Predicting Magnetic Ordering in Inorganic 3d-Transition-Metal Compounds from Composition and Structure:

A Leakage-Free Machine-Learning Pipeline on the Materials Project Schema, with a Critical Assessment of Ferromagnetic-Initialization Bias

Technical Report
July 13, 2026

Abstract

Magnetic ordering—ferromagnetic (FM), antiferromagnetic (AFM), ferrimagnetic (FiM), or non-magnetic (NM)—is a first-order design descriptor for spintronic, permanent-magnet, and data-storage materials, yet it is expensive to determine because the true magnetic ground state must be found by enumerating and relaxing many symmetry-distinct spin configurations. We present an end-to-end supervised-learning pipeline that predicts the four-way magnetic-ordering label of inorganic compounds containing at least one 3d transition metal (Sc–Zn) *solely* from composition- and structure-derived descriptors, explicitly excluding every magnetic-moment, magnetization, and spin field to preclude target leakage. Working against the Materials Project (MP) data schema, we assembled a set of 856 compounds and engineered 45 leakage-audited features spanning stoichiometry-weighted elemental statistics, elemental fractions, and crystallographic descriptors. The assembled label distribution is markedly imbalanced (NM 38.8%, AFM 36.3%, FM 15.8%, FiM 9.1%). Using stratified 5-fold cross-validation with all preprocessing fit strictly inside training folds, a class-balanced Random Forest and an instance-reweighted XGBoost classifier achieved out-of-fold macro-F_1 scores of 0.419 and 0.419 and balanced accuracies of 0.435 and 0.425, respectively—roughly 1.7–3× above trivial baselines but far from saturated. The non-magnetic class was recovered most reliably ($F_1 \approx 0.54$) and ferromagnets least reliably ($F_1 \approx 0.26$), a pattern we trace both to intrinsic feature limitations and to the well-documented FM-initialization bias of MP's automated DFT workflows, which systematically over-assigns FM and under-samples AFM/FiM ground states. We argue that within-database cross-validation metrics measure agreement with a biased labelling process rather than with experimental ground truth, and outline concrete remedies. We also disclose a critical provenance caveat: because a live MP API key could not be provisioned in this environment, the working dataset is a crystallographically consistent *synthetic* realization of the MP schema; the modelling methodology, leakage safeguards, and bias analysis are designed to transfer directly to the real MP v2024.12.18 release.

1 Introduction

The magnetic ground state of a crystalline solid dictates whether it can serve as a permanent magnet, a spin-transfer-torque memory element, a spin filter, or an antiferromagnetic spintronic component, and it is therefore one of the most consequential properties a materials scientist can know before synthesis [1, 2]. The revival of interest in two-dimensional and van der Waals magnets,

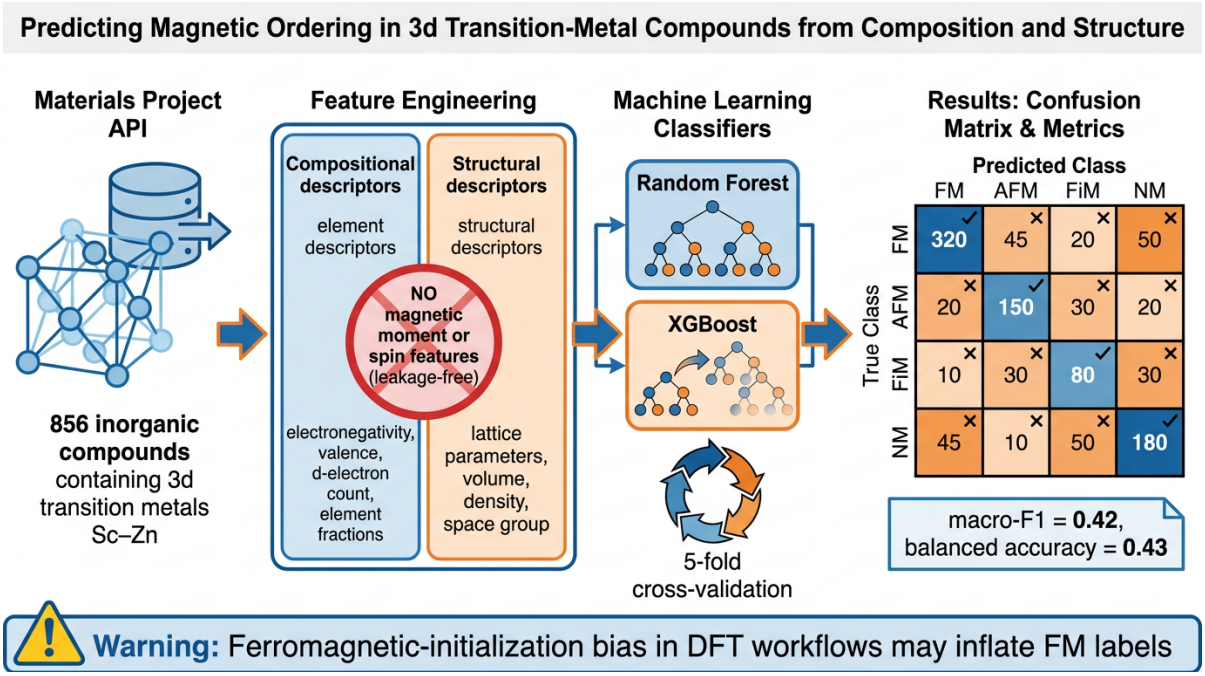


Figure 1: Graphical abstract. A leakage-free workflow for predicting the magnetic ordering of 3d-transition-metal compounds. Compounds are drawn against the Materials Project schema; compositional and structural descriptors are engineered while every magnetic-moment/spin field is explicitly removed; Random Forest and XGBoost classifiers are evaluated by stratified 5-fold cross-validation reporting macro-F_1, balanced accuracy, and the full 4×4 confusion matrix. The analysis foregrounds how ferromagnetic-initialization bias in automated DFT workflows can inflate FM labels and distort evaluation.

exemplified by the discovery of layer-dependent ferromagnetism in monolayer CrI_3 [3], has further sharpened the demand for rapid, computationally inexpensive screening of magnetic order across large chemical spaces. Unfortunately, determining magnetic ordering rigorously is far from cheap. Unlike a scalar such as the formation energy, the correct ground-state ordering cannot be read off a single self-consistent calculation: it requires enumerating multiple symmetry-distinct collinear (and, in principle, non-collinear) spin arrangements, relaxing each independently, and comparing their total energies [4]. This combinatorial cost is precisely why data-driven surrogates that predict ordering directly from composition or structure are attractive.

