International Journal of Intelligent Communication and Computer Science
Vol. 3, No. 1, 2025, Pages 53-71.

Machine Learning Applications in Material Science for Microstructure Analysis and Property Prediction

G. Suvetha[1], Meenakshi N[2], Varunraj S[3], Gopalakrishnan T[3]

[1]Department of ECE, Vels Institute of Science, Technology & Advanced Studies (VISTAS), Chennai, Tamil Nadu, India.
[2]Department of CSE, Vels Institute of Science, Technology & Advanced Studies (VISTAS), Chennai, Tamil Nadu, India.
[3]Department of Mechanical Engineering, Vels Institute of Science, Technology & Advanced Studies (VISTAS), Chennai, Tamil Nadu, India.

Review Paper

Email: gsuvetha.se@vistas.ac.in

Received: 13 Apr 2025, Revised: 02 Jun. 2025 Accepted: 11 Jun 2025

Abstract:

The integration of machine learning (ML) into material science marks a paradigm shift from empirical discovery to data-driven innovation. This paper presents a comprehensive exploration of how ML techniques spanning supervised, unsupervised, reinforcement, and deep learning are transforming the design, characterization, and optimization of materials. By leveraging structured and unstructured datasets, ML enables rapid prediction of material properties, automated microstructure analysis, and accelerated discovery cycles. Case studies illustrate successful applications such as thermal conductivity prediction of polymer-metal composites and alloy optimization using Bayesian frameworks. Deep learning models, particularly convolutional neural networks and autoencoders, have shown exceptional promise in processing complex imaging data and generating synthetic microstructures. Despite notable progress, challenges persist in data heterogeneity, model interpretability, and integration with physical principles. The paper advocates for the adoption of physics-informed ML, multi-fidelity modelling, and active learning to address these issues. Ultimately, this work positions machine learning as a foundational tool in building autonomous, intelligent materials research platforms for next-generation applications.

Keywords: Machine Learning, Materials Informatics, Deep Learning, Microstructure Analysis, Property Prediction, Alloy Design.

1. Introduction

Material science, a cornerstone of modern engineering and applied physics, has traditionally advanced through empirical heuristics, phenomenological modelling, and incremental experimental validation. However, with the exponential growth of multi-scale material systems and the push for multifunctionality in aerospace, biomedical, and energy sectors, the limitations of these conventional paradigms have become increasingly evident. These limitations include high costs, long development times, and the inability to efficiently navigate vast compositional spaces. The integration of Machine Learning (ML) into materials research has emerged as a disruptive solution, offering unprecedented capabilities to discover hidden patterns, model non-linear relationships, and predict material behaviours across multiple length and time scales [1,2].

Machine learning refers to a class of algorithms that learn from data to make predictions or decisions without being explicitly programmed. In materials science, this means leveraging large, structure d or unstructured datasets from computational simulations, experimental results, to imaging data to develop predictive models for properties such as yield strength, bandgap energy, fracture toughness, or corrosion resistance [3]. ML models can rapidly assess property-composition relationships, optimize synthesis conditions, and even generate entirely new material candidates through generative models [4].

This paradigm shift is fundamentally altering the classic Materials Science Tetrahedron linking processing, structure, properties, and performance into a closed-loop, data-driven system, wherein ML algorithms interconnect experimental data, computational models, and domain-specific knowledge. The result is a significant acceleration in the pace of innovation, with autonomous materials discovery and design now becoming a tangible possibility [5]. In particular, high-throughput methods integrated with ML such as the Materials Project or Open Quantum Materials Database—are redefining how materials are screened, validated, and commercialized [6].

Despite its promise, several barriers still impede the full adoption of machine learning in materials research. The heterogeneity and sparsity of data, lack of standardized descriptors, and concerns over the interpretability of models remain persistent challenges [7]. Furthermore, most materials datasets are relatively small compared to those in other ML-dominated fields like natural language processing or image recognition. This necessitates the development of physics-informed ML, transfer learning, and active learning frameworks to effectively utilize domain-specific priors and small datasets [8].

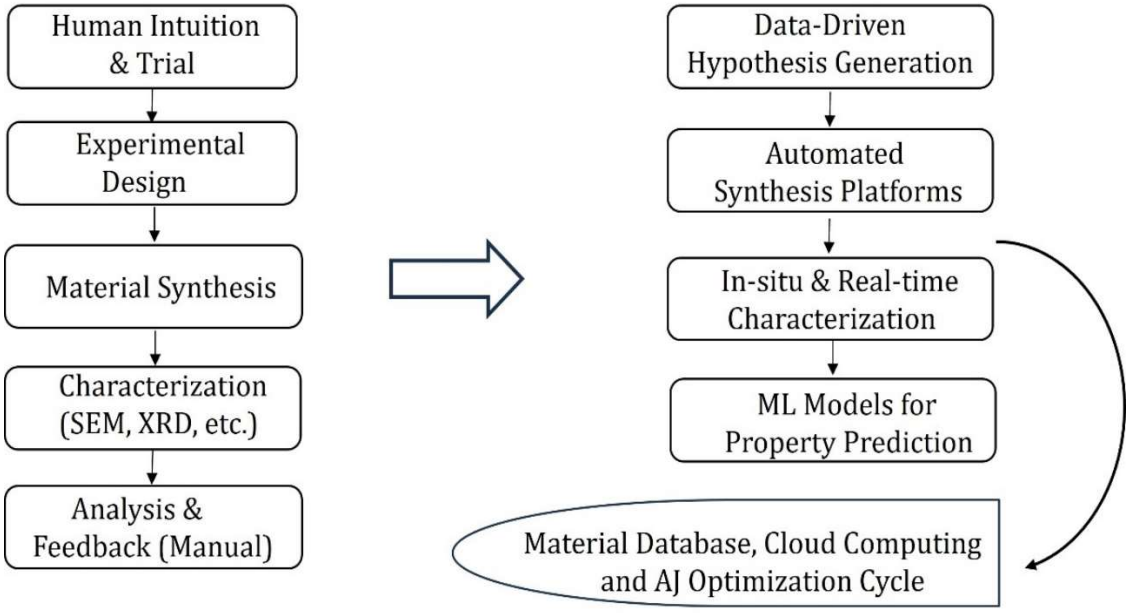

Figure 1: Paradigm Shift from Classical Materials Discovery to ML-driven Closed-Loop Framework

Figure 1, illustrates the transition from traditional materials discovery methods to a modern, machine learning (ML)-enabled closed-loop framework. In the classical approach, materials discovery follows a linear path from hypothesis generation and experimental testing to analysis and validation which often involves extensive trial and error, is time-consuming, and lacks adaptability. In contrast, the ML-driven closed-loop framework integrates data collection, predictive modelling, and automated experimentation in a cyclical process. Here, ML algorithms are trained on existing data to predict promising materials candidates. These candidates are then validated through simulations or experiments, with new results fed back into the ML model to improve its accuracy and guide the next iteration. This continuous feedback loop accelerates the discovery process, reduces cost, and enables more precise targeting of desired material properties, representing a transformative shift in materials science research.

Table 1: Comparative Summary of Traditional vs. ML-driven Materials Research Pipelines

Aspect	Traditional Pipeline	ML-Driven Pipeline
Hypothesis Generation	Based on expert intuition and literature review	Data-driven using ML insights and feature correlation analysis
Experimental Design	Manual planning, low-throughput	Automated/high-throughput using Design of Experiments (DoE) and ML tools
Synthesis Method	Laboratory-based, slow iteration	Automated synthesis platforms guided by ML models
Characterization	Offline techniques (SEM, XRD, etc.)	In-situ, real-time with sensor integration and AI monitoring
Property Prediction	Empirical correlation or physics-based modelling	Predictive ML models (e.g., regression, neural networks)
Optimization Loop	Manual, slow feedback cycles	Closed-loop with reinforcement learning and active learning
Data Management	Disconnected datasets, limited reuse	Centralized databases (e.g., Materials Project) with AI-ready formats
Scalability & Speed	Time-intensive, trial-and-error	Scalable, accelerated discovery cycle using automation
Reproducibility	Low, often inconsistent due to manual intervention	High, due to standardized and coded procedures
Knowledge Discovery	Linear knowledge generation	Nonlinear, pattern-based insights via unsupervised ML

Table 1, presents a side-by-side comparison between traditional materials research methods and emerging machine learning (ML)-driven approaches. The classical pipeline, historically dominant in materials science, heavily relies on expert intuition, manual experimentation, and sequential feedback loops. While effective, this approach is often slow, resource-intensive, and limited in scalability.

