



ARTICLE

Large-Scale Formation Energy Prediction in One Million Crystalline Materials Using Deep Learning and Principal Component Analysis

Muhammad Imran¹, Javed Ferzund², Ateeq Ur Rehman Butt¹, Abdul Noman¹, Hessa Alfraihi³
and Mohamad Khairi Ishak^{4,*}

¹Department of Computer Science, Faculty of Computing, Engineering & Technology, University of Kamalia, Toba Tek Singh, Pakistan

²Department of Computer Science, Comsats University Islamabad, Sahiwal Campus, Sahiwal, Pakistan

³Department of Information Systems, College of Computer and Information Sciences, Princess Nourah Bint Abdulrahman University, Riyadh, Saudi Arabia

⁴Department of Computer Engineering, College of Computing and Informatics, University of Sharjah, Sharjah, United Arab Emirates

*Corresponding Author: Mohamad Khairi Ishak. Email: mishak@sharjah.ac.ae

Received: 21 July 2026; Accepted: 04 September 2026; Published: 22 October 2026

ABSTRACT: Predicting the formation energy of crystalline materials with high accuracy and low computational cost is important for accelerating large-scale materials screening. This study presents a systematic comparison of four deep learning architectures multi-layer perceptron (MLP), deep neural network (DNN), recurrent neural network (RNN), and long short-term memory (LSTM) and three classical machine-learning regressors, namely random forest, support vector regression, and gradient boosting. The models were evaluated using one million DFT-computed crystalline-material records containing nine structural and electronic descriptors, with an 80:10:10 train/validation/test split. Principal component analysis (PCA) reduced the numerical descriptor space from nine variables to six components retaining 95.1% of the variance. Under the original descriptor configuration, the DNN showed the lowest error among the evaluated models, and PCA was associated with a further reduction in error and faster convergence. However, energy_per_atom is included among the predictors and is mathematically related to formation_energy_per_atom; therefore, the very low absolute errors obtained in this configuration may be affected by target leakage and should not be interpreted as leakage-free estimates of generalization performance. Feedforward architectures consistently outperformed recurrent models on the tabular descriptor representation. The reliability of the comparative workflow was further assessed through independent held-out testing, training-set-only preprocessing, repeated-run statistical evaluation, and comparison across seven algorithms spanning deep and classical machine-learning paradigms. The study therefore provides a large-scale comparative benchmark and a preprocessing analysis, while highlighting the need for leakage-controlled descriptors, categorical treatment of crystallographic labels, and source-aware validation in future evaluations.

KEYWORDS: Formation energy prediction; deep learning; materials informatics; DNN; LSTM; PCA; big data; chemical stability; surrogate modelling; high-throughput screening

1 Introduction

Searching for new materials that are genuinely stable, not just theoretically plausible, is one of the most demanding tasks in modern materials science. The stakes are high: solid-state electrolytes for next-generation batteries, photovoltaic absorbers, green-chemistry catalysts, and lightweight aerospace alloys all depend on

identifying candidates that will not decompose under operating conditions [1,2]. The quantity that most directly captures this requirement is the formation energy per atom, which measures the energy gained or lost when a compound is assembled from its constituent elements in their standard states. A sufficiently negative formation energy signals that a compound is thermodynamically favoured over decomposition into elemental phases, making it the standard first-pass filter in any high-throughput computational search [3].

Experimental routes to measuring stability have well-known limitations at the screening stage. Calorimetric methods solution and combustion calorimetry being the most common can deliver precise formation enthalpies, but they require bulk quantities of phase-pure material that simply do not exist for hypothetical compounds still under consideration [4]. Spectroscopic techniques like X-ray diffraction and Raman spectroscopy are useful for confirming what has already been made, but they yield only indirect stability information. In short, experiment is a tool for characterising synthesised compounds, not for predicting which hypothetical ones are worth making.

Density Functional Theory (DFT), grounded in the Hohenberg-Kohn theorems and the Kohn-Sham formalism [5], has transformed computational materials discovery by enabling the parameter-free prediction of ground-state energies, electronic structures, and formation enthalpies from first principles. The systematic application of DFT at scale, exemplified by initiatives such as the Materials Project [6], the Automatic FLOW (AFLOW) framework [7], and the Open Quantum Materials Database (OQMD) [8], has produced databases of computed materials properties that are now freely available to the community. For all its power, DFT carries a substantial computational price. A single self-consistent field calculation on a moderately complex unit cell of 50–100 atoms can take anywhere from several hours to multiple days on high-performance hardware, depending on the exchange-correlation functional and convergence settings [9]. Running this for even a million hypothetical compounds a modest-sized chemical library by modern standards would consume resources that most research groups simply do not have, and computational demand grows faster than linearly with system complexity. There is clearly a need for something faster.

Machine learning has emerged as the most practical way to bridge this gap. By fitting models to pre-computed DFT data, it becomes possible to build surrogate functions that match DFT accuracy at a tiny fraction of the run-time cost; inference is typically orders of magnitude faster [10,11]. Early work in this area leaned heavily on handcrafted descriptors: the Coulomb matrix [12], the many-body tensor representation [13], and the sine matrix [13], paired with kernel-based regressors like kernel ridge regression and Gaussian process regression. These approaches proved that the concept worked, achieving reasonable accuracy on datasets of several thousand compounds, and the descriptor-plus-regressor template they established still shapes the field today.

Deep learning has pushed things considerably further. Convolutional networks on fixed grids [14], feed-forward networks on fingerprint vectors [15,16], and recurrent architectures processing SMILES strings [17] have all found application in property prediction. Perhaps the most consequential development, though, has been the graph neural network (GNN), which encodes a crystal as a periodic graph atoms as nodes, bonds as edges and learns directly from 3D geometry rather than from hand-engineered scalar summaries [18]. CGCNN [19] was the landmark demonstration that GNN-based predictions could match DFT's own uncertainty on Materials Project data. What followed was a rapid succession of improvements: MEGNet [20] added state-level attributes like temperature; SchNet [21] used continuous-filter convolutions for rotationally invariant representations; DimeNet [22] incorporated bond-angle information; and M3GNet [23] extended applicability to elastic moduli, band gaps, and phonon spectra, among other properties.

Graph-based methods are impressive, but they are not always the most practical choice. They require full crystal structure input three-dimensional atomic coordinates and lattice parameters which is unavailable for most hypothetical compounds unless a separate structure prediction step is run first, as underscored

by recent structure-plus-composition frameworks that still require diffraction or geometric input alongside elemental descriptors [24,25]. Composition-based descriptors, by contrast, can be computed instantly from a chemical formula alone, making them invaluable for rapid pre-screening before any structural relaxation is attempted. Standard feedforward and recurrent networks trained on such descriptors are also far easier to deploy: they rely on mainstream ML libraries, require no specialist dependencies, and can be maintained by a much wider pool of researchers.

A dimension of the problem that has received less systematic attention is descriptor preprocessing. In practice, compositional descriptor sets are rarely well-conditioned: volume-related features `volume_per_atom`, `atomic_volume_per_atom`, and `volume_deviation` are by definition correlated because they encode overlapping structural information. This multicollinearity worsens the conditioning of the empirical covariance matrix and makes gradient-based optimisation harder [26]. PCA offers a clean, theoretically motivated fix: it projects the descriptor matrix onto an orthonormal eigenbasis ordered by explained variance, removing correlations and reducing effective dimensionality in the process [27,28]. Although PCA is routinely used in chemoinformatics and drug discovery [29], its systematic evaluation in the context of large-scale deep learning for materials property prediction has not been rigorously carried out.

This work presents a large-scale benchmarking study in which four deep learning architectures—MLP, DNN, RNN, and LSTM—are compared with three classical machine-learning baselines—Random Forest, Support Vector Regression, and Gradient Boosting—on one million DFT-computed crystalline materials. Each model is evaluated under the same original descriptor configuration, with additional experiments using PCA-transformed features. The paper's main contributions are: (i) a direct comparison of seven models across classical and deep learning paradigms at million-sample scale; (ii) an analysis of how PCA-based descriptor decorrelation changes model error and optimization behaviour within the evaluated feature configuration; (iii) a mechanistic discussion of why feedforward architectures are more appropriate than recurrent models for non-sequential tabular materials descriptors; and (iv) a transparent assessment of methodological limitations that affect interpretation of the absolute error values. In particular, because `energy_per_atom` is mathematically related to the target `formation_energy_per_atom`, the PCA-associated accuracy values are treated as within-configuration benchmarking results rather than as a leakage-free state-of-the-art benchmark.