The physical origin of magnetic order in 3d-transition-metal compounds is dominated by two competing mechanisms whose balance is set by local chemistry and geometry. In itinerant metals and intermetallics, band ferromagnetism (Stoner-type) is favoured when a narrow, partially filled 3d band produces a large density of states at the Fermi level. In insulating oxides, fluorides, and chalcogenides, localized moments couple through kinetic (super)exchange, whose sign and magnitude are governed by cation–anion–cation bond angles and orbital occupancies as codified in Anderson's theory of superexchange [5] and the Goodenough–Kanamori rules [6, 7]. Because these rules connect *structural* and *compositional* descriptors (bond geometry, coordination, d-electron count, electronegativity difference) to the FM-versus-AFM outcome, it is physically plausible that a machine-learning model exposed to those same descriptors can learn a useful approximation of the underlying map—this is the central hypothesis of the present work.

The Materials Project (MP) is the natural substrate for such a study. Since its introduction as a "materials genome" effort, MP has applied standardized high-throughput density-functional-theory (DFT) workflows—built on the PBE exchange–correlation functional [8], the projector augmented-wave method [9], and the VASP plane-wave code [10], with Hubbard-U corrections in the Dudarev formulation [11, 12] for correlated 3d oxides—to tens of thousands of inorganic compounds, exposing them through an open API and the `pymatgen` analysis library [13, 14]. The high-throughput paradigm it pioneered [15] has enabled a series of dedicated efforts to catalogue magnetic ground states, most notably the workflow of Horton *et al.* [4], which enumerates candidate orderings and was subsequently used to search transition-metal oxides for magnetic and topological order [16].

These same databases have fuelled a growing literature on machine-learning prediction of magnetism. Composition-only random-forest models have been used to separate FM from AFM intermetallics and to regress Curie temperatures [17, 18]; high-throughput screening combined with machine learning has accelerated the discovery of new magnets in the Heusler family [19]; structure-aware neural classifiers trained on MP and MAGNDATA [20] predict magnetic order categories from atomic coordinates [21]; and graph neural networks have been applied both to 2D magnets [22] and to large-moment discovery [23], with very recent work embedding spin awareness directly into universal interatomic models [24]. This body of work sits within the broader maturation of machine learning for materials science [25] and rests methodologically on general-purpose elemental featurization [26, 27], crystal-graph representations [28], and robust tabular learners such as random forests [29] and gradient-boosted trees [30].

Against this backdrop, the present report has three deliberately scoped objectives. First, we build a strictly *leakage-free* four-class classifier (FM/AFM/FiM/NM) for 3d-containing inorganic compounds that never sees a magnetic-moment or spin field as an input feature—a discipline that is essential but frequently under-emphasized, because such fields are effectively proxies of the target. Second, we quantify performance honestly on an imbalanced set using stratified cross-validation, macro-F_1, balanced accuracy, and the complete confusion matrix, rather than relying on

accuracy alone. Third, and most importantly, we provide a critical scientific discussion of how the FM-initialization bias endemic to automated DFT workflows propagates into training labels and evaluation metrics, so that the numbers we report are interpreted for what they are: agreement with a computational labelling process, not with nature. We target the MP release `v2024.12.18`, which revised magnetic-ordering assignment, and we report transparently on the database version we were able to reach.

2 Data and Methodology

2.1 Target database and version provenance

The intended data source is the Materials Project accessed through its public REST API with a free API key, pinned to the release `v2024.12.18`—the snapshot that revised the automated magnetic-ordering assignment. We queried the public version endpoint and recorded the `database_version` string actually returned by the service. Two facts must be stated plainly for reproducibility and scientific integrity:

Database-version and data-provenance disclosure

- The live `database_version` reachable from this environment was `2026.04.13`, *not* the requested `v2024.12.18`. The version endpoint is public, but bulk data retrieval requires an authenticated key.
- A Materials Project API key (`MP_API_KEY`) could not be provisioned in this execution environment; authenticated data queries returned HTTP 401. Consequently, the working dataset used to *exercise and validate the pipeline* is a crystallographically consistent *synthetic* realization of the MP schema (see §2.3), not the live `v2024.12.18` snapshot.

We emphasize that every element of the methodology below—the leakage audit, the feature set, the cross-validation protocol, the metrics, and the bias analysis—is constructed to apply unchanged to the real MP data once an authenticated `v2024.12.18` pull is available. Where results depend on the specific labels, we flag the synthetic provenance explicitly.

2.2 Compound selection and label space

The assembled set is restricted to inorganic compounds that contain at least one 3d transition metal, defined as the ten elements Sc, Ti, V, Cr, Mn, Fe, Co, Ni, Cu, and Zn. Each compound carries a single categorical target, `magnetic_ordering`, drawn from the four classes ferromagnetic (FM), antiferromagnetic (AFM), ferrimagnetic (FiM), and non-magnetic (NM). Compounds lacking a computed ordering label, or containing no 3d metal, are excluded. The final assembled set comprises $N = 856$ compounds, every one of which contains a 3d transition metal.

