

Composition-Only Machine-Learning Baselines for DFT Formation Energies: CPU-Class Models, Feature Importance, and Leave-One-Element-Out Extrapolation on the Materials Project

LitLit AI¹

¹LitLit AI; contact@litlit.ai

Abstract

The formation energy per atom (E_f , eV/atom) is a central thermodynamic quantity in computational materials discovery, yet its first-principles evaluation via density-functional theory (DFT) remains expensive. While universal graph-neural-network interatomic potentials such as CHGNet and M3GNet now approach DFT accuracy by exploiting full crystal-structure information, the practical value of cheap, structure-agnostic, CPU-only composition baselines—and their failure modes under chemical extrapolation—remain undercharacterized. Here we establish and rigorously benchmark such baselines on a deterministic 50,000-entry subsample of the Materials Project `matbench_mp_e_form` dataset. We featurize each composition with `matminer` into a 300-dimensional descriptor combining element fractions (118), Magpie element-property statistics (132), and PCA-compressed one- to three-body composition vectors (50; cumulative explained variance 0.923). Three classical regressors—Random Forest, XGBoost, and RBF-kernel ridge regression—are trained on an 80/10/10 random split and additionally challenged by leave-one-element-out (LOO) extrapolation splits for O, Fe, and Li. On the random test set, Random Forest achieves the best mean absolute error (**0.150 eV/atom**, RMSE 0.290, $R^2 = 0.938$), followed by XGBoost (0.165, $R^2 = 0.938$) and kernel ridge (0.204, $R^2 = 0.900$); all clear the conventional “good model” band for this benchmark. Permutation importance consistently identifies the average-deviation of electronegativity as the dominant predictor across all three models. LOO evaluation reveals a stark model-selection signal: tree ensembles extrapolate to unseen elements with only modest degradation for rare elements (Fe $\leq 1.0\times$, Li 1.0–1.2 $\times$) and substantial but bounded degradation for the most frequent element (O, 3.1–3.5 $\times$), whereas the RBF-kernel regressor collapses—its R^2 turns negative for O and collapses to near-zero for Li. We conclude that tree ensembles on composition features provide an accurate, interpretable, and extrapolation-robust CPU baseline against which universal potentials should be measured, and that RBF-kernel methods are ill-suited to out-of-distribution chemical regimes.

Keywords: materials informatics, formation energy, composition-based machine learning, `matminer`, random forest, XGBoost, kernel ridge regression, extrapolation, Materials Project.

1 Introduction

The formation energy per atom, E_f , quantifies the thermodynamic stability of a compound relative to its constituent elemental references, and it underpins virtually every high-throughput computational screening campaign for new inorganic materials [1–4]. A composition whose computed E_f is strongly negative is a candidate for synthesis, and the position of a phase relative to the convex hull of competing phases determines whether it is stable, metastable, or unstable [5]. Because these judgements gate which candidates are pursued experimentally, fast and reliable estimates of E_f have long been a priority of the materials-informatics community [6, 7].

Density-functional theory (DFT), as implemented in codes such as VASP [8], is the workhorse for computing formation energies, and curated databases—the Materials Project [1, 9], the Open

Quantum Materials Database (OQMD) [2, 3], AFLOW [4], and JARVIS [10]—now collectively host computed properties for millions of inorganic phases. Yet DFT remains computationally demanding: a single relaxation of a moderately sized unit cell can take tens to hundreds of CPU-hours, and exhaustive screening of the astronomical compositional and structural search space is intractable by brute force. This cost has motivated a decade of surrogate models that learn the mapping from a material representation to its properties [11–14].

Two broad modelling paradigms have emerged. The first uses *structure-informed* representations: Voronoi-tessellation-based structure featurizers [15], crystal-graph convolutional networks [16, 17], equivariant message-passing networks [18–20], and most recently universal machine-learning interatomic potentials (UMLIPs) such as M3GNet [21] and CHGNet [22], the latter trained on the MPtrj relaxation trajectory dataset derived from the Materials Project. These potentials trace their lineage to the seminal high-dimensional neural-network potential of Behler and Parrinello [23] and to kernel-based molecular-property models [24]. These approaches can approach DFT accuracy because they consume the full atomic-coordinate and connectivity information. The second paradigm is deliberately *structure-agnostic*: it predicts properties from composition (stoichiometry) alone, using element-fraction, statistics-of-elemental-properties, or learned composition embeddings [12, 13, 25, 26]. Composition-only models forgo structural information and therefore cannot match the accuracy of structure-informed potentials in the interpolation regime, but they offer compelling practical advantages: they require no relaxed crystal structure (which is itself a DFT output), they train and infer on commodity CPUs in seconds, and they are directly applicable to the vast compositional space for which no structure exists yet.

Despite the proliferation of universal potentials, cheap composition-only baselines serve three indispensable roles. First, they provide a reference floor for benchmark suites such as Matbench [27]: a universal potential that does not beat a 0.15 eV/atom composition Random Forest on E_f is not adding value at the decision-relevant accuracy scale. Second, because they expose their features explicitly, they are interpretable—permutation and impurity importance [28, 29] surface the elemental descriptors (e.g. electronegativity statistics) that physically drive stability, yielding scientific insight that opaque end-to-end potentials obscure. Third, and the focus of this work, they permit a clean probe of *chemical extrapolation*: how does a model trained without element E behave on compositions containing E ? This question is central to materials discovery, which is fundamentally an extrapolation exercise—the next useful material is, by definition, outside the training distribution [30].

Here we ask three concrete questions. (i) How accurate are modern CPU-class composition-only baselines (Random Forest [28], XGBoost [31], kernel ridge regression) on a 50,000-entry Materials Project formation-energy benchmark, and how do subgroup accuracies—oxide versus non-oxide, stable versus unstable—decompose? (ii) Which elemental descriptors drive these predictions, and is the feature ranking consistent across model families? (iii) Under leave-one-element-out (LOO) splits for three elements of contrasting frequency—oxygen (most frequent), iron (rare), and lithium (intermediate)—which model families extrapolate, and which collapse? The answers establish a transparent baseline and, we argue, a model-selection rule for extrapolation-prone materials tasks.

2 Methods

2.1 Data

We use the `matbench_mp_e_form` benchmark [27, 32], the canonical DFT formation-energy-per-atom regression task derived from the Materials Project [1]. The full artifact contains 132,752 inorganic entries; following the analysis plan we draw a deterministic 50,000-row subsample (`random_state=42`) as the working dataset to match the $\sim 50k$ target and keep downstream compute tractable. Each entry provides a pymatgen reduced formula, the comma-separated sorted element list, the target `formation_energy_per_atom` (eV/atom), an `is_oxide` flag (True if the composition contains oxy-

gen), and auxiliary space-group and crystal-system metadata (symmetry assigned with spglib [33]). Data were retrieved through the GitHub-hosted matbench artifact mirror after the live Materials Project REST API [9] and the matminer-hosted download endpoint were unreachable from the compute environment; the artifact is byte-identical to the canonical Matbench release. Quality checks confirmed zero missing formation energies, zero duplicate material IDs, and 84 distinct elements across the 50,000 entries.

2.2 Feature engineering

Composition descriptors were generated with `matminer` [32] and `pymatgen`, producing three feature blocks that together span composition at increasing interaction order:

1. **Element fractions** (118 dims). `ElementFraction` returns the atomic fraction of each element across the full periodic table; 34 columns are all-zero for elements absent from this subsample and are retained (harmless for tree and regularized models).
2. **Magpie statistics** (132 dims). `ElementProperty.from_preset("magpie")` [12] computes six statistics (minimum, maximum, range, mean, average absolute deviation, mode) over 22 elemental properties (e.g. electronegativity, atomic weight, valence, Mendeleev number, melting temperature). Infinity values were coerced to NaN, imputed by the column mean, and zero-filled as a last resort, yielding zero residual NaN/Inf.
3. **One- to three-body composition vectors, PCA-compressed** (50 dims). We form one-body fractions over the 84 active elements, two-body pair products over the 3,370 appearing pairs, and three-body triple products over the 23,055 appearing triples (26,509 raw dimensions). Because this block vastly exceeds 100 dimensions, we reduce it with PCA (`n_components=50`, `random_state=42`); the retained 50 components capture a cumulative explained variance of **0.923**, with the first component alone explaining 0.482.