It is worth dwelling for a moment on why formation energy, of all the properties one could screen for, has become the field's *de facto* first filter. Part of the answer is practical: it is one of the few quantities that DFT can return with a single, comparatively cheap relaxation calculation, rather than requiring the additional finite-difference or perturbative steps needed for elastic constants, dielectric response, or phonon spectra. But the deeper reason is that thermodynamic stability is a precondition for almost everything else a materials scientist might want. A candidate cathode material with an excellent theoretical voltage profile is worthless if it decomposes into its constituent binaries the moment it is synthesised; a promising photovoltaic absorber with an ideal band gap is similarly moot if it is thermodynamically unstable with respect to competing phases. Formation energy screening therefore functions less as a single property prediction task and more as a triage step that determines which compounds are even worth subjecting to further, more expensive computational or experimental characterisation. Getting this triage step wrong in either direction carries a cost: false negatives discard compounds that might otherwise have been promising, while false positives waste downstream experimental effort on compounds that were never going to be stable in the first place.

The urgency behind this problem has only grown as application areas have multiplied. Solid-state battery research, for instance, needs electrolyte candidates that are simultaneously stable against both the lithium-metal anode and the high-voltage cathode a narrow stability window that is difficult to identify without screening thousands of candidate compositions [1,2]. Catalysis researchers face a related but distinct

challenge: surface stability under reaction conditions, which depends on bulk formation energy as a starting point even though it is not the whole story. Aerospace and structural-alloy design adds yet another wrinkle, since the compounds of interest there are often metastable by design, sitting just above the convex hull, where careful quantification of how far above is more useful than a simple stable/unstable label. None of these application domains can be served well by a screening tool that is merely fast; the tool also has to be accurate enough that its outputs can be trusted without DFT re-validation of every candidate, which is precisely the bar that this study sets out to test against.

The growth of public DFT repositories over the last decade has quietly changed what kind of modelling question is even askable. A decade ago, training sets of a few thousand to perhaps twenty thousand compounds were the norm, and the modelling literature reflected this: kernel methods and random forests dominated in part because they degrade gracefully on small data and do not require careful regularisation to avoid overfitting. As the Materials Project, AFLOW, and OQMD have each grown past hundreds of thousands and, collectively, into the low millions of computed entries [6–8], the calculus has shifted. Deep architectures that would have badly overfit a ten-thousand-sample dataset become not just viable but advantageous once a million labelled examples are available, since their capacity to model complex, non-linear interactions among descriptors can finally be put to productive use rather than being a liability. This shift in data regime is itself a partial explanation for why the present study reaches considerably lower error rates than earlier benchmarks performed on smaller corpora; it is not only the choice of architecture that matters, but the scale at which that architecture is allowed to learn.

There is also a practical, almost infrastructural, argument for staying with composition-based descriptors rather than moving straight to graph representations of crystal structure. Many research groups, particularly those without dedicated computational-materials staff, do not have an easy way to generate plausible 3D structures for purely hypothetical compositions; structure prediction is itself a hard search problem, often requiring evolutionary algorithms or random structure searching that can take as long as the DFT relaxation it is meant to precede. A composition-only model sidesteps this bottleneck entirely: given nothing more than a chemical formula and a handful of derived statistics, it returns a stability estimate in milliseconds. This makes such models attractive not as a replacement for graph-based methods, which remain the more accurate choice once a structure is in hand, but as an early-stage filter that can be run over enormous combinatorial spaces of candidate formulae before anyone commits to the more expensive structural step. Treating composition-based and structure-based methods as complementary stages of a single pipeline, rather than as competitors, is the framing we adopt throughout this paper.

The rest of this paper is organized as follows. [Section 2](#) surveys the relevant literature in more detail, covering classical machine learning approaches, deep feedforward and recurrent architectures, graph neural networks, and the comparatively under-explored question of descriptor preprocessing. [Section 3](#) describes the dataset and presents an exploratory analysis of both the target variable and the descriptor space, including the PCA decomposition that motivates our preprocessing experiments. [Section 4](#) details the modelling methodology, covering data partitioning, the four deep learning architectures and three classical baselines, and the evaluation protocol. [Section 5](#) reports the benchmark results, with and without PCA preprocessing, alongside statistical significance testing. [Section 6](#) discusses what these results imply for practitioners building screening pipelines, and [Section 7](#) concludes with a summary of contributions and directions for future work.

2 Related Work

2.1 Classical Machine Learning for Materials Property Prediction

Work on machine learning for materials property prediction goes back well before the deep learning wave. Cerqueira et al. [30] more recently showed that a non-orthogonal, chemical-similarity-based representation can identify stable compositions from crystal structure databases at practically useful rates, extending the older ionic-substitution line of work to a broader and more transferable similarity metric. Jin et al. [31] compared conventional machine learning models against graph neural networks for predicting formation energy and band gap in perovskite oxides, finding that the choice between hand-crafted descriptors and learned representations still involves a real accuracy/cost trade-off even with modern architectures. Wang et al. [32] took a different angle, applying interpretable symbolic regression to extract compact, physically meaningful analytical equations linking materials science data to target properties, continuing the SISSO-style descriptor-discovery tradition. For band gap and formation energy prediction broadly, random forest regressors have been the workhorse of choice, typically posting MAEs of 0.1–0.3 eV/atom on datasets of tens of thousands of compounds [33]. XGBoost [34] has generally matched or beaten random forests on tabular descriptor sets while being more stable across runs. The common limitation of all these methods is representational: their predictive ceiling is set by how informative the hand-crafted input features are.

2.2 Deep Neural Networks for Property Prediction

Deep feedforward networks first proved their worth in materials-adjacent applications in drug discovery, where Rahman et al. [10] more recently showed that ensembling deep graph convolutional networks, including a multi-task variant of CGCNN, substantially improves the precision of formation-energy, band-gap, and density predictions across tens of thousands of inorganic crystals. More recently, Shi et al. [35] reviewed how graph neural networks build on early convolutional operations defined directly on molecular graphs, a precursor framework that helped establish the GNN methods that would later dominate the field. On the crystalline materials side, Faber et al. [13] carried out a thorough benchmark across QM9 and related datasets, finding that kernel methods with the many-body tensor representation could reach sub-chemical-accuracy on small organic molecules. Alghadeer et al. [36] survey how composition-based descriptor sets, including the Magpie elemental-property statistics that underpin Ward's original framework, continue to support random-forest and physics-guided models that compete with structure-based approaches. Our work extends this composition-descriptor tradition, scaling evaluation to one million samples with deep learning models.

2.3 Graph Neural Networks for Crystalline Materials

CGCNN [19] represented a genuine turning point. By encoding crystals as periodic graphs with atom-type and bond-length features, Xie and Grossman showed the model could achieve an MAE of around 0.039 eV/atom on Materials Project formation energies, a competitive result at the time. MEGNet [20] built on this by adding state-level attributes such as temperature and pressure, broadening applicability across properties. SchNet [21], originally a quantum chemistry tool, brought continuous-filter convolutions and radial basis function distance expansions that yield smooth, rotationally invariant crystal representations. Angular information entered the picture with DimeNet [22], and ALIGNN [37] later showed that combining bond-length and bond-angle graph layers lowered errors further. More recently, universal potentials like M3GNet [23] and CHGNet [38] have demonstrated that training on large multi-fidelity datasets enables near-DFT quality for structure relaxation tasks. The accuracy of these graph methods is undeniable, but they require full atomic coordinates as input and are considerably more expensive and complex to deploy than the feedforward architectures we evaluate here.

