crystals

MDPI

Review

A Comprehensive Review of Machine-Learning Approaches for Crystal Structure/Property Prediction

Mostafa Sadeghian *, Arvydas Palevicius and Giedrius Janusas *

Faculty of Mechanical Engineering and Design, Kaunas University of Technology, Studentu 56, 51424 Kaunas, Lithuania
* Correspondence: mostafa.sadeghian@ktu.edu (M.S.); giedrius.janusas@ktu.lt (G.J.)

Abstract

Crystal Property Prediction (CPP) and Crystal Structure Prediction (CSP) play an important role in accelerating the design and discovery of advanced materials across various scientific disciplines. Traditional computational approaches to CSP/CPP often face challenges such as high computational costs, limited scalability, and difficulties in exploring complex energy surfaces. In recent years, the combination of machine learning (ML) has emerged as a powerful approach to overcome these limitations, offering data-driven methods that enhance prediction accuracy while significantly reducing computational expenses. This review provides a comprehensive overview of the evolution of CSP and CPP methodologies, with particular emphasis on the transition from classical optimization algorithms to modern ML-based methods. Various supervised and unsupervised ML algorithms applied in this field are discussed in detail. Beyond structure and property prediction, recent advancements in ML-based modeling of crystal defects are also reviewed. Moreover, several recent case studies on CSP/CPP are presented to demonstrate the practical effectiveness of ML approaches. Finally, the review discusses current challenges and provides recommendations for future research in ML-based investigations of CSP/CPP.

Keywords: machine learning; crystal structure prediction; crystal property prediction; crystal defect

check for updates

Academic Editor: Andreas Hermann

Received: 20 September 2025
Revised: 21 October 2025
Accepted: 24 October 2025
Published: 28 October 2025

Citation: Sadeghian, M.; Palevicius, A.; Janusas, G. A Comprehensive Review of Machine-Learning Approaches for Crystal Structure/Property Prediction. *Crystals* **2025**, *15*, 925. https://doi.org/10.3390/cryst15110925

1. Introduction

1.1. Importance of Crystal Structure/Property Prediction in Material Science

Crystals are solids in which atoms, ions, or molecules are arranged in a regular, repeating lattice. Their geometric symmetry gives rise to distinctive physical and chemical behaviors. These characteristics control how crystals react to light, electric fields, pressure, temperature, and other environmental factors. The significant effects, including the piezoelectric and pyroelectric responses, birefringence, and light dispersion, are caused by their internal symmetry. Due to their exceptional purity and electrical, magnetic, optical, and acoustic characteristics, crystals are widely used in fields such as pharmaceuticals, electronics, superconductors, and explosives [1].

Crystal structure prediction (CSP) and crystal property prediction (CPP) play a vital role in modern materials science. They provide essential structural information for studying the electronic, magnetic, and optical properties of materials [2]. The main goal of CSP is to determine the most stable atomic arrangement of a material based solely on its chemical composition [3]. In addition to predicting the most stable structures, CSP also investigates metastable configurations that may exhibit new or desirable properties [4]. By exploring a wide range of chemical compositions, CSP facilitates the discovery of new compounds [5].

In CSP, identifying the most stable atomic configuration is achieved by locating the lowest-energy structure on the potential energy surface, as determined by quantum-mechanical calculations [6]. However, the number of possible atomic arrangements increases rapidly with the number of atoms in the unit cell, making exhaustive searches computationally infeasible. Evaluating these configurations typically relies on first-principles calculations such as density functional theory (DFT), which provide high accuracy but are computationally expensive. For this reason, CSP is often applied to relatively small systems [7].

Accurate prediction of crystal structures and their properties is crucial because these properties directly influence the physical, chemical, mechanical, and electronic behavior of materials. Parameters such as formation energy, band gap, thermal conductivity, and elastic moduli determine the stability, performance, and potential applications of a material in real-world technologies. For instance, materials used in batteries, photovoltaics, thermoelectrics, or aerospace applications must meet strict property requirements. Reliable prediction of these characteristics enables efficient materials screening and design, reducing experimental cost and development time while supporting innovation across diverse scientific and industrial fields [8–11].

1.2. Traditional CSP/CPP Analyses and Their Limitations

Conventional CSP/CPP approaches are those that do not incorporate ML techniques. These methods generally consist of three key steps: (i) generating candidate structures, (ii) optimizing these structures, and (iii) searching for the most stable configurations. Conventional CSP/CPP methods are mainly based on first-principles calculations such as DFT. Although DFT provides high accuracy, it is often computationally expensive. To address this, hybrid approaches combining DFT with crystal chemical principles and the bond valence method have been proposed. For example, the authors of ref. [12] showed that this combination can significantly reduce computational cost while maintaining good accuracy. However, such conventional approaches are still less efficient than modern ML-based methods [13,14]. Typically, initial structures are produced via random sampling while enforcing symmetry and interatomic distance constraints. These structures are evaluated using DFT calculations, often coupled with global optimization algorithms, to locate low-energy arrangements. The main distinction between different CSP/CPP methods lies in the optimization strategies employed, which include Random Search, Particle Swarm Optimization (PSO), Genetic Algorithms (GA), Bayesian Optimization (BO), Simulated Annealing (SA), and Template-Based Methods.

Random search is a basic CSP/CPP approach that relies on extensive random sampling to identify low-energy structures [15]. Implementations such as ab initio random structure searching (AIRSS) [16] and special quasi-random structures (SQS) [17] have successfully predicted new phases of hydrogen [18], nitrogen [19], lithium [20], and complex alloys [21]. However, its efficiency decreases rapidly with system size, though enhancements such as geometric constraints [22] and parallel computation [23] have been introduced to improve performance.

PSO [24] iteratively refines structures based on collective and individual performance. The CALYPSO [25] has enabled discoveries in battery, superconductor, photovoltaic, and electronic materials [25,26]. Though simple and requiring few parameters, PSO may still become trapped in local minima when exploring complex energy landscapes.

GA, based on evolutionary processes, evolves structures through operations such as selection, crossover, and mutation. The USPEX tool has successfully predicted materials such as Sr_5P_3 electrides, TiN_2, H_3S superconductors, and transparent sodium phases. Despite their effectiveness in navigating complex energy landscapes, GA methods remain computationally intensive, particularly for large systems using DFT. The convergence toward the global minimum can also be slow and sensitive to population size and modification parameters [27,28].

BO, combining Bayesian inference and Gaussian process regression, accelerates structure searches by constructing surrogate models and acquisition functions. BO has been applied to clusters such as Cu_{15} [29], Cu-Ni [30], and C_{24} [31]. However, the computational cost of updating surrogate models and dealing with uncertainties remains a challenge [32].

SA [32] allows temporary transitions to higher energy states to avoid local minima and improve exploration of the potential energy surface. It has been applied to various systems [33,34]. Its performance, however, depends strongly on the choice of parameters such as the initial temperature and cooling rate, often requiring manual tuning and expert knowledge.

Other methods, such as Ion substitution, predict new structures by replacing ions in known lattices based on empirical rules such as the Goldschmidt tolerance factor [35]. Advanced probabilistic models have also been developed to quantify substitution probabilities using crystallographic databases, enabling the discovery of new ternary nitrides and oxides.

While traditional CSP/CPP methods have achieved notable successes, they still face serious limitations in computational efficiency and scalability. The most critical challenges include the rapidly increasing configurational search space with system size and the high computational cost of repeated DFT evaluations during structure optimization. Consequently, these methods are often restricted to small or moderately complex systems, highlighting the need for more efficient and scalable computational strategies in crystal structure and property prediction.

1.3. ML in CSP/CPP Analyses

The increasing complexity of materials design and the limitations of conventional CSP/CPP methods have motivated the development of more advanced and scalable computational approaches. Recent efforts focus on enhancing efficiency, accuracy, and generalizability while reducing dependence on repetitive calculations [11,36]. Among these efforts, ML has become a powerful complement to traditional methods. By learning relationships between atomic structures, energies, and properties from existing datasets, ML can provide fast and accurate surrogate models that accelerate structure prediction, optimization, and property estimation. This method allows the exploration of vast chemical and structural spaces that were previously computationally inaccessible [13,14].

The historical evolution of computational methodologies in crystal and molecular analysis highlights a transition from direct structure–property mappings to data-based, predictive methods (Figure 1). In the early stages of this development, computational approaches primarily relied on local optimization techniques to correlate given structures with their corresponding properties. As methodologies advanced, compositional variables were introduced into the modeling process, enabling the use of global optimization algorithms to simultaneously predict structural configurations and material properties from elemental inputs. Recently, ML has played an important role in combining various physicochemical datasets into predictive modeling workflows. This current phase emphasizes robust data representation, learning approaches (supervised and unsupervised), and careful model selection and validation. Therefore, these improvements show a shift from trial-and-error methods to smart, algorithm-based ways of discovering and studying materials [11].

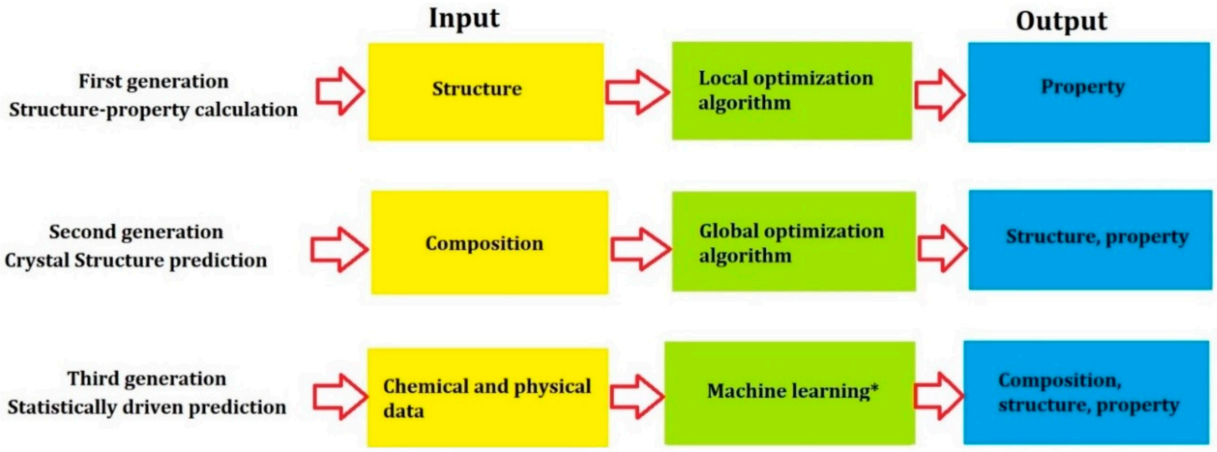
Input
Output
First generation
Structure-property calculation
Structure
Local optimization algorithm
Property
Second generation
Crystal Structure prediction
Composition
Global optimization algorithm
Structure, property
Third generation
Statistically driven prediction
Chemical and physical data
Machine learning*
Composition, structure, property

Figure 1. The evolution of CSP/CPP studies over time (Notes "*" indicates that ML includes: (i) data collection: experiment, simulation, databases; (ii) representation: optimize format, remove noise, extract features; (iii) type of learning: supervised, semi-supervised, unsupervised; (iv) model selection: cross-validation, ensembles, anomaly checks) [11].

Beyond improving efficiency, ML introduces a fundamental change in materials research. However, several challenges remain, including limited and uneven data availability, the generalization of models across diverse materials classes, and ensuring that predicted structures are experimentally achievable [37]. Addressing these issues will be essential for developing reliable, interpretable, and fully combined ML-assisted CSP/CPP studies, which are discussed in the following.

2. ML Algorithms for CSP/CPP Analyses

ML is a branch of artificial intelligence that encompasses a suite of algorithms designed to autonomously obtain knowledge and refine task performance through experiential learning from data, without the necessity of explicit programming. Conventionally, these algorithms are categorized into two principal methods based on the nature of data supervision: supervised and unsupervised learning. The former uses labeled datasets to guide model training, whereas the latter facilitates the discovery of structures within unlabeled data. The following section explains ML algorithms applied to CSP/CPP analyses.

2.1. Supervised Learning

Supervised learning aims to develop predictive models by training on labeled datasets, where each instance consists of an input and a corresponding target output. The problem is treated as either a classification task (such as distinguishing between crystalline and amorphous phases) or a regression (such as predicting solubility, melting points, or lattice energies). Input features in crystallization applications may include molecular descriptors (2D and 3D), structural fingerprints, or image-derived features such as pixel intensities and morphological patterns. In some applications, descriptors often capture electronic, geometric, and topological properties; in image-based tasks, convolutional neural networks (CNNs) are commonly used to extract relevant patterns. The choice of model architecture, whether traditional algorithms (such as support vector machines, random forests (RF), and gradient boosting) or DL approaches (such as CNNs and fully connected neural networks), is guided by the data and problem complexity. Recently, supervised learning supports a wide range of applications in crystal structure analysis, including property prediction and automated phase classification, among others [38].

2.1.1. Linear Regression

Regression models are essential tools for predicting continuous properties in crystal-related studies, such as solubility, melting points, lattice energies, and crystal propensity. While linear regression provides a useful baseline, more advanced methods (including Gaussian process regression, kernel-based techniques, ensemble models, and DL regressors) offer greater capacity to model complex, nonlinear relationships. Also, regularization approaches such as Ridge regression and LASSO help address challenges such as multicollinearity and overfitting (particularly in high-dimensional data). These modern regression methods are increasingly enabling data-based discovery via crystal structure and property (CSP and CPP) analyses [38,39].

2.1.2. k-Nearest Neighbor (k-NN) Algorithm and Naive Bayes (NB)

The k-NN algorithm is a simple and effective nonparametric method used for both classification and regression. It operates on the principle that data points close to each other in the feature space tend to have similar properties or labels. The algorithm assigns labels based on the majority of the nearest neighbors (for classification) or averages their values (for regression). However, in high-dimensional data, its performance may decline due to the challenges associated with too many features [40].

NB is a simple and fast model that uses Bayes' rule. It assumes each feature works independently when making predictions. Despite this simplifying assumption, it often performs well, especially on high-dimensional datasets or when training data is limited. Also, NB has been applied in the related studies for feature selection, outcome prediction, image classification, and quality control using Raman spectroscopy. Its simplicity, efficiency, and robustness make it a practical choice for many CSP/CPP-related ML studies [41].

2.1.3. Support Vector Machine (SVM)

SVMs are powerful supervised learning algorithms used for both classification and regression, particularly effective in high-dimensional spaces and in solving ill-posed problems. In chemical and materials sciences, SVMs have been successfully applied to predict solubility, lipophilicity, melting points, partition coefficients, and cocrystal formation. By constructing an optimal hyperplane that maximizes the margin between classes, SVMs improve generalization. For nonlinear problems, kernel methods (such as linear, polynomial, and radial basis function kernels) enable efficient modeling of complex relationships without explicit high-dimensional transformations. Moreover, the SVM method extends to regression via Support Vector Regression (SVR), making it highly suitable for CSP/CPP-related applications [42,43].

2.1.4. Decision Trees and Ensemble Methods

Decision trees are easy to understand because they follow simple rules step by step. This makes it feel clear why a certain prediction was made. They are robust to outliers and missing data, and offer a good balance between accuracy and interpretability. Decision trees are built by repeatedly splitting the data in a way that increases the purity of each resulting node. It uses criteria, such as information gain. Ensemble methods, such as RF, further enhance performance by combining multiple trees. Also, it uses techniques such as feature randomization to improve generalization and reduce overfitting [42].

2.1.5. Artificial Neural Networks (ANNs)

ANNs are a type of ML model capable of learning complex, nonlinear relationships from data. They consist of interconnected layers of neurons (input, hidden, and output) that process and transform information through weighted connections and activation functions.

ANNs are widely used in predicting physicochemical properties, crystal density, solubility, and analyzing spectral data. Also, common architectures include feedforward neural networks for modeling nonlinear static relationships, recurrent neural networks for capturing temporal dynamics in crystallization processes, and CNNs for image-based applications, such as classifying crystallographic images. Despite their power, ANNs require large datasets and careful regularization to avoid overfitting, making them particularly valuable in data-rich, complex CSP/CPP studies [9,44].

