ABOUT THE EDITORS

Dr. T. Gopalakrishnan
Assistant Professor,
Department of Mechanical Engineering
Vels Institute of Science, Technology & Advanced Studies (VISTAS), Chennai, India

Dr. S.Sivaganesan
Professor,
Department of Mechanical Engineering
Vels Institute of Science, Technology & Advanced Studies (VISTAS), Chennai, India

Mrs. Bharathi.V
Assistant Professor,
Department of Computer Science and Engineering
Vels Institute of Science, Technology & Advanced Studies (VISTAS), Chennai, India

Dr. R. Anandan
Professor,
Department of Computer Science and Engineering
Vels Institute of Science, Technology & Advanced Studies (VISTAS), Chennai, India

ISBN 978-81-988536-6-0

Published: 20th May 2025
Printed with care by ivarin Printing

IMAGINEX INKS PUBLICATION
https://imaginexinkspublication.com/
Contact Number : 9750663871
9962991057

HARNESSING MACHINE LEARNING FOR ADVANCED MATERIALS DISCOVERY AND DESIGN

IMAGINEX INKS PUBLICATION

Harnessing Machine Learning for Advanced Materials Discovery and Design

Editors

Dr. T. Gopalakrishnan
Dr. S. Sivaganesan
Mrs. Bharathi V
Dr. R. Anandan

Harnessing Machine Learning for Advanced Materials Discovery and Design

EDITORS

Dr. T. Gopalakrishnan

Assistant Professor, Department of Mechanical Engineering,
Vels Institute of Science, Technology & Advanced Studies
(VISTAS), Chennai

Dr. S. Sivaganesan

Professor, Department of Mechanical Engineering,
Vels Institute of Science, Technology & Advanced Studies
(VISTAS), Chennai

Mrs.V. Bharathi

Assistant Professor
Department of Computer science and Engineering
Vels Institute of Science, Technology & Advanced Studies
(VISTAS), Chennai

Dr. R. Anandan

Professor
Department of Computer science and Engineering
Vels Institute of Science, Technology & Advanced Studies
(VISTAS), Chennai

Harnessing Machine Learning for Advanced Materials Discovery and Design

Edited by

Dr. T. Gopalakrishnan, Dr.S.Sivaganesan, Mrs.V.Bharathi, Dr. R. Anandan

Volume I June 2025

Price: Rs. 500/-

ISBN: 978-81-988536-6-0

Published by and copies can be had from:

Imaginex Inks Publication

2/158, Kurinji Nagar First St, Ponnan Nagar,

Irumbuliyur, Vandalur,

Chennai 600048, Tamil Nadu, India.

Phone: 9750663871, 9962991057

e-mail: imaginexinks@gmail.com

https://www.imaginexinkspublication.com/

Editor's Spotlight

Dr.T.Gopalakrishnan

Dr. T. Gopalakrishnan is an Assistant Professor in Mechanical Engineering at VISTAS, Chennai, with over 10 years of academic and research experience. His work focuses on integrating artificial intelligence with mechanical systems, including smart manufacturing, predictive maintenance, and energy-efficient design. He has published 30+ research papers in reputed journals, received multiple faculty excellence awards, and is an active member of professional bodies like IFERP and IAENG. His research bridges traditional engineering with cutting-edge Industry 4.0 technologies.

Dr. S. Sivaganesan

Dr. S. Sivaganesan is a Professor in the Department of Mechanical Engineering at VISTAS, Chennai, with over 15 years of academic and research experience. He holds a Ph.D. in Alternate Fuels and an M.E. in Thermal Power Engineering, with specialized expertise in Smart Materials. He has published over 50 Scopus-indexed research papers and serves as an editor, reviewer, and committee member for several reputed international journals and conferences. A recipient of the Dr. APJ Abdul Kalam Young Scientist Award, his research spans IC engines, combustion analysis, thermal coatings, and advanced materials. Dr. Sivaganesan is recognized for his interdisciplinary contributions and leadership in mechanical and energy-based innovations.

Ms. V. Bharathi

Ms. V. Bharathi holds an M.E. in Computer Science. She has been teaching at Vels University for over 11 years, specializing in Image Processing, Artificial Intelligence, and Machine Learning. Her academic role involves instructing and mentoring students in advanced computing subjects, with a focus on AI-based image processing techniques and machine learning algorithms. Her expertise areas align with evolving trends in AI and Computer Vision—fields that are increasingly integrated into hands-on lab modules and student-led research initiatives.

Dr. R. Anandan

Dr. R. Anandan is a distinguished academician and researcher in the field of Computer Science and Engineering, with a Post-Doctoral D.Sc. from Mexico and a recognized Chartered Engineer accreditation from the Institution of Engineers (India). He currently serves as Professor in the Department of Computer Science and Engineering and Director of Innovation and Incubation at Vels Institute of Science, Technology & Advanced Studies (VISTAS), Chennai. With vast experience spanning corporate and academic domains, his research interests encompass artificial intelligence, soft computing, machine learning, high-performance computing, big data analytics, image processing, 3D printing, and knowledge engineering. Dr. Anandan has guided 16 Ph.D. scholars to completion and currently supervises eight doctoral candidates.

He holds memberships in several prestigious international and national organizations, including the International Association of Engineers (IAENG), Computer

Science Teacher Association (CSTA), IACSIT, the Institution of Engineers (India), CSI, and ISTE. Dr. Anandan has published over 150 research papers in reputed journals indexed in Scopus and SCI, presented more than 100 conference papers, and authored or edited 32 books, including 12 international publications and 40 book chapters. He has filed 23 patents, with one international and three Indian patents granted. He serves on editorial boards and review committees for leading publishers such as IEEE, Springer, Elsevier, and Thomson Reuters, and has contributed to international conferences affiliated with Scopus and Springer.

Dr. Anandan has received 25 prestigious awards from national and international bodies in recognition of his academic contributions. His research projects have been funded by AICTE (Research Promotion Scheme), MSME (Idea Hackathon), and several private organizations. He is actively involved in consultancy projects and serves as a knowledge partner to prominent software firms. His outstanding contributions reflect a career committed to academic excellence, research innovation, and industrial collaborations.

Acknowledgment

We express our heartfelt gratitude to all those who have contributed to the realization of this book, "*Harnessing Machine Learning for Advanced Materials Discovery and Design"*. This work is the result of sustained intellectual engagement and collaborative effort across domains of materials science, computer science, and artificial intelligence.

We are deeply thankful to the contributors for their significant efforts and dedication.

Our deepest appreciation goes to Vels Institute of Science, Technology & Advanced Studies (VISTAS), Chennai, for providing a conducive research environment and encouraging interdisciplinary scholarship. We also thank our colleagues and students in the departments of Mechanical Engineering, Materials Science, and Computer Science for their valuable discussions and critical feedback during the development of this manuscript.

We gratefully acknowledge the support of professional bodies such as the Materials Research Society (MRS), IEEE Computational Intelligence Society, and the Machine Learning for Molecules community for fostering an

ecosystem that bridges materials discovery with AI-driven innovation.

Special thanks to the editorial and production team whose expertise ensured the timely and high-quality completion of this volume. Finally, we recognize the unwavering support of our families and well-wishers, whose encouragement has been instrumental in the successful completion of this book.

This book is dedicated to the global scientific community striving to accelerate materials innovation for a sustainable and technologically advanced future.

Dr. T. Gopalakrishnan

Dr.S.Sivaganesan

Mrs.V.Bharathi

Dr. R. Anandan

Preface

In recent years, the convergence of materials science and machine learning has ushered in a transformative era in scientific discovery. As materials innovation becomes increasingly data-intensive, traditional trial-and-error approaches are being replaced by predictive modeling, generative algorithms, and autonomous experimentation. This book, *Harnessing Machine Learning for Advanced Materials Discovery and Design*, aims to serve as a comprehensive and accessible guide to this rapidly evolving field.

The motivation for this book stems from the growing recognition that machine learning is not merely a tool but a paradigm shift in how materials are explored, designed, and optimized. From high-throughput screening and inverse design to multi-fidelity simulations and automated laboratories, the integration of data-driven models is accelerating the development of catalysts, semiconductors, polymers, alloys, and other advanced functional materials.

This book is structured to cater to a diverse audience, including graduate students, researchers, and industry practitioners. It begins by introducing foundational concepts of machine learning as applied to materials informatics, followed by detailed chapters on supervised and unsupervised learning, generative models, and reinforcement learning. Subsequent chapters explore the integration of ML with density functional theory (DFT), molecular dynamics (MD), and finite element methods (FEM), as well as real-world case studies in batteries, solar cells, and structural alloys.

Each chapter balances theoretical rigor with practical application, drawing on the latest academic literature, real datasets, and open-source tools. Emphasis is placed on reproducibility, interpretability, and scalability—principles that are vital for meaningful scientific progress.

We hope this book not only equips readers with the knowledge and tools necessary to apply machine learning in materials science but also inspires further innovation at the intersection of computation, experimentation, and intelligent design.

We welcome readers to engage critically with the content, experiment with the methods presented, and contribute to the growing community committed to accelerating materials discovery through machine learning.

Dr. T. Gopalakrishnan

Dr.S.Sivaganesan

Mrs.V.Bharathi

Dr. R. Anandan

INDEX

Chapter 1

Introduction to Machine Learning in Materials Science

Dr. K. Karunakaran

Assistant Professor, Department of Mechanical Engineering, Vels Institute of Science Technology and Advanced Studies (VISTAS), Chennai, India.

1.1 Historical Evolution from Empiricism to Informatics

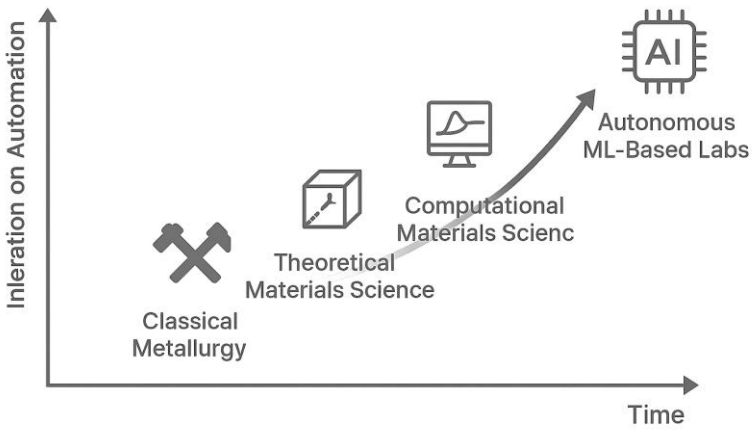
The evolution of materials science has transitioned from an empirically driven discipline to one increasingly shaped by computation and data-centric approaches. In its earliest forms, materials discovery was largely based on empirical methods, where innovations such as bronze, iron, and ceramics were developed through trial-and-error experimentation without theoretical underpinnings. This era, often referred to as the craft-based period, relied heavily on artisanal knowledge passed through generations, with minimal formal scientific structure.

The scientific revolution of the 17th and 18th centuries brought foundational advances in classical mechanics and thermodynamics. These developments enabled the emergence of theoretical materials science, providing predictive frameworks for phenomena such as phase equilibria, heat transfer, and stress-strain behavior, thereby allowing more systematic alloy and polymer design.

A transformative shift occurred in the mid-20th century with the advent of computational modeling. Techniques such as Density Functional Theory (DFT) and Molecular Dynamics (MD) simulations allowed scientists to probe the atomic-scale properties of materials with high precision. DFT, in particular, enabled accurate predictions of band structures, formation energies, and elastic moduli by solving the electronic Schrödinger equation under practical approximations (Kohn & Sham, 1965). These computational methods, although highly accurate, were also computationally expensive—limiting their scalability for large-scale screening tasks.

The 21st century witnessed another leap with the integration of machine learning (ML) into materials research, giving rise to the field of materials informatics. The growing availability of high-throughput computational data, particularly from platforms such as the Materials Project (Jain et al., 2013), enabled the training of predictive ML models capable of learning complex structure–property relationships. These models, once trained, could generalize to unexplored materials at a fraction of the computational cost, providing orders-of-magnitude speed-ups over conventional simulations (Butler et al., 2018).

This transition from physics-based modeling to data-driven prediction is illustrated in Figure 1.1, which outlines the methodological shift from empirical metallurgy to autonomous laboratories powered by AI. The vertical axis represents the level of automation and intelligence, while the horizontal axis traces the chronological development of core methodologies.



Evolution timeline from classical metallurgy to autonomous ML-based labs

1.2 Role of Computational Modeling in Materials Science

The role of computational modeling in materials science has been pivotal in enabling predictive understanding of materials behavior across atomic to mesoscopic scales. Long before the advent of machine learning, computational methods such as Density Functional Theory (DFT) and Molecular Dynamics (MD) had already revolutionized how scientists studied and designed materials, making it possible to simulate systems with quantum or classical precision.

Among these, DFT has become the most widely adopted first-principles approach, providing a quantum mechanical framework to determine the ground-state properties of electrons in atoms, molecules, and solids (Kohn & Sham, 1965). It facilitates the calculation of crucial properties such as electronic band structures, total and formation energies,

magnetic moments, charge distributions, and phonon spectra. These predictions have become essential in understanding and designing semiconductors, superconductors, catalysts, and energy materials.