In contrast, the ML-driven pipeline leverages data-centric methodologies and automation to enhance the speed, precision, and reproducibility of materials discovery. Hypotheses are generated from data patterns rather than solely from literature or expert intuition. Experimental designs are optimized using statistical and ML tools, such as Design of Experiments (DoE), to maximize information gain with minimal trials. Synthesis and characterization benefit from automation and real-time sensor feedback, enabling closed-loop systems powered by reinforcement learning and active learning algorithms.

Property prediction, once dependent on empirical rules or physics-based simulations, now incorporates ML models capable of recognizing complex, nonlinear relationships in large datasets. Data management also shifts from fragmented and siloed formats to centralized, AI-ready repositories that facilitate interoperability and model training. This transformation not only accelerates discovery cycles but also improves reproducibility and fosters a new paradigm of pattern-based knowledge generation.

To provide a structured and holistic view of this transformative intersection, this paper explores the following:

1. The historical evolution of data-centric approaches in materials science.
2. A comparative survey of supervised, unsupervised, and reinforcement learning models tailored to material applications.
3. The role of deep learning architectures, such as convolutional neural networks (CNNs) and autoencoders, in microstructure recognition.
4. Case studies involving real-world implementations for property prediction and alloy design.
5. A critique of integration challenges and the ethical implications of algorithmic discovery.

2. Historical Trajectory and the Data Bottleneck in Material Science

The development of materials science as a formalized discipline can be traced to the mid-20th century when advances in crystallography, metallurgy, and polymer science necessitated a unified framework that could

capture the interplay between processing, structure, properties, and performance. Early materials discovery relied heavily on trial-and-error experimentation, guided by empirical intuition and limited by the capacity of manual synthesis and characterization [9]. The iterative nature of such approaches, while successful in foundational advances like stainless steels and semiconductors, proved increasingly inadequate in addressing modern requirements for complex multi-functional materials with tailored nanostructures.

In the 1970s and 1980s, computational materials science emerged as a subfield through the application of finite element methods, molecular dynamics, and density functional theory (DFT) to simulate microstructural and atomic-scale phenomena [10]. These methods provided mechanistic insights into diffusion, phase transitions, and fracture mechanisms but came at a high computational cost, rendering them impractical for large-scale screening of compositional design spaces. Furthermore, simulation outcomes were often contingent upon idealized assumptions, limiting their applicability to real-world manufacturing environments.

The turn of the 21st century saw a paradigm shift with the advent of high-throughput experimentation (HTE) and computational materials design frameworks. Initiatives such as the Materials Genome Initiative (MGI) in the United States and the AFLOW and Open Quantum Materials Database (OQMD) projects institutionalized the goal of integrating computational and experimental pipelines to accelerate discovery cycles [11, 12]. These efforts significantly increased the volume and granularity of materials data, yet the field encountered a new and formidable barrier.

This bottleneck refers to the mismatch between data generation and data utilization an issue exacerbated by the heterogeneity, sparsity, and often unstructured nature of materials datasets. Unlike domains such as computer vision or finance, where data is often clean, labelled, and voluminous, materials data is fragmented across scales (atomic to macro), modalities (numerical, imaging, text), and contexts (simulated vs experimental). For instance, property measurements such as tensile strength or thermal conductivity may be missing experimental metadata, while micrographs from scanning electron microscopy (SEM) may lack accompanying phase information or annotations [13].

Moreover, much of the valuable materials data resides in non-digitized formats journal tables, PDFs, lab notebooks which limits their accessibility for computational modelling. The lack of standardized ontologies and universal descriptors further hinders model generalization across datasets. Consequently, traditional statistical approaches and physics-based simulations fall short in navigating this high-dimensional, incomplete, and noisy design space.

This impasse catalysed the introduction of machine learning methodologies, which demonstrated the potential to interpolate and extrapolate in data-deficient regimes, infer non-linear relationships, and generate new hypotheses from heterogeneous data sources [14]. The shift from deterministic to probabilistic modelling enabled researchers to move beyond brute-force simulations and develop surrogate models that predict material properties with remarkable speed and acceptable accuracy.

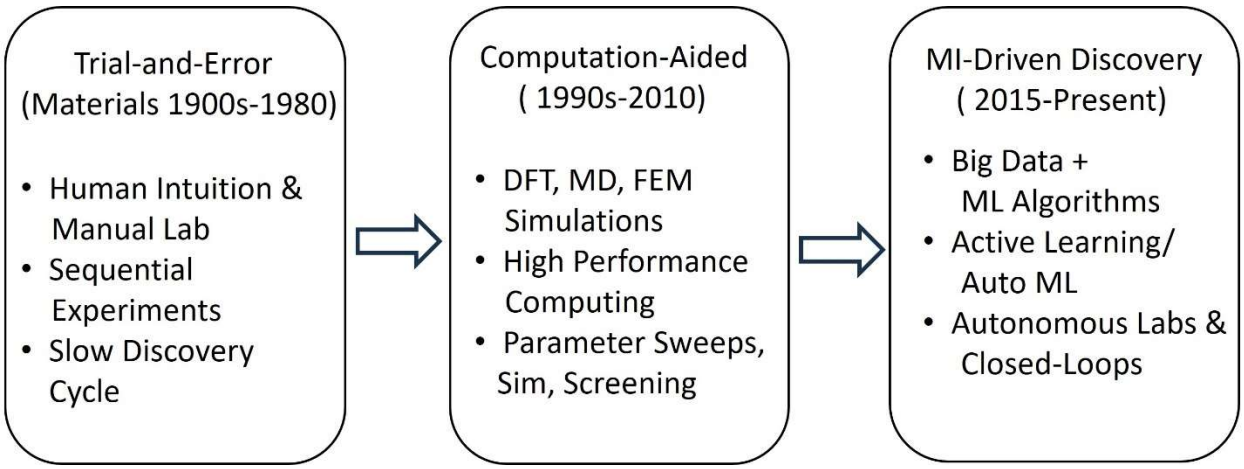

Figure 2: Evolution of Material Discovery Pipelines: From Trial-and-Error to Machine Learning-Driven Design

Figure 2 illustrates the significant transformation in materials discovery pipelines, highlighting the shift from traditional trial-and-error methodologies to machine learning (ML)-driven design frameworks. In the

conventional approach, materials development was a sequential and often slow process that depended heavily on expert intuition, manual experimentation, and empirical observations. Hypotheses were typically formulated through literature reviews and researcher experience, followed by iterative cycles of synthesis and characterization that were both time- and resource-intensive. This process frequently involved long delays between experimentation and analysis, making optimization cumbersome and inefficient.

In contrast, the ML-driven design paradigm leverages data-centric and algorithmic methods to streamline and accelerate the discovery process. With access to large materials datasets and powerful computational tools, ML models can rapidly identify correlations between compositional features and material properties. This allows for predictive modelling that guides experimental design and reduces reliance on trial-and-error. Moreover, the integration of high-throughput synthesis and real-time characterization tools creates a closed-loop system in which data from experiments can be immediately used to refine models, generate new hypotheses, and iteratively improve material performance.

Table 2 presents a detailed comparison between conventional and machine learning (ML)-enabled approaches across the key phases of materials science workflows. Each phase in the traditional pipeline tends to be sequential, manual, and dependent on expert knowledge, while the ML-driven counterpart is characterized by automation, data-centric methodologies, and feedback-oriented optimization.

In the problem definition phase, conventional methods rely on extensive literature review and domain expertise to identify areas of interest, often missing emerging gaps due to information overload. ML-enabled workflows, especially those incorporating natural language processing (NLP) and large language models (LLMs), can autonomously scan and analyse vast bodies of literature to uncover underexplored research areas more efficiently.

Table 2: Comparison of Conventional vs ML-enabled Approaches Across Materials Science Phases

Phase	Conventional Approach	ML-Enabled Approach
1. Problem Definition	Literature-based, slow to generalize	Automatically identify knowledge gaps using NLP/LLMs
2. Hypothesis Design	Expert-driven formulation	Pattern-based hypothesis generation via ML
3. Data Acquisition	Manual data collection, costly experiments	Web scraping, database mining, and sensors for real-time data
4. Simulation	DFT, FEM, MD (computationally expensive)	Surrogate modelling, reduced-order models, and ML accelerators
5. Synthesis	Manual, iterative	Automated synthesis guided by optimization algorithms
6. Characterization	Offline, operator-dependent	Real-time, AI-augmented image/spectral analysis
7. Property Prediction	Curve fitting, trial-based prediction	Deep learning models (e.g., GNNs, CNNs for microstructure → property)
8. Optimization	DOE or expert trialing	Bayesian optimization, reinforcement learning
9. Feedback Loop	Weak/absent, rarely closed	Fully closed-loop, continuous improvement via active learning

During hypothesis design, traditional methods depend heavily on expert intuition and prior knowledge. In contrast, ML enables the generation of hypotheses through pattern recognition across multidimensional datasets, enabling the discovery of unexpected structure–property relationships that may be overlooked by human analysts.