Table 1 shows an overview of selected ML methods, highlighting their advantages and limitations, which aid researchers in selecting appropriate algorithms for specific tasks in CSP/CPP.

Table 1. Some advantages, challenges, and applications of ML models [1,9,42,44,45].

ML Algorithm	Advantages	Challenges	Typical Use Cases/Scenarios to Avoid
Linear Regression	Simple, transparent, and easy to interpret; computationally efficient; provides direct insights into variable relationships.	Poor prediction for nonlinear problems; sensitive to outliers; unreliable when predictors are colinear	Use for: Quick baseline modeling, trend estimation, or problems with clear linear relationships. Avoid for: Highly nonlinear systems or datasets with strong multicollinearity.
k-NN	Nonparametric and intuitive; adapts naturally to nonlinear data; no training phase required.	Slow for large datasets; suffers from dimensionality; sensitive to outliers	Use for: Pattern recognition, anomaly detection, and small to medium datasets with local similarity patterns. Avoid for: Real-time applications or large-scale data requiring fast predictions.
NB	Fast and efficient; works well with small or sparse data; effective for categorical features.	Assumes feature independence; low accuracy probability outputs; requires representative training data	Use for: Text classification, spam filtering, or quick baseline models. Avoid for: Tasks involving dependent or continuous variables.
SVM	High accuracy in high-dimensional spaces; robust to overfitting; effective with limited data	Not suitable for very large datasets; poor performance with overlapping classes; sensitive to kernel selection	Use for: Image recognition, bioinformatics, or moderate-size, high-dimensional problems. Avoid for: Very large or noisy datasets.
Decision Trees	No need for normalization; deals with missing data effectively; easy to visualize and interpret	Prone to overfitting; sensitive to training data variations; longer training times	Use for: Feature selection and interpretable classification tasks. Avoid for: Small, noisy datasets or where stability is critical.
ANNs	Highly flexible; captures complex nonlinear relationships; scales well to large datasets.	Requires large training data; computationally expensive; low interpretability	Use for: Predicting complex material or process properties, image/spectra analysis, or multi-output regression. Avoid for: Small datasets or when interpretability is required.

2.2. Unsupervised Learning

Unsupervised learning algorithms operate without labeled data, making them highly valuable in situations where manual data labeling is not practical. These methods are commonly used for dimensionality reduction, exploratory data analysis, clustering, detecting data that does not fit the pattern, and finding hidden patterns in the data. Clustering groups data points based on feature similarity, but another method, called association rule learning, identifies patterns and relationships between variables through if–then rules. Although unsupervised learning offers flexibility and scalability, it typically faces challenges such as higher computational complexity and reduced accuracy compared to supervised approaches (due to the lack of explicit labels to guide model training). These methods are essential for pattern discovery and hypothesis generation in large, complex datasets [42,44].

2.2.1. Dimensionality Reduction: Principal Component Analysis (PCA) and Nonnegative Matrix Factorization (NMF)

PCA is a widely used unsupervised dimensionality reduction technique that transforms high-dimensional data into a lower-dimensional space while preserving maximal variance by projecting data onto orthogonal principal components (which capture the most significant patterns in the dataset). In chemistry and materials science, PCA is extensively applied to spectral data for tasks such as solvent selection, purity assessment, and monitoring crystallization processes. Preprocessing steps such as standardization are critical to ensuring balanced feature contributions. Complementary to PCA, NMF offers an alternative approach that decomposes data into nonnegative components, providing interpretable parts-based representations. Both PCA and NMF play an important role in exploratory data analysis and feature extraction, helping reveal hidden structures in complex datasets [42,44,45].

2.2.2. Principal Component Regression (PCR)

PCR reduces the number of features. Then it builds a regression model. This helps when many features are related to each other. Such cases are common in IR and Raman spectroscopy. Also, PCR applies PCA to transform the original variables into orthogonal principal components, which are then used as inputs for the regression model. Unlike Partial Least Squares Regression (PLSR), PCR constructs components based solely on the predictor variables, without considering the target variable during component extraction. While this approach helps mitigate collinearity and noise, it often requires more components than PLSR to achieve comparable predictive performance. PCR remains a valuable tool in materials and CSP/CPP studies for analyzing complex, correlated data [42,44,45].

3. ML Approaches for CSP/CPP Analyses

3.1. Background and Recent Trends in ML for CSP/CPP Analyses

CSP/CPP is a fundamental and challenging topic in materials science with significant implications for the discovery and design of new materials. Traditional CSP/CPP methods based on first-principles calculations and global optimization are computationally intensive. They are often impractical for complex materials, such as HEAs, high-entropy oxides (HEOs), organic crystals, and perovskites [46–48]. In recent years, ML approaches have emerged as a good alternative. They can enable faster predictions by learning structure–property relationships from data [23,49]. Various ML models, including gradient boosting algorithms [50], GNNs [23,47], transfer learning methods [46], and metric learning techniques [49], have been successfully applied to predict crystal structures among diverse material analyses.

Some software, such as CrySPAI [51] and CrySPY [52], uses ML with optimization and fast workflows. They are useful tools for speeding up CSP. These approaches have shown advantages such as reduced computational cost, accuracy, and broader applicability to new material domains [23,46–53]. Nevertheless, current ML-based CSP methods still face limitations, including data scarcity, model generalizability across chemical spaces, and the integration of domain knowledge to guide predictions [49,53]. Despite these challenges, ML based CSP is playing an increasingly critical role in the discovery of new materials and in enabling the design of functional materials for applications in energy, electronics, and related fields [47,51,53].

In recent years, ML has significantly advanced CSP by incorporating physical quantities (such as energy, forces, and magnetic moments) into neural networks and training them on large-scale datasets [54,55]. ML contributes to CSP in different areas. ML-based structural representation methods transform complex crystal geometries into high-dimensional feature vectors that are invariant to rotation, translation, and atom indexing [9,56,57], improving the classification of candidate structures [57]. Moreover, ML models (particularly GNNs) can rapidly predict properties such as formation energy and band gap [58,59], and when integrated with global optimization methods, efficiently guide the search for low-energy structures [60]. Also, ML-based force fields provide near first-principles accuracy at significantly reduced computational cost, enabling large-scale structure optimization [61,62]. Lastly, generative models such as VAEs, GANs, and diffusion models can explore chemical space more broadly than traditional approaches [63].

An important step in ML-based CSP is converting raw atomic configurations into numerical descriptors that capture structural and chemical features. These descriptors must satisfy physical invariance, index invariance, and sufficient discriminatory power [9,64]. Two principal strategies are used: 3D voxel representations, which reconstruct atomic arrangements through neural encoders and decoders [65]; and matrix-based representations, more commonly used in GNNs and force fields, which organize lattice parameters and elemental properties into structured arrays [8,66]. Initial atomic features are typically derived from elemental data [8,9] and refined through models such as CGCNN [9], MPNN [67], MEGNet [8], and ALIGNN [68]. To describe local chemical environments, bonding features such as the Behler–Parrinello symmetry functions [68,69] and SOAP descriptors [70] are widely used. These are particularly useful for accurately modeling interatomic interactions.

These features can then be used to quickly predict properties. GNNs trained on large datasets can identify low-energy candidates efficiently [9,68,70]. For final structure refinement, ML-based force fields such as MEGNet [8], CHGNet [54], NequIP [71], and MACE [72] offer accurate alternatives to traditional DFT, and have been applied to systems ranging from crystalline materials [73] to biological molecules [74,75]. Integration of ML force fields with evolutionary algorithms (such as CALYPSO and SPINNER) has successfully identified global minimum structures in complex systems such as boron clusters and ternary compounds [76,77]. Despite these advancements, challenges remain (particularly in data efficiency, model generalization, and interpretability), which are being addressed through strategies such as transfer learning, explainable AI, and model compression [78–80].

Generative models extend the frontier of CSP/CPP by learning structure–property relationships and directly generating new materials. VAEs such as iMatGen [81] and FTCP [82] enable diversified sampling of candidate structures with desired properties. GANs, such as crystal GAN [83] and CCDCGAN [84], produce realistic and composition-aware materials. Matter Gen [85], a diffusion model customized for crystal periodicity, has outperformed earlier models in generating stable and novel crystal structures. These end-to-end models automate structure generation, identification, and refinement, reducing human bias (but further work is needed to improve their controllability and robustness).

Selecting the appropriate CSP/CPP method depends on the system's complexity and computational constraints. Evolutionary algorithms such as those in USPEX are well-suited for exploring large and complex energy landscapes, while simpler methods such as AIRSS offer a fast, low-cost alternative for early-stage exploration. Particle swarm optimization in

CALYPSO allows for quick convergence, and flexible tools such as GASP and MAISE adapt to a range of applications. For high-cost evaluations, Bayesian approaches such as GOFEE and BEACON efficiently reduce the number of calculations needed. Generative models (including iMatGen and Crystal GAN) enable the design of new materials by learning from complex data distributions. For accurate property prediction, graph-based models such as SCCOP and GN-OA are effective, and large language models such as LLaMA-2 are emerging tools for using big data. Beginner-friendly platforms include USPEX and AIRSS, while Matter Gen and IM2ODE serve specialized needs in novel design and constrained systems. Although conventional CSP/CPP methods are reliable and based on physical theory [2,86], they often require extensive computational resources. In contrast, ML-based CSP/CPP models can predict crystal structures or their properties within seconds to minutes, whereas traditional approaches often require days or weeks. This advantage stems from their ability to learn from data, rapidly identify candidate structures, and perform efficient optimization [87–89]. This makes ML a powerful alternative in modern materials discovery.

3.2. Review of ML-Based Methods for CSP/CPP

ML has become a powerful tool for accelerating CSP/CPP, allowing researchers to explore complex composition/structure/property relationships that are often difficult to capture with traditional ab initio or empirical approaches. In recent years, its applications have expanded from small, single-phase systems to multicomponent and high-entropy materials, where data-driven prediction is increasingly combined with physical constraints to enhance efficiency and interpretability. Early research in this area has demonstrated how ML can reduce computational time, predict structural stability, and generate candidate configurations with accuracy approaching that of DFT but at a fraction of its cost.

ML approaches have increasingly demonstrated their capability not only to predict crystal structures and properties but also to guide successful experimental validation. For example, Mansouri Tehrani et al. [90] used ML-guided screening to identify ultra-incompressible, superhard transition metal borides, several of which were subsequently synthesized and experimentally confirmed through hardness measurements. Similarly, Xue et al. [91] employed an ML model to accelerate the search for $BaTiO_3$-based piezoelectric materials with large recoverable strain, and their predictions were later verified experimentally. The authors of ref. [92] demonstrated an experimentally validated inverse design method, where ML predictions of multi property Fe–Co–Ni alloys were confirmed through high-throughput synthesis and testing. The close agreement between predicted and measured values highlights the strong connection between ML-based design and experimental validation. Furthermore, addressing real-world challenges in inverse materials design (such as data scarcity, model transferability, and discrepancies between computational conditions and experimental synthesis) would provide a more balanced and application-oriented perspective. In another notable case, Balachandran et al. [93] demonstrated that ML-assisted predictions of thermoelectric materials closely matched measured experimental properties, which emphasize the reliability of ML in practical materials discovery. Despite these examples, achieving inverse materials design (where desired properties directly guide the discovery and synthesis of new compounds) remains a significant challenge. As emphasized by Ramprasad et al. [94] and Butler et al. [11], major obstacles include limited and biased experimental datasets, the metastability of predicted structures, and synthesis feasibility constraints. Overcoming these challenges by combining ML-based prediction, automated synthesis, and high-throughput experimental feedback will be crucial for realizing the full potential of ML in crystal structure and property prediction. In ref. [95], several ML models were trained on caffeine cocrystal data, with logistic regression achieving the highest accuracy (97.1%). The predicted cocrystals were experimentally confirmed through molecular interaction and cocrystallization studies with compounds such as catechin and catechol. It (ref. [95]) shows how ML predictions can guide experimental synthesis and help avoid time-consuming trial-and-error experiments.

One of the most important studies in this field was conducted by Liu et al. [48], who developed a classification framework for predicting the crystal structure types of HEOs, materials known for their compositional disorder and structural complexity. Their approach followed a clear ML workflow involving data preprocessing, feature selection, model training, and interpretability analysis (illustrated in Figure 2). Their dataset included 112 experimentally reported HEOs described by seven thermodynamic and structural descriptors, including the anion-to-cation radius ratio (r_A/r_C), differences in Pauling and Mulliken electronegativities, atomic size mismatch, and mixing entropy. By comparing several supervised models (such as Naïve Bayes, Random Forest (RF), and Extreme Gradient Boosting (XGBoost)), they found that XGBoost achieved the highest test accuracy (97.73%) and F1 score (0.975). SHAP analysis indicated that the ionic radius ratio (r_A/r_C) was the key factor controlling the prediction, suggesting that basic physical descriptors can still provide strong predictive ability when combined with advanced ML models. While this study established the potential of ML in distinguishing among fluorite, rock salt, and spinel structures, subsequent research aimed to go beyond classification toward structure generation and optimization.

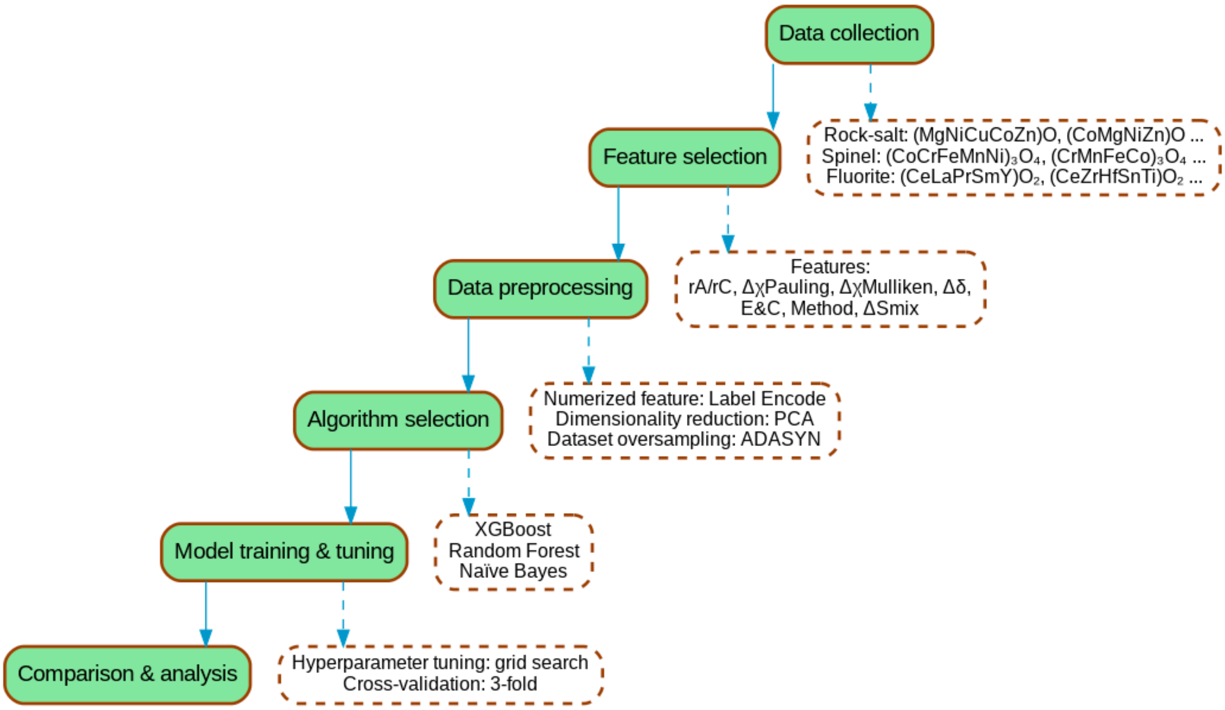
Data collection
Rock-salt: (MgNiCuCoZn)O, (CoMgNiZn)O ...
Spinel: (CoCrFeMnNi)$_3$O$_4$, (CrMnFeCo)$_3$O$_4$ ...
Fluorite: (CeLaPrSmY)O$_2$, (CeZrHfSnTi)O$_2$ ...
Feature selection
Features:
rA/rC, ΔχPauling, ΔχMulliken, Δδ,
E&C, Method, ΔSmix
Data preprocessing
Numerized feature: Label Encode
Dimensionality reduction: PCA
Dataset oversampling: ADASYN
Algorithm selection
XGBoost
Random Forest
Naïve Bayes
Model training & tuning
Hyperparameter tuning: grid search
Cross-validation: 3-fold
Comparison & analysis

Figure 2. General scheme of ML methods for CSP analysis (of HEOs) [48].

