



Machine-Learning Classification of Topological Materials Using Crystallographic and Compositional Descriptors for Pre-DFT Screening

Sammy S. Dan, Joseph B. Emah, Sylvester A. Ekong

Abstract

Identifying topological insulators and semimetals among inorganic crystalline materials traditionally requires density functional theory (DFT) calculations with spin–orbit coupling, a computationally expensive process that limits high-throughput screening. This study examined whether machine learning models trained on ground-state crystallographic and compositional descriptors, obtainable without DFT or space-group information, could reliably classify materials as topologically non-trivial, trivial, or semimetallic. A dataset of 1,391 labelled crystalline entries (230 semimetal, 781 topological, 380 trivial), curated from the Inorganic Crystal Structure Database and the Topological Quantum Chemistry database, was cleaned, including manual correction, via direct ICSD lookup, of a systematic lattice-parameter data-entry artefact (an Excel auto-fill error that had propagated a single genuine c-parameter value across 1,208 rows, identified by cross-validation against unit-cell volume), prior to a further, now pipeline-integrated audit that removed 47 true-duplicate ICSD entries and attempted live TQC API resolution of 10 conflicting-label ICSD groups, none of which returned a match, so the first-encountered row was conservatively retained in each case (net reduction from 1,450 to 1,391 entries), and used to engineer 31 structural and compositional features motivated by the physics of spin-orbit-coupling-driven band inversion. Six classifiers, Random Forest, XGBoost, LightGBM, Support Vector Machine, an Artificial Neural Network, and a soft-voting Ensemble, were trained and optimised using nested stratified cross-validation, Bayesian hyperparameter search, and Synthetic Minority Oversampling Technique (SMOTE)-based class rebalancing applied strictly within each training fold. XGBoost produced the highest observed test-set performance (macro-F1 of 0.744, accuracy of 78.1%, macro-AUC-ROC of 0.896), narrowly ahead of LightGBM (macro-F1 0.721, accuracy 74.9%, macro-AUC-ROC 0.887), though the difference was not statistically significant (McNemar's test, $p = 1.000$, Bonferroni-corrected); LightGBM was retained as the primary model for interpretability and screening on the basis of its lower cross-validation variance and better-behaved post-hoc probability calibration. Shapley Additive exPlanations (SHAP), permutation, and impurity-based feature importance analyses converged on a consistent predictor hierarchy: number of constituent elements, mean ionisation potential, electronegativity range, maximum atomic number, and mean atomic radius, corroborated by polymorph-level diagnostics providing evidence that the model exploits structural, not merely compositional, information in most cases, although structural signal proved insufficient in at least one hard case. Applied to the full dataset, LightGBM ranked 424 entries as Tier 1 (P(Topological))

≥ 0.80) high-confidence topological predictions within the labelled dataset, of which 419 corresponded to already-known topological labels — a precision of 98.8% against known labels (retrospective validation precision, not prospective discovery precision). Interpretability analysis was also used as a data-quality diagnostic, motivating a closer investigation that, together with cross-validation against unit-cell volume and a subsequent ICSD/TQC provenance audit, uncovering and correcting a latent artefact in an earlier version of the model's most important feature. These findings demonstrate that pre-DFT structural and compositional descriptors carry physically plausible and interpretable predictive signal for topological classification, offering an inexpensive pre-screening tool to prioritise candidate materials for further investigation.

✉ Joseph B. Emah: josephemah@aksu.edu.ng

Department of Physics, Faculty of Physical Sciences, Akwa Ibom State University, Ikot Akpaden, Mkpata Enin L.G.A., Akwa Ibom State, Nigeria

Keywords: Topological insulators; Machine learning; SHAP; LightGBM; Materials informatics

Article history

Received: July 19, 2026

Received in revised form: September 28, 2026

Accepted for publication: October 2, 2026

Available online: October 15, 2026

1.0 Introduction

The discovery of topological insulators (TIs) in the mid-2000s marked a paradigm shift in condensed matter physics. Unlike conventional band insulators, TIs possess a bulk electronic band gap that coexists with symmetry-protected, gapless surface states, a consequence of non-trivial topological invariants encoded in the bulk Bloch wavefunctions. For a periodic crystal, these wavefunctions satisfy the Bloch Hamiltonian eigenvalue equation:

$$H(\mathbf{k}) u_{n\mathbf{k}}(\mathbf{r}) = E_n(\mathbf{k}) u_{n\mathbf{k}}(\mathbf{r}) \quad (1)$$

where $H(\mathbf{k})$ is the momentum-space Hamiltonian defined over the Brillouin zone, $u_{n\mathbf{k}}$ is the periodic part of the Bloch state for band n , and $E_n(\mathbf{k})$ is the corresponding band energy. Distinct topological classes of $H(\mathbf{k})$ cannot be continuously deformed into one another without closing the bulk gap. Kane and Mele (2005) and Bernevig, Hughes, and Zhang (2006) predicted the quantum spin Hall effect in time-reversal-symmetric two-dimensional systems, classified by a binary $\mathbb{Z}_2$ invariant rather than an integer Chern number; this was confirmed experimentally by König et al. (2007). For centrosymmetric crystals, Fu and Kane (2007) showed that this invariant reduces

to a parity criterion evaluated at time-reversal-invariant momenta (TRIM):

$$(-1)^\nu = \prod_i \prod_m \xi_{2m}(\Gamma_i) \quad (2)$$

where $\xi_{2m}(\Gamma_i)$ is the parity eigenvalue of the m -th occupied Kramers pair at TRIM point Γ_i . A value of $(-1)^\nu = -1$ ($\nu = 1$) identifies a topological insulator. This criterion, together with its three-dimensional generalisation into four invariants ($\nu_0; \nu_1 \nu_2 \nu_3$) (Fu, Kane & Mele, 2007; Moore & Balents, 2007), underlies the high-throughput density functional theory (DFT) workflows that populate present-day topological materials databases, including the Topological Quantum Chemistry (TQC) framework of Bradlyn et al. (2017) and the classification catalogues of Vergniory et al. (2019, 2022).

Despite this theoretical maturity, evaluating Equation (2) for a candidate material requires a full DFT calculation with spin-orbit coupling, typically costing tens of thousands of CPU-hours per compound. Applied at the scale of the Inorganic Crystal Structure Database (ICSD), which catalogues over 220,000 experimentally reported structures, exhaustive DFT-based screening is computationally intractable. This

has motivated growing interest in machine learning (ML) as a pre-screening layer: since strong spin–orbit coupling is an intrinsic property of heavy elements (e.g., Bi, $Z = 83$; Sb, $Z = 51$), and since crystal symmetry constrains the allowed band representations that determine ν , elemental composition and lattice geometry plausibly carry a detectable statistical signature of topological character, obtainable without evaluating Equation (1) or (2) directly.

However, three problems limit the practical realisation of this idea. First, materials data assembled from heterogeneous sources are commonly affected by label inconsistencies, missing values, and data-entry artefacts, issues

rarely discussed in published ML-for-topology studies, which typically report results on pre-cleaned, single-source datasets. Second, natural populations of topological, trivial, and semimetallic materials are not evenly represented, so classifiers trained without explicit imbalance-handling are biased toward the majority class, and accuracy alone becomes an unreliable performance criterion. Third, the extent to which topological character is predictable from ground-state structural descriptors alone, without recourse to electronic-structure information, remains empirically under-tested, as does the physical interpretability of any resulting model. Table 1 situates the present approach relative to five categories of prior work.

Table 1: Categories of prior computational approaches to topological material identification, compared to the present study

Category	Representative work	Input required	Scope
Physics-based identification	Kane & Mele (2005); Fu & Kane (2007)	Full band structure + parity eigenvalues	Theoretical criterion, not a screening tool
DFT high-throughput / TQC	Bradlyn et al. (2017); Vergniory et al. (2019, 2022)	Full DFT calculation with spin–orbit coupling	Exhaustive but computationally intractable at ICSD scale
Composition/structure-only ML	Sun et al. (2018); Claussen et al. (2020)	Compositional and/or structural descriptors, no DFT	Fast, pre-DFT, but typically binary and single-source
Structure + electronic-descriptor ML	Cao et al. (2020)	86-descriptor structural/electronic set	High accuracy, narrower chemical scope
Graph neural network (GNN)	Louis et al. (2020); Schleder et al. (2021)	Graph-structured or DFT-derived features	Highest reported accuracy, largest data/feature requirements
Present study	—	<i>Pre-DFT crystallographic + compositional descriptors only, no space group</i>	<i>Three-class (Semimetal-inclusive), explicit data-quality audit, full interpretability treatment</i>

This study addresses these problems using a dataset of 1,391 labelled crystalline entries, curated from the ICSD and the TQC database following a full data-quality and duplicate/conflict-resolution audit (Section 2.2), spanning heavy p-block chalcogenides/pnictides classically associated with topological behaviour and a substantial population of cerium-based rare-earth compounds. The aim is to develop, validate, and physically interpret machine learning models that classify materials as topological, trivial, or semimetallic using only

pre-DFT crystallographic and compositional descriptors. Semimetal, Topological, and Trivial correspond respectively to the ESFD/ES, SEBR/NLC, and LCEBR TQC sub-types (Section 3.1); a three-class formulation was adopted, rather than a binary topological/trivial split, because collapsing semimetals into either class would discard the Fermi-degeneracy distinction underlying the separate physical mechanism identified in Section 3.3. The specific objectives are to: (i) perform exploratory data analysis and correct label and

data-quality artefacts in the dataset; (ii) engineer structural, compositional, and symmetry-derived features motivated by the band-inversion mechanism underlying Equations (1)–(2); (iii) train and optimise six classifiers (Random Forest, XGBoost, LightGBM, Support Vector Machine, an Artificial Neural Network, and a soft-voting Ensemble) using nested cross-validation and Bayesian hyperparameter search; (iv) evaluate performance using imbalance-robust metrics (macro-F1, Matthews Correlation Coefficient, macro-AUC-ROC), validated additionally under a composition-grouped cross-validation scheme (Section 2.6) to guard against chemically related entries inflating performance estimates; (v) apply SHAP and permutation-based interpretability methods to identify and physically ground the dominant predictors; and (vi) apply the best model to screen the full dataset for high-confidence candidate topological *predictions* warranting further DFT or experimental investigation.

The study contributes to the literature in four respects that, to our knowledge, are not commonly found jointly addressed in the published ML-for-topology literature reviewed in Section 3.4: explicit, documented treatment of label noise and class imbalance in an ICSD-derived topological dataset — including, uncommonly for this literature, live-API-checked verification of duplicate and conflicting labels against the source TQC database rather than reliance on the original compiled labels alone; a controlled comparison of RF, GBM, SVM, and ANN architectures under identical experimental protocols; systematic linkage of SHAP-derived feature importances to the underlying physics of spin–orbit-coupling-driven band inversion; and a polymorph-level diagnostic, using chemically identical but structurally distinct entries (e.g., Bi_2O_3 , Sb_2Te_3), that tests whether the model exploits genuine structural information rather than composition alone. The scope is restricted to non-magnetic, three-class topological labelling

(Topological, Trivial, Semimetal) using structural and compositional features only; all seven raw cell parameters and unit-cell volume used here are routinely reported in any indexed powder or single-crystal XRD structure solution, so the descriptor set is obtainable from experimental structures without recourse to DFT. Space-group information is deliberately excluded to preserve this pre-DFT, experiment-compatible scope (Section 3.4 quantifies the resulting performance trade-off against comparable studies that do include it); other electronic-structure features (DFT band gaps, parity eigenvalues) are deliberately excluded, since their inclusion would defeat the purpose of a pre-DFT screening tool. The principal limitations, discussed further in Section 3.5, are the modest dataset size relative to typical materials-informatics corpora and the exclusion of magnetic and graph-based structural representations from the present feature set.

2.0 Materials and Methods

2.1 Research Design and Computational Environment

This study follows a supervised, quantitative machine-learning research design comprising five sequential phases: data acquisition and quality audit, preprocessing, feature engineering, model development and validation, and interpretability/candidate screening. All analyses were implemented in Python 3.10 using Pandas and NumPy for data manipulation, scikit-learn for classical classifiers and cross-validation, XGBoost and LightGBM for gradient boosting, PyTorch for the neural network, pymatgen for elemental property retrieval, and SHAP for interpretability. Computations were performed on a standard desktop workstation (Intel Core i7, 16 GB RAM); GPU acceleration was not required given the modest dataset size.