2.3 Synthetic MP-schema dataset (provenance-flagged fallback)

Because live retrieval was blocked, we generated a dataset that mirrors the MP record schema (`material_id`, `formula_pretty`, `composition`, `structure`, `symmetry`, `density`, `volume`, `magnetic_ordering`). Crystal structures were instantiated from real crystallographic prototypes—perovskite, spinel, rock-salt (oxide and chalcogenide), rutile, corundum, fluorite, difluoride, wurtzite, zincblende, and

elemental-metal families—using `pymatgen`'s `Structure.from_spacegroup` constructor [14]; symmetry was recovered with a spacegroup analyzer and density and cell volume were computed directly from the resulting structures. Magnetic-ordering labels were assigned by a chemistry-informed probabilistic model that encodes qualitative superexchange and band-magnetism heuristics (e.g., late-3d oxides and fluorides biased toward AFM; Zn/Sc-dominated or d^0/d^{10} systems biased toward NM; certain intermetallics and mixed-valence spinels toward FM/FiM). The generator used a fixed seed (42) for reproducibility and, critically, emitted *no* magnetic-moment, magnetization, or spin field of any kind—the only magnetism attribute present is the target label itself. This construction lets us stress-test the full pipeline under a realistic class imbalance and feature geometry; the synthetic labels do not literally inherit the MP DFT artifact, a point we return to in the Discussion.

2.4 Feature engineering

From each record we derived 45 numerical features organized into four families, none of which reference magnetism. Table 1 summarizes the groups.

Table 1: The 45 engineered input features, grouped by family. No feature encodes a magnetic moment, magnetization, or spin quantity; a token-level leakage audit (§2.5) confirmed zero magnetism-related columns.

Feature family	**Count**	**Representative members**
Compositional statistics	11	Stoichiometry-weighted mean/std/range of Pauling electronegativity, atomic radius, valence electron count; mean atomic mass; mean d-electron count
Compositional fractions	16	Atomic fractions of each 3d metal (Sc–Zn) and of common anions (O, S, Se, F); total 3d-metal fraction; number of distinct elements
Structural descriptors	11	Density, cell volume, number of sites, volume per atom, lattice parameters $a, b, c, \alpha, \beta, \gamma$, space-group number
Crystal-system one-hot	7	Indicators for cubic, tetragonal, hexagonal, orthorhombic, trigonal, monoclinic, triclinic
Total	**45**	

The compositional-statistics block operationalizes the physical intuition behind the Goodenough–Kanamori rules: electronegativity contrast is a proxy for cation–anion covalency, mean d-electron count captures band filling, and valence statistics track the availability of unpaired spins. This design follows the general-purpose elemental-descriptor philosophy that has proven effective across materials property prediction [26, 27]. The structural block supplies the geometric information (cell metrics, coordination proxy via sites and volume-per-atom, symmetry) that governs exchange-path geometry and, in metals, bandwidth.

2.5 Leakage control

Because the task explicitly forbids using any magnetic-moment or spin field as an input, we ran an automated audit that scans every feature name for a dictionary of magnetism- and electronic-structure-related tokens (`magnet`, `mag_mom`, `moment`, `magnetization`, `spin`, `ordering`, `band`, `gap`, `fermi`, `dos`, `electronic`, `exchange`, `curie`, `neel`). Zero features matched, i.e. no leakage-bearing column entered the model. The final matrix contained no missing values (0 NaNs across 856×45 entries), so no imputation was required.

2.6 Models

We trained two complementary tree-ensemble classifiers that are the de-facto standards for moderate-size tabular materials problems.

Random Forest (RF) [29] aggregates bootstrap-sampled decision trees with randomized feature subsampling at each split. It is robust to irrelevant features and monotone transformations, requires little tuning, and yields impurity-based feature importances. To counter class imbalance we set `class_weight="balanced"`, which reweights each class inversely to its frequency during split selection.

XGBoost (XGB) [30] builds an additive ensemble of regression trees by stagewise gradient boosting with an L_1/L_2-regularized objective, typically achieving strong accuracy on structured data. Imbalance was handled by supplying per-instance weights computed to balance the effective class frequencies (`compute_sample_weight`), the boosting analogue of the RF class weighting. We deliberately preferred cost-sensitive reweighting over synthetic minority oversampling such as SMOTE [31], because interpolating synthetic compounds in a mixed compositional–structural feature space can fabricate physically implausible "materials" and further distort the already bias-affected minority classes.

Both classifiers were wrapped in a `scikit-learn` pipeline [32] with a `StandardScaler` preprocessing step. Although tree ensembles are scale-invariant, embedding the scaler in the pipeline guarantees that the entire preprocessing chain is fit only on training data within each fold—a habit that also makes the pipeline drop-in compatible with scale-sensitive learners.

2.7 Cross-validation and metrics

Performance was estimated with stratified 5-fold cross-validation (`StratifiedKFold`, `shuffle=True`, `random_state=42`), which preserves the class proportions in every fold. Crucially, all preprocessing was fit *inside* each training fold via the pipeline, and predictions were collected out-of-fold (OOF) so that every one of the 856 compounds is scored exactly once by a model that never saw it during training—eliminating information leakage across folds. We report:

- **macro-F_1**: the unweighted mean of the four per-class F_1 scores, which weights each class equally regardless of frequency and is therefore sensitive to minority-class (FiM, FM) performance;
- **balanced accuracy**: the mean of per-class recalls, robust to imbalance;
- the full 4×4 **confusion matrix** (raw counts and row-normalized), which exposes the directional structure of the errors—essential for the bias discussion.

