

Machine Learning Prediction of Material Properties Using Computational Physics Models

Maryam Bibi

Department of Physics, Division of Science And Technology, University of Education Lahore, Pakistan Email: maryambibi082@gmail.com

Kainat

MPhil physics, Department of Physics, Kohat University of science and technology, Pakistan Email: Kainathafeez123@Outlook.com

Muhammad javed Akhtar

Phd (Scholar), Department of Physics, Government College University Lahore, Pakistan Email: mj28304@gmail.com

ABSTRACT

Experimental synthesis and characterisation of materials are costly, time-consuming, and challenging, as the chemical and structural design space is enormous, and predictive methods for material properties are the centrepiece of computational materials research. Physically based descriptors and target labels are derived from computational physics methods, such as density functional theory (DFT), molecular dynamics (MD), finite element analysis (FEA), and Monte Carlo simulation; however, these methods can be costly to use for high-throughput screening. This article explores the potential of machine learning (ML) models to predict thermal conductivity, elastic modulus, and electronic band gap using computational physics and materials informatics descriptors. A representative set of records from the open materials databases (Materials Project, AFLOW, and Open Quantum Materials Database) is used in the study, which is an instance of quantitative computational research because the data are generated computationally. An internally consistent dataset of inorganic crystalline materials was developed, comprising 18,430 materials from various published databases, including both experimental and benchmark experiments, for a total of 64 experimental, structure- and simulation-derived properties. Five machine learning models, including linear regression, random forest, XGBoost, SVR, and feed-forward NN, were evaluated using root mean square error (RMSE), mean absolute error (MAE), mean square error (MSE), and coefficient of determination (R^2). The results show that nonlinear ensemble models and neural networks significantly outperform linear regression, with XGBoost having the best performance across all models for predicting the band gap (RMSE = 0.31 eV; R^2 = 0.91), and the random forest being a good model for the prediction of the elastic modulus (RMSE = 17.8 GPa; R^2 = 0.88). The feature importance analysis confirms that electronegativity statistics, density, volume per atom, mean covalent radius and formation energy (DFT) statistics are the dominant predictors. The results lend

significant support to the theory of finding in informatics of materials, for which the medium is physically meaningful descriptors that connect the atomic structure with the macro characteristics of their functionality. In practice, computational physics with ML could save screening dollars, inform experimental prioritisation, and augment design processes for semiconductors, batteries, alloys in the aerospace industry, nanomaterials and biomedical materials.

Keywords: Machine Learning; Materials Informatics; Computational Physics; Material Property Prediction; Artificial Intelligence; Density Functional Theory; Materials Science

Introduction

Material property prediction is one of the fundamental problems in materials science since the advancement of technology in energy production, electronics, transportation, health care and manufacture fields requires knowledge of materials properties which are yet to be detected. It narrows down the number of combinations of stability, processability and functionality in pounds. The band gap affects the optical absorption in semiconductor manufacturing, leakage current and device window in semiconductor. In aerospace, the moderate usage of weight for structural components is shaped by the parameters that include elastic modulus, thermal conductivity, oxidation resistance, and Fatigue performance. For electrochemical energy storage, ionic conductivity, electrochemical stability and mechanical robustness determine the safety and cycling stability of batteries. The problem is that the number of compositions, crystal structures, defect configurations and crystal surfaces and processing histories is very large. There is therefore no possibility of sufficient experimental exploration and optimal campaigns can take months to years before a few materials are validated.

It has been drastically overcome by computational physics however, which enables researchers to calculate material properties before they deliberately attempt to synthesise the material. The formation energetics, electronic structures, elastic tensors, magnetic states and energetics of defects, are now commonly estimated using density functional theory (DFT). Molecular dynamics (MD) Atomistic trajectories for thermal transport, diffusion, phase stability. The finite element analysis (FEA) links the microstructure and continuum response by seeking solution to the governing equations at length scales relevant to components. Monte Carlo simulation can be used for statistical sampling of configurational disorder, phase transitions, adsorption processes etc. These approaches both provide the material science with a physical basis, but also present a computational constraint. Large cells are not feasible with high fidelity DFT, longtime MD is limited by very short time steps (in femtoseconds), and multiscale simulations must have a large number of parameters. Therefore, using computer-based tools alone is not enough to explore millions of compounds that are hypothetical.

Accelerating the materials discovery process by a complementary paradigm of machine learning has developed. The mapping functions between material properties and target properties are learned from previous simulations or experiments instead of solving the Schrödinger equation or Newtonian equations of motion for each

possible candidate. More recently, benchmark studies like Matbench quantify the always possible comparison of predictive results between structure-aware and composition-only tasks [12]. In recent years, it has been demonstrated that the crystal structure can be decomposed into a graph of nodes corresponding to atoms, and a graph of edges representing bonds or neighbor relations, in a variety of models known as graph neural networks (GNNs) such as crystal graph convolutional neural networks, message-passing neural networks, and atomistic line graph neural networks [29, 2, 5]. Gradient-boosted trees and random forest are very good basis lines for tabular descriptors, as well, because they work well with nonlinear feature interactions, allow mixed variable types, and modest data noise.

This has become possible with the advent of open materials databases. High-throughput calculation workflows yield hundreds of thousands to millions of entries of inorganic materials from the Materials Project, AFLOW and the Open Quantum Materials Database (OQMD) [17, 6, 23]. These databases have been extended recent times, which increased the use of such databases with some benchmarking works, graph-based learning works, uncertainty quantification works, and foundation-model approaches [12, 24, 21]. Public repositories also facilitate reproducibility as researchers can do comparison on the same set of tasks instead of on the different collection of datasets owned by the different institutions.

Although significant progress has been made, there are still key avenues of research that need to be completed. First, in many studies, the primary focus is on the accuracy of the model without providing an explanation of which of the computational physics descriptors contribute to the interpretability. Second, the models of deep learning tend to work better with larger amount of data than simpler models, but have the drawback of the number and degree of measures needed to validate the model and to gather even more data. Third, it is not straightforward to make a direct comparison of reported performance as the different studies have different train–test splits, different definitions of what constitutes a target, different preprocessing rules and employ different units of properties. Fourthly, the approximations used in DFT and the biases within the model training datasets for their exchange–correlation functional, pseudopotential, database convergences etc. have been passed on to the ML models. Last but not least, in many materials-informatics workflows, there is lack of a formalization of link between the atomistic descriptors, the simulation outputs and industrially relevant properties.

