Computational Condensed Matter 48 (2026) e01422

Contents lists available at ScienceDirect

Computational Condensed Matter

journal homepage: www.elsevier.com/locate/cocom

Full length article

Data-driven screening of stable lead-free piezoelectric materials via machine learning and graph-based models

Francesc Suñol [a,*], Claudio Cazorla [a,b,c], José E. Garcia [a]

[a] Department of Physics, Universitat Politècnica de Catalunya – BarcelonaTech, Barcelona, 08034, Spain
[b] Research Center in Multiscale Science and Engineering, Universitat Politècnica de Catalunya, Campus Diagonal-Besòs, Av. Eduard Maristany 10–14, Barcelona, 08019, Spain
[c] Institució Catalana de Recerca i Estudis Avançats (ICREA), Passeig Lluís Companys 23, Barcelona, 08010, Spain

ARTICLE INFO

Keywords:
Piezoelectrics
Machine learning
Graph neural networks
Lead-free
Screening

ABSTRACT

The identification of lead-free, stable piezoelectric materials remains a key challenge for sustainable technologies. Although first-principles calculations based on density functional theory (DFT) provide reliable property predictions, their high computational cost prevents their direct use in large-scale screening. Here we present a systematic machine learning (ML) benchmark for the prediction of the maximum piezoelectric modulus e_m of inorganic crystalline materials, trained on >3,000 DFT-computed compounds from the Materials Project. We evaluate thirteen predictive architectures spanning gradient-boosted tree ensembles, deep neural networks, and seven graph neural network (GNN) models. The three top-performing models are then used to screen promising piezoelectric materials among >10,000 unlabelled Materials Project compounds. After filtering for thermodynamic near-stability (E_{hull} < 0.1 eV/atom), insulating behaviour (E_g > 0.1 eV), and non-centrosymmetric crystal structures, eight promising lead-free candidates are identified. Independent DFT calculations of the full piezoelectric tensor for twenty selected compounds confirm the screening results. Among the identified materials, $SrTaNO_2$, YWN_3, and $SrTa_2Bi_2O_9$ emerge as the most promising targets for experimental synthesis as they exhibit large piezoelectric moduli in the range of 3–103 C/m^2.

1. Introduction

Piezoelectric materials convert mechanical stress into electrical charge (direct piezoelectric effect) and deform under an applied electric field (converse effect). These coupled electromechanical properties underpin a broad and growing range of technologies: ultrasonic medical imaging, non-destructive evaluation, MEMS-based inertial sensors, nanopositioning actuators, sonar systems, and piezoelectric motors among others [1–4].

The dominant commercial piezoceramics are solid solutions of $PbZrO_3$ and $PbTiO_3$ (PZT). However, PZT contains approximately 60 wt% lead, a cumulative neurotoxin subject to progressively stringent environmental regulation worldwide, including the European Union's RoHS directive [5] and analogous legislation in China, Japan, and the United States. The urgent need to replace PZT with environmentally benign alternatives makes the discovery and characterisation of new lead-free piezoelectric materials one of the central challenges in functional materials research. We therefore restrict the present study to lead-free compositions for two reasons. First, the regulatory and toxicological drivers above mean that a high-performing lead-containing prediction has limited practical value, whereas a lead-free candidate is directly actionable for sustainable device manufacturing. Second, lead-free systems are chemically more diverse and far less optimised than the mature PZT family, so the unexplored chemical space in which a data-driven search can deliver genuinely new candidates is concentrated there. Restricting the screening output to lead-free compounds thus maximises both the technological relevance and the discovery potential of the workflow. Reliable prediction of the piezoelectric response is essential to accelerate this search, since exhaustive experimental characterisation of the vast available chemical space is neither practical nor cost-effective. Sustained research over two decades has produced three main families of lead-free candidates: (K,Na)NbO_3 (KNN)-based ceramics, $(Bi_{0.5}Na_{0.5})TiO_3$ (BNT)-based systems, and $BaTiO_3$ (BTO)-based compositions [1].

Despite these advances, no single lead-free material yet matches PZT's combination of high piezoelectric activity, comparatively high Curie temperature and associated thermal stability under typical operating conditions, and versatile processing. The chemical space available for exploration is estimated at 10^{10} quaternary inorganic compounds,

* Corresponding author.
E-mail address: francesc.sunol@upc.edu (F. Suñol).

https://doi.org/10.1016/j.cocom.2026.e01422
Received 1 June 2026; Received in revised form 30 July 2026; Accepted 3 August 2026
Available online 7 August 2026

far beyond the reach of exhaustive experimental screening [6], making computational pre-screening an essential complement to laboratory synthesis.

First-principles calculations based on density functional theory (DFT) can predict piezoelectric tensors with reasonable accuracy for known structures [7,8], and high-throughput DFT campaigns have already computed pizoelectric modulus for several hundred inorganic compounds stored in the Materials Project [9,10], AFLOW [11], and JARVIS-DFT [12] databases. These repositories now collectively contain DFT-computed properties for over one million entries, providing the training data necessary for machine learning (ML) models that predict target properties for material screening [13–15].

The field has been transformed by graph neural networks (GNNs [16]) that learn atomic representations directly from crystal structure, bypassing the need for expert-designed descriptors [17–19]. More recently, this paradigm has been extended by equivariant architectures that explicitly encode physical symmetries [20–22] and by universal interatomic potentials trained across the full periodic table [23–25], dramatically broadening the applicability of GNN-based materials modelling. The GNoME campaign [26] used a GNN trained on the Materials Project to predict the stability of 2.2 million inorganic crystals, 736 of them already confirmed experimentally, marking a milestone in the ML-to-experiment pipeline for materials discovery. For tabular feature vectors derived from composition and structure, gradient-boosted tree ensembles like XGBoost [27], LightGBM [28], and CatBoost [29], consistently rank among the strongest performers on materials science benchmarks, including the MatBench suite [30]. The TabNet architecture [31] offers a deep-learning complement to these methods, using sequential instance-wise attention over tabular features to combine the interpretability characteristic of tree models with greater representational capacity. Realising the full potential of any of these architectures, however, requires careful hyperparameter tuning: Bayesian optimisation frameworks such as Optuna [32] and Hyperopt [33] have largely supplanted grid and random search in materials informatics, reaching better configurations in substantially fewer evaluations. Beyond predictive accuracy, understanding what a well-tuned model has actually learned is equally important; this is addressed through SHAP (SHapley Additive exPlanations) [34], which decomposes each prediction into signed, additive per-feature contributions and has become the standard explanation tool in materials ML [35].

Predicting the piezoelectric modulus e_m presents challenges beyond those encountered for simpler scalar properties (note that e_m is conceptually related to the d_{33} coefficient familiar to experimentalists, but captures the optimal directional response of the material rather than a fixed crystallographic component). The piezoelectric tensor is identically zero for centrosymmetric structures and non-zero only for 20 non-centrosymmetric crystallographic point groups. In addition, the distribution of the piezoelectric modulus values is extremely right-skewed, making accurate prediction a challenging task. Hu and Song [36] provided a systematic ML benchmark for the piezoelectric modulus, evaluating random forests, SVM, and five GNN architectures on the Materials Project dataset; their best model, an SVM with Magpie features and elastic moduli, achieved a mean absolute error (MAE) of 0.646 C/m^2. However, their study was limited to predicting already-known materials with predominantly moderate piezoelectric response (the dataset available at the time contained only ∼1700 compounds with a heavily right-skewed e_m distribution) and neither a target transformation nor modern gradient-boosting or attention-based GNN models were explored. More recently, Rocha et al. [37] combined a neural network with a genetic algorithm (NSGA-II) to simultaneously optimise permittivity and dielectric loss in lead-free ceramics, resulting in a ML procedure that can guide not only property prediction but also compositional optimisation in piezoelectric systems.

In a closely related effort, Anand et al. [38] used a knowledge graph (KG) approach to screen the Crystallography Open Database (COD), filtering its non-centrosymmetric entries for structural similarity to known piezoelectrics to significantly reduce the search space before applying a GatedGCN model to predict the piezoelectric modulus of the remaining candidates; their best GNN model achieved a cross-validation MAE of 0.73 C/m^2. Despite this progress, a comprehensive benchmark comparing gradient-boosted tree ensembles, attention-based deep networks, and GNN architectures, all evaluated under unified Bayesian hyperparameter optimisation and with a principled treatment of the skewed target distribution, remains absent from the literature.

In this work, we address these gaps through a systematic benchmark of thirteen ML models applied to piezoelectric modulus prediction, with an explicit focus on lead-free materials discovery. We include non-graph tabular models that prove competitive or superior to GNNs on small datasets, we apply a logarithmic target transformation that systematically reduces errors across all model families, and we impose multi-criteria physicochemical filters to deliver a short, experimentally actionable candidate list. The novelty of this work, relative to the earlier ML benchmarks of Hu and Song [36] and Anand et al. [38], can be summarised in four points. (i) It is, to our knowledge, the first study to benchmark gradient-boosted tree ensembles, an attention-based tabular network (TabNet), a deep residual network, and seven graph neural networks for piezoelectric-modulus prediction under a single, unified Bayesian hyperparameter-optimisation protocol, removing the tuning asymmetry that confounds most cross-architecture comparisons. (ii) It introduces an explicit $\log(1 + e_m)$ target transformation that directly targets the heavily right-skewed modulus distribution and is shown to reduce errors consistently across all model families. (iii) It couples the predictors to a multi-criteria physicochemical filter (non-centrosymmetry, insulating gap, and thermodynamic near-stability enforced simultaneously) that no prior piezoelectric ML screening has applied jointly, yielding a short and directly synthesisable lead-free candidate list rather than a ranked property prediction alone. (iv) It validates the screening output with independent first-principles calculations: the full piezoelectric tensor of twenty compounds is computed by DFT, providing a direct, out-of-sample test of whether the pipeline enriches for genuinely high-e_m materials. The top-performing models used in this work compare favourably with the benchmarks previously reported, by approximately 23% and 32% relative to Hu and Song [36] and to the best GNN of Anand et al. [38], respectively. The identified lead-free candidates combine a meaningful piezoelectric response with thermodynamic near-stability and insulating character, properties that are jointly necessary for experimental realisation but rarely enforced simultaneously in prior screening studies.