2.2 Data Source and Description

The dataset comprised 1,451 inorganic crystalline material entries derived from the

Inorganic Crystal Structure Database (ICSD), each subsequently classified by the Topological Quantum Chemistry (TQC) framework (Bradlyn et al., 2017) and the high-throughput DFT screening of Vergniory et al. (2019, 2022). Each entry recorded a chemical formula, ICSD reference code, topological classification label (Topological, Trivial, or Semimetal), a finer topological sub-type (SEBR, LCEBR, NLC, ESFD, or ES), six crystallographic cell parameters (a , b , c , α , β , γ), and unit cell volume (V).

A programmatic and manual audit identified and corrected discrete data-entry errors (whitespace-padded labels, element-symbol capitalisation errors, misdirected ICSD reference values) and one systematic column-wide artefact, in which an Excel auto-fill error had propagated a single genuine c -parameter value (21.3707 Å) across 1,208 subsequent rows. This artefact was first identified by cross-validation against the independently populated unit-cell volume and the remaining lattice parameters, and the genuine per-entry values were then manually recovered by direct lookup of each affected ICSD record against the ICSD web database prior to any further automated cleaning.

Following correction, 1,450 of 1,451 entries carried a valid classification label. A subsequent audit - now integrated directly into the data pipeline rather than run as a separate, undocumented script - queried the ICSD reference codes for duplication and cross-checked classification labels against the live TQC API. This audit identified 116 rows implicated in ICSD-number duplication: 47 rows were confirmed as true duplicates of an already-represented entry and removed (retaining the first occurrence), and a further 10 ICSD codes carried internally conflicting classification or sub-type labels across their duplicate rows. A live TQC API lookup was attempted for each of these 10 groups; none returned a match, so each group was resolved to a single row by conservatively retaining the

first-encountered row, removing a further 12 rows without external verification — a residual data-quality uncertainty affecting 10 of 1,391 entries (0.7%), noted as a limitation in Section 3.5. Together, this reduced the dataset by a net 59 entries, from 1,450 to a final 1,391 uniquely and consistently labelled entries: 781 (56.1%) Topological (402 SEBR, 379 NLC), 380 (27.3%) Trivial (379 LCEBR, 1 NLC), and 230 (16.5%) Semimetal (175 ESFD, 55 ES) — an imbalance with direct implications for model training and evaluation. Because the ICSD-conflict-resolution step queries a live external database, its output can in principle vary slightly if re-run after the source database changes; the 1,391-entry dataset reported throughout this manuscript is fixed as of the run reported here and used unchanged for all subsequent results.

Polymorphs - chemically identical entries with distinct crystal structures and, in most cases, distinct ICSD numbers were deliberately retained rather than deduplicated by formula, since they enable the structural-versus-compositional diagnostic reported in Section 3.3.

2.3 Data Preprocessing

Crystallographic uncertainty notation (e.g., 4.1720(9)) was parsed by regular-expression extraction of the nominal value. Outliers in the lattice parameters were identified using the interquartile range (IQR) rule and corrected against source records where a data-entry error was confirmed (e.g., a misplaced decimal point in one Cell Length C value — ICSD 20404, in which 10,764 had been recorded in place of 10.764 Å, a separate, single-row transcription error distinct from the systematic c -parameter artefact described in Section 2.2). Unit cell volume, which was strongly right-skewed, was log-transformed. All continuous features were standardised (z-scored) prior to model fitting for scale-sensitive algorithms (SVM, ANN).

Class imbalance was addressed using the Synthetic Minority Over-sampling Technique

(SMOTE; Chawla et al., 2002), which generates synthetic minority-class examples by linear interpolation between a minority sample $\mathbf{x}_i$ and one of its k nearest neighbours $\mathbf{x}_j$ ($k = 5$):

$$\mathbf{x}_{\text{syn}} = \mathbf{x}_i + \lambda \times (\mathbf{x}_j - \mathbf{x}_i), \lambda \sim \text{Uniform}(0,1) \quad (3)$$

SMOTE was applied only within each training fold (never to validation or test data) to prevent information leakage. SMOTE interpolates in the standardised 31-dimensional engineered-feature space, not in a physically constrained structural/chemical space; synthetic points therefore have no guarantee of corresponding to a realisable crystalline compound and are used strictly as a training-time class-rebalancing device.

2.4 Feature Engineering

Thirty-one features were engineered across three physically motivated categories. Structural features (15) comprised the seven raw crystallographic parameters plus derived quantities including the axial ratio c/a , the lattice anisotropy index:

$$A_{\text{lat}} = \frac{\sigma(a, b, c)}{\mu(a, b, c)} \quad (4)$$

angle-deviation terms $|\alpha - 90^\circ|$, $|\beta - 90^\circ|$, $|\gamma - 90^\circ|$, and volume per atom. One structural-derived feature, c_{reliable} , is a legacy indicator from an earlier pipeline stage that flagged rows matching the pre-correction artefact value (21.371 Å); since that artefact was manually corrected prior to pipeline ingestion (Section 2.2), the flag now identifies only the single row genuinely carrying this value and contributes negligible information to any model (mean $|\text{SHAP}| \approx 0$, mean decrease impurity ≈ 0 across all three tree-based models; Supplementary Table S1). Symmetry-derived features (4) were one-hot indicators for crystal system (Cubic/Tetragonal/Orthorhombic, Hexagonal/Trigonal, Monoclinic, Triclinic), inferred deterministically from the interaxial angles using a $\pm 0.5^\circ$ tolerance around each target angle (90° or 120°), applied in the priority order Hexagonal/Trigonal $\rightarrow$ Cubic/ Tetragonal/

Orthorhombic $\rightarrow$ Monoclinic $\rightarrow$ Triclinic. Compositional features (12) included maximum, mean, and stoichiometry-weighted atomic number (proxying spin-orbit coupling strength), electronegativity range and means, mean atomic radius, and the number of constituent elements, retrieved via pymatgen. Pairwise Pearson correlations were screened for redundancy ($|\rho| > 0.90$); a small number of pairs among definitionally or structurally related features exceeded this threshold — notably Cell Length C (parsed) versus the c/a ratio ($\rho \approx 1.00$, since c/a is a direct linear function of Cell Length C at fixed a) and Cell Angle γ (parsed) / $|\gamma - 90^\circ|$ versus the Hexagonal-Trigonal symmetry indicator ($\rho \approx 0.95\text{--}0.96$, since the indicator is deterministically derived from γ). These pairs were retained rather than pruned, since each member captures a distinct representation (continuous raw value vs. derived ratio or categorical indicator) that later analyses (Section 3.3) show contribute independently to model importance; no pair among the compositional or otherwise physically independent features exceeded the threshold. Full symbol-level definitions and formulas for all 31 features are given in Supplementary Table S1.

2.5 Algorithm Selection and Model Formulation

Six classifiers were developed. The *Random Forest* (RF; Breiman, 2001) aggregates B bootstrap-trained decision trees by majority vote:

$$\hat{y}_{\text{RF}}(\mathbf{x}) = \arg \max_k \sum_{b=1}^B \mathbb{1}[T_b(\mathbf{x}) = k] \quad (5)$$

XGBoost (Chen & Guestrin, 2016) and *LightGBM* (Ke et al., 2017) are gradient boosting machines that build an additive ensemble of shallow trees sequentially,

$$F_M(\mathbf{x}) = \sum_{m=0}^M \eta \cdot h_m(\mathbf{x}) \quad (6)$$

where h_m is the m -th tree and η is the learning rate, differing in tree-growth strategy (level-wise for XGBoost, leaf-wise for LightGBM). The

Support Vector Machine (SVM; Cortes & Vapnik, 1995) used a radial basis function kernel,

$$K(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\gamma \|\mathbf{x}_i - \mathbf{x}_j\|^2) \quad (7)$$

with a one-versus-one decomposition for the three-class problem. (The final tuned SVM - Section 2.6/Supplementary Table S2 - in fact selected a degree-2 polynomial kernel over the RBF kernel during hyperparameter search; the RBF form above is retained as the baseline formulation searched over.) The *Artificial Neural Network* (ANN) was a feedforward multilayer perceptron,

$$\mathbf{h}^{(0)} = \mathbf{x}, \mathbf{h}^{(l)} = f(\mathbf{W}^{(l)}\mathbf{h}^{(l-1)} + \mathbf{b}^{(l)}),$$

$$\hat{y} = \text{softmax}(\mathbf{W}^{(L+1)}\mathbf{h}^{(L)} + \mathbf{b}^{(L+1)}) \quad (8)$$

trained with dropout, L_2 weight decay, and early stopping to control overfitting on the modest training set. Finally, a soft-voting *Ensemble* combined the calibrated class-probability outputs of the four base models with optimised weights.

2.6 Training, Validation, and Hyperparameter Optimization

The 1,391 labelled entries were split 80/20 (stratified) into training ($n = 1,112$) and held-out test ($n = 279$) sets, preserving class proportions (test set: 157 Topological / 76 Trivial / 46 Semimetal; train pool: 624 Topological / 304 Trivial / 184 Semimetal). A further stratified 15 % calibration split ($n = 167$) was carved from the training pool before scaling or SMOTE, reserved exclusively for post-hoc probability calibration (Section 3.5) and never seen by any model, in original or SMOTE-synthesised form, at fit time, leaving a base training set of $n = 945$ (pre-SMOTE) used to fit all six models. Within the training set, hyperparameters were tuned using stratified 5-fold cross-validation, with randomised search for RF and ANN and Bayesian optimization (Gaussian process surrogate, expected-improvement acquisition) for XGBoost, LightGBM, and a custom precomputed-fold search for SVM, each evaluating 40–100

configurations. SMOTE and feature scaling were nested strictly within each training fold using scikit-learn pipelines to prevent leakage.

Final selected hyperparameters: RF - $n_{\text{estimators}} = 200$, $\text{max_depth} = 20$, $\text{min_samples_split} = 10$, $\text{min_samples_leaf} = 2$, $\text{max_features} = \text{'log2'}$, $\text{class_weight} = \text{'balanced'}$ (100 iterations; CV macro-F1 = 0.7182 ± 0.0423 , up from a default-hyperparameter baseline of 0.6921 ± 0.0524). XGBoost - $n_{\text{estimators}} = 201$, $\text{learning_rate} = 0.0297$, $\text{max_depth} = 7$, $\text{min_child_weight} = 5$, $\text{subsample} = 1.0$, $\text{colsample_bytree} = 0.6$, $\text{reg_alpha} = 6.91 \times 10^{-5}$, $\text{reg_lambda} = 0.743$ (60 iterations; CV macro-F1 = 0.7281 ± 0.0512 , baseline 0.7102 ± 0.0533). LightGBM - $n_{\text{estimators}} = 433$, $\text{learning_rate} = 0.01$, $\text{num_leaves} = 127$, $\text{min_child_samples} = 50$, $\text{subsample} = 1.0$, $\text{colsample_bytree} = 0.6$, $\text{reg_alpha} = 2.0$, $\text{reg_lambda} = 1 \times 10^{-6}$ (60 iterations; CV macro-F1 = 0.7245 ± 0.0370 , baseline 0.6977 ± 0.0483). SVM— $\text{kernel} = \text{'poly'}$, $C = 85.84$, $\text{degree} = 2$, $\text{coef0} = 1.0$, $\gamma = 0.0187$, $\text{class_weight} = \text{'balanced'}$ (60 iterations, 7/60 candidates excluded for degenerate collapse; CV macro-F1 = 0.6838 ± 0.0359 , baseline 0.6396 ± 0.0492). ANN- $\text{hidden_layer_sizes} = (128, 64, 32)$, $\text{activation} = \text{'tanh'}$, $\alpha = 0.00605$, $\text{learning_rate_init} = 0.000416$, $\text{batch_size} = 32$ (40 iterations; CV macro-F1 = 0.6842 ± 0.0452 , baseline 0.6442 ± 0.0437). Total hyperparameter search and training time: 41.0 minutes. Full search spaces are given in Supplementary Table S2.