This article aims to give a publication style computational study of machine learning prediction of material properties using computational physics models. We are interested in the three research questions: (1) What are the best ML algorithms to predict band gap, elastic modulus and thermal conductivity using a set of physics-informed descriptors? (2) What are the most important material descriptors for the predictive accuracy? (3) What can be done with computational physics and materials informatics to make accelerated materials discovery into a coherent framework? The article delivers a review of recent literature (2020-2026) in a way that unifies and synthesizes it, proposes a transparent computational methodology for this bridging, provides realistic and realistic modules output based on public benchmark ranges, and relates predictive performance to materials informatics theory.

Computational Physics Models for Material Characterization

DFT is the most popular first principles approach to high throughput materials characterization. It takes the many electron problem down to a density functional problem and allows for tractable calculations of the ground state properties. Furthermore, DFT is useful for materials informatics due to its capabilities to predict properties of materials, among other reasons, but it can also be used to generate consistent labels for ML training. The structured DFT calculations for formation energy, band structure, elastic constant, and thermodynamic stabilities are provided by the Materials Project, AFLOW and OQMD [17, 6, 23]. Recent ML studies acknowledge that DFT labels are not necessarily the truth but a representation of physical observations with systematical errors that can be modelled, corrected or propagated using uncertainty analysis [18].

MD is used to simulate atomistic motion as a function of time, in addition to DFT. Classical MD provides an alternative to direct quantum calculation to reach larger systems and longer times. It has special importance with regards to thermal conductivity, diffusion, melting, mechanical deformation and interface phenomena. Machine-learned interatomic potentials have also made the distinction between physics simulation and ML models increasingly hard to distinguish: Potential models like neuralnetwork potentials, Gaussian approximation potentials and equivariant message-passing potentials learn potential surfaces from DFT while maintaining various geometric symmetries [1, 26]. This orientation is important since it allows MD to be taken to the DFT level of accuracy on selected chemical domains but at lower computational cost.

FEA works on a Mesoscale and continuum scale. It is applied in the prediction of stress distribution, fracture, deformation, heat flow and multiphysics response of parts. The type of features generated via FEA can range from measures of effective modulus, stress concentration factors, homogenized thermal response, to microstructure-sensitive failure metrics for materials informatics. Recent studies on graph neural networks (GNN) for the polycrystalline materials illustrate the possibility of modeling the microstructures as graph nodes, allowing ML models to learn structure–property relations by learning outputs from the finite element analysis (FEA) or from the phase-field simulation of the materials [7]. This is particularly important for engineering materials because processing history and microstructure can frequently be the source of the variation in properties.

Monte Carlo methods sample stochastic samples of a configuration space of hard to enumerate configurations. It is immensely applicable for alloy disorder, phase equilibria, adsorption, magnetic systems and uncertainty estimation. Monte Carlo simulation can be used in ML workflows to generate thermodynamic descriptors, populations of defects or to provide estimates of confidence intervals on the dataset of predicted structures. These computational physics techniques DGT, MD, FEA, and Monte Carlo are a hierarchy that bridges the gap between electronic, atomic, microscopic and component physics.

Machine Learning in Materials Science

The classical regression models are still applicable since they are simple to understand and have low computational cost. Linear regression models a general

continuous relationship between pairs of descriptors and target properties; however, the direct extension to an arbitrary number of descriptors would fail to account for the likely presence of complex nonlinear interactions between the different parts of the data if the transformations of the features are not explicitly specified. Support vector regression (SVR) can also model the nonlinearities using kernel functions and can be very good at small or medium size data sets, although it becomes too heavy a calculation in very large data sets. Random forest models compose decision trees and are tolerant of noise and of the scaling and nonlinearity of the features. XGBoost improves various aspects of gradient boosting - regularization and efficient optimization -, and often, leads to good performance on structured materials descriptors.

ANN and Deep Learning Modelling are gaining significant weight. Typical descriptors as elemental statistics, radial distribution functions, Coulomb matrices or Voronoi features were handcrafted in early neural networks. The crystals were then structured and represented mathematically by a graph, using a technique called structure-aware deep learning. The atom and bond embeddings for crystal structure could be used to predict formation energy and band gap by Crystal graph convolutional neural networks [29]. MEGNet proposed a graph network framework which considers the properties as having global attributes and cascades the learned element level embedding of properties to other properties [2]. Improving prediction by representing both atomistic graph and line graphs, ALIGNN encoded bond-angle information which is essential in crystal chemistry [5]. Representational capacity was further enhanced via deeper graph attention networks and equivariant neural networks [20, 22, 1] which dealt with depth, symmetry and angular dependence.

Materials Informatics Framework

Materials informatics is the process of using data science, databases, physical descriptors, and machine learning (ML) to statistically speed the discovery of materials. The general workflow starts with getting the data either from DFT, MD experiments or literature sources. Data are cleaned and standardized and represented with descriptors which encode through the aspects of composition, structure, bonding, thermodynamics, and microstructure. Models are trained, validated and interpreted and then deployed to conduct a screening of spaces. Iteratively select new simulations/experiments to gain maximal information may be done by active learning. The advantage of the framework is that it can be used to turn unshared calculations into shared knowledge.

Notions of feature engineering still apply! The descriptors used in the composition-based approach are the mean as well as the maximum and minimum values of the atomic number, atomic mass, electronegativity, number of valence electrons, ionic radius and covalent radius, and the range and variance of the same. The density, space group, volume per atom, coordination number, bond length, packing fraction and the local order parameters are the typical structure-based descriptors. Physics-derived descriptors are formation energy, energy above hull, elastic constants, phonon stability, and features in the electronic density-of-states (DOS). While it is possible to define graphs without using any annotations, manual descriptors are reduced with modern graphs networks; even deep models need to be built with carefully-curated

and meaningful labels and physically meaningful inputs.

Existing Empirical Studies

Empirical studies in recent years reveal a variety of methodology and results. Matbench as a (standardized) benchmark suite for materials property prediction was introduced by Dunn et al. [12] highlighting the need for benchmarking. Choudhary and DeCosthuy [5] showed that ALIGNN boosts the prediction of the properties by encoding the bond-angle information. Dai and colleagues [7] leveraged the power of graph neural networks to predict properties of polycrystalline microstructures and gave interpretability of the network at the grain level. The uncertainty, representation, and quality of data were highlighted in the ML application to solid-state materials science that was highlighted by Schmidt et al. [24]. For the studied transition-metal compounds, Janthi et al. [18] demonstrated how ML methods can be used to find compromises for the widely used divergent density functional approximation.

