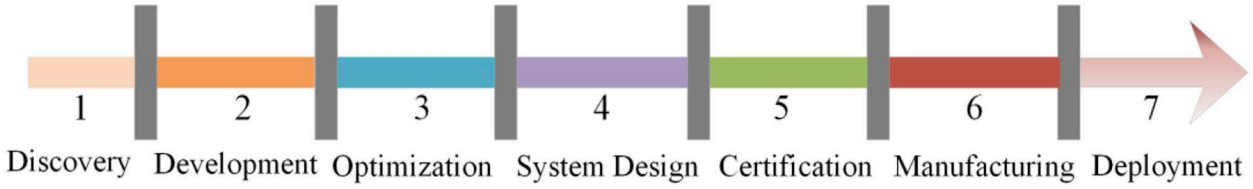
Materials discovery and design using machine learning by Yue Liu, Tianlu Zhao ,Wangwei Ju , Siqi Shi

This review paper provides a comprehensive overview of the application of machine learning techniques in the field of materials science. It covers various methods used for materials discovery and design, discusses the challenges and opportunities in the field, and summarizes recent advances and trends. **Review papers like this are valuable for understanding the current state of research and identifying areas for future exploration.**

Important pointers from the paper:

- Traditional experiments and computational modelling often consume tremendous time and resources and are limited by their experimental conditions and theoretical foundations. For example, the **time frame for discovering new materials** is remarkably long, typically **approximately 10-20 years from initial research to first use**

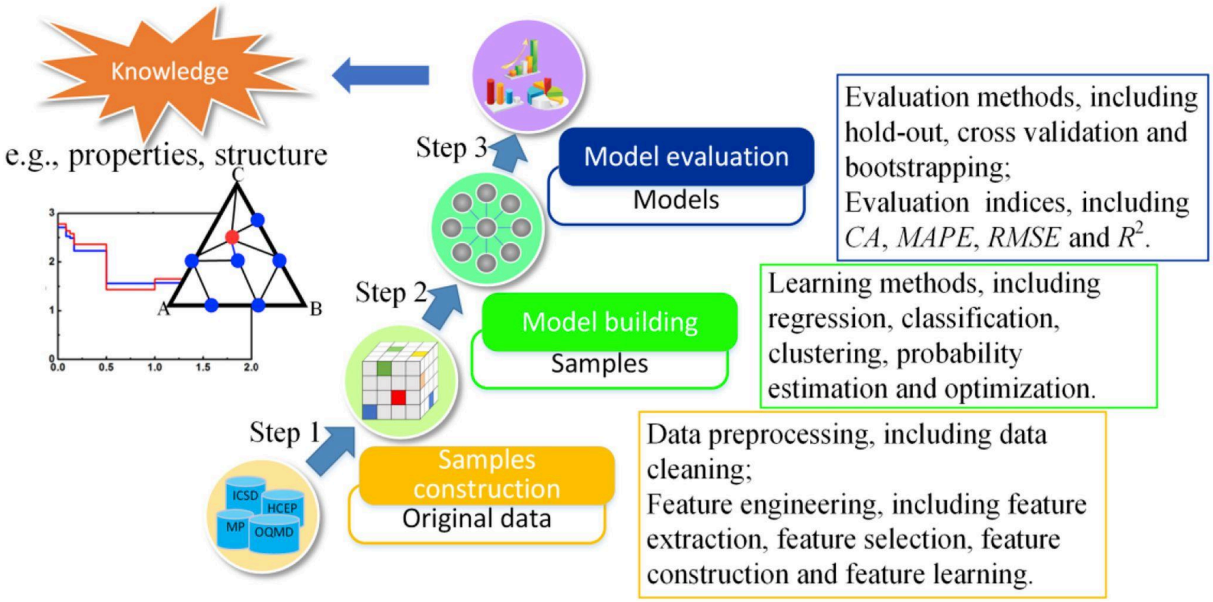


- Machine learning is a method of automating analytical model building. Using algorithms that iteratively learn from data, machine learning allows computers to find hidden insights without being explicitly programmed where to look. The **pioneering** applications of machine learning in materials science can be traced back to the **1990s**, when machine learning methods such as symbol methods and artificial neural networks (ANNs) were employed to **predict the corrosion behavior and the tensile and compressive strengths of the fiber/matrix** interfaces in ceramic-matrix composites.

- Subsequently, machine learning has been used to address various topics in materials science, such as **new materials discovery and material property prediction.**

- A classical definition of machine learning is as follows: , where P, T and E denote performance, task and experience, respectively. The main interpretation is that a computer program is said to learn from experience E with respect to some class of tasks T and a performance measure P if its performance on tasks in T, as measured by P, improves with experience E

- The **construction of a machine learning system is divided into three steps:** sample construction, model building and model evaluation.