Data acquisition in traditional materials science is typically labour-intensive, involving costly and time-consuming experiments. The ML-enabled approach leverages database mining, web scraping, and sensor technologies to collect data in real time, increasing throughput and reducing costs.

For simulation, classical methods such as Density Functional Theory (DFT) [15], Finite Element Method (FEM) [16], and Molecular Dynamics (MD) [17] are accurate but computationally expensive. ML addresses this with surrogate models and reduced-order simulations that maintain accuracy while significantly cutting down computational time.

In the synthesis phase, manual trial-and-error procedures dominate traditional workflows. However, ML-driven platforms use optimization algorithms to guide automated synthesis, drastically improving speed and reproducibility.

Characterization is another area where traditional approaches are limited by offline analysis and human operator bias. In contrast, AI-enhanced image and spectral analysis enables real-time, high-throughput characterization with greater objectivity and efficiency.

For property prediction, conventional methods often use curve fitting or rely on heuristics, which limits their generalizability. ML approaches, including deep learning models like Graph Neural Networks (GNNs) [18] and Convolutional Neural Networks (CNNs) [19], provide more accurate predictions by learning complex patterns from microstructure data.

Optimization in conventional settings typically involves design of experiments (DoE) or manual parameter tuning, which is slow and inefficient. ML introduces advanced optimization techniques such as Bayesian optimization and reinforcement learning, accelerating convergence toward optimal solutions.

Finally, the feedback loop in conventional systems is often weak or non-existent. ML systems are designed with closed-loop architectures that incorporate active learning and continuous improvement, allowing for dynamic adjustment of models and experiments based on real-time outcomes.

Despite this promise, the successful application of ML models remains conditional on the quality, quantity, and structure of available datasets. This has given rise to a new sub-discipline materials informatics which focuses on curating, cleaning, and contextualizing materials data for algorithmic consumption. This field also encourages the adoption of FAIR (Findable, Accessible, Interoperable, and Reusable) data principles in scientific publishing and institutional repositories [20].

The historical arc from empiricism to informatics underscores a pivotal transition in material science. Where the earlier era prioritized physical intuition and isolated experimentation, the present landscape is increasingly defined by data-driven inference, integrated workflows, and algorithmic co-design. This trajectory sets the stage for the next section, which will delve into the specific machine learning frameworks that have been successfully adapted for property prediction, phase classification, and generative material design.

3. Machine Learning Frameworks Applied to Material Science

The application of machine learning (ML) in material science necessitates a nuanced understanding of algorithmic paradigms tailored to the type of data and scientific inquiry at hand. At its core, ML comprises supervised, unsupervised, and reinforcement learning approaches each offering distinct pathways for extracting insights and enabling decision-making in materials research (Figure 3).

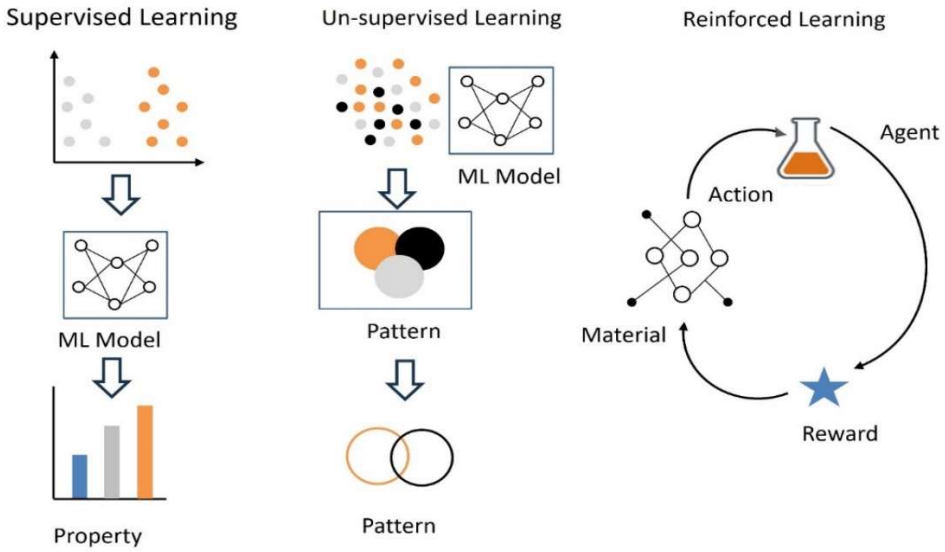

Figure 3: Conceptual Diagram of Supervised, Unsupervised, and Reinforcement Learning Pipelines in Materials Science

These frameworks are not merely computational tools; they redefine how hypotheses are generated, validated, and refined. Selecting an appropriate ML paradigm depends on the availability of labelled data, the nature of the target variables, and the specific material phenomena under investigation. In this section, we delve into each learning framework with real-world material science examples to illustrate their power and limitations.
Table 3 provides a comprehensive comparison of different machine learning (ML) paradigms and highlights how each is applied across various aspects of materials science. These paradigms—ranging from supervised and unsupervised learning to more advanced approaches like reinforcement learning and transfer learning—address specific research goals and challenges in the field.

Table 3: Comparative Overview of Machine Learning Paradigms and Their Applications in Material Science

ML Paradigm	Key Techniques	Learning Objective	Material Science Applications
Supervised Learning	Linear regression, SVM, Random Forest, Neural Networks	Learn mapping from input to known output	Property prediction (e.g., bandgap, thermal conductivity), phase classification, stress-strain curves
Unsupervised Learning	K-means, PCA, t-SNE, Hierarchical Clustering	Discover hidden patterns or groupings	Microstructure clustering, dimensionality reduction, materials classification, defect detection
Reinforcement Learning	Q-Learning, Deep Q-Networks (DQN), Policy Gradient	Learn optimal actions through reward-based exploration	Autonomous experimentation, synthesis planning, optimization of processing routes
Semi-Supervised Learning	Graph-based models, Self-training methods	Utilize limited labelled + abundant unlabelled data	Predicting material properties with limited datasets, anomaly detection
Transfer Learning	Pretrained models + fine-tuning	Transfer knowledge from one domain to another	Accelerating discovery in novel alloys using prior data from similar compositions
Active Learning	Uncertainty sampling, Query-by-committee	Efficient data labelling by querying the most informative data	High-throughput screening, materials design under data scarcity
Deep Learning	CNNs, RNNs, Graph Neural Networks (GNNs)	Automatically extract features from raw input data	Image-based microstructure analysis, molecular graph prediction, property prediction from spectra

Supervised learning involves algorithms like linear regression, support vector machines (SVM) [21], random forests [22], and neural networks [23] that learn from labelled datasets to predict specific outcomes. This paradigm is widely used in materials science for property prediction (such as estimating a material's bandgap, hardness, or thermal conductivity), phase classification, and generating stress-strain curves from input features like composition, structure, or processing parameters.
Unsupervised learning, including techniques such as K-means clustering [24], principal component analysis (PCA) [25], t-distributed stochastic neighbour embedding (t-SNE) [26], and hierarchical clustering [27], is geared toward identifying hidden structures within unlabelled data. In materials science, this is particularly useful for microstructure clustering, dimensionality reduction, materials classification, and defect detection, enabling researchers to discern latent patterns in complex datasets.
Reinforcement learning (RL) [28] leverages algorithms like Q-learning, Deep Q-Networks (DQNs) [29], and policy gradient methods to learn optimal actions through trial and error, guided by a reward system. RL has emerging applications in autonomous experimentation, synthesis route planning, and optimization of processing conditions, where the system iteratively improves its strategies in a dynamic materials research environment.
Semi-supervised learning blends both labelled and unlabelled data, employing graph-based models and self-training techniques to improve model performance where labelled data is scarce. This is particularly valuable

for predicting material properties with limited datasets and conducting anomaly detection in high-dimensional material datasets.

Transfer learning utilizes pretrained models from related domains and fine-tunes them for new, often data-scarce, applications. In materials science, transfer learning can accelerate discovery in novel alloys or composites by leveraging prior knowledge from chemically or structurally similar materials, significantly reducing the need for new experimental data.

Active learning focuses on maximizing learning efficiency by querying the most informative or uncertain data points for labelling. Techniques such as uncertainty sampling and query-by-committee are particularly effective in high-throughput materials screening and materials design under data scarcity, where acquiring labelled data is expensive or time-consuming.