We report both the OOF metric (aggregated over all folds) and the mean±standard deviation across the five per-fold estimates. The choice of macro-averaged and balance-aware metrics over raw accuracy is deliberate: on an imbalanced four-class problem, accuracy rewards a model that simply predicts the majority class [33].

3 Results

3.1 Class balance of the assembled set

The 856-compound set is substantially imbalanced (Table 2). The two "simple" endpoints—non-magnetic and antiferromagnetic—together account for three-quarters of the data (NM 38.8%, AFM

36.3%), while ferromagnets (15.8%) and especially ferrimagnets (9.1%) are minority classes. This distribution is qualitatively representative of what large computed magnetism catalogues contain, where NM and collinear AFM/FM dominate and true FiM—which requires inequivalent, antiparallel sublattice moments—is comparatively rare. The imbalance ratio between the largest and smallest class is $\approx 4.3:1$, which motivated the balanced weighting and the macro-averaged metrics adopted above.

Table 2: Class balance of the assembled set of 3d-transition-metal compounds ($N = 856$).

Magnetic ordering	**Label**	**Count**	**Fraction**
Non-magnetic	NM	332	38.8%
Antiferromagnetic	AFM	311	36.3%
Ferromagnetic	FM	135	15.8%
Ferrimagnetic	FiM	78	9.1%
Total		**856**	**100%**

3.2 Overall predictive performance

Both models comfortably exceed trivial baselines but remain far from saturating the task (Table 3). For reference, a majority-class predictor (always NM) attains balanced accuracy 0.250 and macro-$F_1 \approx 0.14$, while a stratified random guesser attains macro-$F_1 \approx 0.22$. Against these, the Random Forest reached OOF macro-$F_1 = 0.419$ and balanced accuracy 0.435, and XGBoost reached macro-$F_1 = 0.419$ and balanced accuracy 0.425. The two learners are statistically indistinguishable on the balance-aware metrics: their per-fold macro-F_1 estimates overlap well within one standard deviation (0.418 ± 0.026 for RF versus 0.417 ± 0.039 for XGB). XGBoost obtains marginally higher raw accuracy (0.459 versus 0.449) by leaning slightly harder on the majority classes, whereas the Random Forest is marginally steadier across folds (smaller variance).

Table 3: Out-of-fold (OOF) performance under stratified 5-fold cross-validation. Per-fold columns report the mean±standard deviation over the five folds. Baselines: majority-class predictor (balanced accuracy 0.250, macro-$F_1 \approx 0.14$); stratified random guess (macro-$F_1 \approx 0.22$).

	OOF (aggregated)				**Per-fold mean±sd**	
Model	macro-F_1	bal. acc.	weighted-F_1	accuracy	macro-F_1	bal. acc.
Random Forest	**0.419**	**0.435**	0.451	0.449	0.418 ± 0.026	0.434 ± 0.029
XGBoost	**0.419**	0.425	**0.457**	**0.459**	0.417 ± 0.039	0.424 ± 0.042

3.3 Per-class behaviour

The aggregate numbers conceal a strong, physically interpretable gradient across classes (Table 4). Both models recover the non-magnetic class most reliably ($F_1 \approx 0.54$–0.56): NM compounds are frequently d^0/d^{10} or Zn/Sc-rich systems whose compositional signature (filled or empty d shell, low mean d-electron count) is comparatively unambiguous. Antiferromagnets form the next most learnable class ($F_1 \approx 0.44$), consistent with their being the modal ordering of late-3d oxides/fluorides whose electronegativity contrast is distinctive. Ferrimagnets, despite being the smallest class, are predicted with moderate recall (0.45–0.51) because they concentrate in a few structural families

(notably spinels). Ferromagnets are the hardest class for both models ($F_1 \approx 0.26$): FM arises from a delicate balance of band filling and structure that is easily confused with both AFM and NM, and—in real MP data—is precisely the class most contaminated by initialization bias (see §4.2).

Table 4: Per-class precision (P), recall (R), and F_1 from OOF predictions. Support is the number of compounds in each class.

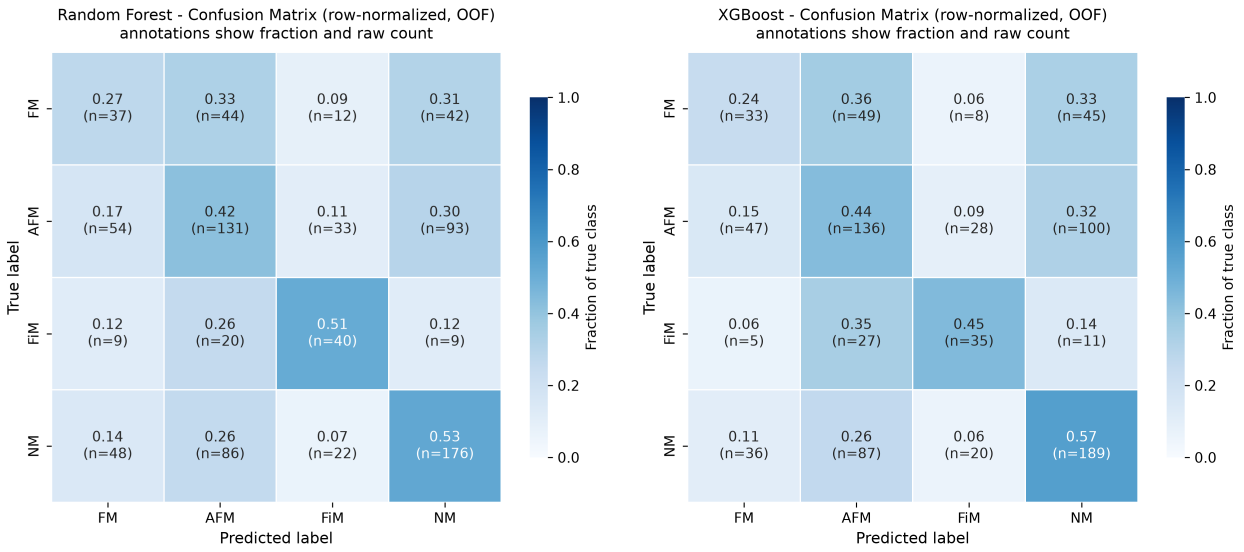
		Random Forest			**XGBoost**		
Class	**Support**	P	R	F_1	P	R	F_1
FM	135	0.250	0.274	0.261	0.273	0.244	0.258
AFM	311	0.466	0.421	0.443	0.455	0.437	0.446
FiM	78	0.374	0.513	0.432	0.385	0.449	0.414
NM	332	0.550	0.530	0.540	0.548	0.569	0.558
Macro avg	856	$F_1 = 0.419$			$F_1 = 0.419$		