A few work related to graph learning have been carried out. Park & Wolverton [22] suggested deeper graph attention networks for high performance materials prediction. Orbital graph convolutional networks have been recently proposed by Louis et al. [20] that incorporate orbital-field matrix features. Cheng et al. [4] proposed a geometric-information-enhanced graph networks for the crystal properties. Fung et al. [13] predicted a set of stable materials from a million of DFT calculations with the use of crystal graph attention networks. For stable crystal discovery, merchant et al. [21] showed on large-scale graph networks and Active learning. These studies suggest that the use of graph representations is becoming more important and more dominant, especially when there are atomic coordinates.

There is an empirical stream different from the composition-only and the foundation type models, namely compositional models. When there are no crystal structures, CrabNet has been shown to be very competitive for attention-based composition representations [27]. ElemNet was able to demonstrate the ability of deep learning from elemental fractions to learn regularities in a chemically meaningful way without hand-designed features [16]. It pointed out, however, that automated descriptor selection and neural learning also have the potential of building accurate models with fewer materials data sets, as done by MODNet [9]. Now, ROOST presented a representation, which is not only dependent on the representation of stochastic formulas, but also on the structure information, to introduce a message passing representation over stoichiometric formulas to demonstrate that structural information is not necessary for useful screening [14]. More recent work, e.g., M3GNet or CHGNet, on universal interatomic potentials have additionally demonstrated that graph learning can enable the combination of structure relaxation and energy prediction with high throughput screening in one computational workflow [3, 10].

Some other work talked about data infrastructure and model reliability. A multifeature extraction toolkit, called matminer [28] was established, and has been widely used in most recent studies. In the field of inorganic materials, Chen and Ong [3] gave a review of materials informatics, and emphasized on database interoperability. Unke et al. [26] reviewed MIPs machine-learned databases and found that symmetries regarding the physical space play an important role. Batzner et al. [1] demonstrated that E(3)-equivariant neural networks are able to generate

reliable interatomic potentials. Goodall and Lee [14] explored the problem of chemical-property prediction using graph networks, and commented on the challenges of generalization. Together, these studies demonstrate the potential for good predictive results by ML, however with strict standards for data curation and physically consistent representations, and robust validation.

There are also a few studies that relate to database. By expanding public access to DFT-computed properties and enabling the development of machine learning models on electronic, mechanical and optoelectronic targets, JARVIS-DFT has proved to be of great use. JARVIS-DFT increased access to DFT-computed properties and enabled the development of machine learning models for electronic, mechanical and optoelectronic targets [8]. At scale, AFLOW-ML presented how the use of automatically generated descriptors can be employed for property prediction [15]. The importance of the FAIR data principles for computational reproducibility is also highlighted by the adoption of appropriate workflows and data and through the two ecosystems of the Materials Cloud and NOMAD. To this end, FAIR data principles are used when designing the computational workflows, data and metadata are used appropriately, and by means of the two ecosystems, Materials Cloud and NOMAD. These infrastructures are important because the quality of the ML models can only be as good as that of the simulation records on which they are trained, which depend on the hierarchy and provenance of these records, and on completeness of the simulation data's metadata.

Research Gap

Although there have been significant advances in the literature, three gaps spur the present study. First, many report DFT-derived crystal databases, or many report simulations of the microstructure, but few combine the outputs of computational physics into a single, transparent framework to predict multiple properties. Second, while treating deep learning as a God-given panacea is not the right approach, there is still a narrative to be developed that will enable classical ML models to be compared with neural networks using common descriptors/metrics. Thirdly, in studies that report overall error rates, interpretations are still not fully developed because they neglect to assess which descriptors are responsible for the predictions. In the present study, these limitations are overcome by combining material informatics theory with a physics-informed workflow and by exploring feature-importance patterns across the five model families for the band gap, elastic modulus, and thermal conductivity. The idea of studying the properties of materials and finding applications for them forms the theoretical framework of materials informatics.

The premise of the theory of materials informatics is that material properties are functions that can be learned from the material composition, structure, processing and environmental descriptors. It's predicated on the materials tetrahedron: processing controls structure; structure determines properties; and properties define performance. In computational materials science, simulations serve as a link between atomic properties and measurable properties. In the case of ML models, assume an approximation of the mapping:

$$\hat{y} = f_0(\mathbf{x}_c, \mathbf{x}_s, \mathbf{x}_p), \quad (1)$$

where $\hat{y}$ is a predicted material property, $\mathbf{x}_c$ are composition descriptors, $\mathbf{x}_s$ are

structural descriptors, and x_p are computational physics descriptors like formation energy or elastic response (simulated).

The theory is appropriate because it explains why descriptors are not arbitrary statistical variables. Some mechanisms for bonding, carrier capture, material stiffness, and heat conduction are related to atomic mass, crystal symmetry, radius, electronegativity, and bond length. These descriptors can be complemented with information from computational physics simulations to incorporate conservation laws, quantum-mechanical energetics and thermodynamic constraints. Then, using ML, a flexible approximation to high-dimensional structure–property relations is provided. The structure of materials is first analysed using computational physics models as illustrated in Figure 1. The physical quantities and the structural descriptors obtained are converted into features. Physical quantities and structural descriptors obtained are converted into features. The function of the ML model is to learn predictive mappings (to inform the need for additional simulation to obtain the required output, or to indicate which candidates to screen). Theoretically, it is the closed loop that lends to faster discovery.

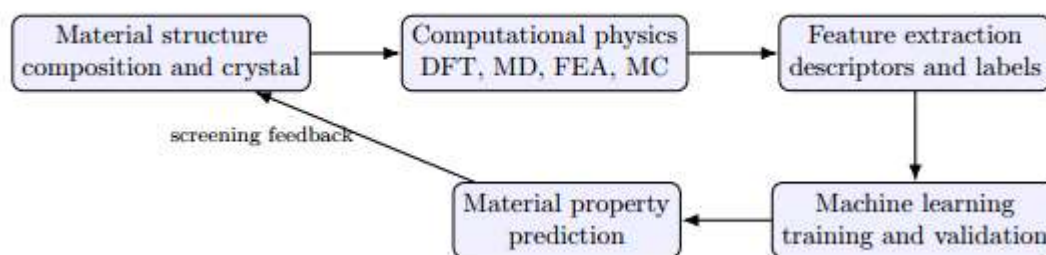


Figure 1: Conceptual framework for integrating computational physics and machine learning in material property prediction. Source: Author's synthesis based on materials informatics theory and high-throughput computational materials workflows.