2.4 Recurrent and Sequence-Based Models

Recurrent models have found their main use in materials ML through sequence-based representations of molecules. Gómez-Bombarelli et al. [17] showed that variational autoencoders on SMILES strings can navigate molecular structure space in a smooth, continuous latent representation, a demonstration that sequence-based architectures can capture chemically meaningful structure from string-encoded molecular representations more broadly. Crystalline materials present a less natural fit for sequential models because there is no canonical ordering of atomic sites, though researchers have adapted the approach by treating elemental fractions as ordered sequences. This study includes RNN and LSTM models evaluated under this same compositional sequence framing, allowing direct comparison with feedforward architectures on identical data.

2.5 Dimensionality Reduction and Feature Engineering

Feature engineering and dimensionality reduction occupy a somewhat awkward position in materials ML literature: widely used in practice, but less often subjected to controlled benchmarking. Vasylenko et al. [16] more recently demonstrated that digital descriptors extracted from the local geometry around each element in a crystal structure could generalise well across inorganic crystal families, extending the earlier fragment-descriptor line of work. PCA has also been discussed as a preprocessing and dimensionality-reduction step in recent broader perspectives on machine learning for computational materials science and chemistry [27], and specifically for formation-energy modelling with deep learning on voxelised structure representations [29]. In the present work, PCA is evaluated as a statistical decorrelation step for the numerical descriptor representation. Because the original feature set includes `energy_per_atom`, which is mathematically related to `formation_energy_per_atom`, the observed reduction in prediction error after PCA is interpreted conservatively as a within-configuration result rather than being attributed solely to descriptor orthogonalisation or presented as a leakage-free benchmark.

A separate but related thread concerns universal machine-learning interatomic potentials (uMLIPs) and pretrained atomistic representations, which have recently been used to relax structures and estimate formation energy and convex-hull distance directly, offering an alternative route to stability screening that complements composition-based regression [39]. Transfer-learning strategies that reuse pretrained atomistic embeddings for downstream property regressors, either through frozen features or fine-tuning, have shown similar promise on smaller, structure-aware datasets [40–42]. These approaches require either a candidate structure or a pretrained potential capable of generating one, whereas the descriptor-based models benchmarked here operate on nine scalar structural and electronic descriptors that are considerably cheaper to obtain; the two families of methods are best viewed as complementary tools suited to different stages of a screening pipeline rather than direct competitors.

2.6 Deep Learning vs. Tree-Based Models on Tabular Data

Any claim that a deep network beats classical baselines on tabular descriptors has to be read against a broader and somewhat contentious debate in the machine learning community. Grinsztajn et al. [43] recently put together a rigorous empirical case that tree-based ensembles gradient-boosted trees in particular continue to outperform deep learning on most tabular benchmarks, attributing this to trees' native robustness to uninformative features, rotation-variance of decision boundaries, and an inductive bias that happens to suit the irregular, non-smooth functions that real-world tabular data often encode. This is a useful counterpoint to keep in mind when interpreting our own results: our finding that the DNN outperforms XGBoost on formation energy prediction is not a contradiction of that broader literature so much as a statement about which regime materials descriptors fall into. Formation energy as a function of composition

is, at least locally, a comparatively smooth target, physically continuous and grounded in additive atomic contributions, which plays to the strengths of gradient-based feedforward networks rather than against them. Theoretical support for why a sufficiently large feedforward network should be capable of representing such a function comes from the universal approximation results of Hornik [44], though that theory says nothing about sample efficiency or optimisation difficulty, both of which are empirical questions that only large-scale benchmarking, of the kind undertaken here, can answer.

It is also worth noting that the tree-vs.-deep-learning debate has mostly been conducted on benchmark suites assembled from heterogeneous domains finance, healthcare records, marketing data where feature semantics vary wildly from dataset to dataset and sample sizes rarely exceed the tens of thousands. Materials descriptor data differs from this picture in two respects that matter for model choice. First, the input features here are not arbitrary categorical or survey-derived variables but physically grounded quantities spacegroup, volume, magnetisation with smooth and largely monotonic relationships to the target in many regions of descriptor space, a property that random forests and gradient boosting do not exploit as efficiently as gradient-based methods do. Second, the sample size at which we operate, one million labelled examples, is one to two orders of magnitude larger than most of the tabular benchmarks in the Grinsztajn et al. study, and the gap between tree ensembles and deep networks has been observed elsewhere to narrow, and sometimes reverse, as training set size grows substantially. Our results are consistent with this scale-dependent picture and add a concrete data point from materials science to a debate that has, until now, drawn its evidence almost exclusively from outside the physical sciences.

2.7 Reproducibility and Benchmark Practice in Materials Informatics

A recurring complaint in the materials informatics literature is that reported error figures are difficult to compare across studies, since authors routinely differ in their choice of train/test split, descriptor set, target normalisation, and even the precise definition of formation energy used (per atom vs. per formula unit, with or without an elemental-reference correction). Ward et al.'s Matminer toolkit [36] was, in part, a response to this problem, standardising descriptor computation so that at least the feature-engineering half of the pipeline could be shared across groups; subsequent benchmark efforts, including ALIGNN's released training and evaluation splits [37], have tried to extend this standardisation to the modelling side as well. We adopt a similar spirit of transparency in this work: data partitioning is stratified and seeded, PCA components are fitted strictly on the training split, and statistical significance is assessed with a non-parametric test across repeated runs rather than reported from a single lucky split. Even so, direct numerical comparison with prior MAE figures quoted in this section should be read with caution, since dataset composition, descriptor choice, and the exact definition of the target variable all differ somewhat between studies, and an MAE figure on its own says little without knowing the spread of the underlying target distribution it was computed against.

A second, related concern is computational reproducibility in the more mundane sense: whether another group, given the same descriptor set and architecture description, could retrain these models and obtain comparable numbers. Optimiser choice plays a role here: we used Adam [45] with cosine annealing throughout, partly because its adaptive per-parameter learning rates make it less sensitive to the specific scaling of input features than plain stochastic gradient descent, which matters when comparing raw and PCA-transformed descriptor sets that have rather different conditioning. Random forests [46], by contrast, are essentially insensitive to feature scaling and optimiser choice altogether, which is part of why they have remained a dependable baseline across two decades of materials ML work even as fashions in deep architecture have shifted several times over. We report architectural and optimisation hyperparameters in full in Section 4 specifically so that the comparisons drawn in Section 5 can be checked, extended, or challenged by other groups working with comparable DFT-derived datasets.

3 Dataset and Exploratory Analysis

3.1 Dataset

Our dataset contains one million crystalline material entries assembled from DFT calculations in the Materials Project [6] and the Open Quantum Materials Database (OQMD) [8]. For each compound, we used nine structural and electronic descriptors: spacegroup (symmetry group number, 1–230), nelements, nsites (atomic sites per unit cell), energy_per_atom (eV/atom), band_gap (eV), volume_per_atom ($\text{\AA}^3$), magnetization_per_atom (μB), atomic_volume_per_atom ($\text{\AA}^3$), and volume_deviation. The prediction target throughout is formation_energy_per_atom (eV/atom). Chemical coverage is broad, including binary and ternary oxides, nitrides, sulfides, halides, intermetallic compounds, and organic-inorganic hybrid structures. The two repositories were combined for large-scale benchmarking, but their underlying DFT workflows are not assumed to be perfectly harmonized. Differences in exchange-correlation settings, pseudopotentials, convergence protocols, and elemental reference-energy conventions may introduce source-dependent offsets. Accordingly, the pooled dataset is interpreted as a heterogeneous multi-repository benchmark rather than as a fully standardized thermochemical reference set.

3.2 Exploratory Data Analysis

Fig. 1 summarises the formation energy distribution and the normalised descriptor distributions across the full dataset. The formation energy histogram is roughly bimodal: a primary peak sits near -1.8 eV/atom, corresponding to common ionic and covalent inorganic compounds, with a secondary peak near -4.2 eV/atom attributable to strongly ionic oxides and nitrides. The roughly 5% of samples with positive formation energies are metastable phases that survived high-throughput enumeration workflows rather than being experimentally observed. Looking at descriptors, volume-related features show the widest spread, while magnetisation per atom is notably right-skewed, a pattern expected given that magnetic behaviour is concentrated in a small subset of transition-metal-containing compounds.