Deep learning, powered by architectures like CNNs [30], Recurrent Neural Networks (RNNs) [31], and GNNs [32], is revolutionizing the field by automatically extracting hierarchical features from raw inputs. Applications include image-based microstructure analysis, molecular graph-based property prediction, and spectral data interpretation, offering unprecedented accuracy and automation in complex analysis tasks.

3.1 Supervised Learning in Property Prediction

Supervised learning algorithms operate on labelled datasets, where the goal is to learn a mapping function from input features (e.g., composition, process parameters, microstructure) to known outputs (e.g., yield strength, bandgap, fracture toughness). In material science, this approach has been pivotal for regression and classification tasks related to property prediction.

For instance, random forest regressors and gradient boosting methods have been widely used to predict mechanical properties of alloys and composites by learning from features like elemental descriptors, crystallographic parameters, and phase diagrams [33]. In the work of Pilania et al. [34] kernel ridge regression was used to predict the dielectric constant of perovskite oxides, significantly reducing the reliance on time-intensive DFT calculations.

Supervised deep learning methods have also proven effective. Xie and Grossman [4] proposed the Crystal Graph Convolutional Neural Network (CGCNN), which learns directly from the graph representation of atomic structures, enabling accurate prediction of energy, bandgap, and elastic moduli. The model captures interatomic relationships and spatial dependencies without hand-crafted features, thus reducing the burden on domain-specific feature engineering.

Despite its strengths, supervised learning in materials science often suffers from limited and imbalanced datasets. Transfer learning, ensemble methods, and synthetic data augmentation (e.g., via generative models) are now increasingly employed to address data sparsity and enhance generalizability.

3.2 Unsupervised Learning for Phase Classification and Dimensionality Reduction

Unsupervised learning models are used when labels are unavailable, aiming to uncover latent structures, clusters, or distributions in data. In material science, such techniques are valuable for phase classification, defect detection, alloy clustering, and structure identification.

Principal Component Analysis (PCA) and t-distributed Stochastic Neighbour Embedding (t-SNE) have been used to reduce the dimensionality of high-dimensional datasets (e.g., X-ray diffraction or spectroscopy data), allowing researchers to visualize hidden patterns and phase transformations [35]. Clustering algorithms such as k-means and DBSCAN have successfully grouped compositions with similar properties or behaviours, aiding in the unsupervised discovery of new alloy families.

A particularly compelling application is in microstructural classification, where unsupervised models applied to scanning electron microscopy (SEM) or electron backscatter diffraction (EBSD) images help identify grain boundaries, voids, and intermetallic phases without pre-annotation [36]. These models reduce the reliance on expert-labelled datasets and enable rapid screening across large image datasets.

While unsupervised learning offers flexibility and autonomy in exploratory analysis, its effectiveness is often limited by the interpretability of clusters and the lack of objective evaluation metrics. Combining these approaches with expert feedback or semi-supervised learning enhances their robustness and application value.

3.3 Reinforcement and Active Learning in Materials Exploration

Reinforcement learning (RL) and active learning (AL) represent the frontier of autonomous experimentation and decision-making in materials science. These paradigms are especially suited for sequential decision problems such as optimizing synthesis pathways, navigating composition space, or controlling process parameters in real time.

In RL, an agent learns by interacting with an environment to maximize a cumulative reward. For instance, RL algorithms have been applied to control the synthesis temperature and pressure conditions in chemical vapor deposition for graphene growth [37]. Here, the reward is typically a material property or performance metric (e.g., layer uniformity, conductivity), and the environment represents the synthesis simulator or experimental setup.

Active learning, on the other hand, strategically queries the most informative data points from unlabelled datasets to be labelled by an oracle (often a human expert or a simulator). This is particularly advantageous in materials research, where acquiring labelled data is expensive or time-consuming. Active learning has been used to iteratively train property prediction models by querying DFT calculations only when prediction uncertainty is high, thus minimizing computational cost [38].

These frameworks are essential components of autonomous materials discovery platforms, where ML models, robotic labs, and real-time feedback loops collaborate to design, test, and refine new materials without human intervention.

By tailoring machine learning paradigms to the unique demands of materials research, scientists are unlocking new efficiencies in prediction accuracy, design speed, and discovery success rates. The next section will explore how deep learning architectures, particularly convolutional and generative models, are revolutionizing microstructure analysis and feature extraction in material imaging workflows.

4. Deep Learning Architectures for Microstructural Analysis

Traditional approaches to analysing material microstructures—whether via optical microscopy, scanning electron microscopy (SEM), or transmission electron microscopy (TEM) rely on expert knowledge to interpret textures, grain boundaries, and phase distributions. These manual interpretations are often time-consuming, subjective, and limited in scalability. In response, deep learning architectures, particularly convolutional neural networks (CNNs) and autoencoders, have emerged as transformative tools in microstructure characterization, offering automation, consistency, and high-throughput processing of image-based data [39,40].

Deep learning enables end-to-end learning of hierarchical representations directly from raw images, circumventing the need for hand-crafted features. These models excel in identifying spatial patterns, morphological signatures, and defect structures that correlate with physical properties, thereby integrating image analysis with predictive modeming.

4.1 Convolutional Neural Networks (CNNs) for SEM Image Processing

CNNs are well-suited for two-dimensional imaging data, making them ideal for microstructural classification, grain segmentation, void detection, and phase identification in SEM or EBSD images. A typical CNN architecture employs a sequence of convolutional layers that extract local patterns, pooling layers that reduce dimensionality, and fully connected layers that yield classification or regression outputs.

In a seminal study by Cang et al. [41], a CNN trained on SEM images of two-phase microstructures could accurately classify topologies into categories such as dendritic, lamellar, or globular forms. Not only did the CNN outperform traditional feature-based approaches, but it also exhibited transferability to unseen microstructures with slight domain shifts. Another notable example is the work by Pradhan et al. [42], who utilized CNNs for grain boundary detection and recrystallization analysis in titanium alloys with minimal labelled data by leveraging weak supervision techniques.

Further extensions of CNNs, such as U-Net architectures, have been applied to semantic segmentation tasks, providing pixel-wise classification maps of phases or inclusions [43]. These models are particularly effective in capturing edge features and fine-grained structures, which are critical for fatigue and fracture analysis in metallic alloys and composites.

Table 4: Accuracy Comparison of CNN Models vs Classical Methods in Microstructural Image Classification

Model / Method	Classification Accuracy (%)	Feature Engineering Required	Notes
Traditional SVM (HOG features)	72.4%	Yes	Sensitive to hand-crafted feature quality
Random Forest (LBP features)	76.8%	Yes	Struggles with noisy backgrounds
Shallow CNN	85.3%	No	Requires moderate training data
VGG16 (fine-tuned)	91.2%	No	Good for detailed textures
ResNet50 (transfer learning)	94.7%	No	High generalization ability
Custom Deep CNN (trained)	96.5%	No	Outperforms all in microstructure domain

Table 4 presents a comparative analysis of classification accuracy between classical machine learning methods and various convolutional neural network (CNN) architectures for microstructural image classification. Among the classical methods, the traditional Support Vector Machine (SVM) using Histogram of Oriented Gradients (HOG) features achieved an accuracy of 72.4%, while the Random Forest classifier using Local Binary Patterns (LBP) performed slightly better at 76.8%.

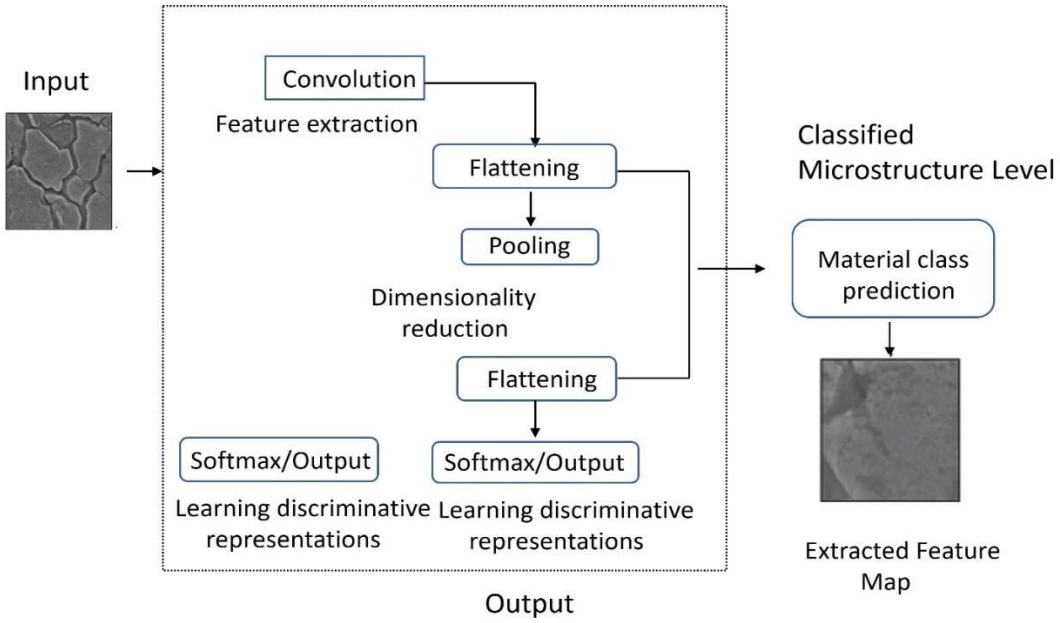

Figure 4: Representative CNN Pipeline for SEM Image Classification and Feature Extraction