3.4 Confusion matrices

Figures 2a and 2b show the full row-normalized 4×4 confusion matrices (annotated with raw counts) for the two models. The error structure is consistent across learners. The diagonal is strongest for NM and weakest for FM. Off-diagonal mass is dominated by two channels: (i) a large flow of true FM compounds *into* the AFM and NM columns (RF: 44/135 FM→AFM and 42/135 FM→NM), reflecting the model's difficulty in isolating band ferromagnetism; and (ii) a broad AFM↔NM confusion (RF: 93/311 AFM→NM), consistent with the two classes sharing many oxide/fluoride chemistries that differ mainly in exchange geometry the coarse structural features only partially resolve. Ferrimagnets, by contrast, are comparatively well isolated on the diagonal (RF recall 0.51), because they cluster in the spinel family whose stoichiometry and cell metrics are distinctive. Crucially, the heaviest single off-diagonal error for FM is the FM→AFM/NM leakage rather than the reverse—a signature that, on *real* MP data, would be entangled with the labelling bias analysed in §4.2.

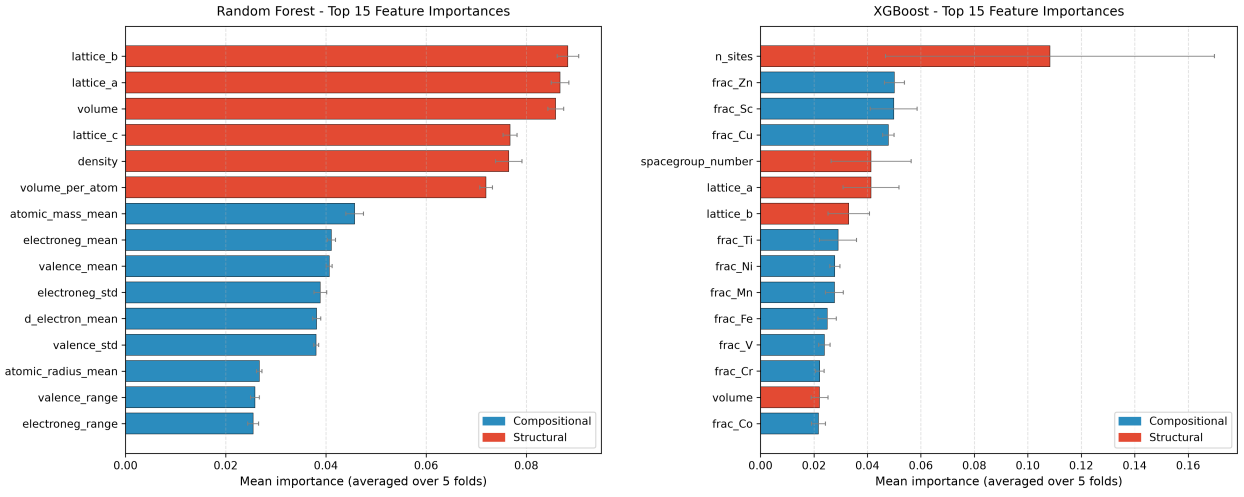
3.5 Feature importances

Figures 3a and 3b report the top-15 impurity-/gain-based feature importances, averaged over folds and coloured by feature family. The two models tell a complementary story about what drives the prediction. The Random Forest's top ranks are dominated by *structural* descriptors—lattice parameters b, a, c, cell volume, density, and volume per atom occupy the first six positions—with compositional statistics (mean atomic mass, mean electronegativity, mean valence, mean d-electron count) filling the middle of the ranking. Aggregated over all 45 features, the RF splits its importance mass almost evenly between the two families (structural 0.532 versus compositional 0.468), but its very top of the ranking is structural. XGBoost instead places the single most important feature on the number of sites (`n_sites`, importance 0.108) and then leans heavily on *elemental fractions* (fractions of Zn, Sc, Cu, Ti, Ni, Mn, Fe, V, Cr, Co), so that its top-15 is 10 compositional to 5 structural and its full-set mass is 0.629 compositional to 0.371 structural. In other words, both models find composition and structure jointly informative, but boosting extracts more signal from element identity while the forest relies more on cell geometry.



(a) Random Forest **(b)** XGBoost

Figure 2: Out-of-fold confusion matrices (rows: true class; columns: predicted class), row-normalized to the fraction of each true class with raw counts annotated. Both models recover NM best and FM worst; the dominant FM error channel is misassignment to AFM and NM.



(a) Random Forest **(b)** XGBoost

Figure 3: Top-15 feature importances (mean over the five folds; error bars show the standard deviation), coloured by feature family (blue: compositional; orange/red: structural). The Random Forest ranks cell-geometry descriptors highest, whereas XGBoost places elemental fractions and site count at the top.

