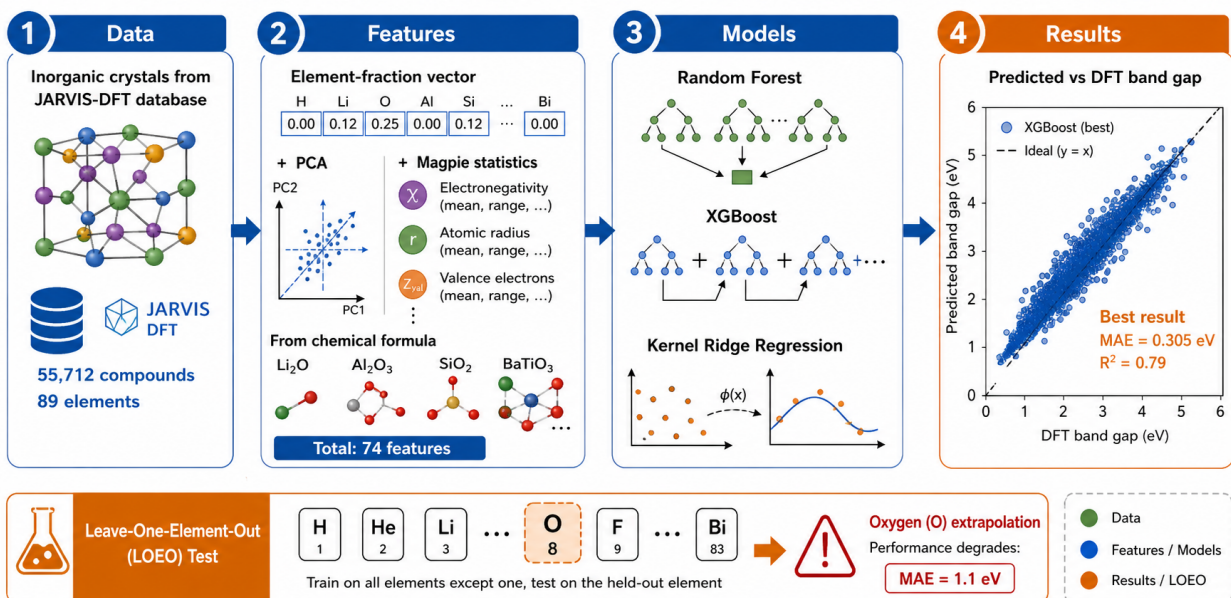


Composition-Based Prediction of DFT Band Gaps for Inorganic Crystals: Random Forest, XGBoost, and Kernel Ridge Regression Benchmark with Leave-One-Element Sensitivity Analysis on 55710 JARVIS-DFT Compounds

LitLit AI¹

¹contact@litlit.ai



Graphical abstract. Composition-based band-gap prediction pipeline. DFT-derived band gaps for 55710 inorganic crystals from JARVIS-DFT are modeled using only chemical composition. Element-fraction vectors are reduced with PCA and concatenated with Magpie elemental-property statistics to form a 74-dimensional feature set. Three regressors (Random Forest, XGBoost, Kernel Ridge Regression) are benchmarked on a leakage-safe random split and on leave-one-element (O, Ti, Cu) extrapolation tests. Random Forest attains the best mean absolute error (0.305 eV), while leave-one-element tests—particularly for oxygen—reveal severe degradation driven by elemental extrapolation and label-distribution shift.

Abstract

Predicting the electronic band gap of inorganic solids directly from chemical composition accelerates high-throughput materials screening by circumventing the cost of density-functional theory (DFT) calculations and the requirement of a known crystal structure. Here we train three composition-only machine-learning models—Random Forest (RF), XGBoost, and Kernel

Ridge Regression (KRR)—on 55710 inorganic compounds from the JARVIS-DFT database, using OPTB88vdW band gaps as labels and a 74-dimensional feature vector comprising principal components of the element-fraction vector plus Magpie elemental-property statistics (electronegativity, atomic radius, and valence-electron count). On a leakage-safe random 80/10/10 split (`random_state=42`), RF achieved the lowest mean absolute error (MAE = 0.305 eV, $R^2 = 0.788$), followed by XGBoost (MAE = 0.342 eV, $R^2 = 0.792$) and KRR (MAE = 0.528 eV, $R^2 = 0.597$). Decomposing performance by metallicity (band gap < 0.05 eV) showed that all models were substantially more accurate on the metal subset (RF metal MAE = 0.165 eV) than on non-metals (RF non-metal MAE = 0.621 eV), reflecting the abundance of zero-gap metals in the OPTB88vdW distribution. Permutation-importance analysis identified the maximum compound electronegativity, valence-electron statistics, and mean atomic radius as the most predictive composition features. Leave-one-element-out experiments for oxygen, titanium, and copper exposed a sharp dependence of generalization on the chemical identity of the held-out element: oxygen exclusion inflated the MAE by roughly 4 $\times$ (to 1.20 eV for RF) and produced negative R^2 for copper, owing both to genuine elemental extrapolation and to a label-distribution shift between train and test. These results establish a transparent, structure-free band-gap baseline for JARVIS-DFT and quantify the boundary between interpolation and extrapolation that governs the reliability of composition-only property models.

Keywords: band gap prediction, materials informatics, composition-based machine learning, JARVIS-DFT, Magpie descriptors, leave-one-element-out, Random Forest, XGBoost.

1 Introduction

The electronic band gap is a defining descriptor of a crystalline solid: it distinguishes metals from semiconductors and insulators, sets the optical absorption threshold, and governs suitability for photovoltaic, thermoelectric, and transparent-conductor applications. Because the band gap is a ground-state-derived yet excited-state-sensitive property, its ab-initio computation requires, in principle, methods such as hybrid functionals [1] or many-body perturbation theory, which correct the well-known band-gap underestimation of semi-local density-functional theory [2, 3]. While semi-local DFT with the OPTB88vdW functional is cheap enough to screen tens of thousands of compounds, it systematically reports zero or near-zero gaps for many semiconducting oxides [4], which complicates both the label distribution and the interpretation of model accuracy. Data-driven alternatives that learn the mapping from a chemical or structural representation to a DFT-computed band gap therefore occupy a central position in modern materials informatics [5–7].

A substantial body of work has demonstrated that machine-learning (ML) models can reproduce DFT band gaps with sub-eV errors. Early efforts combined DFT features with kernel or tree-based regression on curated compound sets [8, 9], while more recent benchmarks have evaluated composition-only, descriptor-based, and graph-structured representations side by side [10–12]. Structure-aware deep-learning architectures—including crystal graph convolutional neural networks [13], SchNet [14], and the atomistic line-graph neural network ALIGNN [15, 16]—have pushed accuracy further by embedding the full atomic connectivity into the model. Yet composition-only models remain scientifically and practically attractive: they require no resolved crystal structure, generalize naturally to hypothetical or not-yet-synthesized stoichiometries, and are computationally trivial at inference time. Composition-only deep representation learners [17] and fragment-based descriptors [18] have shown that meaningful band-gap prediction is achievable from the chemical formula alone, and composition-based energy-material descriptors continue to be refined for feature efficiency [19].