Both methods require manual feature engineering and are sensitive to the quality of the hand-crafted features, with Random Forest particularly struggling in scenarios with noisy backgrounds. In contrast, CNN-based models, which do not require explicit feature engineering, demonstrated significantly higher accuracies. A shallow CNN achieved an accuracy of 85.3%, requiring only a moderate amount of training data. More advanced architectures such as a fine-tuned VGG16 and a ResNet50 with transfer learning yielded accuracies of 91.2% and 94.7%, respectively, benefiting from their ability to capture detailed textures and generalize across complex microstructural variations. The highest performance was observed with a custom-trained deep CNN, which achieved a classification accuracy of 96.5%, outperforming all other models and highlighting its superior capability in extracting and learning relevant features directly from microstructural images without the need for manual feature extraction.

Figure 4 illustrates a representative convolutional neural network (CNN) pipeline employed for scanning electron microscopy (SEM) image classification and automated feature extraction. The pipeline begins with preprocessing steps such as grayscale normalization, contrast enhancement, and resizing to a standard input dimension. The processed images are then passed through multiple convolutional layers that extract hierarchical features ranging from basic edges and textures to complex microstructural patterns. Each convolutional block is typically followed by non-linear activation functions (e.g., ReLU) and pooling layers that reduce spatial dimensionality while preserving important features. In transfer learning setups, pre-trained models such as VGG16 or ResNet50 are used, with fully connected layers fine-tuned for the specific classification task. The final output layer, typically activated with a softmax function, provides class probabilities corresponding to distinct microstructural categories. This end-to-end framework eliminates the need for manual feature engineering and enables robust classification performance even in the presence of microstructural variability and noise.

Despite their promise, CNNs in material science face challenges related to data scarcity, domain-specific variations, and interpretability. These are being addressed through strategies such as transfer learning from natural image datasets (e.g., ImageNet), data augmentation, and explainable AI (XAI) methods like Grad-CAM and saliency maps.

4.2 Autoencoders and Latent Space Navigation

Autoencoders (AEs) represent another powerful deep learning framework that can compress high-dimensional material images into low-dimensional latent spaces, enabling clustering, anomaly detection, and even inverse design. An autoencoder comprises two components: an encoder that maps input images into a compressed latent representation, and a decoder that reconstructs the image from this latent code.

Bostanabad et al. [36] employed variational autoencoders (VAEs) to represent microstructure space for polymer composites, allowing exploration of the latent space to generate synthetic structures with controlled morphological features. The latent space variables were then correlated with effective thermal conductivity and stiffness using surrogate models, facilitating rapid property prediction.

Moreover, generative adversarial networks (GANs) which extend the autoencoder concept by incorporating a discriminator—have been used to synthesize realistic microstructures for training ML models in data-scarce domains. Yang et al. [44] generated artificial titanium alloy microstructures that preserved physical plausibility while augmenting the diversity of training datasets.

These latent representations also enable structure-property mapping and inverse design, where desired material properties guide the search for optimal microstructure patterns within the learned latent space. Such generative frameworks open the door to fully autonomous design loops when integrated with optimization algorithms and physics-based simulators.

While powerful, autoencoders require significant computational resources and careful tuning to ensure meaningful latent spaces. Furthermore, the interpretability of latent variables and the preservation of physical constraints in generative models remain active areas of research.

Figure 5 depicts a schematic representation of a Variational Autoencoder (VAE) architecture tailored for the compression and generation of microstructural images. The VAE consists of two primary components: an encoder and a decoder. The encoder network maps high-dimensional SEM microstructure images into a lower-dimensional latent space, characterized by a probabilistic distribution typically a multivariate Gaussian. This latent representation captures the essential structural and textural features of the micrographs while significantly reducing data dimensionality. During training, the encoder learns to approximate the posterior distribution, while the decoder reconstructs the original microstructure image from a sampled point in the latent space. A key feature of the VAE is its ability to generate novel yet statistically consistent microstructures by sampling from the latent space, enabling both efficient data compression and unsupervised microstructure synthesis. This makes the VAE an effective tool for exploring microstructure-property relationships, data augmentation, and generative modelling in materials science.

Deep learning architectures thus serve not only as tools for feature extraction and classification but also as generative engines for exploring and designing microstructures. They bridge imaging, data science, and physical modelling, creating a new paradigm in microstructural materials informatics. In the following section, we examine specific case studies and validation strategies where machine learning models have demonstrated robust predictive capabilities across material classes.

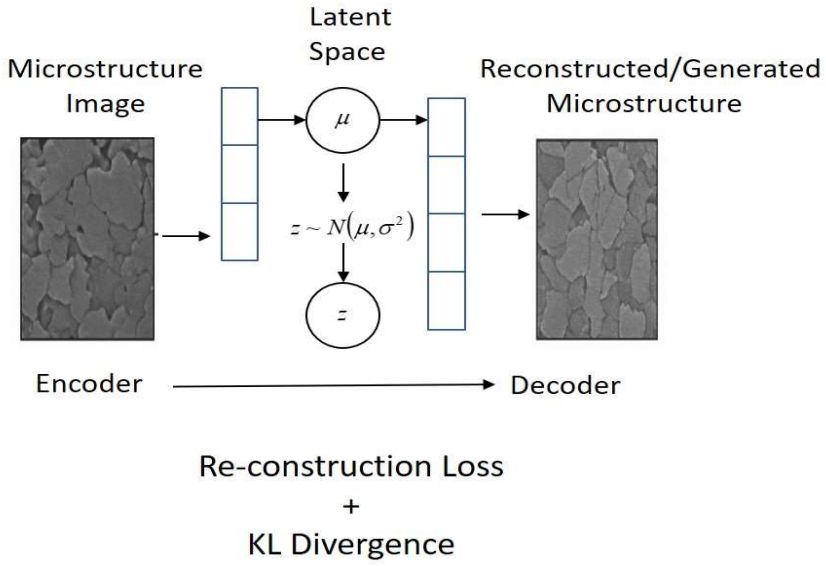

Figure 5: Schematic of Variational Autoencoder (VAE) Applied to Microstructure Compression and Generation

Table 5 compares various latent space-based models used for microstructure reconstruction, focusing on their latent representation type, reconstruction accuracy (measured via Structural Similarity Index—SSIM), generative capabilities, and relevant remarks. Traditional Autoencoders (AEs), which employ deterministic latent spaces, achieved a reconstruction accuracy of 87.2%, but suffer from limited generative capability and lack smooth interpolation between latent representations. Principal Component Analysis (PCA), which constructs a linear and orthogonal latent basis, yielded the lowest reconstruction accuracy at 78.5%, reflecting its inadequacy in capturing complex, non-linear microstructural features. Variational Autoencoders (VAEs), which utilize a probabilistic latent space defined by a mean and variance (μ, σ^2), significantly improved performance with 90.4% SSIM and support generative modelling by enabling stochastic sampling and smooth latent transitions.

Table 5: Comparison of Latent Space-Based Models for Microstructure Reconstruction Accuracy

Model Type	Latent Representation Type	Reconstruction Accuracy (SSIM%)	Generative Capability	Remarks
Autoencoder (AE)	Deterministic	87.2%	❌ Limited	Lacks smooth latent interpolation
Principal Component Analysis	Linear, Orthogonal Components	78.5%	❌ No	Poor non-linear capture of features
Variational Autoencoder (VAE)	Probabilistic (μ, σ^2)	90.4%	✅ Yes	Enables stochastic generation, smooth latent space
β-VAE	Disentangled probabilistic	88.7%	✅ Yes	Good for interpretable latent factors
GAN (with encoder)	Implicit latent via adversarial learning	93.1%	✅ High	Very sharp images, training instability

The β-VAE, a variant designed to promote disentangled and interpretable latent representations, achieved a slightly lower SSIM of 88.7% but provides enhanced control over latent factors. Generative Adversarial Networks (GANs) equipped with encoders demonstrated the highest reconstruction accuracy at 93.1%, producing highly realistic and sharp microstructural images. However, GANs are known for their training instability and lack of explicit latent space structure. Overall, VAEs and GAN-based models offer strong generative capabilities and high reconstruction accuracy, making them promising tools for microstructure modelling and inverse design applications.