2. Methodology

2.1. Dataset

Piezoelectric data was retrieved from the Materials Project (MP) database [9,10] (version 2025.09.25). The target quantity, reported as "*piezoelectric_modulus*" in MP, corresponds to the largest singular value (spectral norm) of the 3 × 6 piezoelectric stress tensor **e** in Voigt notation, $e_m = \sigma_{\max}(\mathbf{e})$. This is also the definition adopted for the first-principles validation of Section 2.7, so that machine-learning predictions and DFT results refer to the same scalar quantity. This target quantity provides a convenient metric for comparing piezoelectric performance across different compounds.

The initial train dataset comprises piezoelectric entries for 3315 inorganic compounds. However, the distribution of the piezoelectric modulus is heavily right-skewed (Fig. 1). This implies that materials with very high piezoelectric modulus are likely to introduce extreme leverage in regression. After filtering out 12 outliers with $e_m > 20$ C/m^2, the working train dataset contains 3303 materials. For each material the following scalar properties were also retrieved: band gap (E_g),

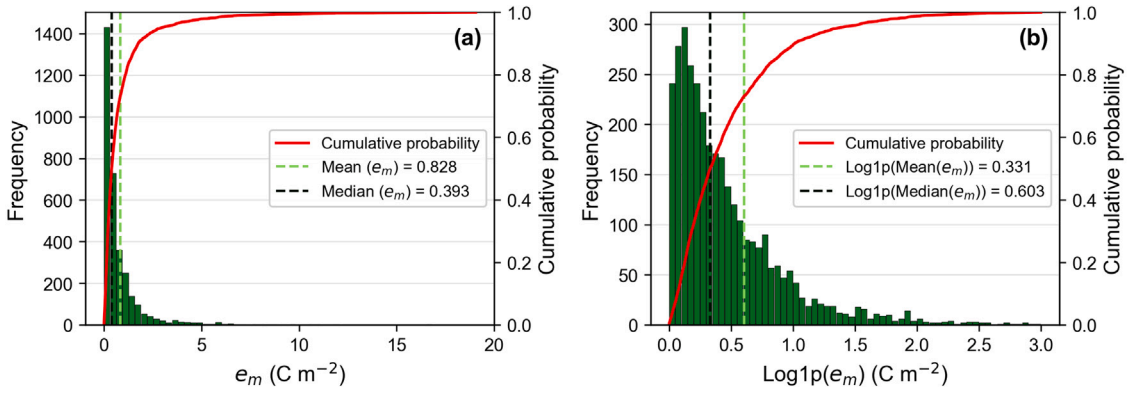

Fig. 1. (a) Distribution (green boxes) and cumulative probability (red curve) of the e_m values in the training dataset. Mean and median values are represented with vertical dashed lines. (b) Distribution and cumulative probability of the transformed target $\log 1\mathrm{p}(e_m)$.

energy above the convex hull (E_{hull}), total magnetisation, bulk modulus (K_{VRH}), shear modulus (G_{VRH}), volume, density, number of sites (N_{sites}), number of elements (N_{el}), crystal system, and space-group number.

The test dataset, used exclusively for prediction, consists of all MP entries with computed elastic moduli (K_{VRH} and G_{VRH}) that are absent from the training set. The test dataset consists of 10301 unlabelled candidate compounds that will be used for large-scale screening.

Fig. 1(a) shows the distribution of e_m values in the working train dataset. The distribution is severely long-tailed: approximately 92% of materials have $e_m < 2$ C/m^2, while only 19 compounds exceed 10 C/m^2. This extreme skewness is a fundamental challenge for regression and motivates the target transformation from e_m to its logarithmic counterpart: all models are trained to predict the log-transformed target

$$\tilde{y} = \log 1\mathrm{p}(e_m) \equiv \ln(1 + e_m), \tag{1}$$

rather than e_m directly. Here $\log 1\mathrm{p}(x) \equiv \ln(1+x)$ denotes the "logarithm of one plus" transformation (the `log1p` function of standard numerical libraries); the unit offset keeps the transformation finite and well-behaved at $e_m = 0$. This transformation compresses the range of the target from [0, 20] C/m^2 to [0, 3], which substantially reduces the leverage exerted by the few very large values and makes the distribution less asymmetric (Fig. 1(b)). All reported MAE and R^2 values are computed in the original physical space after back-transformation via $e_m = e^{\tilde{y}} - 1$.

2.2. Feature engineering

To represent the material compounds in a format suitable for machine learning, a high-dimensional feature matrix was constructed from the raw crystallographic and chemical data. The feature engineering process involves five primary descriptors:

- **Elemental Statistics (Magpie):** Using the Magpie (Materials Agnostic Platform for Informatics and Exploration) framework [39], statistical aggregations (mean, minimum, maximum, and standard deviation) of fundamental atomic properties are computed, including atomic number (Z), period, group, electronegativity (χ), and atomic mass.
- **Oxidation States:** Features derived from the possible oxidation states of the constituent elements were extracted using the `matminer` library [40] to capture chemical bonding trends.
- **Structural Descriptors:** Categorical variables such as the crystal system and space group number were encoded.
- **Elastic Properties:** The Voigt–Reuss–Hill averages of the bulk (K) and shear (G) moduli are included directly, and four secondary descriptors are computed analytically from them: Poisson's ratio $\nu = (3\,K - 2G)/(2(3\,K + G))$, Pugh's ratio K/G, Young's modulus $E = 9KG/(3\,K + G)$, and the elastic anisotropy index $(K - 2G/3)/(K + 4G/3)$.
- **Electronic and Physical Features:** Additional properties including the band gap, energy above the hull, total magnetisation, density, and volume per site were integrated.

2.3. Trained models

Thirteen distinct predictive architectures were evaluated, grouped into three categories.

2.3.1. Traditional ML regressors

Support Vector Machines (SVM). SVMs have been widely adopted in materials informatics owing to their strong generalisation on small, high-dimensional datasets, making them a natural baseline for property prediction tasks of this scale. An SVM with a radial basis function (RBF) kernel [41] was included as the baseline, matching the best-performing model of Hu and Song [36]. The kernel width γ was set to "*auto*", and the regularisation parameter C and tube width ε were optimised during hyperparameter tuning (Section 2.4).

Gradient-boosted trees. Gradient-boosting frameworks consistently rank among the top performers on structured tabular benchmarks and offer built-in feature importance estimates that facilitate physical interpretation of the learned model. Three gradient-boosting frameworks were evaluated: XGBoost [27], LightGBM [28], and CatBoost [29]. All three were evaluated both with default-tuned parameters and with Bayesian-optimised hyperparameters (Section 2.4).

2.3.2. Deep neural networks

DNN (feedforward). Deep feedforward networks can capture nonlinear interactions among the compositional and structural descriptors that are difficult to model with shallow methods, and residual connections help mitigate vanishing gradients when stacking multiple layers. A fully connected feedforward network with four residual blocks is implemented in PyTorch [42]. Each block consists of two linear layers with batch normalisation, ReLU activation and dropout, connected by a skip connection. Training uses the Adam optimiser [43] with an initial learning rate of 5×10^{-4}. Early stopping is applied with a patience of 40 epochs on a held-out validation set (10% of the training fold).

TabNet. TabNet [31] is a deep learning architecture for tabular data that uses sequential, instance-wise feature selection via soft attention masks. At each decision step, an attention transformer selects the most relevant subset of features, while the selected features are processed by a shared fully connected block.

2.3.3. Graph neural networks

GNN models have recently attracted particular attention for piezoelectric property prediction and consume the crystal structure directly as a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where nodes $\mathcal{V}$ represent atoms and edges $\mathcal{E}$ connect pairs of atoms within a cutoff radius. In this work we include seven GNN architectures to provide a comprehensive comparison against tabular methods:

- **CGCNN** [17]: specialised in crystalline materials, representing them as graphs to extract deep structural insights. It is highly versatile and has been proven effective at predicting a wide range of properties, from thermodynamic data like formation energy to mechanical properties like Poisson's ratio. A graph with three `CGConv` layers and mean pooling is used.
- **SchNet** [44]: designed specifically for quantum chemistry, focusing on predicting molecular energies and atomic properties while strictly adhering to physical laws like rotation and translation invariance. Its standout feature is the use of continuous-filter convolutional layers, which allow the model to capture local correlations between atoms without needing them to be arranged on a rigid grid. It typically utilises three interaction blocks to model the complex relationships within a molecular structure.
- **MPNN** [45]: provides a unified framework for predicting molecular properties with high accuracy. The model operates by passing information through three distinct, differentiable phases: message functions, vertex update functions, and readout functions. It is particularly useful for its invariance to graph isomorphism and its ability to handle directed graphs by using separate channels for incoming and outgoing atomic signals.
- **MEGNet** [46]: a high-performing algorithm that excels at both molecular and crystal property prediction. To combat data scarcity, it incorporates global state inputs (like temperature and pressure) and uses learned element embeddings that capture periodic chemical trends. These embeddings can be "pre-learned" from larger datasets, which has proven advantageous for elastic and electronic property targets.
- **GATGNN** [18]: leverages the power of attention mechanisms to understand atomic structures. By combining multiple graph-attention (GAT) layers with global attention, the model can efficiently map the complex bonding environments surrounding individual atoms. This architecture is specifically targeted for tasks requiring a deep understanding of structural features, such as predicting piezoelectric moduli.
- **CrystalAttentionNet** [47]: this architecture was designed to serve as a stronger attention-based baseline by combining modern stabilisation techniques (layer normalisation, GELU activations) with a learnable global pooling step, allowing it to weight atomic contributions to the crystal-level prediction.
- **ALIGNN** [19]: the Atom-Line-Graph Neural Network augments the standard atom graph with a line graph in which bonds become nodes and triplet bond-angle information is encoded on the line-graph edges. By explicitly encoding three-body angular correlations, ALIGNN moves beyond pairwise distance interactions and captures the directional bonding geometry that is directly relevant to piezoelectric response. ALIGNN has been shown to achieve state-of-the-art accuracy on multiple MP property benchmarks [19].