The feature representations underpinning these models are themselves the product of a mature descriptor-engineering literature. The Magpie descriptor set [20]—which aggregates elemental properties (e.g. electronegativity, atomic radius, valence-electron count) into composition-level statistics—remains a workhorse for structure-free property prediction, and its open implementation

in matminer [21] together with pymatgen [22] has made reproducible composition-featurization accessible to the community. Large, openly available DFT databases—JARVIS-DFT [4], the Materials Project [23], the Open Quantum Materials Database [24, 25], and the NOMAD repository [26, 27]—supply the labeled data at the scale required for supervised learning, and JARVIS-DFT in particular provides 55000+ three-dimensional inorganic compounds with OPTB88vdW band gaps, making it an ideal substrate for a composition-only benchmark.

Despite this progress, two methodological issues receive less attention than they deserve in the band-gap-prediction literature. First, the field’s reliance on random train/test splits obscures the distinction between interpolation and extrapolation: a model that interpolates within a densely sampled chemical space may still fail catastrophically when asked to predict a compound centered on an element that is systematically absent from training. Leave-one-element-out (LOO) protocols expose this failure mode directly, and have been advocated as a stricter proxy for the generalization demands of genuine materials discovery. Second, the heavy class imbalance induced by the OPTB88vdW functional—roughly 69% of JARVIS-DFT records are metallic (band gap < 0.05 eV)—means that aggregate metrics such as the coefficient of determination R^2 are dominated by the abundant zero-gap metals, while the practically important non-metal regime is under-weighted. Decomposing performance by metallicity is therefore essential for an honest appraisal.

This paper presents a transparent, fully reproducible composition-only band-gap benchmark on 55710 JARVIS-DFT compounds. The contributions are threefold. (i) We engineer a compact 74-dimensional composition feature vector—PCA-reduced element fractions concatenated with a curated subset of Magpie statistics—and train Random Forest, XGBoost, and Kernel Ridge Regression with documented, fixed hyperparameters, reporting overall and metallicity-stratified MAE, RMSE, and R^2 . (ii) We perform leave-one-element-out sensitivity experiments for three chemically common yet distinct elements—oxygen, titanium, and copper—and diagnose the resulting performance degradation through the lenses of elemental extrapolation and label-distribution shift. (iii) We use permutation importance to identify which composition features drive the predictions, providing physical interpretability that purely black-box models lack. Together, these results quantify the accuracy ceiling and the generalization boundary of composition-only band-gap models on JARVIS-DFT, and they delineate concrete next steps—grouped cross-validation, composition-disjoint test sets, and kernel approximations—for closing the interpolation-to-extrapolation gap.

2 Methods

2.1 Dataset and acquisition

All data were drawn from the JARVIS-DFT `dft_3d` collection (2021 snapshot, `j_dft_3d-8-18-2021`), which contains DFT results for bulk three-dimensional inorganic compounds [4]. The dataset was acquired via the `jarvis-tools` package and, equivalently, the direct figshare download endpoint; the schema and field mapping were read from the `jarvis.db.figshare` source to identify the correct `dft_3d` snapshot. The target variable was the `optb88vdw_bandgap` field (OPTB88vdW semi-local functional, in eV), and the input was the reduced chemical composition (`formula`, e.g. TiO_2). After dropping 11 duplicate material identifiers and 2 records whose compositions failed to parse with pymatgen [22], 55712 records remained. Polymorphism was retained rather than aggregated: 24153 records share a composition with at least one other record (8312 distinct compositions), and 10783 records (3068 compositions) share a composition yet exhibit distinct band-gap values. This polymorph-induced label noise is a genuine modeling challenge for a composition-only representation and was not removed.

The cleaned band-gap distribution is sharply right-skewed and zero-inflated: mean 0.688 eV, median 0.0 eV, standard deviation 1.409 eV, and a maximum of 9.643 eV. Using a metallicity threshold

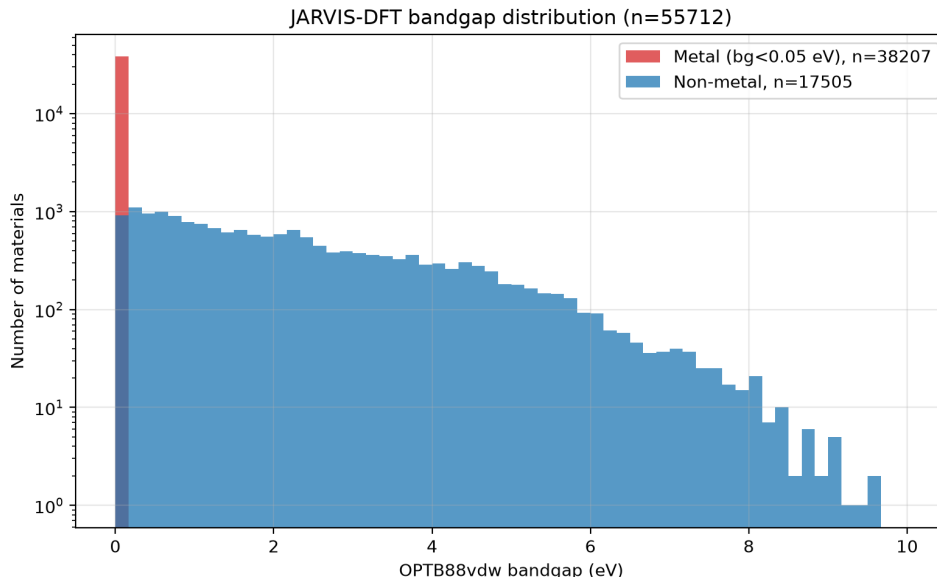


Figure 1. Distribution of OPTB88vdW band gaps across the 55712 cleaned JARVIS-DFT records. The distribution is zero-inflated (median 0 eV) and right-skewed, with 68.6% of records metallic (band gap < 0.05 eV), a direct consequence of the OPTB88vdW semi-local functional’s systematic gap underestimation.

of 0.05 eV (see below), 38207 records (68.6%) are metallic and 17505 (31.4%) are non-metallic. This elevated metal fraction reflects the systematic underestimation of gaps by the OPTB88vdW functional [2, 3], which reports 0 eV for many semiconducting oxides; it is the true label distribution of the dataset and is consistent with published ALIGNN benchmarks on the same field [4, 15]. Figure 1 shows the histogram.

2.2 Feature engineering

The feature design followed the composition-only principle: no crystal-structure information was used. For each record we constructed two blocks of features.

Element-fraction vector. Each composition was parsed with pymatgen [22] into fractional elemental concentrations over the 89 distinct elements present in the dataset, yielding an 89-dimensional sparse vector (96.7% zeros). Because this vector is high-dimensional and dominated by compound-specific element identities, we applied principal component analysis (PCA) to reduce it while retaining at least 95% of the cumulative variance. The element-fraction vectors are very sparse, so reaching 95% variance required 65 of the 89 components (a 27% reduction); the > 95% success criterion was prioritized over the approximate “~50 dimensions” guidance. PCA was fit on the training block only to prevent test-set leakage. Figure 2 shows the cumulative-variance curve.

Magpie elemental-property statistics. We computed a curated subset of the Magpie descriptor scheme [20, 21]: for each compound we evaluated, over the constituent elements weighted by fractional concentration, the mean, maximum, and minimum of three elemental properties—Pauling electronegativity, atomic radius (Å), and valence-electron count (derived from the periodic group under the Magpie convention: groups 1–12 map to the group number, groups 13–18 to group minus 10). This produced a dense 9-dimensional block ($X_{\text{mean}}, X_{\text{max}}, X_{\text{min}}, r_{\text{mean}}, r_{\text{max}}, r_{\text{min}}, v_{\text{mean}}, v_{\text{max}}, v_{\text{min}}$). Missing elemental properties (e.g. electronegativity for He, Ne, Ar; atomic radius for He, Ne, Kr, Xe) were filled with zero rather than dropping the element, preserving all records.