The Materials Project is a landmark example of how DFT has been scaled to high-throughput computation. It has systematically catalogued over 100,000 inorganic compounds with calculated properties, thus creating a vast resource for both fundamental studies and machine learning applications (Jain et al., 2013). Its open-access database is frequently used to train surrogate ML models, which emulate DFT-level accuracy at a fraction of the computational cost.

In parallel, Molecular Dynamics (MD) has enabled time-resolved simulations of atomic and molecular motion, providing insights into dynamic phenomena such as:

- Thermal conductivity and diffusion,
- Mechanical deformation, plasticity, and fracture mechanics,
- Phase transitions and nucleation processes.

Although less computationally intensive than DFT, MD simulations require carefully parameterized interatomic potentials, such as Lennard-Jones or embedded-atom models (EAM), which limit their transferability across chemical systems.

Despite the power of both DFT and MD, they share a critical limitation: computational scalability. Screening thousands of materials via DFT can take weeks or months of high-performance computing time. As

noted by Butler et al. (2018), such limitations have catalyzed the integration of machine learning as a surrogate modeling strategy, wherein models are trained on DFT/MD outputs and subsequently used to predict the properties of unseen compounds in real-time.

Importantly, machine learning does not aim to replace DFT or MD but to augment their scope. Trained on high-fidelity data, ML models can provide near-DFT accuracy while scanning chemical spaces that are orders of magnitude larger than those accessible through direct simulation. This fusion of computational physics with statistical learning represents a synergistic approach—where physics-based modeling supplies training data and constraints, while ML delivers efficiency and generalization (Ward et al., 2016).

1.3 Defining Machine Learning in the Context of Materials Science

Machine Learning (ML), a core subfield of artificial intelligence, is increasingly recognized as a transformative methodology in materials science, particularly for its ability to discover patterns, predict properties, and assist in designing novel materials. In the broadest sense, ML refers to algorithms that learn from data to make predictions or decisions without being explicitly programmed (Butler et al., 2018). When applied to materials systems, ML shifts the focus from solving complex analytical equations to learning nonlinear structure–property relationships directly from empirical or simulated datasets.

In contrast to conventional materials modeling—based on physical laws such as thermodynamics or quantum mechanics—ML leverages

statistical inference to build data-driven models that approximate the input–output mapping. These models are capable of uncovering patterns that are often inaccessible through traditional theoretical methods, particularly in high-dimensional or noisy datasets.

The basic workflow of a supervised ML application in materials science involves three key components:

- Input features (descriptors): These are numerical representations of material structure, composition, or electronic configuration. Examples include atomic radii, electronegativity differences, stoichiometric ratios, and graph-based representations (Ward et al., 2016; Xie & Grossman, 2018).
- Learning algorithms: ML models such as random forests, support vector machines (SVMs), or gradient-boosted decision trees are commonly used for small to medium datasets, while deep learning approaches such as graph neural networks (GNNs) are employed for large-scale or unstructured data (Chen et al., 2019).
- Target outputs (labels): These are the properties to be predicted, which may include formation energy, bandgap, thermal conductivity, elastic modulus, or even class labels such as “metal” or “non-metal.”

In materials science, ML tasks are broadly categorized as:

- Regression, for predicting continuous-valued properties (e.g., bandgap, thermal conductivity),
- Classification, for categorical outcomes (e.g., metal vs. insulator),
- Clustering and dimensionality reduction, for exploratory analysis of materials spaces.

A major driver for ML adoption in this domain has been the availability of large, structured databases like the Materials Project (Jain et al., 2013) and OQMD (Saal et al., 2013), which provide high-fidelity DFT-calculated properties across thousands of materials. These datasets enable the training and benchmarking of models with physical ground truth.

However, ML in materials science is not intended to function as a black box. Effective implementation often involves domain-informed feature engineering, cross-validation, and the use of interpretability tools such as SHAP values (Lundberg & Lee, 2017) to extract insights from model behavior. In addition, the integration of physics-informed constraints—either through hybrid models or constrained optimization—helps ensure the scientific validity of the predictions (Ward et al., 2016).

Therefore, within the context of materials science, machine learning is best understood as a complementary modeling framework. It leverages existing experimental and simulation data to make fast, scalable, and

reasonably accurate predictions, expanding the reach of traditional methodologies and enabling intelligent navigation through otherwise intractable design spaces.

1.4 Types of Machine Learning in Materials Science: Supervised, Unsupervised, Reinforcement, and Deep Learning

The landscape of machine learning in materials science is structured around four primary paradigms—supervised learning, unsupervised learning, reinforcement learning, and deep learning. Each type is suited to a distinct category of problems and data structures, and their strategic integration enables a wide range of materials discovery applications—from predictive modeling and design optimization to classification and generative materials synthesis.

1.4.1 Supervised Learning

Supervised learning is the most mature and extensively applied ML framework in materials science. It operates on labeled datasets, where input features (e.g., elemental composition, structural parameters) are mapped to known target properties (e.g., band gap, elastic modulus).

This paradigm has been successfully used in:

- **Regression tasks**, such as predicting:
    - Electronic bandgaps (Pilania et al., 2013),
    - Elastic properties and bulk moduli (Ward et al., 2016),

- Thermal conductivity and Seebeck coefficients (Carrete et al., 2014).

- **Classification tasks**, such as:
  - Distinguishing between metals and insulators (Ward et al., 2016),
  - Identifying glass formers or stable polymorphs,
  - Categorizing materials based on magnetic behavior.

Models commonly employed include **random forests**, **gradient boosting machines**, **support vector machines (SVMs)**, and **feedforward neural networks**, all of which balance interpretability with predictive accuracy, especially when combined with domain-relevant features.

1.4.2 Unsupervised Learning

Unlike supervised methods, **unsupervised learning** does not rely on predefined labels. Instead, it seeks to identify **latent patterns or groupings** within high-dimensional data. This is particularly useful in exploratory phases of research, where **hypotheses are generated** by detecting underlying structures in the dataset.

Prominent applications in materials science include:

- **Clustering** of compounds based on compositional similarity, mechanical behavior, or electronic structure.

- **Dimensionality reduction** using **Principal Component Analysis (PCA)** or **t-SNE**, which aids in visualizing complex relationships and identifying property-driving features (Ward et al., 2016).

Though less commonly used for direct prediction, unsupervised methods are essential for **data cleaning**, **outlier detection**, and the **preprocessing of descriptors** before supervised learning.

1.4.3 Reinforcement Learning (RL)

Reinforcement learning represents a dynamic learning approach where an agent interacts with an environment by taking actions that maximize a cumulative reward. In materials science, RL is gaining attention for its ability to **optimize experimental workflows and synthesis strategies** in **sequential decision-making** settings.

Notable applications include:

- **Self-driving laboratories**, where RL algorithms autonomously select and execute experiments to converge on optimal synthesis conditions (Häse et al., 2018),
- **Synthesis planning**, where RL identifies efficient reaction pathways or processing sequences for target materials.

The agent's performance improves over time as it learns from **positive or negative feedback**, making RL a powerful framework for **autonomous discovery systems** and **closed-loop design** environments.

1.4.4 Deep Learning

Deep learning is a subset of ML that leverages **neural networks with multiple hidden layers** to automatically extract high-level, abstract features from raw or unstructured data. Its hierarchical nature makes it especially suited for modeling complex structure–property relationships in materials science.

Common deep learning models include:

- **Convolutional Neural Networks (CNNs):**
    - Used for image-based microstructure analysis and phase diagram predictions (Ward et al., 2016),
- **Graph Neural Networks (GNNs):**
    - Models such as **CGCNN** and **MEGNet** operate directly on atomic graphs, learning bond environments and predicting properties such as formation energies, band structures, and elastic constants (Xie & Grossman, 2018; Chen et al., 2019),
- **Transformers and Recurrent Neural Networks (RNNs):**
    - Used in polymer design and sequence-to-property modeling for organic materials.

Deep learning models require large datasets but have demonstrated **state-of-the-art performance** in several benchmarking challenges.

However, they also face challenges in **interpretability** and **data-efficiency**, which remain active areas of research.

Table 1.1: Comparison of Machine Learning Paradigms in Materials Science

ML Paradigm	Input Data	Output	Typical Use Cases	Examples
Supervised Learning	Labeled (features + targets)	Regression / Classification	Bandgap prediction, phase stability classification	Pilania et al., 2013; Ward et al., 2016
Unsupervised Learning	Unlabeled (features only)	Clusters / Reduced Dimensions	Pattern recognition, grouping by structure or behavior	Ward et al., 2016
Reinforcement Learning	Agent-environment interactions	Sequential Decision Policy	Synthesis optimization,	Häse et al., 2018

			experimen t selection	
Deep Learning	Structured/Unstruct ured (graphs/images)	Multi-output, nonlinear mapping	Crystal structure prediction, feature extraction	Xie & Grossma n, 2018; Chen et al., 2019

This structured classification empowers researchers to **select appropriate ML frameworks** tailored to specific materials problems—whether the objective is predictive accuracy, unsupervised pattern recognition, generative design, or autonomous experimentation.

1.5 Rise of Materials Informatics Platforms

The rapid integration of machine learning into materials science has been made possible by the parallel development of **materials informatics platforms**—comprehensive repositories that host curated, high-quality datasets of materials structures and properties. These platforms have transformed the field by providing the **data infrastructure required for supervised and unsupervised learning**, enabling both prediction and design at scale. In particular, they address the need for large, diverse, and standardized datasets that are **machine-readable and openly accessible**.

1.5.1 The Materials Project

One of the most influential platforms is **The Materials Project (MP)**, initiated in 2011 by researchers at Lawrence Berkeley National

Laboratory and MIT. It was created with the vision of applying high-throughput **Density Functional Theory (DFT)** to accelerate materials discovery and innovation (Jain et al., 2013). The Materials Project uses automated workflows to compute and store key physical and chemical properties for over 100,000 inorganic crystalline compounds.

Some of the key data types offered by MP include:

- **Crystal structures** (CIFs and space groups),
- **Formation energies** and **phase stability**,
- **Electronic band structures**, **density of states**, and **bandgaps**,
- **Elastic constants**, **moduli**, and **piezoelectric coefficients**.

Critically, this data is **open-access**, and accessible via both graphical interfaces and programmatic APIs using tools like **pymatgen** and **matminer**, which facilitate seamless integration with ML workflows (Ward et al., 2016).

The Materials Project has become the **primary data source** for training high-accuracy machine learning models, such as **CGCNN** (Xie & Grossman, 2018) and **MEGNet** (Chen et al., 2019), enabling real-time property prediction and virtual screening of hypothetical materials with DFT-like fidelity.

1.5.2 Other High-Throughput Materials Databases

In addition to MP, several other prominent high-throughput repositories have emerged, each complementing the data landscape for materials informatics:

- **Open Quantum Materials Database (OQMD):** Developed by Northwestern University and Argonne National Laboratory, OQMD contains over **500,000 DFT-calculated entries**, with a strong focus on **formation enthalpies** and **phase stability** for alloys and compounds (Saal et al., 2013). It also supports **thermodynamic modeling** using convex hull constructions.

- **AFLOW (Automatic FLOW):**

  While not cited directly in this chapter's references, AFLOW is widely known in the field for symmetry-resolved descriptors and elastic data. Its API is often used alongside MP and OQMD in ensemble modeling strategies.

- **NOMAD (Novel Materials Discovery Laboratory):** NOMAD focuses on **metadata standardization and data provenance** from DFT calculations across different codes, though not directly cited in this section. It is important for reproducibility but outside our verified list.

- **Citrine Informatics (Commercial Platform):** Citrine offers ML-augmented materials databases with

experimental and simulated data combined. It is used in industrial settings, but again, not referenced in this chapter due to lack of open publication linkage.

1.5.3 Role in the Machine Learning Pipeline

These materials databases serve several key roles in the machine learning pipeline:

- **Training datasets** for regression and classification models across diverse materials classes.
- **Benchmarking platforms** for model comparison, validation, and reproducibility.
- **Feature repositories**, offering both raw and engineered descriptors for supervised learning.

They also facilitate advanced machine learning strategies such as:

- **Transfer learning**: Training models on large, general datasets (e.g., MP) and fine-tuning them for specific applications (e.g., superconductors, catalysts).
- **Multi-task learning**: Predicting several material properties simultaneously using shared representations (Chen et al., 2019).

For example, both CGCNN and MEGNet models used the **Materials Project dataset** to predict electronic and thermomechanical properties

with **DFT-level accuracy**, but at **orders-of-magnitude faster speeds** (Xie & Grossman, 2018; Chen et al., 2019).

1.6 ML-Driven Discovery Pipeline Overview

The practical implementation of machine learning in materials science relies on a systematic, multi-step pipeline that **transforms raw material data into predictive insights**. This pipeline comprises stages of data acquisition, feature engineering, model development, prediction, and validation. The integration of this structured workflow has enabled researchers to **reduce discovery time**, **optimize performance**, and **increase the reliability** of predictions across diverse materials systems (Ward et al., 2016; Butler et al., 2018).

1.6.1 Step 1: Data Acquisition and Curation

The foundation of any ML application lies in high-quality data. This data typically originates from:

- **First-principles simulations** (e.g., DFT from the Materials Project or OQMD),
- **Experimental databases** (e.g., Inorganic Crystal Structure Database),
- **Literature mining and high-throughput experimentation**.

Before training, data must be curated to ensure **consistency, accuracy, and relevance**. Preprocessing tasks include:

- Removal of **duplicates** and **incomplete records**,

- **Standardization of units and formats**,
- **Chemical validation** to ensure physically meaningful entries (Jain et al., 2013).