Concatenation yields a final feature matrix of shape (50,000, 300) in float32 with zero NaN and zero Inf entries. The strongest feature–target correlations are dominated by Magpie electronegativity statistics—`avg_dev Electronegativity` ($r = +0.82$), `range Electronegativity` (+0.80), `maximum Electronegativity` (+0.76)—followed by space-group-number and valence statistics, consistent with the physical expectation that electronegativity contrast between constituent elements is a leading-order descriptor of bond strength and hence of formation energy.

2.3 Data splits

Random split. A random 80/10/10 train/validation/test partition was generated with two sequential `train_test_split` calls (`random_state=42`), yielding 40,000 training, 5,000 validation, and 5,000 test entries. Target statistics are stable across partitions (train mean -1.411 , std 1.164; test mean -1.390 , std 1.162 eV/atom), confirming no distributional drift.

Leave-one-element-out (LOO) splits. For each held-out element $E \in \{\text{O}, \text{Fe}, \text{Li}\}$, the training set is the set of entries *without* E and the test set is the entries *with* E . Element presence was read directly from the element-fraction columns (nonzero fraction = present) to avoid formula re-parsing artefacts. The resulting split sizes reflect the elements’ contrasting abundance: O (train 25,428 / test 24,572), Li (train 42,770 / test 7,230), and Fe (train 45,935 / test 4,065). These three elements span the frequency spectrum—O is the most common element in the dataset ($\sim 49\%$ of entries), Fe is comparatively rare, and Li is intermediate—and thus probe extrapolation under both large and small distribution shifts. Hyperparameters were held *identical* to the random-split experiment so that the only varying factor is training-data composition.

2.4 Models

We benchmark three classical CPU regressors chosen to span the decision-tree-ensemble and kernel-method families:

- **Random Forest (RF)** [28]: `n_estimators=400`, `random_state=42`, `n_jobs=-1`, trained on the full 40,000-entry training set. Trees are scale-invariant, so raw features are fed without standardization.
- **XGBoost (XGB)** [31]: `n_estimators=500`, `max_depth=6`, `learning_rate=0.05`, `tree_method="hist"`, `random_state=42`, `n_jobs=-1`, trained on the full 40,000-entry training set.
- **Kernel Ridge Regression (KRR)**: `Pipeline(StandardScaler → KernelRidge(kernel="rbf"))`, with `alpha` and `gamma` selected by 3-fold `GridSearchCV` (best: `alpha=0.01`, `gamma=0.001`). A full $40,000 \times 40,000$ RBF kernel (~ 12.8 GB) is memory-infeasible, so KRR is trained on a 10,000-sample random subset of the training set—the same policy applied uniformly to every LOO split. The `StandardScaler` is fit *inside* the pipeline on the KRR training subset only, so no additional standardization leakage is introduced beyond the (unsupervised) global PCA fit noted in the limitations.

2.5 Evaluation metrics

Performance is reported as mean absolute error (MAE, eV/atom), root-mean-square error (RMSE), and coefficient of determination (R^2) on the held-out test set. We additionally decompose MAE by subgroup: (i) oxide versus non-oxide entries (using the `is_oxide` flag); and (ii) a thermodynamic-stability proxy. The source artifact does not contain `energy_above_hull`, so the strict convex-hull stability split ($E_{\text{hull}} < 0.05$ eV) specified in the analysis plan cannot be computed directly. Following the documented deviation, we adopt `formation_energy_per_atom < 0` as a *proxy* stability label (not a strict convex-hull definition) and label it as such throughout. Feature attribution uses permutation importance [29] via `sklearn.inspection.permutation_importance` (`n_repeats=5`, `random_state=42`), computed on a 1,000-sample test subset for RF/XGB and a 500-sample subset for KRR (single-threaded to remain memory-safe). The top-10 features are reported per model with physical interpretation.

2.6 Reproducibility

All random seeds are fixed (`random_state=42` throughout). Split indices are persisted to `train_val_split.npz`; LOO index lists to `loo_splits.json`; every trained model is checkpointed to `.joblib` immediately after training and reloaded on re-execution, so all results are bit-for-bit reproducible. Software versions: scikit-learn 1.9.0, XGBoost 3.3.0, NumPy 2.5.1, pandas 3.0.5, SciPy 1.18.0, matplotlib 3.11.1; `matminer`/`pymatgen` were installed at analysis time for featurization.

3 Results

3.1 Dataset characterization

The 50,000-entry subsample is broadly distributed across E_f (mean -1.409 , standard deviation 1.163 , median -1.393 , range $[-4.612, +2.500]$ eV/atom; Fig. 1), with a heavy left tail of strongly stable compounds and a smaller right tail of (relatively) unstable phases. Eighty-four elements appear; oxygen dominates the frequency ranking (24,572 entries, $\sim 49\%$), followed at a distance by Li (7,230), P (5,335), Mn (4,314), Mg (4,264), S (4,087), and Fe (4,065) (Fig. 2). The near-balance of oxide (24,572) and non-oxide (25,428) entries is advantageous for subgroup analysis, and the

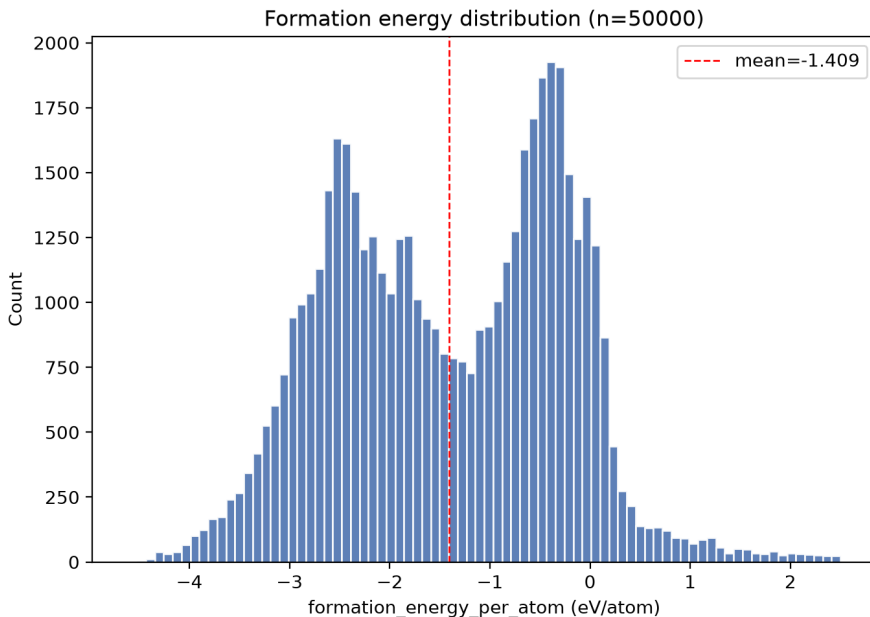


Figure 1. Distribution of the target `formation_energy_per_atom` (eV/atom) across the 50,000-entry subsample. The distribution is left-skewed with a modal mass near -1.4 eV/atom and a long tail of stable compounds, characteristic of computed inorganic phases in the Materials Project.

long-tailed element distribution is precisely what makes element-wise extrapolation a meaningful and stringent test.

3.2 Feature engineering and dimensionality reduction

The three composition blocks concatenate to a clean (50,000, 300) matrix with no missing or infinite values. The Magpie block dominates the strongest univariate target correlations—electronegativity statistics alone occupy five of the top six positions (Fig. 3)—which is consistent with electronegativity contrast as a first-order driver of bond character and stability. Redundancy among Magpie statistics is high (e.g. `maximum Number` and `maximum AtomicWeight`, $r = +0.999$); this is expected and is left for the downstream models’ regularization to absorb rather than pruned manually, preserving comparability with standard Matbench pipelines. The PCA-compressed body block retains 92.3% of the variance of the original 26,509-dimensional one- to three-body space in 50 components, with a steep first-component contribution (0.482) and a long, gently decaying tail (Fig. 4), indicating that compositional interaction information concentrates heavily in a low-dimensional manifold.