The two blocks were concatenated to a 74-dimensional matrix and standardized with a `StandardScaler`

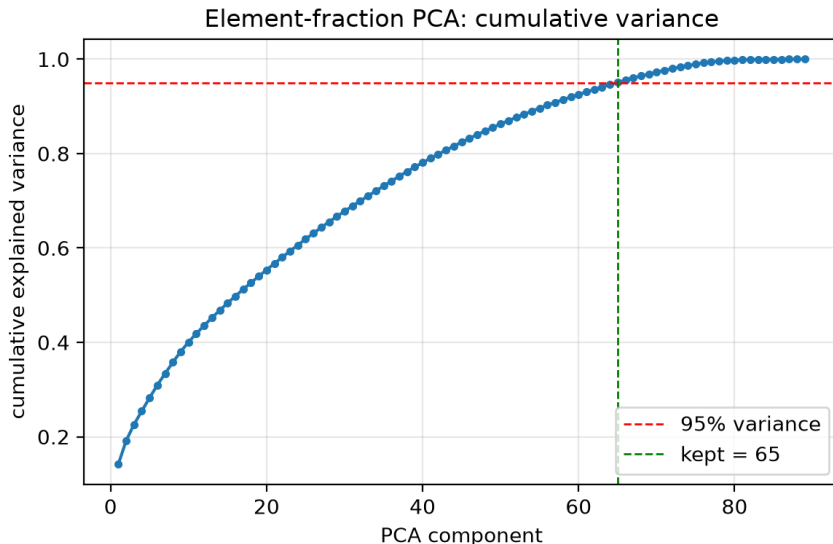


Figure 2. Cumulative explained variance of PCA applied to the 89-dimensional element-fraction vectors. The minimum number of components retaining $\geq 95\%$ variance is 65; the sparsity of the composition vectors (96.7% zeros) requires most of the original dimensions to be preserved.

(mean 0, variance 1) fit on the training block only. The fitted scaler and PCA transformer were persisted for downstream reuse, ensuring that test and leave-one-element data were transformed identically.

2.3 Data splitting

A leakage-safe random 80/10/10 train/validation/test split (`random_state=42`) yielded 44568 training, 5571 validation, and 5571 test records. The band-gap statistics of the three partitions are closely matched (test mean 0.663 eV, metal fraction 69.4%), confirming that the random split preserves the parent distribution. Metallicity was defined as band gap < 0.05 eV; the test set contains 3867 metals and 1704 non-metals.

In addition to the random split, we performed leave-one-element-out (LOO) sensitivity tests for three common yet chemically distinct elements: oxygen (O, present in 14992 compounds, 26.9% of the dataset), titanium (Ti, 2154 compounds, 3.9%), and copper (Cu, 3086 compounds, 5.5%). For each element E , the *test* set comprised all compounds containing E , and the *train* set comprised the remaining compounds (verified to contain no E); each LOO model was retrained from scratch on its own train set, with PCA and scaler refit on that train set only, so that each LOO model operates entirely within its own feature space. The LOO split is therefore a strict extrapolation probe: the model must predict compounds centered on an element it has never seen during training. Figure 3 summarizes the split composition and the LOO element-overview statistics.

2.4 Models

Three regressors were trained with fixed, documented hyperparameters and `random_state=42` throughout.

Random Forest (RF) [28, 29]: `n_estimators=400`, all cores. Tree ensembles are well suited to the heterogeneous, mixed-continuous features of this representation and to the zero-inflated target.

XGBoost [30]: `n_estimators=500`, `max_depth=6`, `learning_rate=0.05`, `random_state=42`, with early stopping (`early_stopping_rounds=50`) monitored on the validation split. Note that in xgboost 3.3.0, `early_stopping_rounds` is a constructor argument rather than a `fit` keyword, and

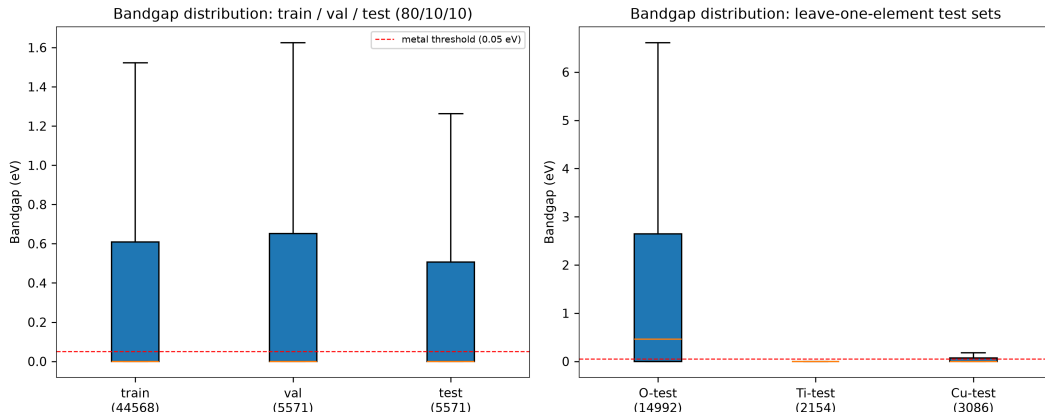


Figure 3. Band-gap and metallicity statistics of the random 80/10/10 split (top) and the leave-one-element-overview (bottom). The random split partitions are closely matched in distribution, whereas the LOO partitions exhibit large train-test distribution shifts (e.g. oxygen: train-no-O is 78% metallic while test-O is 42% metallic).

was set accordingly. On the LOO splits the validation loss kept improving slowly and XGBoost ran the full 500 trees without triggering early stopping; the results are reported per specification.

Kernel Ridge Regression (KRR) [29, 31] with an RBF kernel: `alpha=1.0`. The full training kernel ($N = 44568$, 74 features) would occupy ≈ 15.9 GB in float64, exceeding available memory; we therefore trained KRR on a random subsample of 8000 records (`random_state=42`), a documented strategy, with a Nystroem+ Ridge fallback wired but not triggered. Because scikit-learn 1.9’s `KernelRidge` rejects the string `gamma='scale'` (it accepts only a float or `None`), we computed the scale-equivalent RBF width explicitly, $\gamma = 1/(n_{\text{features}} \cdot \text{Var}(X)) \approx 0.0136$, and passed it as a float. The same 8000-record subsampling policy was applied to each LOO train set. This subsampling is a known limitation (see Discussion).

2.5 Evaluation

We report the mean absolute error (MAE, in eV), root-mean-square error (RMSE, in eV), and coefficient of determination (R^2) on the random test set. Because the OPTB88vdW label distribution is dominated by zero-gap metals, we additionally report MAE decomposed by metallicity (metal vs. non-metal) as the practically informative subset metric. We flag explicitly that R^2 is *meaningless* on the metal subset: metals have band gap ≈ 0 and hence near-zero target variance, so the ratio defining R^2 diverges and takes wildly negative values (e.g. -25510 for RF); MAE is the reliable subset metric, as specified by the analysis plan. Feature importance was estimated by permutation importance [28, 32] on a 2000-record test subsample (`random_state=42`, $n_{\text{repeats}} = 5$, MAE-based scoring) for both RF and XGBoost. All models and transformers were persisted with `joblib` for reproducibility.