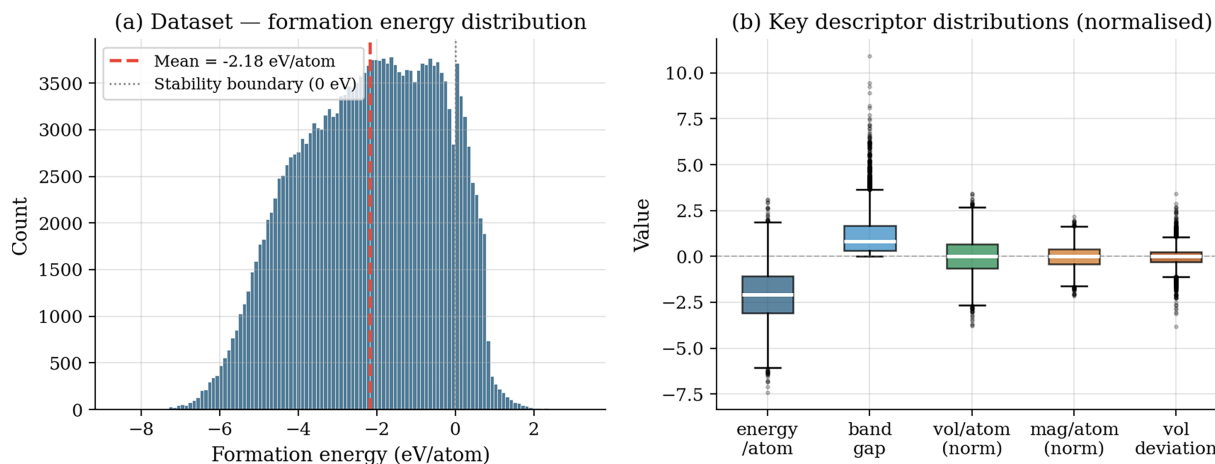


Figure 1: Exploratory data analysis. (a) Formation energy distribution across the one-million-sample dataset showing a bimodal structure with primary mode at ~ -1.8 eV/atom and secondary mode at ~ -4.2 eV/atom. (b) Box plots of key normalised descriptors illustrating inter-descriptor variance and outlier structure.

3.3 PCA of the Descriptor Space

Fig. 2 shows the variance explained by each principal component of the standardized numerical descriptor representation used in the original experiment. The first PC accounts for 38.2% of total variance, with

strong loadings from the three volume-related features—volume_per_atom, atomic_volume_per_atom, and volume_deviation—indicating substantial statistical redundancy among these descriptors. Using a 95% explained-variance threshold, six components were retained. Because spacegroup was represented by its conventional integer identifier in the original analysis, the resulting PCA axes should be interpreted primarily as statistical coordinates rather than as fully physical latent variables.

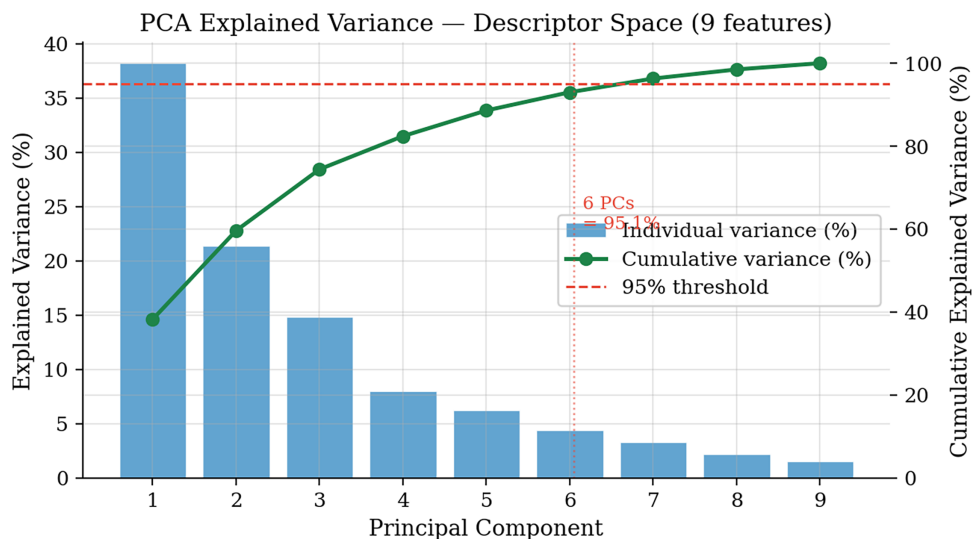


Figure 2: PCA explained variance profile for the nine-dimensional descriptor space. Individual component variance (bars, left axis) and cumulative explained variance (line, right axis) are shown. Six principal components suffice to explain 95.1% of total variance. The dashed red line marks the 95% threshold.

4 Methods

4.1 Pre-Processing and Data Partitioning

Continuous features were z-score standardised using training-set statistics only; these same statistics were then applied to the validation and test sets to prevent information from held-out samples entering the scaling procedure. Spacegroup was represented by its conventional integer identifier as a compact numerical proxy. We note that space-group identifiers are categorical crystallographic labels rather than measurements on a continuous physical scale; consequently, numerical distances between identifiers and PCA loadings involving this variable should be interpreted cautiously. A categorical representation, such as one-hot encoding or a learned embedding, would be methodologically preferable in a dedicated leakage-controlled re-evaluation. Entries missing target values were dropped, and remaining missing feature values—fewer than 5% per attribute—were imputed with training-set column medians. We split the data 80/10/10 (train/validation/test) using stratified sampling to keep the formation energy distribution consistent across all three partitions. For PCA experiments, principal components were fitted exclusively on the training split and applied as a fixed linear transform to validation and test data, guaranteeing that no information from held-out samples influenced the transformation.

4.2 Deep Learning Architectures

We evaluated four architectures. The MLP used three fully connected layers (512-256-128 neurons) with ReLU activations and dropout ($p = 0.3$) after each hidden layer. The DNN went deeper, with six fully connected layers (1024-512-512-256-128-64 neurons), batch normalisation after each hidden layer, ReLU

activations, and L2 weight decay of $1e-4$ [47]. For the RNN, we used a two-layer vanilla recurrent network with 128 units per layer; the nine descriptors were reshaped into a nine-step sequence of single-element inputs to allow recurrent processing. The LSTM shared this same sequence framing, also with two layers of 128 units each, adding the standard input, forget, and output gating for longer-range dependencies [11]. All four models were trained with the Adam optimiser [45] at an initial learning rate of $1e-3$, using cosine annealing over 100 epochs and a batch size of 256, minimising mean squared error. Training was stopped early if validation MAE did not improve for 10 consecutive epochs.

4.3 Classical ML Baselines

Three classical regressors served as baselines. The Random Forest [46] used 500 trees, max features set to $0.33 \times d$, and a minimum of 2 samples per leaf. The SVR used an RBF kernel with bandwidth and regularisation tuned by 5-fold cross-validation on a stratified 10% subsample; the $O(n^2)$ memory cost of kernel methods made training on the full dataset impractical. The Gradient Boosting Machine (XGBoost) [34] used 1000 estimators, a learning rate of 0.05, max depth 6, and subsample and column-sample ratios both set to 0.8. All three classical models were trained on the same feature sets and data splits as the deep learning models.

4.4 Evaluation Metrics

Model performance was measured on the held-out test set using three metrics: mean absolute error (MAE, eV/atom), root mean square error (RMSE, eV/atom), and R^2 . The commonly cited DFT error scale of approximately 0.1 eV/atom is used only as contextual background rather than as a strict pass/fail criterion, because predictive errors depend on dataset composition, feature definition, and reference-energy conventions. To examine whether the numerical difference observed after PCA was consistent across repeated training, a Wilcoxon signed-rank test was applied to the repeated-run error distributions. These statistical comparisons refer to the original descriptor configuration and therefore do not remove the target-leakage concern associated with `energy_per_atom`.