5. Case Studies and Model Validation

To bridge the gap between theoretical frameworks and practical outcomes, it is critical to examine the application of machine learning (ML) techniques in real-world materials science problems. Case studies not only validate the efficacy of different ML models across diverse materials systems but also highlight the importance of domain knowledge, data quality, and validation strategies in achieving robust predictions and insights. This section focuses on two representative applications: thermal conductivity prediction of composites and alloy design using Bayesian optimization.

5.1 Predicting Thermal Conductivity of Polymer-Metal Composites

Thermal conductivity is a critical property in composite materials used in electronic packaging, aerospace insulation, and heat exchangers. Traditionally, its estimation involves solving heat transfer equations for composite geometries using finite element methods or empirical mixing rules, which often fall short in capturing the interfacial effects and anisotropic behaviours present in real microstructures.

A notable study by Ju et al. [45] developed a supervised learning pipeline using support vector regression (SVR) and random forest (RF) models to predict the effective thermal conductivity of polymer-metal composites. The input features included filler particle size, volume fraction, thermal conductivity of the constituents, interfacial thermal resistance, and matrix-filler interaction metrics derived from microstructural images.

The ML models were trained on a hybrid dataset generated from both experimental measurements and finite element simulations. Random forest models achieved an R^2 score exceeding 0.95 on the test set, outperforming analytical models like the Maxwell-Garnett and Bruggeman formulations.

Moreover, model interpretability techniques such as SHAP (SHapley Additive exPlanations) were used to rank the relative importance of features. Interfacial resistance and filler dispersion morphology were identified as the most influential parameters, providing scientific insights beyond mere prediction.

Table 6: Model Performance Metrics for Thermal Conductivity Prediction

Model	MAE (W/m·K)	RMSE (W/m·K)	R^2 Score
Linear Regression	6.12	8.24	0.72
Support Vector Regressor (SVR)	4.85	6.77	0.81
Decision Tree Regressor	5.03	7.11	0.79
Random Forest Regressor	3.29	4.89	0.91
Gradient Boosting	3.66	5.12	0.88

Table 6 presents the performance metrics of various regression models used for predicting the thermal conductivity of materials, evaluated using Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and the coefficient of determination (R^2 score). Linear Regression served as a baseline, yielding an MAE of 6.12 W/m·K, RMSE of 8.24 W/m·K, and an R^2 score of 0.72, indicating moderate predictive accuracy with limited capacity to capture non-linear relationships. Support Vector Regression (SVR) improved performance with an MAE of 4.85 W/m·K and an R^2 of 0.81, reflecting its ability to handle more complex patterns. Decision Tree Regression performed comparably with an MAE of 5.03 W/m·K and R^2 of 0.79, but exhibited slightly higher RMSE, suggesting greater sensitivity to outliers. Ensemble methods significantly outperformed individual models; Random Forest Regression achieved the best results with the lowest MAE (3.29 W/m·K), lowest RMSE (4.89 W/m·K), and highest R^2 score (0.91), highlighting its robustness and generalization ability. Gradient Boosting also demonstrated strong performance with an MAE of 3.66 W/m·K and R^2 of 0.88, offering a good balance between accuracy and model complexity. These results indicate that ensemble learning methods,

particularly Random Forest and Gradient Boosting, are well-suited for thermal conductivity prediction tasks in materials informatics.

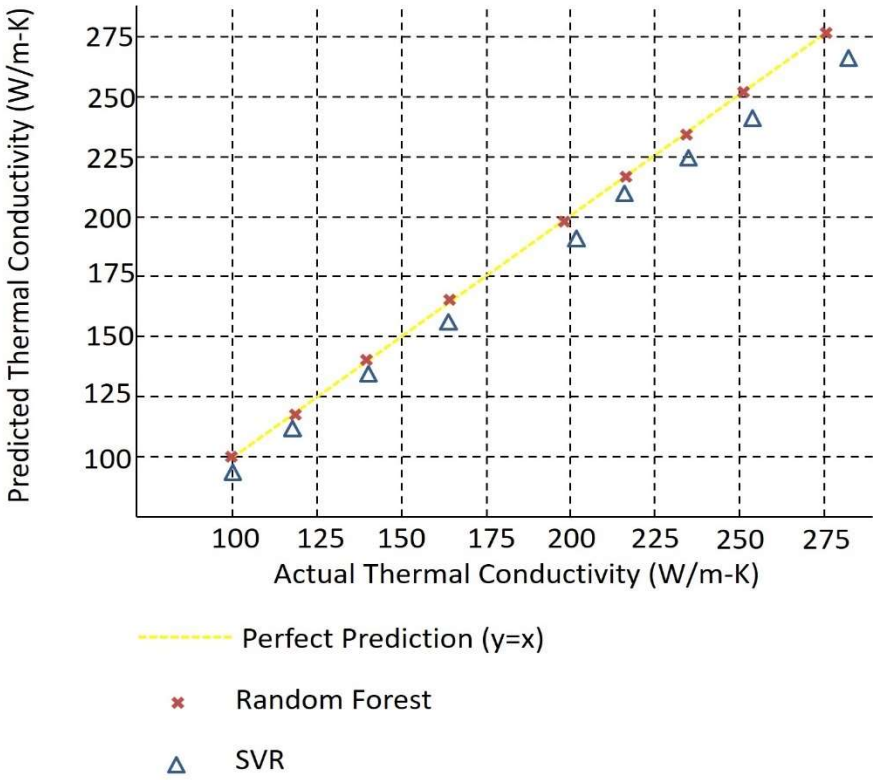

Figure 6: Actual vs Predicted Thermal Conductivity Using Random Forest and SVR Models

Figure 6 presents a comparison of actual versus predicted thermal conductivity values using the Random Forest and Support Vector Regression (SVR) models. Each data point represents a material sample, plotted to assess how closely the model predictions align with ground truth measurements. The Random Forest model demonstrates superior predictive accuracy, with most predictions clustering tightly around the ideal diagonal line, indicating minimal error. In contrast, the SVR model also performs well but shows slightly greater deviation, particularly for higher conductivity values. This visualization highlights the robustness and generalization capability of ensemble-based methods like Random Forest over kernel-based approaches in modelling complex structure–property relationships in materials science.

This study illustrates how ML can uncover structure-property relationships that are difficult to model analytically, especially when microstructural complexity plays a dominant role in effective performance.

5.2 Alloy Design through Bayesian Optimization

Designing new high-performance alloys involves exploring a vast compositional design space. The combinatorial explosion of possible element combinations, heat treatment schedules, and processing parameters makes exhaustive experimentation infeasible. Bayesian optimization (BO) offers a solution by iteratively selecting the most promising candidates based on uncertainty-aware surrogate models.

In a pioneering work by Lookman et al. [38], BO was applied to design NiTi-based shape memory alloys with target transformation temperatures and elastic moduli. A Gaussian process regression (GPR) model was trained on a sparse dataset of experimental alloy compositions and their corresponding properties. The acquisition function used for exploration was the Expected Improvement (EI), which balances the trade-off between sampling unexplored regions and refining existing knowledge.

Over successive iterations, the algorithm efficiently converged toward alloy compositions with optimal properties. Experimental validation confirmed the accuracy of the model's predictions, with some newly suggested alloys outperforming those in the original dataset.

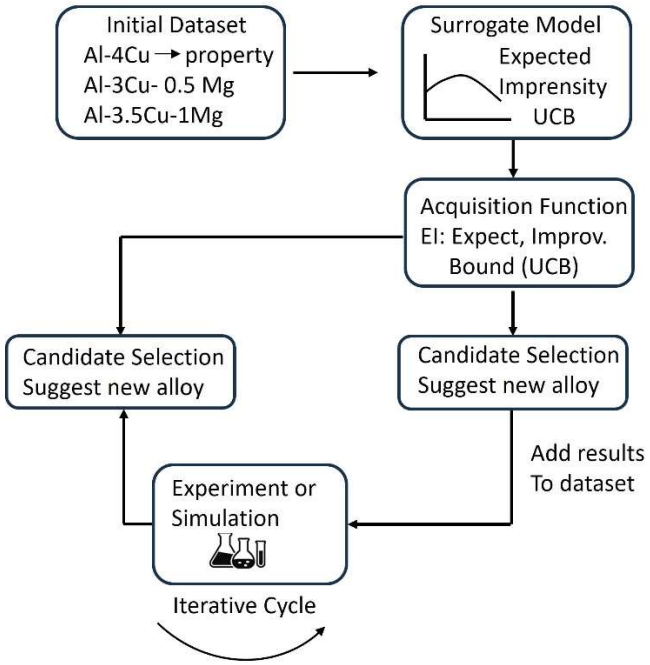

Figure 7: Bayesian Optimization Workflow for Alloy Design

Figure 7 illustrates the Bayesian optimization workflow applied to alloy design. The process begins with an initial dataset of alloy compositions and their corresponding measured or simulated properties. A surrogate model commonly a Gaussian Process is trained to approximate the structure–property relationship, capturing both predictions and associated uncertainties. Based on this model, an acquisition function selects the next alloy composition to evaluate, balancing exploration of uncertain regions with exploitation of high-performing candidates. The selected composition is then evaluated through experiments or high-fidelity simulations, and the resulting data is fed back into the model to update its predictions. This iterative loop continues until convergence criteria are met or optimal material properties are achieved. The Bayesian optimization framework significantly reduces the number of costly experiments required and enables efficient navigation of vast compositional design spaces.