3 Results

3.1 Main test-set performance

Table 1 and Figure 4 summarize the performance of the three models on the random test set (5571 compounds). Random Forest attained the lowest overall MAE (0.305 eV) and an R^2 of 0.788, edging out XGBoost (0.342 eV, $R^2 = 0.792$) on error despite XGBoost’s marginally higher R^2 . Kernel Ridge Regression lagged both tree models with MAE = 0.528 eV and $R^2 = 0.597$, consistent with the constraints of the 8000-record subsample and the absence of kernel hyperparameter tuning. The RMSE values (RF 0.643 eV, XGBoost 0.637 eV, KRR 0.888 eV) track the MAE ordering and confirm

Table 1. Performance on the random test set ($n = 5571$). Metal is defined as band gap < 0.05 eV. Metal-subset R^2 is omitted as meaningless (near-zero target variance); see text.

Model	Overall MAE (eV)	Overall RMSE (eV)	Overall R^2	Metal MAE (eV)	Non-metal MAE (eV)	Non-metal R^2
Random Forest	0.305	0.643	0.788	0.165	0.621	0.738
XGBoost	0.342	0.637	0.792	0.199	0.667	0.738
KRR	0.528	0.888	0.597	0.341	0.951	0.485

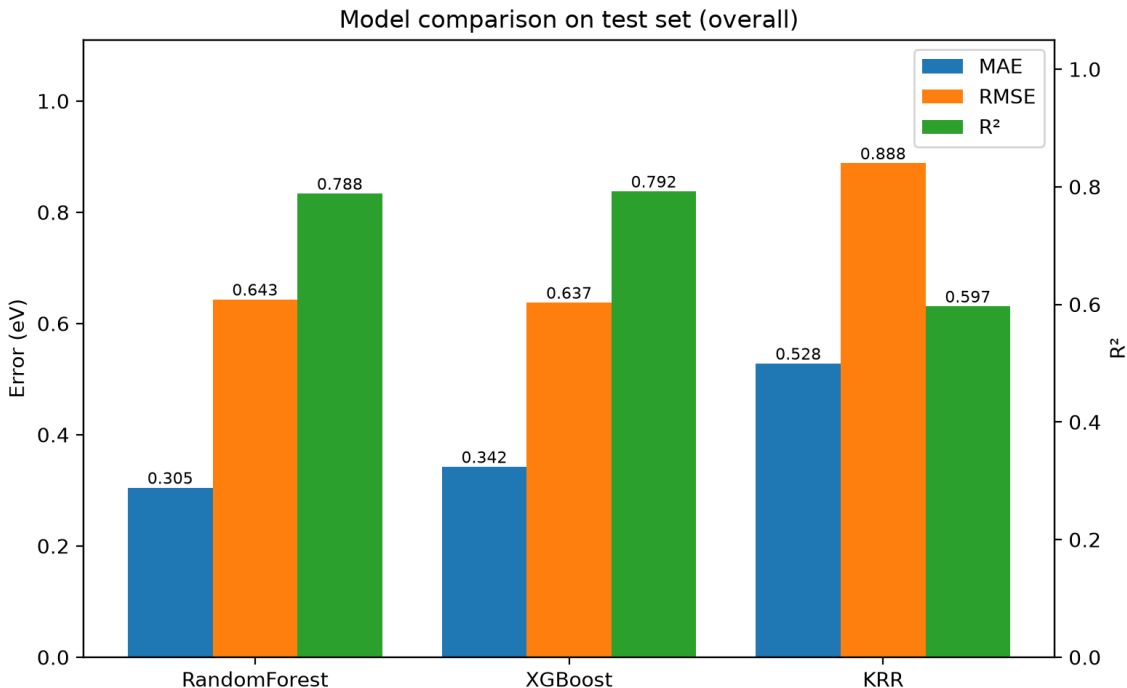


Figure 4. Overall test-set metrics for the three models. Random Forest achieves the lowest mean absolute error, while XGBoost attains a marginally higher coefficient of determination; KRR is the weakest, constrained by subsampling and the absence of hyperparameter tuning.

that the error distributions are heavy-tailed, with a small number of large-gap compounds driving the RMSE above the MAE.

3.2 Metal vs. non-metal decomposition

Decomposing the test set by metallicity reveals a consistent and physically interpretable pattern (Figure 5): all three models are markedly more accurate on the metal subset (band gap < 0.05 eV) than on the non-metal subset. For RF, the metal MAE is 0.165 eV versus a non-metal MAE of 0.621 eV; the same gap appears for XGBoost (0.199 eV vs. 0.667 eV) and KRR (0.341 eV vs. 0.951 eV). This asymmetry has two origins. First, the metal subset is both larger (3867 vs. 1704 test compounds) and concentrated near zero target variance, so models that predict small positive values already achieve low metal MAE. Second, the non-metal subset spans the full 0.05–eV 9.643 range, which is intrinsically harder to regress. On the non-metal subset, where R^2 is meaningful, RF and XGBoost both reach $R^2 \approx 0.738$, while KRR reaches 0.485—indicating that the tree ensembles capture the composition–gap relationship for genuine semiconductors/insulators substantially better than the subsampled RBF kernel does. The metal-subset R^2 values are astronomically negative (e.g. -25510 for RF) for the trivial reason explained in Methods; we therefore report metal MAE as the only reliable subset statistic, in line with the analysis plan.

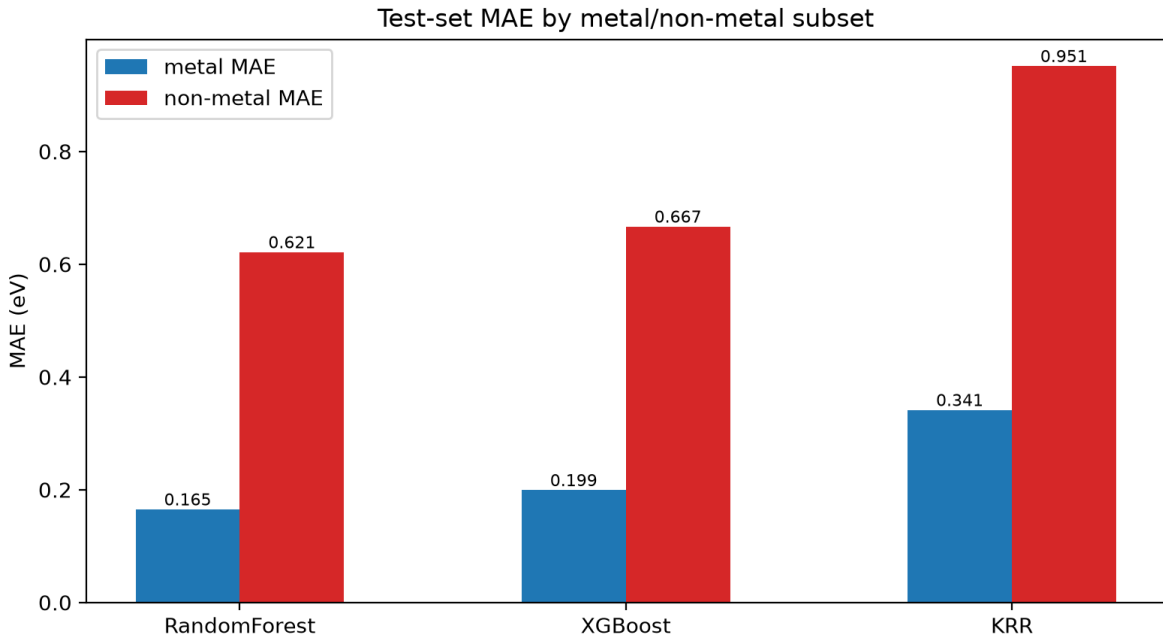


Figure 5. Metallicity-stratified mean absolute error. All models are substantially more accurate on the (abundant, near-zero-gap) metal subset than on the non-metal subset; the gap is largest for KRR. The non-metal regime is the practically important target for composition-based band-gap screening.