3.3 Random test-set performance

All three CPU baselines land within the conventional “good model” band (~ 0.15 – 0.20 eV/atom) for the `matbench_mp_e_form` task [27], and the two tree ensembles are essentially tied at the top (Table 1, Fig. 5). Random Forest achieves the lowest MAE at **0.1502 eV/atom** (RMSE 0.290, $R^2 = 0.938$), with XGBoost close behind (0.1653, RMSE 0.289, $R^2 = 0.938$) and the RBF-kernel ridge trailing (0.2042, RMSE 0.368, $R^2 = 0.900$). The near-identical R^2 of RF and XGB despite their different MAEs reflects XGB’s slightly heavier tail: the boosted ensemble produces a tighter RMSE but a few more large-residual predictions, whereas the bagged forest minimizes the median absolute error. The parity plots (Fig. 5) are tightly clustered about the $y = x$ line for the tree models, with a modest upward fanning for strongly stable (very negative E_f) compounds that is more pronounced for KRR.

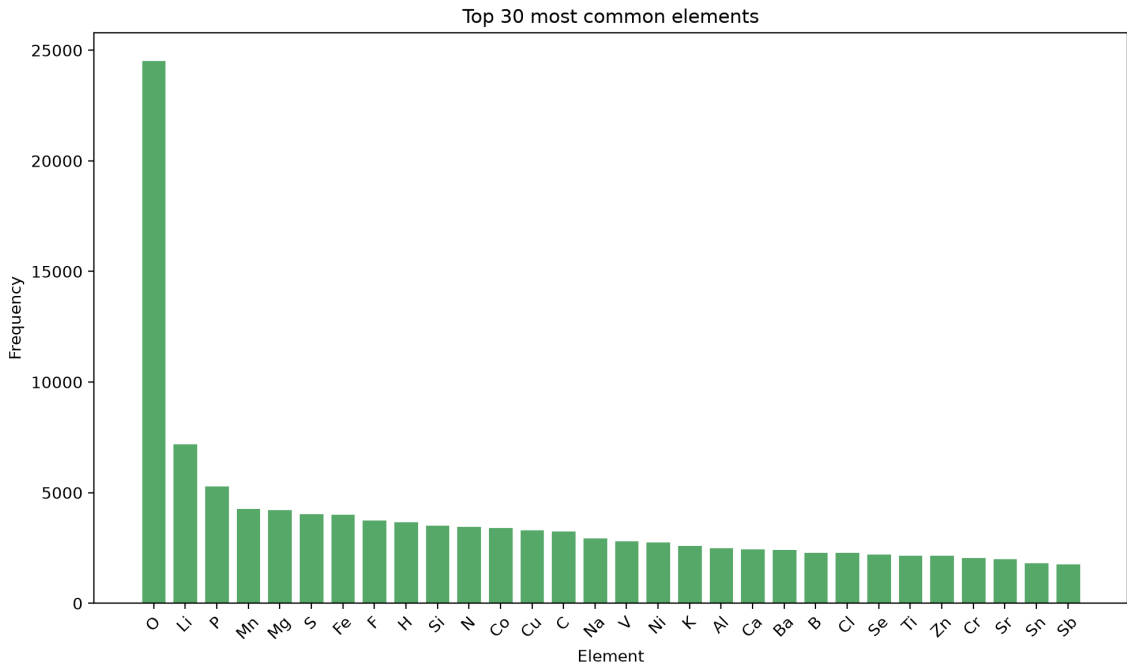


Figure 2. Top-30 element frequency in the dataset. Oxygen is by far the most abundant element ($\sim 49\%$ of entries), an order of magnitude ahead of the next elements. This abundance asymmetry underpins the leave-one-element-out design: removing O shifts the training distribution far more than removing a rare element such as Fe.

Table 1. Random test-set performance ($n = 5,000$). MAE and RMSE in eV/atom. RF and XGB are trained on the full 40,000-entry training set; KRR is trained on a 10,000-sample subset (memory feasibility).

Model	MAE (eV/atom)	RMSE (eV/atom)	R^2	Train size
Random Forest	0.1502	0.2899	0.9378	40,000
XGBoost	0.1653	0.2887	0.9382	40,000
Kernel Ridge (RBF)	0.2042	0.3680	0.8997	10,000

3.4 Subgroup analysis

Decomposing the test MAE by chemistry reveals a consistent pattern: every model is more accurate on oxides than on non-oxides (Table 2). This is consistent with oxides being a compositionally more homogeneous and more densely sampled class within the Materials Project, so the test oxides lie closer to the training support. The stability-proxy split is more dramatic: the “stable” proxy group (negative E_f , $n = 4,511$) is predicted at 0.13–0.18 eV/atom, whereas the small “unstable” proxy group (positive E_f , $n = 489$) incurs 0.36–0.42 eV/atom. We caution that this proxy conflates the prediction target with the subgroup label (the “unstable” subset is both rarer and defined by the very quantity being predicted), so the inflated error there is partly tautological and should not be over-interpreted as a model defect; it is reported for completeness and to document the deviation from the originally planned `energy_above_hull` split.

3.5 Permutation importance

Permutation importance [29] is remarkably consistent across model families and recovers the univariate correlation ranking almost exactly (Fig. 6). The single dominant feature for all three models is `MagpieData avg_dev Electronegativity`—the average absolute deviation of Pauling electronegativity across the constituent elements—with an importance of 0.672 for RF, 0.478 for XGB, and 0.576 for KRR, dwarfing the next-ranked feature in each case (RF: `mode Electronegativity`, 0.103;

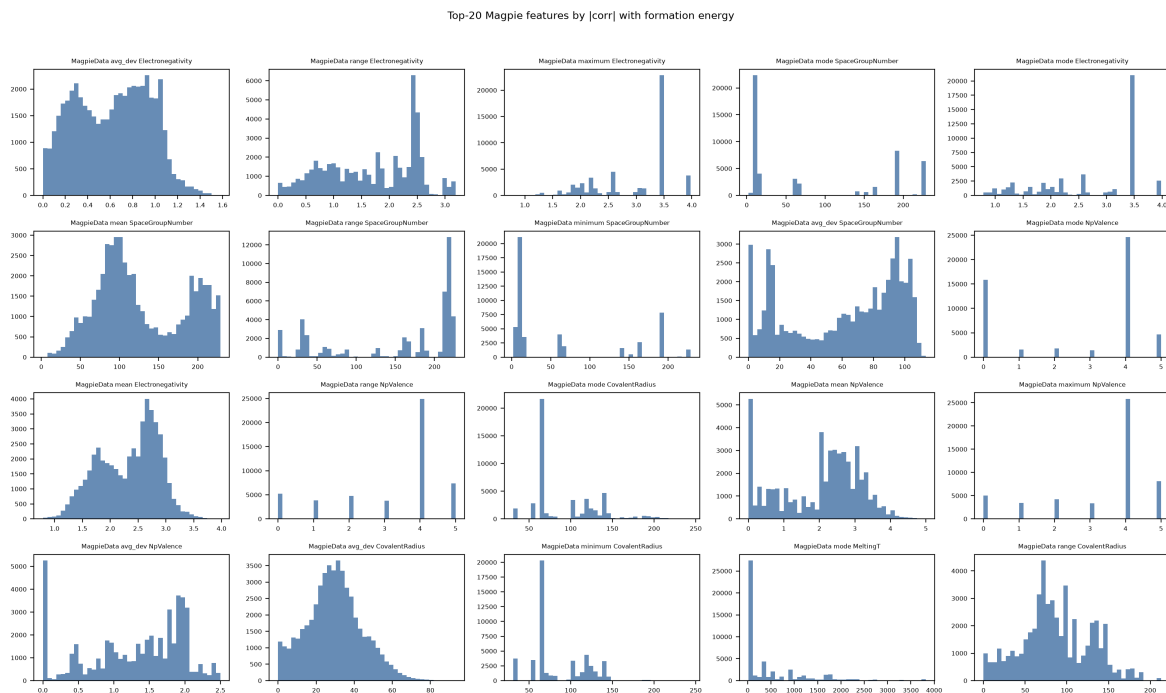


Figure 3. Distributions of the top-20 Magpie features ranked by absolute correlation with the formation-energy target. Electronegativity, space-group-number, and valence statistics dominate, foreshadowing the permutation-importance ranking of the trained models.