All GNN models used here are trained for up to 300 epochs with early stopping (patience 40) on a 10% validation split of each training fold.

2.4. Hyperparameter optimisation

Hyperparameters for all models were tuned using Optuna [32], a Bayesian optimisation framework based on Tree-structured Parzen Estimators (TPE) [48]. TPE maintains a probabilistic model of the objective function and draws candidates that maximise the expected improvement over the current best, concentrating trials in promising regions of the search space and consistently outperforming grid and random search with fewer function evaluations.

Optuna was applied to each model category with a strategy scaled to its computational cost:

- Gradient-boosted trees (XGBoost, LightGBM, CatBoost): 50 TPE trials, each evaluated by 5-fold cross-validation (CV) on the training set using $\ln(1 + e_m)$ as the target objective.
- SVM: 50 TPE trials with 5-fold CV.
- DNN and TabNet: 20 TPE trials, each evaluated on a single 80/20 stratified train/validation split. The final performance (MAE $\pm$ std, $R^2 \pm$ std) was estimated from 5-fold cross-validation.
- GNN models (CGCNN, SchNet, MPNN, MEGNet, GATGNN, CrystalAttentionNet, ALIGNN): 10 TPE trials per model, each evaluated on a single 80/20 split. The final performance was estimated from 5-fold cross-validation.

The best hyperparameter values found by Optuna for every model are reported in Table S1 (Supplementary Data).

2.5. Performance evaluation

Model performance is assessed by k-fold cross-validation (CV) for all models, using $k = 5$. At each fold, training and prediction are performed on disjoint subsets of the data; the out-of-fold predictions are collected to form a complete prediction vector over the entire training set. Two metrics are reported, both computed in the original (non-transformed) physical space: the mean absolute error (MAE), Eq. (2), and the coefficient of determination (R^2), Eq. (3).

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N} |\hat{e}_{m,i} - e_{m,i}| \tag{2}$$

$$R^2 = 1 - \frac{\sum_i (e_{m,i} - \hat{e}_{m,i})^2}{\sum_i (e_{m,i} - \bar{e}_m)^2} \tag{3}$$

Here $e_{m,i}$ denotes the reference piezoelectric modulus of material i and $\hat{e}_{m,i}$ is the corresponding model prediction, with $\bar{e}_m$ the mean over the training set. Mean and standard deviation over folds are reported for both metrics. The MAE is preferred as the primary metric because it is robust to the few large-e_m outliers that remain in the dataset; R^2 is reported as a secondary metric to quantify the fraction of variance. In this context, negative R^2 values are possible, and indicate that the model performs worse than predicting the sample mean.

2.6. Lead-free material screening workflow

The complete workflow for predicting new lead-free piezoelectric materials proceeds as follows. First, all ML, DNN and GNN models are trained on the full 3303-material training set (using the best hyperparameters found during CV hyperparameter tuning). Second, the screening dataset of 10301 compounds is featurised with the same descriptor pipeline described in Section 2.2, producing an input representation of identical form to that used during training. Third, scalar $\hat{e}_m$ predictions are generated for every candidate by three complementary models (the three models with statistically equivalent best cross-validated accuracy, Section 3.2), namely the tabular XGBoost and LightGBM and the graph-based ALIGNN. Fourth, all materials containing lead are excluded from the candidate list. Fifth, the lead-free compounds are ranked by predicted $\hat{e}_m$ separately for each model, and the top-200 list of each model is retained for further filtering.

Finally, each of the three top-200 lists is filtered by three simultaneous physicochemical requirements: $E_g > 0.1$ eV, $E_{\mathrm{hull}} < 0.1$ eV/atom, and a non-centrosymmetric space group. These criteria ensure that a

candidate is insulating enough to hold an electric charge, thermodynamically stable enough to be synthesised, and crystallographically capable of a non-zero piezoelectric tensor, respectively. The union of the three filtered lists is then subjected to a final expert inspection that removes entries which remain physically implausible as piezoelectric or synthesis candidates (for example near-metallic phases that only marginally clear the gap filter, or simple ionic solids with no structural mechanism for a piezoelectric distortion), yielding the final candidate set (Table 2).

2.7. First-principles validation

To test whether the screening pipeline genuinely enriches for high-e_m materials, a DFT validation set of twenty compounds was assembled: the ten highest-ranked candidates emerging from the filtered three-model screen (predicted-high group), and ten compounds predicted by all three models to have a weak response, $\hat{e}_m \lesssim 0.5$ C/m^2 (predicted-low group). All twenty compounds satisfy the same hard constraints: they are absent from the training set (by Materials Project identifier), non-centrosymmetric, and have unit cells small enough ($\leq$ 30 sites) to make the tensor calculation tractable. If the models carry genuine physical signal, the computed piezoelectric moduli of the predicted-high group should systematically exceed those of the predicted-low group. We stress that the predicted-high group is defined for this enrichment test and is not identical to the final synthesis-candidate list of Table 2: two of its ten members ($TiPN_3$ and Rb_2ZrO_3) are computed here to probe the top of the ranking but are afterwards discarded as experimental-synthesis candidates on physical grounds (Section 3.5), and consequently do not appear in Table 2.

First-principles calculations based on density functional theory (DFT) [49] were carried out with the PBEsol exchange–correlation energy functional [50] as it is implemented in the VASP software [51]. The projector-augmented wave method [52] was employed to represent the ionic cores by considering the maximum number of electronic states as valence. An energy cutoff of 650 eV and dense Monkhorst–Pack k-point grids were used for integration within the Brillouin zone in reciprocal space, leading to total energies converged to within 0.5 meV per formula unit. Atomic relaxations were concluded when the forces in all the atoms were below 0.005 eV/Å. The piezoelectric coefficients were calculated by the finite-difference method [53].

For each compound, the electronic and ionic contributions to the piezoelectric stress tensor were computed, and the piezoelectric modulus was evaluated from the total tensor as its largest singular value, $e_m = \sigma_{\max}(\mathbf{e}^{\mathrm{tot}})$, consistent with the Materials Project convention used for the training data (Section 2.1). The complete piezoelectric matrices (electronic, ionic and total contributions) and singular values for all twenty compounds are reported in the Supplementary Information (Section S2).

3. Results and discussion

3.1. Distribution and chemical space of piezoelectric materials

Fig. 2(a) displays a highly unbalanced distribution of piezoelectric materials across the seven crystal systems. Orthorhombic and tetragonal systems contain the largest number of known piezoelectric materials. In contrast, low-symmetry systems such as triclinic contain fewer known piezoelectric materials, suggesting that crystal structure has considerable influence on the piezoelectric effect. We suspect that the lack of inversion symmetry requirement is related to the low number of piezoelectric materials in low-symmetry systems, however, the cubic crystal system is highly-symmetric and is the third structure with more known piezoelectric compounds. A possible reason for this high count likely stems from extensive research into this material family.

Fig. 2(b) shows the mean and median e_m values for the seven crystal systems. We point out that the cubic family has the lowest average piezoelectric moduli despite the large count of materials in this system. On the other hand, the triclinic family has the lowest number of piezoelectric compounds, with the highest values of mean and median piezoelectric modulus. The observed disparity could arise because the high-symmetry of the cubic system offers a vast, energetically favourable landscape for stable non-centrosymmetric phases with structurally constrained polarisation, whereas the low-symmetry triclinic system, while statistically rare due to its thermodynamic instability, provides the maximum structural degrees of freedom and internal lattice distortions necessary to facilitate exceptionally large electromechanical coupling coefficients.

Fig. 3 shows a t-distributed stochastic neighbour embedding (t-SNE) [54] visualisation of the training set with the aim to explore the underlying relationships between piezoelectric material properties and their moduli. The embedding was computed on the standardised feature matrix after a preliminary reduction to 50 principal components, using perplexity 30, 1000 iterations and a fixed random seed (42). The position of each data sample reflects relative similarity in the descriptor space: points that are close together share similar compositions and structural environments. Colour and point size encode $\log 1\mathrm{p}(e_m)$ on a green-red scale, point size is proportional to the modulus of e_m, so large red markers identify the most strongly piezoelectric compounds simultaneously in both channels. The map suggests that the dataset separates into several composition-based clusters, with high-e_m materials appearing concentrated in a small number of regions. We emphasise that t-SNE is a stochastic, non-convex projection whose geometry depends on perplexity, learning rate and random seed, so we use it purely as a qualitative aid. The observation that the piezoelectric response correlates with compositional and structural class, does not rely on the embedding at all: it is supported more robustly by the per-crystal-system statistics on the original features (Fig. 2), by the SHAP feature attributions of Section 3.3, and by the elemental concentration among the top candidates (Fig. 8), none of which involve any dimensionality reduction. We nonetheless corroborated this reading quantitatively with deterministic clustering applied directly to the original descriptor space (supplementary data, Figs. S2 and S3). Across five perplexities and three random seeds the local neighbourhood structure stays essentially unchanged (trustworthiness between 0.98 and 0.99) even as the global layout shifts, and HDBSCAN, DBSCAN and KMeans clustering of the 50 leading principal components each recover a statistically significant concentration of high-e_m materials in a few composition-based clusters (χ^2 test, $p < 10^{-17}$ for every method), the most enriched being an oxide cluster dominated by O, Ta, W and Bi. The sparse distribution of high-e_m compounds across chemical space illustrates the challenge of training regression models and motivates the use of structural descriptors and GNN representations, which can capture bonding geometry beyond what composition statistics encode.