Li et al. [47] extended this direction by developing a graph neural network (GNN)-based model for predicting thermodynamically stable inorganic structures. In their model, atoms were represented as nodes and the bonds between them as weighted edges, capturing both the composition and the geometry of the crystal structure. The GNN predicted formation energies, which were combined with Lennard–Jones potential functions to represent interatomic interactions. The BO framework was then applied to minimize formation energy and potential energy simultaneously, enabling efficient discovery of low-energy configurations. Their hybrid GNN–BO model reduced computational time by over an order of magnitude relative to DFT-based searches while maintaining comparable accuracy. Similarly, Chang et al. [46] proposed the Shotgun CSP method, which combines transfer learning with structure generation. Using a Crystal Graph Convolutional Neural Network (CGCNN) pre-trained on the Materials Project database and fine-tuned on partially optimized crystal data, they achieved 93.3% prediction accuracy on 90 benchmark structures without performing iterative DFT relaxations. These studies showed that deep graph models, especially those combined with optimization algorithms, can accelerate CSP workflows and maintain high predictive precision.

Another approach for efficient structure discovery was explored by Kusaba et al. [49], who employed a metric-learning framework for element substitution. Instead of directly predicting energies or structure types, their model learned a similarity metric between known and unknown crystals, achieving 96.4% accuracy in identifying isomorphic structures. This similarity information was then used to select template structures from a database, followed by element substitution and local relaxation to generate new candidate crystals. By using existing structural knowledge, the method effectively reduced computational cost and expanded the search space toward previously unexplored compositions. In the domain of perovskite materials, Jarin et al. [53] applied RF, SVM, neural networks (NN), and genetic algorithm-enhanced neural networks (GA-NN) to classify structures and predict lattice parameters using descriptors such as ionic radius, electronegativity, and Goldschmidt tolerance factor. Their GA-Support Vector Regression model achieved 88% accuracy for structure classification and 95% for lattice parameter prediction, underscoring how hybrid optimization models can detect small but important structural changes in complex oxide systems.

In high-entropy alloys (HEAs), Dey et al. [50] developed an improved ensemble-learning approach to predict both phase stability and crystal structure, training on datasets of 1345 samples for phase prediction and 705 for structural prediction. Among five boosting algorithms (such as AdaBoost, XGBoost, LightGBM, CatBoost, and Gradient Boosting), LightGBM achieved the best performance with 90.07% accuracy in classifying BCC, FCC, and multiphase HEAs. These results confirmed the strong potential of boosting algorithms to handle nonlinear, high-dimensional feature spaces typical of HEA systems. However, such datasets often suffer from class imbalance, a limitation addressed by Hareharen et al. [96], who used the Synthetic Minority Oversampling Technique (SMOTE) to balance HEA phase data. Their XGBoost model improved phase classification accuracy from 80% to 84% and achieved a more balanced F1 score across classes. In predicting crystal structures (BCC, FCC, FCC+BCC), the Random Forest model initially yielded 94% accuracy, but after applying SMOTE, XGBoost achieved 93% with improved reliability across categories. These studies illustrate how data balancing and feature engineering significantly enhance the generalizability of ML models in alloy systems.

While ensemble and metric-learning approaches focus primarily on supervised prediction, deep learning has enabled more expressive structural representations through graph-based and transformer-based models. Feng and Tian [97] introduced the Multiplex Graph Neural Network (MP-GNN), which represents crystals using multiple graph layers to capture both local atomic coordination and global structural topology. This approach achieved lower mean absolute errors than CGCNN and MEGNet when predicting formation energies, band gaps, and elastic moduli, showing superior generalization to out-of-distribution data. Similarly, transformer-based models have recently been applied to encode atomic environments as universal atomic embeddings (UAEs), which can be integrated with existing GNNs. Using this approach, the Crystal Transformer model improved formation energy prediction by 14% in CGCNN and 18% in ALIGNN, and further increased accuracy by 34% in MEGNet for perovskite datasets. Such hybrid graph–transformer models have improved the scalability and accuracy of ML-based property prediction, bridging the gap between structural realism and computational efficiency.

Building on these advancements, recent studies have focused on using large-scale datasets and refined graph architectures for more accurate mechanical property prediction. The authors of ref. [98] used ML to predict mechanical properties of inorganic crystals—including bulk modulus, shear modulus, and Young's modulus—using a dataset called MPCR. Among the tested models, the Atomistic Line Graph Neural Network (ALIGNN) gave the most accurate results. They also introduced Universal Atomic Embeddings (UAEs), generated by a model named Crystal Transformer, to better represent atomic features. Using these features improved the accuracy of formation energy predictions by 14% in CGCNN and 18% in ALIGNN. On a separate perovskite dataset, UAEs further increased accuracy by 34% in MEGNet and 16% in CGCNN, showing good generalization and strong potential for discovering new materials.

Zhou et al. [99] further advanced graph-based property prediction by introducing the Crystal Gated Graph Attention Network (CGGAT). This model integrates graph attention with gated convolutional layers and includes both bond-length and bond-angle information through hierarchical message passing. Trained on over 120,000 crystal structures from the Materials Project and JARVIS-DFT databases, CGGAT achieved mean absolute errors of 0.020 eV/atom for formation energy, 0.042 and 0.069 $\log_{10}$(GPa) for bulk and shear moduli, and 0.273 eV for the PBE band gap, outperforming existing models such as CGCNN, MEGNet, and ALIGNN. These results demonstrate the benefit of incorporating bond-angle information and attention mechanisms in enhancing crystal structure representation and property prediction accuracy. While such GNN architectures focus on accurate property prediction, recent advances have extended ML applications toward crystal structure generation through generative and diffusion-based frameworks.

Similar to discriminative models, generative and diffusion-based ML approaches have emerged as a transformative step in CSP. Nouira et al. [83] proposed CrystalGAN, a generative adversarial network (GAN) that learns to generate ternary hydride structures from binary training data while preserving physical and chemical constraints. Although pioneering, CrystalGAN primarily explored the feasibility of deep generative learning in crystal systems. Luo et al. [100] advanced this concept by developing the Conditional Crystal Diffusion Variational Autoencoder (Cond-CDVAE), a generative model capable of producing crystal structures conditioned on user-defined variables such as chemical composition and external pressure. Trained on an extensive dataset of over 670,000 locally stable structures collected from the Materials Project and CALYPSO databases, the model successfully reproduced 59.3% of 3547 unseen experimental structures and achieved 83.2% accuracy for systems containing fewer than 20 atoms per unit cell. Importantly, the Cond-CDVAE model eliminates the need for iterative DFT relaxations by directly generating nearly stable crystal configurations. Cai-Yuan Ye et al. [101] further improved this approach by introducing Con-CDVAE—a conditional diffusion variational autoencoder that uses property information, such as formation energy, band gap, and crystal system—into the latent space to guide the generation process. Using a two-stage training strategy and testing on the MP20, MP40, and OQMD datasets, the model successfully generated structures consistent with target properties and achieved over 94% success in DFT relaxation, confirming its ability to produce physically realistic and property-aligned crystals. Similarly, Jiao et al. [102] proposed DiffCSP, an equivariant diffusion-based model that ensures rotational and translational invariance during structure generation, achieving DFT-consistent accuracy across benchmark datasets. Zeni et al. [85] developed MatterGen, a diffusion-based generative model that produced twice as many novel, low-energy inorganic materials as prior models and achieved structures closer to DFT minima. These advances show that generative models can now produce physically realistic crystal structures, representing a major step from property prediction toward automatic crystal design.

Using ML together with CSP workflows has made the process faster and more flexible. Yamashita et al. [52] developed CrySPY, an ML-accelerated CSP platform that combines predictive models within evolutionary algorithms, accelerating search convergence. Wang et al. [51] expanded this framework with CrySPAI, which combines DFT, deep neural networks, and adaptive model retraining to form a self-improving prediction system. Similarly, Wang et al. [103] proposed MAGUS (Machine-Learning-And-Graph-Theory-Assisted Universal Structure Searcher), which combines graph-theoretical decomposition with active-learning interatomic potentials. MAGUS achieved the same final stable structures as pure DFT searches for systems such as boron, lithium, and magnesium silicate, while reducing self-consistent-field computations by more than 70%. Figure 3 illustrates the overall CSP method used by them. It shows how the hybrid method works by combining fast ML selection with detailed DFT calculations, demonstrating that the method is both efficient and practical. Li et al. [104] further demonstrated how ML classifiers can filter low-energy structures in ternary alloys such as NiNbX (X = Al, Si, Ti), achieving over 80% accuracy in identifying stable candidates and enabling efficient DFT validation. These approaches illustrate the evolution toward closed-loop ML–CSP systems that continuously refine their predictions and adapt to new data.

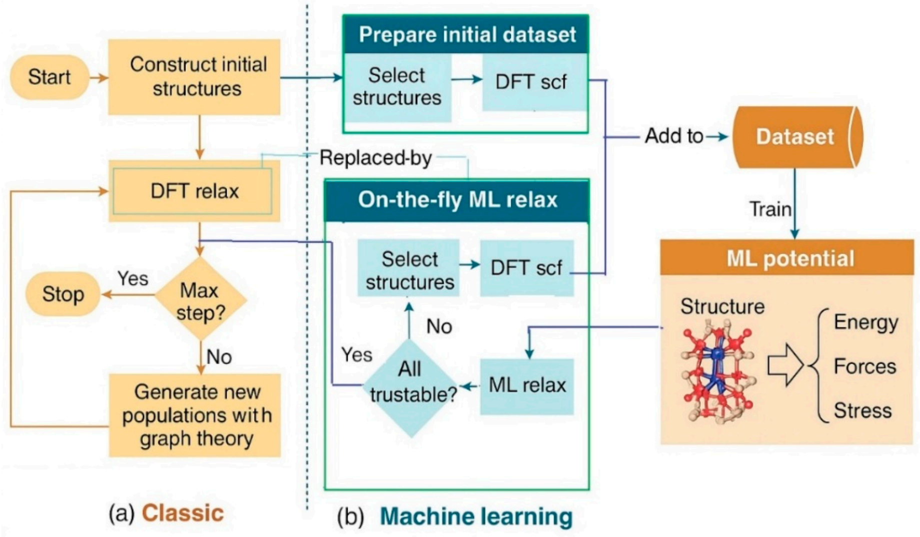

Figure 3. Overview of the ML-accelerated workflow for CSP/CPP combining generation [103].

Beyond inorganic systems, ML has also been applied to predict structure–property relationships in organic crystals and amorphous alloys. Shiraki et al. [105] used regression-based ML to relate molecular structures to crystal packing and dielectric constants in organic crystals, providing predictive insights for materials used in capacitors and ferroelectric memories. In a study on aluminum-based amorphous alloys, the authors of ref. [49] developed ML models for predicting fracture strength and Young's modulus using features such as atomic radius mismatch and valence electron concentration, finding that RF models yielded the lowest prediction errors and strongest generalization. In another effort, the authors of ref. [98] introduced the mechanical properties of crystalline materials repository (MPCR), a large dataset with over 100,000 DFT-computed entries. Using this database, GNN-based models outperformed traditional algorithms in predicting elastic moduli, highlighting the value of combining structural information with large-scale data. According to the same study, incorporating transformer-based atomic representations into the ALIGNN model increased predictive accuracy by 14–34%, depending on the dataset. These developments point to a strengthening link between data curation, structural representation, and predictive modeling in materials informatics.

In ref. [50], boosting-based classifiers were applied to predict the crystal structures of HEAs. These ensemble learning methods improve prediction accuracy by combining several simple models, usually decision trees, into a stronger one (which shows in Table 2). XGBoost and LightGBM are built for speed, working with large datasets and high-dimensional features. CatBoost is especially good at dealing with categorical data. AdaBoost and Gradient Boosting enhance model performance by iteratively correcting errors from previous models [106,107]. Together, these methods enable accurate and computationally efficient prediction of complex crystal structures [50].

Table 2. Comparative performance of different ML models using boosting techniques for CSP of HEAs, as reported by Dey et al. [50].

Classifier	Precision (%)	Accuracy (%)	F1 Score	Recall (%)
AdaBoost	68.93	67.92	0.681	67.56
LightGBM	90.07	90.07	0.899	90.08
XGBoost	86.12	85.45	0.859	85.86
Gradient Boosting	75.14	75.81	0.754	75.94
CatBoost	76.95	79.60	0.783	79.98

Traditionally, stability evaluation relied on DFT to compute formation energies, while structure search was performed using algorithms such as random structure searching (AIRSS) [108], SA [32], evolutionary algorithms in USPEX [109], and particle swarm optimization in CALYPSO [24]. Although highly accurate, these approaches are computationally demanding. ML-based methods provide efficient alternatives by approximating energy landscapes and actively exploring configuration space, reducing the need for extensive DFT calculations. Consequently, modern ML–CSP frameworks such as MAGUS, CrySPY, and Cond-CDVAE combine supervised prediction, generative modeling, and optimization into unified, self-adapting systems.

ML applications in CSP and CPP have evolved from isolated property-prediction tasks to comprehensive, data-driven discovery frameworks. Most studies still employ supervised learning due to the availability of labeled data linking atomic configurations with computed properties, but unsupervised and generative methods are gaining importance for exploring new chemical spaces. Across these approaches, ensemble models like XGBoost remain effective for interpretable and small datasets, while graph and diffusion models dominate when structural complexity increases. Integrated ML pipelines now represent the State-of-the-Art, providing autonomous and scalable routes for structure prediction that combine speed, accuracy, and interpretability. These advances mark a transition from traditional, DFT-heavy materials exploration toward intelligent, self-learning systems capable of guiding future materials discovery.

The most recent Cambridge Crystallographic Data Center (CCDC) CSP Blind Test, reported by Hunnisett et al. [110], provided a comprehensive evaluation of current structure generation methods across seven target systems of increasing complexity [1]. The study demonstrated that while modern CSP approaches can accurately reproduce experimental structures for relatively simple and semi-flexible molecules, they continue to face challenges with highly flexible, multicomponent pharmaceutical systems. Notably, the seventh blind test introduced a PXRD-assisted challenge for the first time, showing that CSP methods can reconstruct a crystal structure from low-quality powder diffraction data, a significant step toward practical industrial application. The authors also explored the prediction of cocrystal stoichiometry, finding that several methods correctly identified the most stable 1:1 and 2:1 compositions, although accurate energetic ranking of these forms remained difficult. Among the twenty-seven participating teams, several incorporated AI-enhanced workflows and ML interatomic potentials to accelerate structure generation and energy evaluation. These AI-CSP approaches achieved competitive accuracy while significantly reducing computational cost, showing the growing influence of data-driven strategies in CSP research.

From a computational science perspective, CSP/CPP can be viewed as a global optimization problem on a complex potential energy surface PES. This challenge can be divided into two main parts: first, how to evaluate the stability of the sampled structures, and second, how to efficiently explore the configuration space. Over the past few decades, advances in computing power and algorithm development have greatly improved our ability to address both issues. For evaluating stability, first-principles methods such as DFT are commonly used to calculate the energy or enthalpy of candidate structures and assess their thermodynamic stability. To explore the configuration space, a variety of algorithms (typically combined with local structure relaxations) have been introduced and implemented in specialized software. These methods have achieved numerous successes across physics, chemistry, and materials science. Examples include the AIRSS method [108], which employs random structure searching and successfully identified the high-pressure structure of SiH_4 [111]; SA, applied by Doll, Schön, and Jansen to determine the structures of lithium fluoride [112] and boron nitride [33]; the minima hopping method developed by Goedecker [113], which revealed the structure of silicon clusters [114]; the evolutionary algorithm used in USPEX [109,115], which discovered unexpected phases such as transparent dense sodium [116]; and the CALYPSO code [25], which applies particle swarm optimization to predict high-pressure phases of materials like lithium [117]. Conventional computational methods for CSP/CPP and their applications are summarized in Table 3 The methods such as USPEX [109], XtalOPT [118], AIRSS [108], CALYPSO [24], GASP [119], AGA [120], MUSE [121], IM2ODE [122], SYDSS [15], and MAISE [123] are all considered conventional CSP/CPP techniques. They primarily rely on classical optimization algorithms such as evolutionary algorithms, random searches, GA, particle swarm optimization, and differential evolution. These methods are based on physical and chemical principles and do not involve ML. Instead, they explore the potential energy surface using rule-based strategies to identify stable crystal structures. Despite being computationally intensive, they have demonstrated strong reliability and are widely used in CSP/CPP, especially for complex materials systems.