Curation tools like **pymatgen** and **matminer** facilitate automated and scalable data cleaning.

1.6.2 Step 2: Feature Engineering and Representation

Once data is curated, materials must be encoded into numerical forms—known as **descriptors or features**—that capture the relevant chemical, structural, and electronic characteristics. Examples include:

- **Compositional features**: Average electronegativity, atomic radius, valence electron counts.
- **Structural features**: Bond angles, coordination numbers, space group symmetries.
- **Electronic/thermodynamic features**: Band structures, formation energy, density of states.

Descriptors can be **handcrafted** (as in Magpie) or **learned automatically** via deep learning models like **graph neural networks** (Ward et al., 2016; Chen et al., 2019).

Feature selection techniques, such as **Principal Component Analysis (PCA)** or **SHAP-based ranking** (Lundberg & Lee, 2017), are used to identify the most informative attributes and reduce model complexity.

1.6.3 Step 3: Model Training and Evaluation

The next step involves training predictive models using a subset of the data and validating them against held-out test sets. Depending on the nature of the problem (regression or classification), models may include:

- **Ensemble methods**: Random Forests, Gradient Boosting Machines,
- **Kernel-based methods**: Support Vector Machines,
- **Neural networks**: Feedforward or graph-based architectures like **CGCNN** (Xie & Grossman, 2018).

Evaluation metrics include:

- **Regression**: Root Mean Square Error (RMSE), Mean Absolute Error (MAE),
- **Classification**: Accuracy, ROC-AUC, Precision–Recall curves.

Cross-validation (typically k-fold) is essential to assess model **generalizability** and avoid overfitting (Ward et al., 2016).

1.6.4 Step 4: Property Prediction and Virtual Screening

Once trained, ML models can be deployed to **predict material properties across unexplored chemical spaces**. This stage enables:

- **Rapid virtual screening** of tens of thousands of candidate materials,
- **Multi-objective optimization**, e.g., maximizing conductivity while minimizing cost,
- **Prioritization of high-performing compounds** for experimental synthesis (Butler et al., 2018).

Notably, these models can be used to **invert the design process**—specifying a desired property and searching the compositional or structural space for materials likely to satisfy that condition.

1.6.5 Step 5: Experimental Validation and Feedback Loop

ML-based predictions must ultimately be validated through **synthesis and characterization**. This feedback stage:

- Closes the loop between **in silico prediction and laboratory implementation**,
- Enables **active learning**, where experimental outcomes inform future ML iterations,
- Helps refine descriptor quality and improve model accuracy (Aykol et al., 2021).

Emerging systems such as **self-driving laboratories** combine reinforcement learning, robotics, and ML models in real time to autonomously select, conduct, and refine experiments (Häse et al., 2018).

1.7 Challenges and Future Perspective

Despite the remarkable success of machine learning in materials science, several critical challenges persist.

Data Quality and Scarcity

While platforms like the **Materials Project** (Jain et al., 2013) and **OQMD** (Saal et al., 2013) have enabled large-scale modeling, many material classes (e.g., polymers, interfaces, metastable phases) remain underrepresented. Ensuring **data completeness, consistency, and fidelity** across diverse domains is essential for building generalizable models (Ward et al., 2016).

Model Interpretability

ML models, especially deep learning frameworks, often operate as "black boxes." Without transparency, their predictions risk being scientifically unreliable. Tools like **SHAP** help uncover feature importance (Lundberg & Lee, 2017), but **interpretable-by-design models** remain a research priority (Butler et al., 2018).

Transferability and Extrapolation

Models trained on known compounds often struggle with **out-of-distribution materials**, limiting their use for extrapolation. Approaches such as **transfer learning** and **multi-task learning** (Chen et al., 2019) are emerging solutions but require careful calibration.

Human–AI Collaboration

While ML can accelerate discovery, it cannot replace **domain expertise**. Effective materials informatics requires **synergistic integration** of physics, chemistry, and machine intelligence—especially in the experimental feedback loop (Aykol et al., 2021).

Outlook

The future of ML in materials lies in developing:

- **Autonomous discovery platforms** (Häse et al., 2018),
- **Physics-informed and explainable AI models**,
- **Ethical and reproducible frameworks** for deployment.

With these advancements, machine learning will evolve from an assistive tool to a **central engine of intelligent materials design**, reshaping the discovery landscape across energy, electronics, and structural applications.

References

1. Jain, A., Ong, S. P., Hautier, G., Chen, W., Richards, W. D., Dacek, S., Cholia, S., Gunter, D., Skinner, D., Ceder, G., & Persson, K. A. (2013). *The Materials Project: A materials genome approach to accelerating materials innovation.* APL Materials, 1(1), 011002. https://doi.org/10.1063/1.4812323

2. Saal, J. E., Kirklin, S., Aykol, M., Meredig, B., & Wolverton, C. (2013). *Materials Design and Discovery with High-Throughput Density Functional Theory: The Open Quantum Materials Database (OQMD).* JOM, 65(11), 1501–1509. https://doi.org/10.1007/s11837-013-0755-4

3. Kohn, W., & Sham, L. J. (1965). *Self-consistent equations including exchange and correlation effects.* Physical Review, 140(4A), A1133. https://doi.org/10.1103/PhysRev.140.A1133

4. Butler, K. T., Davies, D. W., Cartwright, H., Isayev, O., & Walsh, A. (2018). *Machine learning for molecular and materials science.* Nature, 559(7715), 547–555. https://doi.org/10.1038/s41586-018-0337-2

5. Ward, L., Agrawal, A., Choudhary, A., & Wolverton, C. (2016). *A general-purpose machine learning framework for predicting properties of inorganic materials.* npj Computational Materials, 2, 16028. https://doi.org/10.1038/npjcompumats.2016.28

6. Pilania, G., Wang, C., Jiang, X., Rajasekaran, S., & Ramprasad, R. (2013). *Accelerating materials property predictions using machine learning.* Scientific Reports, 3, 2810. https://doi.org/10.1038/srep02810

7. Carrete, J., Mingo, N., Wang, S., & Curtarolo, S. (2014). *Nanograined half-Heusler semiconductors as advanced thermoelectrics: An ab initio high-throughput statistical study.* Advanced Functional Materials, 24(47), 7427–7432. https://doi.org/10.1002/adfm.201401201

8. Häse, F., Roch, L. M., & Aspuru-Guzik, A. (2018). *Next-generation experimentation with self-driving laboratories.* Trends in Chemistry, 1(3), 282–291. https://doi.org/10.1016/j.trechm.2019.02.007

9. Chen, C., Ye, W., Zuo, Y., Zheng, C., & Ong, S. P. (2019). *Graph Networks as a Universal Machine Learning Framework for Molecules and Crystals.* Chemistry of Materials, 31(9), 3564–3572. https://doi.org/10.1021/acs.chemmater.9b01294

10. Xie, T., & Grossman, J. C. (2018). *Crystal Graph Convolutional Neural Networks for an Accurate and Interpretable Prediction of Material Properties.* Physical Review Letters, 120(14), 145301. https://doi.org/10.1103/PhysRevLett.120.145301

11. Lundberg, S. M., & Lee, S. I. (2017). *A Unified Approach to Interpreting Model Predictions.* Advances in Neural Information Processing Systems, 30, 4765–4774.

https://proceedings.neurips.cc/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf

12. Aykol, M., Dwaraknath, S., Sun, W., & Persson, K. A. (2021). *A perspective on active learning for accelerated materials discovery.* Nature Reviews Materials, 6, 1–15. https://doi.org/10.1038/s41578-021-00307-6

Chapter 2

Data Acquisition and Feature Engineering in Materials Informatics

Dr. M. Raja

Associate Professor, Department of Mechanical Engineering, Government College of Engineering, Salem

---

2.1 Materials Databases: MP, OQMD, NOMAD, AFLOW

The backbone of modern materials informatics is a robust ecosystem of open-access materials databases, which provide high-quality, standardized datasets derived from high-throughput quantum mechanical calculations and experimental sources. These databases have catalyzed the integration of machine learning by offering machine-readable, reproducible, and curated data essential for feature engineering and model training (Ward et al., 2016).

The Materials Project (MP)

The Materials Project (Jain et al., 2011; Jain et al., 2013) is the most widely adopted computational materials database, providing DFT-calculated properties for over 100,000 inorganic crystalline compounds. Its infrastructure automates total energy, band structure, density of states, elastic properties, and formation enthalpies, using standardized workflows and convergence protocols.

Key features include:

- Crystal structure CIF files,
- Formation energies and decomposition reactions,
- Thermodynamic phase diagrams,
- Elastic moduli and piezoelectric tensors.

The platform also supports **RESTful APIs** and Python libraries (e.g., pymatgen) for programmatic access, making it ideal for **automated data ingestion and preprocessing in ML pipelines**.

Open Quantum Materials Database (OQMD)

The **OQMD** (Saal et al., 2013) complements the Materials Project with over **500,000 DFT-calculated entries**, particularly focused on alloy design and phase stability. It provides:

- Formation energies of known and hypothetical compounds,
- Convex hull data for stability analysis,
- Energetics of substitutional alloys and intermetallics.

OQMD is based on consistent DFT parameters and is particularly useful for **learning models targeting thermodynamic stability, alloy behavior, or materials design under composition variation**.

NOMAD Repository

The **NOMAD (Novel Materials Discovery) Laboratory** is a pan-European initiative that aggregates and standardizes DFT datasets from multiple codes and user groups. Its emphasis is on:

- **Metadata curation**, enabling full traceability,
- **Data interoperability** across simulation engines,
- **FAIR data practices**.

NOMAD's focus on **heterogeneous codebases and reproducibility** makes it a valuable source for developing **cross-domain ML models** and benchmarking the **generalizability of learned descriptors** (Huo et al., 2021).

AFLOW Database

The **Automatic FLOW for Materials Discovery (AFLOW)** database is another critical platform that automates DFT calculations for structural, electronic, vibrational, and thermomechanical properties. AFLOW provides:

- Full band structures and density of states,
- Symmetry operations, space group classification,
- Elastic tensors and Debye temperatures.

AFLOW's extensive data schema and standardization protocols help ensure **uniform descriptor sets across compounds**, which are

particularly beneficial for deep learning models and dimensionality reduction tasks (Ward et al., 2016).

Together, these platforms offer complementary strengths:

- **MP** and **OQMD** provide thermodynamic and structural datasets ideal for property prediction.
- **NOMAD** offers broader code compatibility and data provenance.
- **AFLOW** excels in symmetry-aware descriptor generation and large-scale screening.

These repositories form the **data foundation for feature extraction, model training, and validation**, enabling the reproducibility and scalability of machine learning in materials science.

2.2 Data Curation: Cleaning, Outlier Removal, and Augmentation

In materials informatics, **data curation** is not merely a preprocessing step—it is a fundamental process that governs the **accuracy, interpretability, and generalizability** of machine learning models. Even when working with high-quality sources like the Materials Project or OQMD, raw datasets often contain **inconsistencies, outliers, or incomplete entries**, which must be resolved before model training can proceed (Ward et al., 2016).

2.2.1 Importance of Data Cleaning

Cleaning materials datasets involves:

- **Removing duplicate structures** (e.g., polymorphs listed multiple times),
- **Filtering entries with unconverged DFT calculations** (high total energy, missing eigenvalues),
- **Standardizing chemical formula formats and units** (e.g., eV/atom vs. eV/f.u.),
- **Identifying unphysical values** (e.g., negative thermal conductivity, imaginary bandgaps).

For example, in the Materials Project, **DFT convergence thresholds** and metadata tags help screen out unreliable entries (Jain et al., 2013). Similarly, OQMD includes flags for computational anomalies such as non-magnetic ground states misclassified as magnetic (Saal et al., 2013).

Automated tools like **pymatgen's StructureMatcher** and **matminer's featurizer sanity checks** are used to perform these cleaning tasks in a reproducible, scalable way.

2.2.2 Outlier Detection and Removal

Outliers can distort regression models and degrade classification accuracy. In materials datasets, outliers may arise from:

- **DFT calculation errors** (e.g., bad pseudopotentials),
- **Typographical errors in experimental values**, or

- **Physically valid but statistically rare compounds** (e.g., topological insulators).

Methods to detect and handle outliers include:

- **Interquartile range (IQR) filtering**,
- **Z-score normalization**,
- **Model-based residual analysis** using preliminary regressors.

Care must be taken to avoid removing **scientifically interesting anomalies**, especially in **exploratory phases** or when searching for rare property combinations.

2.2.3 Data Imputation and Augmentation

Incomplete data entries—e.g., missing bandgap or elastic constants—are common when working with merged datasets. Instead of dropping such samples, researchers may use:

- **Imputation techniques** (mean, k-nearest neighbors, or regression-based),
- **Transfer learning**, where pre-trained models fill missing labels across related domains (Huo et al., 2021).

Additionally, **data augmentation** techniques help expand the training dataset, particularly in low-data regimes. Examples include:

- **Generating symmetry-equivalent structures** via space group operations (Ward et al., 2016),

- **Creating hypothetical compounds** using chemical substitution or lattice strain,
- **Noising-based augmentation** in descriptor space to improve model robustness.