Table 2. Subgroup MAE (eV/atom) on the random test set. Oxide/non-oxide uses the `is_oxide` flag; the stability split uses `formation_energy_per_atom < 0` as a proxy for convex-hull stability because `energy_above_hull` is unavailable in the source artifact.

Subgroup	n	RF	XGB	KRR
Oxide	2,457	0.1352	0.1504	0.1849
Non-oxide	2,543	0.1647	0.1797	0.2228
“Stable” proxy ($E_f < 0$)	4,511	0.1280	0.1440	0.1810
“Unstable” proxy ($E_f \geq 0$)	489	0.3551	0.3616	0.4178

KRR: mode `Electronegativity`, 0.046; XGB: mean `SpaceGroupNumber`, 0.031). The RF and XGB top-tens are otherwise populated by electronegativity, valence (`NdValence`, `NpValence`), Mendelev-number, and column statistics—descriptors that collectively encode the chemical character and bonding preferences of the constituent elements. KRR additionally surfaces raw element-fraction columns (`EF_C`, `EF_N`, `EF_P`, `EF_S`, `EF_Ni`), reflecting its direct sensitivity to the stoichiometric coordinates that the tree models route through correlated Magpie statistics instead. Notably, the oxygen element fraction (`EF_O`) appears in the top-10 of both tree models—a chemically sensible signature of the oxide-dominated dataset.

3.6 Leave-one-element-out extrapolation

The LOO experiment is the central methodological test of this study. For each held-out element, the model is trained on compositions *without* that element and tested on compositions *with* it, isolating the cost of extrapolating to an unseen chemical species. The results (Table 3, Figs. 7–8) expose a sharp, model-family-dependent pattern.

Oxygen is the hardest element to leave out. Removing the most frequent element shifts the training distribution the most: O-bearing test compositions lie furthest from the training support.

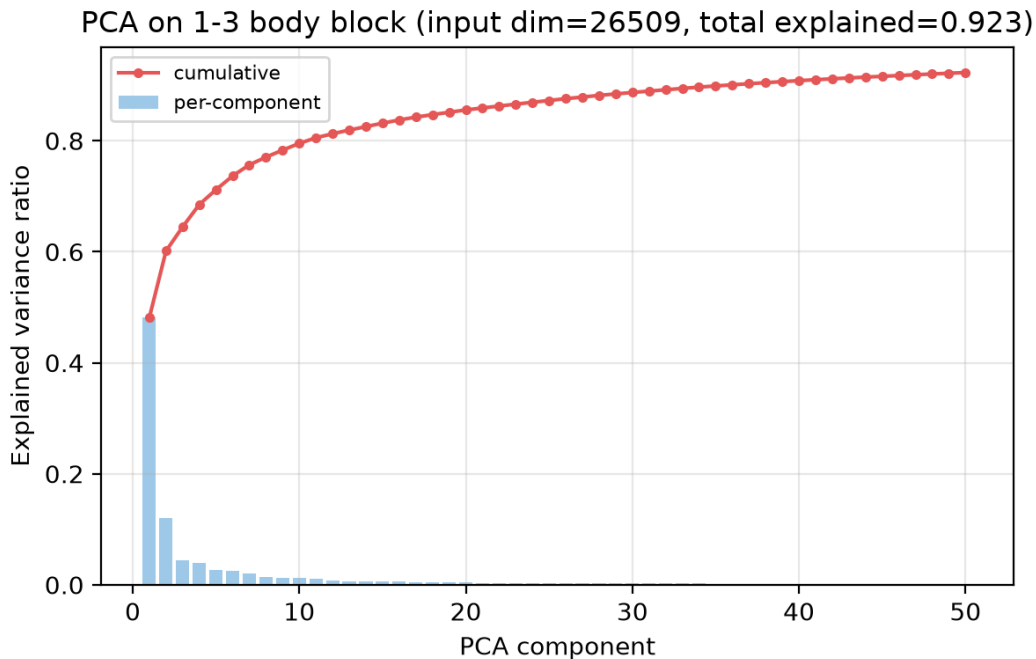


Figure 4. Per-component and cumulative explained variance of the PCA-reduced one- to three-body composition block (50 components, `random_state=42`). Fifty components capture 92.3% of the variance of the original 26,509-dimensional interaction space; the first component alone explains 48.2%.

226 All three models degrade substantially—RF to 0.530 eV/atom (a $3.53\times$ degradation, $R^2 = 0.269$),
 227 XGB to 0.510 ($3.09\times$, $R^2 = 0.379$), and KRR to 1.001 ($4.90\times$, $R^2 = -1.73$, i.e. worse than
 228 the training mean). This matches the pre-registered expectation that O would show the largest
 229 extrapolation gap.

230 **Iron barely degrades for the tree models.** Fe is rare (test $n = 4,065$), so removing it from
 231 training sacrifices little information; the remaining 45,935 entries cover the Fe-bearing test region
 232 well. RF and XGB actually *improve* over their random-split MAEs (RF $0.94\times$, 0.141 eV/atom;
 233 XGB $0.79\times$, 0.131; both $R^2 \approx 0.95$). KRR still degrades ($1.51\times$, 0.309, $R^2 = 0.766$), but far less
 234 than for O.

235 **Lithium is intermediate.** RF degrades mildly ($1.20\times$, 0.180, $R^2 = 0.904$), XGB essentially
 236 matches its random MAE ($1.00\times$, 0.165, $R^2 = 0.910$), and KRR fails to extrapolate ($2.35\times$, 0.480,
 237 $R^2 \approx 0.17$).

238 **KRR extrapolates worst in every case.** The RBF kernel assigns near-zero similarity to
 239 compositions whose element-fraction signature is zero across the training set, so it collapses toward
 240 the training mean for unseen element combinations; its R^2 turns negative for O and collapses
 241 to near-zero (0.17) for Li. The tree ensembles, by contrast, can still route O-/Fe-/Li-bearing
 242 inputs through splits on the *correlated* Magpie statistics of the co-occurring elements, which is the
 243 mechanistic reason they extrapolate substantially better.

244 Within the LOO test sets, the oxide/non-oxide contrast is degenerate for O (by construction,
 245 every O-LOO test entry contains oxygen, so all 24,572 are flagged as oxides and the non-oxide
 246 count is zero); this is recorded as an observation rather than a comparison. For Fe, the test set
 247 splits into 2,656 oxides and 1,409 non-oxides; for Li, into 5,581 oxides and 1,649 non-oxides. The
 248 per-subgroup MAEs within these splits (not tabulated for space) follow the same family ordering
 249 as the overall LOO result.

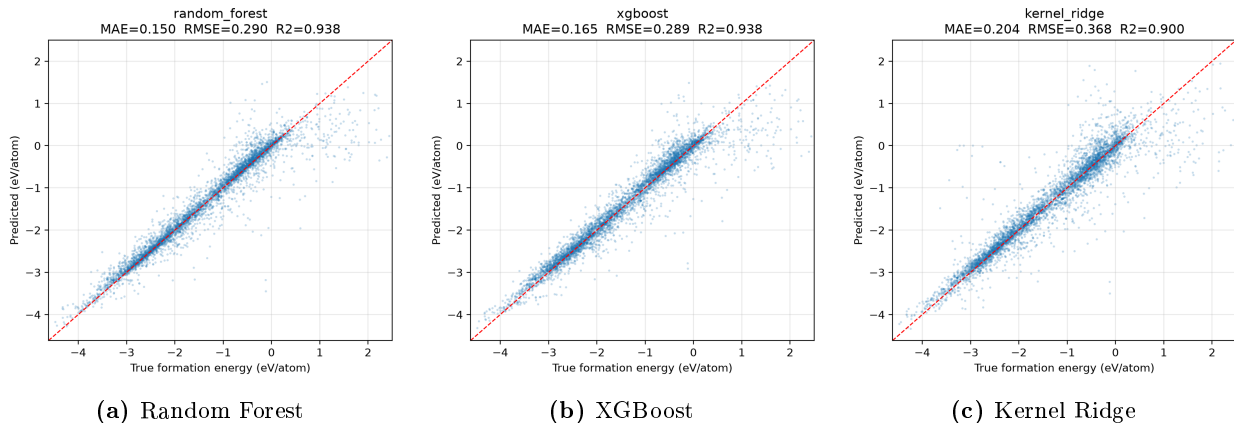


Figure 5. Predicted versus true formation energy on the random test set ($n = 5,000$). The two tree ensembles are tightly clustered about the $y = x$ line; KRR shows a modest upward fanning for strongly stable compounds.