Table 3. Conventional computational methods for CSP/CPP and their applications.

Software	**Applications**	**Methods**
USPEX [109]	NaCl [124], W-B [86]	Evolutionary algorithm
XtalOPT [118]	NaH_n [125], H_2O [126]	Evolutionary algorithm
AIRSS [108]	SiH_4 [111], $NH_{3\pm x}$ [127]	Random search
CALYPSO [24]	Li [117], LaH_{10} [128], P	Particle swarm optimization
GASP [119]	Li-Be [129], Li-Si [130]	Evolutionary algorithm
AGA [120]	Zr-Co [131], MgO-SiO_2 [132]	Adaptive GA
MUSE [121]	IrB_4 [133], $NbSe_2$ [134]	Evolutionary algorithm
IM2ODE [122]	TiO_2 [135], 2D SiS [136]	Differential evolution
SYDSS [15]	H_2O-NaCl [15], Cl-F [137]	Random search
MAISE [123]	Fe-B [138], $NaSn_2$ [139]	Evolutionary algorithm

Beyond structure prediction, another study [99] investigated the use of ML for predicting the mechanical and thermodynamic properties of HEAs, including yield strength, Young's modulus, ductility, thermal conductivity, and melting temperature. They constructed a dataset by combining experimental values and calculated features based on elemental descriptors. To train and evaluate the models, they used several ML algorithms, including RF, SVR, and Gradient Boosting Regression (GBR). The authors reported that among the tested models, the RF algorithm achieved the highest accuracy in predicting most of the target properties. Feature importance analysis showed that factors such as atomic size difference, electronegativity difference, and valence electron concentration had a significant influence on the prediction outcomes. This study demonstrated that ML methods can effectively identify key factors governing HEA properties and provide reliable predictions to guide the design of new alloy compositions.

Table 4 shows ML-based approaches for CSP/CPP and representative use cases. All of these methods apply ML to either generate, optimize, or analyze crystal structures. They perform better than traditional CSP/CPP methods in speed and scalability, often finding new materials with less computational cost. These approaches show the growing role of artificial intelligence in materials discovery and design.

Table 4. ML-based approaches for CSP/CPP and representative use cases.

Software	Applications	Methods
FTCP [82]	$Au_2Sc_2O_3$ [82], $Y_2Zn_2As_2O_3$ [82]	VAE
GN-OA [23]	Tested on typical compounds [23]	Optimization algorithms and GNNs
MAGUS [103]	WN_6 [140], HeH_2O [141]	BO and GA
SCCOP [87]	B-C-N [142], $AgBiS_2$ [143]	Simulated annealing and GNNs
iMatGen [81]	V-O [81]	VAE
CrystalGAN [83]	Pd-Ni-H [83], Mg-Ti-H [83]	GAN
CCDCGAN [84]	$MoSe_2$ [84]	GAN
MatterGen [85]	V-Sr-O [85]	Diffusion model
UniMat [144]	Tested on typical compounds [144]	Diffusion model
DiffCSP [102]	Tested on typical compounds [102]	Diffusion model
LLaMA-2 [145]	Tested on typical compounds [145]	Large language-based model
GOFEE [146]	C_{24} [31], Carbon clusters [31]	Large language-based model
BEACON [29,30]	Cu_{15} [29], CuNi clusters [30]	BO
CrySPY [52]	Y_2Co_{17} [147], Al_2O_3 [147]	VAE

The authors of ref. Y [103] compares two types of MAGUS searches (which are computational runs using the MAGUS code to find the most stable crystal structures), including one using only DFT and another accelerated by ML for three materials: boron, lithium, and magnesium silicate ($MgSiO_3$). From Figure 4, it can be seen that the ML-assisted searches achieved the same correct structures with far fewer DFT calculations.

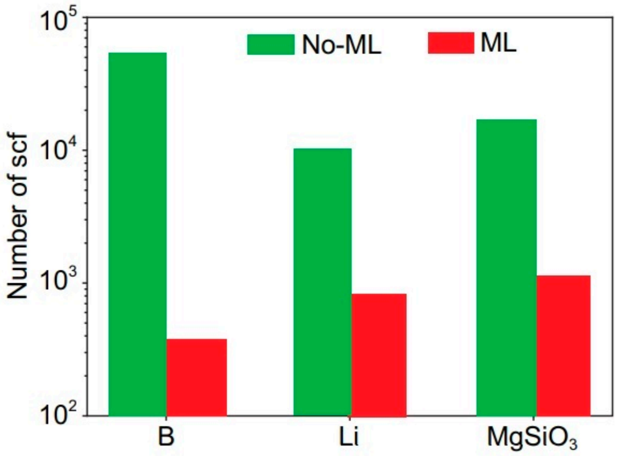

Figure 4. Comparison of computational efficiency between conventional DFT-based searches and those obtained by machine-learning methods [103].

Figure 5 compares the ML predictions with the experimentally measured phase behavior of $BaTiO_3$-based piezoelectrics reported by Xue et al. [91]. The authors synthesized the composition BZT-m50-n30 using a conventional powder processing method and determined its transition temperatures from dielectric permittivity (ε) versus temperature (T) measurements for nine compositions to construct the phase diagram. As shown in Figure 5A, the solid lines represent the phase boundaries predicted by the Bayesian linear-regression model, while the dots indicate the experimental data, showing excellent agreement between prediction and measurement (The solid lines show the mean phase boundaries, and the dashed lines mark the 95% confidence intervals.). Also, Figure 5B illustrates the updated Bayesian model after the new experimental data were added and the model retrained with 20 data points. The retrained model again identified BZT-m50-n30 as the optimal composition and showed markedly reduced uncertainty bands, indicating higher confidence and model refinement. The predicted phase-boundary displacement remained nearly constant (dx = 0.1458), corresponding to a 15% improvement over the previous best sample, BZT-m10-n20. Although not aimed at direct CSP, Ref. [91] clearly shows the power of combining ML with experimental synthesis and dielectric characterization to optimize ferroelectric crystal systems efficiently.

From the literature, it can be noticed that most ML applications in the fields of CSP/CPP mainly use supervised learning models. This preference is mainly related to the availability of labeled datasets in materials science, where each atomic configuration is associated with a corresponding physical or chemical property, such as formation energy, bandgap, conductivity, or elastic modulus. Supervised learning algorithms are well-suited to these scenarios, as they are designed to learn mappings from input features to known output targets. The suitability of supervised learning is further reinforced by the clearly defined objectives in CSP and CPP tasks, which typically involve the prediction of specific, quantifiable properties. Also, this clarity enables the use of well-established evaluation metrics (such as root mean squared error and mean absolute error) to assess and compare model performance. Moreover, supervised models often demonstrate higher predictive accuracy and generalizability when trained on sufficiently diverse and representative datasets. Algorithms such as ANNs, SVMs, and Gaussian process regression have shown considerable success in capturing complex structure–property relationships in crystalline materials.

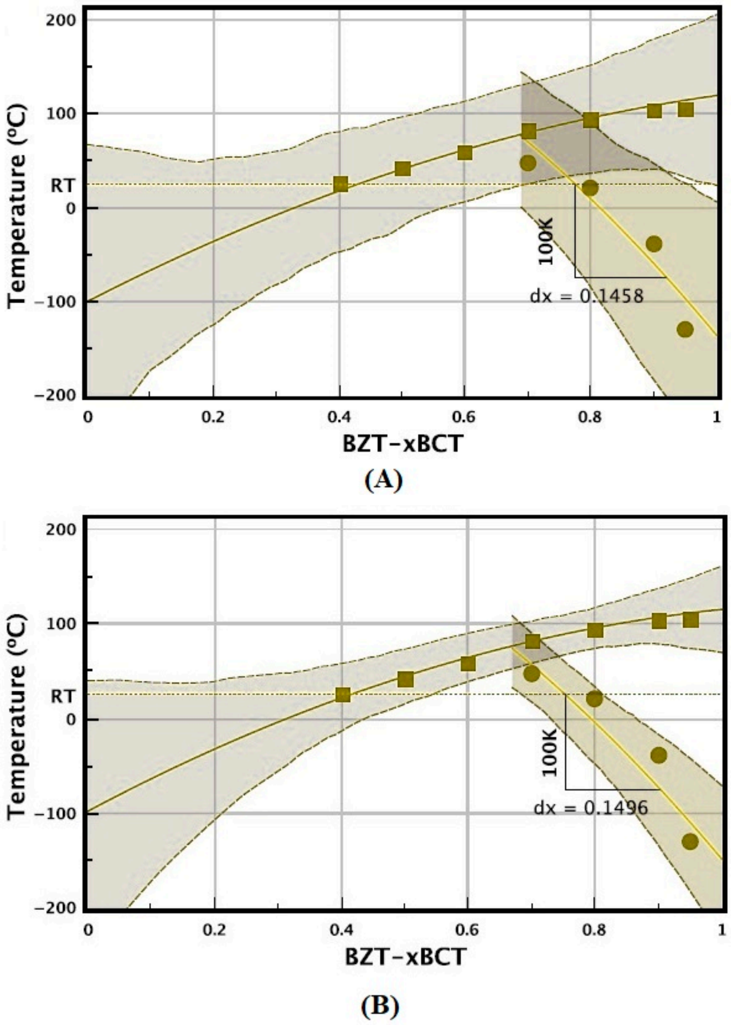

Figure 5. Comparison of (**A**) the predicted (with solid lines) with experimental (dots) phase diagram for BZT-m50-n30 using Bayesian linear regression. (**B**) Updated Bayesian linear regression method after augmenting the experimental BZT-m50-n30 data [91].

In comparison, unsupervised learning techniques, including clustering methods and dimensionality reduction approaches, are less frequently applied in this context. These methods are generally intended for exploratory data analysis, structural classification, or feature extraction in the absence of labeled outputs. While valuable for identifying patterns or correlations within materials datasets, they do not inherently support direct property prediction. Consequently, the structured nature of the data, the specificity of prediction goals, and the demand for quantitative accuracy collectively establish supervised learning as the dominant methodological approach in ML-based CSP and CPP research.

3.3. Improving the Accuracy of ML Potentials for CSP and CPP

ML interatomic potentials (MLIPs) have emerged as powerful tools for predicting crystal structures and their associated properties with near DFT accuracy but at a fraction of the computational cost. However, the reliability and generalization of these models strongly depend on the quality of training data, descriptor selection, and model architecture. Enhancing their accuracy requires a multi-faceted approach. First, the training dataset should be diverse and chemically representative, including not only equilibrium structures but also strained, defective, and metastable configurations. Second, incorporating physically informed descriptors (such as the smooth overlap of atomic positions (SOAP), symmetry functions, or moment tensor potentials) can improve sensitivity to subtle structural differences. Third, hybrid architectures combining neural networks with explicit physics constraints (such as energy conservation and symmetry) have shown promise in increasing both robustness and interpretability. Active learning methods, where the model iteratively queries the most informative data points (such as uncertain configurations), also help reduce bias and improve generalization. Moreover, recent efforts in uncertainty quantification and ensemble learning allow researchers to assess confidence in ML predictions, making these models more reliable for practical applications. Finally, advancing the accuracy of ML potential will enable accurate prediction of stable crystal phases and their properties, facilitating materials discovery in fields such as semiconductors, energy storage, and pharmaceuticals.

Recent studies have also shown that fixing data imbalance is important, especially when modeling complex materials such as HEAs. This comes alongside improvements in descriptors and model design. For instance, Hareharen et al. [96] proposed the use of SMOTE to overcome skewed class distributions in phase and crystal structure datasets. By generating synthetic data for underrepresented classes such as hexagonal close-packed phases, the model achieved significant improvements in accuracy and generalizability across various classifiers. This method demonstrates that data balancing and strategies can complement physical descriptor engineering, ensuring the ML model is exposed to sufficient diversity across structural motifs. Such techniques are especially relevant when experimental or simulated datasets are limited. When using active learning and uncertainty quantification, oversampling further enhances the robustness of MLIPs and their capability to predict not only common but also rare or metastable phases in high-dimensional compositional spaces.

4. ML Approaches for Identification and Modeling of Crystal Defects

4.1. Fundamentals and Development of ML-Based Interatomic Potentials for Crystal Defect Analyses

Crystalline materials inherently contain various types of defects that have profound effects on their physical and mechanical properties. These defects are traditionally categorized based on their dimensionality: point defects (vacancies, interstitials, substitutional atoms), line defects (dislocations), planar defects (grain boundaries, stacking faults), and volumetric defects (voids, cracks) [148]. The schematic illustration of a common crystal defect is shown in Figure 6. These defects arise from imperfections during crystal growth, external stresses, irradiation, or thermally based processes. Understanding and accurately modeling these defects is crucial because they control key material behaviors such as strength, ductility, conductivity, and diffusion mechanisms [149,150].

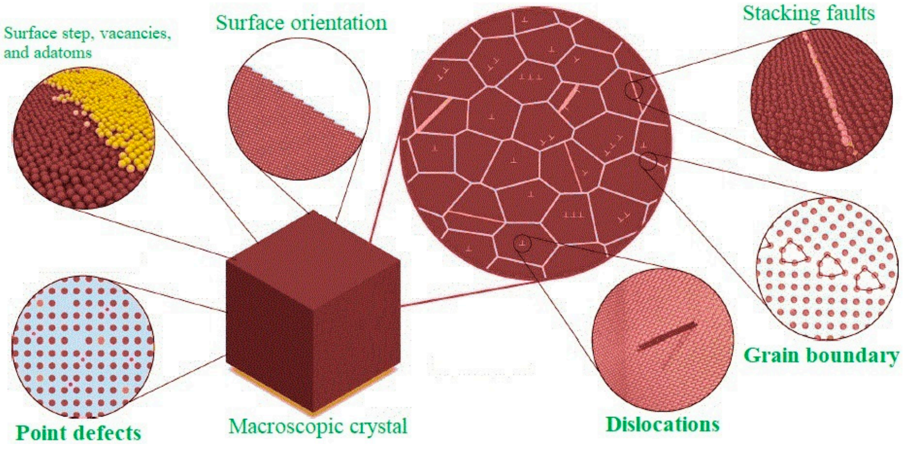

Figure 6. Schematic illustration of common crystal defects [148].

Historically, the detection and characterization of defects in atomistic simulations relied on heuristic-based visualization and classification techniques. Tools such as OVITO [151] and polyhedral template matching (PTM) [152] allow researchers to analyze atomistic configurations and identify defect structures. More advanced approaches, such as in silico microscopy [153], have emerged to extract complex microstructural features, especially when time evolution and kinetic behavior are involved. However, these conventional methods are often limited by manual intervention and lack generalization to diverse defect types. In this context, ML techniques offer significant advantages by enabling automated and data-based defect identification with higher scalability and accuracy [154–156].

A particularly transformative development in this field is the emergence of machine-learning interatomic potentials (MLIAPs). These ML-based potentials, including Neural Network Potentials (NNP) [157,158], Gaussian Approximation Potentials [159], Moment Tensor Potentials [160], Spectral Neighbor Analysis Potentials [161], and Physically Informed Neural Networks (PINN) [162], are trained on DFT data to accurately predict atomic interactions [13,163]. The training datasets typically include configurations such as bulk phases, surfaces, stacking faults, grain boundaries [164–166], and dislocation cores [167,168].

MLIAPs provide several advantages over conventional interatomic potentials [169,170]. They enable large-scale simulations of defect dynamics and kinetics, which are computationally infeasible with DFT alone. For example, MLIAPs have accurately reproduced screw dislocation core structures and mobility in BCC metals such as Fe and W [171], as well as grain boundary energies in systems such as Si and Al [166,172]. Furthermore, defect kinetics such as kink-pair nucleation and vacancy migration barriers can be computed with high fidelity using active learning methods [173,174]. The flexibility of these models allows for on-the-fly refinement of the training dataset using methods such as outlier detection and Bayesian active learning, significantly enhancing generalization to complex or unseen defect structures [156,175,176].