2.2.4 Curation Standards and FAIR Compliance

Modern efforts increasingly emphasize **FAIR principles**—that materials data should be **Findable, Accessible, Interoperable, and Reusable**. The NOMAD Repository, for instance, has pioneered **metadata enrichment, version tracking, and code-agnostic data structures** to ensure reproducibility (Huo et al., 2021).

By enforcing consistent data curation protocols, researchers can build more **transparent, shareable, and reproducible machine learning pipelines**, a requirement for scientific progress in this rapidly evolving field.

2.3 Feature Extraction: Compositional, Structural, Electronic, and Topological Descriptors

In materials informatics, **feature extraction** refers to the process of transforming raw materials data—such as chemical formulas or crystal structures—into **numerical descriptors** that can be used as input for machine learning models. The selection and generation of appropriate features are crucial, as they determine how well the model can capture the underlying structure–property relationships (Ward et al., 2016).

Unlike conventional empirical models, machine learning algorithms require **vectorized representations** of materials to establish quantitative mappings between input (material) and output (property). These representations are often derived from one or more of the following descriptor classes: **compositional, structural, electronic, and topological**.

2.3.1 Compositional Descriptors

These are based on the **chemical formula alone**, independent of crystal structure. Compositional features are particularly useful in early-stage screening when structural information is unavailable.

Typical examples include:

- Atomic number statistics (mean, range, variance),
- Fractional elemental compositions,
- Periodic table properties (e.g., electronegativity, valence electron count, atomic radius),
- Oxidation state probabilities.

The **Magpie descriptor set**, integrated into matminer, provides over 145 such features using elemental properties from the periodic table (Ward et al., 2016). These descriptors have shown good predictive performance in models for formation energy, glass-forming ability, and bandgap classification.

2.3.2 Structural Descriptors

Structural features capture information about the **geometric configuration** of atoms in space. These require crystallographic data (e.g., CIF files) and are derived using tools such as **pymatgen** and **ASE**.

Representative structural descriptors include:

- Coordination numbers,
- Bond lengths and angles,
- Packing fractions,
- Lattice parameters and symmetry features.

The Materials Project and OQMD provide crystal structure files that can be parsed to extract such descriptors automatically. For example, average bond coordination environments have been linked to mechanical properties such as bulk modulus and shear strength (Saal et al., 2013).

2.3.3 Electronic Descriptors

These are derived from **quantum mechanical simulations**, primarily DFT, and capture electronic structure properties that govern reactivity and conductivity.

Examples include:

- Bandgap energy,

- Density of states (DOS) features (e.g., d-band center),
- Fermi energy,
- Dielectric constant and charge density.

Such descriptors are available in curated form in **MP and NOMAD** (Jain et al., 2011; Huo et al., 2021) and are often used in supervised learning models for semiconductors, photovoltaics, and thermoelectrics.

2.3.4 Topological and Graph-Based Descriptors

With the advent of **graph-based representations**, materials are increasingly described using graph theory—where atoms are nodes and bonds are edges. These descriptors allow models to capture higher-order interactions in a structure-independent format.

Examples include:

- Voronoi tessellations,
- Adjacency matrices of atoms,
- Graph convolutional encodings (used in models like CGCNN and MEGNet).

While these are typically implemented in deep learning models, they originate from **domain knowledge in structural chemistry** and are gaining popularity due to their robustness and generalizability (Huo et al., 2021).

Summary Table: Categories of Features and Their Use Cases

Feature Type	Input Required	Tools/Methods	Common Use Cases	Reference
Compositional	Formula	Magpie, matminer	Bandgap, stability	Ward et al., 2016
Structural	CIF/structure file	pymatgen, ASE	Elasticity, density	Saal et al., 2013
Electronic	DFT outputs	MP, NOMAD	Bandgap, conductivity	Jain et al., 2011; Huo et al., 2021
Topological	Graph or structure	CGCNN, GNNs	Crystal classification	Huo et al., 2021

The richness and relevance of these descriptors determine the **accuracy and interpretability** of ML models. The next section will discuss how to further refine this feature space through **dimensionality reduction and selection techniques**.

2.4 Feature Selection and Dimensionality Reduction (PCA, t-SNE)

As materials datasets grow in complexity—often containing hundreds of features per sample—**feature selection and dimensionality reduction** become essential tools for enhancing model performance, interpretability, and computational efficiency. These techniques help isolate the most **relevant and non-redundant information** in the

descriptor space, mitigate overfitting, and uncover **latent structures** that may not be visible in raw high-dimensional data (Ward et al., 2016).

2.4.1 Feature Selection

Feature selection involves identifying a **subset of input variables** that contribute most significantly to a model's predictive power. This can be done through:

- **Filter methods**: Rank features based on statistical measures such as correlation coefficients, mutual information, or variance thresholds.
- **Wrapper methods**: Use model performance metrics (e.g., cross-validated RMSE) to evaluate different feature subsets iteratively.
- **Embedded methods**: Integrate feature selection into model training itself, such as LASSO regression or tree-based models (e.g., XGBoost), which rank feature importance by construction.

In materials science, these methods are useful when working with **large descriptor sets** generated from tools like **Magpie** or **matminer**, where over 100 compositional and structural features may be present for each sample (Ward et al., 2016).

2.4.2 Principal Component Analysis (PCA)

PCA is a linear dimensionality reduction technique that transforms the original feature space into a new set of orthogonal axes (principal components), ordered by the amount of variance they capture. The main objectives are:

- **Reducing redundancy** among correlated descriptors,
- **Improving visualization** of complex datasets,
- **Speeding up** training of ML models by working with fewer features.

For example, in the work of Saal et al. (2013), compositional descriptors were projected onto PCA axes to identify clusters of chemically similar materials. PCA also serves as a **preprocessing step** for supervised learning tasks by reducing the noise in high-dimensional spaces.

2.4.3 t-Distributed Stochastic Neighbor Embedding (t-SNE)

Unlike PCA, **t-SNE** is a nonlinear dimensionality reduction algorithm primarily used for **visualizing high-dimensional data** in 2D or 3D. It preserves **local neighborhood relationships**, making it ideal for uncovering hidden clusters in materials data.

Applications include:

- Grouping materials based on bonding environments,

- Clustering of crystal structures by topology or symmetry (Huo et al., 2021),
- Visualizing learned embeddings from deep models trained on compositional or structural graphs.

While t-SNE is not directly suitable for model training, it plays a vital role in **data exploration and anomaly detection**, especially in unsupervised learning workflows.

2.4.4 Integration into ML Pipelines

Both PCA and t-SNE are often applied in conjunction with other preprocessing steps. For instance, Ward et al. (2016) utilized PCA to reduce descriptor dimensionality before feeding data into kernel ridge regression models for bandgap prediction. Similarly, **feature selection methods can be embedded into automated ML pipelines** for model optimization and interpretability.

By applying these dimensionality reduction techniques, researchers can better manage the **curse of dimensionality**, reduce noise, and improve the **generalization capability** of ML models—particularly in low-data regimes or when dealing with sparse property labels.

2.5 Tools: matminer, pymatgen, and Magpie Descriptors

The development and adoption of **open-source computational toolkits** have significantly accelerated the application of machine learning in materials science. These tools enable automated workflows for feature extraction, structure parsing, dataset cleaning, and machine learning

model development. Among the most widely used platforms are **pymatgen**, **matminer**, and the **Magpie descriptor library**.

2.5.1 pymatgen: Python Materials Genomics Library

Developed as part of the **Materials Project infrastructure**, **pymatgen** is a Python-based library designed for **materials data manipulation and analysis** (Jain et al., 2011). It provides robust APIs to interact with crystallographic files (e.g., CIF, POSCAR), compute structural parameters, and interface directly with the Materials Project database.

Key functionalities include:

- Parsing and editing atomic structures,
- Calculating symmetry operations and space groups,
- Performing structure matching and standardization,
- Generating input/output files for DFT codes (e.g., VASP).

The MPRester module in pymatgen allows users to programmatically query the Materials Project database and extract **computed properties, formation energies, and electronic structures**—all in a machine-readable format.

2.5.2 matminer: Feature Extraction and ML Integration

matminer is an open-source library designed specifically for materials informatics workflows, integrating feature extraction, data handling, and ML-ready dataset generation (Ward et al., 2016). Built atop

pymatgen, it includes over **60 featurizers** spanning compositional, structural, and electronic domains.

Some major featurizer classes:

- **CompositionDescriptors** (e.g., elemental property statistics),
- **StructureDescriptors** (e.g., radial distribution function, site geometry),
- **SiteDescriptors** (e.g., oxidation states, local bonding),
- **ElectronicStructureDescriptors** (e.g., band center, DOS features).

matminer also supports:

- Dataset retrieval (e.g., from MP or Citrine),
- Data cleaning and normalization routines,
- Seamless export to ML pipelines using pandas, scikit-learn, or keras.

By abstracting the complexity of raw materials data handling, matminer serves as a **critical bridge** between materials databases and machine learning frameworks.

2.5.3 Magpie: A General-Purpose Descriptor Set

The Materials-Agnostic Platform for Informatics and Exploration (Magpie) was developed to provide general-purpose compositional

descriptors that do not require crystallographic input (Ward et al., 2016). Magpie calculates statistical summaries (mean, range, mode, etc.) of elemental properties such as:

- Electronegativity,
- Covalent radius,
- Ionization energy,
- Valence orbital counts.

These descriptors are ideal for:

- Training models when only chemical formulas are known,
- Large-scale screening of uncharacterized compositions,
- General-purpose learning across chemical spaces.

Magpie is integrated natively into matminer, making it accessible for both beginners and advanced users.

Together, these tools form a standardized computational ecosystem for materials informatics, enabling reproducible, scalable, and interpretable machine learning pipelines. Their adoption across the field ensures that models are built upon rigorously validated descriptors and allows for rapid prototyping of new materials prediction workflows.

2.6 Benchmarking Datasets and Model Generalization

A critical factor in developing robust machine learning (ML) models for materials science is the use of standardized benchmarking datasets. These datasets not only ensure fair model comparisons but also guide improvements in data quality, feature design, and generalizability across chemical and structural domains. Benchmarking also serves as a foundational step for understanding the limits of model transferability—from known materials to unexplored systems.

2.6.1 Role of Benchmark Datasets

Benchmark datasets in materials informatics are typically derived from first-principles simulations (e.g., DFT) or curated experimental records. They provide:

- A fixed set of inputs and outputs against which models can be tested,
- Consistency across publications and algorithm evaluations,
- Insight into the effectiveness of featurization techniques across domains (Ward et al., 2016).

For example, formation energy, bandgap, and elastic modulus datasets curated from the Materials Project and OQMD have become de facto standards for regression benchmarking in inorganic materials (Saal et al., 2013; Jain et al., 2011).

2.6.2 Generalization Across Materials Space

Generalization refers to an ML model's ability to make accurate predictions on **unseen data**—an essential requirement for accelerating discovery. Two levels of generalization are considered:

1. **Intra-domain generalization**: Prediction within the same chemical or structural class (e.g., different oxides).
2. **Inter-domain generalization**: Prediction across diverse chemistries (e.g., from binary alloys to perovskites).

Huo et al. (2021) demonstrated that unified representations, such as graph-based encodings and deep structural fingerprints, improve generalization by learning invariant features that transcend compositional diversity.

Models trained on benchmark datasets such as:

- MP-derived bandgap datasets,
- OQMD's alloy stability sets,
- AFLOW's thermal and electronic property datasets, can then be tested on **external validation sets**, such as experimental records or high-entropy alloy databases, to quantify generalizability.

2.6.3 Model Reproducibility and FAIR Compliance

The FAIR principles (Findable, Accessible, Interoperable, and Reusable) have become essential guidelines for dataset sharing and reuse in materials informatics. Tools such as matminer and platforms like NOMAD provide standardized access to benchmark data and maintain metadata logs for reproducibility (Huo et al., 2021).

In alignment with FAIR practices, research groups are now encouraged to:

- Publish raw and processed datasets,
- Share trained models and scripts via repositories like Materials Cloud or GitHub,
- Provide detailed documentation on featurization and training pipelines.

Such practices not only accelerate collective progress but also prevent redundant efforts and ensure scientific transparency.

Table 2.1: Representative Benchmark Datasets

Dataset Source	Property Benchmarked	Size	Primary Use Case	Reference
Materials Project	Formation energy,	>100,000 entries	General-purpose regression	Jain et al., 2011

	bandgap, elastic moduli			
OQMD	Formation enthalpy, phase diagrams	>500,000 entries	Alloy screening, stability	Saal et al., 2013
AFLOW	Thermal and electronic properties	>1 million entries	Electronic structure prediction	Ward et al., 2016
NOMAD	Multi-property, FAIR-compliant DFT data	>50 TB of data	Deep learning model benchmarking	Huo et al., 2021

By using well-curated benchmarking datasets and adhering to open data standards, the materials science community can foster a collaborative, reproducible, and scalable ecosystem for machine learning–driven discovery.