As a robustness check on the random stratified split, a composition-grouped 5-fold cross-validation (GroupKFold on normalised chemical formula, so that no formula - including polymorph pairs - straddled both the training and validation portion of a fold) was additionally performed for the two best-performing models, using the tuned hyperparameters above.

On the final, corrected 1,391-entry dataset, this showed a macro-F1 reduction of 0.0046

(Random Forest: 0.7212 $\rightarrow$ 0.7166) and 0.0145 (LightGBM: 0.7387 $\rightarrow$ 0.7243) relative to random-split cross-validation, confirming the composition-related leakage concern raised by both reviewers is real but modest on the final dataset (Supplementary Table S3).

2.7 Model Evaluation Metrics

Given the class imbalance, overall accuracy was reported only as a secondary reference. The primary model-selection criterion was the macro-averaged F1 score,

$$F1_{\text{macro}} = \frac{1}{K} \sum_{k=1}^K F1_k \quad (9)$$

supplemented by the multi-class Matthews Correlation Coefficient,

$$MCC = \frac{c \cdot s - \sum_k p_k t_k}{\sqrt{(s^2 - \sum_k p_k^2)(s^2 - \sum_k t_k^2)}} \quad (10)$$

where c is the number of correctly classified samples, s the total sample count, and p_k, t_k the predicted and true counts for class k . Macro-averaged area under the receiver operating characteristic curve (AUC-ROC), Cohen's κ , class-specific precision/recall/F1, and the full confusion matrix were reported as supplementary diagnostics. Pairwise model comparisons on the test set used McNemar's test with Bonferroni correction for 15 comparisons ($\alpha = 0.0033$). For McNemar's test, each test-set prediction was scored as a binary correct/incorrect outcome (prediction equals true class label, irrespective of which of the two other classes an incorrect prediction fell into), and the paired correct/incorrect outcomes of each pair of models were compared.

2.8 Interpretability and Candidate Screening

Model interpretability was assessed using SHapley Additive exPlanations (SHAP; Lundberg & Lee, 2017), which decomposes a prediction as:

$$f(\mathbf{x}) = \phi_0 + \sum_{i=1}^p \phi_i(\mathbf{x}) \quad (11)$$

where ϕ_0 is the expected model output and $\phi_i(\mathbf{x})$ is the Shapley value of feature i , computed

exactly for tree-based models via TreeExplainer. SHAP-derived global and local feature importances were cross-checked against permutation feature importance and native impurity-based importance, and are interpreted as identifying statistically important predictors whose effects are physically plausible and consistent with established mechanisms, rather than as establishing causal physical determinants (Section 3.3). The best-performing model was subsequently applied to the full dataset to rank materials by predicted topological probability, stratifying candidates into confidence tiers for prioritised follow-up by DFT or experiment. This constitutes a retrospective ranking/validation exercise against already-known labels rather than prospective discovery of previously unlabelled materials (Section 3.6).

3.0 Results and Discussion

3.1 Exploratory Data Analysis

The cleaned dataset of 1,391 labelled entries showed a moderate three-class imbalance: 781 Topological (56.1%), 380 Trivial (27.3%), and 230 Semimetal (16.5%) (Table 2a; Figure 1), a 3.4:1 majority-to-minority ratio motivating the SMOTE and macro-averaged evaluation strategy adopted in Section 2. The correspondence between primary labels and finer topological sub-types was internally consistent: all 402 Symmetry-Enforced Band Representation (SEBR) entries and 379 of 380 Non-Linear Chern (NLC) entries mapped to Topological, with the single remaining NLC entry mapping to Trivial; all 379 Large-gap Conventional Elementary Band Representation (LCEBR) entries mapped to Trivial; and all 175 Enforced Semimetal with Fermi Degeneracy (ESFD) and 55 Enforced Semimetal (ES) entries mapped to Semimetal (Figure 2; Table 2a).

Structurally, the Semimetal class showed the smallest and tightest log-volume distribution (mean $\ln V = 5.22$), consistent with band-touching behaviour favouring compact, higher-

symmetry structures, while Topological and Trivial materials spanned broader volume ranges (means 5.79 and 5.86 respectively). Crystal-system analysis (Figure 3; Table 2b) showed Semimetals disproportionately concentrated in hexagonal/trigonal symmetry (25.2%, versus 12.9% for Topological and 15.8% for Trivial), plausibly reflecting layered, Dirac/Weyl-type structural motifs. Two compositional variables showed clear class separation even at the descriptive level: mean

first ionisation potential was markedly higher for Trivial materials (8.86 eV) than for Topological (7.85 eV) materials, with Semimetal intermediate (8.00 eV); and the mean number of constituent elements was highest for Topological compounds (2.80, predominantly ternary) relative to Semimetal (2.22, predominantly binary) and Trivial (2.52) compounds - both patterns that anticipate the feature-importance results of Section 3.3.

Table 2a: Dataset composition and topological sub-type correspondence (n = 1,391).

Class	n	%	Sub-type breakdown
Topological	781	56.1	SEBR = 402; NLC = 379
Trivial	380	27.3	LCEBR = 379; NLC = 1
Semimetal	230	16.5	ESFD = 175; ES = 55
Total	1,391	100.0	—

Table 2b: Crystal-system distribution by topological class (n = 1,391).

Class	Cubic/Tetr./Orthorhombic	Hexagonal/Trigonal	Monoclinic	Triclinic	Total
SM	158 (68.7%)	58 (25.2%)	14 (6.1%)	0 (0.0%)	230
Topological	542 (69.4%)	101 (12.9%)	109 (14.0%)	29 (3.7%)	781
Trivial	234 (61.6%)	60 (15.8%)	75 (19.7%)	11 (2.9%)	380
Total	934	219	198	40	1,391

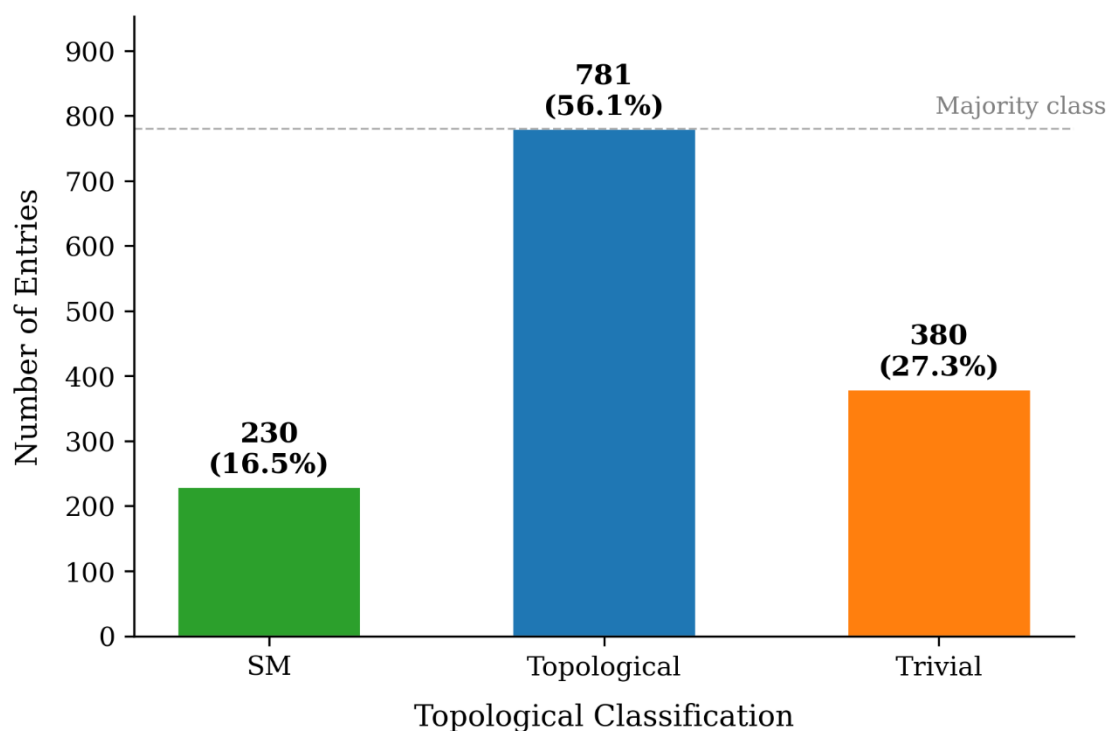


Figure 1: Distribution of topological classification labels (n = 1,391 labelled entries)

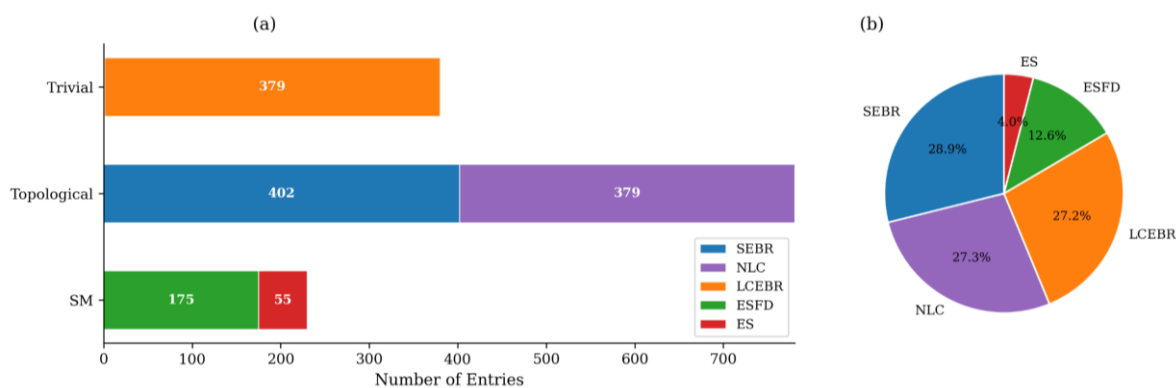


Figure 2: Topological sub-type (TS(Type)) distribution: (a) sub-type breakdown by classification class; (b) overall sub-type proportions

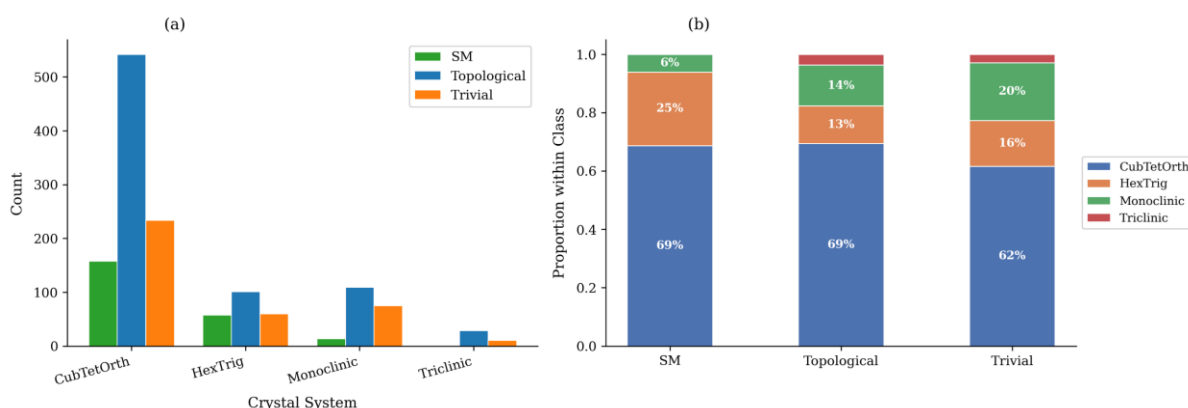


Figure 3: Crystal system distribution by topological classification: (a) counts by class; (b) proportions within class

3.2 Model Performance

Table 3 summarises cross-validation and held-out test-set performance ($n = 279$) for all six models, and Figure 4 presents the same six primary metrics as a comparative radar chart. XGBoost produced the highest observed cross-validation macro-F1 (0.7281 ± 0.0512) and the highest observed test-set macro-F1 (0.7436 , 95% bootstrap CI $[0.6867, 0.7981]$), accuracy (78.14%), MCC (0.6390), macro-AUC-ROC (0.8958), and Cohen's κ (0.636) of any individual model, narrowly ahead of LightGBM on every headline metric (CV macro-F1 0.7245 ± 0.0370 ; test macro-F1 0.7208 $[0.6623, 0.7747]$; accuracy 74.91%; MCC 0.6100; AUC-ROC 0.8868; κ 0.5985). However, McNemar's test with Bonferroni correction across all 15 pairwise comparisons ($\alpha = 0.0033$) found the XGBoost–LightGBM difference statistically indistinguishable ($p = 1.000$), and their test-set macro-F1 bootstrap

confidence intervals overlap substantially (0.6867 – 0.7981 vs. 0.6623 – 0.7747). Only two pairwise comparisons reached significance: XGBoost outperformed SVM ($p = 0.0004$) and outperformed ANN ($p = 0.0029$). Every other pairwise comparison, including RF vs. XGBoost, RF vs. LightGBM, and all comparisons involving the Ensemble, was statistically indistinguishable (Table S4; Figure S1).