In conclusion, MLIAPs are transforming atomistic modeling of crystal defects by combining the accuracy of ab initio methods with the scalability needed for mesoscale simulations. The choice of ML architecture (NNP, GAP, PINN, etc.) depends on the material system and target properties, but training dataset design remains the critical factor determining model success. Inputs to these models include DFT energies, atomic forces, and symmetry-aware structural descriptors (such as SOAP, moment tensors), while outputs include energy landscapes [177], elastic constants [177–179], defect formation [180] and migration energies [181], and thermal transport properties [182]. Future directions include extending these models to even more complex systems such as solid–liquid interfaces [183], chemically complex alloys [184], and magnetically ordered materials [185].

4.2. Advanced Applications and Case Studies in ML-Based Crystal Defects Analyses

Recent studies increasingly demonstrate that ML plays a crucial role in modeling, analyzing, and predicting crystal defects across a wide range of materials and scales. This growing importance results from the need to overcome the inherent limitations of conventional simulation techniques, particularly when addressing complex defect structures or large-scale microstructural features.

One of the most promising developments in this area is the use of ML interatomic potentials (MLIAPs), which connect the gap between quantum-level accuracy and large-scale simulation efficiency. Freitas and Cao [186] provided a comprehensive overview of this emerging field, emphasizing that MLIAPs can replicate quantum mechanical accuracy while scaling efficiently to large atomic systems. They demonstrated that these potentials enable the study of extended defects (such as dislocations and grain boundaries) that are often inaccessible to standard ab initio methods. To further enhance the performance of MLIAPs in defective systems, they proposed methods such as active learning, γ-surface-inspired training configurations, and outlier detection strategies. Similarly, Tanaka et al. [187] applied deep-learning (DL) techniques (specifically a U-net-based convolutional neural network) for the analysis of dynamic transmission electron microscopy (TEM) video data in TWIP steels. Their model achieved pixel-wise classification of dislocations and stacking faults over time, allowing for quantitative tracking of defect evolution under strain. This study revealed grain size dependent shifts in deformation mechanisms and proved that ML based segmentation can obtain results not possible with manual analysis, both in resolution and in temporal coverage. In a more application-oriented study, Klunnikova et al. [156] trained ANNs to predict defect generation during sapphire crystal growth. By incorporating process parameters such as temperature gradients, power inputs, and crucible conditions, the model achieved over 94% accuracy in forecasting defect zones. Their work shows how ML can improve process control and quality assurance in industrial-scale crystal fabrication. Also, Hu [188] reviewed crystal defect energetics in metallic alloys and highlighted the limitations of DFT for large systems, especially for extended defects. He suggested using ML-based approximate models, which offer DFT-like precision while significantly reducing computational cost. These hybrid models are especially useful for predicting migration energies and stress responses in alloys under realistic conditions.

Further extending ML's ability to dynamic environments, Gutierrez [189] demonstrated the use of a Gaussian Approximation Potential (GAP) to model radiation-induced damage in crystalline molybdenum exposed to neutron bombardment. Combining molecular dynamics simulations with a defect-fingerprinting tool, the study tracked the formation of point defects (particularly Frenkel pairs) across a range of primary knock-on atom (PKA) energies. While both GAP and the Embedded Atom Method (EAM) produced similar defect counts, GAP more accurately reproduced atomic configurations such as crowdions, in agreement with DFT and experimental ion-mixing data. This validates the power of ML-based potential for accurate radiation damage prediction.

In addition to these studies, Alarfaj et al. [176] proposed a deep-learning framework for the automated detection of defective crystal structures in silicon nitride. Their multitask dense neural network simultaneously performed classification and regression, distinguishing among four categories: pristine, no-defect, random displacement defective, and vacancy defective crystals. Using a dataset of 16,000 crystal structures (30% of which contained significant defects) and optical images captured under cobalt-blue illumination, the model achieved an accuracy of 97% and a precision of 96%. This study illustrates ML's potential for defect identification and quality improvement in industrial ceramics.

Finally, Goryaeva et al. [171] developed ML-based interatomic potentials for modeling dislocations and defect clusters in body-centered cubic (bcc) iron and tungsten. Using both linear and a novel quadratic mapping approach, termed quadratic noise ML, their framework enabled large-scale simulations of defect cluster stability and vacancy formation free energies. The results demonstrated near DFT-level accuracy, underscoring ML's potential for exploring defect energetics in metallic systems.

Overall, these studies highlight how ML-driven approaches (from neural networks and deep-learning segmentation to advanced interatomic potentials) are transforming the field of crystal defect analysis. They not only accelerate simulations and improve predictive accuracy but also enable experimental validation and process optimization. As models become more data efficient and physically informed, ML is expected to play a significant role in designing defect-tolerant materials and optimizing microstructures for next-generation applications.

4.3. Improving the Accuracy of ML Potentials for Crystal Defects

Accurate modeling of crystal defects requires accurate representation of atomic interactions. Machine-learning interatomic potentials (MLIAPs) have emerged as powerful tools in this area, but ensuring their reliability for complex defect structures remains a challenge. This is mainly because many crystal defects, such as dislocations with kink-pairs or extended grain boundaries, involve atomic environments that are too complex or large to be directly included in training datasets with DFT calculations. One strategy to improve MLIAP accuracy is to include physically motivated surrogate configurations (atomic structures that resemble the environment around defects but are still small enough for DFT). For example, generalized stacking fault configurations have been shown to significantly improve the description of dislocation cores [190]. Also, another recent approach is to use ML algorithms to detect outliers (atomic environments in defect structures that are not represented in the training data). These tools help in evaluating the reliability of a model for a given defect configuration [191]. A more advanced method is active learning, where the potential is improved automatically during simulations. If the model faces unknown atomic environments, it runs new DFT calculations and also updates the training data immediately. This has been successfully applied, for example, in simulations of vacancy migration in aluminum [192]. Finally, it is important to balance model complexity and computational cost. While more complex models can improve accuracy, they may also increase simulation time unnecessarily. Methods, such as Pareto optimization [193,194], can help to choose models that give the best balance for defect applications. These methods, such as data design, outlier detection, active learning, and optimization can make MLIAPs more accurate and practical for simulating crystal defects.

5. Possible Future Research Opportunities

5.1. Future Directions in CSP Analysis with ML

Current studies on CSP, such as Liu et al. [48], Li et al. [47], and Chang et al. [46], still have some limits. They often cover only a small part of chemical space, depend strongly on DFT data, and sometimes produce results that are difficult to interpret physically. These issues can be reduced by using new types of algorithms and better links between simulation and experiment. Future work can use GNNs such as Crystal Graph Convolutional Network (CGCNN), MEGNet, or ALIGNN together with diffusion generative models such as DiffCSP or MatterGen. These models can design new crystal structures by learning atomic patterns directly from data. For example, a diffusion model trained on the Materials Project database can generate new, stable crystal structures that have not yet been tested. Active learning or BO can help select the most useful new data points for DFT validation. This can improve accuracy and efficiency and allow researchers to measure how much each new structure reduces overall uncertainty in the model. Physics Informed Neural Networks can also be used to include basic physical laws, such as energy conservation and symmetry, directly in the training process. This helps avoid unrealistic structures and gives more reliable energy–volume relationships.

Reinforcement Learning (RL) offers another direction for automatic structure exploration. An RL agent can learn to adjust atomic positions and lattice parameters to reduce formation energy, performing a type of self-learning CSP. The efficiency of this process can be measured by comparing how quickly the model reaches a stable structure compared to random or evolutionary search methods.

Experimentally, predictions can be tested by high-throughput synthesis and X-ray diffraction (XRD) or in situ neutron diffraction to check lattice constants and phase stability. Other methods, such as XPS or Raman spectroscopy, can measure bond lengths and bonding environments. Comparing predicted and experimental values of formation enthalpy, lattice strain, and thermal expansion will help measure model accuracy and physical validity.

5.2. Future Directions in CPP Analysis with ML

ML models for property prediction, such as Zhou et al. [99], MEGNet [8], ALIGNN [68], and CHGNet [54], reach good accuracy but usually focus on one property and do not capture how different properties are related. This can be improved by combining models, adding uncertainty measures, and using experimental validation. Future work can apply multitask learning models such as shared-layer GNNs or multitask transformers like the Crystal Transformer to predict several properties at once (for example, formation energy, band gap, dielectric constant, and bulk modulus). This makes it possible to find connections between properties, such as how atomic distortions change both mechanical strength and electronic behavior. Correlation maps or feature analysis can then be used to identify the most important factors.

Uncertainty quantified models, including Bayesian neural networks or Monte Carlo dropout GNNs, can estimate confidence intervals for each result. This helps identify where the model is less reliable and where new data are needed. Combining data-driven and physics-based features (such as electronegativity, ionic radius, and bond topology) makes results easier to interpret. Algorithms such as Gradient Boosting Decision Trees and RF can be used with deep features to improve generalization and prevent overfitting.

Experimental techniques can help check and improve these models. For example, ultraviolet photoelectron spectroscopy and optical absorption tests can measure band gaps. Nanoindentation and Brillouin spectroscopy can measure elastic moduli and hardness. Hall effect and impedance spectroscopy can measure carrier mobility and dielectric constant. Using these experimental values to retrain the models can improve their accuracy and make them more reliable for real materials. The benefit can be measured by comparing prediction errors, such as RMSE, before and after retraining.

Reinforcement learning or BO can also be used for inverse design, where the goal is to find structures or compositions with target properties. For instance, an RL model can adjust the elements in a perovskite to reach a higher dielectric constant while keeping the structure stable. This approach improves prediction accuracy and helps measure how efficiently AI can search for optimal materials.

5.3. Future Directions in Crystal Defect Modeling and Analysis with ML

ML for crystal defect modeling, as shown by Freitas and Cao [186], Tanaka et al. [187], and Gutierrez [189], has made good progress but still faces limits due to small datasets, static geometries, and weak transferability between materials. These problems can be improved with better algorithms and integration with experiments.

To describe atomic-scale defects more accurately, machine-learned interatomic potentials (MLIPs) such as GAP, Moment Tensor Potential (MTP), and NNP can be combined with active learning. This allows the model to learn dynamically from new configurations during molecular dynamics simulations. It can then calculate defect formation energy, migration energy, and diffusion coefficients with near DFT accuracy but at a much lower cost. Graph-based models like CHGNet, NequIP, and MACE can describe both local and long-range atomic interactions around defects. They can produce detailed energy maps for dislocation motion or vacancy diffusion and allow precise measurement of defect mobility, activation energy, and clustering probability. For image-based analysis, CNNs, U-Net, or Vision Transformers can be used to segment and classify TEM or SEM images. Spatio-temporal deep-learning models such as ConvLSTM or TimeSformer can track how dislocations or voids evolve during deformation, providing measurable values like strain-rate sensitivity, defect growth speed, and nucleation rate. Automated image analysis can quantify these features with high precision, far better than manual inspection.

Computational predictions should be linked to in situ experimental methods such as 4D-STEM, atom probe tomography (APT), or X-ray topography to verify defect structures and motion. This integration can reveal how local strain or temperature affects defect stability and can improve understanding of how microscopic defects influence macroscopic mechanical properties. Digital twin systems, as suggested by Klunnikova et al. [156] and Gutierrez [189], can also be used for crystal growth or radiation studies. These systems can use data from sensors (such as temperature, stress, and growth rate) to predict defect formation in real time. Outputs such as dislocation density, grain boundary energy, and vacancy concentration can then be measured during experiments and used to continuously update and validate the models.

6. Conclusions

The application of ML to analyze the crystal structure, property, and defect prediction has fundamentally changed computational materials science. ML models have overcome many of the drawbacks of conventional techniques by making it possible to predict structural configurations and material properties more quickly and accurately. In this paper, first, the importance of crystal structure/property prediction in material science is discussed. Also, traditional CSP/CPP analyses and their limitations are mentioned. Then, the advantages of ML in CSP/CPP investigations are explained. Different ML algorithms for CSP/CPP analyses have also been introduced. In addition, a brief background and recent trends in ML for CSP/CPP analyses are presented. Some related recent studies are reviewed. Furthermore, suggestions for improving the accuracy of ML potential for CSP and CPP are presented. Additionally, ML approaches for the identification and modeling of crystal defects are introduced, and some recent related studies are reviewed. Finally, suggestions for improving the accuracy of ML potential for crystal defects are presented. The incorporation of ML into the study of the crystal structure, property, and defect prediction can open new ways for understanding or the discovery of novel materials or material behavior.

Author Contributions: Conceptualization, M.S. and A.P.; methodology, M.S. and G.J.; formal analysis, M.S. and G.J.; investigation, M.S.; writing—original draft preparation, M.S.; writing—review and editing, A.P. and G.J.; visualization, M.S.; supervision, A.P. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: No new data were created or analyzed in this study. Data sharing is not applicable to this article.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Lu, M.; Rao, S.; Yue, H.; Han, J.; Wang, J. Recent Advances in the Application of Machine Learning to Crystal Behavior and Crystallization Process Control. *Cryst. Growth Des.* **2024**, *24*, 5374–5396. [CrossRef]
2. Oganov, A.R.; Pickard, C.J.; Zhu, Q.; Needs, R.J. Structure prediction drives materials discovery. *Nat. Rev. Mater.* **2019**, *4*, 331–348. [CrossRef]
3. Oganov, A.R.; Lyakhov, A.O.; Valle, M. How Evolutionary Crystal Structure Prediction Works—And Why. *Acc. Chem. Res.* **2011**, *44*, 227–237. [CrossRef]
4. Needs, R.J.; Pickard, C.J. Perspective: Role of structure prediction in materials discovery and design. *APL Mater.* **2016**, *4*, 053210. [CrossRef]