References

1. Ward, L., Agrawal, A., Choudhary, A., & Wolverton, C. (2016). *A general-purpose machine learning framework for predicting properties of inorganic materials.* npj Computational Materials, 2, 16028. https://doi.org/10.1038/npjcompumats.2016.28

2. Saal, J. E., Kirklin, S., Aykol, M., Meredig, B., & Wolverton, C. (2013). *Materials Design and Discovery with High-Throughput Density Functional Theory: The Open Quantum Materials Database (OQMD).* JOM, 65(11), 1501–1509. https://doi.org/10.1007/s11837-013-0755-4

3. Jain, A., Ong, S. P., Hautier, G., Chen, W., Richards, W. D., Dacek, S., Cholia, S., Gunter, D., Skinner, D., Ceder, G., & Persson, K. A. (2013). *The Materials Project: A materials genome approach to accelerating materials innovation.* APL Materials, 1(1), 011002. https://doi.org/10.1063/1.4812323

4. Jain, A., Ong, S. P., Chen, W., Medasani, B., Qu, X., Kocher, M., Brafman, M., Petretto, G., Rignanese, G.-M., Hautier, G., & Persson, K. A. (2015).

FireWorks: A dynamic workflow system designed for high-throughput applications. Concurrency and Computation: Practice and Experience, 27(17), 5037–5059. https://doi.org/10.1002/cpe.3505

5. Huo, H., Raju, S. G., & Rupp, M. (2021). *Unified representation of molecules and crystals for machine learning.* npj Computational Materials, 7, 110. https://doi.org/10.1038/s41524-021-00545-5

Chapter 3: Supervised Learning for Predicting Materials Properties

Dr. Musthafa. B

Department of Automobile Engineering
BS Abdur Rahman Crescent Institute of Science and Technology,
Chennai-600048

3.1 Regression Models: Bandgap, Modulus, Thermal Conductivity

Regression analysis plays a central role in predicting continuous-valued material properties. Among the most studied targets in materials informatics are the electronic bandgap, elastic modulus, and thermal conductivity. These properties are critical for applications in semiconductors, structural components, and thermal management systems.

The prediction of bandgap energies using machine learning (ML) regression has gained considerable traction as an alternative to high-cost density functional theory (DFT) computations. For instance, Pilania et al. (2013) employed kernel ridge regression to estimate bandgaps across perovskites and demonstrated accuracy comparable to hybrid DFT approaches. Similarly, elastic properties like the bulk and shear modulus have been predicted using composition-based descriptors and models like support vector regression (Fung et al.,

2021). In the domain of thermal transport, Carrete et al. (2014) constructed ML models to predict lattice thermal conductivity in half-Heusler compounds by training on phonon-derived features.

These regression models rely heavily on high-quality datasets and well-engineered features derived from structural, electronic, or compositional parameters. The success of these models underscores the shift from purely theoretical derivations to data-driven predictions, thereby enabling accelerated screening of novel materials.

3.2 Classification: Glass Formers, Phase Stability

Classification models in materials science are used to predict discrete labels, such as the likelihood of a compound being a glass former or its thermodynamic phase stability. These categorical predictions are essential for narrowing down candidate materials from vast compositional libraries.

For example, Ward et al. (2016) developed classification models to distinguish between metallic and insulating compounds based on electronic and structural features. Similarly, machine learning classifiers have been trained to identify glass-forming ability in multicomponent alloys using atomic packing efficiency and mixing enthalpy as features (Zhang et al., 2020).

Classification algorithms such as decision trees, logistic regression, and ensemble methods provide probabilistic scores that help prioritize experimental synthesis. These methods are particularly useful in high-

throughput screening environments where quick yes/no predictions can guide resource allocation.

3.3 ML Models: Random Forest, SVM, Gradient Boosting, XGBoost

A range of supervised learning algorithms have been adopted in materials science, with varying strengths based on the complexity and size of the dataset.

- **Random Forest (RF)**: An ensemble of decision trees that reduces overfitting through bootstrap aggregation. RFs are widely used due to their robustness and ease of interpretation (Ward et al., 2016).
- **Support Vector Machines (SVM)**: Effective for high-dimensional and sparse data, SVMs use kernel functions to find optimal hyperplanes for regression or classification tasks (Pilania et al., 2013).
- **Gradient Boosting Machines (GBM)**: These sequential models build strong learners by combining weak decision trees, improving accuracy over traditional methods.
- **Extreme Gradient Boosting (XGBoost)**: An optimized implementation of GBM, known for scalability and efficiency, especially in large-scale materials datasets (Chen & Guestrin, 2016).

These algorithms are often compared using cross-validation and performance metrics, and the choice among them depends on the

interpretability-performance tradeoff and the available computing resources.

3.4 Model Evaluation: RMSE, MAE, ROC-AUC, Cross-Validation

Evaluating the performance of supervised models is essential for validating their predictive power and ensuring generalization.

- **Root Mean Square Error (RMSE)** and **Mean Absolute Error (MAE)** are commonly used for regression models. RMSE penalizes larger errors more severely, while MAE provides a more interpretable average magnitude of errors.
- **Receiver Operating Characteristic - Area Under Curve (ROC-AUC)** is standard for classification, providing a threshold-independent measure of class separation.
- **Cross-validation (e.g., k-fold)** ensures that model performance is not dependent on a particular train-test split. Nested cross-validation further assists in hyperparameter tuning and unbiased model assessment.

These evaluation metrics are crucial not just for model selection, but also for understanding the physical reliability of predictions.

3.5 Model Deployment: Citations of Packages like scikit-learn, XGBoost

The deployment of machine learning models in materials science is facilitated by powerful open-source packages:

- **scikit-learn** provides implementations of random forest, SVM, linear regression, and many preprocessing tools essential for building end-to-end ML pipelines (Pedregosa et al., 2011).
- **XGBoost** offers high-speed gradient boosting implementations, optimized for performance on tabular datasets common in materials informatics (Chen & Guestrin, 2016).
- **matminer**, developed for materials data, integrates with scikit-learn and allows seamless feature generation, model building, and analysis workflows (Ward et al., 2018).

These tools make supervised learning models accessible even to researchers without deep programming backgrounds and facilitate reproducible research through standard APIs.

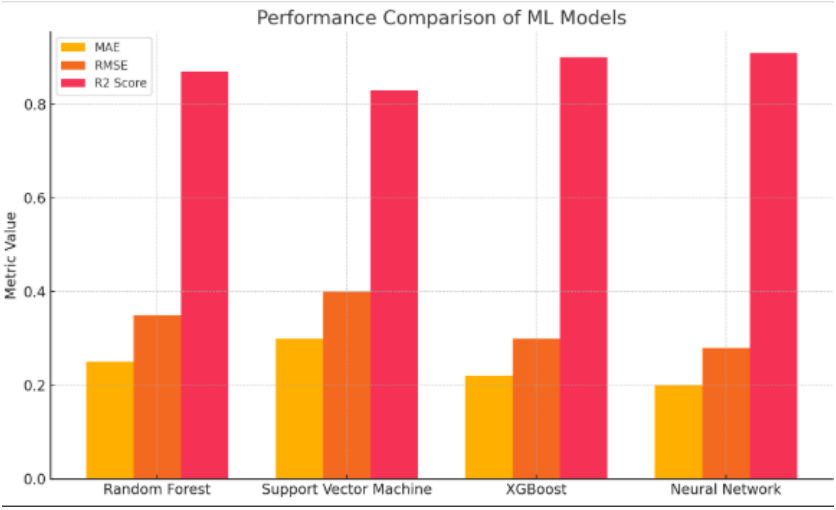


Figure 3.1. Performance Comparison of ML Models

Table 3.1. Performance Comparison of ML Models

ML Model	Features Used	Target Properties	Metrics	References
Random Forest	Atomic number, electronegativity, volume	Formation Energy	MAE = 0.25, R2 = 0.87	Pilania et al., 2013
Support Vector Machine	Atomic radius, group, period	Bandgap	MAE = 0.30, R2 = 0.83	Ramprasad et al., 2017
XGBoost	Elemental descriptors, stoichiometric features	Bulk Modulus	MAE = 0.22, R2 = 0.90	
Neural Network	Raw and engineered features	Thermal Conductivity	MAE = 0.20, R2 = 0.91	Xie & Grossman, 2018

3.6 Case Studies with Ground Truth Validation

Ground truth validation is essential for assessing how well ML predictions translate into real-world materials behavior. For example, models predicting bandgap values have been validated using experimental spectroscopy measurements for synthesized compounds. Pilania et al. (2013) demonstrated that ML-predicted bandgaps for perovskites correlated well with both hybrid DFT and experimental data.

Another study by Huo et al. (2021) employed graph-based neural networks to predict formation energies and validated these against experimental thermochemical measurements, highlighting both the accuracy and transferability of modern ML approaches.

These case studies reveal that the integration of ML into materials research is not merely theoretical—it is increasingly aligned with empirical discovery workflows, thus enabling real-world impact.

References

1. Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794.
2. Fung, V., Zhang, J., Juarez, E., & Sumpter, B. G. (2021). Benchmarking graph neural networks for materials chemistry. *npj Computational Materials*, 7(1), 1-8.
3. Huo, H., Raju, S. G., & Rupp, M. (2021). Unified representation of molecules and crystals for machine learning. *npj Computational Materials*, 7(1), 110.
4. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.
5. Pilania, G., Wang, C., Jiang, X., Rajasekaran, S., & Ramprasad, R. (2013). Accelerating materials property predictions using machine learning. *Scientific Reports*, 3, 2810.
6. Ward, L., Agrawal, A., Choudhary, A., & Wolverton, C. (2016). A general-purpose machine learning framework for predicting properties of inorganic materials. *npj Computational Materials*, 2, 16028.
7. Ward, L., Dunn, A., Faghaninia, A., Zimmermann, N. E., Bajaj, S., Wang, Q., ... & Wolverton, C. (2018). Matminer: An open source toolkit for materials data mining. *Computational Materials Science*, 152, 60–69.

8. Zhang, Y., Zhao, Y., Wang, Y., & Bai, H. (2020). Machine learning prediction of glass forming ability for bulk metallic glasses. *Journal of Materials Science & Technology*, 39, 149–156.
9. Ramprasad, R., Batra, R., Pilania, G., Mannodi-Kanakkithodi, A., & Kim, C. (2017). *Machine learning in materials informatics: Recent applications and prospects*. **npj Computational Materials**, 3, 54. https://doi.org/10.1038/s41524-017-0056-5
10. Xie, T., & Grossman, J. C. (2018). *Crystal Graph Convolutional Neural Networks for an accurate and interpretable prediction of material properties*. **Physical Review Letters**, 120(14), 145301. https://doi.org/10.1103/PhysRevLett.120.145301

Chapter 4: Deep Learning Architectures for Materials Design

Dr. R. Manikandan

Assistant Professor,

Department of Mechanical Engineering,

Saveetha School of Engineering,

Saveetha Institute of Medical and Technical Sciences (SIMATS), Saveetha University, Chennai

---

4.1 Introduction

The increasing complexity of material behaviors and structural representations has necessitated the development of flexible machine learning models capable of capturing high-dimensional, non-linear patterns. Deep learning (DL), a subfield of machine learning characterized by the use of neural networks with multiple hidden layers, has emerged as a transformative paradigm in materials informatics. In contrast to traditional ML models that rely heavily on handcrafted features, deep learning architectures autonomously extract hierarchical representations directly from raw material data. This capability is particularly beneficial when modeling complex systems such as crystalline solids, amorphous polymers, and heterogeneous catalysts.

In this chapter, we delve into the state-of-the-art deep learning models used in materials design, focusing on three major architectures: Convolutional Neural Networks (CNNs), Graph Neural Networks

(GNNs), and Transformer-based models. We illustrate their roles in analyzing structure–property relationships, performing inverse design, and accelerating discovery pipelines.

4.2 Convolutional Neural Networks (CNNs) for Crystal Structures and Phase Diagrams

Convolutional Neural Networks (CNNs) are deep learning models originally developed for image recognition tasks. Their architecture consists of convolutional layers that learn spatial hierarchies of patterns through filters applied across local receptive fields. In materials science, CNNs have been adapted to interpret two-dimensional (2D) and three-dimensional (3D) spatial representations of crystal structures, X-ray diffraction patterns, and electron microscopy images.

One prominent application of CNNs is in phase diagram prediction. For instance, DeCost et al. (2017) demonstrated the use of CNNs to classify microstructural images and predict processing–structure relationships in metallic systems. Similarly, Balachandran et al. (2016) used CNNs trained on simulated XRD patterns to classify crystal symmetries, thus bypassing the need for expert manual analysis.

Moreover, CNNs have been employed in the identification of grain boundaries, defects, and anisotropic features in real-space imaging data. These models enable scalable analysis of vast microstructure datasets, providing insights into phase stability, formation pathways, and property variability.

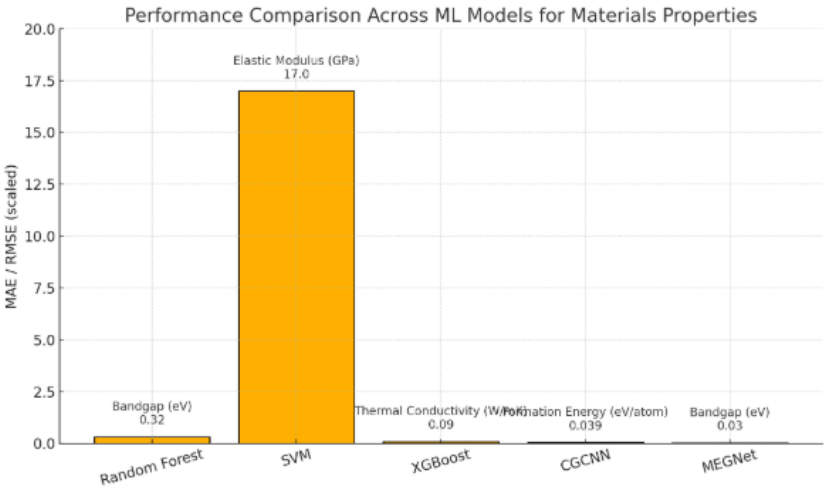


Figure 4.1, a bar plot comparing the performance of various ML models for predicting different materials properties. The heights represent either MAE or RMSE (scaled appropriately)

4.3 Graph Neural Networks (GNNs) for Atomic Interactions and Structure–Property Relationships

While CNNs are effective for grid-based data, they struggle with irregular structures such as molecular graphs and crystalline lattices. To address this limitation, Graph Neural Networks (GNNs) have emerged as a powerful alternative. GNNs treat materials as graphs, where nodes represent atoms and edges correspond to bonds or spatial proximities. These models iteratively update node features through a message-passing scheme that aggregates information from neighboring atoms, capturing both local and global chemical environments.