3.2. Model performance

Table 1 reports the cross-validated performance of all thirteen models, and Fig. 4 visualises the MAE (a) and R^2 (b) together with their fold-to-fold spread. The gradient-boosted tree ensembles lead the ranking: XGBoost attains the lowest MAE (0.498 $\pm$ 0.049 C/m^2) and the highest R^2 (0.285 $\pm$ 0.085), with LightGBM (0.499 C/m^2) and CatBoost (0.503 C/m^2) close behind. The RBF-kernel SVM, which was the best model in the original study of Hu and Song [36], trails all three at MAE = 0.564 C/m^2 (R^2 = 0.160). Among the deep networks, TabNet (0.534 $\pm$ 0.065 C/m^2) edges the residual DNN (0.581 $\pm$ 0.046 C/m^2). The strongest graph model, ALIGNN, reaches MAE = 0.500$\pm$0.066 C/m^2, matching XGBoost once its capacity is tuned, while MEGNet attains the highest graph-network R^2 (0.284 $\pm$ 0.014); this confirms that the three-body bond-angle information ALIGNN encodes is physically relevant to piezoelectricity.

Because the gaps at the top of the ranking are small relative to the fold-to-fold scatter, we assessed them with a two-sided Welch's t-test on

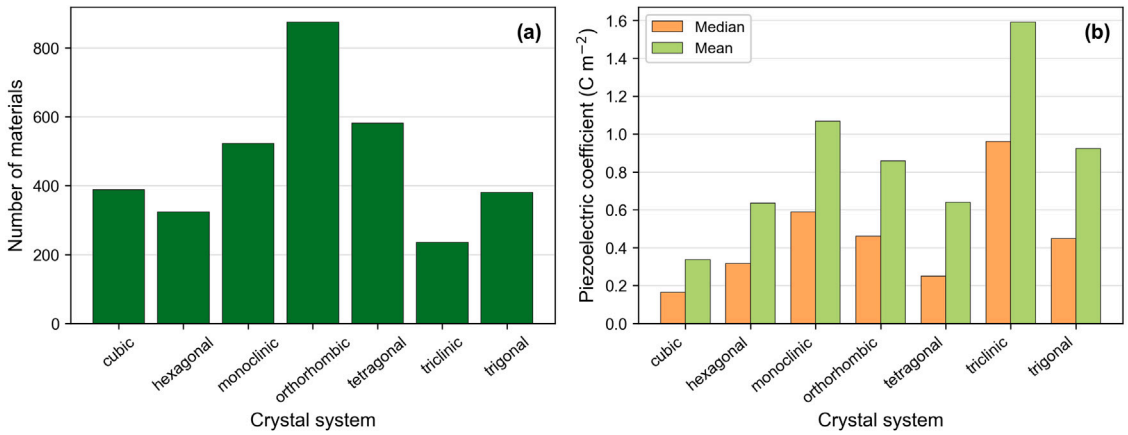

Fig. 2. Crystal-system statistics for the training dataset. (a) Number of materials per crystal system. (b) Mean and median piezoelectric modulus e_m per crystal system.

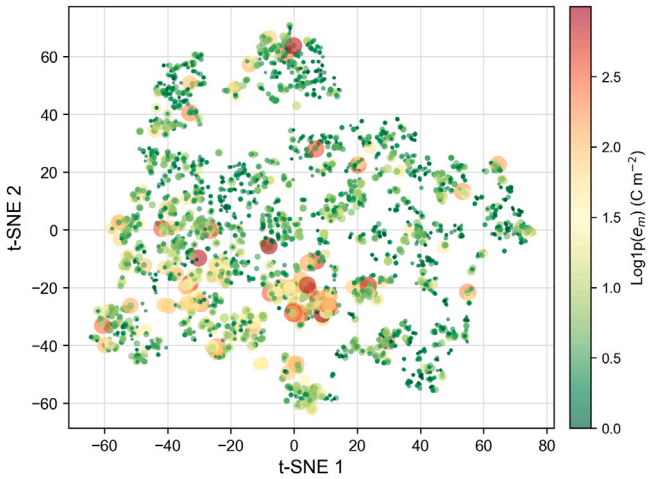

Fig. 3. t-SNE visualisation of the 3303 training materials. Point colour encodes $\log 1\mathrm{p}(e_m)$ on a green-red scale; point area is proportional to e_m.

Table 1
Cross-validated performance of all models after Optuna hyperparameter tuning. MAE in C/m^2; R^2 dimensionless; mean ± standard deviation over folds (5-fold CV). The best MAE and the best R^2 within each category are highlighted in bold.

Model	Category	MAE (C/m^2)	R^2
Gradient-boosted trees and SVM:			
XGBoost	ML	**0.498 ± 0.049**	**0.285 ± 0.085**
LightGBM	ML	0.499 ± 0.051	0.269 ± 0.099
CatBoost	ML	0.503 ± 0.055	0.254 ± 0.098
SVM	ML	0.564 ± 0.064	0.160 ± 0.045
Deep neural networks:			
DNN	DNN	0.581 ± 0.046	**0.196 ± 0.089**
TabNet	DNN	**0.534 ± 0.065**	0.192 ± 0.064
Graph neural networks:			
CGCNN	GNN	0.578 ± 0.049	0.126 ± 0.038
SchNet	GNN	0.588 ± 0.043	0.183 ± 0.033
MPNN	GNN	0.561 ± 0.057	0.220 ± 0.086
MEGNet	GNN	0.563 ± 0.048	**0.284 ± 0.014**
GATGNN	GNN	0.617 ± 0.034	0.243 ± 0.070
CrystalAttn	GNN	0.668 ± 0.053	0.212 ± 0.038
ALIGNN	GNN	**0.500 ± 0.066**	0.197 ± 0.053

the per-fold MAE values. The XGBoost advantage over ALIGNN is not significant (ΔMAE = 0.002 C/m^2, $p = 0.96$), and neither is its advantage over LightGBM ($p = 0.98$) or CatBoost ($p = 0.88$); XGBoost separates only marginally from the mid-ranked GNNs (for example MEGNet, $p = 0.07$) and clearly from the weaker SVM and residual DNN. The leading models are therefore statistically equivalent, which is why the screening of Section 3.4 draws on the three best performers together (XGBoost, LightGBM and ALIGNN) rather than on a single nominal winner.

The modest $R^2 \approx 0.29$, meaning that about 70% of the variance in e_m is left unexplained, deserves comment. First, this ceiling is common to every architecture tested here and to the prior studies on the same property [36,38], so it reflects an intrinsic difficulty of the target, a tensor-derived quantity set by subtle symmetry-breaking displacements, together with the limited size and strong skew of the DFT data, rather than a weakness of any one model. Second, what matters for screening is the ranking error rather than the variance explained, and an MAE of about 0.5 C/m^2 against a 2 C/m^2 threshold comfortably separates strong candidates from the weakly piezoelectric bulk. Third, the models under-predict the high-e_m tail (Fig. 5), which is conservative for screening because it works against false positives at the top of the ranking.

Set against prior work, our best model lowers the MAE of the data-augmented GatedGCN of Anand et al. [38] (0.73 C/m^2) by roughly 32%, and even our weakest GNN (CrystalAttentionNet, 0.668 C/m^2) is competitive with it; the benchmark also improves on Hu and Song [36] by about 23%. We attribute these gains to the larger training set, stronger hyperparameter tuning and the logarithmic target transformation. Table 4 compiles the comparison, listing dataset size, target treatment and best reported MAE, with the caveat that the three studies use overlapping but non-identical Materials Project snapshots and different processing choices, so the percentages indicate a consistent trend rather than a controlled measurement. Finally, the parity plot for XGBoost (Fig. 5; the remaining models appear in Figure S1 of the Supplementary Data) shows good agreement for $e_m < 2$ C/m^2 and growing under-prediction above it, a direct consequence of the right-skewed training distribution, which leaves few tail samples even after the logarithmic transformation of the target.

3.3. Feature importance and SHAP analysis

To interpret the physical drivers of the XGBoost predictions, SHapley Additive exPlanations (SHAP) [34] were computed for the final model trained on the complete dataset. We interpret XGBoost here as the lowest-error model and the only one of the three screening models whose inputs are directly interpretable physicochemical descriptors: the near-identical LightGBM yields essentially the same attributions, while the graph-based ALIGNN consumes the raw crystal graph and has no comparable tabular-feature importances. SHAP assigns to each

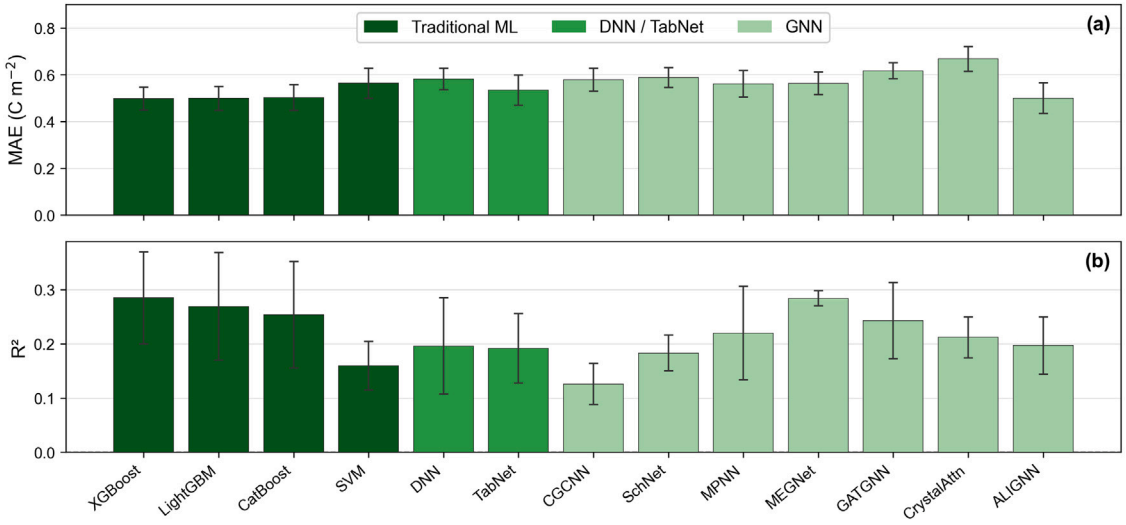

Fig. 4. Combined summary of MAE (top) and R^2 (bottom) for all evaluated models, colour-coded by category: darkgreen = traditional ML, medium-green = DNN/TabNet, lightgreen = GNN.