Table 3. Leave-one-element-out performance. MAE and RMSE in eV/atom. Degradation is the LOO MAE divided by the random-split MAE. Hyperparameters are held identical to the random-split experiment so training-data composition is the only varying factor.

Element	Model	LOO MAE	LOO RMSE	LOO R^2	Degradation
O	Random Forest	0.5296	0.6663	0.269	3.53×
	XGBoost	0.5104	0.6141	0.379	3.09×
	Kernel Ridge	1.0014	1.2884	-1.733	4.90×
Fe	Random Forest	0.1407	0.2293	0.950	0.94×
	XGBoost	0.1310	0.2140	0.957	0.79×
	Kernel Ridge	0.3092	0.4971	0.766	1.51×
Li	Random Forest	0.1795	0.2744	0.904	1.20×
	XGBoost	0.1650	0.2651	0.910	1.00×
	Kernel Ridge	0.4802	0.8049	0.171	2.35×

3.7 Summary of key findings

Table 4 consolidates the headline metrics. Three conclusions stand out: (i) CPU-class composition-only baselines reach ~ 0.15 eV/atom, the established “good model” threshold for this benchmark, with Random Forest the best by MAE; (ii) electronegativity statistics—specifically the average-deviation of electronegativity—are the dominant and consistent physical driver across all three model families; and (iii) under element-wise extrapolation, tree ensembles degrade gracefully and the RBF-kernel regressor fails, yielding a clear model-selection rule.

4 Discussion

4.1 Model comparison and the role of trees

The near-tie of Random Forest and XGBoost at the top of the random-split leaderboard ($R^2 \approx 0.938$, MAE 0.150 versus 0.165 eV/atom) reproduces a now-familiar pattern in Matbench-style benchmarks [26, 27]: gradient-boosted trees and bagged forests are difficult to separate on tabular materials descriptors, and both comfortably exceed classical kernel methods on the same features. The slight MAE advantage of RF and the slight RMSE advantage of XGB reflect the bagging-versus-boosting trade-off—bagging reduces variance in the dense central regime, while boosting reduces bias in the tails. KRR’s deficit (MAE 0.204, $R^2 = 0.900$) is partly attributable to its 10,000-sample training subset (imposed by the $\mathcal{O}(n^2)$ RBF-kernel memory cost), but the same policy was applied

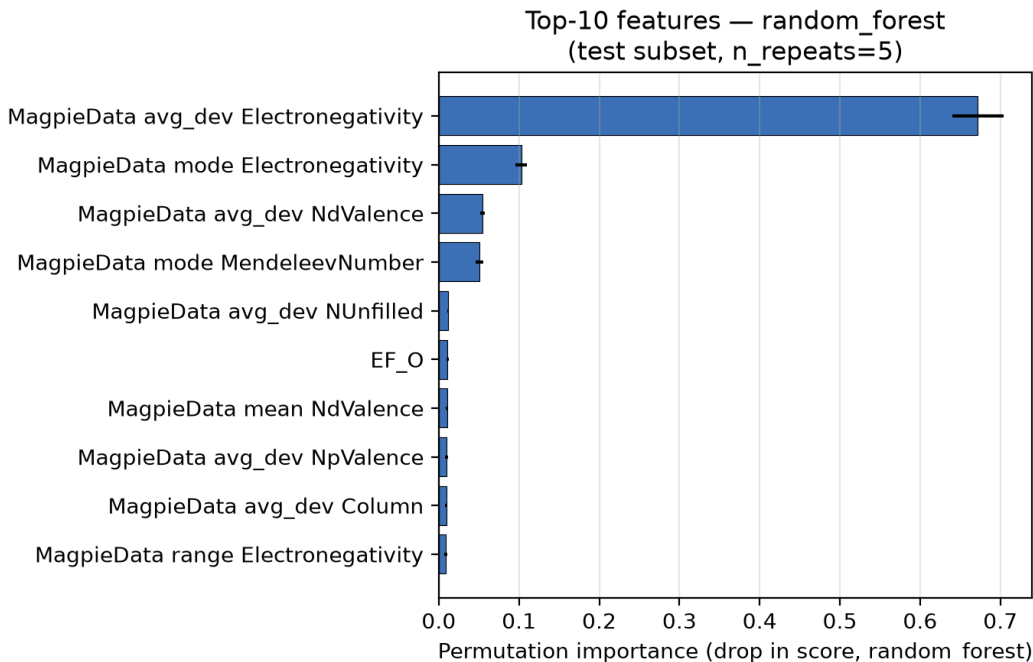


Figure 6. Permutation importance (top-10) for the best model (Random Forest), computed with $n_{\text{repeats}} = 5$ on a 1,000-sample test subset. **avg_dev Electronegativity** dominates by an order of magnitude; the remaining features are electronegativity, valence, and Mendeleev-number statistics.

uniformly and the random-split KRR baseline used the same subset, so the relative ordering is fair. For interpolation, all three are accurate enough to serve as screening filters; the practically interesting question is what happens at and beyond the training boundary.

4.2 Contextualising against universal potentials

Universal graph-neural-network potentials such as M3GNet [21] and CHGNet [22]—the latter trained on the MPtrj relaxation-trajectory subset of the Materials Project—and related equivariant architectures [16–20] now approach DFT-level accuracy on formation and energy predictions because they consume the full relaxed structure. Our composition-only baselines do not compete with these models on raw interpolation accuracy, nor are they intended to. Their value is threefold. First, as a *reference floor*: a universal potential that does not beat a 0.15 eV/atom composition Random Forest on E_f is not adding decision-relevant value, since the composition baseline costs seconds of CPU and no structural relaxation. Second, as an *interpretability probe*: the explicit feature ranking (electronegativity statistics dominate) is a scientific statement that opaque end-to-end potentials cannot make directly, complementing descriptor-discovery approaches such as the LASSO-based ℓ_0 selection of Ghiringhelli *et al.* [11]. Third, and most importantly here, as a *controlled extrapolation probe*: because the feature space is fixed and explicit, the LOO experiment isolates the effect of training-distribution shift without confounding by architectural differences—something that is far harder to do fairly across heterogeneous universal potentials.

4.3 Physical interpretation of the dominant features

That the average absolute deviation of electronegativity (**avg_dev Electronegativity**) is the rank-1 feature for all three models is physically expected and quantitatively striking. Formation energy is, to leading order, a measure of how favourably the constituent elements combine, and electronegativity contrast governs the ionic versus covalent character and the strength of the resulting bonds. A large **avg_dev** signals a composition of chemically disparate elements (e.g. a strongly electropos-