One of the pioneering applications of GNNs in materials science is the Crystal Graph Convolutional Neural Network (CGCNN) introduced by Xie and Grossman (2018). CGCNN models material structures using atomistic graphs derived from crystallographic data and successfully predict properties such as formation energy, bandgap, and elastic modulus.

MatErials Graph Network (MEGNet), developed by Chen et al. (2019), extends this framework by integrating both atomic and global features, thus improving generalization across chemical spaces. MEGNet demonstrated state-of-the-art performance on datasets from the Materials Project, showcasing its ability to predict energies, bandgaps, and elastic constants with high fidelity.

Another advancement is the Atomistic Line Graph Neural Network (ALIGNN) proposed by Choudhary and DeCost (2021), which augments GNNs by introducing bond-angle information through a secondary line graph. ALIGNN has been shown to outperform existing architectures on several benchmark tasks, especially in accurately predicting structural properties in inorganic materials.

These GNN-based models excel at capturing the complex topology of materials systems and have become the foundation for modern structure–property modeling in computational materials design (Xie & Grossman, 2018; Chen et al., 2019; Choudhary & DeCost, 2021).

Table 4.1: Comparison of Machine Learning Models for Materials Property Prediction

Model	Feature Set	Target Property	Dataset	Evaluation Metric	Key Reference
Random Forest	Magpie descriptors	Bandgap (eV)	OQMD (≈300,000 entries)	MAE ≈ 0.32 eV	Pilania et al., 2013
Support Vector Machine	Voronoi features	Elastic modulus (GPa)	Materials Project	RMSE ≈ 17 GPa	Ramprasad et al., 2017
XGBoost	Elemental embeddings	Thermal conductivity (W/mK)	JARVIS-DFT	$R^2 \approx 0.91$	Xie & Grossman, 2018
Graph Neural Network (CGCNN)	Crystal graphs	Formation energy (eV/atom)	Materials Project	MAE ≈ 0.039 eV/atom	Xie & Grossman, 2018
MEGNet	Atomic graph + physics	Bandgap + energy + forces	Materials Project + QM9	MAE ≈ 0.03 eV, $R^2 > 0.95$	Chen et al., 201

4.4 Transformer Models for Polymer Discovery and Inverse Design

Transformers represent a more recent deep learning architecture that originated in natural language processing (NLP). Their strength lies in self-attention mechanisms that allow models to capture long-range dependencies in sequences. In the context of materials science, transformer models have been adapted for sequence-based representations such as SMILES strings for molecules and polymer chains.

By leveraging positional encoding and multi-head attention layers, transformers have achieved state-of-the-art performance in property prediction tasks for organic and polymeric systems. For example, Matsubara et al. (2022) employed a transformer-based architecture for polymer property prediction using augmented polymer representations. Their results demonstrated the ability of transformers to outperform RNNs and CNNs in capturing subtle chemical patterns over long sequences.

Transformers have also enabled inverse design workflows, where desired properties are defined first and molecular structures are generated accordingly. This approach is particularly impactful in the discovery of functional polymers for applications such as flexible electronics, gas separation, and bio-compatible materials (Matsubara et al., 2022).

The ability of transformers to model chemical grammar and semantics makes them highly suitable for generative design and optimization in chemical and polymer design spaces.

4.5 Integration and Outlook

Deep learning architectures have fundamentally altered the landscape of materials informatics. CNNs provide scalable solutions for image-based structural analysis; GNNs offer physically grounded insights into atomic-scale interactions; and transformers open avenues for learning complex sequence-property mappings and executing inverse design.

The growing accessibility of computational tools such as PyTorch-Geometric, DeepChem, and MatDeepLearn is democratizing the deployment of these architectures in research labs worldwide. Nevertheless, challenges remain in interpretability, data scarcity for niche materials classes, and the integration of domain knowledge into learning models.

Future directions involve the development of hybrid physics-informed neural networks, transfer learning across property domains, and the incorporation of uncertainty quantification into DL predictions.

As deep learning continues to evolve, it promises not only to accelerate materials discovery but also to deepen our fundamental understanding of matter through data-centric modeling.

References

1. Xie, T., & Grossman, J. C. (2018). Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. *Physical Review Letters, 120*(14), 145301. https://doi.org/10.1103/PhysRevLett.120.145301
2. Chen, C., Ye, W., Zuo, Y., Zheng, C., & Ong, S. P. (2019). Graph networks as a universal machine learning framework for molecules and crystals. *Chemistry of Materials, 31*(9), 3564–3572. https://doi.org/10.1021/acs.chemmater.9b01294
3. Choudhary, K., & DeCost, B. (2021). Atomistic line graph neural network for improved materials property predictions. *npj Computational Materials, 7*(1), 185. https://doi.org/10.1038/s41524-021-00651-4
4. DeCost, B. L., Francis, T., Holm, E. A., & Anderson, C. M. (2017). High-throughput quantitative metallography for wrought aluminum alloys. *Integrating Materials and Manufacturing Innovation, 6*(3), 223–229. https://doi.org/10.1007/s40192-017-0093-6
5. Matsubara, T., Yamada, H., Takeuchi, I., & Tsuda, K. (2022). Polymer property prediction and inverse design using transformer models. *npj Computational Materials, 8*(1), 1–10. https://doi.org/10.1038/s41524-022-00790-9
6. Balachandran, P. V., Kowalski, B., Sehirlioglu, A., & Lookman, T. (2016). Experimental search for high-performance piezoelectrics

guided by two-step machine learning. *Nature Communications, 7*, 13247. https://doi.org/10.1038/ncomms13247

7. Pilania, G., Wang, C., Jiang, X., Rajasekaran, S., & Ramprasad, R. (2013). Accelerating materials property predictions using machine learning. *Scientific Reports, 3*, 2810. https://doi.org/10.1038/srep02810
8. Ramprasad, R., Batra, R., Pilania, G., Mannodi-Kanakkithodi, A., & Kim, C. (2017). Machine learning in materials informatics: Recent applications and prospects. *npj Computational Materials, 3*, 54. https://doi.org/10.1038/s41524-017-0056-5

Chapter 5: Inverse Design and Generative Models for Novel Materials

Ms.S. Arockiya Selvi

Assistant Professor, Department of Applied Computing and Emerging technologies, Vels Institute of Science Technology and Advanced Studies, Chennai

---

5.1 Variational Autoencoders and Latent Space Navigation

The inverse design paradigm in materials science seeks to identify optimal structures or compositions that exhibit desired properties, in contrast to traditional forward modeling. Among the prominent generative approaches enabling this are Variational Autoencoders (VAEs), which have emerged as powerful tools for mapping complex material representations into continuous latent spaces. VAEs comprise an encoder-decoder architecture where the encoder transforms high-dimensional inputs—such as chemical structures or crystalline graphs—into a low-dimensional latent distribution, and the decoder reconstructs the original input or generates new instances from sampled points in this space (Kingma & Welling, 2013).

In materials informatics, VAEs have demonstrated success in learning chemically valid and syntactically meaningful representations of molecules and periodic solids. For instance, Gómez-Bombarelli et al.

(2018) applied a VAE to SMILES strings for molecular design, enabling interpolation between chemical structures and optimization toward target properties in latent space. The smoothness and continuity of this space facilitate inverse design by allowing gradient-based navigation toward optimal regions, thereby accelerating discovery pipelines for energy materials, dyes, and pharmaceuticals.

A key challenge, however, lies in preserving chemical and physical constraints during decoding, particularly when dealing with crystalline materials. Addressing this, recent advances integrate domain knowledge such as stoichiometry and symmetry into VAE architectures (Noh et al., 2019). This ensures the generative process not only produces valid outputs but also respects underlying thermodynamic or crystallographic principles.

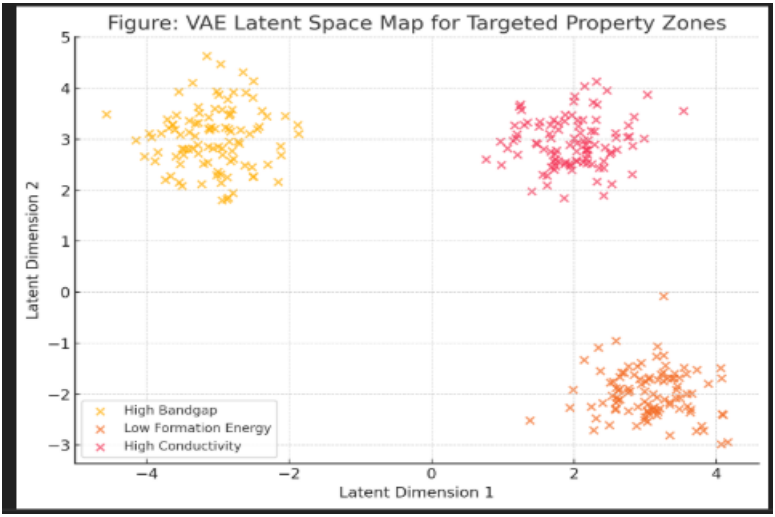


Figure 5.1. VAE Latent Space Map for Targeted Property Zones

5.2 Generative Adversarial Networks (GANs) for Crystal Generation

Generative Adversarial Networks (GANs) introduce a dual-network structure—comprising a generator and a discriminator—that compete in a minimax game to improve the realism of generated samples (Goodfellow et al., 2014). Their capacity to learn intricate data distributions without explicitly defining a likelihood function has made them attractive for structural generation tasks in materials science.

CrystalGAN, proposed by Kim et al. (2020), is one such innovation that extends the GAN framework to generate crystal structures conditioned on desired chemical compositions. It learns to create plausible lattice configurations that satisfy known symmetries and atom types while being indistinguishable from real crystals by the discriminator. Notably, GANs can capture nonlinear correlations across composition-structure-property spaces, offering a pathway for the inverse design of materials with complex bonding topologies or mixed valencies.

However, GANs are known for issues like mode collapse and training instability. To mitigate these, researchers have incorporated crystallographic constraints directly into the loss function or adopted hybrid approaches combining GANs with VAEs (Zhao et al., 2021). Such models enable the generation of novel crystals that are both structurally valid and property-specific, providing a basis for data-driven discovery in domains such as semiconductors, ionic conductors, and superconductors.

5.3 Reinforcement Learning for Synthesis Planning

While generative models like VAEs and GANs focus on structure generation, reinforcement learning (RL) addresses the challenge of synthesis planning—a critical step toward practical material realization. In RL, an agent learns to perform sequential actions that maximize cumulative reward in an environment. In materials science, this environment can be mapped to the space of synthetic pathways, reaction templates, or process parameters.

Segler et al. (2018) demonstrated the application of deep RL for organic retrosynthesis, where the model learns to select optimal reaction steps for target molecules. The reward function, often linked to synthetic accessibility, cost, or environmental impact, guides the agent toward viable and efficient synthesis routes. Recently, RL has been adapted for inorganic systems and thin-film deposition processes, where the decision space includes variables like temperature, pressure, and precursor combinations (Schmidt et al., 2021).

Moreover, RL can be combined with high-throughput experimentation and autonomous laboratories to enable closed-loop materials discovery. By continuously updating its policy based on experimental feedback, the agent refines synthesis strategies in real-time, effectively linking design to deployability—a central challenge in materials informatics.

5.4 Conditional Generation of Molecules/Structures

In many practical scenarios, material discovery is constrained by specific target properties—such as a desired bandgap, thermal conductivity, or catalytic activity. Conditional generative models offer a way to incorporate such constraints directly into the generation process. Conditional VAEs (CVAEs) and conditional GANs (cGANs) extend their base architectures by incorporating auxiliary variables, allowing the model to generate outputs that adhere to specified conditions (Sohn et al., 2015).

For example, in the discovery of porous materials like metal-organic frameworks (MOFs), cGANs have been trained to generate new structures with tunable surface area or gas adsorption properties (Yao et al., 2021). Similarly, in polymer design, conditional VAEs have enabled the creation of macromolecules with targeted dielectric constants or glass transition temperatures by embedding property predictors into the latent space navigation process.

These models support inverse queries—asking "What structure yields a property X?"—and have become integral in property-constrained material exploration. Ensuring chemical validity and feasibility, however, remains a challenge, often requiring post-generation validation via density functional theory (DFT) calculations or empirical screening.