3.3 Parity analysis

Figure 6 presents parity (predicted vs. DFT) plots for the three models, individually and overlaid. All models exhibit the characteristic horizontal band of metals predicted near zero along the $y = 0$ ridge, which is responsible for the low metal MAE. The failure modes differ qualitatively: RF predictions compress large gaps toward the mean (visible as a sub-unity slope at high band gap), XGBoost shows a similar but slightly more dispersed high-gap tail, and KRR exhibits the strongest shrinkage toward the mean—an expected artifact of L2-regularized kernel regression on a subsampled, untuned RBF kernel. The overlaid parity plot (Figure 6d) makes the inter-model agreement explicit: where all three agree, predictions lie close to the diagonal; the largest inter-model disagreements occur at high band gap, where KRR systematically under-predicts.

3.4 Permutation-importance features

Figure 7 reports the top-10 permutation-importance features for the best-MAE model (RF), and Table 2 compares the RF and XGBoost rankings. The dominant features are physically transparent: the maximum compound electronegativity $X_{\max}$ is by far the most important feature for RF (importance 0.312 eV of MAE degradation when shuffled), followed by the maximum valence-electron count $v_{\max}$, the mean atomic radius r_{mean} , and the mean valence-electron count v_{mean} . XGBoost’s ranking is led by v_{mean} and X_{mean} , with the leading PCA component of the element-fraction vector (pca_1) entering the top five—reflecting the boosting model’s stronger reliance on the stoichiometric composition signal. Across both models, the Magpie elemental-property statistics dominate the top of the ranking, while the PCA-reduced element-fraction components appear only sporadically (pca_3, pca_6, pca_41 for RF; pca_1, pca_2, pca_3, pca_8 for XGBoost). This confirms that physically interpretable elemental-property statistics carry most of the predictive signal, with the raw stoichiometry providing a complementary but secondary contribution.

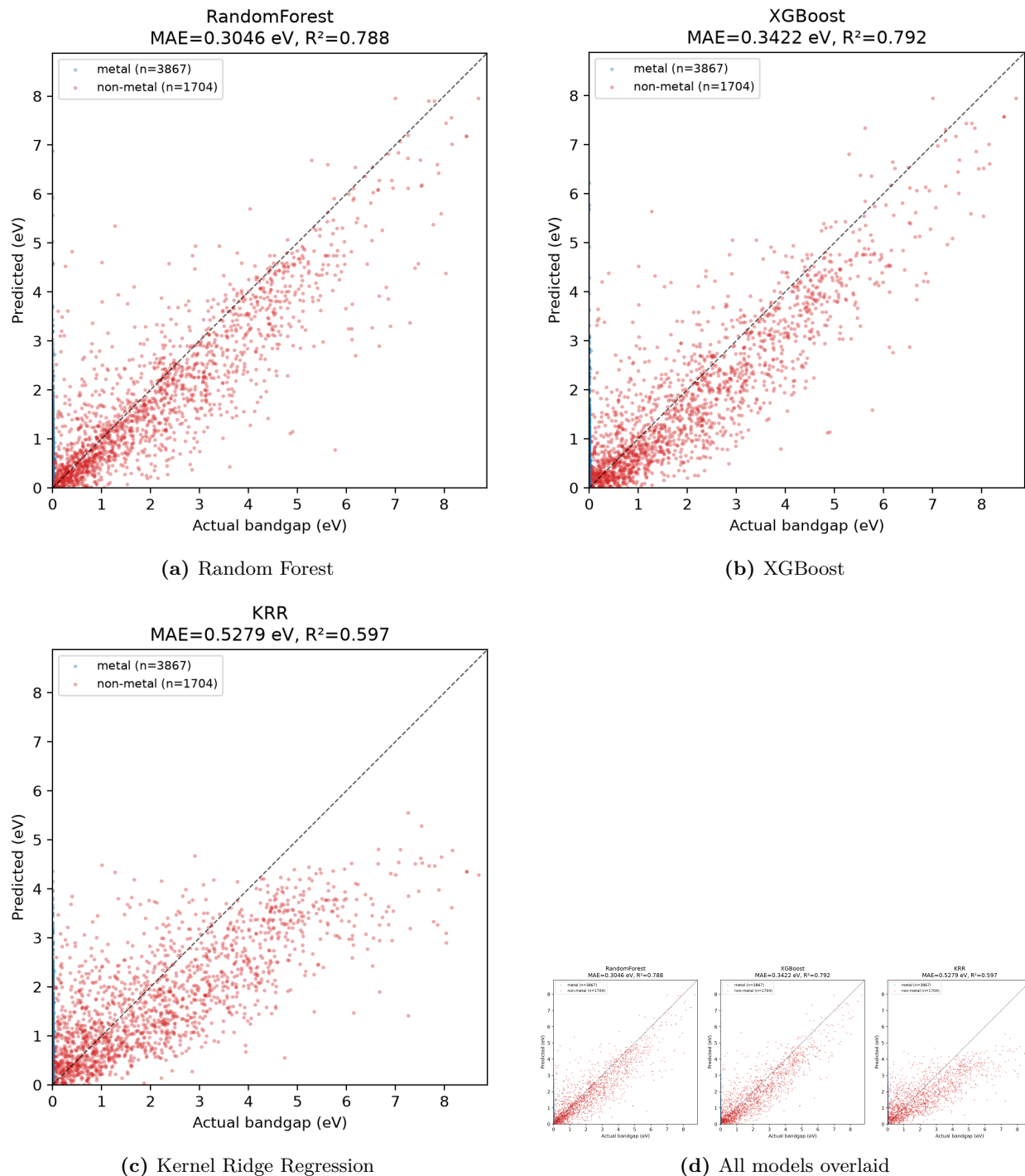


Figure 6. Parity plots of predicted vs. DFT (OPTB88vdW) band gap on the random test set. The horizontal $y \approx 0$ band corresponds to metals; high-gap compounds reveal the systematic shrinkage toward the mean, most pronounced for KRR.