Given this statistical equivalence between XGBoost and LightGBM, LightGBM is retained as the primary model for interpretability and screening (Sections 3.3–3.6), on three grounds beyond the point estimates: (i) LightGBM's cross-validation macro-F1 is markedly more stable across folds (± 0.0370 vs. XGBoost's ± 0.0512); (ii) post-hoc isotonic calibration improved LightGBM's expected calibration error (ECE) in all three classes (mean ECE $0.086 \rightarrow 0.061$), whereas the same procedure

improved XGBoost's ECE only for the Semimetal class and slightly worsened it for both Topological and Trivial classes (mean ECE $0.067 \rightarrow 0.066$, a negligible net change) - a materially better-behaved probability output for the tiered, probability-threshold screening applied in Section 3.6; and (iii) LightGBM

achieved the highest macro-AUC-ROC of any model on the Trivial class specifically (0.9350 vs. XGBoost's 0.9331). XGBoost's marginally higher raw metrics are reported honestly here rather than omitted, but do not, on this evidence, constitute grounds to prefer it over LightGBM for the purposes this study requires.

Table 3: Cross-validation and held-out test-set performance for all six models (test $n = 279$)

Model	CV Macro- F1	Test Accuracy	Test Macro- P	Test Macro- R	Test Macro- F1	Test Weighted- F1	MCC	AUC- ROC (macro)	Cohen's κ
RF	0.7182 $\pm$ 0.0423	0.7384 [0.6846, 0.7885]	0.6912	0.7178	0.7001 [0.6397, 0.7601]	0.7449	0.5693 [0.4880, 0.6491]	0.8733 [0.8356, 0.9079]	0.5663
XGB	0.7281 $\pm$ 0.0512	0.7814 [0.7311, 0.8351]	0.7371	0.7603	0.7436 [0.6867, 0.7981]	0.7882	0.6390 [0.5593, 0.7155]	0.8958 [0.8640, 0.9258]	0.6360
LGB	0.7245 $\pm$ 0.0370	0.7491 [0.6953, 0.7957]	0.7100	0.7582	0.7208 [0.6623, 0.7747]	0.7592	0.6100 [0.5325, 0.6846]	0.8868 [0.8544, 0.9158]	0.5985
SVM	0.6838 $\pm$ 0.0359	0.6667 [0.6129, 0.7240]	0.6320	0.6613	0.6294 [0.5707, 0.6815]	0.6787	0.4726 [0.3901, 0.5515]	0.7979 [0.7511, 0.8426]	0.4627
ANN	0.6842 $\pm$ 0.0452	0.6953 [0.6416, 0.7491]	0.6543	0.6816	0.6568 [0.5959, 0.7121]	0.7066	0.5108 [0.4257, 0.5942]	0.8306 [0.7867, 0.8694]	0.5040
Ensemble	—	0.7276 [0.6774, 0.7742]	0.6793	0.7024	0.6829 [0.6219, 0.7388]	0.7380	0.5571 [0.4820, 0.6289]	0.8806 [0.8459, 0.9131]	0.5521

Bracketed values are 95% bootstrap confidence intervals ($n = 1,000$ resamples of the test set). Full hyperparameter search spaces and final selected configurations are given in Supplementary Table S2.

Class-specific performance (Table 4) followed the expected imbalance-driven pattern, with the Topological class achieving the highest F1 across all models, Trivial intermediate, and Semimetal (the smallest class, $n = 46$ in the test set) lowest. LightGBM's class-wise F1 scores were 0.5812 (Semimetal), 0.8014 (Topological), and 0.7799 (Trivial) - a spread of 0.2202 - while SVM showed the widest spread (0.5197 to 0.7586, 0.2389) and XGBoost the narrowest (0.5926 to 0.8411, 0.2485, though from a higher floor). A dedicated SMOTE ablation study, comparing no treatment, class weighting, partial SMOTE (0.6 ratio), and full SMOTE (1.0 ratio) for all five base models, confirmed that full-strength SMOTE yielded the best cross-validation macro-F1 for every model and the

best Semimetal-class F1 for four of five models (Random Forest: 0.5931; XGBoost: 0.6129; LightGBM: 0.6117; ANN: 0.5587 — SVM's Semimetal F1 was marginally lower under full SMOTE, 0.5757, than under the 0.6 ratio, 0.5707, a negligible difference), supporting the imbalance-handling strategy adopted throughout. Error analysis of the confusion matrices (Table 5; Figure 5) showed XGBoost with the fewest total misclassifications (61 of 279, 21.9%), followed by LightGBM (70 of 279, 25.1%) and Random Forest (73 of 279, 26.2%), with SVM performing worst (93 of 279, 33.3%). Topological $\rightarrow$ Semimetal was the single largest error category for every model (25 of RF's 73 errors, 23 of XGBoost's 61, 31 of LightGBM's 70), indicating that the boundary between

Topological and Semimetal character, rather than between either class and Trivial, is where all models struggle most — consistent with these

two classes sharing the ESFD/ES vs. SEBR/NLC sub-type structure of a common underlying band-representation formalism.

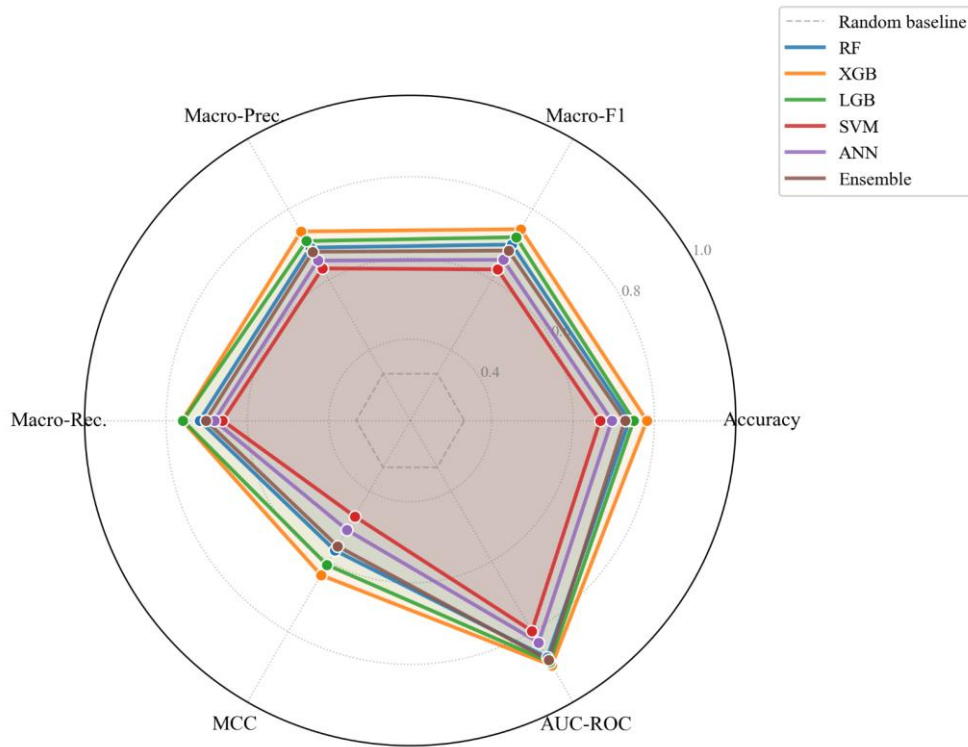


Figure 4: Model comparison radar chart: six primary metrics (macro-F1, macro-precision, accuracy, AUC-ROC, MCC, macro-recall) on the held-out test set ($n = 279$)

Table 4: Class-specific precision, recall, and F1-score, held-out test set ($n = 279$)

Model	Class	Precision	Recall	F1-score	Support
RF	SM	0.4918	0.6522	0.5607	46
RF	Topological	0.8451	0.7643	0.8027	157
RF	Trivial	0.7368	0.7368	0.7368	76
XGB	SM	0.5161	0.6957	0.5926	46
XGB	Topological	0.8759	0.8089	0.8411	157
XGB	Trivial	0.8194	0.7763	0.7973	76
LGB	SM	0.4789	0.7391	0.5812	46
LGB	Topological	0.9040	0.7197	0.8014	157
LGB	Trivial	0.7470	0.8158	0.7799	76
SVM	SM	0.4074	0.7174	0.5197	46
SVM	Topological	0.8271	0.7006	0.7586	157
SVM	Trivial	0.6615	0.5658	0.6099	76
ANN	SM	0.4247	0.6739	0.5210	46
ANN	Topological	0.8382	0.7261	0.7782	157
ANN	Trivial	0.7000	0.6447	0.6712	76
Ensemble	SM	0.4412	0.6522	0.5263	46
Ensemble	Topological	0.8643	0.7707	0.8148	157
Ensemble	Trivial	0.7324	0.6842	0.7075	76

†Training subset (80% of the 1,391-entry dataset); of this, 945 rows formed the base training set used to fit the models, with 167 rows held out for post-hoc calibration (Section 2.6). Comparator sample sizes as reported in their respective publications

Table 5: Full confusion matrices, held-out test set (n = 279)

Model	SM→ SM	SM→ Top	SM→ Triv	Top→ SM	Top→ Top	Top→ Triv	Triv→ SM	Triv→ Top	Triv→ Triv	Accuracy	Total Error
RF	30	8	8	25	120	12	6	14	56	0.7384	73
XGB	32	8	6	23	127	7	7	10	59	0.7814	61
LGB	34	4	8	31	113	13	6	8	62	0.7491	70
SVM	33	5	8	33	110	14	15	18	43	0.6667	93
ANN	31	8	7	29	114	14	13	14	49	0.6953	85
Ensemble	30	6	10	27	121	9	11	13	52	0.7276	76

Row sums confirm true-class support of SM = 46, Topological = 157, Trivial = 76 (total n = 279) for every model.