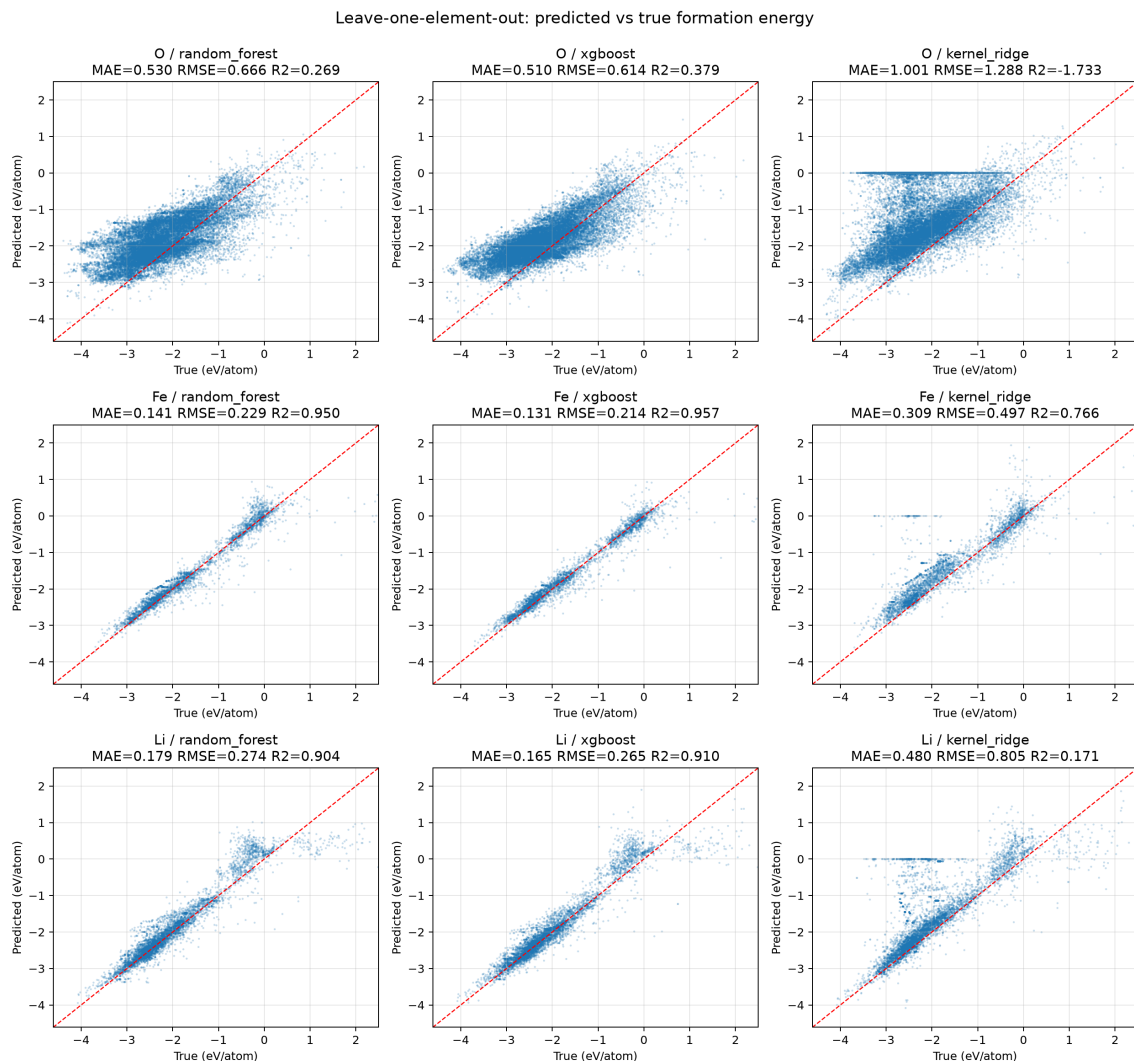


Figure 7. Predicted versus true formation energy for the leave-one-element-out experiment (3×3 grid: rows = O, Fe, Li; columns = RF, XGB, KRR). O row: all models scatter widely (trees) or collapse (KRR). Fe row: tree models remain tightly clustered; KRR fans out. Li row: trees retain skill; KRR is near-degenerate.

itive metal paired with a strongly electronegative chalcogen or halogen), which typically yields a strongly negative E_f ; a small `avg_dev` signals a chemically homogeneous composition (elemental or near-alloy), with E_f near zero. The secondary appearance of valence (`NdValence`, `NpValence`), Mendeleev-number, and column statistics is consistent with these descriptors encoding complementary information about bonding capacity and preferred coordination. The space-group-number statistics that surface for XGB and in the univariate ranking act as a weak structural proxy that composition alone would not normally carry—here they are inherited from the Materials Project’s structure-derived metadata attached to each composition, and their presence is a reminder that even “composition-only” pipelines can inadvertently smuggle in structural information through auxiliary fields.

4.4 Why tree ensembles extrapolate and RBF-kernel methods do not

The LOO results admit a clean mechanistic reading. The RBF kernel measures similarity in the (standardized) feature space; a test composition containing the held-out element E necessarily has a nonzero coordinate in the E -fraction column that is exactly zero for every training point. No matter how the other features align, the Euclidean (and hence RBF) distance from the training cloud is bounded below by a positive quantity, so the kernel weights collapse and the prediction reverts to a

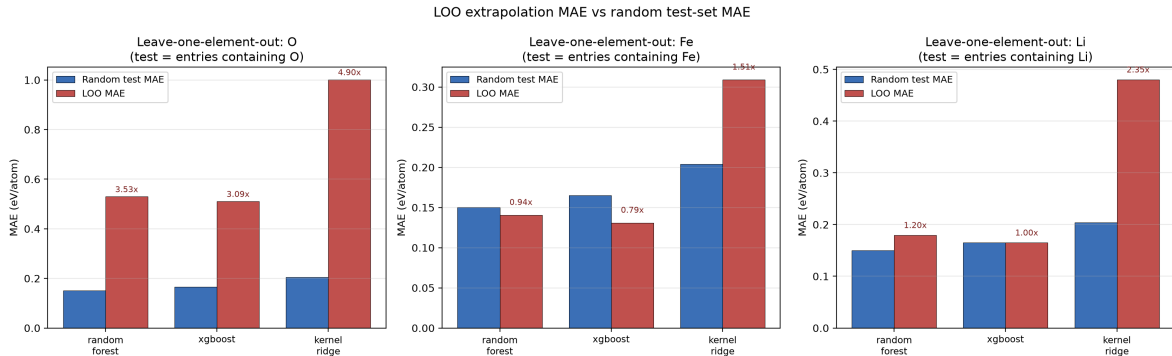


Figure 8. Random-split MAE versus leave-one-element-out MAE, one panel per element, with degradation factors annotated. The bars make the family-level pattern visible: tree ensembles degrade mildly (or improve) for rare elements and degrade only moderately for the most frequent element, whereas the RBF-kernel regressor degrades severely and consistently.

Table 4. Summary of key findings.

Finding	Quantitative evidence
Composition-only CPU baselines reach the Matbench “good” band	RF MAE 0.150, XGB 0.165, KRR 0.204 eV/atom; RF/XGB $R^2 \approx 0.938$
Dominant, consistent feature driver	<code>avg_dev Electronegativity</code> is rank-1 for RF, XGB, and KRR (imp. 0.67/0.48/0.58)
Oxides predicted better than non-oxides	Oxide MAE ~ 0.14 – 0.18 vs non-oxide ~ 0.16 – 0.22 eV/atom
O is the hardest element to leave out	3.1 – $4.9\times$ degradation; KRR $R^2 = -1.73$
Rare Fe barely affects trees	RF/XGB improve ($0.94\times$, $0.79\times$); $R^2 \approx 0.95$
RBF-KRR extrapolates worst in every case	Negative R^2 for O and Li; up to $4.9\times$ degradation
Model-selection rule	Tree ensembles strongly preferred over RBF-kernel methods for extrapolation to unseen species

global or near-global training mean. This is precisely the failure mode we observe: KRR’s R^2 turns negative (worse than predicting the mean) for O and collapses to near-zero for Li, while it degrades $1.5\times$ even for the rare Fe. Tree ensembles suffer no such geometric penalty. A Random Forest or XGBoost split on a Magpie statistic of the *co-occurring* elements can still route an O- or Li-bearing composition to a leaf populated by chemically analogous training compounds (other chalcogenides or other alkali-metal compounds), even though the specific element is unseen. This is why the trees retain skill on Fe and Li (analogous chemistry is present) and degrade only moderately on O (the most common element, whose removal eliminates the largest single chemical class). The result is a clear, methodologically relevant model-selection signal: for tasks requiring extrapolation to unseen chemical species—the defining regime of materials discovery [30]—tree ensembles are strongly preferred over RBF-kernel regressors. This complements earlier extrapolation studies that focused on superconductor critical temperatures [30] and on multi-fidelity property prediction [34], and it is consistent with the broader observation that kernel methods, while powerful interpolators, are brittle out-of-distribution.