1. Sample construction

In materials science, the original data are collected from computational simulations and experimental measurements. These **data are typically incomplete, noisy and inconsistent**, and thus, **data cleaning should be performed** when constructing a sample from the original data. It is important to use a **proper feature selection method** to determine the subset of attributes to be used in the final simulation.

2. Model building

For typical research in materials science, **complex relationships** usually exist between the conditional factors and the target attributes, which traditional methods have difficulty handling. however, a **machine learning method** can be used to model the relationships between conditional factors and decision attributes based on a given sample. This is where machine learning plays a role and where the "core" algorithms lie.

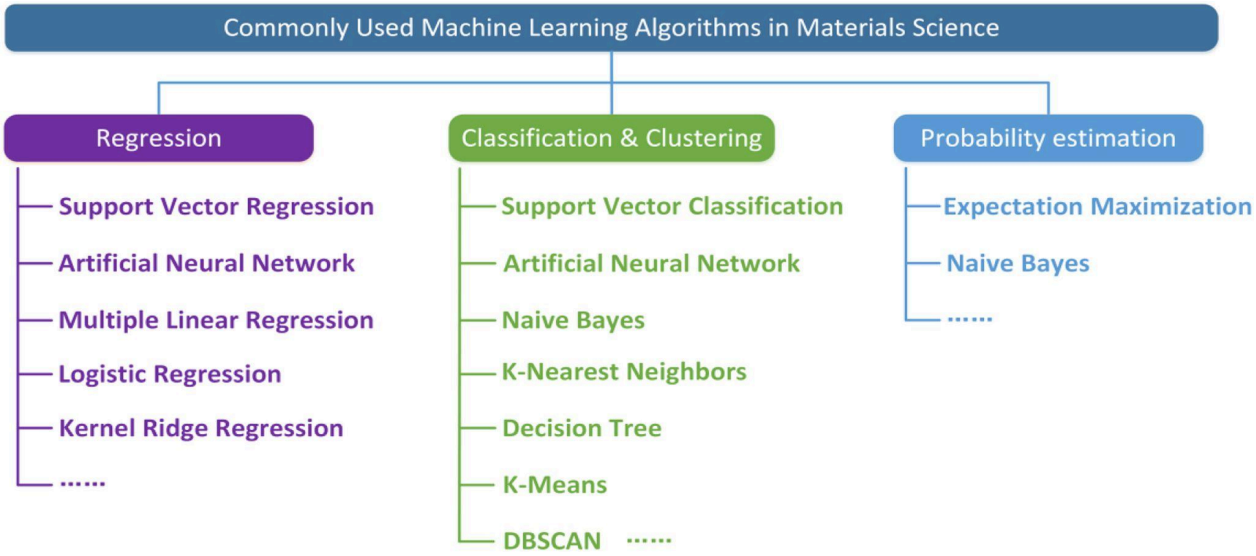
3. Model evaluation

A data-driven model should **achieve good performance** not only on existing data but also on **unseen data.**

Table 1
The comparison of three evaluation methods.

Method	Advantages	Disadvantages	Applicable situation
Hold-out	Low computational complexity.	The proper relative proportions of training/testing data are difficult to determine; The volume of the training data is smaller than that of the original dataset.	The data volume is sufficient.
Cross-validation LOOCV	Not greatly influenced by changes in the volume of training data.	The computational complexity is high, especially on a large dataset; The volume of the training data is smaller than that of the original dataset.	The data volume is sufficient. The data volume is small, and the training and testing data can be partitioned effectively.
Bootstrapping	Effective partitioning of training/testing data.	The distribution of the training data differs from that of the original dataset.	The data volume is small, and the training and testing data are difficult to properly partition.

- Commonly used **machine learning algorithms in materials science** can be divided into four categories: probability estimation, regression, clustering, and classification. Specifically, probability estimation algorithms are mainly used for **new materials discovery**, whereas regression, clustering and classification algorithms are used for **material property prediction** on the macro- and micro-levels.



- The applications of machine learning in materials discovery and design can be divided into three main classes: material property prediction, new materials discovery and various other purposes. In research on material property prediction, regression analysis methods are typically used, and both macroscopic and microscopic properties can be predicted. The main idea underlying the application of machine learning in new materials discovery is to use a probabilistic model to screen various combinations of structures and components and finally to select a material with good performance from the candidate set by means of density functional theory (DFT)-based validation. In addition, machine learning is also used for other purposes in materials science, such as process optimization and density function approximation.