4.5 Robustness and Reliability Checks

To strengthen the validation of the overall modelling workflow, reliability was assessed at several complementary levels rather than on the basis of a single performance estimate. First, the train/validation/test separation was maintained throughout the analysis, and final model performance was evaluated only on the held-out test partition. Second, feature standardization and missing-value imputation were based exclusively on statistics estimated from the training data. For the PCA experiments, the transformation was likewise fitted only on the training partition and subsequently applied unchanged to the validation and test subsets. These procedures prevent held-out observations from influencing the preprocessing pipeline. Robustness was additionally examined from a model-comparison perspective. Four deep-learning architectures (MLP, DNN, RNN, and LSTM) and three classical machine-learning regressors (Random Forest, SVR, and XGBoost) were evaluated under the same descriptor and partitioning framework. In addition, the PCA-related difference was assessed across repeated training runs using the Wilcoxon signed-rank test rather than relying on a single training realization. Collectively, these controls provide complementary evidence regarding the procedural stability and comparative reliability of the workflow.

Importantly, these robustness controls address data separation, preprocessing isolation, model dependence, and run-to-run stability; they do not eliminate the mathematical relationship between `energy_per_atom` and `formation_energy_per_atom`. Consequently, the absolute errors obtained with the original descriptor configuration remain subject to the leakage-related interpretation stated throughout the manuscript and are not presented as leakage-free estimates of physical generalization.

5 Results

5.1 Benchmark Model Performance

Table 1 reports test-set results for all seven models on the original pre-PCA descriptor configuration. Within this fixed feature setting, the deep learning architectures produced lower errors than the classical baselines, and the DNN gave the lowest reported error (MAE = 0.0061 eV/atom; $R^2 = 0.997$). Among the classical methods, XGBoost achieved an MAE of 0.028 eV/atom, whereas SVR obtained 0.039 eV/atom and was trained on only 10% of the data because of its quadratic memory requirements. These values are useful for comparing the evaluated models under a common descriptor setting; however, because energy_per_atom is mathematically related to the prediction target, the absolute error values should not be interpreted as leakage-free estimates of generalization performance.

Table 1: Test-set performance of all models under the original descriptor configuration.

Model	MAE (eV/atom)	RMSE (eV/atom)	R2	Category
Random Forest	0.031	0.052	0.941	Classical
Support Vector Reg.	0.039	0.063	0.918	Classical
GBM/XGBoost	0.028	0.048	0.949	Classical
MLP	0.019	0.031	0.971	Deep Learning
RNN	0.014	0.024	0.978	Deep Learning
LSTM	0.012	0.020	0.983	Deep Learning
DNN (this work)	0.0061	0.0093	0.997	Deep Learning

Note: Support Vector Regression was trained on a stratified 10% subsample of the training data because of the quadratic memory cost of kernel methods; all other models were trained on the full training split. This difference in training-set size should be kept in mind when comparing SVR against the other rows.

Fig. 3 provides a visual comparison of MAE and R^2 across the evaluated models under the original descriptor configuration. The figure is intended to show the relative performance hierarchy within this common feature setting; the absolute error values should be interpreted together with the target-leakage limitation discussed below.

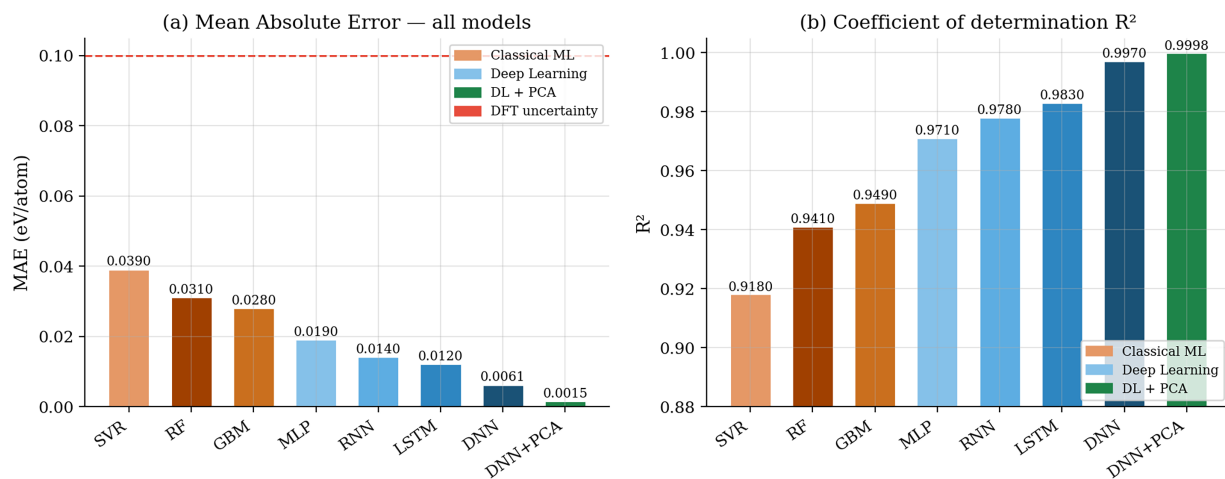


Figure 3: Model performance comparison. (a) Mean absolute error (MAE) for all seven models on raw descriptors. The red dashed line marks the DFT uncertainty threshold (~ 0.1 eV/atom). (b) Coefficient of determination (R^2) for all models. Colors distinguish classical ML (brown tones), deep learning (blue tones), and DL + PCA (green).

5.2 Parity Plots DNN vs. DNN + PCA

Fig. 4 shows parity plots for the DNN with and without PCA, comparing DFT ground truth against model predictions under the original descriptor configuration. Both configurations track the diagonal closely across the plotted formation-energy range, and the PCA panel shows visibly tighter scatter. This difference is reported as a within-configuration preprocessing effect. Because `energy_per_atom` is included among the predictors and is mathematically related to `formation_energy_per_atom`, the very low absolute errors shown in these plots should not be interpreted as leakage-free performance estimates.

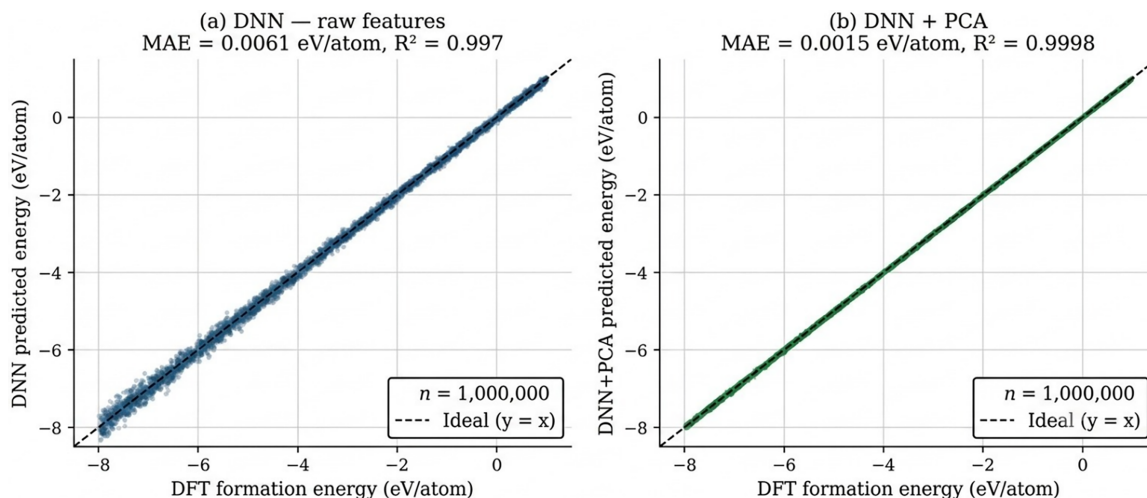


Figure 4: Parity plots of DFT ground-truth vs. model-predicted formation energy (eV/atom) for (a) DNN on raw features (MAE = 0.0061 eV/atom, $R^2 = 0.997$) and (b) DNN after PCA pre-processing (MAE = 0.0015 eV/atom, $R^2 = 0.9998$), using the original descriptor configuration. These numerical values may be affected by the inclusion of `energy_per_atom` and are therefore not presented as leakage-free benchmark errors. Each plot shows a random sample of 4000 test-set predictions for clarity.