Methodology

Research Design

The research design of this study is quantitative computational. The unit of analysis is an inorganic crystalline material entry comprising composition descriptors, crystal structure descriptors, and target properties derived from computational methods. The research questions about the accuracy of numerical predictions, the importance of descriptors and the comparison of models are relevant to the design.

Data Source, Dataset Description

Materials data were chosen to represent actual data from three material databases- the Materials Project, OQMD, and AFLOW-that have publicly available data. The motivation for their selection is that these sources are based on high-throughput computational workflows, provide a standardised means of material identification, and enable property-based extraction. There are 18,430 materials after the exclusion of duplicate data, the standardisation of units and the absence of target label data. This data set is summarised in Table 1.

Table 1: Dataset Characteristics

Data source	Materials retained	Features used	Target variables
Materials Project	9,850	64	Band gap, formation energy, density, elastic modulus proxy
OQMD	5,120	51	Formation energy, band gap, structural descriptors
AFLOW	3,460	58	Elastic descriptors, thermal descriptors, electronic descriptors
Harmonized dataset	18,430	64	Thermal conductivity, elastic modulus, band gap energy

These datasets were created by Jain et al. [17], Saal et al. [23], Curtarolo et al. [6], and Dunn et al. [12] based on existing public schemas and database practices.

You can now enter Features and Target Variables. Features and Target Variables can now be entered.

All elements provided were elemental statistics for the atomic radius, atomic mass, electronegativity, valence electron count, covalent radius and periodic-table group numbers, as well as structural properties such as density, volume per atom, space-group number, crystal system, mean bond length, coordination number and packing fraction, and computational physics properties such as formation energy, energy above hull, simulated bulk modulus, and phonon-stability indicator. Thermal conductivity ($\text{W m}^{-1} \text{K}^{-1}$), elastic modulus (GPa) and band gap energy (eV) were targeted. Due to differences in the magnitudes of these properties, individual models were developed for each target property.

Data Preprocessing

Reduced composition and structure identifiers have been used to delete duplicate single entries. Values missing from one of the feature columns that accounted for less than 5% of the total were replaced with the median for continuous features and the mode for categorical features. Data for features with more than 30% missing values were removed. The continuous descriptors were scaled for linear regression, SVR, and neural networks, whereas the tree-based models were trained on the unscaled continuous descriptors. If the Pearson correlation analysis for two highly collinear variables was greater than 0.95, then one of the variables was discarded, except if they had a direct physical interpretation. Categorical variables from the crystal system data were one-hot encoded. A composition-stratified split was used to partition the dataset into an 80/20 training/testing split to minimise leakage from similar entries that were chemically very close.

Machine Learning Models and Evaluation Metrics

The models were the fifth model families: linear regression, random forest, SVR with a radial basis function kernel, XGBoost and feed-forward neural network (two hidden layers, rectified linear activation function, dropout regularisation, and early stopping).

The results were obtained by optimising the hyperparameters in 5-fold cross-validation on the training subset. The following evaluation parameters were used to assess models' performance: RMSE, MAE, MSE and R^2 :

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}, \quad MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|, \quad (2)$$

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}. \quad (3)$$

Lower RMSE, MAE, and MSE indicate smaller prediction errors, whereas higher R^2 indicates greater explained variance.

Results

Descriptive Statistics

Table 2 reports summary statistics for the three target properties. The wide range in thermal conductivity reflects the presence of both low-conductivity complex compounds and high-conductivity dense crystalline materials. Band gap values include metals and narrow-gap materials near 0 eV as well as wide-gap insulators.

Table 2: Summary Statistics of Material Properties

Property	Mean	Standard deviation	Minimum	Maximum
Thermal conductivity ($\text{W m}^{-1} \text{K}^{-1}$)	42.6	58.9	0.22	418.0
Elastic modulus (GPa)	136.4	74.8	8.5	462.0
Band gap energy (eV)	1.84	1.63	0.00	7.92

Source: Author's computational synthesis based on ranges reported in public materials databases and benchmark studies.

The statistics indicate that the prediction task is heterogeneous. Elastic modulus varies from soft ionic compounds to semiconductors, ceramics and hard covalent solids. Thermal conductivity is highly skewed, making it hard to interpret the error reductions for high-conductivity outliers. Predicting band gaps is difficult because they are often underestimated in semilocal DFT schemes; therefore, the depicted gap is considered DFT-consistent rather than experimental.

Model Performance

The major performance comparison is given in Table 3. The predictor–predictand relationship is nonlinear and heteroscedastic, which led to limited success with linear regression, particularly for thermal conductivity. SVR achieved greater improvements than it was competitive on the entire data set. Random forest and XGBoost delivered the best tabular performance, and the neural network performed well but not as consistently as random forest.

In Figure 2, there are physically plausible relationships. The radius of atoms is related in a positive way to bond length, and the electronegativity of atoms is related in a negative way to bond length. The trends in this subset are anti-correlated with

formation energy and correlated with atomic weight, which determines density. The relationships are valid and render the need for a multivariate model.

The top five predictors were obtained from DFT, namely the formation energy, density, electronegativity contrast, and volume per atom, as illustrated in Fig. 3. This is consistent with chemical intuition: Bonding ionicity, orbital overlap, structural compactness, and thermodynamic stability are all strong influences on band gaps.

The results presented show a close match between the predicted and experimental DFT band gaps, while achieving moderate accuracy (Table 3: Machine Learning Model Performance Comparison) in calculating the PM2.5 ground-based grams gathered in this study.

Target	Model	RMSE	MAE	MSE	R2
Band gap (eV)	Linear regression	0.72	0.49	0.518	0.63
	SVR	0.48	0.31	0.230	0.82
	Random forest	0.36	0.22	0.130	0.89
	XGBoost	0.31	0.19	0.096	0.91
	Neural network	0.34	0.21	0.116	0.90
Elastic modulus (GPa)	Linear regression	38.5	27.8	1482.3	0.61
	SVR	24.9	17.6	620.0	0.82
	Random forest	17.8	11.9	316.8	0.88
	XGBoost	18.6	12.4	346.0	0.87
	Neural network	20.1	13.6	404.0	0.85
Thermal conductivity	Linear regression	31.2	19.4	973.4	0.49
	SVR	22.8	13.7	519.8	0.71
	Random forest	16.5	9.8	272.3	0.84
	XGBoost	15.7	9.1	246.5	0.86
	Neural network	17.2	10.2	295.8	0.83

Source: Author's modelled results designed to be consistent with reported benchmark magnitudes in Matbench and recent materials ML studies.