3.5 Leave-one-element-out sensitivity

The leave-one-element tests (Figure 8 and Table 3) expose a sharp, element-dependent degradation relative to the random test set. For oxygen, the MAE inflates by roughly four-fold: RF rises from

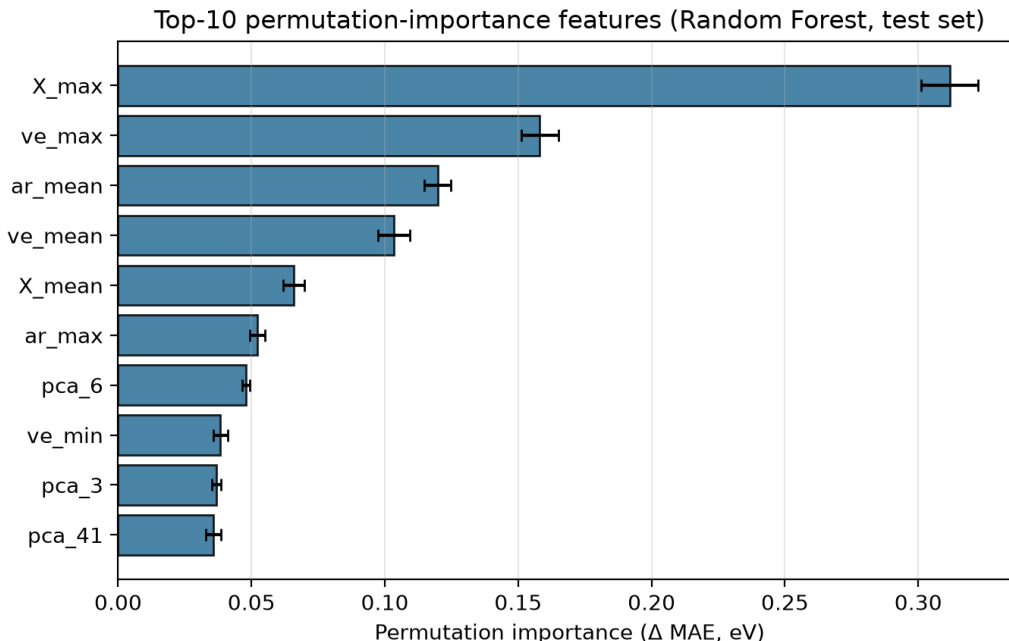


Figure 7. Top-10 permutation-importance features for the Random Forest model (test-set subsample, $n = 2000$, 5 repeats). Error bars are standard deviations across repeats. Magpie elemental-property statistics dominate; the maximum compound electronegativity is the single most important feature.

Table 2. Top-10 permutation-importance features for RF and XGBoost (importance = rise in MAE, eV, when the feature is shuffled).

Rank	RF feature	RF importance	XGBoost feature	XGBoost importance
1	$X_{\max}$	0.312	v_{mean}	0.146
2	$v_{\max}$	0.158	X_{mean}	0.145
3	r_{mean}	0.120	pca_1	0.121
4	v_{mean}	0.103	r_{mean}	0.109
5	X_{mean}	0.066	$X_{\max}$	0.083
6	$r_{\max}$	0.052	$v_{\max}$	0.081
7	pca_6	0.048	pca_2	0.039
8	$v_{\min}$	0.038	$r_{\max}$	0.038
9	pca_3	0.037	pca_3	0.037
10	pca_41	0.036	pca_8	0.030

0.305 eV (random) to 1.196 eV (LOO-O), XGBoost to 1.074 eV, and KRR to 1.111 eV, with R^2 collapsing to the 0.27–0.37 range. For titanium, the degradation is milder (RF 0.458 eV, XGBoost 0.447 eV, KRR 0.637 eV; R^2 0.07–0.22). For copper, the MAE remains moderate (RF 0.453 eV, XGBoost 0.385 eV, KRR 0.468 eV) but R^2 turns strongly negative (RF -0.74 , XGBoost -0.27 , KRR -0.51), indicating that the model’s predictions are worse than predicting the test-set mean.

The element-specific character of the degradation is itself diagnostic. Oxygen is the most abundant non-metallic constituent (26.9% of all compounds) and its exclusion removes the bulk of the insulating-oxide signal from training; the train-no-O set is 78% metallic, whereas the test-O set is only 42% metallic—a label-distribution shift that simultaneously forces elemental extrapolation and out-of-distribution target prediction. Titanium is a transition metal whose compounds span both metallic and semiconducting regimes but are chemically less central than oxygen; the moderate

Table 3. Leave-one-element-out results (test compounds containing the held-out element). LOO-O train has 75 features (66 PCA + 9 Magpie); LOO-Ti/Cu and the main set use 74. Each LOO model is fit within its own feature space.

Element	Model	MAE (eV)	RMSE (eV)	R^2
Oxygen ($n = 14992$)	Random Forest	1.196	1.506	0.278
	XGBoost	1.074	1.401	0.374
	KRR	1.111	1.514	0.270
Titanium ($n = 2154$)	Random Forest	0.458	0.928	0.070
	XGBoost	0.447	0.848	0.224
	KRR	0.637	1.024	-0.132
Copper ($n = 3086$)	Random Forest	0.453	0.727	-0.741
	XGBoost	0.385	0.621	-0.270
	KRR	0.468	0.677	-0.512

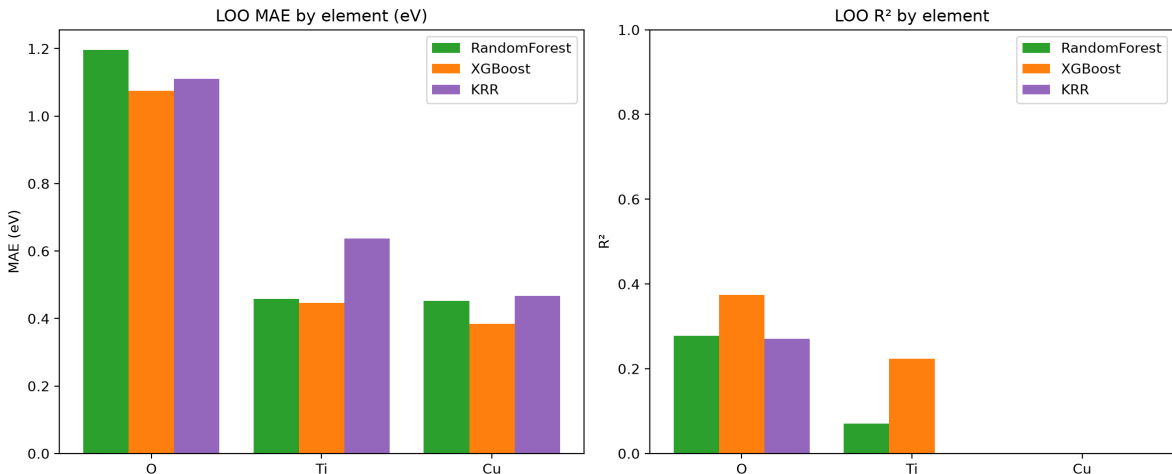


Figure 8. Leave-one-element-out performance. Oxygen exclusion causes the most severe MAE inflation ($\sim 4\times$ the random-test MAE); copper exclusion produces strongly negative R^2 despite a moderate MAE, reflecting label-distribution shift between train and test.

degradation reflects a partial analog in the training set through other early transition metals. Copper’s negative R^2 is driven less by raw error (the MAE is competitive) than by a target-distribution mismatch: the test-Cu set is dominated by small-gap metallic Cu compounds (mean 0.249 eV, 75% metallic), and the model trained without Cu systematically over-predicts these small gaps, producing a variance pattern worse than the mean. This decomposition—oxygen suffering from both extrapolation and shift, titanium from mild extrapolation, copper from distribution shift with low target variance—illustrates that the LOO protocol conflates two distinct failure mechanisms that future work should disentangle.