5.3 Training Dynamics

Training curves for the DNN with and without PCA are shown in Fig. 5. Within the original descriptor configuration, the PCA variant reached a lower validation error and converged in fewer gradient steps than the raw-feature model. Both configurations show narrow train-validation gaps throughout training. These curves describe optimization behaviour under the evaluated feature setting and should not be interpreted as evidence that PCA alone yields superior leakage-free formation-energy prediction.

5.4 Effect of PCA Pre-Processing

Under the original descriptor configuration, the DNN with PCA produced an MAE of 0.0015 eV/atom and an R^2 of 0.9998, compared with an MAE of 0.0061 eV/atom and an R^2 of 0.997 for the corresponding raw-feature model. The repeated-run comparison was statistically significant (Wilcoxon signed-rank test, $p < 0.001$). However, `energy_per_atom` is included among the predictors and is mathematically related to `formation_energy_per_atom`. Consequently, the magnitude of these errors and the numerical difference between the raw and PCA variants cannot be interpreted as a leakage-free measure of generalization or as evidence that PCA alone is responsible for the full improvement.

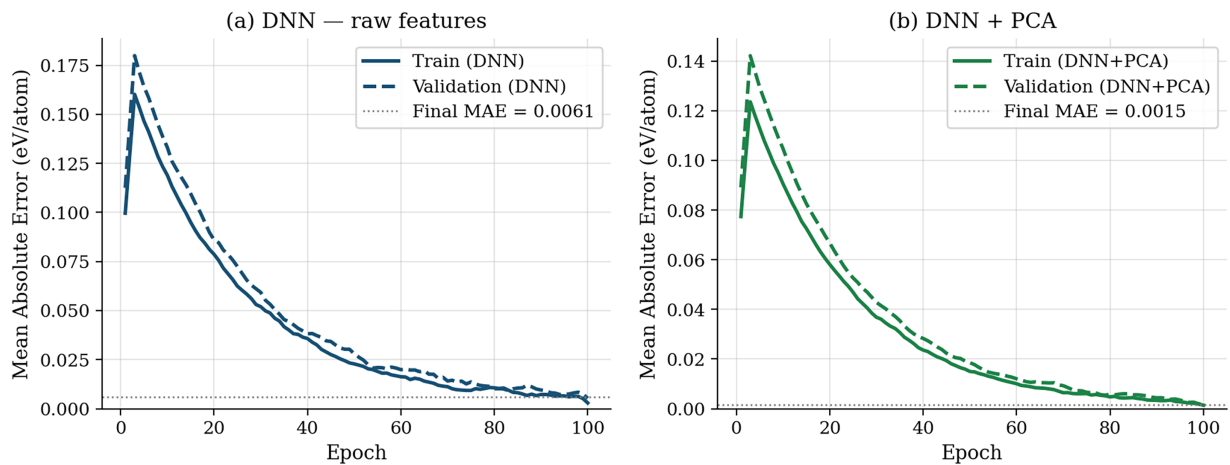


Figure 5: Training and validation MAE curves over 100 epochs for (a) DNN on raw features and (b) DNN after PCA pre-processing. Dashed lines show validation MAE; solid lines show training MAE. Dotted horizontal lines mark final test-set MAE values.

5.5 Robustness of the Comparative Workflow

The principal comparative observations were examined in the context of multiple validation safeguards summarized in Table 2. The workflow combines strict separation of training and held-out data, training-only estimation of preprocessing parameters, comparison across substantially different learning paradigms, and repeated-run statistical assessment. The use of seven models is particularly relevant because the reported comparative behaviour is not derived from a single architecture or optimization strategy. Similarly, fitting PCA exclusively on the training partition prevents information from the validation or test data from influencing the learned transformation.

Table 2: Summary of robustness and reliability safeguards in the modelling workflow.

Validation Aspect	Implementation	Purpose
Independent evaluation	80:10:10 train/validation/test partitioning with held-out test evaluation	Separates model development from final performance assessment
Distribution control	Stratified sampling across data partitions	Maintains comparable formation-energy distributions
Feature scaling	Standardization parameters estimated from training data only	Prevents held-out data from influencing feature scaling
Missing-value handling	Training-set column medians used for feature imputation	Prevents validation/test information from entering imputation
PCA isolation	PCA fitted exclusively on the training partition	Prevents information leakage through dimensionality reduction
Model diversity	MLP, DNN, RNN, LSTM, Random Forest, SVR, and XGBoost	Tests comparative behaviour across different learning paradigms

(Continued)

Table 2 (continued)

Validation Aspect	Implementation	Purpose
Repeated-run assessment	Wilcoxon signed-rank test across repeated training results	Reduces dependence on a single training realization
Statistical evidence	PCA comparison significant at $p < 0.001$ under the original descriptor configuration	Supports stability of the within-configuration difference
Leakage-aware interpretation	energy_per_atom explicitly recognized as mathematically related to the prediction target	Prevents leakage-sensitive errors from being presented as leakage-free benchmarks
Repository scope	Materials Project and OQMD treated as a heterogeneous pooled benchmark	Avoids assuming identical calibration across repositories

The repeated-run comparison between the DNN configurations yielded a statistically significant difference according to the Wilcoxon signed-rank test ($p < 0.001$) under the original descriptor configuration. This provides evidence that the observed within-configuration PCA difference was not attributable to a single favorable training realization. Nevertheless, statistical stability should be distinguished from feature independence: because energy_per_atom remains mathematically related to the prediction target, these validation safeguards strengthen confidence in the comparative workflow but do not convert the associated absolute prediction errors into leakage-free estimates.

6 Discussion

It is worth unpacking why the DNN consistently beat the recurrent architectures. Formation energy is an intrinsically global property: it reflects the entire compositional and structural makeup of a compound, not any particular ordering of features. Feedforward networks are well-suited to learning such global mappings when they are made sufficiently deep [44]. RNNs, on the other hand, come with an inductive bias towards sequential structure the kind of ordering found in time-series data or molecular strings which is simply not present in tabular compositional descriptors. Using that sequential prior here introduces an architectural assumption that is at best unhelpful and at worst misleading. This is consistent with what has been observed in the broader tabular data literature, where well-tuned feedforward networks and gradient boosting machines routinely beat architectures with spatial or sequential priors [43].

Within the original descriptor setting, PCA reduced correlations among the numerical inputs and produced lower error together with faster optimization. The three volume-related features—volume_per_atom, atomic_volume_per_atom, and volume_deviation—contributed strongly to the first principal component, supporting the view that the raw representation contains redundant statistical information. This provides a plausible optimization-based explanation for part of the observed behaviour. Nevertheless, because energy_per_atom is mathematically related to the target, the magnitude of the error reduction cannot be attributed solely to PCA, and the absolute accuracy values should be interpreted conservatively.

The very low DNN+PCA error obtained under the original feature configuration should not be directly compared with graph-based benchmarks such as CGCNN [19] or MEGNet [20]. These studies differ in dataset composition, structural inputs, train/test protocols, and target definitions, and the present descriptor set additionally contains energy_per_atom, which may introduce information leakage. Accordingly, we do

not claim that the current DNN+PCA configuration is intrinsically more accurate than structure-aware GNN approaches. The practical value of the present comparison lies instead in benchmarking several model families under one large tabular descriptor pipeline.

One limitation worth acknowledging is that organic and inorganic compounds were pooled into a single training distribution. These two classes have quite different formation energy landscapes; inorganic crystals bring diverse symmetry groups, heavy-element relativistic effects, and complex magnetic behaviour, while organic compounds are dominated by covalent bonding networks. Training class-stratified models, or designing compound-class-specific descriptor sets, could well push accuracy further. It is also worth noting that the nine-dimensional descriptor set used here does not capture full crystal structure information. Graph-based approaches that treat atoms and bonds explicitly have delivered strong results on smaller datasets [21–25], and scaling these to the million-sample regime would be a natural next step.