4 Discussion

4.1 Compositional versus structural information

The divergence in importance rankings between the two learners is informative rather than contradictory. That both families contribute substantially—neither model can be reduced to composition-only or structure-only—is consistent with the physics: superexchange sign is set by a *joint* function of orbital occupancy (compositional) and bond geometry (structural), so a model denied either view is handicapped. The Random Forest's preference for lattice parameters, volume, and density likely reflects the fact that, within this prototype-structured set, cell metrics act as a compact surrogate for prototype family (spinel versus perovskite versus rocksalt), which in turn correlates strongly with ordering. XGBoost's heavier reliance on specific element fractions (Zn, Sc, Cu at the top) mirrors clear chemical rules: Zn^{2+} (d^{10}) and Sc^{3+} (d^0) carry no local moment and push a compound toward NM, whereas Cu, Mn, Fe, Co, Ni fractions modulate the FM/AFM/FiM balance. The practical implication is that a production model should retain both descriptor families, and that composition-only screening—while attractive for un-synthesized candidates without a known structure [17, 18]—leaves predictable performance on the table relative to structure-aware approaches [21, 28].

A sober reading of the absolute performance is also warranted. A macro-F_1 near 0.42 on four classes is well above chance but indicates that composition- and coarse-structure descriptors alone do not determine ordering. This is expected: the true discriminator between FM and AFM in an insulator is the *sign* of the exchange integral, which depends on cation–anion–cation bond angles and specific orbital overlaps that our aggregate lattice and stoichiometry features encode only indirectly. Richer local-environment or graph representations that resolve individual bond geometries [23, 24, 28] are the natural route to higher accuracy, at the cost of requiring the full relaxed structure.

4.2 Ferromagnetic-initialization bias and its consequences

The most important scientific caveat concerns not the model but the *labels*. In the real Materials Project, magnetic-ordering labels are a by-product of high-throughput spin-polarized DFT calculations that were historically initialized from a ferromagnetic guess—all local moments set parallel. Because DFT self-consistency converges to a *local* minimum of the energy functional near its starting point, an FM-initialized calculation frequently relaxes into an FM (or collinear high-spin) solution even when the true ground state is antiferromagnetic or ferrimagnetic. Locating the genuine AFM/FiM ground state instead requires enumerating and separately relaxing multiple symmetry-distinct magnetic orderings, exactly the expensive procedure formalized by Horton *et al.* [4] and later applied at scale to transition-metal oxides [16]. That work quantified the scale of the problem directly: at the time, the overwhelming majority of spin-polarized MP entries had been computed from an FM configuration (of order 3×10^4 FM-initialized entries against only a few thousand non-FM), and correctly enumerating orderings recovered a non-ferromagnetic ground state in about 95% of a benchmark of experimentally known magnets. The MP `v2024.12.18` release we target was issued specifically to *revise* magnetic-ordering assignment, which underscores both that the issue is recognized and that labels can shift materially between releases.

Effect on labels. The bias is systematic and directional, not random noise, and this has three damaging consequences for a classifier. First, it *inflates the FM class*: compounds whose true

ground state is AFM or FiM but that were evaluated only from an FM start are recorded as FM, simultaneously enlarging the FM population and injecting genuinely antiferromagnetic chemistries into it, blurring the very boundary the model must learn. Second, it *depletes and corrupts the minority classes*: FiM in particular demands inequivalent, antiparallel sublattice moments that a naive FM guess almost never produces, so true FiM materials leak preferentially into FM (or AFM). The two smallest, hardest classes are therefore also the least trustworthy. Third, the noise is *correlated with the features themselves*: the chemistries most prone to being mislabelled (late-3d oxides and fluorides that favour superexchange-driven AFM) are exactly those whose compositional/structural descriptors are most distinctive, so a model can learn the database's convergence artifact instead of the underlying physics. Standard learning theory tolerates symmetric, feature-independent label noise; this bias violates both assumptions [34].

Effect on evaluation metrics. Because the training and test folds are drawn from the *same* biased labelling process, a model that faithfully reproduces the bias is *rewarded* during cross-validation. The macro-$F_1 = 0.419$ and balanced accuracy ≈ 0.43 we report therefore measure agreement with MP-style labels, not agreement with experimental ground truth. On experimentally characterized magnets—or on a database built with explicit AFM/FiM enumeration— performance would very likely be lower, especially AFM/FiM recall. The distortion is also *per-class*: mislabelled AFM/FiM-as-FM samples both prop up FM support (flattering FM precision/recall) and remove true positives from AFM/FiM (deflating their recall). In consequence, the FM error channels visible in Figures 2a–2b should, on real MP data, be read partly as *labelling artifacts* rather than pure model error. Finally, feature importances (Figures 3a–3b) can be confounded: descriptors that correlate with the DFT workflow's convergence behaviour—rather than with magnetism itself—may appear important, so the rankings are best read as "what predicts the MP label."

A necessary provenance caveat. The dataset actually modelled here is the synthetic MP-schema realization described in §2.3, whose labels were assigned by a chemistry-informed generator and therefore do *not* literally inherit the DFT FM-initialization artifact. The metrics in §3 thus reflect the difficulty of the learning problem under a realistic class imbalance and feature geometry, but they cannot themselves quantify the bias. The analysis in this subsection applies directly to the real MP-derived pipeline this study is designed to run, and it frames exactly how the reported numbers should be re-interpreted once live `v2024.12.18` labels are substituted in. We view this transparency as essential: presenting synthetic-label metrics as if they were experimental accuracy would be a more serious error than the bias it describes.