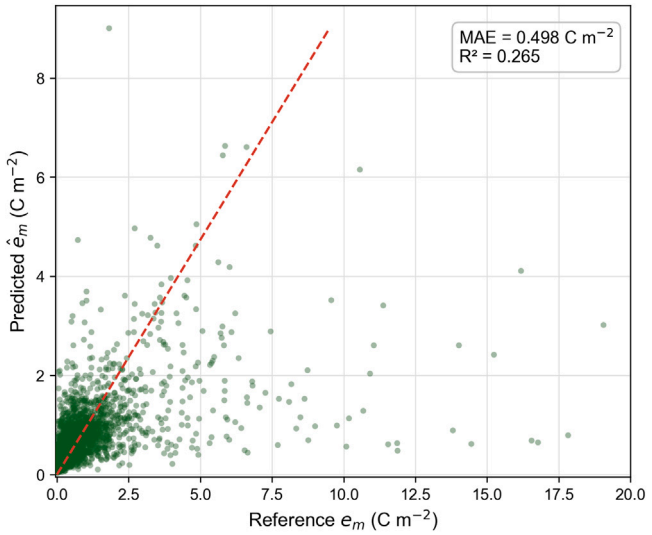

Fig. 5. Parity plot of e_m reference values vs. predicted $\hat{e}_m$ for the XGBoost model. Dashed line is the ideal 1:1 correspondence; Inset text reports the MAE and R^2 computed in the original physical space.

feature a contribution (the SHAP value) that quantifies how much that feature shifts the prediction away from the dataset mean, with the sign indicating the direction of the shift. Fig. 6 ranks the top 25 features by their mean absolute SHAP value, which measures the average magnitude of each feature's contribution to the prediction across all training samples. Fig. 7 complements this global ranking with a beeswarm plot in which each point represents one training material; the horizontal position is the SHAP value and the colour encodes the feature value, revealing not only the importance of each feature but also the direction of its effect.

The space-group number is the most important feature, with a complex effect visible in Fig. 7: materials with both low and high space-group numbers show negative SHAP contributions. This is due to the scattered ordering of the 230 crystallographic space-group numbers across the crystal systems (only 138 space-group numbers lack inversion symmetry and allow a non-zero piezoelectric tensor). The energy above the convex hull (E_{hull}) ranks second. The beeswarm plot in Fig. 7 reveals that materials with high E_{hull} (red points) consistently receive positive SHAP contributions, i.e., they are predicted to have higher e_m. At first glance this appears counter-intuitive, since thermodynamically unstable phases are generally considered less experimentally accessible. The physical explanation is that non-centrosymmetric structures are often metastable with respect to higher-symmetry, centrosymmetric polymorphs: the structural distortion that breaks inversion symmetry and enables piezoelectricity is frequently associated with a shallow energy landscape and a modest positive E_{hull}. It is nonetheless important to note that E_{hull} is a predictor, not a design target: candidates with very large E_{hull} (> 0.1 eV/atom) are unlikely to be synthesisable and should be de-prioritised. Unit-cell volume, volume per site, and density collectively occupy ranks 3–5, making volumetric descriptors the dominant compositional signal after symmetry. In the beeswarm plot, large volumes and low densities are associated with negative SHAP contributions, while compact structures with high densities tend to receive positive contributions. This is consistent with the microscopic origin of piezoelectricity: a high packing density favours strong overlap of electronic wave functions between neighbouring ions, which enhances the Born effective charges that determine the piezoelectric coupling constant [55].

3.4. Predicted lead-free piezoelectric candidates

The three models (XGBoost, LightGBM, and ALIGNN) were independently applied to the 10,301 unlabelled MP compounds, and each produced a ranked top-200 list of predicted lead-free candidates. Every $\hat{e}_m$ reported here is a ranking score with an empirical uncertainty of ± 0.50 C/m^2 (the cross-validated MAE), not a quantitative prediction, and the systematic under-prediction of the high-e_m tail (Section 3.2) makes these scores conservative rather than optimistic. The two tree ensembles largely agree (137 of 200 compounds in common), whereas ALIGNN, which sees the crystal structure directly as a graph, selects a substantially different subset (9 and 12 compounds in common); a candidate flagged by all three models therefore carries a particularly robust consensus signal.

Each top-200 list was reduced by the three physicochemical criteria of Section 2.6 (non-centrosymmetric space group, $E_g > 0.1$ eV and $E_{\text{hull}} < 0.1$ eV/atom), and the union of the filtered lists, after the physical-plausibility inspection of Section 2.6, was assessed compound-by-compound (Table 2). Three of the retained candidates are flagged by all three models: $SrTaNO_2$ (mp-756088), an oxynitride perovskite whose N/O anion ordering breaks inversion symmetry and which lies essentially on the convex hull ($E_{\text{hull}} = 0.003$ eV/atom);

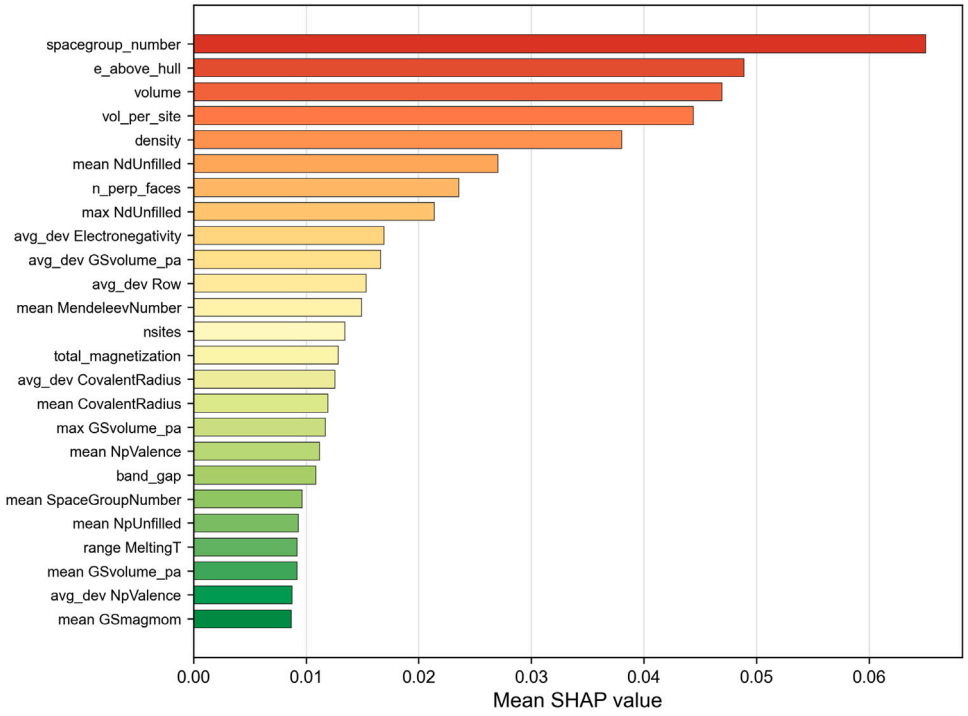

Fig. 6. Mean absolute SHAP values for the top 25 features, providing a global feature importance ranking. Colour encodes rank (red = most important).

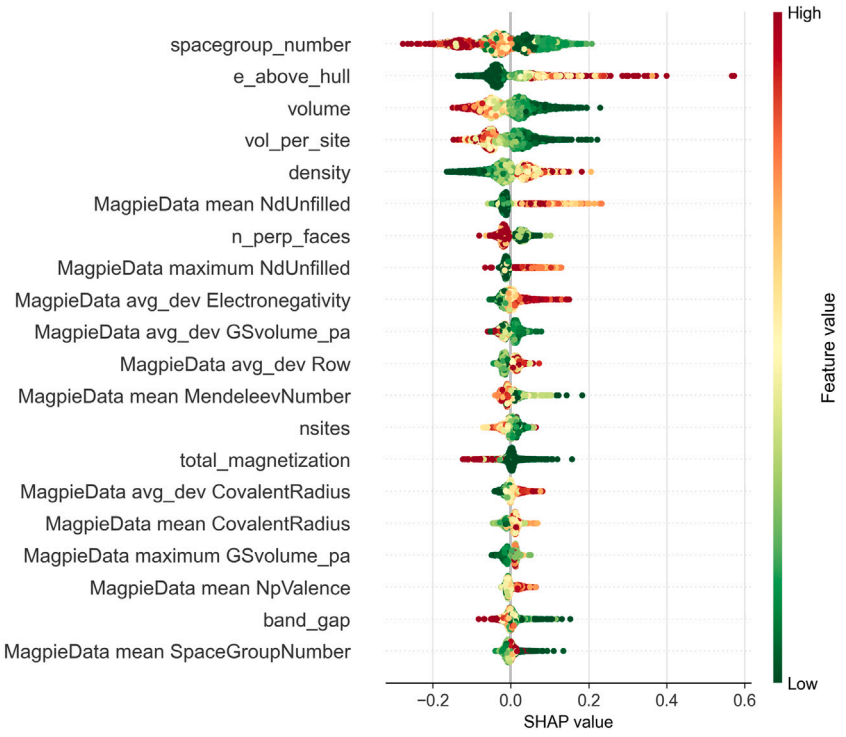

Fig. 7. SHAP beeswarm summary for the top 20 features of the XGBoost model. Each point represents one training sample; horizontal position gives the SHAP value ($\log 1\mathrm{p}(\hat{e}_m)$ space); colour encodes the normalised feature value. Features are ranked by mean |SHAP| (most important at top).