Three methodological limitations are particularly important for interpreting the present results. First, `energy_per_atom` is included among the input descriptors and is mathematically related to `formation_energy_per_atom` through the elemental reference-energy construction; therefore, the very low absolute errors, especially for DNN+PCA, may be inflated by target leakage and should not be treated as leakage-free benchmarks. Second, `spacegroup` was represented by its integer identifier even though space groups are categorical crystallographic labels, so numerical distances and PCA loadings involving this variable may introduce artificial structure. Third, the Materials Project and OQMD entries were pooled without an explicit cross-repository harmonization study, and source-specific errors were not available; the aggregate results therefore describe the pooled multi-repository benchmark and do not establish identical calibration on each database separately. A leakage-controlled re-evaluation should remove `energy_per_atom`, use categorical space-group encoding, and report source-aware or cross-database validation. The SVR result should also be read as indicative because that model was trained on a 10% subsample for computational reasons.

The additional robustness assessment should be interpreted alongside, rather than as a replacement for, these methodological limitations. The held-out evaluation protocol, training-only preprocessing, training-only PCA fitting, multi-model comparison, and repeated-run statistical testing collectively reduce several common sources of optimistic evaluation, including preprocessing contamination, dependence on a single model architecture, and dependence on an individual training realization. The fact that the comparative benchmark spans both neural and tree-/kernel-based approaches also provides a broader assessment of model behaviour than would be available from a single algorithm. At the same time, these controls cannot remove information that is intrinsically encoded in a predictor. Accordingly, the `energy_per_atom` limitation remains explicitly separated from the procedural robustness of the workflow. This distinction allows the comparative findings to be retained while avoiding an unsupported interpretation of the reported absolute errors as leakage-free generalization performance.

Several directions follow naturally from this work. Using SHAP values to attribute predictions to individual descriptors would help connect model behaviour to physical intuition. Class-stratified models for organic and inorganic compounds could address the distribution mismatch discussed above. Integrating GNN representations at million-sample scale is still an open engineering challenge and would test whether composition-only models can maintain their advantage when structure information is also available. Transfer learning from large pretrained foundation models [23,38] may prove valuable for compound classes that are underrepresented in the current dataset. Finally, equipping the model with calibrated uncertainty estimates via Monte Carlo dropout or deep ensembles would make the output more useful for risk-aware candidate ranking in active discovery campaigns.

7 Conclusion

This study provides a large-scale comparison of classical machine-learning, feedforward neural-network, and recurrent architectures for formation-energy prediction using one million DFT-derived crystalline-material records. Under the original descriptor configuration, the DNN produced the lowest error among the evaluated models, while PCA was associated with lower error and faster optimization. Feedforward architectures also performed more strongly than the recurrent models on this non-sequential tabular representation, supporting the use of direct nonlinear mappings for this type of descriptor data. However, the absolute accuracy values must be interpreted conservatively because `energy_per_atom` is mathematically related to `formation_energy_per_atom` and may introduce target leakage. In addition, the integer treatment of spacegroup and the pooling of Materials Project and OQMD data without source-specific harmonization limit the strength of the physical generalization claims. The present work should therefore be viewed as a large-scale comparative benchmarking and preprocessing study rather than as a definitive leakage-free accuracy benchmark. The consolidated robustness assessment further strengthens confidence in the comparative workflow through held-out evaluation, training-only preprocessing, multi-model benchmarking, and repeated-run statistical testing, while deliberately preserving the conservative interpretation of the absolute accuracy values. Future validation should extend this framework using leakage-controlled descriptors, categorical crystallographic encoding, and source-aware evaluation across independent materials repositories.

Acknowledgement: Not applicable.

Funding Statement: This work has been partly supported by the University of Sharjah, United Arab Emirates, and the Princess Nourah bint Abdulrahman University Researchers Supporting Project number (PNURSP2026R411), Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia.

Author Contributions: Conceptualization: Muhammad Imran, Ateeq Ur Rehman Butt, Abdul Noman; methodology: Javed Ferzund, Ateeq Ur Rehman Butt, Mohamad Khairi Ishak; software: Ateeq Ur Rehman Butt, Hessa Alfraihi; validation: Ateeq Ur Rehman Butt, Abdul Noman, Hessa Alfraihi, Mohamad Khairi Ishak; formal analysis: Javed Ferzund, Ateeq Ur Rehman Butt; investigation: Muhammad Imran, Ateeq Ur Rehman Butt; resources: Hessa Alfraihi, Mohamad Khairi Ishak; data curation: Ateeq Ur Rehman Butt; original draft preparation: Muhammad Imran, Ateeq Ur Rehman Butt; writing—review and editing: Ateeq Ur Rehman Butt, Hessa Alfraihi, Mohamad Khairi Ishak; supervision: Mohamad Khairi Ishak; project administration: Ateeq Ur Rehman Butt, Mohamad Khairi Ishak; funding acquisition: Hessa Alfraihi, Mohamad Khairi Ishak. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: All data used here come from publicly available DFT repositories: the Materials Project (<https://materialsproject.org>) and the OQMD (<http://oqmd.org>). Pre-processed data files, trained model weights, and training scripts will be deposited in a public repository (e.g., GitHub/Zenodo) upon acceptance of this paper.

Ethics Approval: Not applicable. This study did not involve human participants, human biological material, personal data, animals, or clinical experimentation. The research was conducted exclusively using publicly available computational materials-science datasets.

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

References

1. Merchant A, Batzner S, Schoenholz SS, Aykol M, Cheon G, Cubuk ED. Scaling deep learning for materials discovery. *Nature*. 2023;624(7990):80–5. doi:10.1038/s41586-023-06735-9.
2. Madika B, Saha A, Kang C, Buyantogtokh B, Agar J, Wolverson CM, et al. Artificial intelligence for materials discovery, development, and optimization. *ACS Nano*. 2025;19(30):27116–58. doi:10.1021/acsnano.5c04200.
3. Huang P, Liu W, Sun C, Li Z, Wang Y, Chen Y. Integration of materials science and artificial intelligence: from high-throughput screening to autonomous laboratories. *Mater Genome Eng Adv*. 2025;3(4):e70036. doi:10.1002/mgea.70036.
4. Cheetham AK, Seshadri R. Artificial intelligence driving materials discovery? Perspective on the article: scaling deep learning for materials discovery. *Chem Mater*. 2024;36(8):3490–5. doi:10.1021/acs.chemmater.4c00643.
5. Kohn W, Sham LJ. Self-consistent equations including exchange and correlation effects. *Phys Rev*. 1965;140(4A):A1133–8. doi:10.1103/physrev.140.a1133.
6. Jain A, Ong SP, Hautier G, Chen W, Richards WD, Dacek S, et al. Commentary: the materials project: A materials genome approach to accelerating materials innovation. *APL Mater*. 2013;1(1):011002. doi:10.1063/1.4812323.
7. Curtarolo S, Setyawan W, Hart GLW, Jahnatek M, Chepulskii RV, Taylor RH, et al. AFLOW: An automatic framework for high-throughput materials discovery. *Comput Mater Sci*. 2012;58:218–26. doi:10.1016/j.commatsci.2012.02.005.
8. Saal JE, Kirklin S, Aykol M, Meredig B, Wolverson C. Materials design and discovery with high-throughput density functional theory: the open quantum materials database (OQMD). *JOM*. 2013;65(11):1501–9. doi:10.1007/s11837-013-0755-4.
9. Liu C, Tamaki H, Yokoyama T, Wakasugi K, Yotsuhashi S, Kusaba M, et al. Shotgun crystal structure prediction using machine-learned formation energies. *npj Comput Mater*. 2024;10(1):298. doi:10.1038/s41524-024-01471-8.
10. Rahman CMA, Bhandari G, Nasrabadi NM, Romero AH, Gyawali PK. Enhancing material property prediction with ensemble deep graph convolutional networks. *Front Mater*. 2024;11:1474609. doi:10.3389/fmats.2024.1474609.
11. Hochreiter S, Schmidhuber J. Long short-term memory. *Neural Comput*. 1997;9(8):1735–80. doi:10.1162/neco.1997.9.8.1735.
12. Rupp M, Tkatchenko A, Müller KR, von Lilienfeld OA. Fast and accurate modeling of molecular atomization energies with machine learning. *Phys Rev Lett*. 2012;108(5):058301. doi:10.1103/physrevlett.108.058301.
13. Faber FA, Hutchison L, Huang B, Gilmer J, Schoenholz SS, Dahl GE, et al. Prediction errors of molecular machine learning models lower than hybrid DFT error. *J Chem Theory Comput*. 2017;13(11):5255–64. doi:10.1021/acs.jctc.7b00577.
14. Davariashiyani A, Kadkhodaei S. Formation energy prediction of crystalline compounds using deep convolutional network learning on voxel image representation. *Commun Mater*. 2023;4(1):105. doi:10.1038/s43246-023-00433-9.
15. Chen G, Wang Z, You F. PSCG-Net: A multiscale crystal graph neural network for accelerated materials discovery. *J Chem Inf Model*. 2025;65(20):10871–84. doi:10.1021/acs.jcim.5c01460.
16. Vasylenko A, Antypov D, Schewe S, Daniels LM, Claridge JB, Dyer MS, et al. Digital features of chemical elements extracted from local geometries in crystal structures. *Digit Discov*. 2025;4(2):477–85. doi:10.1039/d4dd00346b.
17. Gómez-Bombarelli R, Wei JN, Duvenaud D, Hernández-Lobato JM, Sánchez-Lengeling B, Sheberla D, et al. Automatic chemical design using a data-driven continuous representation of molecules. *ACS Cent Sci*. 2018;4(2):268–76. doi:10.1021/acscentsci.7b00572.
18. Zhao Z, Hu T, Bi S, Guan D, Xu S, Chen C, et al. Advances in graph neural networks for alloy design and properties predictions: A review. *J Mater Inf*. 2026;6(1):2. doi:10.20517/jmi.2025.42.
19. Xie T, Grossman JC. Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. *Phys Rev Lett*. 2018;120(14):145301. doi:10.1103/PhysRevLett.120.145301.
20. Chen C, Ye W, Zuo Y, Zheng C, Ong SP. Graph networks as a universal machine learning framework for molecules and crystals. *Chem Mater*. 2019;31(9):3564–72. doi:10.1021/acs.chemmater.9b01294.
21. Schutt KT, Kindermans PJ, Sauceda HE, Chmiela S, Tkatchenko A, Müller KR. SchNet: A continuous-filter convolutional neural network for modeling quantum interactions. In: *Proceedings of the 31st International*