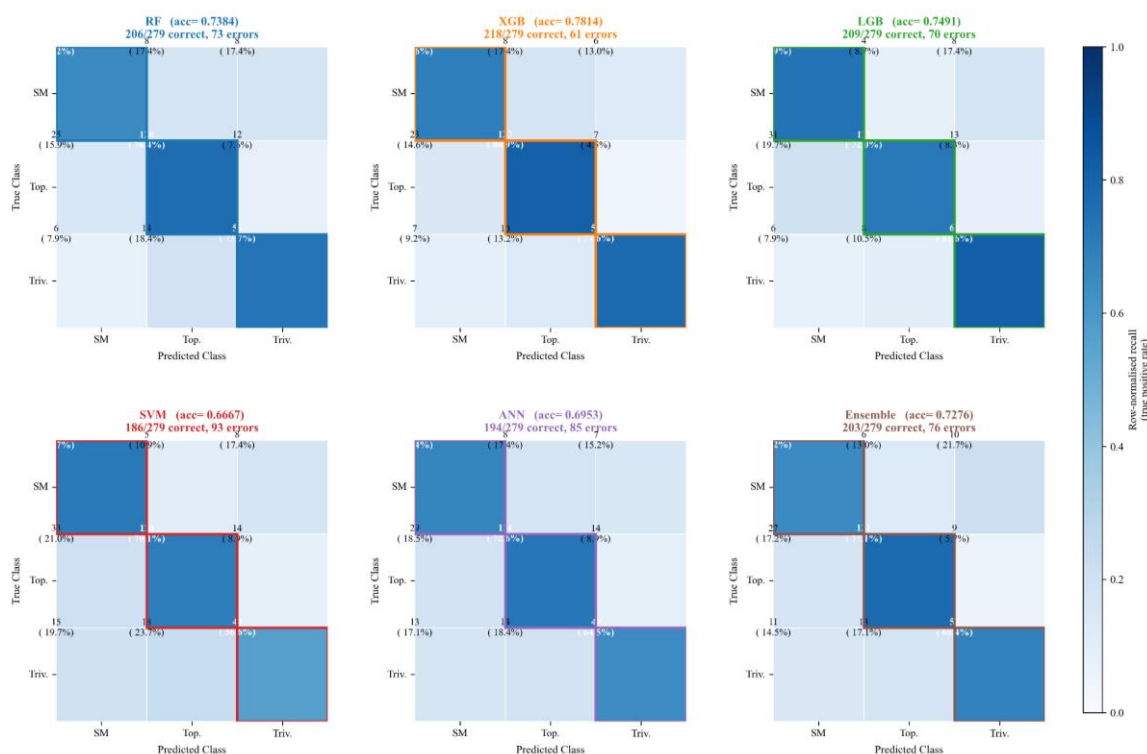


Figure 5: Confusion matrices for all six models on the held-out test set (n = 279), row-normalised recall shading with correct-prediction diagonal highlighted

3.3 Feature Importance and Physical Interpretation

Global SHAP analysis of the LightGBM model (Figure 6a) identified a consistent predictor hierarchy: number of constituent elements (mean $|\text{SHAP}| = 0.355$), mean ionisation potential (0.242), electronegativity range (0.242, effectively tied with ionisation potential), maximum atomic number (0.155), the Z^4 -scaled spin-orbit-coupling proxy (0.122), lattice anisotropy (0.120), and mean atomic radius (0.101). This ranking was stable across all three tree-based models (Figure 6b; Spearman ρ between SHAP rankings of 0.941–0.983 across model pairs), and agreement between SHAP and

permutation feature importance was similarly strong for LightGBM ($\rho = 0.7495$, $p < 0.001$), indicating convergence on genuine signal rather than model-specific artefacts.

The class-specific SHAP decompositions were physically coherent. For the Semimetal class (Figure 7a), low number of constituent elements was overwhelmingly dominant (mean $|\text{SHAP}| = 0.527$), with lattice anisotropy a distant second (0.214): binary compounds present a more restricted set of possible band-crossing configurations at the Fermi level, favouring the accidental or symmetry-enforced degeneracies characteristic of semimetals, whereas additional chemical degrees of freedom in ternary and

higher-order compounds tend to open a gap. For the Topological class (Figure 7b), number of constituent elements again led (0.507, in the opposite direction to its Semimetal effect), followed by electronegativity range (0.385) and maximum atomic number (0.253); high maximum atomic number (heavy elements, strong spin–orbit coupling) pushed predictions toward Topological, consistent with spin–orbit coupling strength scaling approximately as Z^4 and its established contributory (not sole determining) role in band inversion. For the Trivial class (Figure 7c), mean ionisation potential dominated by a wide margin (0.465): high ionisation potential indicates tightly bound, less polarisable valence electrons that favour

conventional, trivially gapped band structures, while lower ionisation potential favours the delocalised, hybridisation-prone electronic structure associated with band inversion. We note that number of constituent elements, the single most important predictor overall, is a compositional-complexity proxy rather than a direct physical determinant of band topology; its predictive value most plausibly reflects the correlation between chemical complexity and the specific chemical families (binary p-block pnictides/chalcogenides vs. ternary/quaternary cerium intermetallics) that dominate this dataset's class composition, rather than a general causal rule.

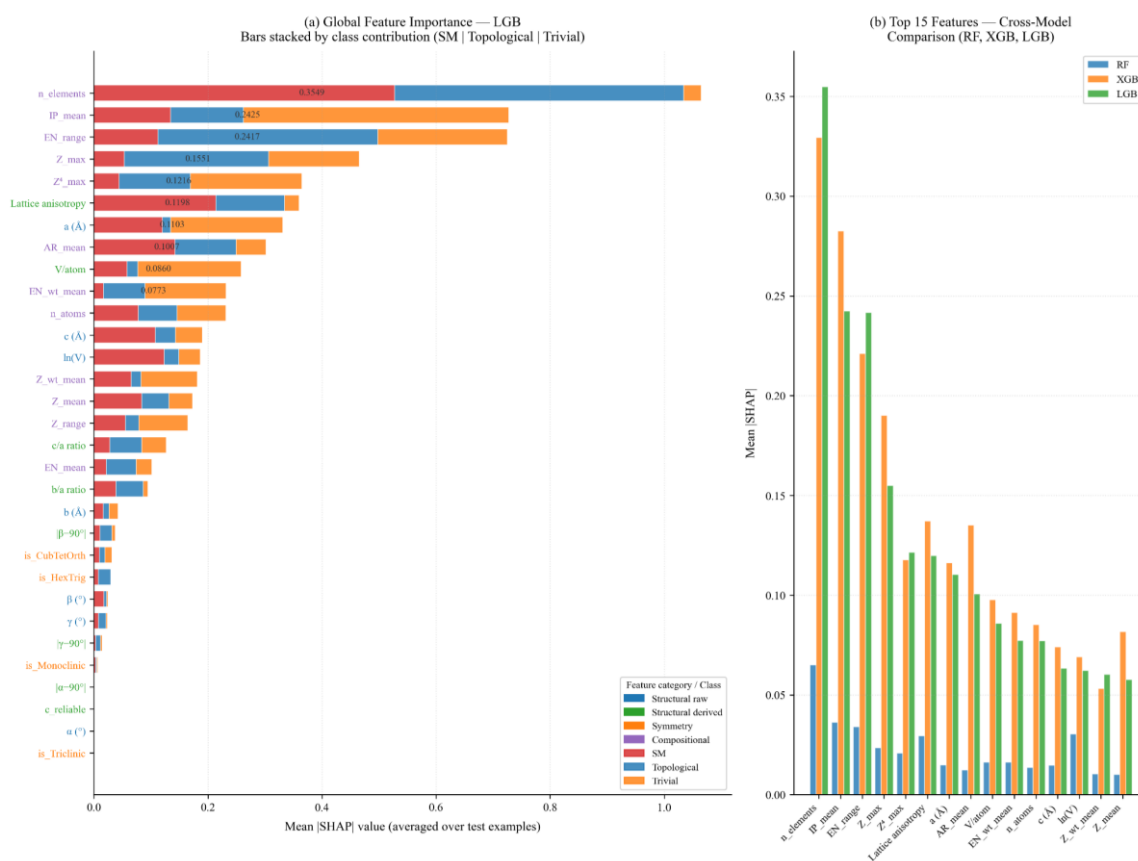
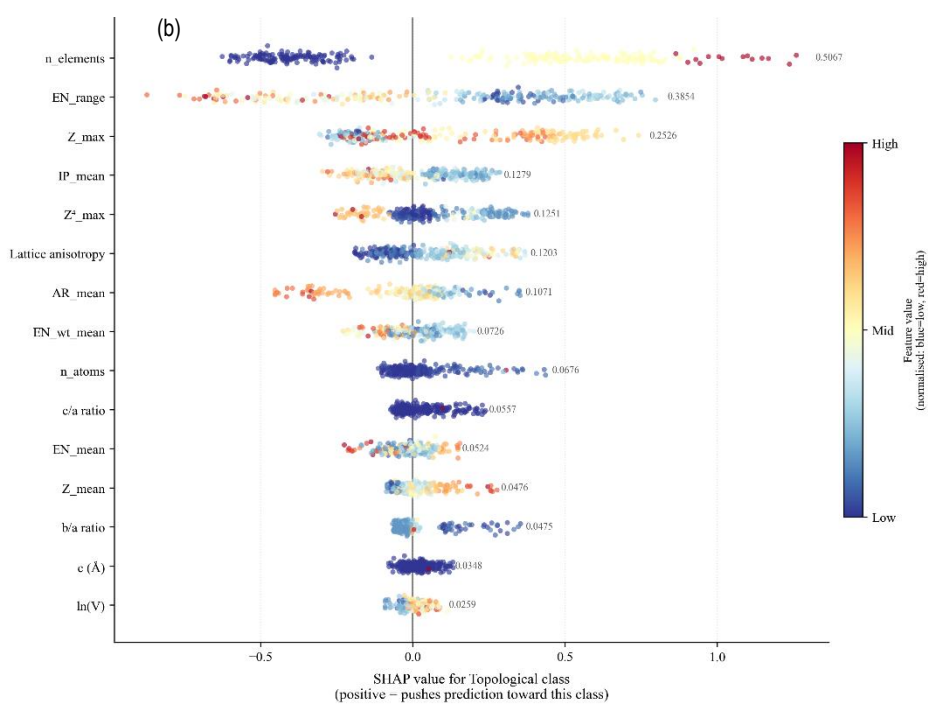
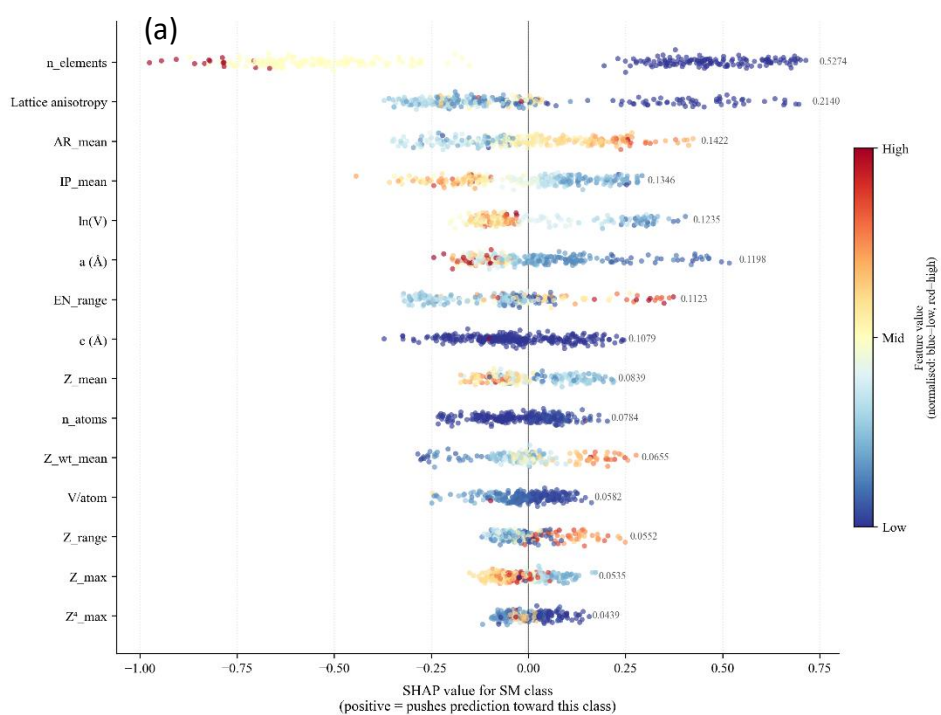


Figure 6: Global SHAP feature importance, LightGBM model: (a) mean |SHAP| value per feature, stacked by per-class contribution; (b) top-15 feature importance comparison across the three tree-based models (Random Forest, XGBoost, LightGBM)