5. Rosen, A.S.; Fung, V.; Huck, P.; O'Donnell, C.T.; Horton, M.K.; Truhlar, D.G.; Persson, K.A.; Notestein, J.M.; Snurr, R.Q. High-throughput predictions of metal–organic framework electronic properties: Theoretical challenges, graph neural networks, and data exploration. *Npj Comput. Mater.* **2022**, *8*, 112. [CrossRef]
6. Kruglov, I.A.; Yanilkin, A.V.; Propad, Y.; Mazitov, A.B.; Rachitskii, P.; Oganov, A.R. Crystal structure prediction at finite temperatures. *Npj Comput. Mater.* **2023**, *9*, 197. [CrossRef]
7. Wu, X.; Kang, F.; Duan, W.; Li, J. Density functional theory calculations: A powerful tool to simulate and design high-performance energy storage and conversion materials. *Prog. Nat. Sci. Mater. Int.* **2019**, *29*, 247–255. [CrossRef]
8. Chen, C.; Ye, W.; Zuo, Y.; Zheng, C.; Ong, S.P. Graph Networks as a Universal Machine Learning Framework for Molecules and Crystals. *Chem. Mater.* **2019**, *31*, 3564–3572. [CrossRef]
9. Xie, T.; Grossman, J.C. Crystal Graph Convolutional Neural Networks for an Accurate and Interpretable Prediction of Material Properties. *Phys. Rev. Lett.* **2018**, *120*, 145301. [CrossRef]
10. Jain, A.; Ong, S.P.; Hautier, G.; Chen, W.; Richards, W.D.; Dacek, S.; Cholia, S.; Gunter, D.; Skinner, D.; Ceder, G.; et al. Commentary: The Materials Project: A materials genome approach to accelerating materials innovation. *APL Mater.* **2013**, *1*, 011002. [CrossRef]
11. Butler, K.T.; Davies, D.W.; Cartwright, H.; Isayev, O.; Walsh, A. Machine learning for molecular and materials science. *Nature* **2018**, *559*, 547–555. [CrossRef]
12. Xu, P.; Wang, H.; Ren, L.; Tu, B.; Wang, W.; Fu, Z. Theoretical study on composition-dependent properties of ZnO·nAl2O3 spinels. Part I: Optical and dielectric. *J. Am. Ceram. Soc.* **2021**, *104*, 5099–5109. [CrossRef]
13. Nyangiwe, N.N. Applications of density functional theory and machine learning in nanomaterials: A review. *Next Mater.* **2025**, *8*, 100683. [CrossRef]
14. Yang, Z.; Liu, X.; Zhang, X.; Huang, P.; Novoselov, K.S.; Shen, L. Modeling crystal defects using defect informed neural networks. *Npj Comput. Mater.* **2025**, *11*, 229. [CrossRef]
15. Domingos, R.; Shaik, K.M.; Militzer, B. Prediction of novel high-pressure H_2O-NaCl and carbon oxide compounds with a symmetry-driven structure search algorithm. *Phys. Rev. B* **2018**, *98*, 174107. [CrossRef]
16. Lu, Z.; Zhu, B.; Shires, B.W.B.; Scanlon, D.O.; Pickard, C.J. Ab initio random structure searching for battery cathode materials. *J. Chem. Phys.* **2021**, *154*, 174111. [CrossRef]
17. Zunger, A.; Wei, S.H.; Ferreira, L.G.; Bernard, J.E. Special quasirandom structures. *Phys. Rev. Lett.* **1990**, *65*, 353–356. [CrossRef]
18. Pickard, C.J.; Needs, R.J. Structure of phase III of solid hydrogen. *Nat. Phys.* **2007**, *3*, 473–476. [CrossRef]
19. Pickard, C.J.; Needs, R.J. High-Pressure Phases of Nitrogen. *Phys. Rev. Lett.* **2009**, *102*, 125702. [CrossRef]
20. Pickard, C.J.; Needs, R.J. Dense Low-Coordination Phases of Lithium. *Phys. Rev. Lett.* **2009**, *102*, 146401. [CrossRef]
21. Zhang, X.; Wang, H.; Hickel, T.; Rogal, J.; Li, Y.; Neugebauer, J. Mechanism of collective interstitial ordering in Fe–C alloys. *Nat. Mater.* **2020**, *19*, 849–854. [CrossRef] [PubMed]
22. Falls, Z.; Avery, P.; Wang, X.; Hilleke, K.P.; Zurek, E. The XtalOpt Evolutionary Algorithm for Crystal Structure Prediction. *J. Phys. Chem. C* **2021**, *125*, 1601–1620. [CrossRef]
23. Cheng, G.; Gong, X.-G.; Yin, W.-J. Crystal structure prediction by combining graph network and optimization algorithm. *Nat. Commun.* **2022**, *13*, 1492. [CrossRef] [PubMed]
24. Wang, Y.; Lv, J.; Zhu, L.; Ma, Y. CALYPSO: A method for crystal structure prediction. *Comput. Phys. Commun.* **2012**, *183*, 2063–2070. [CrossRef]
25. Wang, Y.; Lv, J.; Zhu, L.; Ma, Y. Crystal structure prediction via particle-swarm optimization. *Phys. Rev. B* **2010**, *82*, 094116. [CrossRef]
26. Lv, J.; Xu, M.; Lin, S.; Shao, X.; Zhang, X.; Liu, Y.; Wang, Y.; Chen, Z.; Ma, Y. Direct-gap semiconducting tri-layer silicene with 29% photovoltaic efficiency. *Nano Energy* **2018**, *51*, 489–495. [CrossRef]
27. Oganov, A.R.; Glass, C.W. Crystal structure prediction using ab initio evolutionary techniques: Principles and applications. *J. Chem. Phys.* **2006**, *124*, 244704. [CrossRef]
28. Wang, J.; Hanzawa, K.; Hiramatsu, H.; Kim, J.; Umezawa, N.; Iwanaka, K.; Tada, T.; Hosono, H. Exploration of Stable Strontium Phosphide-Based Electrides: Theoretical Structure Prediction and Experimental Validation. *J. Am. Chem. Soc.* **2017**, *139*, 15668–15680. [CrossRef]
29. Kaappa, S.; del Río, E.G.; Jacobsen, K.W. Global optimization of atomic structures with gradient-enhanced Gaussian process regression. *Phys. Rev. B* **2021**, *103*, 174114. [CrossRef]
30. Kaappa, S.; Larsen, C.; Jacobsen, K.W. Atomic Structure Optimization with Machine-Learning Enabled Interpolation between Chemical Elements. *Phys. Rev. Lett.* **2021**, *127*, 166001. [CrossRef]
31. Bisbo, M.K.; Hammer, B. Global optimization of atomic structure enhanced by machine learning. *Phys. Rev. B* **2022**, *105*, 245404. [CrossRef]
32. Wille, L.T. Searching potential energy surfaces by simulated annealing. *Nature* **1987**, *325*, 374. [CrossRef]
33. Doll, K.; Schön, J.C.; Jansen, M. Structure prediction based on ab initio simulated annealing for boron nitride. *Phys. Rev. B* **2008**, *78*, 144110. [CrossRef]

34. Timmermann, J.; Lee, Y.; Staacke, C.G.; Margraf, J.T.; Scheurer, C.; Reuter, K. Data-efficient iterative training of Gaussian approximation potentials: Application to surface structure determination of rutile IrO_2 and RuO_2. *J. Chem. Phys.* **2021**, *155*, 244107. [CrossRef]
35. Fischer, C.C.; Tibbetts, K.J.; Morgan, D.; Ceder, G. Predicting crystal structure by merging data mining with quantum mechanics. *Nat. Mater.* **2006**, *5*, 641–646. [CrossRef]
36. Gubernatis, J.E.; Lookman, T. Machine learning in materials design and discovery: Examples from the present and suggestions for the future. *Phys. Rev. Mater.* **2018**, *2*, 120301. [CrossRef]
37. Curtarolo, S.; Setyawan, W.; Hart, G.L.W.; Jahnatek, M.; Chepulskii, R.V.; Taylor, R.H.; Wang, S.; Xue, J.; Yang, K.; Levy, O.; et al. AFLOW: An automatic framework for high-throughput materials discovery. *Comput. Mater. Sci.* **2012**, *58*, 218–226. [CrossRef]
38. Aslam, A.; Saeed, S.; Kanwal, S.; Tchier, F. Investigating hexagonal closed packed crystal lattice through QSPR modeling via linear regression analysis and Topsis. *Phys. Scr.* **2024**, *99*, 025201. [CrossRef]
39. Yin, Y.; Wang, A.; Sun, Z.; Xin, C.; Jin, G. Machine learning regression model for predicting the band gap of multi-elements nonlinear optical crystals. *Comput. Mater. Sci.* **2024**, *242*, 113109. [CrossRef]
40. Kongsompong, S.; E-kobon, T.; Chumnanpuen, P. K-Nearest Neighbor and Random Forest-Based Prediction of Putative Tyrosinase Inhibitory Peptides of Abalone Haliotis diversicolor. *Molecules* **2021**, *26*, 3671. [CrossRef]
41. Leitherer, A.; Ziletti, A.; Ghiringhelli, L.M. Robust recognition and exploratory analysis of crystal structures via Bayesian deep learning. *Nat. Commun.* **2021**, *12*, 6234. [CrossRef]
42. Mobarak, M.H.; Mimona, M.A.; Islam, M.A.; Hossain, N.; Zohura, F.T.; Imtiaz, I.; Rimon, M.I.H. Scope of machine learning in materials research—A review. *Appl. Surf. Sci. Adv.* **2023**, *18*, 100523. [CrossRef]
43. Horak, J.; Vrbka, J.; Suler, P. Support Vector Machine Methods and Artificial Neural Networks Used for the Development of Bankruptcy Prediction Models and their Comparison. *J. Risk Financ. Manag.* **2020**, *13*, 60. [CrossRef]
44. Kutsukake, K. Review of machine learning applications for crystal growth research. *J. Cryst. Growth* **2024**, *630*, 127598. [CrossRef]
45. Xiouras, C.; Cameli, F.; Quilló, G.L.; Kavousanakis, M.E.; Vlachos, D.G.; Stefanidis, G.D. Applications of Artificial Intelligence and Machine Learning Algorithms to Crystallization. *Chem. Rev.* **2022**, *122*, 13006–13042. [CrossRef] [PubMed]
46. Chang, L.; Tamaki, H.; Yokoyama, T.; Wakasugi, K.; Yotsuhashi, S.; Kusaba, M.; Oganov, A.R.; Yoshida, R. Shotgun crystal structure prediction using machine-learned formation energies. *Npj Comput. Mater.* **2024**, *10*, 298. [CrossRef]
47. Li, L.; Shen, J.; Xiao, Q.; He, C.; Zheng, J.; Chu, C.; Chen, C. Stable crystal structure prediction using machine learning-based formation energy and empirical potential function. *Chin. Chem. Lett.* **2025**, *36*, 110421. [CrossRef]
48. Liu, J.; Wang, A.; Gao, P.; Bai, R.; Liu, J.; Du, B.; Fang, C. Machine learning-based crystal structure prediction for high-entropy oxide ceramics. *J. Am. Ceram. Soc.* **2024**, *107*, 1361–1371. [CrossRef]
49. Kusaba, M.; Liu, C.; Yoshida, R. Crystal structure prediction with machine learning-based element substitution. *Comput. Mater. Sci.* **2022**, *211*, 111496. [CrossRef]
50. Dey, D.; Das, S.; Pal, A.; Dey, S.; Raul, C.K.; Mandal, P.; Chatterjee, A.; Chatterjee, S.; Ghosh, M. Improved machine learning framework for prediction of phases and crystal structures of high entropy alloys. *J. Alloys Metall. Syst.* **2025**, *9*, 100144. [CrossRef]
51. Wang, Z.; Chen, Z.; Yuan, Y.; Wang, Y. CrySPAI: A New Crystal Structure Prediction Software Based on Artificial Intelligence. *Inventions* **2025**, *10*, 26. [CrossRef]
52. Yamashita, T.; Shinichi, K.; Nobuya, S.; Hiori, K.; Kei, T.; Hikaru, S.; Takumi, S.; Futoshi, U.; Koji, T.; Takashi, M.; et al. CrySPY: A crystal structure prediction tool accelerated by machine learning. *Sci. Technol. Adv. Mater. Methods* **2021**, *1*, 87–97. [CrossRef]
53. Jarin, S.; Yuan, Y.; Zhang, M.; Hu, M.; Rana, M.; Wang, S.; Knibbe, R. Predicting the Crystal Structure and Lattice Parameters of the Perovskite Materials via Different Machine Learning Models Based on Basic Atom Properties. *Crystals* **2022**, *12*, 1570. [CrossRef]
54. Deng, B.; Zhong, P.; Jun, K.; Riebesell, J.; Han, K.; Bartel, C.J.; Ceder, G. CHGNet as a pretrained universal neural network potential for charge-informed atomistic modelling. *Nat. Mach. Intell.* **2023**, *5*, 1031–1041. [CrossRef]
55. Merchant, A.; Batzner, S.; Schoenholz, S.S.; Aykol, M.; Cheon, G.; Cubuk, E.D. Scaling deep learning for materials discovery. *Nature* **2023**, *624*, 80–85. [CrossRef]
56. Deringer, V.L.; Bartók, A.P.; Bernstein, N.; Wilkins, D.M.; Ceriotti, M.; Csányi, G. Gaussian Process Regression for Materials and Molecules. *Chem. Rev.* **2021**, *121*, 10073–10141. [CrossRef]
57. Damewood, J.; Karaguesian, J.; Lunger, J.R.; Tan, A.R.; Xie, M.; Peng, J.; Gómez-Bombarelli, R. Representations of Materials for Machine Learning. *Annu. Rev. Mater. Res.* **2023**, *53*, 399–426. [CrossRef]
58. Yeo, B.C.; Nam, H.; Nam, H.; Kim, M.-C.; Lee, H.W.; Kim, S.-C.; Won, S.O.; Kim, D.; Lee, K.-Y.; Lee, S.Y.; et al. High-throughput computational-experimental screening protocol for the discovery of bimetallic catalysts. *Npj Comput. Mater.* **2021**, *7*, 137. [CrossRef]
59. Rittiruam, M.; Noppakhun, J.; Setasuban, S.; Aumnongpho, N.; Sriwattana, A.; Boonchuay, S.; Saelee, T.; Wangphon, C.; Ektarawong, A.; Chammingkwan, P.; et al. High-throughput materials screening algorithm based on first-principles density functional theory and artificial neural network for high-entropy alloys. *Sci. Rep.* **2022**, *12*, 16653. [CrossRef]

60. Szymanski, N.J.; Rendy, B.; Fei, Y.; Kumar, R.E.; He, T.; Milsted, D.; McDermott, M.J.; Gallant, M.; Cubuk, E.D.; Merchant, A.; et al. An autonomous laboratory for the accelerated synthesis of novel materials. *Nature* **2023**, *624*, 86–91. [CrossRef]
61. Chmiela, S.; Vassilev-Galindo, V.; Unke, O.T.; Kabylda, A.; Sauceda, H.E.; Tkatchenko, A.; Müller, K.-R. Accurate global machine learning force fields for molecules with hundreds of atoms. *Sci. Adv.* **2023**, *9*, eadf0873. [CrossRef]
62. Sauceda, H.E.; Gálvez-González, L.E.; Chmiela, S.; Paz-Borbón, L.O.; Müller, K.-R.; Tkatchenko, A. BIGDML—Towards accurate quantum machine learning force fields for materials. *Nat. Commun.* **2022**, *13*, 3733. [CrossRef]
63. Noh, J.; Gu, G.H.; Kim, S.; Jung, Y. Machine-enabled inverse design of inorganic solid materials: Promises and challenges. *Chem. Sci.* **2020**, *11*, 4871–4881. [CrossRef]
64. Behler, J. Constructing high-dimensional neural network potentials: A tutorial review. *Int. J. Quantum Chem.* **2015**, *115*, 1032–1050. [CrossRef]
65. Hoffmann, J.; Maestrati, L.; Sawada, Y.; Tang, J.; Sellier, J.M.; Bengio, Y. Data-driven approach to encoding and decoding 3-d crystal structures. *arXiv* **2019**, arXiv:1909.00949. [CrossRef]
66. Schütt, K.T.; Sauceda, H.E.; Kindermans, P.J.; Tkatchenko, A.; Müller, K.R. SchNet—A deep learning architecture for molecules and materials. *J. Chem. Phys.* **2018**, *148*, 241722. [CrossRef]
67. Gilmer, J.; Schoenholz, S.S.; Riley, P.F.; Vinyals, O.; Dahl, G.E. Neural message passing for quantum chemistry. In Proceedings of the International Conference on Machine Learning, Sydney, Australia, 6–11 August 2017; pp. 1263–1272.
68. Choudhary, K.; DeCost, B. Atomistic Line Graph Neural Network for improved materials property predictions. *Npj Comput. Mater.* **2021**, *7*, 185. [CrossRef]
69. Behler, J. Atom-centered symmetry functions for constructing high-dimensional neural network potentials. *J. Chem. Phys.* **2011**, *134*, 074106. [CrossRef]
70. Bartók, A.P.; De, S.; Poelking, C.; Bernstein, N.; Kermode, J.R.; Csányi, G.; Ceriotti, M. Machine learning unifies the modeling of materials and molecules. *Sci. Adv.* **2017**, *3*, e1701816. [CrossRef]
71. Batzner, S.; Musaelian, A.; Sun, L.; Geiger, M.; Mailoa, J.P.; Kornbluth, M.; Molinari, N.; Smidt, T.E.; Kozinsky, B. E(3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials. *Nat. Commun.* **2022**, *13*, 2453. [CrossRef]
72. Batatia, I.; Kovacs, D.P.; Simm, G.; Ortner, C.; Csányi, G. MACE: Higher order equivariant message passing neural networks for fast and accurate force fields. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 11423–11436.
73. Gale, J.D.; LeBlanc, L.M.; Spackman, P.R.; Silvestri, A.; Raiteri, P. A Universal Force Field for Materials, Periodic GFN-FF: Implementation and Examination. *J. Chem. Theory Comput.* **2021**, *17*, 7827–7849. [CrossRef]
74. Cole, D.J.; Horton, J.T.; Lauren, N.; Kurdekar, V. The Future of Force Fields in Computer-Aided Drug Design. *Future Med. Chem.* **2019**, *11*, 2359–2363. [CrossRef] [PubMed]
75. Robustelli, P.; Piana, S.; Shaw, D.E. Developing a molecular dynamics force field for both folded and disordered protein states. *Proc. Natl. Acad. Sci. USA* **2018**, *115*, E4758–E4766. [CrossRef] [PubMed]
76. Kang, S.; Jeong, W.; Hong, C.; Hwang, S.; Yoon, Y.; Han, S. Accelerated identification of equilibrium structures of multicomponent inorganic crystals using machine learning potentials. *Npj Comput. Mater.* **2022**, *8*, 108. [CrossRef]
77. Tong, Q.; Xue, L.; Lv, J.; Wang, Y.; Ma, Y. Accelerating CALYPSO structure prediction by data-driven learning of a potential energy surface. *Faraday Discuss.* **2018**, *211*, 31–43. [CrossRef]
78. Cheng, Y.; Wang, D.; Zhou, P.; Zhang, T. Model Compression and Acceleration for Deep Neural Networks: The Principles, Progress, and Challenges. *IEEE Signal Process. Mag.* **2018**, *35*, 126–136. [CrossRef]
79. Zhang, Q.-s.; Zhu, S.-c. Visual interpretability for deep learning: A survey. *Front. Inf. Technol. Electron. Eng.* **2018**, *19*, 27–39. [CrossRef]
80. Zhuang, F.; Qi, Z.; Duan, K.; Xi, D.; Zhu, Y.; Zhu, H.; Xiong, H.; He, Q. A Comprehensive Survey on Transfer Learning. *Proc. IEEE* **2021**, *109*, 43–76. [CrossRef]
81. Noh, J.; Kim, J.; Stein, H.S.; Sanchez-Lengeling, B.; Gregoire, J.M.; Aspuru-Guzik, A.; Jung, Y. Inverse Design of Solid-State Materials via a Continuous Representation. *Matter* **2019**, *1*, 1370–1384. [CrossRef]
82. Ren, Z.; Tian, S.I.P.; Noh, J.; Oviedo, F.; Xing, G.; Li, J.; Liang, Q.; Zhu, R.; Aberle, A.G.; Sun, S.; et al. An invertible crystallographic representation for general inverse design of inorganic crystals with targeted properties. *Matter* **2022**, *5*, 314–335. [CrossRef]
83. Nouira, A.; Sokolovska, N.; Crivello, J.-C. Crystalgan: Learning to discover crystallographic structures with generative adversarial networks. *arXiv* **2018**, arXiv:1810.11203.
84. Fung, V.; Zhang, J.; Hu, G.; Ganesh, P.; Sumpter, B.G. Inverse design of two-dimensional materials with invertible neural networks. *Npj Comput. Mater.* **2021**, *7*, 200. [CrossRef]
85. Zeni, C.; Pinsler, R.; Zügner, D.; Fowler, A.; Horton, M.; Fu, X.; Shysheya, S.; Crabbé, J.; Sun, L.; Smith, J. Mattergen: A generative model for inorganic materials design. *arXiv* **2023**, arXiv:2312.03687.
86. Zhao, C.; Duan, Y.; Gao, J.; Liu, W.; Dong, H.; Dong, H.; Zhang, D.; Oganov, A.R. Unexpected stable phases of tungsten borides. *Phys. Chem. Chem. Phys.* **2018**, *20*, 24665–24670. [CrossRef]