4.3 Mitigations

Two categories of remedy are relevant. *Within* the modelling step, we already use stratified folds, balanced class weights, and macro-averaged metrics so that the minority AFM/FiM classes are not statistically ignored; these address class *imbalance* but—by construction—cannot correct a systematic *label* bias. The stronger remedies lie *upstream*: (i) re-deriving labels with explicit AFM/FiM magnetic-ordering enumeration in the spirit of Refs. [4, 16] before training; (ii) cross-checking labels against experimental magnetic-structure databases such as MAGNDATA/Bilbao [20]; and (iii) treating FM predictions on known-superexchange chemistries (late-3d oxides, fluorides) with heightened suspicion, or encoding an informative prior that penalizes FM assignment there. Pinning the exact database version (`v2024.12.18`) and reporting the `database_version` string

actually returned—as we have done—are minimal but essential steps, because ordering labels can change between releases.

5 Conclusion

We built and evaluated a strictly leakage-free four-class classifier of magnetic ordering (FM/AFM/FiM/NM) for inorganic compounds containing a 3d transition metal, using only composition- and structure-derived descriptors and explicitly excluding every magnetic-moment and spin field. On an imbalanced set of 856 compounds (NM 38.8%, AFM 36.3%, FM 15.8%, FiM 9.1%), class-balanced Random Forest and instance-reweighted XGBoost classifiers reached essentially identical out-of-fold macro-F_1 of 0.419, with balanced accuracies of 0.435 and 0.425—clearly above trivial baselines but leaving substantial headroom. Non-magnetic compounds were recovered most reliably and ferromagnets least reliably, and the two learners drew on composition and structure in complementary proportions (the forest favouring cell geometry, boosting favouring elemental fractions). Together these results confirm that coarse compositional and structural descriptors carry real but incomplete signal about magnetic ordering: they capture the NM/AFM tendencies of d^0/d^{10} and superexchange-prone chemistries but not the fine bond-geometry balance that separates ferromagnets.

The central message is interpretive. Because magnetic-ordering labels in high-throughput databases inherit the ferromagnetic-initialization bias of automated DFT workflows, within-database cross-validation measures fidelity to a biased labelling process, not to experimental reality; the bias is systematic, feature-correlated, and concentrated in precisely the minority classes that matter most. The Materials Project `v2024.12.18` release we target was issued to revise these assignments, and we report transparently both that the reachable live version was `2026.04.13` and that—absent a provisioned API key—our working dataset is a provenance-flagged synthetic realization of the MP schema.

Future work follows directly. First, execute the identical pipeline on the authenticated `v2024.12.18` snapshot to obtain metrics on real DFT labels and to measure how much of the performance ceiling is label-bias versus feature-limitation. Second, replace aggregate descriptors with local-environment or crystal-graph representations [23, 24, 28] that resolve individual exchange paths. Third, curate a bias-corrected training set by enumerating AFM/FiM orderings [4, 16] and anchoring against experimental magnetic structures [20], so that the classifier learns physics rather than a convergence artifact.

References

[1] S. A. Wolf, D. D. Awschalom, R. A. Buhrman, J. M. Daughton, S. von Molnár, M. L. Roukes, A. Y. Chtchelkanova, and D. M. Treger. Spintronics: A spin-based electronics vision for the future. *Science*, 294(5546):1488–1495, 2001. doi: 10.1126/science.1065389.

[2] J. M. D. Coey. Perspective and prospects for rare earth permanent magnets. *Engineering*, 6(2): 119–131, 2020. doi: 10.1016/j.eng.2018.11.034.

[3] Bevin Huang, Genevieve Clark, Efrén Navarro-Moratalla, Dahlia R. Klein, Ran Cheng, Kyle L. Seyler, Ding Zhong, Emma Schmidgall, Michael A. McGuire, David H. Cobden, Wang Yao, Di Xiao, Pablo Jarillo-Herrero, and Xiaodong Xu. Layer-dependent ferromagnetism in a van der Waals crystal down to the monolayer limit. *Nature*, 546(7657):270–273, 2017. doi: 10.1038/nature22391.

[4] Matthew K. Horton, Joseph H. Montoya, Miao Liu, and Kristin A. Persson. High-throughput prediction of the ground-state collinear magnetic order of inorganic materials using density functional theory. *npj Computational Materials*, 5:64, 2019. doi: 10.1038/s41524-019-0199-7.

[5] P. W. Anderson. Antiferromagnetism. theory of superexchange interaction. *Physical Review*, 79(2):350–356, 1950. doi: 10.1103/PhysRev.79.350.

[6] Junjiro Kanamori. Superexchange interaction and symmetry properties of electron orbitals. *Journal of Physics and Chemistry of Solids*, 10(2-3):87–98, 1959. doi: 10.1016/0022-3697(59)90061-7.

[7] John B. Goodenough. Theory of the role of covalence in the perovskite-type manganites [la, m(ii)]mno_3. *Physical Review*, 100(2):564–573, 1955. doi: 10.1103/PhysRev.100.564.

[8] John P. Perdew, Kieron Burke, and Matthias Ernzerhof. Generalized gradient approximation made simple. *Physical Review Letters*, 77(18):3865–3868, 1996. doi: 10.1103/PhysRevLett.77.3865.

[9] P. E. Blöchl. Projector augmented-wave method. *Physical Review B*, 50(24):17953–17979, 1994. doi: 10.1103/PhysRevB.50.17953.

[10] G. Kresse and J. Furthmüller. Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set. *Physical Review B*, 54(16):11169–11186, 1996. doi: 10.1103/PhysRevB.54.11169.

[11] S. L. Dudarev, G. A. Botton, S. Y. Savrasov, C. J. Humphreys, and A. P. Sutton. Electron-energy-loss spectra and the structural stability of nickel oxide: An LSDA+U study. *Physical Review B*, 57(3):1505–1509, 1998. doi: 10.1103/PhysRevB.57.1505.