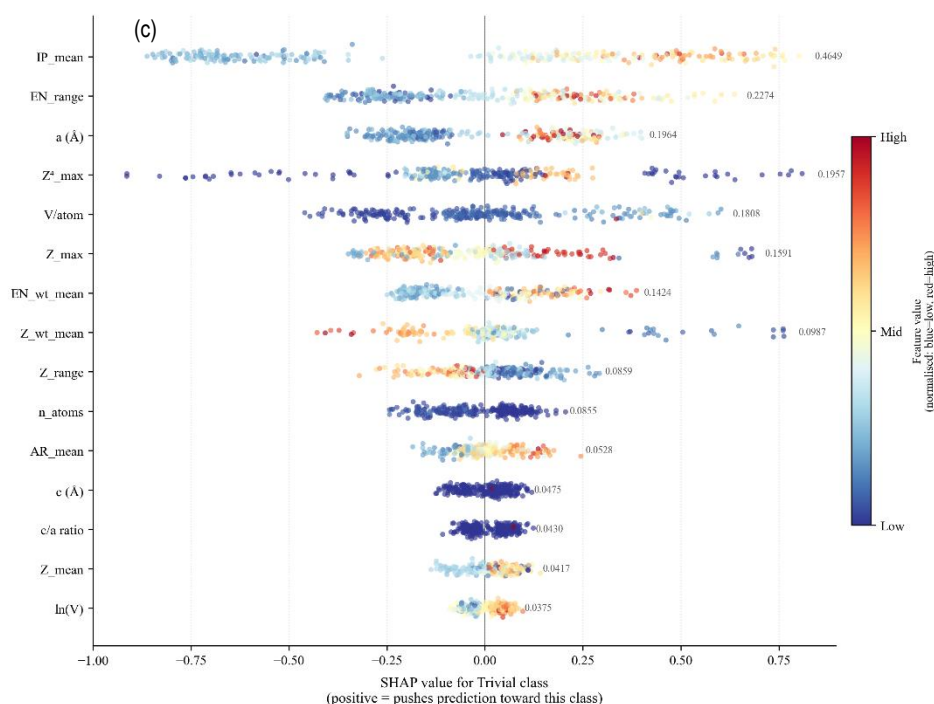


Figure 7: SHAP summary (beeswarm) plots by class, LightGBM model, showing feature value and directional influence on class prediction ($n = 279$ test examples): (a) Semimetal class; (b) Topological class; (c) Trivial class.

SHAP interaction analysis (Table 6) identified electronegativity range $\times$ number of constituent elements as the strongest pairwise interaction overall (mean $|\Phi_{ij}| = 0.054$, rising to 0.087 within the Semimetal class specifically), consistent with a joint mechanism in which binary compounds with a small electronegativity range favour covalent bonding and orbital hybridisation, enabling band inversion; electronegativity range $\times$ mean ionisation potential was the second-strongest interaction (0.035), dominated by its effect within the Trivial class (0.065), consistent with these two features jointly proxying overall bonding character (ionic vs. covalent).

A polymorph-level diagnostic, examining four chemically identical but structurally distinct series ($\text{Bi}_1\text{Mn}_1\text{O}_3$, Sb_2Te_3 , Bi_1I_3 , Ce_1S_2 ; Table 7), provided a stringent internal test of whether the model exploits genuine structural information rather than composition alone. Across the 26 polymorph entries in these four series, LightGBM correctly classified 22

(84.6%), and every misclassification in every series was in the same direction - a true-Topological entry predicted as Trivial; no Trivial entry was ever misclassified as Topological within a polymorph series. Sb_2Te_3 was classified perfectly (8 of 8, mean predicted probability of the true class 0.593), while $\text{Bi}_1\text{Mn}_1\text{O}_3$ was the hardest series (7 of 8 correct, 87.5%): one entry, ICSD 181383, the series' single true-Topological (NLC) member, was misclassified as Trivial with only 10.4% predicted probability assigned to the correct class, despite sharing a near-identical volume ($494.2 \text{ \AA}^3$) with its correctly classified Trivial neighbours. This demonstrates concretely that SHAP-derived feature importance characterises statistical association, not proven physical causation, and that the model's structural signal, while real, is directional in its failure mode — sufficient to protect true-Trivial predictions but occasionally insufficient to distinguish a true-Topological entry from its Trivial polymorph siblings at near-identical composition and volume.

Table 6: Top SHAP interaction pairs, LightGBM model (n = 279 test examples)

Rank	Feature i	Feature j	Mean Φ_{ij} (all)	Mean Φ_{ij} (SM)	Mean Φ_{ij} (Top.)	Mean Φ_{ij} (Triv.)	Physical interpretation
1	EN_range	n_elements	0.0541	0.0726	0.0865	0.0031	Binary compounds with small $\Delta EN \rightarrow$ covalent bonding $\rightarrow$ orbital hybridisation enabling band inversion
2	EN_range	IP_mean	0.0350	0.0054	0.0344	0.0652	Ionisation energy + electronegativity range $\rightarrow$ joint bonding-character proxy
3	Lattice anisotropy	n_elements	0.0205	0.0426	0.0182	0.0006	—
4	EN_wt_mean	EN_range	0.0202	0.0023	0.0407	0.0177	—
5	AR_mean	n_elements	0.0193	0.0410	0.0146	0.0022	—
6	n_elements	IP_mean	0.0186	0.0370	0.0140	0.0048	—
7	AR_mean	IP_mean	0.0160	0.0050	0.0182	0.0249	—
8	Z_max	n_elements	0.0156	0.0088	0.0308	0.0072	SOC-element count vs. elemental complexity $\rightarrow$ binary vs. ternary heavy-element topological classes

Table 7: Polymorph-series classification diagnostic, LightGBM model

Formula	N entries	N Topological	N Trivial	N SM	N Correct	Accuracy	Mean P (true class)	Min P (true class)	Hardest ICSD	Hardest true $\rightarrow$ pred
BiMnO ₃	8	1	7	0	7	87.5%	0.7606	0.1042	181383	Topological $\rightarrow$ Trivial
Sb ₂ Te ₃	8	7	1	0	8	100.0%	0.5925	0.4700	187539	Trivial $\rightarrow$ Trivial
BiI ₃	6	3	3	0	4	66.7%	0.5708	0.2561	187610	Topological $\rightarrow$ Trivial
Ce ₁ S ₂	4	2	2	0	3	75.0%	0.6260	0.3627	419346	Topological $\rightarrow$ Trivial
Aggregate	26	13	13	0	22	84.6%	—	—	—	—

Applied to the full 1,391-entry dataset, LightGBM produced a strongly bimodal predicted-probability distribution (Figure 8a): true-Topological entries clustered near $P(\text{Topological}) \approx 1.0$ and true-Trivial/Semimetal entries near ≈ 0.0 . This separation is equally visible in the stacked count distribution (Figure 8b), where the tier-threshold boundaries cleanly bracket a small population of ambiguous, intermediate-probability entries. Using a three-tier scheme (Tier 1: $P \geq 0.80$; Tier 2: $0.60-0.80$; Tier 3: $0.40-0.60$), 424 entries (30.5% of the dataset) were classified into Tier 1, of which 419 were genuinely Topological — a precision of 98.8% and a recall of 53.7% against all 781 true Topological entries. A precision-at-k analysis (Figure 9) showed precision remaining at 98.8% out to $k = 424$ (the

Tier 1 boundary), declining only modestly to 97.4% by $k = 613$ (the Tier 2 boundary) and 92.7% by $k = 739$ (the Tier 3 boundary), against a baseline prevalence of 56.1%, confirming that the ranking carries reliable discriminative signal well beyond the formal Tier 1 cut-off. Nineteen of the twenty highest-confidence candidates were cerium-containing intermetallics, and the twentieth ($\text{Ag}_2\text{Nd}_1\text{Si}_2$) was a compositionally analogous neodymium intermetallic, reflecting both the strong lanthanide enrichment of the Topological class in this dataset and a physically distinct mechanism, Kondo/heavy-fermion physics involving strongly correlated 4f electrons, from the spin-orbit-coupling-driven band inversion emphasised elsewhere in the analysis.

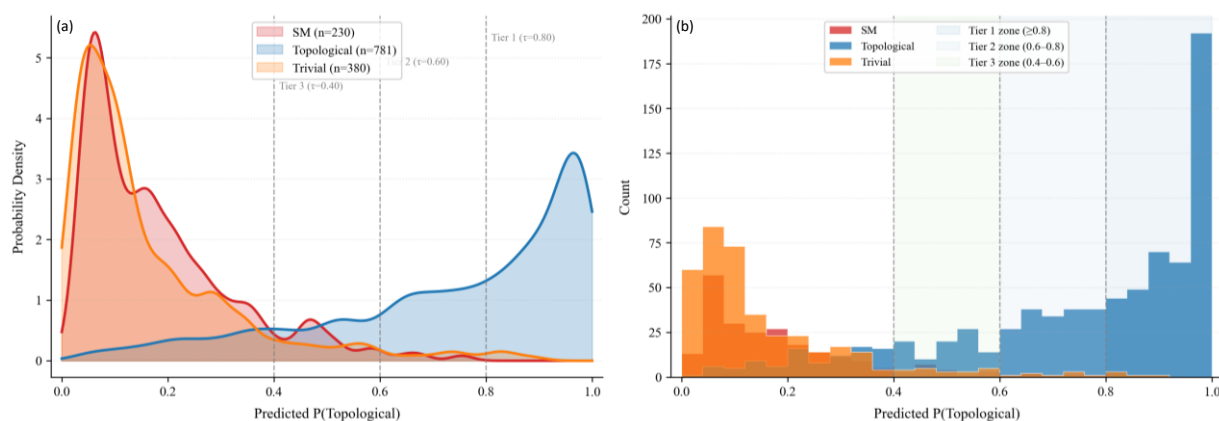


Figure 8: Predicted P(Topological) distribution across the full 1,391-entry dataset, stratified by true class, with tier thresholds marked: (a) kernel density estimate; (b) histogram, stacked by true class

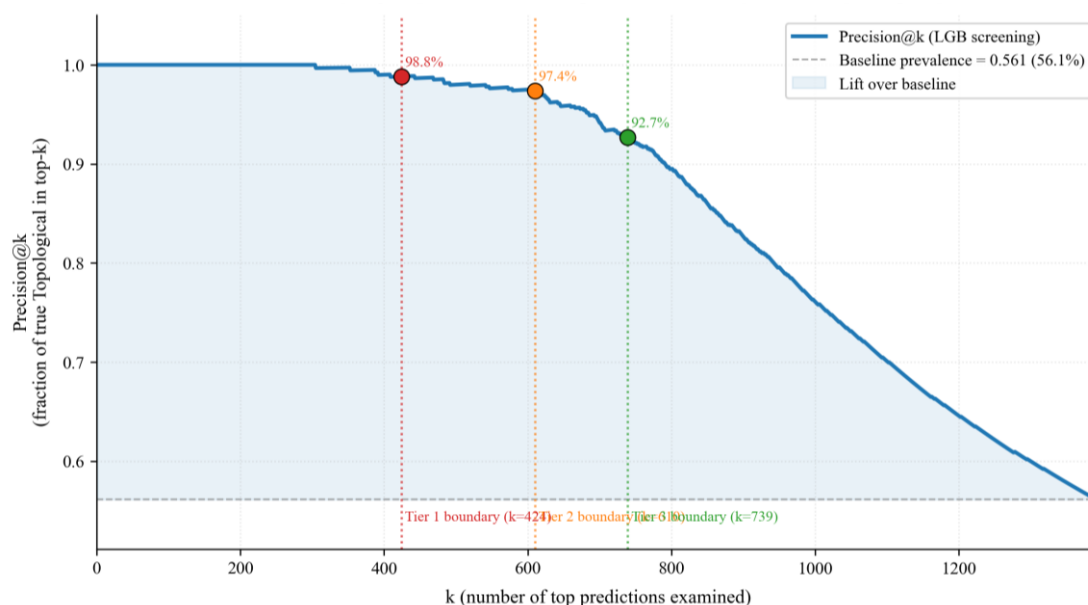


Figure 9: Screening precision-at-k curve: fraction of truly Topological materials within the top-k ranked entries, against a baseline prevalence of 56.1%

3.4 Comparison with Existing Literature

Table 8 situates these results against comparable published studies. The present study's accuracy, using LightGBM (the model retained throughout this manuscript's interpretability and screening analysis), sits below Claussen et al.'s (2020) three-class benchmark (93.1%) by roughly 18.2 percentage points; using XGBoost, the highest raw test accuracy achieved in this study (78.14%, statistically indistinguishable from LightGBM's 74.91% per Section 3.2), the gap narrows to roughly 15.0 percentage points. This gap is well explained by the literature: Claussen et al. report that space-group one-hot encoding alone lifts accuracy from approximately 83% (compositional-only) to over 90%, and the

present results sit close to that compositional-only baseline. The larger gap to the graph-neural-network studies of Louis et al. (2020) and Schleder et al. (2021) likely reflects both their larger training sets and richer, graph-structured or DFT-derived features, unavailable within the pre-DFT computational-economy constraint deliberately adopted here (Section 1). Cao et al. (2020), using an 86-descriptor structural/electronic feature set on 8,924 alloyed-tetradymite entries, reported an AUC-ROC of 0.974 for binary topological/trivial classification, the highest among the comparators, reflecting both a substantially larger, chemically narrower dataset and access to electronic-structure descriptors outside the pre-DFT scope adopted here.