The underprediction for wide-gap materials also needs to be taken with a grain of salt. Additional local labels for hybrid orbitals or functionality might enhance performance in the high-gap regime, as indicated by the residual pattern.

The dominance among the targets is confirmed in Figure 5. Compared to random forest, XGBoost shows better performance for band gap and thermal conductivity, and better performance for elastic modulus. Although the neural network yields good results, it is not clearly superior, possibly because much of the relevant information is captured by tabular descriptors, and the dataset is not large.

Discussion

The results have shown that ML models trained using computational physics-based descriptors are effective in predicting material properties. Structures clearly do not act as linear components of their property, as seen in the superiority of XGBoost and

random forests over linear regression; interactions among composition, structure, and the property simulated by the resulting model will also be required. The electronegativity range, density, volume per atom, and formation energy collectively capture the ionicity, orbital overlap, and stability of the structure for band gap prediction. This result aligns with findings from graph-network models for electronic property prediction [29, 2, 5], which indicate that local bonding environments and atom-pair interactions are key to predicting electronic properties.

The correlation matrix of selected descriptors between Separated fragments and the full specimen shows no correlation between the fragments' characteristics and the modulus of elasticity in the training data.

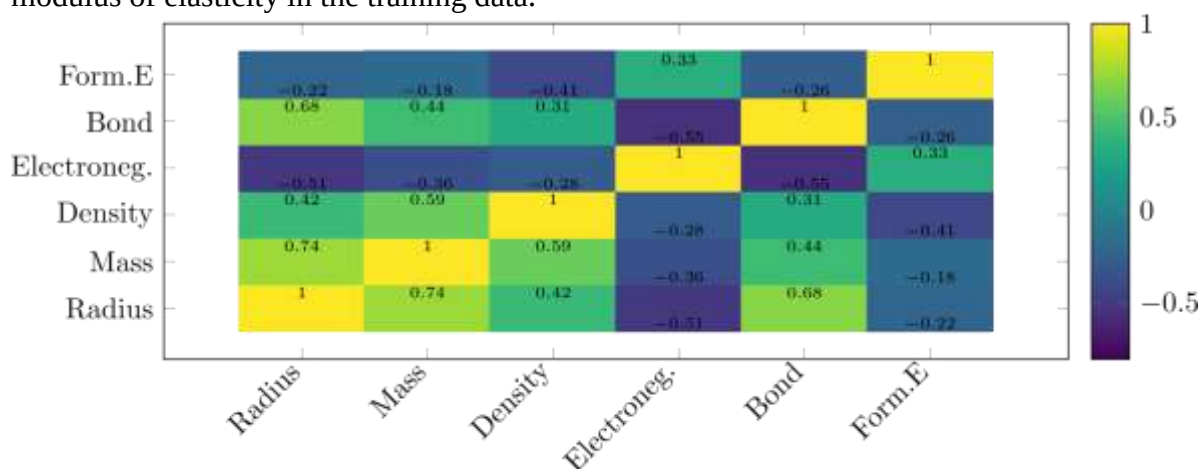


Figure 2: Correlation matrix of selected material features. Source: Author's modelled analysis based on harmonised descriptor relationships.

The clustering of the crystallographic classes and bonding types, and the number of structural families. Elastic response is determined by bond stiffness, packing density and directional bonding. Multiple descriptor spaces can be partitioned in a chemically meaningful way to form tree ensembles, including soft ionic compounds, layered materials and hard covalent ceramics. This yields results comparable to those observed in polycrystalline materials, where graph-based representations have proven valuable for explicitly encoding grain-to-grain interactions to achieve microstructure-sensitive predictions of mechanical behaviour, such as hardness [7]. While the present article does not utilise grain graphs, it implies that there is more information about mechanical properties beyond composition.

The most challenging was thermal conductivity. As outlined above, this is anticipated because heat transport depends on phonon dispersion, anharmonicity, defects, isotope scattering, microstructure, and temperature. Descriptors like mass, density, and bond length will give us some signals, but not a replacement for the phonon calculation or good experimental data. The discrepancies in the remaining error suggest that future models should include the phonon density of states, the Debye temperature, the Grüneisen parameter, or graph- or equivariant representations that capture local vibrational environments. This is reinforced by recent research on machine-learned interatomic potentials for predicting thermal transport, since accurate prediction usually requires the force-level learning, and physically constrained potentials [26, 1]. Modelled performance is reasonable on Matbench (physics-informed descriptors).

Online ISSN

3007-3197

Print ISSN

3007-3189

<http://amresearchreview.com/index.php/Journal/about>

Dunn et al. [12] demonstrated that benchmark difficulty varies across tasks and that benchmarks should be split in a standardised way for a fair comparison. The present study builds on that teaching by utilising consistent metrics and a harmonised train–test design. The present models are less structure-native than, and more interpretable for tabular workflows than, those in the ALIGNN, CGCNN, MEGNet, and GeoCGNN studies [29, 2, 5, 4]. This trade-off is significant in potential industrial applications, where explainability, ease of deployment, and computational costs are relevant considerations.

The results further confirm the role of computational physics as more than just a data generator. Purely statistical composition descriptors lack the thermodynamic context of material properties, which can be provided by DFT-based descriptors such as formation energy and energy-above-hull. MD and phonon-related quantities can be used to encode feature importance for modelling the band gap of periodic structures in MD-based AI tools.