Table 7: Iterative Improvement in Target Property with Each Optimization Cycle

Iteration	Suggested Alloy Composition	Predicted Property (e.g., Yield Strength in MPa)	Measured Property	Improvement (%)
0 (Baseline)	Al-4Cu	—	250 MPa	—
1	Al-4.5Cu-0.2Mg	268 MPa	263 MPa	+5.2%
2	Al-5Cu-0.5Mg	280 MPa	276 MPa	+4.9%
3	Al-5.2Cu-0.7Mg-0.1Zn	290 MPa	288 MPa	+4.3%
4	Al-5.3Cu-0.9Mg-0.15Zn-0.05Si	298 MPa	296 MPa	+2.8%
5	Al-5.4Cu-1.0Mg-0.2Zn-0.05Si	301 MPa	300 MPa	+1.4%

Table 7 summarizes the iterative improvement in the target property—specifically, yield strength—across successive optimization cycles in alloy design. Starting from a baseline alloy composition of Al-4Cu with a measured yield strength of 250 MPa, each subsequent iteration suggests modified alloy compositions aimed at enhancing mechanical performance. Predicted and experimentally measured values consistently demonstrate progressive improvement. For example, the first iteration, Al-4.5Cu-0.2Mg, showed a measured yield strength increase of 5.2% relative to the baseline. Subsequent iterations continue this upward trend, reaching a measured yield strength of 300 MPa by the fifth iteration, corresponding to a cumulative improvement of 20% from the baseline. Notably, the magnitude of improvement per cycle decreases over time, indicating convergence toward an optimal composition. This iterative workflow, likely guided by an optimization algorithm such as Bayesian optimization, effectively explores compositional space and refines alloy formulations to maximize target properties.

Beyond prediction, the framework also guided materials synthesis, linking data-driven design with experimental realization. This closed-loop model demonstrates the practical impact of ML in materials R&D workflows, significantly reducing discovery time and resource consumption.

5.3 Model Validation Strategies in Material Science

A critical aspect of ML deployment in materials science is model validation. Given the scarcity and heterogeneity of datasets, conventional validation protocols from other ML domains must be adapted. Strategies include:

1. Cross-validation with stratified sampling to ensure that rare compositions or phases are not underrepresented in training and testing splits.
2. Domain-aware performance metrics, such as relative error with respect to physically meaningful baselines (e.g., deviation from DFT predictions rather than absolute RMSE).
3. Physics-informed sanity checks, where models are assessed for consistency with known laws (e.g., non-negativity of predicted conductivity, monotonicity with volume fraction).

Additionally, multi-fidelity validation which integrates low-fidelity simulations with high-fidelity experiments has become increasingly popular to reduce validation costs while maintaining model reliability [46].

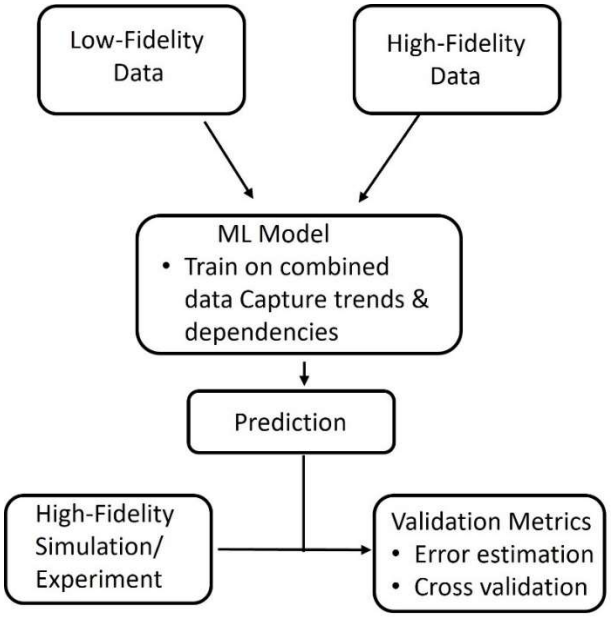

Figure 8: Framework for Multi-Fidelity Validation Using ML in Materials Science

Figure 8 illustrates a multi-fidelity validation framework integrating machine learning (ML) techniques within materials science workflows. This framework combines data and predictions from multiple sources of varying fidelity such as high-accuracy but expensive experimental measurements, intermediate-fidelity simulations,

and lower-fidelity computational models to improve the reliability and efficiency of material property predictions. ML models are trained and validated using this heterogeneous data, leveraging lower-fidelity sources to guide exploration and higher-fidelity data to refine and calibrate predictions. By systematically incorporating uncertainties associated with each data source, the framework enables robust decision-making and accelerates materials discovery while minimizing costly experimental efforts. This multi-fidelity approach is especially valuable for complex materials systems where direct high-fidelity data acquisition is challenging.
Such holistic validation approaches ensure that ML models in materials science are not just statistically accurate but also physically interpretable and experimentally actionable.

6. Conclusion

ML is fundamentally reshaping the landscape of materials science by shifting the traditional trial-and-error paradigm toward a data-driven, predictive, and highly efficient discovery framework**.** From supervised learning models used to predict thermal conductivity to deep learning architectures applied in microstructural analysis, ML enables rapid and accurate insights that were previously difficult or impossible to achieve through conventional methods.
Notable advances include the use of CNNs for automated interpretation of microstructural images, autoencoders for uncovering latent representations of complex material features, and Bayesian optimization for guiding the design of novel alloy compositions. These applications demonstrate that ML not only accelerates materials discovery but also deepens scientific understanding—particularly when integrated with domain expertise and physical principles.
Despite its promise, key challenges persist, notably in areas such as data quality, model interpretability, and the incorporation of governing physical laws. However, emerging strategies—including physics-informed machine learning, active learning, and multi-fidelity modelling are actively addressing these limitations. As the field progresses toward autonomous research platforms and closed-loop experimentation, ML is poised not to replace traditional materials science, but to augment and empower it.
In essence, machine learning is not merely an enhancement; it represents a transformative redefinition of how materials are designed, characterized, and deployed in the modern scientific era.

References

1. Butler, Keith T., Daniel W. Davies, Hugh Cartwright, Olexandr Isayev, and Aron Walsh. "Machine learning for molecular and materials science." *Nature* 559, no. 7715 (2018): 547-555.
2. Ramprasad, Rampi, Rohit Batra, Ghanshyam Pilania, Arun Mannodi-Kanakkithodi, and Chiho Kim. "Machine learning in materials informatics: recent applications and prospects." *npj Computational Materials* 3, no. 1 (2017): 54.
3. Ward, Logan, Ankit Agrawal, Alok Choudhary, and Christopher Wolverton. "A general-purpose machine learning framework for predicting properties of inorganic materials." *npj Computational Materials* 2, no. 1 (2016): 1-7.
4. Xie, Tian, and Jeffrey C. Grossman. "Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties." *Physical review letters* 120, no. 14 (2018): 145301.
5. Jha, Dipendra, Logan Ward, Arindam Paul, Wei-keng Liao, Alok Choudhary, Chris Wolverton, and Ankit Agrawal. "Elemnet: Deep learning the chemistry of materials from only elemental composition." *Scientific reports* 8, no. 1 (2018): 17593.
6. Jain, Anubhav, Shyue Ping Ong, Geoffroy Hautier, Wei Chen, William Davidson Richards, Stephen Dacek, Shreyas Cholia et al. "Commentary: The Materials Project: A materials genome approach to accelerating materials innovation." *APL materials* 1, no. 1 (2013).
7. Schmidt, Jonathan, Mário RG Marques, Silvana Botti, and Miguel AL Marques. "Recent advances and applications of machine learning in solid-state materials science." *npj computational materials* 5, no. 1 (2019): 83.
8. Daw, Arka, Anuj Karpatne, William D. Watkins, Jordan S. Read, and Vipin Kumar. "Physics-guided neural networks (pgnn): An application in lake temperature modeling." In *Knowledge guided machine learning*, pp. 353-372. Chapman and Hall/CRC, 2022.
9. Callister Jr, William D., and David G. Rethwisch. *Materials science and engineering: an introduction*. John wiley & sons, 2020.