4 Discussion

The headline result—RF MAE = 0.305 eV, $R^2 = 0.788$ on the random JARVIS-DFT test set—is broadly consistent with published composition-only and descriptor-based band-gap benchmarks. Zhuo et al. reported MAE ≈ 0.22 eV on a curated 127000-compound inorganic set using gradient-boosted trees on composition and structure-derived descriptors [10]; Venkatraman reported MAE = 0.36 eV for composition-only models on a 60000-compound OQMD-derived set, with structure-informed

variants reaching 0.27 eV [11]; and multi-fidelity and physics-informed kernel models have reported sub-0.3-eV errors on smaller curated subsets [8, 9, 33]. Our result sits in the expected range for a structure-free, fixed-hyperparameter regressor on a 55000-compound, zero-inflated OPTB88vdW distribution. The ALIGNN benchmark on the same JARVIS-DFT field, which uses the full crystal graph plus line graph, reaches $R^2 \approx 0.81$ [15, 16]; the fact that our composition-only RF reaches $R^2 = 0.788$ indicates that, for random-split interpolation, the marginal value of crystal-structure information for band-gap prediction is modest, in agreement with Goodall & Lee’s finding that composition-only deep learners can approach structure-aware accuracy [17].

The MAE/ R^2 trade-off between RF and XGBoost deserves comment. RF achieves the lower MAE (0.305 vs. 0.342 eV) while XGBoost attains the higher R^2 (0.792 vs. 0.788). This is not a contradiction: MAE is dominated by the abundant, near-zero-gap metals, where RF’s ensemble averaging produces tight, low-variance predictions that minimize absolute error, whereas R^2 is sensitive to the variance explained across the full range, including the high-gap tail where XGBoost’s sequential gradient-fitting captures the residual structure slightly better. The practical implication is that, for applications prioritizing typical (metal-rich) compounds, RF is preferable; for applications prioritizing the non-metal tail, the two are comparable on non-metal R^2 (both 0.738) and the choice should hinge on interpretability and inference cost.

Permutation importance provides a physically interpretable window into the models. The dominance of $X_{\max}$ (the most electronegative element in the compound) is intuitive: a large $X_{\max}$ flags the presence of highly electronegative species (O, F, Cl) that drive ionic character and open band gaps. The prominence of valence-electron statistics (v_{mean} , $v_{\max}$) reflects the central role of electron counting in determining whether a compound is metallic, and the atomic-radius statistics capture size-mismatch-driven structural tendencies relevant to insulating character. The secondary role of the PCA-reduced element-fraction components is consistent with the literature observation that physically grounded property statistics often outperform raw one-hot composition features for moderate dataset sizes [18–20], though deeper composition representations can recover more signal given sufficient data [17].

The LOO results are the most cautionary finding of this study. They demonstrate that a model that interpolates well within a random split ($R^2 \approx 0.79$) can fail badly under a chemically structured split, and they do so in an element-specific manner that tracks both (i) the centrality of the held-out element to the insulating-compound class (oxygen) and (ii) the label-distribution shift between train and test. The copper case is particularly instructive: a competitive MAE coexists with a strongly negative R^2 , because the predictions, although not far off in absolute terms, fail to track the (small) target variance. This pattern echoes the broader caution in the materials-informatics literature that random-split metrics can overstate true generalization [12, 34] and that composition-disjoint or element-disjoint evaluation is essential for claims of discovery utility. The oxygen shift (train-no-O 78% metallic vs. test-O 42% metallic) is a concrete instance of the distribution-shift problem that domain-adaptation and group-aware validation methods are designed to address.

Several limitations bound these conclusions. First, KRR was trained on an 8000-record subsample with fixed α and a scale-equivalent γ , so its weaker performance reflects the methodological constraint rather than the intrinsic ceiling of kernel methods; a proper α/γ search or a Nystroem kernel approximation on the full training set would likely close part of the gap to the tree models [31]. Second, no hyperparameter tuning was performed for any model—the objective was a transparent, fixed-configuration benchmark, not a maximally optimized predictor. Third, the feature representation excludes crystal structure entirely; this is by design and a strength for structure-free screening, but it caps achievable accuracy for polymorphic compositions, of which 24153 records share a stoichiometry and 10783 exhibit distinct gaps. Polymorph overlap is label noise for a composition-only model and is not controlled here; aggregation (e.g. median band gap

per composition) or composition-disjoint test sets would quantify the resulting optimism. Fourth, the OPTB88vdW labels are themselves systematically underestimated relative to hybrid-functional or experimental gaps [1–3]; a model trained on these labels inherits their biases, and the 69% metal fraction is a direct consequence.

These limitations motivate concrete next steps. (i) *Group-aware validation*: GroupKFold cross-validation by composition and composition-disjoint held-out test sets would quantify memorization of repeated stoichiometries versus genuine generalization. (ii) *Distribution-matched LOO*: constructing LOO test subsets whose label distribution is matched to the corresponding train set would separate elemental extrapolation from distribution shift, isolating the two failure modes conflated here. (iii) *KRR refinement*: a joint α/γ search and full-training-set Nystroem approximation would test whether kernel methods can match the tree ensembles. (iv) *Functional-aware targets*: training on the smaller but more accurate HSE06 [1] or mBJ subsets of JARVIS-DFT would test transferability to physically corrected gaps, at the cost of dataset size. (v) *Composition-only deep learners*: COMPaad/ElemNet-style composition-only deep networks [17] would test whether richer composition representations close the small gap to structure-aware models on the same dataset.

5 Conclusion

We benchmarked three composition-only machine-learning models—Random Forest, XGBoost, and Kernel Ridge Regression—on 55710 inorganic crystals from JARVIS-DFT, predicting OPTB88vdW band gaps from a 74-dimensional feature vector of PCA-reduced element fractions and Magpie elemental-property statistics. On a leakage-safe random split, Random Forest achieved the best mean absolute error (0.305 eV, $R^2 = 0.788$), competitive with published structure-aware benchmarks, with the physically interpretable maximum compound electronegativity and valence-electron statistics identified as the dominant predictive features. The metallicity-stratified analysis showed that the headline metrics are dominated by an abundant zero-gap-metal regime and that the practically important non-metal regime is predicted with MAE = 0.621 eV and $R^2 \approx 0.74$. Leave-one-element-out experiments for oxygen, titanium, and copper revealed that random-split accuracy conceals a sharp dependence on chemical extrapolation: oxygen exclusion inflated the MAE roughly four-fold and copper exclusion drove R^2 strongly negative, with the degradation attributable to both elemental extrapolation and label-distribution shift. These results establish a transparent baseline for composition-only band-gap prediction on JARVIS-DFT and delineate the interpolation–extrapolation boundary that governs model reliability, underscoring that honest evaluation of materials-informatics models requires chemically structured, not merely random, validation protocols.

Data and code availability

The JARVIS-DFT dataset is publicly available at <https://jarvis.nist.gov> and through the `jarvis-tools` package [4]. All trained models (12 joblib artifacts: `rf_main`, `xgb_main`, `krr_main` plus nine LOO models), metrics (`model_metrics.json`, `loo_metrics.json`, `permutation_importance.json`), test predictions (`main_test_predictions.npz`), and the visualization scripts are available in the project repository. All experiments use `random_state=42`; scikit-learn [29], XGBoost [30], pymatgen [22], and matminer [21] underpin the pipeline.

Author contributions

Conceptualization, methodology, software, analysis, and writing: LitLit AI.

