courville, *Deep Learning*, MIT Press, Cambridge, MA, USA **2016**.
[340] G. Niculescu-Mizil, R. Caruana, in *Proceedings of ICML* **2005**, 625. PMLR.
[341] C. Molnar, *Interpretable Machine Learning*, Lulu Press, Morrisville, NC, USA **2020**.
[342] A. Liaw, M. Wiener, *R News* **2002**, *2*, 18.
[343] T. G. Dietterich, *Lecture Notes in Computer Science* **2000**, *1857*, 1.
[344] C. Blundell, J. Cornebise, K. Kavukcuoglu, D. Wierstra, in *Proceedings of ICML* **2015**, 1613. PMLR.
[345] A. Zunger, *Nat. Rev. Chem.* **2018**, *2*, 0121.
[346] R. Hyndman, G. Athanasopoulos, *Forecasting: Principles, Practice*, 3rd ed., OTexts, Melbourne Australia **2021**.
[347] S. Raschka, *arXiv* **2018**, arXiv:1811.12808.
[348] H. Meyer, C. Reudenbach, T. Hengl, M. Katurji, T. Nauss, *Environmental Modelling & Software* **2018**, *101*, 1.
[349] C. X. Ling, J. Huang, H. Zhang, in *Proceedings of IJCAI* **2003**, 519. PMLR.
[350] B. Lakshminarayanan, A. Pritzel, C. Blundell, *Advances in Neural Information Processing Systems* **2017**, *30*, 6402.
[351] A. M. Deml, R. O'Hayre, C. Wolverton, V. Stevanović, *Phys. Rev. B* **2016**, *93*, 085142.
[352] C. C. Fischer, K. J. Tibbetts, D. Morgan, G. Ceder, *Nat. Mater.* **2006**, *5*, 641.
[353] B. Meredig, A. Agrawal, S. Kirklin, J. E. Saal, J. W. Doak, A. Thompson, K. Zhang, A. Choudhary, C. Wolverton, *Phys. Rev. B* **2014**, *89*, 094104.
[354] R. Ramakrishnan, P. O. Dral, M. Rupp, O. A. von Lilienfeld, *Scientific Data* **2014**, *1*, 140022.
[355] S. van der Walt, S. C. Colbert, G. Varoquaux, *Comput. Sci. Eng.* **2011**, *13*, 22.