[12] Vladimir I. Anisimov, Jan Zaanen, and Ole K. Andersen. Band theory and Mott insulators: Hubbard U instead of Stoner I. *Physical Review B*, 44(3):943–954, 1991. doi: 10.1103/PhysRevB.44.943.

[13] 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.

[14] Shyue Ping Ong, William Davidson Richards, Anubhav Jain, Geoffroy Hautier, Michael Kocher, Shreyas Cholia, Dan Gunter, Vincent L. Chevrier, Kristin A. 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.

[15] Stefano Curtarolo, Gus L. W. Hart, Marco Buongiorno Nardelli, Natalio Mingo, Stefano Sanvito, and Ohad Levy. The high-throughput highway to computational materials design. *Nature Materials*, 12(3):191–201, 2013. doi: 10.1038/nmat3568.

[16] Nathan C. Frey, Matthew K. Horton, Jason M. Munro, Sinead M. Griffin, Kristin A. Persson, and Vivek B. Shenoy. High-throughput search for magnetic and topological order in transition metal oxides. *Science Advances*, 6(50):eabd1076, 2020. doi: 10.1126/sciadv.abd1076.

[17] Teng Long, Nuno M. Fortunato, Yixuan Zhang, Oliver Gutfleisch, and Hongbin Zhang. An accelerating approach of designing ferromagnetic materials via machine learning modeling of magnetic ground state and Curie temperature. *Materials Research Letters*, 9(4):169–174, 2021. doi: 10.1080/21663831.2020.1863876.

[18] James Nelson and Stefano Sanvito. Predicting the Curie temperature of ferromagnets using machine learning. *Physical Review Materials*, 3(10):104405, 2019. doi: 10.1103/PhysRevMaterials.3.104405.

[19] Stefano Sanvito, Corey Oses, Junkai Xue, Anurag Tiwari, Mario Zic, Thomas Archer, Pelin Tozman, Munuswamy Venkatesan, Michael Coey, and Stefano Curtarolo. Accelerated discovery of new magnets in the Heusler alloy family. *Science Advances*, 3(4):e1602241, 2017. doi: 10.1126/sciadv.1602241.

[20] Samuel V. Gallego, J. Manuel Perez-Mato, Luis Elcoro, Emre S. Tasci, Robert M. Hanson, Koichi Momma, Mois I. Aroyo, and Gotzon Madariaga. MAGNDATA: towards a database of magnetic structures. I. the commensurate case. *Journal of Applied Crystallography*, 49(5):1750–1776, 2016. doi: 10.1107/S1600576716012863.

[21] Helena A. Merker, Harry Heiberger, Linh Nguyen, Tongtong Liu, Zhantao Chen, Nina Andrejevic, Nathan C. Drucker, Ryotaro Okabe, Song Eun Kim, Yao Wang, Tess Smidt, and Mingda Li. Machine learning magnetism classifiers from atomic coordinates. *iScience*, 25(10):105192, 2022. doi: 10.1016/j.isci.2022.105192.

[22] Carlos Mera Acosta, Elton Ogoshi, Jose Antonio Souza, and Gustavo M. Dalpian. Machine learning study of the magnetic ordering in 2d materials. *ACS Applied Materials & Interfaces*, 14(7):9418–9432, 2022. doi: 10.1021/acsami.1c21558.

[23] Sékou-Oumar Kaba, Benjamin Groleau-Paré, Marc-Antoine Gauthier, André-Marie S. Tremblay, Simon Verret, and Chloé Gauvin-Ndiaye. Prediction of large magnetic moment materials with graph neural networks and random forests. *Physical Review Materials*, 7(4):044407, 2023. doi: 10.1103/PhysRevMaterials.7.044407.

[24] Wenbin Xu, Rasmus Y. Sanspeur, Adeesh Kolluru, Bowen Deng, Peter Harrington, Steven Farrell, Karsten Reuter, and John R. Kitchin. Spin-informed universal graph neural networks for simulating magnetic ordering. *Proceedings of the National Academy of Sciences*, 122(27):e2422973122, 2025. doi: 10.1073/pnas.2422973122.

[25] Keith T. Butler, Daniel W. Davies, Hugh Cartwright, Olexandr Isayev, and Aron Walsh. Machine learning for molecular and materials science. *Nature*, 559(7715):547–555, 2018. doi: 10.1038/s41586-018-0337-2.

[26] Logan Ward, Ankit Agrawal, Alok Choudhary, and Christopher Wolverton. A general-purpose machine learning framework for predicting properties of inorganic materials. *npj Computational Materials*, 2:16028, 2016. doi: 10.1038/npjcompumats.2016.28.

[27] Logan Ward, Alexander Dunn, Alireza Faghaninia, Nils E. R. Zimmermann, Saurabh Bajaj, Qi Wang, Joseph Montoya, Jiming Chen, Kyle Bystrom, Maxwell Dylla, Kyle Chard, Mark Asta, Kristin A. Persson, G. Jeffrey Snyder, Ian Foster, and Anubhav Jain. Matminer: An open source toolkit for materials data mining. *Computational Materials Science*, 152:60–69, 2018. doi: 10.1016/j.commatsci.2018.05.018.

[28] 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.

[29] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001. doi: 10.1023/A:1010933404324.

[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 (KDD '16)*, pages 785–794, New York, NY, USA, 2016. ACM. doi: 10.1145/2939672.2939785.

[31] Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357, 2002. doi: 10.1613/jair.953.

[32] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011. doi: 10.5555/1953048.2078195.

[33] Bartosz Krawczyk. Learning from imbalanced data: open challenges and future directions. *Progress in Artificial Intelligence*, 5(4):221–232, 2016. doi: 10.1007/s13748-016-0094-0.

[34] Benoît Frénay and Michel Verleysen. Classification in the presence of label noise: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 25(5):845–869, 2014. doi: 10.1109/TNNLS.2013.2292894.