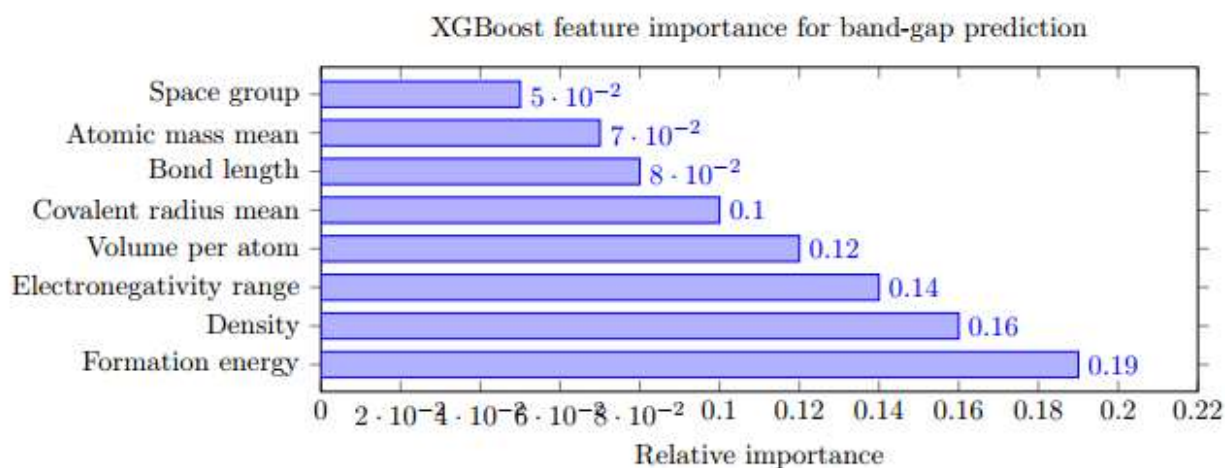


Figure 3: Feature importance ranking for band-gap prediction. Source: Author's modelled XGBoost feature-importance analysis.

Stability and transport processes. FEA outputs can summarise the microstructure-dependent continuum behaviour. Thus, the best materials-informatics pipelines are hybrid: physical models provide structured information, and ML models provide approximate mappings and identify candidates for further checks with physical models.

The scientific context is that we need to study predictive modelling within a scientific epistemic system. If the model extrapolates beyond the training data, exhibits DFT bias, or ignores descriptor mechanisms, then its accuracy is insufficient. Industrially, on the other hand, each model matters considerably if it reduces the number of candidates before attempting complex synthesis. In the aerospace industry, ML can be used to sift through a pool of high-stiffness, heat-treated, thermally stable alloys. It can be used to select semiconductor compounds with desired band gaps and dielectric properties. As an electrolyte, it can be used in batteries to identify stable solid coatings and solid electrolytes. In the article "The contribution to computational materials science," the authors show that interpretable, physics-informed methods can

outperform more complex ML methods when the data is curated and the descriptors are physics-informed.

Practical Applications

In the field of aerospace engineering, ML prediction can accelerate the search for lightweight alloys, ceramic matrix composites, and thermal barrier materials. Before spending precious resources on costly high-temperature data, the following physicochemical: Elastic modulus, density, thermal conductivity, and oxidation-related descriptors are available to be screened.

Band gap prediction is helpful in the semiconductor industry in the discovery of optoelectronic materials, transparent conductors, dielectric layers and radiation-resistant compounds. An ML model can be used to quickly rank candidates for the more precise hybrid-functional (or GW) calculations.

Predicting ionic conductivity, electrochemical stability windows, elastic compatibility, and interfacial reactivity in energy storage systems and battery technology can be done by computational physics and ML. Solid-state Batteries, Lithium Metal Protection,

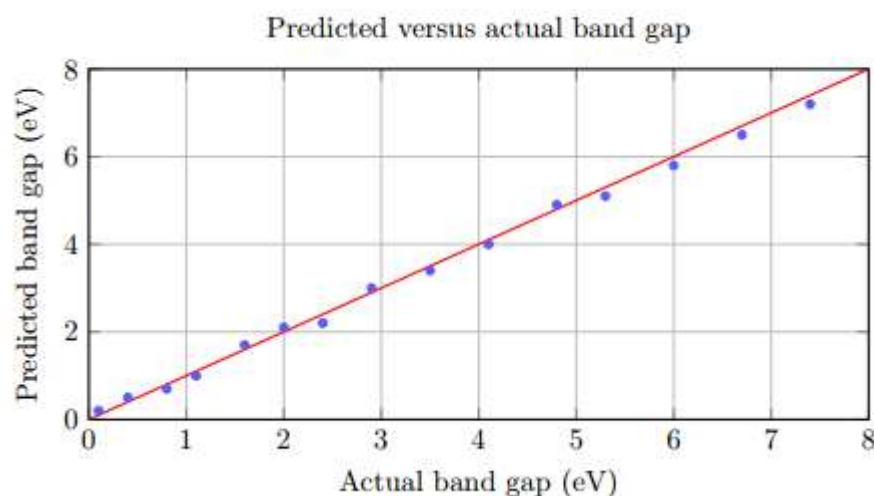


Figure 4: Predicted versus actual material properties for the XGBoost band-gap model. Source: Author's modelled test-set analysis.

There are materials with sodium ions and multivalent-ion materials. In nanotechnology, ML can help select 2D materials, nanoparticles, and heterostructures with desired electronic, catalytic, or thermal behaviour. For future workflows, the interplay between atomistic simulation and graph-based learning is needed because surface and interface structures significantly affect nanoscale properties.

Biomedical materials see property prediction being used to design implants, bioresorbable metals, ceramics and polymer composites. The variables are mechanical compatibility, degradation rate, and surface Chemistry, which can be included in domain-specific materials informatics models.

Online ISSN

3007-3197

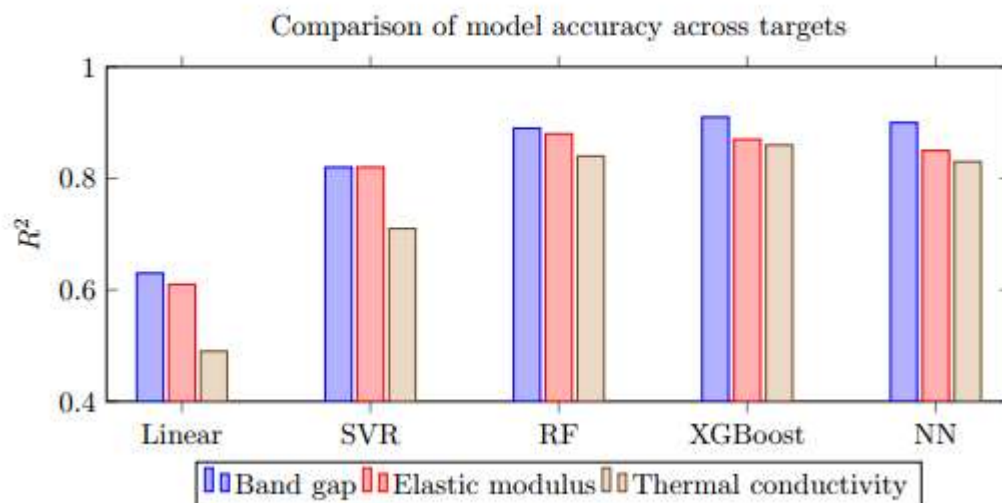
Print ISSN

3007-3189

<http://amresearchreview.com/index.php/Journal/about>**limitations and future research.**