- The application of machine learning in material property prediction

  The application of machine learning in material property prediction offers an efficient **alternative to traditional methods** such as computational simulations and experimental measurements, which are often **complex, time-consuming, and resource-intensive**. The basic idea is to leverage machine learning algorithms to map the nonlinear relationships between material properties and their influencing factors using existing empirical data. The process generally involves three steps: **feature engineering to identify relevant attributes, model training to establish the relationship between these attributes and material properties, and using the trained model for prediction**.

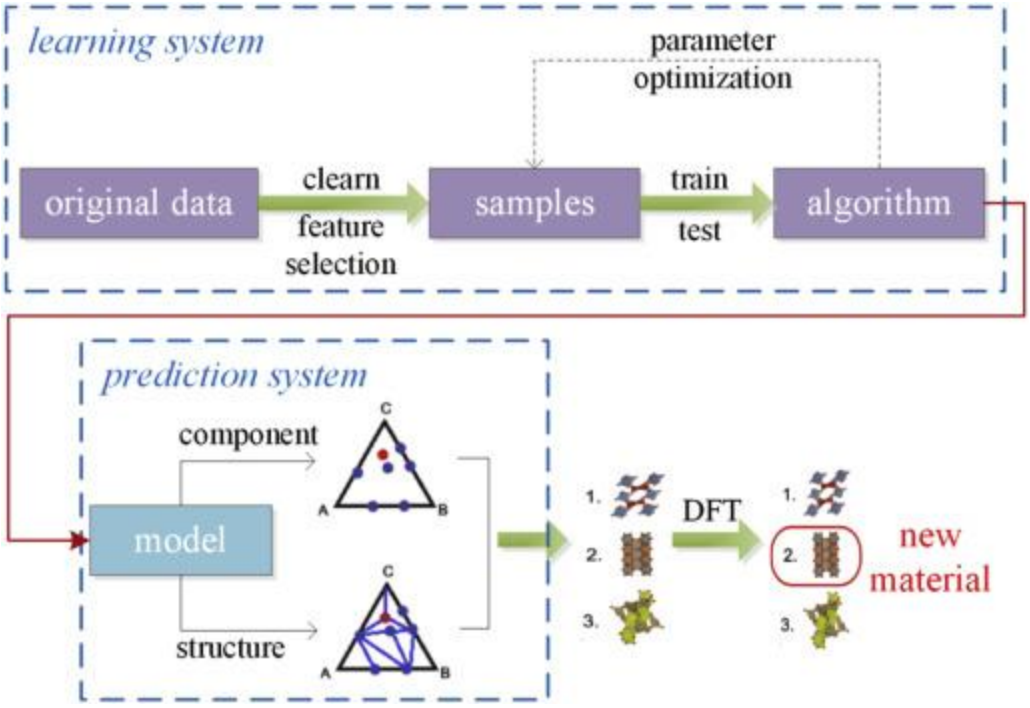
  An example is the **Property-Labelled Materials Fragments (PLMF) tool,** which uses gradient boosting decision trees (GBDT) to predict properties like band gap energy and thermomechanical attributes of inorganic crystals. The **model's accuracy** is evaluated using techniques like **fivefold cross-validation, ROC curves, RMSE, MAE, and R^2**. Machine learning applications in this field can be categorized into macroscopic performance prediction and microscopic property prediction, depending on the scale of the analysis. The development of intelligent and high-performance

prediction models through machine learning is crucial for reducing the time and computational costs associated with materials discovery and design.

- **Back propagation ANNs (BP-ANNs)** have been used to predict material behaviors such as temperature responses and tensile, elongation, wastage, corrosion and compressive properties.
- **RBF-ANNs** are another type of ANN that, by combining the ANN concept with the radial basis function, can fundamentally overcome the problem of local minima and also have the advantage of a high convergence rate.
- ANN modelling has found a place in other applications, such as the prediction of **melting points** , the **density and viscosity of biofuel compounds** , **excited-state energies** , diffusion barriers and **other functional properties**
- A **common criticism of ANNs** is that they require a highly diverse training dataset with sufficient representative examples for property prediction in order to capture the underlying structure to a sufficient extent that their results can be generalized to new cases. Furthermore, an inherent deficiency of neural networks is that the learned knowledge is concealed in a large number of connections, which leads to poor comprehensibility, i.e., **poor transparency of knowledge and poor explainability.**
- **Compared with ANNs, SVM models are more suitable for application to small samples** and can successfully overcome the problems of "the curse of dimensionality" and "overlearning". SVM models can also be employed to predict ionic conductivities, glass transition temperatures and various behaviors of functional materials
- Recently, machine learning methods have also been applied in material property prediction for lithium-ion batteries. it was found that **DBSCAN** can be used to recognize lattice sites, determine the site type, and identify Li hopping events.
- In addition, ANNs have been successfully used for the **prediction of vacancy migration and formation energies, electron affinities , vacancy migration energies and potential energies ,** thereby fostering a deeper physical understanding of the microscopic properties of complex chemical systems.