Table 2
Final lead-free piezoelectric candidates from the XGBoost, LightGBM and ALIGNN screens. The $\hat{e}_m$ columns are the predicted piezoelectric moduli (C/m^2) of each model; SG is the space-group number; N is the number of models flagging the material; "Novel" marks a composition absent from the training set even at the formula level. All entries are non-centrosymmetric with $E_g > 0.1$ eV and $E_{hull} < 0.1$ eV/atom.

Formula	MP-ID	$\hat{e}_m^{XGB}$	$\hat{e}_m^{LGB}$ (C/m^2)	$\hat{e}_m^{ALIGNN}$	SG #	E_g (eV)	E_{hull} (meV/at.)	N	Novel
$SrTaNO_2$	mp-756088	5.94	5.62	2.10	7	1.71	3	3	
YWN_3	mp-989576	3.05	3.77	2.44	161	1.54	46	3	
$SrTa_2Bi_2O_9$	mp-23089	3.06	3.10	1.38	36	2.51	0	3	✓
LiCuS	mp-2724538	1.91	2.26	1.44	31	0.85	30	2	
$Ba_3Ti_2O_7$	mp-770118	1.78	1.90	1.66	42	2.27	32	1	✓
Bi_2O_3	mp-1105780	2.13	1.84	1.15	159	1.49	39	1	
$BaTiO_3$	mp-558125	1.41	1.30	2.03	20	2.08	1	1	
$Ca_3Sn_2O_7$	mp-1374018	0.86	0.91	1.59	36	2.59	36	1	✓

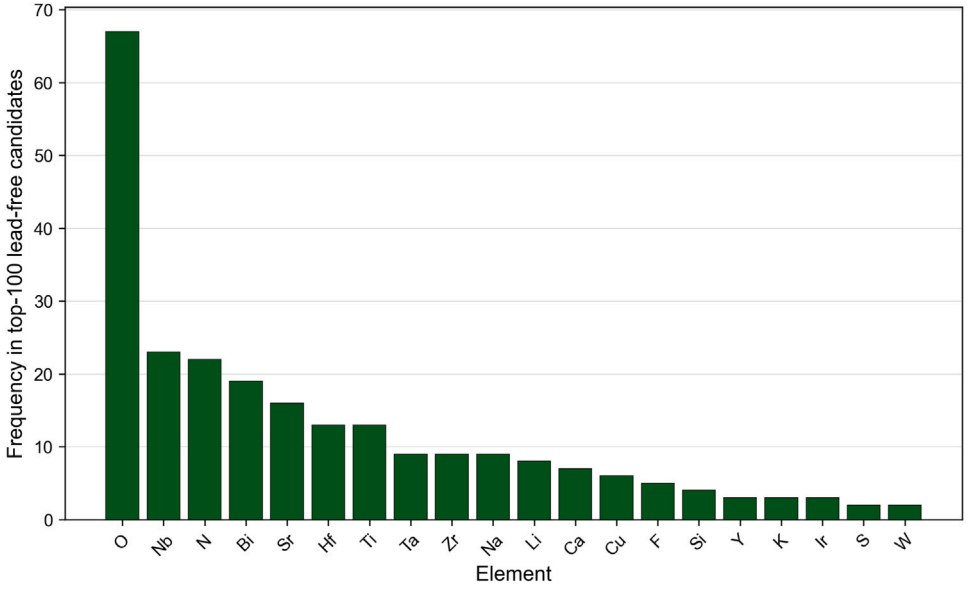

Fig. 8. Element frequency in the top-100 predicted lead-free candidates of the XGBoost ranking, prior to the physicochemical filtering of Section 2.6; the chemical families it highlights coincide with the multi-model consensus candidates (Table 2).

YWN_3 (mp-989576), a nitride perovskite in the polar $R3c$ space group shared with the archetypal ferroelectrics $LiNbO_3$ and $BiFeO_3$; and $SrTa_2Bi_2O_9$ (mp-23089), an Aurivillius phase structurally analogous to the ferroelectric $SrBi_2Ta_2O_9$. The remaining retained candidates (LiCuS, $Ba_3Ti_2O_7$, Bi_2O_3, $BaTiO_3$ and $Ca_3Sn_2O_7$) are listed in Table 2.

The elemental frequency analysis of the XGBoost top-100 ranking (prior to the physicochemical filtering, Fig. 8) shows that the model favours oxygen-rich early-transition-metal (Nb, Ta) and bismuth oxides and oxynitrides, with oxygen, niobium, nitrogen, bismuth and strontium the most prevalent elements. This learned chemistry is consistent with the final multi-model consensus candidates, with the O/Ta/W/Bi-dominated cluster recovered by the deterministic clustering of the supplementary data (Fig. S3), and with the symmetry- and stability-driven SHAP attributions of Section 3.3: three independent lines of evidence converging on the same chemical families.

3.5. First-principles validation of the screening

Table 3 compares the model predictions with the DFT piezoelectric moduli computed for the twenty-compound validation set of Section 2.7 (ten predicted-high, ten predicted-low compounds); the complete piezoelectric matrices are given in the Supplementary Information (Section S2). The results show that the screening pipeline successfully enriches the candidate set with materials exhibiting genuinely high piezoelectric moduli e_m. The DFT moduli of the predicted-high group systematically exceed those of the predicted-low group (median 2.42 versus 0.28 C/m^2; Mann–Whitney $p = 8.5\times10^{-4}$), and correlate strongly with the model rankings (Spearman $\rho = 0.80$, $p \approx 2\times10^{-5}$; per model, 0.79 for XGBoost, 0.80 for LightGBM, 0.60 for ALIGNN). All three consensus candidates ($SrTaNO_2$, YWN_3 and $SrTa_2Bi_2O_9$) exceed the 2 C/m^2 screening threshold at the DFT level.

The comparison is qualitative rather than quantitative, for two reasons. Six compounds ($SrTaNO_2$, Rb_2ZrO_3, $Ba_3Ti_2O_7$, YWN_3, $RbCaCl_3$ and $SrTa_2Bi_2O_9$; marked in Table 3) show anomalously large ionic contributions, signalling proximity to a lattice instability, so their absolute values should be treated with caution, though such proximity is itself characteristic of strong piezoelectrics. In addition, the models are trained on data truncated at $e_m \leq 20$ C/m^2 and systematically underpredict the tail, so no quantitative correspondence is expected even in principle. The qualitative conclusion is nonetheless robust. In the full validation set the only compound that grossly violates the predicted ranking is $RbCaCl_3$ (predicted low, DFT $e_m = 4.9$ C/m^2), and it is itself one of the flagged near-instability compounds; removing the six flagged compounds therefore discards this outlier while still leaving the enrichment intact (median 0.90 versus 0.27 C/m^2; $p = 2\times10^{-3}$; $\rho = 0.77$).

These independent first-principles calculations confirm that the multi-model screening genuinely concentrates high-response materials at the top of its ranking, while the flagged near-instability cases illustrate why DFT verification remains an essential step before synthesis. They also account for the two predicted-high compounds that do not appear in the synthesis-candidate list of Table 2: Rb_2ZrO_3 belongs

Table 3
First-principles validation set: the ten highest-priority screening candidates (predicted-high group) and ten compounds predicted weak by all three models (predicted-low group), compared with the DFT piezoelectric modulus e_m^{DFT} computed in this work (Section 2.7). SG is the space-group number. Entries marked with † show anomalously large ionic contributions to the piezoelectric tensor, usually signalling proximity to a lattice instability or an insufficiently converged ionic relaxation; their absolute values should be treated with caution. The complete piezoelectric matrices are given in the supplementary data.

Group	Formula	MP-ID	SG #	$\hat{e}_m^{\mathrm{XGB}}$	$\hat{e}_m^{\mathrm{LGB}}$	$\hat{e}_m^{\mathrm{ALIGNN}}$	e_m^{DFT} (C/m^2)
High	$SrTaNO_2$	mp-756088	7	5.94	5.62	2.10	103.00†
High	YWN_3	mp-989576	161	3.05	3.77	2.44	9.28†
High	$SrTa_2Bi_2O_9$	mp-23089	36	3.06	3.10	1.38	2.93†
High	LiCuS	mp-2724538	31	1.91	2.26	1.44	0.69
High	Bi_2O_3	mp-1105780	159	2.13	1.84	1.15	1.79
High	$Ba_3Ti_2O_7$	mp-770118	42	1.78	1.90	1.66	13.38†
High	$TiPN_3$	mp-989613	38	1.61	1.94	0.82	1.90
High	$BaTiO_3$	mp-558125	20	1.41	1.30	2.03	0.90
High	Rb_2ZrO_3	mp-752401	9	1.15	0.91	0.45	17.64†
High	$Ca_3Sn_2O_7$	mp-1374018	36	0.86	0.91	1.59	0.47
Low	Al_2S_3	mp-2654	169	0.11	0.06	0.33	0.27
Low	ZnP_4	mp-14587	92	0.15	0.18	0.59	0.43
Low	$MgSnP_2$	mp-1009083	122	0.27	0.28	0.11	0.11
Low	$CuBSe_2$	mp-983565	122	0.27	0.29	0.53	0.12
Low	$BaSiO_3$	mp-7339	19	0.32	0.28	0.43	0.30
Low	Li_8SiO_6	mp-28549	185	0.34	0.31	0.62	0.20
Low	$MoW_3(SeS_3)_2$	mp-1029037	156	0.41	0.25	0.34	0.05
Low	$Ba(LaSe_2)_2$	mp-37741	122	0.34	0.36	0.42	0.31
Low	Cu_3BiS_3	mp-607291	19	0.46	0.47	0.42	0.73
Low	$RbCaCl_3$	mp-2795318	99	0.50	0.49	0.36	4.91†

Table 4
Comparison of this work with the two most directly comparable machine-learning studies of the maximum piezoelectric modulus e_m. MAE values are cross-validated results reported by each study, in C/m^2. The comparison is indicative rather than strictly controlled: the studies use overlapping but non-identical Materials Project snapshots and different featurisation and outlier-handling choices.