87. Li, C.-N.; Liang, H.-P.; Zhang, X.; Lin, Z.; Wei, S.-H. Graph deep learning accelerated efficient crystal structure search and feature extraction. *Npj Comput. Mater.* **2023**, *9*, 176. [CrossRef]
88. Podryabinkin, E.V.; Tikhonov, E.V.; Shapeev, A.V.; Oganov, A.R. Accelerating crystal structure prediction by machine-learning interatomic potentials with active learning. *Phys. Rev. B* **2019**, *99*, 064114. [CrossRef]
89. Tong, Q.; Gao, P.; Liu, H.; Xie, Y.; Lv, J.; Wang, Y.; Zhao, J. Combining Machine Learning Potential and Structure Prediction for Accelerated Materials Design and Discovery. *J. Phys. Chem. Lett.* **2020**, *11*, 8710–8720. [CrossRef]
90. Mansouri Tehrani, A.; Oliynyk, A.O.; Parry, M.; Rizvi, Z.; Couper, S.; Lin, F.; Miyagi, L.; Sparks, T.D.; Brgoch, J. Machine Learning Directed Search for Ultraincompressible, Superhard Materials. *J. Am. Chem. Soc.* **2018**, *140*, 9844–9853. [CrossRef]
91. Xue, D.; Balachandran, P.V.; Yuan, R.; Hu, T.; Qian, X.; Dougherty, E.R.; Lookman, T. Accelerated search for BaTiO3-based piezoelectrics with vertical morphotropic phase boundary using Bayesian learning. *Proc. Natl. Acad. Sci. USA* **2016**, *113*, 13301–13306. [CrossRef]
92. Padhy, S.P.; Chaudhary, V.; Lim, Y.-F.; Zhu, R.; Thway, M.; Hippalgaonkar, K.; Ramanujan, R.V. Experimentally validated inverse design of multi-property Fe-Co-Ni alloys. *iScience* **2024**, *27*, 109723. [CrossRef]
93. Balachandran, P.V.; Kowalski, B.; Sehirlioglu, A.; Lookman, T. Experimental search for high-temperature ferroelectric perovskites guided by two-step machine learning. *Nat. Commun.* **2018**, *9*, 1668. [CrossRef]
94. Ramprasad, R.; Batra, R.; Pilania, G.; Mannodi-Kanakkithodi, A.; Kim, C. Machine learning in materials informatics: Recent applications and prospects. *Npj Comput. Mater.* **2017**, *3*, 54. [CrossRef]
95. Syed, T.A.; Ansari, K.B.; Banerjee, A.; Wood, D.A.; Khan, M.S.; Al Mesfer, M.K. Machine-learning predictions of caffeine co-crystal formation accompanying experimental and molecular validations. *J. Food Process Eng.* **2023**, *46*, e14230. [CrossRef]
96. Hareharen, K.; Panneerselvam, T.; Raj Mohan, R. Improving the performance of machine learning model predicting phase and crystal structure of high entropy alloys by the synthetic minority oversampling technique. *J. Alloys Compd.* **2024**, *991*, 174494. [CrossRef]
97. Feng, H.; Tian, H. Improving Crystal Property Prediction from a Multiplex Graph Perspective. *J. Chem. Inf. Model.* **2024**, *64*, 7376–7385. [CrossRef]
98. Jin, L.; Du, Z.; Shu, L.; Cen, Y.; Xu, Y.; Mei, Y.; Zhang, H. Transformer-generated atomic embeddings to enhance prediction accuracy of crystal properties with machine learning. *Nat. Commun.* **2025**, *16*, 1210. [CrossRef]
99. Zhou, Y.; Li, Q.; Zhou, W.; Zang, H.; Xu, L.; Ren, Y.; Xu, J.; Zhan, S.; Ma, W. Reinforce crystal material property prediction with comprehensive message passing via deep graph networks. *Comput. Mater. Sci.* **2024**, *239*, 112958. [CrossRef]
100. Luo, X.; Wang, Z.; Gao, P.; Lv, J.; Wang, Y.; Chen, C.; Ma, Y. Deep learning generative model for crystal structure prediction. *Npj Comput. Mater.* **2024**, *10*, 254. [CrossRef]
101. Ye, C.-Y.; Weng, H.-M.; Wu, Q.-S. Con-CDVAE: A method for the conditional generation of crystal structures. *Comput. Mater. Today* **2024**, *1*, 100003. [CrossRef]
102. Jiao, R.; Huang, W.; Lin, P.; Han, J.; Chen, P.; Lu, Y.; Liu, Y. Crystal structure prediction by joint equivariant diffusion. *Adv. Neural Inf. Process. Syst.* **2023**, *36*, 17464–17497.
103. Wang, J.; Gao, H.; Han, Y.; Ding, C.; Pan, S.; Wang, Y.; Jia, Q.; Wang, H.-T.; Xing, D.; Sun, J. MAGUS: Machine learning and graph theory assisted universal structure searcher. *Natl. Sci. Rev.* **2023**, *10*, nwad128. [CrossRef]
104. Li, C.-N.; Liang, H.-P.; Zhao, B.-Q.; Wei, S.-H.; Zhang, X. Machine learning assisted crystal structure prediction made simple. *J. Mater. Inform.* **2024**, *4*, 15. [CrossRef]
105. Shiraki, Y.; Kaneko, H. Correlations between the constituent molecules, crystal structures, and dielectric constants in organic crystals. *Chemom. Intell. Lab. Syst.* **2025**, *261*, 105376. [CrossRef]
106. Khan, A.A.; Chaudhari, O.; Chandra, R. A review of ensemble learning and data augmentation models for class imbalanced problems: Combination, implementation and evaluation. *Expert Syst. Appl.* **2024**, *244*, 122778. [CrossRef]
107. Nguyen, N.; Ngo, D. Comparative analysis of boosting algorithms for predicting personal default. *Cogent Econ. Financ.* **2025**, *13*, 2465971. [CrossRef]
108. Pickard, C.J.; Needs, R.J. Ab initio random structure searching. *J. Phys. Condens. Matter* **2011**, *23*, 053201. [CrossRef]
109. Glass, C.W.; Oganov, A.R.; Hansen, N. USPEX—Evolutionary crystal structure prediction. *Comput. Phys. Commun.* **2006**, *175*, 713–720. [CrossRef]
110. Hunnisett, L.M.; Nyman, J.; Francia, N.; Abraham, N.S.; Adjiman, C.S.; Aitipamula, S.; Alkhidir, T.; Almehairbi, M.; Anelli, A.; Anstine, D.M.; et al. The seventh blind test of crystal structure prediction: Structure generation methods. *Acta Crystallogr. Sect. B Struct. Sci. Cryst. Eng. Mater.* **2024**, *80*, 517–547. [CrossRef]
111. Pickard, C.J.; Needs, R.J. High-Pressure Phases of Silane. *Phys. Rev. Lett.* **2006**, *97*, 045504. [CrossRef]
112. Doll, K.; Schön, J.C.; Jansen, M. Global exploration of the energy landscape of solids on the ab initio level. *Phys. Chem. Chem. Phys.* **2007**, *9*, 6128–6133. [CrossRef]
113. Goedecker, S. Minima hopping: An efficient search method for the global minimum of the potential energy surface of complex molecular systems. *J. Chem. Phys.* **2004**, *120*, 9911–9917. [CrossRef]

114. Goedecker, S.; Hellmann, W.; Lenosky, T. Global Minimum Determination of the Born-Oppenheimer Surface within Density Functional Theory. *Phys. Rev. Lett.* **2005**, *95*, 055501. [CrossRef]
115. Lyakhov, A.O.; Oganov, A.R.; Stokes, H.T.; Zhu, Q. New developments in evolutionary structure prediction algorithm USPEX. *Comput. Phys. Commun.* **2013**, *184*, 1172–1182. [CrossRef]
116. Ma, Y.; Eremets, M.; Oganov, A.R.; Xie, Y.; Trojan, I.; Medvedev, S.; Lyakhov, A.O.; Valle, M.; Prakapenka, V. Transparent dense sodium. *Nature* **2009**, *458*, 182–185. [CrossRef]
117. Lv, J.; Wang, Y.; Zhu, L.; Ma, Y. Predicted Novel High-Pressure Phases of Lithium. *Phys. Rev. Lett.* **2011**, *106*, 015503. [CrossRef]
118. Zhang, W.; Oganov, A.R.; Goncharov, A.F.; Zhu, Q.; Boulfelfel, S.E.; Lyakhov, A.O.; Stavrou, E.; Somayazulu, M.; Prakapenka, V.B.; Konôpková, Z. Unexpected Stable Stoichiometries of Sodium Chlorides. *Science* **2013**, *342*, 1502–1505. [CrossRef] [PubMed]
119. Lonie, D.C.; Zurek, E. XtalOpt: An open-source evolutionary algorithm for crystal structure prediction. *Comput. Phys. Commun.* **2011**, *182*, 372–387. [CrossRef]
120. Baettig, P.; Zurek, E. Pressure-Stabilized Sodium Polyhydrides: NaH$_n$ ($n > 1$). *Phys. Rev. Lett.* **2011**, *106*, 237002. [CrossRef] [PubMed]
121. Hermann, A.; Ashcroft, N.W.; Hoffmann, R. High pressure ices. *Proc. Natl. Acad. Sci. USA* **2012**, *109*, 745–750. [CrossRef]
122. Pickard, C.J.; Needs, R.J. Highly compressed ammonia forms an ionic crystal. *Nat. Mater.* **2008**, *7*, 775–779. [CrossRef] [PubMed]
123. Liu, H.; Naumov, I.I.; Hoffmann, R.; Ashcroft, N.W.; Hemley, R.J. Potential high-Tc superconducting lanthanum and yttrium hydrides at high pressure. *Proc. Natl. Acad. Sci. USA* **2017**, *114*, 6990–6995. [CrossRef] [PubMed]
124. Tipton, W.W.; Hennig, R.G. A grand canonical genetic algorithm for the prediction of multi-component phase diagrams and testing of empirical potentials. *J. Phys. Condens. Matter* **2013**, *25*, 495401. [CrossRef] [PubMed]
125. Feng, J.; Hennig, R.G.; Ashcroft, N.W.; Hoffmann, R. Emergent reduction of electronic state dimensionality in dense ordered Li-Be alloys. *Nature* **2008**, *451*, 445–448. [CrossRef]
126. Tipton, W.W.; Bealing, C.R.; Mathew, K.; Hennig, R.G. Structures, phase stabilities, and electrical potentials of Li-Si battery anode materials. *Phys. Rev. B* **2013**, *87*, 184114. [CrossRef]
127. Wu, S.Q.; Ji, M.; Wang, C.Z.; Nguyen, M.C.; Zhao, X.; Umemoto, K.; Wentzcovitch, R.M.; Ho, K.M. An adaptive genetic algorithm for crystal structure prediction. *J. Phys. Condens. Matter* **2014**, *26*, 035402. [CrossRef]
128. Zhao, X.; Nguyen, M.C.; Zhang, W.Y.; Wang, C.Z.; Kramer, M.J.; Sellmyer, D.J.; Li, X.Z.; Zhang, F.; Ke, L.Q.; Antropov, V.P.; et al. Exploring the Structural Complexity of Intermetallic Compounds by an Adaptive Genetic Algorithm. *Phys. Rev. Lett.* **2014**, *112*, 045502. [CrossRef]
129. Umemoto, K.; Wentzcovitch, R.M.; Wu, S.; Ji, M.; Wang, C.-Z.; Ho, K.-M. Phase transitions in MgSiO3 post-perovskite in super-Earth mantles. *Earth Planet. Sci. Lett.* **2017**, *478*, 40–45. [CrossRef]
130. Liu, Z.-L. Muse: Multi-algorithm collaborative crystal structure prediction. *Comput. Phys. Commun.* **2014**, *185*, 1893–1900. [CrossRef]
131. Li, X.; Wang, H.; Lv, J.; Liu, Z. Phase diagram and physical properties of iridium tetraboride from first principles. *Phys. Chem. Chem. Phys.* **2016**, *18*, 12569–12575. [CrossRef]
132. Liu, Z.-L.; Jia, H.; Li, R.; Zhang, X.-L.; Cai, L.-C. Unexpected coordination number and phase diagram of niobium diselenide under compression. *Phys. Chem. Chem. Phys.* **2017**, *19*, 13219–13229. [CrossRef]
133. Zhang, Y.-Y.; Gao, W.; Chen, S.; Xiang, H.; Gong, X.-G. Inverse design of materials by multi-objective differential evolution. *Comput. Mater. Sci.* **2015**, *98*, 51–55. [CrossRef]
134. Chen, H.-Z.; Zhang, Y.-Y.; Gong, X.; Xiang, H. Predicting New TiO$_2$ Phases with Low Band Gaps by a Multiobjective Global Optimization Approach. *J. Phys. Chem. C* **2014**, *118*, 2333–2337. [CrossRef]
135. Yang, J.-H.; Zhang, Y.; Yin, W.-J.; Gong, X.G.; Yakobson, B.I.; Wei, S.-H. Two-Dimensional SiS Layers with Promising Electronic and Optoelectronic Properties: Theoretical Prediction. *Nano Lett.* **2016**, *16*, 1110–1117. [CrossRef] [PubMed]
136. Olson, M.A.; Bhatia, S.; Larson, P.; Militzer, B. Prediction of chlorine and fluorine crystal structures at high pressure using symmetry driven structure search with geometric constraints. *J. Chem. Phys.* **2020**, *153*, 094111. [CrossRef] [PubMed]
137. Hajinazar, S.; Thorn, A.; Sandoval, E.D.; Kharabadze, S.; Kolmogorov, A.N. MAISE: Construction of neural network interatomic models and evolutionary structure optimization. *Comput. Phys. Commun.* **2021**, *259*, 107679. [CrossRef]
138. Kolmogorov, A.N.; Shah, S.; Margine, E.R.; Bialon, A.F.; Hammerschmidt, T.; Drautz, R. New Superconducting and Semiconducting Fe-B Compounds Predicted with an Ab Initio Evolutionary Search. *Phys. Rev. Lett.* **2010**, *105*, 217003. [CrossRef] [PubMed]
139. Shao, J.; Beaufils, C.; Kolmogorov, A.N. Ab initio engineering of materials with stacked hexagonal tin frameworks. *Sci. Rep.* **2016**, *6*, 28369. [CrossRef]
140. Xia, K.; Gao, H.; Liu, C.; Yuan, J.; Sun, J.; Wang, H.-T.; Xing, D. A novel superhard tungsten nitride predicted by machine-learning accelerated crystal structure search. *Sci. Bull.* **2018**, *63*, 817–824. [CrossRef]
141. Liu, C.; Gao, H.; Wang, Y.; Needs, R.J.; Pickard, C.J.; Sun, J.; Wang, H.-T.; Xing, D. Multiple superionic states in helium–water compounds. *Nat. Phys.* **2019**, *15*, 1065–1070. [CrossRef]