- The application of machine learning in discovering new materials -

  The machine learning system for discovering new materials includes **two parts, i.e., a learning system and a prediction system**. The learning system performs the operations of data cleaning, feature selection, and model training and testing. The prediction system applies the model that is obtained from the learning system for component and structure prediction. New materials are often "predicted" through a suggestion-and-test approach: candidate structures are selected by the prediction system through **composition recommendation and structure recommendation, and DFT calculations are used to compare their relative stability.**



- The application of machine learning for various other purposes

  Battery monitoring- By using voltage and discharge efficiency as the input variables, this machine learning system not only can generate an estimate of how much residual battery power is available but also can provide users with additional useful information, such as an estimated travel distance at a given speed.

- Several common challenges and potential solutions in the application of machine learning (ML) to materials science:
  1. **Sample Construction**: Issues arise from the sources of data, construction of feature vectors, and determination of sample size. Materials data often come from various sources and lack a unified format. The accuracy of ML models heavily depends on feature vectors, which are specific to different applications. The determination of sample size is crucial, as it impacts the ability to uncover inherent patterns in the data.
  2. **Generalization Ability**: This refers to a model's capability to predict new, unseen examples. Improving generalization involves balancing the complexity of the model to avoid under-fitting (insufficient learning) and over-fitting (excessive complexity). The paper highlights the importance of considering sample quality, size, and the chosen training algorithm to enhance generalization ability.
  3. **Understandability**: Many ML models are considered "black boxes," meaning their internal workings are not easily interpretable. The text suggests two solutions: developing more interpretable algorithms or extracting intelligible knowledge from less interpretable ones. Improving the understandability of ML models is crucial, especially in materials research, to ensure that the models provide meaningful insights.
  4. **Usability**: The complexity of using ML methods in materials science can be a barrier. This complexity includes the need for professional knowledge in processes like dimension

reduction and parameter determination. The text highlights the importance of optimizing parameters and suggests that improving usability is an urgent challenge.

5. **Learning Efficiency**: While not a significant concern now due to relatively small datasets, learning efficiency will become critical as materials science moves into the era of "big data." The text suggests that high-performance computing techniques like parallel and cloud computing may be necessary to address future challenges in learning efficiency.

Summary :

The paper titled "Materials Discovery and Design Using Machine Learning" by Yiu Lie provides a comprehensive review of the application of machine learning (ML) in the field of materials science. It highlights how ML techniques are revolutionizing the way new materials are discovered, characterized, and designed. The review covers the fundamental concepts of ML, its integration with materials informatics, and the various algorithms and models used for predicting material properties, optimizing material performance, and accelerating the discovery process.

1. **Introduction to ML in Materials Science:**
   - ML is used to analyze large datasets of materials properties to identify patterns and correlations that traditional methods might miss.
   - The integration of ML with materials science has the potential to significantly reduce the time and cost associated with discovering new materials.

2. **Key ML Techniques:**
   - Supervised learning, unsupervised learning, reinforcement learning, and deep learning are discussed in the context of their applications in materials science.
   - Techniques like decision trees, random forests, support vector machines (SVM), and neural networks are highlighted as particularly useful for predicting material properties.

3. **Applications of ML in Materials Science:**
   - ML is applied to predict properties like electronic structure, mechanical properties, phase stability, and chemical reactivity.
   - The paper reviews the use of ML in designing materials with specific properties, optimizing manufacturing processes, and discovering new materials for energy storage, catalysis, and other applications.

4. **Challenges and Limitations:**
   - The review discusses challenges such as the need for large, high-quality datasets, the complexity of feature selection, and the interpretability of ML models.
   - The paper emphasizes the importance of developing more interpretable and explainable ML models to gain deeper insights into material behavior.

5. **Conclusion/Recommendations:**

- The paper concludes by stressing the importance of collaboration between data scientists and materials scientists to fully realize the potential of ML in materials discovery and design.
- It recommends focusing on improving the quality and availability of materials data, developing more advanced ML algorithms tailored to materials science, and enhancing the interpretability of ML models.
- The paper suggests that future research should aim at integrating ML with high-throughput experiments and simulations to create more robust and predictive models for materials discovery.