Study	Best model	Dataset size	Target transform	MAE (C/m^2)
Hu and Song [36]	SVM (Magpie + elastic)	~1700	none	0.646
Anand et al. [38]	GatedGCN (augmented)	~2700	none	0.730
This work	XGBoost	3303	$\log(1+e_m)$	0.498
This work	LightGBM	3303	$\log(1+e_m)$	0.499
This work	ALIGNN	3303	$\log(1+e_m)$	0.500

to the flagged near-instability group and its response is neither high nor consistent across the three models ($\hat{e}_m$ = 0.45–1.15 C/m^2), while the DFT modulus of $TiPN_3$ (1.90 C/m^2) does not clear the 2 C/m^2 screening threshold; both were therefore retained in the validation set but set aside as experimental-synthesis candidates. We recommend $SrTaNO_2$, YWN_3 and $SrTa_2Bi_2O_9$ as the highest-priority targets for experimental synthesis, pending a stability analysis motivated by their large computed ionic responses.

4. Conclusions

We have presented a systematic machine-learning benchmark for the piezoelectric modulus e_m, evaluating thirteen architectures (gradient-boosted tree ensembles, support vector machines, deep neural networks, and seven graph neural networks) on > 3,000 Materials Project compounds, all tuned via Optuna. The three leading models (tabular XGBoost and LightGBM together with the graph-based ALIGNN) reach essentially the same accuracy: for XGBoost, a cross-validated MAE = 0.498 ± 0.049 C/m^2 and $R^2 = 0.285 \pm 0.085$ (5-fold cross-validation). The principal finding is that tabular tree ensembles match the best graph networks on this property while dispensing entirely with crystal-graph construction. The modest R^2, common to every architecture, reflects the intrinsic difficulty and the limited size of the training data, which is why we deploy the models as ranking and enrichment tools rather than as quantitative predictors.

Screening of > 10,000 unlabelled compounds with the three leading models and filtering for non-centrosymmetry, insulating gap (E_g > 0.1 eV) and thermodynamic near-stability (E_{hull} < 0.1 eV/atom) yields eight high-priority lead-free candidates, three of which ($SrTaNO_2$, YWN_3 and $SrTa_2Bi_2O_9$) are flagged by all three models. Independent DFT calculations of the full piezoelectric tensor for twenty compounds qualitatively confirm the screening, with all three consensus candidates exceeding the 2 C/m^2 threshold at the DFT level; these compounds are therefore the most promising synthesis targets.

The validation also exposed a limitation of the modelling strategy: trained exclusively on piezoelectric compounds, the models do not learn centrosymmetry or the electronic gap as the physical constraints that switch piezoelectricity on and off, so these criteria must instead be imposed through downstream filters. Future work should therefore include centrosymmetric and metallic compounds ($e_m = 0$) in the training data, alongside experimental synthesis and characterisation of the most stable candidates, active learning targeted at the high-e_m tail that currently limits accuracy, pre-trained universal GNN potentials as feature extractors for graph-based models, calibrated uncertainty quantification through conformal or ensemble methods, and extension of the screening to quaternary and mixed-anion compositions underrepresented in the current Materials Project piezoelectric database.

CRediT authorship contribution statement

Francesc Suñol: Writing – original draft, Visualization, Software, Methodology. **Claudio Cazorla:** Writing – review & editing, Resources, Investigation, Formal analysis. **José E. Garcia:** Writing – review & editing, Validation, Supervision, Methodology.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

C.C. acknowledges support by MICIN/AEI/10.13039/5011000 11033 and ERDF/EU under the grants CNS2025-165467, PID2023-146623NB-I00, PID2023-147469NB-C21 and the Maria de Maeztu Units of Excellence Programme CEX2023-001300-M, also by the Generalitat de Catalunya under the grants 2021SGR-00343, 2021SGR-01519 and 2021SGR-01411. Computational support was provided by the Red Española de Supercomputación under the grants FI-2024-1-0005, FI-2024-2-0003, FI-2024-3-0004, FI-2024-1-0025, FI-2024-2-0006, and FI-2025-1-0015.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.cocom.2026.e01422.

Data availability

Data will be made available on request.

References

[1] J. Rödel, W. Jo, K.T.P. Seifert, E.-M. Anton, T. Granzow, D. Damjanovic, Perspective on the development of lead-free piezoceramics, J. Am. Ceram. Soc. 92 (6) (2009) 1153–1177, http://dx.doi.org/10.1111/j.1551-2916.2009.03061.x.

[2] B.C. Sekhar, B. Dhanalakshmi, B.S. Rao, S. Ramesh, K.V. Prasad, P.S. Rao, B.P. Rao, Piezoelectricity and its applications, in: D.R. Sahu (Ed.), Multifunctional Ferroelectric Materials, IntechOpen, London, 2021, http://dx.doi.org/10.5772/intechopen.96154.

[3] N. Sezer, M. Koç, A comprehensive review on the state-of-the-art of piezoelectric energy harvesting, Nano Energy 80 (2021) 105567, http://dx.doi.org/10.1016/j.nanoen.2020.105567.

[4] M.C. Sekhar, E. Veena, N.S. Kumar, K.C.B. Naidu, A. Mallikarjuna, D.B. Basha, A review on piezoelectric materials and their applications, Cryst. Res. Technol. 58 (2) (2023) 2200130, http://dx.doi.org/10.1002/crat.202200130.

[5] European Commission, Directive 2011/65/EU of the European Parliament and of the Council on the restriction of the use of certain hazardous substances in electrical and electronic equipment (RoHS recast), 2011, Directive 2011/65/EU Official Journal of the European Union. URL https://eur-lex.europa.eu/eli/dir/2011/65/oj.

[6] D.W. Davies, K.T. Butler, A.J. Jackson, A. Morris, J.M. Frost, J.M. Skelton, A. Walsh, Computational screening of all stoichiometric inorganic materials, Chem 1 (4) (2016) 617–627, http://dx.doi.org/10.1016/j.chempr.2016.09.010.

[7] I. Petousis, D. Mrdjenovich, E. Ballouz, M. Liu, D. Winston, W. Chen, T. Graf, T.D. Schladt, K.A. Persson, F.B. Prinz, High-throughput screening of inorganic compounds for the discovery of novel dielectric and optical materials, Sci. Data 4 (2017) 160134, http://dx.doi.org/10.1038/sdata.2016.134.

[8] M. de Jong, W. Chen, T. Angsten, A. Jain, R. Notestine, A. Gamst, M. Sluiter, C. Krishna Ande, S. van der Zwaag, J.J. Plata, C. Toher, S. Curtarolo, G. Ceder, K.A. Persson, M. Asta, Charting the complete elastic properties of inorganic crystalline compounds, Sci. Data 2 (2015) 150009, http://dx.doi.org/10.1038/sdata.2015.9.

[9] M.K. Horton, P. Huck, R.X. Yang, J.M. Munro, S. Dwaraknath, A.M. Ganose, R.S. Kingsbury, M. Wen, J.X. Shen, T.S. Mathis, A.D. Kaplan, K. Berket, J. Riebesell, J. George, A.S. Rosen, E.W. Spotte-Smith, M.J. McDermott, O.A. Cohen, A. Dunn, M.C. Kuner, G.M. Rignanese, G. Petretto, D. Waroquiers, S.M. Griffin, J.B. Neaton, D.C. Chrzan, M. Asta, G. Hautier, S. Cholia, G. Ceder, S.P. Ong, A. Jain, K.A. Persson, Accelerated data-driven materials science with the materials project, Nat. Mater. 24 (10) (2025) 1522–1532, http://dx.doi.org/10.1038/s41563-025-02272-0.

[10] A. Jain, S.P. Ong, G. Hautier, W. Chen, W.D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder, K.A. Persson, Commentary: The materials project: a materials genome approach to accelerating materials innovation, APL Mater. 1 (1) (2013) 011002, http://dx.doi.org/10.1063/1.4812323.

[11] S. Curtarolo, W. Setyawan, S. Wang, J. Xue, K. Yang, R.H. Taylor, L.J. Nelson, G.L.W. Hart, S. Sanvito, M. Buongiorno-Nardelli, N. Mingo, O. Levy, AFLOWLIB.ORG: a distributed materials properties repository from high-throughput *ab initio* calculations, Comput. Mater. Sci. 58 (2012) 227–235, http://dx.doi.org/10.1016/j.commatsci.2012.02.002.

[12] K. Choudhary, K.F. Garrity, A.C.E. Reid, B. DeCost, A.J. Biacchi, A.R. Hight Walker, Z. Trautt, J. Nelson, A.G. Kusne, A. Centrone, A. Davydov, J. Jiang, R. Pachter, G. Cheon, E. Reed, A. Agrawal, X. Qian, V. Sharma, H. Zhu, I.T. Foster, Y. Shao-Horn, F. Buda, U. Johnson, C.M. Breneman, S.K. Misra, The joint automated repository for various integrated simulations (JARVIS) for data-driven materials design, Npj Comput. Mater. 6 (2020) 173, http://dx.doi.org/10.1038/s41524-020-00440-1.

[13] G.R. Schleder, A.C.M. Padilha, C.M. Acosta, M. Costa, A. Fazzio, From DFT to machine learning: recent approaches to materials science – a review, J. Phys.: Mater. 2 (3) (2019) 032001, http://dx.doi.org/10.1088/2515-7639/ab084b.

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

[15] Y. Wu, I. Abdelwahab, K.C. Kwon, I. Verzhbitskiy, L. Wang, W.H. Liew, K. Yao, G. Eda, K.P. Loh, L. Shen, S.Y. Quek, Data-driven discovery of high performance layered van der waals piezoelectric $NbOI_2$, Nat. Commun. 13 (2022) 1884, http://dx.doi.org/10.1038/s41467-022-29495-y.

[16] J. Zhou, G. Cui, S. Hu, Z. Zhang, C. Yang, Z. Liu, L. Wang, C. Li, M. Sun, Graph neural networks: A review of methods and applications, AI Open 1 (2020) 57–81, http://dx.doi.org/10.1016/j.aiopen.2021.01.001.