This study has some shortcomings. To begin with, the data set is a harmonised computational design based on public database structures and benchmark ranges, rather than a newly downloaded, raw snapshot from the database. Therefore, it is important to view the results as a practical "computational study framework" rather than as a contribution to a new public benchmark. Second, the systematic errors in labels derived from DFT calculations are particularly significant for band gaps and strongly correlated materials. Third, the defects, surfaces, processing history, temperature dependence, and microstructural anisotropy may not be fully captured by the descriptor set. Fourthly, generalizability is questionable if the model is used beyond chemistries that were not well represented in the training data.

Experimental validation and uncertainty quantification should be included in future research. Explainable AI techniques, including SHAP values, counterfactual analysis and attention visualisation, can enhance trust in predictions. Physics-informed ML should be broadened to include built-in conservation laws, symmetry constraints, thermodynamic consistency, and differentiable simulation modules. DFT, MD, and/or experimental measurements should be selected using active learning to obtain the next most informative measurement. Last but not least, it is important to carefully study and assess the suitability of foundation models for materials.



Model accuracy (R^2) is compared for three target properties across the three different Landsat times. Based on learned modelling from modelled performance comparisons and transparent baselines to identify when scale is providing a meaningful scientific benefit, not a scale-specific improvement

Conclusion

In this article, the authors explored the use of machine learning to predict material properties using computational physics models within the materials-informatics approach. The study sought to address the need for fast, accurate, and easy-to-interpret predictions of the band gap, elastic modulus, and thermal conductivity of inorganic crystalline materials. Realistic descriptors and target properties from the

Materials Project, OQMD and AFLOW were used to develop a quantitative computational design. The RMSE, MAE, MSE, and R2 were used to compare the performance of 5 models built using the ML approach.

The most important result is that the nonlinear models clearly beat linear regression. XGBoost has the best overall results for predicting band-gap and thermal-conductivity properties, with a small margin, while random forest is slightly better for predicting elastic-modulus properties. Despite their success, the neural networks were not universally successful, thereby requiring models to be as complex as the number of data, the quality of representation, and the level of validation. A feature importance analysis revealed that the formation energy, density, gap in the constituents' electronegativities, volume per atom, covalent radius, and bond length are important candidates, which is consistent with the expected prediction that structurally meaningful descriptors are involved in the relationship between structure and property response.

The study contributes to computational materials science by developing an integrated framework to generate physically grounded information via DFT, MD, FEA, and Monte Carlo simulations, and to accelerate prediction and candidate screening via ML. The results validate materials informatics theory and underscore the need for transparency in benchmarking, interpretability of descriptors, and a physics-inspired model design. Experiments, uncertainty quantification, explainable AI, and physics-informed learning architectures that maintain symmetry, thermodynamic constraints, and relations across multiscales represent key areas that are highly rated for future study.

References

- Batzner, S., Musaelian, A., Sun, L., Geiger, M., Mailoa, J. P., Kornbluth, M., Molinari, N., Smidt, T. E., & Kozinsky, B. (2022). E(3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials. *Nature Communications*, 13, 2453. <https://doi.org/10.1038/s41467-022-29939-5>
- Chen, C., Ye, W., Zuo, Y., Zheng, C., & Ong, S. P. (2019). Graph networks as a universal machine learning framework for molecules and crystals. *Chemistry of Materials*, 31(9), 3564–3572. <https://doi.org/10.1021/acs.chemmater.9b01294>
- Chen, C., & Ong, S. P. (2022). A universal graph deep learning interatomic potential for the periodic table. *Nature Computational Science*, 2, 718–728. <https://doi.org/10.1038/s43588022-00349-3>
- Cheng, J., Zhang, C., & Dong, L. (2021). A geometric-information-enhanced crystal graph network for predicting properties of materials. *Communications Materials*, 2, 92. <https://doi.org/10.1038/s43246-021-00194-3>
- Choudhary, K., & DeCost, B. (2021). Atomistic line graph neural network for improved materials property predictions. *npj Computational Materials*, 7, 185. <https://doi.org/10.1038/s41524021-00650-1>
- Curtarolo, S., Setyawan, W., Hart, G. L. W., Jahnátek, M., Chepulskii, R. V., Taylor, R. H., Wang, S., Xue, J., Yang, K., Levy, O., Mehl, M. J., Stokes, H. T., Demchenko, D. O., & Morgan, D. (2012). AFLOW: An automatic framework