10. Martin, Richard M. *Electronic structure: basic theory and practical methods*. Cambridge university press, 2020.
11. Jain, Anubhav, Shyue Ping Ong, Geoffroy Hautier, Wei Chen, William Davidson Richards, Stephen Dacek, Shreyas Cholia et al. "Commentary: The Materials Project: A materials genome approach to accelerating materials innovation." *APL materials* 1, no. 1 (2013).
12. Curtarolo, Stefano, Wahyu Setyawan, Gus LW Hart, Michal Jahnatek, Roman V. Chepulskii, Richard H. Taylor, Shidong Wang et al. "AFLOW: An automatic framework for high-throughput materials discovery." *Computational Materials Science* 58 (2012): 218-226.
13. Kalidindi, Surya R., and Marc De Graef. "Materials data science: current status and future outlook." *Annual Review of Materials Research* 45, no. 1 (2015): 171-193.
14. Agrawal, Ankit, and Alok Choudhary. "Perspective: Materials informatics and big data: Realization of the "fourth paradigm" of science in materials science." *Apl Materials* 4, no. 5 (2016).
15. Orio, Maylis, Dimitrios A. Pantazis, and Frank Neese. "Density functional theory." *Photosynthesis research* 102 (2009): 443-453.
16. Dhatt, Gouri, Emmanuel Lefrançois, and Gilbert Touzot. *Finite element method*. John Wiley & Sons, 2012.
17. Hollingsworth, Scott A., and Ron O. Dror. "Molecular dynamics simulation for all." *Neuron* 99, no. 6 (2018): 1129-1143.
18. Wu, Zonghan, Shirui Pan, Fengwen Chen, Guodong Long, Chengqi Zhang, and Philip S. Yu. "A comprehensive survey on graph neural networks." *IEEE transactions on neural networks and learning systems* 32, no. 1 (2020): 4-24.
19. Li, Zewen, Fan Liu, Wenjie Yang, Shouheng Peng, and Jun Zhou. "A survey of convolutional neural networks: analysis, applications, and prospects." *IEEE transactions on neural networks and learning systems* 33, no. 12 (2021): 6999-7019.
20. Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg et al. "The FAIR Guiding Principles for scientific data management and stewardship." *Scientific data* 3, no. 1 (2016): 1-9.
21. Hearst, Marti A., Susan T. Dumais, Edgar Osuna, John Platt, and Bernhard Scholkopf. "Support vector machines." *IEEE Intelligent Systems and their applications* 13, no. 4 (1998): 18-28.
22. Breiman, Leo. "Random forests." *Machine learning* 45 (2001): 5-32.
23. Abdi, Hervé, Dominique Valentin, and Betty Edelman. *Neural networks*. No. 124. Sage, 1999.
24. Kodinariya, Trupti M., and Prashant R. Makwana. "Review on determining number of Cluster in K-Means Clustering." *International Journal* 1, no. 6 (2013): 90-95.
25. Abdi, Hervé, and Lynne J. Williams. "Principal component analysis." *Wiley interdisciplinary reviews: computational statistics* 2, no. 4 (2010): 433-459.
26. Koolstra, Kirsten, Peter Börnert, Boudewijn Lelieveldt, Andrew Webb, and Oleh Dzyubachyk. "t-Distributed stochastic neighbour embedding (t-SNE) as a tool for visualizing the encoding capability of magnetic resonance fingerprinting (MRF) dictionaries." In *Proceedings of the 27th annual meeting of ISMRM, Montréal*. 2019.
27. Murtagh, Fionn, and Pedro Contreras. "Algorithms for hierarchical clustering: an overview." *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 2, no. 1 (2012): 86-97.
28. Wiering, Marco A., and Martijn Van Otterlo. "Reinforcement learning." *Adaptation, learning, and optimization* 12, no. 3 (2012): 729.
29. Huang, Yanhua. "Deep Q-networks." *Deep reinforcement learning: fundamentals, research and applications* (2020): 135-160.
30. Ketkar, Nikhil, Jojo Moolayil, Nikhil Ketkar, and Jojo Moolayil. "Convolutional neural networks." *Deep learning with Python: learn best practices of deep learning models with PyTorch* (2021): 197-242.
31. Medsker, Larry R., and Lakhmi Jain. "Recurrent neural networks." *Design and Applications* 5, no. 64-67 (2001): 2.
32. Corso, Gabriele, Hannes Stark, Stefanie Jegelka, Tommi Jaakkola, and Regina Barzilay. "Graph neural networks." *Nature Reviews Methods Primers* 4, no. 1 (2024): 17.
33. Ward, Logan, Ankit Agrawal, Alok Choudhary, and Christopher Wolverton. "A general-purpose machine learning framework for predicting properties of inorganic materials." *npj Computational Materials* 2, no. 1 (2016): 1-7.

34. Pilania, Ghanshyam, Chenchen Wang, Xun Jiang, Sanguthevar Rajasekaran, and Ramamurthy Ramprasad. "Accelerating materials property predictions using machine learning." *Scientific reports* 3, no. 1 (2013): 2810.
35. Seko, Atsuto, Hiroyuki Hayashi, Keita Nakayama, Akira Takahashi, and Isao Tanaka. "Representation of compounds for machine-learning prediction of physical properties." *Physical Review B* 95, no. 14 (2017): 144110.
36. Bostanabad, Ramin, Yichi Zhang, Xiaolin Li, Tucker Kearney, L. Catherine Brinson, Daniel W. Apley, Wing Kam Liu, and Wei Chen. "Computational microstructure characterization and reconstruction: Review of the state-of-the-art techniques." *Progress in Materials Science* 95 (2018): 1-41.
37. Rajak, Pankaj, Aravind Krishnamoorthy, Ankit Mishra, Rajiv Kalia, Aiichiro Nakano, and Priya Vashishta. "Autonomous reinforcement learning agent for chemical vapor deposition synthesis of quantum materials." *npj Computational Materials* 7, no. 1 (2021): 108.
38. Lookman, Turab, Prasanna V. Balachandran, Dezhen Xue, and Ruihao Yuan. "Active learning in materials science with emphasis on adaptive sampling using uncertainties for targeted design." *npj Computational Materials* 5, no. 1 (2019): 21.
39. DeCost, Brian L., and Elizabeth A. Holm. "A computer vision approach for automated analysis and classification of microstructural image data." *Computational materials science* 110 (2015): 126-133.
40. Lubbers, Nicholas, Turab Lookman, and Kipton Barros. "Inferring low-dimensional microstructure representations using convolutional neural networks." *Physical Review E* 96, no. 5 (2017): 052111.
41. Cang, Ruijin, Yaopengxiao Xu, Shaohua Chen, Yongming Liu, Yang Jiao, and Max Yi Ren. "Microstructure representation and reconstruction of heterogeneous materials via deep belief network for computational material design." *Journal of Mechanical Design* 139, no. 7 (2017): 071404.
42. Padhan, Manas Kumar, Akshay Rai, and Mira Mitra. "Prediction of grain size distribution in microstructure of polycrystalline materials using one dimensional convolutional neural network (1D-CNN)." *Computational Materials Science* 229 (2023): 112416.
43. Bessa, Miguel A., Ramin Bostanabad, Zeliang Liu, Anqi Hu, Daniel W. Apley, Catherine Brinson, Wei Chen, and Wing Kam Liu. "A framework for data-driven analysis of materials under uncertainty: Countering the curse of dimensionality." *Computer Methods in Applied Mechanics and Engineering* 320 (2017): 633-667.
44. Yang, Zijiang, Xiaolin Li, L. Catherine Brinson, Alok N. Choudhary, Wei Chen, and Ankit Agrawal. "Microstructural materials design via deep adversarial learning methodology." *Journal of Mechanical Design* 140, no. 11 (2018): 111416.
45. Ju, Zhaoqiang, Kai Guo, and Xiaojing Liu. "Modelling thermal conductivity on salt-affected soils and its modification." *International Journal of Thermal Sciences* 185 (2023): 108071.
46. Mortazavi, Bohayra. "Recent advances in machine learning-assisted multiscale design of energy materials." *Advanced Energy Materials* 15, no. 9 (2025): 2403876.