142. Li, C.; Liang, H.; Duan, Y.; Lin, Z. Machine-learning accelerated annealing with fitting-search style for multicomponent alloy structure predictions. *Phys. Rev. Mater.* **2023**, *7*, 033802. [CrossRef]
143. Liang, H.-P.; Geng, S.; Jia, T.; Li, C.-N.; Xu, X.; Zhang, X.; Wei, S.-H. Unveiling disparities and promises of Cu and Ag chalcopyrites for thermoelectrics. *Phys. Rev. B* **2024**, *109*, 035205. [CrossRef]
144. Yang, S.; Cho, K.; Merchant, A.; Abbeel, P.; Schuurmans, D.; Mordatch, I.; Cubuk, E.D. Scalable diffusion for materials generation. *arXiv* **2023**, arXiv:2311.09235. [CrossRef]
145. Gruver, N.; Sriram, A.; Madotto, A.; Wilson, A.G.; Zitnick, C.L.; Ulissi, Z. Fine-tuned language models generate stable inorganic materials as text. *arXiv* **2024**, arXiv:2402.04379. [CrossRef]
146. Bisbo, M.K.; Hammer, B. Efficient Global Structure Optimization with a Machine-Learned Surrogate Model. *Phys. Rev. Lett.* **2020**, *124*, 086102. [CrossRef] [PubMed]
147. Terayama, K.; Yamashita, T.; Oguchi, T.; Tsuda, K. Fine-grained optimization method for crystal structure prediction. *Npj Comput. Mater.* **2018**, *4*, 32. [CrossRef]
148. Cai, W.; Nix, W.D. *Imperfections in Crystalline Solids*; MRS-Cambridge Materials Fundamentals; Cambridge University Press: Cambridge, UK, 2016. [CrossRef]
149. Wang, J.; Zhao, X.; Zou, G.; Zhang, L.; Han, S.; Li, Y.; Liu, D.; Fernandez, C.; Li, L.; Ren, L.; et al. Crystal-defect engineering of electrode materials for energy storage and conversion. *Mater. Today Nano* **2023**, *22*, 100336. [CrossRef]
150. Zhang, X.; Zhu, D.; Zhang, C.; Zhou, X.; Wu, H.-H.; Wang, F.; Wang, S.; Wu, G.; Gao, J.; Zhao, H.; et al. A review of crystal defect-induced element segregation in multi-component alloy steels. *Prog. Nat. Sci. Mater. Int.* **2024**, *34*, 840–858. [CrossRef]
151. Stukowski, A. Visualization and analysis of atomistic simulation data with OVITO–the Open Visualization Tool. *Model. Simul. Mater. Sci. Eng.* **2010**, *18*, 015012. [CrossRef]
152. Britton, D.; Hinojos, A.; Hummel, M.; Adams, D.P.; Medlin, D.L. Application of the polyhedral template matching method for characterization of 2D atomic resolution electron microscopy images. *Mater. Charact.* **2024**, *213*, 114017. [CrossRef]
153. Wang, F.; Kurc, T.; Widener, P.; Pan, T.; Kong, J.; Cooper, L.; Gutman, D.; Sharma, A.; Cholleti, S.; Kumar, V.; et al. High-Performance Systems for in Silico Microscopy Imaging Studies. In *Data Integration in the Life Sciences*; Springer: Berlin/Heidelberg, Germany, 2010; pp. 3–18.
154. Khan, N.; Asad, H.; Khan, S.; Riccio, A. Towards defect-free lattice structures in additive manufacturing: A holistic review of machine learning advancements. *J. Manuf. Process.* **2025**, *144*, 1–53. [CrossRef]
155. Wu, L.; Wang, W.; Shi, Z.; Zhang, H.; Ke, L.; Liang, X.; Tian, D.; Zhang, H.; Bi, H.; Chen, W.; et al. Rapid identification of defects in doped organic crystalline films via machine learning-enhanced hyperspectral imaging. *Chem. Eng. J.* **2025**, *513*, 162696. [CrossRef]
156. Klunnikova, Y.V.; Anikeev, M.V.; Filimonov, A.V.; Kumar, R. Machine learning application for prediction of sapphire crystals defects. *J. Electron. Sci. Technol.* **2020**, *18*, 100029. [CrossRef]
157. Rodriguez-Guillen, D.; Díez, A.; Andrés, M.V.; Velazquez-Ibarra, L. Inverse design of photonic crystal fibers for dispersion engineering using neural networks. *Opt. Commun.* **2025**, *587*, 131891. [CrossRef]
158. Ding, C.-j.; Wang, X.-y.; Li, X.-y.; Yang, W.-s.; Li, X.-l.; Zhang, Y.-g.; Xu, Y.-c.; Liu, C.-s.; Wu, X. Machine learning-based interatomic potential for simulating irradiation damage mechanisms in ZrC. *J. Mater. Sci. Technol.* **2026**, *242*, 75–91. [CrossRef]
159. Luo, Y.; Byggmästar, J.; Daymond, M.R.; Béland, L.K. Interatomic force fields for zirconium based on the embedded atom method and the tabulated Gaussian Approximation Potential. *Comput. Mater. Sci.* **2024**, *233*, 112730. [CrossRef]
160. Chen, H.; Yuan, D.; Geng, H.; Hu, W.; Huang, B. Development of a machine-learning interatomic potential for uranium under the moment tensor potential framework. *Comput. Mater. Sci.* **2023**, *229*, 112376. [CrossRef]
161. Williams, L.; Sargsyan, K.; Rohskopf, A.; Najm, H.N. Active learning for SNAP interatomic potentials via Bayesian predictive uncertainty. *Comput. Mater. Sci.* **2024**, *242*, 113074. [CrossRef]
162. Keshavarz, S.; Mao, Y.; Reid, A.C.E.; Agrawal, A. Advancing material simulations: Physics-Informed Neural Networks and Object-Oriented Crystal Plasticity Finite Element Methods. *Int. J. Plast.* **2025**, *185*, 104221. [CrossRef]
163. Hussein, R.; Schmidt, J.; Barros, T.; Marques, M.A.L.; Botti, S. Machine-learning correction to density-functional crystal structure optimization. *MRS Bull.* **2022**, *47*, 765–771. [CrossRef]
164. Li, X.; Mai, Y.; Meng, H.; Bi, H.; Ng, C.H.; Teo, S.H.; Lan, C.; Zhang, P.; Li, S. Machine learning quantification of grain boundary defects for high efficiency perovskite solar cells. *Adv. Compos. Hybrid Mater.* **2024**, *7*, 241. [CrossRef]
165. Zhang, S.; Wang, L.; Zhu, G.; Diehl, M.; Maldar, A.; Shang, X.; Zeng, X. Predicting grain boundary damage by machine learning. *Int. J. Plast.* **2022**, *150*, 103186. [CrossRef]
166. He, Z.; Bi, S.; Asare-Yeboah, K. Study of Grain Boundary: From Crystallization Engineering to Machine Learning. *Coatings* **2025**, *15*, 164. [CrossRef]
167. Salmenjoki, H.; Papanikolaou, S.; Shi, D.; Tourret, D.; Cepeda-Jiménez, C.M.; Pérez-Prado, M.T.; Laurson, L.; Alava, M.J. Machine learning dislocation density correlations and solute effects in Mg-based alloys. *Sci. Rep.* **2023**, *13*, 11114. [CrossRef]

168. Deng, F.; Wu, H.; He, R.; Yang, P.; Zhong, Z. Large-scale atomistic simulation of dislocation core structure in face-centered cubic metal with Deep Potential method. *Comput. Mater. Sci.* **2023**, *218*, 111941. [CrossRef]
169. Yang, Z.; Wang, X.; Li, Y.; Lv, Q.; Chen, C.Y.-C.; Shen, L. Efficient equivariant model for machine learning interatomic potentials. *Npj Comput. Mater.* **2025**, *11*, 49. [CrossRef]
170. Wang, G.; Wang, C.; Zhang, X.; Li, Z.; Zhou, J.; Sun, Z. Machine learning interatomic potential: Bridge the gap between small-scale models and realistic device-scale simulations. *iScience* **2024**, *27*, 109673. [CrossRef]
171. Goryaeva, A.M.; Dérès, J.; Lapointe, C.; Grigorev, P.; Swinburne, T.D.; Kermode, J.R.; Ventelon, L.; Baima, J.; Marinica, M.-C. Efficient and transferable machine learning potentials for the simulation of crystal defects in bcc Fe and W. *Phys. Rev. Mater.* **2021**, *5*, 103803. [CrossRef]
172. Borges, Y.; Huber, L.; Zapolsky, H.; Patte, R.; Demange, G. Insights from symmetry: Improving machine-learned models for grain boundary segregation. *Comput. Mater. Sci.* **2024**, *232*, 112663. [CrossRef]
173. Wang, X.; Valdevit, L.; Cao, P. Neural network for predicting Peierls barrier spectrum and its influence on dislocation motion. *Acta Mater.* **2024**, *267*, 119696. [CrossRef]
174. Zhang, L.; Csányi, G.; van der Giessen, E.; Maresca, F. Efficiency, accuracy, and transferability of machine learning potentials: Application to dislocations and cracks in iron. *Acta Mater.* **2024**, *270*, 119788. [CrossRef]
175. Barros de Moraes, E.A.; D'Elia, M.; Zayernouri, M. Machine learning of nonlocal micro-structural defect evolutions in crystalline materials. *Comput. Methods Appl. Mech. Eng.* **2023**, *403*, 115743. [CrossRef]
176. Alarfaj, A.A.; Hosni Mahmoud, H.A. Feature Fusion Deep Learning Model for Defects Prediction in Crystal Structures. *Crystals* **2022**, *12*, 1324. [CrossRef]
177. Honrao, S.; Anthonio, B.E.; Ramanathan, R.; Gabriel, J.J.; Hennig, R.G. Machine learning of ab-initio energy landscapes for crystal structure predictions. *Comput. Mater. Sci.* **2019**, *158*, 414–419. [CrossRef]
178. Sidnov, K.; Konov, D.; Smirnova, E.A.; Ponomareva, A.V.; Belov, M.P. Machine Learning-Based Prediction of Elastic Properties Using Reduced Datasets of Accurate Calculations Results. *Metals* **2024**, *14*, 438. [CrossRef]
179. Kholtobina, A.; Lončarić, I. Exploring elastic properties of molecular crystals with universal machine learning interatomic potentials. *Mater. Des.* **2025**, *254*, 114047. [CrossRef]
180. Kazeev, N.; Al-Maeeni, A.R.; Romanov, I.; Faleev, M.; Lukin, R.; Tormasov, A.; Castro Neto, A.H.; Novoselov, K.S.; Huang, P.; Ustyuzhanin, A. Sparse representation for machine learning the properties of defects in 2D materials. *Npj Comput. Mater.* **2023**, *9*, 113. [CrossRef]
181. Dembitskiy, A.D.; Humonen, I.S.; Eremin, R.A.; Aksyonov, D.A.; Fedotov, S.S.; Budennyy, S.A. Benchmarking machine learning models for predicting lithium ion migration. *Npj Comput. Mater.* **2025**, *11*, 131. [CrossRef]
182. Wei, H.; Bao, H.; Ruan, X. Perspective: Predicting and optimizing thermal transport properties with machine learning methods. *Energy AI* **2022**, *8*, 100153. [CrossRef]
183. Chen, H.; Cai, J.; Zhang, Y.; Lv, X.; Hu, W.; Huang, B. Development of machine learning potentials for Ce-Ti and Ce-Ta binary systems and studies of the liquid-solid interfaces. *Corros. Sci.* **2025**, *246*, 112766. [CrossRef]
184. Thampiriyanon, J.; Khumkoa, S. Machine Learning–Based Prediction of Complex Combination Phases in High-Entropy Alloys. *Metals* **2025**, *15*, 227. [CrossRef]
185. Jang, Y.; Kim, C.H.; Go, A. Classification of magnetic order from electronic structure by using machine learning. *Sci. Rep.* **2023**, *13*, 12445. [CrossRef]
186. Freitas, R.; Cao, Y. Machine-learning potentials for crystal defects. *MRS Commun.* **2022**, *12*, 510–520. [CrossRef]
187. Tanaka, M.; Sasaki, K.; Punyafu, J.; Muramatsu, M.; Murayama, M. Machine-learning-aided analysis of relationship between crystal defects and macroscopic mechanical properties of TWIP steel. *Sci. Rep.* **2025**, *15*, 14435. [CrossRef] [PubMed]
188. Hu, Y.-J. First-principles approaches and models for crystal defect energetics in metallic alloys. *Comput. Mater. Sci.* **2023**, *216*, 111831. [CrossRef]
189. Domínguez-Gutiérrez, F.J.; Byggmästar, J.; Nordlund, K.; Djurabekova, F.; von Toussaint, U. Computational study of crystal defect formation in Mo by a machine learning molecular dynamics potential. *Model. Simul. Mater. Sci. Eng.* **2021**, *29*, 055001. [CrossRef]
190. Dragoni, D.; Daff, T.D.; Csányi, G.; Marzari, N. Achieving DFT accuracy with a machine-learning interatomic potential: Thermomechanics and defects in bcc ferromagnetic iron. *Phys. Rev. Mater.* **2018**, *2*, 013808. [CrossRef]
191. Goryaeva, A.M.; Lapointe, C.; Dai, C.; Dérès, J.; Maillet, J.-B.; Marinica, M.-C. Reinforcing materials modelling by encoding the structures of defects in crystalline solids into distortion scores. *Nat. Commun.* **2020**, *11*, 4691. [CrossRef]
192. Vandermause, J.; Torrisi, S.B.; Batzner, S.; Xie, Y.; Sun, L.; Kolpak, A.M.; Kozinsky, B. On-the-fly active learning of interpretable Bayesian force fields for atomistic rare events. *Npj Comput. Mater.* **2020**, *6*, 20. [CrossRef]

193. Asilian Bidgoli, A.; Rahnamayan, S.; Erdem, B.; Erdem, Z.; Ibrahim, A.; Deb, K.; Grami, A. Machine learning-based framework to cover optimal Pareto-front in many-objective optimization. *Complex Intell. Syst.* **2022**, *8*, 5287–5308. [CrossRef]
194. Tan, C.S.; Gupta, A.; Ong, Y.-S.; Pratama, M.; Tan, P.S.; Lam, S.K. Pareto optimization with small data by learning across common objective spaces. *Sci. Rep.* **2023**, *13*, 7842. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.