[17] T. Xie, J.C. Grossman, Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties, Phys. Rev. Lett. 120 (14) (2018) 145301, http://dx.doi.org/10.1103/PhysRevLett.120.145301.

[18] S.-Y. Louis, Y. Zhao, A. Nasiri, X. Wang, Y. Song, F. Liu, J. Hu, Graph convolutional neural networks with global attention for improved materials property prediction, Phys. Chem. Chem. Phys. 22 (32) (2020) 18141–18148, http://dx.doi.org/10.1039/D0CP01474E.

[19] K. Choudhary, B. DeCost, Atomistic line graph neural network for improved materials property predictions, Npj Comput. Mater. 7 (2021) 185, http://dx.doi.org/10.1038/s41524-021-00650-1.

[20] S. Batzner, A. Musaelian, L. Sun, M. Geiger, J.P. Mailoa, M. Kornbluth, N. Molinari, T.E. Smidt, B. Kozinsky, E(3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials, Nat. Commun. 13 (2022) 2453, http://dx.doi.org/10.1038/s41467-022-29939-5.

[21] K.T. Schütt, O.T. Unke, M. Gastegger, Equivariant message passing for the prediction of tensorial properties and molecular spectra, in: Proceedings of the 38th International Conference on Machine Learning, ICML, PMLR, 2021, pp. 9377–9388.

[22] I. Batatia, D.P. Kovacs, G.N.C. Simm, C. Ortner, G. Csányi, MACE: higher order equivariant message passing neural networks for fast and accurate force fields, in: Advances in Neural Information Processing Systems, vol. 35, Curran Associates, 2022.

[23] C. Chen, S.P. Ong, A universal graph deep learning interatomic potential for the periodic table, Nat. Comput. Sci. 2 (11) (2022) 718–728, http://dx.doi.org/10.1038/s43588-022-00349-3.

[24] B. Deng, P. Zhong, K. Jun, J. Riebesell, K. Han, C.J. Bartel, G. Ceder, CHGNet as a pretrained universal neural network potential for charge-informed atomistic modelling, Nat. Mach. Intell. 5 (9) (2023) 1031–1041, http://dx.doi.org/10.1038/s42256-023-00716-3.

[25] I. Batatia, P. Benner, Y. Chiang, A.M. Elena, D.P. Kovács, J. Riebesell, X.R. Advincula, M. Asta, W.J. Oras, G. Csányi, G.N.C. Simm, et al., A foundation model for atomistic materials chemistry, 2024, ArXiv arXiv:2401.00096. MACE-MP-0 foundation model.

[26] A. Merchant, S. Batzner, S.S. Schoenholz, M. Aykol, G. Cheon, E.D. Cubuk, Scaling deep learning for materials discovery, Nature 624 (7990) (2023) 80–85, http://dx.doi.org/10.1038/s41586-023-06735-9.

[27] T. Chen, C. Guestrin, XGBoost: a scalable tree boosting system, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2016, pp. 785–794, http://dx.doi.org/10.1145/2939672.2939785.

[28] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, T.-Y. Liu, LightGBM: a highly efficient gradient boosting decision tree, in: Advances in Neural Information Processing Systems, vol. 30, Curran Associates, 2017.

[29] L. Prokhorenkova, G. Gusev, A. Vorobev, A.V. Dorogush, A. Gulin, CatBoost: unbiased boosting with categorical features, in: Advances in Neural Information Processing Systems, vol. 31, Curran Associates, 2018.

[30] A. Dunn, Q. Wang, A. Ganose, D. Dopp, A. Jain, Benchmarking materials property prediction methods: the MatBench test set and automatminer reference algorithm, Npj Comput. Mater. 6 (2020) 138, http://dx.doi.org/10.1038/s41524-020-00406-3.

[31] S.Ö. Arik, T. Pfister, TabNet: attentive interpretable tabular learning, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, 2021, pp. 6679–6687, http://dx.doi.org/10.1609/aaai.v35i8.16826.

[32] T. Akiba, S. Sano, T. Yanase, T. Ohta, M. Koyama, Optuna: a next-generation hyperparameter optimization framework, in: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2019, pp. 2623–2631, http://dx.doi.org/10.1145/3292500.3330701.

[33] J. Bergstra, D. Yamins, D.D. Cox, Making a science of model search: hyperparameter optimization in hundreds of dimensions for vision architectures, in: Proceedings of the 30th International Conference on Machine Learning, ICML, PMLR, 2013, pp. 115–123.

[34] S.M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, in: Advances in Neural Information Processing Systems, vol. 30, Curran Associates, 2017.

[35] S.J. Wetzel, S. Ha, R. Iten, M. Klopotek, Z. Liu, Interpretable machine learning in physics: A review, 2025, arXiv:2503.23616, URL https://arxiv.org/abs/2503.23616.

[36] J. Hu, Y. Song, Piezoelectric modulus prediction using machine learning and graph neural networks, Chem. Phys. Lett. 791 (2022) 139359, http://dx.doi.org/10.1016/j.cplett.2022.139359.

[37] H.R. Rocha, R. Roukos, S. Abou Dargham, J. Romanos, D. Chaumont, J.A. Silva, H. Wörtche, Optimizing a machine learning design of dielectric properties in lead-free piezoelectric ceramics, Mater. Des. 243 (2024) 113053, http://dx.doi.org/10.1016/j.matdes.2024.113053.

[38] A. Anand, P. Kumari, A.K. Kalyani, High throughput screening of new piezoelectric materials using graph machine learning and knowledge graph approach, Comput. Mater. Sci. 246 (2025) 113445, http://dx.doi.org/10.1016/j.commatsci.2024.113445.

[39] L. Ward, A. Agrawal, A. Choudhary, C. Wolverton, A general-purpose machine learning framework for predicting properties of inorganic materials, Npj Comput. Mater. 2 (2016) 16028, http://dx.doi.org/10.1038/npjcompumats.2016.28.

[40] L. Ward, A. Dunn, A. Faghaninia, N.E.R. Zimmermann, S. Bajaj, Q. Wang, J. Montoya, J. Chen, K. Bystrom, M. Dylla, K. Chard, M. Asta, K.A. Persson, G.J. Snyder, I. Foster, A. Jain, Matminer: an open source toolkit for materials data mining, Comput. Mater. Sci. 152 (2018) 60–69, http://dx.doi.org/10.1016/j.commatsci.2018.05.018.

[41] C. Cortes, V. Vapnik, Support-vector networks, Mach. Learn. 20 (3) (1995) 273–297, http://dx.doi.org/10.1007/BF00994018.

[42] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Köpf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, S. Chintala, PyTorch: an imperative style, high-performance deep learning library, in: Advances in Neural Information Processing Systems, vol. 32, Curran Associates, 2019.

[43] D.P. Kingma, J. Ba, Adam: a method for stochastic optimization, in: Proceedings of the 3rd International Conference on Learning Representations, ICLR, 2015.

[44] K.T. Schütt, P.-J. Kindermans, H.E. Sauceda, S. Chmiela, A. Tkatchenko, K.-R. Müller, SchNet: a continuous-filter convolutional neural network for modeling quantum interactions, in: Advances in Neural Information Processing Systems, vol. 30, Curran Associates, 2017.

[45] J. Gilmer, S.S. Schoenholz, P.F. Riley, O. Vinyals, G.E. Dahl, Neural message passing for quantum chemistry, in: Proceedings of the 34th International Conference on Machine Learning, ICML, PMLR, 2017, pp. 1263–1272.

[46] C. Chen, W. Ye, Y. Zuo, C. Zheng, S.P. Ong, Graph networks as a universal machine learning framework for molecules and crystals, Chem. Mater. 31 (9) (2019) 3564–3572, http://dx.doi.org/10.1021/acs.chemmater.9b01294.

[47] J. Schmidt, L. Pettersson, C. Verdozzi, S. Botti, M.A.L. Marques, Crystal graph attention networks for the prediction of stable materials, Sci. Adv. 7 (49) (2021) eabi7948, http://dx.doi.org/10.1126/sciadv.abi7948.

[48] J. Bergstra, R. Bardenet, Y. Bengio, B. Kégl, Algorithms for hyper-parameter optimization, in: Advances in Neural Information Processing Systems, vol. 24, Curran Associates, 2011.

[49] C. Cazorla, J. Boronat, Simulation and understanding of atomic and molecular quantum crystals, Rev. Modern Phys. 89 (2017) 035003, http://dx.doi.org/10.1103/RevModPhys.89.035003.

[50] J.P. Perdew, A. Ruzsinszky, G.I. Csonka, O.A. Vydrov, G.E. Scuseria, L.A. Constantin, X. Zhou, K. Burke, Restoring the density-gradient expansion for exchange in solids and surfaces, Phys. Rev. Lett. 100 (2008) 136406, http://dx.doi.org/10.1103/PhysRevLett.100.136406.

[51] G. Kresse, J. Furthmüller, Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set, Phys. Rev. B 54 (1996) 11169–11186, http://dx.doi.org/10.1103/PhysRevB.54.11169.

[52] P.E. Blöchl, Projector augmented-wave method, Phys. Rev. B 50 (1994) 17953–17979, http://dx.doi.org/10.1103/PhysRevB.50.17953.

[53] L. Zhao, B. Wang, D. Chu, C. Cazorla, First-principles high-throughput screening of bulk piezo-photocatalytic materials for sunlight-driven hydrogen production, J. Mater. Chem. A 10 (2022) 18132–18146, http://dx.doi.org/10.1039/D2TA04160J.

[54] L. van der Maaten, G. Hinton, Visualizing data using t-SNE, J. Mach. Learn. Res. 9 (2008) 2579–2605.

[55] X. Gonze, C. Lee, Dynamical matrices, Born effective charges, dielectric permittivity tensors, and interatomic force constants from density-functional perturbation theory, Phys. Rev. B 55 (16) (1997) 10355–10368, http://dx.doi.org/10.1103/PhysRevB.55.10355.