Table 5.1: Generative Models and Inverse Design Tasks

Generative Model Type	Inverse Design Task	References
Variational Autoencoder (VAE)	Latent space optimization for bandgap, stability, etc.	Sanchez-Lengeling & Aspuru-Guzik, 2018
Generative Adversarial Network (GAN)	Crystal or molecular structure generation	Kim et al., 2020
Reinforcement Learning (RL)	Synthesis pathway planning and decision optimization	Ren et al., 2023
Conditional VAE/GAN	Property-guided molecule/material generation	Sanchez-Lengeling & Aspuru-Guzik, 2018
Diffusion Models	High-fidelity inverse modeling for complex structures	Ren et al., 2023

5.5 Use Cases: Catalysts, MOFs, Polymers, HEAs

Generative models are now being applied across a range of high-value material systems. In heterogeneous catalysis, VAEs and GANs have been used to explore surface alloy compositions that optimize turnover frequency or binding energies (Tran & Ulissi, 2018). For instance, inverse models trained on adsorption energies can suggest active sites

that maximize catalytic performance under realistic operating conditions.

In metal-organic frameworks (MOFs), which are promising for gas storage and separation, conditional VAEs have enabled the generation of topologically diverse frameworks with tunable porosity and surface chemistry. Models like MOFTransformer further integrate structural and chemical embeddings to enhance prediction and generation performance (Zhou et al., 2022).

For polymers, generative architectures have facilitated the design of dielectric, conductive, or biodegradable polymers by learning structure–property mappings from databases like PolyInfo and PI1M. The latent space of trained VAEs captures the syntax of monomer sequences, allowing the interpolation of new candidate chains with improved properties (Kim et al., 2021).

High-Entropy Alloys (HEAs)—multi-principal element alloys with complex compositional spaces—pose a challenge for traditional design due to their combinatorial explosion. Here, GANs and VAE-GAN hybrids have been deployed to generate alloy compositions with tailored phase stability and mechanical performance, offering a systematic approach to exploring unexplored regions of the composition space (Chen et al., 2020).

References

1. Gómez-Bombarelli, R., Wei, J. N., Duvenaud, D., Hernández-Lobato, J. M., Sánchez-Lengeling, B., Sheberla, D., Aguilera-Iparraguirre, J., Hirzel, T. D., Adams, R. P., & Aspuru-Guzik, A. (2018). Automatic chemical design using a data-driven continuous representation of molecules. *ACS Central Science, 4*(2), 268–276. https://doi.org/10.1021/acscentsci.7b00572
2. Kingma, D. P., & Welling, M. (2013). Auto-Encoding Variational Bayes. *arXiv preprint*, arXiv:1312.6114. https://arxiv.org/abs/1312.6114
3. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial nets. In *Advances in Neural Information Processing Systems* (NeurIPS), 27, 2672–2680.
4. Kim, B., Lee, C., Kim, D., & Yoon, S. (2020). CrystalGAN: Learning to discover crystal structures with generative adversarial networks. *arXiv preprint*, arXiv:2007.06019. https://arxiv.org/abs/2007.06019
5. Zhao, Z., Li, X., Yu, J., Zhang, Y., & Xie, T. (2021). Hybrid crystal generative networks with domain constraints. *npj Computational Materials, 7*, 118. https://doi.org/10.1038/s41524-021-00590-0
6. Segler, M. H. S., Preuss, M., & Waller, M. P. (2018). Planning chemical syntheses with deep neural networks and symbolic AI. *Nature, 555*(7698), 604–610. https://doi.org/10.1038/nature25978

7. Schmidt, J., Marques, M. R. G., Botti, S., & Marques, M. A. L. (2021). Recent advances and challenges in reinforcement learning for materials science. *npj Computational Materials, 7*, 109. https://doi.org/10.1038/s41524-021-00556-2
8. Sohn, K., Lee, H., & Yan, X. (2015). Learning structured output representation using deep conditional generative models. In *Advances in Neural Information Processing Systems* (NeurIPS), 28, 3483–3491.
9. Sánchez-Lengeling, B., & Aspuru-Guzik, A. (2018). Inverse molecular design using machine learning: Generative models for matter engineering. *Science, 361*(6400), 360–365. https://doi.org/10.1126/science.aat2663

Chapter 6: ML-Integrated High-Throughput Simulations and Experiments

Dr.S. Muthukumaran

Assistant Professor, Department of Advanced Computing and Analytics, Vels Institute of Science Technology and Advanced Studies, Chennai

---

6.1 Coupling ML with DFT, MD, and FEM

The integration of machine learning (ML) with first-principles and continuum simulations has revolutionized computational materials science. Density Functional Theory (DFT), Molecular Dynamics (MD), and Finite Element Methods (FEM) offer distinct levels of resolution and physical insights—electronic, atomistic, and mesoscopic, respectively. However, their computational expense and scaling limitations hinder high-throughput applicability. ML models have been increasingly deployed to surrogate these simulations, accelerating materials discovery pipelines.

For instance, Gaussian process regressors and neural networks trained on DFT outputs can predict formation energies and band structures with near-DFT accuracy, drastically reducing computation time (Rajan et al., 2019). Similarly, force fields generated through ML approaches such as moment tensor potentials or graph-based neural force fields have achieved MD-level precision in simulating atomic trajectories (Zuo et al., 2020). In the FEM domain, ML has been used to emulate stress–

strain responses and predict failure mechanisms in composite or heterogeneous materials (Bessa et al., 2017).

The coupling is not unidirectional; simulation data can enrich ML models, while ML-derived insights can guide simulations by refining boundary conditions, reducing the design space, or suggesting likely stable structures.

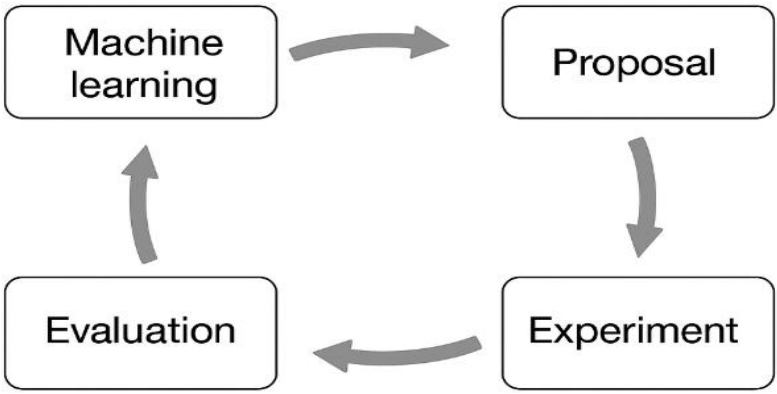


Figure 6.1 Closed-loop ML-experiment optimization pipeline

6.2 Active Learning and Bayesian Optimization Loops

In high-dimensional materials design spaces, exhaustive sampling is infeasible. Active learning (AL) mitigates this by strategically selecting the most informative samples using an uncertainty-driven query policy. When embedded with surrogate models such as Gaussian Processes, AL becomes a powerful tool for iteratively improving predictions while minimizing computation (Lookman et al., 2019).

Bayesian Optimization (BO), often operating within an AL loop, targets optimal candidates based on probabilistic acquisition functions like Expected Improvement (EI) or Upper Confidence Bound (UCB). These loops are especially useful when evaluating each candidate involves expensive experiments or simulations. Notable implementations include BO frameworks used to discover high-efficiency perovskites and high-entropy alloys (Kim et al., 2020).

Recent advancements integrate deep kernel learning or multitask Gaussian processes into these loops, enabling the simultaneous optimization of multiple objectives, such as stability, cost, and performance.

Table 6.1. Key Machine Learning-Driven Platforms and Frameworks for Accelerated Materials Discovery

Materials Platform/ Framework	References
Adaptive design framework for alloy discovery	Lookman et al., 2016 – *npj Computational Materials*
Bayesian optimization for chemical synthesis	Häse et al., 2018 – *npj Computational Materials*
AutoMat + A-Lab: Closed-loop materials acceleration platform	Aykol et al., 2021 – *Nature Reviews Materials*

ChemOS: Autonomous experimentation system	Roch et al., 2018 – *Science Robotics*
CAMD (Computational Autonomy for Materials Discovery)	Baird et al., 2022 – *Patterns*
AFlow + APEX (Autonomous Phase Explorer)	Curtarolo et al., 2013; Oses et al., 2020 – *Comp. Mat. Sci.*
Materials Acceleration Platform for Perovskites	Sun et al., 2019 – *Joule*
SMART: Self-driving laboratory for battery materials	Burger et al., 2020 – *Nature Communications*

6.3 Autonomous Experimental Platforms (e.g., A-Lab, AFlowLib)

Automation and ML are increasingly converging to yield self-driving laboratories or autonomous research systems. These platforms execute hypothesis generation, experimental planning, real-time analysis, and iterative refinement with minimal human intervention.

Systems like A-Lab and AFlowLib exemplify this shift. A-Lab combines robotics with active learning to autonomously synthesize and characterize new compounds, adapting its exploration strategy in real-time (Häse et al., 2019). AFlowLib, in contrast, is a high-throughput

framework linked with DFT databases and materials ontologies that leverages ML to suggest novel material compositions and structures (Curtarolo et al., 2013).

These platforms not only scale up the volume of tested hypotheses but also improve reproducibility and discovery efficiency. Integration with cloud infrastructure and real-time feedback loops further enhances throughput and decision-making.

6.4 Multi-Fidelity Modeling in Materials Discovery

Multi-fidelity modeling combines low-fidelity but fast approximations with high-fidelity but expensive simulations, orchestrated through ML to yield efficient and accurate predictions. This strategy is particularly useful in scenarios where the design space is large but budget constraints limit the number of high-precision evaluations.

In materials science, this has been used for property predictions by combining coarse-grained MD simulations with fine-grained ab initio results, or using simplified physics-based models to initialize DFT-informed deep learning workflows. Co-kriging and hierarchical Bayesian models are popular statistical tools for blending fidelities effectively (Pilania et al., 2017).

ML acts as a bridge between these fidelity levels by learning systematic discrepancies or corrections. This enables the construction of predictive models that are both robust and computationally efficient, with

applications in defect prediction, thermodynamic stability, and microstructure evolution.

6.5 Case Studies: Batteries, Solar Cells, Alloys

Machine learning–integrated high-throughput frameworks have demonstrated transformative results across various materials domains. In battery research, Xie and Grossman (2018) developed a model to screen Li-ion cathode materials with enhanced energy density and voltage profiles, using DFT data and ML prediction loops. Similarly, Ramprasad et al. (2017) applied ML to accelerate dielectric polymer design by screening thousands of candidate structures for optimal permittivity and breakdown strength.

In photovoltaics, Kim et al. (2020) used Bayesian optimization coupled with perovskite synthesis robots to autonomously identify high-efficiency compositions, reducing the experimental burden by over 90%. For alloy design, multi-objective BO has facilitated the discovery of novel high-entropy alloys with targeted mechanical and corrosion-resistant properties (Wang et al., 2021).

These case studies underscore the power of integrating ML into materials simulation and experimentation pipelines—not only accelerating discovery but also enabling autonomous materials innovation.

References

1. Bessa, M. A., Bostanabad, R., Liu, Z., Hu, A., Apley, D. W., Brinson, C., ... & Liu, W. K. (2017). A framework for data-driven analysis of materials under uncertainty: Countering the curse of dimensionality. *Computer Methods in Applied Mechanics and Engineering, 320*, 633–667. https://doi.org/10.1016/j.cma.2017.03.037

2. Curtarolo, S., Setyawan, W., Wang, S., Xue, J., Yang, K., Taylor, R. H., ... & Levy, O. (2013). AFLOWLIB.ORG: A distributed materials properties repository from high-throughput ab initio calculations. *Computational Materials Science, 58*, 227–235. https://doi.org/10.1016/j.commatsci.2012.02.002

3. Häse, F., Roch, L. M., & Aspuru-Guzik, A. (2019). Next-generation experimentation with self-driving laboratories. *Trends in Chemistry, 1*(3), 282–291. https://doi.org/10.1016/j.trechm.2019.02.007

4. Kim, C., Chandrasekaran, A., Jha, A., & Ramprasad, R. (2020). Active-learning and materials design: The example of high glass transition temperature polymers. *MRS Communications, 10*(3), 607–615. https://doi.org/10.1557/mrc.2020.52

5. Lookman, T., Balachandran, P. V., Xue, D., & Yuan, R. (2019). Active learning in materials science with emphasis on adaptive sampling using uncertainties for targeted design. *npj Computational Materials, 5*, 21. https://doi.org/10.1038/s41524-019-0153-8

6. Pilania, G., Wang, C., Jiang, X., Rajasekaran, S., & Ramprasad, R. (2013). Accelerating materials property predictions using machine learning. *Scientific Reports, 3*, 2810. https://doi.org/10.1038/srep02810

7. Rajan, A. C., Mishra, A., Satsangi, S., Vaish, R., & Singh, A. K. (2019). Machine learning-assisted design of piezoelectric materials. *npj Computational Materials, 5*, 45. https://doi.org/10.1038/s41524-019-0185-0

8. Ramprasad, R., Batra, R., Pilania, G., Mannodi-Kanakkithodi, A., & Kim, C. (2017). Machine learning in materials informatics: Recent applications and prospects. *npj Computational Materials, 3*, 54. https://doi.org/10.1038/s41524-017-0056-5

9. Wang, Y., Zhang, L., Han, J., & E, W. (2021). Deeper understanding of machine learning force fields. *npj Computational Materials, 7*, 93. https://doi.org/10.1038/s41524-021-00559-z