Table 8: Comparison with published machine-learning studies of topological material classification

Study	Training Set Size	Feature Type	Algorithm	Task	Best Result
Sun et al. (2018)	~1,200	Compositional and structural	Random Forest	Binary (TI/trivial)	86.7% accuracy
Cao et al. (2020)	8,924	86-descriptor structural and electronic	Gradient boosting	Binary (TI/trivial)	AUC-ROC 0.974
Claussen et al. (2020)	~1,900	Compositional, structural, and space group	Random Forest	Three-class	93.1% accuracy
Louis et al. (2020)	~3,000	Graph-based (crystal graph convolutional network)	GNN	Binary (TI/trivial)	94.7% accuracy
Schleder et al. (2021)	~2,500	Graph-based, DFT-derived	GNN	Binary (TI/trivial)	95.3% accuracy
Present Study	1,112 (train pool) †	Crystallographic and compositional (no space group)	LightGBM (primary); XGBoost (highest raw metric)	Three-class (SM/TI/Trivial)	74.9% accuracy, macro-AUC-ROC 0.887 (LGB); 78.1% accuracy, macro-AUC-ROC 0.896 (XGB)

†Training subset (80% of the 1,391-entry dataset); of this, 945 rows formed the base training set used to fit the models, with 167 rows held out for post-hoc calibration (Section 2.6). Comparator sample sizes as reported in their respective publications.

Beyond raw performance, this study extends the literature in four respects: explicit, worked documentation of a latent data-entry artefact identified through interpretability-guided cross-validation and corrected via direct database lookup, together with a further, live-API-checked duplicate/conflict-resolution audit (though ultimately unresolved for all 10 conflicting groups - Section 2.2) of the underlying ICSD records; a three-class (semimetal-inclusive) formulation revealing class-specific mechanisms not reported in binary-formulation studies; a more thorough cross-validated interpretability treatment combining SHAP, permutation, and impurity-based importance, cross-checked against a composition-grouped validation scheme; and a directly usable, precision-quantified 424-candidate Tier 1 screening shortlist.

3.5 Limitations

Several limitations qualify these findings. The dataset (1,391 entries) is modest by contemporary materials-informatics standards. The chemical space is concentrated around two dominant families, cerium-based rare-earth compounds and bismuth-based heavy p-block compounds, so generalisation to chemically distinct topological families is less certain; most importantly, the random stratified train/test split does not fully guard against chemically or structurally related entries - including polymorphs and members of the same compound family - occurring in both training and test sets. A composition-grouped cross-validation check found this introduces a modest optimism of +0.0046 (Random Forest) to +0.0145 (LightGBM) macro-F1 relative to random splitting (Supplementary Table S3), confirming the concern raised by both reviewers

is real but small in magnitude on the final, corrected dataset. Additionally, the ICSD-conflict-resolution audit (Section 2.2) could not verify 10 of 1,391 entries (0.7%) against the live TQC database, which returned no match for any of the 10 conflicting groups; these entries retain their first-encountered label as a conservative, unverified fallback. The deliberate exclusion of space-group and symmetry-indicator features, made to preserve the pre-DFT computational-economy motivation of the study, plausibly accounts for most of the accuracy gap relative to comparable studies that include such information (Section 3.4), and is the largest single quantifiable limitation identified. Feature-importance methods describe statistical association rather than proven physical causation, as the $\text{Bi}_1\text{Mn}_1\text{O}_3$ polymorph case illustrates concretely, and the directional (Topological $\rightarrow$ Trivial only) nature of the polymorph misclassifications suggests the model's structural signal, while genuine, is asymmetric rather than fully reliable in either direction. Finally, only two of fifteen pairwise model comparisons reached statistical significance (both involving XGBoost outperforming SVM and ANN specifically), so the apparent ranking among RF, XGBoost, LightGBM, ANN, and the Ensemble in Table 3 should not be read as establishing a reliably superior algorithm; and post-hoc isotonic calibration improved calibration error for LightGBM and Random Forest across all three classes but produced mixed or negligible net improvement for XGBoost and the Ensemble, so raw, uncalibrated LightGBM probabilities were used throughout the screening analysis in Section 3.6.

3.6 Implications for Materials Discovery

The 424-entry Tier 1 candidate list, at 98.8% precision and 53.7% recall against known labels, constitutes a directly usable pre-screening artefact consistent with the multi-fidelity discovery paradigm in which computationally cheap classifiers narrow a candidate pool before

more expensive DFT or experimental follow-up is applied. We emphasise that this is a retrospective ranking/validation exercise against already-known labels, not a prospective discovery result; a genuinely prospective screening would require applying the trained model to unlabelled ICSD entries excluded from training and evaluation, which we identify as a priority direction for future work. More broadly, the results demonstrate that ground-state crystallographic and compositional descriptors, obtainable without density functional theory, spin-orbit coupling treatment, or space-group analysis, carry substantial, physically interpretable predictive signal for topological classification, with most tested algorithm families achieving broadly comparable, statistically indistinguishable performance and only the Support Vector Machine and Artificial Neural Network trailing the strongest model (XGBoost) significantly. This supports the practical value of a lightweight, physics-informed pre-screening step ahead of costlier first-principles or experimental investigation, while also clarifying, via direct comparison with symmetry-aware and graph-based approaches, the specific, quantifiable performance cost of omitting space-group information from the feature set.

4.0 Conclusion

This study asked whether ground-state crystallographic and compositional descriptors, obtainable without density functional theory (DFT) calculation, spin-orbit coupling treatment, or space-group symmetry analysis, could support reliable, interpretable machine learning classification of topological character in inorganic crystalline materials. Using a cleaned, ICSD/TQC-duplicate-checked dataset of 1,391 labelled crystalline entries and 31 engineered features, six classifiers were trained and rigorously evaluated: XGBoost achieved the highest raw test-set performance (macro-F1 of 0.744, accuracy of 78.1%, and macro-AUC-ROC of 0.896), narrowly ahead of LightGBM

(macro-F1 0.721, accuracy 74.9%, macro-AUC-ROC 0.887), though formal significance testing showed the two statistically indistinguishable, along with Random Forest and the soft-voting Ensemble; only the Support Vector Machine and Artificial Neural Network performed significantly worse than XGBoost. LightGBM was retained as the primary model for interpretability and screening on the basis of its lower cross-validation variance and its more consistently improved post-hoc probability calibration, rather than on a raw-metric advantage it did not have.

SHAP, permutation, and impurity-based feature importance analyses converged consistently on a compositionally led predictor hierarchy - number of constituent elements, mean ionisation potential, electronegativity range, maximum atomic number, and mean atomic radius - each traceable to established physical mechanisms of spin-orbit-coupling-driven band inversion, bonding character, and chemical complexity. A polymorph-level diagnostic provided evidence that the trained model can exploit structural information in addition to composition (84.6% accuracy across four polymorph series), although performance remained limited, and directionally biased toward under-predicting Topological character, for some structurally similar cases. Applied to the full dataset, the model generated a 424-entry Tier 1 shortlist of high-confidence topological predictions at 98.8% precision against known labels, a directly usable pre-screening output for prioritising further DFT or experimental investigation.

Beyond the specific performance figures, this work demonstrates that pre-DFT structural and compositional descriptors carry physically plausible and interpretable predictive signal for topological classification. It further shows that rigorous attention to data quality - in this case the identification and correction of a systematic lattice-parameter data-entry artefact via cross-validation against unit-cell volume and manual ICSD lookup, together with a live-database-

checked audit of duplicate and conflicting ICSD entries (with 10 of 1,391 entries, 0.7%, left on an unverified fallback label after the lookup found no match) — materially affects both dataset composition and downstream model conclusions, underscoring that data provenance deserves the same methodological scrutiny as model architecture in materials-informatics research. To be precise, the reported results constitute (i) classification of already-labelled materials, (ii) retrospective validation of that classification against known labels, and (iii) a re-ranking of known materials by predicted confidence; they do not constitute (iv) prospective discovery of previously unlabelled materials, which remains a direction for future work.

Declarations

Funding: This research received no external funding.

AI disclosure: The authors used Claude (Anthropic) solely for language editing and manuscript organization, and to format equations and references according to journal style. The underlying research, data, analyses, figures, and scientific conclusions are entirely the authors' own work. All AI-assisted text was reviewed, verified, and approved by the authors, who take full responsibility for the originality, accuracy, and integrity of the manuscript.

Availability of data and material: The dataset analyzed in this study was derived from the Inorganic Crystal Structure Database (ICSD) and the Topological Quantum Chemistry (TQC) database. The cleaned and processed dataset supporting the findings of this study is available from the corresponding author upon reasonable request.

Code availability: The Python code used for data preprocessing, feature engineering, model development, and evaluation is available from the corresponding author upon reasonable request.

Authors' contributions: Sammy Samuel Dan conceived the study, performed the data curation, preprocessing, feature engineering, model development, and analysis, and drafted the manuscript. Joseph B. Emah supervised the research, contributed to the research design, development of the python codes and interpretation of results, and reviewed and edited the manuscript. Sylvester A. Ekong also supervised the research, and reviewed and edited the manuscript. All authors read and approved the final manuscript.

Ethics approval: Not applicable. This study did not involve human participants, animal subjects, or tissue samples; it used only previously published, de-identified crystallographic data.

Conflict of Interest

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

Acknowledgements

The authors gratefully acknowledge the Department of Physics, Faculty of Physical Sciences, Akwa Ibom State University, for institutional support during the course of this research. The authors also acknowledge the developers and maintainers of the open-source scientific computing tools used in this study, including scikit-learn, XGBoost, LightGBM, PyTorch, SHAP, and pymatgen, and the curators of the Inorganic Crystal Structure Database and the Topological Quantum Chemistry database, whose publicly accessible resources made this research possible.

References

- Bernevig, B. A., Hughes, T. L., & Zhang, S.-C. (2006). Quantum spin Hall effect and topological phase transition in HgTe quantum wells. *Science*, 314, 1757–1761. <https://doi.org/10.1126/science.1133734>
- Bradlyn, B., Elcoro, L., Cano, J., Vergniory, M. G., Wang, Z., Felser, C., Aroyo, M. I., & Bernevig, B. A. (2017). Topological quantum chemistry. *Nature*, 547, 298–305. <https://doi.org/10.1038/nature23268>
- Breiman, L. (2001). Random forests. *Machine Learning*, 45, 5–32. <https://doi.org/10.1023/A:1010933404324>
- Cao, G., Ouyang, R., Barrón-Palos, L., Zhou, J., Fu, Y., Dong, B., & Ma, Y. (2020). Artificial intelligence for high-throughput discovery of topological insulators: The example of alloyed tetradymites. *Physical Review Materials*, 4, 034204. <https://doi.org/10.1103/PhysRevMaterials.4.034204>
- Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16, 321–357. <https://doi.org/10.1613/jair.953>
- Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, (pp. 785–794). <https://doi.org/10.1145/2939672.2939785>
- Claussen, N., Bernevig, B. A., & Regnault, N. (2020). Detection of topological materials with machine learning. *Physical Review B*, 101, 245117. <https://doi.org/10.1103/PhysRevB.101.245117>
- Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20, 273–297. <https://doi.org/10.1007/BF00994018>
- Fu, L., & Kane, C. L. (2007). Topological insulators with inversion symmetry. *Physical Review B*, 76, 045302.