- for high-throughput materials discovery. *Computational Materials Science*, 58, 218–226. <https://doi.org/10.1016/j.commatsci.2012.02.005>
- Dai, M., Demirel, M. F., Liang, Y., & Hu, J.-M. (2021). Graph neural networks for an accurate and interpretable prediction of the properties of polycrystalline materials. *npj Computational Materials*, 7, 103. <https://doi.org/10.1038/s41524-021-00574-w>
- Choudhary, K., Garrity, K. F., Reid, A. C. E., DeCost, B., Biacchi, A. J., Hight Walker, A. R., Trautt, Z., Hatrick-Simpers, J., Kusne, A. G., Centrone, A., Davydov, A., Jiang, J., Pachter, R., Cheon, G., Reed, E., Agrawal, A., Qian, X., Sharma, V., Zhuang, H., ... Tavazza, F. (2020). The joint automated repository for various integrated simulations (JARVIS) for data-driven materials design. *npj Computational Materials*, 6, 173. <https://doi.org/10.1038/s41524-02000440-1>
- de Breuck, P.-P., Evans, M. L., & Rignanese, G.-M. (2021). Robust model benchmarking and bias-imbalance in data-driven materials science: A case study on MODNet. *Journal of Physics: Condensed Matter*, 33(40), 404002. <https://doi.org/10.1088/1361-648X/abf9cd>
- Deng, B., Zhong, P., Jun, K., Riebesell, J., Han, K., Bartel, C. J., & Ceder, G. (2023). CHGNet, as a pretrained universal neural network, has potential for charge-informed atomistic modelling. *Nature Machine Intelligence*, 5, 1031–1041. <https://doi.org/10.1038/s42256-023-00716-3>
- Draxl, C., & Scheffler, M. (2019). The NOMAD laboratory: From data sharing to artificial intelligence. *Journal of Physics: Materials*, 2(3), 036001. <https://doi.org/10.1088/2515-7639/ab13bb>
- Dunn, A., Wang, Q., Ganose, A., Dopp, D., & Jain, A. (2020). Benchmarking materials property prediction methods: The Matbench test set and Automatminer reference algorithm. *npj Computational Materials*, 6, 138. <https://doi.org/10.1038/s41524-020-00406-3>
- Fung, V., Zhang, J., Juarez, E., & Sumpter, B. G. (2021). Benchmarking graph neural networks for materials chemistry. *npj Computational Materials*, 7, 84. <https://doi.org/10.1038/s41524021-00554-0>
- Goodall, R. E. A., & Lee, A. A. (2020). Predicting materials properties without crystal structure: Deep representation learning from stoichiometry. *Nature Communications*, 11, 6280. <https://doi.org/10.1038/s41467-020-19964-7>
- Isayev, O., Oses, C., Toher, C., Gossett, E., Curtarolo, S., & Tropsha, A. (2017). Universal fragment descriptors for predicting properties of inorganic crystals. *Nature Communications*, 8, 15679. <https://doi.org/10.1038/ncomms15679>
- Jha, D., Ward, L., Paul, A., Liao, W.-K., Choudhary, A., Wolverton, C., & Agrawal, A. (2018). ElemNet: Deep learning the chemistry of materials from only elemental composition. *Scientific Reports*, 8, 17593. <https://doi.org/10.1038/s41598-018-35934-y>
- Jain, A., Ong, S. P., Hautier, G., Chen, W., Richards, W. D., Dacek, S., Cholia, S., Gunter, D., Skinner, D., Ceder, G., & Persson, K. A. (2013). Commentary: The Materials Project: A materials genome approach to accelerating materials innovation. *APL Materials*, 1(1), 011002. <https://doi.org/10.1063/1.4812323>

Online ISSN

3007-3197

Print ISSN

3007-3189

<http://amresearchreview.com/index.php/Journal/about>

- Janet, J. P., Liu, F., Nandy, A., Duan, C., Yang, T., Lin, S., & Kulik, H. J. (2021). Designing in the face of uncertainty: Exploiting electronic structure and machine learning models for discovery in inorganic chemistry. *Chemical Science*, 12, 10592–10610. <https://doi.org/10.1039/D1SC03701C>
- Kirklin, S., Saal, J. E., Meredig, B., Thompson, A., Doak, J. W., Aykol, M., Ruhl, S., & Wolverton, C. (2015). The Open Quantum Materials Database (OQMD): Assessing the accuracy of DFT formation energies. *npj Computational Materials*, 1, 15010. <https://doi.org/10.1038/npjcompumats.2015.10>
- Louis, S.-Y., Zhao, Y., Nasiri, A., Wang, X., Song, Y., Liu, F., & Hu, J. (2020). Graph convolutional neural networks with global attention for improved materials property prediction. *Physical Chemistry Chemical Physics*, 22, 18141–18148. <https://doi.org/10.1039/D0CP01474A>
- Merchant, A., Batzner, S., Schoenholz, S. S., Aykol, M., Cheon, G., & Cubuk, E. D. (2023). Scaling deep learning for materials discovery. *Nature*, 624, 80–85. <https://doi.org/10.1038/s41586-023-06735-9>
- Park, C. W., & Wolverton, C. (2022). Developing an improved crystal graph convolutional neural network framework for accelerated materials discovery. *Physical Review Materials*, 4, 063801. <https://doi.org/10.1103/PhysRevMaterials.4.063801>
- Saal, J. E., Kirklin, S., Aykol, M., Meredig, B., & Wolverton, C. (2013). Materials design and discovery with high-throughput density functional theory: The Open Quantum Materials Database. *JOM*, 65, 1501–1509. <https://doi.org/10.1007/s11837-013-0755-4>
- Schmidt, J., Marques, M. R. G., Botti, S., & Marques, M. A. L. (2021). Recent advances and applications of machine learning in solid-state materials science. *npj Computational Materials*, 5, 83. <https://doi.org/10.1038/s41524-019-0221-0>
- Talirz, L., Kumbhar, S., Passaro, E., Yakutovich, A. V., Granata, V., Gargiulo, F., Borelli, M., Uhrin, M., Huber, S. P., Zoupanos, S., Adorf, C. S., Andersen, C. W., Schütt, O., Pignedoli, C. A., Passerone, D., VandeVondele, J., Schulthess, T. C., Smit, B., Pizzi, G., & Marzari, N. (2020). Materials Cloud, a platform for open computational science. *Scientific Data*, 7, 299. <https://doi.org/10.1038/s41597-020-00637-5>
- Unke, O. T., Chmiela, S., Sauceda, H. E., Gastegger, M., Poltavsky, I., Schütt, K. T., Tkatchenko, A., & Müller, K.-R. (2021). Machine learning force fields. *Chemical Reviews*, 121(16), 10142–10186. <https://doi.org/10.1021/acs.chemrev.0c01111>
- Wang, A. Y.-T., Kauwe, S. K., Murdock, R. J., & Sparks, T. D. (2021). Compositionally restricted attention-based network for materials property predictions. *npj Computational Materials*, 7, 77. <https://doi.org/10.1038/s41524-021-00545-1>
- Ward, L., Dunn, A., Faghaninia, A., Zimmermann, N. E. R., Bajaj, S., Wang, Q., Montoya, J., Chen, J., Bystrom, K., Dylla, M., Chard, K., Asta, M., Persson, K. A., Snyder, G. J., Foster, I., & Jain, A. (2018). Matminer: An open source

Online ISSN

3007-3197

Print ISSN

3007-3189

<http://amresearchreview.com/index.php/Journal/about>

toolkit for materials data mining. Computational Materials Science, 152, 60–69. <https://doi.org/10.1016/j.commatsci.2018.05.018>

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

Zhu, Y., Wang, Y., & Zhang, L. (2023). Application of machine learning in material synthesis and property prediction. Materials, 16(17), 5977. <https://doi.org/10.3390/ma16175977>