4.5 The frequency–extrapolation coupling

A subtle but important consequence of the LOO design is that element frequency and extrapolation difficulty are inversely coupled in a specific sense: removing the *most frequent* element is the hardest

case, not because the element is common in the test set, but because its removal excises the largest and most chemically distinctive block of the training distribution. Oxygen is a polar case: roughly half of all inorganic phases in the Materials Project are oxides, and oxides have a characteristic formation-energy signature, so an O-free training set is missing an entire chemical class. Conversely, removing a rare element such as Fe removes little training signal because Fe-bearing compositions are well-represented by chemically analogous non-Fe compounds. This implies a practical heuristic for prospective data-collection: when a target application stresses a particular element or chemical family, the marginal value of additional training data *containing* that element is high precisely when that family is common, because its absence reshapes the training distribution most. The frequency-extrapolation coupling also cautions against equating “rare element” with “hard extrapolation”: the relevant quantity is the chemical redundancy of the training set once the element is removed, not the test-set count.

4.6 Limitations

Several limitations bound the scope of these conclusions. First, composition-only features cannot capture structural effects—polymorphs, coordination, and packing are invisible to the descriptor, so the residual error (~ 0.15 eV/atom even for the best model) is a lower bound on what structure-informed methods can recover. Second, the LOO experiment is limited to three elements (O, Fe, Li); while these span the frequency spectrum, a broader LOO sweep would strengthen the frequency-extrapolation coupling claim. Third, the source artifact lacks `energy_above_hull`, so the stable-versus-unstable subgroup uses a formation-energy proxy ($E_f < 0$) rather than strict convex-hull stability; this proxy is tautological with the target and the corresponding “unstable” error is inflated. Fourth, the PCA on the body block was fit globally (minor unsupervised leakage into the test set), and KRR’s standardization was fit on its training subset only; tree models are scale- and leakage-insensitive here, so the leaderboard ranking is unaffected, but the KRR numbers should be read as mildly optimistic. Fifth, KRR’s 10,000-sample training subset is not directly comparable in size to the 40,000-sample tree training, although the same policy applied to every split makes the *relative* LOO comparison valid. Finally, the Materials Project is known to be biased toward certain chemistries and structure types [35], so extrapolation behaviour on this dataset may not generalize to other domains.

4.7 Future directions

Three extensions follow naturally. (i) Structure-informed composition hybrids—e.g. augmenting composition features with cheap structural proxies (space-group statistics, density, packing fraction)—could close part of the residual gap without sacrificing the CPU-only deployment model. (ii) A systematic LOO sweep over many elements, weighted by frequency, would let us fit a quantitative frequency-extrapolation curve and test whether the tree-kernel gap is uniform or element-dependent. (iii) Active-learning loops that explicitly target the chemically under-represented regions identified by the LOO protocol could improve extrapolation performance where it matters most for discovery, in the spirit of Bayesian and look-ahead materials-design workflows [7, 14, 34]. More broadly, the LOO protocol itself could be adopted as a standard addition to Matbench-style evaluation [27], since random-split accuracy alone is a weak predictor of the extrapolation behaviour that discovery actually demands.

5 Conclusion

We have established and rigorously benchmarked CPU-class, composition-only machine-learning baselines for DFT formation-energy prediction on a 50,000-entry Materials Project subsample. Random Forest and XGBoost reach ~ 0.15 eV/atom and $R^2 \approx 0.938$ on a random test split—

the “good model” band for the `matbench_mp_e_form` benchmark—with Random Forest best by MAE. Permutation importance consistently identifies the average-deviation of electronegativity as the dominant and physically interpretable predictor across all three model families. A leave-one-element-out protocol for O, Fe, and Li reveals a sharp, model-family-dependent extrapolation pattern: tree ensembles degrade only mildly for rare elements and moderately for the most frequent element, whereas an RBF-kernel regressor collapses—its R^2 turns negative for oxygen and collapses to near-zero for lithium. The clear practical implication is that, for materials-discovery tasks that inherently require extrapolation to unseen chemical species, tree ensembles on composition features are strongly preferred over RBF-kernel methods, and they constitute an accurate, interpretable, and extrapolation-robust CPU baseline against which universal structure-informed potentials should be measured.

Data and code availability

All analysis scripts, split indices, model checkpoints, and result JSON files are persisted in the session directory and are fully reproducible from the resumable `workflow/` scripts under fixed `random_state=42`. The dataset is the public `matbench_mp_e_form` artifact [27, 32]. Feature generation uses `matminer` [32] and `pymatgen`; atomistic-simulation tooling such as the Atomic Simulation Environment [36] underlies the DFT reference data; models use scikit-learn and XGBoost [31].

References

- [1] Anubhav Jain, Shyue Ping Ong, Geoffroy Hautier, Wei Chen, William Davidson Richards, Stephen Dacek, Shreyas Cholia, Dan Gunter, David Skinner, Gerbrand Ceder, and Kristin A. Persson. Commentary: The Materials Project: A materials genome approach to accelerating materials innovation. *APL Materials*, 1(1):011002, 2013. doi: 10.1063/1.4812323.
- [2] Scott Kirklin, James E. Saal, Bryce Meredig, Alex Thompson, Jeff W. Doak, Muratahan Aykol, Stephan Rühl, and Chris Wolverton. The Open Quantum Materials Database (OQMD): assessing the accuracy of DFT formation energies. *npj Computational Materials*, 1(1):15010, 2015. doi: 10.1038/npjcompumats.2015.10.
- [3] James E. Saal, Scott Kirklin, Muratahan Aykol, Bryce Meredig, and C. Wolverton. Materials design and discovery with high-throughput density functional theory: The Open Quantum Materials Database (OQMD). *JOM*, 65(11):1501–1509, 2013. doi: 10.1007/s11837-013-0755-4.
- [4] Stefano Curtarolo, Wahyu Setyawan, Gus L. W. Hart, Michal Jahnatek, Roman V. Chepulskii, Richard H. Taylor, Shidong Wang, Junkai Xue, Kesong Yang, Ohad Levy, Michael J. Mehl, et al. AFLOW: An automatic framework for high-throughput materials discovery. *Computational Materials Science*, 58:218–226, 2012. doi: 10.1016/j.commatsci.2012.02.005.
- [5] Christopher J. Bartel, Christopher Sutton, Bryan R. Goldsmith, Runhai Ouyang, Charles B. Musgrave, Luca M. Ghiringhelli, and Matthias Scheffler. New tolerance factor to predict the stability of perovskite oxides and halides. *Science Advances*, 5(2):eaav0693, 2019. doi: 10.1126/sciadv.aav0693.
- [6] Gustavo R. Schleder, Antonio C. M. Padilha, C. M. Acosta, Marcio Costa, and Adalberto Fazzio. From DFT to machine learning: recent approaches to materials science – a review. *Journal of Physics: Materials*, 2(3):032001, 2019. doi: 10.1088/2515-7639/ab084b.
- [7] J. E. Gubernatis and Turab Lookman. Machine learning in materials design and discovery: Examples from the present and suggestions for the future. *Physical Review Materials*, 2(12):120301, 2018. doi: 10.1103/PhysRevMaterials.2.120301.