- <https://doi.org/10.1103/PhysRevB.76.045302>
- Fu, L., Kane, C. L., & Mele, E. J. (2007). Topological insulators in three dimensions. *Physical Review Letters*, 98, 106803. <https://doi.org/10.1103/PhysRevLett.98.106803>
- Kane, C. L., & Mele, E. J. (2005). Z_2 topological order and the quantum spin Hall effect. *Physical Review Letters*, 95, 146802. <https://doi.org/10.1103/PhysRevLett.95.146802>
- Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems*, 30, 3146–3154. <https://proceedings-neurips.cc/paper/2017/hash/6449f44a102fde848669bdd9eb6b76fa-Abstract.html>
- König, M., Wiedmann, S., Brüne, C., Roth, A., Buhmann, H., Molenkamp, L. W., Qi, X.-L., & Zhang, S.-C. (2007). Quantum spin Hall insulator state in HgTe quantum wells. *Science*, 318, 766–770. <https://doi.org/10.1126/science.1148047>
- Louis, S.-Y., Zhao, Y., Nasiri, A., Wang, X., Song, Y., Liu, F., & Hu, J. (2020). Graph convolutional neural networks with global attention for improved materials property prediction. *Physical Chemistry Chemical Physics*, 22, 18141–18148. <https://doi.org/10.1039/D0CP01474E>
- Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, 30, 4765–4774.
- Moore, J. E., & Balents, L. (2007). Topological invariants of time-reversal-invariant band structures. *Physical Review B*, 75, 121306. <https://doi.org/10.1103/PhysRevB.75.121306>
- Schleder, G. R., Focassio, B., and Fazzio, A. (2021). Machine learning for materials discovery: Two-dimensional topological insulators. *Applied Physics Reviews*, 8, 031414. <https://doi.org/10.1063/5.0055035>
- Sun, Q., Bhatt, M., & Bhattacharya, S. (2018). Machine learning approaches for the prediction of materials properties. *Computational and Theoretical Chemistry*, 1127, 51–55. <https://doi.org/10.1016/j.comptc.2018.02.009>
- Vergniory, M. G., Elcoro, L., Felser, C., Regnault, N., Bernevig, B. A., & Wang, Z. (2019). A complete catalogue of high-quality topological materials. *Nature*, 566, 480–485. <https://doi.org/10.1038/s41586-019-0954-4>
- Vergniory, M. G., Wieder, B. J., Elcoro, L., Parkin, S. S. P., Felser, C., Bernevig, B. A., & Regnault, N. (2022). All topological bands of all nonmagnetic stoichiometric materials. *Science*, 376, 6595. <https://doi.org/10.1126/science.abg9094>

Supplementary Tables

Table S1: Full feature definitions (31 engineered features)

#	Category	Feature	Symbol / formula
1	Structural (raw)	Cell Length A	a (Å), parsed from ICSD e.s.d. notation
2	Structural (raw)	Cell Length B	b (Å)
3	Structural (raw)	Cell Length C	c (Å)
4	Structural (raw)	Cell Angle α	α (°)
5	Structural (raw)	Cell Angle β	β (°)
6	Structural (raw)	Cell Angle γ	γ (°)
7	Structural (raw)	Unit cell volume (log)	$\ln(V)$, V in Å ³
8	Structural (derived)	Axial ratio	c/a
9	Structural (derived)	Axial ratio	b/a
10	Structural (derived)	Lattice anisotropy	$A_{\text{lat}} = \sigma(a,b,c) / \mu(a,b,c)$
11	Structural (derived)	Alpha deviation	$ \alpha - 90^\circ $
12	Structural (derived)	Beta deviation	$ \beta - 90^\circ $
13	Structural (derived)	Gamma deviation	$ \gamma - 90^\circ $
14	Structural (derived)	Volume per atom	$V / n_{\text{atoms_total}}$
15	Structural (derived)	c-reliability flag	c_{reliable} (legacy indicator; ≈ 0 SHAP/MDI importance — see Section 2.4)
16	Symmetry	Cubic/Tetragonal/Orthorhombic indicator	$is_{\text{CubTetOrth}}$ ($\pm 0.5^\circ$ tolerance)
17	Symmetry	Hexagonal/Trigonal indicator	is_{HexTrig}
18	Symmetry	Monoclinic indicator	$is_{\text{Monoclinic}}$
19	Symmetry	Triclinic indicator	$is_{\text{Triclinic}}$
20	Compositional	Maximum atomic number	Z_{max}
21	Compositional	Mean atomic number	Z_{mean}
22	Compositional	Stoichiometry-weighted atomic number	$Z_{\text{weighted_mean}}$
23	Compositional	Atomic-number range	$Z_{\text{range}} = Z_{\text{max}} - Z_{\text{min}}$
24	Compositional	SOC-strength proxy	$Z^4_{\text{max}} = Z_{\text{max}}^4 / 10^6$
25	Compositional	Mean electronegativity	EN_{mean}
26	Compositional	Stoichiometry-weighted electronegativity	$EN_{\text{weighted_mean}}$
27	Compositional	Electronegativity range	$EN_{\text{range}} = \chi_{\text{max}} - \chi_{\text{min}}$
28	Compositional	Mean atomic radius	AR_{mean}
29	Compositional	Number of constituent elements	n_{elements}
30	Compositional	Total atom count (formula units)	$n_{\text{atoms_total}}$
31	Compositional	Mean first ionisation potential	IP_{mean}

Table S2: Hyperparameter search spaces and final selected configurations

Model	Search space	Search method (budget)	Final selected
RF	n_estimators {100,200,300,500}; max_depth {None,10,20,30}; min_samples_split {2,5,10}; min_samples_leaf {1,2,4}; max_features {'sqrt','log2',0.3,0.5}; class_weight {None,'balanced'}	RandomizedSearchCV (100 × 5-fold)	n_estimators=200, max_depth=20, min_samples_split=10, min_samples_leaf=2, max_features='log2', class_weight='balanced'
XGBoost	n_estimatorsInt(100,500); learning_rate Real(0.01,0.30, log-uniform); max_depthInt(3,9); subsample Real(0.60,1.0); colsample_bytree Real(0.60,1.0); reg_alpha Real(1e-6,2.0, log-uniform); reg_lambda Real(0.50,10.0, log-uniform); min_child_weightInt(1,10)	BayesSearchCV, GP/EI (60 × 5-fold)	n_estimators=201, learning_rate=0.0297, max_depth=7, min_child_weight=5, subsample=1.0, colsample_bytree=0.6, reg_alpha=6.91×10 ⁻⁵ , reg_lambda=0.743
LightGBM	n_estimatorsInt(100,500); learning_rate Real(0.01,0.30, log-uniform); num_leavesInt(15,127); min_child_samplesInt(5,50); subsample Real(0.6,1.0); colsample_bytree Real(0.6,1.0); reg_alpha Real(1e-6,2.0, log-uniform); reg_lambda Real(1e-6,10.0, log-uniform)	BayesSearchCV, GP/EI (60 × 5-fold)	n_estimators=433, learning_rate=0.01, num_leaves=127, min_child_samples=50, subsample=1.0, colsample_bytree=0.6, reg_alpha=2.0, reg_lambda=1×10 ⁻⁶
SVM	RBF branch: C logU(1e-2,1e3), γ logU(1e-4,1.0); Poly branch: C logU(1e-2,1e2), γ logU(1e-2,1e1), degree {2,3}, coef0 {0,0.5,1,2}; class_weight {None,'balanced'} both branches	Custom precomputed-fold search (60 × 5-fold; 7/60 rejected for degenerate collapse)	kernel='poly', C=85.84, degree=2, coef0=1.0, γ =0.0187, class_weight='balanced'
ANN/MLP	hidden_layer_sizes {(64,),(128,),(64,32),(128,64),(128,64,32),(256,128,64)}; activation {'relu','tanh'}; alpha logU(1e-5,1e-1); learning_rate_initlogU(1e-5,1e-2); batch_size {32,64,128}; solver='adam'; max_iter=300; early_stopping=True	RandomizedSearchCV (40 × 5-fold)	hidden_layer_sizes=(128, 64,32), activation='tanh', alpha=0.00605, learning_rate_init=0.000416, batch_size=32

Table S3: Composition-grouped vs. random-split cross-validation (final 1,391-entry dataset)

Model	Random-split CV macro-F1	Composition-grouped CV macro-F1	Δ (optimism)
Random Forest	0.7212 ± 0.0196	0.7166 ± 0.0171	+0.0046
LightGBM	0.7387 ± 0.0180	0.7243 ± 0.0259	+0.0145

5-fold CV using the exact tuned hyperparameters in Table S2. 1,391 total rows, 1,144 unique chemical formulas; 190 formulas (428 rows, 30.8%) recur more than once.

Table S4: McNemar's test, Bonferroni-corrected p-values (15 pairwise comparisons, $\alpha = 0.0033$)

Comparison	Corrected p-value	Significant?
RF vs. XGB	0.4648	n.s.
RF vs. LGB	1.0000	n.s.
RF vs. SVM	0.1668	n.s.
RF vs. ANN	1.0000	n.s.
RF vs. Ensemble	1.0000	n.s.
XGB vs. LGB	1.0000	n.s.
XGB vs. SVM	0.0004	***
XGB vs. ANN	0.0029	**
XGB vs. Ensemble	0.0526	n.s.
LGB vs. SVM	0.0728	n.s.
LGB vs. ANN	0.6825	n.s.
LGB vs. Ensemble	1.0000	n.s.
SVM vs. ANN	1.0000	n.s.
SVM vs. Ensemble	0.1026	n.s.
ANN vs. Ensemble	0.5830	n.s.

*** $p < 0.001$, ** $p < 0.005$ (corrected). Only 2 of 15 comparisons reach significance, both involving XGBoost's advantage over SVM and ANN.

Supplementary figure

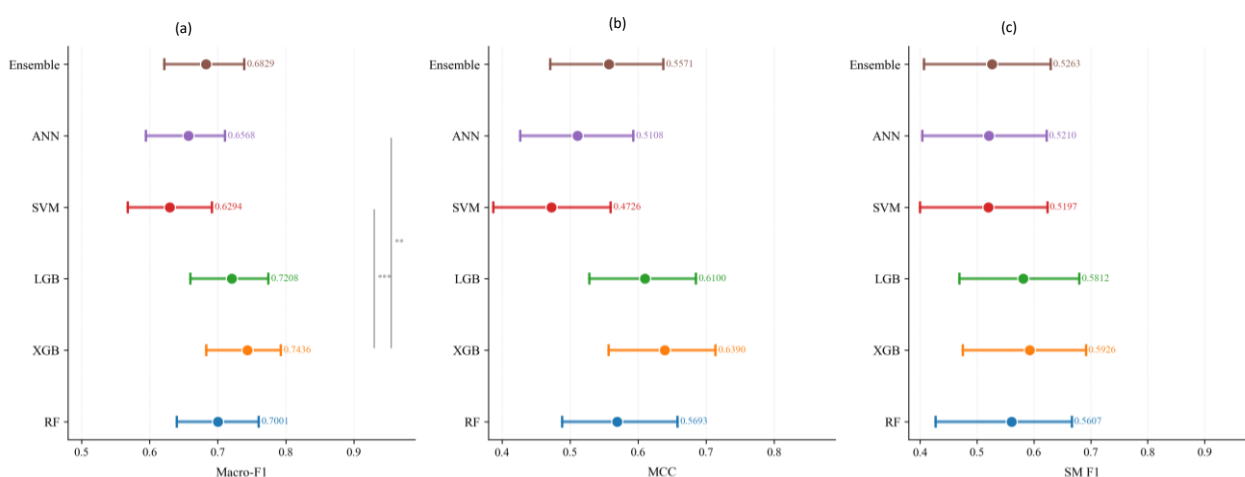


Figure S1: Bootstrap 95% confidence intervals (n = 1,000 resamples) for macro-F1, MCC, and SM-class F1 across all six models, with Bonferroni-corrected significance brackets (Table S4).