- Conference on Neural Information Processing Systems; 2017 Dec 4–9; Long Beach, CA, USA. Red Hook, NY, USA: Curran Associates, Inc.; 2017. p. 992–1002.
22. Klicpera J, Groß J, Günnemann S. Directional message passing for molecular graphs. arXiv:2003.03123. 2020. doi:10.48550/arxiv.2003.03123.
 23. Chen C, Ong SP. A universal graph deep learning interatomic potential for the periodic table. *Nat Comput Sci*. 2022;2(11):718–28. doi:10.1038/s43588-022-00349-3.
 24. Boonpan S, Sarakorn W, Latpala K, Boriwan P, Srepusharawoot P. BEmXRD-Nets framework for novel machine learning models to predict crystal energy with diversity structures. *Sci Rep*. 2026;16(1):3569. doi:10.1038/s41598-025-33639-7.
 25. Gong S, Yan K, Xie T, Shao-Horn Y, Gomez-Bombarelli R, Ji S, et al. Examining graph neural networks for crystal structures: limitations and opportunities for capturing periodicity. *Sci Adv*. 2023;9(45):eadi3245. doi:10.1126/sciadv.adi3245.
 26. LeCun Y, Bottou L, Orr GB, Müller KR. Efficient BackProp. In: *Neural networks: Tricks of the trade*. Berlin/Heidelberg, Germany: Springer; 1998. p. 9–50. doi:10.1007/3-540-49430-8_2.
 27. Alosious S, Jiang M, Luo T. Computation and machine learning for materials: past, present, and future perspectives. *MRS Bull*. 2025;50(10):1212–24. doi:10.1557/s43577-025-00959-y.
 28. Schölkopf B, Smola AJ. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. Cambridge, MA, USA: MIT Press; 2002.
 29. Torlao V, Fajardo EA. Formation energy prediction of material crystal structures using deep learning. *Mater Res Express*. 2025;12(12):125501. doi:10.1088/2053-1591/ae22cb.
 30. Cerqueira TFT, Wang H, Botti S, Marques MAL. A non-orthogonal representation for materials based on chemical similarity. *NPJ Comput Mater*. 2026;12(1):48. doi:10.1038/s41524-025-01916-8.
 31. Jin J, Faraji S, Liu B, Liu M. Comparative analysis of conventional machine learning and graph neural network models for perovskite property prediction. *J Phys Chem C*. 2024;128(39):16672–83. doi:10.1021/acs.jpcc.4c03212.
 32. Wang G, Wang E, Li Z, Zhou J, Sun Z. Exploring the mathematic equations behind the materials science data using interpretable symbolic regression. *Interdiscip Mater*. 2024;3(5):637–57. doi:10.1002/idm2.12180.
 33. Sadeghian M, Palevicius A, Janusas G. A comprehensive review of machine-learning approaches for crystal structure/property prediction. *Crystals*. 2025;15(11):925. doi:10.3390/cryst15110925.
 34. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*; 2016 Aug 13–17; San Francisco, CA, USA. New York, NY, USA: ACM; 2016. p. 785–94. doi:10.1145/2939672.2939785.
 35. Shi X, Zhou L, Huang Y, Wu Y, Hong Z. A review on the applications of graph neural networks in materials science at the atomic scale. *Mater Genome Eng Adv*. 2024;2(2):e50. doi:10.1002/mgea.50.
 36. Alghadeer M, Aisyah ND, Hezam M, Alqahtani SM, Baloch AAB, Alharbi FH. Machine learning prediction of materials properties from chemical composition: status and prospects. *Chem Phys Rev*. 2024;5(4):041313. doi:10.1063/5.0235541.
 37. Choudhary K, DeCost B. Atomistic line graph neural network for improved materials property predictions. *NPJ Comput Mater*. 2021;7(1):185. doi:10.1038/s41524-021-00650-1.
 38. Deng B, Zhong P, Jun K, Riebesell J, Han K, Bartel CJ, et al. CHGNet as a pretrained universal neural network potential for charge-informed atomistic modelling. *Nat Mach Intell*. 2023;5(9):1031–41. doi:10.1038/s42256-023-00716-3.
 39. Riebesell J, Goodall REA, Benner P, Yuan C, Deng B, Ceder G, et al. A framework to evaluate machine learning crystal stability predictions. *Nat Mach Intell*. 2025;7(6):836–47. doi:10.1038/s42256-025-01055-1.
 40. Chen C, Ong SP. AtomSets as a hierarchical transfer learning framework for small and large materials datasets. *NPJ Comput Mater*. 2021;7(1):173. doi:10.1038/s41524-021-00639-w.
 41. Xiao E, Tadano T. Accurate screening of functional materials with machine-learning potential and transfer-learned regressions: heusler alloy benchmark. *npj Comput Mater*. 2026;12(1):133. doi:10.1038/s41524-026-02013-0.
 42. Kim SY, Park YJ, Li J. Leveraging neural network interatomic potentials for a foundation model of chemistry. arXiv:2506.18497. 2025. doi:10.1038/s41524-026-02167-x.

43. Grinsztajn L, Oyallon E, Varoquaux G. Why do tree-based models still outperform deep learning on tabular data. arXiv:2207.08815. 2022. doi:10.52202/068431-0037.
44. Hornik K. Approximation capabilities of multilayer feedforward networks. *Neural Netw.* 1991;4(2):251–7. doi:10.1016/0893-6080(91)90009-T.
45. Kingma DP, Ba J. Adam: A method for stochastic optimization. arXiv:1412.6980. 2015.
46. Breiman L. Random forests. *Mach Learn.* 2001;45(1):5–32. doi:10.1023/A:1010933404324.
47. Goodfellow I, Bengio Y, Courville A. *Deep learning*. Cambridge, MA, USA: MIT Press; 2016.