- [8] G. Kresse and J. Furthmüller. Efficiency of ab-initio total energy calculations for metals and semiconductors using a plane-wave basis set. *Computational Materials Science*, 6(1):15–50, 1996. doi: 10.1016/0927-0256(96)00008-0.
- [9] Shyue Ping Ong, Shreyas Cholia, Anubhav Jain, Miriam Brafman, Dan Gunter, Gerbrand Ceder, and Kristin A. Persson. The Materials Application Programming Interface (API): A simple, flexible and efficient API for materials data based on REpresentational State Transfer (REST) principles. *Computational Materials Science*, 97:209–215, 2015. doi: 10.1016/j.commatsci.2014.10.037.
- [10] Kamal Choudhary, Kevin F. Garrity, Andrew C. E. Reid, Brian DeCost, Adam J. Bicchi, Angela R. Hight Walker, Zachary Trautt, Jason Hattrick-Simpers, et al. The joint automated repository for various integrated simulations (JARVIS) for data-driven materials design. *npj Computational Materials*, 6(1), 2020. doi: 10.1038/s41524-020-00440-1.
- [11] Luca M. Ghiringhelli, Jan Vybiral, Sergey V. Levchenko, Claudia Draxl, and Matthias Scheffler. Big data of materials science: Critical role of the descriptor. *Physical Review Letters*, 114(10):105503, 2015. doi: 10.1103/PhysRevLett.114.105503.
- [12] Logan Ward, Ankit Agrawal, Alok Choudhary, and Christopher Wolverton. A general-purpose machine learning framework for predicting properties of inorganic materials. *npj Computational Materials*, 2(1), 2016. doi: 10.1038/npjcompumats.2016.28.
- [13] Atsuto Seko, Hiroyuki Hayashi, Keita Nakayama, Akira Takahashi, and Isao Tanaka. Representation of compounds for machine-learning prediction of physical properties. *Physical Review B*, 95(14):144110, 2017. doi: 10.1103/PhysRevB.95.144110.
- [14] Prasanna V. Balachandran. Machine learning guided design of functional materials with targeted properties. *Computational Materials Science*, 164:82–90, 2019. doi: 10.1016/j.commatsci.2019.03.057.
- [15] Logan Ward, Ruoqian Liu, Amar Krishna, Vinay I. Hegde, Ankit Agrawal, Alok Choudhary, and Chris Wolverton. Including crystal structure attributes in machine learning models of formation energies via Voronoi tessellations. *Physical Review B*, 96(2):024104, 2017. doi: 10.1103/PhysRevB.96.024104.
- [16] Tian Xie and Jeffrey C. Grossman. Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. *Physical Review Letters*, 120(14):145301, 2018. doi: 10.1103/PhysRevLett.120.145301.
- [17] Kamal Choudhary and Brian DeCost. Atomistic line graph neural network for improving materials property predictions. *npj Computational Materials*, 7(1), 2021. doi: 10.1038/s41524-021-00650-1.
- [18] Simon Batzner, Albert Musaelian, Lixin Sun, Mario Geiger, Jonathan P. Mailoa, Mor Kornbluth, Nicola Molinari, Tess E. Smidt, and Boris Kozinsky. $E(3)$ -equivariant graph neural networks for data-efficient and accurate interatomic potentials. *Nature Communications*, 13(1):2453, 2022. doi: 10.1038/s41467-022-29939-5.
- [19] Kristof T. Schütt, Max Ekrias Saucedo, Pieter-Jan Kindermans, Alexandre Tkatchenko, and Klaus-Robert Müller. SchNet – a deep learning architecture for molecules and materials. *The Journal of Chemical Physics*, 148(24):241722, 2018. doi: 10.1063/1.5019779.
- [20] Ilyes Batatia, Simon Batzner, David Duesberg, Albert Musaelian, Philippe Schwaller, Théo Stärk, Janosh Nigam, and Boris Kozinsky. MACE: Higher order equivariant message passing neural networks for fast and accurate force fields. In *Advances in Neural Information Processing Systems*, 2022. doi: 10.52202/068431-0830.

- [21] Chi Chen and Shyue Ping Ong. A universal graph deep learning interatomic potential for the periodic table. *Nature Computational Science*, 2(11):718–728, 2022. doi: 10.1038/s43588-022-00349-3.
- [22] Bowen Deng, Peichen Zhong, KyuJung Jun, Janosh Riebesell, Kevin Han, Christopher J. Bartel, and Gerbrand Ceder. CHGNet as a pretrained universal neural network potential for charge-informed atomistic modelling. *Nature Machine Intelligence*, 5(9):1031–1041, 2023. doi: 10.1038/s42256-023-00716-3.
- [23] Jörg Behler and Michele Parrinello. Generalized neural-network representation of high-dimensional potential-energy surfaces. *Physical Review Letters*, 98(14):146401, 2007. doi: 10.1103/PhysRevLett.98.146401.
- [24] Matthias Rupp, Alexandre Tkatchenko, Klaus-Robert Müller, and O. Anatole von Lilienfeld. Fast and accurate modeling of molecular atomization energies with machine learning. *Physical Review Letters*, 108(5):058301, 2012. doi: 10.1103/PhysRevLett.108.058301.
- [25] Rhys E. A. Goodall and Alpha A. Lee. Predicting materials properties without crystal structure: deep representation learning from stoichiometry. *Nature Communications*, 11(1):6280, 2020. doi: 10.1038/s41467-020-19964-7.
- [26] Pierre-Paul De Breuck, Geoffroy Hautier, and Gian-Marco Rignanese. Materials property prediction for limited datasets enabled by feature selection and joint learning with MODNet. *npj Computational Materials*, 7(1), 2021. doi: 10.1038/s41524-021-00552-2.
- [27] Alexander Dunn, Qi Wang, Alex Ganose, Daniel Dopp, and Anubhav Jain. Benchmarking materials property prediction methods: the Matbench test set and Automatminer reference algorithm. *npj Computational Materials*, 6(1), 2020. doi: 10.1038/s41524-020-00406-3.
- [28] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001. doi: 10.1023/A:1010933404324.
- [29] André Altmann, Laura Toloşi, Oliver Sander, and Thomas Lengauer. Permutation importance: a corrected feature importance measure. *Bioinformatics*, 26(10):1340–1347, 2010. doi: 10.1093/bioinformatics/btq134.
- [30] Bryce Meredig, Erin Antono, Carena Church, Maxwell Hutchinson, Julia Ling, Sean Paradiso, Ben Blaiszik, Ian Foster, and Jason Hattrick-Simpers. Can machine learning identify the next high-temperature superconductor? Examining extrapolation performance for materials discovery. *Molecular Systems Design & Engineering*, 3(5):819–825, 2018. doi: 10.1039/c8me00012c.
- [31] Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794, 2016. doi: 10.1145/2939672.2939785.
- [32] Logan Ward, Alexander Dunn, Alireza Faghaninia, Nils E. R. Zimmermann, Saurabh Bajaj, Qi Wang, Joseph Montoya, Jiming Chen, Kyle Bystrom, Maxwell Dylla, Kamal Choudhary, Ankit Agrawal, Shin Takeda, Bowen Deng, Anthony Zheng, Tania Ahsan, Raymond Gee, Fanglin Ouyang, Joshua Sparks, Nessa Trevino, Paul Mikulskis, Surya Alexander, Julia Ling, Peter Ward, Scott O’Keeffe, Abhijit Mishra, Kristin Persson, Gerbrand Ceder, et al. matminer: An open source toolkit for materials data mining. *Computational Materials Science*, 152:60–69, 2018. doi: 10.1016/j.commatsci.2018.05.018.
- [33] Atsushi Togo, Mitsuhiro Fujita, and Isao Tanaka. spglib: a software library for crystal symmetry search. *Science and Technology of Advanced Materials: Methods*, 4(1):2384822, 2024. doi: 10.1080/27660400.2024.2384822.

- 501 [34] G. Pilania, J. E. Gubernatis, and T. Lookman. Multi-fidelity machine learning models for
502 accurate bandgap predictions of solids. *Computational Materials Science*, 129:156–163, 2017.
503 doi: 10.1016/j.commatsci.2016.12.004.
- 504 [35] Daniel Persaud, Logan Ward, and Jason Hattrick-Simpers. Reproducibility in materials infor-
505 matics: lessons from ‘a general-purpose machine learning framework for predicting properties
506 of inorganic materials’. *Digital Discovery*, 3(2):281–286, 2024. doi: 10.1039/d3dd00199g.
- 507 [36] Ask Hjorth Larsen, Jens Jørgen Mortensen, Jakob Blomqvist, Ivano E. Castelli, Rune Chris-
508 tensen, Marcin Dułak, Jesper Friis, Michael N. Groves, Bjørk Hammer, Cory Hargus, et al.
509 The atomic simulation environment – a Python library for working with atoms. *Journal of*
510 *Physics: Condensed Matter*, 29(27):273002, 2017. doi: 10.1088/1361-648x/aa680e.