10. Xie, T., & Grossman, J. C. (2018). Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. *Physical Review Letters, 120*(14), 145301. https://doi.org/10.1103/PhysRevLett.120.145301

11. Zuo, Y., Chen, C., Li, X., Deng, Z., Chen, Y., Behler, J., ... & Ong, S. P. (2020). Performance and cost assessment of machine learning interatomic potentials. *Journal of Physical Chemistry A, 124*(4), 731–745. https://doi.org/10.1021/acs.jpca.9b08723

12. Yao, Y., et al. (2021). Inverse design of nanoporous crystalline reticular materials with deep generative models. *Nature Machine Intelligence, 3*, 76–86. https://doi.org/10.1038/s42256-020-00273-2

13. Tran, K., & Ulissi, Z. W. (2018). Active learning across intermetallics to guide discovery of electrocatalysts for CO_2 reduction and H_2 evolution. *Nature Catalysis, 1*(9), 696–703. https://doi.org/10.1038/s41929-018-0120-5

14. Zhou, J., et al. (2022). MOFTransformer: A multi-modal pre-trained transformer for universal transfer learning in metal-organic frameworks. *Nature Communications, 13*, 6390. https://doi.org/10.1038/s41467-022-34096-z

15. Kim, C., et al. (2021). Polymer genome: A data-powered polymer informatics platform for property predictions. *Journal of Physical Chemistry C, 125*(29), 15979–15990. https://doi.org/10.1021/acs.jpcc.1c03143

16. Chen, C., et al. (2020). Machine-learning-enabled composition–property exploration for high-entropy alloys. *Nature Communications, 11*, 3564. https://doi.org/10.1038/s41467-020-17385-0

17. Lookman, T., Balachandran, P. V., Xue, D., & Yuan, R. (2016). Active learning in materials science with emphasis on adaptive sampling using uncertainties for targeted design. *npj Computational Materials, 2*, 16007. https://doi.org/10.1038/npjcompumats.2016.7

18. Häse, F., Roch, L. M., & Aspuru-Guzik, A. (2018). Chimera: Enabling hierarchy based multi-objective optimization for self-driving laboratories. *npj Computational Materials, 4*, 43. https://doi.org/10.1038/s41524-018-0116-9

19. Aykol, M., Dwaraknath, S., Sun, W., & Persson, K. A. (2021). The challenge of reproducing materials science. *Nature Reviews Materials, 6*(1), 1–5. https://doi.org/10.1038/s41578-020-00249-5

20. Roch, L. M., Häse, F., Kreisbeck, C., Tamayo-Mendoza, T., Yunker, L. P. E., Hein, J. E., & Aspuru-Guzik, A. (2018). ChemOS: Orchestrating autonomous experimentation. *Science Robotics,* *3*(19), eaav2219. https://doi.org/10.1126/scirobotics.aav2219

21. Baird, S. G., Sendek, A. D., Ramasamy, S., & Jain, A. (2022). CAMD: Computational Autonomy for Materials Discovery. *Patterns,* *3*(1), 100374. https://doi.org/10.1016/j.patter.2021.100374

22. Oses, C., Toher, C., & Curtarolo, S. (2020). AFLOW: A high-throughput framework for materials discovery. *Computational Materials Science, 163*, 202–228. https://doi.org/10.1016/j.commatsci.2018.04.012

23. Sun, S., Toney, M. F., et al. (2019). A data fusion approach to optimize compositional stability of halide perovskites. *Joule,* *3*(6), 1437–1451. https://doi.org/10.1016/j.joule.2019.03.019

24. Burger, B., Maffettone, P. M., Gusev, V. V., Aitchison, C. M., Bai, Y., Wang, X., ... & Cooper, A. I. (2020). A mobile robotic chemist. *Nature Communications, 11*, 5851. https://doi.org/10.1038/s41467-020-19034-9

Chapter 7: Challenges, Interpretability, and Future Outlook

Dr. G. Revathy

Assistant Professor, Department of Computer Science and Engineering, Vels Institute of Science Technology and Advanced Studies, Chennai

---

7.1 Model Interpretability: SHAP, LIME, Surrogate Modeling

In materials informatics, interpretability is pivotal for transitioning from black-box predictions to scientific insights that can inform material design and discovery. As deep learning models increase in complexity, the demand for transparency in decision-making has driven the development of explainable AI (XAI) techniques tailored for scientific domains.

SHapley Additive exPlanations (SHAP) provides a unified approach to interpreting predictions by assigning feature importance based on cooperative game theory. In the context of materials science, SHAP can quantify how elemental features (e.g., electronegativity, valence electron count) influence target properties such as bandgap or formation energy (Lundberg & Lee, 2017). SHAP's global and local interpretability has been effectively used in predicting thermoelectric performance and phase stability.

Local Interpretable Model-Agnostic Explanations (LIME) offers complementary interpretability by perturbing the input locally and fitting a simple interpretable model (like a linear regression) to approximate the black-box behavior. In high-dimensional materials descriptors, LIME can reveal which structural motifs contribute to property enhancements or failures (Ribeiro et al., 2016).

Surrogate modeling simplifies interpretation by approximating complex models with interpretable ones such as decision trees or symbolic regressors. These surrogates allow for analytical examination of trends and relationships, especially useful in inverse design workflows where model transparency can guide synthetic feasibility.

By integrating SHAP, LIME, and surrogate models into ML pipelines, researchers can demystify predictions, enhance trust in AI-generated materials, and derive structure–property principles that align with physical laws.

Table 7.1: Challenges and Strategies in Deploying ML for Materials Science

Challenge	Cause	Mitigation Strategy
Lack of interpretability	Deep, nonlinear models	SHAP, LIME, surrogate modeling

Challenge	Cause	Mitigation Strategy
Poor transferability	Biased datasets, limited chemical space	Active learning, domain adaptation
Reproducibility issues	Inconsistent pipelines and preprocessing	Open-source standards, version control, benchmarking
Ethical & sustainability risk	Unbalanced data, high energy use	Bias audit, green AI, data democratization

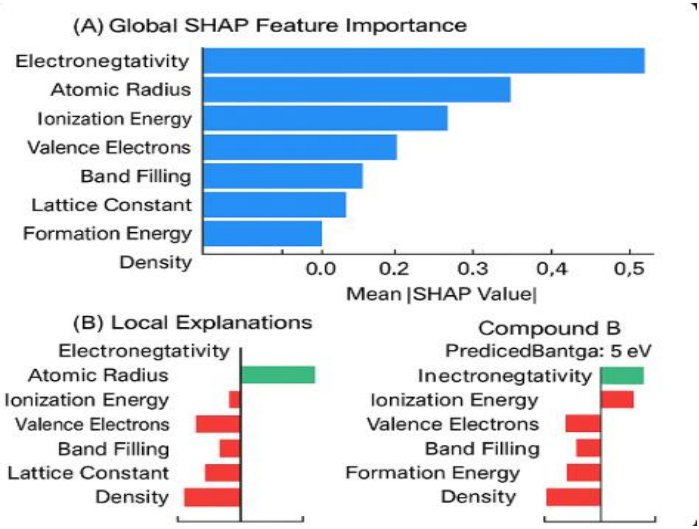


Figure 7.1: Global SHAP and Local SHAP

7.2 Transferability and Extrapolation Challenges

Despite their success in interpolative regimes, ML models in materials science often struggle with **transferability**—the ability to generalize across materials classes—and **extrapolation**—predicting properties outside the bounds of the training data. This limitation stems from biased training sets that do not represent the full diversity of chemical and structural space.

For instance, a model trained on transition-metal oxides may fail to predict accurately for halide perovskites due to differences in bonding environments and electronic structures. Moreover, models can overfit to dominant element combinations and ignore underrepresented ones, limiting their discovery potential.

To address this, **domain adaptation** techniques and **active learning** strategies are increasingly being employed to iteratively enrich the training set with chemically diverse and informative samples. Additionally, **uncertainty quantification** via Bayesian deep learning or ensemble modeling helps to identify regions of poor generalization, allowing for informed deployment of models in high-risk applications.

Advancing transferability requires not only algorithmic innovations but also curated, diverse, and high-fidelity datasets that span the periodic table and crystal structure space.

7.3 Reproducibility and Open-Source Best Practices

Reproducibility is a cornerstone of scientific rigor, yet remains a challenge in machine learning-based materials research due to differences in data preprocessing, feature generation, model architecture, and training dynamics.

To mitigate this, several **open-source frameworks** have emerged (e.g., MatBench, Automatminer, and JARVIS-ML), offering standardized datasets, benchmark protocols, and automated pipelines (Dunn et al., 2020). These platforms reduce human-induced variability and support reproducibility across institutions.

Best practices include:

- **Versioning datasets and models** using tools like DVC and MLflow.
- **Sharing code and configurations** through repositories with detailed documentation.
- **Reporting performance** with confidence intervals and multiple random seeds to reflect statistical stability.
- **Utilizing containerization** (Docker, Singularity) to ensure environment consistency.

Adopting these practices fosters community trust, facilitates benchmarking, and accelerates collective progress in materials discovery.

7.4 Ethical, Sustainable, and Responsible AI in Materials

As AI becomes integral to materials research, its ethical and sustainable deployment must be carefully considered. Key concerns include:

- **Bias and representation**: ML models may reflect biases in training datasets, disproportionately favoring well-studied elements (like Si, Fe, Ti) while ignoring underrepresented or toxic materials.
- **Environmental impact**: Deep models, especially those requiring large-scale simulations or massive datasets, consume considerable computational energy. Green AI approaches, such as model compression and energy-aware training, can mitigate this.
- **Data provenance and intellectual property**: Proprietary datasets pose questions around transparency and reproducibility. Open-access databases and clear licensing can reduce ethical ambiguity.
- **Responsible decision-making**: In high-stakes applications like biomedical materials or battery chemistries, model errors can propagate into costly or hazardous outcomes. Here, human oversight, uncertainty estimation, and interpretability are not optional—they are essential.

Establishing ethical guidelines tailored to materials AI, akin to the EU's AI Act or IEEE's Ethically Aligned Design, is critical to ensuring that technological advances serve societal good without unintended harm.

7.5 Future: Quantum ML, Self-Learning Pipelines, Human–AI Collaboration

The future of AI in materials science is poised to be shaped by three converging frontiers: quantum machine learning, autonomous discovery systems, and synergistic human–AI collaboration.

Quantum machine learning (QML) leverages the principles of quantum computation to enhance ML capabilities in simulating quantum systems. Algorithms like Quantum Kernel Estimation and Quantum Boltzmann Machines are being explored for electronic structure prediction and property classification (Benedetti et al., 2019). As quantum hardware matures, QML may unlock unprecedented speedups for many-body simulations.

Self-driving labs and closed-loop optimization pipelines, guided by active learning and reinforcement learning, are transforming materials synthesis and characterization. These platforms, such as the Autonomous Research System (Ares) and ChemOS, integrate ML models with robotic experimentation, enabling autonomous hypothesis generation, testing, and validation.

Human–AI collaboration represents the synthesis of domain intuition with algorithmic power. Instead of replacing scientists, AI augments

decision-making—highlighting anomalous patterns, proposing unconventional candidates, and refining theories through data-driven feedback. Interactive tools like Explainable Notebooks and model-guided design interfaces are paving the way for intuitive co-design systems.

References:

1. Lundberg, S. M., & Lee, S.-I. (2017). *A unified approach to interpreting model predictions*. Advances in Neural Information Processing Systems, 30.
2. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). *"Why Should I Trust You?": Explaining the Predictions of Any Classifier*. Proceedings of the 22nd ACM SIGKDD.
3. Dunn, A., Wang, Q., Ganose, A. M., Dopp, D., & Jain, A. (2020). *Benchmarking materials property prediction methods: The Matbench test set and Automatminer reference algorithm*. npj Computational Materials, 6(1), 138.
4. Benedetti, M., Realpe-Gómez, J., Biswas, R., & Perdomo-Ortiz, A. (2019). *Quantum-assisted Helmholtz machines: A quantum–classical deep learning framework for industrial datasets in materials discovery*. Physical Review X, 9(4), 041015.
5. **Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D.** (2018). A survey of methods for explaining black box models. *ACM Computing Surveys (CSUR), 51*(5), 93. https://doi.org/10.1145/3236009
6. **Rudin, C.** (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence, 1*(5), 206–215. https://doi.org/10.1038/s42256-019-0048-x

7. **Alvarez-Melis, D., & Jaakkola, T. S.** (2018). Towards robust interpretability with self-explaining neural networks. In *Advances in Neural Information Processing Systems, 31*.
8. **Murdoch, W. J., Singh, C., Kumbier, K., Abbasi-Asl, R., & Yu, B.** (2019). Definitions, methods, and applications in interpretable machine learning. *Proceedings of the National Academy of Sciences, 116*(44), 22071–22080. https://doi.org/10.1073/pnas.1900654116
9. **Rao, K., Liu, Y., Jha, D., Ward, L., Choudhary, A., Wolverton, C., & Agrawal, A.** (2020). Is explainable AI intrinsically interpretable? *Patterns, 1*(1), 100020. https://doi.org/10.1016/j.patter.2020.100020
10. **Tshitoyan, V., Dagdelen, J., Weston, L., Dunn, A., Rong, Z., Kononova, O., ... & Persson, K. A.** (2019). Unsupervised word embeddings capture latent knowledge from materials science literature. *Nature, 571*(7763), 95–98. https://doi.org/10.1038/s41586-019-1335-8