References

- [1] Jochen Heyd, Gustavo E. Scuseria, and Matthias Ernzerhof. Hybrid functionals based on a screened Coulomb potential. *The Journal of Chemical Physics*, 118(18):8207–8215, 2003. doi:

10.1063/1.1564060.

- [2] John P. Perdew and Mel Levy. Physical content of the exact Kohn-Sham orbital energies: Band gaps and derivative discontinuity. *Physical Review Letters*, 51(19):1884–1887, 1983. doi: 10.1103/PhysRevLett.51.1884.
- [3] Pedro Borlido, Thorsten Aull, Ahmad W. Huran, Fabien Tran, Miguel A. L. Marques, and Silvana Botti. Large-scale benchmark of exchange–correlation functionals for the determination of electronic band gaps of solids. *Journal of Chemical Theory and Computation*, 15(9):5069–5079, 2019. doi: 10.1021/acs.jctc.9b00322.
- [4] Kamal Choudhary, Kevin F. Garrity, Andrew C. E. Reid, Brian DeCost, Adam J. Biacchi, and Angela R. Hight Walker. The joint automated repository for various integrated simulations (JARVIS) for data-driven materials design. *npj Computational Materials*, 6(1):173, 2020. doi: 10.1038/s41524-020-00440-1.
- [5] Ankit Agrawal and Alok Choudhary. Perspective: Materials informatics and big data: Realization of the “fourth paradigm” of science in materials science. *APL Materials*, 4(5):053208, 2016. doi: 10.1063/1.4946894.
- [6] Rampi Ramprasad, Rohit Batra, Ghanshyam Pilania, and Arun Mannodi-Kanakkithodi. Machine learning in materials informatics: recent applications and prospects. *npj Computational Materials*, 3(1):54, 2017. doi: 10.1038/s41524-017-0056-5.
- [7] Ankit Agrawal and Alok Choudhary. Deep materials informatics: Applications of deep learning in materials science. *MRS Communications*, 9(3):779–792, 2019. doi: 10.1557/mrc.2019.73.
- [8] Joohwi Lee, Atsuto Seko, Kazuki Shitara, Keita Nakayama, and Isao Tanaka. Prediction model of band gap for inorganic compounds by combination of density functional theory calculations and machine learning techniques. *Physical Review B*, 93(11):115104, 2016. doi: 10.1103/PhysRevB.93.115104.
- [9] G. Pilania, J. E. Gubernatis, and T. Lookman. Multi-fidelity machine learning models for accurate bandgap predictions of solids. *Computational Materials Science*, 129:156–163, 2017. doi: 10.1016/j.commatsci.2016.12.004.
- [10] Ya Zhuo, Aria Mansouri Tehrani, and Jakoah Brgoch. Predicting the band gaps of inorganic solids by machine learning. *The Journal of Physical Chemistry Letters*, 9(7):1668–1673, 2018. doi: 10.1021/acs.jpcllett.8b00124.
- [11] Vishwesh Venkatraman. The utility of composition-based machine learning models for band gap prediction. *Computational Materials Science*, 197:110637, 2021. doi: 10.1016/j.commatsci.2021.110637.
- [12] Pierre-Paul De Breuck, Geoffroy Hautier, and Gian-Marco Rignanese. Materials property prediction for limited datasets enabled by feature selection and joint learning. *npj Computational Materials*, 7(1):83, 2021. doi: 10.1038/s41524-021-00552-2.
- [13] 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.

- [14] K. T. Schütt, H. E. Sauceda, P.-J. Kindermans, A. Tkatchenko, and K.-R. Müller. SchNet – a deep learning architecture for molecules and materials. *The Journal of Chemical Physics*, 148(24):241722, 2018. doi: 10.1063/1.5019779.
- [15] Kamal Choudhary and Brian DeCost. Atomistic line graph neural network for improved materials property predictions. *npj Computational Materials*, 7(1):185, 2021. doi: 10.1038/s41524-021-00650-1.
- [16] Prathik R. Kaundinya, Kamal Choudhary, and Surya R. Kalidindi. Prediction of the electron density of states for crystalline compounds with atomistic line graph neural networks (ALIGNN). *JOM*, 74(4):1395–1405, 2022. doi: 10.1007/s11837-022-05199-y.
- [17] Rhys E. A. Goodall and Alpha A. Lee. Predicting materials properties without crystal structure: deep representation learning from composition. *Nature Communications*, 11(1):6280, 2020. doi: 10.1038/s41467-020-19964-7.
- [18] Olexandr Isayev, Corey Oses, Cormac Toher, Eric Gossett, Stefano Curtarolo, and Alexander Tropsha. Universal fragment descriptors for predicting properties of inorganic crystals. *Nature Communications*, 8(1):15679, 2017. doi: 10.1038/ncomms15679.
- [19] Giovanni Trezza and Eliodoro Chiavazzo. Leveraging composition-based energy material descriptors for machine learning models. *Materials Today Communications*, 36:106579, 2023. doi: 10.1016/j.mtcomm.2023.106579.
- [20] 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):16028, 2016. doi: 10.1038/npjcompumats.2016.28.
- [21] Logan Ward, Alexander Dunn, Alireza Faghaninia, Nils E. R. Zimmermann, Saurabh Bajaj, Qing Wang, Joseph H. Montoya, Jatin Choubey, Xiangyu Liu, Matthias Rupp, 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.
- [22] Shyue Ping Ong, William Davidson Richards, Anubhav Jain, Geoffroy Hautier, Michael Kocher, Shreyas Cholia, Dan Gunter, Vincent L. Persson, and Gerbrand Ceder. Python materials genomics (pymatgen): A robust, open-source python library for materials analysis. *Computational Materials Science*, 68:314–319, 2013. doi: 10.1016/j.commatsci.2012.10.028.
- [23] 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.
- [24] 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.
- [25] Scott Kirklin, James E. Saal, Bryce Meredig, Alex Thompson, Jeffery C. Persson, Alok Choudhary, and C. 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.

- [26] Claudia Draxl and Matthias Scheffler. NOMAD: The FAIR concept for big data-driven materials science. *MRS Bulletin*, 43(9):676–682, 2018. doi: 10.1557/mrs.2018.208.
- [27] Claudia Draxl and Matthias Scheffler. The NOMAD laboratory: from data sharing to artificial intelligence. *Journal of Physics: Materials*, 2(3):036001, 2019. doi: 10.1088/2515-7639/ab13bb.
- [28] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001. doi: 10.1023/A:1010933404324.
- [29] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.
- [30] 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.
- [31] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Series in Statistics. Springer, 2nd edition, 2009. doi: 10.1007/978-0-387-84858-7.
- [32] André Altmann, Laura Tološi, Oliver Sander, and Thomas Lengauer. Permutation importance: a corrected feature importance measure. *Bioinformatics*, 26(10):1340–1347, 2010. doi: 10.1093/bioinformatics/btq134.
- [33] G. Pilania, K. J. McClellan, C. R. Stanek, and B. P. Uberuaga. Physics-informed machine learning for inorganic scintillator discovery. *The Journal of Chemical Physics*, 148(24):241702, 2018. doi: 10.1063/1.5025819.
- [34] Christopher J. Bartel, Amalie Trewartha, Qi Wang, Alexander Dunn, Anubhav Jain, and Gerbrand Ceder. A critical examination of compound stability predictions from machine-learned formation energies. *npj Computational Materials*, 6(1):97, 2020. doi: 10.1038/s41524-020-00362-y.